Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/SurveyMethodology).

Survey Methodology

A Journal Published by Statistics Canada

Volume 40, Number 2, December 2014

Contents

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Survey Methodology, December 2014
Vol. 40, No. 2, p. 135
Statistics Canada, Catalogue No. 12-001-X

Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2016 Waksberg Award.

This issue of *Survey Methodology* opens with the fourteenth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Steve Heeringa (Chair), Cynthia Clark, Louis-Paul Rivest and J.N.K. Rao for having selected Constance Citro as the author of this year's Waksberg Award paper.

2014 Waksberg Invited Paper

Author: Constance F. Citro

Constance F. Citro is director of the Committee on National Statistics (CNSTAT), a position she has held since May 2004. She previously served as acting chief of staff (December 2003-April 2004) and as senior study director (1986-2003). She began her career with CNSTAT in 1984 as study director for the panel that produced *The Bicentennial Census: New Directions for Methodology* in 1990. Dr. Citro received her B.A. in political science from the University of Rochester, and her M.A. and Ph.D. in political science from Yale University. Prior to joining CNSTAT, she held positions as vice president of Mathematica Policy Research, Inc., and Data Use and Access Laboratories, Inc. She was an American Statistical Association (ASA)/National Science Foundation (NSF)/Census research fellow in 1985-1986, and is a fellow of the ASA and an elected member of the International Statistical Institute. For CNSTAT, she directed evaluations of the 2000 census, the Survey of Income and Program Participation, microsimulation models for social welfare programs, and the NSF science and engineering personnel data system, in addition to studies on institutional review boards and social science research, estimates of poverty for small geographic areas, data and methods for retirement income modeling, and a new approach for measuring poverty. She coedited the 2nd-5th editions of *Principles and Practices for a Federal Statistical Agency*, and contributed to studies on measuring racial discrimination, expanding access to research data, the usability of estimates from the American Community Survey, the National Children's Study research plan, and the Census Bureau's 2010 census program of experiments and evaluations.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 137-161
Statistics Canada, Catalogue No. 12-001-X

From multiple modes for surveys to multiple data sources for estimates

Constance F. Citro[1]

Abstract

Users, funders and providers of official statistics want estimates that are "wider, deeper, quicker, better, cheaper" (channeling Tim Holt, former head of the UK Office for National Statistics), to which I would add "more relevant" and "less burdensome". Since World War II, we have relied heavily on the probability sample survey as the best we could do - and that best being very good - to meet these goals for estimates of household income and unemployment, self-reported health status, time use, crime victimization, business activity, commodity flows, consumer and business expenditures, et al. Faced with secularly declining unit and item response rates and evidence of reporting error, we have responded in many ways, including the use of multiple survey modes, more sophisticated weighting and imputation methods, adaptive design, cognitive testing of survey items, and other means to maintain data quality. For statistics on the business sector, in order to reduce burden and costs, we long ago moved away from relying solely on surveys to produce needed estimates, but, to date, we have not done that for household surveys, at least not in the United States. I argue that we can and must move from a paradigm of producing the best estimates possible from a survey to that of producing the best possible estimates to meet user needs from multiple data sources. Such sources include administrative records and, increasingly, transaction and Internet-based data. I provide two examples - household income and plumbing facilities - to illustrate my thesis. I suggest ways to inculcate a culture of official statistics that focuses on the end result of relevant, timely, accurate and cost-effective statistics and treats surveys, along with other data sources, as means to that end.

Key Words: Surveys; Administrative records; Total error; Big data; Income; Housing.

1 Introduction

Tim Holt, former head of the United Kingdom Office for National Statistics and former president of the Royal Statistical Society, once ticked off five formidable challenges for official statistics - namely, to be "wider, deeper, quicker, better, cheaper" (Holt 2007) - to which I would add "less burdensome" and "more relevant". In my view, to respond adequately to one or more, let alone all seven, of these challenges, official statistical offices need to move from the probability sample survey paradigm of the past 75 years to a mixed data source paradigm for the future. Some offices have made that move for most of their statistical programs (see, e.g., Nelson and West (2014) about the extensive use of register-based statistics in Denmark), and almost all offices have made that move for some of their programs, but there are programs not very far along this path. In the case of U.S. household statistical programs, there is a ways to go.

Such a move should not simply elevate another data source as the be all and end all of official statistics in place of the probability sample survey. The 2011 German Republic census - the first census taken in that country since 1983 - provides a useful reminder of the dangers in such an approach. The census results indicated that the administrative records on which Germany based official population statistics for a period of several decades overestimated the population because of failing to adequately record foreign-

1. Constance F. Citro, Director, Committee on National Statistics, U.S. National Academy of Sciences/National Research Council. E-mail: ccitro@nas.edu.

born emigrants (see http://www.nytimes.com/2013/06/01/world/europe/census-shows-new-drop-in-germanys-population.html?_r=0 [November 2014]).

My thesis is that official statistical programs must start with user needs for information for policy development, program evaluation, and understanding societal trends, and work backwards from concepts to appropriate data sources. Such sources may very likely include probability surveys but may also include one or more alternative kinds of data. My thesis is a truism in one sense, but people whose lives are devoted to perfecting a particular tool for data collection may too often see everything as in need of that tool, rather than considering the most cost-effective way to obtain statistics that policy makers, researchers, and other data users want.

I little doubt that Joe Waksberg, whom I was honored to know through his service on a Committee on National Statistics (CNSTAT) Panel on Decennial Census Methodology in the mid-1980s, would approve of my topic. Joe was not only an uncommonly gracious and charming human being, but also a problem-solver and innovator of the first order. Joe stressed "the importance of examining not only what you are asked, but also what you think the analyst has in mind" (Morganstein and Marker 2000). Joe invariably thought outside the box to identify data sources and models that addressed the underlying information need rather than worked from an a priori concept of what tools were appropriate.

In the following text, I briefly review the rise and benefits of probability sampling for official statistics in the United States in Section 2 and the growing threats to the relevance, accuracy, timeliness, cost-effectiveness and public acceptability of survey-based estimates in Section 3. In Section 4 and Section 5, I consider the strengths and weaknesses of administrative records and other non-probability-survey data sources that may be valuable, singly and in combination, for official statistics. In Section 6, I offer examples of ripe opportunities in the United States to transform ongoing household survey programs to use multiple data sources to provide information of greater value. I conclude in Section 7 by enumerating barriers to moving to a multiple data sources paradigm and suggest ways to lower those barriers.

I focus on what I know best - namely, U.S. official statistics and household statistics programs in particular. I hope that readers from other countries, other statistical programs and other agencies will find analogies in their own work. I critique the survey paradigm from a goal of improving official statistics, remaining deeply appreciative of the value of probability surveys, alone and combined with other data sources, and deeply admiring of the important work of statistical agencies in service to the public good (see National Research Council 2013c).

2 The rise of probability sampling in official U.S. statistics

It is not an exaggeration to say that large-scale probability surveys were the 20th-century answer to the need for wider, deeper, quicker, better, cheaper, more relevant and less burdensome official statistics. Such surveys provided information with known precision in contrast to non-probability surveys; and they provided detailed information at greatly reduced cost and increased timeliness compared with censuses. Duncan and Shelton (1978) and Harris-Kojetin (2012) review the rise of probability sampling in U.S. official statistics.

It was not clear at the time when the theory and practice of modern probability sampling was being developed in the 1930s in the United States that probability surveys would gain such widespread

acceptance. The arrival of Jerzy Neyman in the mid-1930s gave a boost to the work of W. Edwards Deming, Calvin Dedrick, Morris Hansen and colleagues at the Census Bureau who were developing the needed theory for sampling of finite populations. Small-scale sample surveys in the 1930s at universities and federal agencies on such topics as consumer purchases, unemployment, urban housing and health provided proofs of concept and practical tips.

Table 2.1
Selected ongoing U.S. statistical agency probability surveys, by year begun

Decade and Year/Type of Survey	Repeated Cross-Sectional Household Survey	Repeated Cross-Sectional Business Establishment Survey	Panel Person Survey
1940	1940 - Current Population Survey (CPS) 1947 - CPS Annual Social and Economic Supplement (CPS/ASEC)	1946 - Monthly Wholesale Trade Survey	
1950	1950 - Consumer Expenditure Survey (CE) 1955 - National Survey of Fishing, Hunting, and Wildlife-Associated Recreation 1957 - National Health Interview Survey (NHIS)	1953 - Advance Monthly Retail Sales Survey 1953 - Business R&D and Innovation Survey (BRDIS) 1959 - Building Permits Survey	
1960	1960 - Decennial Census Long-Form Sample (became American Community Survey in 2005)	1965 - National Hospital Care Survey	1966-1990 - National Longitudinal Survey of Older Men
1970	1972 - National Crime Victimization Survey (NCVS) 1973 - American Housing Survey (AHS); 1973 - National Survey of College Graduates (NSCG) 1979 - Residential Energy Consumption Survey (RECS)	1975 - Farm Costs and Returns Survey and Cropping Practices and Chemical Use Surveys (combined in Agricultural Resource Management Survey in 1996) 1979 - Commercial Buildings Energy Consumption Survey (CBECS)	1972-1986 - National Longitudinal Survey of High School Class of 72 1973-present - Survey of Doctorate Recipients (SDR) 1979-present - National Longitudinal Survey of Youth (NLSY79)
1980	1983 - Survey of Consumer Finances (SCF)	1985 - Manufacturing Energy Consumption Survey (MECS)	1984-present - Survey of Income and Program Participation (SIPP)
1990	1991 - Medicare Current Beneficiary Survey (MCBS)	1996 - Agricultural Resource Management Survey (ARMS)	1997-present - National Longitudinal Survey of Youth (NLSY97)
2000	2005 - American Community Survey (ACS)		2001-2008 - Early Childhood Longitudinal Study (Birth Cohort)

Notes: Current survey name is used; periodicity of interviewing for repeated cross-sectional and panel surveys varies; some repeated cross-sectional surveys have panel component (rotation groups); length of panel surveys (how many years respondents are in sample) varies.
Source: Compiled by author.

The federal government's young statistical Turks still had to surmount hurdles in the bureaucracy up to the White House before they could move sampling into the mainstream of federal statistics. Thus, "old timers" at the Census Bureau were skeptical about the possibility of using survey methods to get information on unemployment and politicians were divided about whether they wanted the estimates (Anderson 1988). In 1937, a major breakthrough occurred when a two percent sample of households on nonbusiness postal routes, designed by Dedrick, Hansen and others, estimated a much higher - and more credible - number of unemployed than a "complete" census of all residential addresses that was conducted on a voluntary basis. Picking up on that effort, from 1940-1942, the Works Progress Administration fielded the sample-based Monthly Report on the Labor Force, the forerunner to the Current Population Survey (CPS). The CPS continues to this day as the source of official monthly estimates of U.S. unemployment conducted by the Census Bureau and published by the Bureau of Labor Statistics (BLS).

Another breakthrough occurred when the Census Bureau, which struggled for decades to respond to demands for added questions on the decennial census without turning the instrument into a nightmare for respondents and interviewers, asked six questions on a five percent sample basis in the 1940 census. The success of sampling led to a decision to administer two-fifths of the questions in the 1950 census to a sample, and subsequent censuses followed suit. Table 2.1 lists selected ongoing U.S. household surveys, business surveys and panel surveys and when they began. The variety of subjects covered and the longevity of these surveys attest to the dominance and value of the sample survey paradigm in U.S. official statistics.

3 Chinks in the armor: Rising threats to the survey paradigm

Probability surveys are indispensable tools for official statistical agencies and others for many kinds of measures - for example, to track such phenomena as public approval of the U.S. president or expressed feelings of well-being. Moreover, probability surveys with a primary purpose to measure constructs, like household income, that could be obtained from other sources, have two major advantages: (1) they can obtain a wide variety of covariates for use in analysis of the primary variable(s) of interest, and (2) they are under the control of the survey designer. Yet threats to the probability survey paradigm are snowballing in ways that bode ill for the future. Manski (2014) goes so far as to accuse statistical agencies of sweeping major problems with their data under the rug and markedly understating the uncertainty in their estimates. He labels survey nonresponse as an example of "permanent uncertainty".

3.1 Characterizing survey quality

A typology of errors and other problems that can compromise the quality of survey estimates is essential for understanding and improving official statistics. A seminal paper in developing data quality frameworks was Brackstone (1999). Most recently, Biemer, Trewin, Bergdahl and Lilli (2014) reviewed the literature on systematic quality frameworks, noting, in particular, the six dimensions proposed by Eurostat (2000): relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability (across time and geography), and coherence (consistent standards). Iwig, Berning, Marck and Prell (2013) reviewed quality frameworks from Eurostat, the Australian Bureau of Statistics, the UK Office for National Statistics, Statistics Canada, and other organizations and developed questions based on six

quality dimensions of their devising - relevance, accessibility, coherence, interpretability, accuracy, and institutional environment - for U.S. statistical agencies to use to assess the utility of administrative records. Daas, Ossen, Tennekes and Nordholt (2012) constructed a framework for evaluating the use of administrative records to produce census data for the Netherlands.

Biemer et al. (2014) went further by using the Eurostat framework (combining comparability and coherence into a single dimension) as the basis for designing, testing and implementing a system of numerical assessments for evaluating and continually improving data product quality at Statistics Sweden. For a full assessment, it would also be necessary to evaluate quality dimensions against cost and respondent burden. Usefully for my purposes, Biemer et al. decomposed the dimension of "accuracy", conceived of as total survey error (or total product error for non-survey-based statistical programs such as national accounts), into sampling error and seven types of nonsampling error: (1) frame error, including undercoverage and overcoverage and missing or erroneous auxiliary variables on the frame; (2) nonresponse error (unit and item); (3) measurement error (overreporting, underreporting, other); (4) data processing error; (5) modeling/estimation error, such as from fitting models for imputation or adjusting data values to conform to benchmarks; (6) revision error (the difference between preliminary and final published estimates); and (7) specification error (the difference between the true, unobservable variable and the observed indicator). For ongoing surveys, I would add *outmoded construct error*, which is related to but different from specification error. For example, the Census Bureau's regular money income concept for official household income and poverty estimates from the CPS Annual Social and Economic Supplement (ASEC) has become progressively outdated due to changing U.S. tax and transfer programs (see, e.g., Czajka and Denmead 2012; National Research Council 1995).

3.2 Four sources of error in U.S. household statistics

3.2.1 Frame deficiencies

Obtaining a comprehensive, accurate frame for surveys can be as difficult as obtaining responses from sample cases drawn from the frame and, in many instances, the difficulties have persisted and even grown over time. Joe Waksberg would resonate to the problem of frame deficiencies: not only did he, with Warren Mitofsky, develop the random digit dialing (RDD) method for generating frames and samples for high-quality residential telephone surveys in the 1970s (see Waksberg 1978; Tourangeau 2004), but he also saw the beginnings of the method's decline in popularity because of such phenomena as cell-phone-only households.

A commonly used frame for U.S. household surveys is the Census Bureau's Master Address File (MAF) developed for the decennial census. The past few censuses have obtained increasingly good net coverage of residential addresses on the MAF, particularly for occupied units (Mule and Konicki 2012). The persistent problem for household surveys is undercoverage of individual members within sampled units. Coverage ratios (i.e., estimates before ratio adjustment to population controls) in the March 2013 CPS, for example, are only 85 percent for the total population, and there are marked differences among men and women, older and younger people, and whites and minorities, with coverage ratios as low as 61 percent for black men and women ages 20-24 (see http://www.census.gov/prod/techdoc/cps/cpsmar13.pdf [November 2014]). No systematic study of the time series of coverage ratios for U.S. household surveys has been conducted, but there is evidence that ratios have been getting worse.

While useful to correct coverage errors for age, gender, race and ethnicity groups, the current household survey ratio adjustments undoubtedly fail to correct for other consequential coverage differences. (The ratio-adjustment controls, in one of the least controversial and most long-standing uses of administrative records in U.S. household surveys, derive from population estimates developed from the previous census updated with administrative records and survey data.) Thus, everything that is known about undercount in the U.S. decennial census indicates that, holding race and ethnicity constant, socioeconomically disadvantaged populations are less well counted than others (see, e.g., National Research Council 2004, App. D). It is unlikely that household surveys perform any better - for example, Czajka, Jacobson and Cody (2004) find that the Survey of Income and Program Participation [SIPP] substantially underrepresents high-income families compared with the Survey of Consumer Finances [SCF], which includes a list sample of high-income households drawn from tax records. Factoring in differential socioeconomic coverage, Shapiro and Kostanich (1988) estimate from simulations that poverty is significantly biased downward for black males in the CPS/ASEC. On the other hand, by comparison with the 2000 census long-form sample, Heckman and LaFontaine (2010) find that survey undercoverage in the 2000 CPS October educational supplement contributes little to underestimates of high school completion rates; other factors are more important.

3.2.2 Unit response in secular decline

A study panel of the (U.S.) National Research Council (2013b) recently completed a comprehensive review of causes and consequences of household survey unit nonresponse, documenting the well-known phenomenon that the public is becoming less available and willing to respond to surveys, even from well-trusted official statistical agencies. In the United States, there was evidence as early as the 1980s that response rates had been declining from almost the beginning of the widespread use of probability sample surveys (see, e.g., Steeh 1981; Bradburn 1992). De Leeuw and De Heer (2002) estimated a secular rate of decline in survey cooperation of 3 percentage points per year from examining ongoing surveys in 16 Western countries from the mid-1980s through the late 1990s. The cooperation rate measures the response of eligible sample cases actually contacted; response rates (there are several accepted variations) have broader denominators, including eligible cases that were not reached (National Research Council 2013c, pp. 9-12). National Research Council (2013b: Tables 1-2, 104) provides initial or screener response rates to a range of U.S. official surveys for 1990/91 (after response rates had already fallen significantly for many surveys) and 2007/2009, which make clear that the problem is not going away.

It was long assumed that lower response rates even with nonresponse weighting adjustments inevitably entailed bias in survey estimates. Recent research (see, e.g., Groves and Peytcheva 2008) finds that the relationship between nonresponse and bias is complex and extraordinary efforts to increase response can inadvertently increase bias by obtaining greater response from only some groups and not others (see, e.g., Fricker and Tourangeau 2010). It would be foolhardy, however, for official statistical agencies to assume that increasing nonresponse has no or little effect on the accuracy of estimates, particularly when unit nonresponse is coupled with item nonresponse. For example, nonrespondents to health surveys are estimated to have poorer health on average than respondents and nonrespondents to volunteering surveys are estimated to be less likely to volunteer than respondents (National Research Council 2013b, pp. 44-45). Moreover, there has been little research on the effects of nonresponse on bivariate or multivariate

associations or on variance, except for the obvious - and not unimportant - effect that unit nonresponse reduces effective sample size.

3.2.3 Item response often low and declining

Neither sample surveys nor censuses can be expected to obtain answers from unit respondents to every item on a questionnaire. U.S. census practice has long been to edit some items for consistency, but until mid-twentieth century, there were no adjustments for item nonresponse - tables included rows labeled "no response" or similar wording. The first use of imputation occurred in 1940 when Deming developed a "cold deck" procedure to impute age by randomly selecting a value for age from an appropriate deck of cards selected according to what other information was known about the person for whom age was missing. Beginning in 1960, with the advent of high-speed computers, "hot deck" imputation methods were used to impute missing values for many census items (Citro 2012). The hot deck procedure uses the latest value for the previously processed person or household stored in a matrix and, consequently, does not have to assume that data are missing completely at random (MCAR), although it does have to assume that data are missing at random (MAR) within the categories defined by variables in the hot deck matrix. Model-based methods of imputation have been developed that do not require such strong assumptions as MAR or MCAR (see National Research Council 2010b), but they are not widely used in U.S. household surveys. Two exceptions are in the Survey of Consumer Finances (SCF) (Kennickell 2011) and the Consumer Expenditure (CE) Interview Survey (Passero 2009).

Whatever the method, imputation has the advantage of creating a full data record for every respondent, which facilitates multivariate analysis and forestalls the likelihood that researchers will use different methods for treating missing data that give different results. Yet imputation may introduce bias into estimates, and the significance of any bias will likely be magnified by the extent of missing data. So it is troubling that nonresponse has been increasing for important items on household surveys, such as income, assets, taxes and consumer expenditures, which require respondents to supply dollar amounts - for example, Czajka (2009:Table A-8) compares item imputation rates for total income and several sources of income for the CPS/ASEC and SIPP for 1993, 1997 and 2002 - a full one-third of income is currently imputed on the CPS/ASEC, up from about one-quarter in 1993 - and SIPP is not much better. Clearly, with such high imputation rates, careful evaluation of the effects of imputation procedures is imperative to carry out. Hoyakem, Bollinger and Ziliak (2014), for example, estimate that the hot deck imputation procedure for earnings in the CPS/ASEC has consistently underestimated poverty by an average of one percentage point, based on evaluating missing earnings in both the CPS/ASEC and Social Security earnings records.

3.2.4 Measurement error problematic and not well studied

Even with complete reporting, or, more commonly, adjustments for unit and item nonresponse, there will still be error in survey estimates from inaccurate reporting by respondents due to guessing at the answer, deliberately failing to provide a correct answer, or not understanding the intent of the question. While acknowledged by statistical agencies, the extent of measurement error is typically less well studied than is sampling error or the extent of missing data. Many measurement error studies compare aggregate estimates from a survey with similar estimates from another survey or an appropriate set of administrative

records, adjusted as far as possible to be comparable. It is not possible to sort out from these studies the part played by measurement error in comparison with other factors, but the results indicate the magnitude of problems. Some studies are able to match individual records and thereby examine components of measurement error.

Significant measurement error is known to affect key socioeconomic estimates produced from U.S. household surveys. Thus, a legion of studies have documented net underestimation of U.S. household income in survey after survey and, even more troubling, a decline in completeness of reporting, even after imputation and weighting. Fixler and Johnson (2012, Table 2), for example, estimated that between 1999 and 2010, mean and median estimates from the CPS/ASEC fell progressively below the National Income and Product Account (NIPA) estimates due to such factors as: (1) underrepresentation of very high-income households in the CPS/ASEC sample; (2) nonreporting and underreporting by those high-income households that are included; and (3) nonreporting and underreporting by middle and lower income households. Studies of individual income sources find even worse error. Meyer and Goerge (2011), for example, by matching Supplemental Nutrition Assistance Program (SNAP) records in two states find that almost 35 percent and 50 percent, respectively, of true recipients do not report receiving benefits in the American Community Survey (ACS) and the CPS/ASEC. Similarly, Meyer, Mok and Sullivan (2009) document large and often increasing discrepancies between survey estimates and appropriately adjusted administrative records estimates of income recipients and total amounts for many sources.

Wealth is notoriously difficult to measure in household surveys, and many do not attempt to do so. Czajka (2009, pp. 143-145) summarizes research on the quality of SIPP estimates of wealth by comparison with the SCF and the Panel Study of Income Dynamics (PSID). Greatly simplifying the findings, SIPP historically has been fairly effective in measuring liabilities, such as mortgage debt, and the value of such assets as owned homes, vehicles, and savings bonds. SIPP has done poorly in measuring the value of assets held mostly by higher income households, such as stocks, mutual funds, and IRA and KEOGH accounts, whereas the PSID has done somewhat better. On net, SIPP significantly underestimates net worth.

A National Research Council (2013a) study of the BLS CE Interview and Diary Surveys found differential quality of reporting of various expenditure types compared with appropriately adjusted personal consumption expenditure (PCE) estimates from the NIPA. Bee, Meyer and Sullivan (2012, Table 2) also find declines in reporting for some expenditures - for example, gasoline reporting in the CE household estimate declined from over 100 percent of the comparable PCE estimate in 1986 to just under 80 percent in 2010, while reporting on furniture and furnishings declined from 77 percent to 44 percent over a comparable period.

4 What can be done?

Survey researchers have not been idle in the face of multiple and increasing threats to the survey paradigm. For at least the last 15 years, they have actively worked on ways to reduce or compensate for coverage error, unit and item nonresponse, measurement error, and, more recently, burden on respondents. Strategies have included: (1) spending more on case completion (although budget constraints limit the viability of this strategy); (2) using paradata and auxiliary information for more effective unit nonresponse

bias identification and adjustment; (3) employing more sophisticated missing data adjustments that do not assume MAR; (4) using adaptive design methods to optimize the cost and quality of response; (5) using multiple frames to reduce coverage error (e.g. cell-phone and land-line frames for telephone surveys); (6) using multiple modes to facilitate more cost-effective response as in the ACS, which recently added an Internet response option to its mail, CATI and CAPI options; (7) reducing burden by optimizing follow-up calls and visits; and (8) describing the needs for the survey data. In the United States, data users are often recruited to make the case to Congress and other stakeholders. For example, the Association of Public Data Users, the Council of Professional Associations on Federal Statistics and the Population Association of America frequently mobilize data users on behalf of statistical agency programs.

My thesis is that these steps, while laudable and necessary, are not sufficient to restore the probability survey-based paradigm for official statistics on households or other types of respondents. I propose, instead, that statistical agencies consistently begin by determining policymakers' and public needs and work backwards to identify appropriate data sources to serve those needs in the most cost-effective and least burdensome manner possible. This multiple sources paradigm should apply to all statistical programs, whether traditionally based on a survey, administrative records, or another source.

Some important statistical programs, such as the NIPAs and the Consumer Price Index (see Horrigan 2013) in the United States and other countries, have for decades used multiple data sources. One reason is that these programs are built around a widely accepted conceptual framework that determines required elements to constitute an acceptable set of estimates. It is not acceptable to omit one or more components of income from the NIPAs simply because data are not available from a single source. Moreover, because key NIPA estimates are periodically revised to add data, improve methodology and refine concepts, there is a built-in positive bias to search for new and improved data sources to fill gaps and improve accuracy. The U.S. economic censuses also use multiple sources, specifically, income tax records for sole proprietors and very small employers together with surveys for larger companies. U.S. household statistics programs, in contrast, have most closely adhered to the probability sample survey paradigm. Moreover, because long intervals typically occur between revisions to household survey concepts and design, the surveys too often fall behind in their ability to serve policymakers and the public, when the use of additional data sources could make possible significant improvements.

5 Which data sources to bolster surveys?

For decades after the introduction of probability sampling in official statistics, the only alternative source was administrative records - from various levels of government, depending on a country's governmental structure (federal, state and local in the United States), and from nongovernmental entities (e. g., employer payroll records or hospital admission records). And a number of national statistical agencies around the world began to incorporate administrative records into their programs - from using them in an ancillary way to moving census and survey programs lock, stock and barrel to an administrative records-based paradigm.

Technological innovations in the 1970s and 1980s led to some additional data sources - such as records of expenditures at checkouts (made possible by the development of bar codes and scanners) and aerial and satellite images for categorizing land use - becoming at least potentially available for official statistics. But the landscape of data sources was still relatively contained. Beginning in the 1990s, the advent of the

Internet and high-speed distributed computing technology unleashed a mind-boggling array of new data sources, such as data from traffic camera feeds, tracking of cell phone locations, search terms used on the Web and postings on social media sites. The challenge for statistical agencies is to classify and evaluate all of these data sources in ways that help agencies determine their usefulness.

5.1 Is "Big Data" a useful concept?

Many new types of data that have become available in the past 15 or so years are often very large in size, leading to the use of the term "big data". I argue that this buzz phrase does little, if anything, to assist statistical agencies to determine appropriate combinations of data for their programs. In computer science, "big data is high volume, high velocity and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization" (Laney 2001). These properties are not inherent in any particular type of data or in any particular platform, such as the Internet. Instead, what qualifies as "big data" is a changing target, as advances are made in high-speed computing and data analysis techniques. In today's computing environment, census, survey, and administrative records data rarely qualify as "big", although they may have done so in an earlier era. People today tend to classify as "big" the data streams from cameras, sensors, and largely free-form interactions with the Internet, such as social media postings. In the future, many of these kinds of data may no longer fit under this rubric. In regard to the Internet, moreover, it not only generates a great deal of today's "big data", but also provides ordinary-size data in a more accessible way - for example, access to public opinion polls or to local property records.

I would argue that statistical agencies will most often want to be and should be "close followers" rather than leaders in using big data. It seems to me most appropriate for academia and the private sector to be out front in tackling the uses of data that are so voluminous and of such high velocity and variety that they require big leaps forward to develop new forms of processing and analysis. Statistical agencies should be alert to developments in the field of big data that promise benefits for their programs down the road and may be well advised to support research in this area to help ensure that applications that are relevant to their programs emerge. Principally, however, I believe that statistical agency resources are best used primarily for working with data sources that offer more immediately useful benefits.

Groves (2011) has attempted to move toward a more relevant classification for statistical agencies than that between "big data" and all other data, by distinguishing between what he terms "designed data" that are "produced to discover the unmeasured" and "organic data" that are "produced auxiliary to processes, to record the process". Keller, Koonin and Shipp (2012) list examples of data sources under Groves' two headings. Their list of designed data includes: administrative data (e.g., tax records); federal surveys; censuses of population; and "other data collected to answer specific policy questions". Their list of organic data includes: location data (cell phone "externals", E-ZPass transponders, surveillance cameras); political preferences (voter registration records, voting in primaries, political party contributions); commercial information (credit card transactions, property sales, online searches, radio-frequency identification); health information (electronic medical records, hospital admittances, devices to monitor vital signs, pharmacy sales); and other organic data (optical, infrared and spectral imagery, meteorological measurements, seismic and acoustic measurements, biological and chemical ionizing radiation). Not mentioned under either category are such data as Facebook or Twitter postings, although they might fall under the broad rubric of "online searches".

Whether the two-part classification in Keller et al. (2012) is all that more useful than "big data" for statistical agency purposes is a question. For example, classifying voter registration records or electronic health records as organic data and not as designed administrative data seems to miss ways in which they differ from such sources as online searches and ways in which they are similar to federal and state government administrative records. Moreover, even organic data are "designed", if only minimally, in the sense that the provider has specified some parameters, such as 140 characters for a Twitter post or a particular angle of vision for a traffic camera. Nonetheless, the designed versus organic distinction does point to a useful dimension, which is the degree to which statistical agencies have ready access to, control changes to, and are able readily to understand the properties of a data source.

5.2 Dimensions of data sources: Illustrations for four major categories

Coming up with satisfactory nomenclature and evaluation criteria that can help statistical agencies assess the potential usefulness of alternative data sources for their programs, with the goal of becoming as familiar with the error properties of alternative sources as they are with total error for surveys, is not going to happen without considerable effort by statistical agencies around the world (Iwig et al. 2013 and Daas et al. 2012 are examples of such efforts). I do not pretend that I can come close to that goal in this paper. My goal is more modest - namely, to provide some illustrations so that those who are wedded to a probability survey paradigm (or an administrative records paradigm) can see that the task of understanding alternative data sources is both feasible and desirable. I provide illustrations for four data sources ranging from traditional to cutting-edge:

(1) Surveys and censuses, or a collection of data obtained from responses of individuals, who are queried on one or more topics as designed by the data collector (statistical agency, other government agency, and academic or commercial survey organization) according to principles of survey research with the goal of producing generalizable information for a defined population.

(2) Administrative records or a collection of data obtained from forms designed by an administrative body according to law, regulation, or policy for operating a program, such as paying benefits to eligible recipients or meeting payroll. Administrative records are usually ongoing and may be operated by government agencies, or non-governmental organizations.

(3) Commercial transaction records, or a collection of data obtained from electronic capture of purchases (e.g., groceries, real estate) initiated by a buyer but in a form determined by a seller (e.g., bar-coded product information and prices recorded by check-out scanners or records of product and price information for Web sales, such as through Amazon).

(4) Interactions of individuals with the WorldWide Web by using commercially provided tools, such as a Web browser or social media site. This category covers a wide and ever-changing array of potential data sources for which there are no straightforward classifications. One defining characteristic is that individuals providing information, such as a Twitter post, act as autonomous agents: they are not asked to respond to a questionnaire or required to supply administrative information but, instead, are choosing to initiate an interaction.

I first rank each source on the following two dimensions, which relate to the framework in Biemer et al. (2014). The rank I assign assumes there have been as yet no proactive steps by a statistical agency to

boost the ranking (e.g., by embedding staff in an administrative agency to become deeply familiar with the agency's records). The two dimensions are:

(1) Degree of accessibility to and control by national statistical agency: high (statistical agency designs the data source and controls changes to it); medium (statistical agency has authority to use the data source and influence on changes to it); low (statistical agency must arrange to obtain the data source on the terms of the provider and has little or no influence on changes to it). Gradations can be added to each of these categories depending, for example, on how strong an agency's authority is to acquire a set of administrative records.

(2) Degree to which components of error can be identified and measured: high, as in designed surveys and censuses; medium, as in public and private sector administrative records; and low, as in streams of data from autonomous choices of individuals.

I further identify aspects of data quality for each source, following Biemer et al. (2014). I also indicate variations for most of the dimensions depending on the provider, such as national statistical agency, other unit of national government, other level of government, academic institution, or commercial entity. Table 5.1 provides all of this information as best I can.

An ideal source for statistical agency use, other things equal, is one that is provided, designed, and controlled by the agency, and for which errors can be identified and measured and are generally under control, such as a high-quality probability survey mounted by the agency. At the other extreme is a data source that is controlled by one or more private companies (e.g., scanner data) or, perhaps, by hundreds or thousands of local governments (e.g., traffic cameras), where the data result from autonomous choices or uncontrolled movements, and where it is difficult to conceptualize, much less measure, errors in the data source. Yet when considering a statistical agency's responsibility to provide relevant, timely, accurate statistics for policymakers and the public for which costs and respondent burden are minimized, there may well be non-survey data sources that warrant the effort to make them usable for statistical purposes. I argue that the threats to the survey paradigm reviewed above make it imperative to consider alternative data sources because surveys are no longer always and everywhere demonstrably the superior choice to other sources - they are not always "high" on the dimensions in Table 5.1.

I further argue that government administrative records, which, as Table 5.1 indicates, more often have desirable properties for official statistics compared with other non-survey data sources, should be a prime candidate for statistical agencies to incorporate as extensively as possible into their survey programs if they have not already done so. Administrative records are generated according to rules - rules about the eligible population, who must file what information, what action by the pertinent administrative body is taken on the basis of the information (e.g., tax refund, benefit payment), and so on. This fact should make it possible, with requisite effort, for a statistical agency to become as familiar with administrative records error structures as they are with total survey error. Couper (2013) provides a useful discussion somewhat like mine. He pokes holes in the ability of organic data sources to be as useful as they are often touted to be, much less to be suitable to replace probability surveys, but he warns survey researchers that they ignore organic data sources at their peril. Ironically, his conclusion to make some use of organic sources is strengthened because of his error in classifying administrative records as organic data. They are properly classified as designed data, even though not designed by a statistical agency.

Table 5.1
Ranking (HIGH, MEDIUM, LOW, VERY LOW, or VARIES) of four data sources on dimensions for use in official statistics

Dimension/ Data Source	Census/Probability Survey (e.g., CPS/ASEC, ACS, NHIS - see Table 2.1)	Administrative Records (e.g., income taxes, Social Security, unemployment, payroll)	Commercial Transaction Records (e.g., scanner data, credit card data)	Individual Interactions with the Internet (e.g., Twitter postings; Google search term volumes)
Degree of Control by/ Accessibility to Statistical Agency	HIGH (survey conducted for statistical agency); MEDIUM to LOW (survey conducted for private organization)	HIGH to MEDIUM (national agency records); MEDIUM to LOW (state or local records); MEDIUM to LOW (commercial records)	MEDIUM to LOW	VERY LOW
Degree of Ability of Statistical Agency to Identify/Assess Properties/ Errors	HIGH (survey conducted for statistical agency); VARIES (survey conducted for private org., depends on documentation and transparency)	HIGH to MEDIUM (national agency records); MEDIUM to LOW (state or local records); MEDIUM to LOW (commercial records)	MEDIUM (to the extent that records follow accepted standards, e.g., for bar coding and pricing information)	VERY LOW
Data Quality Attributes (Biemer et al. 2014)				
Relevance for Policy and Public - Concepts and Measures	HIGH for survey conducted for statistical agency, assuming well designed and up to date in concepts and measures; VARIES for surveys for private organizations	VARIES across and within records systems (e.g., records of benefit payment may be highly relevant, while family composition information may use a different concept)	VARIES	VARIES, but VERY LOW at the present state of the art of acquiring, evaluating, and analyzing these kinds of data
Relevance - Useful Covariates	HIGH for most surveys	VARIES, but rarely as high as for most surveys	VARIES, but rarely as high as for most surveys	VARIES, but typically LOW
Frequency of Data Collection	Weekly to every few years (every decade for the U.S. population census); Some private surveys, such as election polls, may run daily	Generally records are updated frequently (e.g., daily) and continually	Generally records are updated frequently (e.g., at moment of transaction or daily) and continually	Interactions are captured instantaneously
Timeliness of Release	VARIES, depending on effort of statistical agency or private organization, but some lag from the reference period for responses is inevitable	VARIES, but some lag from the reference date to when records are acquired by statistical agency likely	VARIES, but likely to be long lags in acquiring proprietary data by statistical agency	VARIES, but likely to be long lags (although MIT Billion Prices Project has worked out very timely access for prices on the Internet; see bpp.mit.edu)
Comparability and Coherence	HIGH across time and geography within survey (except when deliberately changed or if societal change that affects measurement is not taken into account); VARIES among surveys	HIGH within records system (changes to government records generally heralded by legal/ regulation/policy change, changes to commercial records likely opaque); VARIES among records systems	HIGH within records system (changes generally opaque to statistical agency); VARIES among records systems	VERY LOW, in that vendors (e.g.,Twitter) may add/subtract features or drop an entire product; Changes generally opaque to statistical agency; Initiators of interactions may have very different frames of reference

Accuracy (Components of Error)*				
Dimension/ Data Source	**Census/Probability Survey (e.g., CPS/ASEC, ACS, NHIS - see Table 2.1)**	**Administrative Records (e.g., income taxes, Social Security, unemployment, payroll)**	**Commercial Transaction Records (e.g., scanner data, credit card data)**	**Individual Interactions with the Internet (e.g., Twitter postings; Google search term volumes)**
Frame Error	VARIES, can be significant undercoverage and overcoverage	Frame is usually well defined by law, regulation, or policy; Problem for statistical agency use is that frame may not be comprehensive	Frame is ill-defined for statistical agency purposes, in that represents whoever had a purchase scanned by a specified vendor or used a specific credit card for a purchase during a specified time; Poses significant challenge to statistical agency to determine appropriate use	Frame is ill-defined for statistical agency purposes, in that represents whoever, decided to, for example, set up Twitter account or conduct Google search during a specified time; Poses significant challenge to statistical agency to determine appropriate use
Nonresponse (unit and item)	VARIES, can be significant	VARIES (e.g., Social Security records likely to include almost all eligible people, but income tax records likely to reflect evasion, in terms of failure to file a return or concealing some income)	NOT APPLICABLE, in that "respondents" are self selected; Statistical agency challenge is to determine appropriate use that does not need to assume a probability mechanism	NOT APPLICABLE, in that "respondents" are self selected; Statistical agency challenge is to determine appropriate use that does not need to assume a probability mechanism
Measurement Error	VARIES within surveys by item and among surveys for comparable items; Often not well assessed, even for statistical agency surveys	VARIES among record systems and within record systems by item depending on centrality of the item to program operation (e.g., benefit payment item likely more accurate than items obtained from beneficiaries, such as employment status)	NOT APPLICABLE to data source as such, although any characteristics added by the vendor from another source may/may not be valid; Statistical agency challenge is to not introduce measurement error by inappropriate use of the data	NOT APPLICABLE to data source as such, although any characteristics added by the vendor from another source may/may not be valid; Statistical agency challenge is to not introduce measurement error by inappropriate use of the data
Data Processing Error	VARIES (e.g., may be data capture or recoding errors), but is usually under good statistical control, although harder to assess for private organization surveys	VARIES (e.g., may be keying or coding errors), likely to be under better control for key variables (e.g., benefit payments) than for other variables, but hard for statistical agency to assess	VARIES (e.g., may be errors in assigning bar codes or prices), likely to be under good control, but hard for statistical agency to assess	NOT APPLICABLE, in that error is not defined, although there may be occasional problems of the sort that, say, a day's worth of Twitter posts is overwritten and lost
Modeling/ Estimation Error	Bias from such processes as weighting and imputation VARIES; Often intense effort by statistical agency to design well initially but not to revisit to ascertain continued validity of procedures	NOT APPLICABLE (usually), in that records are "raw" data, except perhaps for some recoded variables, but bias may be introduced by statistical agency reprocessing	NOT APPLICABLE (usually), in that records are "raw" data, except perhaps for some recoded or summarized variables, but bias may be introduced by statistical agency reprocessing	NOT APPLICABLE (usually), in that records are "raw" data, but statistical agency reprocessing may introduce significant bias (e.g., by using the word "fired" as always indicating unemployment in analyzing Twitter posts)

Accuracy (Components of Error)* (CON'T)				
Dimension/ Data Source	**Census/Probability Survey (e.g., CPS/ASEC, ACS, NHIS - see Table 2.1)**	**Administrative Records (e.g., income taxes, Social Security, unemployment, payroll)**	**Commercial Transaction Records (e.g., scanner data, credit card data)**	**Individual Interactions with the Internet (e.g., Twitter postings; Google search term volumes)**
Specification Error	VARIES (e.g., self-reported health status may validly indicate respondent's perception but not necessarily diagnosed physical or mental health); May change over time (e.g., as word usage changes among the public)	VARIES; can be significant when administrative records concept differs from what statistical agency needs (e.g., rules for reporting earnings on tax forms may leave out such components as cafeteria benefits)	VARIES; can be low or high depending on how well the data correspond to statistical agency needs	VARIES, but likely significant at the present state of the art of acquiring, evaluating, and analyzing these kinds of data that arise from relatively free-form choices of autonomous individuals
Burden*	VARIES, can be high	NO ADDITIONAL BURDEN from statistical agency on relevant population (e.g., beneficiaries), but burden on administrative agency	NO ADDITIONAL BURDEN from statistical agency on relevant population (e.g., shoppers), but burden on vendor	NO ADDITIONAL BURDEN from statistical agency on relevant population (e.g., Twitter posters), but burden on vendor
Cost*	VARIES, can be high; Statistical agency bears full costs of design, collection, processing, estimation	VARIES, but could be lower than comparable survey because administrative agency bears data collection costs, but statistical agency likely incurs costs of special processing/ handling	VARIES as for administrative records, but vendor likely to want payment; Statistical agency likely incurs costs of special processing/ handling/ analyzing	VARIES as for administrative records, but vendor likely to want payment; Additional statistical agency costs for processing/analyzing unstructured data may be high

*Direction of scale changes; that is "high" is undesirable and "low" is desirable.
Note: Excludes revision error from the Biemer et al. (2014) classification.
Source: Author's rough assessment.

5.3 Uses of administrative records for household survey-based programs

Household survey respondents have demonstrated time and time again that their responses to many important questions on income, wealth, expenditures, and other topics are not very accurate. Use of administrative records has the potential in many instances to remedy this situation. An alternative strategy of many U.S. household survey programs has been to encourage the respondents themselves to consult their own records, such as tax returns, when answering questions on income and similar topics. Certainly, answers are likely to be more accurate when records are consulted, as Johnson and Moore (no date) find in a comparison of income tax records with SCF responses for the 2000 tax year. However, the strategy itself appears to be largely an exercise in futility. The same study of the SCF by Johnson and Moore reports that only ten percent of households with an adjusted gross income of less than $50,000 consulted records and that only 22 percent of higher income households did so. See National Research Council (2013a, pp. 89-91) and Moore, Marquis and Bogen (1996) for similar findings about the difficulties of getting respondents to consult records.

Turning to strategies for statistical agencies to work with administrative data directly, I identify eight ways in which administrative records can contribute to household survey data quality: (1) assist in evaluation of survey data quality, by comparison with aggregate estimates, appropriately adjusted for differences in population universes and concepts, and by exact matches of survey and administrative records; (2) provide control totals for adjusting survey weights for coverage errors; (3) provide supplemental sampling frames for use in a multiple frame design; (4) provide additional information to append to matched survey records to enhance the relevance and usefulness of the data; (5) provide covariates for model-based estimates for smaller geographic areas than the survey can support directly; (6) improve models for imputations for missing data in survey records; (7) replace "no" for survey respondents who should have reported an item, replace "yes" for survey respondents who should not have reported an item, and replace reported values for survey respondents who misreport an item; and (8) replace survey questions and use administrative records values directly. In a longer unpublished version of this article, I provide some current and potential examples of each type of use and identify benefits, confidentiality and public perception concerns, and limitations and feasibility issues for each use generically and specifically for U.S. household surveys on such topics as income, assets and expenditures. My bottom line is that the benefits should outweigh the drawbacks, given a sustained program to integrate administrative records systems with statistical programs.

5.4 Potential uses of non-traditional data sources

Having previously indicated that data from sources other than surveys and administrative records are problematic in a number of ways for official statistics, I would be remiss not to discuss briefly why such data appear to be so attractive. Private companies have very different loss functions from statistical agencies - they are seeking an edge over competitors. Data that are more timely and that identify ways to increase sales and profits are likely useful to a private company, even if they do not cover a population completely or have other drawbacks for official statistics. From this perspective, the kinds of experiments that a Google does, using its own "big data", on ways to increase ad views are good investments (see, e.g., McGuire, Manyika and Chui 2012). Similarly, program agencies at all levels of government, often working with academic centers, are putting together and analyzing their own and other data in innovative ways to identify patterns, "hot spots", and the like, not only for improving their programs and planning new services, but also for prioritizing resources and improving response in real time (see, e.g., the Center for Urban Science and Progress at New York University (http://cusp.nyu.edu/); and the Urban Center for Computation and Data at the University of Chicago (https://urbanccd.org))

Statistical agencies need, above all, sources of data that cover a known population with error properties that are reasonably well understood and that are not likely to change under their feet - characteristics that are not inherent in such data sources as autonomous interactions with websites on the Internet. There are, however, at least two ways in which household survey-based statistical agency programs could obtain an "edge" from non-traditional sources: one is to improve timeliness for preliminary estimates of key statistics; and the other is to provide leading indicators of social change (e.g., the emergence of new occupations and fields of training) that alert statistical agencies to needed changes in their concepts and measures.

6 From data needs to data sources: Two U.S. examples

For concreteness, I offer two U.S. examples - household income and housing unit characteristics - where I believe it is possible and incumbent on statistical agencies to turn survey programs into multiple sources programs to best meet user needs. The U.S. Office of Management and Budget (2014) has taken a positive step in this direction in a recent memorandum asserting that statistical uses of federal agency administrative records are a positive good and outlining step to institutionalize their use.

6.1 Household income

Official statistics on the distribution of household income are among the most important indicators of economic well-being that are regularly produced by national statistical offices, and they are even more important in light of today's debates about rising inequality and related topics. Yet it is abundantly clear that the quality of household income measures obtained from responses to U.S. surveys is significantly impaired by coverage error, unit nonresponse, item nonresponse and misreporting. Moreover, the concept of regular money income for U.S. surveys is out of date with respect to the complex and continually evolving ways in which households obtain resources for everyday consumption and savings. It seems imperative for the U.S. statistical system to improve its flagship income estimates from CPS/ASEC, SIPP, and, to the extent feasible, the ACS by moving from relying largely on survey responses to an approach that integrates survey and administrative records data. The Census Bureau is implementing new and modified questions to better measure retirement income and other sources in the CPS/ASEC, consequent to a major review of income measurement in that survey by Czajka and Denmead (2012) and a report on cognitive testing of changes to the ASEC questionnaire (Hicks and Kerwin 2011). The Census Bureau also recently implemented a major redesign of SIPP, using event history calendar methods and annual interviews in place of interviews every four months to reduce burden and costs, with effects on quality to be evaluated (see https://www.census.gov/programs-surveys/sipp/about/re-engineered-sipp.html [November 2014]). There is a process in place for review of questions on the ACS, although as yet questions on income have not been tackled. The flagship surveys would be markedly improved if, in addition to continued standard questionnaire research to identify ways to reduce burden, clarify question meaning, and facilitate response to the income questions to the extent possible, the following four steps were taken:

(1) The U.S. Census Bureau and Bureau of Economic Analysis (BEA) were to agree on - and periodically revisit and update as appropriate - a contemporary concept of regular household income on which to base estimates from the CPS/ASEC, SIPP and ACS, and the personal income series in the NIPAs, which are developed largely from administrative records. The surveys and NIPAs currently have conceptual differences, such as in the treatment of retirement benefits, which should be reconciled. Using an integrated concept of household income would make both the personal income accounts and household surveys more useful for analyzing trends from macro and micro perspectives.

(2) The Census Bureau was to conduct research on the likely benefits from implementing socioeconomic survey weight adjustments in addition to demographic weight adjustments.

Assuming a benefit, the Census Bureau would next identify appropriate sources, which could be income tax records or the SCF, to adjust weights in the CPS, SIPP and the ACS to capture coverage differences by broad socioeconomic class.

(3) The Census Bureau was to move strategically, source by source, to improve imputations of income receipt and amounts in the CPS/ASEC and SIPP by using administrative records values. The Census Bureau already has access to many records and is working to obtain additional records (e.g., SNAP records from states) as part of 2020 census planning.

(4) The Census Bureau was to move - carefully, in consideration of the added hurdles for use of administrative records in the United States - toward the Statistics Canada model, whereby respondents can skip entire blocks of income questions by permitting access to their income tax and other administrative records (see http://www.statcan.gc.ca/eng/survey/household/5200 [November 2014]).

I do not mean to underestimate the difficulties of the steps outlined above for U.S. income statistics. These difficulties, in no particular order, include: (1) legal and bureaucratic impediments to obtaining ready access to administrative records, which are orders of magnitude greater for records held by state agencies because of differences in state laws, policies and data standards and systems; (2) respondent consent considerations, particularly if values from records are substituted for questions; (3) perceptions of "big brother" and threats to privacy, which may limit the accessibility of microdata for research and policy analysis; (4) lack of resources for statistical agencies to initiate such activities as redesign of imputation systems; (5) adverse effects on timeliness to the extent that records lag in their availability to statistical agencies, which could be addressed by issuing preliminary estimates followed by final estimates when sufficient administrative data become available; (6) insufficient knowledge of the error structures of records, which could lead to nasty surprises; (7) differences in concepts between records and survey measures that are not readily addressed (e.g., earnings reported to the IRS are not gross earnings but earnings subject to tax); (8) additional burdens on already-stretched-thin statistical agency headquarters staff; (9) the need to rewrite processing systems to link multiple streams of data and conduct all needed matching, reconciling and estimation on a timely basis; (10) the distrust of many U.S. microdata users, who seem to prefer a single-source data set, such as a survey, regardless of inaccuracies in the data, to a multiple-source data set, which may include model-based values for some variables; and (11) the hesitation of statistical agency staff, who often seem to believe that it is improper to use, say, administrative records to impute income receipt to a respondent who did not indicate receipt or to use administrative records to substitute for some questions or improve some imputations, unless this can be done for all items. In planning for the 2020 census, the Census Bureau is considering limited use of administrative records for nonresponse follow-up, which could be a model for selected use of records in household surveys. Although formidable, none of these difficulties are insurmountable. A well-articulated, staged, strategic plan for taking a multiple sources approach could empower statistical agencies to work toward quality gains for income estimates and achieve at the same time a reduction in respondent burden and potentially a reduction in costs for key survey programs.

6.2 Housing characteristics, including plumbing facilities

Originating in the New Deal's concern with poor housing quality for much of the nation, the 1940 U.S. decennial census included a few questions on the characteristics of housing units. That concern was well founded - the 1940 census found, for example, that 45 percent of housing units lacked complete plumbing facilities (hot and cold piped water, flush toilet, shower or bathtub). See https://www.census.gov/hhes/www/housing/census/historic/plumbing.html [November 2014]. Housing questions grew in number and were included on censuses through 2000. When the American Community Survey came on-line, it included the housing questions previously on the long-form sample. A much smaller biannual American Housing Survey (AHS) collects an even wider range of information about housing and neighborhoods.

The major reason to investigate ways to move the ACS housing questions from a survey-based program to a survey-plus-alternative-data-sources-based program is respondent burden, both actual and perceived, which in the current political climate in the United States threatens the viability of the ACS. Because the ACS is in the field with a large sample of about 280,000 households every month, instead of once every ten years as for the census long-form sample that it replaced, the survey generates a small but continuous stream of complaints to members of Congress, which have led to congressional hearings. The Census Bureau has identified four items on the ACS that give rise to the most complaints - income, disability, time of leaving for work and plumbing facilities (see http://www.census.gov/acs/www/Downloads/operations_admin/2014_content_review/ACSContentReviewSummit.pdf[November 2014]). The questions on plumbing facilities in the census long-form sample were also regularly the butt of jokes and complaints. In fact, people answer these questions quite completely (see http://www.census.gov/acs/www/methodology/item_allocation_rates_data/ [October 2014]), yet the questions continue to be resented and sometimes not well understood (see Woodward, Wilson and Chestnut 2007). Moreover, an examination of the full ACS questionnaire suggests that many households experience a substantial burden from the total set of about 30 housing questions, particularly homeowners with a mortgage.

The Census Bureau responded to the concerns about ACS burden by cutting back the number of follow-up calls and visits (see Zelenak and Davis 2013), establishing a "respondent advocate", and giving the public information about the rationale for the questions. Yet the U.S. House of Representatives passed an appropriations bill May 30, 2014, that, if enacted, would turn the ACS into a voluntary instead of mandatory survey. While good quality data could likely be collected given enough follow-up, the costs of the ACS would increase substantially (see Griffin 2011). The Census Bureau recently asked federal agencies about legislative or regulatory justification for each and every question, with the real possibility that some questions will be dropped (see http://www.census.gov/acs/www/about_the_survey/acs_content_review/[November 2014]). Plumbing facilities might seem to be a good candidate for deletion from the ACS, given that only 0.4 percent of U.S. housing units lacked complete plumbing facilities in 2012 (From http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_1YR_DP04&prodType=table [November 2014]). However, that small percentage is concentrated in particular areas, such as Native American reservations and rural areas. Moreover, deleting any question on the ACS seems a drastic step to take without first exploring whether alternative sources might provide the data.

In fact, there are housing items on the ACS questionnaire that could very likely be obtained from a variety of other sources, attached to the Census Bureau's Master Address File (MAF), and be available for inclusion in the ACS and other surveys that use the MAF as a sampling frame. Alternative sources include local government administrative records of taxes assessed, year built, and other characteristics of properties, which are increasingly being compiled by commercial vendors, thereby reducing the need to interact with the thousands of individual governments in the United States. They also include sources like Google Street View for exterior property characteristics, realtor websites for housing value and interior characteristics (e.g., number of rooms), smart meters for utility costs (in use in some areas and likely to spread in the future), and mortgage databases held by federal agencies and commercial vendors. Housing characteristics that rarely change can also be drawn from the previous census long-form samples. Plumbing facilities is a prime example - once a house is plumbed, it is almost never unplumbed (even though at times the plumbing may not be functional).

These alternative sources will vary in how easily they are acquired and evaluated, the actual or perceived lack of threat to privacy and confidentiality they pose, and the extent to which they cover all or most of the country. Development of an augmented Master Address and Housing Unit File (MAHUF) that can serve the ACS and other Census Bureau statistical programs will take time, and, for some items (e.g., plumbing facilities), it may be necessary to use a separate (longer) version of the questionnaire in selected geographic areas that appends the relevant items. All of this will be messy, but the long-term potential payoffs are substantial. To move toward an augmented MAHUF, the Census Bureau can benefit from work of the Office of Policy Development & Research in the U.S. Department of Housing and Urban Development to streamline the lengthy AHS questionnaire by using other sources of data for many housing and neighborhood characteristics in place of survey questions; see http://www.huduser.org/portal/datasets/ahs.html#planning [November 2014].

7 Challenges and strategies for effecting paradigm change

I have argued for a paradigm change in which statistical agencies design and update their flagship programs by determining the best combination of data sources and methods to serve user needs in a topic area of ongoing importance. I use U.S. household surveys as an example, where the evidence is strong that relying on survey responses alone will not suffice to serve critical needs for high-quality information on income, expenditures, and related subjects. I expect it is also true that the use of administrative records alone, as in some countries with detailed population registers, may not provide sufficiently complete and high-quality information in the absence of regular efforts to review the quality of the register data and augment and correct them with information from other sources, such as surveys. As a case in point, Axelson, Homberg, Jansson, Werner and Westling (2012) describe the utility of surveys for evaluating the quality of housing and household data from a new dwelling register that was constructed for the 2011 census in Sweden.

I close by listing factors that make paradigm change difficult, countered by ways to effect the change I recommend and ingrain it in statistical agency culture. The U.S. and other statistical systems have admirable records of innovation in many aspects of their programs, but changing paradigms is always difficult, as was evident in the battle to introduce probability sampling to official U.S. statistics in the

1930s. It is particularly hard to rethink long-lived, ongoing, statistical programs with which both the producer agency and the user base are comfortable.

Factors that can impede change include: (1) inertia, particularly when a program was originally innovative and very well designed, so it can coast on its earlier success; (2) becoming out of touch with stakeholders' changing needs, which can be exacerbated when an agency views itself as the only source of needed data and not in competition; (3) fear of undercutting existing programs combined with fear of "not-invented here"; (4) inadequate ongoing evaluation of data quality in all of its dimensions; and (5) constrained staff and budget resources, coupled with an understandable reluctance of agency staff or their established user base to cut back on one or another long-standing statistical series in order to make important advances in other series.

Yet there are many outstanding examples of important innovation in U.S. and other nation's statistical agencies, so clearly there are ways to overcome the constraints listed above to effect paradigm change. The essential ingredient for paradigm change, I believe, is leadership buy-in and continued support at the top of a statistical agency, proactively deployed to garner buy-in at all levels of the agency. For an outstanding example of such leadership, see the discussion in National Research Council (2010a) of the role of Morris Hansen and his colleagues in reengineering what had been an enumerator-based census into a mailout/mailback census. The reengineering effort was initiated and sustained on the basis of evidence of substantial interviewer bias and variance for important data items. There was also concern that it could become more difficult to recruit enumerators as women moved into the work force.

Specific steps for agency leadership to get behind for the specific purpose of inculcating the use of multiple data sources for ongoing official statistical programs include (see Prell, Bradsher-Fredrick, Comisarow, Cornman, Cox, Denbaly, Martinez, Sabol and Vile (2009), who conducted case studies of successful statistical uses of administrative records in the United States, for similar conclusions): (1) setting clear expectations and goals for staff, such as the expectation that statistical programs will, as a matter of course, combine such sources as surveys and administrative records in the interests of relevant, accurate and timely data produced cost-effectively and with minimal respondent burden; (2) according a prominent role to subject-matter specialists - to interface with outside users and inside data producers; (3) staffing operational programs with expertise in all relevant data sources, which includes putting specialists in survey design and specialists in administrative records or other data sources on an equal footing; (4) providing for rotation of assignments, including internal rotations, rotations among statistical agencies, rotations with data user organizations and rotations with sources of alternative data sources; (5) carving out resources for continued evaluation; and (6) treating organizations with alternative data sources that play important roles in statistical programs as partners. On this last point, see, e.g., Hendriks (2012, p. 1473), who, in discussing the experiences of Statistics Norway with their first register-based census in 2011, stresses that "The three C's of register based statistics (in order to achieve data quality) are Co-operation, Communication and Coordination."

Statistical agencies have shown the ability to make far-reaching changes in response to threats to established ways of doing business. The second half of the 20th century gave us the probability survey paradigm in response to the increasing costs and burden of conducting full enumerations and the flaws of non-probability designs. The 21st century can surely give us the paradigm of using the best source(s), including surveys, administrative records and other sources, to respond to policy and public needs for relevant, accurate, timely and cost-effective official statistics.

Acknowledgements

This paper is based on the author's years of experience at the Committee on National Statistics, but the views expressed are her own and should not be assumed to represent the views of CNSTAT or the National Academy of Sciences. The author thanks John Czajka, David Johnson and Rochelle Martinez for helpful comments on an earlier draft. A longer version of this paper is available from the author on request.

References

Anderson, M.J. (1988). *The American Census: A Social History*. New Haven. CT: Yale University Press.

Axelson, M., Homberg, A, Jansson, I., Werner, P. and Westling, S. (2012). Doing a register-based census for the first time: The Swedish experience. *Paper presented at the Joint Statistical Meetings*, San Diego, CA (August). Statistics Sweden, Stockholm.

Bee, A., Meyer, B.D. and Sullivan, J.X. (2012). The validity of consumption data: Are the consumer expenditure interview and diary surveys informative? *NBER Working Paper No. 18308*. Cambridge, MA: National Bureau of Economic Research.

Biemer, P., Trewin, D., Bergdahl, H. and Lilli, J. (2014). A system for managing the quality of official statistics, with discussion. *Journal of Official Statistics*, 30(3, September), 381-442.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25(2), 139-149.

Bradburn, N.H. (1992). A response to the nonresponse problem. 1992 AAPOR Presidential Address. *Public Opinion Quarterly*, 56(3), 391-397.

Citro, C.F. (2012). *Editing, Imputation and Weighting*. Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey, Second Edition, M. J. Anderson, C.F. Citro and J.J. Salvo, eds, 201-204. Washington, DC: CQ Press.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Keynote presentation at the *5th European Survey Research Association Conference*. Ljubliana, Slovenia. http://www.europeansurveyresearch.org/sites/default/files/files/Couper%20keynote.pdf [July 2014].

Czajka, J.L. (2009). SIPP data quality. Appendix A in *Reengineering the Survey of Income and Program Participation*. National Research Council. Washington, DC: The National Academies Press.

Czajka, J.L. and Denmead, G. (2012). Income measurement for the 21st century: Updating the current population survey. Washington, DC: *Mathematica Policy Research*. Available: http://www.mathematica-mpr.com/~/media/publications/PDFs/family_support/income_measurement_21 century.pdf [July 2014].

Czajka, J.L., Jacobson, J.E. and Cody, S. (2004). Survey estimates of wealth: A comparative analysis and review of the Survey of Income and Program Participation. *Social Security Bulletin*, 65(1). Available: http://www.ssa.gov/policy/docs/ssb/v65n1/v65n1p63.html [July 2014].

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. and Nordholt, E.S. (2012). Evaluation of the quality of administrative data used in the Dutch virtual census. *Paper presented at the Joint Statistical Meetings*, San Diego, CA (August). Methodology Sector and Division of Social and Spatial Statistics, Statistics Netherland, The Hague.

De Leeuw, E.D. and De Heer, W. (2002). *Trends in Household Survey Nonresponse: A Longitudinal and International Comparison*. R.M. Groves, D.A. Dillman, J. L. Eltinge and R.J.A. Little, eds. Survey Nonresponse, 41-54. New York: Wiley.

Duncan, J. W. and Shelton, W. C. (1978). *Revolution in United States Government Statistics 1926–1976*. Office of Federal Statistical Policy and Standards, U.S. Department of Commerce. Washington, DC: Government Printing Office.

Eurostat. (2000). Assessment of the quality in statistics. *Doc. Eurostat/A4/Quality/00/General/Standard report*. Luxembourg (April 4-5). Available: http://www.unece.org/fileadmin/DAM/stats/documents/2000/11/metis/crp.3.e.pdf [July 2014].

Fixler, D. and D.S. Johnson (2012). Accounting for the distribution of income in the U.S. National Accounts. *Paper prepared for the NBER Conference on Research in Income and Wealth*, September 30. Available: http://www.bea.gov/about/pdf/Fixler_Johnson.pdf.

Fricker, S. and R. Tourangeau (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, 74(5), 935-955.

Griffin, D. (2011). Cost and workload implications of a voluntary American community survey. *U.S. Census Bureau,* Washington, DC (June 23).

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(9), 861-871. Special 75th Anniversary Issue.

Groves, R.M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72(2), 167-189.

Harris-Kojetin, B. (2012). *Federal Household Surveys*. Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey, Second Edition, M. J. Anderson, C.F. Citro and J.J. Salvo, eds, 226-234. Washington, DC: CQ Press.

Heckman, J. J. and LaFontaine, P.A. (2010). The American high school graduation rate: trends and levels. *NBER Working Paper 13670*. Cambridge, MA, National Bureau of Economic Research. Available: http://www.nber.org/papers/w13670 [July 2014].

Hendriks, C. (2012). Input data quality in register based statistics-The Norwegian experience. Proceedings of the *International Association of Survey Statisticians-JSM 2012*, 1473-1480. Paper presented at the Joint Statistical Meetings, San Diego, CA (August). Statistics Norway, Kongsvinger, Norway.

Hicks, W. and Kerwin, J. (2011). Cognitive testing of potential changes to the Annual Social and Economic Supplement of the Current Population Survey. *Report to the U.S. Census Bureau*, Westat, Rockville, MD (July 25).

Holt, D.T. (2007).The official statistics Olympics challenge: Wider, deeper, quicker, better, cheaper. *The American Statistician*, 61(1, February), 1-8. With commentary by G. Brackstone and J.L. Norwood.

Horrigan, M.W. (2013). Big data: A BLS perspective. *Amstat News*, 427(January), 25-27.

Hoyakem, C., Bollinger, C. and Ziliak, J. (2014). The role of CPS nonresponse on the level and trend in poverty. *UKCPR Discussion Paper Series*, DP 2014-05. Lexington, KY: University of Kentucky Center for Poverty Research.

Iwig, W., Berning, M., Marck, P. and Prell, M. (2013). Data quality assessment tool for administrative data. Prepared for a subcommittee of the *Federal Committee on Statistical Methodology*, Washington, DC (February).

Johnson, B and Moore, K. [no date]. Consider the source: Differences in estimates of income and wealth from survey and tax data. Available: http://www.irs.gov/pub/irs-soi/johnsmoore.pdf [July 2014].

Keller, S.A., Koonin, S.E. and Shipp, S. (2012). Big data and city living - what can it do for us? *Statistical Significance*, 9(4), 4-7, August.

Kennickell, A. (2011). Look again: Editing and imputation of SCF panel data. *Paper prepared for the Joint Statistical Meetings*, Miami, FL (August).

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *META Group [now Gartner] Research Note*, February 6. See: http://goo.gl/Bo3GS [July 2014].

Manski, C.F. (2014). Communicating uncertainty in official economic statistics. *NBER Working Paper No. 20098*. Cambridge, MA: National Bureau of Economic Research.

McGuire, T., Manyika, J. and Chui, M. (2012). Why big data is the new competitive advantage. *Ivey Business Journal* (July-August).

Meyer, B. D. and Goerge, R.M. (2011). Errors in survey reporting and imputation and their effects on estimates of Food Stamp Program participation. Working Paper. *Chicago Harris School of Public Policy*, University of Chicago.

Meyer, B.D., Mok, W. K.C. and Sullivan, J.X. (2009). The under-reporting of transfers in household surveys: Its nature and consequences. *NBER Working Paper No. 15181*. Cambridge, MA: National Bureau of Economic Research.

Moore, J.C., Marquis, K.H. and Bogen, K. (1996). The SIPP cognitive research evaluation experiment: Basic results and documentation. *SIPP Working Paper No. 212*. U.S. Census Bureau, Washington, DC (January). Available: http://www.census.gov/sipp/workpapr/wp9601.pdf [July 2014].

Morganstein, D. and Marker, D. (2000). A conversation with Joseph Waksberg. *Statistical Science*, 15(3), 299-312.

Mule, T. and Konicki, S. (2012). *2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Housing Units in the United States*. U.S. Census Bureau, Washington, DC.

National Research Council (1995). *Measuring Poverty: A New Approach*. Washington, DC: National Academy Press.

National Research Council (2004). *The 2000 Census: Counting Under Adversity*. Washington, DC: The National Academies Press.

National Research Council (2010a). *Envisioning the 2010 Census*. Washington, DC: The National Academies Press.

National Research Council (2010b). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press.

National Research Council (2013a). *Measuring What We Spend: Toward a New Consumer Expenditure Survey*. Washington, DC: The National Academies Press.

National Research Council (2013b). *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: The National Academies Press.

National Research Council (2013c). *Principles and Practices for a Federal Statistical Agency*. Washington, DC: The National Academies Press.

Nelson, N. and West, K. (2014). Interview with Lars Thygesen. *Statistical Journal of the IAOS*, 30, 67-73.

Passero, B. (2009). The impact of income imputation in the Consumer Expenditure Survey. *Monthly Labor Review* (August), 25-42.

Prell, M., Bradsher-Fredrick, H., Comisarow, C., Cornman, S., Cox, C., Denbaly, M., Martinez, R.W., Sabol, W. and Vile, M. (2009). Profiles in success of statistical uses of administrative records. Report of a subcommittee of the *Federal Committee on Statistical Methodology*, U.S. Office of Management and Budget, Washington, DC.

Shapiro, G.M. and Kostanich, D. (1988). High response error and poor coverage are severely hurting the value of household survey data. *Proceedings of the Section on Survey Research Methods*, 443-448, American Statistical Association, Alexandria, VA. Available: http://www.amstat.org/sections/srms/Proceedings/papers/1988_081.pdf [July 2014].

Steeh, C.G. (1981). Trends in nonresponse rates, 1952-1979. *Public Opinion Quarterly*, 45, 40-57.

Tourangeau, R. (2004). Survey research and societal change. *Annual Review of Psychology*, 55, 775-801.

U.S. Office of Management and Budget. (2014). *Guidance for Providing and Using Administrative Data for Statistical Purposes*. Memorandum M-14-06. Washington, DC.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Woodward, J., Wilson, E. and Chesnut, J. (2007). *Evaluation Report Covering Facilities - Final Report. 2006 American Community Survey Content Test Report H.3.U.S.* Census Bureau. Washington, DC: U.S. Department of Commerce. January.

Zelenak, M.F. and M.C. David (2013). *Impact of Multiple Contacts by Computer-Assisted Telephone Interview and Computer-Assisted Personal Interview on Final Interview Outcome in the American Community Survey*. U.S. Census Bureau, Washington, DC.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 163-188
Statistics Canada, Catalogue No. 12-001-X

Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers

Brady T. West and Michael R. Elliott[1]

Abstract

Survey methodologists have long studied the effects of interviewers on the variance of survey estimates. Statistical models including random interviewer effects are often fitted in such investigations, and research interest lies in the magnitude of the interviewer variance component. One question that might arise in a methodological investigation is whether or not different groups of interviewers (e.g., those with prior experience on a given survey vs. new hires, or CAPI interviewers vs. CATI interviewers) have significantly different variance components in these models. Significant differences may indicate a need for additional training in particular subgroups, or sub-optimal properties of different modes or interviewing styles for particular survey items (in terms of the overall mean squared error of survey estimates). Survey researchers seeking answers to these types of questions have different statistical tools available to them. This paper aims to provide an overview of alternative frequentist and Bayesian approaches to the comparison of variance components in different groups of survey interviewers, using a hierarchical generalized linear modeling framework that accommodates a variety of different types of survey variables. We first consider the benefits and limitations of each approach, contrasting the methods used for estimation and inference. We next present a simulation study, empirically evaluating the ability of each approach to efficiently estimate differences in variance components. We then apply the two approaches to an analysis of real survey data collected in the U.S. National Survey of Family Growth (NSFG). We conclude that the two approaches tend to result in very similar inferences, and we provide suggestions for practice given some of the subtle differences observed.

Key Words: Interviewer variance; Bayesian analysis; Hierarchical generalized linear models; Likelihood ratio testing.

1 Introduction

Between-interviewer variance in survey methodology (e.g., West, Kreuter and Jaenichen 2013; West and Olson 2010; Gabler and Lahiri 2009; O'Muircheartaigh and Campanelli 1998; Biemer and Trewin 1997; Kish 1962) occurs when survey responses nested within interviewers are more similar than responses collected from different interviewers. Between-interviewer variance can increase the variance of survey estimates of means, and may arise due to correlated response deviations introduced by an interviewer (e.g., Biemer and Trewin 1997), given the complexity of survey questions (e.g., Collins and Butcher 1982) or interactions between the interviewer and the respondent (e.g., Mangione, Fowler and Louis 1992), or nonresponse error variance among interviewers (West et al. 2013; Lynn, Kaminska and Goldstein 2011; West and Olson 2010).

Survey research organizations train interviewers to eliminate this component of variance in survey estimates, as it is sometimes larger than the component of variance due to cluster sampling (Schnell and Kreuter 2005). In reality, an interviewer variance component can never be equal to 0 (which would imply that means on the variable of interest are identical across interviewers), but survey managers aim to minimize this component via specialized interviewer training. For example, interviewers may practice the

1. Brady T. West, Survey Methodology Program, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI, 48106. E-mail: bwest@umich.edu; Michael R. Elliott, Survey Methodology Program, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI, 48106. E-mail: mrelliot@umich.edu.

administration of selected questions under the direct supervision of training staff, and then receive feedback on any variance in administration that is noted by the staff (in an effort to standardize the administration; see Fowler and Mangione 1990). In some non-interpenetrated designs, where interviewers are generally assigned to work exclusively in a single primary sampling area (e.g., the U.S. National Survey of Family Growth; see Lepkowski, Mosher, Davis, Groves and Van Hoewyk 2010), interviewer effects and area effects are confounded, preventing estimation of the variance in survey estimates that is uniquely attributable to the interviewers. Elegant interpenetrated sample designs (Mahalanobis 1946) enable interviewers to work in multiple sampling areas, and in these cases, cross-classified multilevel models can be used to estimate the components of variance due to interviewers and areas (e.g., Durrant, Groves, Staetsky and Steele 2010; Gabler and Lahiri 2009; Schnell and Kreuter 2005; O'Muircheartaigh and Campanelli 1999; O'Muircheartaigh and Campanelli 1998).

In general, estimating the overall magnitude of interviewer variance in the measures of a given survey variable or data collection process outcome is a useful exercise for survey practitioners. If random subsamples of sample units are assigned to interviewers following an interpenetrated design, one can estimate the component of variance due to interviewers and subsequently the unique effects of interviewers on the variance of an estimated survey mean (e.g., Groves 2004, p. 364). Large estimates can indicate potential measurement difficulties that certain interviewers are experiencing, or possible differential success in recruiting particular types of sampled units. Given a relatively large estimate of an interviewer variance component and an appropriate statistical test indicating that the component is significantly larger than zero (or "non-negligible", given that variance components technically cannot be exactly equal to zero; see Zhang and Lin 2010), survey managers can use various methods to compute predictions of the random effects associated with individual interviewers, and identify interviewers who may be struggling with particular aspects of the data collection process.

While the estimation of interviewer variance components and subsequent adjustments to interviewer training and data collection protocols have a long history in the survey methodology literature (see Schaeffer, Dykema and Maynard 2010 for a recent review), no studies in survey methodology to date have examined the alternative approaches that are available to survey researchers for *comparing* variance components in two independent groups of survey interviewers. In general, alternative statistical approaches are available for estimating interviewer variance components, and estimates (and corresponding inferences about the variance components) may be sensitive to the estimation methods that a survey researcher employs. The same is true for survey researchers who may desire to compare the variance components associated with different groups of interviewers, for various reasons (e.g., identifying groups that need more training or more optimal modes for certain types of questions): different statistical approaches to performing these kinds of comparisons exist, and inferences about the differences may be sensitive to the approach used. With this paper, we aim to evaluate alternative frequentist and Bayesian approaches to making inference about the differences in variance components between two independent groups of survey interviewers, and provide practical guidance to survey researchers interested in this type of analysis.

The paper is structured as follows. In Section 2, we introduce the general modeling framework that enables these comparisons of interviewer variance components for both normal and non-normal (e.g., binary, count) survey variables, and review existing literature comparing the frequentist and Bayesian approaches to estimation and inference, highlighting the advantages and disadvantages of each approach.

We then present a simulation study in Section 3, evaluating the ability of the two approaches to efficiently estimate differences in variance components between two hypothetical groups of interviewers. Section 4 applies the two approaches to real survey data collected in the U.S. National Survey of Family Growth (NSFG) (Lepkowski et al. 2010; Groves, Mosher, Lepkowski and Kirgis 2009). Finally, Section 5 offers concluding thoughts, suggestions for practitioners, and directions for future research. We include SAS, R, and WinBUGS code that readers can use to implement the two approaches in the Appendix.

2 Alternative approaches for comparing variance components in Hierarchical Generalized Linear Models

We first consider a general class of models that survey researchers can employ to compare variance components in different groups of interviewers. Hierarchical Generalized Linear Models (HGLMs) are flexible analytic tools that can be used to model observations on both normal and non-normal (e.g., binary, count) survey variables of interest, where observations nested within the same interviewer cannot be considered independent (Raudenbush and Bryk 2002; Goldstein 1995). We consider alternative approaches to making inferences about interviewer variance components in a specific class of HGLMs, where the interviewer variance components for two independent groups of interviewers defined by a known interviewer characteristic need not be equal. This type of HGLM can be written as

$$\begin{aligned} & g\left(E\left[y_{ij} \mid u_i\right]\right) = \beta_0 + \beta_1 I\left(Group = 1\right)_i + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i \\ & u_{i(1)} \sim N\left(0, \tau_1^2\right),\ u_{i(2)} \sim N\left(0, \tau_2^2\right), \end{aligned} \tag{2.1}$$

where $g\left(x\right)$ is the link function relating a transformation of the expected value of the dependent variable, y_{ij}, to the linear combination of the fixed and random effects (e.g., $g\left(x\right) = log\left[x/\left(1-x\right)\right]$ for an assumed Bernoulli distribution [binary outcome], $g\left(x\right) = log\left(x\right)$ for an assumed Poisson distribution [count outcome]), i is an index for the interviewer, j is an index for the respondent nested within an interviewer, and $I\left(\bullet\right)$ represents an indicator variable, equal to 1 if the condition inside the parentheses is true and 0 otherwise. The random interviewer effects from Group 1, $u_{i(1)}$, are assumed to follow a normal distribution with mean 0 and variance τ_1^2, while the random interviewer effects from Group 2, $u_{i(2)}$, are assumed to follow a normal distribution with mean 0 and variance τ_2^2. Other distributions may be posited for the random effects, and the general model in (2.1) can accommodate over-dispersion in the observed dependent variable relative to the posited distribution for that variable. The key aspect of the specification in (2.1) is that random effects for different groups of interviewers have *different* variances. The fixed effect parameter β_1 in (2.1) represents a fixed effect of Group 1 on the outcome relative to Group 2 in the HGLM, and fixed effects of other covariates can easily be included. Similarly, additional subgroups of interviewers can be considered by including additional random effects $u_{i(k)}$, for $k > 2$. Analytic interest lies in the magnitude of the difference in the variance components.

Models of the form in (2.1) can be applied when methodological studies are designed to compare two different groups of interviewers in terms of their variance components. For example, there exists a debate

in the survey methodology literature regarding whether interviewers should use standardized or conversational interviewing. Proponents of standardized interviewing argue that all interviewers should administer surveys in the exact same way, allowing respondents to interpret questions as they see fit (e.g., Fowler and Mangione 1990). Other research has shown that more flexible interviewing using a conversational style may increase respondent understanding of survey questions and reduce measurement error (e.g., Schober and Conrad 1997). To test a hypothesis that one interviewing style results in lower between-interviewer variance, a researcher might randomize interviewers to two groups trained in the two different styles, collect survey data on a variety of variables, and then fit model (2.1), including indicator variables for the two groups of interviewers. This same approach could be used to compare the interviewer variance components in two groups of interviewers randomly assigned to different data collection modes (e.g., CAPI vs. CATI). To date, no published studies have attempted these kinds of comparisons, but they are important for understanding the overall impacts of these design decisions on the mean squared error (MSE) of survey estimates.

Frequentist approaches to the estimation of parameters in HGLMs rely on various numerical or theoretical approaches to approximating complicated likelihood functions, especially for models such as (2.1) that involve complex random effects structures (e.g., Faraway 2006, Chapter 10; Molenberghs and Verbeke 2005). In general, inferences are based on these approximate likelihood-based approaches, which include residual pseudo-likelihood (which is different from the pseudo-maximum likelihood estimation approach developed by Binder (1983) for design-based analyses of data from complex sample surveys), penalized quasi-likelihood, and maximum likelihood based on a Laplace approximation. Previous work has found favorable simulation results for the residual pseudo-likelihood approach, which indicate nearly unbiased estimation of the variance components in an HGLM as compared to maximum likelihood using Laplace approximation or adaptive quadrature (Pinheiro and Chao 2006). These findings are similar to the case of restricted maximum likelihood (REML) estimation in a model for a normally distributed outcome variable. For binary outcome variables, marginal or penalized quasi-likelihood techniques can lead to downward bias in parameter estimates and convergence problems, and fully Bayesian approaches may have favorable properties in this case (Browne and Draper 2006; Rodriguez and Goldman 2001). We therefore consider the residual pseudo-likelihood approach in the simulations and applications presented in this study, and contrast this approach with a fully Bayesian approach.

There are two approaches available for making inference about differences in variance components in the frequentist setting. The first approach involves testing the null hypothesis that $\tau_1^2 = \tau_2^2,$ versus the alternative hypothesis that $\tau_1^2 \neq \tau_2^2.$ Conceptually, this is a simple hypothesis test to perform using frequentist methods, as the null hypothesis defines an equality constraint rather than setting a parameter to a value on the boundary of a parameter space. The model under the null hypothesis is nested within the model under the alternative hypothesis, where $\tau_2^2 = \tau_1^2 + k.$ The null hypothesis can thus be rewritten as $k = 0,$ versus the alternative that $k \neq 0.$ A test statistic is computed by fitting a constrained version of the model in (2.1), with the random effect variance components in the two groups specified as equal, and then fitting the model with the more general form in (2.1). The positive difference in the approximate -2 log-likelihood values of these two models is then computed, and referred to a chi-square distribution with one degree of freedom.

The second approach involves computing the difference of the pseudo-ML estimates, $\hat{\tau}_1 - \hat{\tau}_2$, and an associated 95% Wald-type confidence interval for the difference, given by $\hat{\tau}_1 - \hat{\tau}_2 \pm 1.96\sqrt{\hat{\text{var}}(\hat{\tau}_1) + \hat{\text{var}}(\hat{\tau}_2) - 2\hat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_2)}$. This interval requires asymptotic estimates of the variances and covariances of the two estimated variance components, which are computed based on the Hessian (second derivative) matrix of the objective function used for the maximum likelihood estimation procedure. If the resulting Wald interval includes zero, one would conclude that there is not enough evidence against the null hypothesis. Confidence intervals for differences in variance components can also be computed using inversions of profile likelihood tests (e.g., Viechtbauer 2007), although standard software does not include options for implementing this procedure (to our knowledge).

These two frequentist approaches to making inference about differences in interviewer variance components do have limitations. When the number of interviewers in each group is small (say, less than 30; see Hox (1998) for discussion), asymptotic results for the likelihood ratio test (Zhang and Lin 2010) may no longer hold. Frequentist (maximum likelihood) methods also tend to overstate the precision of estimates, given that they ignore the uncertainty in estimates of the variance components (Carlin and Louis 2009, p. 335-336), which is especially problematic for small samples (Goldstein 1995, p. 23). Bayesian approaches allow analysts to place prior distributions on variance components to reflect this uncertainty, unlike frequentist approaches. Furthermore, Molenberghs and Verbeke (2005, p. 277) argue that likelihood ratio tests should not be used to test hypotheses when models are fitted using pseudo-likelihood methods. Approximate maximum likelihood estimation methods can also lead to invalid (i.e., negative) estimates of variance components in these models when variance components are very small. Software that does not use estimation procedures constraining these variance components to be greater than zero generally responds to this problem by setting negative estimates of variance components equal to zero (with no accompanying standard error), which prevents computation of the Wald-type confidence interval described above.

A Bayesian approach to fitting the HGLMs described in (2.1) uses the MCMC-based Gibbs sampler and the adaptive rejection sampling methodology (Gilks and Wild 1992) to simulate draws from the posterior distribution for the parameters in the model defined in (2.1). In general, the posterior distributions for the parameters in an HGLM are not of known distributional forms and need to be simulated (Gelman, Carlin, Stern and Rubin 2004, Section 16.4). Diffuse, non-informative priors for the fixed effects and the variance components in (2.1) can be specified for the simulations, to let the data provide the most information about the posterior distributions of the parameters (Gelman and Hill 2007; Gelman 2006, Section 7). This approach enables inferences based on simulated draws from the marginal posterior distributions of the two fixed effect parameters, the two variance parameters, the random interviewer effects, and any functions of these parameters. This study focuses on the marginal posterior distribution of the difference in the random effect variances for two groups of interviewers defined by a known interviewer-level characteristic, computed using the simulated draws of the two variance components.

Given that traditional hypothesis tests are not meaningful in the Bayesian setting, Bayesian inference will focus on the difference in the interviewer variance components. Inference for the difference is based on several thousand draws of the two variance components from the joint posterior distribution estimated using the Gibbs sampler. For each draw d of the two variance components, the difference in the variance

components, defined as $\tau_1^{2(d)} - \tau_2^{2(d)}$, can be computed. Inferences will then be based on the marginal distribution of these differences, ignoring the draws of the random interviewer effects and the other nuisance parameters. The median and the 0.025 and 0.975 quantiles (for a 95% credible set) of the simulated differences of the two variance components will be computed based on the effective number of simulation draws of the two variance components from the estimated joint posterior distribution. In a given analysis, several thousand draws from the posterior distribution can be generated using the Gibbs sampler, with a large number of initial draws discarded as burn-in draws, and the effective number of simulation draws will be computed based on the number of burn-in draws (Gelman and Hill 2007, Chapter 16). If the resulting 95% credible set includes 0, there will be evidence in favor of the two groups having equal variance components. If the 95% credible set does not include 0, there will be evidence in favor of the two groups having different variances, with a positive median suggesting that group 1 has the higher variance component. Inference for the two fixed effects can follow a similar approach.

Focusing on draws of the two variance components from the full joint posterior distribution (and their differences) and ignoring draws of the random interviewer effects and the fixed effects has the effect of integrating these other parameters out of the joint posterior distribution. This Bayesian approach therefore provides a convenient methodology for simulating draws from the marginal distribution of a complicated parameter (the difference between the two variance components) and computing a 95% credible set for that parameter. While such estimates can also be obtained in the frequentist approach, as noted previously, the Bayesian approach does not require asymptotic assumptions and incorporates the variability in the estimated variance components into the computation of the 95% credible sets via the simulated draws.

Multiple (typically three) Markov chains can be run in parallel in the iterative Gibbs sampling algorithm to simulate random walks through the space of the joint posterior distribution. The Gelman-Rubin $\hat{R}$ statistic, representing (approximately) the square root of the variance of the mixture of the chains divided by the average within-chain variance (Gelman and Rubin 1992), can be used to assess convergence (or mixing) of the chains for each parameter. Values less than 1.1 on this statistic can be considered as evident of convergence of the chains for a given parameter. Posterior draws of the parameters can be pooled from the three chains to generate the final effective sample size of draws used for inferences.

The Bayesian approach outlined above is also not without limitations. The selection of the prior distributions used to compute the posterior distribution for the parameters in (2.1) is essentially arbitrary, and depends on the choices of a given analyst and the amount of prior information available. Furthermore, the choice of the prior distribution can become crucial when there is a small number of interviewers (say, less than 20), where different priors can lead to very different inferences regarding the variance components (Lambert, Sutton, Burton, Abrams and Jones 2005); the use of prior information about the variance components can increase efficiency relative to the use of non-informative priors in these cases. Model misspecification is also a distinct possibility depending on the survey variable being modeled, which is also a limitation of the frequentist approach. Computational demand may also be an issue with the Bayesian (Gibbs sampling) approach (Browne and Draper 2006), especially if one desires comparisons of interviewer variance components for a large number of survey variables (with potentially different distributions) and there are a relatively large number of interviewers; this may not be as problematic with recent advances in hardware speed and algorithm efficiency. Finally, analysts may not be comfortable

with the available software for Bayesian approaches, so there may be a learning curve associated with implementation of this approach.

Several previous articles have compared these alternative frequentist and Bayesian approaches using simulation studies. Chaloner (1987) considered one-way ANOVA models with random effects for unbalanced data (similar to the case in this study, where interviewers have different workloads), and found lower empirical MSE values for posterior modes of the variance components when following the Bayesian approach and using non-informative priors than for the frequentist (maximum likelihood) approach. Van Tassell and Van Vleck (1996) reported that the Gibbs sampler (using either informative or non-informative prior distributions) and REML both produce empirically unbiased estimates of variance components that tend to be extremely similar. Browne and Draper (2006) also found that both approaches can lead to unbiased estimates, with the more "automatic" nature of frequentist approaches being an attractive feature. In the context of predicting means for small areas using models with random area effects, Singh, Stukel and Pfeffermann (1998) reported that Bayesian MSE approximations for the predictions have good frequentist properties, but that the Bayesian method tends to produce larger frequentist biases and prediction MSEs than frequentist methods. Farrell (2000) found that the Bayesian approach resulted in slightly more accurate predictions of small area proportions, with little differences in coverage rates or bias between the two approaches. Ugarte, Goicoa and Militino (2009) also found that the two approaches performed quite similarly in an application involving the detection of high-risk areas for disease. These authors point out that the relative computational simplicity of the frequentist approach is attractive in light of these findings. In general, based on the literature in this area, we anticipate similar performance of the two methods in the case of comparing interviewer variance components, and we evaluate this expectation using a simulation study (Section 3).

While there exist many software procedures for fitting multilevel models and estimating variance components using both frequentist and Bayesian methods (see West and Galecki 2011 for a review), the frequentist approach to the specific comparison of variance components discussed in this paper is only readily implemented in the GLIMMIX procedure of SAS/STAT (SAS 2010), through the COVTEST statement with the HOMOGENEITY option (which assumes that a GROUP variable has been specified in the RANDOM statement, indicating different groups of clusters with random effects arising from different distributions). We are not aware of any other procedures that readily implement the frequentist comparison approach at the time of this writing. Example code that can be used for fitting these models using the GLIMMIX procedure is available in the Appendix. The Bayesian approach to comparing the variance components can be implemented in the BUGS (Bayesian Inference using Gibbs Sampling) software (see References Section for more details). We also include example code that implements this approach by calling WinBUGS from R in the Appendix.

3 Simulation Study

We conducted a small simulation study to examine the empirical properties of these two alternative approaches. Data on two hypothetical survey variables of interest (one normally distributed, one Bernoulli distributed) were simulated according to the following two super-population models:

$$y_{ij} = 45 + 5 \times I\left(Group = 2\right)_i + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i + \varepsilon_{ij}$$
$$u_{i(1)} \sim N\left(0,1\right),\ u_{i(2)} \sim N\left(0,2\right),\ \varepsilon_{ij} \sim N\left(0,64\right) \tag{3.1}$$

$$P\left(y_{ij} = 1\right) = \frac{\exp\left[-1 + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i\right]}{1 + \exp\left[-1 + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i\right]} \tag{3.2}$$
$$u_{i(1)} \sim N\left(0,0.03\right),\ u_{i(2)} \sim N\left(0,0.13\right).$$

The notation used here is consistent with that used in (2.1). Values on the second Bernoulli variable were generated for hypothetical cases according to the logistic regression model specified in (3.2). To obtain the observed Bernoulli variable, a random draw was obtained from a UNIFORM(0,1) distribution, and the variable was set to 1 if the random draw was less than or equal to the predicted probability, and 0 otherwise. For one hypothetical group of interviewers at a time, random interviewer effects were drawn, and values for cases within each interviewer were then generated according to the specified model.

We generated 200 samples of hypothetical cases and simulated data for each variable, with 50 hypothetical interviewers in one group collecting data from 50 hypothetical cases each ($n = 2{,}500$ for each group of interviewers). We then generated an additional 200 samples in a small-sample scenario, with 20 interviewers in each group collecting data from ten hypothetical cases each ($n = 200$ for each group of interviewers). The choices of the variance components in (3.1) correspond to intra-interviewer correlations of 0.015 and 0.030 for the two hypothetical groups of interviewers, while the choices of the variance components in (3.2) correspond to intra-interviewer correlations of 0.009 and 0.038. All of these values would be considered plausible in a face-to-face or telephone survey setting (West and Olson 2010). The known differences in variance components between the groups are therefore 1 for the normal variable, and 0.1 for the Bernoulli variable.

Given these known values for the interviewer variance components in the hypothetical population, we applied each method described in Section 2 [using diffuse, non-informative, uniform priors for the variance components, per recommendations of Gelman (2006, Section 7)] to each hypothetical sample. We computed the following empirical measures for comparison purposes: 1) the empirical and relative bias of the estimator; 2) the empirical MSE of the estimator; 3) the "frequentist" coverage of the 95% Wald-type intervals (when using the frequentist approach) and the 95% credible sets (when using the Bayesian approach); and 4) the average widths of the 95% Wald-type intervals and the credible sets. The number of Wald-type intervals that could not be computed due to estimated variance components of 0 (with no accompanying standard errors) was also recorded in each case. All simulations were performed using SAS, R, and BUGS, and simulation code is available upon request.

Table 3.1 presents the results of the simulation study. The results suggest that for moderate-to-large samples of interviewers and respondents, both approaches yield estimators of the difference in variance components that have fairly small bias, as anticipated. The frequentist approach was found to yield estimators with smaller empirical MSE values; this is not entirely surprising, given the additional variability in the Bayesian estimates introduced by accounting for uncertainty in the prior distributions of the parameters with non-informative priors. The use of more informative priors may improve the efficiency of the Bayesian estimates. In the large sample setting, the 95% confidence intervals and credible sets computed for the difference in variance components appear to have acceptable coverage properties, with the Bayesian approach having slight under-coverage.

Table 3.1
Results of simulation study comparing the empirical properties of the frequentist and Bayesian approaches to making inference about the differences in interviewer variance components.

Sample Sizes		Frequentist Approach	Bayesian Approach
	Normal Y		
	Empirical Bias	-0.0498	-0.0189
	Relative Bias	-4.98%	-1.89%
	Empirical MSE	0.6546	0.8134
	95% CI/CS Coverage	0.960	0.920
	Mean 95% CI/CS Width	3.1689	3.6283
50 interviewers / group	% of Wald CIs Invalid	0.0%	--
50 cases / interviewer	**Bernoulli Y**		
($n = 2,500$ / group)	Empirical Bias	-0.0020	-0.0046
	Relative Bias	-2.0%	-4.6%
	Empirical MSE	0.0029	0.0033
	95% CI/CS Coverage	0.938	0.940
	Mean 95% CI/CS Width	0.2142	0.2372
	% of Wald CIs Invalid	11.5%	--
	Normal Y		
	Empirical Bias	-0.2341	-0.3508
	Relative Bias	-23.41%	-35.08%
	Empirical MSE	6.9873	6.2869
	95% CI/CS Coverage	1.000	0.995
	Mean 95% CI/CS Width	16.6313	18.3574
20 interviewers / group	% of Wald CIs Invalid	54.0%	--
10 cases / interviewer	**Bernoulli Y**		
($n = 200$ / group)	Empirical Bias	-0.0348	-0.0196
	Relative Bias	-34.8%	-19.6%
	Empirical MSE	0.0345	0.0861
	95% CI/CS Coverage	1.000	0.980
	Mean 95% CI/CS Width	1.2604	1.7970
	% of Wald CIs Invalid	65.5%	--

Notably, 11.5% of the 95% Wald-type confidence intervals could not be computed when analyzing the binary outcome for the larger samples, due to one of the estimated variance components being equal to zero (with no standard error). This "failure" rate for the Wald intervals became much worse for both variables in the smaller samples, where both methods also produced inefficient estimates with a negative bias. The frequentist approach can therefore provide an estimate of the difference and associated confidence intervals that work well in larger samples with normally distributed variables, but in small samples or even moderate-to-large samples with non-normal variables, the simple Wald-type intervals that can be computed using standard software may fail a fairly substantial fraction of the time. This is due to

the fact that the Hessian matrix is not invertible when an estimated variance component is set to zero (i.e., the likelihood can't be approximated by a quadratic). Collectively, these simulation results therefore suggest that: 1) both approaches will perform similarly well when applied to real survey data with moderate-to-large samples of interviewers and respondents; 2) the Bayesian approach may be the better option if intervals (or credible sets) for the difference are desired; and 3) caution is advised when applying either method to relatively small samples of interviewers and respondents.

4 Application: The U.S. National Survey of Family Growth (NSFG)

We now apply the frequentist and Bayesian approaches to real survey data collected in the seventh cycle of the NSFG (June 2006 – June 2010). The original design of this cycle of the NSFG (Groves et al. 2009) called for 16 quarters of data collection from a continuous sample that was nationally representative when it was completed in June 2010. The data analyzed in this paper were collected from a national sample of 11,609 females between the ages of 15 and 44, by 87 female interviewers (with varying sample sizes for each interviewer). For more details on the design and operation of the seventh cycle of the NSFG, see Lepkowski et al. (2010) or Groves et al. (2009).

Each of the 87 interviewers has information available on her age (47.1% are age 55 or greater), years of experience (43.7% have five or more years of experience), number of children (33.3% have two or more children), marital status (19.5% have never been married), other employment (46.0% have other jobs), college education (57.5% completed a four-year college degree), previous experience working on NSFG (82.8% have worked on previous cycles), and ethnicity (81.6% are white). These observable interviewer-level characteristics will be used to divide the interviewers into two groups (in the absence of an ideal randomized experiment, like that described in Section 2).

Each of the 11,609 female respondents has their parity (or count of live births) and an indicator of current sexual activity (indicated by at least one current male partner or at least one male partner in the past 12 months) measured and available for analysis. While these measures seem fairly simple, the concepts being measured may be communicated differently by different interviewers (resulting in interviewer variance). The primary analytic question is whether these different groups of female interviewers have significantly different variance components for these particular survey variables.

We first consider an HGLM for the parity variable. Let Y be a Poisson random variable with parameter λ. We allow for overdispersion (or extra-Poisson dispersion) in Y, which is quite common in count variables (for example, the mean parity for the sample of 11,609 females is 1.19, and the variance of the measured parity values is 1.99). Following Hilbe (2007) and Durham, Pardoe and Vega (2004), we let $\lambda = r\mu$, where r is a GAMMA$\left(\alpha^{-1},\alpha^{-1}\right)$ random variable. It then follows that Y has a negative binomial distribution with mean μ and scale parameter α:

$$E(Y) = E(\lambda) = E(r\mu) = \mu E(r) = \mu$$

$$\text{var}(Y) = E(\lambda) + \text{var}(\lambda) = E(r\mu) + \text{var}(r\mu) = \mu E(r) + \mu^2 \text{var}(r) = \mu(1 + \alpha\mu)$$

We specify an HGLM for the observed value of parity on female respondent j interviewed by interviewer i, y_{ij}, as follows:

$$
\begin{aligned}
&y_{ij} \sim Poisson\left(\lambda_i\right),\ \lambda_i = r_i\mu_i \\
&r_i \sim Gamma\left(\alpha^{-1},\alpha^{-1}\right) \\
&\log\left(\mu_i\right) = \beta_0 + \beta_1 I\left(Group = 1\right)_i + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i \\
&u_{i(1)} \sim N\left(0,\tau_1^2\right), \qquad u_{i(2)} \sim N\left(0,\tau_2^2\right).
\end{aligned} \tag{4.1}
$$

In this multilevel negative binomial regression model, $\exp\left(\beta_0\right)$ represents the expected parity for Group 2, $\exp\left(\beta_1\right)$ represents the expected multiplicative change in parity for Group 1 relative to Group 2, $u_{i(1)}$ is a random effect associated with interviewer i in Group 1, and $u_{i(2)}$ is a random effect associated with interviewer i in Group 2.

Next, we consider an HGLM for the binary indicator of current sexual activity. Let $z_{ij} = 1$ if a female respondent j indicates current sexual activity to interviewer i, and 0 otherwise. We specify the following model for this binary indicator:

$$
\begin{aligned}
&z_{ij} \sim Bernoulli\left(p_i\right) \\
&\ln\left[p_i/\left(1-p_i\right)\right] = \beta_0 + \beta_1 I\left(Group = 1\right)_i + u_{i(1)} I\left(Group = 1\right)_i + u_{i(2)} I\left(Group = 2\right)_i \\
&u_{i(1)} \sim N\left(0,\tau_1^2\right), \qquad u_{i(2)} \sim N\left(0,\tau_2^2\right).
\end{aligned} \tag{4.2}
$$

In this model, $\exp\left(\beta_0\right)$ represents the expected odds of current sexual activity for Group 2, $\exp\left(\beta_1\right)$ represents the expected multiplicative change in the odds of current sexual activity for Group 1 relative to Group 2, $u_{i(1)}$ is a random effect associated with interviewer i in Group 1, and $u_{i(2)}$ is a random effect associated with interviewer i in Group 2.

We fit models (4.1) and (4.2) using the two approaches described in Section 2. For the frequentist approach, based on recommendations from the literature discussed in Section 2, we estimated the parameters in these models using residual pseudo-likelihood (RPL) estimation, as implemented in the GLIMMIX procedure in the SAS/STAT software. All frequentist analyses presented in this section were repeated using adaptive quadrature to approximate the likelihood functions, and the primary results did not change; in addition, the use of adaptive quadrature led to longer estimation times.

For the Bayesian approach, the following non-informative prior distributions for these parameters were used. These prior distributions were selected based on a combination of estimates from initial naïve model fitting, and recommendations from Gelman and Hill (2007) and Gelman (2006, Section 7) for proper but non-informative prior distributions for variance parameters in hierarchical models with a reasonably large number (i.e., more than five) of groups (or interviewers, in the present context):

$$
\begin{aligned}
&\beta_0 \sim N\left(0,100\right) \qquad \beta_1 \sim N\left(0,100\right) \\
&\tau_1^2 \sim Uniform\left(0,10\right) \qquad \tau_2^2 \sim Uniform\left(0,10\right) \\
&\ln(\alpha) \sim N\left(0,100\right).
\end{aligned}
$$

The non-informative priors for the fixed effects and the (natural log transformed) scale parameter for the negative binomial count variable (parity) indicate an expectation that these parameters will be somewhere in the range (-10, 10), while the non-informative priors for the variance components are uniform distributions on the range (0, 10). Given initial naive estimates of the fixed effects ranging

between -1 and 1 and initial estimates of the (untransformed) scale parameter and variance components ranging between 0 and 5, these priors are all fairly diffuse, expressing little prior knowledge about these parameters and letting the available NSFG data provide the most information. Prior studies comparing interviewer variance components for similar count variables could also be used in general applications of this technique to specify more informative prior distributions. It is also important to note that the BUGS software uses inverse-variances for the normal distribution, meaning that 0.01 and inverses of the variance components will be specified in the normal distribution functions (example WinBUGS code used for the analyses is available in the Appendix).

Table 4.1 presents descriptive statistics for the interviewers in each of the groups defined by the eight interviewer-level characteristics. These descriptive statistics include the number of interviewers in each group (out of 87 total), and the mean, standard deviation (SD) and range for the number of cases (sample sizes) assigned to each interviewer.

Table 4.1
Descriptive statistics for the NSFG interviewers in each group defined by the eight interviewer-level characteristics

	Number of Interviewers	Total Sample Size	Mean Sample Size	SD of Sample Sizes	Range of Sample Sizes
Age (Years)					
< 54	46	5,888	128.00	113.29	(18, 554)
55+	41	5,721	139.54	132.67	(12, 532)
Experience					
< 5 Years	49	6,062	123.71	126.65	(12, 554)
5+ Years	38	5,547	145.97	116.71	(18, 507)
No. of Children					
< 2	58	7,756	133.72	113.28	(18, 532)
2+	29	3,853	132.86	140.53	(12, 554)
Ever Married					
Yes	70	9,923	141.76	129.00	(17, 554)
No	17	1,686	99.18	83.49	(12, 377)
Other Job					
No	47	5,406	115.02	95.49	(12, 532)
Yes	40	6,203	155.08	145.92	(17, 554)
College Degree					
No	37	4,528	122.38	87.97	(18, 409)
Yes	50	7,081	141.62	142.71	(12, 554)
NSFG Before					
No	15	1,155	77.00	39.17	(20, 166)
Yes	72	10,454	145.19	130.29	(12, 554)
Ethnicity					
Other	16	1,781	111.31	75.53	(20, 297)
White	71	9,828	138.42	130.35	(12, 554)

The descriptive statistics in Table 4.1 indicate substantial variance in the sizes of the samples assigned to the interviewers. A modeling approach treating interviewer effects as fixed would probably not make sense for these data, given the small sample sizes for some of the interviewers (which could lead to unstable estimates for particular interviewers). Instead, a modeling approach that borrows information across interviewers (treating interviewer effects as random) would lead to more stable estimates of means for each interviewer. We also note that for three of the observable interviewer features (Ever Married, NSFG Before, and Ethnicity), one of the two groups has less than 20 interviewers, which is not ideal for reliable estimation of variance components (Hox 1998). In light of the simulation results for smaller sample sizes (Section 3), we consider the impacts of these small sample sizes in our analyses.

Simple examinations of the distributions of the means of observed parity measures for the interviewers in each group are presented in Figure 4.1 below, to obtain an initial sense of the magnitude of interviewer variance in each of the groups. Figure 4.1 presents side-by-side box plots of the interviewer means on the parity variable for each group, with the means weighted by assigned sample sizes, along with the overall distribution of the 11,609 parity measures in the complete data set.

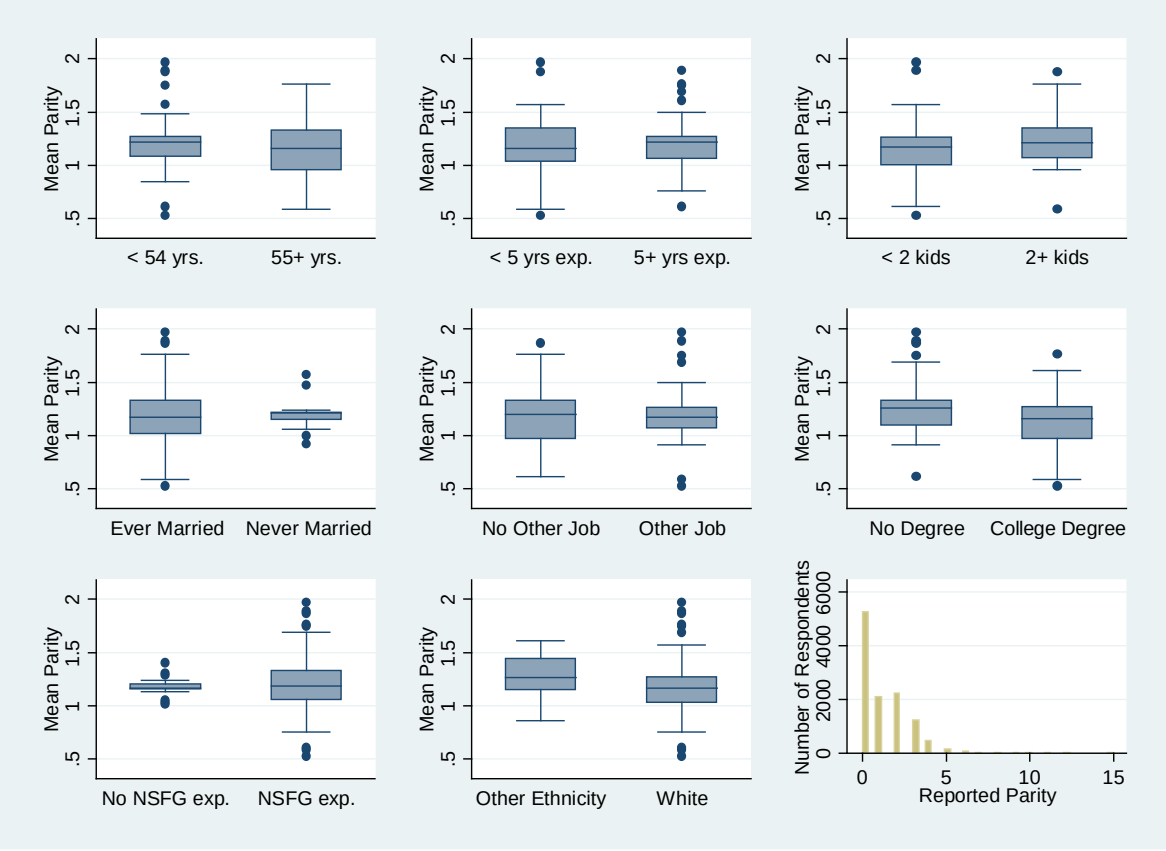

Figure 4.1 Distributions of observed means on parity for interviewers in each group, with interviewer means weighted by assigned sample size, along with the overall distribution of the reported parity measures.

The distributions of the means of measured parity values for the interviewers in Figure 4.1 provide an initial sense of groups that tend to differ in terms of the interviewer variance components. The group of interviewers that has never been married appears to have reduced variance, as does the group that has no

prior experience working on the NSFG. The box plots also suggest that the groups do not vary substantially in terms of parity means, which is reassuring (i.e., different groups of interviewers do not produce different marginal means for the estimate of interest). Finally, the distribution of observed parity values for all 11,609 respondents has the expected appearance for a variable measuring a count of relatively rare events (live births), with mean 1.19 and variance 1.99.

We next consider the distributions of the proportions of females indicating current sexual activity among the interviewers in each group (Figure 4.2).

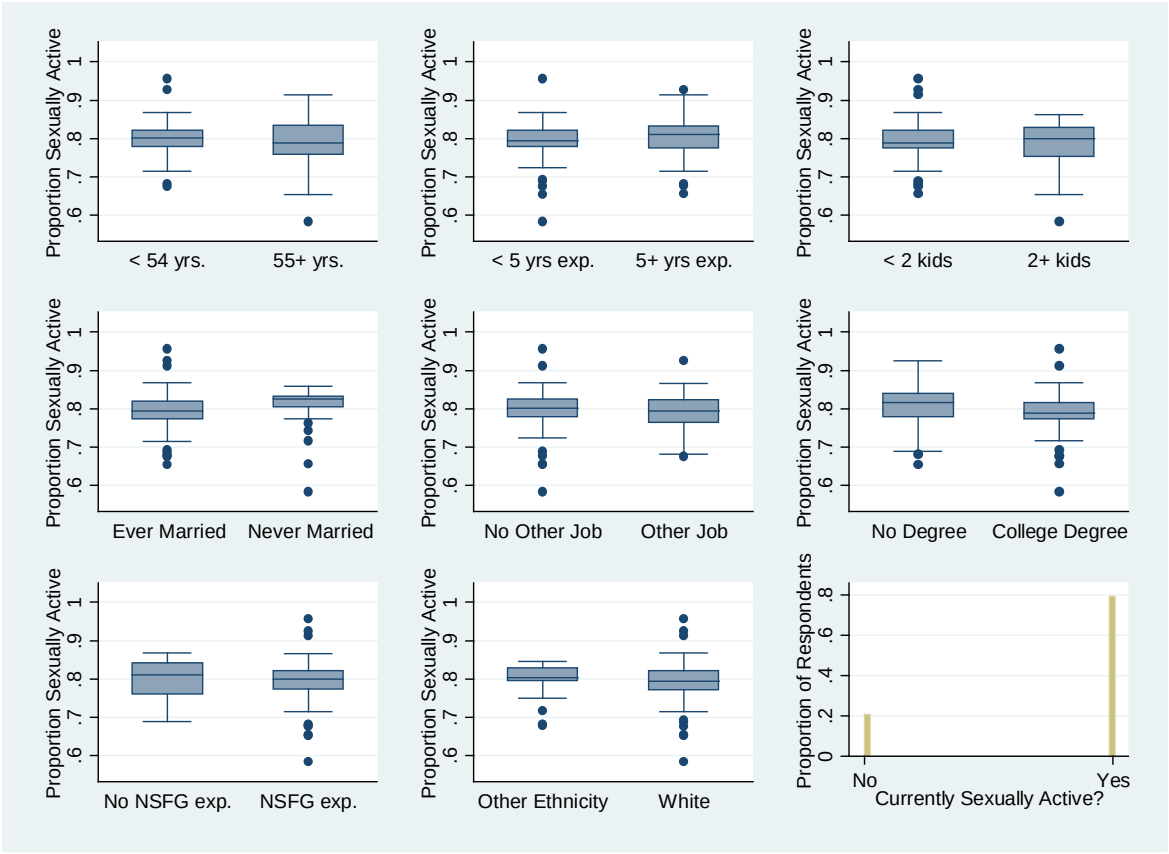

Figure 4.2 Distributions of observed proportions of female respondents indicating current sexual activity for interviewers in each group, with interviewer means weighted by assigned sample size, along with the overall distribution of the sexual activity indicator.

We see less evidence of differences in interviewer variance between the groups in general for this proportion, relative to average parity. Approximately 80% of the female respondents indicated that they were currently in a sexually active relationship.

Table 4.2 presents estimates of the parameters in each of the negative binomial models for the measured parity variable based on the two alternative analytic approaches. This table also presents results of the likelihood ratio tests comparing the two interviewer variance components (for each pair of groups) when following the frequentist approach, and 95% credible sets for the difference in the two variance components when following the Bayesian approach.

Table 4.2
Parameter estimates in the negative binomial regression models for parity and comparisons of the interviewer variance components following the alternative frequentist and Bayesian analytic approaches.

	Frequentist Approach (SAS PROC GLIMMIX)				Bayesian Approach (WinBUGS)			
Interviewer Group Variable	$\hat{\beta}_0$ **(SE)/** $\hat{\beta}_1$ **(SE)**	$\hat{\alpha}$ **(SE)**	$\hat{\tau}_1^2$ **(SE)/** $\hat{\tau}_2^2$ **(SE)**	**Likelihood Ratio Test:** $\tau_1^2 = \tau_2^2$	$\hat{\beta}_0$ **(SD)/** $\hat{\beta}_1$ **(SD)**	$\hat{\alpha}$ **(SD)**	$\hat{\tau}_1^2$ **(SD)/** $\hat{\tau}_2^2$ **(SD)**	**95% CS:** $\tau_1^2 - \tau_2^2$
Age								
(1 = <54 years,	0.185(0.031)/	0.538	0.026(0.009)/	χ_1^2 =0.03,	0.183(0.033)/	0.685	0.025(0.010)/	(-0.026,
2 = 55+ years)	-0.007(0.043)	(0.018)	0.024(0.008)	p= 0.873	-0.003(0.046)	(0.024)	0.024(0.009)	0.028)
Experience								
(1 = <5 years,	0.201(0.033)/	0.537	0.024(0.008)/	χ_1^2 =0.04,	0.197(0.034)/	0.694	0.024(0.009)/	(-0.032,
2 = 5+ years)	-0.036(0.044)	(0.018)	0.027(0.010)	p= 0.835	-0.031(0.045)	(0.027)	0.027(0.011)	0.024)
Number of Kids								
(1 = <2,	0.254(0.036)/	0.537	0.023(0.007)/	χ_1^2 =0.01,	0.253(0.038)/	0.692	0.023(0.007)/	(-0.032,
2 = 2+)	-0.109(0.044)	(0.018)	0.022(0.009)	p= 0.926	-0.109(0.045)	(0.025)	0.023(0.012)	0.024)
Ever Married								
(1 = Yes,	0.184(0.029)/	0.537	0.030(0.008)/	χ_1^2 =5.41,	0.181(0.037)/	0.694	0.030(0.008)/	(0.002,
2 = No)	-0.001(0.039)	(0.018)	0.000(N/A)*	p= 0.020	0.004(0.045)	(0.025)	0.003 (0.007)	0.048)
Other Job								
(1 = Yes,	0.186(0.031)/	0.538	0.022(0.009)/	χ_1^2 =0.15,	0.188(0.032)/	0.688	0.020(0.010)/	(-0.036,
2 = No)	-0.009(0.043)	(0.018)	0.027(0.008)	p= 0.699	-0.010(0.044)	(0.025)	0.028(0.010)	0.021)
College Degree								
(1 = Yes,	0.242(0.031)/	0.538	0.023(0.008)/	χ_1^2 < 0.01,	0.240(0.032)/	0.693	0.024(0.009)/	(-0.025,
2 = No)	-0.108(0.042)	(0.018)	0.022(0.008)	p= 0.963	-0.106(0.044)	(0.024)	0.021(0.010)	0.030)
NSFG Before								
(1 = Yes,	0.174(0.035)/	0.537	0.031(0.008)/	χ_1^2 = 8.26,	0.169(0.036)/	0.692	0.030(0.008)/	(0.006,
2 = No)	0.010(0.043)	(0.018)	0.000(N/A)*	p= 0.004	0.013(0.045)	(0.026)	0.001(0.005)	0.050)
Ethnicity								
(1 = White,	0.217(0.046)/	0.537	0.027(0.007)/	χ_1^2 =0.38,	0.220(0.051)/	0.690	0.026(0.008)/	(-0.045,
2 = Other)	-0.044(0.052)	(0.018)	0.018(0.011)	p= 0.536	-0.050(0.058)	(0.025)	0.020(0.017)	0.027)

* PROC GLIMMIX indicated that the estimated variance-covariance matrix of the random effects was not positive definite, and the estimate was set to zero because the RPL estimate of the variance component was negative. The same result occurred when using adaptive quadrature instead of RPL.

Notes: Estimates following Bayesian approach are medians of draws from posterior distributions. SE = Asymptotic SE. SD = SD of draws from posterior distribution. CS = Credible Set.

Consistent with our simulation study in Section 3, the results in Table 4.2 show that it is not uncommon for the frequentist approach to yield negative estimates of interviewer variance components (which causes SAS PROC GLIMMIX to set the estimates equal to zero, and not report estimated standard errors for the estimates), especially for groups with smaller samples of interviewers. In two cases, this results in a significant likelihood ratio test statistic, which would suggest that the two variance components are different. In contrast, the Bayesian approach produces very small estimates of the variance components, and a 95% credible set for the difference in the variance components. For example, in the cases of marital status and prior NSFG experience, we see estimates that are consistent with Figure 4.1, suggesting that there is significantly lower variance in the parity measures among the never-married group of interviewers and the inexperienced group of interviewers. The credible sets for the differences in these two cases agree with the frequentist tests, but the lower limits of these sets are very close to zero, suggesting that the differences, while significant, may not be very strong. We view this as an advantage of the Bayesian approach.

The Bayesian approach yields only slightly larger standard errors (or posterior standard deviations) for the parameter estimates in nearly all cases, reflecting the uncertainty in the parameter estimates that is accounted for by the prior distributions. The use of non-informative priors in this case, which would result in a posterior distribution that is dominated by the likelihood function, is the likely reason for the similarity in these measures of uncertainty, and more informative priors may increase the efficiency of the Bayesian estimates. Estimates of the individual parameters and corresponding inferences about them are generally quite similar when following the two approaches, as suggested by the literature in Section 2, and the estimated fixed effects suggest that the different groups of interviewers do not have a tendency to collect different measures on the parity variable. Interestingly, both approaches agree that interviewers with fewer children and/or a four-year college degree have a tendency to collect lower measures on the parity variable, but these differences could certainly be due to other covariates not accounted for in these analyses. Finally, we see slightly different estimates of the negative binomial scale parameter when following the two approaches. This is to be expected, as the Bayesian approach uses the medians of posterior distributions while the frequentist approach uses the modes of likelihood functions. In addition, the posterior distributions are not exactly equal to the likelihood functions when proper priors are utilized. The frequentist estimates of the scale parameter were much closer to the Bayesian estimates when using adaptive quadrature with five quadrature points to approximate the negative binomial likelihoods (results not shown); frequentist inferences for the other parameters did not change when using this alternative estimation method.

We repeated these analyses for the binary indicator of current sexual activity. Table 4.3 presents the estimated parameters in the multilevel logistic regression models following each of the two approaches. Consistent with Figure 4.2, these analyses reveal no evidence of differences between the various groups of interviewers in the variance components or the expected values of this outcome. Inferences were once again quite similar when following the two approaches, and the variances of the estimated variance components were once again slightly larger when following the Bayesian approach.

Table 4.3
Parameter estimates in the logistic regression models for current sexual activity and comparisons of the interviewer variance components following the alternative frequentist and Bayesian analytic approaches.

	Frequentist Approach (SAS PROC GLIMMIX)			Bayesian Approach (WinBUGS)		
Interviewer Group Variable	$\hat{\beta}_0$ **(SE)/** $\hat{\beta}_1$ **(SE)**	$\hat{\tau}_1^2$ **(SE)/** $\hat{\tau}_2^2$ **(SE)**	**Likelihood Ratio Test:** $\tau_1^2 = \tau_2^2$	$\hat{\beta}_0$ **(SD)/** $\hat{\beta}_1$ **(SD)**	$\hat{\tau}_1^2$ **(SE)/** $\hat{\tau}_2^2$ **(SE)**	**95% CS:** $\tau_1^2 - \tau_2^2$
Age						
(1 = <54 years,	1.333 (0.066) /	0.008 (0.013) /	$\chi_1^2 = 2.05$,	1.344 (0.055) /	0.009 (0.013) /	(-0.107,
2 = 55+ years)	0.032 (0.076)	0.045 (0.024)	$p = 0.153$	0.024 (0.066)	0.046 (0.028)	0.016)
Experience						
(1 = <5 years,	1.378 (0.064) /	0.004 (0.017) /	$\chi_1^2 = 1.52$,	1.384 (0.051) /	0.005 (0.017) /	(-0.087,
2 = 5+ years)	-0.050 (0.073)	0.037 (0.020)	$p = 0.217$	-0.061 (0.064)	0.039 (0.023)	0.024)
Number of Kids						
(1 = <2,	1.362 (0.080) /	0.022 (0.015) /	$\chi_1^2 = 0.16$,	1.363 (0.059) /	0.024 (0.016) /	(-0.094,
2 = 2+)	-0.015 (0.088)	0.033 (0.024)	$p = 0.689$	-0.012 (0.070)	0.037 (0.030)	0.037)
Ever Married						
(1 = Yes,	1.387 (0.130) /	0.020 (0.012) /	$\chi_1^2 = 0.58$,	1.398 (0.090) /	0.021 (0.013) /	(-0.180,
2 = No)	-0.045 (0.134)	0.048 (0.041)	$p = 0.447$	-0.053 (0.097)	0.051 (0.055)	0.035)
Other Job						
(1 = Yes,	1.374 (0.043) /	0.026 (0.016) /	$\chi_1^2 = 0.01$,	1.381 (0.045) /	0.029 (0.019) /	(-0.055,
2 = No)	-0.046 (0.072)	0.024 (0.020)	$p = 0.927$	-0.051 (0.065)	0.022 (0.022)	0.063)
College Degree						
(1 = Yes,	1.388 (0.051) /	0.016 (0.014) /	$\chi_1^2 = 0.60$,	1.394 (0.052) /	0.014 (0.016) /	(-0.079,
2 = No)	-0.063 (0.071)	0.035 (0.022)	$p = 0.439$	-0.072 (0.064)	0.038 (0.024)	0.033)
NSFG Before						
(1 = Yes,	1.363 (0.103) /	0.020 (0.012) /	$\chi_1^2 = 1.20$,	1.381 (0.113) /	0.021 (0.013) /	(-0.301,
2 = No)	-0.012 (0.111)	0.069 (0.055)	$p = 0.273$	-0.024 (0.118)	0.083 (0.084)	0.019)
Ethnicity						
(1 = White,	1.354 (0.077) /	0.024 (0.014) /	$\chi_1^2 = 0.05$,	1.365 (0.080) /	0.025 (0.015) /	(-0.131,
2 = Other)	-0.004 (0.088)	0.032 (0.031)	$p = 0.816$	-0.013 (0.088)	0.032 (0.044)	0.044)

Notes: Estimates following Bayesian approach are medians of draws from posterior distributions. SE = Asymptotic SE. SD = SD of draws from posterior distribution. CS = Credible Set.

5 Concluding Remarks

This paper has considered frequentist and Bayesian methods for comparing the interviewer variance components for non-normally distributed survey items between two independent groups of survey interviewers. The methods are based on a flexible class of hierarchical generalized linear models (HGLMs) that allow the variance components for two mutually exclusive groups of interviewers to vary, and alternative inferential approaches based on those models. Results from a simulation study suggest that the two approaches have little empirical bias, comparable empirical MSE values and good coverage for moderate-to-large samples of interviewers and respondents. Analyses of real data from the U.S. National Survey of Family Growth (NSFG) suggest that inferences based on the two approaches tend to be quite similar. We find the similar performance of these two approaches to be good news for survey researchers, in that frequentists and Bayesians alike have tools available to them for analyzing this problem that will lead to similar conclusions.

There are some subtle distinctions between the two approaches that emerged in the analyses, mainly related to sample sizes and estimates of variance components that are extremely small or equal to zero.

These issues warrant further discussion, given their implications for survey practice. The Bayesian approach illustrated here is capable of accommodating uncertainty in the estimation of variance components when forming credible sets and does not rely on asymptotic theory, but we found that inferences about differences in variance components between a number of different subgroups of NSFG interviewers (each of moderate size) did not vary from those that would be made using frequentist approaches. Whether or not we would see the same results for even smaller groups of interviewers requires future investigation; the simulation study presented in Section 3 suggested that neither method performs well in a context where two groups of 20 interviewers collect data from 10 respondents each. An initial application of these two methods to data from the first quarter of data collection in this cycle of the NSFG (with about 20 interviewers in each of two groups interviewing about 20 respondents each on average) yielded findings similar to those reported here for larger samples, with some evidence of the Bayesian approach being more conservative (West 2011).

In general, the Bayesian approach provides a more natural form of inference for this problem, indicating a range of values for the difference in which approximately 95% of differences will fall. This may appeal to certain consumers of a given survey's products, as opposed to the simple p-value for a likelihood ratio test, which does not give users a sense of the range of possible differences. In the frequentist setting, the likelihood ratio test may be the only method of inference available if the pseudo maximum likelihood point estimate for one or more of the variance components is zero, with no corresponding standard error (preventing computation of Wald-type intervals). This situation was observed in both the simulations and the NSFG analyses, especially for groups with smaller samples of interviewers; given the reliance of likelihood ratio tests on asymptotic theory, the Bayesian approach may be better a better choice for smaller samples. The performance of the Bayesian approach is not ideal, however, for very small samples, as illustrated in the simulation study in Section 3.

We noted two significant differences between subgroups of interviewers in the NSFG data, and in each of these cases, the group with the smaller variance had an estimated variance component set to zero (with no standard error computed) when using the frequentist approach. The resulting inferences based on these estimates (where likelihood values were computed using the estimates of zero for the subgroups in question when performing the likelihood ratio tests) agreed with the Bayesian approach. We remind readers using frequentist methods that small samples of interviewers or extremely small amounts of variance among interviewers for particular variables may lead to negative maximum likelihood estimates of variance components, which can be problematic for the interpretation of interviewer variance for individual groups. Some software procedures capable of fitting multilevel models (e.g., the gllamm procedure in Stata, or the lmer() function in R) constrain variance components to be greater than zero during estimation to prevent this problem, which can increase estimation times. Other software procedures (like GLIMMIX in SAS) will simply fix these negative estimates to be zero, and fail to compute an estimated standard error. While these variance components technically cannot be equal to zero, we suggest interpreting these findings as evidence that there in negligible variance among the interviewers in a particular group. Bates (2009) argues against the use of standard errors for making inferences about variance components in the frequentist setting, especially when variance components are close to zero, instead suggesting that the profiled deviance function should be used to visualize the precision of the estimates. Both this approach and the Wald approach to computing confidence intervals will still be limited by smaller samples.

We do not see an empirical problem with using these zero estimates to perform the likelihood ratio tests demonstrated here for comparing groups of interviewers, given that Bayesian draws of the variance components in these groups would also be very small. However, in the case of estimating interviewer variance for single groups, examination of the sensitivity of Bayesian inferences to choices of different prior distributions for the variance components should be performed when variance components close to zero are expected, or the number of interviewers is relatively small (Browne and Draper 2006; Lambert et al. 2005). Furthermore, if survey researchers are interested in *predicting* random interviewer effects in the case where interviewer variance components are expected to be close to zero, both frequentist and Bayesian methods perform very poorly, and prediction is not recommended in this case (Singh et al. 1998, p. 390). See Savalei and Kolenikov (2008) for more discussion of the zero variance issue.

This study was certainly not without limitations. We acknowledge that the design of the NSFG, where interviewers are typically assigned to work in a single primary sampling area, did not allow for interpenetrated assignment of sampled cases to interviewers. As a result, disentangling interviewer effects from effects of the primary sampling areas is difficult. The methodologies illustrated in this paper can easily incorporate additional interviewer- or area-level covariates in an effort to "explain" variance among interviewers or areas due to observable covariates. The question of how to estimate interviewer variance in the presence of a strictly non-interpenetrated sample design needs more research in general, and we did not address this open question in this paper. As mentioned in Section 1, interpenetrated sample designs have been used in recent studies to disentangle interviewer and area effects. Future studies should examine the ability of the two approaches reviewed in this paper to detect differences in interviewer variance components when using cross-classified multilevel models that also include the effects of areas in an interpenetrated sample design.

On a similar note, we did not account for any of the complex sampling features of the NSFG (i.e., weighting or stratified cluster sampling) in the analyses. The theory that underlies the estimation of parameters in multilevel models in the presence of survey weights calls for weights for both the respondents and the higher-level clusters, which in this case would be interviewers (Rabe-Hesketh and Skrondal 2006; Pfefferman, Skinner, Holmes, Goldstein and Rasbash 1998). The analyses presented here effectively assume that we have a sample of interviewers from some larger population that was selected with equal probability, and that all respondents within each interviewer had equal weight. Methods outlined by Gabler and Lahiri (2009) might prove useful for addressing this limitation, and analysts could also include fixed effects of survey weights or stratification codes in the models proposed here. We leave these extensions for future research.

Finally, this paper also did not consider another rich aspect of the Bayesian approach, in that posterior draws of the 87 random interviewer effects in the models were also generated by the BUGS Gibbs sampling algorithm. These draws would enable survey managers to make inferences about the effects specific interviewers are having on particular survey measures. Consistent and regular updating of these posterior distributions as data collection progresses would enable survey managers to intervene when the posterior distributions for particular interviewers suggest that these interviewers are having non-zero effects on the survey measures.

Acknowledgements

The authors are grateful for support from a contract with the National Center for Health Statistics that enabled the seventh cycle of the National Survey of Family Growth (contract 200-2000-07001).

Appendix

A.1 Example Code

We provide example code for fitting the types of models discussed in the paper using SAS PROC GLIMMIX below. In this code, PARITY and SEXMAIN are the count and binary variables, respectively, measured on NSFG respondents, FINAL_INT_ID is a final interviewer ID code, and INT_NVMARRIED is an indicator variable for whether or not an interviewer has never been married. The ASYCOV option will print asymptotic estimates of the variances and covariances of the estimated variance components.

```
/* marital status */

proc glimmix data = bayes.final_analysis asycov;
   class final_int_id int_nvmarried;
   model parity = int_nvmarried / dist = negbin link = log solution cl;
   random int / subject = final_int_id group = int_nvmarried;
   covtest homogeneity / cl (type = plr);
   nloptions tech=nrridg;
run;

proc glimmix data = bayes.final_analysis asycov;
   class final_int_id int_nvmarried;
   model sexmain (event = "1") = int_nvmarried / dist = binary link = logit
solution cl;
   random int / subject = final_int_id group = int_nvmarried;
   covtest homogeneity / cl (type = plr);
   nloptions tech=nrridg;
run;
```

We also provide example WinBUGS code for fitting the models using the Bayesian approaches discussed below. We call the WinBUGS code from the R software. In this code, LOWAGE.G is an interviewer-level indicator (with 87 values) for being in the younger interviewer age group, and HIGHAGE.G is an indicator for being in the older group. The full code, including code creating the variables used below, is available from the authors upon request.

```
# load necessary packages for using BUGS from R

library(arm)
library(R2WinBUGS)

############## Parity Analyses
```

```
# BUGS file for Age Group and Parity (age_nb.bug)

model {
   for (i in 1:n){
      parity[i] ~ dpois(lambda[i])
      lambda[i] <- rho[i]*mu[i]
      log(mu[i]) <- b0[intid[i]]
      rho[i]~dgamma(alpha,alpha)
   }

   for (j in 1:J){
      b0[j] ~ dnorm(b0.hat[j], tau.b0[highage.g[j]+1])
      b0.hat[j] <- beta0 + beta1*lowage.g[j]
   }

   beta0 ~ dnorm(0,0.01)
   beta1 ~ dnorm(0,0.01)
   alpha <- exp(logalpha)
   logalpha ~ dnorm(0,0.01)

   for (k in 1:2){
      tau.b0[k] <- pow(sigma.b0[k], -2)
      sigma.b0[k] ~ dunif(0,10)
   }
}

# Simulations for Parity/Age Group model in BUGS

n <- length(parity)
J <- 87
age.data <- list("n", "J", "parity", "intid", "highage.g", "lowage.g")
age.inits <- function(){
   list (b0=rnorm(J), beta0=rnorm(1), beta1=rnorm(1), sigma.b0=runif(2),
logalpha=rnorm(1))}
age.parameters <- c("b0", "beta0", "beta1", "sigma.b0", "alpha")
age.1 <- bugs(age.data, age.inits, age.parameters, "age_nb.bug", n.chains = 3,
n.iter=5000, debug=TRUE,
bugs.directory="C:/Users/bwest/Desktop/winbugs14/WinBUGS14")

attach.bugs(age.1)

# for tables of results and inference

resultsmat <- cbind(numeric(6),numeric(6),numeric(6),numeric(6))

resultsmat[1,1] <- quantile(beta0,0.5)
resultsmat[1,2] <- sd(beta0)
resultsmat[1,3] <- quantile(beta0,0.025)
resultsmat[1,4] <- quantile(beta0,0.975)

resultsmat[2,1] <- quantile(beta1,0.5)
resultsmat[2,2] <- sd(beta1)
resultsmat[2,3] <- quantile(beta1,0.025)
resultsmat[2,4] <- quantile(beta1,0.975)

resultsmat[3,1] <- quantile(sigma.b0[,1]^2,0.5)
resultsmat[3,2] <- sd(sigma.b0[,1]^2)
```

```
resultsmat[3,3] <- quantile(sigma.b0[,1]^2,0.025)
resultsmat[3,4] <- quantile(sigma.b0[,1]^2,0.975)

resultsmat[4,1] <- quantile(sigma.b0[,2]^2,0.5)
resultsmat[4,2] <- sd(sigma.b0[,2]^2)
resultsmat[4,3] <- quantile(sigma.b0[,2]^2,0.025)
resultsmat[4,4] <- quantile(sigma.b0[,2]^2,0.975)

resultsmat[5,1] <- quantile(1/alpha,0.5)
resultsmat[5,2] <- sd(1/alpha)
resultsmat[5,3] <- quantile(1/alpha,0.025)
resultsmat[5,4] <- quantile(1/alpha,0.975)

vardiff <- sigma.b0[,1]^2 - sigma.b0[,2]^2
resultsmat[6,1] <- quantile(vardiff,0.5)
resultsmat[6,2] <- sd(vardiff)
resultsmat[6,3] <- quantile(vardiff,0.025)
resultsmat[6,4] <- quantile(vardiff,0.975)

resultsmat

############## Current Sexual Activity Analyses

# BUGS file for Age Group and Sexual Activity (age_bin.bug)

model {
   for (i in 1:n){
      sexmain[i] ~ dbern(p[i])
      logit(p[i]) <- b0[intid[i]]
   }

   for (j in 1:J){
      b0[j] ~ dnorm(b0.hat[j], tau.b0[highage.g[j]+1])
      b0.hat[j] <- beta0 + beta1*lowage.g[j]
   }
   beta0 ~ dnorm(0,0.01)
   beta1 ~ dnorm(0,0.01)

   for (k in 1:2){
      tau.b0[k] <- pow(sigma.b0[k], -2)
      sigma.b0[k] ~ dunif(0,10)
   }
}

# Simulations for Parity/Age Group model in BUGS

n <- length(sexmain)
J <- 87
age.data <- list("n", "J", "sexmain", "intid", "highage.g", "lowage.g")
age.inits <- function(){
   list (b0=rnorm(J), beta0=rnorm(1), beta1=rnorm(1), sigma.b0=runif(2))}
age.parameters <- c("b0", "beta0", "beta1", "sigma.b0")
age.1 <- bugs(age.data, age.inits, age.parameters, "age_bin.bug", n.chains =
3, n.iter=5000, debug=TRUE,
bugs.directory="C:/Users/bwest/Desktop/winbugs14/WinBUGS14")

attach.bugs(age.1)
```

References

Bates, D. (2009). Assessing the precision of estimates of variance components. *Presentation to the Max Planck Institute for Ornithology*, Seewiesen, July 21, 2009. Presentation can be downloaded from http://lme4.r-forge.r-project.org/slides/2009-07-21-Seewiesen/4PrecisionD.pdf.

Biemer, P.P. and Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. Chapter 27 of *Survey Measurement and Process Quality*, Editors Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz and Trewin. Wiley-Interscience, 603-632.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Browne, W.J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3), 473-514.

BUGS, http://www.mrc-bsu.cam.ac.uk/bugs/welcome.html.

Carlin, B.P. and Louis, T.A. (2009). *Bayesian Methods for Data Analysis*. Chapman and Hall / CRC Press.

Chaloner, K. (1987). A Bayesian approach to the estimation of variance components for the unbalanced one-way random model. *Technometrics*, 29(3), 323-337.

Collins, M. and Butcher, B. (1982). Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, 25, 39-58.

Durham, C.A., Pardoe, I. and Vega, E. (2004). A methodology for evaluating how product characteristics impact choice in retail settings with many zero observations: An application to restaurant wine purchase. *Journal of Agricultural and Resource Economics*, 29(1), 112-131.

Durrant, G.B., Groves, R.M., Staetsky, L. and Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*, 74, 1-36.

Faraway, J.J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall / CRC Press: Boca Raton, FL.

Farrell, P.J. (2000). Bayesian inference for small area proportions. Sankhya: The *Indian Journal of Statistics*, Series B (1960-2002), 62(3), 402-416.

Fowler, F.J. and Mangione, T.W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park, CA: Sage.

Gabler, S. and Lahiri, P. (2009). On the definition and interpretation of interviewer variability for a complex sampling design. *Survey Methodology*, 35(1), 85-99.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Anaylsis*. Chapman and Hall / CRC Press.

Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel / Hierarchical Models*. Cambridge University Press.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-511.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41, 337-348.

Goldstein, H. (1995). *Multilevel Statistical Models, Second Edition*. Kendall's Library of Statistics 3, Edward Arnold.

Groves, R.M. (2004). *Survey Errors and Survey Costs (2nd Edition)*. In chapter 8: The Interviewer as a Source of Survey Measurement Error. Wiley-Interscience.

Groves, R.M., Mosher, W.D., Lepkowski, J.M. and Kirgis, N.G. (2009). Planning and development of the continuous National Survey of Family Growth. National Center for Health Care Statistics. *Vital Health Statistics*, 1(48).

Hilbe, J.M. (2007). *Negative Binomial Regression*. Cambridge University Press.

Hox, J. (1998). *Multilevel Modeling: When and Why*. In I. Balderjahn, R. Mathar and M. Schader (Eds.). Classification, data anaylsis, and data highways. New York: Springer-Verlag, 147-154.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Lambert, P.C., Sutton, A.J., Burton, P.R., Abrams, K.R. and Jones, D.R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15), 2401-2428.

Lepkowski, J.M., Mosher, W.D., Davis, K.E., Groves, R.M. and Van Hoewyk, J. (2010). The 2006-2010 National Survey of Family Growth: sample design and analysis of a continuous survey. National Center for Health Statistics, *Vital and Health Statistics*, 2(150), June 2010.

Lynn, P., Kaminska, O. and Goldstein, H. (2011). Panel attrition: how important is it to keep the same interviewer? *ISER Working Paper Series*, Paper 2011-02.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mangione, T.W., Fowler, F.J. and Louis, T.A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics*, 8(3), 293-307.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, Berlin.

O'Muircheartaigh, C. and Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society*, Series A, 161 (1), 63-77.

O'Muircheartaigh, C. and Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, 162(3), 437-446.

Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60(1), 23-40.

Pinheiro, J.C. and Chao, E.C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15, 58-81.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society*, Series A, 169, 805-827.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA.

Rodriguez, G. and Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society, Series A*, 164(2), 339-355.

SAS Institute, Inc. (2010). Online Documentation for the GLIMMIX Procedure.

Savalei, V. and Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, 13(2), 150-170.

Schaeffer, N.C., Dykema, J. and Maynard, D.W. (2010). *Handbook of Survey Research, Second Edition*. In Interviewers and Interviewing. J.D. Wright and P.V. Marsden (eds). Bingley, U.K.: Emerald Group Publishing Limited.

Schnell, R. and Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21(3), 389-410.

Schober, M. and Conrad, F. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602.

Singh, A.C., Stukel, D.M. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, Series B*, 60(2), 377-396.

Ugarte, M.D., Goicoa, T. and Militino, A.F. (2009). Empirical bayes and fully bayes procedures to detect high-risk areas in disease mapping. *Computational Statistics and Data Analysis*, 53, 2938-2949.

Van Tassell, C.P. and Van Vleck, L.D. (1996). Multiple-trait Gibbs sampler for animal models: Flexible programs for Bayesian and likelihood-based (co)variance component inference. *Journal of Animal Science*, 74, 2586-2597.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26, 37-52.

West, B.T. (2011). Bayesian analysis of between-group differences in variance components in hierarchical generalized linear models. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: *American Statistical Association*, 1828-1842.

West, B.T. and Galecki, A.T. (2011). An overview of current software procedures for fitting linear mixed models. *The American Statistician*, 65(4), 274-282.

West, B.T., Kreuter, F. and Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? *Journal of Official Statistics*, 29(2), 277-297.

West, B.T. and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74(5), 1004-1026.

Zhang, D. and Lin, X. (2010). Variance component testing in generalized linear mixed models for longitudinal / clustered data and other related topics. Random Effect and Latent Variable Model Selection. *Springer Lecture Notes in Statistics*, Volume 192.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 189-209
Statistics Canada, Catalogue No. 12-001-X

Bagging non-differentiable estimators in complex surveys

Jianqiang C. Wang, Jean D. Opsomer and Haonan Wang[1]

Abstract

Bagging is a powerful computational method used to improve the performance of inefficient estimators. This article is a first exploration of the use of bagging in survey estimation, and we investigate the effects of bagging on non-differentiable survey estimators including sample distribution functions and quantiles, among others. The theoretical properties of bagged survey estimators are investigated under both design-based and model-based regimes. In particular, we show the design consistency of the bagged estimators, and obtain the asymptotic normality of the estimators in the model-based context. The article describes how implementation of bagging for survey estimators can take advantage of replicates developed for survey variance estimation, providing an easy way for practitioners to apply bagging in existing surveys. A major remaining challenge in implementing bagging in the survey context is variance estimation for the bagged estimators themselves, and we explore two possible variance estimation approaches. Simulation experiments reveal the improvement of the proposed bagging estimator relative to the original estimator and compare the two variance estimation approaches.

Key Words: Bootstrap; Distribution function; Quantile estimation.

1 Introduction

Bagging, short for "bootstrap aggregating", is a resampling method originally introduced to improve "weak" learning algorithms. Bagging was proposed by Breiman (1996), who heuristically demonstrated how it improved the performance of tree-based predictors. Since then, bagging has been applied to a wide range of settings and analyzed by many authors. Bühlmann and Yu (2002) showed the smoothing effect of bagging and its variations on hard-decision classification algorithms, and formalized the notion of "unstable predictors". Chen and Hall (2003) derived theoretical results on bagging estimators defined by estimating equations. Buja and Stuetzle (2006) considered bagging U-statistics, and claimed that bagging "often but not always decreases variance, whereas it always increases bias". Friedman and Hall (2007) examined the impact of bagging on nonlinear estimators. More recently, Hall and Robinson (2009) discussed the effects of bagging on cross-validation choice of smoothing parameters, and presented intriguing results on improving the order of the cross-validation selected kernel bandwidth by bagging.

The aforementioned literature studied the effects of bagging on various estimators, especially nonlinear, non-differentiable estimators, under the *iid* (independent and identically distributed) sampling assumption. For dependent data, Lee and Yang (2006); Inoue and Kilian (2008) studied the effects of bagging on economic time series. The former authors studied the bagging effect on non-differentiable predictors like sign functions and quantiles, and the latter focused on bagging pretest predictors with application to U.S. consumer price inflation forecasting.

As this brief literature review shows, bagging is a promising method used to improve the efficiency of estimators. To date, however, bagging for survey estimators has not been considered. This article is a first exploration of the use of bagging in the survey context, including an evaluation of the potential efficiency gain, a number of theoretical results, and a discussion of implementation and variance estimation issues.

1. Jianqiang C. Wang, Hewlett-Packard Labs, Palo Alto, CA 94304. Jean D. Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523. E-mail: jopsomer@stat.colostate.edu; Haonan Wang, Department of Statistics, Colorado State University, Fort Collins, CO 80523.

Corresponding with general survey practice, we will only consider estimators that can be written as functions of Horvitz-Thompson (HT) estimators. More specifically, we will consider the following three types of estimators. Firstly, many commonly used survey estimators can be written as differentiable functions of HT estimators. For instance, the Hajek estimator, ratio estimator, generalized regression estimator can all be regarded as differentiable functions of HT estimators. Secondly, there are other survey estimators that are non-differentiable, including the Dunstan and Chambers estimator (Dunstan and Chambers 1986), the Rao-Kovar-Mantel estimator (Rao, Kovar, and Mantel 1990), the endogenous post-stratification estimator (Breidt and Opsomer 2008), and estimators of low-income proportion (Berger and Skinner 2003), among others. Thirdly, other estimators are only defined as solutions to weighted estimating equations. For more information on estimating equations in the survey context, see Godambe and Thompson (2009); Fuller (2009), and references therein.

While bagging can be considered a type of replication method, it is quite different from bootstrapping and other replication methods that are designed for variance estimation. Unlike these other methods, bagging is introduced to improve the actual estimator itself. The bagging method can be naturally embedded in large-scale complex surveys, since we can take advantage of replication weights that are readily available in many practical surveys. In this paper, we will show how replicates created for bootstrap variance estimation can be modified and used in bagging the original estimator. Unfortunately, one difficulty in implementing bagging in surveys is the lack of a design-based variance estimator. We will discuss a number of proposals on how to estimate the variance of bagged survey estimators, but further work is still required in this area.

The remainder of this paper is organized as follows. We define our target survey estimators and introduce the bagged version of each estimator in Section 2. We then present the theoretical properties of the bagged estimators in Section 3. Section 4 shows how to use survey replicates to implement bagged versions of estimators, and addresses variance estimation for the resulting bagged estimators. We report on simulation results in Section 5, and conclude the paper with some final remarks in Section 6.

2 Bagging survey estimators

2.1 General approach

In this section, we discuss the implementation of bagging in the context of survey estimation. We first introduce necessary notation. Let U represent a finite population of size N, in which each element $i \in U$ is associated with a vector of measurements, $\mathbf{y}_i$, in the q-dimensional Euclidean space $\mathbb{R}^q$. The sampling design $p(\ )$ is used to draw a random sample $A \subseteq U$ of sample size n. We denote by $\mathcal{Y} = \{\mathbf{y}_i \mid i \in A\}$ the collection of sample observations. Here, the sampling design could be simple random sampling without replacement (SRSWOR), Poisson sampling or a complex design with stratification and/or multiple stages. Under each design, the probability of an element i being included in the sample is denoted by π_i.

The population mean of the measurement vector $\mathbf{y}$ is denoted by $\boldsymbol{\mu}$. It can be estimated by the Horvitz-Thompson (HT) estimator defined as

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i \in A}\frac{\mathbf{y}_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{y}_i}{\pi_i} I_i, \tag{2.1}$$

where I_i is the **sample membership indicator** for the i-th element. More generally, let θ denote a population quantity of interest, and $\hat{\theta}(\mathcal{Y})$ is the estimator of θ based on the sample observations $\mathcal{Y}$. The estimator $\hat{\theta}(\mathcal{Y})$ will be abbreviated as $\hat{\theta}$ when there is no confusion. As noted in the previous section, we assume that $\hat{\theta}$ can be written as a function of simpler estimators of the form (2.1).

In its most general form, the bagging algorithm for survey estimation is as follows:

1. For $b = 1, 2, \ldots, B$:
   a. Draw resample A_b from the random sample A, and denote the observations in the resample as $\mathcal{Y}_b^* = \{\mathbf{y}_i \mid i \in A_b\}$.
   b. Calculate the parameter estimate based on the resample A_b, denoted by $\hat{\theta}(\mathcal{Y}_b^*)$.

2. Average over the replicated estimates $\hat{\theta}(\mathcal{Y}_1^*)$, $\hat{\theta}(\mathcal{Y}_2^*)$, …, $\hat{\theta}(\mathcal{Y}_B^*)$ to obtain the bagged survey estimator,

$$\hat{\theta}_{bag} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(\mathcal{Y}_b^*). \tag{2.2}$$

In the bagging literature, the resamples A_b are often referred to as **bootstrap samples** (Breiman 1996), and we will do the same here despite the fact that we will not use them for variance estimation.

In the algorithm, the bootstrap samples could be drawn according to the sampling design rather than the empirical distribution of the sample observations, which is more commonly used in the ordinary bagging literature (Breiman 1996) and equivalent to simple random sampling (with or without replacement). For example, if the sample A is drawn using stratified or cluster sampling, such design scheme could be taken into consideration when selecting the resamples. More generally in the survey context, step 1 of the proposed bagging algorithm can be treated in the framework of two-phase sampling: the first phase corresponds to the original sample A and the second phase to the resample A_b. Thus the classical expansion estimator for two-phase designs Särndal, Swensson and Wretman (1997) is implemented in calculating the replicated estimator $\hat{\theta}(\mathcal{Y}_b^*)$. In the resample A_b, the pseudo inclusion probability for the i-th element is $\pi_i^* = \pi_i \pi_{i|A}$ where $\pi_{i|A} = \Pr(i \in A_b \mid i \in A)$ is the inclusion probability of the i-th element in resample A_b given that it is included in sample A. Hence, the bagged estimator is an approximation to the expectation of the two-phase estimator with respect to the second sampling phase, which is also referred to as **bootstrap expectation** in ordinary bagging methods (Bühlmann and Yu 2002). Although a general design for the bootstrap samples is possible, in the theoretical portions of this article we will restrict ourselves to SRSWOR. To broaden the scope of our discussion, in the variance estimation and numerical section, we introduce the case in which the bootstrap samples are drawn by stratified SRSWOR with the same strata as the original sample A, which is a useful and realistic extension.

As an example, we consider the HT estimator as defined in (2.1). The bootstrap resampling from the realized sample A is drawn under SRSWOR of size k. Under this resampling scheme, the replicated sample estimator is defined as

$$\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)=\frac{1}{N}\sum_{i\in A_b}\frac{\mathbf{y}_i}{\pi_i^*}, \tag{2.3}$$

where the pseudo inclusion probability $\pi_i^*=\pi_i\pi_{i|A}=k\pi_i/n$. Then the bagged version of the classical π^*-estimator can be calculated using (2.2). Straightforward calculation shows that the bagged estimator is identical to the original HT estimator if all SRSWOR samples of size k are enumerated in calculating (2.2). The same result holds for any other linear survey estimator. In general, the calculation of the bagged estimator $\hat{\theta}_{bag}$ is not as easy. In the rest of this section, we will focus on such calculations for the three types of nonlinear survey estimators discussed in Section 1.

2.2 Bagging differentiable survey estimators

For the survey estimators that are differentiable functions of HT estimators, the population quantity of interest can also be written as a differentiable function of population means; that is, $\theta_d=m(\boldsymbol{\mu})$, where $m(\cdot)$ is a known differentiable function. The subscript "d" stands for **differentiable** in contrast to **non-differentiable** (θ_{nd}) and **estimating equation** (θ_{ee}) coming later. A direct plug-in estimator of θ_d, based on sample observations $\mathcal{Y}$, can be written as

$$\hat{\theta}_d=m(\hat{\boldsymbol{\mu}}), \tag{2.4}$$

where $\hat{\boldsymbol{\mu}}$ is defined in (2.1). Thus, the replicated sample version of $\hat{\theta}_d$ can be expressed as

$$\hat{\theta}_d\left(\mathcal{Y}_b^*\right)=m\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)\right),$$

where $\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)$ is defined by (2.3). Then the bagged estimator of θ_d, denoted by $\hat{\theta}_{d,bag}$, is defined using (2.2).

2.3 Bagging explicitly defined non-differentiable estimators

As an example of this type of estimators, consider the estimation of the proportion of households with income below the poverty line for a population. Such quantity can be written as $(1/N)\sum_{i=1}^{N}I(y_i\le\lambda_N)$, where y_i is the income value for the i-th household in the population, and λ_N is the population poverty line. It can be seen that this quantity of interest is the mean of indicator kernel functions, and the kernel function is non-differentiable with respect to λ_N. Here, we consider a more general class in which the kernel is an arbitrary non-differentiable but bounded function. This type of population quantity can be expressed as

$$\theta_{nd}=\frac{1}{N}\sum_{i=1}^{N}h(\mathbf{y}_i-\boldsymbol{\lambda}_N),$$

where $\boldsymbol{\lambda}_N$ is an unknown population parameter, for example, the mean, a quantile or other population quantity, and $h(\mathbf{y}-\boldsymbol{\lambda}):\mathbb{R}^p\to\mathbb{R}$ is a non-differentiable function of $\boldsymbol{\lambda}$. The population quantity θ_{nd} generalizes the notion of the proportion below an estimated level and resembles the general form of a U-statistic.

Wang and Opsomer (2011) studied a class of U-statistics-like estimators, namely, non-differentiable survey estimators,

$$\hat{\theta}_{nd} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right), \tag{2.5}$$

where $\hat{\boldsymbol{\lambda}}$ is a design-based estimator of $\boldsymbol{\lambda}_N$. In the non-survey context, estimators of this type are regarded as "non-differentiable functions of the empirical distribution" (Bickel, Götze and van Zwet 1997). The study of appropriate bootstrap procedures for such estimators was carried out by Beran and Srivastava (1985) and Dümbgen (1993), among others. We define the replicated version of $\hat{\theta}_{nd}$ based on resample A_b as

$$\hat{\theta}_{nd}\left(\mathcal{Y}_b^*\right) = \frac{1}{N}\sum_{i\in A_b}\frac{1}{\pi_i^*}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right),$$

where $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ solely depends on the bootstrap resample A_b, and the bagged estimator is then defined by averaging replicated estimators. Suppose that the resampling process is SRSWOR of size k, and every subsample is selected in calculating the bagging estimator, then the bagging estimator takes the following form after manipulation,

$$\hat{\theta}_{nd,bag} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i\binom{n-1}{k-1}}\sum_{A_b\ni i}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right), \tag{2.6}$$

which replaces $h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right)$ in (2.5) by a "smoothed" quantity $\sum_{A_b\ni i}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\Big/\binom{n-1}{k-1}$, by averaging the "jumps" in the estimator. Very often, variance reduction can be achieved by this replacement. The summand $\sum_{A_b\ni i}h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\Big/\binom{n-1}{k-1}$ is the **bootstrap expectation** of $h\left(\mathbf{y}_i - \cdot\right)$ and can be approximated using the convolution of $h\left(\mathbf{y}_i - \cdot\right)$ with the sampling distribution of $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$. Study of the theoretical aspects of $\hat{\theta}_{nd,bag}$ is deferred until Section 3.

2.4 Bagging estimators defined by non-differentiable estimating equations

Finally, we explain how to bag estimators defined by non-differentiable estimating equations. For ease of presentation, we consider a one-dimensional parameter of interest. The population parameter θ_{ee} of interest is defined as

$$\theta_{ee} = \inf\left\{\gamma : S(\gamma) \geq 0\right\},$$

where

$$S(\gamma) = \frac{1}{N}\sum_{i=1}^{N}\psi\left(y_i - \gamma\right),$$

and $\psi(\cdot)$ is a non-differentiable real function. We can estimate the population parameter θ_{ee} by $\hat{\theta}_{ee}$, where

$$\hat{\theta}_{ee} = \inf\left\{\gamma : \hat{S}(\gamma) \geq 0\right\}$$

with

$$\hat{S}(\gamma) = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\psi(y_i - \gamma).$$

A frequently encountered estimator of this type is the sample quantile defined by inverting the sample cumulative distribution function (Francisco and Fuller 1991), where $\psi(y_i - \gamma) = I_{(y_i \leq \gamma)} - \alpha$ for the α-quantile.

Conceptually, there are two versions of bagging $\hat{\theta}_{ee}$, one is to solve the "bagged estimating equation" defined by bagging the score function, and another is to average over resampled estimates of $\hat{\theta}_{ee}$. Similarly to the discussion in Section 2.1, the first version results in an estimator equivalent to the original estimator, because the bootstrap expectation of bootstrap samples of $\hat{S}(\gamma)$ is equal to $\hat{S}(\gamma)$ for fixed γ. We therefore only consider the latter version. To define the bagged estimating equation estimator, we first define the replicated score function $\hat{S}_b(\gamma)$ based on resample A_b as

$$\hat{S}_b(\gamma) = \frac{1}{N}\sum_{i\in A_b}\frac{1}{\pi_i^*}\psi(y_i - \gamma).$$

Then the replicated estimator based on A_b is defined as $\hat{\theta}_{ee}(\mathcal{Y}_b^*) = \inf\left\{\gamma : \hat{S}_b(\gamma) \geq 0\right\}$. Thus the overall bagging estimator is defined as

$$\hat{\theta}_{ee,bag} = \frac{1}{\binom{n}{k}}\sum\hat{\theta}_{ee}(\mathcal{Y}_b^*), \tag{2.7}$$

where the average is over all possible without-replacement samples of size k selected from A. Chen and Hall (2003) discussed bagging estimators defined by nonlinear estimating equations under the *iid* setup, and they stated that bagging does not always improve the precision of estimators under study.

3 Theoretical results

We begin by briefly describing the asymptotic analysis of the bagging estimators under general sampling design from a finite population, i.e. the design-based setting. We do this under the usual increasing-population framework, where we consider an increasing sequence of nested populations, say U_N, $N = 1,2,\ldots$, with finite population means $\boldsymbol{\mu}_N$. Associated with the sequence of populations is a sequence of sampling designs used to draw random sample $A_N \subseteq U_N$ of sample size n_N, with associated inclusion probabilities π_{iN}. As commonly done in the survey literature, we suppress the N subscript in the sample A, the sample size n and the inclusion probabilities π_i. For the sake of brevity, only design-based asymptotic results for bagging differentiable estimator $\hat{\theta}_d$ and non-differentiable $\hat{\theta}_{nd}$ are provided. The formal assumptions under which the results are obtained and the theorems for differentiable and non-

differentiable estimators are in Appendix A.1. The main result we are able to obtain in this design-based setting is that, if we are starting from a design-consistent estimator and we let the number of bootstrap samples k grow with n, the bagged versions of the estimators are also design consistent. This is clearly a key property of these estimators, since there would be no reason to consider them unless they satisfied this design consistency.

Unfortunately, the above design-based results are quite limited and in particular, do not provide an asymptotic distribution with which one might be able to perform inference, another highly desirable property of survey estimators. We therefore also consider a model-based setting, under which we are able to obtain an asymptotic variance approximation. In presenting model-based results, we assume the sampling design selecting the original sample A is an equal probability design, and the population characteristics can be regarded as an *iid* sample from a superpopulation distribution. Under this framework, the bagging estimator can be treated as a U-statistic. Thus we can apply the theory on U-statistics to obtain asymptotic expansion of bagging estimators. The analysis parallels that of Bühlmann and Yu (2002) and Buja and Stuetzle (2006). For the current paper, we restrict ourselves to bootstrap samples of size k where k is bounded and fixed. Under this asumption, the bagging estimators can be regarded as fixed-degree U-statistics, for which asymptotic theory has been well developed. A more interesting case is when the resample size k grows with sample size n, and this leads to infinite-degree U-statistics. Infinite-degree U-statistics have applications in studying the Kaplan-Meier estimator and m-out-of-n bootstrap estimators, and the readers are referred to Frees (1989); Heilig (1997); Heilig and Nolan (2001), and the references therein on their statistical properties. Schick and Wefelmeyer (2004) studied the statistical properties of infinite-degree U-statistics constructed from moving averages of innovations in time series. The study of bagging estimators by viewing them as infinite-degree U-statistics is out of the scope of the current paper, and hence we limit ourselves to the case of fixed and bounded bootstrap sample size in the model-based case.

We first consider bagged estimator (2.5). Under SRSWOR, estimator (2.5) can be simplified to

$$\hat{\theta}_{nd} = \frac{1}{n}\sum_{i\in A} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\right)$$

and the bagged version of $\hat{\theta}_{nd}$ is defined as

$$\hat{\theta}_{nd,bag} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\binom{n-1}{k-1}} \sum_{A_b \ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) \tag{3.1}$$

where $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ only depends on resample A_b. For ease of presentation, we take $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ as the sample mean. In this case, straightforward algebra reveals that

$$\hat{\theta}_{nd,bag} = \frac{1}{\binom{n}{k}} \sum_{A_b \in \mathcal{A}} \left\{ \frac{1}{k}\sum_{i\in A_b} h\left(\frac{k-1}{k}\mathbf{y}_i - \frac{1}{k}\sum_{j\neq i}\mathbf{y}_j \right) \right\},$$

where $\mathcal{A}$ is the collection of subsets of size k from set $\{1,2,\ldots,n\}$. The estimator $\hat{\theta}_{nd,bag}$ is a degree-k U-statistic with kernel

$$g(y_1,\ldots,y_k)=\frac{1}{k}\sum_{i=1}^{k}h\left(\frac{k-1}{k}\mathbf{y}_i-\frac{1}{k}\sum_{\substack{j=1\\ j\neq i}}^{k}\mathbf{y}_j\right)$$

provided that k remains finite.

One can see that the bagging estimator $\hat{\theta}_{nd,bag}$ is a symmetric statistic of $\mathbf{y}_i$, and standard theory on symmetric statistics (Lee 1990) applies. The results are stated in Theorem 1, with assumptions and proofs in Appendix A.2.

Theorem 1 *Under Assumptions M.1-M.4 on the superpopulation distribution, sampling and resampling designs,*

$$\mathrm{AV}\left(\hat{\theta}_{nd,bag}\right)^{-1/2}\left(\hat{\theta}_{nd,bag}-\theta_{nd,\infty}\right)\xrightarrow{p}N(0,1), \tag{3.2}$$

where the limiting value $\theta_{nd,\infty}=\lim_{n\to\infty}\mathrm{E}\left[h\left(\mathbf{y}_i-\hat{\boldsymbol{\lambda}}\right)\right]$, *the asymptotic variance*

$$\mathrm{AV}\left(\hat{\theta}_{nd,bag}\right)=\frac{1}{n}\mathrm{Var}\left[u(\mathbf{y}_i)\right]+\frac{(k-1)^2}{n}\mathrm{Var}\left[v(\mathbf{y}_i)\right]+\frac{2(k-1)}{n}\mathrm{Cov}\left[u(\mathbf{y}_i),v(\mathbf{y}_i)\right], \tag{3.3}$$

and

$$u(\mathbf{y})=\mathrm{E}\left[h\left(\mathbf{y}-\hat{\boldsymbol{\lambda}}(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y})\right)\right],$$
$$v(\mathbf{y})=\mathrm{E}\left[h\left(\mathbf{y}_1-\hat{\boldsymbol{\lambda}}(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y})\right)\right].$$

As indicated by (3.3), the asymptotic variance of the bagging estimator depends on unknown functions $u(\mathbf{y})$ and $v(\mathbf{y})$, which are expectations of $h(\cdot)$ with respect to the superpopulation distribution. In $u(\mathbf{y})$ and $v(\mathbf{y})$, $\hat{\boldsymbol{\lambda}}(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1},\mathbf{y})$ is calculated from $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1}$ together with an arbitrary vector $\mathbf{y}$. The expectation is with respect to the distribution of *iid* random vectors $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_{k-1}$. This high-dimensional expectation is difficult to calculate and may not have an explicit expression in general. The exact form of $u(\cdot)$ and $v(\cdot)$ can not be obtained but can be approximated via a resampling-based approach. The unknown functions $u(\cdot)$ and $v(\cdot)$ are defined as expectations of respective quantities with respect to the superpopulation distribution, which can be approximated by the expectation with respect to the empirical distribution.

The model-based asymptotic variance can be estimated along with the process of bagging. We can calculate integrands $h\left(\mathbf{y}-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*,\mathbf{y}\right)\right)$ and $h\left(\mathbf{y}_1-\hat{\boldsymbol{\lambda}}\left(\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*,\mathbf{y}\right)\right)$ based on each bootstrap sample, with $\mathbf{y}$ denoting where we want to evaluate $u(\cdot)$ and $v(\cdot)$, and $\mathbf{y}_1^*,\mathbf{y}_2^*,\ldots,\mathbf{y}_{k-1}^*$ denoting resampled values. Then we can average each quantity to approximate the expectation. Finally, the variance can be estimated by computing the sample variance of the expectations evaluated at each of the sample points. For nonsmooth estimators like the ones we are dealing with, it is often recommended to use smoothed bootstrap in variance approximation (Efron 1979; Davison and Hinkley 1997). We apply the smoothed

bootstrap and add a small amount of noise to each resampled value to smooth the underlying function. The detailed algorithm will be explained in Section 5 through an example.

We now study the model-based result of bagging estimators defined by estimating equations (2.7). A special case in this framework is bagging sample quantiles, which was studied by Knight and Bassett (2002). Knight and Bassett (2002) considered both bootstrap and SRSWOR for resampling, and studied the effects of bagging on the remainder term in the Bahadur representation of quantiles (Bahadur 1966). We take a slightly different perspective and treat the bagging estimator as a U-statistic. Assumptions and proof are again in Appendix A.2. Note that Assumption M.5 requires that the non-differentiable estimating function have a smooth limit. In the next theorem, we provide linearization of the bagging estimating equation estimator and give an expression for the asymptotic variance.

Theorem 2 *Under Assumptions M.1-M.3 and M.5, the following asymptotic result holds for the bagged estimating equation estimator (2.7),*

$$\mathrm{AV}\left(\hat{\theta}_{ee,bag}\right)^{-1/2}\left(\hat{\theta}_{ee,bag}-\theta_{ee,\infty}\right)\xrightarrow{p}N(0,1), \tag{3.4}$$

where $\theta_{ee,\infty}$ *denotes the asymptotic limit of population quantity* θ_{ee}*, the asymptotic variance of* $\hat{\theta}_{ee,bag}$ *is*

$$\mathrm{AV}\left(\hat{\theta}_{ee,bag}\right)=\frac{k^2}{n}\mathrm{Var}\left[u(y_i)\right], \tag{3.5}$$

and

$$u(y)=\mathrm{E}\ \inf\left\{\gamma:\frac{1}{k}\sum_{i=1}^{k-1}\psi(y_i-\gamma)+\frac{1}{k}\psi(y-\gamma)\geq 0\right\}. \tag{3.6}$$

As we saw for the bagged estimator (3.1), the asymptotic results in Theorem 2 involve an unknown function. This function can again be computed using resampling that takes advantage of the available replicate samples.

4 Variance Estimation

While the model-based approach makes it possible to obtain asymptotic distributions and hence perform inference that is asymptotically correct, we are most interested here in the design-based applications of bagging. In the design-based context, the construction of the bagging estimator can be naturally combined with the variance estimation of the original statistic, by taking advantage of the replication samples released by the statistical agencies. In this article, we take stratified simple random sampling as a specific example, with a bootstrap sampling design of stratified SRSWOR.

We begin by applying a version of the Rao and Wu (1988) bootstrap procedure to estimate the variance of the survey estimators prior to bagging. Let N_h, n_h and k_h denote the population size, sample size and sub-sample size in the h-th stratum, $h=1,2,\ldots,H$. Here, B bootstrap samples are drawn by stratified simple random sample without replacement of size k_h for computing the bootstrap variance of the original statistic and the bagging estimator. For each bootstrap sample, we assign a weight of

$$\frac{N_h}{N}\left(1-k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\right)\frac{1}{n_h}+\frac{N_h}{N}k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\frac{1}{k_h}$$

to each sampled element in the h-th stratum, and

$$\frac{N_h}{N}\left(1-k_h^{1/2}\left(n_h-1\right)^{-1/2}\left(1-\frac{n_h}{N_h}\right)^{1/2}\right)\frac{1}{n_h}$$

to the nonsampled elements. We then use the ordinary variance of the replicated sample estimators as variance estimator. The aforementioned weighting scheme is algebraically identical to equation 4.1 of Rao and Wu (1988), in which the finite population correction is incorporated into replication weights. The resampling variance estimator derived from the weighting method reduces to ordinary variance estimator under stratified SRSWOR and guarantees design unbiasedness. In order to combine bagging with bootstrap variance estimator, we use the same bootstrap samples to construct the bagging estimators for the population quantities of interest.

Under the design-based framework, no analytic variance estimator is available for the bagged estimator in general. For now, we would suggest the following two variance estimation approaches in practice:

(Var. 1) Use the estimated variance of the original estimator even though the bagged estimator may have a smaller variance. This method provides confidence intervals of the same width but outperforms the original confidence interval in having larger coverage rate.

(Var. 2) Multiply the estimated variance of the original estimator by an adjustment factor accounting for the likely improvement in efficiency. One possible choice for such a factor is the efficiency gain assuming the sample is an *iid* sample from an infinite superpopulation. The factor can be determined by using the results of Theorems 1 and 2, or by a nonparametric bootstrap experiment. One such possible bootstrap procedure is double bootstrap, which is implemented by drawing ordinary bootstrap resamples to estimate the variance of the original estimator, and another level of SRSWOR resamples to determine the variance of the bagging estimator. One can estimate the ratio of the variance of bagging estimator to original estimator using these nested bootstrap samples, and multiply the design variance of the original estimator by this ratio.

We will explore both approaches in the simulations in Section 5, but this is clearly an area in which further research is warranted.

5 Simulations

To evaluate the practical behavior of bagging in the survey context, we generate a finite population of size $N=2,000$ with three strata. The size of each stratum is denoted as N_h with $h=1,2,3,$ and the stratum proportions are fixed at $\left(N_1,N_2,N_3\right)/N=\left(0.5,0.3,0.2\right).$ The distribution of the target variable y_i

within each stratum is $y_{1i} \sim \left|N(-1,1)\right|$, $y_{2i} \sim \Gamma(1,1)$ and $y_{3i} \sim \left|N(3,2)\right|$. An auxiliary variable x_i is generated via $x_i = A_0 + A_1 y_i + A_2\left(G_i - \alpha/\beta\right)$ where $A_0 = A_1 = 2$, $A_2 = 1$, $\alpha = 2$, $\beta = 1$ and $G_i \overset{iid}{\sim} \Gamma(2,1)$. We repeatedly draw samples of size n using stratified simple random sampling from the population of interest and the sample size allocation is $(n_1, n_2, n_3)/n = (0.3, 0.3, 0.4)$. In this set-up, the design is clearly informative, because the observations are not *iid* in the overall population and are correlated with the inclusion probabilities.

We are interested in three population quantities: a population α-quantile, a population proportion below a given fraction of a population quantile (see Berger and Skinner 2003, for an example) and the Rao-Kovar-Mantel (RKM) estimator of the distribution function (Rao et al. 1990). The former is an example of a non-differentiable estimating equation-based estimator, while the latter two are explicitly defined non-differentiable estimators. The sample estimator of the quantile is found by inverting the estimated cumulative distribution function. The sample estimator of the proportion below a given fraction of a population quantile is the HT estimator of the proportion of observations below the sample median of a variable of interest times a constant c,

$$\hat{\theta}_{pr} = \frac{1}{N}\sum_{i\in A}\frac{1}{\pi_i}\mathrm{I}_{\left(y_i \le c\hat{\theta}_{\mathrm{med}}\right)},$$

where $\hat{\theta}_{\mathrm{med}}$ denotes the sample median of the y_i. The design-based RKM difference estimator based on a ratio model is

$$\hat{\theta}_{\mathrm{RKM}} = \frac{1}{N}\left\{\sum_{i\in A}\frac{1}{\pi_i}\mathrm{I}_{(y_i \le t)} + \sum_{i=1}^{N}\mathrm{I}_{\left(\hat{R}x_i \le t\right)} - \sum_{i\in A}\frac{1}{\pi_i}\mathrm{I}_{\left(\hat{R}x_i \le t\right)}\right\}, \tag{5.1}$$

where $\hat{R}$ denotes the estimated ratio between y and x.

The design variance of these non-differentiable estimators is somewhat cumbersome to estimate. For variance and interval calculations for sample quantiles, the readers are referred to Francisco and Fuller (1991), Sitter and Wu (2001), and references therein. For proportion below an estimated level, see Shao and Rao (1993) and Berger and Skinner (2003).

The design variances of the original estimators $\hat{\theta}_{qr}$, $\hat{\theta}_{pr}$ and $\hat{\theta}_{\mathrm{RKM}}$, are estimated via the without-replacement bootstrap procedure described in the previous section. We employ a bootstrap sample size of $k_h = n_h/2$. The so-constructed bagging estimators are often referred to as subagging estimators (Bühlmann and Yu 2002). It was established that without-replacement samples of size $n/2$ produces similar results to with replacement samples of size n in bagging (Buja and Stuetzle 2006; Friedman and Hall 2007). We apply the two variance approaches for bagging estimators proposed in the previous section, i.e. one identical to that of unbagged estimator (Var. 1) and another one that multiplies the original variance estimate by a model-based adjustment factor (Var. 2). The factor is determined by double bootstrap on one particular sample. In principle, one should repeat the exercise for each sample, but this is precluded by the heavy computational burden. The confidence intervals of all three estimators are constructed by normal approximation. The confidence intervals for the proportion and the RKM estimator are constructed by normal approximation on *logit* transformed scale, $\log\left[\hat{\theta}/\left(1-\hat{\theta}\right)\right]$ or $\log\left[\hat{\theta}_{bag}/\left(1-\hat{\theta}_{bag}\right)\right]$, and then back transformation (Agresti 2002; Korn and Graubard 1998).

Table 5.1 summarizes the bias, standard deviation and MSE ratio of the original and bagged sample quantiles and Table 5.2 examines the variance estimators and confidence intervals. The sample sizes are chosen to be $n = 100$ and 200. From Table 5.1, we can see that the bagged quantile estimator is more efficient than the original estimator since the MSE ratio is less than one in this simulation experiment. The smoothing effects of bagging generally become more prominent as we decrease the sample size. In Table 5.2, we compare the two confidence intervals with bagging point estimator to that of original confidence intervals. As expected, the confidence interval constructed via method 1 has the same length and higher coverage than the original. In this example, the confidence intervals via method 2 are narrower but maintain coverage level close to nominal.

Table 5.1
Bias, standard deviation and MSE ratios of sample quantiles and bagged sample quantiles; population size $N = 2,000$, number of bootstraps $B = 2,000$, and results are from $2,000$ simulations

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
α	**0.2**	**0.3**	**0.5**	**0.7**	**0.8**	**0.2**	**0.3**	**0.5**	**0.7**	**0.8**
$\text{bias}\left(\hat{\theta}_{qt}\right)$	0.002	0.008	0.000	-0.005	-0.035	-0.008	0.005	0.006	0.007	-0.005
$\text{bias}\left(\hat{\theta}_{qt,bag}\right)$	0.018	0.019	-0.001	-0.007	-0.043	-0.006	0.009	0.005	0.006	-0.022
$\text{sd}\left(\hat{\theta}_{qt}\right)$	0.093	0.124	0.149	0.181	0.212	0.070	0.076	0.103	0.136	0.148
$\text{sd}\left(\hat{\theta}_{qt,bag}\right)$	0.089	0.112	0.138	0.167	0.197	0.065	0.073	0.099	0.127	0.139
$\frac{MSE_p\left(\hat{\theta}_{qt,bag}\right)}{MSE_p\left(\hat{\theta}_{qt}\right)}$	0.946	0.844	0.859	0.854	0.875	0.866	0.924	0.919	0.862	0.912

Table 5.2
Relative bias, coverage probability and confidence interval width of bootstrap variance estimators for sample quantiles and unadjusted $\left(\hat{V}_1\right)$ and adjusted $\left(\hat{V}_2\right)$ variance estimators for bagged sample quantiles; simulation setting is the same as in Table 5.1

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
α	**0.2**	**0.3**	**0.5**	**0.7**	**0.8**	**0.2**	**0.3**	**0.5**	**0.7**	**0.8**
$\frac{\text{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{qt}\right)\right]}{V\left(\hat{\theta}_{qt}\right)}$	1.208	1.091	1.099	1.135	1.205	1.067	1.117	1.093	1.098	1.180
$\frac{\text{E}\left[\hat{V}_1\left(\hat{\theta}_{qt,bag}\right)\right]}{V\left(\hat{\theta}_{qt,bag}\right)}$	1.327	1.325	1.279	1.331	1.402	1.224	1.217	1.188	1.273	1.326
$\frac{\text{E}\left[\hat{V}_2\left(\hat{\theta}_{qt,bag}\right)\right]}{V\left(\hat{\theta}_{qt,bag}\right)}$	1.307	1.217	1.196	1.184	1.383	1.245	1.249	1.392	1.107	1.104
C.P.(C.I.)	0.944	0.934	0.924	0.928	0.922	0.938	0.951	0.942	0.935	0.950
$\text{C.P.}\left(\text{C.I.1}_{\cdot bag}\right)$	0.950	0.946	0.938	0.938	0.939	0.942	0.950	0.946	0.943	0.954
$\text{C.P.}\left(\text{C.I.2}_{\cdot bag}\right)$	0.949	0.934	0.932	0.929	0.938	0.944	0.952	0.958	0.927	0.936
Width(C.I.)										
$\text{Width}\left(\text{C.I.1}_{\cdot bag}\right)$	0.386	0.492	0.597	0.729	0.880	0.277	0.309	0.414	0.544	0.612
$\text{Width}\left(\text{C.I.2}_{\cdot bag}\right)$	0.383	0.472	0.577	0.688	0.874	0.279	0.313	0.448	0.508	0.559

Tables 5.3 and 5.4 summarize design-based results on the low-income proportion estimator. Based on the MSE ratio, we can see that the bagging estimator is uniformly more efficient than the original estimator, and the MSE of bagging estimator is less than 50% of that of original estimator in a few cases (see $c = 1.2$). The likely reason for this is that the estimator involves two "levels" of non-differentiability: the sample median being a non-differentiable estimator, whose efficiency gain was shown in Table 5.1, and the low-income proportion being a non-differentiable function of the sample median. The "jumps" in the estimators are smoothed out by bagging, resulting in a more stable estimator. The confidence interval comparison in Table 5.4 leads to results similar to the quantile case.

Table 5.3
Bias, standard deviation and MSE ratio of estimated proportion below a constant c multiplied by estimated median and the bagged proportion estimator; population size $N = 2,000,$ number of bootstraps $B = 2,000,$ and results are from $2,000$ simulations

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
c	**0.2**	**0.4**	**0.6**	**1.2**	**1.5**	**0.2**	**0.4**	**0.6**	**1.2**	**1.5**
$\text{bias}\left(\hat{\theta}_{pr}\right)$	-0.002	-0.002	-0.003	0.011	0.006	0.000	-0.002	-0.005	-0.004	-0.004
$\text{bias}\left(\hat{\theta}_{pr,bag}\right)$	-0.004	-0.004	-0.007	0.017	0.009	-0.001	-0.005	-0.009	-0.001	-0.004
$\text{sd}\left(\hat{\theta}_{pr}\right)$	0.034	0.039	0.038	0.034	0.046	0.023	0.027	0.026	0.026	0.036
$\text{sd}\left(\hat{\theta}_{pr,bag}\right)$	0.031	0.035	0.031	0.020	0.034	0.022	0.025	0.022	0.017	0.029
$\dfrac{MSE_p\left(\hat{\theta}_{pr,bag}\right)}{MSE_p\left(\hat{\theta}_{pr}\right)}$	0.861	0.821	0.709	0.538	0.581	0.883	0.860	0.783	0.434	0.671

Table 5.4
Relative bias, coverage probability and confidence interval width of bootstrap variance estimators for sample proportions and unadjusted $\left(\hat{V}_1\right)$ and adjusted $\left(\hat{V}_2\right)$ variance estimators for bagged sample proportions; simulation setting is the same as in Table 5.3. We use "C.I.T." to denote confidence intervals obtained with logit transformation

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
c	**0.2**	**0.4**	**0.6**	**1.2**	**1.5**	**0.2**	**0.4**	**0.6**	**1.2**	**1.5**
$\dfrac{\text{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{pr}\right)\right]}{V\left(\hat{\theta}_{pr}\right)}$	1.122	1.191	1.325	1.472	1.281	1.140	1.191	1.251	1.350	1.217
$\dfrac{\text{E}\left[\hat{V}_1\left(\hat{\theta}_{pr,bag}\right)\right]}{V\left(\hat{\theta}_{pr,bag}\right)}$	1.323	1.471	1.959	4.095	2.307	1.293	1.428	1.766	3.064	1.821
$\dfrac{\text{E}\left[\hat{V}_2\left(\hat{\theta}_{pr,bag}\right)\right]}{V\left(\hat{\theta}_{pr,bag}\right)}$	1.240	0.963	1.190	1.174	1.149	1.145	1.262	1.319	2.039	1.524
C.P.(C.I.T.)	0.969	0.970	0.984	0.991	0.980	0.964	0.974	0.977	0.983	0.946
$\text{C.P.}\left(\text{C.I.T.1}_{\cdot bag}\right)$	0.979	0.983	0.995	0.998	0.995	0.974	0.980	0.988	0.998	0.976
$\text{C.P.}\left(\text{C.I.T.2}_{\cdot bag}\right)$	0.976	0.944	0.973	0.922	0.942	0.962	0.969	0.968	0.993	0.957
Width(C.I.T)										
$\text{Width}\left(\text{C.I.T.1}_{\cdot bag}\right)$	0.144	0.166	0.168	0.157	0.197	0.098	0.115	0.114	0.113	0.149
$\text{Width}\left(\text{C.I.T.2}_{\cdot bag}\right)$	0.139	0.134	0.131	0.085	0.140	0.093	0.108	0.099	0.092	0.136

Tables 5.5 and 5.6 summarize the design-based results on the RKM estimator. Again, we observe the efficiency gain by applying the bagging method, and the gain is between 2% and 12%. Both variance estimators of the bagging quantity perform quite well. Both versions of confidence intervals for bagging estimators have actual coverage rates close to 95%, and the confidence intervals using the adjustment factor approach (Var. 2) are slightly shorter than method 1.

Table 5.5
Bias, standard deviation and MSE ratios of RKM estimator and bagging RKM estimator (5.1); population size $N = 2,000,$ number of bootstraps $B = 2,000,$ and results are from $2,000$ simulations

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
t	**0.5**	**1.5**	**2.5**	**3.5**	**4.5**	**0.5**	**1.5**	**2.5**	**3.5**	**4.5**
$\text{bias}\left(\hat{\theta}_{\text{RKM}}\right)$	0.000	0.000	0.000	0.000	0.000	-0.001	0.001	0.000	0.000	0.001
$\text{bias}\left(\hat{\theta}_{\text{RKM},bag}\right)$	-0.001	0.000	-0.001	0.000	0.000	-0.001	0.001	0.000	0.001	0.001
$\text{sd}\left(\hat{\theta}_{\text{RKM}}\right)$	0.043	0.044	0.030	0.015	0.012	0.030	0.030	0.020	0.011	0.009
$\text{sd}\left(\hat{\theta}_{\text{RKM},bag}\right)$	0.042	0.042	0.028	0.014	0.012	0.030	0.029	0.019	0.011	0.009
$\frac{MSE_p\left(\hat{\theta}_{\text{RKM},bag}\right)}{MSE_p\left(\hat{\theta}_{\text{RKM}}\right)}$	0.965	0.911	0.877	0.914	0.917	0.976	0.928	0.917	0.918	0.981

Table 5.6
Relative bias, coverage probability and confidence interval width of bootstrap variance estimators for the RKM estimator (5.1) and unadjusted $\left(\hat{V}_1\right)$ and adjusted $\left(\hat{V}_2\right)$ variance estimators for bagging RKM estimators; simulation setting is the same as in Table 5.5

	$n = 100,\ k = 50$					$n = 200,\ k = 100$				
t	**0.5**	**1.5**	**2.5**	**3.5**	**4.5**	**0.5**	**1.5**	**2.5**	**3.5**	**4.5**
$\frac{\text{E}\left[\hat{V}_{boot}\left(\hat{\theta}_{\text{RKM}}\right)\right]}{V\left(\hat{\theta}_{\text{RKM}}\right)}$	1.081	1.192	1.078	1.082	1.078	1.016	1.045	1.138	1.121	1.016
$\frac{\text{E}\left[\hat{V}_1\left(\hat{\theta}_{\text{RKM},bag}\right)\right]}{V\left(\hat{\theta}_{\text{RKM},bag}\right)}$	1.115	1.324	1.183	1.198	1.156	1.038	1.138	1.223	1.210	1.062
$\frac{\text{E}\left[\hat{V}_2\left(\hat{\theta}_{\text{RKM},bag}\right)\right]}{V\left(\hat{\theta}_{\text{RKM},bag}\right)}$	1.087	1.117	0.962	1.042	1.019	1.009	1.083	1.106	1.118	1.002
C.P.(C.I.)	0.958	0.963	0.955	0.956	0.959	0.954	0.956	0.966	0.964	0.948
$\text{C.P.}\left(\text{C.I.1}_{\cdot bag}\right)$	0.958	0.968	0.958	0.967	0.964	0.958	0.964	0.970	0.970	0.956
$\text{C.P.}\left(\text{C.I.2}_{\cdot bag}\right)$	0.957	0.954	0.937	0.951	0.950	0.955	0.958	0.959	0.960	0.948
Width(C.I.)										
$\text{Width}\left(\text{C.I.1}_{\cdot bag}\right)$	0.171	0.183	0.116	0.074	0.052	0.122	0.122	0.083	0.049	0.034
$\text{Width}\left(\text{C.I.2}_{\cdot bag}\right)$	0.169	0.168	0.105	0.069	0.049	0.120	0.120	0.079	0.047	0.033

In the context of nonsmooth estimators such as those considered here, it is often recommended that one uses a smoothed bootstrap instead of the simple bootstrap in variance estimation. We considered perturbing each resampled observation y_{hi}^* in the h-th stratum to obtain,

$$\tilde{y}_{hi}^* = \bar{y}_h + \left(1+\sigma_Z^2\right)^{-1/2}\left(y_{hi}^* - \bar{y}_h + s_h Z^*\right), \tag{5.2}$$

where $\bar{y}_h$, s_h denote the sample mean and standard deviation of the original sample stratum, y_{hi}^* denotes the originally resampled value and Z^* denotes random noise with $Z^* \overset{iid}{\sim} N\left(0,\sigma_Z^2\right)$. The variance of Z^* controls the amount of smoothing. We applied this method to quantile estimation and the proportion below an estimated level, but it did not appear to improve the performance of the estimation procedure. One possible explanation is that noise contamination "jitters" duplicated observations arising from with-replacement sample and stabilizes subsequent variance estimator to some extent. Since we used without-replacement sampling, this problem was already mostly avoided. More careful study is necessary to understand the effect of smoothing in the context.

6 Conclusions

In this article, we have explored the use of bagging procedures for nonlinear and non-differentiable survey estimators. We presented theoretical results on bagging estimator both under design-based and model-based framework. The bagging estimator can be treated as the expectation of a two-phase estimator conditioning on the first phase, and this expectation smoothes out "jumps" in the non-differentiable estimator. The empirical study has revealed the potential of bagging non-differentiable survey estimators, and while the relative performance of bagging varies from one scenario to another, the results are certainly promising.

How to estimate the variance of bagged survey estimators remains an open question when the sampling design is a general complex design. We have proposed two ideas for variance estimation for practical use, but further theoretical study of variance estimation under design-based framework is certainly warranted.

Appendix

A.1 Design-based theory

Assumptions D.1-D.6 are used to show the design-based results given below (Theorems 3 and 4). Assumption D.1 specifies moment conditions on the study variable $\mathbf{y}_i$, and Assumption D.2 specifies conditions on the second order inclusion probability of the sampling design. Assumption D.3 guarantees that the size of each resample converges to infinity in the limit. Assumption D.4 specifies smoothness conditions on $m(\cdot)$ in the differentiable estimator. Assumptions D.5-D.6 are used to show the design consistency of bagging non-differentiable survey estimators.

(D.1) The study variable $\mathbf{y}_i$ has finite $2+\delta$ population moment for arbitrarily small $\delta > 0$,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \| \mathbf{y}_i^{2+\delta} \| \; < \; \infty,$$

where each element of $\mathbf{y}_i^{2+\delta}$ is the original element raised to the power of $2+\delta$ and $\|\cdot\|$ denotes Euclidean norm.

(D.2) For all N, $\min_{i\in U_N} \pi_i \geq \pi_N^* > 0$, where $N\pi_N^* \to \infty$, and

$$\lim_{N\to\infty} \sup\, n \cdot \max \left| \pi_{ij} - \pi_i \pi_j \right| < \infty,$$

where π_{ij} denotes the joint inclusion probability of elements i, j.

(D.3) The resampling process generating A_b is SRSWOR of size k, with $k = O\left(n^{\kappa}\right)$, $\kappa \in (0,1]$. Further, every bootstrap resample of size k is used in calculating the bagged estimator.

(D.4) The function $m(\cdot)$ is differentiable and has nontrivial continuous second derivative in a compact neighborhood of $\boldsymbol{\mu}_N$.

(D.5) The estimator $\hat{\boldsymbol{\lambda}}$ is $\sqrt{n}$-consistent for the population target $\boldsymbol{\lambda}_N$, $\lim_{N\to\infty} \boldsymbol{\lambda}_N = \boldsymbol{\lambda}_\infty$ and the estimator $\hat{\boldsymbol{\lambda}}$ is a symmetric statistic.

(D.6) The function $h(\cdot)$ is bounded and the population quantity is "compactly differentiable in a weak sense" (Dümbgen 1993). There exists a function $g(\cdot)$ such that,

$$\sup_{\mathbf{s}\in C_{\mathbf{s}}} \left| \frac{1}{N}\sum_{i=1}^{N} h\left(\mathbf{y}_i - \boldsymbol{\lambda}_\infty - N^{-\alpha}\mathbf{s}\right) - \frac{1}{N}\sum_{i=1}^{N} h\left(\mathbf{y}_i - \boldsymbol{\lambda}_\infty\right) - g\left(\boldsymbol{\lambda}_\infty\right) N^{-\alpha}\mathbf{s} \right| \to 0,$$

where $C_{\mathbf{s}}$ is a large enough compact set in $\mathbb{R}^p$, $0 < \alpha \leq 1/2$ and $g\left(\boldsymbol{\lambda}_\infty\right)$ is bounded.

The following theorem gives several asymptotic approximations for the bagged estimator, depending on the rate of convergence of k relative to n. In all three cases, the bagged estimator is design consistent. Intuitively speaking, the bagging estimator behaves like the original estimator when the resample size k is large (approaches infinity no slower than $n^{1/2}$), but converges at a different speed when the resample size is small.

Theorem 3 *Under Assumptions D.1-D.4, the bagged differentiable estimator* $\hat{\theta}_{d,bag}$ *admits the following second-order expansion,*

$$\hat{\theta}_{d,bag} - \theta_d = \begin{cases} \left\{m'\left(\boldsymbol{\mu}_N\right)\right\}^T \left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_N\right) + o_p\left(n^{-1/2}\right), & \text{for } \kappa > 1/2 \\ \left\{m'\left(\boldsymbol{\mu}_N\right)\right\}^T \left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_N\right) + \\ \quad \dfrac{1}{2\binom{n}{k}} \sum \left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right) - \boldsymbol{\mu}_N\right)^T m''\left(\boldsymbol{\mu}_N\right)\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right) - \boldsymbol{\mu}_N\right) + o_p\left(n^{-1/2}\right), & \text{for } \kappa = 1/2 \\ \dfrac{1}{2\binom{n}{k}} \sum \left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right) - \boldsymbol{\mu}_N\right)^T m''\left(\boldsymbol{\mu}_N\right)\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right) - \boldsymbol{\mu}_N\right) + o_p\left(k^{-1}\right), & \text{for } \kappa < 1/2 \end{cases}$$

where $\kappa > 0$ *is such that the resample size* $k = O\left(n^{\kappa}\right)$.

Proof of Theorem 3:

The proof easily follows from a Taylor expansion of the individual resample-based estimator $m\left(\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)\right)$ around $\boldsymbol{\mu}_N$. The linear expansion term reduces to $\left\{m'\left(\boldsymbol{\mu}_N\right)\right\}^T \left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_N\right)$ based on an earlier argument. Under D.1 and D.3, the quadratic term has the same order as the SRSWOR variance of $\hat{\boldsymbol{\mu}}\left(\mathcal{Y}_b^*\right)$ and hence is $o_p\left(1/k\right)$.

Next, Theorem 4 gives the design consistency of the non-differentiable bagged estimator.

Theorem 4 *Under Assumptions D.1-D.3 and D.5-D.6, the bagged non-differentiable estimator* $\hat{\theta}_{nd,bag}$ *is design consistent for its population target* θ_{nd}, *i.e.,* $\hat{\theta}_{nd,bag} - \theta_{nd} = o_p\left(1\right)$.

Proof of Theorem 4:

We can establish that $\left(1/N\right)\sum_{i \in A}\left(1/\pi_i\right)h\left(\mathbf{y}_i - \boldsymbol{\lambda}_N\right)$ is design consistent for θ_{nd} as a result of D.2 and the fact that $h(\cdot)$ is bounded (D.6). Then it suffices to show that $\hat{\theta}_{nd,bag} - \left(1/N\right)\sum_{i \in A}\left(1/\pi_i\right)h\left(\mathbf{y}_i - \boldsymbol{\lambda}_N\right) = o_p\left(1\right)$, or

$$\frac{1}{N}\sum_{i \in A}\frac{1}{\pi_i}\left\{\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni i} h\left(\mathbf{y}_i - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right) - h\left(\mathbf{y}_i - \boldsymbol{\lambda}_N\right)\right\} = o_p\left(1\right)$$

following (2.6). We can establish that the collection of resample-based estimators $\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)$ are uniformly contained in a neighborhood of $\boldsymbol{\lambda}_N$, or, $\sup_{A_b}\left|\hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right) - \boldsymbol{\lambda}_N\right| = O\left(N^{-\alpha}\mathbf{s}\right)$ for some $\alpha > 0$. Then we can apply D.6 to conclude the design consistency of the bagging estimator.

A.2 Model-based theory

Assumptions M.1-M.4 are used to show the model-based results (Theorems 1 and 2). Assumption M.1 specifies superpopulation distribution of population characteristics $\mathbf{y}_i$. Assumptions M.2 and M.3 assume simple random without replacement sampling for both the design and the resampling process. Assumption M.5 is needed for showing the model-based asymptotic results for the bagging estimator defined by estimating equations.

(M.1) The sequence of population characteristics $\mathbf{y}_i$ constitute an *iid* sample from a probability distribution with density $f_Y(\mathbf{y})$.

(M.2) The sampling design is ignorable, or equivalently, the sampled and unsampled observations are subject to the same distribution.

(M.3) The resampling process generating A_b is SRSWOR of size k, where the bootstrap sample size k is bounded. Further, every bootstrap resample of size k is used in calculating the bagged estimator.

(M.4) The function $h(\cdot)$ is bounded.

(M.5) Let $S_\infty(\gamma) = \mathrm{E}\psi(y_i - \gamma)$ be a continuous function of γ, and $\theta_{ee,\infty}$ be the smallest root of $S_\infty(\gamma) = 0$; for an arbitrary y in the support of the random variable y_i, the quantity

$$\inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k-1}\psi(y_i - \gamma) + \frac{1}{k}\psi(y - \gamma) \geq 0\right\}$$

belongs to a compact set with probability 1.

Proof of Theorem 1:

The bagging estimator $\hat{\theta}_{nd,bag}$ is a symmetric statistic, provided that $\hat{\boldsymbol{\lambda}}$ is symmetric (Lee 1990). We can project it onto a single dimension, say, $\mathbf{y}_1$. But projections onto other observations are equivalent due to symmetry,

$$\begin{aligned}
&\mathrm{E}\left\{\hat{\theta}_{nd,bag}\middle|\mathbf{y}_1 = \mathbf{y}\right\} \\
&= \mathrm{E}\left\{\frac{1}{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni 1} h\left(\mathbf{y}_1 - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\middle|\mathbf{y}_1 = \mathbf{y}\right\} + \mathrm{E}\left\{\frac{n-1}{n}\frac{1}{\binom{n-1}{k-1}}\sum_{A_b \ni \{i,1\}, i \neq 1} h\left(\mathbf{y}_1 - \hat{\boldsymbol{\lambda}}\left(\mathcal{Y}_b^*\right)\right)\middle|\mathbf{y}_1 = \mathbf{y}\right\} \\
&= \frac{1}{n}u(\mathbf{y}) + \frac{k-1}{n}v(\mathbf{y}).
\end{aligned}$$

Then we can derive the following linearization of bagging estimator using the theory of symmetric statistics,

$$\hat{\theta}_{nd,bag} - \theta_{nd,\infty} = \frac{1}{n}\sum_{i=1}^{n}\left\{u\left(\mathbf{y}_i\right) - \theta_{nd,\infty}\right\} + \frac{k-1}{n}\sum_{i=1}^{n}\left\{v\left(\mathbf{y}_i\right) - \theta_{nd,\infty}\right\} + o_p\left(n^{-1/2}\right),$$

where $u(\cdot)$, $v(\cdot)$ and $\theta_{nd,\infty}$ are defined in Theorem 1. The asymptotic variance (3.3) can be easily derived given the *iid* sampling assumption.

Proof of Theorem 2:

The bagged estimator defined in (2.7) can be treated as a one-sample k-th order U-statistic, with kernel function

$$h\left(y_1, y_2, \ldots, y_k\right) = \inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k}\psi\left(y_i - \gamma\right) \geq 0\right\}.$$

We can directly apply a well-known formula for linearizing U-statistic (Serfling 1980 and van der Vaart 1998, p. 161) to obtain the linearization

$$\hat{\theta}_{ee,bag} - \theta_{ee,\infty} = \frac{k}{n}\sum_{i=1}^{n}\left\{u\left(y_i\right) - \theta_{ee,\infty}\right\} + o_p\left(n^{-1/2}\right),$$

where

$$\begin{aligned} u\left(y\right) &= \mathrm{E}\ h\left(y, y_1, y_2, \ldots, y_{k-1}\right) \\ &= \mathrm{E}\ \inf\left\{\gamma : \frac{1}{k}\sum_{i=1}^{k-1}\psi\left(y_i - \gamma\right) + \frac{1}{k}\psi\left(y - \gamma\right) \geq 0\right\}. \end{aligned}$$

The bagged estimating equation estimator (2.7) can be linearized as

$$\hat{\theta}_{ee,bag} - \theta_{ee,\infty} = \frac{k}{n}\sum_{i=1}^{n}\left\{u\left(y_i\right) - \theta_{ee,\infty}\right\} + o_p\left(n^{-1/2}\right). \tag{A.1}$$

The asymptotic variance of $\hat{\theta}_{ee,bag}$ can be directly obtained from linearization (A.1).

References

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition, New York: John Wiley and Sons.

Bahadur, R.R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37, 577-580.

Beran, R. and Srivastava, M. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, 13, 95-115.

Berger, Y.G. and Skinner, C.J. (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 52 (4), 457-468.

Bickel, P., Götze, F. and van Zwet, W. (1997). Resampling fewer than n observations: gains, losses and remedies for losses. *Statistica Sinica*, 7, 1-31.

Breidt, F. and Opsomer, J. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Annals of Statistics*, 36, 403-427.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30 (4), 927-961.

Buja, A. and Stuetzle, W. (2006). Observations on bagging. *Statistica Sinica*, 16 (2), 323-351.

Chen, S.X. and Hall, P. (2003). Effects of bagging and bias correction on estimators defined by estimating equations. *Statistica Sinica*, 13 (1), 97-109.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95, 125-140.

Dunstan, R. and Chambers, R.L. (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics*, 35, 276-280.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1-26.

Francisco, C.A. and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

Frees, E.W. (1989). Infinite order U-statistics. *Scandinavian Journal of Statistics*, 16, 29-45.

Friedman, J.H. and Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137 (3), 669-683.

Fuller, W. (2009). *Sampling Statistics*. John Wiley and Sons.

Godambe, V. and Thompson, M. (2009). Estimating functions and survey sampling. In C. Rao and D. P. (editors) (Eds.), *Handbook of Statistics, vol. 29: Sample Surveys: Inference and Analysis*, 669-687. Elsevier/North-Holland.

Hall, P. and Robinson, A. (2009). Reducing variability of crossvalidation for smoothing parameter choice. *Biometrika*, 96, 175-186.

Heilig, C.M. (1997). *An Empirical Process Approach to U-processes of Increasing Degree*. Ph. D. thesis, University of California, Berkeley.

Heilig, C.M. and Nolan, D. (2001). Limit theorems for the infinite-degree U-process. *Statistica Sinica*, 11, 289-302.

Inoue, A. and Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, 103 (482), 511-522.

Knight, K. and Bassett, J.G. (2002). Second order improvements of sample quantiles using subsamples. Unpublished manuscript.

Korn, E.L. and Graubard, B.I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 24 (2), 193-201.

Lee, A.J. (1990). *U-statistics: Theory and Practice*. Marcel Dekker Inc.

Lee, T.-H. and Yang, Y. (2006). Bagging binary and quantile predictors for time series. *Journal of Econometrics*, 135 (1-2), 465-497.

Rao, J., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Rao, J. and Wu, C. (1988). Resampling inference with complex surveys. *Journal of the American Statistical Association*, 83 (401), 231-241.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1997). *Model Assisted Survey Sampling*. Springer-Verlag Inc (Berlin; New York).

Schick, A. and Wefelmeyer, W. (2004). Estimating invariant laws of linear processes by U-statistics. *The Annals of Statistics*, 32, 603-632.

Sering, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons.

Shao, J. and Rao, J. (1993). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhya B,* 55, 393-414.

Sitter, R.R. and Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics & Probability Letters*, 52 (4), 353-358.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, J.C. and Opsomer, J.D. (2011). On the asymptotic normality and variance estimation of nondifferentiable survey estimators. *Biometrika*, 98, 91-106.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 211-230
Statistics Canada, Catalogue No. 12-001-X

Fractional hot deck imputation for robust inference under item nonresponse in survey sampling

Jae Kwang Kim and Shu Yang[1]

Abstract

Parametric fractional imputation (PFI), proposed by Kim (2011), is a tool for general purpose parameter estimation under missing data. We propose a fractional hot deck imputation (FHDI) which is more robust than PFI or multiple imputation. In the proposed method, the imputed values are chosen from the set of respondents and assigned proper fractional weights. The weights are then adjusted to meet certain calibration conditions, which makes the resulting FHDI estimator efficient. Two simulation studies are presented to compare the proposed method with existing methods.

Key Words: EM algorithm; Kullback-Leibler information; Missing at random (MAR); Multiple imputation.

1 Introduction

Imputation is a popular method of compensating for item non-response in sample surveys. Let y be the study variable subject to non-response and $\mathbf{x}$ be the vector of auxiliary variables fully observed. A model on the conditional distribution $f(y|\mathbf{x})$ is often used to generate imputed values for missing y_i. Such model-based imputation method is well developed in the literature. Multiple imputation of Rubin (1987) is a Bayesian approach of model-based imputation. Monte Carlo EM of Wei and Tanner (1990) can be treated as a frequentist's approach of model-based imputation. Kim (2011) proposed parametric fractional imputation to handle multivariate missing data.

However, the model-based imputation method that generates imputed values from $f(y|\mathbf{x})$ is not a hot deck imputation in the sense that artificial values are constructed after the imputation. A desirable property of hot deck imputation is that all imputed values are observed values. For example, imputed values for categorical variables will also be categorical with the same number of categories as observed for the respondents. For this reason, hot deck imputation is the most popular imputation method, especially in household surveys. Nearest neighbor imputation method is also a hot deck imputation. Chen and Shao (2001), Beaumont and Bocci (2009), Kim, Fuller and Bell (2011) investigated nearest neighbor imputation in the context of survey sampling. Durrant (2009), Haziza (2009) and Andridge and Little (2010) provided comprehensive overviews of the hot-deck imputation methods in survey sampling.

Fractional hot deck imputation was proposed by Kalton and Kish (1984) to achieve efficiency in hot deck imputation. Kim and Fuller (2004) and Fuller and Kim (2005) provided a rigorous treatment of fractional hot deck imputation and discussed variance estimation. However, their approach is only applicable when $\mathbf{x}$ is categorical. For continuous covariate case, predictive mean matching can be treated as a nearest neighbor imputation method using the predicted value obtained from $f(y|\mathbf{x})$ but its statistical properties are not fully addressed in the literature.

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011. E-mail: jkim@iastate.edu; Shu Yang, Department of Statistics, Iowa State University, Ames, IA 50011.

In this paper, we propose a new fractional hot deck imputation (FHDI) method based on a parametric model of $f(y|\mathbf{x})$ that allows continuous covariates. The proposed method has several advantages over the existing methods. First, it is a hot deck imputation preserving the correlation structure between the items. Second, it is robust in that the resulting estimator is less sensitive against the failure of the assumed model $f(y|\mathbf{x})$. Third, it provides consistent variance estimators for various parameters without requiring the congeniality condition of Meng (1994). Multiple imputation, however, requires the congeniality condition for the validity of variance estimation. When the congeniality condition does not hold, multiple imputation often leads to conservative inference, which in turn reduces test powers. See Section 5.2 for more details.

The paper is organized as follows. Section 2 describes the basic setup. The proposed method is presented in Section 3. The robustness of FHDI is discussed in Section 4. Results from two simulation studies are presented in Section 5 before some concluding remarks are made in Section 6.

2 Basic setup

Consider a finite population of N elements identified by a set of indices $U=\{1,2,\ldots,N\}$ with N known. Associated with each unit i in the population are study variables, $\mathbf{x}_i$ and y_i, with $\mathbf{x}_i$ always observed and y_i subject to non-response. Let A denote the set of indices for the elements in a sample selected by a probability sampling mechanism. We are interested in estimating η, defined as a (unique) solution to the population estimating equation $\sum_{i=1}^{N} U(\eta;\mathbf{x}_i,y_i)=0$. For example, a population mean can be obtained by letting $U(\eta;\mathbf{x}_i,y_i)=\eta-y_i$. Under complete response, a consistent estimator of η is obtained by solving

$$\sum_{i\in A} w_i U(\eta;\mathbf{x}_i,y_i)=0, \tag{2.1}$$

where $w_i=\{Pr(i\in A)\}^{-1}$ is the inverse of the first-order inclusion probability of unit i. Binder and Patak (1994) and Rao, Yung and Hidiroglou (2002) considered the asymptotic properties of the estimator obtained from (2.1). Under the existence of missing data, we define

$$\delta_i=\begin{cases}1 & \text{if } y_i \text{ is observed}\\ 0 & \text{otherwise.}\end{cases}$$

A consistent estimator of η is then obtained by taking the conditional expectation and solving

$$\sum_{i\in A} w_i\left[\delta_i U(\eta;\mathbf{x}_i,y_i)+(1-\delta_i)E\{U(\eta;\mathbf{x}_i,Y)\,|\,\mathbf{x}_i,\delta_i=0\}\right]=0 \tag{2.2}$$

for η. Estimating equation (2.2) is sometimes referred to as expected estimating equation (Wang and Pepe 2000).

To compute the conditional expectation in (2.2), we assume that the finite population at hand is a realization from an infinite population, called superpopulation. In the superpopulation model, we often postulate a parametric conditional distribution of y given $\mathbf{x}$, $f(y|\mathbf{x};\theta)$, which is known up to the parameter θ with parameter space Ω. Under the specified model, we can compute a consistent estimator

$\hat{\theta}$ of θ and then use a Monte Carlo method to evaluate the conditional expectation in (2.2) given the estimate $\hat{\theta}$. If the response mechanism is missing at random (MAR) or ignorable in the sense of Rubin (1976), we can approximate the expected estimating equation in (2.2) by

$$\sum_{i\in A} w_i \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \frac{1}{m} \sum_{j=1}^{m} U\left(\eta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{2.3}$$

where

$$y_i^{*(1)}, \ldots, y_i^{*(m)} \overset{i.i.d.}{\sim} f\left(y_i \mid \mathbf{x}_i; \hat{\theta}\right).$$

Often, we use the maximum likelihood estimator $\hat{\theta}$, which solves

$$S\left(\theta\right) = \sum_{i\in A} w_i \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) = 0, \tag{2.4}$$

where $S\left(\theta; \mathbf{x}, y\right) = \partial \log f\left(y \mid \mathbf{x}; \theta\right) / \partial\theta$. Note that we use the sampling weights w_i in the score equation (2.4). Thus, we are implicitly assuming that the imputation model, the model for generating the imputed values, is the model about the finite population values $f\left(y_i \mid \mathbf{x}_i\right)$, not the model about the sample values. Thus, we allow that the sampling mechanism can be informative in the sense of Pfeffermann (2011). Multiple imputation, on the other hand, uses the sample model, $f_s\left(y_i \mid \mathbf{x}_i\right) \equiv f\left(y_i \mid \mathbf{x}_i, i \in A\right)$, to generate the imputed values and often assumes that the sampling mechanism is non-informative. Thus, in multiple imputation, MAR is assumed for the sample at hand, while, in fractional imputation, MAR is assumed for the population. Under informative sampling design, generating imputed values from the sample model $f_s\left(y_i \mid \mathbf{x}\right)$ does not necessarily lead to valid inference even when sample MAR condition holds. See Section 8.4 of Kim and Shao (2013) for further discussion of MAR under informative sampling.

To compute the conditional expectation in (2.2) efficiently, the parametric fractional imputation (PFI) of Kim (2011) can be used. In PFI, the imputed values are generated from a suitable proposal distribution $h\left(y \mid \mathbf{x}_i\right)$ and then the imputed estimating equation (2.3) is changed to

$$\sum_{i\in A} w_i \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j=1}^{m} w_{ij}^{*} U\left(\eta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{2.5}$$

where

$$w_{ij}^{*} = \frac{f\left(y_i^{*(j)} \mid \mathbf{x}_i; \hat{\theta}\right) \Big/ h\left(y_i^{*(j)} \mid \mathbf{x}_i\right)}{\sum_{k=1}^{m} \left\{ f\left(y_i^{*(k)} \mid \mathbf{x}_i; \hat{\theta}\right) \Big/ h\left(y_i^{*(k)} \mid \mathbf{x}_i\right) \right\}}. \tag{2.6}$$

The choice of the proposal distribution $h\left(\cdot\right)$ is somewhat arbitrary. We will discuss a particular choice that may lead to a robust estimation.

The consistency of the resulting estimator $\hat{\eta}$ from (2.3) or (2.5) can be established under the assumption that the conditional distribution $f\left(y \mid \mathbf{x}; \theta\right)$ is correctly specified (by similar argument in the proof of Corollary II.2 of Andersen and Gill (1982) and its proof is skipped here). In this paper, we consider an alternative approach of fractional imputation that is more robust against the failure of the assumption on the imputation model.

3 Proposed method

We first consider a particular fractional hot deck imputation method, called **full fractional imputation**, where the imputed values are taken from the set of respondents denoted as $A_R = \{i \in A; \delta_i = 1\}$. That is, the j-th imputed value of missing y_i, denoted by $y_i^{*(j)}$, is equal to the j-th value of y among the set in A_R. We propose a fractional hot deck imputation approach that makes use of the parametric model assumption $f(y \mid \mathbf{x}; \theta)$. If all of the elements in A_R are selected as the imputed values for missing y_i, we can treat $\{y_j; j \in A_R\}$ as a realization from $f(y_j \mid \delta_j = 1)$ and fractional weight assigned to donor y_j for the missing item y_i is, by choosing $h(y_j \mid \mathbf{x}_i) = f(y_j \mid \delta_j = 1)$ in (2.6),

$$\begin{aligned} w_{ij}^* &\propto f(y_j \mid \mathbf{x}_i, \delta_i = 0; \hat{\theta}) \big/ f(y_j \mid \delta_j = 1) \\ &\propto f(y_j \mid \mathbf{x}_i, \hat{\theta}) \big/ f(y_j \mid \delta_j = 1), \end{aligned} \tag{3.1}$$

with $\sum_{j;\delta_j=1} w_{ij}^* = 1$, and $\hat{\theta}$ being the MLE obtained from (2.4). The second line follows from the MAR assumption. Furthermore, we can write

$$\begin{aligned} f(y_j \mid \delta_j = 1) &= \int f(y_j \mid \mathbf{x}, \delta_j = 1) f(\mathbf{x} \mid \delta_j = 1) d\mathbf{x} \\ &= \int f(y_j \mid \mathbf{x}) f(\mathbf{x} \mid \delta_j = 1) d\mathbf{x} \\ &\cong \frac{1}{N_R} \sum_{k=1}^{N} \delta_k f(y_j \mid \mathbf{x}_k), \end{aligned} \tag{3.2}$$

where the second equality follows from the MAR assumption, and the last (approximate) equality follows by approximating the integral by the population empirical distribution, and N_R is the number of respondents in the population. Using the survey weights, we can approximate

$$f(y_j \mid \delta_j = 1) \cong \frac{\sum_{k \in A_R} w_k f(y_j \mid \mathbf{x}_k)}{\sum_{k \in A_R} w_k}$$

and the fractional weights in (3.1) are computed from

$$w_{ij}^* \propto \frac{f(y_j \mid \mathbf{x}_i; \hat{\theta})}{\sum_{k \in A_R} w_k f(y_j \mid \mathbf{x}_k; \hat{\theta})} \tag{3.3}$$

with $\sum_{j \in A_R} w_{ij}^* = 1$. In (3.3), the point mass w_{ij}^* assigned to donor y_j for missing unit i is expressed by the ratio of the density $f(y \mid \mathbf{x})$. Thus, for each missing unit i, $n_R = |A_R|$ observations are used as donors for the hot deck imputation using w_{ij}^* as the fractional weights. Such fractional imputation can be called full fractional imputation (FFI) because there is no randomness due to the imputation mechanism. The FFI estimator of η, defined by $\sum_{i=1}^{N} U(\eta; \mathbf{x}_i, y_i) = 0$, is then computed by solving

$$\sum_{i \in A} w_i \left\{ \delta_i U(\eta; \mathbf{x}_i, y_i) + (1 - \delta_i) \sum_{j \in A_R} w_{ij}^* U(\eta; \mathbf{x}_i, y_j) \right\} = 0, \tag{3.4}$$

where w_{ij}^* is defined in (3.3). Note that the imputed estimating equation (3.4) is a good approximation to the expected estimating equation in (2.2).

In survey sampling, an imputed data set with a large imputation size may not be desirable. Thus, instead of taking all the observations in A_R as donors for each missing item, a subset of A_R can be selected to reduce the size of the donor set of missing y_i. Thus, the selection of the donors is viewed as a sampling problem and we use an efficient sampling design and weighting techniques to obtain efficient imputation estimators. For the donor selection mechanism, efficient sampling designs, such as a stratified sampling design or systematic Proportional-to-Size (PPS) sampling, can be used to select donors of size m. A systematic PPS sampling for fractional hot deck imputation can be described as follows:

1. Within each i with $\delta_i = 0,$ sort the donors in the full respondent set $\{y_j; \delta_j = 1\}$ in ascending order as $y_{(1)} \le \cdots \le y_{(r)}$ and use $w_{i(j)}^*$ to denote the fractional weight associated with $y_{(j)}$. That is, $w_{i(j)}^* = w_{ik}^*$ for $y_{(j)} = y_k$.

2. Partition $[0,1]$ by $\left\{I_j \equiv \left[\sum_{k=0}^{j} w_{i(j)}^*, \sum_{k=0}^{j+1} w_{i(j)}^*\right), j = 1,\ldots,r-1\right\},$ where $w_{i(0)}^* = 0.$

3. Generate $u \sim \text{uniform}(0,1/m)$ and let $u_k = u + k/m,\ \ k = 0,\ldots,m-1.$ For $k = 0,\ldots,m-1,$ if $u_k \in I_j$ for some $0 \le j \le r-1,$ include j in the sample D_i.

After we select D_i from the complete set of respondents, the selected donors in D_i are assigned with the initial fractional weights $w_{ij0}^* = 1/m$. The fractional weights are further adjusted to satisfy

$$\sum_{i \in A} w_i \left\{(1-\delta_i) \sum_{j \in D_i} w_{ij,c}^* \mathbf{q}(\mathbf{x}_i, y_j)\right\} = \sum_{i \in A} w_i \left\{(1-\delta_i) \sum_{j \in A_R} w_{ij}^* \mathbf{q}(\mathbf{x}_i, y_j)\right\}, \tag{3.5}$$

for some $\mathbf{q}(\mathbf{x}_i, y_j)$, and $\sum_{j \in D_i} w_{ij,c}^* = 1$ for all i with $\delta_i = 0,$ where w_{ij}^* is the fractional weights for FFI method, as defined in (3.3). Regarding the choice of the control function $\mathbf{q}(\mathbf{x}, y)$ in (3.5), we can use $\mathbf{q}(\mathbf{x}, y) = (y, y^2)'$, which keeps the empirical distributions of y for D_i and A_R as close as possible in the sense that the first and second moment of y are the same. Other choices can also be considered. See Fuller and Kim (2005).

The problem of adjusting the initial weights to satisfy certain constraints is often called calibration and the resulting fractional weights can be called calibrated fractional weights. Using the idea of regression weighting, the final calibration fractional weights that satisfy (3.5) and $\sum_j w_{ij,c}^* = 1$ can be computed by

$$w_{ij,c}^* = w_{ij0}^* + w_{ij0}^* \Delta \left(\mathbf{q}_{ij}^* - \overline{\mathbf{q}}_{i\cdot}^*\right), \tag{3.6}$$

where $\mathbf{q}_{ij}^* = \mathbf{q}(\mathbf{x}_i, y_j),\ \overline{\mathbf{q}}_{i\cdot}^* = \sum_{j \in A_R} w_{ij0}^* \mathbf{q}_{ij}^*,$

$$\Delta = \left\{C_q - \sum_{i \in A} w_i (1-\delta_i) \sum_{j \in A_R} w_{ij0}^* \mathbf{q}_{ij}^*\right\}^T \left\{\sum_{i \in A} w_i (1-\delta_i) \sum_{j \in A_R} w_{ij0}^* \left(\mathbf{q}_{ij}^* - \overline{\mathbf{q}}_{i\cdot}^*\right)^{\otimes 2}\right\}^{-1}$$

and $C_q = \sum_{i\in A} w_i \left\{(1-\delta_i)\sum_{j\in A_R} w_{ij}^* \mathbf{q}(\mathbf{x}_i, y_j)\right\}$. Here, $B^{\otimes 2}$ denotes BB^T. Some of the fractional weights computed by (3.6) can take negative values. If that happens, algorithms alternative to regression weighting should be used. For example, consider entropy weighting, where the fractional weights of the form

$$w_{ij,c}^* = \frac{w_{ij}^* \exp\left(\Delta \mathbf{q}_{ij}^*\right)}{\sum_{k\in A_R} w_{ik}^* \exp\left(\Delta \mathbf{q}_{ik}^*\right)} \tag{3.7}$$

are approximately equal to the regression fractional weights in (3.6) and are always positive. Once the calibration fractional weights are obtained, the FHDI estimator of η is then computed by solving

$$\sum_{i\in A} w_i \left\{\delta_i U(\eta; \mathbf{x}_i, y_i) + (1-\delta_i)\sum_{j\in D_i} w_{ij,c}^* U(\eta; \mathbf{x}_i, y_j)\right\} = 0. \tag{3.8}$$

For variance estimation, a replication method can be used. See Appendix A.1 for a brief discussion of the replication variance estimator for the proposed method.

Furthermore, the proposed method can handle non-ignorable non-response under the correct specification of the response model. See Appendix A.3 for the extension to non-ignorable non-response case.

4 Robustness

We now discuss the robustness of the proposed method against a small departure from the assumed parametric model. The robustness feature in our proposed estimator is defined to be robust against imputation model misspecification, a small exponential tilting of the true model. For simplicity of the presentation, assume that the sampling design is simple random sampling and the realized sample is a random sample from the superpopulation model.

We assume that the true model $g(y|x)$ does not belong to $\{f(y|x;\theta); \theta\in\Omega\}$. However, we can still specify a working model $f(y|x;\theta)$ and compute the MLE of θ. It is well known (White 1982) that the MLE converges to θ^*, the minimizer of the Kullback-Leibler information

$$K(\theta) = E_g\left[\log\left\{\frac{g(Y|x)}{f(Y|x;\theta)}\right\}\right]$$

for $\theta\in\Omega$. Sung and Geyer (2007) discussed the asymptotic properties of the Monte Carlo MLE of θ under missing data.

To formally discuss robustness, suppose that the true distribution $g(y|x)$ belongs to the neighborhood

$$\mathcal{N}_\varepsilon = \left\{g; D(g,f) < \frac{1}{2}\varepsilon^2\right\} \tag{4.1}$$

for some radius $\varepsilon > 0$, where

$$D(g,f) = \int log\left(\frac{g}{f}\right) g\, dy, \tag{4.2}$$

is the Kullback-Leibler distance measure. The neighborhood (4.1) can be characterized in the following way. Let $z(x,y,\theta)$ be a function of x, y and θ, standardized to satisfy $E_{Y|x}(z)=0$ and $Var_{Y|x}(z)=1$, and define

$$g(y\,|\,x)=f(y\,|\,x;\theta)\exp\{\varepsilon z(x,y,\theta)-\kappa(x,\theta)\}, \tag{4.3}$$

where

$$\kappa=\log\left(E_{Y|x}\left[\exp\{\varepsilon z(x,Y,\theta)\}\right]\right).$$

For small $\varepsilon>0$ it can be shown that

$$\kappa\cong D(g,f)\cong\frac{1}{2}\varepsilon^2. \tag{4.4}$$

Equation (4.3) represents an extensive set of distributions close to $f(y\,|\,x;\theta)$ created by varying $z(x,y,\theta)$ over different standardized functions, where z and ε contain some geometric interpretation which represent the direction and magnitude of the misspecification respectively. For p-dimension parameter θ, we can specify the directions of the misspecification as

$$\left(z_1,z_2,\ldots,z_p\right)^T=I_\theta^{-1/2}s(x,y,\theta),$$

where $s(x,y,\theta)=\partial\log f(y\,|\,x;\theta)/\partial\theta$ and I_θ is the information matrix for θ. Represent $z(x,y,\theta)$ as

$$z(x,y,\theta)=\lambda^T I_\theta^{-1/2}s(x,y,\theta),$$

where $\sum_{i=1}^{p}\lambda_i^2=1$, then $z(x,y,\theta)$ satisfies the standardization criterion of $E_{Y|x}(z)=0$ and $Var_{Y|x}(z)=1$. See Copas and Eguchi (2001) for further discussion of this expression.

Let $w_{ij,g}^*$ be the fractional weight of the form (3.3) using the true density g and $w_{ij,f}^*$ be the corresponding fractional weight using the "working density" f. By the special construction of the weights, we can establish

$$w_{ij,g}^*\cong w_{ij,f}^*+\varepsilon\lambda^T I_\theta^{-1/2}\frac{\partial}{\partial\theta}\left(w_{ij,f}^*\right). \tag{4.5}$$

Proof of (4.5) is given in Appendix A.2. Thus

$$\begin{aligned}\sum_i w_i\sum_j w_{ij,g}^*U(\eta;x_i,y_j)\cong\;&\sum_i w_i\sum_j w_{ij,f}^*U(\eta;x_i,y_j)\\&+\varepsilon\lambda^T I_\theta^{-1/2}\sum_i w_i\sum_j\frac{\partial}{\partial\theta}\left(w_{ij,f}^*\right)U(\eta;x_i,y_j).\end{aligned} \tag{4.6}$$

For small ε, we have

$$\sum_i w_i\sum_j w_{ij,g}^*U(\eta;x_i,y_j)\cong\sum_i w_i\sum_j w_{ij,f}^*U(\eta;x_i,y_j),$$

and so the resulting estimator of η from $\sum_i w_i\sum_j w_{ij,f}^*U(\eta;x_i,y_j)=0$ will be close to the true value η_0.

5 Simulation study

We performed two simulation studies. In Section 5.1, we compared the performance of the proposed method with some other imputation methods in a correctly specified model and a misspecified model, respectively, with ignorable missing data. In Section 5.2, we compared the statistical power of a test based on FHDI versus MI.

5.1 Simulation one

The first simulation study tested the performance of the proposed method under the setup of ignorable missing data. We used two sets of models to generate the observations. In model A, $y_i = 0.5x_i + e_i$, where $x_i \sim \exp(1)$, $e_i \sim N(0,1)$, with x_i and e_i being independent. In model B, $y_i = 0.5x_i + e_i$, where $x_i \sim \exp(1)$, $e_i \sim \{\chi^2(2) - 2\}/2$, with x_i and e_i being independent. Random samples of size $n = 200$ were separately generated from the two models. In addition to (x_i, y_i), we also generated δ_i from Bernoulli(π_i), where $\pi_i = \{1 + \exp(-0.2 - x_i)\}^{-1}$. Variable x_i was always observed but variable y_i was observed if and only if $\delta_i = 1$. The overall response rates were about 65% in both cases. We used $B = 2,000$ Monte Carlo samples in the simulation.

From each of the Monte Carlo samples, one generated from model A and the other generated from model B, we computed the following eight estimators:

1. Full sample estimator (Full) that is computed using the full sample.

2. Predictive Mean Matching (PMM) is a semi-parametric imputation method, which fills in a value randomly from observations that are closest to the predicted value obtained from $f(y \mid \mathbf{x})$. The PMM was implemented using "mice.impute.pmm" function in R.

3. Multiple imputation (MI) estimator with imputation size $m = 10$, where the imputed values are generated from the normal-theory regression model, as considered in Schenker and Welsh (1988).

4. Parametric fractional imputation (PFI) estimator without calibration with imputation size $m = 10$.

5. Parametric fractional imputation (PFI_cal) estimator with calibration with imputation size $m = 10$. The fractional weights are computed using the calibration method in (3.6) with $\mathbf{q} = (y, y^2)$.

6. Full fractional imputation (FFI) estimator using the full set of respondents as imputation values, i.e. the imputation size $m = n_R$, where n_R is the size of A_R.

7. Fractional hot deck imputation (FHDI) estimator without calibration using a small subset of respondents of size $m = 10$ as imputation values.

8. Fractional hot deck imputation (FHDI_cal) estimator with calibration using a small subset of respondents of size $m=10$ as imputation values. The fractional weights are computed using the calibration method in (3.6) with $\mathbf{q}=\left(y,y^2\right)$.

Multiple imputation is an approach of generating imputed values with simplified variance estimation. In this procedure, Bayesian methods of generating imputed values are considered, where $m>1$ imputed values are generated from the posterior predictive distribution. Using the imputed values $\mathbf{y}^{*(1)},\ldots,\mathbf{y}^{*(m)}$, the multiple imputation estimator of η, denoted by $\hat{\eta}_{MI}$ is

$$\hat{\eta}_{MI}=\frac{1}{m}\sum_{k=1}^{m}\hat{\eta}^{(k)}$$

where $\hat{\eta}^{(k)}$ is the complete response estimator applied to the k-th imputed data set. Rubin's formula can be used for variance estimation in MI,

$$\hat{V}_{MI}\left(\hat{\eta}_{MI}\right)=W_m+\left(1+\frac{1}{m}\right)B_m, \tag{5.1}$$

where $W_m=m^{-1}\sum_{k=1}^{m}\hat{V}^{(k)}$, $B_m=\left(m-1\right)^{-1}\sum_{k=1}^{m}\left(\hat{\eta}^{(k)}-\hat{\eta}_{MI}\right)^2$, and $\hat{V}^{(k)}$ is the variance estimator of $\hat{\eta}^{(k)}$ under complete response applied to the k-th imputed data set.

In both models, we used the normal density with mean $\beta_0+\beta_1 x$ and variance σ^2 as the working model for imputation. Thus, the working model is the true model in model A but not true in model B.

We considered three parameters: $\theta_1=E\left(Y\right)$, the population mean of y, $\theta_2=Pr\left(Y<1\right)$, the proportion of Y less than one, and θ_3, the 0.5 quantile of Y. In estimating θ_2 under full sample, we used $\hat{\theta}_{2,n}=n^{-1}\sum_{i=1}^{n}I\left(y_i<1\right)$. In estimating θ_3 under full sample, we used $\hat{\theta}_{3,n}=\hat{F}^{-1}\left(p\right)=\inf\left\{y:\hat{F}\left(y\right)>p\right\}$, where $\hat{F}\left(y\right)=n^{-1}\sum_{i=1}^{n}I\left(y_i<y\right)$ and $p=0.5$.

Table 5.1 and Table 5.2 show Monte Carlo means, standardized variance (Std Var) and standardized mean squared errors (Std MSE) of the eight estimators under model A and under model B, respectively. The standardized variance (mean squared error) is calculated as the ratio of variance (mean squared error) and the variance (mean squared error) of the full sample estimator multiplied by 100, which measures the increased variance (mean squared error) due to imputation relative to the full sample estimator. As for the Monte Carlo means (4th column), the imputation estimators are all unbiased for estimating θ_1, θ_2, and θ_3 under model A. Under model B, PMM, MI, PFI, PFI_cal for estimating θ_3 have much larger biases in absolute values than FFI, FHDI, and FHDI_cal under model misspecification in this simulation. Regarding the standardized variance and standardized mean squared error (5th and 6th column), PFI is more efficient than FHDI. The reason is that in PFI, the imputed values are generated according to the conditional distribution $f\left(y\,|\,x\right)$ directly; whereas in FHDI, the imputed values can be taken from respondents with dominantly large fractional weights. The effective imputation data size is determined by the imputed observations with large fractional weights, which also contribute to the loss of efficiency. FHDI loses efficiency in order to gain robustness. Lastly, FHDI with $m=10$ has slightly larger standardized variance for θ_2 than FFI, because of the additional variability due to the sampling procedure. Comparing PFI with PFI_cal and FHDI with FHDI_cal, the calibration step improves the efficiency a little bit. The PMM shows the largest variance in all scenarios.

Table 5.1
Monte Carlo mean, standardized variance and standardized mean squared error of point estimators in Model A of Simulation one.

Model	Parameter	Method	Mean	Std Var	Std MSE
A	μ_y	Full	0.50	100	100
		PMM	0.50	175	175
		MI ($m=10$)	0.50	135	135
		PFI ($m=10$)	0.50	130	130
		PFI cal ($m=10$)	0.50	130	130
		FFI ($m=n_R$)	0.50	130	130
		FHDI ($m=10$)	0.50	156	156
		FHDI cal ($m=10$)	0.50	130	130
	$Pr(Y<1)$	Full	0.68	100	100
		PMM	0.68	168	167
		MI ($m=10$)	0.68	112	112
		PFI ($m=10$)	0.68	110	110
		PFI cal ($m=10$)	0.68	109	109
		FFI ($m=n_R$)	0.68	130	130
		FHDI ($m=10$)	0.68	137	136
		FHDI cal ($m=10$)	0.68	132	132
	Quantile	Full	0.47	100	100
		PMM	0.47	184	184
		MI ($m=10$)	0.47	111	111
		PFI ($m=10$)	0.47	111	111
		PFI cal ($m=10$)	0.47	111	111
		FFI ($m=n_R$)	0.47	135	135
		FHDI ($m=10$)	0.47	142	142
		FHDI cal ($m=10$)	0.47	141	141

Table 5.2
Monte Carlo mean, standardized variance and standardized mean squared error of point estimators in Model B of Simulation one.

Model	Parameter	Method	Mean	Std Var	Std MSE
B	μ_y	Full	0.50	100	100
		PMM	0.50	172	172
		MI ($m=10$)	0.50	131	131
		PFI ($m=10$)	0.50	131	131
		PFI cal ($m=10$)	0.50	128	128
		FFI ($m=n_R$)	0.50	127	127
		FHDI ($m=10$)	0.50	147	147
		FHDI cal ($m=10$)	0.50	127	127
	$Pr(Y<1)$	Full	0.75	100	100
		PMM	0.75	166	166
		MI ($m=10$)	0.73	140	170
		PFI ($m=10$)	0.73	138	168
		PFI cal ($m=10$)	0.73	137	169
		FFI ($m=n_R$)	0.75	137	137
		FHDI ($m=10$)	0.75	145	145
		FHDI cal ($m=10$)	0.75	140	141
	Quantile	Full	0.26	100	100
		PMM	0.24	191	198
		MI ($m=10$)	0.31	122	159
		PFI ($m=10$)	0.31	123	160
		PFI cal ($m=10$)	0.31	122	159
		FFI ($m=n_R$)	0.26	135	135
		FHDI ($m=10$)	0.26	144	144
		FHDI cal ($m=10$)	0.26	139	139

For variance estimation, we considered replication variance estimation for FFI and FHDI, particularly the delete-1 Jackknife variance estimation, which is described in Appendix A.1. We also considered variance estimation in MI, which uses Rubin's formula (5.1).

Table 5.3 shows the Monte Carlo relative biases of the variance estimators, which is calculated as $\left[E_{MC}\left\{\hat{V}\right\}-V_{MC}\left\{\hat{\theta}\right\}\right]\Big/V_{MC}\left\{\hat{\theta}\right\}$, where $E_{MC}\left\{\hat{V}\right\}$ is the Monte Carlo mean of variance estimates $\hat{V}$, and $V_{MC}\left\{\hat{\theta}\right\}$ is the Monte Carlo variance of the point estimates $\hat{\theta}$. The relative bias of the variance estimator in FFI and FHDI is reasonably small for all parameters considered in both models, suggesting that the replication variance estimator is valid. The relative bias and $t-$statistics of variance estimator in MI are small for θ_1 but quite large for θ_2 even when the working model is true (model A). Rubin's formula is based on the following decomposition,

$$V\left(\hat{\theta}_{MI}\right)=V\left(\hat{\theta}_n\right)+V\left(\hat{\theta}_{MI}-\hat{\theta}_n\right), \tag{5.2}$$

where $\hat{\theta}_n$ is the full sample estimator of η. Basically, the W_m term in (5.1) estimates $V\left(\hat{\theta}_n\right)$ and the $\left(1+m^{-1}\right)B_m$ term in (5.1) estimates $V\left(\hat{\theta}_{MI}-\hat{\theta}_n\right)$. The decomposition (5.2) holds when $\hat{\theta}_n$ is the MLE of θ, which is the congeniality condition of $\hat{\theta}_n$ (Meng 1994). For general case, we have

$$V\left(\hat{\theta}_{MI}\right)=V\left(\hat{\theta}_n\right)+V\left(\hat{\theta}_{MI}-\hat{\theta}_n\right)+2Cov\left(\hat{\theta}_{MI}-\hat{\theta}_n,\hat{\theta}_n\right) \tag{5.3}$$

and Rubin's variance estimator can be biased if $Cov\left(\hat{\theta}_{MI}-\hat{\theta}_n,\hat{\theta}_n\right)\neq 0$. The congeniality condition holds true for estimating the population mean; however, it does not hold for the method of moments estimator of $Pr\left(Y<1\right)$. Note that the imputed estimator of $\theta_2=Pr\left(Y<1\right)$ can be expressed as

$$\hat{\theta}_{2,I}=n^{-1}\sum_{i=1}^{n}\left[\delta_i I\left(y_i<1\right)+\left(1-\delta_i\right)E\left\{I\left(y_i<1\right)\mid x_i;\hat{\mu},\hat{\sigma}\right\}\right]. \tag{5.4}$$

Thus, the imputed estimators of θ_2 "borrows strength" by making use of extra information associated with $f\left(y\mid x\right)$. That is, the normality of $f\left(y\mid x\right)$ is used in computing the conditional expectation in (5.4), which improves the efficiency of the imputed estimator for θ_2. The same phenomenon also holds for θ_3. In Table 5.1, the increase of variance due to imputation for MI with $m=10$ is about 35 % for θ_1 but only 12% and 11% for θ_2 and θ_3, respectively, which shows the phenomenon of "borrowing strength" for estimating θ_2 and θ_3 thanks to the use of extra information in the imputation stage. Thus, when the congeniality conditions do not hold, the imputed estimator improves the efficiency but Rubin's variance estimator does not recognize this improvement.

Table 5.3
Monte Carlo relative bias of the replication variance estimator in Simulation one.

Model	Parameter	Method	R.B. (%)
*A	$V(\hat{\theta}_1)$	MI ($m=10$)	-2.33
		FFI ($m=n_R$)	-0.80
		FHDI_cal ($m=10$)	-0.80
	$V(\hat{\theta}_2)$	MI ($m=10$)	8.20
		FFI ($m=n_R$)	-5.01
		FHDI_cal ($m=10$)	-5.12
	$V(\hat{\theta}_3)$	MI ($m=10$)	19.84
		FFI ($m=n_R$)	4.50
		FHDI_cal ($m=10$)	3.78
*B	$V(\hat{\theta}_1)$	MI ($m=10$)	2.60
		FFI ($m=n_R$)	-0.56
		FHDI_cal ($m=10$)	-0.56
	$V(\hat{\theta}_2)$	MI ($m=10$)	-3.33
		FFI ($m=n_R$)	-1.89
		FHDI_cal ($m=10$)	-3.25
	$V(\hat{\theta}_3)$	MI ($m=10$)	-8.99
		FFI ($m=n_R$)	3.50
		FHDI_cal ($m=10$)	3.80

5.2 Simulation two

Simulation two tested the power of the proposed method in a hypothesis test using the null model as the imputation model. Samples of bivariate data (x_i, y_i) of size $n=100$ were generated from

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \left(x_i^2 - 1\right) + e_i \tag{5.5}$$

where $(\beta_0, \beta_1, \beta_2) = (0, 0.9, 0.06)$, $x_i \sim N(0,1)$, $e_i \sim N(0, 0.16)$, with x_i and e_i being independent. The variable x_i is always observed but the probability that y_i responds is 0.5. Monte Carlo samples were generated independently for $B=10,000$ times. We are interested in testing $H_0: \beta_2 = 0$ from the respondents. We compared FHDI with MI using the same imputation size $m=30$. The imputation model is the null model,

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

That is, the imputation model uses extra information of $\beta_2 = 0$. From the imputed data, we fit model (5.5) and computed the power of a test $H_0: \beta_2 = 0$ at the significant level of 0.05. In addition, we also considered the complete case (CC) method that only uses the respondents for regression.

Table 5.4 shows the Monte Carlo mean and variance of the point estimators, relative bias of the variance estimator and the Monte Carlo power of testing $H_0: \beta_2 = 0$. In each Monte Carlo sample, we constructed a 95% Wald confidence interval of β_2 as $\left(\hat{\beta}_2 - 1.96\hat{V}^{1/2}, \hat{\beta}_2 + 1.96\hat{V}^{1/2}\right)$ and reject the null hypothesis if $\beta_2 = 0$ does not fall in the Wald confidence interval. The Monte Carlo power is calculated as

the relative frequency of rejecting the null hypothesis among the Monte Carlo samples. From the second column, FHDI and MI estimators are biased for β_2, as expected since in imputation the imputation model is the null model and it is slightly different from the true model that generated sample. The bias of FHDI is smaller than that of MI because of the robustness of FHDI discussed in Section 4. In MI, 50% of the imputed MI data comes from the null model and the other 50% from the true model, so the slope β_2 is attenuated to zero by half of the true slope. In FHDI, though we used the null model to calculate the fractional weights, the imputed data come from the true model which reduces the bias. Moreover, MI provides more efficient point estimators than the CC method but variance estimation is very conservative (about 180% overestimation). Because of the serious positive bias of MI variance estimator, the statistical power of the test based on MI is actually lower than the CC method. On the other hand, FHDI also provides more efficient point estimators than the CC method and variance estimation is essentially unbiased, the statistical power of the test based on FHDI is higher than the CC method.

Table 5.4
Simulation results based on 10,000 Monte Carlo samples in Simulation two.

Method	$E\left(\hat{\beta}_2\right)$	$V\left(\hat{\beta}_2\right)$	R.B.$\left(\hat{V}\right)$	Power
FHDI	0.046	0.00146	0.02	0.314
MI	0.028	0.00056	1.81	0.044
CC	0.060	0.00234	-0.01	0.285

6 Concluding remarks

We have proposed a fractional hot deck imputation method that uses a parametric model for $f\left(y \mid \mathbf{x}\right)$ when $\mathbf{x}$ contains continuous components. The proposed method provides robust estimation for the parameters in the sense that the imputation model is not necessarily equal to the data-generating model. The price we pay in the FHDI is the loss of efficiency in point estimation. Under our first simulation, the FHDI estimator for $P\left(Y<1\right)$ has the second largest variance but the smallest mean squared error when the working model is not true, as compared with other estimators.

The loss of efficiency mainly comes from the fact that the fractional weights are more variable than those under the PFI method because some of $\mathbf{x}_j$ are not useful in imputing y_i. That is, the value of $f\left(y_i \mid \mathbf{x}_j;\hat{\theta}\right)$ can be very small. The fractional hot deck imputation under a small imputation size (e.g. $m=10$) does not increase the variance significantly, as can be seen in Table 5.1 under model A.

The proposed fractional imputation method can actually be used to develop a single imputation method by applying FHDI with $m=1$, which selects an imputed value with probability proportional to the fractional weight for each missing unit. In this case, the FHDI can be used to develop a single imputation that is still robust against model misspecification. However, weighting calibration cannot co-exist with single imputation. Calibration constraints can still be achieved by employing the balanced imputation method as discussed in Chauvet, Deville and Haziza (2011) or the rejective Poisson sampling of Fuller (2009). Further investigation along this direction will be a topic of future research.

Acknowledgements

We thank two anonymous referees and the associate editor for very helpful comments. This research was partially supported by a grant from NSF (MMS-121339) and by the Cooperative Agreement between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

Appendix

A.1 Replication variance estimation

For variance estimation, replication methods can be used. Let $w_i^{[k]}$ be the k-th replication weights such that

$$\hat{V}_{rep} = \sum_{k=1}^{L} c_k \left(\hat{Y}^{[k]} - \hat{Y} \right)^2$$

is consistent for the variance of $\hat{Y} = \sum_{i \in A} w_i y_i$, where L is the replication size, c_k is the k-th replication factor that depends on the replication method and the sampling mechanism, and $\hat{Y}^{[k]} = \sum_{i \in A} w_i^{[k]} y_i$ is the k-th replicate of $\hat{Y}$. In delete-1 jackknife variance estimation, $L = n$ and $c_k = (n-1)/n$.

To apply the replication method in FFI, we first apply the replication weights $w_i^{[k]}$ in (2.4) to compute $\hat{\theta}^{[k]}$. Once $\hat{\theta}^{[k]}$ is obtained, we use the same imputed values to compute the initial replication fractional weights

$$w_{ij}^{*[k]} \propto w_j^{[k]} w_j^{-1} f\left(y_j \mid x_i; \hat{\theta}^{[k]} \right) \Bigg/ \left\{ \sum_{l \in A_R} w_l^{[k]} f\left(y_j \mid x_l; \hat{\theta}^{[k]} \right) \right\}, \tag{A.1}$$

with $\sum_{j \in A_R} w_{ij}^{*[k]} = 1$. The variance of $\hat{\eta}_{FFI}$, computed from (3.4), is then computed by

$$\hat{V}_{rep} = \sum_{k=1}^{L} c_k \left(\hat{\eta}_{FFI}^{[k]} - \hat{\eta}_{FFI} \right)^2,$$

where $\hat{\eta}_{FFI}^{[k]}$ comes from solving

$$\sum_{i \in A} w_i^{[k]} \left\{ \delta_i U\left(\eta; \mathbf{x}_i, y_i \right) + \left(1 - \delta_i \right) \sum_{j \in A_R} w_{ij}^{*[k]} U\left(\eta; \mathbf{x}_i, y_j \right) \right\} = 0,$$

and $w_{ij}^{*[k]}$ is defined in (A.1).

We now discuss replication variance estimation of the FHDI estimator $\hat{\eta}_{FHDI}$ computed from (3.8). Define $d_{ij} = 1$ if $j \in D_i$ and $d_{ij} = 0$ otherwise. Note that $\hat{\eta}_{FHDI}$ is computed via two steps: in the first step, a systematic PPS sampling is used with the selection probability proportional to the fractional weights from the FFI method. In the second step, the calibration weighting method using the constraint (3.5) with

$\sum_{j\in A_R} d_{ij} w^*_{ij,c} = 1$ is used. Thus, the replicate fractional weights are also computed in two steps. Firstly, the initial replication fractional weight for $w^*_{ij0} = 1/m$ is then given by

$$w^{*[k]}_{ij0} = \frac{d_{ij}\left(w^{*[k]}_{ij} / w^*_{ij}\right)}{\sum_{l\in A_R} d_{il}\left(w^{*[k]}_{il} / w^*_{il}\right)}, \tag{A.2}$$

where w^*_{ij} is the fractional weight for FFI defined in (2.6) and $w^{*[k]}_{ij}$ is the k-th replication fractional weight for FFI defined in (A.1). Secondly, the replication fractional weights are adjusted to satisfy the calibration constraints. The calibration equation for replication fractional weights corresponding to (3.5) is then

$$\sum_{i\in A} w^{[k]}_i \left\{(1-\delta_i)\sum_{j\in D_i} w^{*[k]}_{ij,c}\mathbf{q}\left(\mathbf{x}_i, y_j\right)\right\} = \sum_{i\in A} w^{[k]}_i \left\{(1-\delta_i)\sum_{j\in A_R} w^{*[k]}_{ij}\mathbf{q}\left(\mathbf{x}_i, y_j\right)\right\} \tag{A.3}$$

and $\sum_{j\in D_i} w^{*[k]}_{ij,c} = 1$. Either regression weighting or entropy weighting can be used to obtain the replication fractional weights satisfying the constraints. Once the replicate fractional weights are obtained, the replicate estimate $\hat{\eta}^{[k]}$ is computed by solving

$$\sum_{i\in A} w^{[k]}_i \left\{\delta_i U\left(\eta; x_i, y_i\right) + \left(1-\delta_i\right)\sum_{j\in A_R} w^{*[k]}_{ijc} U\left(\eta; x_i, y_j\right)\right\} = 0.$$

The replication variance estimator of $\hat{\eta}$, computed from (3.8), is given by

$$\hat{V}_{rep}\left(\hat{\eta}\right) = \sum_{k=1}^{L} c_k \left(\hat{\eta}^{[k]} - \hat{\eta}\right)^2.$$

Because $\hat{\eta}$ is a smooth function of $\hat{\theta}$, the consistency of $\hat{V}_{rep}(\hat{\eta})$ follows directly from the standard argument of the replication variance estimation (Shao and Tu 1995).

A.2 Proof of Equation (4.5)

Using

$$\frac{g\left(y_j \mid x_i\right)}{g\left(y_j \mid x_k\right)} = \frac{f\left(y_j \mid x_i\right)}{f\left(y_j \mid x_k\right)} \exp\left(\varepsilon\Delta_{ik|j} - \kappa\left(x_i\right) + \kappa\left(x_k\right)\right)$$

where $\Delta_{ik|j} = z\left(x_i, y_j; \theta\right) - z\left(x_k, y_j; \theta\right)$. Based on Taylor linearization and the fact of (4.4), we have

$$\frac{g\left(y_j \mid x_i\right)}{g\left(y_j \mid x_k\right)} \cong \frac{f\left(y_j \mid x_i\right)}{f\left(y_j \mid x_k\right)}\left\{1 + \varepsilon\Delta_{ik|j}\right\}.$$

If we know the true density, the correct fractional weights in (3.3) can be expressed by

$$
\begin{aligned}
w_{ij,g}^{*} &\propto \frac{g\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k g\left(y_j \mid x_k\right)} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{g(y_j \mid x_k)}{g(y_j \mid x_i)}\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)} \exp\left(\varepsilon \Delta_{ki|j} - \kappa\left(x_i\right) + \kappa\left(x_k\right)\right)\right\}} \\
&\cong \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\left(1+\varepsilon \Delta_{ki|j}\right)\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \sum_{k;\delta_k=1} w_k \left[\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\left\{z\left(x_k, y_j\right) - z\left(x_i, y_j\right)\right\}\right]} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \lambda^T I_\theta^{-1/2} \sum_{k;\delta_k=1} w_k \left(\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\left\{s\left(x_k, y_j\right) - s\left(x_i, y_j\right)\right\}\right)} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\} + \varepsilon \lambda^T I_\theta^{-1/2} \sum_{k;\delta_k=1} w_k \frac{\partial}{\partial \theta}\left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}} \\
&\propto \frac{1}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}} \left[1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\sum_{k;\delta_k=1} w_k \frac{\partial}{\partial \theta}\left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}}{\sum_{k;\delta_k=1} w_k \left\{\frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}}\right] \\
&\propto \frac{f\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)} \left[1 - \varepsilon \lambda^T I_\theta^{-1/2} \frac{\frac{\partial}{\partial \theta}\left\{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}\right\}}{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}}\right] \\
&= \frac{f\left(y_j \mid x_i\right)}{\sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta}\left\{\frac{1}{\sum_{k;\delta_k=1} w_k \frac{f\left(y_j \mid x_k\right)}{f\left(y_j \mid x_i\right)}}\right\} \\
&= a_{ij} + \varepsilon \lambda^T I_\theta^{-1/2} \frac{\partial}{\partial \theta} a_{ij},
\end{aligned}
$$

where $a_{ij} = f\left(y_j \mid x_i\right) \Big/ \sum_{k;\delta_k=1} w_k f\left(y_j \mid x_k\right)$ and $a_{i+} = \sum_{j;\delta_j=1} a_{ij}$. So, $w^*_{ij,f} = a_{ij} \big/ a_{i+}$ and

$$
\begin{aligned}
w^*_{ij,g} &\cong \frac{a_{ij} + \varepsilon\lambda^T I_\theta^{-1/2} \dfrac{\partial}{\partial\theta} a_{ij}}{a_{i+} + \varepsilon\lambda^T I_\theta^{-1/2} \dfrac{\partial}{\partial\theta} a_{i+}} \\
&= \frac{a_{ij}}{a_{i+}} \left(1 + \frac{\varepsilon\lambda^T I_\theta^{-1/2} \dfrac{\partial}{\partial\theta} a_{ij}}{a_{ij}} \right) \left(\frac{a_{i+}}{a_{i+} + \varepsilon\lambda^T I_\theta^{-1/2} \dfrac{\partial}{\partial\theta} a_{i+}} \right) \\
&\cong \frac{a_{ij}}{a_{i+}} \left(1 + \varepsilon\lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta} \log a_{ij} \right) \left(1 - \varepsilon\lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta} \log a_{i+} \right) \\
&\cong \frac{a_{ij}}{a_{i+}} + \varepsilon\lambda^T I_\theta^{-1/2} \frac{a_{ij}}{a_{i+}} \left(\frac{\partial}{\partial\theta} \log a_{ij} - \frac{\partial}{\partial\theta} \log a_{i+} \right) \\
&= \frac{a_{ij}}{a_{i+}} + \varepsilon\lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta} \left(\frac{a_{ij}}{a_{i+}} \right) \\
&= w^*_{ij,f} + \varepsilon\lambda^T I_\theta^{-1/2} \frac{\partial}{\partial\theta} \left(w^*_{ij,f} \right),
\end{aligned}
$$

which proves (4.5).

A.3 Extension to a non-ignorable missing case

We consider an extension of the proposed method to a non-ignorable missing case. Under the non-ignorable missing assumption, both the conditional model $f\left(y \mid x\right)$ and the response probability model $P\left(\delta = 1 \mid \mathbf{x}, y\right)$ are needed to evaluate the expected estimating function in (4.6). Let the response probability model be given by $Pr\left(\delta_i = 1 \mid \mathbf{x}_i, y_i\right) = \pi\left(\mathbf{x}_i, y_i; \phi\right)$, for some ϕ with a known $\pi(\cdot)$ function. We assume that the parameters are identifiable as discussed in Wang, Shao and Kim (2013).

In PFI, according to Kim and Kim (2012), the MLE $\left(\hat{\theta}, \hat{\phi}\right)$ can be obtained by solving

$$
\sum_{i \in A} w_i \left\{ \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) + \left(1 - \delta_i\right) \sum_{j=1}^{m} w^*_{ij} S\left(\theta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{A.4}
$$

and

$$
\sum_{i \in A} w_i \left\{ \delta_i S\left(\phi; \mathbf{x}_i, y_i\right) + \left(1 - \delta_i\right) \sum_{j=1}^{m} w^*_{ij} S\left(\phi; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0, \tag{A.5}
$$

where $S\left(\theta; \mathbf{x}, y\right) = \partial \log f\left(y \mid \mathbf{x}; \theta\right) \big/ \partial\theta$, $S\left(\phi; \mathbf{x}, y\right) = \partial \log \pi\left(\mathbf{x}, y; \phi\right) \big/ \partial\phi$, and the fractional weights are given by

$$
w^*_{ij}\left(\theta, \phi\right) = \frac{f\left(y_i^{*(j)} \mid \mathbf{x}_i; \theta\right) \left\{ 1 - \pi\left(\mathbf{x}_i, y_i^{*(j)}, \phi\right) \right\} \Big/ h\left(y_i^{*(j)} \mid \mathbf{x}_i\right)}{\sum_{k=1}^{m} \left[f\left(y_i^{*(k)} \mid \mathbf{x}_i; \theta\right) \left\{ 1 - \pi\left(\mathbf{x}_i, y_i^{*(k)}, \phi\right) \right\} \Big/ h\left(y_i^{*(k)} \mid \mathbf{x}_i\right) \right]}. \tag{A.6}
$$

The solution to (A.4) and (A.5) can be obtained via the EM algorithm. In the EM algorithm, the E-step computes the fractional weights in (A.6) using the current parameter values and the M-step updates the parameter value $\hat{\theta}^{(t+1)}$ and $\hat{\phi}^{(t+1)}$ by solving

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j=1}^{m} w_{ij}^{*} \left(\hat{\theta}^{(t)}, \hat{\phi}^{(t)}\right) S\left(\theta; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0,$$

and

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\phi; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j=1}^{m} w_{ij}^{*} \left(\hat{\theta}^{(t)}, \hat{\phi}^{(t)}\right) S\left(\phi; \mathbf{x}_i, y_i^{*(j)}\right) \right\} = 0.$$

In the proposed FFI method, the fractional weights are given by

$$\begin{aligned} w_{ij}^{*} &\propto f\left(y_j \mid \mathbf{x}_i, \delta_i = 0; \theta, \phi\right) \Big/ f\left(y_j \mid \delta_j = 1\right) \\ &\propto f\left(y_j \mid \mathbf{x}_i; \theta\right) \left\{1 - \pi\left(\mathbf{x}_i, y_j; \phi\right)\right\} \Big/ f\left(y_j \mid \delta_j = 1\right), \end{aligned}$$

with $\sum_{j;\delta_j=1} w_{ij}^{*} = 1$. Because

$$\begin{aligned} f\left(y_j \mid \delta_j = 1\right) &= \int \pi\left(\mathbf{x}, y_j\right) f\left(y_j \mid \mathbf{x}\right) f(\mathbf{x}) d\mathbf{x} \\ &\cong \sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j\right) f\left(y_j \mid \mathbf{x}_k\right). \end{aligned} \tag{A.7}$$

The fractional weights can be computed from

$$w_{ij}^{*} \propto \frac{f\left(y_j \mid \mathbf{x}_i; \theta\right) \left\{1 - \pi\left(\mathbf{x}_i, y_j; \phi\right)\right\}}{\sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j; \phi\right) f\left(y_j \mid \mathbf{x}_k; \theta\right)}. \tag{A.8}$$

with $\sum_{j\in A_R} w_{ij}^{*} = 1$.

Thus, we can use the following EM algorithm to obtain the desired parameter estimates.

(I-step) For each missing unit $i \in A_M = \left\{i \in A; \delta_i = 0\right\}$, take m imputed values as $y_i^{(1)}, \ldots, y_i^{(m)}$ from A_R, where $m = r$.

(E-step) The fractional weights are given by

$$w_{ij}^{*(t)} \propto \frac{f\left(y_j \mid \mathbf{x}_i, \hat{\theta}^{(t)}\right) \left\{1 - \pi\left(\mathbf{x}_i, y_j; \hat{\phi}^{(t)}\right)\right\}}{\sum_{k\in A} w_k \pi\left(\mathbf{x}_k, y_j; \hat{\phi}^{(t)}\right) f\left(y_j \mid \mathbf{x}_k; \hat{\theta}^{(t)}\right)}$$

and $\sum_{j=1}^{m} w_{ij}^{*(t)} = 1$.

(M-step) Update the parameter $\hat{\theta}^{(t+1)}$ and $\hat{\phi}^{(t+1)}$ by solving the following imputed score equations,

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\theta; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j\in A_R} w_{ij}^{*(t)} S\left(\theta; \mathbf{x}_i, y_j\right) \right\} = 0,$$

and

$$\sum_{i\in A} w_i \left\{ \delta_i S\left(\phi; \mathbf{x}_i, y_i\right) + \left(1-\delta_i\right) \sum_{j\in A_R} w_{ij}^{*(t)} S\left(\phi; \mathbf{x}_i, y_j\right) \right\} = 0.$$

Note that the I-step does not have to be repeated in the EM algorithm. Once the final parameter estimates are computed, the fractional weights are computed by (A.8), which serve as the selection probabilities for FHDI with a small imputation size m. The same systematic PPS sampling method as discussed in Section 3 can be used to obtain FHDI.

References

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting process: a large sample study. *The Annals of Statistics*, 10, 1100-1120.

Andridge, R.R. and Little, R.J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78, 40-64.

Beaumont, J.F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Binder, D. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

Chauvet, G., Deville, J.-C. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98, 459-471.

Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American Statistical Association*, 96, 260-269.

Copas, J.B. and Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, Series B*, 63, 871-895.

Durrant, G.B. (2009). Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. *International Journal of Social Research Methodology*, 12, 293-304.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Fuller, W.A. and Kim, J.K. (2005). Hot deck imputation for the response model. *Survey Methodology*, 31, 139-149.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of Statistics*, 29, 215-246.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics, Series A*, 13, 1919-1939.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119-132.

Kim, J.K. and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Fuller, W.A. and Bell, W.R. (2011). Variance estimation for nearest neighbor imputation for U.S. census long form data. *Annals of Applied Statistics*, 5, 824-842.

Kim, J.K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Chapman and Hall/CRC.

Kim, J.Y. and Kim, J.K. (2012). Parametric fractional imputation for nonignorable missing data. *Journal of the Korean Statistical Society*, 41, 291-303.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9, 538-573.

Pfeffermann, D. (2011). Modeling of complex survey data: why is it a problem? How should we approach it? *Survey Methodology*, 37, 115-136.

Rao, J.N.K., Yung, W. and Hidiroglou, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *The Indian Journal of Statistics, Series A*, 64, 364-378.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schenker, N. and Welsh, A.H. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics*, 16, 1550-1566.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer.

Sung, Y.J. and Geyer, C.J. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics*, 35, 990-1011.

Wang, C.-Y. and Pepe, M.S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society, Series B*, 62, 509-24.

Wang, S., Shao J. and Kim, J.K. (2013). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*. In press.

Wei, G.C. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699-704.

White, H. (1982). Maximum likelihood estimation of misspecifed models. *Econometrica*, 50, 1-25.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 231-242
Statistics Canada, Catalogue No. 12-001-X

Potential gains from using unit level cost information in a model-assisted framework

David G. Steel and Robert Graham Clark[1]

Abstract

In developing the sample design for a survey we attempt to produce a good design for the funds available. Information on costs can be used to develop sample designs that minimise the sampling variance of an estimator of total for fixed cost. Improvements in survey management systems mean that it is now sometimes possible to estimate the cost of including each unit in the sample. This paper develops relatively simple approaches to determine whether the potential gains arising from using this unit level cost information are likely to be of practical use. It is shown that the key factor is the coefficient of variation of the costs relative to the coefficient of variation of the relative error on the estimated cost coefficients.

Key Words: Optimal allocation; Optimal design; Sample design; Sampling variance; Survey costs.

1 Introduction

Unequal unit costs have been reflected in sample designs by using simple linear cost models. In stratified sampling, a per-unit cost coefficient can sometimes be estimated for each stratum. The resulting allocation of sample to strata is proportional to the inverse of the square root of the stratum cost coefficients (Cochran 1977). In a multistage design the costs of including the units at the different stages of selection can be used to decide the number of units to select at each stage (Hansen, Hurwitz and Madow 1953).

While this theory is well established, unequal costs have not been used extensively in practice (Brewer and Gregoire 2009), perhaps because of a lack of good information on costs, and because of a focus on sample size rather than cost of enumeration. Groves (1989) argued that linear cost models are unrealistic, and that mathematical cost modelling can distract from more important decisions such as the mode of collection, the number of callbacks and how the survey interacts with other surveys conducted by the same organisation. Nevertheless, given the pressures on survey budgets, the final design should reflect costs and variance in a rational way, without being fixated on formal optimality.

Increasing use of computers in data collection is leading to more extensive and useful cost-related information on units on survey frames. In a programme of business surveys conducted by a national statistics institute, most medium and large enterprises will be selected in some surveys at least every year or two. This may provide information on costs for those businesses, for example some businesses may have required extensive follow-up or editing in a previous survey. Direct experience is less likely to be available for any given small business, but datasets of costs could be modelled to give predictions of likely costs.

1. David G. Steel, National Institute for Applied Statistics Research Australia, University of Wollongong, NSW Australia 2522. E-mail: dsteel@uow.edu.au; Robert Graham Clark, National Institute for Applied Statistics Research Australia, University of Wollongong, NSW Australia 2522. E-mail: rclark@uow.edu.au.

Adaptive and responsive survey designs make use of paradata (process data) collected during a survey's operation, and auxiliary data known for the sampling frame (typically from administrative sources), to guide ongoing decisions. These may include the number of callbacks, which respondents to follow up, targeting of incentives, and choice of mode of collection for followup attempts (Groves and Heeringa 2006). In one example discussed by Groves and Heeringa (2006), interviewers designated non-respondents as having either low or high propensity to respond. The latter are less costly to convert to respondents, and a higher sampling fraction was assigned to them in a second phase of the survey. More recently, Schouten, Bethlehem, Beullens, Kleven, Loosveldt, Luiten, Rutar, Shlomo and Skinner (2012, Section 6) suggested that followup in the second phase of a survey should be designed to improve the R-indicator of non-response bias (defined in Schouten, Cobben and Bethlehem 2009; and in Schouten Shlomo and Skinner 2011). Peytchev, Riley, Rosen, Murphy and Lindblad (2010) argued that likely non-responders should be targeted with a different protocol from the very outset of a survey.

Thus, unequal unit costs can arise in practice, either for all units in advance of sampling, or for non-respondents who are to be targeted for followup. In either case, the collection and use of cost information incurs some expense and additional complexity. Moreover, effectively trading off cost and variance is only part of the picture, and response bias must also be considered. It is therefore important to understand whether the potential gains from using this information are worthwhile, particularly as any cost data is likely to be imperfect.

This paper develops relatively simple approximations to the gains arising from using unit level cost information in a model-assisted framework. Section 2 contains notation and some key expressions. Section 3 is concerned with the optimal design when cost parameters are known. Section 4 analyses the use of estimated unit costs, and Section 5 presents examples. Section 6 offers a discussion.

2 Notation and objective criterion

Consider a finite population, U containing N units, consisting of values Y_i for $i \in U$. A sample $s \in U$ is to be selected using an unequal probability sampling scheme with positive probability of selection $\pi_i = P[i \in s]$ for all units $i \in U$. A vector of auxiliary variables $\mathbf{x_i}$ is assumed to be available either for the whole population, or for all units $i \in s$ with the population total, $\mathbf{t_x} = \sum_{i \in U} \mathbf{x_i}$, also known. The auxiliary variables could consist of, for example, industry, region and size in a business survey, or age, sex and region in a household survey.

In the model-assisted approach (see for example Särndal, Swensson and Wretman 1992), the relationship between a variable of interest and the auxiliary variables is captured in a model, typically of the following form in single-stage surveys:

$$\left.\begin{array}{l} E_M[Y_i] = \beta^{\mathrm{T}} \mathbf{x_i} \\ var_M[Y_i] = \sigma^2 z_i \\ Y_i \text{ independent of } Y_j \text{ for all } i \neq j \end{array}\right\} \tag{2.1}$$

where E_M and var_M denote expectation and variance under the model, β is a vector of unknown regression parameters, σ^2 is an unknown variance parameter, and $\mathbf{x_i}$ and z_i are assumed to be known for

all $i \in U$. Let E_p and var_p denote expectation and variance under repeated probability sampling with all population values held fixed.

The generalized regression estimator is a widely used model-assisted estimator of t_y :

$$\hat{t}_y = \sum_{i \in s} \pi_i^{-1} \left(y_i - \hat{\beta}^{\mathbf{T}} \mathbf{x_i} \right) + \hat{\beta}^{\mathbf{T}} \mathbf{t_x} \tag{2.2}$$

where $\hat{\beta}$ may be a weighted or unweighted least squares estimate of the regression coefficients of y_i on $\mathbf{x_i}$ using sample data. Estimators can also be constructed for nonlinear extensions to model (2.1), but in practice the linear model is almost always used.

The **anticipated variance** of $\hat{t}_y$ is defined by $E_M var_p \left[\hat{t}_y - t_y \right]$, and is approximated by

$$E_M var_p \left[\hat{t}_y \right] \approx \sigma^2 \sum_{i \in U} \left(\pi_i^{-1} - 1 \right) z_i \tag{2.3}$$

for large samples (Särndal et al. 1992, formula 12.2.12, p. 451) under model (2.1). Model-assisted designs and estimators should minimise $E_M var_p \left[\hat{t}_y \right]$ subject to approximate design unbiasedness, $E_p \left[\hat{t}_y \right] = t_y$. Even if the model is incorrect, (2.2) remains approximately design-unbiased, although it will no longer have the lowest possible large sample anticipated variance. The anticipated variance has been used to motivate model-assisted sample designs in one stage (Särndal et al. 1992) and two stage sampling (Clark and Steel 2007; Clark 2009). One advantage of using the anticipated variance for this purpose is that it depends only on the selection probabilities and a small number of model parameters, which can be roughly estimated when designing the sample. In contrast, $var_p \left[\hat{t}_y \right]$ typically depends on the population values of y_i and on joint probabilities of selection, both of which are difficult to quantify in advance.

The cost of enumerating a sample is assumed to be $C = \sum_{i \in s} c_i$ where c_i is the cost of surveying a particular unit i. The values of c_i are usually assumed to be known. Typically c_i are also assumed to be constant for all units in the population, or constant within strata. With the generalization that c_i may be different for every unit i, the cost C depends on the particular sample s selected. The expected cost is $E_p [C] = \sum_{i \in U} \pi_i c_i$. The aim is to minimise the anticipated variance (2.3) subject to a constraint on the expected enumeration cost,

$$\sum_{i \in U} \pi_i c_i = C_f . \tag{2.4}$$

There will also be fixed costs that are not affected by the sample design and so do not have to be included here.

Some notation for population variances and covariances is needed. Consider the pairs (u_i, v_i), and let $S_{uv} = N^{-1} \sum_{i \in U} (u_i - \overline{u})(v_i - \overline{v})$ denote their population covariance, and $S_u^2 = N^{-1} \sum_{i \in U} (u_i - \overline{u})^2$ denote the population variance of u_i $(i = 1, \ldots, N)$. Let $\overline{u}$ and $\overline{v}$ be the population means of u_i and v_i. The population coefficient of variation of u_i is $C_u = S_u / \overline{u}$. The population relative covariance of (u_i, v_i) is $C_{u,v} = S_{uv} / \overline{u}\,\overline{v}$. A useful result is

$$\sum_{i \in U} u_i v_i = N \overline{u}\,\overline{v} \left(1 + C_{u,v} \right). \tag{2.5}$$

3 Optimal design with known cost and variance parameters

3.1 Optimal Model-Assisted Design

The values of $\left(\pi_i : i \in U\right)$ which minimise (2.3) subject to (2.4) are

$$\pi_i = C_f \frac{z_i^{1/2} c_i^{-1/2}}{\sum_{j \in U} z_j^{1/2} c_j^{1/2}} \propto z_i^{1/2} c_i^{-1/2} \tag{3.1}$$

and the resulting anticipated variance is

$$AV_{opt} = E_M var_p\left[\hat{t}_y\right] = \sigma^2 C_f^{-1}\left(\sum_{i \in U} c_i^{1/2} z_i^{1/2}\right)^2 - \sigma^2 \sum_{i \in U} z_i. \tag{3.2}$$

This can be easily derived using Lagrange multipliers or the Cauchy-Schwarz Inequality, and generalizes Särndal et al. (1992, Result 12.2.1, p. 452) to allow for unequal costs. Higher probability of selection is given to units which have higher unit variance or lower cost. However the square roots of z_i and c_i in (3.1) means that probabilities of selection do not vary dramatically in many surveys.

For the special case of stratified sampling where $c_i = \overline{c}_h$ and $z_i = \overline{z}_h$ for units i in stratum h, (3.1) becomes the usual optimal stratified allocation with $\pi_i \propto \sqrt{\overline{z}_h / \overline{c}_h}$, so that $n_h \propto N_h \sqrt{\overline{z}_h / \overline{c}_h}$.

It is assumed that the last term of (3.2), which represents the finite population correction, is negligible. Applying (2.5) gives:

$$AV_{opt} \approx \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}\left(1 + C_{\sqrt{c},\sqrt{z}}\right)^2}{\left(1 + C_{\sqrt{c}}^2\right)\left(1 + C_{\sqrt{z}}^2\right)} \tag{3.3}$$

where $C_{\sqrt{c}}$ and $C_{\sqrt{z}}$ refer to the population coefficients of variation of $\sqrt{c_i}$ and $\sqrt{z_i}$, respectively. To make our results interpretable, we will assume that unit costs c_i and variances σz_i are unrelated, so that $C_{\sqrt{c},\sqrt{z}} = 0$. This assumption may not always be satisfied in practice, but any relationship between c_i and z_i will be specific to the particular example, and could be either positive or negative. To identify general principles, it makes sense to ignore any such relationship. In practice, it is often reasonable to also assume that $C_{\sqrt{c}}$ and $C_{\sqrt{z}}$ are small. A Taylor Series expansion then shows that $C_c^2 \approx 4C_{\sqrt{c}}^2$ and $C_z^2 \approx 4C_{\sqrt{z}}^2$. Putting these approximations together, (3.3) becomes

$$AV_{opt} = \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}}{\left(1 + \frac{1}{4}C_c^2\right)\left(1 + \frac{1}{4}C_z^2\right)}. \tag{3.4}$$

See the Appendix for details of these derivations.

Ignoring Costs

If the costs are ignored, then (3.1) suggests that $\pi_i \propto z_i^{1/2}$. To make comparisons for the same expected cost, C_f,

$$\pi_i = C_f \frac{z_i^{1/2}}{\sum_{j\in U} z_j^{1/2} c_j} \tag{3.5}$$

with resulting anticipated variance

$$AV_{nocosts} = \sigma^2 C_f^{-1} \left(\sum_{i\in U} z_i^{1/2} \right) \left(\sum_{i\in U} c_i z_i^{1/2} \right) - \sigma^2 \sum_{i\in U} z_i. \tag{3.6}$$

Applying derivations similar to those used in Section 3.1,

$$AV_{nocosts} \approx \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z}}{\left(1 + \frac{1}{4} C_z^2\right)}. \tag{3.7}$$

See Appendix for details. Comparing (3.7) and (3.4), we see that taking costs into account in the design results in dividing the anticipated variance by $\left(1 + (1/4) C_c^2\right)$.

4 The effect of using estimated cost parameters

In practice, c_i are not known precisely. Suppose that estimates $\hat{c}_i = b_i c_i$ are used instead. Using the auxiliary variable and the estimated costs in the optimal probabilities implies $\pi_i \propto z_i^{1/2} \hat{c}_i^{-1/2}$. To make comparisons for the same expected costs,

$$\pi_i = C_f \frac{z_i^{1/2} \hat{c}_i^{-1/2}}{\sum_{j\in U} z_j^{1/2} \hat{c}_j^{-1/2} c_j}.$$

The resulting anticipated variance is

$$AV_{ests} = \sigma^2 C_f^{-1} \left(\sum_{i\in U} \hat{c}_i^{1/2} z_i^{1/2} \right) \left(\sum_{j\in U} z_j^{1/2} \hat{c}_j^{-1/2} c_j \right) - \sigma^2 \sum_{i\in U} z_i. \tag{4.1}$$

If we assume that the values of b_i are unrelated to the values of c_i and z_i, then

$$AV_{ests} = \sigma^2 C_f^{-1} \left(\sum_{i\in U} c_i^{1/2} z_i^{1/2} \right)^2 N^{-2} \left(\sum_{i\in U} b_i^{-1/2} \right) \left(\sum_{i\in U} b_i^{1/2} \right) - \sigma^2 \sum_{i\in U} z_i, \tag{4.2}$$

see Appendix for details. If the coefficient of variation of b_i is small, then a Taylor Series approximation gives $N^{-2} \sum b_i^{-1/2} \sum b_i^{1/2} \approx 1 + (1/4) C_b^2$. Applying this, and the same approximations as in Subsection 3.1, (4.2) becomes

$$AV_{ests} = \frac{\sigma^2 C_f^{-1} N^2 \overline{c}\,\overline{z} \left(1 + \frac{1}{4} C_b^2\right)}{\left(1 + \frac{1}{4} C_c^2\right) \left(1 + \frac{1}{4} C_z^2\right)}. \tag{4.3}$$

See Appendix for details.

Comparing (4.3) and (3.7), the effect of using estimated cost parameters rather than no costs at all is to multiply the anticipated variance by $\left[1+(1/4)C_b^2\right]/\left[1+(1/4)C_c^2\right]$. Therefore cost information is worth using provided $C_b < C_c$. The coefficient of variation of the error factors has to be less than that of the true unit costs over the population.

5 Examples of cost models

The key quantities determining the usefulness of the unit cost data are C_b and C_c. Optimal designs using unequal cost information are not very common, so there is relatively little literature on the typical values of these measures. Unequal costs may be driven by a variety of factors, including mode effects, geography and willingness to respond, and literature on these issues is helpful to give a rough idea of cost models that may apply in practice.

One reason why unequal per-unit costs may arise is the use of mixed mode interviewing. Different respondents may respond using different modes of collection, for example computer-assisted personal or telephone interviewing, mail or web questionnaires, or face to face interviewer (Dillman, Smyth and Christian 2009). This may be done to reduce cost or to improve response rate, however care must be taken that the approach does not introduce bias due to mode effects. Mode effects may consist of selection effects (which are generally not a problem) and measurement effects (which typically lead to bias), and the two are often hard to disentangle (Vannieuwenhuyze, Loosveldt and Molenbergs 2012). Cost savings from the use of mixed modes could potentially be magnified by incorporating mode costs into the sample design as described in this paper. Groves (1989, p. 538) compares per-respondent costs of telephone interviewing ($38.00) and personal interviewing ($84.90) of the general population. If the preference of all units on a frame was known, and half preferred each mode, this would imply $C_c = 0.38$. Greenlaw and Brown-Welty (2009) compared paper and web surveys, and found per-respondent costs of $4.78 and $0.64, respectively, in a survey of members of a professional association. In a mixed mode option, two thirds of respondents opted for the web option. If preferences are known in advance, then $C_c = 0.76$.

Another reason for varying costs is that some respondents are more difficult to recruit than others, requiring more visits or reminders. Groves and Heeringa (2006, Section 2.2) trialled a survey where interviewers classified non-respondents from the first approach as either likely or unlikely to respond. In subsequent follow-up, the first group had a response rate of 73.7% compared to 38.5% for the second group. This suggests that the per-respondent cost for the second group would be at least 1.9 times higher than the first group. (In fact, the ratio would be higher, because more follow-up attempts would be made for the difficult group.) If 50% of respondents are in both groups, then $C_c = 0.31$.

Geography is another source of differential costs in interviewer surveys. In the Australian Labour Force Survey, costs have been modelled as having a per-block component and a per-dwelling component (Hicks 2001, Table 4.2.1 in Section 4.2) depending on the type of area (15 types were defined). Assuming a constant 10 dwellings sampled per block, the net per-dwelling costs range from $4.98 in Inner City Sydney and Melbourne to $6.71 in Sparse and Indigenous areas. While this is a significant difference in costs across area types, the great majority of the population are in three area types (settled area, outer

growth and large town) where per-dwelling costs vary only between \$5.71 and \$6.07. As a result, C_c is estimated at a very small 0.054.

Table 5.1 shows the approximate percentage improvement in the anticipated variance from using estimated cost information for different values of C_c and C_b, some suggested by these examples. Negative values indicate that the design is less efficient than ignoring costs altogether. The table suggests that cost information is only worthwhile provided there is a fair variation in the unit costs, otherwise the benefit is very small, and can be erased when there is even small imprecision in the estimated costs. Mixed mode surveys have the most potential for exploiting varying unit costs in sample design, but the possibility of measurement bias would need to be carefully assessed in any such approach, using methods such as those in Vannieuwenhuyze, Loosveldt and Molenberghs (2010), Vannieuwenhuyze et al. (2012), Vannieuwenhuyze and Loosveldt (2013) and Schouten, Brakel, Buelens, Laan and Klaus (2013). It might even be possible to incorporate mode effects (or uncertainty about mode effects) into the optimal design via the variance model, and this may be the topic of future research. The findings made in this paper suggest that such an approach is worth considering.

Table 5.1
Percentage improvement in anticipated variance from using estimated cost information compared to no cost information.

Coefficient of Variation of Unit Costs (C_c) (%)	Possible scenario	Coefficient of Variation of Error Factor (C_b) (%)			
		0	10	25	50
5		0.1	-0.2	-1.5	-6.2
10	Interviewer travel due to remoteness	0.2	0.0	-1.3	-6.0
20		1.0	0.7	-0.6	-5.2
30	Response propensity	2.2	2.0	0.7	-3.9
40	Mixed mode (phone/personal int.)	3.8	3.6	2.3	-2.2
50		5.9	5.6	4.4	0.0
75	Mixed mode (paper/web self-complete)	12.3	12.1	11.0	6.8

6 Discussion

Incorporating unequal unit costs can improve the efficiency of sample designs. For the gains to be appreciable, the unit costs need to vary considerably. Even with no estimation error, a coefficient of variation of 50% may lead to a gain of only 6% in the anticipated variance. When this coefficient of variation is 75%, as can happen in a mixed mode survey, the reduction in the anticipated variance (or in the sample size for fixed precision) can be over 12%. Costs will be estimated with some error and this reduces the gain by a factor determined by the relative variation of the relative errors in estimating the costs at the individual level.

Appendix

A.1 Detailed derivations

Lemma 1: Let u_i be defined for $i \in U$. Let $u_i = \overline{u} + \theta e_i$, where $\sum_{i \in U} e_i = 0$ and θ is small. Then:

a. $\overline{\sqrt{u}} = \sqrt{\overline{u}} - \frac{1}{8}\theta^2 \overline{u}^{-3/2} S_e^2 + o\left(\theta^2\right)$.

b. $S_{\sqrt{u}}^2 = \frac{1}{4}\theta^2 \overline{u}^{-1} S_e^2 + o\left(\theta^2\right) = \frac{1}{4}\overline{u}^{-1} S_u^2 + o\left(\theta^2\right)$.

c. $N^{-2}\left(\sum_{i \in U} u_i^{1/2}\right)\left(\sum_{i \in U} u_i^{-1/2}\right) = 1 + \frac{1}{4}\theta^2 \overline{u}^{-2} S_e^2 + o\left(\theta^2\right) = 1 + \frac{1}{4} C_u^2 + o\left(\theta^2\right)$.

d. $C_{\sqrt{u}}^2 = \frac{1}{4}\theta^2 \overline{u}^{-2} S_e^2 + o\left(\theta^2\right) = \frac{1}{4} C_u^2 + o\left(\theta^2\right)$.

The notation $o\left(C_u^2\right)$ can be used in place of $o\left(\theta^2\right)$, since $C_u^2 = \theta^2 C_e^2$. This will be done in the remainder of the Appendix.

Proof:

We start by writing $\overline{\sqrt{u}}$ as a function of θ:

$$\overline{\sqrt{u}} = N^{-1} \sum_{i \in U} \sqrt{u_i} = N^{-1} \sum_{i \in U} \sqrt{\overline{u} + \theta e_i}.$$

Call this $g(\theta)$, then differentiating about $\theta = 0$ gives $g(0) = \sqrt{\overline{u}}$, $g'(0) = 0$ and

$$g''(0) = -\frac{1}{4} N^{-1} \overline{u}^{-3/2} \sum_{i \in U} e_i^2 = -\frac{1}{4} \overline{u}^{-3/2} S_e^2.$$

Hence

$$\overline{\sqrt{u}} = g(\theta) = g(0) + g'(0)\theta + \frac{1}{2} g''(0)\theta^2 + o\left(\theta^2\right) = \sqrt{\overline{u}} - \frac{1}{8}\theta^2 \overline{u}^{-3/2} S_e^2 + o\left(\theta^2\right)$$

which is result a.

Result b is proven using result a:

$$\begin{aligned}
S_{\sqrt{u}}^2 &= N^{-1} \sum_{i \in U} \left(\sqrt{u_i}\right)^2 - \left(N^{-1} \sum_{i \in U} \sqrt{u_i}\right)^2 \\
&= \overline{u} - \left(\overline{\sqrt{u}}\right)^2 \\
&= \overline{u} - \left(\sqrt{\overline{u}} - \frac{1}{8}\theta^2 \overline{u}^{-3/2} S_e^2 + o\left(\theta^2\right)\right)^2 \\
&= \overline{u} - \left(\overline{u} + \frac{1}{64}\theta^4 \overline{u}^{-3} S_e^4 - \frac{1}{4}\theta^2 \overline{u}^{-1} S_e^2 + o\left(\theta^2\right)\right) \\
&= \frac{1}{4}\theta^2 \overline{u}^{-1} S_e^2 + o\left(\theta^2\right) = \frac{1}{4}\overline{u}^{-1} S_u^2 + o\left(\theta^2\right).
\end{aligned}$$

To derive c, we firstly write $N^{-1}\sum_{i\in U} u_i^{-1/2}$ as a function $g()$ of θ and take a Taylor Series expansion:

$$\begin{aligned} N^{-1}\sum_{i\in U} u_i^{-1/2} &= N^{-1}\sum_{i\in U}\left(\overline{u}+\theta e_i\right)^{-1/2} \\ &= g\left(\theta\right) = g(0)+g'(0)\theta+\frac{1}{2}g''(0)\theta^2+o\left(\theta^2\right) \\ &= \overline{u}^{-1/2}+0\theta+\frac{1}{2}\frac{3}{4}\overline{u}^{-5/2}N^{-1}\sum_{i\in U} e_i^2\theta^2+o\left(\theta^2\right) \\ &= \overline{u}^{-1/2}+\frac{3}{8}\overline{u}^{-5/2}S_e^2\theta^2+o\left(\theta^2\right) \end{aligned} \tag{A.1}$$

Note that $N^{-1}\sum_{i\in U} u_i^{1/2} = \overline{\sqrt{u}}$. Multiplying the expression for $\overline{\sqrt{u}}$ in result a and (A.1) gives

$$\begin{aligned} N^{-2}\left(\sum_{i\in U} u_i^{1/2}\right)\left(\sum_{i\in U} u_i^{-1/2}\right) &= \left\{\sqrt{\overline{u}}-\frac{1}{8}\theta^2\overline{u}^{-3/2}S_e^2+o\left(\theta^2\right)\right\}\left\{\overline{u}^{-1/2}+\frac{3}{8}\overline{u}^{-5/2}S_e^2\theta^2+o\left(\theta^2\right)\right\} \\ &= 1+\frac{1}{4}\overline{u}^{-2}S_e^2\theta^2+o\left(\theta^2\right) \\ &= 1+\frac{1}{4}C_u^2+o\left(\theta^2\right) \end{aligned}$$

which is result c.

For result d, firstly note that $\overline{\sqrt{u}} = \sqrt{\overline{u}}+o\left(\theta\right)$ from result a, and so, from a first order Taylor Series,

$$\left(\overline{\sqrt{u}}\right)^{-2} = \left(\sqrt{\overline{u}}\right)^{-2}+o\left(\theta\right) = \overline{u}^{-1}+o\left(\theta\right).$$

Combining this with result b, we obtain

$$\begin{aligned} C_{\sqrt{u}}^2 &= S_{\sqrt{u}}^2\left(\overline{\sqrt{u}}\right)^{-2} \\ &= \left\{\frac{1}{4}\theta^2\overline{u}^{-1}S_e^2+o\left(\theta^2\right)\right\}\left\{\overline{u}^{-1}+o\left(\theta\right)\right\} \\ &= \frac{1}{4}\theta^2\overline{u}^{-2}S_e^2+o\left(\theta^2\right) \\ &= \frac{1}{4}C_u^2+o\left(\theta^2\right) \end{aligned}$$

giving result d.

Derivation of (3.3)

For the special case where $u_i = v_i$, (2.5) becomes

$$\sum_{i\in U} u_i^2 = N\overline{u}^2\left(1+C_u^2\right). \tag{A.2}$$

Applying (2.5),

$$\sum_{i\in U} c_i^{1/2}z_i^{1/2} = N\overline{\sqrt{c}}\ \overline{\sqrt{z}}\left(1+C_{\sqrt{c},\sqrt{z}}\right) \tag{A.3}$$

where $\overline{\sqrt{c}} = N^{-1}\sum_{i\in U}\sqrt{c_i}$ and $\overline{\sqrt{z}} = N^{-1}\sum_{i\in U}\sqrt{z_i}$. Using (A.2), we can express $\overline{\sqrt{c}}$ in terms of $\overline{c}$:

$$\overline{c} = N^{-1}\sum_{i\in U} c_i = N^{-1}\sum_{i\in U}\left(\sqrt{c_i}\right)^2 = \left(\overline{\sqrt{c}}\right)^2\left(1+C^2_{\sqrt{c}}\right). \tag{A.4}$$

Similarly,

$$\overline{z} = \left(\overline{\sqrt{z}}\right)^2\left(1+C^2_{\sqrt{z}}\right). \tag{A.5}$$

Assuming the last term of (3.2) is negligible, applying (A.3), (A.4) and (A.5) gives (3.3).

Derivation of (3.4)

Lemma 1d implies that $C^2_{\sqrt{c}} = (1/4)C^2_c + o\left(C^2_c\right) \approx (1/4)C^2_c$ and $C^2_{\sqrt{z}} = (1/4)C^2_z + o\left(C^2_z\right) \approx (1/4)C^2_z$. Result (3.4) follows from (3.3) by using these approximations, as well as assuming that $C_{\sqrt{c},\sqrt{z}} = 0$.

Derivation of (3.7)

Firstly, $\sum_{i\in U} c_i z_i^{1/2} = N\overline{c}\,\overline{\sqrt{z}}\left(1+C_{c,\sqrt{z}}\right)$, from (2.5), where $C_{c,\sqrt{z}}$ is the population relative covariance between the values of $z_i^{1/2}$ and c_i. It is assumed that the values of c_i and z_i are unrelated, so that $C_{c,\sqrt{z}} = 0$. It is also assumed that the second term of (3.6) is negligible, corresponding to small sampling fraction. Hence (3.6) becomes:

$$AV_{nocosts} = \sigma^2 N^2 C_f^{-1}\overline{c}\left(\overline{\sqrt{z}}\right)^2. \tag{A.6}$$

From (A.5), and Lemma 1d, we have

$$\left(\overline{\sqrt{z}}\right)^2 = \frac{\overline{z}}{1+C^2_{\sqrt{z}}} \approx \frac{\overline{z}}{1+(1/4)C^2_z}.$$

Substituting into (A.6) gives (3.7).

Derivation of (4.2)

Two terms in (4.1) will be simplified using (2.5). Firstly,

$$\begin{aligned}\sum_{i\in U}\hat{c}_i^{1/2} z_i^{1/2} &= \sum_{i\in U} b_i^{1/2} c_i^{1/2} z_i^{1/2} \\ &= N\left(N^{-1}\sum_{i\in U} b_i^{1/2}\right)\left(N^{-1}\sum_{i\in U} c_i^{1/2} z_i^{1/2}\right) + C_{\sqrt{b},\sqrt{cz}}\end{aligned} \tag{A.7}$$

where $C_{\sqrt{b},\sqrt{cz}}$ is the covariance between the population values of $b_i^{1/2}$ and $c_i^{1/2} z_i^{1/2}$. Secondly,

$$\begin{aligned}\sum_{i\in U} z_i^{1/2}\hat{c}_i^{-1/2} c_i &= \sum_{i\in U} b_i^{-1/2} c_i^{1/2} z_i^{1/2} \\ &= N\left(N^{-1}\sum_{i\in U} b_i^{-1/2}\right)\left(N^{-1}\sum_{i\in U} c_i^{1/2} z_i^{1/2}\right) + C_{1/\sqrt{b},\sqrt{cz}}\end{aligned} \tag{A.8}$$

where $C_{1/\sqrt{b},\sqrt{cz}}$ is the covariance between the population values of $b_i^{-1/2}$ and $c_i^{1/2}z_i^{1/2}$.

If we assume that the population values of b_i are unrelated to the values of c_i and z_i, so that $C_{\sqrt{b},\sqrt{cz}} = C_{1/\sqrt{b},\sqrt{cz}} = 0$, and subsitute (A.7) and (A.8) into (4.1), then we obtain (4.2).

Derivation of (4.3)

We can express (4.2) in terms of AV_{opt} which is defined in (3.2), assuming the last term of (3.2) is negligible, corresponding to small sampling fraction:

$$AV_{ests} \approx AV_{opt} N^{-2} \sum_{i\in U} b_i^{-1/2} \sum_{i\in U} b_i^{1/2} \tag{A.9}$$

Lemma 1c implies that

$$N^{-2} \sum_{i\in U} b_i^{-1/2} \sum b_i^{1/2} = 1 + \frac{1}{4} C_b^2 + o\left(C_b^2\right) \approx 1 + \frac{1}{4} C_b^2.$$

Substituting this, and (3.3), into (A.9) gives (4.3).

References

Brewer, K. and Gregoire, T.G. (2009). Introduction to survey sampling. In *Handbook of Statistics 29A: Sample Surveys: Design, Methods and Applications*, eds. Pfeffermann, D. and Rao, C.R., Amsterdam: Elsevier/North-Holland, pp. 9-37.

Clark, R.G. (2009). Sampling of subpopulations in two-stage surveys. *Statistics in Medicine*, 28, 3697-3717.

Clark, R.G. and Steel, D.G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 63-82.

Cochran, W. (1977). *Sampling Techniques*. New York: Wiley, 3rd ed.

Dillman, D., Smyth, J. and Christian, L. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* John Wiley and Sons, 3rd ed.

Greenlaw, C. and Brown-Welty, S. (2009). A comparison of web-based and paper-based survey methods testing assumptions of survey mode and response cost. *Evaluation Review*, 33, 464-480.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. and Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 439-457.

Hansen, M., Hurwitz, W. and Madow, W. (1953). *Sample Survey Methods and Theory Volume 1: Methods and Applications*. New York: John Wiley and Sons.

Hicks, K. (2001). Cost and variance modelling for the 2001 redesign of the Monthly Population Survey. www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.037, *Australian Bureau of Statistics Methodology Advisory Committee Paper*.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.

Särndal, C., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. and Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, 80, 382-399.

Schouten, B., Brakel, J.v.d., Buelens, B., Laan, J.v.d. and Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35 (1), 101-113.

Schouten, B., Shlomo, N. and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 231-253.

Vannieuwenhuyze, J.T. and Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods and Research*, 42, 82-104.

Vannieuwenhuyze, J.T., Loosveldt, G. and Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74, 1027-1045.

Vannieuwenhuyze, J. T., Loosveldt, G., and Molenberghs, G. (2012). A method to evaluate mode effects on the mean and variance of a continuous variable in mixed-mode surveys. *International Statistical Review*, 80, 306-322.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 243-263
Statistics Canada, Catalogue No. 12-001-X

Optimal solutions in controlled selection problems with two-way stratification

Sun Woong Kim, Steven G. Heeringa and Peter W. Solenberger[1]

Abstract

When considering sample stratification by several variables, we often face the case where the expected number of sample units to be selected in each stratum is very small and the total number of units to be selected is smaller than the total number of strata. These stratified sample designs are specifically represented by the tabular arrays with real numbers, called controlled selection problems, and are beyond the reach of conventional methods of allocation. Many algorithms for solving these problems have been studied over about 60 years beginning with Goodman and Kish (1950). Those developed more recently are especially computer intensive and always find the solutions. However, there still remains the unanswered question: In what sense are the solutions to a controlled selection problem obtained from those algorithms optimal? We introduce the general concept of optimal solutions, and propose a new controlled selection algorithm based on typical distance functions to achieve solutions. This algorithm can be easily performed by a new SAS-based software. This study focuses on two-way stratification designs. The controlled selection solutions from the new algorithm are compared with those from existing algorithms using several examples. The new algorithm successfully obtains robust solutions to two-way controlled selection problems that meet the optimality criteria.

Key Words: Cell expectation; Probability sampling; Distance function; Optimum array; Linear programming problem; Simplex method.

1 Introduction

In the term, "Controlled Selection (or Controlled Sampling)", "control" has a broad meaning. The pioneering paper of Goodman and Kish (1950, page 351) defined controlled selection as "...any process of selection in which, while maintaining the assigned probability for each unit, the probabilities of selection for some or all preferred combinations of n out of N units are larger than in stratified random sampling".

The focus in this paper is upon **controls** required in deciding the number of units (e.g., primary sampling units (PSUs)) allocated to each stratum cell in a **two-way stratification design**, where the total number of units to be selected is smaller than the number of strata cells or the expected number of units to be selected from each stratum cell is very small. This assumes that given precision and cost constraints, simply reducing the number of strata cells or increasing the number of the sampled units is not appropriate for the design.

Here **controlled selection** refers to the following two-stage procedure. First, the **controlled selection problem** represented by a tabular array with real numbers formed by the two-way stratification design is solved according to a specified algorithm (or technique). The solution to the problem is a set of feasible arrays with nonnegative integer sample allocation to the cells of each array and probabilities of selection corresponding to each array. Second, a random selection of one of the solution arrays is made using the assigned probabilities. The integer number appearing in each cell of the selected solution array then serves as

1. Sun Woong Kim, Director, Survey & Health Policy Research Center, Professor, Department of Statistics, Dongguk University, 26, 3-Ga, Pil-Dong, Jung-Gu Seoul, South Korea 100-715. E-mail: sunwk@dongguk.edu; Steven G. Heeringa, Senior Research Scientist, Survey Research Center, Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106. E-mail: sheering@isr.umich.edu; Peter W. Solenberger, Applications Programmer Analyst Lead, Survey Research Center, Institute for Social Research, University of Michigan. E-mail: pws@isr.umich.edu.

the number of sample units to be allocated to that cell of the two-way stratification. The key to the controlled selection is the **algorithm** that defines a set of solution arrays that achieve the **controls** to solve the problem.

Many controlled selection techniques have been developed since Goodman and Kish (1950) first described the application of controlled selection to a specific problem of choosing 17 PSU's to represent the North Central States of the United States. Bryant, Hartley and Jessen (1960) proposed a simple method which was applicable in a limited number of sample situations. Raghunandanan and Bryant (1971) generalized their method and Chernick and Wright (1983) suggested an alternative. Jessen (1970) proposed two methods called "method 2" and "method 3", both quite complicated to implement and sometimes failing to provide a solution. Jessen (1978, chapter 11) introduced a simpler algorithm for solving controlled selection problems.

Hess, Riedel and Fitzpatrick (1975) gave a detailed explanation of how to use controlled selection in order to select a representative sample of Michigan's hospitals. Groves and Hess (1975) first suggested a formal computer algorithm for obtaining solutions to controlled selection problems with two-and three-way stratification. Heeringa and Hess (1983) reported the response to Roe Goodman's question: How does a computer solution of highly controlled selection compare with a manual solution? The answer was "For the same sample design, computer generated controlled selection often leads to slightly higher variances than does manual controlled selection; but since the differences in precision are small and manual controlled selection is laborious, computer generated controlled selection is preferred." Lin (1992) improved the algorithm of Groves and Hess (1975) and the software called "PCCONSEL" for their algorithm was presented by Heeringa (1998). Huang and Lin (1998) proposed a more efficient algorithm, which imposes additional constraints in the controlled selection problem with two-way stratification and uses any standard network flow computer package. Hess and Heeringa (2002) summarized investigations on controlled selection over 40 years that have been made at the Survey Research Center, University of Michigan.

Taking a different approach, Causey, Cox and Ernst (1985) proposed an algorithm that applied a transportation model to controlled selection problems with two-way stratification, based on the theory originally suggested in a previous paper of Cox and Ernst (1982). Winkler (2001) developed an integer programming algorithm quite similar to that of Causey et al. (1985). Deville and Tillé (2004) suggested an algorithm called the Cube method.

Following Rao and Nigam (1990, 1992), Sitter and Skinner (1994) applied a linear programming (LP) approach to solve controlled selection problems. Later, Tiwari and Nigam (1998) proposed an LP method that reduces the probabilities of selecting non-preferred samples.

In summary, many different algorithms for controlled selection have been investigated and described in the literature. Those most recently developed are especially **computer-intensive**, since they are highly dependent on available software and high speed computers. However, in spite of this evolution in the algorithms over about 60 years, a question still remains: In what sense are the solutions to a controlled selection problem obtained from those algorithms **optimal**?

In this paper, we define in Section 2 the two-way controlled selection problem and revisit several problems of this type that have appeared in the historical literature. In Section 3, we present the desirable constraints. In Section 4, we introduce our concept of **optimal solutions** to controlled selection problems. In Section 5, we describe the weaknesses in the previous algorithms. In Section 6, we suggest a new algorithm using the LP approach for achieving **optimal solutions** and a new publicly available **software** for implementing the new controlled selection algorithm is presented in Section 7. In Section 8, to show the

robustness of the new algorithm, it is applied to several example controlled selection problems and the results are compared to those obtained using existing algorithms. We conclude in Section 9.

2 Controlled selection problems

In order to select a sample of n units, consider a two-way stratification design classifying a population of N units by two criteria with R and C categories, respectively. The controlled selection problem under two-way stratification is defined by the $R \times C$ tabular array A, which consists of RC cells that have nonnegative real numbers a_{ij}, called the **cell expectations**, representing the expected number of units to be drawn in each cell ij. The standard two-way controlled selection problem is described as in Table 2.1.

Table 2.1
$R \times C$ **Controlled selection problem**

a_{11}	a_{12}	. . . . .	a_{1C}	$a_{1.}$
a_{21}	a_{22}	. . . . .	a_{2C}	$a_{2.}$
.	.		.	.
.	.		.	.
.	.	. . a_{ij} . .	.	$a_{i.}$
.	.		.	.
.	.		.	.
a_{R1}	a_{R2}	. . . . .	a_{RC}	$a_{R.}$
$a_{.1}$	$a_{.2}$	. . $a_{.j}$ . .	$a_{.C}$	$a_{..}(=n)$

The **marginal expectations** $a_{i.}$ and $a_{.j}$ denote the sum of cell expectations in each row category i and each column category j. Hence $a_{..}$ denotes the sum of all cell expectations and equals the total sample size n.

Although Table 2.1 takes a simple two-way tabular form, it should be noted that typically $n < RC$, and furthermore a_{ij} can be very small (e.g., often less than 1). In this case deciding how to allocate n units to cells, that is, how to obtain an $R \times C$ array with cells rounded to a nonnegative integer for each a_{ij}, requires an algorithm to solve the problem.

A variety of controlled selection problems are used as examples in the literature. The first example of a controlled selection problem was the 17×4 array, described by Goodman and Kish (1950, page 356), for allocating 17 PSU's to 68 cells given by 17 strata and 4 groups of North Central States in the United States. The array may be formed as follows. Let N_{ij} denote the number of population elements in each cell ij and let $N_{i.}$ denote the total number of population elements in each stratum. Then $a_{ij} = N_{ij} / N_{i.}$, where some N_{ij} are zero and $0 \le a_{ij} < 1$. All $a_{i.}$ equal the integer 1, whereas $a_{.j}$ are nonintegers sums of the a_{ij} in column j. The problem is therefore one of selecting one PSU per sample stratum (i dimension) and simultaneously controlling the distribution to state groups (j dimension). A total of $n = 17$ PSUs will be selected.

The following paragraphs describe four additional problems found in the literature that will be used in the discussion and comparative evaluations presented in this paper.

Problem 2.1: Jessen (1970)

A 3×3 problem involving two stratifying variables is given by Jessen (1970, page 779). Each cell ij corresponds to one PSU and $N=9$. A sample of size $n=6$ is drawn. $a_{ij}=nX_{ij}/X$, where X_{ij} is a "measure of size" for the PSU in cell ij and $X=\sum_{i=1}^{R}\sum_{j=1}^{C}X_{ij}$. Note that in this problem, $0<a_{ij}<1$, and both $a_{i.}$ and $a_{.j}$ are equal to 2.

Problem 2.2: Jessen (1978)

An extended 4×4 version of Problem 2.1 comes from Jessen (1978, page 375). In this problem, $N=16$ and $n=8$. As in Problem 2.1, both $a_{i.}$ and $a_{.j}$ are equal to 2, but $0\le a_{ij}\le 1$.

Problem 2.3: Causey et al. (1985)

Causey et al. (1985, page 906) describe an 8×3 two-way stratification problem designed to select 10 PSU's, that is, $n=10$. Let $X_{ijq_{ij}}\left(q_{ij}=1,\ldots,r_{ij}\right)$ be some measure of size for the PSU q_{ij} in cell ij. Here $a_{ij}=n\,X_{ijq}/X_q$, where $X_{ijq}=\sum_{q_{ij}=1}^{r_{ij}}X_{ijq_{ij}}$ and $X_q=\sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{q_{ij}=1}^{r_{ij}}X_{ijq_{ij}}$. Note that in this problem, $0\le a_{ij}\le 2$, and most $a_{i.}$ and $a_{.j}$ are noninteger values.

Problem 2.4: Winkler (2001)

Winkler (2001) provides the 5×5 controlled selection problem with two stratifying variables shown in Table 2.2.

The objective in solving this problem is to select $n=37$ sample units from the population of $N=1,251$. The problem definition begins with a 5×5 array with cell population sizes N_{ij}, where some N_{ij} are quite small. The marginal row and column expectations, $a_{i.}$ and $a_{.j}$, are integer-valued and are predetermined using the prior information on precision (e.g., coefficients of variation).

Table 2.2
5×5 Controlled selection problem

2.000	2.483	1.052	0.103	0.362	6
2.182	1.061	1.101	1.046	0.610	6
0.000	1.614	1.914	2.200	1.272	7
0.860	0.377	0.930	2.840	2.993	8
0.958	0.465	2.003	1.811	4.763	10
6	**6**	**7**	**8**	**10**	37

Source: Table 4, Appendix, Winkler (2001). Reproduced with permission.

The cell expectations, a_{ij}, are obtained by applying the generalized iterative fitting procedure (GIFP) of Dykstra (1985a, 1985b) and Winkler (1990) to the initial array. The GIFP is used to ensure that $a_{ij} < N_{ij}$ for the cells with small N_{ij}, when $a_{i.}$ and $a_{.j}$ are given. Note that in the Table 2.2, the a_{ij} are given to 3 decimal places, and $0 \le a_{ij} < 5$.

The common characteristic shared by these controlled selection problems is that, as mentioned above, the total number of selected units is smaller than the number of cells (except for Problem 2.4, where $n = 37 > RC = 25$) and many a_{ij} are less than 1. The algorithms used to solve these problems must enforce some strict constraints described in next section. As described in Section 4, the solution to a controlled selection problem obtained by any algorithm is a set of some $R \times C$ arrays and probabilities of selection corresponding to each array.

3 Desirable constraints

Each controlled selection problem of the form illustrated in Table 2.1 has many possible integer solutions. Let B_k denote one such solution, whose internal entries b_{ijk} are the replacement of the real numbers a_{ij} in the controlled selection problem A by the adjacent nonnegative integers. The entry, b_{ijk}, equals either $\lfloor a_{ij} \rfloor$ or $\lfloor a_{ij} \rfloor + 1$, where [ ] is the greatest integer function. If a_{ij} is a nonnegative integer, $b_{ijk} = a_{ij}$ for all k. The same rule is applied to the marginal expectations. As noted by Jessen (1970) and Causey et al. (1985), we primarily pay attention to B_k that simultaneously satisfy the following **constraints** for all i and j:

$$b_{ijk} \ge 0 \tag{3.1}$$

$$\left| b_{ijk} - a_{ij} \right| < 1 \tag{3.2}$$

$$\left| b_{i.k} - a_{i.} \right| < 1 \text{ and} \tag{3.3}$$

$$\left| b_{.jk} - a_{.j} \right| < 1, \tag{3.4}$$

where $b_{i.k} = \sum_{j=1}^{C} b_{ijk}$ equals either $\lfloor a_{i.} \rfloor$ or $\lfloor a_{i.} \rfloor + 1$, $b_{.jk} = \sum_{i=1}^{R} b_{ijk}$ equals either $\lfloor a_{.j} \rfloor$ or $\lfloor a_{.j} \rfloor + 1$, $\sum_{i=1}^{R} b_{i.k} = a_{..}$ and $\sum_{j=1}^{C} b_{.jk} = a_{..}$.

Consider the set of **all possible arrays**, $\mathfrak{B} = \{B_k, k = 1,\ldots,L\}$, satisfying (3.1) - (3.4). Since a_{ij} is the expectation of the sample allocation to each cell in A, the following **constraints** (3.5) and (3.6) on b_{ijk} in $B_k (\in \mathfrak{B})$ are especially important.

$$E\left(b_{ijk} \middle| i, j\right) = \sum_{B_k \in \mathfrak{B}} b_{ijk}\, p\left(B_k\right) = a_{ij}, \quad i = 1,\ldots,R, \text{ and } \quad j = 1,\ldots,C \tag{3.5}$$

and

$$\sum_{B_k \in \mathfrak{B}} p\left(B_k\right) = 1, \tag{3.6}$$

where $p\left(B_k\right)$, which depends on a specified algorithm for solving the controlled selection problem, is the selection probability of the array B_k and $p\left(B_k\right) \ge 0$.

Note that (3.5) and (3.6) will define a rigorous **probability sampling** method when randomly selecting any array in $\mathfrak{B}$. Also, note that since $\sum_{i=1}^{R}\sum_{j=1}^{C} E\left(b_{ijk}\middle|i,j\right) = a_{..}\sum_{B_k\in\mathfrak{B}} p\left(B_k\right) = a_{..}$, (3.5) implies (3.6) for any controlled selection problem such as those described in Problems 2.1 through 2.4. In addition, as an illustration, when applied to Problem 2.3, where $a_{ij} = n X_{ijq}/X_q$, (3.5) yields

$$E\left(b_{ijk}/X_{ijq}\middle|i,j\right) = a_{ij}/X_{ijq} = n/X_q, \tag{3.7}$$

which indicates the equal allocation for each cell.

4 Optimal solutions

Given the set of L possible arrays in $\mathfrak{B}$, consider the subset $\mathfrak{B}'(\subseteq\mathfrak{B})$ where

$$p\left(B_k\right) > 0.$$

A **solution set** to a controlled selection problem A denoted as

$$\left\{\left(B_k, p\left(B_k\right)\right), B_k \in \mathfrak{B}'\right\}$$

is the set of the arrays that have the required positive selection probabilities $\left(p\left(B_k\right)>0\right)$. This solution set, or simply a "solution" to the controlled selection problem, is usually obtained by an algorithm to control the constraints in (3.1) through (3.6). As described in the introduction, since Goodman and Kish (1950), many algorithms for obtaining solutions to controlled selection problems have been developed.

Until Groves and Hess (1975) suggested a computer algorithm, most solutions were manually obtained in a process that resembled solving a mathematical puzzle. Furthermore, for most problems it is possible that there is more than one solution set that meets the constraints. Since the 1980s, the computer-intensive controlled selection algorithms using transportation theory, network flow, integer programming, and LP have been developed. These algorithms may depend on highly specialized software or may be programmed to run in major software systems.

However, previous solutions ranging from manual to computer-intensive algorithms have rarely been compared empirically using a standard set of performance criteria. Therefore, we begin with the description of a concept called **optimal solution sets**, or more simply, **optimal solutions**.

The controlled selection problem A is only one array, but there may be many possible arrays in $\mathfrak{B}$. Also, only one array B_k from any solution to A is randomly chosen by $p\left(B_k\right)$ as the basis for choosing the stratified sample. In general then, we might define an optimal solution as that satisfying the following **requirements** (R1 and R2):

R1. The solution is obtained based on appropriate and objective measurements of the **closeness** between A and every single array B_k in $\mathfrak{B}$.

R2. The solution, as much as possible, maximizes the probabilities of selection over the **arrays nearest** to A under such measurements as referenced in R1.

The remainder of this section will address how to specify R1 and R2 for optimal solutions. First, in order to define **closeness** in R1, a real number $d(B_k : A)$ representing the distance between A and B_k, can be considered, where d is a distance function that satisfies the following axioms:

(i) $d(B_k, A) > 0$ if $B_k \neq A$; $d(A, A) = 0$;

(ii) $d(B_k, A) = d(A, B_k)$;

(iii) $d(B_k, A) \leq d(B_k, B_k') + d(B_k', A)$ for any $B_k' \in \mathfrak{B}$.

Axiom (iii) is termed the **triangle inequality axiom**. Distance functions satisfying (i), (ii), and (iii) can be defined by using the two-ordered RC-tuples $(a_{11}, a_{12}, \cdots, a_{RC})$ and $(b_{11k}, b_{12k}, \cdots, b_{RCk})$ for A and B_k. We first define the ordinary or **Euclidean distance** (**2-norm distance**):

$$d_2(B_k, A) = \left[\sum_{i=1}^{R} \sum_{j=1}^{C} (b_{ijk} - a_{ij})^2 \right]^{\frac{1}{2}}, \quad k = 1, \ldots, L. \tag{4.1}$$

This function is probably the most familiar measure to define the distance between B_k and A.

Also, we can define the function called the **Chebyshev distance** (**infinite norm distance**):

$$d_\infty(B_k, A) = \max \left\{ |b_{ijk} - a_{ij}| : i = 1, \ldots, R, \; j = 1, \ldots, C \right\}, \quad k = 1, \ldots, L. \tag{4.2}$$

These distance functions give rise to distinct distance spaces. Owing to (3.2), for any B_k, the following holds.

$$0 \leq d_2(B_k, A) < (RC)^{1/2} \tag{4.3}$$

and

$$0 \leq d_\infty(B_k, A) < 1. \tag{4.4}$$

For instance, for the 3×3 array in Problem 2.1 and the 8×3 array in Problem 2.3, $0 < d_2(B_k, A) < 3$ and $0 < d_2(B_k, A) < 4.9$, respectively.

Second, as mentioned in R2, regarding the **arrays nearest** to A under such measurements described in R1, consider the set of arrays in $\mathfrak{B}$ having the minimum d_2 or d_∞ value from A. Let $\mathfrak{B}_2 (\subseteq \mathfrak{B}')$ be the set of the arrays having the minimum d_2 value from A and $\mathfrak{B}_\infty (\subseteq \mathfrak{B}')$ be the set of the arrays having the minimum d_∞ value from A.

Assuming that all possible arrays in $\mathfrak{B}$ are known, we define **optimum arrays** as follows.

Definition. The arrays in $\mathfrak{B}_2 \cup \mathfrak{B}_\infty$ are called the **optimum arrays**.

Note that in the new algorithm for controlled selection to be described in Section 6, d_2 or d_∞ are chosen based on preference. We avoid defining the intersection of $\mathfrak{B}_2$ and $\mathfrak{B}_\infty$ as the optimum arrays because this may exclude the other arrays not in $\mathfrak{B}_2 \cap \mathfrak{B}_\infty$ with the same minimum d_2 (d_∞) value. We illustrate below that there may exist a very small number of optimum arrays relative to the number of all possible arrays in $\mathfrak{B}$ for any A. The details of how to find all possible arrays will be described in Section 6 and Section 7.

Illustrations.

For Problem 2.1 through Problem 2.4, it is noted that $\mathfrak{B}_2 \subseteq \mathfrak{B}_\infty$. Thus, it is possible to use d_∞ only in illustrating the optimum arrays.

1. For Problem 2.1, there are six possible arrays satisfying (3.1), (3.2), (3.3), and (3.4). That is, $\mathfrak{B} = \{B_k, k = 1,\ldots,6\}$, as given in Table 4.1. There exists only one optimum array, B_2, with the minimum value of $d_\infty = 0.5$.

Table 4.1
3×3 **Controlled selection problem, optimum array with** $d_\infty = 0.5$ **and the other arrays**

A	B_1	B_2	B_3	B_4	B_5	B_6
0.8 0.5 0.7	0 1 1	1 0 1	1 1 0	0 1 1	1 0 1	1 1 0
0.7 0.8 0.5	1 0 1	1 1 0	0 1 1	1 1 0	0 1 1	1 0 1
0.5 0.7 0.8	1 1 0	0 1 1	1 0 1	1 0 1	1 1 0	0 1 1
d_∞	0.8	0.5	0.7	0.8	0.8	0.8

2. Problem 2.2 has 30 possible arrays, and there are three optimum arrays, shown in Table 4.2

Table 4.2
4×4 **Optimum arrays with** $d_\infty = 0.6$

0 0 1 1	0 1 1 0	0 1 1 0
1 1 0 0	1 0 0 1	1 0 1 0
0 0 1 1	0 0 1 1	1 0 0 1
1 1 0 0	1 1 0 0	0 1 0 1

3. Problem 2.3 has 141 possible arrays. There are six optimum arrays, where each array has the same $d_\infty = 0.6$. One of them is given in Table 4.3.

Table 4.3
One of six optimum arrays with $d_\infty = 0.6$

1	2	0
1	0	1
0	0	0
1	1	0
1	1	0
0	0	1
0	0	0
0	0	0

4. There are 159 possible arrays for Problem 2.4, and there is only one optimum array given in Table 4.4.

Table 4.4
5×5 **Optimum array with** $d_\infty = 0.517$

2	3	1	0	0
2	1	1	1	1
0	2	2	2	1
1	0	1	3	3
1	0	2	2	5

Accordingly, based on the definition of optimum arrays as well as d_2 and d_∞ satisfying the axioms (i), (ii), and (iii), we suggest the following **specifications** (S1 and S2) of R1 and R2 of optimal solutions:

S1. The solution is based on the values of the distance $d_2\left(d_\infty\right)$ between A and every single array B_k in $\mathfrak{B}$.

S2. The solution maximizes the probabilities of selection of optimum arrays.

S1 and S2 will be the rudiments of a new algorithm presented in Section 6, and in the next section we turn into the discussion on the previous algorithms from the viewpoint of optimal solutions.

5 Non-optimal properties of existing methods

As described in Section 4, the algorithms for controlled selection may be divided into two parts, manual algorithms before 1980s and computer-intensive algorithms since then. For large controlled selection problems with many cells, the latter class of algorithms may be preferred. But when the problem is small, the former can be easily used without the complexity of the latter. Therefore we would not say that the former is always inferior to the latter. More objective criteria for comparing them would be necessary, and the optimal solution may be adopted as one of the better criteria to compare their strengths or weaknesses.

As discussed by Jessen (1978, pages 375-376), the algorithms of Jessen (1970) aim to minimize the number of arrays in a solution set $\mathfrak{B}'$, and the algorithm of Jessen (1978) quite easily achieves that purpose relative to those of Jessen (1970). Thus his algorithms pursue "simplicity" in formulating a solution rather than an optimal solution.

The algorithm of Causey et al. (1985) may give a "partially" optimal solution. Other than the original problem, A, it sequentially creates a small number of new controlled selection problems, and then as a solution it finds only one array $B_k\,(\in\mathfrak{B})$ to be **nearest** to each problem, starting with A. Each problem is regarded as the transportation problem of Cox and Ernst (1982), which is formed by the objective function mimicking the behavior of

$$\sum_{i=1}^{R}\sum_{j=1}^{C}\left|b_{ijk}-a_{ij}\right|^p,\ k=1,\ldots,L,\ 1\le p<\infty. \tag{5.1}$$

Note that since function (5.1) violates the triangle inequality axiom (iii), it is not a distance function. It needs the inclusion of the p-th root to be a distance function. Also, each $p(B_k)$ is calculated by a simple formula. In view of the optimality requirements given by R1 and R2 the Causey et al. algorithm has the following weaknesses: 1) Since other controlled selection problems in addition to the original problem A are involved, it is difficult to obtain the solution consistently based on the **closeness** between the unique A and every individual B_k in $\mathfrak{B}$; 2) The maximization of the probabilities of selection for the **arrays nearest** to A is not guaranteed.

Winkler (2001) presented a modification of the method of Causey et al. (1985). Instead of using the transportation problem, he proposed integer linear programming, resulting in slight changes of the $p(B_k)$. Nevertheless, the Winkler (2001) algorithm is not free from the weaknesses of the Causey et al. (1985) method.

Adopting a network flow problem approach, the Huang and Lin (1998) algorithm imposes the additional subgroup constraints in A, raised by Goodman and Kish (1950). However, it does not attain objectives R1 and R2, just as in Causey et al. (1985) and Winkler (2001), since a new network, instead of a new controlled selection problem, is generated at every iteration, an arbitrary $B_k (\in \mathfrak{B})$ is obtained as a solution to the network, and $p(B_k)$ is calculated by a simple formula.

In contrast, the LP algorithms proposed by Sitter and Skinner (1994) and Tiwari and Nigam (1998) use all possible arrays in $\mathfrak{B}$. Note that finding all those arrays is an important issue, and that $p(B_k)$ for all possible arrays are simultaneously obtained by running the software for LP only once. The key idea underlying the algorithm of Sitter and Skinner (1994) is to use a "loss function" defined by

$$\sum_{i=1}^{R}\left(b_{i.k}-a_{i.}\right)^2+\sum_{j=1}^{C}\left(b_{.jk}-a_{.j}\right)^2. \tag{5.2}$$

In terms of R1 and R2, their algorithm has the following disadvantages: 1) The closeness between A and B_k is not well captured by loss function (5.2). This is because it is not a distance function that satisfies axiom (iii), as the marginal totals are used, instead of the cell entries; 2) Loss function (5.2) is irrelevant to the maximization of the probabilities of selection over the **arrays nearest** to A in Problems 2.1, 2.2, and 2.4, since it is always zero.

The LP method of Tiwari and Nigam (1998) can be used to reduce the selection probabilities of non-preferred arrays (e.g., arrays not containing the PSU corresponding to the cell $ij = 23$ in Problem 2.1), which are initially determined by the samplers. For controlled selection problems with integer margins and without considering the non-preferred arrays, their method will give the same solutions as that of Sitter and Skinner (1994).

The solutions from these previous methods will be compared with those from the proposed method in Section 6, on several examples in Section 8.

6 Suggested method

In this section, we present the details on an algorithm for achieving S1 and S2 of optimal solutions described in Section 4.

6.1 The algorithm

The algorithm has the following **characteristics**: 1) it finds a solution directly based on the values of the distance d_2 (d_∞) between the controlled selection problem A and each individual array B_k in $\mathfrak{B}$; 2) it is computer-intensive, but easily implemented by LP; 3) it is applicable to any type of controlled selection problem with two-way stratification.

The algorithm has five steps. They are as follows:

Step 1. Find the set of all possible arrays, $\mathfrak{B}$, satisfying (3.1) - (3.4) for a given controlled selection problem A. Specifically, if there are any noninteger marginal expectations in A, find all possible roundings of these marginal expectations by adjacent integers, which satisfy (3.3) and (3.4). Those rounded marginal integers will be $[a_{i.}]$ or $[a_{i.}]+1$ $\left([a_{.j}] \text{ or } [a_{.j}]+1\right)$, while the integer marginal expectations will remain, since $[a_{i.}] = a_{i.}$ $\left([a_{.j}] = a_{.j}\right)$. Next, find all possible arrays satisfying (3.1) and (3.2) under the rounded marginal integers and the other marginal integers.

Step 2. Choose either $d_2^*(B_k, A)$ or $d_\infty^*(B_k, A)$ (based on preference) and compute the chosen distance function for each $B_k\,(\in \mathfrak{B})$, where:

$$d_2^*(B_k, A) = d_2(B_k^*, A^*) = \left[\sum_{i=1}^{R}\sum_{j=1}^{C}\left(b_{ijk}^* - a_{ij}^*\right)^2\right]^{\frac{1}{2}} \tag{6.1}$$

$$d_\infty^*(B_k, A) = d_\infty(B_k^*, A^*) = \max\left\{\left|b_{ijk}^* - a_{ij}^*\right| : i = 1,\ldots,R,\ j = 1,\ldots,C\right\}. \tag{6.2}$$

Note that since each of the ij cells in the problem array, A, will receive a minimum allocation equal to $[a_{ij}]$ with certainty the distance functions need only consider the non-integer part of a_{ij}:

$$a_{ij}^* = a_{ij} - \left[a_{ij}\right], \tag{6.3}$$

and the integer difference (either 0 or 1) between the allocated sample size, b_{ijk}, for solution $k = 1,\ldots,L$ and the certainty count for the ij-th cell of A:

$$b_{ijk}^* = b_{ijk} - \left[a_{ij}\right]. \tag{6.4}$$

Step 3. According to the distance function chosen in Step 2, construct the following LP problem consisting of the minimization of the objective function (6.5) or (6.6), which is a linear form, with the linear constraints (6.7) and (6.8):

Minimize

$$OF_1 = \sum_{B_k \in \mathfrak{B}} d_2^*(B_k, A)\, p(B_k) \tag{6.5}$$

or

$$OF_2 = \sum_{B_k \in \mathfrak{B}} d_\infty^*\left(B_k, A\right) p\left(B_k\right) \tag{6.6}$$

subject to

$$\sum_{B_k \in \mathfrak{B}} b_{ijk}^*\, p\left(B_k\right) = a_{ij}^*,\ i = 1,\ldots,R,\ \ j = 1,\ldots,C, \tag{6.7}$$

and

$$p\left(B_k\right) \geq 0,\ k = 1,\ldots,L. \tag{6.8}$$

Step 4. By using an algorithm for LP, solve the LP problem established in Step 3 with respect to L unknown variables

$$\left\{p\left(B_k\right), B_k \in \mathfrak{B}\right\}. \tag{6.9}$$

Step 5. Obtain the solution set $\left\{\left(B_k, p\left(B_k\right)\right), B_k \in \mathfrak{B}'\right\}$ to A consisting of arrays such that $p\left(B_k\right) > 0$ in the solution set to the LP problem obtained in Step 4.

Some remarks to be useful in implementing the algorithm are in order.

Remark 6.1. In Step 2, note that $\left\lceil a_{ij} \right\rceil$ in (6.3) or (6.4) indicates the number of units to be selected with certainty in each cell. Also, note that

$$d_2^*\left(B_k, A\right) = d_2\left(B_k, A\right) \tag{6.10}$$

and

$$d_\infty^*\left(B_k, A\right) = d_\infty\left(B_k, A\right), \tag{6.11}$$

since $b_{ijk}^* - a_{ij}^* = b_{ijk} - a_{ij}$ due to (6.3) and (6.4).

Remark 6.2. In addition to the fact that d_2^* is the natural concept of distance and d_∞^* is the simplest and easiest to compute under the norm, there is sensible advice on the choice of d_2^* or d_∞^* in Step 2. Let D_2 and D_∞ be the sets of the distance values for all possible arrays calculated by d_2^* and d_∞^*, respectively. Let those arrays with the same distance value in D_2 (D_∞) be in the same group. Then logically, d_2^* would cluster possible arrays into many different groups, where the number of groups is larger than in d_∞^*, due to (4.3) and (4.4). Accordingly, when using d_2^* in LP problem, the number of arrays in $\mathfrak{B}$ such that $p(B_k) > 0$ would be larger than in using d_∞^*.

Remark 6.3. It is clear from (6.5) and (6.6) involving the distance values d_2^* or d_∞^* that the solution in Step 5 results in the safe achievement of S1. Furthermore, S2 is achieved efficiently using linear constraints (6.7) and (6.8).

Remark 6.4. In constructing the LP problem in Step 3, the constraints for the cells with $a_{ij}^{*}=0$ can be omitted in (6.7). For example, for the 5×5 controlled selection problem of Problem 2.4, the number of necessary constraints is 23, since two cells have $a_{ij}^{*}=0$. Also, the linear constraint (3.6) is not essential, because it is implied in (6.7).

6.2 Using the simplex method

The LP problem constructed in Step 3 with the system of constraints of RC equations in (6.7) for L nonnegative unknowns in (6.8) is in the "standard form" and no transformation is required.

Supposing that $RC<L$, the number of equations is smaller than the number of unknowns. Consequently it is an LP problem with a standard form, and it can always be solved by the simplex method by transforming with the system of RC constraints in canonical form. To change the system into canonical form, one could arbitrarily choose RC variables among L variables as **basic variables** and then, using a pivot operation, attempt to put the system into canonical form, where each basic variable has coefficient one in one equation and zero in the others, and each equation has exactly one basic variable with coefficient one.

Letting the other $L-RC$ variables except RC variables chosen as basic variables be 0 in the system in canonical form, the initial **basic feasible solution** is obtained. Next, by replacing exactly one basic variable, another basic feasible solution is obtained, and these steps are continued until the minimal value of the objective function is attained by any basic feasible solution. The set of these basic feasible solutions to the LP problem is convex. Many software packages for the simplex method are available for solving the LP problem. See Dantzig (1963) and Thie and Keough (2008, chapter 3) for the details on the simplex method.

6.3 The computational demands of the LP problem

It may be claimed that our algorithm is computationally expensive due to the following burdens:

a. Before solving the LP problem, all possible arrays to the controlled selection problem should be known.

b. The number of unknowns in the LP problem, L, is equal to the number of all possible arrays, which becomes large as RC, the number of cells in the controlled selection problem, increases. Hence, it is not unreasonable that L may be as large as the binomial coefficient

$$\binom{RC}{a_{..}^{*}}, \text{ where } a_{..}^{*}=a_{..}-\sum_{i=1}^{R}\sum_{j=1}^{C}\left[a_{ij}\right]. \tag{6.12}$$

c. If RC is large, it also yields a large number of constraints in (6.7).

Sitter and Skinner (1994), and Tiwari and Nigam (1998) also referred to these potential disadvantages in describing their LP algorithms. However, due to the following reasons, these computational burdens stated in a, b, and c may not be prohibitive in **actual operations**.

First, finding all possible arrays manually might be difficult for any controlled selection problem with a large number of cells, but this task is greatly simplified using an efficient algorithm and the power of modern computers. Using the software described in the next section, they can be easily obtained in seconds even in comparatively large problems such as Problems 2.3 and 2.4.

Second, applying (6.12) to Problems 2.1 through 2.4, respectively yields 84; 11,440; 10,626; and 4,457,400 arrays. However, the actual numbers for L are only 6, 30, 141 and 159, respectively. This is because marginal expectations of both rows and columns are simultaneously matched and some cell expectations are zero. The actual numbers can also be obtained from the software described in the next section.

Third, although the greater RC, the greater the number of constraints in the LP problem, the computational demands may depend on L as well as RC, and more specifically, on the number of basic feasible solutions, possibly denoted by

$$S = \binom{L}{RC}. \tag{6.13}$$

For example, if $L = 1{,}000$ and $RC = 100$, (6.13) gives 6.4E+139, which is an extremely large number. In this case, it is almost impossible to solve the LP problem, since each basic feasible solution should be investigated. But such cases would not happen in practice. According to Ross (2007, pages 221-224), when $RC < L$, the **number of necessary transitions**, say T, moving along the basic feasible solutions in solving the LP problem with standard form is approximately normally distributed with mean $E(T) = \log_e S$ and variance $Var(T) = \log_e S$, where

$$\log_e S \approx RC\left[1 + \log_e\left\{\left(L/RC\right) - 1\right\}\right]. \tag{6.14}$$

When applying this theory to the case of $L = 1{,}000$ and $RC = 100$, approximating both the mean and variance of T by (6.14) becomes 320, and the 95% confidence interval (CI) of T is (285, 355), which is smaller than the expected lower and upper limits.

Table 6.1
Comparison between S and T

	Problem 2.1	Problem 2.2	Problem 2.3	Problem 2.4
L	6	30	141	159
RC^*	9	14	13	23
S	NA	1.5E+8	7.9E+17	3.1E+27
$E(T)$	NA	16	43	64
95% CI of T	NA	(8, 24)	(30, 56)	(48, 80)

Note: NA - not available

Table 6.1 shows the results of the comparison between S and T for the four problems considered above. Note that due to Remark 6.4, RC in (6.13) and (6.14) is replaced by RC^*, that is, the number that results from subtracting the number of cells with $a_{ij}^* = 0$ from RC. The theory on T is not applied to Problem 2.1 because $RC^* > L$.

As shown in the table, the mean or confidence interval bounds of T are considerably smaller than S in each problem. In Section 8, T in Table 6.1 will be compared with the **actual number of transitions**, say t.

7 Software

To take the advantages of the power of modern computing, we have developed a public use SAS-based software called the SOCSLP (Software for Optimal Controlled Selection Linear Programming) for our algorithm to solve controlled selection problems with two-way stratification. The recent version may be downloaded from the URL: http://www.isr.umich.edu/src/smp/socslp.

In using the software, there are no restrictions on the number of all possible arrays that can be considered for the solution. The number of those arrays and the number of constraints that can be solved depend on the memory capacity and the available disk space of the computer.

The two-phase revised simplex method, implemented using SAS/OR LP Procedure, simply "PROC LP", is employed to solve the LP problem. A unique optimal solution to the LP problem is obtained when the objective function is minimized under the given constraints (6.7) through phase 1 and 2 of PROC LP, with the assumption that all unknown variables are nonnegative (6.8).

The software produces much information including the solution set to the controlled selection problem. Also, by choosing a simple option in the software, one array can be randomly selected from the solution set, completing the controlled selection. The SOCSLP is currently available for personal computers, and the details are provided through the User Guide on the website.

8 Comparisons of algorithms

Using the four controlled selection problems given in Section 2, we present some results from the **two methods** using d_2^* and d_∞^* in the new algorithm, and compare the solutions for these two methods to solutions generated under the algorithms previously described by Jessen (1970), Jessen (1978), Causey et al. (1985), Huang and Lin (1998), and Winkler (2001). The solutions from the two methods using d_2^* and d_∞^* were obtained by implementing the SOCSLP, running on the version 9.2 of SAS/OR (2008). Solutions for the algorithm of Sitter and Skinner (1994) using LP were also obtained using PROC LP of the version 9.2 of SAS/OR (2008). Solutions for the other methods are the results as they appeared in the original papers.

The answers to two questions help us compare the algorithms: 1) Are the solutions from the new methods different from those of the previous algorithms described in Section 5? 2) Do the solutions from the new methods give higher probabilities of selection for optimum arrays compared to those generated using the previous methods?

Prior to the comparison of the algorithms, we need to take a look at the results in Table 8.1 obtained from the two methods. In the table, the method using d_2^* and the one using d_∞^* are denoted by N_2 and N_∞, respectively. Since when calculated by d_2^* (d_∞^*), the arrays with the same distance value are in the same group, there would be different groups for all possible arrays (see Remark 6.2). Let G denote the number of the different groups. Also, let OF be the actual value of the objective function (6.5) or (6.6) and t the actual number of T, the number of transitions, introduced in Section 6.3. They are all obtained from the SOCSLP, and t especially indicates the number of iterations in phase 1 and 2 of the PROC LP in the software.

Table 8.1
Results with the new methods

	Problem 2.1		Problem 2.2		Problem 2.3		Problem 2.4	
	N_2	N_∞	N_2	N_∞	N_2	N_∞	N_2	N_∞
G	4	3	9	2	6	2	157	14
OF	1.336	0.620	1.689	0.640	1.582	0.720	1.661	0.701
t	2	2	8	6	18	15	43	41

As seen in the table, most values of G are much smaller than L, the number of all possible arrays given in Table 6.1, except for the case of the large value of "157" for Problem 2.4, which arises simply due to the fact that the a_{ij} are given to three decimal places. When using d_2^*, the values of OF range between 1 and 2, while they are always less than 1, when using d_∞^*. Most values of t do not reach the 95% CI of T shown at the bottom of Table 6.1. Thus, the actual computational demands are less than those expected in the theory.

The solutions from different algorithms for the first three problems are presented in order in Table 8.2 through Table 8.4. Results for Problem 2.4 are simply described below. (The table of solutions to this problem is available on request.) In Table 8.2, the method of Sitter and Skinner (1994), Jessen's (1970) method 2 and method 3 are denoted by SS, $J2$ and $J3$, respectively. The solutions for $J2$ and $J3$ in the table are from Jessen (1970, page 782). The table shows that all methods except Jessen's (1970) method 3 yield the same solution for the 3 x 3 array Problem 2.1. In the common solutions, the probability of selection for the optimum arrays, denoted by $\sum_{B_k \in \mathfrak{B}_\infty} p(B_k)$, is 0.5.

Table 8.2
Comparison of solutions to Problem 2.1

B_k	$p(B_k)$				
	N_2	N_∞	SS	$J2$	$J3$
0 1 1 1 0 1 1 1 0	0.2	0.2	0.2	0.2	0.1
1 0 1 * 1 1 0 0 1 1	0.5	0.5	0.5	0.5	0.4
1 1 0 0 1 1 1 0 1	0.3	0.3	0.3	0.3	0.2
0 1 1 1 1 0 1 0 1					0.1
1 0 1 0 1 1 1 1 0					0.1
1 1 0 1 0 1 0 1 1					0.1
Total	1.0	1.0	1.0	1.0	1.0
Total †	0.5	0.5	0.5	0.5	0.4

Note: * – Optimum array
† – The sum of probabilities of selection for optimum arrays

In Table 8.3, Jessen's (1978) method is denoted by JS. The solution for JS in the table is from Jessen (1978, pages 375-376). As shown in the table, the new methods using d_2^* and d_∞^* have the same solution for the Problem 2.2 4×4 array; however only one-half of the arrays in those solutions overlap with the arrays in the solutions from the methods of Sitter and Skinner (1994) and Jessen (1978). Also, the Sitter and Skinner and Jessen methods provide a lower probability of 0.6 to optimum arrays, whereas the new methods allocate the higher probability of 0.8 to the arrays.

Table 8.3
Comparison of solutions to Problem 2.2

B_k	$p(B_k)$			
	N_2	N_∞	SS	JS
0 0 1 1 0 1 0 1 1 1 0 0 1 0 1 0	0.2	0.2		
0 0 1 1 * 1 1 0 0 0 0 1 1 1 1 0 0	0.2	0.2	0.4	0.2
0 1 1 0 * 1 0 0 1 0 0 1 1 1 1 0 0	0.2	0.2		
0 1 1 0 * 1 0 1 0 1 0 0 1 0 1 0 1	0.4	0.4	0.2	0.4
0 1 1 0 0 0 1 1 1 1 0 0 1 0 0 1			0.2	
0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0			0.2	
0 1 1 0 1 0 0 1 0 1 0 1 1 0 1 0				0.2
0 0 1 1 0 1 0 1 1 0 1 0 1 1 0 0				0.2
Total	1.0	1.0	1.0	1.0
Total †	0.8	0.8	0.6	0.6

See note for Table 8.2.

Problem 2.3, with 141 possible arrays, is considerably larger than the above two problems. The solutions to this problem under the five methods are compared in Table 8.4. In the table, the methods of Causey et al. (1985) and Huang and Lin (1998) are denoted by CA and HU, respectively. The solutions for CA and HU in the table are from Causey et al. (1985, page 906) and Huang and Lin (1998, Figure 3), respectively.

Table 8.4
Comparison of solutions to Problem 2.3

B_k	$p(B_k)$					B_k	$p(B_k)$					B_k	$p(B_k)$				
	N_2	N_∞	SS	CA	HU		N_2	N_∞	SS	CA	HU		N_2	N_∞	SS	CA	HU
0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 1 0 0 0 1 0 0 0	0.2	0.2	0.2			0 2 0 1 0 1 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1			0.11			0 2 0 1 0 1 0 0 0 2 0 0 1 0 0 0 1 0 0 0 1 0 0 1				0.2	
0 2 0 1 0 1 1 0 0 1 0 1 1 0 1 0 0 0 0 1 0 0 0 0	0.1	0.2	0.03			0 2 0 1 0 1 1 0 0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0			0.03			0 2 0 1 0 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 0 0				0.2	0.2
0 2 0 1 0 1 1 0 0 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0	0.1					0 2 0 1 0 1 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 1			0.03			0 2 0 2 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 0				0.2	
0 2 0 2 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 0 0 1	0.1					0 2 0 2 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 0			0.09			0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 0 0 0 1 0 0 0 1					0.2
0 2 0 2 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 1	0.1					0 2 0 2 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0			0.08			0 2 0 2 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0					0.2
1 2 0* 1 0 1 0 0 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0	0.1		0.08			0 2 0 2 0 1 0 0 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 0			0.03								
1 2 0* 1 0 1 0 0 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 0	0.3	0.4	0.2	0.4	0.4	1 2 0 1 0 1 0 0 0 1 0 1 1 0 0 0 0 0 0 1 0 0 0 1			0.06								
0 2 0 2 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 1		0.2				1 2 0 1 0 1 0 0 0 1 0 1 1 1 0 0 1 0 0 0 0 0 0 0			0.06			Total	1.0	1.0	1.0	1.0	1.0
												Total †	0.4	0.4	0.28	0.4	0.4

See note for Table 8.2.

We note that all these methods provide different solutions, and about half of the arrays overlap between the new methods and the method of Sitter and Skinner (1994). Moreover, the solutions from the methods of Causey et al. (1985) and Huang and Lin (1998) are quite unlike the solution from the method using d_∞^*. The

method using d_2^* and Sitter and Skinner's method distribute the probabilities of selection to two optimum arrays, whereas the other three methods just allocate the probability to only one optimum array. Sitter and Skinner's method appears to be less effective in selecting optimum arrays since their method gives the probability of 0.28 to those, while the others give the higher probability of 0.4.

The solutions to Problem 2.4, which is the largest of the given problems, are compared under the four methods (N_2 , N_∞ , SS , and Winkler's (2001) method). Only two arrays, including one optimum, overlap in the solutions, and the two new methods give the same probabilities (0.127 and 0.483) to those arrays. Even when comparing the method using d_∞^* with the methods of Sitter and Skinner (1994) and Winkler (2001), their solutions are very different. Also, the new methods give the same probability of selection of 0.483 to the optimum array, whereas the other previous methods give the lower probabilities of 0.385 and 0.104, respectively.

In summary, it seems that the new methods successfully achieve S1 and S2 of optimal solutions. Note that the new methods consistently give higher probabilities of selection for optimum arrays and that the totals of those probabilities are always the same. The solutions from the new methods are very different from those obtained using previous methods, when the controlled selection problems are not small. This implies that the solutions from the previous methods may be far from optimal under criteria S1 and S2 (R1 and R2).

9 Concluding remarks

In this paper, we introduced the concept of optimal solutions to a controlled selection problem with two-way stratification, and proposed a new algorithm for finding such solutions. The algorithm has been easily and successfully implemented in the new SAS-based software (SOCSLP).

Since an optimal solution is a general idea, it may be adopted as one of the useful criteria for comparing the different algorithms. As shown in the above comparisons, the new algorithm results in solutions to large controlled selection problems that are very different from those derived using previously published methods. It is also likely to yield greater probabilities of selection for optimum arrays as compared to those obtained by the previous methods.

Based on the results for the two-way controlled selection problems, we expect that the suggested method would also contribute to improvements in the properties of solutions to controlled selection problems with three-way or more stratification dimensions.

Acknowledgements

This paper is in honor of I. Hess who dedicated her life to studying controlled selection. The authors wish to thank Jea-Bok Ryu in Chongju University for providing ideas and advice in the early stage of this study. We are also grateful to two anonymous referees, the Editor and the Associate Editor for their valuable comments and suggestions.

References

Bryant, E.C., Hartley, H.O. and Jessen, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.

Causey, B.D., Cox, L.H. and Ernst, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

Chernick, M.R. and Wright, T. (1983). Estimation of a population mean with two-way stratification using a systematic allocation scheme. *Journal of Statistical Planning and Inference*, 7, 219-231.

Cox, L.H. and Ernst, L.R. (1982). Controlled rounding. *INFOR: Information Systems and Operational Research*, 20, 423-432.

Dantzig, G.B. (1963). *Linear Programming and Extensions*. Princeton University Press, Princeton, New Jersey.

Deville, J-C and Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91, 893-912.

Dykstra, R. (1985a). An iterative procedure for obtaining I-projections onto the intersection of convex sets. *Annals of Probability*, 13, 975-984.

Dykstra, R. (1985b). Computational aspects of I-projections. *Journal of Statistical Computation and Simulation*, 21, 265-274.

Goodman, R. and Kish, L. (1950). Controlled selection – a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Groves, R.M. and Hess, I. (1975). An algorithm for controlled selection. In *Probability Sampling of Hospitals and Patients*, *Second Edition*, (Eds., I. Hess, D.C. Ridel and T.B. Fitzpatrick), Health Administration Press, University of Michigan, Ann Arbor, USA.

Heeringa, S.G. (1998). PCCONSEL user guide. In *Controlled Selection Continued*, *2002 Edition*, (Eds., I. Hess and S.G. Heeringa), Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

Heeringa, S.G. and Hess, I. (1983). More on controlled selection. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 106-111.

Hess, I. and Heeringa, S.G. (2002). *Controlled Selection Continued.* Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

Hess, I., Ridel, D.C. and Fitzpatrick, T.B. (1975). *Probability Sampling of Hospitals and Patients*, *Second Edition*. Health Administration Press, University of Michigan, Ann Arbor, USA.

Huang, H.C. and Lin, T.K. (1998). On the two-dimensional controlled selection problem. In *Controlled Selection Continued*, *2002 Edition*, (Eds., I. Hess and S.G. Heeringa), Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, USA.

Jessen, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-796.

Jessen, R.J. (1978). *Statistical Survey Techniques*. New York: John Wiley and Sons.

Lin, T.K. (1992). Some improvements on an algorithm for controlled selection. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 407-410.

Raghunandanan, K. and Bryant, E.C. (1971). Variance in multi-way stratification. *Sankhyā, Series A*, 33, 221-226.

Rao, J.N.K. and Nigam, A.K. (1990). Optimal controlled sampling designs. *Biometrika*, 77, 807-814.

Rao, J.N.K. and Nigam, A.K. (1992). "Optimal" controlled sampling: a unified approach. *International Statistical Review*, 60, 89-98.

Ross, S.M. (2007). *Introduction to Probability Models*. Burlington, MA: Academic Press.

SAS/OR (2008). *User's Guide: Mathematical Programming*. Version 9.2, Cary, NC: SAS Institute Inc.

Sitter, R.R. and Skinner, C.J. (1994). Multi-way stratification by linear programming. *Survey Methodology*, 20 (1), 65-73.

Thie, P.R. and Keough, G.E. (2008). *An Introduction to Linear Programming and Game Theory*, *Third Edition*. Hoboken, New Jersey: John Wiley and Sons.

Tiwari, N. and Nigam, A.K. (1998). On two-dimensional optimal controlled selection. *Journal of Statistical Planning and Inference*, 69, 89-100.

Winkler, W.E. (1990). On Dykstra's iterative fitting procedure. *Annals of Probability*, 18, 1410-1415.

Winkler, W.E. (2001). Multi-way survey stratification and sampling. U.S. Census Bureau, *Statistical Research Division Report RRS 2001/01*. Available from: //www.census.gov/srd/www/byyear.html.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 265-283
Statistics Canada, Catalogue No. 12-001-X

On aligned composite estimates from overlapping samples for growth rates and totals

Paul Knottnerus[1]

Abstract

When monthly business surveys are not completely overlapping, there are two different estimators for the monthly growth rate of the turnover: (i) one that is based on the monthly estimated population totals and (ii) one that is purely based on enterprises observed on both occasions in the overlap of the corresponding surveys. The resulting estimates and variances might be quite different. This paper proposes an optimal composite estimator for the growth rate as well as the population totals.

Key Words: Business surveys; Coefficient of variation; General restriction estimator; Kalman equations; Panels; Variances.

1 Introduction

In many countries a monthly business survey is held for the major Standard Industrial Classification (SIC) codes to estimate the level of the monthly turnover and the change in that level compared to a month or a year ago. When repeatedly sampling a population, a complicating factor is that there are various methods for estimating the (relative) change from a panel with different outcomes especially when the samples on different occasions are not completely overlapping.

Kish (1965), Tam (1984), Laniel (1987), Hidiroglou, Särndal and Binder (1995), Nordberg (2000), Berger (2004), Qualité and Tillé (2008), Wood (2008) and Knottnerus and Van Delden (2012) examined various estimators for the parameter of change in different situations. The main aim of this paper is to derive estimators for a relative change as well as the corresponding population totals that are in line with each other and that have minimum variance property. The derivation of the aligned composite estimators is based on the general restriction (GR) estimator of Knottnerus (2003). Composite estimators for totals and (absolute) changes are also proposed by Särndal, Swensson and Wretman (1992, pages 370-378) but in separate steps. Moreover, this paper focuses on estimators for growth rates because: (i) users of figures from business surveys for a specific SIC code often are more interested in growth rates than in absolute changes, (ii) in practice there might be model-assisted reasons to look at growth rates (auxiliary variables in regression models often explain the different growth rates of the units rather than their different levels), and (iii) growth rates are needed for making an overall index for the (monthly) turnover for each of the major SIC codes. For instance, Smith, Pont and Jones (2003) describe the method of matched pairs to measure a change from month to month, using responses that are common to both periods. The authors use this method for deriving the monthly retail sales index (RSI).

The outline of the paper is as follows. Section 2 briefly describes two methods for estimating a growth rate of the total turnover for enterprises with a certain SIC code. Two examples illustrate the possibly substantial differences between the two approaches. Section 3 discusses the question of which estimation

1. Paul Knottnerus, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. Email: pkts@cbs.nl.

method is to be preferred and explains as to why the difference between the variances of both estimators might be so large. For various situations Section 4 and Section 5 propose an optimal composite estimator. Section 6 discusses some extensions of the aligned composite (AC) estimator for growth rates and totals. Section 7 summarizes the main conclusions and issues to be further investigated.

2 Two estimators for the growth rate of the total turnover

Consider a population of N enterprises $U = \{1, ..., N\}$, and suppose there are no births and deaths in the population. Let Y_i denote the value of the turnover for the i-th enterprise in a given month (say t) and X_i the value of the turnover of that enterprise in month $t-12$. Hence, the variables y and x concern the same variable on two different occasions. Denote their population totals by Y and X, and their population means by $\overline{Y}$ and $\overline{X}$, respectively. That is, $Y = \sum_{i \in U} Y_i$, $X = \sum_{i \in U} X_i$, $\overline{Y} = Y/N$ and $\overline{X} = X/N$. Let s_1, s_2 and s_3 denote three mutually disjoint simple random samples from U without replacement (SRS). Define s_{12} and s_{23} by $s_{12} = s_1 \cup s_2$ and $s_{23} = s_2 \cup s_3$, respectively. Denote the size of s_k by n_k $(k = 1, 2, 3, 12, 23)$ and the corresponding sample means by $\overline{y}_k$ and $\overline{x}_k$. Let the variable x be observed in s_{12} on the first occasion and the variable y in s_{23} on the second occasion. Denote the overlap ratios by λ $(= n_2/n_{12})$ and μ $(= n_2/n_{23})$. The SRS estimators for the population totals Y and X are defined by $\hat{Y}_{SRS} = N\overline{y}_{23}$ and $\hat{X}_{SRS} = N\overline{x}_{12}$, respectively.

Define the growth rate g of the total turnover between the two occasions by $g = G - 1$ with $G = Y/X$. For estimating G there are two options. One of the standard (STN) options is based on the estimated totals on both occasions, that is

$$\hat{G}_{STN} = \frac{\hat{Y}_{SRS}}{\hat{X}_{SRS}} = \frac{\overline{y}_{23}}{\overline{x}_{12}}; \tag{2.1}$$

see Nordberg (2000), Qualité and Tillé (2008) and Knottnerus and Van Delden (2012). Note that the estimator $\hat{g}_{STN} = \hat{G}_{STN} - 1$ for g has the same variance as $\hat{G}_{STN}$. For sufficiently large n this variance can be approximated by using a first-order Taylor series expansion of $\hat{G}_{STN}$. That is,

$$\begin{aligned} \operatorname{var}\left(\hat{G}_{STN}\right) &\approx \frac{1}{\overline{X}^2} \operatorname{var}\left(\overline{y}_{23} - G\overline{x}_{12}\right) \\ &= \frac{1}{\overline{X}^2}\left\{\operatorname{var}\left(\overline{y}_{23}\right) + G^2 \operatorname{var}\left(\overline{x}_{12}\right) - 2G \operatorname{cov}\left(\overline{y}_{23}, \overline{x}_{12}\right)\right\} \\ &= \frac{1}{\overline{X}^2}\left\{\left(\frac{1}{n_{23}} - \frac{1}{N}\right) S_y^2 + G^2\left(\frac{1}{n_{12}} - \frac{1}{N}\right) S_x^2 - 2G\left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right) S_{xy}\right\}, \end{aligned} \tag{2.2}$$

where $S_y^2 = \sum_U \left(Y_i - \overline{Y}\right)^2 / (N-1)$ is the adjusted population variance of the Y_i and S_x^2 that of the X_i while $S_{xy} = \sum_U \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) / (N-1)$ is their adjusted population covariance. Cochran (1977, page 153) suggests as *working rule* to use the large-sample result if the sample size exceeds 30 and the coefficients of variation of the numerator and denominator are less than 10%. For (different) derivations

of the expression for $\text{cov}(\bar{y}_{23}, \bar{x}_{12})$ used in (2.2), see Tam (1984) and Knottnerus and Van Delden (2012). The adjusted population (co)variances can be estimated unbiasedly by the sample (co)variances; recall sample (co)variances s_{yk}^2 and s_{yxk} from sample s_k $(k = 1, 2, 3, 12, 23)$ are defined by

$$s_{yk}^2 = \frac{1}{n_k - 1} \sum_{i \in s_k} (Y_i - \bar{y}_k)^2$$

$$s_{yxk} = \frac{1}{n_k - 1} \sum_{i \in s_k} (Y_i - \bar{y}_k)(X_i - \bar{x}_k).$$

An alternative option for estimating G and g is based on enterprises observed on both occasions in overlap s_2 (OLP). That is,

$$\hat{G}_{OLP} = \frac{\bar{y}_2}{\bar{x}_2} \tag{2.3}$$

For sufficiently large n_2, the well-known approximation for the variance of this estimator is

$$\begin{aligned} \text{var}\left(\hat{G}_{OLP}\right) &\approx \frac{1}{\bar{X}^2} \text{var}\left(\bar{y}_2 - G\bar{x}_2\right) \\ &= \frac{1}{\bar{X}^2}\left(\frac{1}{n_2} - \frac{1}{N}\right) S_{y-Gx}^2, \end{aligned} \tag{2.4}$$

where S_{y-Gx}^2 stands for $S_y^2 + G^2 S_x^2 - 2GS_{xy}$; see Cochran (1977, page 31). In order to get some more insight into the merits of both $\hat{g}_{STN}$ and $\hat{g}_{OLP}$, consider the following examples.

Example 2.1. The data used in this example are panel observations on the turnover of Dutch supermarkets in February 2011 and 2012 from stratum 3 (size class 3). The stratum size is $N = 386$. Furthermore, $n_1 = 15$, $n_2 = 57$ and $n_3 = 17$. For the different samples we have (in thousand euros)

$$\bar{y}_{23} = 97.2,\ \bar{x}_{12} = 89.8,\ s_{y23}^2 = 3,781,\ \text{and}\ s_{x12}^2 = 2,232.$$

The population correlation coefficient $\rho_{xy}\left(= S_{xy}/S_x S_y\right)$ between the Y_i and the X_i is estimated from overlap s_2 by $\hat{\rho}_{xy2} = s_{xy2}/s_{y2}s_{x2} = 0.876$. To avoid negative variance estimates, Knottnerus and Van Delden (2012) propose estimating S_{xy} in (2.2) by $\hat{S}_{xy} = \hat{\rho}_{xy2} s_{x12} s_{y23} = 2,545$. Substituting the above outcomes into (2.1) and (2.2), we obtain $\hat{g}_{STN} = 0.082\ (= 8.2\%)$ and $\hat{\text{var}}(\hat{g}_{STN}) = 0.00324$. Assuming normality and using $u_{0.975} = 1.96$, the 95%-confidence interval is approximately $I_{STN}^{95} \approx (-3.0\%, 19.4\%)$. In contrast, from overlap s_2 we get the estimates

$$\bar{y}_2 = 102.2,\ \bar{x}_2 = 97.3\ \text{and}\ \hat{g}_{OLP} = 0.050\ (= 5.0\%).$$

Substituting the same estimates as before for $\bar{X}$ and the (co)variances of the X_i and Y_i into (2.4) yields $\hat{\text{var}}(\hat{g}_{OLP}) = 0.00166$. Under the normality assumption this yields a smaller 95%-confidence interval $I_{OLP}^{95} \approx (-3.0\%, 13.0\%)$.

Example 2.2. Among the data of Example 2.1 there were three enterprises with extreme g-values of -50%, 133% and -91%. It is beyond the scope of this paper to further analyse or correct these outliers. But

to illustrate the difference between the estimators $\hat{g}_{STN}$ and $\hat{g}_{OLP}$ once more, we simply omit these enterprises so that $n_2 = 54$ instead of $n_2 = 57$. A first result is that estimate $\hat{\rho}_{xy2}$ increases from 0.876 to 0.970. The latter is fairly high in spite of the fact that the coefficient of variation of the growth rates $g_i = (Y_i / X_i - 1)$ is $cv_{g2} = s_{g2} / \overline{g}_2 = 4.1$ which still indicates a rather high volatility among the growth rates in this example. Furthermore, in analogy with the previous example, we get $\hat{g}_{STN} = 0.074$ $(= 7.4\%)$ with $\text{var}(\hat{g}_{STN}) = 0.00251$ and $\hat{g}_{OLP} = 0.039$ $(= 3.9\%)$ with $\text{var}(\hat{g}_{OLP}) = 0.00039$. The corresponding 95%-confidence intervals in this slightly modified example are approximately $I^{95}_{STN} \approx (\text{-}2.4\%, 17.2\%)$ and $I^{95}_{OLP} \approx (0.1\%, 7.7\%)$. Compared to Example 2.1 the interval I^{95}_{OLP} decreased relatively stronger than I^{95}_{STN}.

In addition, Example 2.2 may serve as a warning to be cautious when using sample means as $\overline{y}_{23}$ and $\overline{x}_{12}$ for estimating growth rates because these estimates may lead to unnecessarily large confidence interval around a suboptimal estimate. In the next section we look more closely at the question of what kind of circumstances may lead to a large interval I^{95}_{STN}.

3 Reasons for a large interval I^{95}_{STN}

In order to get more insight into the difference between $\text{var}(\hat{g}_{OLP})$ and $\text{var}(\hat{g}_{STN})$, we assume $n_{12} = n_{23} = n$ and G, $S_{xy} > 0$; hence, $\lambda = \mu = n_2 / n$. Then subtracting (2.4) from (2.2) yields

$$\begin{aligned}
\text{var}(\hat{g}_{STN}) - \text{var}(\hat{g}_{OLP}) &\approx \frac{1}{\overline{X}^2}\left\{2G\left(\frac{1}{n_2} - \frac{\lambda}{n}\right)S_{xy} - \left(\frac{1}{n_2} - \frac{1}{n}\right)\left(S_y^2 + G^2 S_x^2\right)\right\} \\
&= \frac{1}{\lambda n \overline{X}^2}\left\{2G(1-\lambda^2)S_{xy} - (1-\lambda)\left(S_y^2 + G^2 S_x^2\right)\right\} \\
&= \frac{1-\lambda}{\lambda n \overline{X}^2}\left(2G\lambda S_{xy} - S^2_{y-Gx}\right). \qquad (3.1)
\end{aligned}$$

In other words, $\text{var}(\hat{g}_{OLP})$ is smaller than $\text{var}(\hat{g}_{STN})$ when $\lambda > S^2_{y-Gx} / 2GS_{xy}$ provided $S_{xy} > 0$. Assuming $S_y^2 = S_x^2$, Qualité and Tillé (2008) derive a similar result for the parameter of *absolute* change when $\lambda > (1 - \rho_{xy}) / \rho_{xy}$. An anonymous referee pointed out that $\lambda < (1 - \rho_{xy}) / \rho_{xy}$ is a sufficient condition for $\text{var}(\hat{g}_{OLP}) > \text{var}(\hat{g}_{STN})$ because (3.1) can be rewritten as

$$\frac{(1-\lambda)GS_xS_y}{\lambda n \overline{X}^2}\left(2\lambda\rho_{xy} + 2\rho_{xy} - \frac{S_y^2 + G^2 S_x^2}{GS_xS_y}\right) \le \frac{(1-\lambda)GS_xS_y}{\lambda n \overline{X}^2}\left(2\lambda\rho_{xy} + 2\rho_{xy} - 2\right) < 0,$$

provided that $\lambda < (1 - \rho_{xy}) / \rho_{xy}$.

If N is sufficiently large, a weaker condition can be derived under some standard model assumptions. Suppose that the data satisfy the model $Y_i = BX_i + u_i$ with $E(u_i) = 0$, $E(u_i^2) = \sigma^2 X_i^{\delta}$ and $E(u_i u_j) = 0$ $(i \neq j)$; recall X_i is not random in this context. Under this model, we make the (weak) assumptions (i) $G = S_{yx} / S_x^2$ and (ii) $S^2_{y-Gx} = S_y^2 (1 - \rho_{xy}^2)$. To justify these assumptions, recall from regression theory that

$\hat{B} = S_{yx} / S_x^2$ can be seen as the unbiased, consistent estimator for B from an ordinary least squares (OLS) regression of Y_i on X_i and a *constant* $(i = 1,...,N)$. Furthermore, the corresponding OLS estimator $\left(\overline{Y} - \hat{B}\overline{X}\right)$ for the *constant* has zero expectation under the above model while its variance is of order $1/N$. Hence, $0 = \text{plim}\left(\overline{Y} - \hat{B}\overline{X}\right) = \text{plim}\left\{\overline{X}\left(G - \hat{B}\right)\right\}$ as $N \to \infty$ and provided $\overline{X} > c > 0$ for all N, we get the somewhat counterintuitive result $\text{plim}\left(G - \hat{B}\right) = 0$. In fact, it can be shown that

$$G = \overline{Y} / \overline{X} = \hat{B}\left[1 + O_p\left(1/\sqrt{N}\right)\right] = \left(S_{yx} / S_x^2\right)\left[1 + O_p\left(1/\sqrt{N}\right)\right]$$

as $N \to \infty$. This justifies assumption (i); for further details, see the end of this section. Furthermore, $S_y^2\left(1 - \rho_{xy}^2\right)$ can be seen as the (unexplained) variance of the residuals from the OLS regression. However, under the above model assumptions, these residuals are asymptotically equal to $Y_i - GX_i$ from which the *approximate* validity of (ii) follows. In addition, noting that $S_y^2\rho_{xy}^2$ is the so-called *explained* variance of the above OLS regression, it follows from assumption (i) that $S_y^2\rho_{xy}^2 = \hat{B}^2 S_x^2 \approx G^2 S_x^2$. Combining this with assumptions (i) and (ii), we can rewrite (3.1) as

$$\begin{aligned} \text{var}\left(\hat{g}_{STN}\right) - \text{var}\left(\hat{g}_{OLP}\right) &\approx \frac{1-\lambda}{\lambda n \overline{X}^2}\left\{2G^2\lambda S_x^2 - \left(1-\rho_{xy}^2\right)S_y^2\right\} \\ &\approx \frac{(1-\lambda)S_y^2}{\lambda n \overline{X}^2}\left(2\lambda\rho_{xy}^2 - 1 + \rho_{xy}^2\right) \\ &= \frac{(1-\lambda)S_y^2}{\lambda n \overline{X}^2}\left\{\rho_{xy}^2\left(1+2\lambda\right) - 1\right\}. \end{aligned} \tag{3.2}$$

Hence, $\text{var}\left(\hat{g}_{OLP}\right)$ is larger than $\text{var}\left(\hat{g}_{STN}\right)$ when

$$\lambda < \left(1-\rho_{xy}^2\right)/2\rho_{xy}^2 \quad \left[> \left(1-\rho_{xy}\right)/\rho_{xy}\right]. \tag{3.3}$$

Thus for say $\rho_{xy} = 0.9$, $\text{var}\left(\hat{g}_{OLP}\right)$ is under the above model for sufficiently large N larger than $\text{var}\left(\hat{g}_{STN}\right)$ when $\lambda < 0.117$, and for say $\rho_{xy} = 0.75$ when $\lambda < 0.389$. In addition, applying (3.2) to the data in Example 2.1 with $\lambda \approx 57/73 = 0.78$ and $\rho_{xy} = 0.876$ yields as approximation for the difference between both variances 0.0017 which is not very different from the actual difference of 0.0016 (=0.00324-0.00166) in the example. For Example 2.2, taking $\lambda = 54/70 = 0.77$ and $\rho_{xy} = 0.970$, applying (3.2) yields 0.00226 instead of 0.00212 (=0.00251-0.00039) in the example.

Under the above assumptions, it can also be shown that the ratio, say Q, of $\text{var}\left(\hat{g}_{OLP}\right)$ and $\text{var}\left(\hat{g}_{STN}\right)$ can be approximated by

$$Q = \frac{\text{var}\left(\hat{g}_{OLP}\right)}{\text{var}\left(\hat{g}_{STN}\right)} \approx \left(\lambda^{-1} - f\right)\left(1 - f + 2\left(1-\lambda\right)\frac{\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}, \tag{3.4}$$

irrespective of the values of S_y^2 and S_x^2; f stands for n/N. For a proof of (3.4), see Appendix A.1. From (3.4) it can be seen that Q and $\text{var}\left(\hat{g}_{OLP}\right)$ tend to zero as ρ_{xy}^2 tends to unity, provided N is sufficiently large and $\lambda < 1$.

It should be noted that in practice the correlations ρ_{xy} often are rather high by the very nature of the data (Y_i, X_i). That is, a large (small) enterprise in period $(t-12)$ is in most cases still large (small) after 12 months; Knottnerus and Van Delden (2012, page 47) found for various strata an overall mean correlation of 0.90 and a variance of 0.0074. So it appears that $\text{var}(\hat{g}_{STN})$ is more affected by a decrease of λ than $\text{var}(\hat{g}_{OLP})$ unless λ is extremely low because (i) $\text{var}(\hat{g}_{OLP}) = \text{var}(\hat{g}_{STN})$ when $\lambda = 1$ and (ii) Q is large when ρ_{xy}^2 is large. For example, when $\rho_{xy} = 0.9$ and $f = 0.1$ a decrease of λ from 0.9 to 0.5 leads to a decrease of Q from 0.58 to 0.37; recall $Q = 1$ when $\lambda = 1$. This emphasizes once more the importance of avoiding panel attrition when using estimator $\hat{g}_{STN}$ while N is large.

A natural question that remains to be answered is when is N sufficiently large. To answer this question, consider the difference $\Delta \equiv \hat{B} - G$ and its variance, say σ_Δ^2. The difference Δ can be written as

$$\begin{aligned}\Delta &= \frac{S_{xy}}{S_x^2} - \frac{\overline{Y}}{\overline{X}} = \frac{1}{N-1}\sum_{i\in U}\frac{X_i - \overline{X}}{S_x^2}Y_i - \frac{1}{N}\sum_{i\in U}\frac{Y_i}{\overline{X}} \\ &\approx \frac{1}{N}\sum_{i\in U}\left(\frac{X_i - \overline{X}}{S_x^2} - \frac{1}{\overline{X}}\right)Y_i \\ &= \frac{1}{N}\sum_{i\in U} M_i U_i \quad \left(M_i = \frac{X_i - \overline{X}}{S_x^2} - \frac{1}{\overline{X}}\right).\end{aligned}$$

In the second line we assumed $N >> 1$ and in the last line we used the model assumption $Y_i = BX_i + U_i$. Next, assuming $\text{var}(U_i) = \sigma^2 X_i^\delta$, we get

$$\sigma_\Delta^2 \equiv \text{var}\left(\hat{B} - G\right) = \frac{\sigma^2}{N^2}\sum_{i\in U} M_i^2 X_i^\delta.$$

This variance can be estimated by

$$\hat{\sigma}_\Delta^2 = \frac{\hat{\sigma}^2}{Nn_2}\sum_{i\in s_2}\hat{m}_i^2 X_i^{\hat{\delta}},$$

where

$$\hat{m}_i = \frac{X_i - \overline{x}_2}{s_{x2}^2} - \frac{1}{\overline{x}_2}, \qquad \hat{\sigma}^2 = \frac{1}{n_2 - 1}\sum_{i\in s_2}\left(Y_i - \frac{\overline{y}_2}{\overline{x}_2}X_i\right)^2 \Big/ X_i^{\hat{\delta}}$$

and $\hat{\delta}$ is an estimate from the OLS regression

$$\ln\left(Y_i - \frac{\overline{y}_2}{\overline{x}_2}X_i\right)^2 = \alpha + \delta \ln X_i + w_i \qquad (i = 1,...,n_2);$$

units with $Y_i = \overline{y}_2 X_i / \overline{x}_2$ are omitted. Based on $\hat{\sigma}_\Delta^2$, one may call N sufficiently large if the outcome of (3.1) will not severely be affected by replacing G by $G + \hat{\sigma}_\Delta$. In addition, it should be borne in mind that

relationships for very large N are probably still a reasonably appropriate indication for what may occur when N is not very large.

4 Composite estimator for the growth rate

Examining a composite estimator (COM) of the form

$$\hat{g}_{COM} = k\hat{g}_{STN} + (1-k)\hat{g}_{OLP}, \tag{4.1}$$

it follows from minimizing $\text{var}(\hat{g}_{COM})$ with respect to k that

$$k = \frac{\text{var}(\hat{g}_{OLP}) - \text{cov}(\hat{g}_{OLP}, \hat{g}_{STN})}{\text{var}(\hat{g}_{OLP}) + \text{var}(\hat{g}_{STN}) - 2\,\text{cov}(\hat{g}_{OLP}, \hat{g}_{STN})}; \tag{4.2}$$

see also Särndal et al. (1992, page 372). Note that, by construction, $\text{var}(\hat{g}_{COM})$ can not exceed $\min\{\text{var}(\hat{g}_{STN}), \text{var}(\hat{g}_{OLP})\}$.

Using the linearized forms of the estimators $\hat{g}_{OLP}$ and $\hat{g}_{STN}$, we get for their covariance

$$\begin{aligned}\text{cov}(\hat{g}_{OLP}, \hat{g}_{STN}) &\approx \text{cov}\left(\frac{\bar{y}_2 - G\bar{x}_2}{\bar{X}}, \frac{\bar{y}_{23} - G\bar{x}_{12}}{\bar{X}}\right) \\ &= \frac{1}{\bar{X}^2}\left\{\text{cov}(\bar{y}_2, \bar{y}_{23}) - G\,\text{cov}(\bar{y}_2, \bar{x}_{12}) - G\,\text{cov}(\bar{x}_2, \bar{y}_{23}) + G^2\,\text{cov}(\bar{x}_2, \bar{x}_{12})\right\}.\end{aligned}$$

Now using some results from Knottnerus (2003, page 377)

$$\text{cov}(\bar{y}_2, \bar{y}_{23}) = \text{var}(\bar{y}_{23}) \quad \left[= \left(\frac{1}{n_{23}} - \frac{1}{N}\right) S_y^2\right]$$

$$\text{cov}(\bar{x}_2, \bar{y}_{23}) = \text{cov}(\bar{x}_{23}, \bar{y}_{23}) \quad \left[= \left(\frac{1}{n_{23}} - \frac{1}{N}\right) S_{xy}\right],$$

we obtain

$$\text{cov}(\hat{g}_{OLP}, \hat{g}_{STN}) \approx \frac{1}{\bar{X}^2}\left\{\left(\frac{1}{n_{23}} - \frac{1}{N}\right)\left(S_y^2 - GS_{yx}\right) + \left(\frac{1}{n_{12}} - \frac{1}{N}\right)\left(G^2 S_x^2 - GS_{yx}\right)\right\}. \tag{4.3}$$

In practice k can be estimated by replacing all (co)variances in (4.2) by their sample estimates, yielding

$$\hat{k} = \frac{\hat{\text{var}}(\hat{g}_{OLP}) - \hat{\text{cov}}(\hat{g}_{OLP}, \hat{g}_{STN})}{\hat{\text{var}}(\hat{g}_{OLP}) + \hat{\text{var}}(\hat{g}_{STN}) - 2\,\hat{\text{cov}}(\hat{g}_{OLP}, \hat{g}_{STN})} \tag{4.4}$$

To illustrate this approach, consider the following example.

Example 4.1. The data are the same as for Example 2.1. Applying formulas (2.1) - (2.4) and (4.3) to these data yields

$$\hat{g}_{STN} = 0.082 \ (0.00254), \ \hat{g}_{OLP} = 0.050 \ (0.00134), \ \text{and}$$
$$\hat{\text{cov}}\left(\hat{g}_{STN}, \hat{g}_{OLP}\right) = 0.00097.$$

The variances are mentioned between parentheses. Substituting these estimates into (4.4) yields $\hat{k} = 0.191$ and subsequently, $\hat{g}_{COM} = 0.056$ (0.00127). For the ease of exposition, all (co)variances in (4.4) are estimated from overlap s_2, including the estimates of G and $\bar{X}$ in (2.2), (2.4) and (4.3). Furthermore, using these estimates, we found that $\text{var}\left(\hat{g}_{STN}\right) < \text{var}\left(\hat{g}_{OLP}\right)$ and $k > 0.5$ only if $n_2 \leq 12 \ \left(\lambda \leq 0.167\right)$.

For the sake of completeness, we also give an example for the composite estimator for the parameter of absolute change (i.e., $\bar{D} = \bar{Y} - \bar{X}$).

Example 4.2. We use the same data as in Example 2.1. As before all estimates for the (co)variances are based on s_2. Define D_i by $D_i = Y_i - X_i$. Then we have two estimators for the parameter of absolute change

$$\hat{\bar{D}}_{STN} = \bar{y}_{23} - \bar{x}_{12} = 7.35 \quad \text{and} \quad \hat{\bar{D}}_{OLP} = \bar{d}_2 = \bar{y}_2 - \bar{x}_2 = 4.89.$$

For the (co)variances of $\hat{\bar{D}}_{STN}$ and $\hat{\bar{D}}_{OLP}$ we get

$$\begin{aligned}
\hat{\text{var}}\left(\hat{\bar{D}}_{STN}\right) &= \hat{\text{var}}\left(\bar{y}_{23}\right) + \hat{\text{var}}\left(\bar{x}_{12}\right) - 2\,\hat{\text{cov}}\left(\bar{y}_{23}, \bar{x}_{12}\right) \\
&= \left(\frac{1}{n_{23}} - \frac{1}{N}\right) s_{y2}^2 + \left(\frac{1}{n_{12}} - \frac{1}{N}\right) s_{x2}^2 - 2\left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right) s_{xy2} = 23.58 \\
\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) &= \left(\frac{1}{n_2} - \frac{1}{N}\right) s_{y-x,2}^2 = 13.11
\end{aligned}$$

$$\begin{aligned}
\hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right) &= \hat{\text{cov}}\left(\bar{y}_{23} - \bar{x}_{12}, \bar{y}_2 - \bar{x}_2\right) \\
&= \left(\frac{1}{n_{23}} - \frac{1}{N}\right)\left(s_{y2}^2 - s_{xy2}\right) - \left(\frac{1}{n_{12}} - \frac{1}{N}\right)\left(s_{xy2} - s_{x2}^2\right) = 9.46.
\end{aligned}$$

In analogy with (4.4) we now obtain

$$\hat{k} = \frac{\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) - \hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)}{\hat{\text{var}}\left(\hat{\bar{D}}_{OLP}\right) + \hat{\text{var}}\left(\hat{\bar{D}}_{STN}\right) - 2\,\hat{\text{cov}}\left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)} = 0.206$$

and consequently, $\hat{\bar{D}}_{COM} = 5.40\,(12.37)$.

Note that $\hat{g}_{COM}$ can be rewritten as

$$\begin{aligned}
\hat{g}_{COM} &= \hat{g}_{OLP} + \hat{k}\left(\hat{g}_{STN} - \hat{g}_{OLP}\right) \\
&\approx \hat{g}_{OLP} + k\left(\hat{g}_{STN} - \hat{g}_{OLP}\right),
\end{aligned}$$

where we used a first-order Taylor series approximation of $\hat{g}_{COM}$. Therefore, the random character of estimator $\hat{k}$ can be neglected for estimating $\text{var}\left(\hat{g}_{COM}\right)$. The error thus introduced is of order $1/n_2$ as

$n_2 \to \infty$ and $\hat{g}_{COM}$ is asymptotically unbiased. Recall that the standard procedure for estimating the variance of the ratio estimator or the regression estimator is based on a first-order Taylor series approximation as well.

In addition, under the same assumptions as (3.4), it can be shown that for sufficiently large N,

$$k = \left(1 + \frac{2\lambda\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}; \tag{4.5}$$

for a proof of (4.5), see Appendix A.1. From (4.5) it can be seen that k is decreasing in λ. So we have the somewhat counterintuitive result that k is decreasing in λ whereas according to (3.1), ratio Q in (3.4) is a convex function of λ; recall that $\text{var}\left(\hat{g}_{STN}\right) = \text{var}\left(\hat{g}_{OLP}\right)$ and, consequently, $Q = 1$ for $\lambda = 1$ and $\lambda = S_{y-Gx}^2 \big/ 2GS_{xy}$.

5 Aligned composite estimators for growth rates and totals

So far we only looked at growth rates because in practice the estimate $\hat{X}_{SRS}$ for the turnover of 12 months ago can be considered more or less as fixed (i.e., can not be changed anymore). When X refers to the total turnover in month $(t-1)$, it is likely that the figures for the preceding month can still be improved and modified. In such a situation the initial estimate $\hat{X}_{SRS}$ might be revised as well.

Before examining a multivariate composite estimator for growth rates and totals, we first look at a multivariate composite estimator for the parameter of absolute change and the corresponding population means or totals; also see Example 4.2. Define the initial vector estimator $\hat{\theta}_0$ by $\hat{\theta}_0 = \left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}, \bar{x}_{12}\right)'$. Denote the underlying parameter vector to be estimated by $\theta = \left(\theta_1, \theta_2, \theta_3\right)'$. Let V_0 denote the covariance matrix of $\hat{\theta}_0$. In terms of θ the problem is now to find an aligned composite estimator $\hat{\theta}_{AC}$ with elements satisfying the prior restriction $\theta_1 - \theta_2 + \theta_3 = 0$ or, equivalently, $\bar{D} - \bar{Y} + \bar{X} = 0$ or $D - Y + X = 0$. Although there is one restriction in this situation, we treat in this section the somewhat more general case with m restrictions $\left(1 \le m \le 3\right)$. When the prior restrictions are of the linear form $c - R\theta = 0$ where R is a $m \times 3$ matrix of rank m $(m \le 3)$, the optimal unbiased composite estimator for θ is equal to the general restriction (GR) estimator

$$\hat{\theta}_{GR} = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right) \tag{5.1}$$

$$K = V_0 R'\left(RV_0 R'\right)^{-1}$$

$$V_{GR} \equiv \text{cov}\left(\hat{\theta}_{GR}\right) = \left(I_3 - KR\right)V_0, \tag{5.2}$$

where I_3 stands for the 3×3 identity matrix. The estimator $\hat{\theta}_{GR}$ is optimal in the sense that when $\hat{\theta}_0$ follows a multivariate normal distribution $N\left(\theta, V_0\right)$, the likelihood of $\hat{\theta}_0$ attains its maximum, under the constraint $c - R\theta = 0$, for $\theta_{\max} = \hat{\theta}_{GR}$. Moreover, given the form $\hat{\theta}_K = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right)$, it can be shown

that minimizing $\mathrm{tr}\left\{\mathrm{cov}\left(\hat{\theta}_K\right)\right\}$ with respect to the $3\times m$ matrix K leads to (5.2). Recall that this means that for any other matrix K the corresponding covariance matrix $\mathrm{cov}\left(\hat{\theta}_K\right)$ exceeds V_{GR} by a positive semidefinite matrix; see Magnus and Neudecker (1988, pages 255-256). For further details on the GR estimator, see Knottnerus (2003, pages 328-332). To illustrate how (5.1) and (5.2) can be used for obtaining an aligned composite (AC) estimator $\hat{\theta}_{AC}$, consider the following example dealing with the estimation of two population means and their difference.

Example 5.1. We use the same data as in Examples 2.1 and 4.2. The initial vector $\hat{\theta}_0 = \left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}, \bar{x}_{12}\right)'$ is given by $\left(4.89, 97.19, 89.84\right)'$. These estimates do not satisfy the restriction $\theta_1 - \theta_2 + \theta_3 = 0;$ note that $R = (1, -1, 1)$ and $c = 0.$ Most elements of V_0 have already been discussed. Similar to Example 4.2, for element $\mathrm{cov}\left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}\right)$ we get

$$\begin{aligned}\mathrm{cov}\left(\hat{\bar{D}}_{OLP}, \bar{y}_{23}\right) &= \mathrm{cov}\left(\bar{y}_2 - \bar{x}_2, \bar{y}_{23}\right) \\ &= \mathrm{var}\left(\bar{y}_{23}\right) - \mathrm{cov}\left(\bar{x}_{23}, \bar{y}_{23}\right).\end{aligned} \tag{5.3}$$

Each term in (5.3) can be estimated from s_2 as described before. The other covariances in V_0 have a similar form and can be estimated in the same manner. The variance estimates for $\hat{\bar{D}}_{OLP}$, $\bar{y}_{23}$ and $\bar{x}_{12}$ are 13.12, 38.79 and 22.92, respectively. Next, applying (5.1) and (5.2) with K replaced by $\hat{K} = \hat{V}_0 R'\left(R\hat{V}_0 R'\right)^{-1}$, we obtain the following aligned composite AC estimates

$$\hat{\bar{D}}_{AC} = 5.40\ (12.37), \quad \hat{\bar{Y}}_{AC} = 96.28\ (36.32), \quad \text{and} \quad \hat{\bar{X}}_{AC} = 90.88\ (19.75).$$

Between parentheses the variances are mentioned.

Now three remarks are in order. Firstly, $\hat{\bar{D}}_{COM}$ discussed in the preceding section can also be derived from (5.1) and (5.2) by choosing $\hat{\theta}_0 = \left(\hat{\bar{D}}_{STN}, \hat{\bar{D}}_{OLP}\right)'$ with prior restriction $\theta_1 - \theta_2 = 0.$ Secondly, by construction, the estimator $\hat{\bar{D}}_{AC}$ is equal to estimator $\hat{\bar{D}}_{COM}$ and, consequently, they have the same variance. Thirdly, were K known, then the AC estimator would be unbiased. But because K is to be replaced by $\hat{K},$ the AC estimator $\hat{\theta}_{AC}$ is only asymptotically unbiased. The same remark applies to the estimator $\left(I_3 - \hat{K}R\right)\hat{V}_0$ of $\mathrm{cov}(\hat{\theta}_{AC})$. Similar to $\hat{\theta}_{COM}$ described in the preceding section, the bias of $\hat{\theta}_{AC}$ is of order $O(1/n_2);$ for the relationship between $\hat{\theta}_{AC}$ and the regression estimator, see Appendix A.3.

In case of m nonlinear restrictions, say $c - R(\theta) = 0,$ a first-order Taylor series approximation around $\theta = \hat{\theta}_0$ yields $c - R\left(\hat{\theta}_0\right) - D_R\left(\hat{\theta}_0\right)\left(\theta - \hat{\theta}_0\right) = 0$ or, equivalently,

$$c\left(\hat{\theta}_0\right) - D_R\left(\hat{\theta}_0\right)\theta = 0, \quad \text{where} \quad c\left(\hat{\theta}_0\right) = c - R\left(\hat{\theta}_0\right) + D_R\left(\hat{\theta}_0\right)\hat{\theta}_0. \tag{5.4}$$

$D_R(\theta)$ stands for the $m \times 3$ matrix of partial derivatives of $R(\theta)$ $\left(\text{i.e., } D_R(\theta) = \partial R(\theta)/\partial \theta'\right)$. Subsequently, an iterative procedure can be carried out by repeatedly applying (5.1) and (5.2) to the updated linearized versions of the nonlinear restrictions $c - R(\theta) = 0$. This yields

$$\left.\begin{aligned} \hat{\theta}_h &= \hat{\theta}_0 + K_h \hat{e}_h; \\ \hat{e}_h &= c_h - D_h \hat{\theta}_0; \\ K_h &= V_0 D_h' \left(D_h V_0 D_h'\right)^{-1}; \\ \text{cov}(\hat{\theta}_h) &= \left(I_3 - K_h D_h\right) V_0; \\ D_h &= D_R\left(\hat{\theta}_{h-1}\right); \\ c_h &= c - R\left(\hat{\theta}_{h-1}\right) + D_h \hat{\theta}_{h-1} \qquad (h = 1,2,\ldots). \end{aligned}\right\} \tag{5.5}$$

For further details, see Appendix A.2 and Knottnerus (2003, pages 351-354). Note that the first equation can be seen as an update of $\hat{\theta}_0$ rather than of $\hat{\theta}_{h-1}$. This is an important difference with the celebrated Kalman equations; see Kalman (1960). In the present context, the vectors $\hat{\theta}_{h-1}$ are only used in a numerical procedure for finding new (better) Taylor series approximations of the nonlinear restrictions $c - R(\theta) = 0$ around $\theta = \hat{\theta}_{h-1}$ $(h = 1,2,\ldots)$ until convergence is reached. Furthermore, note that $\hat{e}_h$ can be seen as a $m-$vector of restriction errors when substituting $\theta = \hat{\theta}_0$ into the linearized restrictions around $\theta = \hat{\theta}_{h-1}$. To illustrate the use of the Kalman-like equations in (5.5) for deriving aligned composite estimators for growth rates and totals, consider the following example.

Example 5.2. We use the same data as in Example 4.1. The initial vector $\hat{\theta}_0$ is now defined by $\hat{\theta}_0 = \left(\hat{G}_{OLP}, \overline{y}_{23}, \overline{x}_{12}\right)'$ and is given by $(1.050, 97.191, 89.840)'$. These estimates do not satisfy the (nonlinear) prior restriction $\theta_2 - \theta_1\theta_3 = 0$ $(m = 1)$. All elements of V_0 and their estimation have already been discussed. For the $(h+1)$-th recursion $R\left(\hat{\theta}_h\right)$ and the 1×3 matrix D_{h+1} are given by

$$R\left(\hat{\theta}_h\right) = \left(\hat{\theta}_{h2} - \hat{\theta}_{h1}\hat{\theta}_{h3}\right)$$
$$D_{h+1} = \left(-\hat{\theta}_{h3} \quad 1 \quad -\hat{\theta}_{h1}\right),$$

respectively; $\hat{\theta}_{hk}$ is the k-th element of vector $\hat{\theta}_h$ $(1 \le k \le 3)$. Recall V_0 and $\hat{V}_0$ remain unchanged for all recursions. The first recursion from (5.5) yields

$$\hat{\theta}_1 = (1.0544,\ 95.945,\ 91.000)'.$$

The (nonlinear) restriction is almost satisfied, that is, $R\left(\hat{\theta}_1\right) = -0.005$. The second recursion yields the following aligned composite (AC) estimates

$$\hat{G}_{AC} = 1.0544\ (0.00130), \quad \hat{\overline{Y}}_{AC} = 95.947\ (35.55), \quad \text{and} \quad \hat{\overline{X}}_{AC} = 90.998\ (19.85).$$

Between parentheses the variances are mentioned. The (absolute) error of the second restriction further decreased, that is, $R\left(\hat{\theta}_2\right) = -0.001$ and we stopped the recursions. Due to the nonlinearity of the restriction, the estimates of $\hat{G}_{AC}$ and its variance are slightly different from those of $\hat{G}_{COM}$ and its variance in Example 4.1.

It is noteworthy that in Example 5.2 $\hat{G}_{AC}$ is not much different of $\hat{G}_{OLP}$ (=1.050). A related method for estimating totals is the so-called matched pair (MP) method; see Smith et al. (2003, page 269-271). The original MP method is purely based on $\hat{G}_{OLP}$ (in our notation) between months t and $t-1$ and used by ONS for estimating the monthly retail sales index. In a simulation study the authors found that the MP method gives a good performance for the short-term growth rates but for terms of more than 15 months the performance was worsening with respect to the bias. The bias could be corrected by benchmarking to growth rates on a regular basis. Another drawback of the MP method seems to be that a formula for the variance of the MP estimator is (still) lacking. In the next section we describe an extension of the AC estimator for incorporating auxiliary information into the AC estimation procedure.

6 Extensions

In this section we briefly discuss a number of extensions of the AC estimator described in the preceding section. Firstly, we pay attention to the situation whereby regression estimators, say $\hat{\bar{Y}}_{REG,k}$ and $\hat{\bar{X}}_{REG,k}$, are used instead of SRS estimators $\left(k = 2,\ 12 \text{ and } 23\right)$. To avoid a notational burden, we look at the situation with one explanatory variable, say z; a generalization for more auxiliaries is straightforward. Furthermore, for simplicity's sake, we assume that the estimated regression coefficients, denoted by b_{yz2} and b_{xz2}, stem from s_2. In order to derive the aligned composite estimators in this situation, we only need to evaluate (co)variance terms of the form $\operatorname{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{X}}_{REG,l}\right)$ in the different formulas $\left(k,\ l = 2,\ 12 \text{ and } 23\right)$. This evaluation can be done as follows. Replace the Y_i and X_i in the formulas by the corresponding (estimated) residuals from a regression on Z_i and a *constant*. That is,

$$\operatorname{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{X}}_{REG,l}\right) = \operatorname{cov}\left(\bar{y}_k^*, \bar{x}_l^*\right), \tag{6.1}$$

where the (estimated) residual variables Y_i^* and X_i^* are defined by

$$Y_i^* = Y_i - \bar{y}_k - b_{yz2}\left(Z_i - \bar{z}_k\right) = Y_i - b_{yz2}Z_i + const.$$
$$X_i^* = X_i - \bar{x}_l - b_{xz2}\left(Z_i - \bar{z}_l\right) = X_i - b_{xz2}Z_i + const.$$

The term $\operatorname{cov}\left(\bar{y}_k^*, \bar{x}_l^*\right)$ on the right-hand side of (6.1) can be calculated in the same manner as $\operatorname{cov}\left(\bar{y}_k, \bar{x}_l\right)$, discussed in preceding sections; see also formula (A.8) in Appendix A.3 and recall $\operatorname{var}\left(\hat{\bar{Y}}_{REG,k}\right) = \operatorname{cov}\left(\hat{\bar{Y}}_{REG,k}, \hat{\bar{Y}}_{REG,k}\right)$. In addition, the same approach can be applied when use is made of ratio estimators such as $\hat{\bar{Y}}_{R,k} = \bar{y}_k \bar{Z} / \bar{z}_k$ and $\hat{\bar{X}}_{R,l} = \bar{x}_l \bar{Z} / \bar{z}_l$. That is, the residual variables Y_i^* and X_i^* are now to be read as

$$Y_i^* = Y_i - \frac{\overline{y}_2}{\overline{z}_2} Z_i \quad \text{and} \quad X_i^* = X_i - \frac{\overline{x}_2}{\overline{z}_2} Z_i.$$

An alternative option for taking an auxiliary variable into account is to extend both the parameter vector θ and the set of prior restrictions. For instance, in Example 5.1 the parameter θ was implicitly defined by $\theta = \left(\overline{D}, \overline{Y}, \overline{X}\right)'$. When the variable z is observed in samples 12 and 23, the new, extended $\hat{\theta}_0$ is given by

$$\hat{\theta}_0 = \left(\hat{\overline{D}}_{OLP}, \overline{y}_{23}, \overline{x}_{12}, \overline{z}_{23}, \overline{z}_{12}, \overline{z}_2\right)'$$

and the extended set of prior restrictions is

$$\begin{aligned} &\theta_2 - \theta_1 - \theta_3 = 0; \\ &\theta_4 - \theta_5 = 0; \\ &\theta_4 - \theta_6 = 0; \\ &\theta_4 = \overline{Z}. \end{aligned}$$

Hence, the new c is $c = \left(0, 0, 0, \overline{Z}\right)'$. In this way the efficiency of $\hat{\theta}_0$ can be further improved.

Secondly, another extension regards births and deaths. With respect to deaths, the population in period $t-12$ can be divided into two (post)strata: one consisting of the deaths in period t and one consisting of the enterprises existing in periods $t-12$ and t. Using such a poststratification still leads to an asymptotically unbiased estimator for the population mean at period t, provided there are no births. In order to take births into account, one should draw an appropriate sample from this substratum of births especially when the number of births is substantial, and when there are no realistic assumptions with respect to the total turnover in this substratum in month t.

Finally, we examine the situation whereby a combination of quarterly and semesterly data is to be analysed. Suppose that in quarters 2, 4 and 6 semesterly samples are drawn which need not be the same as the quarterly samples in those quarters. In order to explain the AC estimator in this situation, consider six consecutive quarterly SRS estimates for the quarterly means of the turnover, say $\overline{y}_1, \overline{y}_2, \overline{y}_3, \overline{y}_4, \overline{y}_5, \overline{y}_6$, and three semesterly SRS estimates for the semesterly means of turnover, say $\overline{x}_2$, $\overline{x}_4$ and $\overline{x}_6$; note that the subscript refers to the quarter of observation and *not* to a sample set as before. Furthermore, suppose that the following growth ratios are to be estimated: $G_{62} = Y_6/Y_2$, $H_{62} = X_6/X_2$ and $H_{64} = X_6/X_4$ as well as the corresponding quarterly and semesterly totals. In order to obtain a consistent set of estimators for totals (means) and growth rates, define in analogy with the approach in Section 5

$$\hat{\theta}_0 = \left(\hat{G}_{62,OLP}, \hat{H}_{62,OLP}, \hat{H}_{64,OLP}, \overline{y}_1, \overline{y}_2, \overline{y}_3, \overline{y}_4, \overline{y}_5, \overline{y}_6, \overline{x}_2, \overline{x}_4, \overline{x}_6\right)'.$$

The corresponding set of restrictions is

$$\begin{aligned} \theta_9 - \theta_1\theta_5 &= \overline{Y}_6 - G_{62}\overline{Y}_2 = 0 \\ \theta_{12} - \theta_2\theta_{10} &= \overline{X}_6 - H_{62}\overline{X}_2 = 0 \\ \theta_{12} - \theta_3\theta_{11} &= \overline{X}_6 - H_{64}\overline{X}_4 = 0 \\ \theta_4 + \theta_5 - \theta_{10} &= \overline{Y}_1 + \overline{Y}_2 - \overline{X}_2 = 0 \\ \theta_6 + \theta_7 - \theta_{11} &= \overline{Y}_3 + \overline{Y}_4 - \overline{X}_4 = 0 \\ \theta_8 + \theta_9 - \theta_{12} &= \overline{Y}_5 + \overline{Y}_6 - \overline{X}_6 = 0. \end{aligned}$$

The matrix V_0 can be estimated in a similar manner as described in Sections 2 and 4.

7 Conclusions and discussion

This section summarizes a number of conclusions and issues for further research.

When totals of turnover are estimated from a panel in months t and $t-12$, two estimators $\hat{g}_{STN}$ and $\hat{g}_{OLP}$ for the growth rate between these months can be distinguished.

When using $\hat{g}_{STN}$, one should be aware that in practice, $\text{var}(\hat{g}_{OLP})$ might be much smaller than $\text{var}(\hat{g}_{STN})$ especially when the turnover in month $t-12$ and the turnover in month t are highly correlated and the overlap ratios λ and μ are not too small.

The efficiency of $\hat{g}_{STN}$ and $\hat{g}_{OLP}$ can be improved by the composite estimator $\hat{g}_{COM}$ described in Section 4.

Using least squares techniques, an aligned composite vector-estimator $\left(\hat{g}_{AC}, \hat{Y}_{AC}, \hat{X}_{AC}\right)'$ can be derived that obeys the nonlinear restriction for totals and growth rates: $\hat{Y}_{AC} = \left(1+\hat{g}_{AC}\right)\hat{X}_{AC}$.

The AC estimator subject to *linear* restrictions can be extended in several ways: (i) for nonlinear restrictions, (ii) for different data sets such as monthly, quarterly and yearly data, (iii) for births and deaths, (iv) for regression and ratio estimators, and (v) for additional auxiliary variables.

Similar to the regression estimator, the AC estimator is asymptotically unbiased. This remark also applies to the covariance-matrix estimator $\left(I_k - \hat{K}R\right)\hat{V}_0$.

There is not yet an unambiguous answer on the question of to what extent data from the past should be included in the vector estimate $\hat{\theta}_0$ each month. The answer depends upon: (i) the NSI's policy and rules with respect to revision of already published figures, (ii) the fact that from a theoretical viewpoint, the sequence of T monthly SRS estimates $\overline{y}_1, \overline{y}_2, ..., \overline{y}_T$ (included as component in $\hat{\theta}_0$) should be so long that the difference between the two AC estimators of $\overline{Y}_1$, say $\hat{\overline{Y}}_{1AC}^{T}$ and $\hat{\overline{Y}}_{1AC}^{T+1}$, is not substantial, and (iii) the size of the samples. That is, in analogy with the regression estimator or, equivalently, the calibration estimator, the sample sizes should be much larger than the number of (calibration) restrictions. For a simulation study on the variance of the regression estimator and the number of regressors, see Silva and Skinner (1997) and for a relationship between the regression estimator and the GR estimator, see Appendix A.3 and Knottnerus (2003).

In the specific case of estimating mutually aligned totals and changes, additional research is needed for finding: (i) the optimal and practical length for the monthly, quarterly, semesterly and yearly series of SRS estimates to be included in the initial vector $\hat{\theta}_0$ and (ii) a rule of thumb with respect to the number of restrictions compared to the sample sizes in order to find an AC estimator $\hat{\theta}_{AC}$ with an improved efficiency.

Acknowledgements

The views expressed in this paper are those of the author and do not necessarily reflect the policy of Statistics Netherlands. The author would like to thank Harm Jan Boonstra, Arnout van Delden, Sander Scholtus, the Associate Editor and two anonymous referees for their helpful comments and corrections.

Appendix

A.1 Proofs of (3.4) and (4.5)

The proof of (3.4) is as follows. For $n_{12} = n_{23} = n,$ formula (2.2) can be rewritten as

$$\text{var}\left(\hat{g}_{STN}\right) \approx \frac{1}{\bar{X}^2}\left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2 + 2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}\right\} \tag{A.1}$$

Dividing (2.4) by (A.1) yields

$$\begin{aligned} Q = \frac{\text{var}\left(\hat{g}_{OLP}\right)}{\text{var}\left(\hat{g}_{STN}\right)} &\approx \frac{\left(\frac{1}{\lambda n}-\frac{1}{N}\right)S_{y-Gx}^2}{\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2 + 2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}} \\ &= \frac{\left(\lambda^{-1}-f\right)S_{y-Gx}^2}{(1-f)S_{y-Gx}^2 + 2(1-\lambda)GS_{xy}} \\ &= \left(\lambda^{-1}-f\right)\left(1-f+2(1-\lambda)\frac{GS_{xy}}{S_{y-Gx}^2}\right)^{-1} \\ &\approx \left(\lambda^{-1}-f\right)\left(1-f+2(1-\lambda)\frac{\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}. \end{aligned} \tag{A.2}$$

In the last line we used that under the model assumptions mentioned in Section 3, $GS_{xy} \approx \hat{B}^2 S_x^2 = \rho_{xy}^2 S_y^2$ and $S_{y-Gx}^2 \approx \left(1-\rho_{xy}^2\right)S_y^2,$ provided that N is sufficiently large; see also the derivation of (3.2).

Next, under the same assumptions, (4.5) can be derived as follows. Since $n_{12} = n_{23} = n,$ the covariance in (4.3) can be rewritten as

$$\text{cov}\left(\hat{g}_{OLP}, \hat{g}_{STN}\right) \approx \frac{1}{\bar{X}^2}\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-Gx}^2. \tag{A.3}$$

Combining (2.4), (A.1) and (A.3), we can write k in (4.2) as

$$\begin{aligned} k &\approx \frac{\left(\frac{1}{\lambda n}-\frac{1}{n}\right)S_{y-Gx}^2}{\left(\frac{1}{\lambda n}-\frac{1}{n}\right)S_{y-Gx}^2 + 2\left(\frac{1}{n}-\frac{\lambda}{n}\right)GS_{xy}} \\ &= \left(1+\frac{2\lambda GS_{xy}}{S_{y-Gx}^2}\right)^{-1} \approx \left(1+\frac{2\lambda\rho_{xy}^2}{1-\rho_{xy}^2}\right)^{-1}. \end{aligned}$$

Similar to deriving (A.2), we used in the last line $GS_{xy}/S^2_{y-Gx} \approx \rho^2_{xy}/\left(1-\rho^2_{xy}\right)$.

A.2 Derivation of (5.5)

In case of m linear restrictions $c - R\theta = 0,$ matrix K can be found by minimizing

$$\min_K E\left[\left\{\theta-\hat{\theta}_0-K\left(c-R\hat{\theta}_0\right)\right\}'\left\{\theta-\hat{\theta}_0-K\left(c-R\hat{\theta}_0\right)\right\}\right];$$

see Knottnerus (2003, page 330). The solution of this least squares problem is given by

$$\begin{aligned} K &= E\left\{\left(\theta-\hat{\theta}_0\right)\left(c-R\hat{\theta}_0\right)'\right\}\left[\operatorname{cov}\left(c-R\hat{\theta}_0\right)\right]^{-1} \\ &= V_0R'\left(RV_0R'\right)^{-1}. \end{aligned} \tag{A.4}$$

In case of m nonlinear restrictions, the new minimand is

$$E\left[\left\{\theta-\hat{\theta}_0-K\left[c-R\left(\hat{\theta}_0\right)\right]\right\}'\left\{\theta-\hat{\theta}_0-K\left[c-R\left(\hat{\theta}_0\right)\right]\right\}\right].$$

Similarly to (A.4), it can be shown that this minimand attains its minimum for

$$K = E\left\{\left(\theta-\hat{\theta}_0\right)\left[c-R\left(\hat{\theta}_0\right)\right]'\right\}\left[\operatorname{cov}\left\{c-R\left(\hat{\theta}_0\right)\right\}\right]^{-1}. \tag{A.5}$$

Substituting Taylor's linearization $R\left(\hat{\theta}_0\right) \approx R(\theta) + D_R(\theta)\left(\hat{\theta}_0-\theta\right)$ into (A.5), we get the following approximation, say $K_1,$ for K

$$\begin{aligned} K_1 &\approx V_0D_R'(\theta)\left[D_R(\theta)V_0D_R'(\theta)\right]^{-1} \\ &\approx V_0D_R'\left(\hat{\theta}_0\right)\left[D_R\left(\hat{\theta}_0\right)V_0D_R'\left(\hat{\theta}_0\right)\right]^{-1}. \end{aligned} \tag{A.6}$$

Assuming that $\hat{\theta}_0 \sim N\left(\theta, V_0\right),$ the first approximation for the constrained maximum likelihood (ML) solution, say $\hat{\theta}^{(1)}_{ML},$ can be calculated in the standard manner by using the linearized restrictions

$$\hat{\theta}^{(1)}_{ML} = \hat{\theta}_0 + K_1\left\{c\left(\hat{\theta}_0\right) - D_R\left(\hat{\theta}_0\right)\hat{\theta}_0\right\}, \tag{A.7}$$

where $c\left(\hat{\theta}_0\right)$ is defined by (5.4). If $\hat{\theta}^{(1)}_{ML}$ does not satisfy the nonlinear restrictions $c - R(\theta) = 0,$ a better approximation of K might be obtained by replacing $\hat{\theta}_0$ in (A.6) by update $\hat{\theta}^{(1)}_{ML}$ resulting in a new matrix $K_2.$ In turn, in analogy with (A.7) K_2 leads to a better approximation or update of $\hat{\theta}_0,$ say $\hat{\theta}^{(2)}_{ML},$

$$\hat{\theta}^{(2)}_{ML} = \hat{\theta}_0 + K_2\left\{c\left(\hat{\theta}^{(1)}_{ML}\right) - D_R\left(\hat{\theta}^{(1)}_{ML}\right)\hat{\theta}_0\right\},$$

where we used Taylor's linearization of the nonlinear restrictions around $\theta = \hat{\theta}^{(1)}_{ML}.$ Repeating this procedure, we get the following recursions for $\hat{\theta}^{(h)}_{ML}$ or, for short, $\hat{\theta}_h$

$$\hat{\theta}_h = \hat{\theta}_0 + K_h\left\{c_h - D_h\hat{\theta}_0\right\}$$
$$K_h = V_0 D_h'\left[D_h V_0 D_h'\right]^{-1} \qquad (h = 1, 2, ...).$$

For definitions of c_h and D_h, see Section 5; in practice, V_0 should be replaced by its estimate $\hat{V}_0$. By construction, for each h we have

$$\begin{aligned} 0 &= c\left(\hat{\theta}_{ML}^{(h-1)}\right) - D_R\left(\hat{\theta}_{ML}^{(h-1)}\right)\hat{\theta}_{ML}^{(h)} \\ &= c - R\left(\hat{\theta}_{ML}^{(h-1)}\right) + D_R\left(\hat{\theta}_{ML}^{(h-1)}\right)\hat{\theta}_{ML}^{(h-1)} - D_R\left(\hat{\theta}_{ML}^{(h-1)}\right)\hat{\theta}_{ML}^{(h)}; \end{aligned}$$

see (5.4). Hence, when $\hat{\theta}_{ML}^{(h)}$ converges to the (constrained) maximum likelihood solution $\hat{\theta}_{ML}$, $c - R\left(\hat{\theta}_{ML}^{(h-1)}\right)$ converges to zero. Also, assuming K_h converges to say $\hat{K}_{ML}$, the corresponding covariance matrix of $\hat{\theta}_{ML}$, say V_{ML}, can be approximated by

$$V_{ML} \approx \left\{I_k - KD_R(\theta)\right\}V_0,$$

which for sufficiently large h can be estimated by $\hat{V}_{ML} = \left(I_k - K_h D_h\right)\hat{V}_0$; see also Cramer (1986, page 38).

A.3 Regression estimator as GR estimator

Suppose that Y_i and the auxiliary variable Z_i, with known population mean $\overline{Z}$, are observed in s_2. In order to apply the GR estimator to this situation, define

$$\hat{\theta}_0 = \begin{pmatrix} \overline{y}_2 \\ \overline{z}_2 \end{pmatrix}, \quad V_0 = \operatorname{cov}\left(\hat{\theta}_0\right) = \left(\frac{1}{n_2} - \frac{1}{N}\right)\begin{pmatrix} S_y^2 & S_{yz} \\ S_{yz} & S_z^2 \end{pmatrix}.$$

The prior restriction is

$$0 = c - R\theta = \overline{Z} - (0,\ 1)\begin{pmatrix} \theta_y \\ \theta_z \end{pmatrix}.$$

Applying (5.1) and (5.2) to this case yields the following GR estimator

$$\hat{\theta}_{GR} = \hat{\theta}_0 + K\left(c - R\hat{\theta}_0\right) = \begin{pmatrix} \overline{y}_2 \\ \overline{z}_2 \end{pmatrix} + K\left(\overline{Z} - \overline{z}_2\right)$$
$$K = V_0 R'\left(RV_0R'\right)^{-1} = \begin{pmatrix} S_{yz} \\ S_z^2 \end{pmatrix}\frac{1}{S_z^2} = \begin{pmatrix} b_{yz} \\ 1 \end{pmatrix} \qquad \left(b_{yz} = S_{yz}/S_z^2\right)$$
$$V_{GR} = \left(I_2 - KR\right)V_0 = \begin{pmatrix} 1 & -b_{yz} \\ 0 & 0 \end{pmatrix}V_0.$$

Hence, replacing b_{yz} by its estimate $b_{yz2} = s_{yz2}/s_{z2}^2$, we can approximate the first element in $\hat{\theta}_{GR}$ by $\hat{\theta}_{GRy} \approx \overline{y}_2 + b_{yz2}\left(\overline{Z} - \overline{z}_2\right)$ which corresponds to the familiar regression estimator, often denoted by $\hat{\overline{Y}}_{REG}$. For sufficiently large n_2, the variance of $\hat{\overline{Y}}_{REG}$ can be approximated by

$$\begin{aligned}\operatorname{var}\left(\hat{\bar{Y}}_{REG}\right) &\approx \operatorname{var}\left(\hat{\theta}_{GRy}\right) = \left[V_{GR}\right]_{11} = \left(\frac{1}{n_2}-\frac{1}{N}\right)\left(S_y^2 - b_{yz}S_{yz}\right) \\ &= \left(\frac{1}{n_2}-\frac{1}{N}\right)S_e^2; \qquad (A.8) \\ S_e^2 &= \frac{1}{N-1}\sum_{i\in U}\left\{Y_i - \bar{Y} - b_{yz}\left(Z_i - \bar{Z}\right)\right\}^2;\end{aligned}$$

recall from regression theory that $b_{yz}S_{yz} = b_{yz}^2 S_z^2$ and $S_y^2 = b_{yz}^2 S_z^2 + S_e^2$. The variance in (A.8) can be estimated by the well-known variance estimator

$$\hat{\operatorname{var}}\left(\hat{\bar{Y}}_{REG}\right) = \left(\frac{1}{n_2}-\frac{1}{N}\right)s_{\hat{e}2}^2, \quad \text{where} \quad s_{\hat{e}2}^2 = \frac{1}{n_2-1}\sum_{i\in s_2}\left\{Y_i - \bar{y}_2 - b_{yz2}\left(Z_i - \bar{z}_2\right)\right\}^2.$$

Similar results can be derived for more than one auxiliary variable. This illustrates once more that with respect to the bias and the variance approximation the AC estimator strongly resembles the regression estimator or, equivalently, the calibration estimator.

References

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics*, 32, 451-467.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.

Cramer, J.S. (1986). *Econometric Applications of Maximum Likelihood Methods*. Cambridge: Cambridge University Press.

Hidiroglou, M.A., Särndal, C.E. and Binder, D.A. (1995). Weighting and estimation in business surveys. In *Business Survey Methods*, (Eds., B.G. Cox et al.). New York: John Wiley and Sons, Inc.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. Transactions ASME, *Journal of Basic Engineering*, 82, 35-45.

Kish, L. (1965). *Survey sampling*. New York: John Wiley and Sons, Inc.

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. New York: Springer-Verlag.

Knottnerus, P. and Van Delden, A. (2012). On variances of changes estimated from rotating panels and dynamic strata. *Survey Methodology*, 38(1), 43-52.

Laniel, N. (1987). Variances for a rotating sample from a changing population. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 496-500.

Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: John Wiley and Sons, Inc.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Qualité, L. and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34(2), 173-181.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Silva, P.L.D.N. and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23(1), 23-32.

Smith, P., Pont, M. and Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994–2000. *Journal of the Royal Statistical Society D*, 52, 257-295.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288-289.

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24, 53-78.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 285-321
Statistics Canada, Catalogue No. 12-001-X

The estimation of gross flows in complex surveys with random nonresponse

Andrés Gutiérrez, Leonardo Trujillo and Pedro Luis do Nascimento Silva[1]

Abstract

Rotating panel surveys are used to calculate estimates of gross flows between two consecutive periods of measurement. This paper considers a general procedure for the estimation of gross flows when the rotating panel survey has been generated from a complex survey design with random nonresponse. A pseudo maximum likelihood approach is considered through a two-stage model of Markov chains for the allocation of individuals among the categories in the survey and for modeling for nonresponse.

Key Words: Design-based inference; Rotating panel surveys; Gross flows; Markov chains.

1 Introduction

Survey techniques are commonly used in order to estimate some parameters of interest in a finite population. The inference for these parameters is based on the probability distribution induced by the sampling design used to get the sample of individuals. In most of the cases for official statistics, the sample design under consideration is complex in the sense of not providing a simple random sample of the population.

After getting a probabilistic sample, sometimes it is necessary to consider the classification of the individuals in the sample through different categories in one or more nominal variables. This classification can be incorporated in a contingency table in order to summarize two variables or the temporal variations in a single variable at two different periods of time. However, in order to get accurate estimates, it is not advisable to ignore the sampling design in the inference for the parameters of interest.

Another common problem in this type of survey is nonresponse for some sample units, which can rarely be considered random or ignorable. Therefore it is necessary to consider some approach that can compensate for the potentially nonignorable nonresponse. Chen and Fienberg (1974), Stasny (1987) and recently Lu and Lohr (2010) have considered two-stage models in order to classify the individuals in a sample for two different times with nonignorable nonresponse. However, this approach ignored the sampling design that is complex and also informative for most surveys conducted for producing official statistics.

This article considers a common scenario for longitudinal surveys where the main aim is to estimate the number of population individuals belonging to several cells in a contingency table according to the categories of a variable measured at two different points in time. We also consider the modeling of the nonresponse that can affect the estimates if it is ignored. The inferential processes are tied to the complex survey design used to collect the information in the sample.

For instance, in labour force surveys, it is possible to find complex classifications depending on the labour force status of the respondents at two consecutive periods of observation and measurement. The

1. Andrés Gutiérrez, Facultad de Estadística. Universidad Santo Tomás. E-mail: hugogutierrez@usantotomas.edu.co; Leonardo Trujillo, Department of Statistics, Universidad Nacional de Colombia. E-mail: ltrujilloo@unal.edu.co; Pedro Luis do Nascimento Silva, Instituto Brasileiro de Geografia e Estatística (IBGE). E-mail: pedro-luis.silva@ibge.gov.br.

aim is to estimate the number of people that in a past period were working and are still working in the current period of observation. Another possible objective is to estimate the number of people who were unemployed in the last period of observation and are still unemployed in the current period of the survey or the number of people that in the last period of observation were employed and in the current period are unemployed or vice versa. For this example, all the entries on Table 1.1 are considered as parameters of interest. Note that even under a census, the counts in Table 1.1 may not be observable due to nonresponse.

Table 1.1
Parameters of interest in a contingency table corresponding to a labour force survey at two consecutive periods of observation.

Period 1	Period 2			
	Employed	**Unemployed**	**Inactive**	**Total**
Employed	X_{11}	X_{12}	X_{13}	X_{1+}
Unemployed	X_{21}	X_{22}	X_{23}	X_{2+}
Inactive	X_{31}	X_{32}	X_{33}	X_{3+}
Total	X_{+1}	X_{+2}	X_{+3}	X_{++}

Kalton (2009) stated that, in terms of the marginal totals, it is possible to estimate the net flows through a direct comparison between the two periods of observation. Then, it is possible to determine if the unemployment rate increased or decreased and also in what magnitude. For example, comparing that on period 1 there were $X_{1+} = \sum_j X_{1j}$ people employed, whereas on period 2 there were $X_{+1} = \sum_i X_{i1}$ people employed. Nevertheless, a more detailed analysis can be obtained analyzing the gross flows as a decomposition of the net flows. In this way, if the unemployment rate increased one percentage point, it is possible to conclude if this increase was due to the fact that one percentage point of the employed people lost their job or because ten percentage points of the employed people lost their job and nine percentage points of the unemployed people found a new job. This is possible comparing the values X_{ij}.

Also, given that in a complex survey it is possible to have unequal sampling weights and clustering and stratification effects, the likelihood function of the sampling data is difficult to find in an analytical way. Then, using classical methods of maximum likelihood would no longer be convenient for survey data from complex surveys. Then, the standard analyses must be modified to take into account the sampling weights and the sampling effects of a complex survey such as weighted estimation of proportions, variance estimation based on the sampling design and generalized corrections for the design effects (Pessoa and Silva 1998).

Section 2 surveys the basic statistical concepts used in this paper, such as survey estimators, nonresponse and categorical data inference. Section 3 proposes a superpopulation model describing the probabilistic behavior of the assignment of the individuals according to the categories of the variable considered in the survey. This corresponds to a two-stage Markov chain model. Some basic concepts of pseudo-likelihood estimation are also reviewed in Section 3. Then, in Section 4, we propose some estimators for the model parameters and the counts in the gross flows contingency table. These estimators are design-unbiased and the mathematical expressions to estimate their variance are shown in Section 5. Section 6 considers both an empirical application and a Monte Carlo simulation in order to test the

proposed methodology when the data in the survey is obtained under a simple and a complex survey design. Our simulation shows that other methodological approaches lead to biased estimation. Section 7 considers a practical application for estimating gross flows for the *Pesquisa Mensal de Emprego* (PME survey) in Brazil. In Section 8, we highlight the strengths and shortcomings of the proposed method. All the mathematical proofs are presented in the Appendix.

2 Motivation

2.1 Sampling designs and estimators

Consider a finite population as a set of N units, where $N < \infty$, forming the universe of study. N is known as the population size. Each element belonging to the population can be identified with an index k. Let U be the index set given by $U = \{1,...,k,...,N\}$. The selection of a sample $s = \{k_1, k_2, \ldots, k_{n(s)}\}$ is done according to a sampling design defined as the multivariate probability distribution over a support Q in a way that $p(s) > 0$ for every $s \in Q$ and

$$\sum_{s \in Q} p(s) = 1.$$

Under a sampling design $p(\cdot)$, an inclusion probability is assigned to every element in the population in order to denote the probability that the element belongs to the sample. For the k-th element in the population this probability is denoted as π_k and it is known as the first order inclusion probability given by

$$\pi_k = Pr(k \in S) = Pr(I_k = 1) = \sum_{s \ni k} p(s)$$

where I_k is a random variable denoting the membership of the element k to the sample, and the subindex $s \ni k$ refers to the sum over all the possible samples containing the k-th element. Analogously, π_{kl} is known as the second order inclusion probability and it denotes the probability that the elements k and l belong to the sample and it is given by

$$\pi_{kl} = Pr(k \in S; l \in S) = Pr(I_k = 1; I_l = 1) = \sum_{s \ni k,l} p(s).$$

The aim of the sample survey is to study a characteristic of interest y associated with every unit in the population and to estimate a function of interest T, called a parameter.

$$T = f(y_1, \ldots, y_k, \ldots, y_N).$$

This inferential approach is known as design-based inference. Under this approach, the estimates of the parameters and their properties depend directly on the discrete probability measure related to the chosen sampling design and do not take into account the properties of the finite population. Also, the values y_k are taken as the observation for the individual k for the characteristic of interest y. Also, y is considered as a fixed quantity rather than a random variable.

Then, the Horvitz-Thompson (HT) estimator can be defined as:

$$\hat{t}_{y,\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k$$

where $d_k = 1/\pi_k$ is the reciprocal of the first-order inclusion probability and it is known as the expansion factor or basic design weight. The HT estimator is unbiased for the total population $t_y = \sum_U y_k$, (assuming all the first order inclusion probabilities are greater than zero) and its variance is given by

$$Var\left(\hat{t}_{y,\pi}\right) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \tag{2.1}$$

where $\Delta_{kl} = Cov\left(I_k, I_l\right) = \pi_{kl} - \pi_k\pi_l$. If all the second-order inclusion probabilities are greater than zero, an unbiased estimator of (2.1) is given by

$$\widehat{Var}\left(\hat{t}_{y,\pi}\right) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

Gambino and Silva (2009) suggest that in a household survey, the main interest is to focus on characteristics for particular household members that could be related to health variables, educational variables, income/expenses, employment status, etc. In general, the sampling designs used for this kind of survey are complex and use techniques such as stratification, clustering or unequal probabilities of selection. Some of the results from repeated surveys consider the estimation of level at a particular point of time, estimation of changes between two survey rounds and the estimation of the average level parameters over repeated rounds of a survey. Different rotation schemes and the frequency of the survey can affect considerably the precision of the estimators.

2.2 Pseudo-likelihood

Some authors such as Fuller (2009), Chambers and Skinner (2003, p. 179), and Pessoa and Silva (1998, chapter 5) consider the problem where the maximum likelihood estimation is appropriate for simple random samples, as is the case in Stasny (1987), but not for samples resulting from a complex survey design. Under this scheme, it is assumed that the density population function is $f\left(y,\theta\right)$ where the parameter of interest is θ. If there is access to the information for the whole population, through a census, the maximum likelihood estimator of θ can be obtained by maximizing

$$L\left(\theta\right) = \sum_{k\in U}\log f\left(y_k,\theta\right)$$

with respect to θ. We will denote θ_N as the value maximizing the last expression. The likelihood equations for the population are given by

$$\sum_{k\in U}u_k\left(\theta\right) = 0.$$

The u_k are known as *scores* and they are defined as

$$u_k\left(\theta\right) = \frac{\partial\log f\left(y_k,\theta\right)}{\partial\theta}.$$

The pseudo-likelihood approach considers that θ_N is the parameter of interest according to the information collected in a complex sample. If $\sum_{k\in U}u_k\left(\theta\right)$ is considered as the parameter of interest, it is possible to estimate it using a weighted linear estimator

$$\sum_{k\in s} d_k u_k(\theta)$$

where d_k is a sampling design weight such as the inverse of the inclusion probability of the individual k. Then, it is possible to obtain an estimator for θ_N solving the resulting equation system.

Definition 2.1 *A maximum pseudo-likelihood estimator $\hat{\theta}_s$ for θ_N corresponds to the solution of the pseudo-likelihood equations given by*

$$\sum_{k\in s} d_k u_k(\theta) = 0.$$

Using the Taylor linearization method, the asymptotic variance of a maximum pseudo-likelihood estimator based on the sampling design is given by

$$V_p\left(\hat{\theta}_s\right) \approx \left[J(\theta_N)\right]^{-1} V_p\left[\sum_{k\in s} d_k u_k(\theta_N)\right]\left[J(\theta_N)\right]^{-1}$$

where $V_p\left[\sum_{k\in s} d_k u_k(\theta_N)\right]$ is the variance of the estimator for the population total of the *scores* based on the sampling design and

$$J(\theta_N) = \left.\frac{\partial \sum_{k\in U} u_k(\theta)}{\partial\theta}\right|_{\theta=\theta_N}.$$

An estimator for $V_p\left(\hat{\theta}_s\right)$ is given by

$$\hat{V}_p\left(\hat{\theta}_s\right) = \left[\hat{J}\left(\hat{\theta}_s\right)\right]^{-1} \hat{V}_p\left[\sum_{k\in s} d_k u_k\left(\hat{\theta}_s\right)\right]\left[\hat{J}\left(\hat{\theta}_s\right)\right]^{-1}$$

where $\hat{V}_p\left[\sum_{k\in s} d_k u_k\left(\hat{\theta}_s\right)\right]$ is a consistent estimator for the variance of the estimator of the population total of the *scores* and

$$\hat{J}\left(\hat{\theta}_s\right) = \left.\frac{\partial \sum_{k\in s} d_k u_k(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_s}.$$

Then, following Binder (1983), the asymptotic distribution of $\hat{\theta}_s$ is normal since

$$\hat{V}_p\left(\hat{\theta}_s\right)^{-1/2}\left(\hat{\theta}_s - \theta_N\right) \sim N(0,1).$$

These definitions offer a solid background for the correct inference when using large samples as is the case in labour force surveys.

2.3 Nonresponse

Särndal and Lundström (2005) state that nonresponse has been a topic of increasing interest in national statistical offices during the last decades. Also, in the sampling survey literature, the attention to this topic

has increased considerably. Nonresponse is a common non desirable issue in the development of a survey that can affect considerably the quality of the estimates.

Lohr (1999) discusses several types of nonresponse mechanisms:

- The nonresponse mechanism is ignorable when the probability of an individual responding to the survey does not depend on the characteristic of interest. Note that the word "ignorable" makes reference to a model explaining the mechanism.

- On the other hand, the nonresponse mechanism is nonignorable when the probability of an individual responding to the survey depends on the characteristic of interest. For example, in a labour survey, the possibility of response may depend on the labour force classification of the individuals in a household.

Lumley (2009, chapter 9) analyses individual nonresponse with partial data for a respondent considering a design-based approach adjusting the sampling weights. Fuller (2009, chapter 5) considers some imputation techniques for the nonresponse treatment through probabilistic models and sampling weights. Särndal (2011) considers a model-based approach through balanced sets in order to achieve higher representativeness of the estimates. In the same way, Särndal and Lundström (2010) propose a set of indicators in order to judge the effectiveness of auxiliary information in order to control the bias generated by nonresponse. Särndal and Lundström (2005) give a large number of references about nonresponse. These references examine two main complementary aspects in a survey: prevention of the problem of nonresponse (before it happens) and estimation techniques in order to take into account nonresponse in the inference process. This second aspect is known as adjustment for nonresponse.

3 Markov models for contingency tables with nonresponse

Consider the problem of estimating gross flows between two consecutive periods of time using categorical data obtained from a panel survey and under nonresponse. Also, suppose that the outcome of every interview is the classification of the respondent into any of G possible pairwise disjoint categories, and the aim is to estimate the gross flows between these categories using the information from individuals who were interviewed at two consecutive periods of time. Individuals who either did not answer in one or two periods or were excluded or included for only one of the two periods shall not have a definite classification among the categories. Then, there is one group of individuals with classification between the two periods, a group of individuals who only have the information for one of the two periods and a group of individuals who did not respond in any of the two periods of the survey.

For those individuals responding on times $t-1$ and t, the classification data can be summarized in a matrix of dimension $G \times G$. The available information for those individuals not responding the survey at time $t-1$ but responding at time t can be summarized in a column complement; the information for those individuals not responding at time t but responding at time $t-1$ can be summarized in a row complement. Finally, individuals not responding at any of the two times are included in a single cell counting the number of individuals with missing data at both times.

The whole matrix is illustrated in Table 3.1, where N_{ij} $(i,j=1,\ldots,G)$ denotes the number of individuals in the population having classification i at time $t-1$ and classification j at time t, R_i denotes the number of individuals not responding at time t and having classification i at time $t-1$, C_j denotes the number of individuals not responding at time $t-1$ and had classification j at time t, and M denotes the number of individuals in the sample not responding in any of the two times. It is important to mention that this analysis does not take into account nonresponse due to the rotation in the survey; it only takes into account individuals belonging to the matched sample ignoring those individuals not responding because they were not selected in the sample.

Table 3.1
Gross flows at two consecutive periods of time.

Time $t-1$			**Time** t		
	1	**2**	$\cdots$	G	**Row complement**
1	N_{11}	N_{12}	$\ldots$	N_{1G}	R_1
2	N_{21}	N_{22}	$\ldots$	N_{2G}	R_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	N_{G1}	N_{G2}	$\ldots$	N_{GG}	R_G
Column complement	C_1	C_2	$\ldots$	C_G	M

This paper considers ideas from Stasny (1987) and Chen and Fienberg (1974) - in the sense of considering a maximum likelihood approach in contingency tables for partially classified data - and data resulting from a two-stage process as follows:

1. In the first stage (nonobservable), the individuals are located among the cells of a matrix $G \times G$ according to the probabilities of a Markov chain process. Let η_i be the initial probability of an individual being at the category i at the time $t-1$ with $\sum_i \eta_i = 1,$ and p_{ij} be the transition probability from the category i to category j, where $\sum_j p_{ij} = 1$ for every i.

2. In the second stage (observable) of the process, every individual in cell ij can either be nonrespondent at time $t-1,$ losing the classification by row; nonrespondent at time $t,$ losing the classification by column; or nonrespondent at both times, losing both classifications.
   - Let ψ be the initial probability of an individual in cell ij responding at time $t-1$.
   - Let ρ_{RR} be the transition probability of classification of the individual in cell ij responding at time $t-1$ and responding at time t.
   - Let ρ_{MM} be the transition probability of an individual in cell ij being a nonrespondent at time $t-1$ and becoming a nonrespondent at time t.

   These probabilities do not depend on the classification stage of the individual.

Data is observed only after the second stage. The aim is to make inferences for the probabilities in the Markov chain process generating the data but also in the chain generating the nonresponse mechanism. In the context of this two-stage model, the corresponding probabilities are shown in Table 3.2.

Table 3.2
Gross flow probabilities at two consecutive times.

Time $t-1$	**Time** t						
	1	**2**	$\cdots$	j	$\cdots$	G	**Row complement**
1							
2							
$\vdots$							
i				$\{\eta_i p_{ij}\psi\rho_{RR}\}$			$\left\{\sum_j \eta_i p_{ij}\psi(1-\rho_{RR})\right\}$
$\vdots$							
G							
Column complement				$\left\{\sum_i \eta_i p_{ij}(1-\psi)(1-\rho_{MM})\right\}$			$\sum_i \eta_i p_{ij}\sum_j (1-\psi)\rho_{MM}$

In this way, the likelihood function for the observed data under this two-stage model is proportional to

$$\prod_i \prod_j \left[\psi\rho_{RR}\eta_i p_{ij}\right]^{N_{ij}} \times \prod_i \left[\sum_j \psi(1-\rho_{RR})\eta_i p_{ij}\right]^{R_i} \times \prod_j \left[\sum_i (1-\psi)(1-\rho_{MM})\eta_i p_{ij}\right]^{C_j} \times \left[\sum_i \sum_j (1-\psi)\rho_{MM}\eta_i p_{ij}\right]^{M}. \tag{3.1}$$

3.1 Parameters of interest

Data are only observed after the second stage and the aim is to make inferences for both the probabilities at the Markov chain generating the data and the chain generating nonresponse. Under this two-stage model, the probabilities of the matrix of data are shown in Table 3.2 and they constitute some of the parameters of interest.

On the other hand, coming from the non-observable process, it is necessary to consider other parameters of interest as follows. Suppose a finite population U exists, having a classification in two periods of time for all its individuals. This is a non-observable process as, even when census data is obtained, it would be not possible to have a complete classification since not all the individuals will be willing to respond. Considering this non-observable process and assuming that there are G possible classifications at each time, the distribution of the gross flows at the population level are shown in Table 3.3.

X_{ij} is the number of units at the finite population with classification i at time $t-1$ and classification j at time t $(i, j = 1, \ldots, G)$. The population size, N, must satisfy the expression:

$$N = \sum_i \sum_j X_{ij}.$$

Table 3.3
Population gross flows (non-observable process) at two consecutive periods of time.

Time $t-1$	Time t					
	1	**2**	$\cdots$	j	$\cdots$	G
1	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1G}
2	X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2G}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{iG}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
G	X_{G1}	X_{G2}	$\cdots$	X_{Gj}	$\cdots$	X_{GG}

Following the non-observable process from the last section, it is supposed that the vector corresponding to the entries at the last contingency table follows a multinomial distribution with a probability vector containing the values $\{\eta_i p_{ij}\}_{i,j=1,\ldots,G}$. This assumes a superpopulation model where the contingency table counts are considered random. In terms of notation, the probability measure considering these counts will be denoted with the subindex ξ. Then, the probability of classification at cell i,j for the k-th individual is

$$\begin{aligned}
&P_\xi\left(k \text{ has got a classification } i \text{ at } t\text{-}1 \text{ and classification } j \text{ at } t\right)\\
&= P_\xi\left(k \text{ has got a classification } i \text{ at } t\text{-}1\right)\\
&\times P_\xi\left(k \text{ has got a classification } j \text{ at } t \mid k \text{ has got a classification } i \text{ at } t\text{-}1\right)\\
&= \eta_i p_{ij}.
\end{aligned}$$

This treats X_{ij} as a random variable and if the finite population has N individuals, its expected value based on the model is given by

$$E_\xi\left(X_{ij}\right) = N\eta_i p_{ij} = \mu_{ij}. \tag{3.2}$$

Note that this expected value μ_{ij} is one of the most important parameters to be estimated on this paper as it corresponds to the expected value of the gross flows at the population of interest at the two consecutive periods. On the other hand, it is also important to understand that μ_{ij} is a parameter for the two-stage model. Also, the estimators for η_i and p_{ij} are interdependent and determined by the estimations of the defined parameters at the second stage. Let $\boldsymbol{\eta}$ be the vector containing the parameters η_i; and $\mathbf{p}$ be the vector containing the parameters p_{ij}, for every $i,j=1,\ldots,G$. The final parameters of interest are:

- the model parameters, determined by the vector

$$\boldsymbol{\theta}=\left(\psi', \rho'_{RR}, \rho'_{MM}, \boldsymbol{\eta}', \mathbf{p}'\right)';$$

- the expected value vector of the population counts defined as

$$\boldsymbol{\mu}=\left(\mu_{11},\ldots,\mu_{ij},\ldots,\mu_{GG}\right)'.$$

4 Estimation of the parameters of interest

Let N_{ij} be the total number of respondents for the population of interest having a classification i at time $t-1$ and j at time t. Let R_i be the total number of individuals in the population not responding at time t but responding at time $t-1$ with classification i. Let C_j denote the total number of individuals in the population not responding at time $t-1$ but responding at time t with classification j and finally let M be the total number of individuals at the population not responding at any of the two periods of observation. It follows that the total size of the population, N, must satisfy:

$$N = \sum_i \sum_j N_{ij} + \sum_j C_j + \sum_i R_i + M.$$

Defining the following characteristics of interest, it is possible to define the parameters of interest:

$$y_{1ik} = \begin{cases} 1, & \text{if the } k-\text{th individual responds at } t-1 \text{ with classification } i; \\ 0, & \text{otherwise.} \end{cases}$$

$$y_{2jk} = \begin{cases} 1, & \text{if the } k-\text{th individual responds at } t \text{ with classification } j; \\ 0, & \text{otherwise.} \end{cases}$$

Then, the product of these quantities, defined as $y_{1ik}y_{2jk}$, corresponds to a new characteristic of interest taking the value one if the individual has responded at both times and is classified in the cell ij, or zero otherwise. Also,

$$N_{ij} = \sum_{k \in U} y_{1ik} y_{2jk}.$$

Define the following dichotomic characteristics:

$$z_{1k} = \begin{cases} 1, & \text{if the } k-\text{th individual responds at } t-1; \\ 0, & \text{otherwise.} \end{cases}$$

$$z_{2k} = \begin{cases} 1, & \text{if the } k-\text{th individual responds at } t; \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

$$\begin{aligned} R_i &= \sum_{k \in U} y_{1ik}\left(1 - z_{2k}\right) \\ C_j &= \sum_{k \in U} y_{2jk}\left(1 - z_{1k}\right) \\ M &= \sum_{k \in U} \left(1 - z_{1k}\right)\left(1 - z_{2k}\right). \end{aligned}$$

Let w_k denote the weight for the k-th individual corresponding to a specific sampling strategy (sampling design and estimator) in both waves. Then the following expressions represent the estimators of the parameters of interest:

$$\hat{N}_{ij} = \sum_{k\in S} w_k y_{1ik} y_{2jk}$$
$$\hat{R}_i = \sum_{k\in S} w_k y_{1ik}\left(1-z_{2k}\right)$$
$$\hat{C}_j = \sum_{k\in S} w_k y_{2jk}\left(1-z_{1k}\right)$$
$$\hat{M} = \sum_{k\in S} w_k \left(1-z_{1k}\right)\left(1-z_{2k}\right)$$

for N_{ij}, R_i, C_j and M, respectively. Note that an unbiased estimation for the population size is given by

$$\hat{N} = \sum_i \sum_j \hat{N}_{ij} + \sum_j \hat{C}_j + \sum_i \hat{R}_i + \hat{M} = \sum_s w_k v_k$$

where

$$v_k = \sum_i y_{1ik} \sum_j y_{2jk} + \sum_j y_{2jk}\left(1-z_{1k}\right) + \sum_i y_{1ik}\left(1-z_{2k}\right) + \left(1-z_{1k}\right)\left(1-z_{2k}\right).$$

Taking into account the functional form of all the parameters of interest, and noticing that the likelihood function of the model is proportional to (3.1), we arrive at the following result.

Result 4.1 *The log-likelihood for the observed data at the population can be rewritten as*

$$l_U = \sum_{k\in U} f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right) \tag{4.1}$$

where

$$\begin{aligned} & f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right) \\ & = \sum_i \sum_j y_{1ik} y_{2jk} \ln\left(\psi \rho_{RR} \eta_i p_{ij}\right) \\ & + \sum_i y_{1ik}\left(1-z_{2k}\right) \ln\left(\sum_j \psi\left(1-\rho_{RR}\right)\eta_i p_{ij}\right) \\ & + \sum_j y_{2jk}\left(1-z_{1k}\right) \ln\left(\sum_i \left(1-\psi\right)\left(1-\rho_{MM}\right)\eta_i p_{ij}\right) \\ & + \left(1-z_{1k}\right)\left(1-z_{2k}\right) \ln\left(\sum_i \sum_j \left(1-\psi\right)\rho_{MM}\eta_i p_{ij}\right) \end{aligned}$$

where $\mathbf{y}_1$ *is a vector containing the characteristics* y_{1ik}, $\mathbf{y}_2$ *is a vector containing the characteristics* y_{2jk}, $\mathbf{z}_1$ *is a vector containing the characteristics* z_{1k}, *and* $\mathbf{z}_2$ *is a vector containing the characteristics* z_{2k} *(for every* $k = 1,\ldots,N$ *and* $i, j = 1,\ldots,G$*).*

Now, in order to obtain estimators of the parameters, it is necessary to maximize this last function. Using standard techniques of maximum likelihood, the corresponding likelihood equations are given by

$$\sum_{k\in U} \mathbf{u}_k\left(\theta\right) = \mathbf{0}$$

where the vector $\mathbf{u}_k$, commonly known as *scores*, is defined by

$$\mathbf{u}_k(\theta) = \frac{\partial f_k(\theta)}{\partial \theta}.$$

Also, as it is not usual to survey the whole population, a probability sample is selected and the expression $\sum_{k\in U}\mathbf{u}_k(\theta)$ is considered as a population parameter. In this way, considering $w_k = 1/\pi_k$ as the corresponding sampling weights, an unbiased estimator for this sum of *scores* is defined as $\sum_{k\in S} w_k \mathbf{u}_k(\theta)$. The next expression is known as the pseudo-likelihood equation and it is an effective way to find estimators for the model parameters taking into account the sampling weights:

$$\sum_{k\in S} w_k \mathbf{u}_k(\theta) = \mathbf{0}.$$

It is assumed that for the model in this paper, the initial probability of an individual responding at time $t-1$ is the same for all the possible classifications in the survey. Also, the transition probabilities between respondents and nonrespondents do not depend on the classification of the individual in the survey, ρ_{MM} and ρ_{RR}. Considering these assumptions, the following results will let the estimation of the Markov model probabilities take into account the sampling weights.

Result 4.2 *Under the assumptions of the model, the resulting maximum pseudo-likelihood estimators for* ψ, ρ_{RR} *and* ρ_{MM} *are given by*

$$\hat{\psi}_{mpv} = \frac{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j + \hat{M}}$$

$$\hat{\rho}_{RR,mpv} = \frac{\sum_i \sum_j \hat{N}_{ij}}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}$$

$$\hat{\rho}_{MM,mpv} = \frac{\hat{M}}{\sum_j \hat{C}_j + \hat{M}}$$

respectively.

Result 4.3 *Under the assumptions of the model, the resulting maximum pseudo-likelihood estimators for* η_i *and* p_{ij} *are obtained through iteration until convergence of the next expressions*

$$\hat{\eta}_{i,mpv}^{(v+1)} = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}$$

$$\hat{p}_{ij,mpv}^{(v+1)} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}$$

respectively. The superindex (v) *denotes the value of the estimation for the parameters of interest at the* v*-th iteration.*

The results before provide an exhaustive frame for the implementation of the two-stage Markovian model in order to take into account the sampling weights in longitudinal surveys. Another question of interest is how to choose the initial values $\left\{\hat{\eta}_i^{(0)}\right\}$ and $\left\{\hat{p}_{ij}^{(0)}\right\}$. In general, any set of values is valid if they follow the initial restrictions. These are

$$\sum_i \hat{\eta}_i^{(0)} = 1$$

$$\sum_j \hat{p}_{ij}^{(0)} = 1.$$

However, following the guidelines at Chen and Fienberg (1974) and considering the hypothetical case where all of the individuals responded in both periods, then $M = 0, R_i = 0$ (for every $i = 1,\ldots,G$) and $C_j = 0$ (for every $j = 1,\ldots,G$) and their sampling estimations are also null. Given this, and considering the expressions of the resulting estimators, a sensible choice is given by

$$\hat{\eta}_i^{(0)} = \frac{\sum_j \hat{N}_{ij}}{\sum_i \sum_j \hat{N}_{ij}}$$

$$\hat{p}_{ij}^{(0)} = \frac{\hat{N}_{ij}}{\sum_j \hat{N}_{ij}}.$$

Lastly, this iterative approach is commonly implemented for estimation problems by maximum likelihood in contingency tables. However, some approaches for the fit of log-linear models in contingency tables for complex survey designs can be found at Clogg and Eliason (1987), Rao and Thomas (1988), Skinner and Vallet (2010), among others. The next result provides an approach to gross flow estimation considering the sampling weights at both periods of interest.

Result 4.4 *Under the assumptions of the model, a sampling estimator of* μ_{ij} *is*

$$\hat{\mu}_{ij,mpv} = \hat{N}\hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}.$$

5 Properties of the estimators

Following Cassel, Särndal and Wretman (1976), the aim of considering a survey sampling approach is to gather information from just a subset (sample) of units in the finite population that enable us to obtain conclusion for the whole population. During this process, the statistician must face the randomness sources defining the complex stochastic behavior of the inferential process. Although this paper considers the sampling design as the probability measure determining the inference for the parameters and the model, it is necessary to understand that the proposed Markovian model provides another correctly defined measure of probability. Now we obtain some properties of the estimators proposed in the last section.

The aim of this paper is to incorporate the sampling weights in the proposed model and then it is important to get approximately unbiased estimators with respect to the probability measure related to the sampling design for $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$. The following results show some properties of the proposed estimators considered under the complex survey design. In terms of notation, the probability measure induced for the sampling design will be denoted with the subindex p. The following results provide the maximum likelihood estimators for the parameters of interest when instead of getting a sample, the measurement is obtained through a census or complete enumeration of the individuals in the population.

Result 5.1 *Suppose there is complete access to the whole population and the log-likelihood function of the model is given by (4.1), then the maximum likelihood estimators, under the model assumptions are*

$$\psi_U = \frac{\sum_i \sum_j N_{ij} + \sum_i R_i}{\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M}$$

$$\rho_{RR,U} = \frac{\sum_i \sum_j N_{ij}}{\sum_i \sum_j N_{ij} + \sum_i R_i}$$

$$\rho_{MM,U} = \frac{M}{\sum_j C_j + M}$$

$$\eta_{i,U}^{(v+1)} = \frac{\sum_j N_{ij} + R_i + \sum_j \left(C_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j} \tag{5.1}$$

$$p_{ij,U}^{(v+1)} = \frac{N_{ij} + \left(C_j \hat{\eta}_i^{(v)} p_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_j N_{ij} + \sum_j \left(C_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)} \tag{5.2}$$

where (5.1) and (5.2) must be jointly iterated to convergence.

Result 5.2 *Under the model assumptions, a maximum likelihood estimator of* μ_{ij} *is*

$$\mu_{ij,U} = N \times \eta_{i,U} \times p_{ij,U}$$

where N *corresponds to the population size and* $\eta_{i,U}$ *and* $p_{ij,U}$ *are defined by the last result, respectively.*

Note that both $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ can be defined as descriptive population quantities. Based on the inference approach induced by the maximum likelihood method, there exist estimators $\boldsymbol{\theta}_U = \left(\psi'_U, \boldsymbol{\rho}'_{RR,U}, \boldsymbol{\rho}'_{MM,U}, \boldsymbol{\eta}'_U, \mathbf{p}'_U \right)'$ and $\boldsymbol{\mu}_U = \left(\mu_{11,U}, \ldots, \mu_{ij,U}, \ldots, \mu_{GG,U} \right)'$ defined as the corresponding descriptive population quantities making $\boldsymbol{\theta}_{mpv}$ and $\boldsymbol{\mu}_{mpv}$ consistent with regard to the sampling design in

the sense of definition 2 in Pfeffermann (1993). Note also that $\boldsymbol{\theta}_U$ and $\boldsymbol{\mu}_U$ can be calculated only if there is access to the whole finite population.

Following Pessoa and Silva (1998, p. 79), it is possible to assess that under some regularity conditions, it follows that $\boldsymbol{\theta}_U - \boldsymbol{\theta} = o_p(1)$ and $\boldsymbol{\mu}_U - \boldsymbol{\mu} = o_p(1)$. Also, as in many sampling surveys, both the population and the sample size are generally large, then an appropriate estimator of $\boldsymbol{\theta}_U$ is also an appropriate estimator for $\boldsymbol{\theta}$, and an appropriate estimator for $\boldsymbol{\mu}_U$ will be an appropriate estimator for $\boldsymbol{\mu}$.

In the next section, we explore the properties of the estimators proposed above and we discuss about their suitability for our research problem.

5.1 Properties of the count estimators

Result 5.3 *The estimators* $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$, *and* $\hat{N}$ *defined in Section 4 are unbiased with regard to the sampling design.*

The proof is quite immediate. The weighting factor w_k corresponds to the inverse of π_k, the inclusion probability associated to the k-th element. All the estimators are of the Horvitz-Thompson class and therefore are unbiased.

Result 5.4 *Making* $w_k = 1/\pi_k$, *the corresponding variances for* $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$ *and* $\hat{N}$, *are given by*

$$
\begin{aligned}
Var_p\left(\hat{N}_{ij}\right) &= \sum_U\sum_U \Delta_{kl}\frac{y_{1ik}y_{2jk}}{\pi_k}\frac{y_{1il}y_{2jl}}{\pi_l} \\
Var_p\left(\hat{R}_i\right) &= \sum_U\sum_U \Delta_{kl}\frac{y_{1ik}(1-z_{2k})}{\pi_k}\frac{y_{1il}(1-z_{2l})}{\pi_l} \\
Var_p\left(\hat{C}_j\right) &= \sum_U\sum_U \Delta_{kl}\frac{y_{2jk}(1-z_{1k})}{\pi_k}\frac{y_{2jl}(1-z_{1l})}{\pi_l} \\
Var_p\left(\hat{M}\right) &= \sum_U\sum_U \Delta_{kl}\frac{(1-z_{1k})}{\pi_k}\frac{(1-z_{1l})}{\pi_l} \\
Var_p\left(\hat{N}\right) &= \sum_U\sum_U \Delta_{kl}\frac{v_k}{\pi_k}\frac{v_l}{\pi_l}.
\end{aligned}
$$

Unbiased estimators for these variances, respectively, are given by

$$
\begin{aligned}
\widehat{Var}_p\left(\hat{N}_{ij}\right) &= \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{1ik}y_{2jk}}{\pi_k}\frac{y_{1il}y_{2jl}}{\pi_l} \\
\widehat{Var}_p\left(\hat{R}_i\right) &= \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{1ik}(1-z_{2k})}{\pi_k}\frac{y_{1il}(1-z_{2l})}{\pi_l} \\
\widehat{Var}_p\left(\hat{C}_j\right) &= \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{2jk}(1-z_{1k})}{\pi_k}\frac{y_{2jl}(1-z_{1l})}{\pi_l}
\end{aligned}
$$

$$\widehat{Var}_p\left(\hat{M}\right) = \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{\left(1-z_{1k}\right)}{\pi_k}\frac{\left(1-z_{1l}\right)}{\pi_l}$$

$$\widehat{Var}_p\left(\hat{N}\right) = \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{v_k}{\pi_k}\frac{v_l}{\pi_l}.$$

On the other hand, if the w_k correspond to calibration weights, then all the estimators considered are asymptotically unbiased and proofs are given in Deville and Särndal (1992). Their corresponding variances are given by Kim and Park (2010).

5.2 Properties of the model probabilities estimators

Result 5.5 *The first-order Taylor approximation for the estimator ψ_{mpv}, defined at the result 4.2 above, around the point $\left(N_{ij}, R_i, C_j, M\right)$ and $i, j = 1,\ldots,G$, is given by the expression*

$$\begin{aligned}\hat{\psi}_{mpv} &\cong \hat{\psi}_0 \\ &= \psi_U + a_1\sum_i\sum_j\left(\hat{N}_{ij} - N_{ij}\right) + a_1\sum_i\left(\hat{R}_i - R_i\right) \\ &\quad + a_2\sum_j\left(\hat{C}_j - C_j\right) + a_2\left(\hat{M} - M\right)\end{aligned}$$

with

$$a_1 = \frac{\sum_j C_j + M}{\left(\sum_i\sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M\right)^2}$$

$$a_2 = -\frac{\sum_i\sum_j N_{ij} + \sum_i R_i}{\left(\sum_i\sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M\right)^2}.$$

Result 5.6 *The first-order Taylor approximation for the estimator $\hat{\rho}_{RR,mpv}$, defined at the result 4.2 above, around the point $\left(N_{ij}, R_i\right)$ and $i, j = 1,\ldots,G$, is given by the expression*

$$\begin{aligned}\hat{\rho}_{RR,mpv} &\cong \hat{\rho}_{RR,0} \\ &= \rho_{RR,U} + a_3\sum_i\sum_j\left(\hat{N}_{ij} - N_{ij}\right) + a_4\sum_i\left(\hat{R}_i - R_i\right)\end{aligned}$$

with

$$a_3 = \frac{\sum_i R_i}{\left(\sum_i\sum_j N_{ij} + \sum_i R_i\right)^2}$$

$$a_4 = -\frac{\sum_i\sum_j N_{ij}}{\left(\sum_i\sum_j N_{ij} + \sum_i R_i\right)^2}.$$

Result 5.7 *The first-order Taylor approximation for the estimator $\hat{\rho}_{MM,mpv}$, defined at the result 4.2 above, around the point $\left(C_j, M\right)$ and $j = 1,\ldots,G,$ is given by the expression*

$$\hat{\rho}_{MM,mpv} \cong \hat{\rho}_{MM,0}$$
$$= \rho_{MM,U} + a_5 \sum_j \left(\hat{C}_j - C_j \right) + a_6 \left(\hat{M} - M \right)$$

with

$$a_5 = -\frac{M}{\left(\sum_j C_j + M \right)^2}$$
$$a_6 = -\frac{\sum_j C_j}{\left(\sum_j C_j + M \right)^2}.$$

Result 5.8 *The estimators* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *and* $\hat{\rho}_{RR,mpv}$, *are approximately unbiased for* ψ_U, $\rho_{MM,U}$, $\rho_{RR,U}$.

Result 5.9 *The estimators* $\hat{\eta}_{i,mpv}$ *and* $\hat{p}_{ij,mpv}$, *are approximately unbiased for* $\eta_{i,U}$ *and* $p_{ij,U}$.

Result 5.10 *The approximate variances for the estimators* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *and* $\hat{\rho}_{RR,mpv}$, *are given by*

$$AV_p\left(\hat{\psi}_{mpv}\right) = V_p\left(\sum_s \frac{E_k^{\psi}}{\pi_k} \right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{\psi}}{\pi_k} \frac{E_l^{\psi}}{\pi_l}$$
$$AV_p\left(\hat{\rho}_{RR,mpv}\right) = V_p\left(\sum_s \frac{E_k^{RR}}{\pi_k} \right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{RR}}{\pi_k} \frac{E_l^{RR}}{\pi_l}$$
$$AV_p\left(\hat{\rho}_{MM,mpv}\right) = V_p\left(\sum_s \frac{E_k^{MM}}{\pi_k} \right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{MM}}{\pi_k} \frac{E_l^{MM}}{\pi_l}$$

where

$$E_k^{\psi} = a_1\left(2 - z_{2k}\right) + a_2\left(1 - z_{1k}\right)\left(2 - z_{2k}\right)$$
$$E_k^{RR} = a_3 + a_4\left(1 - z_{2k}\right)$$
$$E_k^{MM} = a_5\left(1 - z_{1k}\right) + a_6\left(1 - z_{1k}\right)\left(1 - z_{2k}\right).$$

Result 5.11 *Unbiased estimators for the approximate variances of the estimators* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *and* $\hat{\rho}_{RR,mpv}$, *are given by*

$$\hat{V}\left(\hat{\psi}_{mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{\psi}}{\pi_k} \frac{e_l^{\psi}}{\pi_l}$$
$$\hat{V}\left(\hat{\rho}_{RR,mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{RR}}{\pi_k} \frac{e_l^{RR}}{\pi_l}$$
$$\hat{V}\left(\hat{\rho}_{MM,mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{MM}}{\pi_k} \frac{e_l^{MM}}{\pi_l}$$

respectively, where

$$
\begin{aligned}
e_k^{\psi} &= \hat{a}_1\left(2-z_{2k}\right)+\hat{a}_2\left(1-z_{1k}\right)\left(2-z_{2k}\right)\\
e_k^{RR} &= \hat{a}_3+\hat{a}_4\left(1-z_{2k}\right)\\
e_k^{MM} &= \hat{a}_5\left(1-z_{1k}\right)+\hat{a}_6\left(1-z_{1k}\right)\left(1-z_{2k}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{a}_1 &= \frac{\sum_j \hat{C}_j+\hat{M}}{\left(\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i+\sum_j \hat{C}_j+\hat{M}\right)^2}\\
\hat{a}_2 &= -\frac{\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i}{\left(\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i+\sum_j \hat{C}_j+\hat{M}\right)^2}\\
\hat{a}_3 &= \frac{\sum_i \hat{R}_i}{\left(\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i\right)^2}\\
\hat{a}_4 &= -\frac{\sum_i\sum_j \hat{N}_{ij}}{\left(\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i\right)^2}\\
\hat{a}_5 &= -\frac{\hat{M}}{\left(\sum_j \hat{C}_j+\hat{M}\right)^2}\\
\hat{a}_6 &= -\frac{\sum_j \hat{C}_j}{\left(\sum_j \hat{C}_j+\hat{M}\right)^2}.
\end{aligned}
$$

Result 5.12 *The approximate variances for the estimators* $\hat{\eta}_{i,mpv}$ *and* $\hat{p}_{ij,mpv}$ *are given by*

$$
\begin{aligned}
AV_p\left(\hat{\eta}_{i,mpv}\right) &= \frac{1}{\left(J_{\eta_i}\right)^2}\sum_U\sum_U \Delta_{kl}\frac{u_k\left(\eta_i\right)}{\pi_k}\frac{u_l\left(\eta_i\right)}{\pi_l}\\
AV_p\left(\hat{p}_{ij,mpv}\right) &= \frac{1}{\left(J_{p_{ij}}\right)^2}\sum_U\sum_U \Delta_{kl}\frac{u_k\left(p_{ij}\right)}{\pi_k}\frac{u_l\left(p_{ij}\right)}{\pi_l}
\end{aligned}
$$

where

$$
u_k\left(\eta_i\right) = \frac{\sum_j y_{1ik}y_{2jk}+y_{1ik}\left(1-z_{2k}\right)}{\eta_i}+\sum_j y_{2jk}\left(1-z_{1k}\right)\frac{p_{ij}}{\sum_i \eta_i p_{ij}}+\left(1-z_{1k}\right)\left(1-z_{2k}\right)-1
$$

$$
\begin{aligned}
u_k\left(p_{ij}\right) = \frac{y_{1ik}y_{2jk}}{p_{ij}}+y_{1ik}\left(1-z_{2k}\right)+y_{2jk}\left(1-z_{1k}\right)\frac{\eta_i}{\sum_i \eta_i p_{ij}}+\left(1-z_{1k}\right)\left(1-z_{2k}\right)\eta_i\\
-\frac{1}{\hat{N}}\left(\sum_j \hat{N}_{ij}+\hat{R}_i+\hat{M}\eta_i+\sum_j \hat{C}_j\left(\frac{\eta_i p_{ij}}{\sum_i \eta_i p_{ij}}\right)\right)
\end{aligned}
$$

$$J_{\eta_i} = -\frac{2}{\eta_i^2}\sum_U y_{1ik} + \frac{1}{\eta_i^2}\sum_U y_{1ik} z_{2k} - \sum_U (1-z_{1k})\sum_j \frac{y_{2jk} p_{ij}^2}{\left(\sum_i \eta_i p_{ij}\right)^2}$$

$$J_{p_{ij}} = -\frac{1}{p_{ij}^2}\sum_U y_{1ik} y_{2jk} - \frac{\eta_i^2}{\left(\sum_i \eta_i p_{ij}\right)^2}\sum_U y_{2jk}(1-z_{1k}).$$

Result 5.13 *Unbiased estimators for the approximate variances of the estimators* $\hat{\eta}_{i,mpv}$ *and* $\hat{p}_{ij,mpv}$ *are given by*

$$\hat{V}_p\left(\hat{\eta}_{i,mpv}\right) = \frac{1}{\left(\hat{J}_{\hat{\eta}_i}\right)^2}\sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{u}_k\left(\hat{\eta}_i\right)}{\pi_k}\frac{\hat{u}_l\left(\hat{\eta}_i\right)}{\pi_l}$$

$$\hat{V}_p\left(\hat{p}_{ij,mpv}\right) = \frac{1}{\left(\hat{J}_{\hat{p}_{ij}}\right)^2}\sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{u}_k\left(\hat{p}_{ij}\right)}{\pi_k}\frac{\hat{u}_l\left(\hat{p}_{ij}\right)}{\pi_l}$$

where

$$\hat{u}_k\left(\hat{\eta}_i\right) = \frac{\sum_j y_{1ik} y_{2jk} + y_{1ik}(1-z_{2k})}{\hat{\eta}_i} + \sum_j y_{2jk}(1-z_{1k})\frac{\hat{p}_{ij,mpv}}{\sum_i \hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}} + (1-z_{1k})(1-z_{2k})$$

$$\hat{u}_k\left(\hat{p}_{ij}\right) = \frac{y_{1ik} y_{2jk}}{\hat{p}_{ij,mpv}} + y_{1ik}(1-z_{2k}) + y_{2jk}(1-z_{1k})\frac{\hat{\eta}_{i,mpv}}{\sum_i \hat{\eta}_{i,mpv} p_{ij,,mpv}} + (1-z_{1k})(1-z_{2k})\hat{\eta}_{i,mpv}$$

and

$$\hat{J}_{\hat{\eta}_i} = -\frac{2}{\hat{\eta}_{i,mpv}^2}\sum_U y_{1ik} + \frac{1}{\hat{\eta}_{i,mpv}^2}\sum_U y_{1ik} z_{2k} - \sum_U (1-z_{1k})\sum_j \frac{y_{2jk}\hat{p}_{ij,mpv}^2}{\left(\sum_i \hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}\right)^2}$$

$$\hat{J}_{\hat{p}_{ij}} = -\frac{1}{\hat{p}_{ij,mpv}^2}\sum_U y_{1ik} y_{2jk} - \frac{\hat{\eta}_{i,mpv}^2}{\left(\sum_i \hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}\right)^2}\sum_U y_{2jk}(1-z_{1k}).$$

5.3 Properties of gross flows estimators

Result 5.14 *Under the model assumptions, the first-order Taylor approximation of the gross flows estimator given by* $\hat{\mu}_{ij}$ *and defined in result 4.4, around the point* $(N, \eta_{i,U}, p_{ij,U})$ *and* $i, j = 1, \ldots, G$, *is given by*

$$\begin{aligned}\hat{\mu}_{ij,mpv} &\cong \hat{\mu}_{ij,0} \\ &= \mu_{ij,U} + a_7\left(\hat{N}_{ij} - N_{ij}\right) + a_8\left(\hat{\eta}_{i,mpv} - \eta_{i,U}\right) + a_9\left(\hat{p}_{ij,mpv} - p_{ij,U}\right)\end{aligned}$$

with

$$a_7 = \eta_{i,U} p_{ij,U}$$

$$a_8 = N_{ij} p_{ij,U}$$

$$a_9 = N_{ij} \eta_{i,U}.$$

Result 5.15 *The gross flows estimator* $\hat{\mu}_{ij,mpv}$ *is approximately unbiased for* $\mu_{ij,U}$.

Result 5.16 *The following expression approximate the variance for* $\hat{\mu}_{ij,mpv}$

$$AV_p\left(\hat{\mu}_{ij,mpv}\right) \cong a_7^2 Var_p\left(\hat{N}_{ij}\right) + a_8^2 AV_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 AV_p\left(\hat{p}_{ij}\right). \tag{5.3}$$

Result 5.17 *An approximately unbiased estimator for the asymptotic variance in (5.3) is given by*

$$\hat{V}_p\left(\hat{\mu}_{ij,mpv}\right) = \hat{a}_7^2 \hat{V}_p\left(\hat{N}_{ij}\right) + a_8^2 \hat{V}_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 \hat{V}_p\left(\hat{p}_{ij}\right)$$

with

$$\hat{a}_7 = \hat{\eta}_{i,U} \hat{p}_{ij,U}$$

$$\hat{a}_8 = \hat{N}_{ij} \hat{p}_{ij,U}$$

$$\hat{a}_9 = \hat{N}_{ij} \hat{\eta}_{i,U}.$$

6 Empirical application

We first consider an empirical approach in this section, through simulations that will let us assess some statistical properties such as unbiasedness and efficiency of the proposed estimators. Following the modeling proposed by Stasny (1987), we considered a two-stage simulation as follows:

- Allocation of all the individuals in the population to the different cells of a contingency table. In this first stage, we will define the initial probabilities η_i, p_{ij} and,

- Nonresponse process at two consecutive periods. In this second stage, we will define the initial probabilities ψ, ρ_{RR} and ρ_{MM}.

In the first stage, it was necessary to assume some conditions (non observable process) where the group classification probabilities were established at time $t-1$ and the conditional classification probabilities at time t. In this way, every individual in the population was assumed to be classified in any of three categories: E1, E2 and E3. The state vector at time t was given by

$$\boldsymbol{\eta} = \left(\eta_1, \eta_2, \eta_3\right)' = \left(0.9, 0.05, 0.05\right)'.$$

In this way, there is a classification probability in E1 equals to 0.9 for any individual in the population and classification probabilities in E2 and E3 equal to 0.05. The transition matrix from time $t-1$ to time t is given by

$$\mathbf{P}=\begin{pmatrix}\mathbf{p}_1'\\ \mathbf{p}_2'\\ \mathbf{p}_3'\end{pmatrix}=\begin{pmatrix}p_{11} & p_{12} & p_{13}\\ p_{21} & p_{22} & p_{23}\\ p_{31} & p_{32} & p_{33}\end{pmatrix}=\begin{pmatrix}0.80 & 0.15 & 0.05\\ 0.30 & 0.60 & 0.10\\ 0.10 & 0.10 & 0.80\end{pmatrix}.$$

We assumed that the population size was $N=100,000$ and its size would not change at the two periods of evaluation. In order to classify the individuals at the periods of time we used the R function rmultinom (R Development Core Team 2012). This way, the distribution of gross flows according to equation (3.2) would be given by the values in Table 6.1.

Table 6.1
Expected values under the model ξ for the population gross flows at two consecutive periods.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	72,000	13,500	4,500
E2	1,500	3,000	500
E3	500	500	4,000

6.1 Methodology

We considered for this empirical exercise, a Monte-Carlo with $L=1,000$ simulations. In order to classify the individuals between respondents and nonrespondents at the two periods of time, we used the function rmultinom from the language R. Dichotomic variables y_{1ik}, y_{2jk}, z_{1k} and z_{2k} were created using the function Domains from the library TeachingSampling (Gutiérrez 2009).

For each run of the simulation, a sample of size $n=10,000$ was drawn. We considered a simple random sampling design (SI) along with a complex sampling design inducing unequal inclusion probabilities (π PS). The behavior of the different proposed estimators will be assessed according to their relative bias and relative root mean square error, given by

$$RB=L^{-1}\sum_{l=1}^{L}\frac{\hat{\theta}_l-\theta}{\theta} \qquad \text{and} \qquad RRMSE=\frac{\sqrt{L^{-1}\sum_{l=1}^{L}\left(\hat{\theta}_l-\theta\right)^2}}{\theta}.$$

respectively. In those situations where the vector of inclusion probabilities was unequal, function S.piPS on the library TeachingSampling was used in order to choose a without replacement sample with inclusion probabilities proportional to an auxiliary characteristic assumed known and following a normal distribution with different parameters. The proposed methodology is compared with two other estimators: an estimator taking into account the functional shape of the model but not taking into account the sampling design and a gross flows estimator not taking into account the sampling design but assuming that the nonresponse is ignorable.

The first estimator, that we call the *Design-based estimator*, corresponds to the expressions at results 4.2, 4.3 and 4.4. The second estimator, that we shall call as *Model-based estimator*, correspond to the expressions at result 5.1, being maximum likelihood estimators not considering the sampling weights.

Finally, the third estimator, that we call the *Naive estimator* estimator expands the sampling information to the population and is given by

$$\hat{\mu}_{ij,ING} = \frac{N}{\sum_i \sum_j N_{ij}} N_{ij}.$$

The response probability at time $t-1$ was assumed as $\psi = 0.8$. The response probability at time t for those individuals responding at time $t-1$ was assumed as $\rho_{RR} = 0.9$. Finally, the nonresponse probability at time t for those individuals not responding at time $t-1$ was assumed as $\rho_{MM} = 0.7$.

Based on model ξ, the expected values of the responses are given in Table 6.2.

Table 6.2
Expected values under the model ξ for the response at two consecutive periods.

Time $t-1$	Time t	
	Response	Nonresponse
Response	72,000	8,000
Nonresponse	6,000	14,000

Taking into account the dynamics of the respondents in both periods and assuming that is possible to collect all the population information through a census, we get the classifications given in Table 6.3 below.

Table 6.3
Expected values under the model ξ for the population gross flows (observable process) at two consecutive periods.

Time $t-1$	Time t			
	E1	E2	E3	Row complement
E1	51,840	9,720	3,240	7,200
E2	1,080	2,160	360	400
E3	360	360	2,880	400
Column complement	4,440	1,020	540	14,000

6.2 Results

6.2.1 Simple random sampling: *design-based* and *model-based estimator*

In a first empirical approach, we considered a simple random sampling without replacement as the sampling design. This sampling design induces uniform inclusion probabilities and expansion factors. Under this scenario, the design-based and model-based estimators are the same. Under this scenario the approach shows some strength according to the values of the relative biases that can be considered as negligible. This can be appreciated in Tables 6.4, 6.5 and 6.6.

Table 6.4
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the proposed estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	0.24 (0.094)	-0.35 (0.189)	-0.49 (0.474)
E2	-2.89 (0.158)	-1.89 (0.221)	2.00 (0.980)
E3	-0.63 (0.790)	4.54 (0.822)	-0.84 (0.569)

Table 6.5
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the transition probabilities p_{ij}.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	0.13 (0.284)	-0.39 (0.537)	-1.00 (3.225)
E2	1.70 (1.296)	-2.29 (0.569)	8.64 (0.347)
E3	-6.6 (3.415)	2.09 (1.992)	0.56 (0.158)

Table 6.6
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the initial classification probabilities η_i.

Time $t-1$		
η_1	η_2	η_3
-0.01 (0.094)	-1.42 (0.980)	1.74 (0.790)

Also, the relative bias in percentage for the response probability ψ was -0.23 and the relative root mean square error in percentage was 0.221; for the response probability ρ_{RR}, the bias in percentage was 0.055 and the relative root mean square error in percentage was 0.031; for the nonresponse probability ρ_{MM}, the bias in percentage was -0.192 and and the relative root mean square error in percentage was 0.189. On the other hand, Table 6.7 shows the empirical expected value of the gross flows for the proposed estimator and it can be appreciated that the values are very close to those given on Table 6.1.

Table 6.7
Empirical expected values for the proposed estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	72,085	13,444	4,454
E2	1,504	2,889	535
E3	474	519	4,092

6.2.2 Simple random sampling: *naive estimator*

Under this scenario and considering that this estimator does not take into account the nonresponse process, the values of the relative biases cannot be considered as negligible. This can be appreciated in Table 6.8.

Table 6.8
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the naive estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	-1.21 (4.4)	**10.2** (60.7)	8.34 (25.2)
E2	-0.25 (38.8)	-7.51 (30.9)	1.33 (12.6)
E3	**13.7** (43.3)	-8.54 (46.1)	0.92 (6.9)

Table 6.9 shows the empirical expected values for the naive estimator; compared to the expected values for the model given in Table 6.1 these are not even close.

Table 6.9
Empirical expected values for the naive estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	54,628	760	4,507
E2	1,506	2,079	1,175
E3	1,603	905	32,832

6.2.3 Unequal inclusion probabilities: *design-based estimator*

In a third scenario, we considered a sampling design that induces unequal inclusion probabilities and expansion factors. Under this scenario, the proposed estimators are still unbiased both for the gross flows and for the parameters of the model. The relative biases are shown on Tables 6.10, 6.11 and 6.12.

Table 6.10
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the proposed estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	-0.09 (0.8)	0.25 (3.6)	3.17 (7.9)
E2	0.72 (40.9)	-1.21 (27.2)	-4.62 (71.08)
E3	1.76 (20.4)	-3.19 (22.6)	-0.73 (7.2)

Table 6.11
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the transition probabilities p_{ij}.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	-0.05 (0.7)	0.115 (3.6)	0.47 (7.1)
E2	2.39 (36.0)	-0.13 (18.6)	-6.40 (69.1)
E3	1.15 (24.9)	-5.14 (21.7)	0.49 (3.7)

Table 6.12
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the initial classification probabilities η_i.

Time $t-1$		
η_1	η_2	η_3
-0.02 (1.1)	-0.70 (19.8)	1.13 (6.9)

For the probability of response ψ, the bias in percentage was -0.46 and the relative root mean square error in percentage was 0.6; for the probability of response ρ_{RR}, the bias in percentage was -0.21 and the relative root mean square error in percentage was 0.6; for the probability of nonresponse ρ_{MM}, the bias in percentage was 0.99 and the relative root mean square error in percentage was 1.8. On the other hand, Table 6.13 shows the empirical expected values of the proposed estimator for the population gross flows and these are very close to the values given in Table 6.1.

Table 6.13
Empirical expected values of the design-based estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	71,910	13,505	4,518
E2	1,523	2,972	470
E3	511	479	4,062

6.2.4 Unequal inclusion probabilities: *model-based estimator*

A fourth scenario considers a sampling design inducing unequal inclusion probabilities and expansion factors in the same way as the last scenario. However, we consider estimators not taking into account the sampling design only the model ξ. Under this scenario, estimations are biased for both the gross flows and the model parameters as can be appreciated by the relative biases on Tables 6.14, 6.15 and 6.16.

Table 6.14
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the model-based estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	4.7 (6.1)	4.6 (8.9)	6.3 (10.5)
E2	**-89.0** (126.6)	**-89.5** (125.9)	**-88.4** (126.9)
E3	4.1 (23.8)	-3.7 (26.67)	5.3 (10.4)

Table 6.15
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the transition probabilities p_{ij}.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	0.03 (0.9)	-0.71 (4.1)	1.63 (8.6)
E2	2.77 (35.5)	-1.50 (19.6)	0.70 (70.6)
E3	4.00 (20.8)	**-14.6** (20.1)	1.33 (3.41)

Table 6.16
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the initial probabilities of classification η_i.

Time $t-1$		
η_1	η_2	η_3
4.74 (6.48)	**-89.3** (126.7)	3.95 (11.9)

In the same way, for the response probability ψ, the relative bias in percentage was -0.77 and the relative root mean square error was 1.7; for the response probability ρ_{RR}, the relative bias in percentage was -0.53 and the relative root mean square error was 0.5; for the nonresponse probability ρ_{MM}, the relative bias in percentage was 0.11 and the relative root mean square error was 1.8. On the other hand, Table 6.17 shows the empirical expected values for the model-based estimator for the population gross flows (not considering the sampling design) and these are quite far from the values in Table 6.1, especially for the second category.

Table 6.17
Empirical expected values for the model-based estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	75,438	14,039	4,790
E2	164	315	53
E3	540	443	4,213

6.2.5 Unequal inclusion probabilities: *naive estimator*

In a fifth scenario, we consider a sampling design with unequal inclusion probabilities and expansion factors. Considering the naive estimator, that does not take into account the sampling design nor the respondent model, Table 6.18 shows the relative bias for each cell in the matrix of gross flows. This estimator would be only recommendable if the nonresponse was ignorable and the sampling design would correspond to a simple random sampling design.

Table 6.18
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) for the naive estimator of the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	-28.1 (34.7)	-27.6 (60.0)	-24.5 (41.1)
E2	**497.0** (629.2)	**570.2** (610.3)	**432.7** (686.2)
E3	**-40.5** (44.2)	**-37.0** (47.4)	**-33.0** (33.8)

In order to have a more accurate comparison, it would be possible to calculate the expected values of the gross flows and compare them with the current scenario. Table 6.19 shows the empirical expected values for the naive estimator; compared to the expected values for the model given in Table 6.1 these are not very close and are especially poor for the classifications in the second category.

Table 6.19
Empirical expected values for the naive estimator of the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	51,755	9,849	3,297
E2	9,194	19,838	2,823
E3	279	295	2,665

7 Actual application: estimation of population gross flows for the PME survey

The *Pesquisa Mensal de Emprego* (PME - Brazilian Monthly Labour Survey) is a survey providing monthly indicators about the labour market in the main metropolitan areas in Brazil. Its main aim is to estimate the monthly work force and to evaluate the fluctuations and tendencies of the metropolitan labour market. It is also possible to get indicators regarding the effects of the economic conditions in the labour market and to satisfy important needs for policy planning and socio-economic development. This survey has been conducted since 1980, with some major methodological changes in 1982, 1988, 1993 and 2001 (IBGE 2007).

This section illustrates the use of the proposed estimators and the final results for the PME are shown. We will consider the panel P6 from this survey from November, 2010 to February, 2011 and then from November, 2011 to February, 2012. This window of observation administered 21,374 interviews to different people. We have chosen the first two measurements of the panel (November and December, 2010) in order to implement the proposed estimation procedure for the corresponding gross flows. Following an algorithm using the library TeachingSampling (Gutiérrez 2009), we obtain the classification at panel P6, for the months of November and December, 2010 given in Table 7.1.

Table 7.1
Labour classification and response for the occupation level in the sample of panel P6 of the PME survey.

November 2010	December 2010				
	Employed	Unemployed	Inactive	Not in the labour force	Row complement
Employed	5,231	62	227	10	386
Unemployed	51	183	113	0	28
Inactive	235	93	4,200	12	281
Not in the labour force	2	0	17	1,426	96
Column complement	499	27	372	132	7,691

However, since panel P6 corresponds to a probabilistic complex sample of the metropolitan areas in Brazil, every individual in the panel represents themselves and other additional people in the population. Then, using the proposed estimation procedure in this paper and using the corresponding expansion factors from the survey, we notice that the estimated population values for panel P6 correspond to those obtained in Table 7.2.

Table 7.2
Estimated contingency table for the population showing level of occupation and nonresponse at the two considered measurements for the panel P6 in the PME survey.

November 2010	December 2010				
	Employed	Unemployed	Inactive	Not in the labour force	Row complement
Employed	2,162,635	20,602	76,303	3,074	160,768
Unemployed	16,233	80,169	37,786	0	11,504
Inactive	70,551	31,822	1,707,675	6,018	122,412
Not in the labour force	958	0	7,035	566,530	38,171
Column complement	205,033	9,293	136,146	53,640	3,076,388

Using the estimation procedure proposed in this paper, we computed the estimated population gross flows given in Table 7.3. The corresponding estimators are unbiased under the complex design of the PME survey. According to this, the number of employed people in both periods of measurement is estimated as 3,913,274, whereas the number of inactive people for both periods is estimated as 3,035,463.

Table 7.3
Population estimated gross flows for both periods at the PME survey. Estimated coefficients of variation in percentage are shown in brackets.

November 2010	December 2010			
	Employed	Unemployed	Inactive	Not in the labour force
Employed	3,913,274 (0.2)	36,570 (3.1)	136,102 (1.6)	5,573 (7.2)
Unemployed	29,776 (3.5)	144,253 (1.7)	68,320 (2.1)	0 (-)
Inactive	127,193 (1.6)	56,296 (2.3)	3,035,463 (0.3)	10,872 (6.5)
Not in the labour force	1,727 (17.3)	0 (-)	12,496 (5.8)	1,022,836 (0.5)

The estimates in the last table above are the result of the proposed estimation procedure in this paper. Next, we show the estimated parameters on the first stage of the model, defined as the transition probabilities from one category to another in both observation periods.

Table 7.4
Estimation of the probabilities p_{ij}. Estimated coefficients of variation in percentage are shown in brackets.

November 2010	December 2010			
	Employed	Unemployed	Inactive	Not in the labour force
Employed	0.9564 (0.1)	0.0089 (3.1)	0.0332 (1.6)	0.0013 (7.2)
Unemployed	0.1228 (3.4)	0.5952 (1.1)	0.2819 (2.0)	0 (-)
Inactive	0.0393 (1.5)	0.0174 (2.3)	0.9398 (0.1)	0.0033 (6.5)
Not in the labour force	0.0016 (17.6)	0 (-)	0.0120 (5.8)	0.9862 (0.1)

The initial probabilities of classification on the first period of interest are shown in Table 7.5. It can be noticed that, for this particular survey, the highest classification probabilities can be found for the categories of employed and inactive.

Table 7.5
Estimation of the probabilities η_i. Estimated coefficients of variation in percentage are shown in brackets.

November 2010			
η_1	η_2	η_3	η_4
0.4757 (0.2)	0.0281 (1.2)	0.3755 (0.3)	0.1205 (0.5)

Finally, the general response probability was estimated as $\hat{\psi}_{mpv} = 0.595$ (with an estimated coefficient of variation of 0.1%). That means that the rate of response is around 60%. Also, the transition probability that a nonrespondent in the first period is still a nonrespondent the next time was estimated as $\hat{\rho}_{MM,mpv} = 0.883$ (with an estimated coefficient of variation of 0.1%). The transition probability that a respondent in the first period stays on as a respondent the next time was estimated as $\hat{\rho}_{RR,mpv} = 0.934$ (with an estimated coefficient of variation of 0.1%). In general terms, it is possible to state that a status response of an individual in the first period is not changing significatively by the second.

8 Conclusions

This paper has considered a common problem in survey sampling applications. Using superpopulation Markov chain based models, a new methodology was proposed leading to approximately unbiased estimators of gross flows at different times for the particular case of data coming from complex surveys with unequal sampling weights. Possible applications of the methodology in this paper are broad in the case of, for example, national statistical offices considering complex surveys. Life quality or labour force surveys are usually concerned about the estimation of gross flows. However, the possible extensions of this methodology could be applied to the public policy sector for impact evaluations having a classification of the respondents before and after an intervention.

Also we present a solution to a general problem such as nonignorable nonresponse. Models where the nonresponse is not differentiated at different periods or by classification status were considered. However, in some practical applications, it is possible that this is not the case.

The approach of this paper considers that design weights for units between the two time periods are the same. Further work will try to consider different weights between waves by considering either a two-phase sampling scheme or a calibration approach in two-stages. Indeed, it would be of interest to compare the performance of the methodology given in this paper with the calibration methodology. One could consider the approach of Ash (2005) and Sikkel, Hox and de Leeuw (2008) to calibrate in two periods along with Särndal and Lundström (2005) for handling nonresponse.

Further work will try to extend this methodology for more complex Markov chain models in order to consider different sampling weights. A new definition of parameters in the model will be necessary. Also, this methodology could be extended to the case of gross flows in more than two periods of time where classification errors are taken into account.

Acknowledgements

The authors wish to thank two anonymous referees for their constructive comments on an earlier version of this manuscript which resulted in this improved version. Also, the first author wishes to thank Universidad Santo Tomas for the financial support during his PhD studies. This paper is a result of the PhD thesis of Andrés Gutiérrez at Universidad Nacional de Colombia under supervision of the other two authors.

Appendix

A.1 Mathematical proofs of the results on the paper

In this section, the mathematical proofs for some of the most important results in this paper are included.

Proof of Result 4.1

Proof. Taking logarithm to the likelihood function, and defining it as l, it follows that

$$
\begin{aligned}
l_U &= \ln\left(L_U\right) \\
&= \sum_i \sum_j N_{ij} \ln\left(\psi \rho_{RR} \eta_i p_{ij}\right) + \sum_i R_i \ln\left(\sum_j \psi\left(1-\rho_{RR}\right)\eta_i p_{ij}\right) \\
&\quad + \sum_j C_j \ln\left(\sum_i (1-\psi)(1-\rho_{MM})\eta_i p_{ij}\right) + M \ln\left(\sum_i \sum_j (1-\psi)\rho_{MM}\eta_i p_{ij}\right).
\end{aligned}
$$

Note that $N_{ij} = \sum_{k\in U} y_{1ik} y_{2jk}$, $R_i = \sum_{k\in U} y_{1ik}\left(1-z_{2k}\right)$, $C_j = \sum_{k\in U} y_{2jk}\left(1-z_{1k}\right)$ and $M = \sum_{k\in U}\left(1-z_{1k}\right)\left(1-z_{2k}\right)$. After factorizing the sum over the whole population, the result is finally obtained.

Proof of Result 4.2

Proof. Starting from the definition of pseudo-likelihood and taking into account the model assumptions, it follows that

$$
\begin{aligned}
l_U &= \sum_{k\in U}\Bigg[\sum_i \sum_j y_{1ik} y_{2jk}\left[\ln\left(\psi\right) + \ln\left(\rho_{RR}\right) + \ln\left(\eta_i\right) + \ln\left(p_{ij}\right)\right] \\
&\quad + \sum_i y_{1ik}\left(1-z_{2k}\right)\left[\ln\left(\psi\right) + \ln\left(1-\rho_{RR}\right) + \ln\left(\eta_i\right) + \ln\left(\sum_j p_{ij}\right)\right] \\
&\quad + \sum_j y_{2jk}\left(1-z_{1k}\right)\left[\ln(1-\rho_{MM}) + \ln(1-\psi) + \ln\left(\sum_i \eta_i p_{ij}\right)\right] \\
&\quad + \left(1-z_{1k}\right)\left(1-z_{2k}\right)\left[\ln\left(1-\psi\right) + \ln\left(\rho_{MM}\right) + \ln\left(\sum_i \sum_j \eta_i p_{ij}\right)\right]\Bigg] \\
&= \sum_{k\in U} f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right).
\end{aligned}
$$

The *score* for ψ can be defined as

$$
\begin{aligned}
u_k\left(\psi\right) &= \frac{\partial f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right)}{\partial \psi} \\
&= \frac{(1-\psi)\left(\sum_i \sum_j y_{1ik} y_{2jk} + \sum_i y_{1ik}\left(1-z_{2k}\right)\right) - \psi\left(\sum_j y_{2jk}\left(1-z_{1k}\right) + \left(1-z_{1k}\right)\left(1-z_{2k}\right)\right)}{\psi\left(1-\psi\right)}.
\end{aligned}
$$

Then, for this parameter, the pseudo-likelihood equations are given by

$$
\sum_{k\in S} w_k u_k\left(\psi\right) = 0.
$$

Solving for ψ, it is obtained that

$$
\hat{\psi}_{mpv} = \frac{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_{i\,i} + \sum_j \hat{C}_j + \hat{M}}.
$$

Following an analogous process for the remaining parameters, the result is obtained.

Proof of Result 4.3

Proof. First, it is necessary to warn that the estimation for these parameters is subject to the restrictions $\sum_i \eta_i = 1$ and $\sum_j p_{ij} = 1$. Then, the process must consider the use of Lagrange multipliers. The function to be maximized, including these restrictions, can be expressed as

$$l_U + \lambda_1\left(\sum_i \eta_i - 1\right) + \lambda_2\left(\sum_j p_{ij} - 1\right).$$

Then, the corresponding *score* for η_i is defined by

$$\begin{aligned} u_k(\eta_i) &= \frac{\partial f_k(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2)}{\partial \eta_i} + \frac{\partial \lambda_1\left(\sum_i \eta_i - 1\right)}{\partial \eta_i} \\ &= \frac{\sum_j y_{1ik} y_{2jk} + y_{1ik}(1 - z_{2k})}{\eta_i} + \sum_j y_{2jk}(1 - z_{1k})\frac{p_{ij}}{\sum_i \eta_i p_{ij}} + (1 - z_{1k})(1 - z_{2k}) + \lambda_1. \end{aligned}$$

The last step takes into account the restrictions, since $\sum_i \sum_j \eta_i p_{ij} = \sum_i \eta_i \sum_j p_{ij} = \sum_i \eta_i = 1$. Then, for this parameter, the pseudo-likelihood equations are given by

$$\sum_{k \in S} w_k u_k(\eta_i) = 0.$$

Then, after some algebra, it follows that

$$\eta_i = \frac{\sum_j \sum_s w_k y_{1ik} y_{2jk} + \sum_s w_k y_{1ik}(1 - z_{2k}) + \sum_j \sum_s w_k y_{2jk}(1 - z_{1k})\left(\eta_i p_{ij} \big/ \sum_i \eta_i p_{ij}\right)}{-\sum_s w_k (1 - z_{1k})(1 - z_{2k}) - \lambda_1 \sum_s w_k}.$$

Besides, using the restriction $\sum_i \eta_i = 1$ and adding up over i, it follows that

$$\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j = \left(-\sum_s w_k (1 - z_{1k})(1 - z_{2k}) - \lambda_1 \sum_s w_k\right).$$

Then, we finally obtain that

$$\eta_i = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \eta_i p_{ij} \big/ \sum_i \eta_i p_{ij}\right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}.$$

On the other hand, in order to find the maximum pseudo-likelihood estimator of $\{p_{ij}\}$, the *score* for p_{ij} is defined as

$$\begin{aligned} u_k(p_{ij}) &= \frac{\partial f_k(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2)}{\partial p_{ij}} + \frac{\partial \lambda_2\left(\sum_i p_{ij} - 1\right)}{\partial p_{ij}} \\ &= \frac{y_{1ik} y_{2jk}}{p_{ij}} + y_{1ik}(1 - z_{2k}) + y_{2jk}(1 - z_{1k})\frac{\eta_i}{\sum_i \eta_i p_{ij}} + (1 - z_{1k})(1 - z_{2k})\eta_i + \lambda_2. \end{aligned}$$

Hence,

$$p_{ij} = \frac{\sum_s w_k y_{1ik} y_{2jk} + \sum_s w_k y_{2jk} (1 - z_{1k}) p_{ij} \eta_i / \sum_i \eta_i p_{ij}}{-\sum_s w_k y_{1ik} (1 - z_{2k}) - \sum_s w_k (1 - z_{1k})(1 - z_{2k}) \eta_i - \sum_s w_k \lambda_2}.$$

Using the restriction $\sum_j p_{ij} = 1$ and adding up over j on both sides, it follows that

$$\sum_j \hat{N}_{ij} + \sum_j \hat{C}_j \frac{p_{ij} \eta_i}{\sum_i \eta_i p_{ij}}$$
$$= \left(-\sum_s w_k y_{1ik} (1 - z_{2k}) - \sum_s w_k (1 - z_{1k})(1 - z_{2k}) \eta_i - \sum_s w_k \lambda_2 \right).$$

Then, it follows that

$$p_{ij} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \eta_i p_{ij} / \sum_i \eta_i p_{ij} \right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \eta_i p_{ij} / \sum_i \eta_i p_{ij} \right)}.$$

Now, note that it is impossible to solve the last expression for the $\{p_{ij}\}$ in such a way that the solution is a closed expression. The same happens with the expression for the $\{\eta_i\}$. However, it is possible to use an iterative approach, which has proven to have a fast convergence in maximum likelihood estimation problems for contingency tables. This approach assumes that the maximum pseudo-likelihood estimator can be found after jointly iterating the following expressions at step $(v+1)$, for $v \geq 1$,

$$\hat{\eta}_{i,mpv}^{(v+1)} = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} / \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}$$

$$\hat{p}_{ij,mpv}^{(v+1)} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} / \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} / \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}.$$

This particular iterative procedure was used initially for the formulation of nested likelihood models by Hocking and Oxspring (1971). However, it also appears implemented by Blumenthal (1968), Reinfurt (1970), Chen and Fienberg (1974), Fienberg and Stasny (1983), Stasny (1987), Stasny (1988), and others.

Proof of Result 5.5

Proof. The non-linear estimator $\hat{\psi}_{mpv}$, can be expressed as a function of the estimated totals $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$ and $\hat{M}$ (with $i, j = 1, \ldots, G$). Then,

$$\hat{\psi}_{mpv} = f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right).$$

Finally, the first order Taylor approximation at the point $\left(\hat{N}_{ij} = N_{ij}, \hat{R}_i = R_i, \hat{C}_j = C_j, \hat{M} = M \right)$ is given by

$$\hat{\psi}_{mpv} = \psi_U + a_1 \sum_i \sum_j \left(\hat{N}_{ij} - N_{ij} \right) + a_1 \sum_i \left(\hat{R}_i - R_i \right) + a_2 \sum_j \left(\hat{C}_j - C_j \right) + a_2 \left(\hat{M} - M \right)$$

with

$$a_1 = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{R}_i} \right|_{\substack{\hat{N}_{ij}=N_{ij} \\ \hat{R}_i=R_i \\ \hat{C}_j=C_j \\ \hat{M}=M}} = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{N}_{ij}} \right|_{\substack{\hat{N}_{ij}=N_{ij} \\ \hat{R}_i=R_i \\ \hat{C}_j=C_j \\ \hat{M}=M}} = \frac{\sum_j C_j + M}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M \right)^2}$$

and

$$a_2 = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{C}_j} \right|_{\substack{\hat{N}_{ij}=N_{ij} \\ \hat{R}_i=R_i \\ \hat{C}_j=C_j \\ \hat{M}=M}} = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{M}} \right|_{\substack{\hat{N}_{ij}=N_{ij} \\ \hat{R}_i=R_i \\ \hat{C}_j=C_j \\ \hat{M}=M}} = -\frac{\sum_i \sum_j N_{ij} + \sum_i R_i}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M \right)^2}.$$

Proof of Result 5.8

Proof. Calculating the expected value under the sampling design, it follows that

$$\begin{aligned} AE_p\left(\hat{\psi}_{mpv} \right) &\cong E_p\left(\hat{\psi}_0 \right) \\ &= \psi_U + a_1 \sum_i \sum_j \left(E_p\left(\hat{N}_{ij} \right) - N_{ij} \right) + a_1 \sum_i \left(E_p\left(\hat{R}_i \right) - R_i \right) \\ &\quad + a_2 \sum_j \left(E_p\left(\hat{C}_j \right) - C_j \right) + a_2 \left(E_p\left(\hat{M} \right) - M \right) \\ &= \psi_U . \end{aligned}$$

Following a similar process for the remaining estimators, the result is obtained. This proof is a result of the application of the pseudo-likelihood method that induces unbiased estimations for the population parameters in the model as it is proved on Corollary 1 at Binder (1983, p. 291).

Proof of Result 5.10

Proof. Considering $\hat{\psi}_{mpv}$, replacing the expressions for $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$ and after some algebraic simplifications, the approximate variance can be expressed as

$$AV\left(\hat{\psi}_{mpv} \right) = Var\left(a_1 \sum_i \sum_j \hat{N}_{ij} + a_1 \sum_i \hat{R}_i + a_2 \sum_j \hat{C}_j + a_2 \hat{M} \right) = Var\left(\sum_{k \in S} \frac{E_k^{\psi}}{\pi_k} \right).$$

Initially, we have that

$$E_k^{\psi} = a_1 \sum_i \sum_j y_{1ik} y_{2jk} + a_1 \sum_i y_{1ik} \left(1 - z_{2k} \right) + a_2 \sum_j y_{2jk} \left(1 - z_{1k} \right) + a_2 \left(1 - z_{1k} \right)\left(1 - z_{2k} \right).$$

Then, using that $\sum_i \sum_j y_{1ik} y_{2jk} = \sum_i y_{1ik} = \sum_j y_{2jk} = 1$ and after some algebra, it follows that

$$E_k^{\psi} = a_1\left(2 - z_{2k}\right) + a_2\left(1 - z_{1k}\right)\left(2 - z_{2k}\right).$$

After an analogous process for $\hat{\rho}_{RR,mpv}$ and $\hat{\rho}_{MM,mpv}$, the variance expressions at the heading of this result are obtained.

Proof of Result 5.12

Proof. The proof is obtained following expression (3.3) at Binder (1983), and taking into account that

$$J_{\eta_i} = \frac{\partial \sum_U u_k\left(\eta_i\right)}{\partial \eta_i}$$

$$J_{p_{ij}} = \frac{\partial \sum_U u_k\left(p_{ij}\right)}{\partial p_{ij}}.$$

Also,

$$\frac{\partial u_k\left(\eta_i\right)}{\partial \eta_i} = -\frac{2y_{1ik} - y_{1ik}z_{2k}}{\eta_i^2} - \left(1 - z_{1k}\right)\sum_j \frac{y_{2jk}p_{ij}^2}{\left(\sum_i \eta_i p_{ij}\right)^2}$$

$$\frac{\partial u_k\left(p_{ij}\right)}{\partial p_{ij}} = -\frac{y_{1ik}y_{2jk}}{p_{ij}^2} - \frac{\eta_i^2}{\left(\sum_i \eta_i p_{ij}\right)^2} y_{2jk}\left(1 - z_{1k}\right).$$

Proof of Result 5.16

Proof.

$$\begin{aligned} AV_p\left(\hat{\mu}_{ij,mpv}\right) &= a_7^2 Var_p\left(\hat{N}_{ij}\right) + a_8^2 AV_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 AV_p\left(\hat{p}_{ij}\right) \\ &\quad + 2a_7a_8 Cov\left(\hat{N}_{ij}, \hat{\eta}_{i,mpv}\right) + 2a_7a_9 Cov\left(\hat{N}_{ij}, \hat{p}_{ij}\right) 2a_8a_9 Cov\left(\hat{\eta}_{i,mpv}, \hat{p}_{ij}\right) \\ &\cong a_7^2 Var_p\left(\hat{N}_{ij}\right) + a_8^2 AV_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 AV_p\left(\hat{p}_{ij}\right). \end{aligned}$$

This due to

$$\begin{aligned} Cov\left(\hat{N}_{ij}, \hat{\eta}_{i,mpv}\right) &= E_p\left(\hat{N}_{ij}\hat{\eta}_{i,mpv}\right) - E_p\left(\hat{N}_{ij}\right)E_p\left(\hat{\eta}_{i,mpv}\right) \\ &\cong \hat{N}_{ij,U}\eta_{i,U} - \hat{N}_{ij,U}\eta_{i,U} = 0. \end{aligned}$$

Then, it is possible to get:

$$E_p\left(\hat{N}_{ij}\hat{\eta}_{i,mpv}\right) \cong \hat{N}_{ij,U}, \eta_{i,U}$$

using Taylor linearization for $\left(\hat{N}_{ij,U}, \eta_{i,U}\right)$. The other covariances are obtained in a similar way.

References

Ash, S. (2005). Calibration weights for estimators of longitudinal data with an application to the National Long Term Care Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*. American Statistical Association: Alexandria, VA, 2694–2699.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63, 542-551.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). *Foundations of Inference in Survey Sampling*. New York: Wiley.

Chambers, R.L. and Skinner, C.J. (2003). *Analysis of Survey Data*. John Wiley and Sons, Chichester: UK.

Chen, T. and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Clogg, C.C. and Eliason, S.R. (1987). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-44.

Deville, J. and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fienberg, S.E. and Stasny, E.A. (1983). Estimating monthly gross flows in labour force participation. *Survey Methodology*, 9(1), 77-102.

Fuller, W.A. (2009). *Sampling Statistics*. Wiley.

Gambino, J.G. and Silva, P.L. (2009). Sampling and estimation in household surveys. In D. Pfeffermann and C.R. Rao (Eds.), *Handbook of Statistics*. Vol. 29A. Sample Surveys: Design, Methods and Applications (pp. 407-439). Amsterdam: Elsevier.

Gutiérrez, H.A. (2009). TeachingSampling: Sampling designs and parameter estimation in finite population. R package version 2.0.1.

Hocking, R.R. and Oxspring, H.H. (1971). Maximun likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66, 65-70.

IBGE (2007). *Pesquisa Mensal de Emprego*. Vol. 23, 2nd edition.

Kalton, G. (2009). Designs for surveys over time. In D. Pfeffermann and C.R. Rao (Eds.), *Handbook of Statistics*. Vol. 29A. Sample Surveys: Design, Methods and Applications (pp. 89-108). Amsterdam: Elsevier.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78, 21-39.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Lu, Y. and Lohr, S. (2010). Gross flow estimation in dual frame surveys. *Survey Methodology*, 36(1), 13-22.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R*. New York: Wiley.

Pessoa, D.G.C. and Silva, P.L. (1998). *Análise de Dados Amostrais Complexos*. São Paulo : Associação Brasileira de Estatística.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

R Development Core Team (2012). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J.N.K. and Thomas, D.R. (1988). The analysis of cross-classified data from complex surveys. *Sociological Methodology*, 18, 213-269.

Reinfurt, D.W. (1970). The analyis of categorical data with supplemented margins including applications to mixed models. Unpublished Ph.D dissertation. Department of Biostatistics. University of North Carolina.

Särndal, C.E. (2011). The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection. *Journal of Official Statistics*, 27, 1-21.

Särndal, C.E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* John Wiley and Sons, Chichester: UK.

Särndal, C.E. and Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36(2), 131-144.

Sikkel, D., Hox, J. and de Leeuw, E. (2008). Using auxiliary data for adjustment in longitudinal research. In P. Lynn (Ed), *Methodology of longitudinal surveys*. New York: Wiley. An earlier version is available at http://www.iser.essex.ac.uk/ulsc/mols2006/programme/data/papers/Sikkel.pdf

Skinner, C.J. and Vallet, L.A. (2010). Fitting log-linear models to contingency tables from surveys with complex sampling designs: An investigation of the Clogg-Eliason approach. *Sociological Methods and Research*, 39, 83-108.

Stasny, E.A. (1987). Some Markov-chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 3, 359-373.

Stasny, E.A. (1988). Modeling nonignorable nonresponse in categorical panel data with an example in estimating gross labor-flows. *Journal of Business and Economic Statistics*, 6, 207-219.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 323-334
Statistics Canada, Catalogue No. 12-001-X

Chi-squared tests in dual frame surveys

Yan Lu[1]

Abstract

In order to obtain better coverage of the population of interest and cost less, a number of surveys employ dual frame structure, in which independent samples are taken from two overlapping sampling frames. This research considers chi-squared tests in dual frame surveys when categorical data is encountered. We extend generalized Wald's test (Wald 1943), Rao-Scott first-order and second-order corrected tests (Rao and Scott 1981) from a single survey to a dual frame survey and derive the asymptotic distributions. Simulation studies show that both Rao-Scott type corrected tests work well and thus are recommended for use in dual frame surveys. An example is given to illustrate the usage of the developed tests.

Key Words: Asymptotic properties; Chi-squared tests; Dual frame surveys; First-order corrected test; Second-order corrected test; Simulations.

1 Introduction

A general situation of a dual frame survey is depicted in Figure 1.1, where the union of frame A and frame B is denoted as the union of the three nonoverlapping domains, i.e., $A \cup B = a \cup ab \cup b$. Probability samples are selected independently from these two frames.

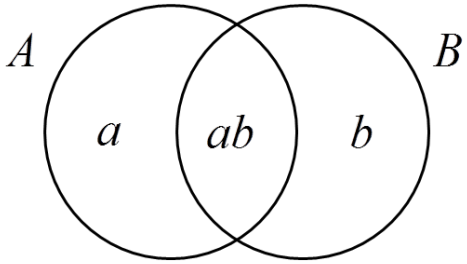

Figure 1.1: Frames A and B are both incomplete but overlapping

A dual frame survey often gives better coverage of the population, and can achieve considerable cost savings. The statistical literature has several methods for cross-sectional analyses of dual-frame survey data, see Hartley (1962, 1974), Fuller and Burmeister (1972), Skinner (1991), Skinner and Rao (1996), Lohr and Rao (2000, 2006), etc. As Rao and Thomas (1988) noted, the need to perform statistical analyses of categorical data is frequently encountered in quantitative sociological research. Pearson's chi-squared test and likelihood ratio test are both well known tests for categorical data. These methods rely on the assumption that data are obtained by simple random sampling (SRS) from one or more large population. Most current surveys have complex designs with stratification and clustering, where the SRS assumption is violated. Wald's test (Wald 1943) is one of the earliest methods proposed to assess model fit in complex designs. Fay (1979, 1985) proposed a jackknifed chi-squared test for use in complex surveys. Both Wald's (1943) and Fay's (1979) procedures require detailed survey information from which the covariance matrix can be estimated. Such detailed information is often not available in practice. Rao and Scott (1981, 1984)

1. Yan Lu, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001. E-mail: luyan@math.unm.edu.

proposed chi-squared tests for goodness of fit and independence in two-way and multi-way tables. Bedrick (1983) and Rao and Scott (1987) also studied the use of limited information on cell and marginal design effects to provide approximate tests. Thomas, Singh and Roberts (1996) described a Monte Carlo study of developed procedures for testing independence in a two-way table.

The research problem in this article arises from categorical data analysis in dual frame surveys. For example, a dual frame may consist of the online membership directories of the American Statistical Association (ASA) and the Institute for Mathematical Statistics (IMS). The overlap domain consists of the statisticians who are members from both societies. One may be interested in testing if the percentage of female in academia is the same across the three domains (domain a: members of ASA only; domain ab: members from both ASA and IMS; domain b: members of IMS only). The tests in a dual frame survey present additional challenges to those from a single frame survey because there are now two samples, each with a possibly complex sampling design and may have an unknown degree of overlap. It is possible to apply a fixed weighting constant for the overlap domain, say 1/2, and consider the union of sample A $\left(\mathcal{S}_A\right)$ and sample B $\left(\mathcal{S}_B\right)$ as a single sample. By doing so, the chi-squared tests for a single frame survey in literature such as Rao and Scott (1981) could be applied. However, this application is based on the assumption that a set of ultimate cell proportions exist for the dual frame structure, which is not necessary true. In this paper, we assume that each domain has their own set of cell proportions, under which Rao and Scott (1981) type estimator is a special case when the three sets of cell proportions in the three domains are all the same. We extend Wald's (1943) test and Rao-Scott first-order and second-order corrected tests (Rao and Scott 1981) from a single survey to a dual frame survey and derive asymptotic distributions.

This paper is organized as follows. Section 2 gives a background of the research. Section 3 proposes several chi-squared tests. Section 4 gives a small simulation study of the proposed chi-squared tests under a simple hypothesis. Section 5 gives a real example study. Finally, we give a summary in Section 6.

2 Background

2.1 Chi-squared tests in a single frame survey

Consider a one-way frequency table with k classes and associated finite population proportions $p_1, p_2, \cdots, p_k$ with $\sum_{i=1}^{i=k} p_i = 1$. Let $n_1, \cdots, n_k$ denote the observed cell frequencies in a sample falling in each of k categories with $\sum_{i=1}^{k} n_i = n$. Under SRS, the Pearson chi-squared statistic for testing simple hypothesis $H_0 : p_i = p_{0i},\ \left(i = 1, \cdots, k\right)$ is given by

$$\tilde{X}^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_{0i}\right)^2}{np_{0i}}. \tag{2.1}$$

For complicated designs, $\tilde{X}^2$ involve noncentral distributions. It is natural to consider a more general statistic

$$X^2 = n\sum_{i=1}^{k} \frac{\left(\hat{p}_i - p_{0i}\right)^2}{p_{0i}}, \tag{2.2}$$

where $\hat{p}_i$ is a consistent estimator of p_i under a specified sampling design $p(s)$.

Let $\hat{\mathbf{p}} = (\hat{p}_1, \cdots, \hat{p}_{k-1})'$ represent the $k-1$ vector of estimated proportions with $\hat{p}_k = 1 - (\hat{p}_1 + \cdots + \hat{p}_{k-1})$; $\mathbf{p}_0$ be the corresponding $k-1$ vector of hypothesized proportions; $\mathbf{V}$ be the $(k-1)\times(k-1)$ covariance matrix of $\hat{\mathbf{p}}$, and $\hat{\mathbf{V}}$ be the estimate of $\mathbf{V}$ obtained from the survey data. The generalized Wald statistic

$$X_W^2 = (\hat{\mathbf{p}} - \hat{\mathbf{p}}_0)' \hat{\mathbf{V}}^{-1} (\hat{\mathbf{p}} - \hat{\mathbf{p}}_0), \tag{2.3}$$

is distributed asymptotically as χ^2_{k-1} under $H_0 : p_i = p_{0i}$, $(i = 1, \cdots, k)$ for sufficiently large n.

Rao and Scott (1981) showed that under H_0, X^2 in (2.2) is distributed asymptotically as a weighted sum $\delta_1 W_1 + \cdots + \delta_{k-1} W_{k-1}$ of $k-1$ independent χ^2_1 random variables W_i, $i = 1, 2, \cdots, k-1$. The δ_is are the eigenvalues of a design effect matrix $\mathbf{P}^{-1}\mathbf{V}$, where $\mathbf{P}$ is the covariance matrix corresponding to SRS when H_0 is true, i.e. $\mathbf{P} = n^{-1}(diag(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0')$. The standard result of Pearson test is recovered under SRS. Let $\hat{\delta}_i$ be an estimate of δ_i and $\hat{\delta}_\cdot = \left(\sum_{i=1}^{k-1} \hat{\delta}_i\right) \Big/ (k-1)$, the Rao-Scott first order corrected test refers $X^2 / \hat{\delta}_\cdot$ to χ^2_{k-1}. When the full estimated covariance matrix $\hat{\mathbf{V}}$ is known, a better approximation to the asymptotic distribution of X^2 is to match the first moment and second moment of the test statistic to a χ^2 distribution. The Rao-Scott (Rao and Scott 1981) second-order corrected test statistic considers $X_S^2 = X^2 \Big/ \left[\hat{\delta}_\cdot (1 + \hat{a}^2)\right]$. This statistic is approximately a chi-squared random variable on $v = (k-1) \Big/ (1 + a^2)$ degrees of freedom, where $\hat{a}$ is an estimate of a with $\hat{a}^2 = \sum_{i=1}^{k-1} \hat{\delta}_i^2 \Big/ \left[(k-1)\hat{\delta}_\cdot^{\,2}\right] - 1$, and $\sum_{i=1}^{k-1} \hat{\delta}_i^2 = n^2 \sum_{i=1}^{k} \sum_{j=1}^{k} \hat{\mathbf{V}}_{ij}^2 / p_{0i} p_{0j}$. If the design effects are all similar, the first and second-order corrections will behave similarly. Otherwise, the second order correction almost always performs better.

2.2 Framework of chi-squared tests and pseudo maximum likelihood estimator in dual frame surveys

The set up in this section follows from Hartley (1962) and Lu and Lohr (2010). Assume there are k categories in both surveys and the same quantities are measured. Let p_{id} be the population proportion of category i in domain d (domain d can be domain a, domain ab or domain b), with $\sum_{i=1}^{k} p_{id} = 1$. Let N_a, N_{ab} and N_b denote the population sizes of the three domains respectively, with $N_a + N_{ab} = N_A$ and $N_b + N_{ab} = N_B$. We consider the common case that N_{ab} is unknown, while N_A and N_B are constants. As a result, $\sum_{i=1}^{k} p_{ia} N_a / N_A + \sum_{i=1}^{k} p_{iab} N_{ab} / N_A = 1$ and $\sum_{i=1}^{k} p_{ib} N_b / N_B + \sum_{i=1}^{k} p_{iab} N_{ab} / N_B = 1$ (see Figure 2.1 for illustration of the proportions). The vector of proportions $\mathbf{p} = (p_1, p_2, \cdots p_{k-1})'$ for the union of the two frames is a function of the parameters p_{ia}, p_{iab}, p_{ib} and N_{ab}. For example, a natural form of p_i is

$$p_i = \frac{N_a}{N} p_{ia} + \frac{N_{ab}}{N} p_{iab} + \frac{N_b}{N} p_{ib}, \qquad \text{for} \qquad i = 1, 2, \cdots, k-1, \tag{2.4}$$

where $N = N_A + N_B - N_{ab}$.

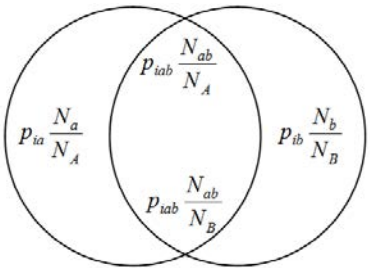

Figure 2.1: Population proportion in domains and frames

In the following, we briefly review the pseudo maximum likelihood estimator that we will use in Section 4 and Section 5. Assume independent simple random samples are taken from frames A and B respectively. The likelihood function is

$$L\left(p_{ia}, p_{iab}, p_{ib}, N_{ab}\right) \propto \prod_i \left(p_{ia}\frac{N_a}{N_A}\right)^{x_{ia}} \times \prod_i \left(p_{iab}\frac{N_{ab}}{N_A}\right)^{x_{iab}^A} \times \prod_i \left(p_{ib}\frac{N_b}{N_B}\right)^{x_{ib}} \times \prod_i \left(p_{iab}\frac{N_{ab}}{N_B}\right)^{x_{iab}^B} \quad (2.5)$$

where x_{ia}, x_{ib} represent the units falling in category i within domain a and domain b respectively; x_{iab}^A and x_{iab}^B represent the units falling in category i within the overlapping domain ab that are originally sampled from frame A and frame B respectively.

For the estimators of complex surveys, the basic idea is to use a working assumption of a multinomial distribution from a finite population to give the form of the estimators and use a design effect to adjust the cell counts to reflect the complex survey design. The pseudo likelihood function is as follows

$$\begin{aligned} L\left(p_{ia}, p_{iab}, p_{ib}, N_{ab}\right) \propto & \prod_i \left(p_{ia}\frac{N_a}{N_A}\right)^{\frac{\tilde{n}_A}{N_A}\hat{X}_{ia}} \prod_i \left(p_{iab}\frac{N_{ab}}{N_A}\right)^{\frac{\tilde{n}_A}{N_A}\hat{X}_{iab}^A} \\ & \times \prod_i \left(p_{ib}\frac{N_b}{N_B}\right)^{\frac{\tilde{n}_B}{N_B}\hat{X}_{ib}} \prod_i \left(p_{iab}\frac{N_{ab}}{N_B}\right)^{\frac{\tilde{n}_B}{N_B}\hat{X}_{iab}^B}, \end{aligned} \quad (2.6)$$

where design effect is defined as $\left\{v\left(\hat{\theta}\right) \text{ from complex survey}\right\} \Big/ \left\{v\left(\hat{\theta}\right) \text{ from SRS of same size}\right\}$, $\tilde{n}_A = n_A \Big/ \left(\text{design effect of } \hat{N}_{ab}^A\right)$, $\tilde{n}_B = n_B \Big/ \left(\text{design effect of } \hat{N}_{ab}^B\right)$, n_A and n_B are the observed sizes of S_A and S_B, and $\hat{X}_{id}$ denote the estimated counts according to the survey design. The pseudo maximum likelihood estimators (PMLEs), found by maximizing (2.6) are $\hat{p}_{ia} = \hat{X}_{ia} \big/ \hat{N}_a$, $\hat{p}_{ib} = \hat{X}_{ib} \big/ \hat{N}_b$, and

$$\hat{p}_{iab} = \frac{\frac{\tilde{n}_A}{N_A}\hat{N}_{ab}^A\hat{p}_{iab}^A + \frac{\tilde{n}_B}{N_B}\hat{N}_{ab}^B\hat{p}_{iab}^B}{\frac{\tilde{n}_A}{N_A}\hat{N}_{ab}^A + \frac{\tilde{n}_B}{N_B}\hat{N}_{ab}^B}, \quad (2.7)$$

where $\hat{p}_{iab}^A = \hat{X}_{iab}^A \big/ \hat{N}_{ab}^A$ and $\hat{p}_{iab}^B = \hat{X}_{iab}^B \big/ \hat{N}_{ab}^B$, and $\hat{N}_{ab,PML}$ is the smaller root of the quadratic function

$$\left[\tilde{n}_A + \tilde{n}_B\right]\hat{N}_{ab,PML}^2 - \left[\tilde{n}_A N_B + \tilde{n}_B N_A + \tilde{n}_A\hat{N}_{ab}^A + \tilde{n}_B\hat{N}_{ab}^B\right]\hat{N}_{ab,PML} + \left[\tilde{n}_A\hat{N}_{ab}^A N_B + \tilde{n}_B\hat{N}_{ab}^B N_A\right] = 0. \quad (2.8)$$

The estimators of the population proportions are

$$\hat{p}_{i,PML} = \frac{\left(N_A - \hat{N}_{ab,PML}\right)\hat{p}_{ia} + \hat{N}_{ab,PML}\hat{p}_{iab} + \left(N_B - \hat{N}_{ab,PML}\right)\hat{p}_{ib}}{N_A + N_B - \hat{N}_{ab,PML}}. \tag{2.9}$$

If SRSs are taken in each frame and $k = 1$, these PMLEs reduced to PMLEs in Skinner and Rao (1996).

3 Chi-squared tests in dual frame surveys

In this section, we consider the case of chi-squared tests in a dual frame survey. Some hypotheses of interest may include: a simple hypothesis $H_0: q_{ia} = p_{ia}N_a/N_A = q^A_{ia0}$, $q^A_{iab} = p_{iab}N_{ab}/N_A = q^A_{iab0}$, $q^B_{iab} = p_{iab}N_{ab}/N_B = q^B_{iab0}$, $q_{ib} = p_{ib}N_b/N_B = p_{ib0}$ (note that q_{ia}, etc., are used to simplify the notations); $H_0: p_{i,PML} = p_{i0,PML}$, in which we test whether the PMLE of proportions from the union of the two frames in (2.9) are some specific values (note that p_i can be estimated by other methods); $H_0: p_{ia} = p_{iab} = p_{ib}$, testing whether the proportions are equal in the three domains; or $H_0: p_{ij} = p_{i+}p_{+j}$, testing independence of the row classification and column classification.

Let $\boldsymbol{\eta} = \left(\mathbf{p}'_a N_a/N_A, \mathbf{p}'_{ab} N_{ab}/N_A, \mathbf{p}'_b N_b/N_B, \mathbf{p}'_{ab} N_{ab}/N_B\right)'$, $\mathbf{p}_a = \left(p_{1a}, p_{2a}, \cdots p_{ka}\right)'$, $\mathbf{p}_b = \left(p_{1b}, p_{2b}, \cdots p_{kb}\right)'$, $\mathbf{p}_{ab} = \left(p_{1ab}, p_{2ab}, \cdots p_{(k-1)ab}\right)'$, and h_i's are continuous functions. A more general hypothesis of interest may be denoted as the following:

$$H_0: h_i(\boldsymbol{\eta}) = 0, \quad i = 1, 2, \cdots r. \tag{3.1}$$

Let η_j be the j-th element of $\boldsymbol{\eta}$ and let $h(\boldsymbol{\eta}) = \left(h_1(\boldsymbol{\eta}), h_2(\boldsymbol{\eta}), \cdots h_r(\boldsymbol{\eta})\right)'$.

Assume that $\partial h_i(\boldsymbol{\eta})/\partial \eta_j$ is continuous in a neighborhood of $\boldsymbol{\eta}$ and that

$$\nabla = \frac{\partial h_i(\boldsymbol{\eta})}{\partial \eta_j} \tag{3.2}$$

has full rank. Also assume

A$_1$. There is a sequence of superpopulations $U_{A1} \subset U_{A2} \subset \ldots \subset U_{At} \subset \ldots$ as defined in Isaki and Fuller (1982).

A$_2$. Let $\tilde{n}_A$ and $\tilde{n}_B$ as defined in Section 2 and assume that $\tilde{n}_A$ and $\tilde{n}_B$ both increase such that $\tilde{n}_A/\tilde{n}_B \to \gamma$ for some $0 < \gamma < 1$.

A$_3$. Let $\pi^A_{it} = p\left(\text{psu } i \text{ is in sample from Frame } A, \text{using population } U_{At}\right)$ and $\pi^A_{ijt} = p\left(\text{psus } i \text{ and } j \text{ are in sample from Frame } A, \text{ using population } U_{At}\right)$ be the inclusion and joint inclusion probabilities for the frame-A sample from population U_{At}, and define π^B_{it}, π^B_{ijt} and U_{Bt} similarly for frame B. Assume there are constants c_1 and c_2 such that

$$0 < c_2 < \pi_{it}^F < c_1 < 1 \tag{3.3}$$

for all i and any superpopulation in the sequence, where F denotes frame A or frame B. Also assume there exists an α_t with $\alpha_t = o(1)$ such that

$$\pi_{it}^F \pi_{jt}^F - \pi_{ijt}^F \le \alpha_t \pi_{it}^F \pi_{jt}^F. \tag{3.4}$$

A$_4$. $N_{ab}/N \to \psi$ for some ψ between 0 and 1.

Theorem 1. *With assumptions $A_1 - A_4$ set out beforehand, we have the following conclusion: $\tilde{n}^{1/2}\mathbf{h}(\hat{\boldsymbol{\eta}})$ is asymptotically normal with mean $\mathbf{0}$ and asymptotic variance $\nabla\Sigma\nabla'$, where Σ is a block-diagonal matrix with blocks Σ_A and Σ_B and $\tilde{n} = \tilde{n}_A + \tilde{n}_B$. Σ_A is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_A$ with $\hat{\boldsymbol{\eta}}_A = \left(\hat{\mathbf{p}}_a' \hat{N}_a/N_A, \hat{\mathbf{p}}_{ab}^{A\prime} \hat{N}_{ab}/N_A\right)'$, Σ_B is the asymptotic covariance matrix of $\tilde{n}^{1/2}\hat{\boldsymbol{\eta}}_B$ with $\hat{\boldsymbol{\eta}}_B = \left(\hat{\mathbf{p}}_b' \hat{N}_b/N_B, \hat{\mathbf{p}}_{ab}^{B\prime} \hat{N}_{ab}/N_B\right)'$ and $\hat{\boldsymbol{\eta}} = \left(\hat{\boldsymbol{\eta}}_A', \hat{\boldsymbol{\eta}}_B'\right)'$.*

Proof. The arguments given in Theorem 1 in Lu and Lohr (2010) show that $\hat{\boldsymbol{\eta}}$ is consistent for $\boldsymbol{\eta}$ and that $\hat{\boldsymbol{\eta}}$ obeys the central limit theorem, as $\tilde{n}_A$ and $\tilde{n}_B$ both increase such that $\tilde{n}_A/\tilde{n}_B \to \gamma$. Thus, since the samples S_A and S_B are selected independently, we have

$$\tilde{n}^{1/2}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) \xrightarrow{d} N(0, \Sigma).$$

$\mathbf{h}(\hat{\boldsymbol{\eta}})$ is consistent for $\mathbf{h}(\boldsymbol{\eta})$ because $\hat{\boldsymbol{\eta}}$ is consistent for $\boldsymbol{\eta}$. Using the delta method, $\tilde{n}^{1/2}\mathbf{h}(\hat{\boldsymbol{\eta}})$ is asymptotically normal with mean $\mathbf{0}$ and asymptotic variance $\nabla\Sigma\nabla'$.

Based on Theorem 1, the following results follow immediately.

Result 1. (Extended Wald Test) If a consistent estimator of the variance Σ is available, by Theorem 1, the generalized Wald statistic can be formed as follows:

$$X_W^2 = \tilde{n}\mathbf{h}(\hat{\boldsymbol{\eta}})' \left(\hat{\nabla}\hat{\Sigma}\hat{\nabla}'\right)^{-1} \mathbf{h}(\hat{\boldsymbol{\eta}}). \tag{3.5}$$

This test statistic is distributed asymptotically as $\chi^2(r)$ under H_0 (refer to equation 3.1), where r is the rank of ∇.

As we have noted previously, the estimate of the variance may be unstable or no closed-form estimate of Σ is available. One way we can modify the statistic in (3.5) is to first act as though the sample is a simple random sample, then modify the reference distribution used in the test to get the correct level. Equation (3.6) gives the modified statistic.

Result 2. Let

$$X_{MW}^2 = \tilde{n}\mathbf{h}(\hat{\boldsymbol{\eta}})' \left(\hat{\nabla}_0\hat{\mathbf{P}}_0\hat{\nabla}_0'\right)^{-1} \mathbf{h}(\hat{\boldsymbol{\eta}}), \tag{3.6}$$

where $\hat{\nabla}_0\hat{\mathbf{P}}_0\hat{\nabla}_0'$ can be any estimate of $\nabla\mathbf{P}\nabla$ that is consistent when H_0 is true. Matrix $\mathbf{P}_0$ is a block diagonal matrix with diagonal blocks: covariance matrix from frame A and covariance matrix from frame

B when H_0 is true and when sampling is SRS. Suppose the matrix ∇ has rank r under the null hypothesis $H_0 : \mathbf{h}(\boldsymbol{\eta}) = 0$. Then $X^2_{MW} \approx \sum_1^r \lambda_{0i} W_i$, where the λ_i's are the eigenvalues of $(\nabla \mathbf{P} \nabla')^{-1}(\nabla \Sigma \nabla')$, $W_1, \ldots, W_r$ are independent χ^2_1 random variables and λ_{0i} is the value of λ_i under H_0.

Result 3. (Extended Rao-Scott first order correction) Suppose matrix ∇ has rank r. Let X^2_{MW} be as defined in (3.6). Under the null hypothesis $H_0 : \mathbf{h}(\boldsymbol{\eta}) = 0$, the statistic $X^2_{MW} / \hat{\lambda}.$ has expectation r, where $\hat{\lambda}. = \sum \hat{\lambda}_i / r$, $\hat{\lambda}_i$ is a consistent estimate of λ_i under H_0. For example, $\hat{\lambda}_i$'s could be the eigenvalues of $\left(\hat{\nabla}_0 \hat{\mathbf{P}}_0 \hat{\nabla}_0'\right)^{-1}\left(\hat{\nabla} \hat{\Sigma} \hat{\nabla}'\right)$.

Result 4. (Extended Rao-Scott second order correction) Suppose matrix ∇ has rank r. Define

$$X^2_S = \frac{X^2_{MW}}{\hat{\lambda}.\left(1 + \hat{a}^2\right)}$$

where $\hat{a}^2 = \sum_{i=1}^{k-1} \hat{\lambda}_i^2 \Big/ \left[(k-1)\hat{\lambda}.^2\right] - 1$ is an estimate of the population value a^2. Under null hypothesis, X^2_S is distributed asymptotically as χ^2_v, a chi-square random variable with degrees of freedom $v = (k-1)/\left(1 + a^2\right)$.

4 Simulations

In this section, a small simulation has been conducted to study the proposed chi-squared tests under a simple hypothesis $H_0 : q_{ia} = p_{ia} N_a / N_A = q^A_{ia0}$, $q^A_{iab} = p_{iab} N_{ab} / N_A = q^A_{iab0}$, $q^B_{iab} = p_{iab} N_{ab} / N_B = q^B_{iab0}$, $q_{ib} = p_{ib} N_b / N_B = p_{ib0}$ to investigate the performance of chi-squared tests proposed in Section 3. We compare the percentages of samples for which the test statistics exceed the critical value to the nominal level $(\alpha = 0.05)$. R (www.r-project.org) is used to perform simulation study and other analysis.

We generated the data following Skinner and Rao (1996), with $\gamma_a = N_a / N$ and $\gamma_b = N_b / N$. A cluster sample from frame A was generated with n_p psus and m observations in each psu, and a simple random sample of n_B observations was generated for frame B. We generated the clustered binary responses for the sample from frame A by generating correlated multivariate normal random vectors and then using the probit function to convert the continuous responses to binary responses. After the sample was generated, we calculated the PML estimators of $\mathbf{p}_{id} N_d / N_A$ and $\mathbf{p}_{id} N_d / N_B$ (see Section 2.2). These estimated proportions were used to compute the chi-squared test statistics. We then compared the percentages of samples for which the test statistics exceed the critical value to the nominal level under different settings.

The simulation study was performed with factors: (1) γ_a : 0.4, (2) γ_b : 0.2, (3) clustering parameter ρ : 0.3, (4) sample sizes: n_p : 10, 30 or 50; m : 3, 5, or 10, n_B : 100, 300 or 500. (5) Simulation runs: 1,000 times for each setting and 100 times when estimating the variance covariance matrix V using bootstrapping. All runs used probability parameters $\mathbf{p}_a$: $(.3,.1,.2,.4)$, $\mathbf{p}_{ab}$: $(.3,.1,.1,.5)$, and

$\mathbf{p}_b$: $(.4,.1,.1,.4)$. Table 4.1 reported the percentages of samples for which the test statistics exceed the critical value.

Table 4.1
Comparison of the actual significance levels (%) among different tests. X^2 is the uncorrected test; X^2_{FC} is the first order corrected X^2 and X^2_{SC} is the second order corrected X^2.

$\tilde{n}_p$	m	n_B	X^2	**Wald**	X^2_{FC}	X^2_{SC}
10	3	100	12.1	17.3	5.6	4.9
30	3	300	13.6	8.4	4.8	4.8
50	3	500	15.5	10.0	6.4	3.6
10	5	100	25.7	13.5	7.5	4.9
30	5	300	29.2	9.3	7.9	5.3
50	5	500	31.5	8.5	8.1	4.9
10	10	100	46.1	21.2	6.6	5.4
30	10	300	50.2	11.5	7.5	5.6
50	10	500	58.7	8.0	9.6	5.1

Table 4.1 indicates that naively using uncorrected X^2 test for complex survey data is dangerous. With increased psu size and number of psu's, the actual significance level even reaches 62.2%. Extended Wald test doesn't perform well since the estimate of the variance may be unstable. Extended first order corrected test is acceptable with actual significance level around 7%. Extended second order corrected tests almost reach the nominal level 5%, for which is the one we recommend to use in a dual frame survey categorical data analysis.

5 Application

In this section, we give a real example to illustrate how to perform the chi-squared tests in a dual frame survey. We consider the hypothesis test $H_0: p_{ia} = p_{iab} = p_{ib}$, testing whether the proportions are equal in the three domains.

5.1 Data description and related PMLEs

Data (Lohr and Rao 2006) were originally collected for a three-frame survey of statisticians, using the online membership directories of the American Statistical Association (ASA), the Institute for Mathematical Statistics (IMS) and the Statistical Society of Canada. We treat the union of online membership directories of ASA and online membership directories of IMS as a dual frame with notation $A \cup B = a \cup ab \cup b$ (A: online membership directories of ASA; B: online membership directories of IMS; domain a: ASA member but not IMS member; domain ab: ASA member and also IMS member; domain b: IMS member but not ASA member). Note that the union of these two frames does not cover the entire population of statisticians. Many statisticians do not belong to either of the two societies, and some statisticians decline to participate in online directories. In the data set, the information of the occupation is a categorical variable with three levels: academia, industry and government. We combine

industry and government to be one level named nonacademia. Together with sex, we have a 2×2 table with four cells: female in academia, female not in academia, male in academia and male not in academia.

At the time of data collection, there were $15,500$ people in American Statistical Association (Frame A) and $4,000$ people in Institute for Mathematical Statistics (Frame B), so $N_A = 15,500$ and $N_B = 4,000.$ A stratified cluster sample of size 500 was taken from frame A, of which 378 observations had information on both responses (sex and occupation). The design had 26 strata constructed by regions or states. Because of the restrictions on access to records, clusters for large states were members whose last name began with the same letter of the alphabet. There are 173 psu's in frame A. A simple random sample of size 140 was taken from frame B, in which 102 records have valid information for both responses. The weighted total of observations from frame A is $10,976.$ We assume that data are missing randomly, so the nonresponse is adjusted by a fraction of $15,500/10,976.$ Table 5.1 lists the number of statisticians falling in each cell within each domain.

Table 5.1
Observed data in domain a and domain ab from frame A (adjusted by a fraction of 15,500/10,976) together with observed data in domain b and domain ab from frame B.

	Domain a		Domain $ab \in A$		Domain b		Domain $ab \in B$	
	Female	Male	Female	Male	Female	Male	Female	Male
Academia	2,425	4,969	302	1,488	10	41	10	33
Nonacademia	1,959	4,091	59	209	0	3	2	3

The estimated design effect of frame A is 1.801209, so the effective sample size of n_A is $\tilde{n}_A = 378/1.8 = 210.$ The effective sample size of $n_{B,eff} = n_B = 102.$ The PMLEs of the estimated proportions by using (2.6) and (2.9) are listed in Table 5.2.

Table 5.2
Estimated proportions from domains and union of two frames.

	Domain a		Domain ab		Domain b		Frame $A \cup B$	
	Female	Male	Female	Male	Female	Male	Female	Male
Academia	0.180	0.370	0.186	0.701	0.185	0.759	0.182	0.452
Nonacademia	0.146	0.304	0.037	0.076	0	0.056	0.116	0.250

5.2 Test the equivalence of proportions across domains

The hypothesis of interest is whether the proportions are equal across the three domains,

$$H_0 : p_{ia} = p_{iab} \quad \text{and} \quad p_{iab} = p_{ib},\ i = 1,2,3. \tag{5.1}$$

In this example, $p_{ia},\ i = 1,2,3,4$ represent the proportion of female in academia, female not in academia, male in academia and male not in academia among ASA members respectively. Similarly define p_{iab} and $p_{ib}.$ $\boldsymbol{\eta}$ (see Section 3) reduces to a 14×1 vector

$$\boldsymbol{\eta} = (p_{1a} N_a/N_A, p_{2a} N_a/N_A, p_{3a} N_a/N_A, p_{4a} N_a/N_A, p_{1ab} N_{ab}/N_A, p_{2ab} N_{ab}/N_A, p_{3ab} N_{ab}/N_A,$$
$$p_{1b} N_b/N_B, p_{2b} N_b/N_B, p_{3b} N_b/N_B, p_{4b} N_b/N_B, p_{1ab} N_{ab}/N_B, p_{2ab} N_{ab}/N_B, p_{3ab} N_{ab}/N_B)'.$$

Since H_0 in (5.1) only involves the simple parameters p_{ia}, p_{iab}, p_{ib} and N_{ab}, we introduce a new vector $\theta = \left(p_{1a}, p_{2a}, p_{3a}, p_{1ab}, p_{2ab}, p_{3ab}, N_{ab}/N_A, p_{1b}, p_{2b}, p_{3b}, p_{1ab}, p_{2ab}, p_{3ab}, N_{ab}/N_B\right)'$.

Let $\Omega = \left(\partial h_i(\boldsymbol{\eta})/\partial\theta_j\right)$ and $\mathbf{D}(\theta) = \left(\partial\boldsymbol{\eta}/\partial\theta_j\right)$. $\mathbf{D}(\theta)$ is found to be a block diagonal matrix with

$$\mathbf{D}_A = \begin{pmatrix}
\frac{N_a}{N_A} & 0 & 0 & 0 & 0 & 0 & -p_{1a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{N_a}{N_A} & 0 & 0 & 0 & 0 & -p_{2a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{N_a}{N_A} & 0 & 0 & 0 & -p_{3a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{N_a}{N_A} & -\frac{N_a}{N_A} & -\frac{N_a}{N_A} & 0 & 0 & 0 & -p_{4a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{N_{ab}}{N_A} & 0 & 0 & p_{1ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{N_{ab}}{N_A} & 0 & p_{2ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_A} & p_{3ab} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & 0 & 0 & -p_{1b} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & 0 & -p_{2b} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_b}{N_B} & 0 & 0 & 0 & -p_{3b} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{N_b}{N_B} & -\frac{N_b}{N_B} & -\frac{N_b}{N_B} & 0 & 0 & 0 & -p_{4b} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & 0 & 0 & p_{1ab} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & 0 & p_{2ab} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{N_{ab}}{N_B} & p_{3ab}
\end{pmatrix}.$$

Notice the relationship between $\hat{p}_{iab}$ and $\hat{p}^A_{iab}$ and $\hat{p}^B_{iab}$ from (2.7), Ω is found to be

$$\Omega = \begin{pmatrix} 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 & 0 & 0 \\ 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & 1-\phi & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & (1-\phi) & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -\phi & 0 & 0 & 0 & 0 & 0 & 0 & -(1-\phi) & 0 \\ 0 & 0 & 0 & 0 & 0 & \phi & 0 & 0 & 0 & -1 & 0 & 0 & (1-\phi) & 0 \end{pmatrix},$$

where $\phi = N_B \tilde{n}_A / \left(N_B \tilde{n}_A + N_A \tilde{n}_B \right)$. It is easy to show that $\nabla = \Omega(\mathbf{D})^{-1}$ (recall that $\nabla = \partial h_i(\boldsymbol{\eta}) / \partial \eta_j$). $\hat{\Sigma}$ is estimated by using a jackknife method by deleting one psu each time from frame A. All the results in Section 3 can be derived. The eigenvalues of $\left(\hat{\nabla}_0 \hat{\mathbf{P}}_0 \hat{\nabla}_0' \right)^{-1} \left(\nabla \Sigma \nabla' \right)$ are very close to each other, which indicates that first order corrected test perform similarly as second order corrected test. The Wald statistic, first order corrected statistic and second order corrected statistic give values of 81.48295, 72.31026 and 70.28581 respectively. Comparing to the critical value with six degrees of freedom $\chi^2(6) = 12.95$, we reject the null hypothesis that the cell proportions (female in academia, female not in academia, male in academia and male not in academia) are the same across the three domains (ASA member only, ASA and IMS member and IMS member only).

6 Conclusions

In this research, we extend Wald's (1943) test and Rao-Scott first-order and second-order corrected tests (Rao and Scott 1981) from a single survey to a dual frame survey and derive the asymptotic distributions. A limited simulation study suggests that second order corrected tests almost reach the nominal level. Although the results in this paper are for dual frame surveys, the methods are general and could be extended to more than two surveys. Our research is done in the context of survey sampling; it also applies to other settings in which data could be combined from two independent sources.

Acknowledgements

The author thanks Dr. Sharon Lohr for her valuable advisement and comments on the manuscript. The author also wants to thank the referees and the associate editor for their very helpful comments and constructive suggestions.

References

Bedrick, E.J. (1983). Adjusted chi-squared tests for cross-classified tables of survey data. *Biometrika*, 70, 591-595.

Fay, R.E. (1979). On adjusting the Pearson chi-square statistic for clustered sampling. In *ASA Proceedings of the Social Statistics Section*, 402-406. American Statistical Association.

Fay, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

Fuller, W.A. and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. In *ASA Proceedings of the Social Statistics Section*, 245-249. American Statistical Association.

Hartley, H.O. (1962). Multiple frame surveys. In *ASA Proceedings of the Social Statistics Section*, 203-206. American Statistical Association.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā, Series C*, 36 (3), 99-118.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Lohr, S.L. and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S.L. and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Lu, Y. and Lohr, S. (2010). Gross flow estimation in dual frame surveys. *Survey Methodology*, vol. 36, 13-22.

Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K. and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.

Rao, J.N.K. and Scott, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, 15, 385-397.

Rao, J.N.K. and Thomas, D.R. (1988). The analysis of cross-classified categorical data from complex sample surveys. *Sociological Methodology*, 18, 213-269.

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. and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Thomas, D.R., Singh, A. and Roberts, G. (1996). Tests of independence on two-way tables under cluster sampling: An evaluation. *International Statistical Review*, 64(3), 295-311.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, 426-482.

Survey Methodology, December 2014
Vol. 40, No. 2, p. 335-346
Statistics Canada, Catalogue No. 12-001-X

Estimation methods on multiple sampling frames in two-stage sampling designs

Guillaume Chauvet and Guylène Tandeau de Marsac[1]

Abstract

When studying a finite population, it is sometimes necessary to select samples from several sampling frames in order to represent all individuals. Here we are interested in the scenario where two samples are selected using a two-stage design, with common first-stage selection. We apply the Hartley (1962), Bankier (1986) and Kalton and Anderson (1986) methods, and we show that these methods can be applied conditional on first-stage selection. We also compare the performance of several estimators as part of a simulation study. Our results suggest that the estimator should be chosen carefully when there are multiple sampling frames, and that a simple estimator is sometimes preferable, even if it uses only part of the information collected.

Key Words: Expansion survey; Hansen-Hurwitz estimator; Horvitz-Thompson estimator; Two-stage sampling.

1 Introduction

When studying a finite population, sometimes no sampling frame covers that population completely, and it is necessary to select samples from two or more sampling frames in order to represent all individuals. Many methods of estimation on multiple sampling frames have been proposed to pool these samples (Hartley 1962; Bankier 1986; Kalton and Anderson 1986; Mecatti 2007; Rao and Wu 2010); see also the review articles by Lohr (2009, 2011) and the referenced articles for a complete picture. Note that the Mecatti method (2007) is inspired by the work of Lavallée (2002, 2007) on the Generalized Weight Share Method. In Section 2, we present different estimation methods for multiple sampling frames.

In Section 3, we are interested in the scenario where two samples are selected using a two-stage design, with common first-stage selection. This framework corresponds to INSEE expansion surveys: an initial sample of dwellings is selected from the communes of the master sample (Bourdalle, Christine and Wilms 2000), and a second sample is selected and surveyed from the communes of the same master sample to target a specific subpopulation. We have two survey measurements from two independent samples at the second stage of the design. We apply estimation methods to multiple sampling frames to pool these two samples. We show that the estimators examined can in this case be calculated conditional on the first stage of selection, which simplifies calculation particularly for Hartley's optimal estimator (1962). In Section 4, we compare the performance of these estimators as part of a simulation study. We present our conclusion in Section 5.

1. Guillaume Chauvet, ENSAI (CREST), Ker Lann Campus, Bruz, France. Email: chauvet@ensai.fr. Guylène Tandeau de Marsac, INSEE, Regional Direction of Lille, France. Email: guylene.tandeau-de-marsac@insee.fr.

2 Estimation for multiple sampling frames

A finite population U upon which is defined a variable of interest y of value y_k for individual k is considered. If a sample S is selected from U with inclusion probabilities π_k, the estimator $\hat{Y} = \sum_{k \in S} \pi_k^{-1} y_k$ proposed by Narain (1951) and Horvitz and Thompson (1952) is unbiased for total $Y = \sum_{k \in U} y_k$ if all probabilities π_k are strictly positive.

We are interested in the scenario where the population is fully covered by two overlapping sampling frames, U_A and U_B. We used Lohr's (2011) notation, namely $a = U_A \setminus U_B$ the domain covered by U_A only; $b = U_B \setminus U_A$ the domain covered by U_B only; $ab = U_A \cap U_B$ the domain covered both by U_A and U_B. A sample S^A is selected in U_A with inclusion probabilities $\pi_k^A > 0$. For any domain $d \subset U_A$, the sub-total $Y_d = \sum_{k \in d} y_k$ is unbiasedly estimated by $\hat{Y}_d^A = \sum_{k \in S_A} d_k^A y_k 1(k \in d)$ with $d_k^A = \left(\pi_k^A\right)^{-1}$. A sample U_B is selected in S^B with inclusion probabilities $\pi_k^B > 0$. For any domain $d \subset U_B$, the sub-total Y_d is unbiasedly estimated by $\hat{Y}_d^B = \sum_{k \in S_B} d_k^B y_k 1(k \in d)$ with $d_k^B = \left(\pi_k^B\right)^{-1}$. The objective is to combine the samples S^A and S^B to get estimation Y as accurate as possible.

2.1 Hartley estimator

Hartley (1962) proposes the class of unbiased estimators

$$\hat{Y}_\theta = \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B, \tag{2.1}$$

with θ one parameter to be determined. The choice $\theta = 1/2$ gives samples S^A and S^B the same weight for the estimation on the intersection domain ab. Hartley (1962) proposes choosing the parameter that minimizes the variance of $\hat{Y}_\theta$. This leads to

$$\theta_{opt} = \frac{Cov\left(\hat{Y}_a^A + \hat{Y}_{ab}^B + \hat{Y}_b^B, \hat{Y}_{ab}^B - \hat{Y}_{ab}^A\right)}{V\left(\hat{Y}_{ab}^B - \hat{Y}_{ab}^A\right)}, \tag{2.2}$$

which can be re-expressed as

$$\theta_{opt} = \frac{V\left(\hat{Y}_{ab}^B\right) + Cov\left(\hat{Y}_{ab}^B, \hat{Y}_b^B\right) - Cov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A\right)}{V\left(\hat{Y}_{ab}^A\right) + V\left(\hat{Y}_{ab}^B\right)} \tag{2.3}$$

when the samples S^A and S^B are independent. As noted by Lohr (2007), the optimal coefficient θ_{opt} may not be between 0 and 1 if a covariance term present in (2.3) is large. To simplify, let us assume that $Cov\left(\hat{Y}_{ab}^B, \hat{Y}_b^B\right) = 0$, which is the case if b and ab are used as strata in the selection of S^B. Then $\theta_{opt} > 1$ if and only if $Cov\left(\hat{Y}^A, \hat{Y}_{ab}^A\right) < 0$. When S^A is selected by simple random sampling, this will be the case, for example, if in U_A the low values of the variable y are concentrated in the domain ab.

In practice, the variance and covariance terms are unknown and must be replaced by estimators, which introduces additional variability. Another disadvantage is that the optimal parameter depends on the

variable of interest considered. If optimal estimators are calculated for different variables of interest, estimations may be internally inconsistent (Lohr 2011).

2.2 Kalton and Anderson estimator

A more general class of estimators is obtained by noting that total Y can be re-expressed as

$$Y = Y_a + \sum_{k\in ab}\theta_k y_k + \sum_{k\in ab}\left(1-\theta_k\right)y_k + Y_b,$$

with θ_k a coefficient specific to the individual k. Kalton and Anderson (1986) propose the choice $\theta_k = \left(d_k^A + d_k^B\right)^{-1} d_k^B$, which leads to the estimator

$$\hat{Y}_{KA} = \sum_{k\in S^A} d_k^A m_k^A y_k + \sum_{k\in S^B} d_k^B m_k^B y_k \tag{2.4}$$

with on one hand $m_k^A = 1$ if $k \in a$ and $m_k^A = \theta_k$ if $k \in ab$, and on the other hand $m_k^B = 1$ if $k \in b$ and $m_k^B = 1-\theta_k$ if $k \in ab$. The estimation weights are the same regardless of the variable of interest, which guarantees internal consistency of the estimations; on the other hand, the Kalton and Anderson estimator is less effective than Hartley's optimal estimator for a given variable of interest. Note that it is a Hansen-Hurwitz (1943) type estimator, which can be re-expressed as $\hat{Y}_{KA} = \sum_{k\in U}\left[W_k / E\left(W_k\right)\right]y_k$ noting $W_k = 1\left(k \in S^A\right) + 1\left(k \in S^B\right)$ the number of times when unit k is selected in the pooled sample $S^A \cup S^B$. In particular this gives $E(W_k) = \pi_k^A + \pi_k^B$.

2.3 Bankier estimator

Bankier (1986) proposes using a Horvitz-Thompson type estimator, calculating the inclusion probabilities in the pooled sample.

$$\pi_k^{HT} \equiv P\left(k \in S^A \cup S^B\right) = \pi_k^A + \pi_k^B - Pr\left(k \in S^A \cap S^B\right).$$

If the samples S^A and S^B are independent, we get $\pi_k^{HT} = \pi_k^A + \pi_k^B - \pi_k^A\pi_k^B$ and the estimator

$$\hat{Y}_{HT} = \sum_{k\in S^A\cup S^B}\frac{y_k}{\pi_k^{HT}} = \sum_{k\in S^A\cap a}\frac{y_k}{\pi_k^A} + \sum_{k\in S^B\cap b}\frac{y_k}{\pi_k^B} + \sum_{k\in\left(S^A\cup S^B\right)\cap ab}\frac{1}{\pi_k^A + \pi_k^B - \pi_k^A\pi_k^B}y_k. \tag{2.5}$$

3 Estimation with common first-stage selection

Here we are interested in the case of two samples selected using a two-stage design, with common first-stage selection. Population U is partitioned to obtain a population $U_I = \{u_1,\ldots,u_M\}$ of M primary sampling units. In the first stage, a sample S_I of primary sampling units (PSU) is selected, with a selection probability π_{Ii} for a PSU u_i. In the second stage, in each primary sampling unit $u_i \in S_I$, the following is selected: a sample S_i^A in $u_i^A \equiv u_i \cap U_A$, with a (conditional) selection probability $\pi_{k|i}^A > 0$ for

$k \in u_i^A$; a sample S_i^B in $u_i^B \equiv u_i \cap U_B$, with a (conditional) selection probability $\pi_{k|i}^B > 0$ for unit $k \in u_i^B$. We make the following hypotheses, which are common for two-stage selection: the second stage of selection in the primary sampling unit u_i depends only on i; between two primary sampling units $u_i \neq u_j \in S_I$, the samples S_i^A and S_j^A (respectively, S_i^B and S_j^B) are conditionally independent to S_I (property of independence). We also assume that within each primary sampling unit $u_i \in S_I$, the sub-samples S_i^A and S_i^B are conditionally independent to S_I.

For a domain $d_1 \subset U_A$, the sub-total Y_{d_1} is estimated by $\hat{Y}_{d_1}^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{d_1,i}^A$ with $d_{Ii} = \left(\pi_{Ii}\right)^{-1}$ the sampling weight of the primary sampling unit u_i, $\hat{Y}_{d_1,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1\left(k \in d_1\right)$ the estimator of the sub-total $Y_{d_1,i} = \sum_{k \in u_i} y_k 1\left(k \in d_1\right)$ over $d_1 \cap u_i$, and $d_{k|i}^A = \left(\pi_{k|i}^A\right)^{-1}$ the sampling weight of k in u_i^A. For a domain $d_2 \subset U_B$, the sub-total Y_{d_2} is estimated by $\hat{Y}_{d_2}^B = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{d_2,i}^B$ with $\hat{Y}_{d_2,i}^B = \sum_{k \in S_i^B} d_{k|i}^B y_k 1\left(k \in d_2\right)$ the estimator of the sub-total $Y_{d_2,i}$ and $d_{k|i}^B = \left(\pi_{k|i}^B\right)^{-1}$ the sampling weight of k in u_i^B. This yields in particular the estimators

$$\hat{Y}_{ab}^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{ab,i}^A \text{ where } \hat{Y}_{ab,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1\left(k \in ab\right), \tag{3.1}$$

$$\hat{Y}_b^A = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{b,i}^A \text{ where } \hat{Y}_{b,i}^A = \sum_{k \in S_i^A} d_{k|i}^A y_k 1\left(k \in b\right), \tag{3.2}$$

$$\hat{Y}_{ab}^B = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{ab,i}^B \text{ where } \hat{Y}_{ab,i}^B = \sum_{k \in S_i^B} d_{k|i}^B y_k 1\left(k \in ab\right). \tag{3.3}$$

3.1 Hartley estimator

The Hartley estimator given in (2.1) may be re-expressed as

$$\hat{Y}_\theta = \sum_{u_i \in S_I} d_{Ii} \hat{Y}_{\theta,i} \tag{3.4}$$

with $\hat{Y}_{\theta,i} = \hat{Y}_{a,i}^A + \theta \hat{Y}_{ab,i}^A + \left(1-\theta\right) \hat{Y}_{ab,i}^B + \hat{Y}_{b,i}^B$ the Hartley estimator of sub-total Y_i over unit primary sampling unit u_i. We get $E\left(\hat{Y}_\theta \mid S_I\right) = \sum_{i \in S_I} d_{Ii} Y_i$, then

$$V\left(\hat{Y}_\theta\right) = V\left(\sum_{i \in S_I} d_{Ii} Y_i\right) + EV\left(\hat{Y}_\theta \mid S_I\right). \tag{3.5}$$

In (3.5), the first term of the right member does not depend on θ. Hartley's optimal estimator can, therefore, be calculated by minimizing the second term only. This gives:

$$\theta_{opt|S_I} = \frac{EV\left(\hat{Y}_{ab}^B \mid S_I\right) + ECov\left(\hat{Y}_{ab}^B, \hat{Y}_b^B \mid S_I\right) - ECov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I\right)}{EV\left(\hat{Y}_{ab}^A \mid S_I\right) + EV\left(\hat{Y}_{ab}^B \mid S_I\right)}, \tag{3.6}$$

which can be estimated by

$$\hat{\theta}_{opt} = \frac{\hat{V}\left(\hat{Y}_{ab}^{B}\right) + \widehat{Cov}\left(\hat{Y}_{ab}^{B}, \hat{Y}_{b}^{B}\right) - \widehat{Cov}\left(\hat{Y}_{a}^{A}, \hat{Y}_{ab}^{A}\right)}{\hat{V}\left(\hat{Y}_{ab}^{A}\right) + \hat{V}\left(\hat{Y}_{ab}^{B}\right)} \tag{3.7}$$

by replacing each variance and covariance term with an unbiased estimator conditional on the first stage.

3.2 Kalton and Anderson estimator

With the sample design considered, we get $d_k^A = d_{Ii} d_{k|i}^A$ for any unit $k \in u_i^A$, and $d_k^B = d_{Ii} d_{k|i}^B$ for any unit $k \in u_i^B$. Therefore, the Kalton and Anderson estimator given in (2.4) can be re-expressed as

$$\hat{Y}_{KA} = \sum_{i \in S_I} d_{Ii} \hat{Y}_{KA,i} \tag{3.8}$$

with $\hat{Y}_{KA,i} = \sum_{k \in S^A} d_{k|i}^A m_{k|i}^A y_k + \sum_{k \in S^B} d_{k|i}^B m_{k|i}^B y_k$ the Kalton and Anderson estimator of the sub-total Y_i, where

$$m_{k|i}^A = \begin{cases} 1 & \text{if } k \in a \cap u_i, \\ \dfrac{d_{k|i}^B}{d_{k|i}^A + d_{k|i}^B} & \text{if } k \in ab \cap u_i, \end{cases} \quad \text{and} \quad m_{k|i}^B = \begin{cases} 1 & \text{if } k \in b \cap u_i, \\ \dfrac{d_{k|i}^A}{d_{k|i}^A + d_{k|i}^B} & \text{if } k \in ab \cap u_i. \end{cases}$$

3.3 Bankier estimator

With the sampling design considered, we get $\pi_k^{HT} = \pi_{Ii}\left(\pi_{k|i}^A + \pi_{k|i}^B - \pi_{k|i}^A \pi_{k|i}^B\right)$ for any $k \in u_i$. Therefore, the Bankier estimator given in (2.5) can be re-expressed as

$$\hat{Y}_{HT} = \sum_{i \in S_I} d_{Ii} \hat{Y}_{HT,i} \tag{3.9}$$

with $\hat{Y}_{HT,i} = \sum_{k \in S_i^A \cup S_i^B} \left(y_k / \pi_{k|i}^{HT}\right)$ the Bankier estimator for the sub-total Y_i, and $\pi_{k|i}^{HT} = \pi_{k|i}^A$ if $k \in a$, $\pi_{k|i}^{HT} = \pi_{k|i}^B$ if $k \in b$, $\pi_{k|i}^{HT} = \pi_{k|i}^A + \pi_{k|i}^B - \pi_{k|i}^A \pi_{k|i}^B$ if $k \in ab$.

Each of the three estimators examined is obtained by applying the estimation method PSU by PSU, conditional on the first stage. This result is particularly attractive for Hartley's optimal method, since the optimal coefficient estimator given in (3.7) only requires variance estimators conditional on the first stage.

4 Simulation study

We are using artificial populations proposed by Saigo (2010). We generate two populations, each containing $M = 200$ primary sampling units grouped in $H = 4$ strata U_{Ih} of size $M_h = 50$. Each primary sampling unit u_{hi} contains $N_{hi} = 100$ secondary units. In each population, we generate for each primary sampling unit $u_{hi} \in U_{Ih}$:

$$\mu_{hi} = \mu_h + \sigma_h v_{hi} \tag{4.1}$$

where the values μ_h and σ_h are those used by Saigo (2010). The term σ_h^2 makes it possible to control dispersion between the primary sampling units. The v_{hi} are iid, generated according to a standard normal distribution $N(0,1)$. For each unit $k \in u_{hi}$, we then generate the value y_k according to the model

$$y_k = \mu_{hi} + \left\{\rho^{-1}(1-\rho)\right\}^{0.5} \sigma_h v_k, \tag{4.2}$$

where the v_k are iid, generated according to standard normal distribution. The variance term in the model (4.2) can give an intra-cluster correlation coefficient approximately equal to ρ. In particular, the larger the ρ coefficient, the less the values y_k are dispersed in the primary sampling units. We use $\rho = 0.2$ for population 1 and $\rho = 0.5$ for population 2, which reflects less dispersion of the variable y in population 2. The sampling frame U_A corresponds to all secondary units, and the corresponding part of u_{hi} is $u_{hi}^A = u_{hi}$, of size $N_{hi}^A = N_{hi}$. For each secondary unit k, a value u_k is generated according to uniform distribution over $[0,1]$. The sampling frame U_B corresponds to the secondary units k such that $u_k \leq 0.5$, and the corresponding part of u_{hi} is $u_{hi}^B = u_{hi} \cap U_B$ of size N_{hi}^B. This gives, therefore, the situation where $ab = U_B$ and $b = \varnothing$. The framework selected in the simulations is the one used in the INSEE household surveys, with expansion to target a specific sub-population. For these surveys, a sample S_I of communes (or groups of communes) is first selected in the first stage. A sub-sample S_i^A of dwellings is then selected in each $u_i \in S_I$; the pooled sample $S^A = \bigcup_{u_i \in S_I} S_i^A$ represents the entire population of dwellings $U_A = U$. A second sub-sample S_i^B of dwellings is then selected from within a sub-population of each $u_i \in S_I$, in order to target a specific sub-population U_B (for example, dwellings located in a Sensitive Urban Area); the pooled sample $S^B = \bigcup_{u_i \in S_I} S_i^B$ represents only the targeted sub-population U_B.

In each of the two populations created, several samplings are taken concurrently; Table 4.1 presents for each population the eight possible combinations of sample sizes per stratum in the first and second stage, as well as the values μ_h and σ_h. In the first stage, we select independently in each stratum U_{Ih}: either a sample S_{Ih} of $m_h = 5$ primary sampling units by simple random sampling; or a sample S_{Ih} of $m_h = 25$ primary sampling units by simple random sampling. In the second stage, we select in each $u_{hi} \in S_{Ih}$: either a sample S_{hi}^A of size $n_{hi}^A = 10$ by simple random sampling in u_{hi}^A; or a sample S_{hi}^A of size $n_{hi}^A = 40$ by simple random sampling in u_{hi}^A. In the second stage, we also select in each $u_{hi} \in S_{Ih}$: either a sample S_{hi}^B of size $n_{hi}^B = 5$ by simple random sampling in u_{hi}^B; or a sample S_{hi}^B of size $n_{hi}^B = 20$ by simple random sampling in u_{hi}^B. Also we note $f_{hi}^A = \left(N_{hi}^A\right)^{-1} n_{hi}^A$ and $f_{hi}^B = \left(N_{hi}^B\right)^{-1} n_{hi}^B$ the sampling rates in u_{hi}^A and u_{hi}^B.

Table 4.1
Parameters used in each stratum to generate both populations and select samples

	Sample Sizes per Stratum			Parameters							
				Stratum 1		Stratum 2		Stratum 3		Stratum 4	
	m_h	n_{hi}^A	n_{hi}^B	μ_h	σ_h	μ_h	σ_h	μ_h	σ_h	μ_h	σ_h
Population 1	5 or 25	10 or 40	5 or 20	200	20	150	15	120	12	100	10
Population 2	5 or 25	10 or 40	5 or 20	200	10	150	7.5	120	6	100	5

For each sample, Hartley's estimator given in (3.4) is calculated with either $\theta = 1/2$ (HART1), or for value of θ the optimal coefficient estimator given in (3.7) (HART2), with

$$\hat{V}\left(\hat{Y}_{ab}^{A}\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} \left(N_{hi}^{A}\right)^2 \frac{1-f_{hi}^{A}}{n_{hi}^{A}\left(n_{hi}^{A}-1\right)} \sum_{k\in S_{hi}^{A}} \left\{y_k 1\left(k\in ab\right) - \overline{y}_{ab;S_{hi}^{A}}\right\}^2,$$

$$\hat{V}\left(\hat{Y}_{ab}^{B}\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} \left(N_{hi}^{B}\right)^2 \frac{1-f_{hi}^{B}}{n_{hi}^{B}\left(n_{hi}^{B}-1\right)} \sum_{k\in S_{hi}^{B}} \left\{y_k 1\left(k\in ab\right) - \overline{y}_{ab;S_{hi}^{B}}\right\}^2,$$

$$\widehat{Cov}\left(\hat{Y}_{a}^{A},\hat{Y}_{ab}^{A}\right) = \sum_{h=1}^{H}\left(\frac{M_h}{m_h}\right)^2 \sum_{u_{hi}\in S_{Ih}} \left(N_{hi}^{A}\right)^2 \frac{1-f_{hi}^{A}}{n_{hi}^{A}\left(n_{hi}^{A}-1\right)} \sum_{k\in S_{hi}^{A}} \left\{y_k 1\left(k\in a\right) - \overline{y}_{a;S_{hi}^{A}}\right\}\left\{y_k 1\left(k\in ab\right) - \overline{y}_{ab;S_{hi}^{A}}\right\},$$

noting $\overline{y}_{d;V}$ the average of variable $y_k 1\left(k\in d\right)$ on a subset $V\subset U$. For each sample, the Kalton and Anderson estimator (KALT) given in (3.8) is also calculated, as well as the Bankier estimator (BANK) given in (3.9), and the Horvitz-Thompson estimator $\hat{Y}^{A}$ based on the single sample S^{A} (HTA). The sampling procedure is repeated 10,000 times. To measure the bias of an estimator $\hat{Y}$, we calculate its relative Monte Carlo bias

$$RB_{MC}\left(\hat{Y}\right) = \frac{E_{MC}\left(\hat{Y}\right) - Y}{Y}\times 100$$

with $E_{MC}\left(\hat{Y}\right) = \left(1/10{,}000\right)\sum_{b=1}^{10{,}000}\hat{Y}_{(b)}$, and $\hat{Y}_{(b)}$ the value of estimator $\hat{Y}$ for sample b. To measure the variability of $\hat{Y}$, we calculate its Monte Carlo mean square error

$$MSE_{MC}\left(\hat{Y}\right) = \frac{1}{10{,}000}\sum_{b=1}^{10{,}000}\left(\hat{Y}_{(b)} - Y\right)^2.$$

The results are given in Table 4.2. As emphasized by a referee, the performances of the HTA estimator do not depend on the sample size n_{hi}^{B} chosen. For consistency, Table 4.2 indicates the results obtained in the simulations with $n_{hi}^{B} = 5$ only. For identical sample sizes m_h and identical n_{hi}^{A}, the same results are reported in the case $n_{hi}^{B} = 20$.

All estimators are virtually unbiased. The HART2 estimator gives better results in terms of mean squared error, as could be expected. The HTA estimator gives almost equivalent results. This result is explained by the fact that the optimal coefficient is near 1 (in the simulations, $\hat{\theta}_{opt}$ is between 0.80 and 1.06), and that in this case, the formula (2.1) shows that the HART2 and HTA estimators are very close: In the appendix we present some general conditions under which this property is approximately checked. Of the three estimators, HART1 yields the best results, with a mean square error lower than or equivalent to that of KALT and BANK in 11 out of 16 cases.

Table 4.2
Relative bias and mean squared error of five estimators

Pop.	m_h	n_{hi}^A	n_{hi}^B	HART1		HART2		KALT		BANK		HTA	
				RB (%)	MSE $\times 10^9$	RB (%)	MSE $\times 10^9$	RB (%)	MSE $\times 10^9$	RB (%)	MSE $\times 10^9$	RB (%)	MSE $\times 10^9$
1	5	10	5	0.05	7.76	0.01	5.70	0.05	7.79	0.06	8.56	0.04	5.75
1	5	10	20	0.01	7.57	-0.05	5.57	0.03	11.36	0.04	12.75	0.04	5.75
1	5	40	5	0.01	5.01	-0.02	4.51	-0.02	4.57	-0.02	4.81	-0.02	4.52
1	5	40	20	0.00	4.65	-0.01	4.33	0.00	4.66	0.00	5.22	-0.02	4.52
1	25	10	5	-0.03	1.19	-0.02	0.78	-0.03	1.20	-0.02	1.34	-0.01	0.78
1	25	10	20	-0.01	1.17	0.00	0.78	-0.03	1.94	-0.03	2.22	-0.01	0.78
1	25	40	5	0.00	0.62	0.01	0.51	0.00	0.52	0.00	0.57	0.01	0.51
1	25	40	20	0.02	0.58	0.01	0.51	0.02	0.58	0.02	0.68	0.01	0.51
2	5	10	5	0.00	3.59	0.01	1.15	0.00	3.56	0.02	4.38	0.01	1.15
2	5	10	20	0.00	3.60	-0.02	1.15	0.00	7.38	0.00	8.76	0.01	1.15
2	5	40	5	0.00	1.48	0.01	1.07	0.00	1.13	0.01	1.35	0.01	1.07
2	5	40	20	0.00	1.49	-0.01	1.09	0.00	1.49	0.00	2.03	0.01	1.07
2	25	10	5	0.00	0.63	0.00	0.14	0.00	0.63	0.00	0.78	0.00	0.14
2	25	10	20	0.00	0.62	0.00	0.13	0.00	1.38	0.00	1.67	0.00	0.14
2	25	40	5	0.00	0.20	0.00	0.12	0.00	0.13	0.00	0.18	0.00	0.12
2	25	40	20	0.00	0.20	0.00	0.12	0.00	0.20	0.01	0.31	0.00	0.12

For each estimator, all other things being equal, the mean square error is lower in population 2 than in population 1. This result comes from the fact that the variance due to the first-stage selection, which is the same for each estimator and is

$$V\left(\sum_{i\in S_I} d_{Ii}Y_i\right)=\sum_{h=1}^{H}M_h^2\left(\frac{1}{m_h}-\frac{1}{M_h}\right)S_{Y;U_{Ih}}^2, \tag{4.3}$$

is larger in population 1: the dispersion term $S_{Y;U_{Ih}}^2=\left(M_h-1\right)^{-1}\sum_{u_i\in U_{Ih}}\left(Y_i-\overline{Y}_{U_{Ih}}\right)^2$ increases with σ_h^2 and, to a lesser degree, increases when ρ decreases. The mean square error decreases for each estimator when the number m_h of primary sampling units selected in each stratum increases, since in this case the common variance term given in (4.3) decreases. Similarly, the mean square error decreases for each estimator when n^A increases, since in this case the variance due to the second stage of selection decreases. For the HART1 and HART2 estimators, the mean square error is stable when n^B increases, and more surprisingly for the KALT and BANK estimators the mean square error increases when n^B increases. This somewhat counterintuitive result is due to the convergence of two facts. On one hand, the contribution of sample S^B to the variance due to the second stage of selection is low: the increase of n^B may reduce this variance, but even in this case, overall reduction of the variance is marginal. On the other hand, with the KALT and BANK estimators, the contribution of sample S^A to the variance due to the second stage of selection increases when n^B increases.

In the case of KALT, the estimator can be re-expressed

$$\hat{Y}_{KA}=\sum_{h=1}^{H}\frac{M_h}{m_h}\sum_{i\in S_{Ih}}\hat{Y}_{KA,i}$$

with

$$\hat{Y}_{KA,i} = \frac{1}{f_{hi}^A} \sum_{k \in S_i^A} m_{k|i}^A y_k + \frac{1}{f_{hi}^A + f_{hi}^B} \sum_{k \in S_i^B} y_k \quad \text{and} \quad m_{k|i}^A = \begin{cases} 1 & \text{if } k \in a \cap u_i, \\ \dfrac{f_{hi}^A}{f_{hi}^A + f_{hi}^B} & \text{if } k \in ab \cap u_i. \end{cases} \tag{4.4}$$

In (4.4), the dispersion of the variable $m_{k|i}^A$ (and therefore, that of $m_{k|i}^A\, y_k$) increases when the factor $f_{hi}^A \Big/ \left(f_{hi}^A + f_{hi}^B\right)$ moves away from 1. This factor is near 1 when f_{hi}^B is small compared to f_{hi}^A (and therefore, if n^B is small compared to n^A), but moves away from 1 when n^B increases. Note that the variance (conditional on S_I) of the second term of $\hat{Y}_{KA,i}$ is equal to

$$V\left(\frac{1}{f_{hi}^A + f_{hi}^B} \sum_{k \in S_i^B} y_k \,\middle|\, S_I \right) = \left(N_{hi}^A\right)^2 N_{hi}^B \times \frac{n_{hi}^B \left(N_{hi}^B - n_{hi}^B\right)}{\left(N_{hi}^B n_{hi}^A + N_{hi}^A n_{hi}^B\right)^2} \times S_{u_{hi}^B}^2$$

with $S_{u_{hi}^B}^2 = \left(N_{hi}^B - 1\right)^{-1} \sum_{k \in u_{hi}^B} \left(y_k - \overline{y}_{u_{hi}^B}\right)^2$. This variance does not necessarily decrease when n_{hi}^B increases. For example, one of the cases considered in the simulations corresponds to $N_{hi}^A = 100$, $N_{hi}^B \simeq 50$ and $n_{hi}^A = 40$. In this case, the term $n_{hi}^B \left(N_{hi}^B - n_{hi}^B\right) \Big/ \left(N_{hi}^B n_{hi}^A + N_{hi}^A n_{hi}^B\right)^2$ attains its maximum value for $n_{hi}^B = 11$.

In the case of BANK, the estimator can be re-expressed

$$\hat{Y}_{HT} = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{i \in S_{Ih}} \hat{Y}_{HT,i}$$

with

$$\hat{Y}_{HT,i} = \sum_{k \in S_i^A \cup S_i^B} \frac{y_k}{\pi_{k|i}^{HT}} \quad \text{and} \quad \pi_{k|i}^{HT} = \begin{cases} f_{hi}^A & \text{if } k \in a, \\ f_{hi}^A + f_{hi}^B \left(1 - f_{hi}^A\right) & \text{if } k \in ab. \end{cases} \tag{4.5}$$

In (4.5), dispersion of the variable $\pi_{k|i}^{HT}$ increases when the factor $f_{hi}^B \left(1 - f_{hi}^A\right)$ increases. This factor is close to 0 when n_{hi}^B (and, therefore, f_{hi}^B) is low, but increases when n_{hi}^B increases.

5 Conclusion

We examined the Hartley (1962), Kalton and Anderson (1986) and Bankier (1986) estimators to pool the samples resulting from two survey waves. More particularly, we studied the case where the first sample represents the entire population (completely representative sample), while the second represents only a part (partially representative sample). Within the framework considered in the simulations (also see the Appendix for a more general framework), using the partially representative sample did not improve accuracy: if its size increases, the accuracy of the estimators in the Hartley class remains stable or improves slightly, while the accuracy of the Kalton and Anderson and Bankier estimators is worsened. Hartley's optimal estimator itself, although more complex to calculate, offers accuracy that is only slightly improved as compared to the classic Horvitz-Thompson estimator calculated on the fully representative sample. Although our simulation study is limited, the results suggest that the estimator should be chosen carefully when there are multiple survey frames, and that a simple estimator is sometimes preferable, even if it uses only part of the information collected.

Acknowledgements

The authors would like to thank an associate editor and referee for their careful reading and comments, which helped to significantly improve the article, and David Haziza for the useful discussions.

Appendix

A1. Comparison of Hartley's optimal estimator and the Horvitz-Thompson estimator

Let us take the framework and notations from Section 4: samples S^A and S^B are selected using a two-stage frame with common first stage selection. Stratified simple random sampling is used at the first stage, and simple random sampling in each primary sampling unit at the second stage. The sampling frame U_A corresponds to the entire population, while the sampling frame U_B covers only part of the population.

With Hartley's optimal estimator, the formula (3.6) gives

$$\theta_{opt|S_I} = \frac{EV\left(\hat{Y}_{ab}^B \mid S_I\right) - ECov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I\right)}{EV\left(\hat{Y}_{ab}^B \mid S_I\right) + EV\left(\hat{Y}_{ab}^A \mid S_I\right)}.$$

After some calculation, we get

$$EV\left(\hat{Y}_{ab}^A \mid S_I\right) = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{u_{hi} \in U_{Ih}} \left(N_{hi}\right)^2 \frac{1-f_{hi}^A}{n_{hi}^A} \left\{ \frac{N_{hi}^B - 1}{N_{hi} - 1} S_{u_{hi}^B}^2 + \frac{N_{hi}^B \left(N_{hi} - N_{hi}^B\right)\left(\overline{y}_{u_{hi}^B}\right)^2}{N_{hi}\left(N_{hi} - 1\right)} \right\}, \tag{A.1}$$

$$-ECov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I\right) = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{u_{hi} \in U_{Ih}} \left(N_{hi}\right)^2 \frac{1-f_{hi}^A}{n_{hi}^A} \left\{ \frac{N_{hi}^B \left(\overline{y}_{u_{hi}^B}\right)\left(N_{hi}\overline{y}_{u_{hi}} - N_{hi}^B \overline{y}_{u_{hi}^B}\right)}{N_{hi}\left(N_{hi} - 1\right)} \right\}$$

with $\overline{y}_{u_{hi}} = \left(N_{hi}\right)^{-1} \sum_{k \in u_{hi}} y_k$, $\overline{y}_{u_{hi}^B} = \left(N_{hi}^B\right)^{-1} \sum_{k \in u_{hi}^B} y_k$ and $S_{u_{hi}^B}^2 = \left(N_{hi}^B - 1\right)^{-1} \sum_{k \in u_{hi}^B} \left(y_k - \overline{y}_{u_{hi}^B}\right)^2$.

The Horvitz-Thompson estimator based on the single sample S^A and Hartley's optimal estimator agree if the coefficient $\theta_{opt|S_I}$ is equal to 1, which is the case if $EV\left(\hat{Y}_{ab}^A \mid S_I\right) = -ECov\left(\hat{Y}_a^A, \hat{Y}_{ab}^A \mid S_I\right)$. This condition will be verified in particular if in (A.1) the terms between the brackets agree for each primary sampling unit u_{hi}. We get therefore $\theta_{opt|S_I} \simeq 1$ if

$$\forall \quad u_{hi} \in U_I \quad \frac{N_{hi}\left(N_{hi}^B - 1\right)}{N_{hi}^B} \frac{S_{u_{hi}^B}^2}{\overline{y}_{u_{hi}^B}\left(N_{hi}\overline{y}_{u_{hi}} - N_{hi}^B \overline{y}_{u_{hi}^B}\right)} + \frac{\left(N_{hi} - N_{hi}^B\right)\overline{y}_{u_{hi}^B}}{N_{hi}\overline{y}_{u_{hi}} - N_{hi}^B \overline{y}_{u_{hi}^B}} \simeq 1. \tag{A.2}$$

Let us suppose that the mean value of y is approximately the same in the frames U_A and U_B for each primary sampling unit, i.e. that $\forall\, u_{hi} \in U_I \quad \overline{y}_{u_{hi}^B} \simeq \overline{y}_{u_{hi}}$. Then, the condition (A.2) will be verified approximately if $\forall\, u_{hi} \in U_I \quad cv_{u_{hi}^B}^2$ is close to 0, with $cv_{u_{hi}^B} = \sqrt{S_{u_{hi}^B}^2} \Big/ \overline{y}_{u_{hi}^B}$.

In summary, the Horvitz-Thompson estimator based on the single sample S^A and Hartley's optimal estimator will be close if within each primary sampling unit u_{hi}: (a) there is not much difference in the mean value of y between the two bases, and (b) the variable y has low dispersion within u_{hi}^B. In the simulations, the condition (a) is approximately met since the distribution of individuals between the sampling frames U_A and U_B is completely random; the condition (b) is approximately met with values of $cv^2_{u_{hi}^B}$ varying from 0.02 to 0.10 for population 1, and from 0.001 to 0.005 for population 2.

References

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, p.1074-1079.

Bourdalle, G., Christine, M. and Wilms, L. (2000). Échantillons maître et emploi. *Série INSEE Méthodes*, 21, p. 139-173.

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, p. 333-362.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, p. 203-206.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, p. 663-685.

Kalton, G. and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, A*, 149, p. 65-82.

Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles (Belgium) and Éditions Ellipses (France).

Lavallée, P. (2007). *Indirect sampling*. New York: Springer.

Lohr, S.L. (2007). Recent developments in multiple frame surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3257-3264.

Lohr, S.L. (2009). Multiple frame surveys. In *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, Eds., D. Pfeffermann and C.R. Rao. Amsterdam: North Holland, Vol. 29A, p. 71-88.

Lohr, S.L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology*, Vol.37 no.2, p. 197-213.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, Vol.33 no.2, p. 151-157.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, p. 169-175.

Rao, J.N.K. and Wu, C. (2010). Pseudo-empirical likelihood inference for dual frame surveys. *Journal of the American Statistical Association*, 105, p. 1494-1503.

Saigo, H. (2010). Comparing four bootstrap methods for stratified three-stage sampling. *Journal of Official Statistics*, Vol. 26, No. 1, 2010, p. 193–207.

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 347-354
Statistics Canada, Catalogue No. 12-001-X

Combining information from multiple complex surveys

Qi Dong, Michael R. Elliott and Trivellore E. Raghunathan[1]

Abstract

This manuscript describes the use of multiple imputation to combine information from multiple surveys of the same underlying population. We use a newly developed method to generate synthetic populations nonparametrically using a finite population Bayesian bootstrap that automatically accounting for complex sample designs. We then analyze each synthetic population with standard complete-data software for simple random samples and obtain valid inference by combining the point and variance estimates using extensions of existing combining rules for synthetic data. We illustrate the approach by combining data from the 2006 National Health Interview Survey (NHIS) and the 2006 Medical Expenditure Panel Survey (MEPS).

Key Words: Synthetic populations; Posterior predictive distribution; Bayesian bootstrap; Inverse sampling.

1 Introduction

Survey agencies often repeatedly draw samples from similar populations and collect similar variables, sometimes even using the same frame. For example, the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES) are both conducted by the U.S. National Center for Health Statistics. These two surveys target the U.S. non-institutionalized population and have a considerable overlap of questions. By combining information from multiple surveys, we hope to obtain more accurate inference for the population than if we use the data from a single survey.

One of the biggest challenges in such combining is the compatibility of multiple data sources. Surveys may use different sampling designs or modes of data collection, which may result in various sampling and nonsampling error properties. Instead of directly pooling the data from multiple surveys for a simple analysis, we need to adjust for the discrepancies among the data to make them comparable.

Various methods for combining data collected in two surveys have been proposed in the survey methodology literature (Hartley 1974; Skinner and Rao 1996; Lohr and Rao 2000; Elliott and Davis 2005; Raghunathan, Xie, Schenker, Parsons, Davis, Dodd and Feuer 2007; Schenker, Gentleman, Rose, Hing and Shimizu 2002; Schenker and Raghunathan 2007; Schenker, Raghunathan and Bondarenko 2009). The most recent papers by Raghunathan et al. (2007) and Schenker et al. (2009) applied model-based approaches. The basic idea for the model-based approach is to fit an imputation model to the data of better quality and use the fitted model to impute the values in the other samples of lower quality. As long as the imputation model is correctly specified, this approach can take advantage of the strengths of the multiple data sources and improve the statistical inference. However, as suggested by Reiter, Raghunathan and Kinney (2006), when the sample is collected using complex sampling designs, ignoring those features could result in biased estimates from the design-based perspective. However, fully accounting for the complex sampling design features in practice is very difficult. For example, both Raghunathan et al. (2007) and Schenker et al. (2009) used a simplified method to adjust for stratification and clustering.

1. Qi, Dong, Google, Inc., 1R4A, Quad 5, Google Inc, 399 N. Whisman Road, Mountain View, CA 94043. E-mail: qdong@google.com; Michael R. Elliott, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109 and Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. E-mail: mrelliot@umich.edu; Trivellore E. Raghunathan, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109 and Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. E-mail: teraghu@umich.edu.

Raghunathan et al. (2007) used a rudimentary concept of design effect and Schenker et al. (2009) used propensity scores to create adjustment subgroups for modeling.

Here we propose a new method for combining multiple surveys that adjusts for the complex sampling design features in each survey. The unobserved population in each survey will be treated as missing data to be multiply imputed. The imputation model will account for complex design features using a recently developed nonparametric synthetic population generation method (Dong, Elliott and Raghunathan 2014). For each survey, the observed data and the multiply imputed unobserved population produce multiple synthetic populations. Once the whole population is generated, the complex sampling design features such as stratification, clustering and weighting will be of no use in the analysis and the synthetic populations can be treated as equivalent simple random samples. Finally, the estimate for the population quantity of interest will be calculated from each synthetic population and then will be combined first within each individual survey and then across multiple surveys.

This paper proceeds as follows: Section 2 summarizes generating synthetic population while accounting for complex sampling design features using the nonparametric approach. Section 3 describes methodology to produce combined estimates from these multiple synthetic populations. In Section 4, we apply the proposed method to combine the 2006 NHIS and the Medical Expenditure Panel Survey (MEPS) to estimate the health insurance coverage rates of the US population. Section 5 concludes with discussion and directions for future research

2 Generating synthetic populations from single survey data that accounts for complex sampling designs

Dong et al. (2014) extended work in the finite population Bayesian bootstrap to develop a non-parametric approach to the generation of posterior predictive distributions. A summary of the algorithm to draw the l-th of $l = 1,...,L$ synthetic populations for stratified, clustered sample designs with unequal probabilities of selection is as follows:

1. Use the Bayesian Bootstrap (BB) (Rubin 1981) to adjust for stratification and clustering. Draw a simple random sample with replacement (SRSWR) of size m_h from the c_h clusters within each stratum $h = 1,...,H$ and calculate bootstrap replicate weights for each of the n_{hi} observations in each cluster as $w^{*(l)} = \left\{ w_{hi}^{*(l)},\ h = 1,\ldots,H,\ i = 1,\ldots,c_h,\ k = 1,\ldots,n_{hi} \right\}$, where $w_{hik}^{*} = w_{hik}\left(\left(1 - \sqrt{\left(m_h / c_h - 1 \right)} \right) + \sqrt{\left(m_h / c_h - 1 \right)} \left(c_h / m_h \right) m_{hi}^{*} \right)$ and m_{hi}^{*} denotes the number of times that cluster $i,\ i = 1,\ldots,c_h$ is selected. To ensure all the replicate weights are non-negative, $m_h \le \left(c_h - 1 \right);$ here and below we take $m_h = \left(c_h - 1 \right).$

2. Use the finite population Bayesian bootstrap (FPBB) (Lo 1986; Cohen 1997) for unequal probabilities of selection to adjust for unequal probabilities of selection. For each cluster i in stratum h of population size N_{hi}, draw a sample of size $N_{hi} - n_{hi}$, denoted by $\left(y_1^{*},\ldots,y_{N_{hi}-n_{hi}}^{*} \right)$, by drawing y_{hik}^{*} from cluster data $\left(y_1,\ldots,y_{n_{hi}} \right)$ with probability

$\frac{w^*_{hik} - 1 + l_{hik,j-1} * (N_{hi} - n_{hi})/n_{hi}}{N_{c_H} - n_{c_H} + (j-1) * (N_{hi} - n_{hi})/n_{hi}}$, where w^*_{hik} is the replicate weight of unit k in cluster i in stratum h, and $l_{hik,j-1}$ is the number of bootstrap selections of y_{hik} among $y^*_1,\ldots,y^*_{j-1}$. Form the FPBB population $y_1,\ldots,y_{n_{hi}},y^*_1,\ldots,y^*_{N_{hi}-n_{hi}}$.

3. Produce F FPBB samples for each BB sample, denoted by $S_{l1},\ldots,S_{lF}, l=1,\ldots,L$. Pool the F FPBB samples to produce one synthetic population, S_l. (Because $N = \sum_h \sum_i N_{hi}$ may be unrealistically large, generating a sample of size $k * n$ for large k is sufficient.)

3 Combining rule for the synthetic populations from multiple surveys

Assume that $Q = Q(Y)$ is the population quantity of interest depending upon the set of variables Y that are collected in multiple surveys: for example, a population mean, proportion or total, a vector of regression coefficients, etc. For simplicity of exposition we assume Q to be scalar. Assume that, using data from a single survey s, we create L synthetic populations, $S_l^{(s)}$, $l=1,\ldots,L$, using the methods summarized in Section 2. Denote $Q_l^{(s)}$ as the corresponding estimate of the population quantity Q obtained from synthetic population l generated using data from survey s (note this estimate can be obtained under a simple random sampling assumption). Dong et al. (2014) shows that, under reasonable asymptotic assumptions (sufficient sample size for the sample quantity of interest to be normally distributed, synthetic populations generated consistent with the survey design),

$$Q \mid S_1^{(s)},\ldots,S_L^{(s)} \dot{\sim} t_{L-1}\left(\bar{Q}_L^{(s)}, \left(1 + L^{-1}\right) B_L^{(s)}\right) \tag{3.1}$$

where $\bar{Q}_L^{(s)} = L^{-1}\sum_{l=1}^{L} Q_l^{(s)}$ is the mean of Q across the L synthetic populations and $B_L^{(s)} = (L-1)^{-1}\sum_{l=1}^{L}\left(Q_l^{(s)} - \bar{Q}_L^{(s)}\right)^2$ is the between-imputation variance. The result follows immediately from Section 4.1 of Raghunathan, Reiter and Rubin (2003), and is based on the standard Rubin (1987) multiple imputation combining rules. The average "within" imputation variance is zero, since the entire population is being synthesized; hence the posterior variance of Q is entirely a function of the between-imputation variance.

The combining rule obtained in (3.1) may not yield valid inference for the parameters of interest for multiple surveys, since the models to generate synthetic populations for the multiple surveys may be different. Thus, a new rule for combining estimates across multiple surveys needs to be developed.

3.1 Normal Approximation when L is large

Let $\bar{Q}_L^{(s)}$ and $B_L^{(s)}$ be the combined estimator of the population quantity of interest and its variance for survey s obtained using the combining formulas for synthetic populations $S_{syn}^{(s)} = \left\{S_l^{(s)}, l=1,\ldots,L\right\}$, $s=1,\ldots,S$ in a single survey setting. When L is large, we have

$$Q \mid S_{syn}^{(1)},...,S_{syn}^{(S)} \dot{\sim} N\left(\bar{Q}_L, B_L\right) \tag{3.2}$$

where $\bar{Q}_L = \sum_{s=1}^{S}\left(\bar{Q}_L^{(s)}/B_L^{(s)}\right)\Big/\sum_{s=1}^{S}\left(1/B_L^{(s)}\right)$ and $B_L = 1\Big/\sum_{s=1}^{S}\left(1/B_L^{(s)}\right)$. Equation (3.2) follows immediately from standard Bayesian results, assuming that 1) the true variance of $\bar{Q}_L^{(s)}$, B_s, can be approximated by $B_L^{(s)}$ obtained from the synthetic populations as in Section 3, i.e., $\left(\bar{Q}_L^{(s)} \mid Q, B_s\right) = \left(\bar{Q}_L^{(s)} \mid Q, B_L^{(s)}\right) \sim N\left(Q, B_L^{(s)}\right)$, 2) each survey is independent, and 3) Q has a non-informative prior $\pi\left(Q \mid B_L^{(s)}\right) \propto 1$.

3.2 T-corrected Distribution for Small/Moderate L

For small to moderate L, the posterior distribution of Q is better approximated by

$$Q \mid S_{syn}^{(1)},...,S_{syn}^{(S)} \dot{\sim} \mathrm{t}_{\nu_L}\left(\bar{Q}_L, \left(1+L^{-1}\right)B_L\right) \tag{3.3}$$

where $\bar{Q}_L$ and B_L are defined as in 3.1, and degrees of freedom $\vartheta_L = (L-1)\Big/\sum_{s=1}^{S}\left(\left(1/b_L^{(s)}\right)\Big/\sum_{s=1}^{S}\left(1/b_L^{(s)}\right)\right)^2$. Details are available in Dong (2012), and follow the extensions of Raghuanthan et al. (2003) that were used to derive the large L results.

4 Combined estimates of health insurance coverage from the NHIS, MEPS and BRFSS

The 2006 NHIS and MEPS data are multistage probability samples that incorporate stratification, clustering and oversampling of some subpopulations (e.g., Black, Hispanic, and Asian in later years). For confidentiality reasons the true strata and PSUs are suppressed. The NHIS is released with 300 pseudo-strata and two pseudo-PSUs per stratum; MEPS, which is a subsample of the households which participate in the NHIS, is released with 203 pseudo-strata and up to three pseudo-PSUs per stratum (Ezzati-Rice, Rohde and Greenblatt 2008; National Center for Health Statistics 2007). The NHIS and MEPS ask one randomly-sampled adult in each household whether they are covered by any health insurance and, if so, whether they are covered by private or government insurance. We consider this trinomial distribution of insurance status in the overall adult population, as well as in subpopulations consisting of males, Hispanics, non-Hispanic whites, and non-Hispanic whites earning between $25,000 and $35,000 per year. We delete the cases with item-missing values and focus on our study on the complete cases. This results in 20,147 and 20,893 cases in the NHIS and MEPS data respectively.

The 2006 BRFSS is obtained via random digit dialing (RDD) using list-assisted sampling, stratified by state. While such designs avoid clustering, unequal probability of selection is introduced because the sample size is roughly equal in each state; in addition only one adult is sampled per household. In contrast to the NHIS and MEPS, the BRFSS only asks whether one is insured or not, so we only calculate the proportion of respondents who are not covered by any insurance. We delete the cases with item-missing

values and focus on our simulation on the complete cases. There are 294,559 complete cases in the 2006 BRFSS data.

We generate the synthetic populations for the three surveys from 200 BB samples, each consisting of 10 FPBB samples of size $5n$ $(B = 200,\ F = 10,\ k = 5)$. We then produce the combined estimates of people's health insurance coverage rates using the combining survey method described above. Since all three surveys have the information about whether people have insurance or not, we can combine the NHIS, BRFSS and MEPS to estimate the proportion of uninsured people. However, the BRFSS does not ask people what type insurance they have (private vs. public). For these proportions, we can only combine the NHIS and MEPS. The results are summarized in Table 4.1. The variance estimates for the combined estimator are much smaller than the ones obtained from the actual data. Specifically, the precision of the estimates obtained from the NHIS is increased by 43% on average, with the largest increase of 98% obtained by combining the NHIS and MEPS. The gains in precision for the MEPS are even more. The average increase in precision for the MEPS is 101%, with the largest increase being 202%. The precision is further increased when we combine all three surveys. For example, for the proportion of people who have no coverage, on average the precision is increased by 5 times for the NHIS, 1.5 times for the BRFSS and 4.2 times for the MEPS. This implies gains in precision by making use of the information from multiple surveys can be significant, and the more information we combine, the larger the gains are in precision.

5 Discussion

In this paper, we propose a new method to combine information from multiple complex surveys. We apply the new method to combine information about health insurance status from the 2006 NHIS, MEPS, and BRFSS. Results show that the combined estimate is more precise compared to the estimates from individual surveys. As previous work has shown (Dong et al. 2014), we have little information loss in the sense that the sampling properties of inferences from the synthetic population and the actual sample are very similar. Thus when we combine the estimates from three samples, the combined estimate is substantially more efficient that the estimates from individual surveys. (We note that this application is primarily for illustrative purposes; similar inferences could be made by computing the design-based estimates and variances for each of the surveys, then applying the combining rule in (3.2) on the design-based estimates.)

This new combining survey method has two major advantages over the existing methods. First, the approach used here to generate synthetic populations, discussed in detail in Dong et al. (2014), accounts for the complex sample design nonparametrically using extensions of finite population Bayesian bootstrap methods. Since the resulting synthetic populations can be analyzed as simple random samples, information from other surveys can be used to adjust for the nonsampling errors and/or filling in the missing variables. Another advantage of this method is it has no limitation on the number of surveys to be combined as long as the surveys have the same underlying population. The proposed method that adjusts for the complex sampling design features can be applied to each survey independently. After the missing information is imputed, regardless the number of surveys to be combined, we only need to combine the estimates from each survey using the combing rule developed in this manuscript. A final advantage of the proposed approach is the ability of the synthetic populations generated by the nonparametric method to preserve the

item-missing values in the actual data. This potentially fills in a gap in the multiple imputation area that existing imputation methods typically ignore the complex sampling design features in the data and impute the missing values as if they are simple random samples. We consider this application in future work.

Table 4.1
Individual and combined estimates for the 2006 NHIS, MEPS and BRFSS.

Domain		Actual Data (Complex Design)			Combined Estimates	
	Types	NHIS	BRFSS	MEPS	NHIS and MEPS	NHIS, BRFSS and MEPS
Whole Population	**Proportion**					
	Private	0.746		0.735	0.741	
	Public	0.075		0.133	0.094	
	Uninsured	0.179	0.154	0.132	0.152	0.153
	Variance					
	Private	2.46E-05		2.78E-05	1.61E-05	
	Public	6.29E-06		1.44E-05	5.35E-06	
	Uninsured	1.84E-05	3.32E-06	1.41E-05	9.80E-06	2.55E-06
Male	**Proportion**					
	Private	0.740		0.735	0.738	
	Public	0.060		0.101	0.074	
	Without	0.200	0.167	0.164	0.181	0.172
	Variance					
	Private	3.32E-05		3.87E-05	2.06E-05	
	Public	6.82E-06		1.53E-05	5.72E-06	
	Uninsured	2.94E-05	8.88E-06	2.64E-05	1.51E-05	5.61E-06
Hispanic	**Proportion**					
	Private	0.494		0.506	0.5014	
	Public	0.096		0.161	0.1157	
	Without	0.410	0.371	0.334	0.3684	0.3689
	Variance					
	Private	1.24E-04		1.73E-04	9.76E-05	
	Public	2.57E-05		8.03E-05	2.66E-05	
	Uninsured	1.23E-04	7.18E-05	1.19E-04	8.71E-05	3.79E-05
Non-Hispanic White	**Proportion**					
	Private	0.805		0.788	0.796	
	Public	0.062		0.116	0.081	
	Without	0.134	0.1059	0.096	0.113	0.107
	Variance					
	Private	2.99E-05		3.35E-05	1.97E-05	
	Public	8.20E-06		1.81E-05	6.86E-06	
	Uninsured	2.02E-05	2.15E-06	1.51E-05	1.02E-05	1.90E-06
Non-Hispanic White & Income [25,000, 35,000)	**Proportion**					
	Private	0.827		0.813	0.821	
	Public	0.039		0.079	0.053	
	Without	0.134	0.173	0.108	0.122	0.154
	Variance					
	Private	1.0E-04		1.39E-04	7.74E-05	
	Public	2.82E-05		6.31E-05	2.52E-05	
	Uninsured	7.24E-05	2.78E-05	8.92E-05	5.14E-05	1.93E-05

References

Cohen, M.P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 635-638.

Dong, Q. (2012). Unpublished PhD thesis, University of Michigan.

Dong Q., Elliott, M.R. and Raghunathan T.E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, 40 (1), 29-46.

Elliott, M.R. and Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Journal of the Royal Statistical Society C*: Applied Statistics, 54, 595-609.

Ezzati-Rice, T.M., Rohde, F. and Greenblatt, J. (2008). Sample design of the medical expenditure panel survey household component, 1998–2007. *Methodology Report No. 22*. Agency for Healthcare Research and Quality, Rockville, MD. Accessed at: http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.pdf, February 2014.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *The Indian Journal of Statistics*, C, 38, 99-118.

Lo, A.Y. (1986). Bayesian statistical inference for sampling a finite population. *Annals of Statistics*, 14, 1226-1233.

Lohr, S.L. and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

National Center for Health Statistics (2007). Data file documentation, National Health Interview Survey, 2006 (machine readable data file and documentation). *National Center for Health Statistics*, Centers for Disease Control and Prevention, Hyattsville, Maryland. Accessed at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2006/srvydesc.pdf, February 2014.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., Xie, D.W., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. and Feuer, D.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102, 474-486.

Reiter, J.P., Raghunathan, T.E. and Kinney, S.K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32, 143-149.

Rubin, D.B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9, 131-134.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schenker, N., Gentleman, J.F., Rose, D, Hing, E. and Shimizu, I.M. (2002). Combining estimates from complementary surveys: A case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*, 117, 393-407.

Schenker, N. and Raghunathan, T.E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, 26, 1802-1811.

Schenker, N., Raghunathan, T.E. and Bondarenko, I. (2009). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine*, 29, 533-545.

Skinner, C.J. and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2014.

T. Adams, *U.S. Census Bureau*
S. Adeshiyan, *U.S. Energy Information Administration*
R. Andridge, *Ohio State University*
P. Ardilly, *INSEE, France*
T. Asparouhov, *Muthén & Muthén*
J. Aston, *University of Warwick*
R. Bautista, *NORC at the University of Chicago*
J.-F. Beaumont, *Statistics Canada*
E. Benhin, *Statistics Canada*
Y.G. Berger, *University of Southampton*
J. Bethlehem, *Statistics Netherlands/Leiden University*
C. Bocci, *Statistics Canada*
J. Breidt, *Colorado State University*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Census Bureau*
R. Chambers, *Centre for Statistical and Survey Methodology*
S. Chaudhury
G. Chauvet, *ENSAI, France*
R. Clark, *NIASRA, University of Wollongong*
G. Datta, *University of Georgia*
T. DeMaio, *U.S. Census Bureau*
J. Dever, *Research Triangle Institute*
J. Eltinge, *U.S. Bureau of Labour Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
D. Haziza, *Université de Montréal*
S. Heeringa, *ISR, U. of Michigan*
K.A. Henry, *Statistics of Income, Internal Revenue Service*
B. Hulliger, *U. of Applied Sciences Northwestern Switzerland*
F. Hutchinson, *Cancer Research Center*
D. Judkins, *Abt Associates*
K. Kadraoui, *Université Laval*
R.J. Karunamuni, *University of Alberta*
D. Kasprzyk, *NORC at the University of Chicago*
J.-K. Kim, *Iowa State University*
P.S. Kott, *RTI International*
P. Lahiri, *JPSM*, *University of Maryland*
P. Lavallée, *Statistics Canada*
L. Lee, *NORC at the University of Chicago*
P. Linde, *Statistics Denmark*
S. Lohr, *Westat*
P. Lynn, *University of Essex*
D.J. Malec, *National Center for Health Statistics*
D. Marker, *Westat*
J. Montaquila, *Westat*
B. Nandram, *Worcester Polytechnic Institute*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
F. Picard, *Statistics Canda*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
E. Robison, *U.S. Bureau of Labor Statistics*
T. Savitsky, *Bureau of Labor Statistics*
J. Shao, *University of Wisconsin*
F.J. Scheuren, *National Opinion Research Center*
P.L.D.N. Silva, *Escola Nacional de Ciências Estatísticas*
C. Skinner, *LSE*
P. Smith, *Office for National Statistics*
D. Steel, *University of Wollongong*
D. Sverchkov, *U.S. Bureau of Labor Statistics*
N. Thomas, *Statisctics Reasearch and Consulting Center, Pfizer*
M. Thompson, *University of Waterloo*
M. Torabi, *Mplus*
D. Toth, *U.S. Bureau of Labor Statistics*
R. Valliant, *University of Maryland*
J. van den Brakel, *Statistics Netherlands*
F. Verret, *Statistics Canada*
B.T. West, *ISR, University of Michigan - Ann Arbor*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
D. Yang, *Bureau of Labor Statistics*
Y. You, *Statistics Canada*
Z. Yu, *U. of Wisconsin Madison*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard University*
S.Z. Zangeneh, *Vaccine and Infectious Disease Division*

Acknowledgements are also due to those who assisted during the production of the 2014 issues: Joana Bérubé of Business Survey Methods Division; the team from Dissemination Division, in particular: Éva Demers-Brett, Chantal Chalifoux, Jacqueline Luffman, Kathy Charbonneau, Lucie Gauthier, Daniel Piché, Jasvinder Jassal, Joseph Prince et Darquise Pellerin; Céline Ethier and Nick Budko of Statistical Research and Innovation Division as well as our partners in the Communications Division.

ANNOUNCEMENTS

Nominations Sought for the 2016 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium and give the 2016 Waksberg Invited Address at the Statistics Canada Symposium to be held in the autumn of 2016. The paper will be published in a future issue of *Survey Methodology* (targeted for December 2016).

The author of the 2016 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2015 to the chair of the committee, Louis-Paul Rivest (Louis-Paul.Rivest@mat.ulaval.ca).

Previous Waksberg Award honorees and their invited papers are:

2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.
2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.
2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.
2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.
2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.
2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.
2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.
2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.
2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.
2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.
2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.
2012 Lars **Lyberg**, "Survey Quality". *Survey Methodology*, vol. 38, 2, 107-130.
2013 Ken **Brewer**, "Three controversies in the history of survey sampling". *Survey Methodology*, vol. 39, 2, 249-262.
2014 Constance F. **Citro**, "From Multiple Modes for Surveys to Multiple Data Sources for Estimates". *Survey Methodology*, vol. 40, 2, 137-161.
2015 Robert M. **Groves**, Manuscript topic under consideration.

Members of the Waksberg Paper Selection Committee (2014-2015)

Louis-Paul Rivest, *Université Laval* (Chair)
J.N.K. Rao, *Carleton University*
Kirk Wolter, *National Opinion Research Center*
Tommy Wright, *U.S. Bureau of the Census*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)
Mary E. Thompson (2011 - 2012)
Steve Heeringa (2012 - 2013)
Cynthia Clark (2013 - 2014)

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 30, No. 2, 2014

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 30, No. 3, 2014

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 42, No. 3, September/septembre 2014

CONTENTS TABLE DES MATIÈRES

Volume 42, No. 4, December/décembre 2014