C.3

SURVEY METHODOLOGY

STATISTICS CANADA
STATISTIQUE CANADA
JUL 11 1995
LIBRARY
BIBLIOTHÈQUE

Catalogue 12-001

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1995
•
VOLUME 21
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 1995 • VOLUME 21 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 1995

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission from Licence Services, Marketing Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

June 1995

Price: Canada: $45.00
United States: US$50.00
Other Countries: US$55.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members
D. Binder
B.N. Chinnappa
G.J.C. Hole
F. Mayda (Production Managef)
C. Patrick
R. Platek (Past Chairman)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistics Canada*
M.J. Colledge, *Australian Bureau of Statistics*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistics Canada*
J.-J. Droesbeke, *Université Libre de Bruxelles*
W.A. Fuller, *Iowa State University*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *University of Maryland*
D. Holt, *University of Southampton*
G. Kalton, *Westat, Inc.*
A. Mason, *East-West Center*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *George Washington University*
J. Sedransk, *State University of New York*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors N. Laniel, M. Latouche, L. Mach, H. Mantel and D. Stukel, *Statistics Canada*

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. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue No. 12-001) is $45 per year in Canada, US $50 in the United States, and US $55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 21, Number 1, June 1995

CONTENTS

In This Issue

This issue of the *Survey Methodology* journal contains a special memorial section in honour of Stanley L. Warner, which includes an introduction by C.-E. Särndal, a bibliography of Warner's principal publications and papers, organized by topic, and three papers dealing with areas in which Warner was a pioneer. The first paper, by Fienberg and Jazairi, summarizes Warner's work in the area of statistically balanced information technology, in which the goal is to develop statistical procedures to ensure that different positions regarding a policy or an issue are fairly and adequately represented in a debate or decision. The other two papers, by Bellhouse, and by Mangat, *et al.*, are in the area of randomized response. The paper by Bellhouse begins with an overview of Warner's contributions to randomized response and then discusses the problem of estimation of a correlation coefficient using data from a randomized response survey. Three randomized response setups are considered: unrelated question, additive constant, and multiplicative constant. The paper by Mangat *et al.* compares the efficiencies of with and without replacement sampling in the context of randomized response. The papers by Fienberg and Jazairi and by Bellhouse are based on presentations given at a special session in memory of Warner at the meetings of the Statistical Society of Canada in Banff in 1994.

The next two papers, by Lavallée and by Kalton and Brick, discuss weighting schemes for cross-sectional estimation in panel surveys.

Lavallée presents the Weight Share Method, used in cross-sectional estimation for longitudinal surveys, in a more general context. He demonstrates the unbiasedness of the method and obtains a general expression for the variance of the estimator of a total. Then the author illustrates the method by applying it in the context of the Survey of Làbour and Income Dynamics of Statistics Canada. The estimation of variance is also discussed.

Kalton and Brick describe weighting schemes for cross-sectional analysis of later waves of a household panel survey using data for all households for whom data are collected. These weighting schemes can accommodate new entrants to the population who move in to live with members of the original population, but not other new entrants. The authors discusses cases where the schemes are optimal as well as further weighting adjustments to compensate for nonresponse and noncoverage.

Small area techniques for estimation of net undercoverage of persons in population censuses are discussed by Dick in the Canadian context and by Kim, Zaslavsky and Blodgett in the U.S. context.

The paper by Dick describes modelling that was done in order to produce estimates of census net undercoverage of persons within age-sex-province categories for the 1991 Canadian Census using data from the Reverse Record Check and the Overcoverage Study. An Empirical Bayes model for direct estimates of adjustment factors is formulated and used to obtain smoothed estimates of those adjustment factors. The smoothed estimates of net missed persons are then raked to match the direct estimates of national age-sex group and provincial totals, which are considered to be of good quality.

Kim, Zaslavsky and Blodgett describe the two analyses performed to test the "synthetic assumption" of homogeneity of undercount rates between parts of different states falling in the same poststratum for the 1990 U.S. Census. In the first analysis, the distributions of five "surrogate variables" that, like undercount, were related to the census-taking process, were investigated using a large extract from the census. In the second analysis, the distribution of undercount was analyzed using the Post Enumeration Survey data.

Breidt presents Markov chain designs for one-per-stratum sampling which includes systematic sampling, stratified simple random sampling and balanced systematic sampling as special cases. He introduces new designs that are shown to be competitive, in terms of the efficiency of the Horvitz-Thompson estimator of a total, with standard one-per-stratum designs under a variety of superpopulation models. Theoretical and numerical comparisons are provided.

Meeden considers the problem of estimation of the median when an auxiliary variable is available. He uses a non-informative Bayesian approach based on a Polya posterior for the ratios of the variable of interest to the covariate. The resulting estimator is empirically compared to a number of alternatives in terms of bias and average absolute error for a variety of real and synthetic populations. The Polya posterior is also to be used to generate interval estimates which are evaluated empirically. Robustness of the procedure to moderate departures from the assumptions is also considered.

Hulliger develops design-based M-estimators for samples with unequal inclusion probabilities. He expresses the Horvitz-Thompson (HT) estimator as a least square functional and then makes it robust against outliers through M-estimators, analogous to the robustification of least square estimators in linear models for infinite populations. He also provides an approximation to the sampling variance of this robustified HT-estimator and its estimate. The results of the Monte-Carlo study confirm that the robustified HT-estimators outperform the HT-estimator in many outlier situations.

Iachan and Kemp describe the sampling designs for two visitor sample surveys of recreational users of parks, a survey of National Park Service area users over a one year period, and a survey of users of three river basin in the Pittsburgh area. The potential problems associated with sampling in both time and space are described, and the ways in which the designs of these two surveys meet these challenges are compared and contrasted.

The Editor

Survey Methodology, June 1995
Vol. 21, No. 1, pp. 3–4
Statistics Canada

Stanley L. Warner

1928-1992

Born and educated in the United States, Stanley Warner received a Ph.D. in Economics from Northwestern University in 1961. In 1971, he moved to Canada, where he was to pursue the rest of his academic career as professor at York University in the Department of Economics and the Faculty of Administrative Studies. He died suddenly in August 1992 at the age of 63.

Stan was a highly original thinker. His statistical research was guided by relevance and common sense. In his memory, a session was organized at the Statistical Society of Canada meetings at Banff in 1994. Two of the papers which follow were presented at that occasion, namely, David Bellhouse: "Estimation of Correlation in Randomized Response" and Stephen E. Fienberg and Nuri Jazairi: "Stanley Warner's Contributions to Statistically Balanced Information Technology." They deal with two areas where Stan made pathbreaking contributions.

Stan Warner is known to many, especially to survey statisticians, as the man who invented randomized response. This technique is used in surveys with sensitive questions as in a survey concerning drug usage among high school students. The objective is to eliminate two embarrassing nonsampling errors: measurement error and nonresponse. Bias caused by such errors is a problem that has haunted statisticians since the beginnings of survey taking. Standard methods exist to "adjust for" such bias. They may reduce but do not eliminate the bias. Stan, disregarding the conventional wisdom, tackled the problem in a completely unorthodox way and came up with an unbiased solution useful at least for some surveys. His seminal article, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", appeared in the *Journal of the American Statistical Association* in 1965.

Randomized response, if carried out according to the intentions, guarantees the anonymity of the respondent: a "yes" answer will not identify the respondent since his or her question is selected at random. Unbiased estimation of the "yes" proportion in the population is nevertheless possible (usually at the price of some increase in variance) because the survey taker knows the number of "yes" responses as well as the probability with which the random choice device selects the sensitive question (rather than its opposite or a completely unrelated question).

The idea struck the imagination of many statisticians. Numerous modifications, refinements and extensions were given, as evidenced through the 1988 bibliography of contributions to randomized response put together by Chaudhuri and Mukerjee. Why this stream of papers? That nonresponse bias was always an important practical problem without a satisfactory solution is not the whole explanation. There is also the fact that it seemed like magic, even to experienced statisticians, that valid answers could be obtained without knowing what question had been asked of the respondent. Of course now that the idea exists, it is not hard to explain; in fact it has become a favorite classroom example, useful even in an elementary statistics course, to show students the powers of statistical reasoning.

Implementing randomized response in practice requires some special arrangements, including a choice device that randomly selects a question. As time went by, Stan realized that the technique had to be adapted to modern low cost data gathering. As late as 1989 at the International Statistical Institute meetings in Paris, he presented a "quick randomized response" version suitable for telephone surveys and touch tone entry, thereby reducing time and cost.

The paper by David Bellhouse traces the development of randomized response in more detail.

Later, Stan's interest focused on statistical procedures for balanced information, which occupied him from about 1975 and on. In the last few years of his life he was working on a book on the topic; the manuscript is now being prepared for publication.

The goal of balanced information technology is to give statistical procedures to ensure that different positions regarding a policy or an issue can be fairly and adequately represented. Stan's first paper on this topic, entitled "Advocate Scoring for Unbiased Information", appeared in the *Journal of the American Statistical Association* in 1975. It deals with the situation in which advocates are each to give *pro* and *con* information regarding an issue to a number of individuals who, after exposure to the information, are to express opinions for or against the issue. Each advocate is charged with using the given data to prepare separate *pro* and *con* cases.

Decisions are frequently made on information provided by advocates; this occurs in government, education, law, *etc*. One can imagine that Stan was concerned with the incomplete and sometimes arbitrary way in which quantitative information is used in decision making of the utmost importance, including political summits, where prestige, mistrust and political consideration would often prevent the parties from meeting on even ground.

The paper by Stephen E. Fienberg and Nuri Jazairi presents the main ideas of this work, which is probably less familiar to statisticians.

Another example of Stan's creative spirit is the fact that he developed, with his wife, a musician, a system for music notation which is in wide use.

I did not know Stan Warner at the time of his original work on randomized response but keep vivid memories of conversations with him later in his career. These occasions were not that numerous; however, each left a strong impression on me. The warmth, modesty and unassuming manner of this highly original man could not fail to make an impression. Stan was a real scholar, not a run-of-the-mill researcher: he did not hesitate to follow the sometimes lonely route laid out to him by belief in ideas that were truly his own.

C.-E. Särndal

Survey Methodology, June 1995
Vol. 21, No. 1, pp. 5–6
Statistics Canada

Principal Publications and Papers of Stanley L. Warner

1928-1992

1. PH.D. THESIS AND RELATED PUBLICATIONS

(1962) *Stochastic Choice of Mode in Urban Travel. A Study in Binary Choice*, Evanston, Illinois: Northwestern University Press.

(1963) Multivariate regression of dummy variates under normality assumptions. *Journal of the American Statistical Association*, 58, 1054-1063.

(1967) Asymptotic variances for dummy variate regression under normality assumptions. *Journal of the American Statistical Association*, 62, 1305-1314.

2. RANDOMIZED RESPONSE

(1965) Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

(1971) The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

(1976) Optimal randomized response models. *International Statistical Review*, 44, 205-212.

(1976) With F. Leysieffer. Respondent jeopardy and optimal randomized response models. *Journal of the American Statistical Association*, 71, 649-656.

(1979) Extended randomized response applications. In *Ethical and Legal Problems in Applied Social Research*, (Eds. R. Boruch, J. Ross and J.C. Cecil), Evanston, Illinois: Northwestern University Press.

(1986) The omitted digit randomized-response model for telephone and other applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.

(1989) Using randomized response for forecasting dimensions of the AIDS problem. Invited keynote presentation. *The Ninth International Symposium on Forecasting*, Vancouver.

(1989) Quick randomized response. *Proceedings of the 47th Session, International Statistical Institute*, Paris, Contributed paper, 431-432.

3. BALANCED INFORMATION

(1975) Advocate scoring for unbiased information. *Journal of the American Statistical Association*, 70, 15-22.

(1977) Advocate scoring design for technological and social policy assessment. *Proceedings of the 41st Session, International Statistical Institute*, New Delhi, Book 3 - Invited Papers, 373-379.

(1979) Subjective information in statistics. *Proceedings of the Business and Economics Section, American Statistical Association*, 558-563.

(1981) Balanced information, the Pickering Airport experiment. *The Review of Economics and Statistics*, LXII, 256-262.

(1984) The overlapping information survey model for evaluating summary information. *Proceedings of the Social Statistics Section, American Statistical Association*, 581-584.

(1985) Applications of the overlapping information model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 401-403.

(1985) The overlapping information model for measuring summary information. *Proceedings of the 45th Session, International Statistical Institute*, Amsterdam, Contributed paper, 49-50.

(1987) Identifying rational opinion-formation with the overlapping model. In *Applied Probability, Stochastic Processes and Sampling Theory*, (Eds. I.B. MacNeill and G.J. Umphrey). Dordrecht, The Netherlands: D. Reidel, 323-329.

(1987) Using test populations to develop balanced agenda. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.

Statistically Balanced Information Technology, (Manuscript of Book).

4. CLINICAL TRIALS

(1981) Post-treatment randomization in clinical trials. *Proceedings, Statistics Eighty One Canada*, Concordia University Canada.

(1981) Post-treatment randomization in clinical research. *Proceedings of the Social Statistics Section, American Statistical Association*, 233-236.

(1982) Post-treatment randomization extensions for medical and educational research. *Proceedings of the Social Statistics Section, American Statistical Association*, 410-413.

(1983) Post-treatment randomization estimates. *Proceedings of the 44th Session, International Statistical Institute*, Madrid, 464-468.

5. OTHERS

(1958) With R.L. Andreano. Professor Bain and barriers to new competition. *Journal of Industrial Economics*, 66-78.

(1965) Cost models, errors in variables and economies of scale in trucking. *The Cost of Trucking, Econometric Analysis*, Dubuque: William C. Brown and Co., 1-46.

(1983) With W.D. Cook and L. Seiford. Preference ranking models: Conditions for equivalence. *Journal of Mathematical Sociology*, 9, 125-137.

Stanley Warner's Contributions to Statistically Balanced Information Technology

STEPHEN E. FIENBERG and NURI JAZAIRI[1]

ABSTRACT

Stanley Warner was widely known for the creation of the randomized response technique for asking sensitive questions in surveys. Over almost two decades he also formulated and developed statistical methodology for another problem, that of deriving balanced information in advocacy settings so that both positions regarding a policy issue can be fairly and adequately represented. We review this work, including two survey applications implemented by Warner in which he applied the methodology, and we set the ideas into the context of current methodological thinking.

KEY WORDS: Advocate scoring; Bayes' Theorem; Embedded experiment; Logistic regression; Survey analysis.

1. INTRODUCTION

Consider some recent controversial public or professional issues such as:

1. Should Canada endorse the North American Free Trade Agreement?
2. Should Quebec secede from the Canadian Federation?
3. Should the American Statistical Association adopt a program to certify statisticians?
4. Should smoking be banned in all restaurants in Ottawa?

The discussions and debates surrounding such issues often reflect highly polarized positions and "pro" and "con" arguments can strongly influence the opinions of individuals in the relevant populations of interest (*e.g.*, Canadian residents, ASA members, those who frequent restaurants in Ottawa). How to think about the presentation of such advocacy information in a balanced fashion is the topic of this paper.

It has often been said that only a small fraction of scientists make a truly novel research contribution once in their lifetime. Far fewer are responsible for multiple innovations. Stanley Warner is well-known for his creation and development of the randomized response model for surveys and that contribution has been widely hailed as a major development in statistics. What is less well known is his truly novel approach to the problem of balanced information in advocacy settings, on which he worked over a period of almost 20 years. As York University colleagues of Warner's at the time of his death in 1992, we know how seriously he took the obligation of statistics and statisticians to deal with such complex problems, and this work is one example of how he attempted to fulfill the obligation.

Our goal in this paper is to reintroduce Warner's ideas on the topic of balanced information in advocacy settings to the profession and to demonstrate how they fit into current survey practice and methodological thinking. In Section 2, we present his basic approach to the advocacy problem and we describe the statistical model he chose to focus upon (Warner 1975). In Sections 3 and 4, we discuss embellishments of the basic approach which he presented in subsequent papers (*e.g.*, see Warner 1981, 1984, 1985, 1987a), and we end by describing how Warner continued to pursue this research program up until the time of his death. In the process, we also stress the importance Warner attached to the application of his ideas.

2. THE BASIC PROBLEM

In a thoughtful, well-argued, yet provocative 1975 paper in the *Journal of the American Statistical Association*, Stanley Warner first presented the issue of measuring the impact of advocacy and balance on public opinion in connection with controversial issues. He did so by asking (and then answering) a pair of interrelated questions:

1. How can we estimate what the population would conclude on issue were each of the members provided with balanced information on the topic?

Warner's idea for answering this question was to use advocates to present summaries of arguments, both "pro" and "con," and to implement this in a factorial experimental design to different samples, and in the process achieve information about a balanced presentation. This then leads rather naturally to the second question:

2. How can we rate or score advocates in such settings?

He developed his formulation to answer the two questions simultaneously and, in doing so, he used *both* economical and statistical arguments. In this paper, we focus on the statistical portion of his arguments and refer the interested reader to Warner's paper for the economic details.

[1] Stephen E. Fienberg, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.; Nuri Jazairi, Department of Economics, York University, North York, Ontario, Canada, M3J 1P3.

Consider a pair of advocates or advocate teams whose role it is to brief individuals on the arguments associated with a controversial issue, H. Let $P(H)$ and $P(\bar{H})$ denote the proportion of the number of subjects in a given population "for" and "against" issue H. Let F_i and A_j denote "pro" and "con" presentations of advocates i and j, respectively for $i, j = 1, 2$. Let $P(H \mid F_i, A_j)$ and $P(\bar{H} \mid F_i, A_j)$ denote the number of subjects "for" and "against" issue H after hearing "pro" case from advocate i and "con" case from advocate j.

Warner defined the "net information" associated with F_i and A_j as

$$I(F_i,A_j) = ln[P(H \mid F_i,A_j)/P(\bar{H} \mid F_i,A_j)] - ln[P(H)/P(\bar{H})]. \quad (1)$$

Formula (1) is, of course, the logarithm of the Bayes factor, or what Good (1950) called the *weight of evidence.* While Warner recognized the evocative nature of the use of Bayes' Theorem here, his approach towards its use was purely frequentist.

Similarly, Warner defined the net information associated with "pro" and "con" cases of F_i and A_j, separately:

$$I(F_i) = ln[P(H \mid F_i)/P(\bar{H} \mid F_i)] - ln[P(H)/P(\bar{H})], \quad (2)$$

$$I(A_j) = ln[P(H \mid A_j)/P(\bar{H} \mid A_j)] - ln[P(H)/P(\bar{H})]. \quad (3)$$

The simplest assumption we can make relating the joint and marginal information quantities is that of independence of the "pro" and "con" cases,

$$I(F_i,A_j) = I(F_i) - I(A_j), \quad (4)$$

for $i = 1, 2$, and $j = 1, 2$. This assumption allows for some direct comparisons and, as we shall see, can be checked empirically.

In order to ensure that the advocates fairly treat both "pro" and "con" positions, Warner proposed to reward them on the basis of the sum of the net information they provided, *i.e.*

$$I(F_i) + I(A_j). \quad (5)$$

Economic theory, Warner argued, suggests that rewarding advocates in this fashion will lead them to at least strive to approximate the "unbiased information" associated with maximization under resource constraints. Thus we need to estimate the quantity in expression (5) along with the posterior odds implied by unbiased information:

$$P(H \mid F',A') \mid P(\bar{H} \mid F',A'). \quad (6)$$

"Balance" in design for data collection was the key to Warner's plan for estimation.

Warner's estimation plan was linked to his application. The controversial issue was the completion of the north-south Spadina Expressway in Toronto (Warner's home city). The original first section of the expressway was constructed in 1966 and, after much debate, the remainder of the project was canceled in 1971. Two years later, in 1973, Warner conducted a survey to learn what proportions of the population of registered voters of Metropolitan Toronto were for or against the original expressway plan. He took a random sample of 1,360 registered voters (1% of the corresponding population) divided into 8 equal subsamples of size 170. Two advocate teams prepared written positions, both "pro" and "con" the expressway, and one of each was included in the mailing. The order of presentation of the two written positions was also varied producing a 2 × 2 × 2 experimental design with the first variable corresponding to who prepared the "pro" brief, the second to who prepared the "con" brief, and the third to the order of presentation ("pro" first or "con" first). Advocates were paid a basic fee and a larger amount was set aside to be paid to the team with the "best combined score." This is an excellent example of a factorial experiment embedded within a survey, and fits well with the spirit of embedding described in Fienberg and Tanur (1988).

Table 1

Sample Preferences for Spadina Expressway After Information by Advocates

Sample	i	j	k	For	Against	Undecided	Total	p_{ijk}	n_{ijk}
1	1	1	1	22	4	1	27	.846	26
2	1	1	2	18	9	2	29	.666	27
3	1	2	1	26	8	0	34	.764	34
4	1	2	2	21	11	1	33	.656	32
5	2	1	1	28	10	1	39	.736	38
6	2	1	2	14	11	1	26	.560	25
7	2	2	1	19	16	1	36	.542	35
8	2	2	2	19	17	2	38	.527	36

Source: Warner (1975).

In the cover letter, Warner asked respondents to return prepaid postcards indicating their preferences after reviewing the briefs. At the cut off date, 262 cards had been returned for a response rate of about 20%. The resulting data, in Table 1, are reproduced from Warner (1975).

Let p_{ijk} be the true proportion of the population "for" the expressway, in group (i, j, k). Then, with an additive term for order of presentation, the model of expression (1) becomes

$$ln[p_{ijk}/(1 - p_{ijk})] = ln[P(H)/P(\bar{H})] + I(F_i) - I(A_j) + D_k. \quad (7)$$

We now recognize expression (7) as a linear logit model, and the sampling scheme as product-binomial (ignoring the correction for the 0.2% sampling fraction). Of course when Warner did this work it preceded the existence of a monograph by Bishop, Fienberg, and Holland (1975), and was virtually concurrent with Nelder and Wedderburn's (1972) paper on generalized linear models. Thus his paper made no reference to the now extensive literature on logit and loglinear models.

To estimate the parameters in expression (7), Warner used weighted least squares, which yields both estimated coefficients and standard errors. Instead of dealing directly with the parameters in expression (7), he redefined them, in part to simplify computation and in part to aid in their interpretation:

$$\beta_1 = ln \frac{P(H)}{P(\bar{H})} + \frac{I(F_1) - I(A_1)}{2} + \frac{I(F_2) - I(A_2)}{2} + \frac{D_1 + D_2}{2}, \quad (8)$$

$$\beta_2 = I(F_1) - I(F_2), \quad (9)$$

$$\beta_3 = I(A_1) - I(A_2), \quad (10)$$

$$\beta_4 = D_1 - D_2. \quad (11)$$

The coefficient β_1 is an "intercept" or normalizing parameter, while β_2, β_3, and $\beta_2 + \beta_3$ measure the performance of the advocate teams, and β_4 measures the order effect. The net information provided by team 1 is $\beta_1 + .5(\beta_2 - \beta_3)$, and that provided by team 2 is $\beta_1 - .5(\beta_2 - \beta_3)$. The difference in net influence is thus $\beta_2 - \beta_3$.

Table 2

Weighted Least Squares Estimates of Theoretical Parameters

Parameter	Estimate	Approx. Std. Error
β_1	.712	.139
β_2	.648	.277
β_3	−.383	.275
β_4	.528	.274
$\beta_2 + \beta_3$	.264	.386
$\beta_1 + .5\beta_2 - .5\beta_3$	1.228	.266
$\beta_1 - .5\beta_2 + .5\beta_3$	.196	.215
$\beta_2 - \beta_3$	1.032	.395

Source: Warner (1975).

We reproduce Warner's estimation results in Table 2. We have double-checked the estimated values in Table 2 using the generalized model routines in $S+$, which utilize a version of iteratively weighted least squares (maximum likelihood in this case). Our logit model computations agree with Warner's to two decimal places. The residual deviance for this model equals 1.95 with 4 d.f., indicative of a remarkably good fit and offering strong support for the reasonableness of the independence assumption of expression (4).

In interpreting the results in Table 2, Warner noted that his economic analysis leads to the conclusion that the overall proportion of the population in favour of H when presented with unbiased information lies between the "pure" estimates for the 2 advocate teams, or in the present instances (.55, .77). These bounds correspond to the estimates in the 2nd and 3rd last lines of the table. As was clear from Table 1, no matter how we combine "pro" and "con" arguments, the majority in each subgroup favored completion of the expressway. Warner observed that we might be tempted to use $\hat{\beta}_1$ to produce a "best estimate" of the value of p corresponding to unbiased information, but he argued for a higher value, since Team 1 is superior to Team 2 in terms of total information, *i.e.*, $\hat{\beta}_2 - \hat{\beta}_3 > 0$. (The superiority of Team 1 is quite evident from a quick examination of Table 1 and does not require the full analysis.)

Warner ended his 1975 paper by pointing out all of the shortcomings of his small experiment, and his initial modelling efforts. What we can observe in retrospect is the way in which he was able to attack a very complicated public policy and survey problem using a simple but ingenuous model, as well as a rigorous estimation scheme built on the solid framework of a factorial experiment embedded in a sample survey, and then actually applying the methodology to produce an answer for a real problem.

It is worth noting that the first version of this paper was submitted for publication to *JASA* in June 1972, before Warner had actually carried out the empirical study on the Spadina Expressway controversy. Over two years passed before he resubmitted a revised version of the paper with the detailed example. Even well-known authors with innovative ideas often struggle to have their work published in major statistical journals, and a compelling empirical application is always of help.

3. EXTENSIONS AND A SECOND APPLICATION

Warner extended his balanced information approach in a second paper (Warner 1981), focusing on yet another application. This paper also signals a substantial change in Warner's thinking about statistics and probability, towards a subjective Bayesian approach and away from the classical approaches that he stressed in his early career. While the reported analyses are still frequentist in nature, Warner used, at least informally, the assessments of prior probabilities in a manner that fits rather naturally with the Bayesian formulation of expression (1) above.

In March 1972, the Canadian Federal Government announced a plan to build a second Toronto International Airport to the east of the city in Pickering, Ontario. This led to considerable controversy. In 1974, the government appointed a 3-person commission of inquiry. Warner carried out a concurrent but independent survey experiment. The question he posed was whether or not the Pickering Airport should be built before the year 2000. The general structure of the experiment was similar to the previous one on the Spadina Expressway controversy, but with some differences:

(i) This time his study population was economists.

(ii) He incorporated 2 "neutral" control sub-samples, which received neither "pro" nor "con" statements.

(iii) Respondents in the 8 experimental subsamples gave probability assessments (instead of 0-1 values) after assessing the advocacy positions. Those in the control groups also gave their probability assessment.

The test population was limited to those economists who belonged to the Canadian Economic Association or who could be identified as professors or lecturers in an economics department in a Canadian university. The survey was done via mail in two stages – the first identified those willing to read detailed briefs and "report opinions regarding an undisclosed federal project," and the second mailing divided those willing to participate into 10 subsamples, corresponding to the 2 × 2 × 2 design of Section 2 plus the 2 control samples consistent of those who were asked for their opinions without briefs. A total of 726 economists participated in the experiment. In Table 3, we provide Warner's summary of the data for the 8 experimental subsamples in which he aggregated the posterior judgments into three groups according to whether they were substantially greater, nearly equal, or substantially less than 0.5. The data have been further aggregated across the 8 experiment groups. The results have been post-stratified according to whether the economists were professors, graduate students, or others. The data on "prior beliefs" come from a combination of the two control groups.

Table 3

Test Population Opinions on Pickering Airport

	Professors		Students		Others		Totals	
	Before Briefs	After Briefs	Before Briefs	After Briefs	Before Briefs	After Briefs	Before Briefs	After Briefs
For	9 (.143)	58 (.266)	9 (.257)	32 (.288)	11 (.180)	71 (.298)	29 (.182)	161 (.284)
Against	32 (.508)	155 (.711)	12 (.343)	72 (.648)	36 (.590)	160 (.672)	80 (.503)	387 (.683)
Undecided	22 (.349)	5 (.023)	14 (.400)	7 (.063)	14 (.230)	7 (.029)	50 (.315)	19 (.033)
Totals	63 (1.000)	218 (1.000)	35 (1.000)	111 (1.000)	61 (1.000)	238 (1.000)	159 (1.000)	567 (1.000)

Source: Warner (1981).

Note that all three groups had substantial negative opinions about the proposed airport, a posteriori, and that the differences in proportions of undecided between the experimental and control subsamples provide evidence that the advocacy briefs affected public opinion on the issue. Warner's formal statistical analysis of the data focused solely on the 8 experimental subsamples and utilized three variants of the formal model in expression (9) and the reparameterization of expressions (10) through (13):

(i) A logit structure similar to that in Warner (1975) based on the aggregation in Table 3, with "undecideds" in effect *imputed* as belonging in either the "pro" or "con" categories with probability 0.5. He called this a *Simple Aggregate Influence* model.

(ii) A more direct approach, which averaged the posterior assessments to get "aggregate proportions" in favor, and then treated these observed proportions as if they were binomial. He called this a *Weighted Aggregate Influence* model.

(iii) A two-stage model, which first used individual-level assessments breaking up the range of 0 to 1 into 17 levels, and then a "variable coefficient" regression model analysis. He referred to this as an *Average Disaggregate Influence* model.

Each analysis involved the use of a different form of weighted least squares to estimate the coefficients of interest.

In Table 4, we provide Warner's estimated coefficients under all three models and analyses. The results are similar across models and we can summarize the findings as follows:

(a) Team 2 clearly presented the strongest case ($\hat{\beta}_2$ and $\hat{\beta}_3$ are both positive and similar for all three columns).

(b) The estimated aggregate influence for Team 2 is $[\hat{\beta}_1 + .5(\hat{\beta}_2 - \hat{\beta}_3)] = -0.688$ corresponding to an estimated proportion in favor of the airport project of $\hat{p} = 0.355$.

(c) The disaggregate influence for Team 2 corresponds to an estimated proportion in favor of the airport project of $\hat{p} = 0.355$.

(d) The effect of order of presentation ($\hat{\beta}_4$) suggests that the brief appearing first in the enclosures had greater impact, and is consistent with the hypothesis that the "previous information favoring one position serves to discount new information against that position."

It turns out that the advisor for Team 1 felt that construction of the airport could not be defended and this seriously handicapped the "pro" efforts of Team 1 (something reflected in the estimates of β_2).

Table 4
Estimated Case Influence for Pickering Airport Experiment

Parameters	Simple Aggregate Influence	Weighted Aggregate Influence	Average Disaggregate Influence
β_1	−.857 (.093)	−.529 (.047)	−.736 (.065)
β_2	.485 (.188)	.337 (.097)	.462 (.132)
β_3	.147 (.188)	.146 (.097)	.187 (.132)
β_4	.313 (.186)	.209 (.095)	.307 (.129)
N	8	8	567

Source: Warner (1981).

4. OVERLAPPING INFORMATION

Warner worried that the information used in the Pickering Airport survey experiment involved an overlap between the "pro" and "con" cases, and there was also an overlap between the prior information available to the respondents and that presented by the advocates. He turned to this question several years later in Warner (1984, 1985), using a formal argument drawn from sampling theory.

Warner's idea was to consider N pieces of independent information being used to influence the proposition in question. Let Y_{ij} be the information content seen by the i-th person in the j-th piece of information. Then the prior odds for the i-th individual is

$$ln[p_i/(1 - p_i)] = \sum_{j \epsilon A(i)} Y_{ij}, \qquad (12)$$

where $A(i)$ is the collection of information seen by the i-th person prior to the presentation of the advocacy arguments. If $A(i)$ is empty, the initial log-odds for individual i should be 0 and thus $p_i = 0.5$.

The presented "pro" and "con" summaries draw on a subset, S, of m out of the N units. Suppose that participants act "rationally" and are not further influenced by data which has been seen before. The added information is then

$$\sum_{j \epsilon S} Y_{ij} - \sum_{j \epsilon A(i) \cap S} Y_{ij}. \qquad (13)$$

If the m units of information are randomly selected without replacement from the total N, then this implies that we can treat the units in $A(i) \cap S$ as having been selected at random without replacement. We can treat the information from these overlapping units as following a hypergeometric distribution, and then we rewrite expression (1) as

$$I_i = m\bar{S}_i - m/N[p_i/(1 - p_i)] + \epsilon, \qquad (14)$$

where I_i is the net information in the summary for the i-th individual, $\bar{S}_i$ is the average of the Y_{ij}'s for those data units $j \epsilon S$, and the error term ϵ has zero conditional expectation, *i.e.*,

$$E[\epsilon \mid [p_i/(1 - p_i)]] = 0. \qquad (15)$$

If we group subjects in an advocacy experiment exposed to the same "pro" and "con" briefs according to the values of p_i, then differences in net information should be related to $[p_i/(1 - p_i)]$ according to equation (14). Warner (1984) did this with additional simplifying assumptions and then analyzed the data from the Pickering experiment using assessments, the control groups and the Team 2 "pro" and "con" group, aggregated according to whether the respondent was a professor, a student, or other. The problem with using the data from the Pickering experiment is that we are in effect matching the individuals in the control group and the experimental group. What we really want is both the prior and the posterior assessments from the same individual (Warner 1987b).

Warner (1985, 1987a) returned to this theme of overlapping information, and he extended the model of expression (14) to take the form:

$$I_i = m\bar{S}_i + D_i r_i(Z_i - U_i), \qquad (16)$$

where we have in effect replaced the coefficient $-m/N$ in expression (14) by $D_i r_i$, where $D_i \geq -1$ is a discount factor and $E(r_i) = m/N$. He then showed how to estimate the coefficients in this "random coefficients" regression model using generalized least squares, under various assumptions about the correlations among the quantities in (16).

He then applied the approach to a new set of data collected for a telephone survey of Carleton University students on the question of whether an elected Canadian Senate would be preferable to the existing appointed Senate. The interviewees were asked for their opinion on the issue expressed as a probability. They were then presented with a 6-sentence summary of a television debate on the topic, and asked to reconsider their probability assessments. Of the 417 participants, 316 gave prior probabilities different from 0 and 1. Of this group, 163 actually changed their assessment, and overall the average log-odds after the summary was virtually the same as it was before, but with a slightly smaller variance. Warner actually fit the model to the data, and the fitted equations were consistent with the notion of partial discounting of the information that they had already seen.

This was essentially Warner's last published contribution to the topic of balanced information assessment. At the time of his death, Warner was hard at work on a book-length manuscript whose title, *Statistically Balanced Information Technology*, suggests that he was attempting to synthesize and extend his ideas on the topic. Unfortunately, we have only been able to locate the early chapters of this book and these include just the introductory ideas on probability and regression that he expected to utilize in the later chapters.

5. FURTHER OBSERVATIONS

Warner's balanced information technology addresses the common problem of adversarial policy advocacy which may give rise to confusion and incorrect decisions because of imbalance in the presentation of the relevant facts. Examples of how the adversarial approach to dispute resolution in a legal setting could have distorting effects on questions of scientific fact are discussed in Fienberg (1989; see especially Appendix H by Vidmar). Among the responses to this situation have been repeated proposals to establish a science court to ensure balance in organizing the information relevant to a factual dispute and reaching decisions. In these proposals, the science court itself is an adversarial system, but based on well-defined procedures for the selection of issues, advocates, and judges designed to ensure impartiality and minimize the effects of personal bias. Warner's approach outlined here is a formal way to achieve precisely this kind of impartial result.

Warner's progression through the various stages of the work on balanced information was paralleled by a shift in his outlook on the foundations of statistics. He was trained as an economist and a classical statistician and his early statistical contributions, including the work on randomized response models, were all set in a frequentist statistical framework. The 1975 paper on advocate scoring represented his first step towards a subjectivist perspective and, with each successive paper, he added further elements of the Bayesian approach. In Warner (1979), Stan articulated this shift in thinking and it is especially apparent in the early chapters of his unpublished book. At the May 1992 annual meeting of the Statistical Society of Canada in Edmonton, his last public lecture, Stan described devices for the solicitation of probabilities that he had been developing for the book.

We can only speculate about how Stan's subjectivist synthesis of balanced information technology would have looked had he been able to complete the book. But given the depth of his commitment to the Bayesian approach and its recent methodological innovations, we expect that it would have included a hierarchical generalized linear model approach and utilized the latest developments in Markov Chain Monte Carlo simulation techniques.

Stanley Warner was constantly using the ideas from his research in the classroom and in reflecting back upon the work described here, he noted:

> "... almost all of the basic elements of an elementary statistics course are to some degree represented in these procedures, and the problems in modeling and design that are suggested could be considered at quite an advanced level", (Warner 1987b).

The statistics profession has lost a true innovator and a great educator. We count ourselves amongst Stanley Warner's students and we continue to learn from his work.

ACKNOWLEDGMENTS

The preparation of this paper was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada to the first author at York University.

REFERENCES

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.

FIENBERG, S.E., and TANUR, J.M. (1988). From the inside out and outside in: Combining experimental and sampling structures. *Canadian Journal of Statistics*, 16, 135-151.

FIENBERG, S.E., (Ed.) (1989). *The Evolving Role of Statistical Assessments as Evidence in the Courts*. New York: Springer-Verlag.

GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

NELDER, J.A., and WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, A, 135, 370-384.

WARNER, S.L. (1975). Advocate scoring for unbiased information. *Journal of the American Statistical Association*, 70, 15-22.

WARNER, S.L. (1979). Subjective information in statistics. *Proceedings of the Business and Economics Section, American Statistical Association*, 558-563.

WARNER, S.L. (1981). Balanced information, The Pickering Airport experiment. *The Review of Economics and Statistics*, LXII, 256-262.

WARNER, S.L. (1984). An overlapping information survey model for evaluating summary information. *Proceedings of the Social Statistics Section, American Statistical Association*, 581-584.

WARNER, S.L. (1985). Applications of the overlapping information model. *Proceedings of the Section on Survey Research Methods, American Statistical Society*, 401-403.

WARNER, S.L. (1987a). Identifying rational opinion-formation with the overlapping information model. In *Applied Probability, Stochastic Processes, and Sampling Theory*. (Eds. I.B. MacNeill and G.J. Umphrey). Dordrecht, The Netherlands: Reidel, 323-329.

WARNER, S.L. (1987b). Using test populations to develop balanced agenda. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 441-443.

WARNER, S.L. (1992). Statistically Balanced Information Technology. Unpublished manuscript.

Survey Methodology, June 1995
Vol. 21, No. 1, pp. 13–19
Statistics Canada

Estimation of Correlation in Randomized Response

D.R. BELLHOUSE[1]

ABSTRACT

Stanley Warner's contributions to randomized response are reviewed. Following this review, a linear model, based on random permutation models, is developed to include many known randomized response designs as special cases. Under this model optimal estimators for finite population variances and covariances are obtained within a general class of quadratic design-unbiased estimators. From these results an estimator of the finite population correlation is obtained. Three randomized response designs are examined in particular: (i) the unrelated questions model of Greenberg *et al.* (1969); (ii) the additive constants model of Pollock and Bek (1976); and (iii) the multiplicative constants model of Pollock and Bek (1976). Simple models for response bias are presented to illustrate the effect of this bias on estimation of the correlation.

KEY WORDS: Additive constants model; Linear models; Multiplicative constants model; Response bias; Unrelated question model; Variance estimation.

1. A BRIEF OVERVIEW OF WARNER'S CONTRIBUTIONS TO RANDOMIZED RESPONSE

Randomized response is a technique used to elicit responses to sensitive questions. It was developed thirty years ago by Stanley Warner (Warner 1965) to estimate a proportion under a simple random sampling design with replacement. The development was a substantial intellectual achievement requiring much originality of thought. How does one get truthful responses to sensitive questions? Warner's solution was to get the response without the interviewer knowing whether the sensitive question had actually been asked. He devised the probabilistic structure to the questioning so that an estimate of the required proportion could be obtained. In Warner's original formulation the population is divided into two mutually exclusive and exhaustive groups, A and B. It is of interest to estimate the proportion π of the population belonging to group A. To do this, a spinner is constructed with a face marked with the letters A and B. The construction is such that the spinner points to the letter A with probability p and to B with probability $1 - p$. The interviewee spins the spinner and is required only to say yes or no according to whether or not the spinner points to the interviewee's correct membership group. The with replacement design allows estimation of π by maximum likelihood.

This very original idea has received substantial attention over the past thirty years. Since Warner's original work, several randomized response techniques have been suggested for the estimation of a proportion or set of proportions as in polytomous data, or for the estimation of a population mean with continuous data. A variation on Warner's original theme is asking the sensitive question or an unrelated question with probabilities p and $1 - p$ respectively. This was originally due to Greenberg *et al.* (1969). Other variations with continuous data include adding a random variable to the response to the sensitive question or multiplying the response by a random variable. The underlying theme to any of these techniques is the masking of the original response in such a way that the sensitive information cannot be attributed to any single respondent but that information on the sensitive attribute can be extracted from the whole sample. A substantial literature, including a monograph by Chaudhuri and Mukerjee (1988), has grown up around these techniques. Nathan (1988) has provided a fairly comprehensive bibliography of this literature. Umesh and Peterson (1991) have given several detailed examples from very diverse areas of the application and applicability of the techniques of randomized response.

With several different randomized response techniques, the question arises as to how to compare the different methods. Minimization of variance cannot be the sole criterion. Each method is designed to protect the privacy of the respondent. A gain in efficiency, in terms of variance, by the choice of different values of the probabilities in the randomizing device, or by the choice of one randomized response method over another, could lead to jeopardizing the privacy of the respondents. In response to this, Leysieffer and Warner (1976) and Warner (1976) formulated natural measures of respondent jeopardy. These measures are related to the probability of the interviewer being able to infer the interviewee's response to the sensitive attribute. The theory of respondent jeopardy is reviewed in Chaudhuri and Mukerjee (1988) and some practical considerations regarding respondent jeopardy are reviewed in Umesh and Peterson (1991).

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, N6A 5B7.

Stanley Warner made two other contributions to the literature of randomized response. The first contribution is directly related to the results obtained here. With the explosion of new ideas and new techniques in randomized response, Warner (1971) formulated a linear model which unified the theory. Most of the randomized response techniques at that time could be put in his linear model framework. The second contribution was in response to the growing use of telephone interviewing. Stem and Steinhorst (1984) described randomized response methods applicable to telephone interviewing and to mail questionnaires. Warner (1986) suggested practical natural randomizing devices, such as the serial numbers on paper money, for use in telephone interviewing.

The major topics in randomized response methodology are: the development of randomized response techniques, the comparison of these techniques through the concept of respondent jeopardy, the construction of reasonable randomizing devices, the development of a unified theory of randomized response, and the validation of randomized response techniques through field studies. Stanley Warner's contributions to randomized response touch on most of these major developments in the subject. Moreover, most of these contributions were substantial and influential. He is the originator of the technique. His original setup of a dichotomous population was quickly generalized to a polytomous one and to populations with continuous measurement. New randomized response techniques continue to be developed. Warner was at the forefront of evaluating randomized response designs through the modeling of respondent jeopardy. His work in the development of a unified linear model for randomized response designs was the foundation on which a unified theory of randomized response has been built.

2. INTRODUCTION TO ESTIMATION OF CORRELATION

Consider a finite population of size N with two measurements of interest x_j and y_j for $j = 1, \ldots, N$. It is of interest to estimate the finite population correlation

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

where $\sigma_{xy} = \sum (x_j - \bar{X})(y_j - \bar{Y})/N$ is the finite population covariance between the variables x and y and where σ_x^2 and σ_y^2 are the finite population variances of the variables x and y respectively. To estimate ρ a sample of fixed size n is chosen with probability $P(s)$ from the finite population where s denotes the set of finite population units chosen for the sample. The expectation operator with respect to the sampling design $P(s)$ is denoted by E_p. Estimators for ρ are obtained by replacing σ_x^2, σ_y^2 and σ_{xy} by their respective estimators, unbiased or biased, optimal in some sense or otherwise.

To illustrate the general results obtained here for estimation of the finite population correlation coefficient, three particular randomized response techniques will be considered:

(i) The unrelated questions model due to Greenberg *et al.* (1969). The sensitive question is asked with probability p and an unrelated question which is not sensitive is asked with probability $1 - p$. For estimation of the mean it is assumed that the finite population mean $\bar{X}$ of the unrelated question is known. For estimation of variance it is also assumed that σ_x^2 is known.

(ii) The additive constants model due to Pollock and Bek (1976). The outcome of a random variable from a known probability distribution is added to the value of the response to the sensitive question.

(iii) The multiplicative constants model due to Pollock and Bek (1976). The value of the response to the sensitive question is multiplied by the outcome of a random variable from a known probability distribution.

Edgell *et al.* (1986) have provided estimators for ρ under the unrelated questions model and the additive constants model.

Most randomized response designs that have been considered have assumed that the sampling design is simple random sampling either with or without replacement. Since the results obtained here are under a fixed size design, the simple random sampling design assumed here is without replacement.

Assume that both x and y are sensitive variables. Consequently, a randomized response technique is used to obtain information on both these variables. Let w_j and z_j, for $j \in s$ be the sampled measurements that are obtained. Let u_j and v_j for $j = 1, \ldots, N$ be the nonsensitive measurements associated with x_j and y_j respectively. Under the unrelated question model (randomized response model (i)) u_j and v_j are the responses to the unrelated questions for the j-th individual. Under the additive constants model or the multiplicative constants model (randomized response models (ii) or (iii)) u_j and v_j are the j-th outcomes of random variables from two, possibly different, known probability distributions.

3. RANDOM PERMUTATION MODELS

Several models for the finite population measurements have been put forward in the survey sampling literature. Here attention is focused on the random permutation models of Rao (1975) and Rao and Bellhouse (1978). One compelling reason for using these models is that the model parameters have a direct interpretation in the finite population of interest since model parameters in random permutation models are also finite population parameters. In the simplest context for random permutation models it is assumed that the N-dimensional vector of finite population measurements is a random permutation of an N-dimensional vector of fixed numbers. Rao (1975) has shown how this assumption leads to a linear model. Bellhouse (1980) extended this model to randomized response designs under unequal probability sampling.

The model and associated designs applicable to unequal probability sampling are not easily applicable to estimation of variances and covariances either with or without a randomized response. Consequently, a special case of the model in Bellhouse (1980) is given here. In the model which follows there are two different expectation operators at work which together yield a composite expectation E_m. These expectation operators are: E_r, the expectation operator with respect to the randomizing device, and E_{rp}, the expectation operator with respect to the random permutation model. The composite expectations $E_m = E_{rp}E_r$ and $E = E_mE_p$. For the random permutation model we assume that the pairs (x_j, y_j), $j = 1, \ldots, N$ are a random permutation of a set of N fixed pairs of numbers, say (p_j, q_j), $j = 1, \ldots, N$. This is a special case of model (4.1) in Rao and Bellhouse (1978); the more general model in Rao and Bellhouse (1978) was used in double sampling and sampling on two occasions. The unrelated questions randomized response model (randomized response model (i)) requires an additional assumption that the quadruples (x_j, y_j, u_j, v_j), $j = 1, \ldots, N$ are a random permutation of a set of N fixed quadruples of numbers, say (p_j, q_j, r_j, t_j), $j = 1, \ldots, N$.

Assume that the randomizing device coupled with the random permutation model leads to the following linear model:

$$\begin{aligned} w_j &= \alpha_1 + \beta_1\bar{X} + e_{1j} \\ z_j &= \alpha_2 + \beta_2\bar{Y} + e_{2j}, \end{aligned} \tag{1}$$

for $j = 1, \ldots, N$ where $\bar{X}$ and $\bar{Y}$ are the finite population means of the x and y measurements respectively and where for $j = 1, \ldots, N$

$$\begin{aligned} &E_m(e_{1j}) = E_m(e_{2j}) = 0, \\ &E_m(e_{1j}^2) = \phi_1\sigma_x^2 + \psi_{01} + \psi_{11}\bar{X} + \psi_{21}\bar{X}^2, \\ &E_m(e_{2j}^2) = \phi_2\sigma_y^2 + \psi_{02} + \psi_{12}\bar{Y} + \psi_{22}\bar{Y}^2, \\ &E_m(e_{1j}\,e_{1k}) = \delta_1\sigma_x^2 + \lambda_1,\ E_m(e_{2j}\,e_{2k}) = \delta_2\sigma_y^2 + \lambda_2, \\ &\qquad\qquad\qquad\qquad \text{for } j \neq k, \\ &E_m(e_{1j}\,e_{2j}) = \phi_3\sigma_{xy} + \psi_3, \quad \text{and} \\ &E_m(e_{1j}\,e_{2k}) = \delta_3\sigma_{xy} + \lambda_3, \quad \text{for } j \neq k. \end{aligned} \tag{2}$$

and all other higher moments are independent of j. In the model given by (1) and (2), the α's, λ's, ϕ's, ψ's and δ's are all known constants. The finite populations variances and covariances of the sensitive questions, σ_x^2, σ_y^2 and σ_{xy} are all unknown.

For the unrelated questions model (randomized response model (i)) assume that the randomizing schemes on the two sensitive questions are independent and that sensitive question i, $i = 1, 2$, is asked with probability p_i and the associated nonsensitive questions with probability $1 - p_i$. Assume further that the sensitive questions are unrelated to the nonsensitive questions so that $\sigma_{xu} = \sigma_{yv} = \sigma_{xv} = \sigma_{yu} = 0$. This assumption is unnecessary under simple random sampling with replacement. When, in addition, a random permutation model is assumed on the quadruple (x_j, y_j, u_j, v_j) then in the model given by (1) and (2):

$$\begin{aligned} &\alpha_1 = (1 - p_1)\bar{U},\ \beta_1 = p_1,\ \alpha_2 = (1 - p_2)\bar{V},\ \beta_2 = p_2, \\ &\phi_1 = p_1,\ \psi_{01} = (1 - p_1)\sigma_u^2 + p_1(1 - p_1)\bar{U}^2, \\ &\qquad\qquad \psi_{11} = -2p_1(1 - p_1)\bar{U},\ \psi_{21} = p_1(1 - p_1), \\ &\phi_2 = p_2,\ \psi_{02} = (1 - p_2)\sigma_v^2 + p_2(1 - p_2)\bar{V}^2, \\ &\qquad\qquad \psi_{12} = -2p_2(1 - p_2)\bar{V},\ \psi_{22} = p_2(1 - p_2), \\ &\delta_1 = -p_1^2/(N - 1),\ \lambda_1 = -(1 - p_1)^2\sigma_u^2/(N - 1), \\ &\delta_2 = -p_2^2/(N - 1),\ \lambda_2 = -(1 - p_2)^2\sigma_v^2/(N - 1), \\ &\phi_3 = p_1p_2,\ \delta_3 = -\phi_3/(N - 1), \\ &\qquad\qquad \psi_3 = (1 - p_1)(1 - p_2)\sigma_{uv}, \\ &\qquad\qquad\qquad \text{and} \quad \lambda_3 = -\psi_3/(N - 1)\,. \end{aligned} \tag{3}$$

Note that the model assumptions require that the finite population variance-covariance matrix of the nonsensitive questions is known as well as the finite population means.

For the additive constants model (randomized response model (ii)) assume that the random variables u and v that are added to the value of the responses to the two sensitive questions are independent with means μ_u and μ_v and variances σ_u^2 and σ_v^2 respectively. When the random permutation model is assumed on the pair (x_j, y_j) then in the model given by (1) and (2):

$$\begin{aligned} &\alpha_1 = \mu_u,\ \beta_1 = 1,\ \alpha_2 = \mu_v,\ \beta_2 = 1, \\ &\phi_1 = \phi_2 = \phi_3 = 1,\ \psi_{01} = \sigma_u^2,\ \psi_{02} = \sigma_v^2, \\ &\delta_1 = \delta_2 = \delta_3 = -1/(N - 1), \\ &\psi_{11} = \psi_{21} = \psi_{12} = \psi_{22} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0. \end{aligned} \tag{4}$$

In the multiplicative constants model, two independent random variables, u and v with means μ_u and μ_v and variances σ_u^2 and σ_v^2 respectively, are multiplied respectively by the value of the response on the x-variable and the y-variable. When the random permutation model is assumed on the pair (x_j, y_j) then in the model given by (1) and (2):

$$\alpha_1 = \alpha_2 = 0, \beta_1 = \mu_u, \beta_2 = \mu_v,$$

$$\phi_1 = \mu_u^2 + \sigma_u^2, \phi_2 = \mu_v^2 + \sigma_v^2, \phi_3 = \mu_v \mu_v,$$

$$\psi_{21} = \sigma_u^2, \psi_{22} = \sigma_v^2,$$

$$\delta_1 = -\mu_u^2/(N-1), \delta_2 = -\mu_v^2/(N-1),$$
$$\delta_3 = -\mu_u\mu_v/(N-1), \text{ and}$$

$$\psi_{01} = \psi_{11} = \psi_{02} = \psi_{12} = \psi_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0. \tag{5}$$

4. ESTIMATION OF VARIANCE AND COVARIANCE

Consider estimation of σ_y^2 so that the appropriate data are z_j for units $j \in s$. The general class of quadratic estimators of σ_y^2 is of the form:

$$e_{bs} = b_{s..} + \sum_{j \in s} b_{sj.}\, z_j + \sum_{j \in s} b_{sjj}\, z_j^2 + \sum\sum_{i \neq j \in s} b_{sij}\, z_i z_j, \tag{6}$$

where the coefficients of the z's are defined for all s, all $j \in s$ and all pairs $(i, j) \in s$.

In the context of randomized response, an estimator e_b in the class defined by (6) is design-unbiased for σ_y^2 if $E_p E_r(e_b) = \sigma_y^2$ and is pm-unbiased if $E(e_b) = \sigma_y^2$. Conditions under which an estimator e_b is pm-unbiased are obtained upon taking the expectation E of (6) under (1) and (2). On equating coefficients in $\bar{Y}^0$, $\bar{Y}^1$, $\bar{Y}^2$ and σ_y^2 four equations in four unknowns are obtained. The solution to these four equations yields the following conditions under which estimators in the class defined by (6) are pm-biased for σ_y^2:

$$E_p\Big(\sum_{j \in s} b_{sjj}\Big) = \frac{\beta_2^2}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = A_2, \tag{7}$$

$$E_p\Big(\sum\sum_{i \neq j \in s} b_{sij}\Big) = -\frac{\beta_2^2 + \psi_{22}}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = -(A_2 + B_2), \tag{8}$$

$$E_p\Big(\sum_{j \in s} b_{sj.}\Big) = \frac{(2\alpha_2\psi_{22} - \beta_2\psi_{12})}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = C_2, \tag{9}$$

and

$$E_p(b_{s..}) = \frac{\lambda_2(\beta_2^2 + \psi_{22}) - (\alpha_2^2\psi_{22} - \alpha_2\beta_2\psi_{12} + \beta_2^2\psi_{02})}{\beta_2^2(\phi_2 - \delta_2) - \delta_2\psi_{22}} = D_2. \tag{10}$$

In order to obtain the optimal estimator we need to define an associated class of quadratic estimators of 0. This is given by

$$e_{cs} = c_{s..} + \sum_{j \in s} c_{sj.}\, z_j + \sum_{j \in s} c_{sjj}\, z_j^2 + \sum\sum_{i \neq j \in s} c_{sij}\, z_i z_j.$$

The conditions for an estimator e_c in this class to be pm-biased for 0 are

$$E_p(c_{s..}) = E_p\Big(\sum_{j \in s} c_{sj.}\Big) = E_p\Big(\sum_{j \in s} c_{sjj}\Big) = E_p\Big(\sum\sum_{i \neq j \in s} c_{sij}\Big) = 0. \tag{11}$$

Derivation of the minimum variance quadratic design-unbiased estimator of σ_y^2 follows along the same lines as that used for the finite population mean by Rao and Bellhouse (1978) for cases without randomized response and by Bellhouse (1980) for cases with randomized response. The covariance $E(e_b e_c)$ under the composite expectation is determined under the model such that only expectations of the form E_p remain to be determined. From this expression the coefficients b are set to make $E(e_b e_c) = 0$ under the conditions in (11). The values of the coefficients b are then determined from the conditions in (7) through (10). From a theorem on minimum variance unbiased estimation of Rao (1952), the resulting estimator is the optimal pm-unbiased estimator of σ_y^2. If there exists a design such that this estimator is also design-unbiased for σ_y^2, then by arguments similar to those given in Theorem (2.4) of Rao and Bellhouse (1978), the estimator is also the optimal design-unbiased estimator of σ_y^2. We present results for pm-unbiased estimators first (Theorems 1 and 2) and then present results for design-unbiased estimators under the three randomized response schemes.

Theorem 1. Under the model defined by (1) and (2) and for any design of fixed size n, the pm-variance of e_b, $E_{rp}[E_p E_r(e_b - \sigma_y^2)^2] = E(e_b - \sigma_y^2)^2$, is minimized for the estimator given by

$$(A_2 + B_2)s_z^2 - B_2\frac{1}{n}\sum_{j \in s} z_j^2 + C_2\bar{z} + D_2, \tag{12}$$

where $\bar{z}$ is the sample mean of the data and

$$s_z^2 = \frac{1}{n-1} \sum_{j \epsilon s} (z_j - \bar{z})^2$$

is the sample variance of the data obtained through randomized response where A_2, B_2, C_2 and D_2 are defined in (7) through (10) respectively.

Proof. Under the model given by (1) and (2) the covariance $E(e_b e_c)$ is algebraically quite lengthy but may be expressed in the following form:

$$\boldsymbol{b}^T G \boldsymbol{c} + H, \tag{13}$$

where $\boldsymbol{b}^T$ is the vector

$$\left[E_p(b_{s..}), E_p\left(\sum_{j \epsilon s} b_{sj.}\right), E_p\left(\sum_{j \epsilon s} b_{sjj}\right),\right.$$
$$\left. E_p\left(\sum\sum_{i \neq j \epsilon s} b_{sij}\right)\right], \tag{14}$$

and $\boldsymbol{c}^T$ is the same as (14) with the b's replaced by c's. The 4×4 matrix G in (13) contains functions of the first order moments of z_j and the second order moments of e_{2j} in (1). The expression H in (13) is a sum of terms of the form

$$\kappa \sum b_{sij} \, c_{skl}, \tag{15}$$

where the summation symbol is up to a quadruple sum, where the subscripts of b could be replaced by a dot (.) and where κ is a function of second through fourth order moments of e_{2j} in (1). Note that these moments are all independent of j. In (15) the sum is a single sum over $j \epsilon s$ when, for example, the subscripts $i = j = k = l$ or when $i = k$ and j and l are replaced by dots. The sum is a double sum over $i \neq k \epsilon s$ when, for example, $i \neq k$ and j and l are replaced by dots. This process continues to the quadruple sum in which $i \neq j \neq k \neq l$. From (11) $E(e_b e_c)$ reduces to 0 if $b_{s..} = h_1$, $b_{sj} = h_2$, $b_{sjj} = h_3$, and $b_{sij} = h_4$, where the h_i are constants. From (7) through (10) and the fact that the design is of fixed size we obtain

$$b_{s..} = D_2,\ b_{sj.} = C_2/n,\ b_{sij} = -\frac{A_2 + B_2}{n(n-1)},\ b_{sjj} = A_2/n,$$

so that the estimator in (12) minimizes the variance in the pm-unbiased class of quadratic estimators of σ_y^2. Q.E.D.

By the same arguments

$$(A_1 + B_1) s_w^2 - B_1 \frac{1}{n} \sum_{j \epsilon s} w_j^2 + C_1 \bar{w} + D_1, \tag{16}$$

is the optimal pm-unbiased estimator for σ_x^2 where

$$A_1 = \frac{\beta_1^2}{\beta_1^2(\phi_1 - \delta_1) - \delta_1 \psi_{21}},$$

$$B_1 = \frac{\beta_1^2 + \psi_{21}}{\beta_1^2(\phi_1 - \delta_1) - \delta_1 \psi_{21}},$$

$$C_1 = \frac{(2\alpha_1 \psi_{21} - \beta_2 \psi_{11})}{\beta_1^2(\phi_1 - \delta_1) - \delta_1 \psi_{21}}, \quad \text{and}$$

$$D_1 = \frac{\lambda_1(\beta_1^2 + \psi_{21}) - (\alpha_1^2 \psi_{21} - \alpha_1 \beta_1 \psi_{11} + \beta_1^2 \psi_{01})}{\beta_1^2(\phi_1 - \delta_1) - \delta_1 \psi_{21}}.$$

The same technique can be used to estimate the covariance σ_{xy}. The general class of quadratic estimators of σ_{xy} is of the form

$$e_{ds} = d_s + \sum_{j \epsilon s} d_{1sj} z_j + \sum_{j \epsilon s} d_{2sj} w_j + \sum\sum_{i \neq j \epsilon s} d_{sij} w_i z_j,$$

where the coefficients of the w's and z's are defined for all s, all $j \epsilon s$ and all pairs $(i, j) \epsilon s$. The result on the covariance is stated without proof in

Theorem 2. Under the model defined by (1) and (2) and for any design of fixed size n, the pm-variance of e_d, $E_{rp}[E_p E_r(e_d - \sigma_{xy})^2] = E(e_d - \sigma_{xy})^2$, is minimized for the estimator given by

$$\frac{s_{wz} - (\psi_3 - \lambda_3)}{\phi_3 - \delta_3}, \tag{17}$$

where

$$s_{wz} = \frac{1}{n-1} \sum_{j \epsilon s} (w_j - \bar{w})(z_j - \bar{z})$$

is the sample covariance between w and z.

An estimator for ρ is obtained from (12), (16) and (17). In the additive constants randomized response model (randomized response model (ii)) the estimator of ρ is given by

$$\hat{\rho}_{ac} = \frac{s_{wz}}{\sqrt{(s_w^2 - \sigma_u^2)(s_z^2 - \sigma_v^2)}}. \tag{18}$$

This is the same as the estimator obtained by Edgell *et al.* (1986). Under the multiplicative constants model (randomized response model (iii)) the estimator reduces to

$$\hat{\rho}_{mc} = \frac{s_{wz}}{\sqrt{s_w^2 - \dfrac{\sigma_u^2/\mu_u^2}{1 + \sigma_u^2/\mu_u^2} \dfrac{1}{n} \sum_{j \in s} w_j^2} \sqrt{s_z^2 - \dfrac{\sigma_v^2/\mu_v^2}{1 + \sigma_v^2/\mu_v^2} \dfrac{1}{n} \sum_{j \in s} z_j^2}}, \tag{19}$$

for $\mu_u \neq 0$ and $\mu_v \neq 0$. When $\mu_u = 0$ the coefficient of $\sum w_j^2$ is $1/n$ and when $\mu_v \neq 0$ the coefficient of $\sum z_j^2$ is $1/n$. The estimator for ρ under the unrelated questions model (randomized response model (i)) is

$$\hat{\rho}_{uq} = \frac{s_{wz} - \dfrac{(1 - p_1)(1 - p_2)}{p_1 p_2} S_{uv}}{\sqrt{\hat{S}_x^2 \hat{S}_y^2}}, \tag{20}$$

where $S_{uv} = N\sigma_{uv}/(N - 1)$ and where

$$\hat{S}_x^2 = s_w^2 - (1 - p_1)\frac{1}{n}\sum_{j \in s} w_j^2 + 2(1 - p_1)\bar{U}\bar{w} - (1 - p_1)\bar{U}^2 - (1 - p_1)\sigma_u^2\left(p_1 + \frac{1 - p_1}{N - 1}\right)$$

and

$$\hat{S}_y^2 = s_z^2 - (1 - p_2)\frac{1}{n}\sum_{j \in s} z_j^2 + 2(1 - p_2)\bar{V}\bar{z} - (1 - p_2)\bar{V}^2 - (1 - p_2)\sigma_v^2\left(p_2 + \frac{1 - p_2}{N - 1}\right) .$$

When $p_1 = p_2$ this may be compared to the estimator in Edgell *et al.* (1986). The resulting estimator for $\hat{\rho}_{uq}$ differs from the estimator in Edgell *et al.* (1986) who assume that $\sigma_{uv} = 0$. They also use biased estimators of σ_x^2 and σ_y^2. Edgell *et al.*'s estimator for σ_y^2 is obtained by writing the design variance of $\bar{z}$ under simple random sampling with replacement as

$$\sigma_z^2/n = \sum_{j=1}^{N} (z_j - \bar{Z})^2/(Nn). \tag{21}$$

The design variance of $\bar{z}$ under the randomizing device is

$$[p_2\sigma_y^2 + (1 - p_2)\sigma_v^2 + p_2(1 - p_2)(\bar{Y} - \bar{V})^2]/n. \tag{22}$$

Expression (22) is found in Greenberg *et al.* (1971). The estimator for σ_y^2 is found by equating (22) to the left hand side of (21), by substituting sample the estimator of σ_z^2 and the randomized response estimator of $\bar{Y}$ in the resulting equation, and then by solving for σ_y^2.

Each of the estimators of the finite population variances and covariance, which are the components of $\hat{\rho}$ in (18), (19) and (20), are design-unbiased under the appropriate randomized response model for any design with joint inclusion probability for units i and j given by $\pi_{ij} = n(n - 1)/[N(N - 1)]$. Consequently, each estimator is the optimal design-unbiased estimator for its finite population parameter counterpart. To obtain the appropriate unbiased estimators in (18), multiply the numerator and denominator each by $(N - 1)/N$. The resulting numerator is design-unbiased for σ_{xy} and the expressions under the square root sign in the denominator of (18) are unbiased for σ_x^2 and σ_y^2. In (19) it is necessary to multiply the numerator and denominator by $(N - 1)/[N\mu_u\mu_v]$ in order to obtain the correct form of the design-unbiased estimators. The correct estimators are obtained in (20) when the multiplier is $(N - 1)/(Np_1p_2)$.

In any of the randomized response designs, the simplest estimate of the variance of $\hat{\rho}$ is the jackknife estimate of variance. Jackknife estimates of variance for $\hat{\rho}$ can be obtained from formulae (4.2.3) or (4.2.5) in Wolter (1985).

5. EFFECT OF RESPONSE BIAS

In the additive constants model, the respondent is asked to add a random variable u to x and an independent random variable v to y. Instead, the respondent may add different independent random variables, say u' and v'. The means and variances of u' and v' may differ from those of u and v. It is reasonable to assume, however, that $\sigma_{u'}^2 \geq \sigma_u^2$ and $\sigma_{v'}^2 \geq \sigma_v^2$. One example in which this situation might occur is the following. The respondent does not want to add on the outcome of a random variable near to the mean of the distribution of the random variable. In this case the distribution of response bias could be modelled by the original distribution with an interval around the mean in which any outcome from the original distribution which falls in the chosen interval is set to one of the end points of the interval. On taking separately the expectations of the numerator and the expression under each of the square root signs in the denominator of (18) the expression

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2 + \sigma_{u'}^2 - \sigma_u^2}\ \sqrt{\sigma_y^2 + \sigma_{v'}^2 - \sigma_v^2}}, \tag{23}$$

is obtained. From (23) it may be noted that the response bias leads to an estimate of correlation lower than the true value.

The multiplicative constants model is the same as the additive constants model with the exception that the responses to the sensitive questions are multiplied by the random variables. As in the response bias model for additive constants, assume that u' and v' are used by the respondent instead of u and v. Then on taking separately the expectations of the numerator and the expressions under each of the square root signs in the denominator of (19) the expression

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2+\frac{\sum_{j=1}^{N}x_j^2}{N\mu_{u'}^2}\,\frac{\sigma_u^2\mu_{u'}^2-\sigma_{u'}^2\mu_u^2}{\sigma_u^2+\mu_u^2}}\sqrt{\sigma_y^2+\frac{\sum_{j=1}^{N}y_j^2}{N\mu_{v'}^2}\,\frac{\sigma_v^2\mu_{v'}^2-\sigma_{v'}^2\mu_v^2}{\sigma_v^2+\mu_v^2}}}, \tag{24}$$

is obtained. If $\mu_u = \mu_{u'}$, $\mu_v = \mu_{v'}$, $\sigma_{u'}^2 \geq \sigma_u^2$ and $\sigma_{v'}^2 \geq \sigma_v^2$, as in the case of the additive constants model, then from (24) the response bias leads to an overestimate of the correlation.

In the unrelated questions model a reasonable model for response bias is to assume that the sensitive questions are answered with probability $p_1' < p_1$ and $p_2' < p_2$. In general the effect of this response bias is dependent on the relative values of the various probabilities, the means and variances of the sensitive questions, and the means and variances of the nonsensitive questions. Under simple random sampling without replacement and the response bias model, the design expectation of the numerator of (20) is given by

$$p_1'p_2'\left[S_{xy}+\frac{(1-p_1')(1-p_2')-(1-p_1)(1-p_2)}{p_1'p_2'}\,S_{uv}\right],$$

which is greater than $p_1'p_2'S_{xy}$. Likewise the design expectation of $\hat{S}_x^2$ in (20) is

$$\begin{aligned} S_x^2[p_1'^2 &+ (N-1)p_1'(p_1-p_1')/N] \\ &+ (p_1-p_1')S_u^2[p_1'-(p_1'+2p_1-2)/N] \\ &+ p_1'(p_1-p_1')(\bar{X}-\bar{U})^2, \end{aligned}$$

which is greater than $p_1'^2\hat{S}_x^2$ when N is large. If $S_{uv} = 0$, then the response bias leads to an underestimate of the correlation.

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Stan Warner and was written for the Stanley Warner Memorial Session at the Statistical Society of Canada meetings in Banff, Alberta, May 1994. The research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

BELLHOUSE, D.R. (1980). Linear models for randomized response designs. *Journal of the American Statistical Association*, 75, 1001-1004.

CHAUDHURI, A., and MUKERJEE, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

EDGELL, S.E., HIMMELFARB, S., and CIRA, D.J. (1986). Statistical efficiency of using two quantitative randomized response techniques to estimate correlation. *Psychological Bulletin*, 100, 251-256.

GREENBERG, B.G., ABUL-ELA, A.A., SIMMONS, W.R., and HORVITZ, D.G. (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., and HORVITZ, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243-250.

LEYSIEFFER, F.W., and WARNER, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.

NATHAN, G. (1988). A bibliography on randomized response: 1965-1987. *Survey Methodology*, 14, 331-346.

POLLOCK, K.H., and BEK, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71, 884-886.

RAO, C.R. (1952). Some theorems on minimum variance unbiased estimation. *Sankhyā* (A), 12, 27-42.

RAO, J.N.K. (1975). On the foundations of survey sampling. In *A Survey of Statistical Design and Linear Models*. (Ed. J.N. Srivastava). Amsterdam: North-Holland, 489-505.

RAO, J.N.K., and BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2, 125-141.

STEM, D.E., and STEINHORST, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the American Statistical Association*, 79, 555-564.

UMESH, U.N., and PETERSON, R.A. (1991). A critical evaluation of the randomized response method. *Sociological Methods and Research*, 20, 104-138.

WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.

WARNER, S.L. (1976). Optimal randomized response models. *International Statistical Review*, 44, 205-212.

WARNER, S.L. (1986). The omitted digit randomized response model for telephone applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag.

Survey Methodology, June 1995
Vol. 21, No. 1, pp. 21–23
Statistics Canada

On Efficiency of Using Distinct Respondents in a Randomized Response Survey

N.S. MANGAT, R. SINGH, S. SINGH, D.R. BELLHOUSE and H.B. KASHANI[1]

ABSTRACT

It is well known that the sample mean based on the distinct sample units in simple random sampling with replacement is more efficient than the sample mean based on all units selected including repetitions (Murthy 1967, pp. 65-66). Seth and Rao (1964) showed that the mean of the distinct units is less efficient than the sample mean in sampling without replacement under the same average sampling cost. Under Warner's (1965) method of randomized response we compare simple random sampling without replacement and sampling with replacement when only the distinct number of units in the sample are considered.

KEY WORDS: Simple random sampling with and without replacement; Inferences with distinct units; Warner's technique.

1. INTRODUCTION

The randomized response (RR) technique to procure trustworthy data for estimating the proportion of the population belonging to a sensitive group was first introduced by Warner (1965). Since then many developments have taken place in this area. Recently, among others, Franklin (1989), Kuk (1990), Mangat and Singh (1990, 1991), Mangat, Singh and Singh (1992) and Mangat (1994) have suggested alternative RR procedures/estimators.

In the usual simple random sampling (SRS) with replacement (WR) surveys, it is well known that the estimator of population mean based on the distinct units is always more efficient than the mean based on all selections (Murthy 1967, pp. 65-66). Also, Seth and Rao (1964) showed that, under the same average cost to sample, sampling without replacement was more efficient than with replacement sampling using the mean of the distinct sample units. This motivated the authors to investigate whether the above observation also holds in the case of Warner's pioneer RR model which is widely used in practice for selecting the respondents in the case of a survey dealing with sensitive characters. To investigate the problem we shall consider the use of four sampling strategies.

1.1 Strategy I

According to this (Warner's) procedure, each respondent included in the sample using the SRSWR method is provided with a suitable randomization device consisting of two statements of the form: (i) "I belong to sensitive group" and (ii) "I do not belong to sensitive group", represented with probabilities p and $(1 - p)$, respectively. The respondent answers "yes" or "no" according to the randomly selected statement and to his actual status with respect to the attribute, without revealing the statement chosen. If n' persons in the sample (including repetitions) answered "yes", Warner's estimator

$$\hat{\pi} = \frac{n'/n - 1 + p}{2p - 1}, \quad p \neq .5, \tag{1}$$

is unbiased for π and its variance is given by

$$V_1(\hat{\pi}) = \frac{\pi(1 - \pi)}{n} + \frac{p(1 - p)}{n(2p - 1)^2}. \tag{2}$$

The value of p should be chosen as close to 1 or 0 as possible without threatening the degree of co-operation by respondents.

1.2 Strategy II

A sample of n respondents is drawn from a finite population of N units using SRSWR but the information from the d distinct units in the sample, $1 \leq d \leq n$, is used in the construction of the estimator. Let d' denote the respondents reporting a "yes" answer in the interview conducted with the RR device. We then consider the following estimator for π:

$$\hat{\pi}_d = \frac{d'/d - 1 + p}{2p - 1}, \quad p \neq .5. \tag{3}$$

Conditional on d distinct units, the resulting sample is a simple random sample without replacement of size d from N units. The estimator $\hat{\pi}_d$ is, therefore, unbiased for the population π.

[1] N.S. Mangat, R. Singh and S. Singh, Punjab Agricultural University, Ludhiana-141004 (India); D.R. Bellhouse, University of Western Ontario, London, Ontario, Canada, N6A 5B7; H.B. Kashani, West Oregon State College, Monmouth, OR 97361, U.S.A.

In order to study the performance of the proposed estimator $\hat{\pi}_d$, we need its variance. We give here the expression for the conditional variance $V_2(\hat{\pi}_d)$ for a given value of d. Thus

$$V_2(\hat{\pi}_d) = \frac{N-d}{N-1}\,\frac{\pi(1-\pi)}{d} + \frac{p(1-p)}{d(2p-1)^2}. \quad (4)$$

If E_1 and V_1 are the expectation and variance over all values of d, then we have $V_{\text{II}}(\hat{\pi}_d) = E_1V_2(\hat{\pi}_d) + V_1E_2(\hat{\pi}_d)$. On using (4) one gets

$$V_{\text{II}}(\hat{\pi}_d) = \left[NE_1\left(\frac{1}{d}\right) - 1\right]\frac{\pi(1-\pi)}{N-1} + \frac{p(1-p)}{(2p-1)^2}E_1\left(\frac{1}{d}\right) \quad (5)$$

since the second term in $V_{\text{II}}(\hat{\pi}_d)$ is zero as $E_2(\hat{\pi}_d) = \pi$.

1.3 Strategy III

The sample of n respondents is selected using SRSWOR (Kim and Flueck 1978). In this case the variance of the estimator $\hat{\pi}$ in (1) can be written by replacing d in (4) by n. Thus we have

$$V_{\text{III}}(\hat{\pi}) = \frac{N-n}{N-1}\,\frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \quad (6)$$

1.4 Strategy IV

Here the estimator is based on a WOR simple random sample of size $E(d)$. This yields the same expected cost for both in SRSWR and SRSWOR. For this scheme the estimator will be

$$\hat{\pi}_E = \frac{d'/E(d) - 1 + p}{2p-1}, \quad p \neq .5$$

with variance

$$V_{\text{IV}}(\hat{\pi}_E) = \frac{N/E(d) - 1}{N-1}\,\pi(1-\pi) + \frac{p(1-p)}{E(d)(2p-1)^2}. \quad (7)$$

2. EFFICIENCY COMPARISONS

It has been shown by Korwar and Serfling (1970) that, for $n \geq 3$,

$$Q - \frac{1}{720N} < E\left(\frac{1}{d}\right) \leq Q$$

where

$$Q = \frac{1}{n} + \frac{1}{2N} + \frac{n-1}{12N^2}.$$

Let us now examine the variance expression in (5). Using Q, it is easily verified that

$$\frac{NE_1(1/d) - 1}{N-1} \leq \frac{1}{n}, \quad (8)$$

in the first term on the right of (5) but that $E_1(1/d) \geq 1/n$ in the second term on the right of (5). Thus the relative efficiency of the SRSWR estimator in (1) using repeated units with respect to the SRSWR estimator in (3) using the distinct number of units will depend on the relative sizes of π and p. This is due to the fact that the repeated units can give rise to different responses because of the randomizing device and hence can provide some additional information. A sufficient condition for the inequality $V_{\text{II}}(\hat{\pi}_d) - V_{\text{I}}(\hat{\pi}) < 0$ to hold is obtained by using $E_1(d) = Q$. Thus we get the condition as

$$\pi(1-\pi) > \frac{n(N-1)(6N+n-1)}{N\{6Nn - 12N - n(n-1)\}}\,\frac{p(1-p)}{(2p-1)^2}. \quad (9)$$

The above inequality is likely to hold for values of p closer to 0 or 1, the situations in which respondent jeopardy would be of concern. For example, if $N = 100$, $n = 10$ and $p = 0.9$, the inequality (9) will hold for $0.236 \leq \pi \leq 0.764$.

Similarly, Strategy II will be inferior to Strategy I if $V_{\text{II}}(\hat{\pi}_d) - V_{\text{I}}(\hat{\pi}) > 0$. Using $E_1(1/d) = Q - 1/720N$ this inequality reduces to

$$\pi(1-\pi) < \frac{n(N-1)\{359N + 60(n-1)\}}{N\{361Nn - 720N - 60n(n-1)\}}\,\frac{p(1-p)}{(2p-1)^2}.$$

This inequality will hold for the example considered for inequality (9) whenever either $\pi \leq 0.234$ or $\pi \geq 0.764$.

On using the Cauchy-Schwarz inequality, $E(1/d) > 1/E(d)$, as in Seth and Rao (1964) we find that $V_{\text{II}}(\hat{\pi}_d) > V_{\text{IV}}(\hat{\pi}_E)$. This implies that Strategy IV is more efficient than Strategy II.

It is trivial to note that Strategy III is more efficient than Strategy I.

We know that $E(1/d) \geq 1/n$. This means $V_{II}(\hat{\pi}_d) > V_{III}(\hat{\pi})$, implying that Strategy III is more efficient than Strategy II.

Since $1/E(d) \geq 1/n$, Strategy III is more efficient than Strategy IV.

The last pair to consider consists of Strategies I and IV. Since $E(1/d) > 1/E(d)$ for $n > 1$, on using (8) we have

$$\frac{N/E(d) - 1}{N - 1} \leq \frac{1}{n}$$

implying that in (7) and (2)

$$\frac{N - E(d)}{N - 1} \frac{\pi(1 - \pi)}{E(d)} \leq \frac{\pi(1 - \pi)}{n}$$

for $n > 1$. Also $1/E(d) \geq 1/n$. This shows that the second term of (7) on the right hand side will be more than the corresponding term of (2). Thus the relative efficiencies of Strategies I and IV depend on relative values of π and p. As a numerical illustration, if $N = 100$, $n = 10$ and $p = 0.9$ then Strategy IV will be more efficient than Strategy I for $0.18 \leq \pi \leq 0.82$.

REFERENCES

FRANKLIN, L.A. (1989). Randomized response sampling from dichotomous populations with continuous randomization. *Survey Methodology*, 15, 225-235.

KIM, J.-I., and FLUECK, J.A. (1978). Modifications of the randomized response technique for sampling without replacement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 346-350.

KORWAR, R.M., and SERFLING, R.J. (1970). On averaging over distinct units in sampling with replacement. *Annals of Mathematical Statistics*, 41, 2132-2134.

KUK, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77, 436-438.

MANGAT, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society*, Series B, 56, 93-95.

MANGAT, N.S., and SINGH, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.

MANGAT, N.S., and SINGH, R. (1991). An alternative approach to randomized response survey. *Statistica*, anno LI, 327-332.

MANGAT, N.S., SINGH, R., and SINGH, S. (1992). An improved unrelated question randomized response strategy. *Calcutta Statistical Association Bulletin*, 42, 277-281.

MURTHY, M.N. (1967). *Sampling Theory and Methods*, Calcutta, India: Statistical Publishing Society.

SETH, G.R., and RAO, J.N.K. (1964). On the comparison between simple random sampling with and without replacement. *Sankhyā* (A), 26, 85-86.

WARNER, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

Survey Methodology, June 1995
Vol. 21, No. 1, pp. 25–32
Statistics Canada

Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method

PIERRE LAVALLÉE[1]

ABSTRACT

Statistical agencies are conducting increasing numbers of longitudinal surveys. Although the main output of these surveys consists of longitudinal data, most of them are also expected to produce reliable cross-sectional estimates. In surveys of individuals and households, population dynamics significantly changes household composition over time. For this reason, methods of cross-sectional estimation must be adapted to the longitudinal aspect of the sample. This paper discusses in a general context the Weight Share method, of which one application is to assign a basic weight to each individual in a household. The variance estimator associated with the Weight Share method is also presented. The weighting of a longitudinal sample is then discussed when a supplementary sample is selected to improve the cross-sectional representativeness of the sample. The paper presents as an application the Survey of Labour and Income Dynamics (SLID) introduced by Statistics Canada in 1994. This longitudinal survey covers individuals' work experience, changes in income and changes in family composition.

KEY WORDS: Weight share method; Longitudinal survey; Cross-sectional estimate; Supplementary sample.

1. INTRODUCTION

Longitudinal surveys, *i.e.* surveys that follow units over time, are steadily gaining importance within statistical agencies. Statistics Canada is currently developing three major longitudinal surveys of individuals: the National Population Health Survey, the National Longitudinal Survey of Children; and the Survey of Labour and Income Dynamics (SLID).

The primary objective of these surveys is to obtain longitudinal data. One of the uses of these data is to study the changes in variables over time (*e.g.*, longitudinal data may be used to analyze the chronic aspect of poverty). A secondary objective is the production of cross-sectional estimates, in other words estimates that represent the population at a given point in time. Although these estimates are far less important than the longitudinal data, to many users they are an essential aspect of the survey. Obtaining a representative cross-sectional view of the current population constitutes a means of measuring changing situations over time. The longitudinal aspect of the survey also improves the accuracy of the measurement of change.

This paper presents an extension of the Weight Share method presented by Ernst (1989). Although the method has been developed in the context of longitudinal household surveys, it is shown that the Weight Share method can be generalized to situations where a population of interest is sampled through the use of a frame which refers to a different population, but linked somehow to the first one. In the context of longitudinal surveys, the frame can be associated to the initial population, while the population of interest can be the population a few years later. The paper also provides a new proof of the unbiasedness of the Weight Share method together with the variance formula and variance estimator to be used with the method.

Using the Weight Share method, the question addressed in this paper is that of ensuring that the longitudinal sample can be used for cross-sectional estimation. The difficulty arises from the fact that, although the longitudinal sample remains constant, distribution of the population (individuals and households) changes over time. At the individual level, these changes are produced by such events as births and deaths, immigration and emigration, and moves within the country. Obviously, the birth or death of an individual also changes household composition; and such events as marriage, divorce, separation, departure of a child and cohabitation, are all factors that affect population distribution within the household. If we are to obtain accurate, unbiased cross-sectional estimates based on a longitudinal sample, we need an estimation method that takes these changes into account.

Our initial topic is the presentation of the Weight Share method in a general context. Secondly, we present the sample design for SLID. This is one of the major longitudinal surveys for which the production of cross-sectional estimates from a longitudinal sample is a significant problem. The survey itself is a typical longitudinal survey of individuals and households. Thirdly, we describe the use of a supplementary sample added to the initial longitudinal sample to improve the cross-sectional representativeness. Fourthly, we present the concept of basic weights, the equivalent, as it were, of sample weights. Finally, we describe the use of the Weight Share method to calculate basic weights for all individuals interviewed in SLID.

[1] Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

2. THE WEIGHT SHARE METHOD IN A GENERAL CONTEXT

The Weight Share method is described in Ernst (1989) in the context of longitudinal household surveys. In the same context, Kalton and Brick (1995) discuss different weighting schemes, including the Weight Share method. Various implications of using the Weight Share method for longitudinal household surveys have been described by Gailly and Lavallée (1993).

We now present this method in a general context that can be applied to several sampling situations where the population of interest needs to be sampled through the use of a frame which refers to a different population, but is linked somehow to the first one. Note that this can be viewed as a form of Network Sampling (see Thompson 1992). For example, one can imagine the need to sample young children where the only available frame is a list of names of parents. The population of interest is really the children but we need to select a sample of parents from the frame in order to obtain the sample of children. Note that the children of a particular family can be sampled through either the father or the mother. Another example is one of business surveys where an incomplete frame of establishments is available. For each selected establishment from the frame, we wish to sample the entire set of establishments belonging to the same enterprise. The missing establishments from the frame are expected to be sampled via the establishments present on the frame.

Suppose that a sample s^A of m^A units is selected from a population U^A of M^A units using some sampling design. Let π_j^A be the selection probability of unit j. We assume $\pi_j^A > 0$ for all $j \in U^A$.

Let U^B be a population of M^B units. This population is divided into N clusters where cluster i contains M_i^B units. For example, in the context of social surveys, the clusters can be households and the units can be the persons within the households. For business surveys, the clusters can be enterprises and the units can be the establishments within the enterprises. From population U^B, we are interested in estimating the total $Y = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$ for some characteristic y.

An important constraint that is imposed in the measurement (or interviewing) process is to consider all units within the same cluster. That is, if a unit is selected in the sample, then every unit of the cluster containing the selected unit will be interviewed. This constraint is one which often arises in surveys for two reasons: cost reductions and the need for producing estimates on clusters. Referring back to the example of social surveys, there is normally a small marginal cost for interviewing all persons within the household. On the other hand, household estimates are often of interest with respect to poverty measures, for example.

We assume that there exists a *link* (or a correspondence) between each unit j of population U^A and at least one unit k of population U^B. Also, each cluster i of U^B has at least one link with a unit j of U^A. The link is identified through an indicator variable l_{jk} where $l_{jk} = 1$ if there is a link between unit $j \in U^A$ and unit $k \in U^B$ and 0 otherwise. All units of population U^A have at least one link with population U^B, *i.e.*, $L_j^A = \sum_{k \in U^B} l_{jk} \geq 1$ for all $j \in U^A$. However, there can be zero, one or more links for a unit k of population U^B, *i.e.*, it is possible to have $L_k^B = \sum_{j \in U^A} l_{jk} = 0$ or $L_k^B = \sum_{j \in U^A} l_{jk} > 1$ for some $k \in U^B$. This is illustrated in Figure 1.

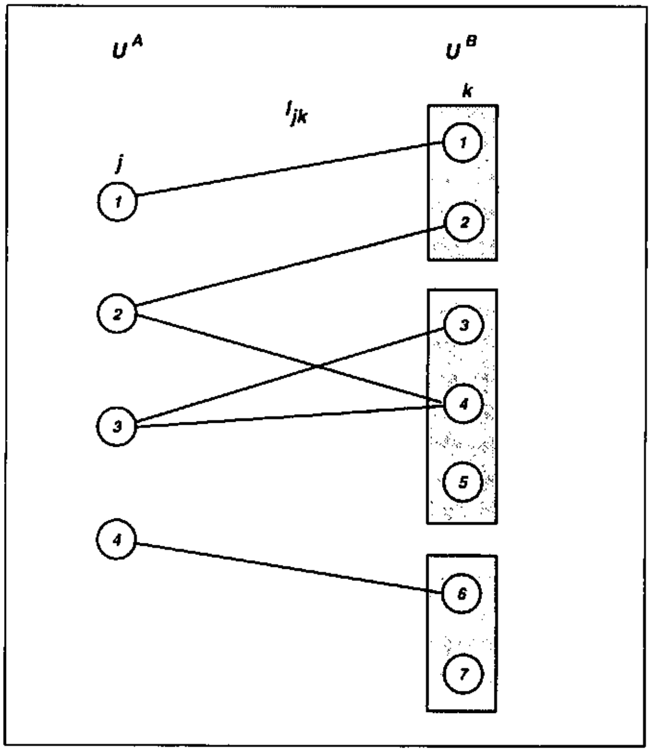

Figure 1. Links between units of populations U^A and U^B.

The estimation process presented now uses the sample s^A together with the links existing between U^A and U^B to obtain an estimation of the total Y belonging to population U^B. The links are in fact utilized as a bridge to go from population U^A to population U^B, and vice versa. Note that in practice, it might not be physically possible to directly select a sample s^B from U^B, as it has been described in the introductory examples.

To estimate the total Y, one can use the estimator

$$\hat{Y} = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik}\, y_{ik}, \tag{1}$$

where n is the number of interviewed clusters and w_{ik} is the weight attached to unit k of cluster i. To obtain unbiased estimates, a possible set of weights could be the inverse of the selection probabilities of the units entering into the estimator $\hat{Y}$. For each unit k of cluster i having a link $l_{j,ik} = 1$ with a unit j in U^A, this is possible since we have $\pi_k^B = \pi_j^A$. However, not all units of U^B necessarily have a link to U^A. Moreover, even if a link exists, it is not guaranteed that the selection probability π_j^A is known when $j \notin s^A$; the sample design used to select s^A could be, for example, a multistage sample design where the ultimate selection probability of each unit j is only known at the end of the selection process. To assign a nonzero weight w_{ik} to each unit k of cluster i entering into $\hat{Y}$, the Weight Share method can be used.

In general, the Weight Share method allocates to each sampled unit a basic weight established from an average of weights calculated within each cluster i entering into $\hat{Y}$. An *initial weight* that corresponds to the inverse of the selection probability is first obtained for unit k of cluster i of $\hat{Y}$ having a link $l_{j,ik} = 1$ with a unit $j \in s^A$. An initial weight of zero is assigned to units not having a link. The *basic weight* is obtained by calculating the mean of the initial weights for the cluster. This weight is finally assigned to all units within the cluster. Note that the fact of allocating the same basic weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters.

Formally, each unit k of cluster i entering into $\hat{Y}$ is assigned an initial weight w'_{ik} as follows:

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}, \tag{2}$$

where $t_j = 1$ if $j \in s^A$ and 0 otherwise. Note that a unit k having no link with any unit j of U^A automatically has an initial weight of zero.

The basic weight w_i is given by

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}}, \tag{3}$$

where $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik}$. The quantity L_{ik} represents the number of links between the units of U^A and the unit k of cluster i of population U^B. The quantity $L_i = \sum_{k=1}^{M_i^B} L_{ik}^B$ then corresponds to the total number of links present in cluster i.

Finally, we assign $w_{ik} = w_i$ for all $k \in i$.

2.1 Unbiasedness of the Weight Share Method

We now show that the estimator $\hat{Y}$ with the Weight Share method is unbiased for Y. Starting with $\hat{Y} = \sum_{i=1}^n w_i \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^n w_i Y_i$, we replace the definition of w_i in $\hat{Y}$ to get

$$\hat{Y} = \sum_{i=1}^n Y_i \left[\frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}} \right] = \sum_{i=1}^n \frac{Y_i}{L_i} \sum_{k=1}^{M_i^B} w'_{ik}.$$

Letting $z_{ik} = Y_i / L_i$ for all $k \in i$, we then have

$$\hat{Y} = \sum_{i=1}^n \sum_{k=1}^{M_i^B} w'_{ik} z_{ik}. \tag{4}$$

Let a single index k be used to identify the m^B units entering into $\hat{Y}(m^B = \sum_{i=1}^n M_i^B)$. By replacing w'_k by its definition (2), we obtain

$$\hat{Y} = \sum_{k=1}^{m^B} w'_k z_k$$

$$= \sum_{k=1}^{m^B} \left[\sum_{j=1}^{M^A} l_{jk} \frac{t_j}{\pi_j^A} \right] z_k.$$

Now since $t_j \neq 0$ only for the units k entering into $\hat{Y}$, we can extend the first summation to all units k in U^B. That is,

$$\hat{Y} = \sum_{k=1}^{M^B} \left[\sum_{j=1}^{M^A} l_{jk} \frac{t_j}{\pi_j^A} \right] z_k.$$

Rearranging $\hat{Y}$, we finally obtain

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{k=1}^{M^B} l_{jk} z_k$$

$$= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j. \tag{5}$$

Now, taking the expectation gives

$$E(\hat{Y}) = \sum_{j=1}^{M^A} \frac{E(t_j)}{\pi_j^A} Z_j$$

$$= \sum_{j=1}^{M^A} Z_j = Z$$

since $E(t_j) = \pi_j^A$.

It suffices now to show that $Z = Y$. First, we have

$$Z = \sum_{j=1}^{M^A} Z_j = \sum_{j=1}^{M^A} \sum_{k=1}^{M^B} l_{jk} z_k = \sum_{k=1}^{M^B} z_k \sum_{j=1}^{M^A} l_{jk}.$$

By rearranging these summations in terms of the N clusters of population U^B, we then obtain

$$Z = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} z_{ik} \sum_{j=1}^{M^A} l_{j,ik} = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} z_{ik} L_{ik}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \frac{Y_i}{L_i} L_{ik} = \sum_{i=1}^{N} Y_i = Y.$$

The unbiasedness of the Weight Share method can also be proved using an approach similar to the one presented by Ernst (1989).

2.2 Variance Estimation

To obtain a variance formula for $\hat{Y}$, we start with equation (5). Since $\hat{Y}$ turns out to be nothing more than a Horvitz-Thompson estimator of Z (see Horvitz and Thompson 1952), the variance of $\hat{Y}$ is then directly given by

$$\mathrm{Var}(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'} \qquad (6)$$

where $\pi_{jj'}^A$ is the joint probability of selecting units j and j' (see Särndal, Swensson and Wretman 1992 for the calculation of $\pi_{jj'}^A$ under various sampling designs).

In practice, equation (6) is simple to use. Initially, it suffices to calculate $z_k = Y_i / L_i$ for each unit $k \in i$. Then, we compute $Z_j = \sum_{k=1}^{M^B} l_{jk} z_k$. All that remains is to plug each Z_j into the variance equation of the Horvitz-Thompson estimator.

The variance $\mathrm{Var}(\hat{Y})$ may be unbiasedly estimated from the following equation:

$$\widehat{\mathrm{Var}}(\hat{Y}) = \sum_{j=1}^{m^A} \sum_{j'=1}^{m^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_{jj'}^A \pi_j^A \pi_{j'}^A} Z_j Z_{j'}. \qquad (7)$$

Another unbiased estimator of the variance $\mathrm{Var}(\hat{Y})$ may be developed in the form of Yates and Grundy (1953). Other variance estimators are available in the literature, such as jackknife variance estimators. A jackknife variance estimator in the context of the SLID sample design is discussed in Section 3.2.3. For further details, see Wolter (1985) and Särndal, Swensson and Wretman (1992).

3. APPLICATION TO SLID

In January 1994, SLID was launched by Statistics Canada. Its aim is to observe individual activity in the labour market over time, and changes in individual income and family circumstances. To repeat, the primary aim of SLID is to provide longitudinal data. However, cross-sectional estimates will also be produced. The target population of SLID is all persons, with no distinction as to age, who live in the provinces of Canada. For operational reasons, the Territories, institutions, Indian reserves and military camps are excluded (see Lavallée 1993).

3.1 Sample Design

3.1.1 Initial Sample

The SLID longitudinal sample was drawn in January 1993 from two groups rotating out of the Canadian Labour Force Survey (LFS), making the sample a sub-sample of the LFS. The longitudinal sample for SLID is made up of close to 15,000 households. A household is defined as any person or group of persons living in a dwelling. It may consist of one person living alone, a group of people who are not related but who share the same dwelling, or it may be a family.

LFS is a continuing survey designed to produce monthly estimates of employment, self-employment and unemployment. This survey uses a stratified multi-stage design which uses an area frame in which dwellings are the final sampling units. All the individuals who are members of households that occupy the selected dwellings make up the LFS sample. In other words, LFS draws a sample of dwellings and all individuals in the households that live in the selected dwellings are interviewed. A six-group rotation plan is used to construct the sample: every month, one group that has been in the sample for six months is rotated out. Each rotation group contains approximately 10,000 households, or approximately 20,000 individuals 16 years old or more. For further details on the LFS sample plan, see Singh *et al.* (1990).

For SLID, the longitudinal sample will not be updated following its selection in January 1993. However, to give the sample some cross-sectional representativeness, *initially-absent individuals* in the population (*i.e.*, individuals who were not part of the population in the year the longitudinal sample was selected) will need to be considered in the sample in January 1994 and later. Initially-absent individuals include *newborns* (births since January 1993) and *immigrants*. Note that this addition to the sample will be cross-sectional in that only the longitudinal individuals will be permanently included in the sample.

Table 1 presents the terminology developed for SLID. After sample selection in January 93 (year 1), the population contains longitudinal individuals and initially-present individuals. In January 94 (year 2), the population contains longitudinal individuals, initially-present individuals and initially-absent individuals. Focusing on the households containing at least one longitudinal individual (*i.e.*, *longitudinal households*), initially-present and initially-absent individuals who join these households are referred to as *cohabitants*.

Table 1

SLID Terminology

Individuals:

Longitudinal individuals: Individuals selected at year 1 in the longitudinal sample.

Initially-absent individuals: Individuals who were not part of the population in the year the longitudinal sample was selected (year 1). It includes in-migrants and newborns.

Initially-present individuals: Individuals who were part of the population of year 1 but were not selected then.

Cohabitants: Initially-absent and initially-present individuals who join a longitudinal household.

In-migrants: Individuals who, in January of year 1, were outside the ten provinces of Canada and individuals in excluded areas (the Territories, institutions, Indian reserves and military barracks).

Newborns: Births since January of year 1.

Households:

Longitudinal households: Households containing at least one longitudinal individual.

SLID will follow individual and household characteristics over time. At the time of each wave of interviews, all the members of a longitudinal household will be interviewed. The composition of the longitudinal households will change over time, as the result of a birth or the arrival of an in-migrant in the household. A part of the selection of initially-absent individuals may be based on individuals who join longitudinal households.

3.1.2 Supplementary Sample

The restriction to initially-absent individuals who join longitudinal households will unfortunately exclude households made up of initially-absent individuals only (*e.g.*, in-migrant families). To offset this shortcoming, one possibility is to draw a *Supplementary Sample*. This sample could be one of dwellings drawn directly from the ongoing LFS at each wave of interviews. Supplementary questions would then be added to the LFS questionnaire to detect households that contain *at least one in-migrant*; the households selected would then be interviewed.

Recalling that the Supplementary Sample is used for the selection of households made up solely of initially-absent individuals (*i.e.*, in-migrants and newborns), restricting this sample to in-migrants only would not cause any representativeness problem. This is because it is highly unlikely that households containing only newborns would be found: each household normally contains at least one adult. The newborns are then already represented in the sample by the longitudinal households. Now, if the Supplementary Sample were to include newborns in addition to in-migrants, significant costs would be added to the survey. This is because the Supplementary Sample would include a complete household for each newborn selected in the Supplementary Sample, producing excessive sample growth and unnecessary costs since the newborns are already represented in the sample.

One other approach different from using the ongoing LFS could be to select the Supplementary Sample by revisiting the dwellings used for the selection of the initial sample. This method offers some practical advantages, one being the facility to go to known addresses. This approach however would bring the problem of new dwellings which were not there in January 1993. These dwellings would have a zero probability of being selected in the Supplementary Sample and a bias would therefore be introduced. This is one reason favouring the first approach, *i.e.*, detecting households that contain at least one in-migrant via the questionnaire of the ongoing LFS.

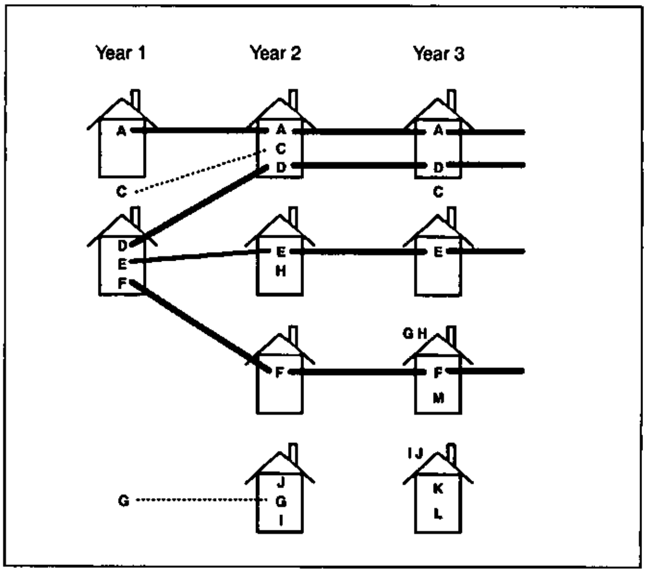

Figure 2. Selection of persons for SLID.

Figure 2 summarizes the longitudinal and cross-sectional selection of individuals. In Figure 2, the letters and houses represent individuals and households, respectively. Individuals A, D, E and F are longitudinal individuals whom we follow over time. Individual C is an initially-present individual, *i.e.*, an individual who was included in the population in year 1 but was not selected then. Initially-absent and initially-present individuals who join a longitudinal household are called cohabitants. In year 2, individual H represents an initially-absent individual who joins the sample as a cohabitant. The fourth house in year 2 represents a household selected for the Supplementary Sample of year 2 and in which individuals I and J are initially-absent individuals (with one of the two being necessarily an in-migrant since the Supplementary Sample is restricted to them). Individual G is an initially-present individual with the same status as C. In year 3, individuals C and H have left their longitudinal households and will therefore not be interviewed. Individuals I and J who were selected in the Supplementary Sample are now replaced with the individuals of the Supplementary Sample of year 3, *i.e.*, individuals K and L. Individual M is an initially-absent individual joining a longitudinal household as a cohabitant. It may finally be noted that, for cross-sectional purposes, a selected household may contain one or more longitudinal individuals, initially-present individuals and initially-absent individuals (newborns and in-migrants).

3.2 Basic Weighting

3.2.1 General Considerations

To produce cross-sectional estimates, the longitudinal sample augmented with initially-absent individuals and initially-present individuals must be weighted. The first step is to obtain a *basic weight* for each individual in each interviewed household. The basic weight is the weight prior to adjustment or post-stratification. It is, so to speak, the equivalent of the sample weight. Note that the basic weights are useful solely for cross-sectional estimation.

The basic weights are obtained from the selection probabilities. As described above, in January 1993 (year 1), we select for SLID a sample $s^{(1)}$ of $m^{(1)}$ individuals from a population $U^{(1)}$ of $M^{(1)}$ individuals. The sample is selected through dwellings which contain households. In other words, the $m^{(1)}$ individuals are obtained by selecting $n^{(1)}$ households from $N^{(1)}$, each household I being selected with probability $\pi_I^{(1)} > 0, I = 1, \ldots, N^{(1)}$. Let $M_I^{(1)}$ be the size of household I so that $M^{(1)} = \sum_{I=1}^{N^{(1)}} M_I^{(1)}$. Also let $\pi_j^{(1)}$ be the selection probability of individual j. This selection probability is retained throughout all waves of the survey.

For a given subsequent wave (which may be defined as year 2), the population U contains the $M^{(1)}$ individuals present at year 1, plus some $M^{(2)}$ initially-absent individuals (*i.e.*, initially absent from the population at year 1). The population of initially-absent individuals is indicated by $U^{(2)}$. Hence, the population $U = U^{(1)} \cup U^{(2)}$ contains $M = M^{(1)} + M^{(2)}$ individuals. Letting $U^{*(2)}$ be the population of $M^{*(2)}$ in-migrants of year 2, we have $U^{*(2)} \subseteq U^{(2)}$ and also $M^{*(2)} \leq M^{(2)}$. In our notation, the asterisk (*) is used to specify that the newborns have been excluded. The individuals of year 2 are contained in N households where household i is of size M_i, $i = 1, \ldots, N$.

For cross-sectional representativeness, some in-migrants are selected from the Supplementary Sample. At year 2, we then select a sample $s^{*(2)}$ of $m^{*(2)}$ individuals from the population $U^{*(2)}$ of $M^{*(2)}$ in-migrants. The Supplementary Sample is selected through households, *i.e.*, the $m^{*(2)}$ individuals are obtained by selecting $n^{*(2)}$ households. Let $\pi_j^{*(2)}$ be the selection probability of the in-migrant j. We assume $\pi_j^{*(2)} > 0$ for $j = 1, \ldots, M^{*(2)}$.

One implication of selecting in-migrants through households is that other individuals (such as newborns, initially-present individuals or longitudinal individuals) can be brought in by the Supplementary Sample by living in the same household as the selected in-migrants. However, since the selection units of the Supplementary Sample are restricted to the in-migrants, these other individuals are not properly selected, say, in the Supplementary Sample, even if they will be interviewed. The selection probabilities of these individuals are in fact not well defined.

The remaining in-migrants selected for cross-sectional representativeness are those individuals who join longitudinal households, who are then considered as cohabitants. As with the newborns and initially-present individuals of the previous paragraph, the addition of cohabitants to longitudinal households brings individuals with non-well defined selection probabilities.

The individuals with non-well defined selection probabilities have entered the survey process in a "non-legitimate" way. They complicate the determination of the basic weights, as their selection probability is not well defined. In order to override this difficulty, the Weight Share method is proposed.

3.2.2 Basic Weight Calculation

The Weight Share method described in Section 2 is now applied to the SLID sample, including the Supplementary Sample. The population U^A is here represented by the union of the two distinct populations $U^{(1)}$ and $U^{*(2)}$, *i.e.*, $U^A = U^* = U^{(1)} + U^{*(2)}$. The sample s^A of $m = m^{(1)} + m^{*(2)}$ individuals corresponds to the union of the two distinct samples $s^{(1)}$ and $s^{*(2)}$. The population U^B is represented by $U = U^{(1)} + U^{(2)}$. The population $U^A = U^*$ excludes the newborns while the population $U^B = U$ includes them. The clusters of population U^B simply correspond to the N households of year 2, and hence $M_i^B = M_i$.

One possible linkage between population U^A and U^B can be established by the same individuals in populations U^A and U^B. That is, $l_{jk} = 1$ if individual j in population U^A corresponds to individual k in population U^B, and $l_{jk} = 0$ otherwise. For each individual k not being a newborn, we then have $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = 1$. On the other hand, for each newborn k, we have $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = 0$ since they are excluded from U^A. We now have $L_i = \sum_{k=1}^{M_i^B} L_{ik} = M_i^*$ where M_i^* is the size of household i excluding the newborns.

Note that this last linkage is only one among several other possibilities. One other possible linkage would be to extend the linkage of the previous paragraph to all other persons within the household. That is, assign $l_{jk} = 1$ for all individuals k (of U^B) belonging to the same household i where individual j (of U^A) now belongs in U^B, and 0 otherwise. In other words, $l_{jk} = 1$ if individuals j and k belongs to household i. For each individual k in household i, we then have $L_{ik} = \sum_{j=1}^{M_i^A} l_{j,ik} = M_i^*$. We also get $L_i = \sum_{k=1}^{M_i^B} L_{ik} = \sum_{k=1}^{M_i^B} M_i^* = M_i^B M_i^*$. One can show that this linkage produces the same basic weighting as the one from the previous paragraph. Because the first linkage corresponds to a more natural way to link the individuals (*i.e.*, by linking only the same individuals between U^A and U^B), we will adopt the linkage proposed in the previous paragraph.

By considering the definition (2) of the initial weight w'_{ik} of individual k in household i, we obtain

$$w'_{ik} = \frac{t_{ik}^{(1)}}{\pi_{ik}^{(1)}} + \frac{t_{ik}^{*(2)}}{\pi_{ik}^{*(2)}}, \tag{8}$$

where $t_{ik}^{(1)} = 1$ if individual k is part of $s^{(1)}$ and 0 otherwise, $t_{ik}^{*(2)} = 1$ if individual k is part of $s^{*(2)}$ and 0 otherwise. This can be written more explicitly as

$$w'_{ik} = \begin{cases} 1/\pi_{ik}^{(1)} & \text{for} \quad k \in s^{(1)} \\ 1/\pi_{ik}^{*(2)} & \text{for} \quad k \in s^{*(2)} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Note that the first line of (9) corresponds to the longitudinal individuals. The second line corresponds to the in-migrants selected through the Supplementary Sample. The third line represents altogether newborns, cohabitants (if the household is a longitudinal household not part of the Supplementary Sample) and/or initially-present individuals (if the household is part of the Supplementary Sample).

The basic weight w_i of household i is obtained from

$$w_i = \frac{\sum_{k=1}^{M_i} w'_{ik}}{\sum_{k=1}^{M_i} L_{ik}} = \frac{1}{M_i^*} \sum_{k=1}^{M_i} w'_{ik}, \tag{10}$$

and finally $w_{ik} = w_i$ for $k \in i$.

Using the basic weights obtained from the Weight Share method, one can estimate the total $Y = \sum_{i=1}^{N} \sum_{k=1}^{M_i} y_{ik}$ of the characteristic y measured at year 2. The estimator used is the one given by equation (1). Using the definitions of the initial weights and the basic weights, $\hat{Y}$ can be rewritten as

$$\hat{Y} = \sum_{k=1}^{m^{(1)}} \frac{z_k^*}{\pi_k^{(1)}} + \sum_{k=1}^{m^{*(2)}} \frac{z_k^*}{\pi_k^{*(2)}}$$

$$= \hat{Z}^{*(1)} + \hat{Z}^{*(2)}, \tag{11}$$

where $z_k^* = \bar{Y}_i^*$ for $k \in i$ with $\bar{Y}_i^* = \sum_{k=1}^{M_i} y_{ik}/M_i^*$. Thus, estimator (11) is simply the sum of two Horvitz-Thompson estimators related to $s^{(1)}$ and $s^{*(2)}$. As shown in Section 2, this estimator is unbiased for Y.

3.2.3 Variance Estimation

The variance formula for $\hat{Y}$ is provided by equation (6). However, assuming that the two samples $s^{(1)}$ and $s^{*(2)}$ are selected independently, we have $\text{Var}(\hat{Y}) = \text{Var}(\hat{Z}^{*(1)}) + \text{Var}(\hat{Z}^{*(2)})$, where each term has the form of equation (6). For SLID, this assumption of independance holds if the selection of the Supplementary Sample is done through LFS.

Considering $\hat{Z}^{*(1)}$, we can re-index the individuals to reflect the fact that the $m^{(1)}$ individuals were selected at year 1 through $n^{(1)}$ households. This gives

$$\hat{Z}^{*(1)} = \sum_{k=1}^{m^{(1)}} \frac{z_k^*}{\pi_k^{(1)}} = \sum_{I=1}^{n^{(1)}} \sum_{j=1}^{M_I^{(1)}} \frac{z_{Ij}^*}{\pi_{Ij}^{(1)}}$$

$$= \sum_{I=1}^{n^{(1)}} \frac{1}{\pi_I^{(1)}} \sum_{j=1}^{M_I^{(1)}} z_{Ij}^* = \sum_{I=1}^{n^{(1)}} \frac{Z_I^{*(1)}}{\pi_I^{(1)}}, \tag{12}$$

since, by selecting complete households $\pi_{Ij}^{(1)} = \pi_I^{(1)}$ for $j \in I$. The variance $\text{Var}(\hat{Z}^{*(1)})$ is then directly obtained as

$$\text{Var}(\hat{Z}^{*(1)}) = \sum_{I=1}^{N^{(1)}} \sum_{I'=1}^{N^{(1)}} \frac{(\pi_{II'}^{(1)} - \pi_I^{(1)}\pi_{I'}^{(1)})}{\pi_I^{(1)}\pi_{I'}^{(1)}} Z_I^{*(1)} Z_{I'}^{*(1)}. \tag{13}$$

Considering $\hat{Z}^{*(2)}$, the individuals can also be re-indexed for consistency with $\hat{Z}^{*(1)}$, although this modification has no effect on the form of $\hat{Z}^{*(2)}$. Following the same steps used for $\text{Var}(\hat{Z}^{*(1)})$, $\text{Var}(\hat{Z}^{*(2)})$ is obtained as

$$\text{Var}(\hat{Z}^{*(2)}) = \sum_{I=1}^{N^{*(2)}} \sum_{I'=1}^{N^{*(2)}} \frac{(\pi_{II'}^{*(2)} - \pi_I^{*(2)}\pi_{I'}^{*(2)})}{\pi_I^{*(2)}\pi_{I'}^{*(2)}} Z_I^{*(2)} Z_{I'}^{*(2)}, \tag{14}$$

where $N^{*(2)}$ is the number of households of year 2 containing at least one in-migrant and $Z_I^{*(2)} = \sum_{j=1}^{M_I^{*(2)}} z_{Ij}^*$. The quantity $M_I^{*(2)}$ represents the number of in-migrants present in household I.

Finally, $\mathrm{Var}(\hat{Y})$ is simply given by

$$\mathrm{Var}(\hat{Y}) = \sum_{I=1}^{N^{(1)}} \sum_{I'=1}^{N^{(1)}} \frac{(\pi_{II'}^{(1)} - \pi_I^{(1)}\pi_{I'}^{(1)})}{\pi_I^{(1)}\pi_{I'}^{(1)}} Z_I^{*\,(1)} Z_{I'}^{*\,(1)}$$

$$+ \sum_{I=1}^{N^{*(2)}} \sum_{I'=1}^{N^{*(2)}} \frac{(\pi_{II'}^{*(2)} - \pi_I^{*\,(2)}\pi_{I'}^{*\,(2)})}{\pi_I^{*\,(2)}\pi_{I'}^{*\,(2)}} Z_I^{*\,(2)} Z_{I'}^{*\,(2)}. \quad (15)$$

The variance (15) may be unbiasedly estimated using the following equation:

$$\widehat{\mathrm{Var}}(\hat{Y}^*) = \sum_{I=1}^{n^{(1)}} \sum_{I'=1}^{n^{(1)}} \frac{(\pi_{II'}^{(1)} - \pi_I^{(1)}\pi_{I'}^{(1)})}{\pi_{II'}^{(1)}\pi_I^{(1)}\pi_{I'}^{(1)}} Z_I^{*\,(1)} Z_{I'}^{*\,(1)}$$

$$+ \sum_{I=1}^{n^{*(2)}} \sum_{I'=1}^{n^{*(2)}} \frac{(\pi_{II'}^{*(2)} - \pi_I^{*\,(2)}\pi^{*\,(2)})}{\pi_{II'}^{*(2)}\pi_I^{*\,(2)}\pi_{I'}^{*\,(2)}} Z_I^{*\,(2)} Z_{I'}^{*\,(2)}. \quad (16)$$

As SLID is in fact a sub-sample from LFS, the jackknife variance estimator developed for LFS (see Singh *et al.* 1990) may also be used, with minor modifications. In general, the jackknife method works as follows: the sample first is divided into random groups (or replicates, according to the LFS terminology). Then, each random group r is removed in turn from the sample and a new estimate $\hat{Y}_{(r)}$ of the total Y is computed. The different estimates $\hat{Y}_{(r)}$ are finally compared to the original estimate $\hat{Y}$ to obtain an estimate of the variance $\mathrm{Var}(\hat{Y})$. For further details on the jackknife method in general, see Särndal, Swensson and Wretman (1992).

Recall that LFS is based on a stratified multi-stage design which uses an area frame. Within each first-stage stratum h, the random groups (or replicates) correspond basically to the primary sampling units (PSUs). To compute the jackknife variance estimate for the estimation of the total Y, the following formula can be used:

$$\widehat{\mathrm{Var}}_J(\hat{Y}) = \sum_h \frac{(R_h - 1)}{R_h} \sum_{r \in h} (\hat{Y}_{(hr)} - \hat{Y})^2, \quad (17)$$

where R_h is the number of replicates in stratum h and $\hat{Y}_{(hr)}$ is the estimate of Y obtained after replicate r in stratum h is removed. For LFS, both $\hat{Y}$ and $\hat{Y}_{(hr)}$ are poststratified based on the integrated approach of Lemaître and Dufour (1987). The plan is to use the same poststratification approach for SLID but this discussion is out of the scope of the present paper.

ACKNOWLEDGEMENTS

The author would like to thank Carl Särndal, M.P. Singh, the Associate Editor and the referees for very useful comments which helped in improving the clarity of the paper. We would like also to gratefully acknowledge Jean-Claude Deville for suggesting the idea of an extension of the Weight Share method to a general context.

REFERENCES

ERNST, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*. (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley and Sons, 135-159.

GAILLY, B., and LAVALLÉE, P. (1993). Insérer des nouveaux membres dans un panel longitudinal de ménages et d'individus: simulations. CEPS/Instead, Document PSELL No. 54, Luxembourg, mai 1993.

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, 663-685.

KALTON, G., and BRICK, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.

LAVALLÉE, P. (1993). Sample representativity for the Survey of Labour and Income Dynamics. Statistics Canada, Research Paper of the Survey of Labour and Income Dynamics, Catalogue No. 93-19, December 1993.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of The Canadian Labour Force Survey*. Statistics Canada, Catalogue No. 71-526.

THOMPSON, S.K. (1992). *Sampling*. New York: John Wiley and Sons.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 235-261.

Weighting Schemes for Household Panel Surveys

GRAHAM KALTON and J. MICHAEL BRICK[1]

ABSTRACT

Household panel surveys often start with a sample of households and then attempt to follow all the members of those households for the life of the panel. At subsequent waves data are collected for the original sample members and for all the persons who are living with the sample members at the time. It is desirable to include the data collected both for the original sample persons and for the persons living with them in making person-level cross-sectional estimates for a particular wave. Similarly, it is desirable to include data for all the households for which data are collected at a particular wave in making household-level cross-sectional estimates for that wave. This paper reviews weighting schemes that can be used for these purposes. These weighting schemes may also be used in other settings in which units have more than one way of being selected for the sample.

KEY WORDS: Cross-sectional estimates; Fair share weighting; Multiplicity weighting; Panel surveys; Weight share method.

1. INTRODUCTION

National panel surveys of household economics have been mounted in many countries in recent years. The U.S. Panel Study of Income Dynamics (PSID), conducted by the Survey Research Center of the University of Michigan, began in 1968 and has been collecting data on an annual basis since that time (Hill 1992), and the British Household Panel Survey began in 1990 (Buck *et al.* 1994). Similar household panel surveys are also in progress or are being planned in most other European countries. The U.S. Bureau of the Census started to conduct the Survey of Income and Program Participation (SIPP) in 1983 (Nelson *et al.* 1985; Kasprzyk 1988; Jabine *et al.* 1990; Citro and Kalton 1993), and Statistics Canada introduced the Survey of Labour and Income Dynamics (SLID) in 1994 (Lavallée *et al.* 1993).

A common feature to most of these household panel surveys is that they start with a national sample of households, and then follow all the members of those households for the life of the panel. Over the course of time, household compositions change in a variety of ways. Some members of original sampled households leave those households to set up on their own or to join other households, as, for example, when a daughter leaves her parental household to get married. New members may join original sampled households, as, for example, when an elderly parent moves in with the family of a child or when a child is born to a household member. In order to be able to describe the economic circumstances of sample members at different points of time, household panel surveys usually collect data not only for the sample members but also for the individuals living with the sample members at the particular point of time. Following Lavallée (1995), these individuals are termed cohabitants in this paper. In other literature, they are often called associated persons or nonsample persons.

As the panel duration increases, the proportion of cohabitants in the sample at a wave rises. For example, in the 1984 SIPP panel, cohabitants comprise about 8.6 percent of the sample after one year and about 12.6 percent of the sample after two years (based on Table 1 in Kasprzyk and McMillen 1987). With a long-term household panel survey, the proportion of cohabitants becomes substantial after several years. The PSID, for example, defines sample members as all persons in the family units sampled in 1968 who are still alive, all the children born to these original sample members since the start of the panel, and the children of such children. In addition, the PSID collects data on the cohabitants who are living with sample members at each individual wave of data collection. Of the 20,535 individuals in interviewed family units in 1992, 41.2 percent were original sample members, 34.6 percent were the children of original sample members born since the start of the panel and children of such children, and 24.2 percent were cohabitants (excluding the Latino sample that was added in 1990) (Hill 1995).

This paper reviews methods of weighting the data collected from both sample persons and cohabitants in order to produce unbiased (or approximately unbiased) estimates of population parameters. In considering the analysis of a household panel survey, three different types of analysis may usefully be distinguished:

[1] Graham Kalton and J. Michael Brick, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.

- Cross-sectional analyses of households at a particular point in time;
- Cross-sectional analyses of individuals at a particular point in time;
- Longitudinal analyses of individuals over a period of time.

Weighting schemes for these three types of analysis are discussed in later sections. Longitudinal analyses of households over a period of time are not treated here because of the problematic nature of this type of analysis caused by changes in household composition (see, for example, Duncan and Hill 1985).

The weighting schemes used in household panel surveys need to account for the fact that households and individuals included in the survey at a particular wave may have more than one route by which they can be selected. At a given wave a household and its members are included in the sample if any of the original households (*i.e.*, households existing at the time of the initial selection) from which the current household has drawn members was selected. With the usual weighting approach, households are assigned weights inversely proportional to their joint selection probabilities, taking account of the different ways they can be selected. However, this approach cannot be applied with most household panel surveys because these joint selection probabilities cannot be determined. The alternative weighting approach reviewed here, termed by Lavallée (1995) the *weight share method*, avoids the need to know the joint selection probabilities of sample elements, but it introduces a random variation into the weights. Since this random variation results in a loss in precision of the survey estimates as compared with the inverse selection probability weighting scheme, this alternative approach should be considered only for situations where the joint selection probabilities cannot be ascertained. This situation often applies in household panel surveys and also in a number of other sample designs where elements can be selected by different routes.

In order to prepare for the discussion of weighting schemes for household panel surveys, the next section elaborates on the household changes that can occur over time, and the types of individuals involved. Sections 3, 4 and 5 then discuss weighting schemes that may be used for the three different forms of analysis described above. These sections deal with weighting schemes for unequal selection probabilities, without the complications of adjustments for nonresponse and noncoverage. The discussion relies heavily on previous work by Ernst (1989), Gailly and Lavallée (1993), Huang (1984), Judkins *et al.* (1984), Lavallée and Hunter (1992), and Little (1989). Section 6 then briefly reviews the issues involved in making adjustments to the weights to compensate for missing data arising from nonresponse and noncoverage. Section 7 presents some concluding remarks, and provides an illustration of another application of the weight share method.

2. CHANGES IN POPULATION AND HOUSEHOLD COMPOSITION OVER TIME

In analyzing a panel survey, it needs to be recognized that survey populations change over time. With household panel surveys it is important to distinguish between changes in population composition and changes in household composition.

The composition of a survey population changes over time because some individuals leave the population, some enter the population, and some may leave and join the population more than once. Individuals leave the population through death, emigration, or entering an institution (for surveys of the noninstitutional population). They enter the population through birth (or reaching the specified minimum age), immigration, and leaving an institution.

Households change composition over time for many different reasons, including deaths, births, marriages and divorces. For example, a household at time 1 may contain several individuals who end up in a number of different households at time 2. These individuals may set up new households on their own, they may join individuals who were in one or more households at time 1, or they may join individuals who were not in the population at time 1. One or more of the individuals may leave the population during the intervening period.

Consider a simple sample design in which households are selected independently at time 1 with equal probability. At time 2, the sample of households comprises all the households that contain one or more individuals from the households sampled at time 1, and the sample of individuals at time 2 comprises all the members of the sampled households at time 2. The samples of households and individuals at time 2 are selected with unequal probabilities. For instance, the selection probability of a household at time 2 that contains individuals from three households at time 1 is three times greater than that of a household at time 2 that contains individuals from only one household at time 1. Similarly, the individuals in that household have three times the probability of selection. Thus weighting schemes that compensate for these unequal selection probabilities are needed for the analysis of the resultant data.

Changes in population composition occur when individuals leave or enter the population. An individual sampled at time 1 who leaves the population before time 2 reduces the sample size for time 2 but does not otherwise affect cross-sectional estimates at time 2. In essence, the sampling frame for the time 2 population is the time 1 population, with the leavers in the intervening period being treated as blanks on the frame. Simply omitting the selected blanks from the time 2 sample causes no bias in the survey estimates (see, for example, Kish 1965). The situation with regard to entrants is, however, less straightforward. The household panel survey enumeration rule described above incorporates new entrants who join households that contain individuals who were eligible for the initial sample into the population for cross-sectional estimates for later time points. However, new entrants who set up their own households are not represented in person-level analyses at later waves of the panel. Equally, households composed of only new entrants are not represented in household-level analyses at later waves.

The failure of household panel surveys to cover households composed of only new entrants presents a problem for cross-sectional analyses of later waves of the panel. If these households and their members constitute a negligible proportion of the population, the solution may be to simply ignore the problem. However, if the proportion is appreciable, as can occur in later waves of a long-term panel, alternative solutions may be called for. One possibility is to add a supplementary sample of new entrants (*e.g.*, immigrants) to the panel, as discussed by Lavallée (1995) for the SLID. This solution is, however, often impracticable. Another solution is to limit the population of inference to persons who were members of the population at the start of the panel. New entrants found living with sample members are then excluded from the sample. This solution provides a clearcut definition of the population of inference. Whether the solution is appropriate depends on whether that definition can adequately satisfy the survey objectives.

Changes in population composition pose problems for longitudinal analyses of individuals. For many purposes, the population of inference is restricted to those who were present in the population throughout the time period of observation specified for the analysis. The inclusion of cohabitants in longitudinal analysis also creates problems. If the time period for the longitudinal analysis starts at the beginning of the panel, the analysis can be restricted straightforwardly to original sample members. If the time period starts later, it is tempting to include both original sample members and cohabitants joining the panel before the start of the analytic time period. However, the usual enumeration rules for household panel surveys specify that data are collected for cohabitants only while they continue to live with original sample members, that is, they are not followed if they cease to live with such persons. Unless the time period is short enough that the number of cohabitants who cease to live with sample persons in that period is negligible, this enumeration rule makes it problematic to include cohabitants in longitudinal analyses. This problem is discussed further in Section 5.

3. CROSS-SECTIONAL ESTIMATES FOR HOUSEHOLDS

This section considers weighting schemes that may be used to produce cross-sectional estimates for households for any wave of a household panel survey after the first. At the first wave a sample of households is selected and all the individuals in the sampled households become panel members to be followed throughout the life of the panel or until they leave the survey population. At a subsequent wave, wave t, the household sample comprises all the households in which panel members reside. Households that consist of new entrants only are not represented in the sample at later waves. Such households are ignored here. Complications of nonresponse are deferred until Section 6.

Consider the estimation of the total Y for all H households in the population at time t:

$$Y = \sum_{i=1}^{H} Y_i. \qquad (3.1)$$

A general estimator for this total can be expressed as

$$\hat{Y} = \sum_{i=1}^{H} w_i Y_i,$$

where w_i is a random variable that takes the value $w_i = 0$ if household i is not in the sample. The expectation of $\hat{Y}$ is

$$E(\hat{Y}) = \sum_{i=1}^{H} E(w_i) Y_i. \qquad (3.2)$$

By comparing equations (3.1) and (3.2), it can be seen that $\hat{Y}$ is unbiased for Y for any weighting scheme for which $E(w_i) = 1$ for all i.

There are many ways to satisfy the condition $E(w_i) = 1$. Three will be treated here. First, consider a standard *inverse selection probability weighting scheme*. The probability of a household being in the sample at time t is the probability of one or more of the households at time 1 from which it has drawn members being selected for the original sample. The probability of household H_i being in the sample at time t is then

$$\begin{aligned} P(H_i) &= P(h_j \cup h_k \cup h_\ell \cup \ldots) \\ &= \Sigma p_j - \Sigma\Sigma p_{jk} + \Sigma\Sigma\Sigma p_{jk\ell} - \ldots, \end{aligned} \qquad (3.3)$$

where $P(h_j \cup h_k \cup h_\ell \cup \ldots)$ is the selection probability of the union of original households h_j, h_k, h_ℓ, *etc.* for the original sample, p_j is the selection probability of original household h_j for the original sample, p_{jk} is the joint selection probability of original households h_j and h_k for the original sample, *etc.* and where households h_j, h_k, h_ℓ, *etc.* each contain at least one member who is currently in household H_i. The weight for each sampled household is then $w_i = 1/P(H_i)$. With this weighting scheme,

$$E(w_i) = P(H_i)[1/P(H_i)] + [1 - P(H_i)]0 = 1,$$

satisfying the condition for an unbiased estimator of a population total.

In practice, the computation $P(H_i)$ will generally not be as complex as equation (3.3) might suggest because the number of original households represented in household H_i is usually small. With, say, two original households involved, $P(H_i)$ reduces to

$$P(H_i) = P(h_1 \cup h_2) = p_1 + p_2 - p_{12}. \tag{3.4}$$

A problem with the application of the inverse selection probability approach is that p_j may be known only for households selected for the original sample, and not for other households. Also the joint probability may not be known. Even when the original sample was selected with equal probabilities, so that all the p_j are the same, the joint probability may depend on the sample design (for instance, whether the two households were in the same segment or not). The difficulty of obtaining $P(H_i)$ is a major drawback with the inverse selection probability approach.

An alternative strategy for developing the weights for time t is to base them only on the selection probabilities of households selected for the original sample, thus avoiding the difficulty in obtaining $P(H_i)$ noted above. One approach is to identify the set of households h_j at time 1 that would result in household H_i being in the sample at time t, and compute the weight for household H_i as

$$w_i = \sum_j \alpha_{ij} w'_{ij}, \tag{3.5}$$

where $w'_{ij} = 1/p_j$ if household h_j, which has at least one member in household H_i, was selected for the original sample and $w'_{ij} = 0$ if not, and where α_{ij} are any set of constants satisfying $\sum_j \alpha_{ij} = 1$.

With this approach,

$$E(w'_{ij}) = p_j(1/p_j) + (1 - p_j)0 = 1,$$

and hence

$$E(w_i) = \sum_j \alpha_{ij} = 1.$$

Thus, the use of weights w_i will yield unbiased estimators of totals for the household population for any choice of constants α_{ij}, provided that $\sum_j \alpha_{ij} = 1$. As indicated above, the principal advantage of this type of scheme is that it requires information only on the initial selection probabilities of the original households that were sampled at time 1, which are known. It does not require information on the initial selection probabilities of the other original households that have members in the current household, which are often not known.

A natural choice of α_{ij} is to make them equal for all the original households that lead to the selection of household H_i at time t. Huang (1984) terms this scheme a multiplicity approach. Here the scheme will be called an *equal household weighting scheme*. With this scheme

$$w_i = \sum_j w'_{ij}/C_i, \tag{3.6}$$

where C_i is the number of original households represented in household H_i at time t.

An alternative version of the above approach is one based on original sample persons rather than households. In this case, let I_{ijk} denote individual k from original household j in household i. Then

$$w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk},$$

where $w'_{ijk} = 1/p_j$ if individual k in household h_j was in the original sample and $w'_{ijk} = 0$ if not, and where the α_{ijk} are any set of constants satisfying $\sum_j \sum_k \alpha_{ijk} = 1$. Since the probability of an individual being selected for the original sample is the same as that of that individual's household,

$$E(w'_{ijk}) = p_j(1/p_j) + (1 - p_j)0 = 1.$$

In this case, the natural choice of the constants α_{ijk} is to make them equal for all members of the current household who were eligible for selection for the original sample. This produces what has been termed the fair share weighting scheme (Huang 1984; Ernst 1989). This scheme is termed here an *equal person weighting scheme*. With this scheme

$$w_i = \frac{1}{M_i} \sum_j M_{ij} w'_{ij},$$

where $w'_{ij} = w'_{ijk}$ is constant for all individuals in household H_i emanating from the same sampled household at time 1, M_{ij} is the number of individuals in household H_i coming from household h_j, and $M_i = \sum_i M_{ij}$ is the number of individuals in household H_i who were eligible for the sample at time 1. The equal person weighting scheme is applied in the SIPP and is proposed for use in the SLID.

Although developed here in terms of persons rather than households, it is readily apparent that the equal person weighting scheme could equally have been generated in terms of households. As shown above, the household weight $w_i = \sum_j \alpha_{ij} w'_{ij}$ satisfies the condition $E(w_i) = 1$ for any set of constants α_{ij} such that $\sum_j \alpha_{ij} = 1$. The equal household weighting scheme chooses $\alpha_{ij} = 1/C_i$, with $\sum_j \alpha_{ij} = 1$. The choice $\alpha_{ij} = M_{ij}/M_i$, with $\sum_j \alpha_{ij} = 1$, leads to the equal person weighting scheme.

It is instructive to compare the inverse selection probability weighting scheme with the equal household and equal person weighting schemes in a simple case. Following Little (1989), consider household H_i selected at time t

with household members coming from two original households. Let p_1 and p_2 denote the selection probabilities for the original households, and let p_{12} denote their joint selection probability. Under the inverse selection probability approach, the household weight is

$$w_i^* = \frac{1}{p_1 + p_2 - p_{12}},$$

as indicated above.

Under the equal person weighting scheme the weight for household H_i depends on which household or households were selected for the original sample:

$w_i = P_1/p_1$ if only household h_1 was selected;

$w_i = P_2/p_2$ if only household h_2 was selected;

$w_i = (P_1/p_1) + (P_2/p_2)$ if both h_1 and h_2 were selected;

where P_1 and P_2 are the proportions of members of household H_i who came from households h_1 and h_2, respectively (excluding any new entrants to the population). The probability of only household h_1 being selected is $(p_1 - p_{12})$, of only household h_2 being selected is $(p_2 - p_{12})$, and of both households being selected is p_{12}. The expected value of the weight conditional on household H_i being in the sample is thus

$$E(w_i \mid H_i \text{ in sample}) =$$

$$\frac{(p_1-p_{12})(P_1/p_1)+(p_2-p_{12})(P_2/p_2)+p_{12}[(P_1/p_1)+(P_2/p_2)]}{p_1+p_2-p_{12}},$$

i.e.,

$$E(w_i \mid H_i \text{ in sample}) = \frac{1}{p_1 + p_2 - p_{12}} = w_i^*.$$

As this result demonstrates, the weight for household H_i varies depending on which original households were selected, but in expectation the weight is the same as that obtained from the inverse selection probability approach.

Results for the expectation of the weight of household H_i under the equal household weighting scheme can be readily obtained as a special case of the above derivation in which $P_1 = P_2 = ½$. In expectation, the weight is the same as that for the inverse selection probability approach.

Given that the weight $w_i = \sum_j \alpha_{ij} w'_{ij}$ satisfies the condition $E(w_i) = 1$ for any set of α_{ij} such that $\sum_j \alpha_{ij} = 1$, the question arises as to the optimal choice of the α_{ij}. One approach is to choose the α_{ij} to minimize the variance of the estimated total $\hat{Y}$.

The variance of $\hat{Y}$ may be expressed as

$$V(\hat{Y}) = VE(\hat{Y} \mid s) + EV(\hat{Y} \mid s), \qquad (3.7)$$

where s denotes the set of households in the sample at time t. Now

$$E(\hat{Y} \mid s) = E\left(\sum_{i=1}^{H} w_i Y_i \mid s\right)$$

$$= \sum^{s} E(w_i \mid H_i) Y_i = \sum^{s} w_i^* Y_i = \hat{Y}^*,$$

where $\hat{Y}^*$ is the standard inverse selection probability estimator. Thus

$$VE(\hat{Y} \mid s) = V(\hat{Y}^*).$$

The first term in equation (3.7) is thus the variance of the standard inverse selection probability estimator, and the second term is the additional variance resulting from the use of weighting schemes from the class (3.5), $w_i = \sum_j \alpha_{ij} w'_{ij}$. The α_{ij} may then be chosen to minimize $EV(\hat{Y} \mid s)$.

Consider

$$V(\hat{Y} \mid s) = V\left(\sum^{H} w_i Y_i \mid s\right)$$

$$= \sum^{s} Y_i^2 V(w_i \mid H_i) +$$

$$\sum_{i \neq i'}\sum Y_i Y_{i'} \mathrm{Cov}(w_i, w_{i'} \mid H_i, H_{i'}).$$

Assuming $\mathrm{Cov}(w_i, w_{i'} \mid H_i, H_{i'}) = 0$,

$$V(\hat{Y} \mid s) = \sum Y_i^2 V(w_i \mid H_i)$$

$$= \sum Y_i^2 [E(w_i^2 \mid H_i) - w_i^{*2}],$$

since, as noted above, $E(w_i \mid H_i) = w_i^*$. Thus, assuming $\mathrm{Cov}(w_i, w_{i'} \mid H_i, H_{i'}) = 0$, $V(\hat{Y} \mid s)$ is minimized when $E(w_i^2 \mid H_i)$ is minimized.

Consider the application of this approach to the simple case discussed above in which H_i is composed of members from two original households and let $w_i = \alpha_i w'_{i1} + (1 - \alpha_i) w'_{i2}$. Then

$$E(w_i^2 \mid H_i) =$$

$$\frac{(p_1-p_{12})\dfrac{\alpha_i^2}{p_1^2} + (p_2-p_{12})\dfrac{(1-\alpha_i)^2}{p_2^2} + p_{12}\left(\dfrac{\alpha_i}{p_1} + \dfrac{1-\alpha_i}{p_2}\right)^2}{p_1+p_2-p_{12}}.$$

Minimizing $E(w_i^2 \mid H_i)$ is equivalent to minimizing

$$\Delta = (p_1 - p_{12}) p_2^2 \alpha_i^2 + (p_2 - p_{12}) p_1^2 (1 - \alpha_i)^2$$

$$+ p_{12}[(p_2 - p_1)\alpha_i + p_1]^2.$$

Then

$$\frac{\partial \Delta}{\partial \alpha_i} = 2(p_1 - p_{12})p_2^2\alpha_i - 2(p_2 - p_{12})p_1^2(1 - \alpha_i) \cdot + 2p_{12}(p_2 - p_1)[(p_2 - p_1)\alpha_i + p_1].$$

Solving $\partial\Delta/\partial\alpha_i = 0$ for α_i gives the optimum α_i as

$$\alpha_{oi} = \left(1 + \frac{p_2 - p_{12}}{p_1 - p_{12}}\right)^{-1}. \tag{3.8}$$

If the original households are selected independently, *i.e.*, $p_{12} = p_1p_2$,

$$\alpha_{oi} = \left[1 + \frac{p_2(1 - p_1)}{p_1(1 - p_2)}\right]^{-1} = \left[1 + \frac{\psi_2}{\psi_1}\right]^{-1}, \tag{3.9}$$

where $\psi_j = p_j/(1 - p_j)$ is the odds of original household h_j being selected.

Irrespective of whether the households are sampled independently, in the special case of an equal probability (epsem) sample of households initially, with $p_1 = p_2$,

$$\alpha_{oi} = \frac{1}{2}.$$

Thus, in the two-household case, the equal household weighting scheme minimizes the variance of the household weights around the inverse selection probability weight when the initial sample is an epsem one.

The optimal choice of α_{oi} given by (3.8) requires knowledge of p_1, p_2 and p_{12}, and that given by (3.9) requires independence and knowledge of p_1 and p_2. If these probabilities were known, then the standard inverse selection probability weight could be employed and would be preferable. In the case of an approximately epsem sample, the equal household weighting scheme should be close to the optimal, at least for the case where the members of the household at time t come from one or two households at the initial wave. This would apply, for instance, in the case of an epsem initial sample, with perhaps a few departures from epsem. With the equal household weighting scheme, when only one of the C_i original households, h_j, represented in H_i was selected for the original sample (as will generally be the case), then the weight for H_i is simply $1/C_ip_j$.

In the case of a non-epsem initial sample, the choice of the α_{ij} would ideally depend on the original household selection probabilities. However, since these probabilities are unknown, that approach cannot be applied. By default, the equal household or equal person weighting schemes may therefore be employed in this case. The use of these schemes (or any scheme with constant α_{ij}'s satisfying $\sum_j\alpha_{ij} = 1$) with a non-epsem initial sample still results in an unbiased estimate $\hat{Y}$. The drawback to these schemes in such a case is only that the α_{ij} are suboptimal in terms of minimizing the variance of $\hat{Y}$.

It should be noted that the equal household weighting scheme requires information on the number of original households h_j contributing members to household H_i at time t. That number may be difficult to determine in some cases. Consider, for example, a household at time t that contains two cohabitants. It may sometimes be difficult to determine whether these two persons were in a single household or in two separate households at the time of the initial sample selection. The equal person weighting scheme has the attractive feature of avoiding the need for Wave 1 household information, except for persons in sampled households at Wave 1. This feature provides an important reason for preferring the equal person to the equal household weighting scheme.

4. CROSS-SECTIONAL ESTIMATES FOR INDIVIDUALS

In producing cross-sectional estimates for individuals for any wave of a household panel survey after the first, it needs to be recognized that some new entrants will have joined the survey population since the start of the panel. New entrants who join households that contain one or more members of the original population can be represented in cross-sectional estimates for later waves, but new entrants living in households that do not contain any members of the original population are not covered (unless a special sample of them can be taken). The former type of new entrants is included in the weighting procedure described below, but the latter type is not.

Let there be N individuals in the population at time t, with N_i individuals in household $H_i (i = 1, 2, \ldots, H)$ and $\sum N_i = N$. The members of household H_i come from households h_j, h_k, h_ℓ, *etc.*, at time 1. Let M_{ij} denote the number of members of household H_i at time t who were in household h_j at the start of the panel. The sum $M = \sum\sum M_{ij}$ is less than the population size at time 1 because of leavers from the population in the period from time 1 to time t, and $M < N$ because of new entrants to the population who are in households containing members from the original population.

Consider now the estimation of a total for the population of individuals at time t:

$$Y = \sum_{i=1}^{H} \sum_{k=1}^{N_i} Y_{ik}. \tag{4.1}$$

where Y_{ik} is the value for individual k in household H_i. As in the household case discussed in the previous section, a general estimator for this total can be expressed as

$$\hat{Y} = \sum_{i=1}^{H} \sum_{k=1}^{N_i} w_{ik} Y_{ik}, \qquad (4.2)$$

where w_{ik} is a random variable that takes the value $w_{ik} = 0$ if individual k in household H_i is not in the sample. The estimator $\hat{Y}$ is unbiased for Y provided that $E(w_{ik}) = 1$ for all i and k.

As noted earlier, there are many ways to satisfy the condition $E(w_{ik}) = 1$. It is instructive to consider three of them. First, let $w_{ik} = 0$ for all individuals not in the original sample. In this case, the estimator $\hat{Y}$ discards cohabitants. Let p_{ik} denote the probability of a member of the original population, individual k residing in household H_i at time t, being selected for the initial sample, and let $w_{ik} = 1/p_{ik}$. Then, for such an individual

$$E(w_{ik}) = p_{ik}(1/p_{ik}) + (1 - p_{ik})0 = 1.$$

With this scheme, all new entrants to the population have $w_{ik} = 0$ with certainty. Thus $\hat{Y}$ in (4.2) provides an unbiased estimator of the total for the original population that is still present at time t, but does not include a component for the new entrants.

Modifications to the above procedure can be made to cover certain types of new entrants. For instance, births to sampled mothers can be included by assigning them the weights of their mothers, or if, as in the SIPP, the survey population is taken to be adults aged 16 and over, those under 16 at the start of the panel can be treated as sampled persons with assigned probabilities, and they can be included in the analyses of later waves after they have attained the age of 16. Such modifications do not, however, handle all types of new entrants. Provided that the proportion of other types of new entrants is small, this deficiency may not be a serious concern.

The weighting scheme that restricts the analysis to original sample persons, plus certain specified new entrants, is employed with the PSID. Its limitation is that it fails to make direct use of data collected for cohabitants. Such data may be used to provide information on the situation of sample persons, but the cohabitants are excluded from the sample for the analysis.

In order to include cohabitants in cross-sectional analyses for time t they need to be assigned positive weights. Noting that the probability of an individual being selected for the sample is the same as that of his or her household, weighting schemes for cross-sectional analyses of individuals at wave t can be obtained directly from those for households given in Section 3. Here we will develop the general strategy of producing weights for cross-sectional analysis at time t based only on the selection probabilities of members of the original sample, thus avoiding the problems with the inverse selection probability approach noted in Section 3.

Let I_{ijk} denote individual k from original household h_j who is now in household H_i. Let w_i denote the weight for every member of household H_i for cross-sectional analyses at time t, and let

$$w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk}$$

where $w'_{ijk} = 1/p_j$ if household h_j was in the original sample and $w'_{ijk} = 0$ if not. Then, as before, $E(w'_{ijk}) = 1$ for members of the original population. New entrants, for whom $p_j = 0$, may be handled by setting $\alpha_{ijk} = 0$. Then

$$E(w_i) = \sum_j \sum_k^{N_i} \alpha_{ijk} E(w'_{ijk}) = \sum_j \sum_k^{M_i} \alpha_{ijk} = 1$$

provided that $\sum_j \sum_k \alpha_{ijk} = 1$. Under this condition $\hat{Y}$ is unbiased for Y.

A natural choice of α_{ijk} is to set $\alpha_{ijk} = 1/M_i$ for all members of the original population. This is the equal person weighting scheme in which every member of household H_i at time t (including new entrants) receives the weight

$$w_i = \sum_j \sum_k w'_{ijk}/M_i.$$

Another choice of the α_{ijk} is that used for the equal household weighting scheme. Let C_i denote the number of original households that have members in household H_i at time t. Then $\sum_j \sum_k \alpha_{ijk} = 1$ can be divided equally between households, with each member of original household h_j being assigned a value of $\alpha_{ijk} = 1/C_i M_{ij}$. Then for original household h_j

$$\sum_k^{M_{ij}} \alpha_{ijk} = 1/C_i.$$

The derivation of the α_{ijk} to minimize the variance of the estimated total $\hat{Y}$ for the population of individuals follows directly from the corresponding derivation for the population of households given in Section 3. The estimated total for the population of individuals is

$$\hat{Y} = \sum_i^s \sum_k^{N_i} w_{ik} Y_{ik} = \sum_i^s \sum_k^{N_i} w_i Y_{ik},$$

since the weights for every individual in sampled household H_i are the same. This estimated total can be expressed as

$$\hat{Y} = \sum_i^s w_i Y_i,$$

where $Y_i = \sum_k Y_{ik}$ is the household total for H_i. Thus $\hat{Y}$ can be expressed as a household total, and the results of Section 3 can be applied directly.

Consider the example from Section 3 in which H_i is composed of members from only two original households, perhaps together with one or more new entrants. In this case the person-level weight $w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk}$ reduces to

$$w_i = \left(\sum_k \alpha_{i1k}\right) w'_{i1} + \left(\sum_k \alpha_{i2k}\right) w'_{i2}$$

$$= \alpha_i w'_{i1} + (1 - \alpha_i) w'_{i2},$$

where $\alpha_i = \sum_k \alpha_{i1k}$. As shown in equation (3.8), the optimum value of α_i is

$$\alpha_{oi} = \left(1 + \frac{p_2 - p_{12}}{p_1 - p_{12}}\right)^{-1}.$$

The individual values α_{ijk} are not needed for computing the w_i; only the original household totals $\sum_k \alpha_{ijk}$ are required. If individual values are needed for the α_{ijk}, they may be simply assigned as $\sum_k \alpha_{ijk}/M_{ij}$.

As in the household case, the optimum weighting α_{oi} requires knowledge of p_1, p_2 and p_{12}. If these probabilities are known, the standard inverse selection probability weight w_i^* can be computed, and would be preferred. In the case of an approximately epsem sample, the equal household weighting scheme should fare well. However, the equal household weighting scheme requires information on the number of original households contributing members to current household H_i, and this information may not always be available. As discussed in Section 3, for this reason the equal person weighting scheme may be preferred.

5. LONGITUDINAL ANALYSES OF INDIVIDUALS

A key analytic advantage of a panel survey is the ability to conduct longitudinal analyses relating variables for the same sampled units measured at different time points. Since all persons in original sampled households are followed throughout the life of the panel or until they leave the survey population, the data they provide may be readily analyzed longitudinally for any time period within the panel's time span (although nonresponse adjustments may be needed for panel attrition). Thus, for example, in a ten-year panel, data for original sampled persons may be analyzed from year 1 to year 10, from year 5 to year 9, or for any other period. New entrants (*e.g.*, births) may be included in the analysis for periods beginning after the start of the panel provided that they are treated as panel members who are followed throughout the panel even when they leave the households of original sampled persons.

Given the weighting schemes described in the previous section, cohabitants can be included in cross-sectional analyses of later waves. These weighting schemes provide a cross-sectional representation of the population at any wave of the panel (apart from new entrants not living with original population members). It is then possible to consider all the sample of original sample members and cohabitants at time t as the initial sample of a new panel that may be used for longitudinal analyses from time t to $(t + k)$. This procedure is, for instance, used in the SIPP, where all original sample members and cohabitants present at the start of the second year of the panel are included in analyses relating to that year.

The limitation to the inclusion of cohabitants in longitudinal analysis is that the following rules used in most household panel surveys specify that cohabitants are dropped from the panel if they cease living with original sample persons. Thus, cohabitants who live with original sample members at the start of the analysis period but who cease to live with them before the end of that period effectively become nonrespondents. If the analysis period is relatively short, the number of such nonrespondents may be small and the risk of serious nonresponse bias may be negligible. If the analysis period is a long one, however, the number of not-followed cohabitants may be appreciable, causing concerns about potential bias. The issue here is one of a trade-off between the reduced variance due to the increase in sample size from including cohabitants in the analysis versus the increased bias resulting from the additional nonresponse caused by failing to follow cohabitants leaving the households of original sample persons.

The additional nonresponse bias can be avoided by changing the following rules to specify that cohabitants are to be followed from the time they join the panel for the rest of the life of the panel, or until they leave the survey population, irrespective of whether they continue to live with original sample members. This change, however, leads to an expanding panel and the need for additional resources. Not only do data need to be collected for cohabitants at waves after they cease to live with sample persons, but data also need to be collected for any persons with whom the cohabitants live at later waves.

6. ADJUSTMENTS TO COMPENSATE FOR NONRESPONSE AND NONCOVERAGE

The discussion thus far has assumed that data are collected for all sampled persons and their cohabitants and that all the target population is covered by the sampling procedures. In practice both these assumptions are violated. Nonresponse is present in nearly all surveys and is of particular concern in household panel surveys, where some sampled households fail to respond at the initial wave and others fail to respond at some of the subsequent waves. The sampling frames used in most surveys are subject to some degree of noncoverage, and in later waves of household panel surveys there is an additional source of noncoverage associated with new entrants to the population who are not living with members of the original population.

In a simple cross-sectional survey, missing data can be classified into item nonresponse, total nonresponse and noncoverage. Imputation procedures can then be used to assign values for item nonresponses, weighting adjustments can be applied to compensate for total nonresponse, and poststratification adjustments can be applied to compensate for nonresponse and noncoverage. The situation is made far more complex in panel surveys by the occurrence of wave nonresponse, which arises when a sampled element responds for some but not all of the waves for which it was eligible. Not only do methods need to be devised to compensate for wave nonresponse, but also the preferred methods of compensation may depend on the type of analysis to be performed, in particular whether cross-sectional or longitudinal analyses are to be conducted.

From one perspective wave nonresponse can be viewed as a set of item nonresponses in the element's longitudinal record, suggesting that imputation may be used to fill in the missing values. Alternatively, it can be treated as total nonresponse, handled by weighting adjustments. The imputation approach is more natural for the creation of a panel file for longitudinal analysis, whereas the weighting approach is more natural for the creation of a cross-sectional file for the analysis of the data collected at a single wave.

The attraction of the imputation approach with a longitudinal file is that it retains all the reported data, whereas the weighting approach discards the reported data for all the elements that fail to provide data for one or more waves for which they were eligible. However, the imputation approach may involve the fabrication of a large amount of data, especially when an element fails to respond at several waves. Thus, for panel files, a compromise solution may be preferred, imputing responses for elements with few missing waves and using weighting adjustments to compensate for those with several missing waves (including total nonrespondents). In the SIPP, for example, imputation is used to assign responses for sample persons with a single missing wave that is bounded on both sides by responding waves, and weighting adjustments are used for all other sample persons with missing waves (Singh *et al.* 1990). Further discussion of methods of handling wave nonresponse in panel files is provided by Lepkowski (1989), Kalton (1986), and Lepkowski *et al.* (1993).

Another complication of some household panel surveys is the occurrence of partial household nonresponse, which occurs when the survey data are collected for some but not all members of a sampled household at a particular wave. The lack of data for one individual in a household means that key household characteristics (*e.g.*, household earnings) cannot be computed. One solution is to drop the household and its responding members from the sample, and use a weighting adjustment. Another is to impute the responses for the nonresponding household members, as is done in SIPP (where they are termed Type *z* nonrespondents). With the latter solution, data are available for all members of responding households, and hence person-level adjustments are unnecessary within responding households.

We now turn to consider the issues involved in dealing with missing data for cross-sectional analyses of a household panel survey. A separate cross-sectional file containing data for all responding households and their members (either deleting the households or imputing values for missing responses in the case of partial household nonresponse) can be created for each wave. Adjustments are then needed to compensate for the nonresponding and noncovered households and persons in each file.

Nonresponding households at wave t can be divided into total nonrespondents and wave nonrespondents. Total nonresponse occurs in a panel survey when a sampled element fails to provide data for any wave. Since it is common practice not to follow up sample households that fail to respond at the initial wave, these households and their members are generally the total nonrespondents. Compensation for total nonrespondents is relatively straightforward. The Wave 1 weights of the responding households at the initial wave can be adjusted using standard nonresponse adjustment methods and the adjusted weights can be used instead of the selection probabilities in developing the cross-sectional weights for later waves. Most nonresponse adjustment methods, such as weighting class adjustments (Kalton and Kasprzyk 1986) and adjustments based on response propensities (Little 1986), are based on the assumption that nonresponse is random within weighting classes or that the probabilities of responding within a class can estimated precisely. Under these conditions, the response mechanism can be treated as an additional stage of sampling. Thus, the selection probabilities, p_j, used to define the weights in equation (3.5) may be redefined as the product of the selection probabilities and the adjustment due to nonresponse. For example, if weighting class adjustments are used, the selection probability of original household h_j multiplied by the weighted response rate for the weighting class in which h_j falls is used instead of the original p_j. The previous results then follow for the weights adjusted for total nonresponse.

The same approach can also be extended to cover weighting adjustments for households responding at the initial wave that lead to no responding households at wave t. In this case, the responding households at the initial wave can be divided into weighting classes based on responses given at that wave, and the weights of households leading to one or more responding households at wave t

can be further adjusted to compensate for those leading to no responding households at wave t. The revised w'_{ij} can then be employed in equation (3.5) and subsequently.

Both the above nonresponse adjustments are applied in relation to the original households. A further type of household nonresponse cannot be handled in this way. This type of nonresponse involves the situation where an original household splits into two or more separate households at wave t, and where some but not all of those households respond at that wave. In this case the adjustment for the nonresponding households needs to be made in relation to the wave t households, H_i, rather than the original households, h_j. If the number of original households having members in each wave t nonresponding household of this type were known, the weights w_i for these households could be computed using the approach described above. Then weighting adjustments could be readily applied within weighting classes of the wave t households to compensate for the nonresponding households. In practice, however, the number of original households having members in a nonresponding household at wave t may often be unknown. One approach for handling this situation is to estimate this number by the average number for responding households at wave t that have similar characteristics to (*e.g.*, they are also splits from original households), and are in the same weighting class as, the nonresponding household. Using such estimated numbers where necessary, the weights w_i can be determined for all nonresponding households of the type being discussed. Standard weighting adjustments can then be applied to the responding households at wave t to compensate for these nonresponding households.

Incomplete coverage of the target population is another nonsampling problem that has been traditionally addressed in surveys by adjusting the sampling weights. For example, poststratification (see, for example, Holt and Smith 1979) and generalized raking procedures (Deville *et al.* 1993) are often used to adjust the weights so that they sum to counts from independent sources not subject to undercoverage. These adjustments may also reduce the sampling errors of the estimates, although bias reduction is often more critical.

The control totals used in most household surveys are counts of the number of persons in classes defined by characteristics such as age, sex and race. This method of reducing undercoverage bias may be fully sufficient when estimates of persons are the only types of statistics to be produced from the survey. However, further steps are needed to calculate household-level weights for producing statistics of household characteristics.

One approach to developing household-level weights when control totals are based on person-level counts is called the principal person method, as described by Alexander (1987). In this method, poststratification adjustments are applied at the person level. One household member is then identified as the principal person and the fully adjusted weight for that person is assigned to be the household weight. Since the person weights are already adjusted to the control totals, the household weight does incorporate some adjustments to reduce coverage bias. For cross-sectional estimation from a household panel survey, the principal person method can readily be used in conjunction with the equal household and person weighting schemes to produce household level weights.

A disadvantage of the principal person method is that estimates of the number of persons calculated using the principal person weight will generally differ from control totals. The estimates may also differ significantly depending on the criteria used to identify the principal person in the household. The method has also been criticized because the weights of the members of the same household differ, even though they were all selected at the same rate as the household. Estimation schemes proposed by Alexander (1987), Lemaître and Dufour (1987), and Zieschang (1990) address these objections by constraining the household weights so that they are consistent with the independent person-level totals while minimizing the distance between the original household weights and the adjusted weights. All three consider variants of a generalized least squares (GLS) algorithm to achieve this objective. Zieschang (1990) shows how GLS can be used to create weights that are consistent with the person controls and force all persons within a household to have the same weight.

The application of GLS methods when the household weights are computed using the equal household and person weighting schemes is relatively straightforward. However, empirical evaluation of the consequences of using these methods is needed. The GLS methods have the unattractive feature that they can result in negative weights. Furthermore, the increase in the variation in the weights arising from the constraints imposed may result in less precise estimates. This concern may be especially important when the variability in the household weights is increased due to their multiple routes for selection and the equal household or person weighting schemes are necessary.

7. SUMMARY AND CONCLUDING REMARKS

This paper has described weighting schemes for cross-sectional analysis of later waves of a household panel survey using data for all households for which and all individuals for whom data are collected. These weighting schemes can accommodate new entrants to the population who move in to live with members of the original population, but not other new entrants.

The usual inverse selection probability weighting scheme requires information on the household selection probabilities of all members of the households sampled at a later wave, as well as the joint selection probabilities of the original households that contribute members to the later wave households. The inverse selection probability weighting scheme can often not be applied because these probabilities are unknown. To deal with this problem, an alternative approach that requires information on only the selection probabilities of sampled original households is described.

This alternative approach produces a class of weighting schemes including the equal person (fair share) scheme used in SIPP and the equal household weighting scheme. All the schemes in this class produce weights that are in expectation equal to those produced by the usual inverse selection probability scheme. The variance in the weights around the inverse selection probability weights gives rise to an increase in the variance of the survey estimates. When the original households are selected with approximately equal probability, the equal household weighting scheme is near optimal for both household and individual level analyses to control this increase in variance.

The alternative class of weighting schemes produces unbiased estimates of population totals for any choice of constant α_{ij} that satisfies the condition $\sum_j \alpha_{ij} = 1$ and for any initial sample design. The equal household and equal person weighting schemes are, however, suboptimal for non-epsem initial samples. One of them may nevertheless be the appropriate scheme for such designs, because the optimal choice of the α_{ij} depends on the unknown initial selection probabilities, and hence cannot be determined. The equal household and equal person weighting schemes have different data requirements, in that the former requires knowledge of the number of Wave 1 households represented in the Wave t household whereas the latter does not. The fact that this information may not always be readily obtainable thus argues in favor of the equal person weighting scheme.

The cross-sectional individual weights for a particular wave can be used as the starting weights for a longitudinal analysis that begins at that wave. This procedure includes cohabitants present at that wave in the longitudinal analysis. However, if cohabitants are not followed when they cease to live with sampled persons, those who leave sample persons before the end of the period of the longitudinal analysis become nonrespondents. Before cohabitants are included in a longitudinal analysis, a check should therefore be made to ensure that their inclusion will not give rise to risks of serious nonresponse bias.

The class of weighting schemes described has a broader range of application than that indicated here. It can in fact be usefully applied in any situation where an inverse selection probability weighting scheme would be appropriate, but where not all the inclusion probabilities and joint inclusion probabilities are known. Consider, for instance, the modified version of the Mitofsky-Waksberg random digit dialing sampling procedure for telephone surveys described by Brick and Waksberg (1991). A sample of telephone numbers (primes) is selected at the first stage of this two-stage sample design. If a prime number is found to be a working residential number, that household is selected and a fixed number of additional telephone numbers in the same 100-bank is selected. The households found at these numbers are then all included in the sample. If a prime number is not a working number, the sampling process stops. With this procedure, the probability of a working residential number being selected depends on the number of working residential numbers in its 100-bank, and hence differs across 100-banks. This probability can be estimated from the sample of telephone numbers in the 100-bank. A complication arises, however, when a sampled household has two or more telephone numbers. In this case, the selection probability of the sampled telephone number can be estimated, but those of the nonsampled numbers cannot. Thus, the standard inverse selection probability weighting scheme cannot be used. However, the alternative weighting scheme described here can be employed.

ACKNOWLEDGEMENTS

We thank the referees for their helpful comments on the earlier version of the paper. The paper reports research undertaken for the U.S. Census Bureau. The views expressed are the authors'. They do not necessarily reflect the views of the Census Bureau.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

BRICK, J.M., and WAKSBERG, J. (1991). Avoiding sequential sampling with random digit dialing. *Survey Methodology*, 17, 27-41.

BUCK, N., GERSHUNY, J., ROSE, D., and SCOTT, J. (Eds.) (1994). *Changing Households: The British Household Panel Survey 1990-1992*. Colchester, U.K.: ESRC Research Centre on Micro-social Change.

CITRO, C.F., and KALTON, G. (1993). *The Future of the Survey of Income and Program Participation*. Washington D.C.: National Academy Press.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DUNCAN, G.J., and HILL, M.S. (1985). Conceptions of longitudinal households: Fertile or futile? *Journal of Economic and Social Measurement*, 13, 361-375.

ERNST, L.R. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley, 139-159.

GAILLY, B., and LAVALLÉE, P. (1993). *Insérer des Nouveaux Membres dans un Panel Longitudinal de Ménages et D'Individus: Simulations*. Walferdange, Luxembourg: CEPS/Instead.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.

HILL, M.S. (1995). Personal Communication.

HOLT, D., and SMITH, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society*, A, 142, 33-46.

HUANG, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 670-675.

JABINE, T.B., KING, K.E., and PETRONI, R.J. (1990). *Survey of Income and Program Participation: Quality Profile*. Washington D.C.: U.S. Bureau of the Census.

JUDKINS, D., HUBBLE, D., DORSCH, J., MCMILLEN, D., and ERNST, L. (1984). Weighting of persons for SIPP longitudinal tabulations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 676-687.

KALTON, G. (1986). Handling wave nonrespone in panel surveys. *Journal of Official Statistics*, 2, 303-314.

KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

KASPRZYK, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*. SIPP Working Paper No. 8830. Washington D.C.: U.S. Bureau of the Census.

KASPRZYK, D., and MCMILLEN, D.B. (1987). SIPP: Characteristics of the 1984 Panel. *Proceedings of the Social Statistics Section, American Statistical Association*, 181-186.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.

LAVALLÉE, P., and HUNTER, L. (1992). Weighting for the Survey of Labour and Income Dynamics. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, 65-75.

LAVALLÉE, P., MICHAUD, S., and WEBBER, M. (1993). The Survey of Labour and Income Dynamics, design issues for a new longitudinal survey in Canada. *Bulletin of the International Statistical Institute*, 49th Session, Contributed Papers, Book 2, 99-100.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

LEPKOWSKI, J. (1989). Treatment of wave nonresponse in panel surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley, 348-374.

LEPKOWSKI, J.M., MILLER, D.P., KALTON G., and SINGH, R. (1993). Imputation for wave nonresponse in the SIPP. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, 99-109.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

LITTLE, R.J.A. (1989). Sampling weights in the PSID: Issues and comments. Panel Study of Income Dynamics Working Paper, Ann Arbor: University of Michigan.

NELSON, D., MCMILLEN, D., and KASPRZYK, D. (1985). *An Overview of the SIPP, Update* 1. SIPP Working Paper No. 8401. Washington D.C.: U.S. Bureau of the Census.

SINGH, R., HUGGINS, V., and KASPRZYK, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys*. SIPP Working Paper No. 9009. Washington D.C.: U.S. Bureau of the Census.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

Modelling Net Undercoverage in the 1991 Canadian Census

PETER DICK[1]

ABSTRACT

In 1991, Statistics Canada for the first time adjusted the Population Estimates Program for undercoverage in the 1991 Census. The Census coverage studies provided reliable estimates of undercoverage at the provincial level and for national estimates of large age - sex domains. However, the population series required estimates of undercoverage for age - sex domains within each province and territory. Since the direct survey estimates for some of these small domains had large standard errors due to the small sample size in the domain, small area modelling techniques were needed. In order to incorporate the varying degrees of reliability of the direct survey estimates, a regression model utilizing an Empirical Bayes methodology was used to estimate the undercoverage in small domains. A raking ratio procedure was then applied to the undercoverage estimates to preserve consistency with the marginal direct survey estimates. The results of this modelling process are shown along with the estimated reduction in standard errors.

KEY WORDS: Small area; Empirical Bayes; Undercoverage.

1. INTRODUCTION AND BACKGROUND

The Census of Canada is conducted every five years; one of its objectives is to provide the Population Estimates Program with accurate baseline counts of the number of persons by age and sex within each province and territory. Unfortunately, not all eligible persons are correctly enumerated by the Census. As part of the evaluation of the Census, Statistics Canada estimates, through two sample surveys, the net number of persons missed by the Census. The estimates are from the Reverse Record Check Study, which estimates the gross number of persons missed by the Census, and the Overcoverage Study, which estimates persons double counted or erroneously included in the final Census count. When combined the figures estimate the net number of people missed by the Census.

These surveys were designed to produce reliable direct estimates for large areas, such as provinces, and for large domains, such as age - sex combinations at the national level. However, the Population Estimates Program requires estimates of missed persons for single year of age for both sexes for each province. However using the direct survey estimate would result in individual estimates having unacceptably high standard errors due to insufficient sample in the small domain. One approach to reducing the variance of the small domain estimates would be to borrow strength from related domains. This approach leads to creating an explicit model for the small domain that can be used to predict the net missed persons in that domain.

The result of modelling the small domain estimates is to produce a series of estimates with a smaller Mean Square Error than the direct estimate. However, as opposed to the direct survey estimate which is design unbiased, the modelling approach will introduce a bias for each estimate. Thus modelling the small domain estimates implies that a trade off is required between reducing the variance of each estimate and the bias introduced through the modelling process. One approach to ensuring that the more reliable direct survey estimates are utilized is to introduce an Empirical Bayes model. This procedure creates an estimate that is a combination of a model estimate and the direct survey estimate weighted by their respective variances. It is an Empirical Bayes estimate instead of a Bayes estimate because underlying parameters are first estimated, then these estimated parameters are considered known in later calculations. Note that since the individual sampling variances are used in the estimation, a more precise direct estimate would contribute much more to the final Empirical Bayes estimate than a similar estimate with low precision. This ensures that the model does not dominate estimates that are already considered reliable. It is also possible to approach this estimation problem through a Hierarchical Bayes methodology: details on this method can be found in Datta, *et al.* (1992). Ghosh and Rao (1994) give an appraisal of both the Hierarchical Bayes and Empirical Bayes approaches to small area estimation.

Outside of Canada, two different approaches to smoothing the Census undercoverage have been described in the literature. In the United States, the net undercoverage in the 1990 American Census was evaluated by means of the Post Enumeration Survey (Hogan 1992). Initially, it was planned to multiply the US Census counts by adjustment factors (the ratio of true population over the enumerated population) for 1,392 *a priori* defined post strata.

[1] Peter Dick, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

These estimated adjustment factors would then be used to adjust the Census count for missed persons. Since some of these 1,392 estimated adjustment factors had high standard errors, it was proposed to smooth the direct estimates through an Empirical Bayes regression model, similar to one proposed by Ericksen and Kadane (1985), and then to rake the smoothed estimates to agree with direct estimates for large geographic regions. However, this approach was criticized by Freedman and Navidi (1992). Eventually, the United States Department of Commerce, the U.S. Census Bureau's parent agency, decided not to proceed with adjusting the Census counts for under-enumeration in July 1991. Consideration was also given in the United States to adjusting the post Censal population estimates for undercoverage in the Census, but the Department of Commerce also rejected this adjustment.

The Australians use a different method than the Americans for estimating the domain totals. Choi, Steel and Skinner (1988) describe a methodology that incorporates the estimates of net undercoverage of the Census into the population estimates but leaves the actual Census counts as enumerated. The under enumeration is estimated through a Post Enumeration Survey (PES) and demographic analysis. The small domain estimates are produced by raking the Census age counts for each sex to the PES estimates for national age/sex totals and part of State/Territory/sex totals.

The procedure proposed for the 1991 Canadian Census combines some of the elements of both the American and Australian approaches. As in the American procedure, a model is postulated for the underlying true adjustment factors and another model is postulated for relating the direct survey estimates to the true underlying adjustment factors. Through Empirical Bayes estimation, a new smoothed adjustment factor is estimated that will have a lower MSE then the direct survey estimate. These smoothed adjustment factors are then converted into estimates of missed persons. The Australian method for constraining the resulting estimates to known marginal totals is then adopted. These final raked estimates are used as the base for the small domain estimates of missed persons. In turn, these estimates are adjusted to account for known demographic principles (See Michalowski 1993). Details on the technical criteria for adjustment of the population estimates can be found in Royce (1992).

This paper is organized as follows. In Section 2, some background information on the two sample surveys is described and the basic Empirical Bayes model is presented. Assumptions and limitations of the model are also discussed and the estimation of the parameters is briefly discussed. In Section 3, the explanatory variables used in the regression model are presented and the model building process is described. The final model is presented and the results displayed. Section 4 presents a discussion on the rationale behind constraining the Empirical Bayes estimates to reliable marginal totals. The final adjusted estimates are then presented. Finally, Section 5 presents some conclusions and topics for further study.

2. MODEL FOR THE ADJUSTMENT FACTORS

2.1 Background and Notation

The model for the adjustment factors requires input data. The actual data originates with two coverage studies: the Reverse Record Check (RRC) and the Overcoverage Study (OCS). The RRC is used to estimate the number of persons missed by the Census while the OCS is used to estimate the number of persons erroneously included in the Census count. These surveys are designed to give reliable estimates of net undercoverage for all provinces, some of the larger metropolitan areas and for some large national domains, such as males aged 20 to 24. Since the surveys are independent, it can be assumed that the variance of net missed persons will be the sum of the two estimated variances from the RRC and the OCS. Further details on these studies can be found in Germain and Julien (1993) and the 1991 Census Technical Report - Coverage (Statistics Canada 1994).

The domains of interest can be defined by partitioning the sample into $p = 1, 2, \ldots, P$ provinces/territories and $a = 1, 2, \ldots, A$ age - sex groups, hence a total of $A \times P$ domains require estimates. Let C_i be the number of persons in the i-th province - age domain enumerated in the Census and T_i be the true population of the same domain. The net number of persons missed in the i-th cell is $M_i = T_i - C_i$. The adjustment factor, Θ_i, is the ratio of the true population in a domain over the Census count, while the undercoverage rate, U_i, the unit that is usually reported in the releases from the coverage studies, is the ratio of missed persons over the true population.

The true adjustment factors, Θ_i, which are the variables that we wish to estimate, can be written as:

$$\Theta_i = \frac{T_i}{C_i} = \frac{M_i + C_i}{C_i}.$$

Undercoverage rates (U_i) which are usually reported in the releases from Statistics Canada, are related to the adjustment factors through the relationship

$$U_i = M_i(M_i + C_i)^{-1} = 1 - \Theta_i^{-1}.$$

In the modelling of the adjustment factors, the creation of ultimate domains is required. These domains are those at which the actual direct survey estimates of the adjustment factors will be produced. There must be an estimate for each province (10) and territory (2), so immediately P is fixed at 12. The age groups were fixed at 4 to create national estimates that have acceptably low standard errors. These age groups are defined for male and female as follows: 0 to 19 years of age; 20 to 29 years of age; 30 to 44 years of age; and 45 years and older. In total there are $12 \times 8 = 96$ direct survey estimates of adjustment factors that have to be fitted into the Empirical Bayes model. Each domain requires, besides the direct estimate of the adjustment factor, an associated estimate of the sampling variance.

2.2 Model and Assumptions

The basic model for the undercount is composed of two distinct parts. The first part describes how the direct survey estimates are related to the true underlying adjustment factors, while the second part models the relationship between the true adjustment factors and a set of explanatory variables. Since the parameters in the regression model are estimated by first estimating the parameters of an assumed underlying prior distribution and then assuming that these estimated parameters are known for any further calculation, this model is known as an Empirical Bayes model (Maritz and Lwin 1989).

The first part of the model, the sampling model, relates the observed adjustment factors to the true adjustment factors. This relationship is assumed true within each domain, and can be expressed as:

the observed adjustment factor =
the true adjustment factor + a random error.

The sampling model is written as follows:

$$F_i = \Theta_i + \epsilon_i : \epsilon_i \sim \text{Normal } (0,\sigma_i^2),$$
$$i = 1, 2, \ldots, n = A \times P,$$

where Θ_i is the true adjustment factor and ϵ_i is a random error component with a variance of σ_i^2. The assumptions underlying this model are:

(a) the sampling errors, ϵ_i, have mean zero;

(b) the sampling variances, σ_i^2, are known in each of the n domains;

(c) since the sample was selected independently within each domain, the covariance between the sampling errors ϵ_i in domain i and ϵ_j in domain j is zero; and

(d) the random errors ϵ_i are normally distributed in each domain.

Further discussion on the assumption of the known sampling variance in each domain is given below.

The second part of the model, the regression model, relates the true adjustment factors to a set of underlying explanatory variables. This model states that:

the true adjustment factor = a linear combination
of explanatory variables + a random error.

The regression model can be written as:

$$\Theta_i = X_i \beta + \delta_i : \delta_i \sim \text{Normal } (0, \tau^2),$$
$$i = 1, 2, \ldots, n = A \times P,$$

where X_i is the i-th row in X, a known $(n \times p)$ matrix of explanatory variables, β is a $(p \times 1)$ vector of unknown regression parameters and δ_i is (a different) random error with a model variance of τ^2. Underlying the system model are the following assumptions:

(a) the model errors, δ_i, have mean zero;

(b) the model variance, τ^2, is constant over all n domains;

(c) the model errors, δ_i, are normally distributed;

(d) the model errors, δ_i, are independent of sampling errors, ϵ_i;

(e) the covariance between different domains is zero (*i.e.*, $\text{Cov}(\delta_i, \delta_j) = 0$).

The problem is to use both the sampling model and the regression model to estimate Θ_i, the true adjustment factors. The conditional expectation for Θ_i given β, σ_i^2, τ^2, F_i can be determined for the joint model. Using standard arguments (Rao 1973), it can be shown that the conditional expectation of Θ_i is:

$$E(\Theta_i \mid \beta,\sigma_i^2,\tau^2,F_i) = (1 - \omega_i)X_i \beta + \omega_i F_i, \qquad (1)$$

where $\omega_i = \tau^2(\tau^2 + \sigma_i^2)^{-1}$.

Equation (1) is the basis for all the estimates that follow, although a few modifications need to be made before applying it to the data. Note that it is basically a weighted average of the direct survey estimate and the regression model estimate of the adjustment factor. Each estimate is weighted according to the precision with which it was estimated. If the sampling error, σ_i^2, is small compared to the model error, τ^2, implying that the direct survey estimate is relatively precise, then the final smoothed estimate will be mainly composed of the direct survey estimate. However, if the direct survey estimate has a large sampling variance relative to the model variance then the final smoothed estimate will be mainly constituted from the best linear unbiased predictor. The amount each estimate contributes to the final smoothed estimate is controlled by the weighting coefficient, ω_i.

Some limitations apply to interpretations that can be made about this model. First, it must be emphasized that this model is purely descriptive; it cannot be considered to be a causal model. Since the primary goal of this model is descriptive, the inferences on the regression parameters, β, while interesting are not of primary importance. Hence, the final regression model when it contains a term, say, on British Columbia renters and not Manitoba renters, is only saying that British Columbia renters explain a significant portion of the variation in adjustment factors in British Columbia while Manitoba renters does not explain a significant portion of the variation in adjustment factors in Manitoba.

As mentioned above, the sampling variances associated with the direct survey estimates of the adjustment factors are considered known in the Empirical Bayes model. However, experience has shown that the directly estimated variances are, in fact, somewhat unstable. In order to create some stability with the estimation of these variances it is proposed to model them. If we consider the design of the two sample surveys, then, under relatively mild assumptions, Dick (1993) has shown that within each domain the variance of the estimate of missed persons is proportional to a power of the Census count. If we add in appropriate normalizing parameters, then this relationship can be written as:

$$\sigma_i^2 C_i^2 = V(M_i) = KC_i^{\gamma},$$

or, as in the form of a regression equation,

$$\mathrm{Log}(V(M_i)) = \alpha + \gamma \log(C_i) + \eta_i \quad \text{with}$$

$$\eta_i \sim N(0, \zeta^2).$$

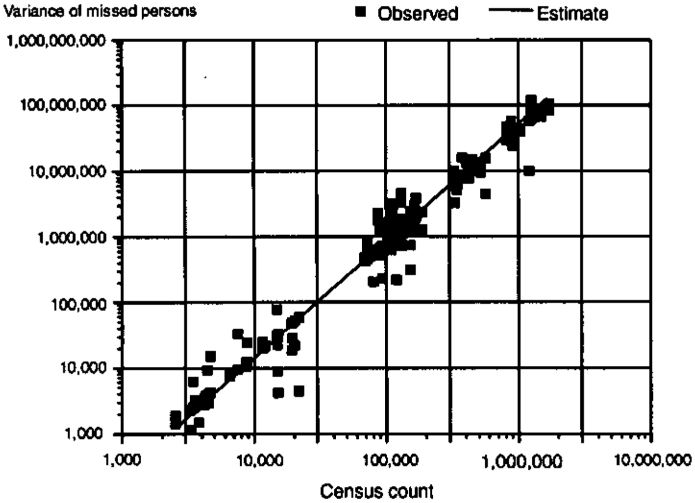

Figure 1. Observed variance *vs.* census.

This model for the sampling variance assumes that the product of the design effect and the undercoverage rate is constant within each domain. As discussed in Dick (1993), this assumption appears to be reasonable. Figure 1 shows the plot of the observed variance of missed persons calculated from the two coverage studies versus the Census count for the 96 domains. The least squares regression line was estimated as

$$\log(\hat{v}(M_i)) = -6.133 + 1.715 \log C_i$$

and is also plotted in Figure 1. A residual analysis (Dick 1993) did not detect any apparent violations of the underlying model assumptions. Since, in addition, the coefficient of determination, the R^2, is 0.943, this model was adopted for producing the sampling variances. The estimated survey variances were calculated for the adjustment factors through

$$\hat{v}(F_i) = \hat{v}(M_i)/C_i^2.$$

It will be assumed that these predicted values for the sampling variances are the actual 'known variances' required for the Empirical Bayes model.

2.3 Parameter Estimation

So far the model has been described in purely Bayesian terms: only the parameter Θ_i is considered unknown. Taking the usual Empirical Bayes approach (Maritz and Lwin 1989), we will assume that all the parameters except β, the regression parameter, are known. The conditional expectation of Θ_i with the regression parameter estimated can be written as

$$\tilde{F}_i^{(\mathrm{eb})} = E(\Theta_i \mid \hat{\beta}, \sigma_i^2, \tau^2, F_i).$$

However, in practice, the model variance, τ^2, is also unknown and must be estimated. The conditional expectation of Θ_i will now change to

$$\hat{F}_i^{(\mathrm{eb})} = E(\Theta_i \mid \hat{\beta}, \sigma_i^2, \hat{\tau}^2, F_i),$$

where the sampling variance, σ_i^2, is still considered known.

To estimate the model variance and the regression coefficients in the Empirical Bayes model, the marginal distribution of the observed adjustment factors, $m(F_i) \sim N(x_i\beta, \tau^2 + \sigma_i^2)$, can be used. Three possible methods were examined for estimating the variance parameter, τ^2, : Method of Moments (MM) as in Fay and Herriot (1979), Maximum Likelihood (MLE) as in the PES in the United States (Hogan 1992) and Restricted Maximum Likelihood (REML).

It is well known that MLE estimation of variance components is biased downwards (Harville 1977). Underestimation of the model variance in the Empirical Bayes model would result in more reliance being placed upon the regression model instead of the direct survey estimate. This is a result we wished to avoid. In Dick (1993), it is shown that there is little difference between the estimates of the model variance from REML or MM. Since the REML has a well understood asymptotic theory, it was adopted for the estimation of the model variance in the Empirical Bayes model.

Harville gives a full account of REML estimation. The basic approach is to first estimate the regression parameter, and then to estimate the model variance from the resulting residuals instead of the actual data. If we let X^* be a matrix of $(n - p)$ linear contrasts such that $E[X^{*\prime}F] = 0$, then Harville shows that the resulting (log) likelihood function, L_{reml}, when maximized with respect to the unknown model variance will give the restricted maximum likelihood estimates.

In the context of the Empirical Bayes model, Harville's approach can be described as follows. First, an initial estimate, usually zero, of the model variance, $\hat{\tau}^2_{(0)}$, is made and then the regression parameter, β, is estimated through weighted least squares:

$$\hat{\beta}_{(1)} = (X' \hat{V}_0^{-1} X)^{-1} X' \hat{V}_0^{-1} F, \tag{2}$$

where $\hat{V}_0 = \text{diag}(\hat{\tau}^2_{(0)} + \sigma_i^2 : i = 1, \ldots, n)$. Using this estimate of $\hat{\beta}_{(1)}$, a new REML estimate of the model variance, $\hat{\tau}^2_{(1)}$, can be made through

$$\hat{\tau}^2_{\kappa+1} = \hat{\tau}^2_\kappa + \left(\frac{\partial L_{\text{reml}}}{\partial \tau^2}\right) [i(\tau^2)]^{-1}, \quad \kappa = 0, 1, \ldots, \tag{3}$$

where, if we set $\hat{P}_\kappa = \hat{V}_\kappa^{-1} - \hat{V}_\kappa^{-1} X (X' \hat{V}_\kappa^{-1} X)^{-1} X \hat{V}_\kappa^{-1}$, we have

$$\frac{\partial L_{\text{reml}}}{\partial \tau^2} = -\frac{1}{2} \text{ trace } \hat{P}_\kappa + \frac{1}{2} (F - X\hat{\beta})' \hat{V}_\kappa^{-1} \hat{V}_\kappa^{-1} (F - X\hat{\beta})$$

and

$$i(\tau^2) = -E\left[\frac{\partial^2 L_{\text{reml}}}{\partial(\partial \tau^2)^2}\right] = \frac{1}{2} \text{ trace } (\hat{P}_\kappa' \hat{P}_\kappa).$$

Note, upon convergence of τ^2 and β, $i(\tau^2)^{-1}$ will be the asymptotic variance of $\hat{\tau}^2$.

By iterating between (2) and (3), new estimates of τ^2 will be used to update the estimate of β, which in turn will be used to update the estimate of τ^2. The iterations then continue until a suitable convergence has been reached: in this case $((\hat{\tau}^2_{\kappa+1}/\hat{\tau}^2_\kappa) - 1) < 10^{-6}$ was used.

Once the estimates for $\hat{\beta}$, the regression parameters, and $\hat{\tau}^2$, the model variance, have been determined, then the final smoothed estimates can be found. Maritz and Lwin (1989) show that the Empirical Bayes, or smoothed, estimate can be written as

$$\hat{F}_i^{eb} = (1 - \hat{\omega}_i) X_i \hat{\beta} + \hat{\omega}_i F_i,$$

where $\hat{\omega}_i = \hat{\tau}^2 (\hat{\tau}^2 + \sigma_i^2)^{-1}$. This is a combination of the original estimate and the regression estimate weighted by their respective variances.

The objective of the smoothing model is to create a series of estimates with smaller MSE than the original estimates. Prasad and Rao (1990), through asymptotic arguments, have suggested using the following estimator for the mean square error:

$$\text{MSE}[\hat{F}_i^{eb}] = \text{MSE}[\tilde{F}_i^{eb}] + \left[\left(\frac{\partial \omega_i}{\partial \tau^2}\right)^2 \omega_i E(\hat{\tau}^2 - \tau^2)^2\right].$$

The mean square error for the Empirical Bayes estimate, using restricted maximum likelihood estimation, has been conjectured by Cressie (1992) to be:

$$\widehat{\text{MSE}}[\hat{F}_i^{eb}] = \widehat{\text{MSE}}(\tilde{F}_i^{eb}) + 2\hat{g}_{3i}(\hat{\tau}^2) = \hat{g}_{1i}(\hat{\tau}^2) + \hat{g}_{2i}(\hat{\tau}^2) + 2\hat{g}_{3i}(\hat{\tau}^2),$$

where

$$\hat{g}_{1i}(\hat{\tau}^2) = \hat{\tau}^2(1 - \hat{\omega}_i)$$

$$\hat{g}_{2i}(\hat{\tau}^2) = (1 - \hat{\omega}_i)^2 X_i' (X' \hat{V}^{-1} X)^{-1} X_i$$

and

$$\hat{g}_{3i}(\hat{\tau}^2) = (1 - \hat{\omega}_i)^2 (\hat{\tau}^2 + \sigma_i^2)^{-1} [i(\tau^2)]^{-1}.$$

The assumed normality of ϵ_i and δ_i is an important assumption in the derivation. Note the value for the sampling variance, σ_i^2, is assumed known.

Prasad and Rao give the following interpretation to each of the three components: $\hat{g}_{1i}(\hat{\tau}^2)$ is the Bayes estimate of the variance, $\hat{g}_{2i}(\hat{\tau}^2)$ is the contribution from estimating the regression parameters and $\hat{g}_{3i}(\hat{\tau}^2)$ is the contribution from estimating the model variance τ^2. An estimate of the component due to the estimation of the sampling variance is *not* available: the additional variance this would add is not known but its absence clearly implies that the MSE is underestimated.

3. EMPIRICAL BAYES LINEAR MODEL

3.1 Explanatory Variables

The Empirical Bayes model described above was fitted to the 96 observed adjustment factors, with the sampling variances estimated using the method described in Section 2. The linear model that was fitted to this data included the following explanatory variables:

(a) An indicator variable for each province/territory.

(b) An indicator variable for each sex.

(c) An indicator variable for each age group.

(d) A variable indicating the percentage of people in the domain that are renters.

(e) A variable indicating the percentage of people in the domain that do not speak either official language.

(f) Various interaction variables including province by renters, province by non-official language, age and sex by renters.

In total, 42 variables were used in the initial regression.

These variables were selected for the initial regression model based, in part, on the experiences of previous RRC studies (Burgess 1988), partly on the results of the 1991 coverage studies (Germain and Julien 1993) and partly on the experiences of the PES in the United States as described in Hogan (1992) and Datta *et al.* (1992). The actual rationale for the variables to be included are as follows:

(a) The province indicator was included as an indication of the difficulty of Census collection within each province. Prior to the 1991 Census, it was assumed that collection would be more difficult in British Columbia and Ontario, and the anecdotal field evidence during collection seemed to support this conjecture.

(b) The age and sex variable were included because of the known differences in undercoverage rates between males and females. The undercoverage, in previous studies, has also shown a marked increase for individuals in their 20's.

(c) Tenure, in effect the percent of renters in each domain, was included because of the experiences in the United States PES, results of previous RRC studies and as a suggestion from the Statistics Canada Statistical Methods Advisory Committee.

(d) The use of non-official language was an attempt to locate the immigrant and minority groups that in the past have tended to have higher undercoverage rates.

(e) The interaction terms were included to further refine the predictive power of the model.

The mean encompasses all those variables that are not included in the model. Note that since indicator variables are used for province, sex and age-sex, one variable has to be excluded in order to avoid a singular design matrix. In effect, the missing variable, say the province indicator for Newfoundland, is included in the mean.

An operational constraint was also placed on the model. The SAS IML program written to estimate the parameters was limited to 4,095 numeric elements in the design matrix, hence with 96 domains, or observations, the model was limited to a maximum of 42 variables.

3.2 Model Building Process

After starting with the full regression model and 42 explanatory variables, a procedure was needed to remove those variables that were not statistically significant. The procedure chosen was to eliminate the least significant variable after each completed estimation cycle. This implies that for the 42 variable model, the variable Female Renters aged 0 to 19 would be eliminated since it has a t-value of 0.05. The regression model was then re-run with the remaining 41 variables. The least significant variable was then eliminated from that model. This procedure is equivalent to the Backward Stepwise Regression described in Draper and Smith (1966, page 167).

The Backward Stepwise Regression method was used to eliminate all variables until all remaining variables had a t-values greater than 2 (in absolute value). However when the final model was examined, it was noticed that a multicollinearity problem existed between the indicator variables for certain provinces and the interaction terms for renters within the same provinces. The implication of this problem is that there are some explanatory variables which are highly correlated with each other. This in turn implies that not all parameters in the model can be estimated precisely. As a rule of thumb Judge *et al.* (1984, page 459) suggest that this can be a problem when the simple correlation between variables is greater than R^2, the coefficient of determination. The final model had a $R^2 = 0.85$ and the simple correlation between the variables in question were all greater than 0.90 (in absolute value, since the correlations were negative).

A solution to this problem was to delete the variables with the lower t-values which turned out to be the provincial indicators. The final model is shown in Table 1 with the estimated coefficients and their t-values. The effect of removing the provincial indicators was to lower the final R^2 from 0.85 to 0.844, thus little predictive power has been lost.

Table 1

Final Estimates of Variables Used in Regression

Category	Variable	Final Estimate ($\hat{\beta}$)	T-Value (absolute value) ($H_0 : \beta = 0$)
Mean	Mean	1.0075	575.72
Age - Sex Combination	Male 20 to 29	0.0563	15.34
	Male 30 to 44	0.0208	5.81
	Female 20 to 29	0.0240	6.49
Sex by Age by Non-Official Language	Female Language 0 to 19	0.0797	2.75
Tenure by Province	British Columbia Renters	0.0449	3.96
	Ontario Renters	0.0804	7.35
	Quebec Renters	0.0255	2.66
	New Brunswick Renters	0.1064	5.61
	Yukon Renters	0.0639	3.80
	Northwest Territories Renters	0.0682	6.22

The final regression model then had various diagnostic tests performed on it. Since the regression is a weighted least squares with a random error term, Lange and Ryan (1989) have suggested using the following form to create standardized residuals:

$$z_i = \frac{\hat{F}_i^{(\mathrm{eb})} - X_i\,\hat{\beta}}{\sqrt{\sigma_i^2 + \hat{\tau}^2}}.$$

The residuals were analyzed using both Q-Q plots and outlier detections methods: no major departures from the assumed distribution of the residuals were detected. More details on the residual analysis can be found in Dick (1993).

Table 2
Direct, Smoothed and Raked Estimates of Adjustment Factors

Sex	Age	Estimate	B.C.	Alta	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld	Yukon	N.W.T.
Male	0-19	Direct	1.017	1.026	1.012	1.029	1.028	1.017	1.022	1.019	1.004	0.999	1.031	1.036
		Smooth	1.019	1.013	1.009	1.013	1.029	1.016	1.027	1.010	1.007	1.006	1.026	1.027
		Raked	1.020	1.016	1.011	1.015	1.031	1.018	1.027	1.013	1.005	1.007	1.029	1.031
	20-29	Direct	1.087	1.036	1.068	1.058	1.113	1.071	1.122	1.063	1.060	1.057	1.098	1.127
		Smooth	1.086	1.056	1.065	1.062	1.104	1.074	1.103	1.064	1.063	1.062	1.094	1.122
		Raked	1.083	1.061	1.073	1.067	1.101	1.079	1.096	1.073	1.041	1.074	1.096	1.127
	30-44	Direct	1.031	1.021	1.028	1.034	1.054	1.047	1.043	1.018	1.025	1.026	1.069	1.080
		Smooth	1.039	1.026	1.028	1.030	1.053	1.041	1.046	1.026	1.028	1.028	1.052	1.059
		Raked	1.038	1.028	1.032	1.032	1.051	1.043	1.043	1.029	1.018	1.033	1.053	1.059
	45 +	Direct	1.019	1.018	1.002	1.014	1.013	1.011	1.014	1.016	1.018	1.016	0.992	1.076
		Smooth	1.017	1.011	1.006	1.009	1.019	1.013	1.019	1.010	1.009	1.009	1.021	1.039
		Raked	1.014	1.010	1.006	1.009	1.016	1.012	1.015	1.010	1.005	1.010	1.019	1.035
Female	0-19	Direct	1.034	1.018	1.017	1.012	1.037	1.029	1.029	1.014	0.995	1.016	1.026	1.054
		Smooth	1.030	1.015	1.013	1.015	1.038	1.023	1.030	1.010	1.006	1.010	1.028	1.061
		Raked	1.032	1.018	1.016	1.017	1.040	1.026	1.030	1.012	1.004	1.013	1.030	1.068
	20-29	Direct	1.068	1.047	1.028	1.020	1.072	1.043	1.070	1.030	1.004	1.041	1.068	1.072
		Smooth	1.058	1.036	1.031	1.029	1.070	1.044	1.071	1.031	1.027	1.033	1.069	1.092
		Raked	1.058	1.041	1.036	1.032	1.070	1.048	1.068	1.037	1.018	1.041	1.072	1.099
	30-44	Direct	1.013	1.009	1.004	1.006	1.027	1.017	1.031	1.019	1.004	1.024	1.031	1.020
		Smooth	1.018	1.008	1.007	1.007	1.030	1.017	1.029	1.010	1.007	1.011	1.028	1.026
		Raked	1.017	1.008	1.007	1.007	1.028	1.017	1.025	1.011	1.004	1.012	1.027	1.026
	45 +	Direct	1.007	1.003	1.018	1.001	1.011	1.011	1.000	1.002	0.993	1.013	1.024	1.007
		Smooth	1.014	1.006	1.010	1.006	1.021	1.015	1.020	1.006	1.005	1.009	1.031	1.026
		Raked	1.008	1.004	1.007	1.004	1.012	1.009	1.011	1.004	1.002	1.006	1.019	1.016

3.3 Estimates of Adjustment Factors

Table 2 shows both the direct survey estimate and the smoothed Empirical Bayes estimate of the adjustment factors. An inspection of the table shows that these estimates are relatively close, reflecting the Empirical Bayes methodology of combining the direct survey estimate with the model estimate. Note that all of the domains that were originally estimated to have overcoverage - shown by an estimated adjustment factors being less than one - have been changed, by the Empirical Bayes estimates to being an estimate of undercoverage. The difference between the two sets of estimated adjustment factors - in absolute terms - differ by under 1% and in the larger provinces by less than 0.5%. However, for some of the smaller provinces and territories the difference between the two estimates can be substantially larger. In the Northwest Territories the change between the directly estimated adjustment factor and the Empirical Bayes estimate is about 2% for 3 age - sex groups and over 3% for another.

The objective of the Empirical Bayes model is to produce estimates with smaller MSE than the survey estimates. From Section 2.2 it can be shown that the variance for the direct survey estimates is calculated from

$$\log \hat{v}(F_i) = -6.133 - 0.285 \log C_i,$$

while the Prasad-Rao MSE, from Section 2.3, is calculated by

$$\widehat{\mathrm{MSE}}[\hat{F}_i^{\mathrm{eb}}] = \widehat{\mathrm{MSE}}(\tilde{F}_i^{\mathrm{eb}}) + 2\hat{g}_{3i}(\hat{\tau}^2) =$$

$$\hat{g}_{1i}(\hat{\tau}^2) + \hat{g}_{2i}(\hat{\tau}^2) + 2\hat{g}_{3i}(\hat{\tau}^2).$$

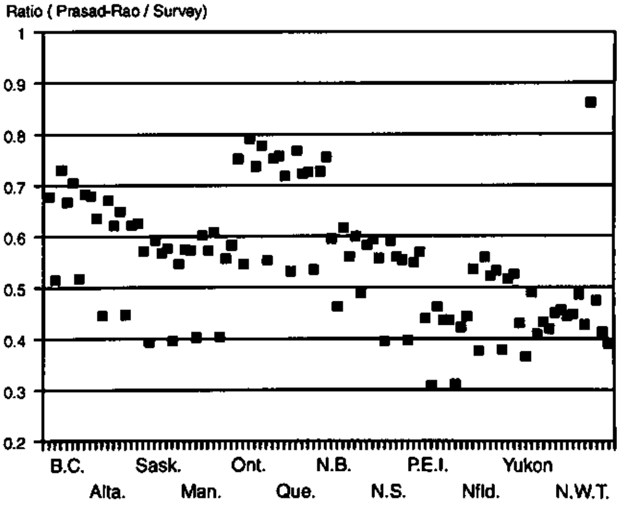

Figure 2. Ratio of root MSE, Prasad-Rao and survey.

Figure 2 plots, for each domain, $R = \sqrt{\widehat{\text{MSE}}[\hat{F}_i^{\text{eb}}]/\hat{v}(F_i)}$ the ratio of the root mean square errors for the Empirical Bayes model and the estimated survey variance (Note that within provinces the domains are ordered as Male aged 0-19, 20-29, 30-44 and 45 and over and Female aged 0-19, 20-29, 30-44 and 45 and over). Clearly, the Empirical Bayes MSE is smaller in all domains. However, in the larger provinces, Ontario and Quebec, the ratio of the root MSEs is only between 0.7 and 0.8. This relatively small gain is a reflection of the large sample sizes in these domains which in turn give a reliable estimate of the variance. The large gains are made in the smaller provinces and territories. For instance, in Prince Edward Island, the ratio of the root MSEs are all smaller than 0.5 showing the large improvement in the estimates. The one outlier is in the Northwest Territories (females aged 0 to 19): the Prasad-Rao MSE appears to have been overestimated in this domain.

4. ADJUSTMENTS MADE TO EMPIRICAL BAYES ESTIMATES

4.1 Rationale and Methodology

The advantage of the Empirical Bayes method is apparent from the above discussion. However, the Empirical Bayes methodology does not preserve the higher level (*i.e.*, the large domain) direct survey estimates that are reliable. By this it is meant that the provincial totals and the age - sex domain totals for the direct survey estimates and the Empirical Bayes estimates are not equal. Since the two surveys were designed to produce estimates at these levels, it is crucial that the Empirical Bayes be consistent with these reliable marginal totals.

To achieve consistency of estimates of missed persons between the reliable provincial and age - sex totals from the direct survey estimates and the final Empirical Bayes estimates, a raking ratio procedure was used. This is basically the method used in Australia to determine their small domain estimates (see Choi *et al.* 1988). This technique re-scales the individual Empirical Bayes estimates to conform to the known provincial and national age - sex totals. Once this procedure has converged, the final estimates will be consistent with the direct survey totals. In terms of a log-linear model, we are using as the main effects (province and age-sex) estimates the results from the two coverage studies and the interaction terms (province by age-sex) estimates from the Empirical Bayes modelling.

Details of the procedure can be described as follows. Assume that we have a matrix of estimated missed persons that has P columns (corresponding to the provinces) and A rows (corresponding to the age-sex groups). First set $F_{\text{pa}} = F_i$, then let $\hat{M}^s_{\text{pa}} = C_{\text{pa}}(F_{\text{pa}} - 1)$ be the direct survey estimate of the number of missed persons in province p and age - sex group a and let $\hat{M}^{(\text{eb})}_{\text{pa}} = \hat{M}^{(0)}_{\text{pa}} = C_{\text{pa}}(\hat{F}^{(\text{eb})}_{\text{pa}} - 1)$ be the Empirical Bayes estimate of missed persons from the Empirical Bayes model. If we let a plus sign (+) represent addition across the variable then the raking estimate can be written for cycles $\kappa = 0, 1, \ldots$ as;

$$\hat{M}^{(2\kappa+1)}_{\text{pa}} = \hat{M}^{(2\kappa)}_{\text{pa}} \left(\sum_{a=1}^{A} \hat{M}^s_{\text{pa}} \Big/ \sum_{a=1}^{A} \hat{M}^{(2\kappa)}_{\text{pa}} \right)$$

and

$$\hat{M}^{(2\kappa+2)}_{\text{pa}} = \hat{M}^{(2\kappa+1)}_{\text{pa}} \left(\sum_{p=1}^{P} \hat{M}^s_{\text{pa}} \Big/ \sum_{p=1}^{P} \hat{M}^{(2\kappa+1)}_{\text{pa}} \right).$$

This procedure will converge to a unique solution. Since this is basically a log-linear model, the underlying assumption is that the relationship determined by the Empirical Bayes model for the interaction between province and age - sex group is valid and will be preserved.

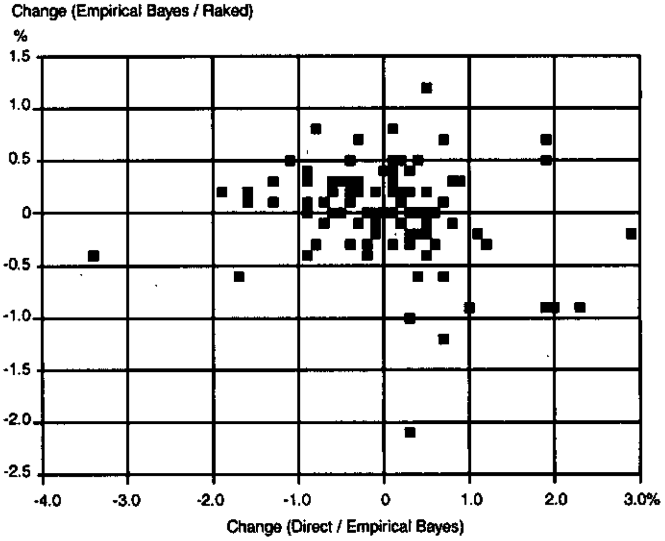

Figure 3. Percent change in estimates of adjustment factors.

Table 2 shows the final raked estimates of the adjustment factors along with both the original survey estimates and the Empirical Bayes estimates. Generally, the impact of raking is to shrink the Empirical Bayes estimate back towards the survey estimate. This is shown in Figure 3. Here two different percent changes in the estimated adjustment factors are plotted. The X-axis shows the percent change between the direct survey estimate and the Empirical Bayes estimate. The Y-axis shows the percent change between the Empirical Bayes estimate and final raked estimate. The plot shows that the two variables are negatively correlated: hence the raking tends to move the Empirical Bayes estimates closer to the original survey estimates.

One draw back of this procedure is that the MSEs of the raked adjustment factors are now very difficult to estimate. Due to the non-linear nature of the raking ratio procedure, a direct calculation is impossible. It is possible to use a Taylor series expansion; however this assumes a large sample size in each domain when in fact we know some domains have very small sample sizes. A possible procedure is to adjust the estimated MSE from the Empirical Bayes estimates and multiply these by the squared ratio of the raked Empirical Bayes estimate over the Empirical Bayes estimate. While this procedure is only a crude approximation, it can at least give some guidance as to the reliability of the individual estimates. This method will ensure that the coefficient of variations calculated for the Empirical Bayes estimates will be retained for the corresponding raked Empirical Bayes estimates. This is the procedure that was used to produce the final MSE estimates for the raked Empirical Bayes estimates of missed persons.

4.2 Detailed Domain Estimates

The Population Estimates Program requires even finer detail than that produced by the various models discussed above. In fact the program needs estimates for single years of age for each sex for each Census Division within each province. Since the Empirical Bayes methodology is limited somewhat by the direct survey results - an estimate with a non-zero standard error is required for each domain - synthetic methods must be used to generate the more detailed estimates.

For the Population Estimates Program, estimates for each province and sex were produced for 9 age groups instead of the 4 age groups used in the Empirical Bayes model. A straight synthetic model, using the raked Empirical Bayes estimates as initial values, was proposed for this stage of estimation. To produce these more detailed estimates, the raked Empirical Bayes estimate was allocated proportionally by Census count across all sub-age groups within each province and sex. Let the final raked estimate in the p-province and the a-th age-sex group be $\hat{M}_{\mathrm{pa}}^{2\kappa+2} = \hat{M}_{\mathrm{pa}}^{rf}$. Also if the a-th age - sex group is composed of Q exclusive sub-age groups then the estimate of the missed persons in the p-th province and the q-th sub-age group within the a-th age - sex group would be

$$\hat{M}_{\mathrm{pa}_q} = \hat{M}_{\mathrm{pa}}^{rf} \left(\frac{C_{\mathrm{pa}_q}}{C_{\mathrm{pa}_+}} \right),$$

where $C_{\mathrm{pa}_+} = C_{\mathrm{pa}} = \sum_{q=1}^{Q} C_{\mathrm{pa}_q}$. This approach guarantees that the estimates from the earlier raked Empirical Bayes output are preserved for the original domain total. The further estimates that are required for the population estimates program use demographic methods. In fact, one of the objectives of the Empirical Bayes procedure is to provide initial estimates for the demographic methods. See Michalowski (1993) for further details.

5. SUMMARY AND CONCLUSIONS

The Empirical Bayes methodology was adopted because it preserves the more reliable estimates from the larger provinces and domains while permitting a model based estimate to dominate if the underlying direct estimate is unreliable. This is in accordance with standard survey methods of using the direct survey estimates as much as possible. The raking ratio procedure used for adjusting the estimates from the Empirical Bayes model was used to ensure consistency with the direct survey results that were known to be reliable.

As for the explicit model used to describe the underlying true adjustment factors, it must be noted that this model is purely descriptive. Its primary function is to use explanatory variables to describe the variation in adjustment factors, taking into account the sampling error associated with each adjustment factor. It would not be prudent to make far-reaching conclusions on the nature of undercoverage from the final set of parameters included in the model.

The main weakness of this approach is with the two variances that are estimated. The assumption of the regression model errors being approximately normally distributed is difficult to assess. In the absence of any real knowledge about the true underlying distributions any assumption about the model variance will be essentially unverifiable. The proposed model variance seems reasonable and diagnostic checks have not revealed any major problems.

The sampling variance model is more problematic. All Empirical Bayes methods assume that this variance is known, when in fact it has to be estimated. Efforts to extend the Prasad-Rao MSE calculation to include the contribution from this estimated parameter have not yielded any new results.

In the future, research will concentrate in working around the problem associated with estimating the sampling variances. Further work needs to be conducted on the Prasad-Rao MSE calculation. In addition, the possibility of using the micro level data from the coverage studies and estimating the undercoverage rates directly through logistic regressions as in Wong and Mason (1985) will be pursued.

Another project would be to examine the implications of recasting the Empirical Bayes model into the standard state space framework (Robinson 1991). Pfeffermann and Burck (1990) have suggested a method for calculating the MSE for a time series placed in a state space model that has to conform to certain periodic benchmarks. The state space formulation would also be useful in explicitly incorporating the demographic methods.

ACKNOWLEDGEMENTS

The author is grateful to D. Binder, M. Hidiroglou, R. Carter, M. Armstrong, J. Tourigny, J.N.K. Rao and especially D. Royce for commenting on an earlier version of this paper. An Associate Editor and two referees also provided comments that improved the final version.

REFERENCES

BURGESS, R.D. (1988). Evaluation of Reverse Record Check estimates of undercoverage in the Canadian Census of Population. *Survey Methodology*, 14, 137-156.

CHOI, C.Y., STEEL, D.G., and SKINNER, T.J. (1988). Adjusting the 1986 Australian Census count for under-enumeration. *Survey Methodology*, 14 173-189.

CRESSIE, N. (1992). REML estimation in Empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.

DATTA, G.S., GHOSH, M., HUANG, E.T., ISAKI, C.T., SCHULTZ, L.K., and TSAY, J.H. (1992). Hierarchical and Empirical Bayes methods for adjustment of census undercount: The 1988 Missouri Dress Rehearsal data. *Survey Methodology*, 18, 95-108.

DICK, J.P. (1993). Procedures used in modelling net undercoverage in the 1991 Census. Internal Statistics Canada memorandum.

DRAPER, N.R., and SMITH, H. (1966). *Applied Regression Analysis*. New York: John Wiley and Sons.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year (with discussion). *Journal of the American Statistical Association*, 84, 927-943.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 82, 269-277.

FREEDMAN, D., and NAVIDI, W. (1992). Should we have adjusted the U.S. Census of 1980? (with discussion). *Survey Methodology*, 18, 3-74.

GERMAIN, M.-F., and JULIEN, C. (1993). Results of the 1991 Census coverage error measurement program. *Proceedings of Seventh Annual Research Conference.* United States Bureau of the Census, 55-70.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.

HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-337.

HOGAN, H. (1992). The 1990 Post-Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.

JUDGE, G.G., GRIFFITHS, W.E., CARTER HILL, R., and LEE, T-C. (1984). *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.

LANGE, N., and RYAN, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17, 624-642.

MARITZ, J.S., and LWIN, T. (1989). *Empirical Bayes Methods (2nd edition)*. London: Chapman and Hall.

MICHALOWSKI, M. (1993). Revised postcensal and intercensal estimates: Canada, provinces and territories, 1971 - 1991. Internal report, Population Estimates Section, Statistics Canada.

PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*. New York: John Wiley and Sons.

ROBINSON, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Sciences*, 6, 15-51.

ROYCE, D. (1992). A comparison of some estimators of a set of population totals. *Survey Methodology*, 18, 109-125.

STATISTICS CANADA (1993). *1991 Census Technical Report: Coverage*. Ottawa: Supply and Services Canada, 1994. 1991 Census of Canada: Catalogue No. 92-341E.

WONG, G.Y., and MASON, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.

Survey Methodology, June 1995
Vol. 21, No. 1, pp. 55–62
Statistics Canada

Between-State Heterogeneity of Undercount Rates and Surrogate Variables in the 1990 U.S. Census

JAY JONG-IK KIM, ALAN ZASLAVSKY and ROBERT BLODGETT[1]

ABSTRACT

As part of the decision on adjustment of the 1990 Decennial Census, the U.S. Census Bureau investigated possible heterogeneity of undercount rates between parts of different states falling in the same adjustment cell or poststratum. Five "surrogate variables" believed to be associated with undercount were analyzed using a large extract from the census and significant heterogeneity was found. Analysis of Post Enumeration Survey on undercount rates showed that more variance was explained by poststratification variables than by state, supporting the decision to use the poststratum as the adjustment cell. Significant interstate heterogeneity was found in 19 out of 99 poststratum groups (mainly in nonurban areas), but there was little if any evidence that the poststratified estimator was biased against particular states after aggregating across poststrata. Nonetheless, this issue should be addressed in future coverage evaluation studies.

KEY WORDS: Poststratification; Influence statistics; Linearization; Synthetic estimation.

1. INTRODUCTION

The Post Enumeration Survey (PES) of the 1990 Decennial Census of the United States was designed to produce coverage estimates for 1,392 poststrata. The nation was first divided into 116 domains, called poststratum groups (PSGs) according to geography, race/Spanish origin and tenure (owner *vs.* renter). With only 4 exceptions, all PSGs are defined within a census division, one of nine contiguous geographic areas each composed of several states. Each PSG was further divided into 12 age-by-sex groups, the poststrata. For example, roughly all Black renters in New York city constitute a PSG and all females, age 0-9, of this PSG make a poststratum (PS). Further details on the PES are in Hogan (1992,1993).

Small area undercount rates were calculated by synthetic estimation; the same adjustment factor was applied to persons from a given PS in all areas. This procedure is accurate under the "synthetic assumption" of homogeneity of undercount rate within a PS. The validity of the synthetic assumption has been hotly debated (Section 2). This paper reports on research conducted as a part of a PES evaluation project (the "P12 project") which investigated heterogeneity within poststrata. In particular, this research focused on the following question: can differences in coverage be identified between parts of a poststratum that fall into different states?

Under the homogeneity assumption, the rates are the same within a PS regardless of state. Thus, this assumption can be tested by comparing rates from state to state within a PS; this test focuses attention on the question of whether synthetic estimation is "unfair" to certain states. The unit of analysis is the intersection of a census block and a PS or PSG, called a block part (BP) for the analysis of the undercount rate data. A census block is a small area bounded by visible features such as streets, streams *etc.* and/or by political boundaries. In urban areas it roughly corresponds to a city block. In fact, most of our analyses are performed on PSGs, since the age-sex breakdown of the PSG did not vary much from state to state. Hence, the analysis focuses on whether BPs differ between states within PSG.

Two distinct analyses were performed. The distributions of five "surrogate variables" were investigated (Section 3), using a large (4.26%) extract from the census. The distribution of undercount was investigated using the much smaller PES data set (Section 4). For more detailed tables and documentation of the project, see Kim (1991).

2. LITERATURE REVIEW

Two key questions have been addressed in the literature on heterogeneity:

1. The empirical question: how much heterogeneity is there, and how can it be described?
2. The theoretical and policy question: what are the implications of heterogeneity for the accuracy of synthetic adjustments and the validity of assessments of these adjustments?

Heterogeneity may be identified and analyzed at many levels of aggregation. Perfect homogeneity of undercount rates for very small domains is numerically impossible,

[1] Jay Jong-Ik Kim, Statistical Research Division, U.S. Bureau of the Census, Suitland, MD 20233, U.S.A.; Alan Zaslavsky, Department of Statistics, Harvard University, Cambridge, MD 02138, U.S.A.; and Robert Blodgett, U.S. Food and Drug Administration, 200 C St., S.W., Washington, DC 20204, U.S.A.

because of discreteness of the true population and the census counts. Indeed, because census errors (omissions or erroneous enumerations) tend to be either independent of each other or positively associated (as when a household with several members is omitted, or when some local characteristic affects an entire block), we would anticipate at least binomial variability in observed undercount rates.

Hengartner and Speed (1993) analyzed 1990 PES data from two sites by fitting models in which the explanatory variables were block and "demoid" (a unit defined by the non-geographic poststratification variables, such as race, sex, age, and tenure). They found that the amount of variance explained by block was slightly greater than the amount explained by demoid; the number of blocks was not much greater than the number of demoids in their data set. In response, Schafer (1993) argued that an estimation scheme involving block effects would not be practical because it would require collecting data from every block.

Heterogeneity of undercount at any level may be defined as excess variability in observed undercount rates at that level over what would be expected as a consequence of variability at a lower level of aggregation. For example, confining our attention to a single poststratum, a set of blocks are heterogeneous if their undercount rates in that poststratum differ more than would be expected if households, including those counted, partially counted, and omitted in the census, had been randomly distributed across the blocks. Similarly, a group of states are heterogeneous (similarly controlling for poststratum) if they differ more than would be expected if blocks, including those with higher and lower undercounts, had been randomly distributed across the states. Several studies have attempted to measure heterogeneity in undercount rates and other census variables. Wachter and Freedman (1992) analyzed a large sample of census data (similar to that considered in Section 3). They estimated the excess variability between "superblocks" over that predicted by a binomial model with uniform rates, for four "artificial population" variables (multi-unit housing rate, nonmailback rate, allocations, and substitutions, described in Section 3). Compared to the greatest possible amount of heterogeneity (if each block were homogeneous), the "excess variability" ranged from around 20% (for multiunit housing) to 2% (for substitutions). Another study by Freedman and Wachter (1993) examined between-state heterogeneity using "artificial populations" based on the same variables and two others, and found substantial variability.

Alho, Mulry, Wurdeman and Kim (1993) used conditional logistic regression models to describe heterogeneity associated with measured covariates that were not captured in the poststratification. Their concern was primarily with reducing the bias of dual system estimates of population, rather than with more accurate small-area estimates.

A controversial topic in evaluation of the proposed adjustment of the 1990 census was the effect of heterogeneity on the accuracy of adjusted population counts obtained by synthetic estimation, and particularly on comparisons of the accuracy of adjusted and unadjusted counts. Wachter and Freedman (1992) argued that because the "synthetic assumption" of uniform coverage within poststrata is demonstrably false, aggregate measures of the accuracy of an adjusted census systematically underestimate error. Because nonuniformity of coverage affects the accuracy of an unadjusted census as well, however, the implications of this conclusion for the appropriateness of adjustment are not obvious.

In one of the earlier "surrogate variables" studies, Isaki, Schultz, Diffendal and Huang (1988) simulated the behavior of synthetic estimators on "artificial populations" which were transformations of the substitution (unit imputation) rate. They found that a synthetic estimator generally did better than "unadjusted" counts.

Schirm and Preston (1987) argued, using analytical calculations and simulation, that synthetic estimation makes estimates for small areas more accurate under plausible conditions, even if the synthetic assumption does not hold. Wolter and Causey (1991) investigated the performance of synthetic estimators and of a single ratio adjustment when the undercount rates are estimated with error, using undercount rates from the 1980 Post-Enumeration Program (PEP) and simulating various levels of sampling error; they estimated "break-even" coefficients of variation at which sampling error in the adjusted counts or proportions would make them less accurate than unadjusted counts or proportions. The conclusions of these articles were criticized by Freedman and Navidi (1992), who gave counterexamples of possible distributions of undercount for which adjustment by synthetic estimation would make population distribution less accurate.

Fay and Thompson (1993) simulated effects of heterogeneity on accuracy of synthetic estimates, using eight surrogate variables (including the five used in this study) and the same data set as analyzed in Section 3. They performed a loss function analysis as in Mulry and Spencer (1993) to compare the accuracy of simulated unadjusted counts to that of synthetically adjusted counts. They found that the effect of ignoring heterogeneity was to underestimate the gain in accuracy due to synthetic adjustment for five of eight variables, and to overestimate it for one variable (unemployment rate), while there was little difference for two other variables (poverty and migration rates).

3. ANALYSIS OF SURROGATE VARIABLES

In the analysis of census data, we selected variables which were available for the entire census and which, like undercount, were descriptive of or related to the census-taking process. The selected surrogates are the allocation rate, mail return rate, multiunit structure rate, mail universe rate (fraction of units receiving mail questionnaire) and substitution rate. The allocation rate is the fraction of households for which imputations were made for item nonresponse, and the substitution rate is the fraction of households which were imputed as a whole because it was determined that a unit was occupied but no interview could be obtained.

Table 1 shows correlations between each of these variables and undercount rate by PSG. These "ecological" correlations (Freedman, Pisani and Purvis 1978, pp. 141-142) of PSG averages differ from those which could be calculated from block-level data. The latter are smaller, possibly because of the noise introduced by random variability in the small populations in each block.

Table 1

Correlation Coefficients between the Surrogate Variable and Undercount Rate by PSG

Variable	Correlation
Allocation Rate	.44
Mail Return Rate	−.57
Multiunit Structure Rate	.39
Mail Universe Rate	.08
Substitution Rate	.47

Applying a naive test which treats the PSGs as independent, each correlation is significant except that for mail universe rate, but the magnitudes of the correlations are not large. To some extent, furthermore, these variables are descriptive of conditions which tend to lead to higher omission rates (allocations due to poor completion of questionnaires, substitutions due to difficulty in obtaining interviews) or to lower omission rates (high mail return rates). On the other hand, difficult census-taking conditions can also lead to erroneous enumerations, so these effects on net undercount are not entirely clear-cut. We do not analyze these variables simply because we believe that they are distributed in exactly the same way as undercount. Rather we hope that by obtaining results on the distributions of a range of different census variables, we may gain some insight into the distribution of undercount.

For the analyses of the surrogate variables, a stratified cluster sample of 1990 Census data was extracted. This sample is composed of 204,394 blocks corresponding to 125,000 block clusters. A block part containing less than ten persons was combined with successive block parts (in order by block number) until a minimum count of ten persons was obtained. This operation was performed to obtain relatively stable rates for the surrogate variables which allows us to analyze the rates themselves.

Surrogate variables are analyzed by logistic regression. Two forms of logistic regression model were used. For the within-PSG analysis, the model for PSG i is

$$log[P_{ij}/(1 - P_{ij})] = A + C_j$$

and for the within-division analysis,

$$log[P_{ij}/(1 - P_{ij})] = A + B_i + C_j,$$

where P_{ij} is the rate for a surrogate variable in the i-th PSG and j-th state, A is the intercept, B_i is the i-th PSG effect and C_j is the j-th state effect. The models used only the 99 PSGs astride two or more states. Models were built for surrogate variables in the 99 PSGs and in each of nine divisions. SAS PROC CATMOD estimated the parameters by maximum likelihood and provided Wald statistics for testing the significance of state effects.

The data were collected with a cluster sample rather than a simple random sample so the test statistics must be divided by a design effect. We estimate a design effect,

$$\hat{D}_{ij} = \frac{\sum_{k=1}^{K_{ij}} n_{ijk}(\hat{p}_{ijk} - \hat{p}_{ij})^2}{K_{ij}\hat{p}_{ij}(1 - \hat{p}_{ij})},$$

where $\hat{p}_{ijk}$ is the rate for the i-th PSG, j-th state and k-th combined BP; n_{ijk} is the size of the combined BP; K_{ij} is the sample number of combined BPs in the i-th PSG in the j-th state and $\hat{p}_{ij}$ is the estimated rate for the i-th PSG and j-th state. The fraction is the ratio of the observed between-block variance to that expected under binomial sampling.

The estimated design effect $\hat{D}_{ij}$ is a measure of within-state within-PSG heterogeneity. The more within-state heterogeneity there is, the greater the sampling variance of the state-level rate and the harder it is to detect a significant difference. The magnitude of the design effect thus affects the statistical power of the hypothesis tests.

The calculated design effect only approximates the required correction. First, $\hat{D}_{ij}$ sums over the combined BPs rather than individual BPs. Second, the sample is a stratified cluster sample, and most or all post-strata span several sampling strata. The formula is only strictly correct for an unstratified sample. Third, the correct effect involves off-diagonal (covariance) as well as on-diagonal (variance) terms, and the off-diagonal terms are omitted. To account for the above, the calculated design effects were multiplied by the judgmentally chosen factor, 1.25.

A design effect was calculated for each surrogate variable and PSG. It is small (around 2) in most PSGs for the allocation and substitution rate. The effect is slightly higher for mail return rate, but it tends to be large (as much as 20) for multiunit structure and mail universe rate, since these characteristics are usually fairly uniform within a block but vary greatly between blocks.

Table 2 summarizes the design-corrected tests for state effects within PSG.

Table 2

Number of PSGs with Significant ($\alpha = .05$) State Effect (Logistic Regression)

Div.	No. Grp	Alloc	Mail Ret	Mult Str	Mail Unv	Sub
1	5	5	5	5	1(1)	3(4)
2	12	11	11	12	7(10)	12
3	16	15	16	16	3(3)	12(12)
4	8	8	8	7	5(6)	5(8)
5	10	10	9	10	4(4)	7(8)
6	15	15	13	15	5(7)	15
7	9	8	9	9	4(4)	8(8)
8	7	7	7	7	2(3)	6(6)
9	17	15	14	14	5(5)	6(12)
Sum	99	94	92	95	36(43)	74(84)

The numbers in ( ) are the number of PSGs for which test statistics are available when they are less than the number of groups.

Nationally, for each surrogate variable the state effect is significant for at least 84% of the PSGs. (The total number of PSGs varies because when a PSG falls entirely within one state or when only one state has non-zero observations for a particular variable, the corresponding model cannot be fit). The results are summarized at the division level. (Divisions 1 through 9 are New England, Mid-Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain and Pacific Divisions.)

Table 3 shows the magnitude of state effects, expressed as χ^2 values of test statistics adjusted for design effect, for three variables having relatively high correlation with the undercount rate. In the table, the χ^2 values have from 1 to 8 degrees of freedom.

Table 3

Magnitude of State Effects with respect to Test Statistics

	Allocation Rate	Mail Return Rate	Substitution Rate
Minimum	4.3	0.28	5.46
25%-ile	27.5	102.83	49.80
50%-ile	68.9	254.49	97.35
75%-ile	140.3	644.05	260.88
Maximum	945.2	8,779.88	1,815.12

In division-level models with state and PSG effects, both the state and PSG effects were significant at the 1% level in every division and for every variable (excluding mail universe rate in two divisions where a test statistic could not be calculated).

4. ANALYSIS OF UNDERCOUNT RATE

The results described above for surrogate variables were obtained early in the census process, but they have limited relevance to homogeneity of undercount itself. After PES data were processed, direct analysis of the distribution of undercount became possible.

The data set for these analyses merged two data sets for the 12,124 PES sample blocks, one for the *E*-sample (Census follow-up) and the other for the *P*-sample (PES). There were 12,124 collection blocks, some of which were split up for tabulation, giving 12,964 tabulation blocks. More importantly, because some of the smaller blocks were combined in the sampling, there were 5,293 block clusters sampled. Correct enumerations and *E*-sample total counts are on the *E*-sample file. The *P*-sample file includes *P*-sample total counts and counts of matches (*P*-sample cases that were included in the Census).

4.1 Variance Explained by State and PSG

For each division, a two-way ANOVA was fitted to undercount rates for state parts. Table 4 shows the ratio of the sum of squares due to PSGs to that due to states within a division.

Table 4

Variance of Undercount Rate Explained by State and PSG

Div.	No. of Groups	No. of States*	SS (Group) / SS (State)	MS (Group) / MS (State)
1	5	6	4.51	5.64
2	12	3	4.88	.89
3	16	9	12.69	6.77
4	8	4	8.73	3.74
5	10	4	8.17	2.72
6	15	5	7.67	2.19
7	9	7	2.78	2.09
8	7	8	1.31	1.53
9	17	5	40.28	10.07

* States include D.C.

The ratio is always greater than one and in Division 9 it is 40.28, showing much larger effects for PSG than for state. The mean square for group also exceeds the mean square for state in each division except Division 2. This supports the decision to use the PS rather than the state as the cell for undercount estimation and adjustment.

4.2 Tests for State Effects on Undercount Rates

Assuming the substitution rate (fraction of units imputed for nonresponse) is negligible, the adjustment factor ($\hat{R}$) for a domain is

$$\hat{R} = \frac{WCE/WE}{WM/WP},$$

and the undercount rate is

$$1 - 1/\hat{R},$$

where *WE* and *WP* are the estimated population sizes weighted up from the *E* and *P*-sample, respectively. *WCE* is the weighted number of correct enumerations and *WM* is the weighted number of matches in the PES.

The statistic for the influence (see Appendix) of the *i*-th BP on the adjustment factor or undercount rate is

$$I_i = \hat{R}\left(\frac{WCE_i}{WCE} + \frac{WP_i}{WP} - \frac{WE_i}{WE} - \frac{WM_i}{WM}\right),$$

where WCE_i, WP_i, WE_i and WM_i are contributions from the *i*-th BP to the totals above.

A linear model was fitted to BP influence statistics to test for state effects. Under the null hypothesis, all the state parts in a PSG have the same undercount rate and the expected mean of the influence statistics for each state is 0 within each PSG. The influence statistics can be analyzed with one way ANOVA within a single PSG or two way ANOVA for all PSGs within a division.

Table 5 summarizes the tests for state effects on linearized statistics within each PSG.

Table 5

Analysis of Linearized Undercount at the PSG Level

Division	Number of PSG	Number of PSG with $P < .05$
1	5	0
2	12	3
3	16	4
4	8	5
5	10	2
6	15	1
7	9	0
8	7	1
9	17	3
Sum	99	19

The tests reveal significant heterogeneity between states in 19 out of 99 groups at the 5% significance level. The magnitude of the estimated state effect ranges from a few percent up to 20%, but the standard errors of these estimates are very large.

Table 6 summarizes the results of these analysis by place type. Place types 0, 1, 2 and 3 are large central cities in a Primary Metropolitan Statistical Area (PMSA), place types 4, 5 and 6 are non-central cities in PMSA with large central cities and place types 7, 8 and 9 are other areas.

The significant results are concentrated in PSGs for small areas (place types 7, 8 and 9). Ten out of 32 such groups show significant interstate heterogeneity at the 5% level. This suggests that the poststratification can be improved in those areas.

Table 6

Summary of Analysis of Linearized Undercount by Place Type

Place Type	Number of PSG	Number of PSG with $P < .05$
0	11	3
1	23	1
2	12	1
3	8	1
4	0	0
5	6	2
6	6	1
7	11	3
8	11	4
9	10	3

Table 7 shows the *F*-statistics and *p*-value for state effect for state × PSG models, once weighted by the size of domain and once without weights.

Table 7

State Effects by Division – Weighted and Unweighted Data

Division	D.F.	Unweighted Models		Weighted Models	
		F	*p*	*F*	*p*
1	5	.57	.72	.40	.85
2	2	4.64	.01	1.72	.18
3	8	.43	.91	.65	.74
4	3	.64	.59	.66	.58
5	3	.66	.58	1.37	.25
6	4	.60	.66	.24	.92
7	6	.39	.88	.22	.97
8	7	.62	.74	.76	.62
9	4	.77	.54	.48	.75

The additive effect of state was significant in only one division ($p = .01$) in the unweighted state × PSG model; when data were weighted by size of domain, the smallest *p*-value for the state effect was .18. In both cases, the most significant effect was observed in Division 2, in which New Jersey appeared to have higher undercount rate, controlling for PSG, than New York. Note that the most undercounted area in New York (New York City) had its own poststrata. In eight out of ten PSGs for which New Jersey and New York could be compared, including nonurban areas, the estimated undercount for New Jersey was larger than that for New York. Elsewhere, because the state effects in different PSGs varied in magnitude and sometimes in sign, and because only within a minority of PSGs in any division were there significant state effects, there was not significant evidence that in the aggregate the poststratification was biased against certain states.

Table 8 shows point estimates of the state effects in linear models for undercount rate by state part in each division, with effects for state and poststratum group. (Effects are centered at zero by division.) In effect, these are estimates of interstate differences after correcting for effects explained by the PSG composition of the different states.

Table 8

Estimated State Effects on Undercount within Division (as percent)

Division 1		**Division 4**		**Division 7**	
CT	−2.42	AL	−2.90	IA	−1.10
ME	.74	KY	1.89	KS	−0.50
MA	−0.48	MS	−0.02	MN	−0.01
NH	−0.14	TN	1.03	MO	−0.66
RI	1.43			NE	1.76
VT	0.90			ND	−0.07
				SD	0.60
Division 2		**Division 5**		**Division 8**	
NJ	4.18	AR	1.44	AZ	2.70
NY	−3.91	LA	−0.71	CO	0.68
PA	−0.26	OK	1.58	ID	−2.24
		TX	−2.30	MT	−1.61
				NV	−0.10
				NM	3.35
				UT	0.08
				WY	−2.84
Division 3		**Division 6**		**Division 9**	
DE	−0.42	IL	0.86	AK	−0.78
DC	2.82	IN	1.12	CA	1.02
FL	−0.88	MI	−0.73	HI	−0.18
GA	−1.43	OH	−0.88	OR	−0.26
MD	−1.32	WI	−0.38	WA	0.18
NC	0.53				
SC	0.70				
VA	−0.11				
WV	0.11				

The root mean square in the analysis of variance for state within division, averaged across all divisions, is 1.72 percent. Recall that only in the unweighted Division 2 analysis were the differences between states significant, it must be emphasized that the estimates in Table 8 do not represent well-measured interstate differences. The fact that the estimated effects are substantial in magnitude but are still not statistically significant tells us that the power of these tests to find interstate differences, given the sample sizes of the PES, is not as great as might be desired.

Another approach to the power problem is to consider the effect of reducing the size of the census extract used in analysis of surrogate variables by a factor of 25, the ratio of the census extract to the PES sample sizes. If we divide by 25 each of the chi-square test statistics summarized in Table 3, then in only 27 out of 99 PSGs would interstate differences have been significant for allocation rate (compared to 94 out of 99 PSGs with the full sample). Similarly, significant differences would have been found for 53 out of 99 PSGs for mail return rate (compared to 92 out of 99 PSGs with the full sample), and for 14 out of 84 for substitution rate (compared to 74 out of 84). Substitution rates are comparable in magnitude to undercount rates; after our hypothetical reduction of sample size, we obtain similar numbers of significant tests for substitution and undercount rates. It is plausible that with a much larger sample we would have found many more significant interstate differences, although one can only speculate on whether they would have been large enough to be of substantive concern.

5. DISCUSSION

This paper evaluates interstate heterogeneity in undercount rate and other census variables in the 1990 Census.

The evaluation used 1990 Census data and 1990 PES data. When this research was first embarked upon, the PES data were unavailable and were not expected to become available for analysis before the scheduled completion date. Surrogate variables from the 1990 Census were tested for significant heterogeneity among states within PSG. At the PSG level, state effect was significant ($\alpha = .05$) for 84%-95% of its PSGs for the various surrogate variables.

ANOVA on linearized undercount based on the PES data at the PSG level showed significant ($\alpha = .05$) state effects for 19 out of 99 PSGs. The significant results were concentrated in the PSGs in non-PMSA areas. Ten out of 32 such PSGs had significant state effects. This suggests that the poststratification in the relatively nonurban areas was not as successful as in the more urbanized areas.

How can we explain the different findings of the two studies? The two data sets had very different sample sizes, *i.e.*, the Census data had 125,000 block clusters but the PES data had 5,293 block clusters. It is therefore not surprising that small differences between states on surrogate variables would be statistically significant although corresponding differences would not be demonstrable with respect to undercount rates.

Furthermore, the correlations between the undercount rate and the surrogate variables are low as shown in Table 1. Therefore, any generalization from surrogate variables to undercount rates is somewhat conjectural. Given the modest correlation between undercount rates and surrogate variables, we prefer to give greater weight to the analysis of the PES data.

We conclude from these data that there are no demonstrable differences in average undercount rate between states within each division, after adjusting for PSG effects. While there is weak evidence for a difference between

New Jersey and New York within the Mid-Atlantic division, this result must be downweighted in the context of the number of divisions (nine) for which the test was performed. We conclude that if adjustment of population counts had been carried out based on the 1990 PES, no state would have been able to show that the poststratification was manifestly unfair in that it underadjusted that state relative to what direct state estimates showed that it deserved.

As the review in Section 2 shows, there is no consensus on whether or not between-state heterogeneity in undercount rates within PSG which is of substantial magnitude, although not large enough to be accurately measured by PES, would systematically affect the gain in accuracy obtained by synthetic adjustment. Nonetheless, the differences between states that were identified in analysis of the PES, together with the ancillary evidence of the surrogate variable analyses, make it appear likely that heterogeneity between states will again be an issue in coverage measurement for the year 2000 census, especially for the larger states for which these coverage differences can be most accurately measured. Fay and Thompson (1993) argue that a coverage measurement sample for 2000 should be designed to support direct (rather than synthetic) estimates of undercount for all states, although a CNSTAT panel (CNSTAT 1994) warns that for some states this could impose a highly inefficient sample allocation. Research over the intervening years must address the development of a combination of sample design and estimation methods that will produce defensible estimates of population by state.

ACKNOWLEDGEMENTS

The authors express their appreciation to the referees, an associate editor and editor for their comments which led to an improved version of this manuscript. This paper reports general results of research undertaken by the authors. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau or Harvard University. Zaslavsky's work was supported by Joint Statistical Agreements 90-23 and 91-31 and a contract between the Census Bureau and the National Opinion Research Center. The third author was with the Bureau of the Census when the research was in progress.

APPENDIX

Testing for Interstate Differences Using Linearized Statistics

A two-way ANOVA for adjustment factors in state parts yields an intuitively meaningful summary of the relative contributions of state and PSG effects to the variation in adjustment factors. Because the sampling unit of the PES is the block cluster rather than the state part, these models do not yield valid statistical tests of the significance of the state effects.

Consider a statistic whose sample estimate for a state or state part is a weighted mean of the sample estimates in each component block or BP. Significance of the state effects for this statistic within a PSG could be evaluated by one-way ANOVA with the included block parts as units (corresponding to PSUs), or aggregated across PSGs by two-way ANOVA for state and PSG effects.

The sample adjustment factor estimate $(WCE/WE)/(WM/WP)$ is a nonlinear function of sample counts. In small primary sampling units (PSUs) such as block parts this nonlinearity may be very noticeable, especially when the number of matches in a PSU is very small or zero so that the sample estimate of the adjustment factor is large or infinite. In this situation, if PSU sample estimates are treated as data, the additive assumptions of ANOVA are violated. Useful tests may be recovered, however, by using a linearized version of the statistic of interest.

Suppose that we are interested in a parameter $Z = f(X)$ where X is a vector of population proportions in certain cells. Let $\bar{x}$, x_i represent the corresponding sample proportions in the entire sample and in PSU i respectively, so $\bar{x} = \sum N_i x_i / \sum N_i$ is a size-weighted average of block cell proportions. Let $f_1(X)$ be the gradient of f at X. Then by Taylor linearization $f(\bar{x}) - f(X) \approx f_1(X)'(\bar{x} - X) = \sum N_i f_1(X)' x_i / \sum N_i - f_1(X)'X$, *i.e.*, we may treat the problem as one of inference regarding the quantities (pseudo-observations) $z_i = f_1(X)' x_i$. Because the approximate (linearized) influence of PSU i on the estimate $f(\bar{x})$, that is, the difference between the estimate with and without PSU i included, is $N_i f_1(X)'(x_i - \bar{x})$, we may describe this as a method based on influence statistics (Hampel *et al.* 1986) or the infinitesimal jackknife (Efron 1982, Chapter 6).

To derive a sensible variance model, suppose that we may regard PSU i as sample (not necessarily SRS) from a superpopulation with cell proportions X_i. A simple model is then, for some covariance matrices U_i and V_i,

superpopulation model:

$$E(X_i) = X, \quad \mathrm{Var}(X_i) = V_i$$

and

sampling model:

$$E(x_i \mid X_i) = X_i, \quad \mathrm{Var}(x_i \mid X_i) = U_i.$$

The sampling covariance U_i will typically be proportional to N_i^{-1}. A plausible and mathematically convenient specification for V_i is $V_i \alpha N_i^{-1}$ (*i.e.*, smaller PSUs more variable than larger ones), so $\mathrm{Var} z_i = \sigma^2 / N_i$ for some constant σ^2. The corresponding linear model weight for PSU i is N_i so the model-based estimate of the mean agrees with the design-based estimate obtained by aggregating the cell counts if N_i is a weighted size measure.

In the case of the adjustment factor $\hat{R} = (WCE/WE)/(WM/WP)$, the pseudo-observations are of the form $z_i = f_1(X)'(x_i - \bar{x}) =$

$$\hat{R}\left(\frac{WCE_i}{WCE} + \frac{WP_i}{WP} - \frac{WE_i}{WE} - \frac{WM_i}{WM}\right),$$

where WCE_i, WP_i, WE_i and WM_i are similar to the above for the i-th BP. We approximate the appropriate weight of a block part by $N_i = (WE_i + WP_i)/2$.

If the variance specifications of the model are inaccurate so there is some heteroscedasticity, or if the distribution is very long-tailed, then there will be a long-tailed distribution of residuals, making the tests at least slightly liberal. Some care must be taken to note the presence of outliers signaling this heteroscedasticity, for example, outlying blocks due to large-scale geocoding errors.

The assumption of approximately independent observations in ANOVA may be violated in two ways. First, the PSUs are not selected by SRS but rather by a geographical stratification somewhat finer than reflected in the post-stratification scheme. To the extent that this geographical stratification reduces the sampling variance of the state effect estimates, inferences under the independence model will be somewhat conservative. Second, there will be correlations between adjustment factors for different block parts from the same block (in multi-PSG models). These will tend to make inferences assuming independence somewhat liberal. On the balance, we regard the tests performed here as useful.

REFERENCES

ALHO, J.M., MULRY, M.H., WURDEMAN, K., and KIM, J. (1991). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.

BUREAU OF THE CENSUS (1990). Sample Selection Procedures for Performing Evaluation Study P12. STSD 1990 Coverage Studies and Evaluation Memorandum Series No. N-1, Memorandum from D. Bateman to L. Iskow and M. Lynch, October 3, 1990.

BUREAU OF THE CENSUS (1991). Request for Block Split Level Data for Performing PES Evaluation Project P12. STSD 1990 Coverage Studies and Evaluation Memorandum Series No. N-2, Memorandum from J. Thompson to A. Jackson, January 30, 1991.

COMMITTEE ON NATIONAL STATISTICS, PANEL TO EVALUATE ALTERNATIVE CENSUS METHODS (1994). *Counting People in the Information Age*. Washington D.C.: National Academy Press.

EFRON, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

FAY, R.E., and THOMPSON, J.H. (1993). The 1990 Post Enumeration Survey Statistical Lessons, in Hindsight. *Proceedings of the 1993 Annual Research Conference*. U.S. Bureau of the Census, 71-91.

FREEDMAN, D.A., and NAVIDI, W.C. (1992). Should we have adjusted the U.S. Census of 1980? *Survey Methodology*, 18, 3-24.

FREEDMAN, D.A., PISANI, R., and PURVIS, R. (1978). *Statistics*. New York: Norton.

FREEDMAN, D.A., and WACHTER, K.W. (1993). Heterogeneity and Census Adjustment for the Inter-Censal Base. Technical Report No. 381, Department of Statistics, University of California at Berkeley.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.

HENGARTNER, N., and SPEED, T.P. (1993). Assessing between-block heterogeneity within the poststrata of the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 88, 1119-1125.

HOGAN, H. (1992). The 1990 Post Enumeration Survey: An overview. *American Statistician*, 46, 261-269.

HOGAN, H. (1993). The 1990 Post Enumeration Survey: Operations and results. *Journal of the American Statistical Association*, 88, 1047-1057.

ISAKI, C.T., SCHULTZ, L.K., DIFFENDAL, G.J., and HUANG, E.T. (1988). On estimating census undercount in small areas. *Journal of Official Statistics*, 4, 95-112.

KIM, J. (1991). 1990 PES Evaluation Project P12: Evaluation of Synthetic Assumption. 1990 Coverage Studies and Evaluation Memorandum Series No. N-4, internal memorandum, U.S. Bureau of the Census.

MULRY, M.H., and SPENCER, B.D. (1993). Accuracy of the 1990 Census and undercount adjustment. *Journal of the American Statistical Association*, 88, 1080-1091.

SCHAFER, J.L. (1993). Comment on Hengartner, N and Speed, T.P.'s Assessing between-block heterogeneity within the poststrata of the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 88, 1125-1127.

SCHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and quality of geographic population distributions. *Journal of the American Statistical Association*, 82, 965-978.

WACHTER, K.W., and FREEDMAN, D.A. (1992). Measuring Local Homogeneity 1990 Census Data. Technical Report, Department of Statistics, University of California at Berkeley.

WOLTER, K.M., and CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *Journal of the American Statistical Association*, 86, 278-284.

Markov Chain Designs for One-Per-Stratum Sampling

F. JAY BREIDT[1]

ABSTRACT

Classical results in finite population sampling tell us that systematic sampling is the most efficient equal-probability one-per-stratum design for certain kinds of autocorrelated superpopulations, but stratified simple random sampling may be much better than systematic sampling if the superpopulation is a trend with uncorrelated errors. What if the superpopulation consists of a trend plus autocorrelated errors? Intuitively, some sort of "compromise" between the two designs might be better than either. Such compromise designs are constructed in this paper and are shown to be examples of Markov chain designs, a wide class of methods for one-per-stratum selection from a finite population. These designs include as special cases systematic sampling, balanced systematic sampling and stratified simple random sampling with one sampling unit per stratum. First and second-order inclusion probabilities are derived for Markov chain designs, yielding the Horvitz-Thompson estimator and its variance. Efficiency of the Horvitz-Thompson estimator is evaluated using superpopulation models. Numerical examples show that new designs considered here can be more efficient than standard designs for superpopulations consisting of trend plus autocorrelated errors. An example of the implementation of Markov chain designs for the 1992 National Resources Inventory in Alaska is given.

KEY WORDS: Balanced systematic sampling; National Resources Inventory; Systematic sampling.

1. INTRODUCTION

A stratified sampling design, in which a finite population is divided into non-overlapping strata and samples are drawn from each stratum, is a common and effective technique for reducing sampling error. In practice, stratified sampling designs with only one sampling unit per stratum are widely used. Examples include stratified simple random sampling and systematic sampling with its variants (*e.g.*, Murthy and Rao 1988).

Systematic samples are susceptible to systematic errors. In large-scale spatial samples, for example, sources of systematic error could include roads, powerlines, irrigation systems, and so forth. A favorite example is the system of "section roads" in areas of the United States covered by the public land survey. This grid-based system is built up from square tracts of land called sections, each one mile on a side, which are often bounded by roads in midwestern agricultural regions. A systematic sampler with a one-mile sampling interval and an unlucky random start might conclude that Iowa is covered by gravel roads!

Systematic sampling does have the advantage of efficiency when the sampled population is positively autocorrelated, as is often the case in temporal and spatial sampling problems, since it forces observations to be as distant and hence as uncorrelated as possible.

Both autocorrelation and systematic error are of concern in the National Resources Inventory (NRI), an area sample of the nonfederal lands in the United States conducted every five years by the Soil Conservation Service of the United States Department of Agriculture. NRI data items, collected by a combination of remote sensing and ground observation, include soil characteristics, land use, agricultural practices, erosion measures, and so on.

The 1992 NRI sample design for the northwestern region of the state of Alaska is a controlled version of one-per-stratum sampling. The region was divided into twenty-minute bands of latitude. Each band was divided into 500,000-acre strata. Each stratum was divided into a 10 × 10 grid of cells indexed by latitude and longitude, and one cell per stratum was selected. Selection moved from east to west across the strata within a particular twenty-minute band. The random numbers which determined the longitude cells of the selected units and the random numbers which determined the latitude cells evolved as two independent Markov chains. (Basic results on Markov chains used in this paper can be found in an introductory text on stochastic processes such as Taylor and Karlin 1984). Details of the design are given in Section 2.

How does this *ad hoc* design compare to more standard one-per-stratum designs? It turns out, as shown in Section 2, that simple Markov chain techniques can describe a broad class of equal-probability designs for one-per-stratum selection from a finite population. This class includes standard techniques such as stratified simple random sampling, systematic sampling and balanced systematic sampling, as well as the Alaska designs described above. It is also easy to generate new designs within this class. This unified treatment of one-per-stratum designs allows for comparisons of efficiency.

[1] F. Jay Breidt, Iowa State University, Department of Statistics, Ames, IA 50011-1210, U.S.A.

First and second-order inclusion probabilities for all of these designs are derived in Section 3, yielding the Horvitz-Thompson estimator and its variance. As in much of the relevant literature (Madow and Madow 1944; Cochran 1946; Sedransk 1969; Bellhouse and Rao 1975; Wolter 1985; Bellhouse 1988; *etc.*) the average design variance of the Horvitz-Thompson estimator is evaluated under a variety of superpopulation models. Compact expressions for model-averaged design variances are obtained. Numerical examples in Section 4 show that designs introduced in this paper can be more efficient than standard one-per-stratum designs for superpopulations consisting of trend plus autocorrelated errors. Discussion follows in Section 5.

Though our motivating example is two-dimensional, one-dimensional designs will be considered throughout. Most proofs and derivations are straightforward and are omitted for brevity.

2. MARKOV CHAIN DESIGNS

Consider the problem of sampling from a finite population of $N = na$ labeled units, denoted by

$$U = \{1, \ldots, N\}$$
$$= \{1, \ldots, a, a + 1, \ldots, 2a, \ldots, (n - 1)a + 1, \ldots, na\}.$$

The value of a study variable $y_k = y_{(i-1)a+j} = y_{ij}$ is associated with each label k; the notation y_k or y_{ij} will be used for both random variables and realizations of random variables.

Here n is the sample size and a is the *sampling interval*. The n subsets

$$\{(i - 1)a + 1, \ldots, (i - 1)a + a\} \quad (i = 1, \ldots, n)$$

will be referred to as *strata*. The goal is to select one unit per stratum. Often, a stratified sampling design is defined to be one in which independent probability samples are selected in each stratum, but the restriction to independence is not used here.

Given a doubly stochastic transition probability matrix P, a *Markov chain sample* is given by

$$s = \{R_1, a + R_2, \ldots, (n - 1)a + R_n\},$$

where $R_1, \ldots, R_n$ is the Markov chain defined by P and $R_1 \sim$ uniform $(1, \ldots, a)$. Formally, then, a *Markov chain design* (MC) is a function $p(\,\cdot\,;P)$ such that

$$p(s;P) = \Pr\{s = \{r_1, a + r_2, \ldots, (n - 1)a + r_n\}\}$$
$$= \Pr\{R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n\}$$
$$= \begin{cases} P_{r_{n-1},r_n} P_{r_{n-2},r_{n-1}} \cdots P_{r_1,r_2}/a, \\ \qquad \text{for} \quad r_1, \ldots, r_n \in \{1, \ldots, a\}, \\ 0, \quad \text{otherwise.} \end{cases}$$

MC designs as defined in this paper are related to the designs given in Chandra, Sampath and Balasubramani (1992), in which a $1 \times N$ vector of initial selection probabilities and a $N \times N$ transition probability matrix of periodicity n determine a without-replacement sampling scheme. Chandra *et al.* focus on producing designs with strictly positive second-order inclusion probabilities. They do not explicitly consider the one-per-stratum designs of this paper, which can be imbedded in their structure in a straightforward way by constructing the appropriate initial probability vector and transition probability matrix.

The following result is useful in deriving the probabilistic features of MC designs.

Result 1 Consider a Markov chain for which the transition probability matrix P is doubly stochastic (*i.e.*, all row sums and all column sums equal one) and R_1 has a discrete uniform distribution, with mass $1/a$ on each of the states $1, \ldots, a$. Then R_i has a discrete uniform distribution on the states $1, \ldots, a$ for all i. In particular, R_i has mean $(a + 1)/2$ and variance $V(R_i) = (a^2 - 1)/12$.

Some special cases of MC designs are of interest.

Stratified simple random sampling. If the transition probability matrix is

$$H = [1/a]_{j,j'=1}^{a},$$

then

$$\Pr\{R_{i'} = j' \mid R_i = j\} = 1/a = \Pr\{R_{i'} = j'\}$$
$$(j,j' = 1, \ldots, a; i < i'),$$

which, together with the Markov property, implies that $R_1, \ldots, R_n$ are probabilistically independent. In this case, the MC design is stratified simple random sampling with one unit per stratum (ST).

Systematic sampling. If the transition probability matrix is I, the $a \times a$ identity matrix, then

$$\Pr\{R_{i'} = j' \mid R_i = j\} = \begin{cases} 1, j = j', \\ 0, j \neq j', \end{cases}$$

so that $R_1, \ldots, R_n$ are deterministically related. Thus,

$$s = \{R_1, a + R_1, \ldots, (n - 1)a + R_1\},$$

and so the MC design is systematic sampling (SY).

Compromise designs. Intuitively, ST and SY are at opposite "extremes" in some sense. If $\rho \in [0,1]$, then

$$G_\rho = \rho H + (1 - \rho) I$$

is doubly stochastic. If $\rho = 0$, the design is SY and if $\rho = 1$, the design is ST. Any other choice of ρ will yield a sequence consisting of "runs" of SY samples. Thus, the class G_ρ includes ST and SY, as well as a continuum of "compromise" MC designs.

Other convex combinations of doubly stochastic matrices could be considered. The class of doubly stochastic matrices is also closed under matrix multiplication, transposition, and row and column permutation, so there are many ways to create MC designs.

Balanced systematic sampling. Murthy (1967, §5.9d) describes a one-per-stratum selection method which he calls *balanced systematic sampling* (BA). This method gives samples

$$s = \{R_1, a + (a + 1 - R_1), \ldots, (n - 2)a + R_1, \\ (n - 1)a + (a + 1 - R_1)\}$$

for n even and

$$s = \{R_1, a + (a + 1 - R_1), \ldots, (n - 2)a + \\ (a + 1 - R_1), (n - 1)a + R_1\}$$

for n odd. An interesting feature of this design is that if n is even and the population is perfectly linear ($y_{ij} = \beta_0 + \beta_1[(i - 1)a + j]$), then the sample mean equals the population mean for any sample. With the transition probability matrix,

$$J = \begin{bmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \end{bmatrix}_{a \times a},$$

BA is a MC design.

Alaska NRI design. As described in Section 1, the 1992 NRI sample design for the northwestern region of the state of Alaska used two independent Markov chains in the controlled selection of latitude and longitude cells. The transition probability matrix for longitude cells, P_{long}, is given in Table 1. This design, henceforth denoted AK, is a MC design since P_{long} is doubly stochastic. Most of the transition probabilities are close to 0.10, so most "step sizes" are approximately equally likely. Note, however, that mass has been removed from on and near the back diagonal and placed in the upper left and lower right corners, so that P_{long} discourages large east to west steps, such as from cell one to cell ten, and discourages small steps, such as from cell ten to cell one. On the other hand, P_{long} encourages steps of around length ten, such as from cell two to cell one, two or three. The realized sample of longitude cells is thus well-dispersed east to west, like a systematic sample would be, but its additional randomness guards against systematic error. Similarly, the Markov chain for latitude cells was set up to give good spatial dispersion north to south.

Table 1

Transition probability matrix for Markov chain sample of longitude cells,1992 National Resources Inventory, Alaska. Entries are the conditional probabilities of selecting cell j' of stratum $i + 1$ given that cell j of stratum i was selected.

Cell j of stratum i	Cell j' of stratum $i + 1$									
	1	2	3	4	5	6	7	8	9	10
1	0.05	0.15	0.15	0.15	0.15	0.15	0.10	0.10	0	0
2	0.15	0.15	0.15	0.10	0.10	0.10	0.10	0.10	0.05	0
3	0.15	0.15	0.10	0.10	0.10	0.10	0.05	0.05	0.10	0.10
4	0.15	0.10	0.10	0.10	0.10	0.10	0.10	0.05	0.10	0.10
5	0.15	0.10	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.15
6	0.15	0.10	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.15
7	0.10	0.10	0.05	0.10	0.10	0.10	0.10	0.10	0.10	0.15
8	0.10	0.10	0.05	0.05	0.10	0.10	0.10	0.10	0.15	0.15
9	0	0.05	0.10	0.10	0.10	0.10	0.10	0.15	0.15	0.15
10	0	0	0.10	0.10	0.15	0.15	0.15	0.15	0.15	0.05

3. HORVITZ-THOMPSON ESTIMATION UNDER MC

Write the population total as

$$t = \sum_U y_k = \sum_{i=1}^{n} \sum_{j=1}^{a} y_{(i-1)a+j} = \sum_{i=1}^{n} \sum_{j=1}^{a} y_{ij}.$$

For all k, the first-order inclusion probabilities of a MC design are given by

$$\pi_k = \Pr\{k \in s\} = \Pr\{R_i = j\} = 1/a$$

and for $k \leq l$, the second-order inclusion probabilities are given by

$$\pi_{kl} = \begin{cases} 1/a, & \text{for } i = i', j = j', \\ 0, & \text{for } i = i', j \neq j', \\ P_{jj'}^{(i'-i)}/a, & \text{for } i < i'. \end{cases}$$

The design-unbiased Horvitz-Thompson estimator (Horvitz and Thompson 1952) for the population total is then

$$\hat{t}_\pi = \sum_s y_k/\pi_k = \sum_{i=1}^{n} \frac{y_{iR_i}}{1/a} = a \sum_{i=1}^{n} \sum_{j=1}^{a} y_{ij} I_{\{R_i=j\}},$$

where

$$I_{\{R_i=j\}} = \begin{cases} 1, & \text{if } R_i = j, \\ 0, & \text{if } R_i \neq j. \end{cases}$$

The design covariances of the indicators $I_{\{R_i=j\}}$ are given by

$$\begin{aligned} C_{\mathrm{MC}}\left(I_{\{R_i=j\}}, I_{\{R_{i'}=j'\}}\right) &= E_{\mathrm{MC}}\left[I_{\{R_i=j\}} I_{\{R_{i'}=j'\}}\right] - \\ &\qquad E_{\mathrm{MC}}\left[I_{\{R_i=j\}}\right] E_{\mathrm{MC}}\left[I_{\{R_{i'}=j'\}}\right] \\ &= \pi_{(i-1)a+j,(i'-1)a+j'} - \\ &\qquad \pi_{(i-1)a+j}\, \pi_{(i'-1)a+j'}, \end{aligned}$$

and so the design variance of $\hat{t}_\pi$ is

$$\begin{aligned} V_{\mathrm{MC}}(\hat{t}_\pi) = a^2 \sum_{i=1}^{n} \sum_{j=1}^{a} \left(\frac{1}{a} - \frac{1}{a^2}\right) y_{ij}\, y_{ij} \qquad (1) \\ + a^2 \sum_{i=1}^{n} \sum_{j=1}^{a} \sum_{j' \neq j} \left[0 - \frac{1}{a^2}\right] y_{ij}\, y_{ij'} \\ + 2a^2 \sum_{i=1}^{n} \sum_{i'>i} \sum_{j=1}^{a} \left[\frac{P_{jj}^{(i'-i)}}{a} - \frac{1}{a^2}\right] y_{ij}\, y_{i'j} \\ + 2a^2 \sum_{i=1}^{n} \sum_{i'>i} \sum_{j=1}^{a} \sum_{j' \neq j} \left[\frac{P_{jj'}^{(i'-i)}}{a} - \frac{1}{a^2}\right] y_{ij}\, y_{i'j'}. \end{aligned}$$

Since the design variance depends on all the values of the study variable in the finite population, (1) is not easily used for comparing designs. Following Cochran (1946), assume that the values of the study variable are generated from the superpopulation model

$$\xi : y_{ij} = \mu_{ij} + e_{ij},$$

where the μ_{ij} are fixed and the e_{ij} are random variables with $E_\xi[e_{ij}] = 0$, $V_\xi(e_{ij}) = \sigma_{ij}^2$ and $C_\xi(e_{ij}, e_{i'j'}) = \sigma_{ij,i'j'}$. Then designs can be compared on the basis of model-averaged design variance.

Proposition 1 Under the superpopulation model ξ, the average design variance of the Horvitz-Thompson estimator is

$$\begin{aligned} E_\xi[V_{\mathrm{MC}}(\hat{t}_\pi)] = a^2 V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right] + \\ (a-1) \sum_{i=1}^{n} \sum_{j=1}^{a} \sigma_{ij}^2 - \sum_{i=1}^{n} \sum_{j=1}^{a} \sum_{j' \neq j} \sigma_{ij,ij'} \\ + 2a \sum_{i=1}^{n} \sum_{i'>i} \sum_{j=1}^{a} \sum_{j'=1}^{a} \sigma_{ij,i'j'} \left[P_{jj'}^{(i'-i)} - \frac{1}{a}\right] \end{aligned}$$

for any MC design. Note that if μ_{ij} is independent of j, then $V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right] = 0$.

The following proposition gives a sufficient condition under which no MC design has worse average design variance than SY.

Proposition 2 Consider an uncorrelated additive model,

$$\xi : y_{ij} = \mu_{ij} + e_{ij} = \alpha_i + \beta_j + e_{ij},$$

where $E_\xi[e_{ij}] = 0$, $V_\xi(e_{ij}) = \sigma_{ij}^2$ and $C_\xi(e_{ij}, e_{i'j'}) = 0$. Then

$$E_\xi[V_{\mathrm{SY}}(\hat{t}_\pi)] \geq E_\xi[V_{\mathrm{MC}}(\hat{t}_\pi)]$$

for all MC designs.

Proof From Proposition 1, the only term of interest is $V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right]$, which under SY is

$$V_{\mathrm{SY}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right] = V_{\mathrm{SY}}\left[\sum_{i=1}^{n} \alpha_i + n\beta_{R_1}\right] = n^2 V\left(\beta_{R_1}\right),$$

while under a general MC design,

$$V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \mu_{iR_i}\right] = \sum_{i=1}^{n} \sum_{i'=1}^{n} C_{\mathrm{MC}}\left(\beta_{R_i}, \beta_{R_{i'}}\right).$$

Since $C_{\mathrm{MC}}\left(\beta_{R_i}, \beta_{R_{i'}}\right) \leq V\left(\beta_{R_1}\right)$, the proposition follows. □

Some specific models are considered in the next five subsections.

3.1 Random Permutation Model

A model for a population in random order is a permutation model, in which a realization of the measurements $y_1, \ldots, y_N$ is given by one of the $N!$ equally likely permutations of N fixed values. This model can be written as

$$\xi_1 : y_{ij} = \bar{y}_U + e_{ij},$$

where $\bar{y}_U = \sum_U y_k / N$. See Rao (1975) for more details. The following result is then a consequence of Theorem 2.1 of Rao and Bellhouse (1978).

Result 2 Under the random permutation model,

$$E_{\xi_1}[V_{\mathrm{MC}}(\hat{t}_\pi)] = (N^2/n)(1 - n/N)\sum_U (y_k - \bar{y}_U)^2/(N - 1)$$

for any MC design.

Thus, the average variance over all permutations is exactly $V_{\mathrm{SI}}(\hat{t}_\pi)$, where SI denotes (unstratified) simple random sampling without replacement. For SY, this result is originally due to Madow and Madow (1944). See also Sedransk (1969).

3.2 Stratification Effects Model

A model for a population with stratification effects is

$$\xi_2 : y_{ij} = \alpha_i + e_{ij},$$

where the α_i are fixed constants and e_{ij} are uncorrelated random variables with mean zero and variance σ^2. Note that if $\alpha_i \equiv \mu$, then ξ_2 is an alternative to ξ_1 as a model for a population in random order.

Result 3 Under the stratification effects model,

$$E_{\xi_2}[V_{\mathrm{MC}}(\hat{t}_\pi)] = na(a - 1)\sigma^2$$

for any MC design.

3.3 Linear Trend Model

A model for a population with a linear trend is

$$\xi_3 : y_{ij} = \beta_0 + \beta_1[(i - 1)a + j] + e_{ij},$$

where β_0 and β_1 are fixed constants and e_{ij} are uncorrelated $(0,\sigma^2)$ random variables.

Result 4 Under the linear trend model ξ_3,

$$E_{\xi_3}[V_{\mathrm{MC}}(\hat{t}_\pi)] = \beta_1^2 a^2 V_{\mathrm{MC}}\left[\sum_{i=1}^{n} R_i\right] + na(a - 1)\sigma^2 \quad (2)$$

for any MC design. Since ξ_3 is additive, no MC design has a larger expected variance under a linear trend model than SY.

The only design-dependent term in (2) is $V_{\mathrm{MC}}\left[\sum_{i=1}^{n} R_i\right]$. Under SY, $\sum_{i=1}^{n} R_i = nR_1$, so that

$$V_{\mathrm{SY}}\left[\sum_{i=1}^{n} R_i\right] = n^2 V(R_1),$$

while under ST,

$$V_{\mathrm{ST}}\left[\sum_{i=1}^{n} R_i\right] = nV(R_1).$$

Under BA, for n even,

$$V_{\mathrm{BA}}\left[\sum_{i=1}^{n} R_i\right] = V_{\mathrm{BA}}\left[\frac{n}{2}R_1 + \frac{n}{2}(a + 1 - R_1)\right] = 0.$$

This implies that if the population is perfectly linear $(\sigma^2 = 0)$, then

$$E_{\xi_3}[V_{\mathrm{BA}}(\hat{t}_\pi)] = 0,$$

so that $\hat{t}_\pi = t$ for all samples, as noted by Murthy (1967, p. 165).

Result 5 Under the linear trend model ξ_3,

$$\begin{aligned} E_{\xi_3}[V_{\mathrm{BA}}(\hat{t}_\pi)] &\le E_{\xi_3}[V_{\mathrm{ST}}(\hat{t}_\pi)] \\ &\le E_{\xi_3}[V_{G_\rho}(\hat{t}_\pi)] \\ &\le E_{\xi_3}[V_{\mathrm{SY}}(\hat{t}_\pi)] = \max_{\mathrm{MC}} E_{\xi_3}[V_{\mathrm{MC}}(\hat{t}_\pi)], \end{aligned} \quad (3)$$

where the middle term is monotone increasing with decreasing $\rho \in [0,1]$. If n is even, the left-hand side of (3) equals $\min_{\mathrm{MC}} E_{\xi_3}[V_{\mathrm{MC}}(\hat{t}_\pi)]$.

3.4 Periodic Population Model

A simple model for a population showing a deterministic periodicity with period p is the sine wave model

$$\xi_4 : y_{ij} = \alpha \sin\left\{\frac{2\pi}{p}[(i - 1)a + j]\right\} + e_{ij},$$

where e_{ij} are uncorrelated random variables with mean zero and variance σ^2.

Result 6 Under the periodic population model ξ_4,

$$E_{\xi_4}[V_{\mathrm{MC}}(\hat{t}_\pi)] = a^2\alpha^2 V_{\mathrm{MC}}\left[\sum_{i=1}^{n} \sin\frac{2\pi}{p}[(i - 1)a + R_i]\right] + na(a - 1)\sigma^2$$

for any MC design.

Denote the sine wave model ξ_4 with $p = a$ by ξ_{4a}. Under ξ_{4a},

$$\sin\left\{\frac{2\pi}{p}[(i - 1)a + j]\right\} = \sin\frac{2\pi j}{a},$$

so that the model is additive and no MC design has larger expected design variance under ξ_{4a} than SY, highlighting the fact that SY is inappropriate for a population containing a periodicity with period equal to the sampling interval (Madow and Madow 1944). This result generalizes as follows.

Result 7 If $\mu_{ij} \equiv \beta_j$ in ξ, then ξ is a model for a population showing a deterministic periodicity with period equal to the sampling interval, a. The model ξ is additive and so no MC design has larger expected design variance under ξ than SY.

3.5 Autocorrelated Model

Beginning with Cochran (1946), many authors have compared ST, SY and simple random sampling under an autocorrelated superpopulation model. See Bellhouse (1988, §4) for a review.

Consider the following autocorrelation model due to Cochran (1946):

$$\xi_5 : y_{ij} = \mu + e_{ij}$$

where $\sigma_{ij,i'j'} = \gamma[(i' - i)a + j' - j]$ for $i' \geq i$.

Result 8 Under the autocorrelated model ξ_5,

$$E_{\xi_5}[V_{\text{MC}}(\hat{t}_\pi)] = na(a-1)\gamma(0) - 2n\sum_{h=1}^{a-1}\gamma(h)(a-h)$$

$$+ 2a\sum_{h=1}^{n-1}\sum_{j=1}^{a}\sum_{j'=1}^{a}\gamma(ha + j' - j)(n-h)\left(P_{jj'}^{(h)} - \frac{1}{a}\right)$$

for any MC design.

Result 9 If, for $h \geq 0$, $\gamma(h)$ is non-negative, non-increasing and convex, *i.e.*,

$$\gamma(h) \geq 0, \gamma(h) \geq \gamma(h+1) \quad \text{and}$$
$$\gamma(h+2) - 2\gamma(h+1) + \gamma(h) \geq 0,$$

then $E_{\xi_5}[V_{\text{SY}}(\hat{t}_\pi)] = \min_{\text{MC}} E_{\xi_5}[V_{\text{MC}}(\hat{t}_\pi)]$.

This result is a corollary of a theorem due to Hájek (1959), given as Theorem 4.1 of Bellhouse (1988); Bellhouse clarified the conditions under which the theorem holds. Hájek's theorem generalized an earlier result due to Cochran (1946), who compared SY, ST and simple random sampling.

4. EFFICIENCY: SOME NUMERICAL EXAMPLES

An important class of models for time series and spatial processes consists of a low-order polynomial trend plus an autocorrelated error sequence. A simple example is

$$\xi_{(\beta,\phi)} : y_{ij} = \beta_0 + \beta_1[(i-1)a + j] + e_{ij},$$

where the autocorrelation structure is that of a first-order autoregressive (AR) model,

$$\sigma_{ij,i'j'} = \gamma[(i' - i)a + j' - j] = \sigma^2\phi^{(i'-i)a+j'-j}$$

for $i' \geq i$ and $|\phi| < 1$. The average design variance under this model is obtained from Results 4 and 8. For different choices of β_1 and ϕ, the ratio of expected design variances,

$$E_\xi[V_{\text{MC}}(\hat{t}_\pi)]/E_\xi[V_{\text{SY}}(\hat{t}_\pi)], \tag{4}$$

is given in Table 2 for various MC designs. Also tabled is the optimal G_ρ design, obtained by minimizing (4) with respect to ρ. Use of this design is only feasible if superpopulation parameters are known, so it is tabled merely as a benchmark and not as a competitor.

When $\beta_1 \neq 0$ and $\phi = 0$, the model is ξ_3 and the tabled values agree with Result 5: SY is the worst MC design and BA is the best, with $G_{1/3}$, $G_{2/3}$ and ST falling between them. Though BA does extremely well for this model, any non-SY MC design would be a good choice.

When $\beta_1 = 0$ and $\phi \neq 0$, $\xi_{(\beta,\phi)}$ is a special case of model ξ_5. For $\phi > 0$, Result 9 and the table agree that SY is most efficient since it makes the sample as "spread out" as possible, but for weak autocorrelation, the other MC designs are competitive. BA is very poor for this model, because the design ensures that every other pair R_i, R_{i+1} will be no more than a units apart. (For the same reason, BA is good for a negatively autocorrelated population.) AK, $G_{1/3}$ and $G_{2/3}$ outperform ST, because each of these designs encourages state transitions of around length a.

Similar results are obtained for the superpopulation model

$$\xi_{(\alpha,\phi)} : y_{ij} = \alpha\sin\frac{2\pi j}{a} + e_{ij},$$

where $\sigma_{ij,i'j'}$ is as above. Table 2 gives the ratio of expected design variances (4) under this model, obtained from Results 6 and 8.

When $\alpha \neq 0$ and $\phi = 0$, the model is ξ_{4a} and SY performs badly, as indicated by Result 7. Even for $\phi \neq 0$, SY performs well only when the periodicity is swamped by highly-correlated noise.

Note that no design dominates Table 2: each of SY, $G_{1/3}$, $G_{2/3}$, ST, BA and AK is the best at least once among those considered. For a moderate trend and high autocorrelation, AK, $G_{1/3}$ and $G_{2/3}$ can beat standard MC designs. Overall, Table 2 suggests that some non-standard MC designs, such as $G_{2/3}$ and AK, do reasonably well for a variety of populations: retaining much of the efficiency of SY against an autocorrelated population, while still guarding against systematic effects in other kinds of populations.

Table 2

Ratio of expected design variance under MC to expected design variance under SY for superpopulation consisting of trend (line with slope β_1 or sine wave with period a and amplitude α) plus autoregressive (AR) errors ($N = 1{,}000$, $\sigma^2 = 100$, $a = 10$). Here G_{ρ^*} is the optimal compromise design, where ρ^* is a function of superpopulation parameters. Ratio for the best realizable design in each row (if not SY) is italicized.

Model		Markov Chain Design						
	ϕ	$G_{1/3}$	$G_{2/3}$	ST	BA	AK	G_{ρ^*}	(ρ^*)
Line + AR $\beta_1 = 0.7$	−0.5	0.2322	0.2085	0.2001	*0.1666*	0.2056	0.2001	(1.0000)
	0.0	0.2220	0.1983	0.1903	*0.1821*	0.1957	0.1903	(1.0000)
	0.1	0.2187	0.1950	0.1871	*0.1825*	0.1921	0.1871	(1.0000)
	0.5	0.1922	0.1702	*0.1645*	0.1754	0.1659	0.1645	(1.0000)
	0.9	0.0980	0.0778	*0.0742*	0.0768	0.0762	0.0742	(1.0000)
Line + AR $\beta_1 = 0.4$	−0.5	0.4504	0.4328	0.4262	*0.3647*	0.4304	0.4262	(1.0000)
	0.0	0.4344	0.4172	0.4114	*0.4054*	0.4153	0.4114	(1.0000)
	0.1	0.4291	0.4121	*0.4065*	0.4085	0.4094	0.4065	(1.0000)
	0.5	0.3853	0.3727	0.3724	0.4116	*0.3667*	0.3719	(0.8320)
	0.9	0.1876	*0.1835*	0.1914	0.2170	0.1848	0.1821	(0.5223)
Line + AR $\beta_1 = 0.1$	−0.5	0.9233	0.9190	0.9163	*0.7941*	0.9175	0.9163	(1.0000)
	0.0	0.9201	0.9177	0.9169	*0.9160*	0.9174	0.9169	(1.0000)
	0.1	0.9191	0.9175	0.9175	0.9349	*0.9156*	0.9174	(0.8156)
	0.5	*0.9160*	0.9289	0.9439	1.0606	0.9185	0.9135	(0.1997)
	0.9	*0.8621*	0.9787	1.0725	1.2710	1.0017	0.7888	(0.0981)
Pure AR	−0.5	0.9978	0.9956	0.9935	*0.8617*	0.9942	0.9935	(1.0000)
	0.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	(---)
	0.1	1.0009	1.0019	1.0028	1.0228	1.0001	1.0000	(0.0000)
	0.5	1.0179	1.0357	1.0536	1.1852	1.0245	1.0000	(0.0000)
	0.9	1.2517	1.4380	1.5814	1.8798	1.4734	1.0000	(0.0000)
Sine + AR $\alpha = 0.1$	−0.5	0.9929	0.9906	0.9884	*0.8578*	0.9892	0.9884	(1.0000)
	0.0	0.9947	0.9946	*0.9945*	0.9950	0.9946	0.9945	(1.0000)
	0.1	0.9955	0.9963	0.9972	1.0175	*0.9945*	0.9954	(0.1925)
	0.5	1.0110	1.0285	1.0462	1.1775	1.0173	0.9977	(0.0364)
	0.9	1.2178	1.3980	1.5371	1.8294	1.4322	0.9999	(0.0018)
Sine + AR $\alpha = 1.0$	−0.5	0.6747	0.6634	0.6586	*0.6008*	0.6604	0.6586	(1.0000)
	0.0	0.6603	0.6499	*0.6464*	0.6770	0.6477	0.6464	(1.0000)
	0.1	0.6554	0.6455	0.6425	0.6863	*0.6421*	0.6425	(1.0000)
	0.5	0.6149	0.6133	0.6196	0.7320	*0.6041*	0.6121	(0.5079)
	0.9	*0.3570*	0.3832	0.4126	0.5527	0.3877	0.3560	(0.2852)
Sine + AR $\alpha = 10.0$	−0.5	0.0668	0.0384	*0.0287*	0.1101	0.0323	0.0287	(1.0000)
	0.0	0.0656	0.0372	*0.0275*	0.1115	0.0311	0.0275	(1.0000)
	0.1	0.0652	0.0368	*0.0271*	0.1115	0.0307	0.0271	(1.0000)
	0.5	0.0622	0.0339	*0.0245*	0.1106	0.0277	0.0245	(1.0000)
	0.9	0.0529	0.0247	*0.0154*	0.1016	0.0187	0.0154	(1.0000)

5. DISCUSSION

The class of Markov chain designs has been defined and shown to include systematic sampling, stratified simple random sampling and balanced systematic sampling as special cases. Some new designs have been introduced (G_ρ, AK) and shown to be competitive with standard one-per-stratum designs under a variety of superpopulation models. In particular, the new designs work well in numerical examples for trending superpopulations with autocorrelated errors. This is the kind of population of concern in many area sampling problems, such as the 1992 National Resources Inventory in Alaska. A two-dimensional MC design implemented for that survey shows that one-dimensional MC designs might be usefully extended to a spatial sampling context, though further work on this extension is necessary.

Further work on variance estimation for MC designs is also needed. Because these are one-per-stratum designs, design-unbiased estimation of the variance of the Horvitz-Thompson estimator is not possible. The problem of variance estimation for one-per-stratum designs, particularly for SY, has received much attention. For example, Wolter (1985) discusses in detail eight different biased variance estimators for SY and evaluates their biases under superpopulation models. Work in this direction for the collapsed strata variance estimator (*e.g.*, Cochran 1977, p. 139) under general MC designs is in progress.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Agriculture's Soil Conservation Service under SCS Cooperative Agreement No. 68-3A75-2-64. Jeff Goebel (Soil Conservation Service) and Wayne Fuller (Iowa State University) developed the MC design for the state of Alaska. The author thanks the anonymous referees for their constructive comments on an earlier version of this paper.

REFERENCES

BELLHOUSE, D.R. (1988). Systematic sampling. In *Handbook of Statistics*. (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6. Amsterdam: North-Holland, 125-145.

BELLHOUSE, D.R., and RAO, J.N.K. (1975). Systematic sampling in the presence of a trend. *Biometrika*, 62, 694-697.

CHANDRA, K.S., SAMPATH, S., and BALASUBRAMANI, G.K. (1992). Markov sampling for finite populations. *Biometrika*, 79, 210-213.

COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: Wiley.

HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Časopis Pro Pěstování Matematiky*, 84, 387-423.

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, 663-685.

MADOW, W.G., and MADOW, L.H. (1944). On the theory of systematic sampling, I. *Annals of Mathematical Statistics*, 15, 1-24.

MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

MURTHY, M.N., and RAO, T.J. (1988). Systematic sampling with illustrative examples. In *Handbook of Statistics*. (Eds. P.R. Krishnaiah and C.R. Rao), (Vol. 6). Amsterdam: North-Holland, 147-185.

RAO, J.N.K. (1975). On the foundations of survey sampling. In *A Survey of Statistical Design and Linear Models.* (Ed. J.N. Srivastava). Amsterdam: North-Holland, 489-505.

RAO, J.N.K., and BELLHOUSE, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Journal of Statistical Planning and Inference*, 2,125-141.

SEDRANSK, J. (1969). Some elementary properties of systematic sampling. *Skandinavisk Aktuarietidskrift*, 1-2, 39-47.

TAYLOR, H.M., and KARLIN, S. (1984). *An Introduction to Stochastic Modeling*. Orlando: Academic Press.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Survey Methodology, June 1995
Vol. 21, No. 1, pp. 71–77
Statistics Canada

Median Estimation Using Auxiliary Information

GLEN MEEDEN[1]

ABSTRACT

The problem of estimating the median of a finite population when an auxiliary variable is present is considered. Point and interval estimators based on a non-informative Bayesian approach are proposed. The point estimator is compared to other possible estimators and is seen to perform well in a variety of situations.

KEY WORDS: Sample survey; Estimation; Median; Auxiliary variable; Quantile; Non-informative Bayes.

1. INTRODUCTION

The problem of estimating a population mean in the presence of an auxiliary variable has been widely discussed in the finite population sampling literature. The ratio estimator has often been used in such situations. For the problem of estimating a population median the situation is quite different. Only recently has this problem been discussed. Chambers and Dunstan (1986) proposed a method for estimating the population distribution function and the associated quantiles. They assumed that the value of the auxiliary variable was known for every unit in the population and their estimator came from a model-based approach. Rao *et al.* (1990) proposed ratio and difference estimators for the median using a design-based approach. Kuk and Mak (1989) proposed two other estimators for the population median. To use the Kuk and Mak estimators one only needs to know the values of the auxiliary variable for the units in the sample and its median for the whole population. The efficiencies of these estimators depend directly on the probability of 'concordance' rather than on the validity of an assumption of linearity between the variable of interest and the auxiliary variable.

Recently Meeden and Vardeman (1991) discussed a non-informative Bayesian approach to finite population sampling. This new approach uses the 'Polya posterior' as a predictive distribution for the unobserved members of the population once the sample has been observed. Often it yields point and interval estimates that are very similar to those of standard frequentist theory. Moreover it can be easy to implement in problems that are difficult for standard theory. In this note we show how this method can be used for the problem of estimating a population median when an auxiliary variable is present and compare it to some of the other proposed methods.

2. ESTIMATING THE MEDIAN

Consider a finite population containing N units. For the unit with label i let y_i denote the characteristic of interest and x_i the auxiliary variable. We assume that both y_i and x_i are real numbers and each is known for every unit in the population. Let s denote a typical sample of size n which was chosen by simple random sampling without replacement. We assume simple random sample for convenience, since in many problems of this type the sampling will often be more purposeful. Before considering the problem of estimating the median of the population we review some well known facts about the problem of estimating the mean.

Consider the super population model where it is assumed that for each i, $y_i = bx_i + u_i e_i$. Here b is an unknown parameter while the u_i's are known constants and the e_i's are independent identically distributed random variables with zero expectations. Since the population mean can be written as $N^{-1}(\sum_{i\epsilon s} y_i + \sum_{j\notin s} y_j)$ we would expect $N^{-1}(\sum_{i\epsilon s} y_i + \hat{b}\sum_{j\notin s} x_j)$ to be a sensible estimate of the mean whenever $\hat{b}$ is a sensible estimate of b. One particular choice of $\hat{b}$ is the weighted least squares estimator where the weights are determined by the u_i's. For example if for all i, $u_i = \sqrt{x_i}$, the resulting estimator is just the usual ratio estimator. While if for all i, $u_i = x_i$, then $\hat{b} = n^{-1}\sum_{i\epsilon s}(y_i/x_i)$ and the resulting estimator is one that was discussed by Basu (1971). (See also Royall (1970).) Using this super population setup it is easy to generate populations where the ratio estimator has smaller mean squared error than the Basu estimator and vice versa. A somewhat limited simulation study on a variety of populations found that the performance of the Basu estimator is quite similar to the performance of the ratio estimator although in the majority of the cases the ratio estimator performs better than the Basu estimator. This is not unexpected, given the wide use of the ratio estimator.

[1] Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

In Meeden and Vardeman (1991) a non-informative Bayesian approach to finite population sampling, based on the Polya posterior, was developed. For the simple problem where no auxiliary variable is present, given the observed values in the sample, it introduces a Polya urn distribution as a pseudo posterior distribution over the unobserved members of the population. This pseudo posterior distribution can be used to obtain point and interval estimates of a variety of population quantities. It is related to the Bayesian bootstrap of Rubin (1981) and the Dirichlet process prior of Ferguson (1973). When estimating the median it yields results similar to those of Binder (1982). A theoretical justification for it is a stepwise Bayes argument which yields the admissibility of the resulting estimators. See for example Meeden and Ghosh (1983). There the admissibility of the Basu estimator was demonstrated. In that case the Basu estimator was shown to arise from a 'posterior' which treats the known and unknown ratios, $r_i = y_i/x_i$ as exchangeable. Note that this is very similar in spirit to the super population model justification for this estimator given above, where the ratios $r_i = y_i/x_i$ were independent and identically distributed. We shall see that the stepwise Bayes logic underlying the Basu estimator for the mean carries over in a straight forward way to point and interval estimators for the median. Unfortunately this is not the case for some of the other estimators. One natural but perhaps naïve estimator which mimics in some sense the ratio estimator of the mean is just the ratio of the median of the y values in the sample to the median of the x values in the sample multiplied by the median of the x values in the population. There is no known model based theory which underlies this estimator as is the case for the ratio estimator of the mean.

In the Bayesian approach to finite population sampling one needs to specify a prior distribution. Then given a sample, inferences are based on the posterior distribution, which is the predictive distribution for the unseen members of the population given the units in the sample. In the stepwise Bayes approach, given the sample one always has a 'posterior' distribution but it does not arise from a single prior distribution. However this 'posterior' distribution can be used in the usual Bayesian manner to find point and interval estimators of parameters of interest. We now will show how the stepwise Bayes model which yields Basu's estimator for the mean can also be used when estimating the median. In this setup, given a sample, the predictive distribution for the unobserved ratios treats the observed and unobserved ratios as 'exchangeable'.

For definiteness suppose our sample contains the first n units of the population. We construct an urn which contains n balls where ball i is given the value of the i-th observed ratio, say r_i. We begin by selecting a ball at random from the urn and the observed value is assigned to the unobserved unit $n + 1$. This ball and an additional ball with the same value is returned to the urn. Another ball is chosen from the urn and its value is assigned to the unobserved unit $n + 2$. This ball and another with the same value are returned to the urn. This process is continued until all of the unobserved units have been assigned a ratio. Once they have all been assigned a value we have observed one realization from our 'posterior' distribution for the unseen ratios given the sample of seen ratios. If in this process the unobserved unit j has been assigned the ratio with value r we then assign its y_j value to be rx_j. Hence using simple Polya sampling we have created a predictive distribution for the unobserved units given the sample. We call this predictive distribution the 'Polya posterior'. It is easy to check that this predictive distribution gives the Basu estimator when estimating the population mean under squared error loss.

Given the sample the 'Polya posterior' yields a predictive distribution for the unobserved members of the population and hence a predictive distribution for the median as well. From the decision theory point of view the usual loss function is absolute error when estimating a median. For this loss function the Bayes estimate is just the median of the posterior or predictive distribution for the population median. If one were using squared error loss for estimating the median then the Bayes estimate is just the mean of the predictive distribution for the population median. The admissibility of these estimators under the appropriate loss function follows from a stepwise Bayes argument in the same way as the proof of admissibility for the Basu estimator of the population mean. In Meeden and Vardeman (1991) and Meeden (1993) the following somewhat surprising fact was noted. For many common distributions the mean of the predictive distribution for the population median performed better than the median of the predictive distribution for the population median under both loss functions. Similar results hold for this problem. Hence our estimator will be the mean of the predictive distribution for the population median even though we will follow standard practice and use absolute error as our loss function. We will denote this estimator by *estpp*. This estimator cannot be found explicitly. However we will find it approximately by simulating observations from the posterior or predictive distribution for the population median. Under the Polya sampling scheme for the ratios described above we can simulate a possible realization of the entire population. For this simulated copy we can then find its median. If we repeat this process R times then we have simulated the predictive distribution of the population median under the 'Polya posterior'. When R is large the mean of these R simulated population medians yields, approximately, the estimate *estpp*.

In what follows we will compare the estimator *estpp* to several other estimators. Another estimator we consider is just the sample median of the y_i's. This ignores the information contained in the auxiliary variable and is used as a bench mark. It will be denoted by *estsm*. Another estimator is the natural analogue of the ratio estimator of the population mean. This is discussed in Kuk and Mak (1989) and denoted by *estrm*. It is just the ratio of the median of the y values to the median of the x values in the sample multiplied by the median of all the x values in the population. They proposed two other estimators for the median. We will consider just the first one and denote it by *estkm*. This estimator has a plausible intuitive justification and can be found in their paper. Rao, Kovar and Mantel (1990) considered a designed based estimator for the median. We will denote this estimator by *estrkm*. Since this estimator can be time consuming to compute we will find it approximately using a method due to Mak and Kuk (1993). Finally we will consider the estimator proposed in Chambers and Dunstan (1986) and denote it by *estcd*. Actually Chambers and Dunstan propose a whole family of estimators and we will only consider one special case which is appropriate when $u_i = \sqrt{x_i}$ in the super population model described at the beginning of this section. We now briefly outline the argument that leads to their estimator of the median. Let F denote the cumulative distribution function associated with the y values of the population. That is F puts mass $1/N$ on each y_i in the entire population. The first step is to get an estimator of $F(t)$ for an arbitrary real number t. If s denotes our sample of size n then given the sample we can write

$$F(t) = N^{-1}\left\{\sum_{i\epsilon s} \Delta(t - y_i) + \sum_{j\notin s} \Delta(t - y_j)\right\}$$

where $\Delta(z)$ is the step function which is one when $z \geq 0$ and zero elsewhere. Since the first sum in the above expression is known once we have observed the sample, to get an estimate of $F(t)$ it suffices to find an estimate of the second sum. Now under our assumed super population model the population ratios $(y_i - bx_i)/\sqrt{x_i}$ are independent and identically distributed random variables. Since after the sample s is observed a natural estimate of b is $\hat{b} = \sum_{i\epsilon s}y_i / \sum_{i\epsilon s}x_i$ one could act as if the n known ratios $(y_i - \hat{b}x_i)/\sqrt{x_i}$ for $i \in s$ are actual observations from this unknown distribution. Under this assumption, for a fixed t and a fixed unit j not in the sample s an estimate of $\Delta(t - y_j)$ is just the number of the n known ratios incorporating $\hat{b}$ less than or equal to $(t - \hat{b}x_j)/\sqrt{x_j}$ divided by n. Finally if we sum over all the unobserved units j these estimates of $\Delta(t - y_j)$ we then have an estimate for the second sum in the above expression for $F(t)$ which then yields an estimate of $F(t)$. Once we can estimate $F(t)$ for any t by say $\hat{F}(t)$ then the estimate of the population median is $\inf\{t:\hat{F}(t) \geq 0.5\}$.

3. THE POPULATIONS

We will compare these estimators using several different populations. We begin with three actual populations. The first is a group of 125 American cities. The x variable is their 1960 populations, in millions, while their y variable is the corresponding 1970 populations, again in millions. The second is a group of 304 American counties. The x variable is the number of families in the counties in 1960, while the y variable is the total 1960 population of the county. Both variables are given in thousands. The third population is 331 large corporations. The x variable is their total sales in 1974 and the y variable their total sales in 1975. The sales are given in billions of dollars. We denote these three populations by *ppcities*, *ppcounties* and *ppsales*. For the three populations the correlations are .947, .998 and .997. These populations were discussed in Royall and Cumberland (1981). Our *ppcounties* is similar to their population Counties60 except we have taken the x variable to be the number of families rather than the number of households.

We have also considered six artificial populations. In each case the auxiliary variable x was chosen first and then the y variable was generated from it. In some cases we followed the super population model described at the beginning of the previous section for some choice of the u_i's. In some other cases we violated the assumption that conditional on the value x_i the mean of y_i is bx_i. In all cases the errors, the e_i's, were independent and identically distributed normal random variables with mean zero and variance one.

In the first population, *ppgamma20*, the x_i's were a random sample from a gamma distribution with shape parameter twenty and scale parameter one. Then given x_i the conditional distribution of y_i was normal with mean $1.2x_i$ and variance x_i, *i.e.*, $u_i = \sqrt{x_j}$.

In the second population, *ppgamma5a*, the x_i's were ten plus a random sample from a gamma distribution with shape parameter five and scale parameter one. Then given x_i the conditional distribution of y_i was normal with mean $3x_i$ and variance x_i.

In *ppgamma5b* the auxiliary variable was the same as in *ppgamma5a*. Then given x_i the conditional distribution of y_i was normal with mean $3x_i$ and variance x_i^2.

In *ppstskew* the auxiliary variable was strongly skewed to the right with mean 42.63, median 39.29 and variance 204.59. Then given x_i the conditional distribution of y_i was normal with mean $x_i + 5$ and variance $9x_i$.

In *ppln* the auxiliary variable was a random sample from a log-normal population with mean and standard deviation (of the log) 4.9 and .586 respectively. Then given x_i the conditional distribution of y_i was normal with mean $x_i + 2 \log x_i$ and variance x_i^2.

In *ppexp* the auxiliary variable was fifty plus a random sample from the standard exponential distribution. Then given x_i the conditional distribution of y_i was normal with mean $80 - x_i$ and variance $(.6 \log x_i)^2$.

All the populations contain 500 units except *ppstskew* which has 1,000. The correlations between the two variables for these last six populations are .76, .87, .41, .61, .58 and −.28 respectively.

In most examples where ratio type estimators are used both the y_i's and x_i's are usually strictly positive. In population *ppstskew* 13 of the 1,000 units have a y value which is negative. In the original construction of population *ppln* quite a few more of the y values were negative. The population was modified so that all the values are greater than zero.

Note that these populations were constructed under various scenarios for the relationship between the x and y variables. *Ppgamma20* and *ppgamma5a* satisfy the assumptions of the super population model leading to *estcd*, while *ppgamma5b* is consistent with the assumptions underlying *estpp*. In *ppstskew* the conditional variance of y_i given x_i is consistent with *estcd* while for the unmodified *ppln* it was consistent with *estpp*. In both these cases the assumption for the conditional expectation is not satisfied. For the populations *ppcounties*, *ppgamma5a* and *ppln* we have plotted y against x and y/x against x. The results are seen in Figures 1 through 3.

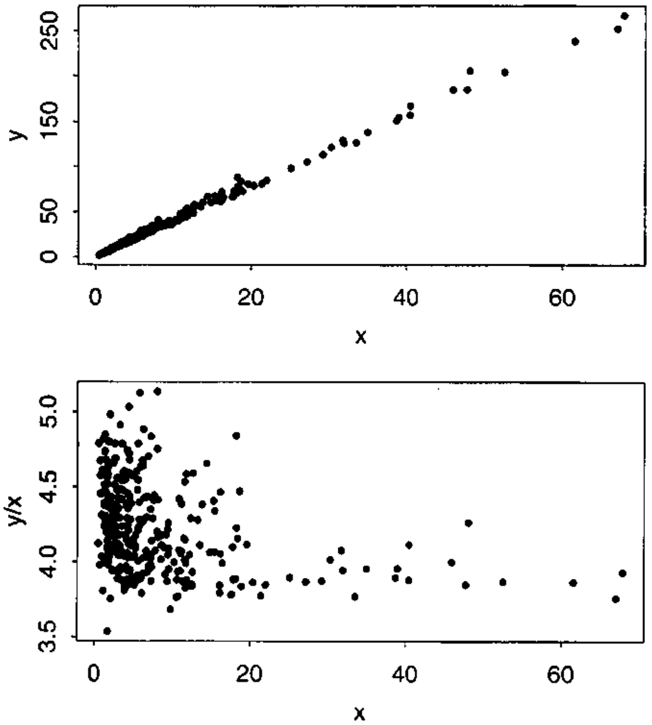

Figure 1. For *ppcounties* the plot of y versus x and y/x versus x where x is the number of families (thousands) living in a county and y is the total population (thousands) of the county for 304 counties.

The estimator *estpp* is based on the assumption that given the sample s our beliefs about the observed ratios, *i.e.*, the ratios y_i/x_i for $i \in s$ and the unobserved ratios, *i.e.*, the ratios y_j/x_j for $j \notin s$ are roughly exchangeable. In particular this means that one's beliefs about a ratio y_j/x_j

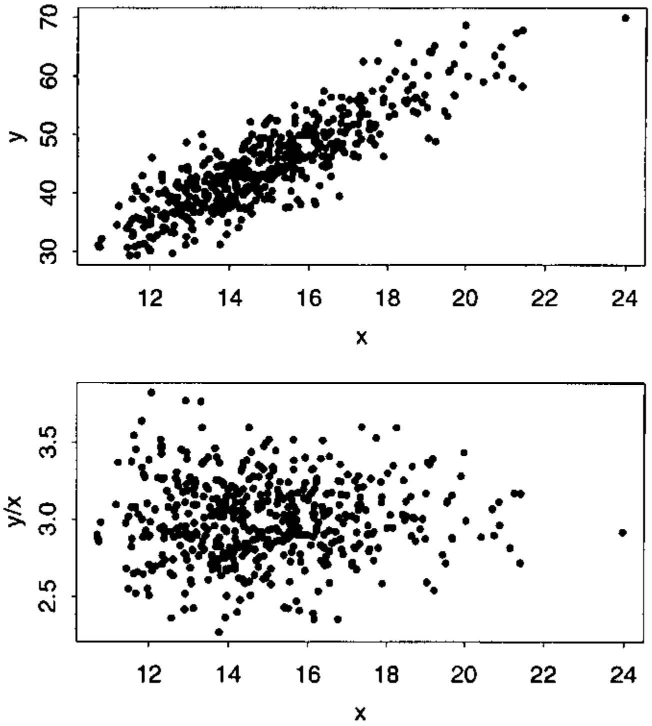

Figure 2. For *ppgamma5a* the plot of y versus x and of y/x versus x.

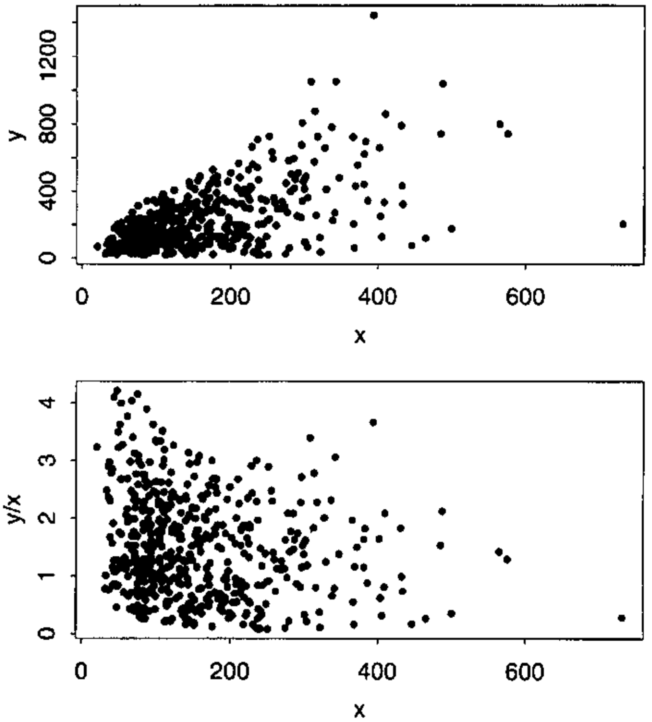

Figure 3. For *ppln* the plot of y versus x and of y/x versus x.

should not depend on the size of x_j. In fact *ppgamma5b* was constructed so that this would indeed be true. On the other hand, under the super population model leading to the estimator *estcd* we would expect the variability of the ratios to get smaller as the size of the x variable increases while the average value of the ratios in any thin vertical strip remains roughly constant as the strip moves to the right. This is seen clearly in the plot of the ratios for population *ppgamma5a*. For the rest of the populations, except for *ppgamma20* the values of the ratios do in fact depend on size of x. This is seen clearly in the plots for *ppcounties* and *ppln*. Hence they should make interesting test cases for the estimator *estpp*. *Ppexp* was included as a test case to see what would happen if the underlying assumptions of *estpp* and *estcd* were strongly violated.

4. SOME SIMULATION RESULTS

To compare the six estimators 500 simple random samples of various sizes were taken from the nine populations. For each sample the values of the six estimators were computed. For the estimator *estpp* this meant finding it approximately by simulating $R = 500$ realizations of the predictive distribution for the population median induced by the 'Polya posterior'. In each case the average value and average absolute error of the estimator were computed. In Table 1 the average values of all the estimators except *estsm* are given. All the estimators are approximately unbiased except in one case, *estcd* for the population *ppln*. We did not include the results for *estsm* since it is well known that it is unbiased. In Table 2 the average absolute error for all six estimators are given. We see from Table 2 that *estcd* and *estpp* are the clear winners. They both perform better than the other four estimators in every case but one. In *ppexp* they are both beaten by *estsm*, but this is one case where neither would be expected to do well. For the first seven populations their performances are nearly identical while for population *ppln* the estimator *estpp* is preferred and for population *ppstskew* the opposite is true.

In practice one often desires interval estimates as well as point estimates for parameters of interest. Kuk and Mak (1989) and Chambers and Dunstan (1986) each suggested possible methods for finding interval estimates based on their estimator using asymptotic theory. But in each case they did not actually find any interval estimators. Meeden and Vardeman (1991) noted how approximate 95% credible regions based on the 'Polya posterior' can be found approximately. If we let $q(.025)$ and $q(.975)$ be the .025 quantile and the .975 quantile of the collection of 500 simulated population medians under the 'Polya posterior' then $(q(.025), q(.975))$ is an approximate 95% credible interval. (See Berger 1985 for the definition of such intervals.) Table 3 gives the average length and relative frequency of coverage for these intervals. We see that for these populations the intervals have reasonable frequentist properties. Perhaps this is not unexpected given the discussion in Meeden and Vardeman (1991). But on the other hand only one of the populations was constructed so that the ratios y_i/x_i are exchangeable. These results suggest that point and interval estimators of the median based on the 'Polya posterior' for the ratios are fairly robust against the exchangeability assumption and should work well in a variety of situations. This will be discussed further in section 5.

Table 1

The Average Value of Five Estimators of the Median for 500 Simple Random Samples

Population (median)	Sample Size	Average Value of the Estimator				
		estrm	*estkm*	*estrkm*	*estcd*	*estpp*
ppcities (1.90)	25	.197	.196	.193	.195	.195
ppsales (1.24)	30	1.21	1.25	1.23	1.25	1.24
ppcounties (18.33)	30	18.21	18.60	18.66	18.26	18.39
ppexp (29.02)	30	29.03	29.05	29.00	29.03	29.05
ppgamma5a	30	43.82	43.88	43.91	43.99	43.89
(43.90)	50	43.90	43.91	43.85	44.06	43.90
ppgamma5b	30	43.84	43.96	44.19	44.15	43.61
(44.17)	50	44.28	44.37	44.18	44.18	43.98
ppgamma20	30	23.47	23.28	23.14	23.46	23.77
(23.15)	50	23.34	23.18	23.17	23.43	23.18
ppln	30	171.15	169.38	168.12	185.01	170.61
(170.25)	50	169.15	167.54	167.65	185.03	169.61
ppstskew	30	43.66	40.27	45.88	45.50	45.11
(46.12)	50	44.04	40.70	46.01	45.43	45.37

Table 2

The Average Absolute Error of Six Estimators of the Median for 500 Simple Random Samples

Population	Sample Size	Average Absolute Error of the Estimator					
		estsm	*estrm*	*estkm*	*estrkm*	*estcd*	*estpp*
ppcities	25	.0326	.0161	.0162	.0155	.0075	.0072
ppsales	30	.1797	.0770	.0797	.0870	.0244	.0245
ppcounties	30	3.12	.586	.964	1.34	.215	.214
ppexp	30	.43	.49	.48	.47	.48	.46
ppgamma5a	30	1.36	.96	1.03	.89	.54	.53
	50	.95	.74	.78	.65	.44	.43
ppgamma5b	30	2.84	2.74	2.71	2.58	2.37	2.38
	50	2.08	2.04	2.01	1.89	1.80	1.85
ppgamma20	30	1.08	1.06	1.05	.88	.67	.64
	50	.94	.77	.78	.73	.51	.49
ppln	30	25.9	25.8	24.2	21.62	21.4	17.0
	50	18.0	20.1	17.9	16.46	17.7	12.7
ppstskew	30	3.86	4.26	6.69	3.21	2.72	3.14
	50	2.92	3.63	5.82	2.55	2.20	2.51

Table 3

The Average Length and Relative Frequency of Coverage for a .95 Credible Interval for the Median Based on the 'Polya Posterior' for 500 Simple Random Samples

Population	Sample Size	Average Length	Frequency of Coverage
ppcities	25	.041	.968
ppsales	30	.141	.964
ppcounties	30	1.44	.994
ppexp	30	2.26	.944
ppgamma5a	30	2.70	.950
	50	2.15	.956
ppgamma5b	30	11.67	.932
	50	8.86	.942
ppgamma20	30	3.24	.960
	50	2.51	.964
ppln	30	84.8	.934
	50	65.4	.956
ppstskew	30	15.52	.936
	50	12.00	.938

5. DISCUSSION

The motivation for the estimator *estpp* is based on the assumption that the population ratios y_i/x_i's are exchangeable. This assumption can be described mathematically in two separate but related ways. The first is the super population model given earlier while the second comes from the 'Polya posterior' which is based on a stepwise Bayes argument and gives a non-informative Bayesian interpretation for the estimator. This second approach can be used no matter what parameter is being estimated. When estimating the mean it leads to Basu's estimator which performs very much like the ratio estimator although the ratio estimator usually does a bit better. When estimating the median it leads to the estimator discussed in this note. Here we have argued that the 'Polya posterior' for the ratios leads to good point and interval estimators for the median when an auxiliary variable is present and seems to be reasonably robust against the assumption that the ratios y_i/x_i's are exchangeable.

Royall and Cumberland (1981) gave an empirical study of the ratio estimator and estimators of its variance. They argued that given a sample an estimate of variance based on the super population model, which leads to the ratio estimator, often made more sense than a design based estimate based on a probability sampling distribution. In Royall and Cumberland (1985), they demonstrated that, conditional on the sample mean of the auxiliary variable, the conditional coverage properties of the usual designed based confidence interval for the population mean were 'hopelessly unreliable'.

We now wish to address the question of the conditional behavior of the intervals for the median based on the Polya posterior which were developed in this note. In the simulation studies given earlier simple random sampling was used for convenience. To get some idea of the conditional behavior of the 'Polya posterior' we considered five of our populations. In each case we ordered the population using the values of the auxiliary variable x. We then took 500 random samples from the first or smallest half of the population, then 500 more random samples from the second or largest half of the population and finally 500 more random samples from the middle third of the population. We then calculated the .95 credible interval for the median based on the 'Polya posterior' which assumes the exchangeability of the ratios y_i/x_i's. In Table 4 we give the results for the 'Polya posterior' estimators for the median. (We also computed the average value and average absolute error of *estcd* for these examples. We did not include these results since they match closely the results of the 'Polya posterior'.) We see that their conditional behavior, at least in these cases, is very much like their unconditional behavior. In short, interval estimates for the median based on the 'Polya posterior' should have reasonable frequentist properties, no matter how the sample was selected, as long the population approximates our beliefs that the ratios are roughly exchangeable.

Table 4

The Average Value and Absolute Error for the Point Estimator and the Average Length and Relative Frequency of Coverage for a .95 Credible Interval for the Median Based on the 'Polya Posterior' for 500 Simple Random Samples from the whole Population, the 'Smallest' Half, the 'Largest' Half and the 'Middle' Third

Population	Sample Size	Where Taken	Average Value	Average Error	Average Length	Frequency of Coverage
ppcities	25	whole	.195	.0072	.041	.968
		smallest ½	.192	.0047	.033	.994
		largest ½	.196	.0078	.048	.988
		middle ⅓	.201	.0114	.055	.922
ppcounties	30	whole	19.4	.220	1.46	.990
		smallest ½	18.6	.305	1.34	.942
		largest ½	18.1	.283	1.59	.954
		middle ⅓	18.5	.252	1.35	.964
ppsales	30	whole	1.24	.0072	.141	.964
		smallest ½	1.24	.027	.153	.966
		largest ½	1.23	.020	.125	.982
		middle ⅓	1.23	.027	.139	.944
ppgamma5a	30	whole	43.9	.53	2.70	.950
		smallest ½	43.8	.55	2.82	.948
		largest ½	44.0	.53	2.55	.940
		middle ⅓	43.9	.47	2.63	.974
ppgamma5b	30	whole	43.6	2.38	11.7	.932
		smallest ½	42.2	2.69	11.6	.890
		largest ½	45.1	2.25	11.2	.950
		middle ⅓	45.2	2.27	11.3	.936

As can be seen by looking at the plots of y_i/x_i versus x_i and our simulation results it does not seem to matter much if the variability in the ratios y_i/x_i's decreases as x_i increases. What is crucial however is that the average value of the ratios in the narrow strip above a small interval of possible x values remains fairly constant as we move the small interval to the right. In Figure 2, the plot of the ratios for *ppgamma5a* is an example of such a plot. In fact this is how the population was constructed, since it satisfies the assumptions underlying *estcd*. In Figures 1 and 3 we see for *ppcounties* and *ppln* that the average value of the ratios in a narrow strip tends to decrease as we move to the right and helps to explain the relatively poorer performance of the 'Polya posterior' estimators in these cases. Overall however, the performance of procedures based on the 'Polya posterior' seem to be reasonably robust against the exchangeability assumption.

As another alternative we could consider a more balanced sampling plan which is based on stratifying the population on the auxiliary variable. For example consider again population *ppgamma5b* where it is ordered on the basis of its x_i values. We constructed ten strata where the first stratum consisted of the units with the fifty smallest x_i values, the second stratum of the units with the next fifty smallest x_i values and so on. We then took 500 stratified random samples of size fifty where five units were chosen at random from each stratum. For these samples the average value of *estpp* was 43.94 and its average absolute error was 1.81. The average length of its corresponding interval estimator was 8.95 with .938 relative frequency of covering the true value. Note that these figures are very similar to those given Tables 1 and 2 when simple random sampling was used.

ACKNOWLEDGEMENTS

Research supported in part by NSF grant SES 9201718.

REFERENCES

BASU, D. (1971). An essay on the logical foundations of survey sampling, part one. In *Foundations of Statistical Inference*. Toronto: Holt, Reinhart and Winston, 203-242.

BERGER, J.O. (1985). *Statistical Decision and Bayesian Analysis*. New York: Springer-Verlag.

BINDER, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B*, 44, 388-393.

CHAMBERS, R.L., and DUNSTAN, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

FERGUSON, T.S. (1973). A Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1, 209-230.

KUK, A.Y.C., and MAK, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, 51, 261-269.

MAK, T.K., and KUK, A. (1993). A new method for estimating finite-population quantiles using auxiliary information. *The Canadian Journal of Statistics*, 21, 29-38.

MEEDEN, G., and GHOSH, M. (1983). Choosing between experiments: applications to finite population sampling. *Annals of Statistics*, 11, 296-305.

MEEDEN, G., and VARDEMAN, S. (1991). A noninformative Bayesian approach to interval estimation in finite population sampling. *Journal of the American Statistical Association*, 86, 972-980.

MEEDEN, G. (1993). Noninformative nonparametric Bayesian estimation of quantiles. *Statistics and Probability Letters*, 16, 103-109.

RAO, J.N.K., KOVAR, J.G., and MANTEL, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

ROYALL, R.M., and CUMBERLAND, W.D. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

ROYALL, R.M., and CUMBERLAND, W.D. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

RUBIN, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

Survey Methodology, June 1995
Vol. 21, No. 1, pp. 79–87
Statistics Canada

Outlier Robust Horvitz-Thompson Estimators

BEAT HULLIGER[1]

ABSTRACT

The Horvitz-Thompson estimator (HT-estimator) is not robust against outliers. Outliers in the population may increase its variance though it remains unbiased. The HT-estimator is expressed as a least squares functional to robustify it through M-estimators. An approximate variance of the robustified HT-estimator is derived using a kind of influence function for sampling and an estimator of this variance is developed. An adaptive method to choose an M-estimator leads to minimum estimated risk estimators. These estimators and robustified HT-estimators are often more efficient than the HT-estimator when outliers occur.

KEY WORDS: Outlier; M-estimator; Adaption; Population mean; Sampling; Sensitivity curve.

1. INTRODUCTION

The mean of a variable over a finite population is an important indicator. Examples are the mean salary of employees in a branch of the economy or the mean yield of corn of the farms in a region. Due to its connection to the sum the mean cannot be easily replaced by other indicators. But the population mean is a sensitive characteristic because a single large observation may determine its value. The Horvitz-Thompson estimator (HT-estimator) is a natural estimator of the population mean if the sample design has unequal inclusion probabilities and is without replacement. It is the sample mean in simple random sampling. It is always unbiased whatever the population distribution of the investigated variable is. But the HT-estimator is not robust against outliers because it is linear in the observed values like its estimand, the population mean. Large observations together with small inclusion probabilities have a particularly large influence on the HT-estimator.

Suppose there is an outlier in a sample. The outlier may be a correct observation from the target population. Discarding such a correct outlier makes the HT-estimator biased. But keeping it with full weight makes the HT-estimator highly variable because typically the outlier would show up only in a few of the possible samples. Thus there is a tradeoff between bias and variance in this case, which in particular includes asymmetric distributions with one heavy tail.

The outlier may also be an incorrect observation, *e.g.*, due to a measurement or coding error or stemming from an element outside the target population. In that case keeping the outlier with full weight may entail a large bias of the HT-estimator in addition to high variability. Thus discarding incorrect outliers reduces both bias and variance.

Since it is often difficult to detect outliers and to decide whether it is correct or not one would like to have estimators that perform well in terms of bias and variance irrespective of the nature and the detection of possible outliers. HT-estimators which are robustified through M-estimators are promising candidates for this difficult task.

In the survey sampling literature the problem of outliers or aberrant values is often treated under the heading "skew populations". Kish (1965, sec. 11.4 B) describes the problem in economical surveys and surveys of individuals. He proposes the formation of separate strata for outliers if possible, truncation, transformation or modelling. The idea of forming a separate class for large units and combining the class means is investigated for example in (Glasser 1962) and (Hidiroglou and Srinath 1981).

The truncation idea is made more precise by the winsorized mean proposed by Searls (1966). Fuller (1991) proposed a preliminary-test-estimator which reduces the impact of the largest data values only when a test for extreme values is significant. Rivest (1993) studied the behavior of various winsorization schemes under simple random sampling. Shoemaker and Rosenberger (1983) derive exact formulae for the expected value and variance of the median and trimmed mean under simple random sampling without replacement. Oehlert (1985) proposes the random average mode estimator to estimate the mean of finite populations in an outlier robust way. Smith (1987) emphasises that it is as important to detect and treat influential observations if the inference is based on the randomisation provided by the sample design as if the observations are considered realisations of random variables. He proposes an influence measure for linear estimators based on case deletion, which involves both the variable of interest and its weight.

The prediction approach in sampling theory uses stochastic models for the population to predict the total of the present realisation. Linear models and (nonrobust) linear estimators are used. Aspects of the sensitivity and robustification against model misspecification are reviewed

[1] Beat Hulliger, Swiss Federal Statistical Office, Schwarztorstrasse 96, CH-3003 Bern, Switzerland.

in (Iachan 1984). Chambers (1986) develops an outlier-robustification of the prediction approach using M-estimators. He distinguishes representative and nonrepresentative outliers in a sample. Representative outliers must be included with full weight in an unbiased estimate of the population mean while nonrepresentative outliers should be downweighted or discarded.

Little and Smith (1987) treat outliers and missing data in certain positive multivariate continuous data by a robustified EM-algorithm. Gwet and Rivest (1992) investigate resistant ratio estimators under simple random sampling without replacement.

M-estimators form a class of flexible and simple robust estimators. An M-estimator T of location is defined implicitely by the estimating equation

$$\sum_{i=1}^{n} \psi(X_i - T) = 0$$

for a predetermined function ψ, *e.g.*, $\psi_{\mathrm{Hub}}(x,k) = \max(-k,\min(k,x))$, where k is a tuning constant. An M-estimator may be written as a functional of the empirical distribution function. The influence function of an estimator is a functional derivative of the estimator (Hampel 1974). It describes the reaction of the estimator to a small contamination in the data. An M-estimator with bounded ψ-function usually has a bounded influence function such that outliers cannot disturb the estimator too much. For the estimation of the mean of asymmetric finite populations M-estimators must be adapted.

In this article we develop design-based M-estimators for samples with unequal inclusion probabilities. The simple linear model which implicitely is the basis of the Horvitz-Thompson strategy is made explicit and the HT-estimator is expressed as a functional of an empirical distribution function which accounts for the complex sample design. This establishes the link to classical robust statistics and allows a straightforward robustification of the HT-estimator (Section 2). We define an influence function for sampling which clarifies the outlier-sensitivity of the HT-estimator and leads to an approximation of the sampling variance of the robustified HT-estimator. An estimator of this variance is presented. In Section (3) we briefly comment on stratification, domains, robust designs and one-step estimators. In Section (4) an adaptive robustification of the HT-estimator is developed. The method chooses from a class of robustified HT-estimators the one which minimizes an estimate of the mean squared error. The resulting estimator is called minimum estimated risk estimator (MER-estimator). A Monte-Carlo simulation is presented in Section 5. Robustified HT-estimators and MER-estimators outperform the HT-estimator in many outlier situations. The premium to pay is a moderate loss of efficiency in situations where the HT-estimator is optimal.

2. ROBUSTIFICATION OF HORVITZ-THOMPSON ESTIMATORS

2.1 The HT-Estimator as a Least Squares Functional

A finite population $U = \{1, \ldots, N\}$ of $0 < N < \infty$ distinct elements is sampled. We are interested in a variable y which takes the values y_i for $i \in U$. The sample design $p(S)$ on the space of samples S of fixed size n has inclusion probabilities $\pi_i = \mathrm{P}[i \in S] = \sum_{S \ni i} p(S)$. These π_i are proportional to some known positive auxiliary variable $x_i\,(i \in U)$. Such sample designs are called IPPS designs (inclusion probability proportional to size) because often x_i is some size measure. Denote by π_{ij} the joint inclusion probability $\mathrm{P}[i \in S, j \in S]$ $(i, j \in U)$. The vector of all y-values is denoted $y_U :\, = (y_1, \ldots, y_N)^T$ and x_U is defined in an analogous way. The vector of the y-values of a sample S is denoted $y_S :\, = (y_{i_1}, \ldots, y_{i_n})^T$ $(i_k \in S)$. The goal is to estimate the population mean of the variable $y : \bar{y}_U :\, = \sum_{i \in U} y_i / N$.

The HT-estimator for $\bar{y}_U$ is $T_{HT} :\, = \sum_{i \in S} y_i / (N\pi_i)$. The variance of T_{HT} is estimated by the well known estimator

$$v_{HT}(T_{HT}) =$$

$$\frac{1}{N^2}\left[\sum_{i \in S} (1 - \pi_i)\,\frac{y_i^2}{\pi_i^2} + \sum_{i \neq j \in S} (1 - \pi_i \pi_j / \pi_{ij}) \frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\right], \tag{1}$$

which is due to Horvitz and Thompson or by the variance estimator due to Yates, Grundy and Sen (see Cochran 1977, p. 261).

The rationale behind the HT-estimator given in the survey sampling literature is that it has sampling variance zero if the inclusion probabilities π_i are exactly proportional to y_i. Then $T_{HT}(y_S) = \bar{y}_U$ for every sample S. The HT-estimator is bias-robust but not variance-robust with respect to deviations from proportionality between y_i and π_i (*cf.* Rao 1966).

How can the HT-estimator be formulated in a way which allows the derivation of an influence function analogue and a variance estimator? The key idea is to express the HT-estimator as a least squares (LS) functional of an estimate of the population distribution function in such a way that the design is incorporated in the estimator of the population distribution function while the proportionality of y_i and x_i is taken up by the LS-functional.

The joint population distribution function of two variables (x_i, y_i) is defined as $F_U(r,t) = \sum_{i \in U} \mathbf{1}\{x_i \leq r\}$ $\mathbf{1}\{y_i \leq t\}/N$, where $\mathbf{1}\{y_i \leq t\} = 1$ if $y_i \leq t$ and 0 elsewhere. There are various possibilities to estimate F_U but the easiest and most generally applicable estimator is the sample distribution function

$$F_S(r,t) = \sum_{i\epsilon S} \frac{1}{\pi_i} 1\{x_i \le r\} 1\{y_i \le t\} \Big/ \sum_{i\epsilon S} \frac{1}{\pi_i}. \qquad (2)$$

The estimator F_S is a distribution function itself.

To derive a LS-functional the following superpopulation model for the proportionality between y_i and x_i is used: We assume that y_U is a vector of realisations of independent random variables Y_i with expectation βx_i and variance $\sigma^2 x_i$.

Definition 1. The LS-estimator $\beta_{LS}(F_S)$ of β in the above model with respect to the sampling distribution function F_S of (x_i,y_i) $(i \in S)$ minimizes $\int (y - \beta x)^2/x \, dF_S(x,y)$ or equivalently solves

$$\sum_{i\epsilon S} \frac{1}{\pi_i} \left(\frac{y_i - \beta x_i}{\sqrt{x_i}} \right) \frac{x_i}{\sqrt{x_i}} = 0. \qquad (3)$$

The following statement is well known and its proof is easy. If S is a sample drawn according to an IPPS sample design with inclusion probabilities $\pi_i = nx_i/\sum_{i\epsilon U} x_i$ $(i \in U)$ then the HT-estimator is $T_{HT} = \bar{x}_U \beta_{LS}(F_S)$, where $\beta_{LS}(F_S)$, the LS-estimator defined by (3):, is given by

$$\beta_{LS}(F_S) = \frac{\sum_{i\epsilon S} y_i/\pi_i}{\sum_{i\epsilon S} x_i/\pi_i}.$$

Note that the expression $T_{HT} = \bar{x}_U \beta_{LS}(F_S) = \bar{x}_U (\sum_{i\epsilon S} y_i/\pi_i)/(\sum_{i\epsilon S} x_i/\pi_i)$ does not depend on the superpopulation model. However the superpopulation model clarifies the role of the HT-estimator: The slope $\beta_{LS}(F_S)$ involved in the HT-estimator is a weighted least squares estimator that incorporates the information in the design through F_S as well as the information in the auxiliary variable through the regression.

2.2 The Robustified HT-Estimator

After the separation of design and auxiliary information and its expression as a LS-functional the robustification of the HT-estimator is analogous to the robustification of LS-estimators in linear models for infinite populations through M-estimators (*cf.* Hampel *et al.* 1986, Chapter 6): The estimating equation (3) now involves some function η which depends on the standardized residuals $(y_i - \beta x_i)/x_i^{1/2}$ and on x_i. For ease of notation denote by a prime the division by $x^{1/2}$ and let $r'(\beta) = (y - \beta x)/x^{1/2}$.

Definition 2. Let $\beta(F_S,\eta)$ be a solution of the equation

$$\sum_{i\epsilon S} \frac{1}{\pi_i} \eta(x_i', r_i'(\beta)) x_i' = 0. \qquad (4)$$

The robustified HT-estimator (RHT-estimator) is

$$T_{RHT}(F_S) := \bar{x}_U \beta(F_S,\eta).$$

$\beta(F_S,\eta)$ is called the slope of the RHT-estimator.

In general useful choices of η are of the form $\eta(x,r) = w(x)\psi(r \cdot u(x))$, where $w(x)$ and $u(x)$ are two weighting functions and ψ is a defining function for a location M-estimator (*cf.* Hampel *et al.* 1986, p. 315). In the following we use the so-called Mallows form, which sets $u(x) \equiv 1$. Mallows-type regression downweights outlying x-values and outlying residuals independently. A well-known example, which also sets $w(x) \equiv 1$, is the Huber-function $\eta(x,r) = \psi_{\text{Hub}}(r,k) = \max(-k, \min(k,r))$ for some constant k. The RHT-estimator with defining function $\eta(x,r) \equiv r \; \forall \; x$ is the HT-estimator. Thus by adjusting the tuning constant k in the Huber-function a smooth transition of estimators from the HT-estimator to more and more robust estimators is possible.

Scale estimates are needed in $w(x)$ and $\psi(r)$ to make $\beta(F_S,\eta)$ scale equivariant. While for the weighting function $w(x_i')$ preliminary scale estimators are available, *e.g.*, the median of the x_i', the scale of the residuals must be estimated simultaneously with the slope β. The median of the absolute residuals may be used. In the following theoretical development (Sections 2.3 to 4) scale is assumed known to simplify the treatment.

The RHT-estimator is a nonparametric estimator. The model $\mathbf{E}y = \beta x$ is merely used to motivate the expression of the HT-estimator as a least squares functional. Neither the HT-estimator nor the RHT-estimator need this model or symmetry of errors with variance proportional to x in order to be applied.

Other formulations of the HT-estimator as least squares functionals may be appropriate in certain conditions. Suppose that in spite of the IPPS-design y_i is not correlated with π_i. Then one would probably choose the unweighted sample mean $\bar{y}_S = \sum_{i\epsilon S} y_i/n$ as an estimator of the population mean (*cf.* Rao 1966). A robustification of $\bar{y}_S$ could be a solution $\hat{\mu}$ of $\sum_{i\epsilon S} \psi(y_i - \mu) = 0$. This is a location M-estimator. If the HT-estimator is in fact appropriate due to the correlation between y_i and π_i then this robustification is not efficient.

A third robustification would assume y_i proportional to x_i but with variance proportional to the square of x_i. This is in fact the situation where the HT-estimator is optimal. The corresponding robustification would be a solution $\hat{\beta}$ of $\sum_{i\epsilon S} \eta(x_i, y_i/x_i - \beta) = 0$. Obviously this robustification does not account for the IPPS-sample design. If the design is put back into the estimating equation by solving $\sum_{i\epsilon S} \eta(x_i, y_i/x_i - \beta)/\pi_i = 0$ then we do not get back the HT-estimator when $\eta(x,r) \equiv r$.

One may argue that in fact the HT-estimator is never used in its pure form for estimating population means. The usual estimator is $(\sum_{i\epsilon S} y_i/\pi_i)/(\sum_{i\epsilon S} 1/\pi_i)$, sometimes called the Hájek-estimator. The estimating equation of the Hájek-estimator, $\sum_{i\epsilon S}(y_i - T)/\pi_i = 0$, makes obvious that the Hájek estimator is not robust against outliers in y.

But the residual $y_i - T$ does not involve the auxiliary variable x_i. Therefore the Hájek-estimator does not suffer from a possible combined effect of large y_i together with small x_i, which may be a leverage point for the regression model underlying the HT-estimator.

2.3 A Sampling Sensitivity Curve

The derivation of an approximate sampling variance of the RHT-estimator (see Section 2.4) uses a finite population analogue to the influence function for infinite populations (Hampel 1974). For finite population sampling with design based inference it is appropriate to develop a sensitivity curve (SC) (*cf.* Hampel *et al.* 1986, p. 93) for $\beta(F,\eta)$ at the population distribution function F_U. In other words, the slope of the RHT-functional is linearized around F_U. Denote by $U+$ the population U augmented by a unit with characteristic (x,y). Denote by $\lambda(\beta,F_U)$ the function $\sum_{i\epsilon U}\eta(x_i',r_i'(\beta))x_i'/N$, such that the defining equation for $\beta(F_U,\eta)$, the M-estimator at the population distribution function, is $\lambda(\beta,F_U) = 0$. Clearly

$$(N+1)[\lambda(\beta(F_{U+},\eta),F_{U+}) - \lambda(\beta(F_U,\eta),F_U)] = 0.$$

Using a linear approximation to $\eta(x,\cdot)$ and neglecting terms in $1/N$ the sensitivity curve of $\beta(F_U,\eta)$ can be isolated from this equation:

$$(N+1)(\beta(F_{U+},\eta) - \beta(F_U,\eta)) \approx$$

$$\frac{\eta(x',r')x'}{\sum_{i\epsilon U}\eta_2(x_i',r_i')x_i'^2/N} =: \mathrm{SC}(x,y,F_U,\eta), \qquad (5)$$

where $\eta_2(x,r) = \partial\eta(x,r)/\partial r$ and both r' and r_i' are evaluated at $\beta(F_U,\eta)$. This SC may be extended to the case of a p-dimensional explanatory variable (*cf.* Hampel *et al.* 1986, p. 316 and Hulliger 1991, p. 183).

Since units usually are not independently included into an IPPS sample, the reaction of the RHT-slope to a particular observation must be investigated by conditioning on a particular sample. The deviation of the estimator $\beta(F_S,\eta)$ at a particular sample S from $\beta(F_U,\eta)$ may be approximated by integrating the SC of $\beta(F,\eta)$ with respect to the sampling distribution function F_S (*cf.* Hampel *et al.* 1986, p. 85):

$$\beta(F_S,\eta) - \beta(F_U,\eta) \approx \int \mathrm{SC}(x,y,F_U,\eta)\,dF_S. \qquad (6)$$

The influence of unit i in sample S may then be defined as the contribution of the unit i to the deviation due to the sample S, *i.e.*,

$$\mathrm{SC}((x_i,\pi_i,y_i) \mid S,F_U,\eta) =$$

$$\frac{\eta(x_i',r_i')x_i'/\pi_i}{(\sum_{j\epsilon S}1/\pi_j)\sum_{j\epsilon U}\eta_2(x_j',r_j')x_j'^2/N}. \qquad (7)$$

The SC may be studied theoretically to discuss the properties of the RHT-estimator and to choose a good η-function. And it may be estimated by replacing the standardization factor $N/(\sum_{j\epsilon U}\eta_2(x_j',r_j')x_j'^2)$ by an appropriate estimator. The estimated SC may be used as a tool for outlier detection.

The influence of unit i in sample S on the HT-estimator is

$$\bar{x}_U\mathrm{SC}((x_i,\pi_i,y_i) \mid S,F_U,\eta \equiv r) =$$

$$(y_i - \beta_{LS}(F_U)x_i)\bigg/\Bigg(1 + \pi_i\sum_{j\epsilon S\setminus i}1/\pi_j\Bigg).$$

This SC is unbounded in y_i such that the HT-estimator is not robust against outlying y_i. The y_i influences the HT-estimator through the residual $y_i - \beta_{LS}(F_U)x_i$. This makes clear why a large y_i combined with a small x_i (or small π_i) has a large influence. If π_i is directly proportional to x_i, as the IPPS design in principle requires, then the SC of the HT-estimator is bounded in x_i. In other words the HT-estimator is robust against outlying x_i. However the bound may be quantitatively too high to be efficient and further downweighting of outlying x_i may be necessary.

2.4 Approximate Expectation and Variance

Along the lines of the proof of proposition 2.1 in Gwet and Rivest (1992) it can be shown that $\beta(F_S,\eta)$ is consistent for $\beta(F_U,\eta)$ in the sense that for a growing and nested sequence of populations and IPPS samples $\lim_{N,n\to\infty}P[\,|\beta(F_S,\eta) - \beta(F_U,\eta)| < \epsilon] = 1\,\forall\epsilon > 0$.

Due to the consistency of $\beta(F_S,\eta)$ the sampling expectation $\mathrm{E}_S\beta(F_S,\eta)$ is approximately $\beta(F_U,\eta)$. Of course $\bar{x}_U\beta(F_U,\eta)$ may be different from the population mean and then $\bar{x}_U\beta(F_S,\eta)$ has a bias as an estimator of $\bar{y}_U$. In particular if the population distribution is not symmetric then $\bar{x}_U\beta(F_S,\eta)$ is in general a biased estimator for $\bar{y}_U$ but nevertheless consistent for $\bar{x}_U\beta(F_U,\eta)$. The important question then is how large is the bias of $\bar{x}_U\beta(F_S,\eta)$, in particular when compared with the variance.

The SC (5) may be used to derive a variance approximation. The derivation is analogous to the case of independent identically distributed random variables with the influence function replaced by the sampling SC. Taking the expectation of the square of (6) one gets after some approximations

$$\mathrm{Var}_S\beta(F_S,\eta)$$

$$\approx \mathrm{E}_S[(\beta(F_S,\eta) - \beta(F_U,\eta))^2]$$

$$\approx \frac{\mathrm{Var}_S(\sum_S\eta(x_i',r_i')x_i'/\pi_i)}{(\sum_{i\epsilon U}\eta_2(x_i',r_i')x_i'^2)^2}$$

$$\approx \frac{\sum_{i\epsilon U}\left(\frac{1}{\pi_i} - 1\right)\eta(x_i',r_i')^2 x_i'^2 + \sum_{i\neq j\epsilon U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)\eta(x_i',r_i')x_i'\eta(x_j',r_j')x_j'}{\sum_{i\epsilon U}\eta_2(x_i',r_i')^2 x_i'^4 + \sum_{i\neq j\epsilon U}\eta_2(x_i',r_i')x_i'^2\eta_2(x_j',r_j')\,x_j'^2}, \tag{8}$$

where r_i' is evaluated at $\beta(F_U,\eta)$. Denote this approximate variance by V_r. An important difference to the case of the asymptotic variance of an M-estimator with independent identically distributed random variables is that the cross-product terms in the numerator of V_r do not vanish. If $\eta(x,r) \equiv r$ then V_r yields the correct variance of the HT-estimator.

2.5 Estimation of the Variance

The numerator of V_r is the variance of $\sum_{i\epsilon S}\eta(x_i',r_i')(\beta(F_U,\eta))x_i'/\pi_i$ which is a HT-estimator apart from the unknown $r_i'(\beta(F_U,\eta))$. Therefore the variance estimator (1) for the HT-estimator may be used. After replacing $\beta(F_U,\eta)$ by the estimator $\beta(F_S,\eta)$, the estimator of the variance of the RHT-estimator becomes

$$v_{rHT} = -\bar{x}_U^2 \frac{\sum_{i\epsilon S}\frac{1}{\pi_i}\eta(x_i',r_i')^2 x_i'^2 + \sum_{i\neq j\epsilon S}\frac{1}{\pi_{ij}}\eta(x_i',r_i')x_i'\eta(x_j',r_j')x_j'}{\sum_{i\epsilon S}\frac{1}{\pi_i}\eta_2(x_i',r_i')^2 x_i'^4 + \sum_{i\neq j\epsilon S}\frac{1}{\pi_{ij}}\eta^2(x_i',r_i')x_i'^2\,\eta_2(x_j',r_j')x_j'^2}. \tag{9}$$

The minus sign in (9) is in order. The (negative) cross-product terms in the numerator usually dominate. Nevertheless v_{rHT} may become negative as can the HT-variance estimator (1) itself (*cf.* Cochran 1977, p. 261). The variance estimator v_{rHT} does not yield the variance estimator (1) if $\eta(x,r) \equiv r$. Of course the Yates-Grundy-Sen estimator may be used to estimate the numerator of V_r. A third variance estimator may be derived by writing the RHT-estimator as a weighted least squares estimator whose weights depend on the estimate (*cf.* Hulliger 1991, p. 166). Since the MER-estimators (*cf.* Section 4) performed slightly better with v_{rHT} than with the other variance estimators the simulations of Section 5 were done with v_{rHT}.

3. EXTENSIONS

3.1 Stratification and Domains

The stratified mean under stratified random sampling is a HT-estimator. The stratified mean may be written as the mean of predicted values under a one-way analysis of variance model. The corresponding robustification is straightforward. It amounts to the separate robustification of the stratum means (Hulliger 1991). However, if the stratum sample size is 1 or 2 no outlier can be down-weighted without the help of further assumptions. Furthermore the biases of the robustified stratum means may add up to a large overall bias (*cf.* Rivest 1993, Section 4). Therefore different robustifications may be appropriate for estimating stratum means and overall means.

This is a general problem for robust estimation in sub-populations (domains) since the definition of an outlier depends on the reference population. An observation may be an outlier in a particular subpopulation but may be harmless in another one. Thus a robust estimator may be suited for one subpopulation but perform poorly in another subpopulation. Often no robustification is needed or wanted for overall means but subpopulation means need to be robustified because of outliers that turn up. Luckily the sample size is often considerably smaller in a subpopulation than in the whole population and then the bias component of the MSE of a robust estimator is often smaller than the variance component. Thus robust estimators may be more efficient than the HT-estimator when used in domain estimation.

3.2 Hansen-Hurwitz Strategy

When sampling is done with replacement and with unequal drawing probabilities the Hansen-Hurwitz estimator is used instead of the HT-estimator. The Hansen-Hurwitz estimator may be robustified analogously to the HT-estimator (see Hulliger 1991, section 4.4) since the underlying model is the same. The variance approximation for the robustified HH-estimator is simpler than for the RHT-estimator because the crossproduct terms vanish due to the drawing with replacement of the Hansen-Hurwitz design.

3.3 Robustified IPPS Design

The ratios y_i/π_i in the HT-estimator act like the summands of an arithmetic mean. Small π_i together with large y_i inflate the HT-estimator. To robustify the design against very large and very small inclusion probabilities we may put $\tilde{\pi}_i = n\tilde{x}_i/\sum_U\tilde{x}_i$, where $\tilde{x}_i = \bar{x}_U + \psi_{\text{Hub}}(x_i - \bar{x}_U,k)$. Thus the auxiliary variable x_i is "Huberised" from its mean to prevent too high and too low values. Now an IPPS sample is drawn with inclusion probabilities $\tilde{\pi}_i$. The HT-estimator is still $T_{HT} = (1/N)\sum_S y_i/\tilde{\pi}_i$ and it is still unbiased. Of course it is not robust against outliers in y and it may loose efficiency if the expectation of the y_i is not proportional to $\tilde{\pi}_i$. The weighted LS-estimator under the superpopulation model for the HT-estimator (see Section 2.1) with inclusion probabilities $\tilde{\pi}_i$ and unmodified auxiliary variable x_i is

$$\beta_{LS}(F_S) = \frac{\sum_S y_i/\tilde{\pi}_i}{\sum_S x_i/\tilde{\pi}_i}, \tag{10}$$

with corresponding estimator for the population mean $\bar{x}_U\beta_{LS}(F_S)$. This β_{LS} may be robustified against outliers in y_i like the HT-estimator. Ratio estimators in IPPS samples are of the same form with the original π_i instead of $\tilde{\pi}_i$. Thus ratio estimators may be robustified analogously to HT-estimators, too (*cf.* Gwet and Rivest 1992).

3.4 One-step Estimators

It is not advisable to express robust estimators as weighted means with fixed weights attached to the observations because the notion and the effect of an outlier depend on the particular domain and variable to be analysed. However, so-called one-step estimators, which are expressed as weighted means, reduce the computational complexity of robust estimators. The one-step RHT-estimator is

$$\bar{x}_U \frac{\sum_{i\in S} w_i y_i' x_i'/\pi_i}{\sum_{i\in S} w_i x_i'^2/\pi_i}, \tag{11}$$

with weights $w_i = \eta(x_i', y_i' - \beta_{LS}x_i')/(y_i' - \beta_{LS}x_i')$. In fact this is the result of the first step of the iteratively reweighted least squares algorithm, which is often used to calculate M-estimators. The one-step RHT-estimator inherits much of the good properties of the fully iterated RHT-estimator and is simpler to implement and faster to compute.

4. MINIMUM ESTIMATED RISK ESTIMATORS

The RHT-estimator is in general biased. A convenient performance criterium is the sampling mean squared error (MSE) $\mathbf{E}_S[(\bar{x}_U\beta(F_S,\eta) - \bar{y}_U)^2]$. For small to moderate samples the gains of RHT-estimators over the HT-estimator are not very sensitive to the particular robustification chosen if there are outliers in the sample (*cf.* Hulliger 1991, Chapter 3). But with well-behaved data or for moderate to large samples the losses in MSE of certain RHT-estimators may be considerable. The question arises how to choose a good RHT-estimator. Minimum estimated risk estimators (MER-estimators), which adapt the tuning constant of a RHT-estimator to the sample, are a possibility. MER-estimators for the expectation of a univariate random variable are investigated in Hulliger (1991, Chapter 2). The idea is to take a simple M-estimator like a Huber M-estimator, to estimate its MSE for a set of tuning constants k, and to choose the tuning constant with least estimated MSE.

Huber's (1964, p. 97) proposal 3 and Jaeckels (1971) adaptive trimmed mean aim at symmetric random variables and therefore use a variance estimate instead of an estimate of the MSE. MER-estimators are similar but their aim is to estimate the mean of asymmetric distributions.

Here we introduce MER-estimators for IPPS designs. Consider a parametric set of functions $\{\eta_k(x,r) : k \in K\}$, where $K \subset \mathbf{R}^p_+$ is the set of parameters. Usually $p = 1$ or 2 to make minimization feasible and to keep the efficiency loss due to the estimation of the nuisance parameter k low. We do not call k a parameter but a tuning constant to avoid any confusion with the concept of parameters in probability distributions. A suitable set of η-functions induces a set $\mathcal{B} := \{\beta(F_S,\eta_k) : k \in K\}$, where $\beta(F_S,\eta_k)$ is the slope of an RHT-estimator. To ensure consistency of the MER-estimator let $\lim_{k\to\infty}\eta_k(x,r) = r \,\forall(x,r)$ such that the HT-estimator is an element of $\mathcal{B}$. The MSE of $\beta(F_S,\eta_k)$ may be estimated by

$$r(F_S,k) = \max(v_r(F_S,k),0) + (\beta(F_S,k) - \beta_{LS}(F_S))^2, \tag{12}$$

where $v_r(F_S,k)$ is the variance estimator (9) or some other estimator of the variance of $\beta(F_S,\eta_k)$. We use $\max(v_r,0)$ in $r(F_S,k)$ because the variance estimator (9) may become negative. Typically the function $r(F_S,k)$ with $k \in \mathbf{R}_+$ has a maximum at or close to $k = 0$ which stems from a large bias. Then it drops to a minimum where bias and variance are both small. For large tuning constants $r(F_S,k)$ approaches the variance of the HT-estimator, usually from below.

Definition 3. Suppose $r(F_S,.)$ has a global minimum at $k_m(F_S) \in K$. Then the MER-estimator of the population mean is $M(F_S) = \bar{x}_U\beta(F_S,\eta_{k_m})$.

MER-estimators with suitable defining functions are scale equivariant and do not need a scale estimator. MER-estimators are in general consistent estimators of the population mean. A proof of the strong consistency of MER-estimators of the expectation of a random variable is in Hulliger (1991, Chapter 2).

Problems with nonuniqueness of the minimum or when the minimum is not attained on K are easily resolved in practice by inspection of the function $r(F_S,k)$. (If there are several global minima choose the one with smallest tuning constant to obtain more robustness.) The bias part of $r(F_S,k)$ involves the slope $\beta_{LS}(F_S)$ of the HT-estimator. By this term the sensitivity of the HT-estimator is transferred to MER-estimators and thus the robustness of RHT-estimators is lost again. But if the MER-estimator should be consistent for the population mean there is no way around a consistent and therefore nonrobust estimator in the bias part of the risk estimator. Nevertheless MER-estimators are quantitatively less sensitive to outliers and more efficient than the HT-estimator if outliers occur (see Section 5).

It is even possible to bound the influence of outliers on the MER-estimator for finite samples without loosing its (asymptotic) consistency. This is achieved by downweighting the bias part in the estimated risk of the HT-estimator in an appropriate way (MER2-estimators, (Hulliger 1991, Paragraph 2.4.1)).

MER-estimators may be more efficient than the HT-estimator because their bias is more than compensated by the variance reduction due to the downweighting of outliers. How much can be gained quantitatively is explored in Section 5.

5. A SMALL SIMULATION STUDY

Simulations with populations of size $N = 128$ and with samples of size $n = 16$ are presented here. The sample design in Dey and Srivastava (1987) is used (Note that there is a factor 2 missing in their formula (2.3)). Dey and Srivastava propose to form $m > n/2$ groups. The group totals $\sum_{U_j} x_i (j = 1, \ldots, m)$ must fulfill the inequality $\sum_{U_j} x_i / \sum_U x_i > (n - 2)/(n(m - 1))$. Thus the group totals are allowed only little variability and the groups are difficult to form in particular for larger samples (Hulliger 1991, p. 179).

The $x_i (i = 1, \ldots, N)$ are independent realisations according to a 5%-scale contaminated exponential distribution with origin at 1, *i.e.*, $(X_i - 1) \sim 0.95\,\mathrm{Exp}(1) + 0.05\,\mathrm{Exp}(3)$, where $\mathrm{Exp}(\theta)$ denotes the exponential distribution function $1 - \exp(-x/\theta)$. The shift $+1$ is introduced to lower the probability of negative responses in the regression through the origin model with symmetric errors.

The first response $y_U^{(1)}$, with acronym GODA, is a realization of independent normal variables distributed as $Y_i \sim \mathcal{N}(100x_i, x_i^2)$. This is the model under which the HT-estimator is optimal (*cf.* Godambe 1955). The response $y_U^{(2)}$ (HTLS) is a realization of independent variables distributed as $Y_i \sim \mathcal{N}(2x_i, x_i/4)$. This is the ideal model that yields the HT-estimator as LS-estimator. A third response $y_U^{(3)}$ (HTG) is created by the model $Y_i \sim 0.95\mathcal{N}(2x_i, x_i/4) + 0.05\mathcal{N}(2x_i, 9x_i/4)$. The residual outliers have 3 times larger scale. The response $y_U^{(4)}$ (HTE) has asymmetric outliers which are not related to the x-variable. The bulk of the data (120 observations) stems from the distribution $Y_i \sim \mathcal{N}(2x_i, x_i/4)$ of $y_U^{(2)}$ but 8 randomly chosen observations stem from Exp(2.5). The population $y_U^{(5)}$ (HMT) stems from a distribution with expectation $0.4 + 0.25x_i$ and has a Gamma distribution with variance proportional to $x^{3/2}$. Thus the variable y has the distribution proposed in Hansen, Madow and Tepping (1983, p. 781). Finally a population $y_U^{(6)}$ (HMTE) is generated with 120 observations from the same distribution as $y_U^{(5)}$ but with 8 randomly chosen observations from the distribution Exp(2). The six populations above are chosen to be realistic. They all use the same population of x-values (see Figure 1).

The RHT estimator in the simulation uses

$$\eta(x_i', r_i') = w(x_i', k_x)\psi_{\mathrm{Hub}}(r_i', k_r\, \mathrm{med}_S\, |\, r_i'\,|),$$

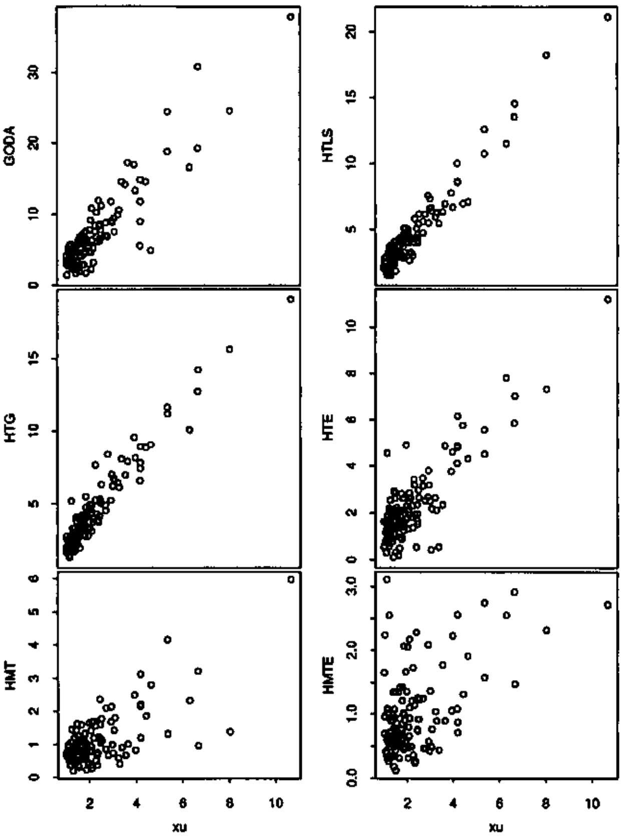

Figure 1. Populations of the Monte-Carlo Study.

with $w(x_i', k_x) = \min(1, k_x\, \mathrm{med}_U\, |\, x_i'\, |\, / \,|\, x_i'\, |)$ and $k_x = k_r = 2$. The weighting function $w(x_i', k_x)$ corresponds to an asymmetric Huber-function $\psi_{a\mathrm{Hub}} = \min(x_i', k_x)$, which downweights large x_i' only. The scale $\mathrm{med}_S\, |\, r_i'\, |$ is the median of the absolute residuals evaluated at the solution of the preceding iteration of the iteratively reweighted least squares algorithm. The MER-estimator uses the same η with tuning constants k_x, k_r evaluated at 20 points which lie on the diagonal of the range of k_x and k_r. S-PLUS functions for the calculation of the estimators may be obtained from the author.

For each of the populations a set of 400 samples was drawn to evaluate the estimators. The obtained precision is sufficient to draw conclusions (see the standard errors of the efficiencies in Table 1).

The results are presented in Table 1. The relative bias of the RHT-estimator is always larger than the relative bias of the MER-estimator. The biases of the two estimators have the same sign, except when they are very small. With the exception of populations HTE and HMTE the variance of the RHT-estimator is larger than the variance of the MER-estimator. While the RHT-estimator looses 9% efficiency at population GODA, where the HT-estimator should be optimal, the MER-estimator looses little. With population HTLS, where the HT-estimator is the least squares estimator, the RHT-estimator looses about 12%.

Table 1

Monte-Carlo simulations with RHT- and MER-estimator

	Populations					
	GODA	HTLS	HTG	HTE	HMT	HMTE
MC-mean of HT	6.996	4.531	4.483	2.271	1.068	0.991
Rel. bias of RHT	−0.002	−0.001	−0.009	−0.009	0.006	−0.052
Rel. bias of MER	0.000	−0.001	−0.007	−0.008	−0.002	−0.035
Rel. SE of HT	0.067	0.041	0.044	0.098	0.107	0.170
Rel. SE of RHT	0.070	0.044	0.040	0.087	0.117	0.144
Rel. SE of MER	0.068	0.042	0.040	0.091	0.107	0.146
Eff. of RHT	0.911	0.876	1.110	1.310	0.827	1.234
Eff. of MER	0.969	0.981	1.158	1.194	0.989	1.284
MC-SE of eff. RHT	0.020	0.017	0.073	0.009	0.018	0.001
MC-SE of eff. MER	0.003	0.009	0.037	0.002	0.013	0.002

NOTE: Relative bias and relative standard error (rel. SE) are biases and standard errors divided by the MC-mean of the HT-estimator. Efficiencies (Eff.) are MSE of the HT-estimator divided by the MSE of the estimator. Estimated standard errors of these Monte-Carlo estimates of efficiency are given in the last two lines.

The efficiency loss of the MER-estimator is once again small. Population HTG contains symmetric residual outliers. The RHT-estimator gains about 11% (but see the error of 7.3%) and the MER-estimator about 16%. Under the asymmetric outliers of population HTE the gain of the RHT-estimator is 31% while the MER-estimator gains 19%. If neither the regression through the origin, nor the symmetry of errors or the proportionality of their variance to the explanatory variable holds, *i.e.*, for population HMT, then the RHT-estimator looses 17% compared with the HT-estimator while the MER-estimator looses practically nothing. If in such a population a few asymmetric outliers turn up like in population HMTE then both robust estimators gain considerably against the HT-estimator, namely 23% and 28% respectively.

In conclusion from this limited simulation the MER-estimator looses little in terms of MSE, compared with the HT-estimator, when there are no outliers in the population. It gains moderately in populations with symmetric outliers and considerably when the outliers are asymmetric. The RHT-estimator looses more under ideal situations than the MER-estimator. The adaptivity of the MER-estimators pays off.

Extensive simulations with infinite populations in Hulliger (1991) confirm these conclusions and show that the gains of robust estimators may be very large for skew populations with outliers. However the possible efficiency gains with robust estimators vanish for large samples since then the bias dominates MSE. On the other hand if the outliers that turn up in a sample are not representative, *e.g.*, if they are uncorrected coding errors, then the robust estimators are much more efficient than the HT-estimator for all sample sizes.

ACKNOWLEDGEMENTS

This article is an outgrowth of the authors Ph.D. thesis at ETH Zürich. The author would like to thank Prof. F.R. Hampel and Prof. H.R. Künsch, for their advice. The author is grateful for the valuable comments of two anonymous referees on drafts of this paper.

REFERENCES

CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

COCHRAN, W.G. (1977). *Sampling Techniques* (3rd. Ed.). New York: Wiley.

DEY, A., and SRIVASTAVA, A.K. (1987). A sampling procedure with inclusion probabilities proportional to size. *Survey Methodology*, 13, 85-92.

FULLER, W.A. (1991). Simple estimators of the mean of skewed populations. *Statistica Sinica*, 1, 137-158.

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *International Statistical Review*, 30, 28-32.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, Series B, 17, 269-278.

GWET, J.-P., and RIVEST, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87, 1174-1182.

HAMPEL, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383-393.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986). *Robust Statistics*. New York: Wiley.

HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-807.

HIDIROGLOU, M.A., and SRINATH, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73-101.

HULLIGER, B. (1991). Nonparametric M-estimation of a population mean. Doctoral Dissertation ETH No. 9443, ETH Zürich.

IACHAN, R. (1984). Sampling strategies, robustness and efficiency: the state of the art. *International Statistical Review*, 52, 209-218.

JAECKEL, L.A. (1971). Robust estimates of location: symmetry and asymmetric contamination. *Annals of Mathematical Statistics*, 42, 1020-1034.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

LITTLE, R.J.A., and SMITH, Ph.J.(1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.

OEHLERT, G.W. (1985). The random average mode estimator. *Annals of Statistics*, 13, 1418-1431.

RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā* A, 28, 47-60.

RIVEST, L.-P. (1993). Winsorization of survey data. *Proceedings of the 49th Session, International Statistical Institute.*

SEARLS, D.T. (1966). An estimator for a population mean which reduces the effect of large observations. *Journal of the American Statistical Association*, 61, 1200-1204.

SHOEMAKER, L.H., and ROSENBERGER, J.L. (1983). Moments and efficiency of the median and trimmed mean for finite populations. *Communications in Statistics, Simulations and Computations*, 12(4), 411-422.

SMITH, T.M.F. (1987). Influential observations in survey sampling. *Journal of Applied Statistics*, 14, 143-152.

STATISTICAL SCIENCES, INC. (1990). *S-PLUS Software*, Seattle: Statistical Science, Inc.

Visitor Sample Surveys

RONALDO IACHAN and SUZANNE S. KEMP[1]

ABSTRACT

This paper discusses the design of visitor surveys. To illustrate, two recent surveys are described. The first is a survey of visitors to National Park Service areas nationwide throughout the year (1992). The second is a survey of recreational users of the three-river basin around Pittsburgh, Pennsylvania, during a twelve-month period. Both surveys involved sampling in time with temporal as well as spatial stratification. Sampling units had the form of site-period pairs for the stage before the final, visitor sampling stage. Random assignment of sample sites to periods permits the computation of unbiased estimates for the temporal strata (*e.g.*, monthly and seasonal estimates) as well as estimates for strata defined by region and by type of use.

KEY WORDS: Recreational user; Sampling in time; Site-period.

1. INTRODUCTION

Surveys of visitors present unique challenges that are rarely discussed in the statistical literature. This paper attempts to fill this gap by describing the design and emphasizing the common features of two surveys recently conducted by the Research Triangle Institute (RTI). We hope that the lessons learned in these efforts will be beneficial to researchers planning similar surveys.

The first survey was a study of visitors to National Park Service (NPS) areas jointly conducted for the National Park Service by RTI and HBRS, Inc. This study involved a probability sample of park visitors that represented visitors to 323 NPS areas nationwide (except Alaska) throughout the year (1992). We will refer to NPS areas as parks for simplicity while pointing out that the NPS areas include locations of historical and cultural parks. The main objective of the NPS study was to assess the visitors' experiences and problems with particular attention to those related to aircraft overflights (*e.g.*, noise and other possible annoyances). A variety of data were also collected in a mail survey for a subsample of selected visitors.

The second survey was a study of recreational users of the Pittsburgh-area three-river basin along the Monongahela, Allegheny and Ohio Rivers in 1992 (or more precisely, between February 1992 and January 1993). This survey was jointly conducted for the Ohio River Valley Sanitation Commission by RTI and Terrestrial Environmental Specialists. The study area included a 40-mile segment of the Ohio River, a 24-mile segment of Monongahela River and a 7-mile segment of the Allegheny River. The primary objective of the Three-River Study was to construct a baseline profile of recreationists in the area and to model the economic value they assign to various activities. Three basic types of recreational activities were distinguished: boating, fishing and park use.

The Three-River Study is the most comprehensive of a series of studies conducted by RTI to assess environmental impact in a number of states. These studies estimate possible reductions in economic or recreational value assigned by actual and potential recreational users to areas that have been or might be affected. While a wider survey of potential users of such areas may consider a telephone sample design, a visitor intercept survey design is found necessary to capture users at a point in time close to actual use.

A discussion of design issues in visitor surveys such as these has been recently provided in Kalton (1991) including issues related to sampling in time and space that are crucial in our framework. In its simplest form, a prototype, two-stage sample design for a visitor survey considers site-period pairs as primary sampling units (PSUs) from which visitors are selected in the second stage. Examples include exit polls (see, for example, Levy 1983), shopping mall intercept surveys (see, for example Sudman 1980) and other transportation and traffic surveys (Gough and Ghangurde 1977; Kish, Lovejoy and Rackow 1961). Among the design issues salient in visitor surveys, the following general problems may be singled out:

- It is desirable to select with greater probabilities those site-periods with larger numbers of visitors; stratification and PPS selection are then effective design features.
- Data collection arguments are key for the specification of the period length and of sampling rates within site-periods; *e.g.*, trade-offs occur between the potential for the field staff to be too busy (short periods, high sampling rates) or not busy enough (long periods, low sampling rates).
- Analytic objectives as well as efficiency suggest temporal stratification dimensions as season, month, weekend versus weekday, and even time-of-day; *e.g.*, the need for seasonal estimates suggests the use of seasons or months as strata for the selection of periods.

[1] Ronaldo Iachan, Research Statistician and Suzanne S. Kemp, Statistician, Research Triangle Institute, 3040 Cornwallis Road, Research Triangle Park, NC 27709-2194, U.S.A.

The two surveys discussed in this paper share the primary objective of characterizing the visitor population in the area or the nation over an entire year. They differ, however, in the priority estimates that lead to the basic design features in each case. In the Three-River Study but not in the NPS study, reasonably precise monthly estimates needed to be computed. For the former study, then, the temporal sampling units - days - were stratified into months. Spatial stratification of the Three-River Study sites was geographic and by recreation type (boating, fishing or park sites).

For the NPS study, primary stratification was by park type. Some park areas needed to be included with certainty into the sample to satisfy legislative requirements. In these and other selected parks, it was further desired to compute park-specific estimates. In these park areas, labelled intensive parks, we then decided to select relatively more site-periods. The initial design optimization problem was how to allocate the sample size to the sampling stages, *i.e.*, to decide how many parks and how many site-periods per park should be selected. Section 3 discusses a solution for this problem which is a function of the intracluster correlations within parks and within periods. The design optimization for the NPS survey also applied to the temporal and spatial strata at the intermediate sampling stages, between park areas at the first stage and site-periods at the penultimate stage (keeping in mind that visitors are selected at the final stage).

These two surveys illustrate issues such as temporal stratification, the choice of appropriate sampling units, and random assignment of spatial units to temporal units. Section 2 outlines the common aspects of the two studies as well as their basic differences. Sections 3 and 4 describe the design of the NPS Visitor Survey and of the Three-River Study, respectively. Section 5 discusses the weighting procedures used for the surveys. A brief overview and some conclusions are presented in Section 6.

2. OVERVIEW OF SAMPLE DESIGNS: PARALLELS AND CONTRASTS

For both surveys, the ultimate visitor samples were selected via intercept sampling as visitors left the sample locations at the selected time periods. Exit interviews were necessary to reflect their attitudes immediately following their recreational experiences. Also, in both studies, visitors were selected from sample site-period pairs. The use of site-period pairs as sampling units dates back to Kish, Lovejoy and Rackow (1961). This sampling unit definition permits the selection of visitors according to a data collection schedule that specifies which sites will be covered at which points in time. Unlike the Three-River Study, the selection of site periods was not the first stage for the NPS Study. The primary sampling units (PSUs) for this study were NPS areas, or parks. The NPS survey involved several stages of selection described in the next section.

Additionally, both studies used temporal frames of days, and eligible data collection periods within days, to permit inferences about the entire year. The designs included the selection of time periods so that each eligible period has a known, positive probability of selection. Although both studies involved temporal frames, the structure of the frames and selection of days for each study were quite different.

For example, the sample for the Three-River Study was selected as twelve independent monthly samples. Each monthly sample has essentially the same, stratified random sampling design but a different sample allocation and different sample sizes were used in different months. This design took into account seasonal variations in recreational patterns, and enabled estimation for each month and stratum (*e.g.*, by type: boating, fishing or park). Both spatial and temporal frames were allowed to vary from month to month. The stratification and allocation for this sample are discussed in Section 4.

In contrast, the temporal frame for the NPS visitor survey first considered two-month blocks for each sample park (PSU). The use of two-month periods as (second-stage) sampling units in time efficiently met the survey objectives for two basic reasons. First, it allows the effective (geographic) concentration of staffing resources and staggered data collection throughout the year. Second, this choice of period permitted capturing seasonal fluctuations in park visitation across the park system, resulting from some parks having relatively higher visitation in the spring, others in the fall months, and so on.

One two-month block was selected for each sample park so that data collection could be effectively concentrated in time. Then, at the next stage of temporal selection, days were selected from within the two-month block for each sample park. Like the parks themselves at the first stage, these two-month blocks were selected with probabilities proportional to size (PPS), with the size measure being the aggregate visitation.

The sample sizes and allocation to the several sampling stages were carefully balanced to minimize clustering effects associated to clusters in time and space. For the Three-River Study, this clustering occurs at the first stage of selection where sampling units are sites and time periods. The allocation also considered the varying sample sizes used in successive, independent monthly samples. For the NPS survey, clusters in time were a result of the two-month blocks and sample days periods selected at different stages. Spatial clusters resulted from the use of parks and park exits as sampling units for this survey.

The next section describes in more detail the design of the NPS survey.

3. NPS SURVEY SAMPLE DESIGN

The design of NPS visitor survey capitalized on auxiliary information of various kinds and sources:

- Information obtained in previous studies (*e.g.*, park rankings based on noise exposure and NPS staff classifications).
- Information available in NPS data bases (*e.g.*, park visitation data by month).
- Information collected from NPS staff specifically for design purposes (*e.g.*, an inventory of the park exits for each sample park and number of vehicles leaving each of the exits).

The next subsections describe the various stages of selection for the visitor intercept survey. This survey component will also be designated the frontcountry survey to distinguish it from a survey of backcountry users that was conducted in tandem in the sample parks. Subsection 3.5 describes the backcountry survey as well as a mail survey administered to a subsample of frontcountry respondents and to sample backcountry users.

The fourth-stage selection for the mail survey involved additional stages (and phases) of selection. For the visitor intercept survey, groups (*i.e.*, vehicles) were selected as an ultimate cluster: all persons in a sample group were solicited for an interview. For the mail survey, groups were subsampled, and one person was subsampled from each subsampled group.

3.1 Frame Construction and First-Stage Sampling

We constructed a sampling frame for the selection of parks by compiling NPS information on park visitation (monthly and annual) and on noise exposure, information that was used in the sample design in two distinct ways: for stratification and for assigning size measures. The latter information was based on two different sources: (a) a previous NPS study which ranked parks according to potential exposure, and (b) a classification of parks performed independently by NPS staff (park superintendents, regional staff *etc.*).

In consultation with NPS, RTI combined these two classifications for noise exposure to construct eleven strata. The stratification partitioned parks into categories - very high, high, low and very low. Strata were divided into two substrata using the rankings in the stratum. (Note that the "medium" stratum was not subdivided due to its small park count.)

In addition to noise exposure, these strata incorporate three classes of parks that deserve separate treatment: (1) urban and suburban park areas, (2) parks with missing data on visitation (needed for PPS selection), and (3) parks whose elongated shapes present unique problems of access and reduce the meaning of prior exposure assessments. These classes were sampled at a much lower rate than the other strata; the lowest sampling rate is in the urban stratum (1 in 79).

The certainty stratum included the seven parks that were mandated by legislation to be included in the study. In addition, it included those parks whose aggregate (annual) visitation rates were so large as to ensure selection into the first-stage sample. The 39 sample parks are listed in Exhibit 1; a 40th selected park (Grand Teton) was dropped from the sample for political reasons.

Design optimization calculations led to a first stage sample size of about 40 sample parks, yielding a total of 405 site-periods (or exit-days in this case) selected across intensive and non-intensive parks. As described in Section 3.3, 15 exit-days were selected in each of the three intensive parks, and 10 exit-days were selected in each of the 36 non-intensive parks in the sample. An accurate optimization would require variance components for the between-park and within-park variances. These variance components were approximated using data from a previous study as well as monthly visitation data available for all NPS areas.

3.2 Selection of Two-Month Blocks

While the selection of two-month blocks was with probability proportional to size (PPS), practical requirements were also taken into consideration.

First, training local park staff immediately before data collection was desired. Geographic clusters of parks were formed so that trainers could visit parks in one cluster in one trip. Specifically, the 39 sample parks were grouped into 14 such clusters. This requirement led to the selection of a two-month period for each park cluster. Thus, fourteen two-month blocks were selected, one for each cluster. The size measure used for each selection was the aggregate visitation over the parks in the cluster.

Exhibit 1 shows the sample parks in each cluster, and the sample two-month block selected for the cluster (*i.e.*, for every park in the cluster). Three strata - groups of clusters and hence of parks - were formed for the selection of two-month blocks:

(a) In the "very-high summer" stratum-1 (with 3 clusters), the frame of two-month blocks contained only the summer-peak period, July-August, which was then selected with certainty for these clusters;

(b) In the "high-summer" stratum-2 (with 7 clusters), the frame of two-month blocks contained 11 overlapping two-month blocks;

(c) In the "low-summer" stratum-3 (with 4 clusters), the frame of two-month blocks contained 5 non-overlapping two-month blocks.

The rationale for this temporal stratification was two-fold: it ensured that the data collection was spread throughout the year, and it distinguished parks where a vast majority of the visitation occurs during the summer from those with a more uniform visitation pattern.

Exhibit 1

Park Clusters, Second-stage Strata and Selected Two-month Blocks

		Selected Period
Stratum 1: Very High Summer*		
Cluster 2:	Mount Rainier, N. Cascades, Olympic	July-August
Cluster 4:	Glacier, Yellowstone	July-August
Cluster 10:	Sleeping Bear, Perry's Victory	July-August
Stratum 2: High Summer		
Cluster 3:	Lassen, Yosemite, Kings Canyon/Sequoia	August-September
Cluster 5:	Dinosaur, Rocky Mtn., Mt. Rushmore	June-July
Cluster 6:	Glen Canyon, Grand Canyon, Walnut Canyon	June-July
Cluster 8:	Bandelier, Lake Meredith	May-June
Cluster 11:	Cape Cod, Delaware Gap, Gettysburg	August-September
Cluster 12:	Shenandoah, Fredericksburg, Assateague	June-July
Cluster 13:	Great Smoky, Cape Hatteras, Fort Sumter	June-July
Stratum 3: Low Summer		
Cluster 1:	Haleakala, Hawaii Volcanoes	March-April
Cluster 7:	Lake Mead, Saguaro, Casa Grande	March-April
Cluster 9:	Hot Springs, Wilson's Creek, Buffalo	October-November
Cluster 14:	Cumberland Isd., Canaveral, Everglades, Gulf Island	March-April

*Certainty selection of July-August for each cluster in this stratum.

We selected two-month blocks with different procedures in the two strata as described below.

(1) For park clusters in the former (high-summer) stratum, eleven overlapping periods were included in the temporal frame; January-February, February-March, . . . , November-December. One such period was then selected with probability proportional to size (PPS). Note that for this stratum each month was included in two frame periods except for January and December. The probability of selection for these two winter months was thus reduced even further (beyond the already small probability assigned to the winter periods with the PPS procedure).

(2) For park clusters in the latter (low-summer) stratum, five non-overlapping two-month periods constituted the temporal frame: January-February, March-April, May-June, September-October, November-December. Note that the two-month summer period, July-August, was excluded from the frame for these parks to ensure the selection of other, non-summer months. For each cluster in this stratum, one of these five two-month periods was selected with PPS.

3.3 Third-stage Sampling

Sampling units at the third stage were exit-day pairs. The third-stage sampling is easier to envisage as the combination of two independent selections: (a) selecting days from the temporal window (2-month block) drawn at the second stage for the park, and (b) selecting exits from a list of exits specified for the park. A final step consisted of the random assignment of sample days to sample exits.

The sample of days was stratified by weekdays versus weekend days (including major holidays). Sample days were selected with equal probabilities within each of these two strata. The sample of exits was selected with probabilities proportional to size (PPS). The size measure assigned to each exit was the relative use of the exit among all the exits listed for the selected park. This usage measure was derived with the aid of local park staff.

The third-stage sample design distinguished two groups of parks designated as intensive (3 parks: Grand Canyon and the two Hawaii parks) and non-intensive (36 remaining parks in the sample). Sample sizes in non-intensive parks were 10 exits and 10 days, and hence 10 exit-day pairs. In the three intensive parks, 15 exits and 15 days were selected. Equal allocation to weekend/weekday strata was used in non-intensive parks: 5 weekend days and 5 weekdays were independently selected in each of these (36) sample parks. The allocation was approximately equal in intensive parks with the selection of 7 weekend days and 8 weekdays.

3.4 Fourth-stage Sampling

At the fourth stage, park visitors were intercepted in the selected exit-days. A systematic random sample of visitors or (visitor groups) exiting the park was selected with a fixed sampling interval for each selected third-stage unit (exit-day). The interval was allowed to vary from day to day to capitalize on the experience of previous days and on the variability in visitation across days and exits. Each visitor found in the selected eligible groups was screened for eligibility, and interviewed if eligible (adult visitor).

3.5 Backcountry and Mail Surveys

The same sample of parks (PSUs) was used for a mail survey of frontcountry and backcountry users in the two-month period selected for the park. The sample for the backcountry survey was restricted to the subset of sample parks with some backcountry use. Within each such sample park, the (third-stage) sampling frame for this survey component was based on backcountry permits issued during the data collection time window (two-month block) established for the park. The ultimate sample of backcountry users was then selected with equal probabilities from permit lists provided by park staff. A subsample of visitors selected in the intercept survey were also selected for the mail survey.

The mail survey sample was based on the same third-stage sample of exit-days selected for the intercept survey. For each selected exit-day, a fixed number of groups was subsampled from groups responding to the intercept survey (this number was 15 in intensive parks and 10 in non-intensive parks). A further stage of subsampling was that of one person from within the respondents in each group subsampled. This subsampling of groups and persons was with stratified random sampling to control the demographic composition of the final sample.

4. SAMPLE DESIGN FOR THREE-RIVER SURVEY

4.1 Frame Construction and Stratification

First, RTI and Terrestrial Environment Specialists (TES) constructed a sampling frame based on an inventory of all sites in the Three-River Study area. Then a stratified multistage sample was selected independently for each of the twelve months of the study. First-stage sampling units were site-periods, and second-stage units were individuals engaged in recreation in the selected site-periods. Temporal and spatial stratification were used for the first-stage sampling of time periods and sites.

Primary stratification along the temporal dimension was by month, and primary stratification of sites was by use type: sites (access points) were classified as boating, fishing, or parks. Each primary stratum was divided into six geographic areas, or pools, defined as the river areas between locks. The fishing stratum was further substratified by the presumed use intensity as low or high use: high-use sites are those below locks and dams.

Each monthly sample of site periods was selected with stratified random sampling. Advantages of selecting independent monthly samples included: monthly and seasonal estimates can be computed, and some design features may be changed from month to month.

For example, this design permitted altering the spatial frame from one month to the next with the addition or deletion of sites. In particular, the boating stratum for the winter months (November through April) was restricted to boat ramps open that season. Further, several fishing sites included in the frame for the first few months were found inaccessible and deleted from the frame for the subsequent monthly samples.

Other design features that changed in successive months include: the second-stage sampling rates for selecting eligible users, and the data collection windows used in the morning, afternoon and evening periods for sites of each type. Varying data collection windows were used in different months of the study. These periods were defined using sunrise and sunset information as well as expected patterns of use of the various types. The winter months (November-April) included two periods per day while the summer months (May-October) included three periods per day.

The temporal (sub)stratification of each monthly sample was by weekend days versus weekdays. It is worth noting that the weekend stratum also included major holidays.

4.2 Sample Selection

As noted above, we selected an independent first-stage sample of site-periods for each of the 12 months of the study. Each monthly sample had two components: (a) a stratified random sample of "*n*" sites, and (b) a stratified random sample of "*n*" periods.

Following selection of the sample periods for each month, the sample sites were randomly assigned to the selected periods. The assignment of sites to periods was entirely at random for the months of February through June but was modified in subsequent months. From July on, a sample of time periods was independently selected for each type stratum with the random assignment taking place within stratum. The allocation of the sample time periods (*e.g.*, the number of morning periods and the number of evening periods included in the sample) varied from stratum to stratum. With this more flexible method, relatively more fishing sites could be assigned to morning period and more boating sites to afternoon periods, for example. Exhibit 2 shows the sample sizes - sites and periods - used in random assignment each month.

Exhibit 2

Sample Sizes Used in Random Assignment of Sites to Periods for each Monthly Sample of Three-River Study

Month	Overall*	Sample Size		Parks	Marinas	Regattas	
		Boating	Fishing			Boating	Marinas
1		8	12	4			
2	20						
3	28						
4	28						
5	36						
6	36						
7		17	22	6	12		
8		10	22	6	6	6+	6+
9		10	14	6	6		
10		10	12	4			
11		8	12	4			
12		8	12	4			

* For these (5) months, the assignment took place for the entire collection of sample sites and periods (NOT blocked by site type). For the remaining months the assignment was within each type stratum, a process which involved the selection of independent samples of time periods for each type.

+ In August, the assignment of regatta sites to periods was performed first, separately. Following the assignment, the sample site-periods associated with regattas were shifted either to the boating or to the marina strata depending on the site type.

Second-stage sampling rates were specified for each of the three primary strata prior to each month of data collection. These stratum-specific rates were determined based on the experience of the previous months, and were distributed to the field interviewers along with the month's data collection schedule. Individuals were selected with systematic random sampling within each site-period.

4.3 Marina Survey and Special Events

A marina survey was conducted in the months of June to September. A sampling frame of 48 marinas was based on the TES inventory that was updated in late May 1992.

The sampling method for the phase-in month of June differed slightly from that used for the subsequent monthly samples for marina sites. The June sample was a supplement of 12 marina sites coupled with a sample of 12 days. The marina samples for the months of July to September were selected considering the marina frame as a fourth stratum. The selection procedures were then similar to those in the other three type strata; specifically, (a) a sample of "*n*" marina sites was selected, (b) a sample of "*n*" time periods was selected, and (c) the sample sites were randomly assigned to the selected periods.

It is worth pointing out that some of the marina sites were also included in the boating stratum. In such cases, two distinct frame units were created for the same site. This situation also arose for some sites used for both boating and fishing, and such sites were included in both strata.

In addition to the marina survey, we identified special events taking place in the study area over the 12-month study period. Most of these events were handled in a way similar to weekends and holidays by assigning them to a stratum to be oversampled. A special category of interest was comprised of the regattas occurring in the summer months. For two monthly samples (July and August), we identified the regatta dates as well as the sites affected by each regatta. We then constructed a separate (fifth) stratum to include these site-periods. The first-stage sample allocation to the regatta stratum reflected the oversampling desired for this stratum. As shown in Exhibit 2, the sampling procedure used to select site-periods (first-stage units) from the regatta stratum also differed to that used in the other four strata. Sample site-periods were directly selected in one step from the subset of site-periods in the stratum, *i.e.*, no random assignment was needed.

5. SURVEY WEIGHTING

5.1 NPS Survey Weighting

Sampling weights were first computed for each of the first three stages of selection. The first-stage sampling weight for each sample park was the reciprocal of the selection probability for the park. The second-stage sampling weight for each sample two-month block was similarly computed. The sum of the first-stage sampling weights overall (or in a stratum) was the number of parks in the frame (or in a stratum).

Third-stage weights for sample exit-days were the product of two factors associated with the selection of days and exits. Note that for each selected park and two-month block, the sum of the former set of weights in a temporal stratum (weekend *vs.* weekdays) is the number of days in the stratum, and the sum of the latter set of weights is the number of exits listed in the park. These weights were adjusted for nonresponse which arose at the third stage because in a few parks, data collection did not take place in some selected exit-days. In a given park with this data collection shortcoming, the sum of the adjusted third-stage weights over the active exit-days was made equal to the sum of the sampling weights over all selected exit-days in the park.

Fourth-stage weights were computed at a group-level and at a person-level. Group-level weights are assigned to all participating groups in a sample exit-day, and have the same value for the groups in the same exit-day. Similarly, person-level weights are assigned to all persons intercepted in a sample exit-day. The fourth-stage sampling weights were computed as the reciprocal of the sampling rate specified for the sample exit-day. These weights were then adjusted for group and person-level nonresponse.

The mail survey sample was based on the same third-stage sample of exit-days selected for the intercept survey. For each selected exit-day in non-intensive parks, 10 groups were first selected with equal probabilities from among the participating groups; then, one person was subsampled from all intercept survey respondents in each selected group. A similar procedure was used in intensive parks with the exception that the number of selected groups per exit-day was 15 rather than 10.

The sampling weight for each mail survey record is the product of WTB = number of intercept respondents in the group, and WTA = number of participating groups/10 [non-intensive parks]. For intensive parks, the denominator of WTA is 15 rather than 10.

These weights were adjusted for mail survey nonresponse using exit-days (within park) as weighting classes. Thus, the sum of the adjusted weights for all respondents coming from the same exit-day is the same as the sum of the (unadjusted) weights for all persons intercepted in the exit-day.

For the backcountry survey, the sampling frame for each eligible park (in the subset of sample parks with backcountry use) was based on lists of permits issued during the data collection period: eligible permits were associated with exit dates in this period. A sample of 5 groups per day was selected within each park from the set of permits linked to that day. One person was subsampled from each group. The weight computation parallels that for the mail survey with

$$\text{WTBACK} = \text{BACKA*BACKB},$$

where for each day:

BACKA = (number of groups linked to the day)/5

BACKB = number of persons in group.

Analysis weights for the backcountry survey resulted from nonresponse adjustments made using days as weighting classes.

5.2 Three-River Survey Weighting

As each monthly sample of site-period units was selected, we computed sampling weights that reflected the selection probabilities for the site-period pairs. Initial weights were the product of two sets of weights computed for each monthly sample: (a) weights assigned to each site in the stratified random sample of sites of the given type (boating, fishing, marina and park), and (b) weights assigned to each period in the stratified random sample of periods. These weights were then inflated to take into account the random assignment of sample sites to periods (or vice-versa). Thus for each month, the sum of the site-period weights was equal to the number of site-period combinations in the frame.

The weight adjustment process started with the sampling weights associated with the selected site-days. An initial adjustment was made to the first-stage weights to account for site-periods that were found ineligible or that had missing sampling forms (*e.g.*, not sent by field staff). For this first-stage adjustment, we used the type-by-month strata as weighting classes. That is, the sum of the adjusted weights over the reduced set was made equal to the sum of the unadjusted weights over the entire sample within the type-by-month class. A final adjustment was made at the respondent-level to reflect (a) the systematic sampling interval used within site-periods, and (b) the survey non-response at the individual level.

As part of the weight check procedures, we computed the sum of the final analysis weights over the entire file, and also by month and by type of site. The weight sum should approximate the estimated total number of recreational users leaving the inventory sites during the data collection time window for each month and type of site.

6. CONCLUSION

This paper described the design of two surveys of recreational users that share a number of useful features. The sample designs include sampling in time as well as in space; site-periods are selected at the stage prior to sampling visitors. A spatial frame is then constructed side by side with a temporal frame. After sites are selected from the former and periods are selected from the latter frame, sample sites are randomly assigned to sample periods (or conversely). Sampling weights need to take into account this additional step of randomization. The findings of the NPS visitor survey are described in the study final report (National Park Service 1994). This analysis included a variety of regression models that investigate the impact of hearing and seeing aircraft flying over NPS areas.

Temporally, both studies represent periods throughout the year, that is, a user will have a positive probability of selection for any time of the year. Both studies also include temporal stratification to reflect patterns of use and increase sampling efficiency.

Spatially, while the sample for the National Park Service study was a national sample of visitors to NPS areas, the sample for the Three-River Study was more restricted in spatial scope. Both studies, however, distinguished users of different types. For the NPS study, backcountry and frontcountry users were selected in two separate (third-stage) strata at the point where the two components branch out with the selection of permits for the backcountry survey. For the Three-River Study, visitors were classified by primary fishing, boating, or park use, and sites were stratified in a similar way. This design permitted the computation of precise estimates by type and by season.

The ultimate sampling unit for a survey of recreational users is the specific visit; thus, visitors may have multiple chances of inclusion in the sample to the extent that they use the target areas multiple times. It is worth noting that this structure is consistent with the objectives of such surveys.

Sampling weights accounted for the selection of time and space units at each stage and also for the random assignment step. The samples were designed to minimize the effects of unequal weighting on survey variances. The potential for severe unequal weighting effects was considered in combining different survey components. Examples in the two surveys include combining:

(a) Backcountry and frontcountry components of NPS mail survey, and

(b) Fishing, boating and park users in the Three-River Study.

Some disadvantages of this type of study design should also be pointed out. While sampling in time and random assignment introduce an element of statistical rigor and extend the range of valid statistical inferences, the methodology may be disrupted if field interviewers change the date assigned for a sample location because access may be difficult in the specified period or for other reasons. Noteworthy examples in the NPS study included hurricanes, park closed due to fugitives from justice, space shuttle launches, and severe snowstorms. While some of these occurrences may be minimized with the temporal stratification and allocation, others are clearly beyond the control of the statistician. However, the sampling statistician should be involved in interviewer training to stress that modifications in the sample schedule should be avoided at all costs, and should monitor any changes that do occur.

ACKNOWLEDGEMENTS

The authors are grateful to the Associate Editor and the referees for comments that helped improve the paper's presentation.

REFERENCES

GOUGH, J.H., and GHANGURDE, P.D. (1977). Survey of Canadian residents returning by land. *Survey Methodology*, 3, 215-231.

KALTON, G. (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17, 183-194.

KISH, L., LOVEJOY, W., and RACKOW, P. (1961). A Multi-stage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.

LEVY, M.R. (1983). The methodology and performance of election day polls. *Public Opinion Quarterly*, 54-67.

NATIONAL PARK SERVICE (1994). Survey of Visitors to National Park Service Areas: Survey Findings. Report 94-2, No. 290940.12.

SUDMAN, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research*, 17, 423-431.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.