Catalogue 12-001 c.3

Survey Methodology

A Journal of Statistics Canada

December 1991 Volume 17 Number 2

STATISTICS CANADA STATISTIQUE CANADA
JAN 30 1992
LIBRARY
BIBLIOTHÈQUE

Statistics Canada Statistique Canada

Canada

Statistics Canada
Social Survey Methods Division

Survey Methodology

A Journal of Statistics Canada

December 1991 Volume 17 Number 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1991

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 Chief, Author Services, Publications Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

December 1991

Price: Canada: $35.00
United States: US$42.00
Other Countries: US$49.00

Catalogue 12-001

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal 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 B.N. Chinnappa, G.J.C. Hole, F. Mayda (Production Manager), R. Platek (Past Chairman), C. Patrick, D. Roy, M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

B. Afonja, *United Nations*
D.R. Bellhouse, *U. of Western Ontario*
D. Binder, *Statistics Canada*
E.B. Dagum, *Statistics Canada*
J.-C. Deville, *INSEE*
D. Drew, *Statistics Canada*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistics Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Laval University*
D.B. Rubin, *Harvard University*
I. Sande, *Bell Communications Research, U.S.A.*
C.E. Särndal, *University of Montreal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
C.M. Suchindran, *University of North Carolina*
J. Waksberg, *Westat Inc.*
K.M. Wolter, *A.C. Nielsen, U.S.A.*

Assistant Editors

J. Gambino, L. Mach, H. Mantel and A. Théberge, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal 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

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social 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 the Survey Methodology Journal (Catalogue No. 12-001) is $35 per year in Canada, US $42 in the United States, and US $49 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 of Statistics Canada
Volume 17, Number 2, December 1991

CONTENTS

Survey Methodology, December 1991
Vol. 17, No. 2, p. 121
Statistics Canada

In This Issue

This issue of Survey Methodology contains papers covering a broad range of topics. In the first paper, Alho investigates different estimators of the variance of the population size estimated using dual registration. The bias of the usual variance estimator, derived under the assumption of homogeneous capture probabilities, is investigated under population heterogeneity and two alternative variance estimators are proposed. The three estimators are applied to occupational disease data from Finland.

Rao and Shimizu compare three procedures for combining independent estimates obtained at successive time periods. They show that all three procedures produce an improvement over the use of the estimate based on only one occasion. The findings are illustrated by applying them to data from the American National Health Discharge Survey.

In their paper, Lavrakas, Settersten and Maier take a descriptive look at the problem of panel attrition in surveys, using data from two surveys carried out using random digit dialing. The paper gives the reader a good introduction to some of the causes of attrition, with suggestions on how its effects can be reduced.

Sutradhar, Dagum and Solomon give an exact test for the presence of significant stable seasonality for time series with seasonal patterns that are stable over time except for possible annual shifts. The assumptions of the standard ANOVA F-test used by the X-11-ARIMA seasonal adjustment method are violated when the residuals are autocorrelated. The exact test, however, takes into account the possibility of autocorrelated residuals. The exact and the standard tests are compared for several Canadian socioeconomic series.

A characteristic of quota sampling is the absence of randomized selection. Some kind of modelling must be imposed to construct estimators. The traditional approach is to use superpopulation modelling, and Deville proposes interesting extensions to this method. An approach suggested by the author is to model the sampling process. Comparisons with random sampling are made.

While household surveys are successfully used to collect data about human populations, they are not suitable for studying the characteristics of mobile human populations, such as visitors to museums or parks, shoppers, *etc.* Kalton describes different sample designs for surveys of flows of human populations and provides a number of examples of such surveys. The examples illustrate that field work considerations play an important role in the choice of a sample design.

Hidiroglou, Choudhry and Lavallée provide a sampling methodology for continuing subannual business surveys. A rotation scheme is suggested to maintain a representative sample through time. The properties of a number of estimators of totals for this sampling methodology have been evaluated in an empirical study which reflects a number of possible survey conditions.

Stasny, Goel and Rumsey use regression models to obtain small area estimates of wheat production when the data come from non-probabilistic sources. A simulation study compares the estimates obtained through this approach with the standard synthetic and direct estimators. Three scaling methods to satisfy additivity constraints are also compared.

In the last paper of this issue, Norris and Paton give an overview of Canada's five year old General Social Survey. They present a brief account of the information needs and discuss the five annual topics addressed by the survey. A description of the survey's methodology and experiences with the use of random digit dialing are presented. The authors' analysis of nonresponse rates over the life of the survey has implications for other telephone surveys.

The Editor

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 123–130
Statistics Canada

Variance Estimation in Dual Registration Under Population Heterogeneity

JUHA M. ALHO[1]

ABSTRACT

The usual dual system estimator for population size can be severely biased, if there is population heterogeneity in the capture probabilities. In this note we investigate the bias of the corresponding variance estimator under heterogeneity. We show that the usual estimator is conservative, *i.e.*, it gives too large values, if the two registration systems are negatively correlated, uncorrelated, or when the correlation is positive, but small. In the case of high positive correlation the usual estimator may yield too low values. Two alternative estimators are proposed. One is conservative under arbitrary heterogeneity. The other is conservative under Gaussian heterogeneity. The methods are applied to occupational disease data from Finland.

KEY WORDS: Capture-recapture; Dual system; Heterogeneity; Occupational diseases.

1. INTRODUCTION

Suppose there are N individuals in a closed population. The problem is to estimate the unknown N using dual registration. We sample twice with n_j individuals captured at the jth time, $j = 1, 2$. Let m be the number captured twice. Define indicator variables u_{ji} and m_i for $i = 1, \ldots, N$ such that $u_{ji} = 1$, if and only if individual i is captured at the jth time only, $j = 1, 2$; and $m_i = 1$, if and only if individual i is captured twice. Otherwise u_{ji} and m_i are zero. Define $n_{ji} = u_{ji} + m_i$ as the indicator of capture at the jth time, $j = 1, 2$. Let $M_i = u_{1i} + u_{2i} + m_i$ indicate capture at least once. Define the individual capture probabilities as $p_{ji} = E[n_{ji}]$, $j = 1, 2$; and $p_{12i} = E[m_i]$. Assume that the probabilities are strictly between zero and one. The fact that the probabilities are allowed to vary by individual indicates that we may have population heterogeneity in the capture probabilities. We complete the definition of the dual registration (or capture-recapture) model by assuming that the captures are independent for each individual, or $p_{12i} = p_{1i}p_{2i}$, and that the multinomial vectors

$$(u_{1i}, u_{2i}, m_i, 1 - M_i) \sim \text{Mult}(1; p_{1i}q_{2i}, p_{2i}q_{1i}, p_{1i}p_{2i}, 1 - \phi_i),$$

where $q_{ji} = 1 - p_{ji}$, $j = 1, 2$, and $\phi_i = p_{1i} + p_{2i} - p_{1i}p_{2i}$, are independent for $i = 1, \ldots, N$.

It is well-known that when capture probabilities do not vary by individual, or $p_{ji} = p_j$, $j = 1, 2$, the maximum likelihood estimator of N is $\hat{N} = n_1n_2/m$ (or more precisely, the largest integer short of this value; *cf.*, Feller 1968, p. 46). This classical estimator can be severely biased under population heterogeneity (Seber 1982, p. 565; Burnham and Overton 1979, Table 4, pp. 931-932). As shown, *e.g.*, in Example 1 below, under homogeneous capture probabilities the asymptotic variance of $\hat{N}$ is $\text{Var}(\hat{N}) = Nq_1q_2/(p_1p_2)$, where $q_j = 1 - p_j$, $j = 1, 2$. Then $\text{Var}(\hat{N})$ can be estimated by $V_1 = n_1n_2u_1u_2/m^3$ (Sekar and Deming 1949, pp. 114-115).

[1] Juha M. Alho, Institute for Environmental Studies and Department of Statistics, University of Illinois at Urbana-Champaign, 1101 W. Peabody Dr., Urbana IL, 61801, U.S.A.

The purpose of this note is to investigate the adequacy of the variance estimator V_1, and compare the bias of V_1 to the bias of $\hat{N}$. One motive for investigating V_1 is that it has not been previously known whether V_1 is adequate in the case in which there is population heterogeneity, but $\hat{N}$ is, nevertheless, consistent. This turns out to be the case. Similarly, it has not been clear when V_1 gives overestimates and thus can lead to valid confidence intervals, despite the bias of $\hat{N}$. This turns out to be possible for one-sided intervals in special circumstances.

In Section 2 we calculate the asymptotic variance of $\hat{N}$, as $N \to \infty$, and derive a conservative estimator V_2 for this variance under arbitrary heterogeneity. In other words, V_2 overestimates the true asymptotic variance. One might hope that an overestimate of variance could compensate for the typically negative bias of $\hat{N}$ and still yield valid confidence intervals. Unfortunately, this appears possible only when the bias of $\hat{N}$ is small, or when N is small. In Section 3 the adequacy of V_1 is studied under Gaussian heterogeneity and an estimator V_3 is derived, which is conservative under this restricted type of heterogeneity. Gaussianity *per se* is not required for the arguments, only that the moments of the pairs (p_{1i}, p_{2i}) agree with those of a bivariate Gaussian distribution. This setup permits the ready examination of the effect of correlation between p_{1i}'s and p_{2i}'s on variance estimation, because correlation is expressible in terms of just one parameter, the ordinary moment correlation coefficient. In Section 4 we compare the bias in variance estimates to the bias of $\hat{N}$ using empirical data relating to the registration of occupational diseases in Finland.

2. BIAS AND VARIANCE UNDER HETEROGENEITY

Define $\bar{p}_{jN}$ as the average probability of capture at the jth time, $j = 1, 2$; and let $\bar{p}_{12N}$ be the average of the products $p_{1i}p_{2i}$, $i = 1, \ldots, N$. Then, $C_N = \bar{p}_{12N} - \bar{p}_{1N}\bar{p}_{2N}$ is the covariance of the pairs (p_{1i}, p_{2i}). Assume that the limits $\bar{p}_{jN} \to \bar{p}_j$, $j = 1, 2$; $\bar{p}_{12N} \to \bar{p}_{12}$, and $C_N \to C$ exist. Then we have that $\hat{N}/N \to \bar{p}_1\bar{p}_2/\bar{p}_{12}$, so $\hat{N}/N - 1 \to -C/\bar{p}_{12}$, as $N \to \infty$. This is the asymptotic bias of the classical estimator under population heterogeneity. Interestingly, it only depends on the first two moments of the distribution of the pairs (p_{1i}, p_{2i}). As is well-known (Sekar and Deming 1949, pp. 105-106; Seber 1982, p. 86), when the covariance is zero ($C = 0$), then the classical estimator is consistent; if $C > 0$, $\hat{N}$ gives an underestimate; and if $C < 0$, it gives an overestimate. As noted above the adequacy of V_1, when the p_{ji}'s vary from one individual to the next but still $C = 0$, is of particular interest.

We shall now calculate the asymptotic variance of the classical estimator under our general heterogeneity model. Note that the finite variance does not exist, because there is a positive probability that $m = 0$. Therefore, "asymptotic variance" properly refers here to the variance of the limiting distribution rather than to limit of the variances, as $N \to \infty$.

Lemma 1. The asymptotic variance of $\hat{N}$ is

$$\mathrm{Var}(\hat{N}) = N\left\{\frac{\bar{p}_1^2\bar{p}_2^2}{\bar{p}_{12}^3} - \frac{\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}^2} - \frac{\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}^2} - \frac{\bar{p}_2^2}{\bar{p}_{12}^2}\bar{S}_1 - \frac{\bar{p}_1^2}{\bar{p}_{12}^2}\bar{S}_2 - \frac{\bar{p}_1^2\bar{p}_2^2}{\bar{p}_{12}^4}\bar{S}_3 + 2\left(\frac{\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}^3}\bar{S}_4 + \frac{\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}^3}\bar{S}_5\right)\right\},$$

where $\bar{S}_j = S_j/N$ for $j = 1, \ldots, 5$, with

$$S_1 = \sum_{i=1}^{N} p_{1i}^2, \quad S_2 = \sum_{i=1}^{N} p_{2i}^2, \quad S_3 = \sum_{i=1}^{N} p_{1i}^2 p_{2i}^2,$$

$$S_4 = \sum_{i=1}^{N} p_{1i}^2 p_{2i}, \quad S_5 = \sum_{i=1}^{N} p_{1i} p_{2i}^2.$$

The proof is sketched in the Appendix. We note that unlike the bias of $\hat{N}$ that depends on the first two moments of the pairs (p_{1i}, p_{2i}) only, $\mathrm{Var}(\hat{N})$ depends on moments up to fourth order. In special cases, such as the ones considered in Example 2 and Proposition 2, a simpler representation is possible.

Example 1. Suppose there is no heterogeneity in the probabilities, or $p_{ji} = p_j, j = 1, 2$. Then $\bar{p}_j = p_j, j = 1, 2$; $\bar{p}_{12} = p_1 p_2$; $\bar{S}_j = p_j^2, j = 1,2$; $\bar{S}_3 = p_1^2 p_2^2$, $\bar{S}_4 = p_1^2 p_2$, and $\bar{S}_5 = p_1 p_2^2$. Hence, the asymptotic variance is $\mathrm{Var}(\hat{N}) = N(1 - p_1 - p_2 + p_1 p_2)/(p_1 p_2) = Nq_1 q_2/(p_1 p_2)$. Consistent estimators for $Np_1 p_2$ and Np_j are m and $n_j, j = 1, 2$. In other words, $Np_j/n_j \to 1, j = 1, 2$, and $Np_1 p_2/m \to 1$, as $N \to \infty$. This gives us V_1 as an estimator for $\mathrm{Var}(\hat{N})$.

Example 2. Suppose that the pairs (p_{1i}, p_{2i}), $i = 1, \ldots, N$, are independent in the sense that the distribution of p_{1i}'s is the same for each distinct value of the p_{2i}'s. Then, $\bar{p}_{12} = \bar{p}_1 \bar{p}_2$, $\bar{S}_3 = \bar{S}_1 \bar{S}_2$, $\bar{S}_4 = \bar{p}_2 \bar{S}_1$, $\bar{S}_5 = \bar{p}_1 \bar{S}_2$. Substituting into the Lemma we get

$$\mathrm{Var}(\hat{N}) = N\left(\frac{1}{\bar{p}_1\bar{p}_2} - \frac{1}{\bar{p}_2} - \frac{1}{\bar{p}_1} - \frac{\bar{S}_1\bar{S}_2}{\bar{p}_1^2\bar{p}_2^2} + \frac{\bar{S}_1}{\bar{p}_1^2} + \frac{\bar{S}_2}{\bar{p}_2^2}\right)$$

$$= N\left(\frac{\bar{q}_1\bar{q}_2}{\bar{p}_1\bar{p}_2} - cv(p_{1i})^2\, cv(p_{2i})^2\right),$$

where $cv(p_{ji}) = (\bar{S}_j - \bar{p}_j^2)/\bar{p}_j$, is the coefficient of variation of the p_{ji}'s, $j = 1, 2$. Obviously, $\mathrm{Var}(\hat{N}) \le N\bar{q}_1\bar{q}_2/(\bar{p}_1\bar{p}_2)$. A comparison with Example 1 shows that V_1 is a conservative estimator of $\mathrm{Var}(\hat{N})$ (*i.e.*, V_1 is asymptotically too large), when p_{1i}'s are independent of p_{2i}'s. Another way of saying this is that, given the means $\bar{p}_j$, $j = 1, 2$, the largest value of the variance is obtained at homogeneity. This is analogous to the variance of the number of successes in Bernoulli trials with variable probabilities of success, *cf.* Feller 1968, pp. 230-231. A comparison with Example 1 shows that V_1 is a conservative estimator of $\mathrm{Var}(\hat{N})$ (*i.e.*, V_1 is asymptotically too large), when the pairs (p_{1i}, p_{2i}) are independent. Note that the independence condition implies that $C = 0$.

When the probabilities are not independent, the classical variance estimator is not guaranteed to be conservative. A conservative estimator exists, however. It is obtained by majorizing $\mathrm{Var}(\hat{N})$ by a quantity that can be estimated in terms of the observable variables. We prove in the Appendix the following general proposition.

Proposition 1. A conservative estimator of $\mathrm{Var}(\hat{N})$ is

$$V_2 = (n_1^2 n_2^2 + n_2^2 m u_1 + n_1^2 m u_2)/m^3,$$

where $u_j = n_j - m, j = 1, 2$.

3. GAUSSIAN HETEROGENEITY

We shall now turn to a special case in which the sample moments of the pairs (p_{1i}, p_{2i}), $i = 1, \ldots, N$, agree with those of a bivariate normal, or Gaussian, distribution. This will permit a much sharper specification of a conservative variance estimator than the one obtained in the general case above. Assume that

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \nu_1^2\mu_1^2 & \rho\nu_1\nu_2\mu_1\mu_2 \\ \rho\nu_1\nu_2\mu_1\mu_2 & \nu_2^2\mu_1^2 \end{bmatrix}\right),$$

where $|\rho| < 1$, and $0 < \mu_j < 1, j = 1, 2$. Note that ν_j's can be interpreted as the coefficients of variation of the distributions of p_{ji}'s. Write $\bar{S}_j = S_j/N$ for $j = 1, \ldots, 5$, as before. Then substitute the moments of the bivariate normal distribution into Lemma 1 as follows,

$$\bar{p}_j = E[X_j] = \mu_j, j = 1, 2;$$

$$\bar{S}_j = E[X_j^2] = \mu_j^2(1 + \nu_j^2), j = 1, 2;$$

$$\bar{p}_{12} = E[X_1X_2] = \mu_1\mu_2(1 + \rho\nu_1\nu_2);$$

$$\bar{S}_3 = E[X_1^2X_2^2] = \mu_1^2\mu_2^2(1 + \nu_1^2 + \nu_2^2 + 4\rho\nu_1\nu_2 + (2\rho^2 + 1)\nu_1^2\nu_2^2);$$

$$\bar{S}_4 = E[X_1^2X_2] = \mu_1^2\mu_2(1 + 2\rho\nu_1\nu_2 + \nu_1^2);$$

$$\bar{S}_5 = E[X_1X_2^2] = \mu_1\mu_2^2(1 + 2\rho\nu_1\nu_2 + \nu_2^2).$$

Straightforward, but slightly tedious calculations prove then the following proposition (details omitted).

Proposition 2. With the above assumptions

$$\mathrm{Var}(\hat{N}) = A_1A_2 + RN,$$

where

$$A_1 = N/(1 + \rho\nu_1\nu_2)^2;$$

$$A_2 = [1 - (\mu_1 + \mu_2)(1 + \rho\nu_1\nu_2 + \mu_1\mu_2(1 + \rho\nu_1\nu_2)^2]/[\mu_1\mu_2(1 + \rho\nu_1\nu_2)^2];$$

$$R = (2\rho\nu_1\nu_2 + 3\rho^2\nu_1^2\nu_2^2 - \rho^2\nu_1^4\nu_2^2 - \rho^2\nu_1^2\nu_2^4 - \nu_1^2\nu_2^2)/(1 + \rho\nu_1\nu_2)^4.$$

We can evaluate the classical variance estimator $V_1 = n_1n_2u_1u_2/m^3$ using this result. Note first that $\{n_1n_2/m\}/A_1 \to 1$, as $N \to \infty$. Similarly, $\{u_1u_2/m^2\}/A_2 \to 1$. This proves the following corollary to Proposition 2: $(V_1 - \mathrm{Var}(\hat{N}))/N \to -R$, as $N \to \infty$. For example, if $\rho = 0$, then $-R = \nu_1^2\nu_2^2$, so that V_1 is seen to overestimate the asymptotic variance. This is in accordance with Example 2.

How reasonable is the assumption of Gaussian moments? Certainly the capture probabilities cannot have strictly Gaussian distributions, because the Gaussian distribution always puts some probability mass outside the unit interval. On the other hand, suppose we generate the p_{ji}'s by taking $\text{logit}(p_{ji}) = a_j + b_j Y_{ji}$, where the pairs (Y_{1i}, Y_{2i}) are a sample from a bivariate normal distribution with mean zero, unit variances, and correlation ρ. If we have the relations $a_j = \text{logit}(\mu_j), j = 1, 2$, and $b_j = v_j(1 + \mu_j)^2$, then the assumption of Gaussian moments is approximately true. In fact, even the distribution of the pairs (p_{1i}, p_{2i}) is in that case approximately bivariate Gaussian.

Let us consider the adequacy of V_1 further, under the Gaussian moments. The fact that probabilities are constrained to be between 0 and 1 means that μ_j's are between zero and one. Moreover, to be sure that most of the probability mass is in the unit square, let us assume that $0 < v_j \leq ½, j = 1, 2$. If μ_j's are close to one, a much smaller upper bound would be needed. Assume now that $\rho \leq 0$. Then, one can show that

$$-R \geq (\rho^2 v_1^4 v_2^2 + \rho^2 v_1^2 v_2^4)/(1 + \rho v_1 v_2)^4 > 0,$$

so that V_1 overestimates $\text{Var}(\hat{N})$ for $\rho \leq 0$ also. Note that by continuity V_1 must overestimate $\text{Var}(\hat{N})$ for some positive values of ρ, as well.

One can show that $R = R(\rho)$ is an increasing function of ρ for at least $\rho > 0$. In the limit we have

$$-R(\rho) \to (-2v_1 v_2 - 2v_1^2 v_2^2 + v_1^4 v_2^2 + v_1^2 v_2^4)/(1 + v_1 v_2)^4,$$

as $\rho \to 1$. When $0 < v_j \leq ½, j = 1, 2$, the smallest value of the above limit occurs at $v_1 = v_2 = ½$. The minimum value is $-152/625 > -1/4$. Consequently, for $\rho > 0$, V_1 can either underestimate or overestimate $\text{Var}(\hat{N})$.

The practical implications of the above results are as follows. First, if $\rho \leq 0$, then $\hat{N}$ is either consistent or it overrestimates N and V_1 gives an overestimate of the variance, so we can calculate a conservative upper confidence limit for N. When $\rho > 0$, $\hat{N}$ gives an underestimate of N. If, in addition, ρ is small, then V_1 gives an overestimate, and we can get a conservative lower confidence limit for N. Obviously, these are rather special circumstances that one would not expect to be of wide practical utility.

Under the present model the asymptotic bias of V_1 is $> -N/4$ for all values of ρ. We can derive a conservative variance estimator by noting that in the Gaussian case the asymptotic relative bias of $\hat{N}$ is $-\rho v_1 v_2/(1 + \rho v_1 v_2) \geq -1/5$. Hence, asymptotically $5\hat{N}/4 \geq N$. A conservative estimator of $\text{Var}(\hat{N})$ is, for example, $V_3 = V_1 + 5\hat{N}/16$. This can be much smaller than V_2 indicating that the Gaussian assumption is a very powerful one.

4. AN APPLICATION TO OCCUPATIONAL DISEASE REGISTRATION DATA

To get an idea of how large the biases may be in practice, let us look at occupational disease data from Finland as an example. The Finnish Register of Occupational Diseases has been in operation since 1964. It is kept by the Institute of Occupational Health in Helsinki. Since 1975 the number of new cases reported to the Register has varied from about 4,000 to over 7,000 annually (0.2 - 0.4 % of the employed population). Noise-induced hearing loss, diseases caused by repetitive of monotonous work (epicondylitis, bursitis, tendinivaginitis), and skin diseases

are the major diagnostic groups (*cf.* Vaaranen *et al.* 1985). The Register can be viewed as a dual registration system, because each case of disease should, under existing regulations, be reported to the Register both from the appropriate insurance company and the examining physician.

It is likely that the probability of reporting a case depends on diagnosis, for example. Indeed, based on data from the year 1981 we get the following statistics. Reports from the insurance companies, $n_1 = 3{,}769$; reports from the physicians, $n_2 = 3{,}053$; and cases reported from both sources, $m = 1{,}591$. Thus the usual dual registration estimate is $\hat{N} = 7{,}232$ with $V_1^{1/2} = 97$, $V_2^{1/2} = 222$, and $V_3^{1/2} = 108.0$. The closeness of V_3 to V_1 is striking. Stratifying the data into four categories by diagnosis (the three diagnostic groups mentioned above, and the remaining "other" category) yields the following estimates. Noise-induced hearing loss: $\hat{N} = 2{,}230$, $V_1^{1/2} = 33.4$, $V_2^{1/2} = 47.2$, and $V_3^{1/2} = 42.6$; diseases caused by repetitive or monotonous work: $\hat{N} = 3{,}572$, $V_1^{1/2} = 201.4$, $V_2^{1/2} = 303.8$, and $V_3^{1/2} = 204.2$; skin diseases: $\hat{N} = 1{,}441$, $V_1^{1/2} = 30.9$, $V_2^{1/2} = 86.2$, and $V_3^{1/2} = 37.5$; other diseases $\hat{N} = 1{,}015$, $V_1^{1/2} = 32.7$, $V_2^{1/2} = 79.1$, and $V_3^{1/2} = 37.2$. Adding the results yields the following estimates for the total number of diseases: $\hat{N} = 8{,}258$, $V_1^{1/2} = 209.0$, $V_2^{1/2} = 340.3$, and $V_3^{1/2} = 215.2$. We see that diseases caused by repetitive or monotonous work are underreported to a particularly great extent.

The analysis was extended further by stratifying the data by diagnosis (4 categories), insurance company (11 categories), and main groups of industry (7 categories). *A priori*, these factors could be thought to have an influence on reporting probabilities. However, the stratification did not alter the point estimate materially. It did increase the estimated standard deviations by over a third, apparently because some of the strata became very small. We conclude that the bias in the point estimator caused by diagnosis is the dominant source of error in the classical estimator in this application.

The same data were further analyzed using a logistic regression technique that allows us to take into account observable population heterogeneity due to both discrete and continuous explanatory variables. In this application age was shown to have an effect on reporting probabilities within the diagnostic groups for one source of information, but not for the other. Therefore, the point estimates remained unchanged and the conclusion regarding the role of diagnosis could not be refuted (Alho 1990).

5. DISCUSSION

Our theoretical results indicate that the usual variance estimator V_1 is conservative when the two registration systems are negatively correlated or independent. By continuity the estimator may be conservative also when the correlation is positive but small. Under high positive correlation V_1 gives too low values. We introduced an alternative estimator V_2, which is conservative under arbitrary population heterogeneity. However, it appears to be unduly conservative in view of the numerical comparisons with V_3, which is guaranteed to be conservative under Gaussian heterogeneity. The closeness of V_3 to V_1 suggests that, in practice, V_1 may be fairly robust against population heterogeneity.

Unfortunately, even the use of the conservative estimator V_2 would not have been sufficient to cover the bias in the classical point estimator in our empirical example. Perhaps this was to be expected, since the bias of $\hat{N}$ and the degree of overestimation provided by V_2 are both of order N. Hence, the use of V_2 inflates the width of a confidence interval by a factor of order $N^{1/2}$ only. Therefore, V_2 can compensate for the bias of $\hat{N}$, if the bias is small, or if N itself is small. Hence, it seems that the successfull application of the dual registration method requires that either we have roughly uncorrelated registration systems, or that the heterogeneity is

observable. In the latter case we may use stratification as suggested already by Sekar and Deming (1949), or logistic regression modeling as suggested by Huggins (1989) and Alho (1990), to adjust for the bias of the classical estimator of population size.

ACKNOWLEDGEMENT

The author would like to thank Bruce Spencer and an anonymous referee for comments that helped to improve the presentation. Part of the empirical results were first presented at the 11th Nordic Conference on Mathematical Statistics in Uppsala, Sweden, in June 1986.

APPENDIX

Proof of Lemma 1. Apply a linear Taylor-series development to $\hat{N} = n_1 n_2/m$ at $E[n_1]$ $E[n_2]/E[m] = N\bar{p}_1\bar{p}_2/\bar{p}_{12}$, or

$$\hat{N} \approx \frac{N\bar{p}_1\bar{p}_2}{\bar{p}_{12}} + \frac{\bar{p}_2}{\bar{p}_{12}}(n_1 - N\bar{p}_1) + \frac{\bar{p}_1}{\bar{p}_{12}}(n_2 - N\bar{p}_2) - \frac{\bar{p}_1\bar{p}_2}{\bar{p}_{12}^2}(m - N\bar{p}_{12}).$$

Hence, we have

$$E\left[\left(\hat{N} - \frac{N\bar{p}_1\bar{p}_2}{\bar{p}_{12}}\right)^2\right] \approx \left(\frac{\bar{p}_2}{\bar{p}_{12}}\right)^2 \mathrm{Var}(n_1) + \left(\frac{\bar{p}_1}{\bar{p}_{12}}\right)^2 \mathrm{Var}(n_2) + \left(\frac{\bar{p}_1\bar{p}_2}{\bar{p}_{12}^2}\right)^2 \mathrm{Var}(m)$$

$$- \frac{\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}} \mathrm{Cov}(n_1, m) - 2\frac{\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}^3} \mathrm{Cov}(n_2, m).$$

Under our independence assumptions $\mathrm{Var}(n_j) = N\bar{p}_j - S_j, j = 1, 2$; $\mathrm{Var}(m) = N\bar{p}_{12} - S_3$, $\mathrm{Cov}(n_1, m) = -S_4 + N\bar{p}_{12}$, and $\mathrm{Cov}(n_2, m) = -S_5 + N\bar{p}_{12}$. Substituting these into the mean squared error gives the result.

Proof of Proposition 1. We ignore the negative term containing S_3 in Lemma 1. Since $0 < p_{ji} < 1$, we have $S_4 < N\bar{p}_{12}$, and $S_4 < S_1$. Therefore,

$$\frac{2\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}^3} S_4 < \frac{\bar{p}_1\bar{p}_2^2}{\bar{p}_{12}^3} N\bar{p}_{12} + \frac{\bar{p}_2^2}{\bar{p}_{12}^2} S_1 + \left(\frac{\bar{p}_1 - \bar{p}_{12}}{\bar{p}_{12}}\right) \frac{\bar{p}_2^2}{\bar{p}_{12}^2} N\bar{p}_{12}.$$

Similarly,

$$\frac{2\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}} S_5 < \frac{\bar{p}_1^2\bar{p}_2}{\bar{p}_{12}^3} N\bar{p}_{12} + \frac{\bar{p}_1^2}{\bar{p}_{12}^2} S_2 + \left(\frac{\bar{p}_2 - \bar{p}_{12}}{\bar{p}_{12}}\right) \frac{\bar{p}_1^2}{\bar{p}_{12}^2} N\bar{p}_{12}.$$

Substituting these bounds to the expression of Lemma 1 we get

$$\mathrm{Var}(\hat{N}) < \frac{\bar{p}_1^2\bar{p}_2^2}{\bar{p}_{12}^3} N + \frac{(\bar{p}_1 - \bar{p}_{12})\bar{p}_2^2}{\bar{p}_{12}^2} N + \frac{(\bar{p}_2 - \bar{p}_{12})\bar{p}_1^2}{\bar{p}_{12}^2} N.$$

Estimating $N\bar{p}_j$ by $n_j, j = 1, 2$; and $N\bar{p}_{12}$ by m we get the result.

REFERENCES

ALHO, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.

BURNHAM, P.K., and OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927-936.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, (Vol. I, 3rd ed.). New York: Wiley.

HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance*, (2nd ed.). New York: Griffin.

SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

VAARANEN, V., VASAMA, M., and ALHO, J. (1985). Occupational diseases in Finland in 1984. Reviews 11, Institute of Occupational Health, Helsinki.

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 131–141
Statistics Canada

Combining Estimates from Surveys

PODURI S.R.S. RAO and I.M. SHIMIZU[1]

ABSTRACT

For estimating the proportion and total of an item for the present occasion, independent estimates at the current and previous occasions are combined through three different procedures. In the first one, trend over the occasions is utilized. For the second one, the One-Way Random Effects Model is employed. The third procedure uses the Empirical Bayes approach. All the three procedures are seen to perform better than the sample estimates obtained from the data of the current occasion alone. Advantages of these methods and their limitations are discussed. All the procedures are illustrated with the data from the National Health Discharge Survey.

KEY WORDS: Trend; Weighted least squares; Random effects; Improved estimation; Biases; Mean square errors.

1. INTRODUCTION

In several national surveys, independent samples are obtained at successive time periods. In this paper, information from the past surveys is utilized to improve the estimates for the current period. For the sake of illustration, we consider the National Health Discharge Survey (NHDS) in the U.S. In this survey, which has been recently redesigned, a three stage sampling design is used with geographical regions as the Primary Sampling Units (PSU's) at the first stage. Hospitals and discharges are selected at the second and third stages respectively. The survey collects information on various items of the patients like age, sex, racial characteristics, length of stay, diagnosis, and surgical and non-surgical procedures. The selected PSU's and hospitals remain in the study for a certain number of years. Independent samples of discharges are obtained every year from the selected hospitals. Shimizu (1987) presents further details of the redesign of the NHDS.

At present, for a given hospital, estimates of the proportions for the different items for the current year are obtained only from the data of this year. National estimates are obtained by suitably weighting these proportions with the reciprocals of the probabilities of selection of the hospitals and the PSU's. However, Bean (1987) found that for most of the items the estimates are somewhat correlated over the years. For the sake of illustration, sample proportions obtained from the NHDS for 1977-86 for Acute Myocardial Infraction (AMI) and Mental Disorders (MDS) are presented in Table 1 for three hospitals and they are exhibited in Figures 1 and 2. Examination of the proportions for these three and 17 more hospitals suggested that the inclusion of past information can increase the precision of the estimates for the current year.

It should be cautioned that the sample proportions in Table 1 or Figures 1 and 2 should not be used to make inferences regarding the increase or decrease of AMI or MDS in the entire population.

[1] Poduri S.R.S Rao, Department of Statistics, Hylan 703, University of Rochester, Rochester NY, 14618 U.S.A., and I.M. Shimizu, National Center for Health Statistics, Office of Research and Methodology 1-68, 3700 East-West Highway, Hyattsville MD, 20782, U.S.A.

Table 1

Data from the National Health Discharge Survey for 1977-86
Sample totals and proportions for Acute Myocardial
Infraction (AMI) and Mental Disorders
(MDS) for three hospitals

Year	No. of discharges N	Sampled No. of discharges n	AMI		MDS	
			Total	Sample proportion	Total	Sample proportion
1977	9,416	276	5	.018	37	.13
1978	10,234	266	7	.026	24	.09
1979	9,354	294	9	.031	39	.13
1980	10,372	327	9	.028	41	.13
1981	10,712	342	8	.023	45	.13
1982	10,683	309	9	.029	43	.14
1983	10,935	360	7	.019	46	.15
1984	10,090	330	6	.018	50	.15
1985	10,431	297	8	.027	41	.14
1986	10,247	264	4	.015	35	.13
1977	6,720	474	9	.019	18	.04
1978	6,710	470	14	.030	25	.05
1979	6,970	495	8	.016	28	.06
1980	6,794	466	14	.030	29	.06
1981	7,055	486	9	.019	34	.07
1982	6,265	442	9	.020	24	.05
1983	6,234	442	10	.023	28	.06
1984	6,221	439	9	.021	15	.03
1985	6,063	375	8	.021	19	.05
1986	5,781	371	4	.011	12	.03
1977	6,400	606	21	.0347	41	.0677
1978	6,286	635	23	.0362	42	.0661
1979	6,494	554	12	.0217	27	.0487
1980	6,813	571	17	.0298	25	.0438
1981	7,430	729	14	.0192	32	.0439
1982	7,267	712	20	.0281	39	.0548
1983	7,110	694	23	.0331	43	.0620
1984	7,268	718	35	.0487	29	.0404
1985	6,716	657	19	.0289	45	.0685
1986	6,464	655	21	.0321	33	.0504

In this article, we examine three procedures for improving the estimates for a specified hospital by utilizing the information from the current and the previous years. In the first method, estimates of the proportions are obtained from the linear trend over the years and the Weighted Least Squares Method. If there is a significant positive or negative trend over the years, this method will have higher precision than the sample estimate of the current period. If the trend is not pronounced, the increase in precision will be negligible, as expected.

For the second procedure, the One-Way Random Effects Model with unequal variances is used to combine the information. Yates and Cochran (1938) and Cochran (1954) suggested this type of procedure for combining information from experiments conducted at different time periods and locations. While the Analysis of Variance (ANOVA) method had been used

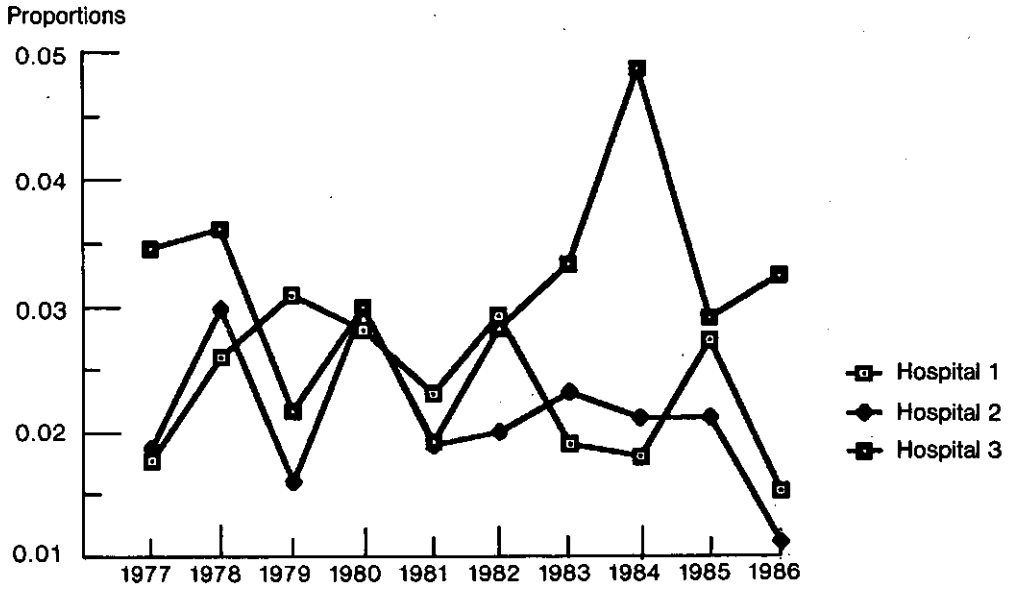


Figure 1. Proportions for AMI: 1977-86

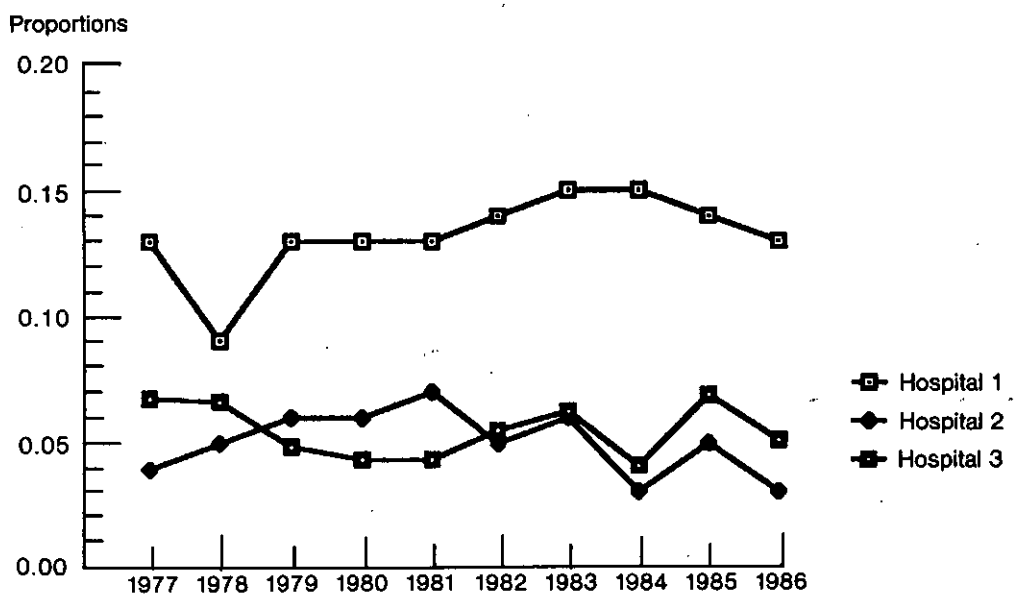


Figure 2. Proportions for MDS: 1977-86

for quite some time for this purpose, C.R. Rao (1970) suggested the Minimum Norm Quadratic Unbiased Estimation (MINQUE) and demonstrated its advantages. P.S.R.S. Rao, Kaplan and Cochran (1981) examined the relative merits of the ANOVA, MINQUE and several related procedures. We have employed the estimation procedures related to these methods. The estimate for the proportion obtained by any of these procedures is a weighted combination of the estimates of the different time periods. The weights depend on both the between and within variances of the time periods. In the third procedure, the Empirical Bayes approach is used to estimate the proportions for the current period.

We denote the above three procedures by TR, VC and EB respectively. The notation is presented in Section 2. The sample estimator for the proportion and its variance are given in Section 3. The above three estimation procedures along with the expressions for their Standard Errors (S.E.'s) are presented in Sections 4, 5 and 6. We have used these expressions to compute for 1986 the sample proportions, the above three types of estimates, and their S.E.'s for 20 hospitals in the NHDS. These estimates for the three hospitals mentioned earlier are presented in Table 2 for AMI and Table 3 for MDS. Results from the entire study are described in Section 7. The final section contains a discussion of the results and topics for further research.

Table 2

Estimates of the Proportions for 1986 and S.E.'s (bottom figures) for Acute Myocardial Infractions (AMI)

Hospital	Sample proportion	Trend estimate	Variance components estimate	Bayes estimate
1	.0152	.0196	.0224	.0224
	.0070	.0046	.0026	.0003
2	.0108	.0162	.0204	.0203
	.0048	.0036	.0031	.0003
3	.0321	.0319	.0304	.0309
	.0060	.0038	.0028	.0037

Table 3

Estimates of the Proportions for 1986 and S.E.'s (bottom figures) for Mental Disorders (MDS)

Hospital	Sample proportion	Trend estimate	Variance components estimate	Bayes estimate
1	.1326	.1431	.1292	.1292
	.0205	.0115	.0060	.0010
2	.0323	.0437	.0500	.0427
	.0087	.0056	.0039	.0057
3	.0504	.0496	.0534	.0523
	.0080	.0049	.0032	.0048

It should be mentioned that for the problem considered in this paper, the samples are drawn independently at the different time periods. Secondly, the population proportions for the previous periods are not known. Because of these reasons, the usual ratio and regression methods cannot be employed to improve the accuracy of the estimators for the current period. For the same reasons, the estimation procedures suggested in the literature for the rotation sampling schemes cannot be used in this situation. In spite of these difficulties, the three methods considered in this paper can be used to estimate the population quantities with a high accuracy. When summary figures at the different periods are available, public and private users can obtain these estimates and their standard errors without much difficulty. These procedures can also be used when there is nonresponse during some years - some of the hospitals do not provide information to the survey during some years.

2. NOTATION

We present in this section the notation for a selected PSU. Let y_{itj} denote the jth observation on the sampled discharge on an item like the number of surgical cases at time $t = (1, 2, \ldots, T)$, from the ith hospital, $i = (1, 2, \ldots, K)$, which has N_{it} discharges. Note that K may change over the years due to nonresponse or the addition of new hospitals.

The total and mean at time t are

$$Y_{it} = \sum_{1}^{N_{it}} y_{itj} \tag{1}$$

and

$$\bar{Y}_{it} = Y_{it}/N_{it}. \tag{2}$$

The total and mean of the sample of size n_{it} from the N_{it} discharges are

$$y_{it} = \sum_{1}^{n_{it}} y_{itj} \tag{3}$$

and

$$\bar{y}_{it} = y_{it}/n_{it}. \tag{4}$$

To estimate the total number and proportion for a specified item, let $y_{itj} = 1$ if the observation belongs to that item, and zero otherwise. With this notation, the total and proportion for an item at time t can be written as A_{it} and $P_{it} = A_{it}/N_{it}$. Note that P_{it} is the same as $\bar{Y}_{it}$.

In the following four sections, for the sake of convenience, we suppress the subscript i and describe the estimators for a given hospital.

3. SAMPLE PROPORTION

An unbiased estimator of the proportion P_t for an item like AMI or MDS is

$$\hat{P}_t = a_t/n_t, \tag{5}$$

where a_t is the number of cases of that item observed in the n_t sample discharges. The variance of $\hat{P}_t$ and its unbiased estimator are

$$V(\hat{P}_t) = \frac{N_t - n_t}{N_t - 1} \frac{P_t(1 - P_t)}{n_t} \tag{6}$$

and

$$v(\hat{P}_t) = (1 - f_t) \frac{\hat{P}_t(1 - \hat{P}_t)}{n_t - 1}, \tag{7}$$

where $f_t = n_t/N_t$. Note that $\hat{P}_t$ is the same as $\bar{y}_t = \sum_1^{n_t} y_{tj}/n_t$.

4. LINEAR TREND

The sample observations $y_{tj}, j = (1, 2, \ldots, n_{tj})$ can be written as

$$y_{tj} = \mu_t + \epsilon_{tj}, \tag{8}$$

where μ_t is the mean for the ith hospital at the tth period, and ϵ_{tj} is the random error with expectation zero and variance $\sigma_t^2 = P_t(1 - P_t)$. Since the samples are drawn independently during each year, the errors ϵ_{tj} are uncorrelated from one year to another.

With the assumption of a linear trend, the sample mean can be expressed as

$$\bar{y}_t = \alpha + \beta x_t + \bar{\epsilon}_t, \tag{9}$$

where $x_t = t$ and $\bar{\epsilon}_t = \sum_1^{n_t} \epsilon_{tj}/n_t$. Further, $V(\bar{\epsilon}_t) = (N_t - n_t)\sigma_t^2/(N_t - 1)n_t = 1/W_t$. Note that with the zero-one notation, $\bar{y}_t$ is the same as $\hat{P}_t$. The WLS estimators of β and α are

$$\hat{\beta} = \frac{\sum W_t(x_t - \bar{x})\bar{y}_t}{\sum W_t(x_t - \bar{x})^2} \tag{10}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{11}$$

where $\bar{x} = \sum W_t x_t / \sum W_t$ and $\bar{y} = \sum W_t \bar{y}_t / \sum W_t$.

Estimator of μ_t is

$$\begin{aligned} \hat{\mu}_t &= \hat{\alpha} + \hat{\beta} x_t \\ &= \bar{y} + \hat{\beta}(x_t - \bar{x}). \end{aligned} \tag{12}$$

This is the Trend Estimator (TR). Estimators of this type for infinite populations have been examined in the literature; see Carroll and Rupert (1988), for instance.

We have obtained the estimate of μ_t from this expression by replacing W_t with $w_t = (1 - f_t)\,\hat{\sigma}_t^2/n_t$, where $\hat{\sigma}_t^2 = \hat{P}_t(1 - \hat{P}_t)$. If it can be assumed that for large N_t the distribution of y_{tj} is normal, $\bar{y}_t$ will be independent of w_t. In this case, the expression in (12) remains unbiased for μ_t. Even if the assumption of normality is not valid, it can be seen that w_t approaches W_t for large n_t and hence the expression in (12) with the estimated weights approaches μ_t.

The variance of the above estimator is

$$V(\hat{\mu}_t) = \frac{1}{\sum W_t} + \frac{(x_t - \bar{x})^2}{\sum W_t(x_t - \bar{x})^2}. \tag{13}$$

We have estimated this variance by replacing W_t by w_t. The bias in the resulting estimator will be small for large n_t.

For the illustration in this article, $t = (1, 2, \ldots, 10)$, that is, $T = 10$. For 1986, we have found the estimate for the proportion of an item and its S.E. from (12) and (13) with $x_t = 10$.

5. VARIANCE COMPONENTS MODEL

Examination of the proportions for the AMI and MDS of the 20 hospitals for the ten years showed no specific linear or nonlinear trend. For all of them the patterns somewhat resembled those of the three hospitals, presented in Figures 1 and 2. These observations indicated that the proportion for AMI or MDS for the current year can be obtained by combining the information from all the ten years. The One-Way Random Effects Model can be used for this purpose.

The model in (8) can be written as

$$\begin{aligned} y_{tj} &= \mu + (\mu_t - \mu) + \epsilon_{tj} \\ &= \mu + \alpha_t + \epsilon_{tj}. \end{aligned} \tag{14}$$

If μ_t is considered to be randomly drawn from a population with mean μ, the random effect α_t will have mean zero and variance σ_α^2. It is assumed to be independent of ϵ_{tj}. The sample mean (proportion) can now be written as

$$\bar{y}_t = \mu + \alpha_t + \bar{\epsilon}_t, \tag{15}$$

where $\bar{\epsilon}_t$ has mean zero and variance $(1 - f_t)\,\sigma_\alpha^2/n_t$. Thus, from (15),

$$V(\bar{y}_t) = \sigma_\alpha^2 + (N_t - n_t)\sigma_t^2/(N_t - 1)n_t = \frac{1}{U_t}. \tag{16}$$

The WLS estimator of μ is

$$\hat{\mu} = \frac{\sum U_t \bar{y}_t}{\sum U_t}. \tag{17}$$

This is the Variance Components Estimator (VC) and its variance is

$$V(\hat{\mu}) = 1/\sum U_t. \tag{18}$$

For obtaining the mean in (17) and its variance in (18), we have replaced σ_t^2 by its estimate $\hat{P}_t(1 - \hat{P}_t)$. Procedures like the ANOVA and MINQUE are available for estimating σ_α^2. The MINQUE depends on the *a priori* values r_t of $(\sigma_t^2/\sigma_\alpha^2)$. A related procedure called the Unweighted Sums of Squares (USS) method does not depend on r_t and it is described below. P.S.R.S. Rao, Kaplan and Cochran (1981) found that this method provides estimates for σ_α^2 comparable to the ANOVA and MINQUE, unless n_t or r_t is very small. The USS is computationally less cumbersome than the MINQUE. With $\bar{y}^* = (\sum \bar{y}_t)/T$, from (15),

$$E[\sum(\bar{y}_t - \bar{y}^*)^2] = (T - 1)\sigma_\alpha^2 + (T - 1)(\sum v_t)/T, \tag{19}$$

where $v_t = (N_t - n_t)P_t(1 - P_t)/(N_t - 1)n_t$. The USS estimator for σ_α^2 is

$$\hat{\sigma}_\alpha^2 = \sum(\bar{y}_t - \bar{y}^*)^2/(T - 1) - (\sum \hat{v}_t)/T, \tag{20}$$

where $\hat{v}_t = (1 - f_t)\hat{P}_t(1 - \hat{P}_t)/(n_t - 1)$. If N_t is large relative to n_t, the sampling fraction f_t can be set to zero. We have estimated U_t from (16) by estimating σ_α^2 from (20) and the second term by $\hat{v}_t$. Utilizing this estimate of U_t, we have estimated μ from (17) and its variance from (18). If σ_α^2 is much larger than v_t, the estimator $\hat{\mu}$ in (17) will be close to $\bar{y}^*$. In this case, estimation of U_t as described above can be expected to have almost no effect on $\hat{\mu}$. Since $\hat{\mu}$ depends only on the relative values of U_t, this conclusion can be expected to be valid even when σ_α^2 is not considerably larger than v_t. Thus, estimation of U_t can be expected to result in only a negligible bias for $\hat{\mu}$.

As is well-known, all the procedures for estimating σ_α^2 unbiasedly can result in negative estimates. In such a case, we have employed the usual practice of substituting a small positive quantity for the negative estimate. In Rao *et al.* (1981) it was found that unless σ_α^2 is very small, this adjustment results in only a negligible bias for $\hat{\sigma}_\alpha^2$ and an insignificant increase in its standard error. Further, unless σ_α^2 is small, the difference in the MSE of $\hat{\mu}$ for the USS and other methods of estimating U_t was found to be negligible.

6. BAYES' ESTIMATOR

The discussion in the beginning of Section (5) suggests that μ_t can be assumed to have a prior distribution with mean μ and variance σ_α^2. With the assumptions that for large N_t the distribution of y_{tj} is normal with mean μ_t and variance σ_t^2, and that the prior distribution of μ_t is also normal, the Bayes' Estimator for μ_t is

$$B_t = E(m_t \mid \bar{y}_t) = (1 - a_t)\bar{y}_t + a_t\mu, \tag{21}$$

where $a_t = v_t/(\sigma_\alpha^2 + v_t)$. The expression for v_t is the same as given in the previous section.

For given $\bar{y}_t$, the variance of the above estimator is

$$V(B_t) = \frac{1}{(1/\sigma_\alpha^2) + (1/v_t)}. \tag{22}$$

With estimates $\hat{\sigma}_\alpha^2$, $\hat{\sigma}_t^2$ and $\hat{\mu}$, the expression in (21) can be written as

$$\hat{B}_t = (1 - \hat{a}_t)\,\bar{y}_t + \hat{a}_t\hat{\mu}, \tag{23}$$

where $\hat{a}_t = \hat{v}_t/(\hat{\sigma}_\alpha^2 + \hat{v}_t)$. This estimator may be called the Empirical Bayes' estimator (EB). Note that $\hat{\mu}$ is obtained from (17) with $\hat{\sigma}_\alpha^2$ and $\hat{v}_t$. The variance of this estimator may be obtained from (22) by replacing σ_α^2 and v_t with their estimates. For obtaining the EB and its variance, we have estimated σ_α^2 and v_t from the USS procedure described in the previous section.

7. PERFORMANCE OF THE ESTIMATORS

We have computed the estimates of P_t for 1986 for the 20 hospitals through the different procedures described in the previous sections. Since the population values of P_t are not known, as described earlier, we have found the S.E.'s for the different procedures by substituting the sample proportion $\hat{P}_t$ in the place of P_t. Since the sample sizes n_t are not small, the resulting biases in estimating the variances or S.E.'s of the estimators can be expected to be small.

For the three hospitals, the estimates of P_t and the S.E.'s of the different procedures are presented in Tables 2 and 3 for AMI and MDS respectively.

As can be seen from these tables, S.E.'s of TR, VC and EB are smaller than the S.E. of the sample proportion. As expected, utilizing the data from the previous periods has helped reduce the S.E. of the estimate for the current period.

Both VC and EB have smaller S.E.'s than TR. However, TR does not require the estimation of σ_α^2. We have found the S.E. of TR to be usually less than 50 percent of the sample proportion.

The EB has smaller S.E. than VC, as expected. Note that VC estimates the overall proportion, whereas EB estimates the proportion of the conditional distribution. The S.E. of the EB becomes close to that of the sample proportion if the sample size is large.

It is interesting to observe from Tables 2 and 3 that for both AMI and MDS the difference between the VC and EB estimates is negligible. The reason for this result is that $\hat{a}_t$ is close to unity, which indicates that σ_α^2 is small relative to v_t.

The estimates for the total number of cases for 1986 and their S.E.'s can be obtained by multiplying the estimates of the proportions in Tables 2 and 3 by the corresponding number of discharges N_t given in Table 1.

8. DISCUSSION

As described in the above section, the results of this investigation recommend the TR, VC or EB methods for estimating the proportions and totals for the current period.

For estimating the S.E.'s of the different procedures, we have utilized the sample proportions. Further investigation is needed to examine the biases and MSE's of these S.E's.

For estimating σ_α^2 and v_t, we have employed the USS. The effects of the ANOVA and the MINQUE procedures for this purpose can also be examined. However, the investigation in Rao *et al.* (1981) showed that different procedures of estimating σ_α^2 may not have a significant effect on the estimation of μ or its S.E.

Further investigation is needed to determine the effect of the different procedures of estimating the variances on the EB for μ_t.

We have substituted a small positive quantity for a negative estimate of σ_α^2. As can be seen, this adjustment may result in a small S.E. for both the VC and EB, and may present too optimistic a view about the estimates of μ and μ_t. Further examination of this problem is needed.

We have assumed a linear model for the proportion. The logit or probit transformation can be used before using this model. However, large population and sample sizes are needed to justify the estimates that can be obtained through these transformations. The estimates proposed in this article can be obtained by the public and private users by using any simple computer program.

Improved estimates for each hospital are considered in this paper. The national estimates for a given item like AMI or MDS can be obtained by suitably weighting the above estimates by the reciprocals of the probabilities with which the hospitals were selected. Such a procedure is expected to improve the precision of the national estimates.

Time series methods like the ARIMA can be used as suggested for instance by Blight and Scott (1973) and Scott and Smith (1977) for estimating the proportions and total numbers. These methods will result in different models for different items. Secondly, the available package programs for these approaches assume large population sizes and equal error variances, and the same sample sizes for all the time periods. Such assumptions are not satisfied for the problem we have considered in this article. As mentioned in Section 1, the TR, VC and EB methods can also be used when there is nonresponse during some years.

ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor and a referee for their helpful comments.

REFERENCES

BEAN, J.A. (1987). NHDS variance and covariance estimation of year to year differences. National Center for Health Statistics, research report.

BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Series B, 35, 61-68.

CARROLL, R.J., and RUPERT, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

COCHRAN, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.

RAO, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.

RAO, P.S.R.S., KAPLAN, J., and COCHRAN, W.G. (1981). Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association*, 76, 89-96.

SCOTT, A.J., and SMITH, T.M.F. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.

SHIMIZU, I.M. (1987). Specifications for the redesigned NHDS sample. National Center for Health Statistics, research report.

YATES, F., and COCHRAN, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Sciences*, 28, 556-580.

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 143–152
Statistics Canada

RDD Panel Attrition in Two Local Area Surveys

PAUL J. LAVRAKAS, RICHARD A. SETTERSTEN, Jr.
and RICHARD A. MAIER, Jr.[1]

ABSTRACT

This paper compares the magnitude and nature of attrition in two separate RDD panel surveys conducted in the City of Chicago (*i.e.* the surveys were independent studies and were not conducted as part of a planned experiment), each with a between-wave lag of approximately one year. For each survey, sampling at Wave 1 was performed via one-stage (*i.e.* simple) random-digit dialing. In Study 1, respondents' names were *not* elicited; thus, when telephone calls were made at Wave 2 of Study 1 interviewers could not ask for respondents by name. Instead, interviewers asked for respondents by using a gender-age identifier. In Study 2, respondent name identifiers were gathered during Wave 1 and were used in Wave 2 re-contact attempts. The magnitude of the attrition in Study 1 (*i.e.* the proportion of Wave 1 respondents not re-interviewed at Wave 2) was 47%, whereas in Study 2 it was 43%: a marginal difference in attrition rates. In both surveys, age, race, education and income were significantly related to attrition. Discussion is presented on the trade-off between minimizing attrition *vs.* minimizing respondent reactivity as potential sources of total survey error. Suggestions for decreasing the size of attrition in RDD panel surveys are discussed.

KEY WORDS: Panel attrition; Random-digit dialing; Telephone surveys.

1. INTRODUCTION AND LITERATURE REVIEW

For the past several decades the problem of panel attrition has received only passing attention in the survey methods literature. Published articles either have addressed techniques which can be employed to minimize the size of the attrition from panel studies (*e.g.* Droege and Crambert 1965; Crider, Willets and Bealer 1971; McAllister, Goe and Bulter 1973; Freedman, Thornton and Camburn 1980; and Burgess 1989) or have addressed statistical techniques that may be used to *adjust* for the effects of panel attrition (*e.g.* Lehnen and Koch 1974; Hausman and Wise 1979; Winer 1983; and Lepkowski 1989).

Few articles have reported on the magnitude and nature of the resulting attrition. And, even fewer have dealt with *random* samples of the public which would allow other researchers to estimate what to expect in future general population surveys. An exception was Sobol's (1959) reporting on the attrition that occurred in a five-wave panel studying economic attitude change. At Wave 1, in 1954, a probability sample of the non-institutionalized urban population of the United States was interviewed (n = 1,150). Subsequent waves were conducted six, 12, 18, and 33 months later. Compared to the original sample, attrition for each subsequent wave was 17%, 26%, 29% and 39%, respectively.

Sobol reported that, in general, "because of canceling variations, the demographic structure . . . after five rounds of interviewing, remained very similar to that of the original [sample]" (p. 52). Yet there were some significant variations, with a disproportionate number of renters, lower income households, residents of large metropolitan areas, younger (under 25 years) and older (over 64 years) adults, and those not interested in the survey subject matter lost to the panel. Winer (1983) reported results of unpublished studies which generally confirmed Sobol's findings.

[1] Paul J. Lavrakas, Richard A. Settersten, Jr. and Richard A. Maier, Jr., Northwestern University Survey Laboratory, 625 Haven St., Evanston IL 60208 – 4150 USA.

It is important to note that in each of these studies interviewers knew Wave 1 respondents' full names. In fact, in their article on techniques to minimize panel attrition, McAllister *et al.* (1973) stressed the importance of gathering detailed information about the respondent's future whereabouts at the end of the interview, including "complete names and addresses of friends and/or relatives ... of the respondent" (p. 416).

Although it can be argued that panel attrition is a serious enough problem to prompt researchers to obtain the full name and other identifying information of each Wave 1 respondent, this approach may cause problems of its own. In those instances where a respondent's name is elicited as part of the Wave 1 interview, an explanation is sometimes given that the name is important because the respondent may/will be called back after some specified time to determine if any changes occurred. This raises concerns about "evaluation apprehension" (*i.e.* reactivity) on the part of respondents (*cf.* Crano and Brewer 1973). Whereas some authors explicitly address the trade-off between attrition and reactivity (*e.g.* Sobol 1959), it is implicit in most other articles, that authors typically regard reactivity as less a problem than attrition.

All of the aforementioned research was conducted with personal interviews. But what of panel attrition when telephone surveying is done, including those studies in which Wave 1 respondents' names are not recorded? In particular, what can be expected by a researcher who plans *a priori* to conduct a panel telephone survey and thus ask respondents for name identifiers *vs.* a researcher who does not gather respondent name identifiers, either because he/she explicitly chooses not to or because a decision is made *post hoc* to convert a cross-sectional telephone survey to a panel after Wave 1 interviewing is complete?

In an attempt provide a preliminary perspective on these issues, the present paper reports findings on the magnitude and the nature of attrition in two RDD (two-wave) panel studies conducted in the City of Chicago, each with a between-wave lag of approximately one year. It should be noted that these two surveys were conducted independently of each other, not as part a planned test of RDD attrition. As such, there are various differences in the substantive focus and specific execution of the two surveys, beyond the fact that in Study 2 a name identifier was known for most respondents whereas in Study 1 it was not. We explicitly acknowledge that these differences in focus and execution somewhat limit the conclusions that can be drawn from the comparison of the two studies.

For both surveys, one stage (*i.e.* simple) random-digit dialing was used to sample Wave 1 respondents. In Study 1, respondent names were not asked as part of the Wave 1 interview and respondents were not told that they would be re-contacted. In Study 2, name identifiers were gathered at the completion of the Wave 1 interviews and were used to reach respondents at Wave 2. Respondents most often did not provide their full names, instead giving their first name only or other name identifier, *e.g.* nickname or initials. (Interviewers did not probe for full names so as to not contribute to possible feelings of paranoia on the part of reluctant respondents.)

When respondents' names are not known, how does one go about re-contacting the original respondent? This was a problem faced in 1979 by the first author when trying to determine the efficacy of creating a panel from a 1977 cross-sectional survey. As nothing was found in the published literature to provide guidance, a pilot-test was conducted with a resulting 50 percent of the 1977 respondents re-interviewed by asking for them by *gender and age*.

The results of this pilot-test were encouraging enough to recommend the procedure for use in the first study reported here. In Study 1, interviewers dialed the same telephone numbers as the Wave 1 completions, verified each number whenever the call was answered, and informed the listener that approximately one year ago a person at the telephone number had completed an interview. The original respondent was identified by *gender* (*e.g.* "a man" or "a woman") and by *age* (*e.g.* "in his early twenties" or "in her late sixties").

In Study 2, a name identifier was known for more than eight out of 10 of the Wave 1 respondents. For these respondents, Wave 2 interviewers asked for the respondent using the name identifier, after first verifying the telephone number. For the respondents with no name identifier, interviewers asked for the respondent by using demographic identifiers, as in Study 1.

In reporting the results from these studies, it is our modest intention to shed preliminary light on the magnitude and nature of attrition in RDD panels. Although the results should not be generalized to a national RDD sample, they are suggestive. Given the prevalence of RDD sampling, we believe it is important to build a knowledge-base about the attrition that can be expected in panel studies where Wave 1 sampling is done via random-digit dialing, especially when researchers have no Wave 1 name identifier for respondents. By doing this, we can better consider strategies to reduce the size and effects of this attrition.

2. STUDY 1

2.1 Methodology

In February, 1983, a city-wide (one-stage) RDD survey was conducted by the Northwestern University Survey Laboratory to gather baseline data for professors who were evaluating a series of community crime prevention programs in Chicago neighborhoods. (The questionnaire took an average of 20 minutes to administer.) Approximately 2,800 telephone numbers were dialed in the process of completing 814 interviews. For each residence contacted, one head-of-household (male or female) was systematically selected as the designated respondent (*cf.* Lavrakas 1987; pp. 99-100). Whenever necessary, Spanish-language questionnaires were administered by bilingual interviewers. Up to seven call-backs were made to hard-to-reach respondents. Of all telephone numbers dialed, 1,247 were found to ring in eligible households (defined by the survey sponsors as English-speaking or Spanish-speaking households with at least one adult 19 years of age or older); those eligibles not interviewed either were unavailable at the time calls were made or refused to participate.

One year later, in February, 1984, the Wave 1 telephone numbers were re-dialed to gather "post-test" data for the evaluation project. In those instances where the telephone was answered within eight call-attempts (across different days and times), the following introduction was read by interviewers:

> Hello, is this _____ ? My name is _____, and I'm calling from Northwestern University. About a year ago (February 1983) we conducted an interview with a _____ at this number. May I please speak with (her/him)?

The interviewer first verified the telephone number and then gave her/his own name. The third *blank* contained pre-recorded Wave 1 demographic information (gender and age) about each respondent: *e.g.* "woman in her mid 30s," or "man in his early 70s." For those few respondents who had not given their year of birth at Wave 1, the third blank simply contained the gender identifier, "woman" or "man."

Once the interviewer was speaking to the original respondent he/she continued with the following explanation, before beginning the interview:

> The information you gave us last year was a big help in understanding the concerns of residents like yourself. We are calling back now to find out some things about the quality of life in Chicago neighborhoods during the past year.

The purpose of this statement was to reinforce the respondent's willingness to cooperate with the Wave 2 interviewer by reminding the respondent of his/her cooperation in the Wave 1 survey.

Coinciding with the purpose of this evaluation project, respondents who had moved or changed their telephone numbers were not interviewed at Wave 2. This was due to the need to interview only those persons who resided at the same address as the previous year, since many of questions dealt with perceived neighborhood change since February, 1983.

2.2 Results

Due to a clerical error in processing the Wave 1 questionnaires and call-records, duplicate or incorrect respondent I.D. numbers were assigned to 17 Wave 1 respondents by the survey sponsors' staff. For the purposes of this paper, these respondents were dropped from our analyses because we could not match correctly their Wave 2 dispositions with their respective Wave 1 data. Thus the following analyses are based on the 797 respondents whose Wave 1/ Wave 2 match was certain.

The magnitude of the attrition. As shown in Table 1, approximately one-half of the Wave 1 sample was re-interviewed (53%). Of the 375 respondents who were "lost" to the panel, the greatest proportion was due to telephone numbers that rang in a new household or in an original household from which the respondent had moved; this accounted for approximately 40% of the attrition. Second most frequent were those persons whose Wave 1 telephone number was no longer in service; this accounted for a fourth of those lost. Next in frequency of those lost were respondents who refused in some way. The fourth most prevalent reason for losing respondents were those who were never home during the Wave 2 field period when their telephone was answered, even after eight call-backs; (these 33 persons were verified to be the original respondent by someone else in their household).

Table 1

Dispostion of Wave 1 Samples for Study 1 (Names not known) and Study 2 (Names known)

Wave 2 disposition	Study 1 - No names		Study 2 - Names	
	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency
		%		%
No Wave 2 contact made				
Non-working, disconnected	95	11.9	94	9.4
Never answered	17	2.1	45	4.5
Contact made				
Completion	422	52.9	572	57.4
Respondent gone from number	165	20.7	163	16.4
Respondent never available	33	4.1	30	3.0
Respondent refusal/partial	37	4.7	57	5.7
"Gatekeeper" refusal	21	2.6	13	1.3
Incapacitated, deceased	3	0.4	13	1.3
Misc. other	4	0.6	10	1.0
Total	**797**	**100.0**	**997**	**100.0**

The nature of the attrition. As shown in Table 2, the group of Wave 1 respondents in Study 1 who were re-interviewed differed significantly on several factors from those who were lost to the panel. In terms of age, those adults less than 30 years of age at Wave 1 were re-interviewed with only 42% success *vs.* adults in the 40-59 year group of whom 60% completed Wave 2 surveys. Blacks were significantly less likely to be re-interviewed than Whites. In terms of household income, those respondents who reported Wave 1 annual household incomes of less than $10,000 were re-interviewed with only 44% success *vs.* those with incomes over $20,000 of whom 63% were re-interviewed. Married respondents were more successfully re-interviewed (57%) than those not married (49%). Sixty-two percent of home owners were re-interviewed compared with 47% of renters. The longer one had lived in the neighborhood and the more likely one reported at Wave 1 that he/she would not move, the more likely he/she was re-interviewed.

3. STUDY 2

3.1 Methodology

During November and December of 1983, a city-wide (one-stage) RDD survey was conducted by the Northwestern University Survey Laboratory for professors who were examining economic well-being/hardship among Chicago families. (The questionnaire took an average of 20 minutes to administer.) Approximately 3,900 telephone numbers were dialed in the process of completing 997 interviews. For each residence contacted, one head-of-household (male or female) was systematically selected as the designated respondent. Up to 20 call-backs were made to increase the likelihood of completing interviews with hard-to-reach respondents. In total, 1,659 eligible households were reached; those eligibles not interviewed either were unavailable at the time calls were made or refused to participate.

Sixteen months later (Spring 1985), all 997 telephone numbers were re-dialed to gather Wave 2 data. Unlike Study 1, in which respondents were not tracked if they had moved or changed their telephone numbers, an effort was made to find respondents whenever possible, although this effort resulted in only few successes as the respondent's full name (first and last) was typically not available. As in Wave 1 of Study 2, at least 20 call-backs were used with the hardest-to-reach respondents.

More than 80% of respondents had given a name identifier at Wave 1. This information was used by interviewers as follows:

> Hello, is this _____ ? My name is _____ , and I'm calling from Northwestern University. About 16 months ago, in late 1983, we conducted an interview with a (man/woman) named _____ at this number. May I please speak with (her/him)?

As with Study 1, the interviewer first verified the telephone number and then gave her/his own name. The third blank contained the pre-recorded name identifier given by the respondent at Wave 1. Those respondents who did not give a name at Wave 1 were asked for by using the same procedure used in Study 1 (*i.e.* asking by gender and age, or by gender only).

3.2 Results

The magnitude of the attrition. As shown in Table 1, nearly six in 10 of the Wave 1 sample were re-interviewed (57.4%). Overall, the pattern of Wave 2 dispositions in Study 2 was very similar to what was observed in Wave 2 of Study 1. Of the 425 respondents who were "lost"

Table 2

Respondent Characteristics of Wave 2 Re-interviews for Study 1 (Names not known) and Study 2 (Names known)

Respondent characteristic	Percentage Re-interviewed at Wave 2	
	Study 1	Study 2
Gender:		
Females	55	56
Males	49	60
Age:		
< 30 years	42*	--
30-39 years	54	--
40-59 years	60	--
>59 years	55	--
< 34 years	--	54*
35-49 years	--	64
50-64 years	--	61
> 64 years	--	50
Race:		
Asian	68***	42*
Black	49	53
Hispanic	44	62
White	58	61
Education:		
Not high school graduate	45**	49**
High school graduate	57	58
Some college	50	53
College graduate	60	64
Graduate school	--	68
Household income:		
< $10,000	44**	--
$10,000-$19,999	55	--
$20,000-$29,999	63	--
$30,000 or more	63	--
< $12,000	--	50**
$12,000-$17,999	--	59
$18,000-$23,999	--	66
$24,000 or more	--	64
Marital status:		
Married	57*	57*
Divorced	--	68
Separated	--	43
Single	--	57
Widowed	--	51
Not married	49	--
Residential status:		
Own	62***	59
Rent	47	56
Residential tenure in neighborhood:		
< 3 years	44***	--
3-9 years	55	--
10 or more years	56	--
Likelihood of moving in next 2 years:		
Definitely will	37***	--
Probably will	47	--
Probably will not	57	--
Definitely will not	60	--

Note: Chi-square tests of significance were employed.
*** $p < .001$
** $p < .01$
* $p < .05$

to the panel, the greatest proportion (nearly four in 10) was associated with numbers that rang in a new household or in an original household from which the respondent had moved with no new number available. Second most frequent were those persons whose Wave 1 telephone number was no longer in service, accounting for nearly one in four of those lost to the panel. Next in frequency were respondents who refused in some way. The fourth most prevalent reason for losing Wave 1 respondents were those persons whose original telephone numbers were never answered at Wave 2.

The nature of the attrition. As shown in Table 2, the group that was interviewed at Wave 2 of Study 2 differed significantly on several factors from the group which was lost, with patterns similar to what was observed in Study 1. In terms of age, those adults less than 34 years of age and those more than 64 years of age at Wave 1 were least likely to be re-interviewed. Asians and Blacks were less likely to be re-interviewed than were Hispanics and Whites. In terms of education, those with less formal education were least likely to be re-interviewed. Those respondents who reported Wave 1 annual household incomes of less than $12,000 were re-interviewed with only 50% success *vs.* those with incomes over $24,000 of whom 64% were re-interviewed. Divorced respondents were most successfully re-interviewed (68%), whereas those who said they were separated at Wave 1 were least likely to be re-interviewed (43%).

4. DISCUSSION

4.1 Summary of Findings

The two independent studies reported here were two-wave RDD telephone surveys, one with a 12 month lag between waves and the other with a 16 month lag. In Study 1, where names were not known for use at Wave 2, attrition was 47.1%. In Study 2, where name identifiers from Wave 1 were known for 83% of the respondents, attrition was somewhat less, at 42.6%.

This marginal difference in attrition rates (χ^2 (1) = 3.51, $p < .10$) is best considered within the following contextual differences between the studies: Study 1 respondents were **not** explicitly told at Wave 1 that they would be called back a year later and, thus, their names were not asked. In Study 2, respondents were told that they would be called back at some future time. Given the particular nature of the research in Study 1, no effort was made to track Wave 2 respondents who had moved or changed their telephone number. On the other hand, an effort was made to do this in Study 2, although with little success. Study 1 employed a Spanish-language version of the questionnaire; in Study 2, Hispanics who could not speak English were not interviewed.

In both studies, the vast majority of those lost to the panel were respondents who could not be reached via their Wave 1 telephone number, either because the number reached an entirely new residence, the respondent had moved from the household, or the number was no longer in service.

Taken together, the findings of these two telephone studies are fairly consistent with past findings from in-person surveys (*e.g.* Sobol 1959) in identifying the types of persons most likely to be lost in panel studies. In both Study 1 and 2, younger and older adults, non-Whites, the less educated, and those with lower income were less likely to be re-interviewed than other demographic subgroups.

4.2 Implications

Given the cost/benefit attraction of RDD surveys, added to the analytic benfits associated with panel studies, it is worthwhile to consider options that may improve the representativeness

of the final panel in surveys that use RDD for Wave 1 sampling. But before discussing these considerations, the issue of asking for respondents' names in telephone surveys merits further discussion.

Asking for respondents' names at Wave 1. As mentioned above, the issue is purportedly one of increasing the likelihood of reaching and, thus, re-interviewing the respondent at Wave 2 *vs.* the possibility of creating an evaluation apprehension effect (Crano and Brewer 1973) which may bias Wave 2 data. Yet, more than this trade-off enters into consideration.

The issues of confidentiality and informed consent also come into play: it is common practice in academic survey research for a survey organization to never provide respondent telephone numbers to anyone, with the possible exception of the sponsor, and only when he/she is planning a panel study or conducting follow-up interviews with respondents who have explicitly given permission for this. This practice follows from the reasoning that an assurance of confidentiality given to Wave 1 respondents is not violated when respondents are called back as part of the *same* on-going research. The fact that so few Wave 1 respondents refuse to participate at Wave 2, coupled with the observation that it is demographically predictable who is most likely to refuse at Wave 2, provides strong support for the conclusion that calling respondents back without having asked their permission at Wave 1 is *not* a problem.

When a telephone survey sponsor can pay for the expense of tracking respondents who have moved, it appears logical to record respondents' full names at Wave 1, since those who have moved may be tracked through telephone directories; calling new numbers given by telephone company recordings; or, even by calling former neighbors to get a forwarding telephone number in those cases where a respondent's address is also known and a reverse-telephone directory is used. But if respondents will not be tracked at Wave 2, how useful is it to be able to ask for the respondent by name?

It cannot be denied that interviewers say they prefer it. That is, most interviewers feel more comfortable asking for "John" or "John Smith" *vs.* asking for "a man in his mid-50s." Yet the marginal difference in attrition rates in the two studies reported here, even considering the four-month longer lag time between waves in Study 2 which gathered name identifiers at Wave 1, does not provide compelling evidence of the advantage of names. We acknowledge that an unfortunate limitation of our paper is that other differences in these two RDD panel surveys may have contributed to the observed differential in attrition rates: *e.g.* Wave 2 callbacks were greater in Study 2 (eight in Wave 2 of Study 1 *vs.* 20 in Wave 2 of Study 2). Thus, this issue will remain unresolved until more controlled research is conducted.

Given the current state of knowledge, we believe that it remains the responsibility of the individual researcher using an RDD panel to weigh the competing tensions of possibly biasing measures of the phenomenon under investigation by alerting respondents that they will be "measured" again (*i.e.* the "reactivity" effect) *vs.* the possibility of experiencing slightly less attrition by asking for names at the time of the Wave 1 interview.

Considerations to minimize attrition effects. Some suggestions can be considered in the attempt to minimize the effects of RDD panel attrition.

Sobol (1959) suggested the possibility of a Wave 1 *over-sampling* of those types of respondents who were most likely to be lost in subsequent waves. At first, this suggestion may sound appealing. This initial appeal follows the reasoning that if one knows who is most likely to be lost, then one can project an over-sampling of those groups at Wave 1. As was shown in Sobol's work, and as found in the two studies reported here, one could estimate what types of persons should be over-sampled at Wave 1; *e.g.* older and younger adults. Over-sampling could be accomplished through the use of a screening procedure introduced late in the Wave 1 field period; (although this clearly would increase Wave 1 total survey costs).

Although it is possible to over-sample, is it also desirable? In asking this question, one is ultimately asking whether the resulting panel is more than just an *on-the-surface* demographic match of the population of interest. In other words, is it enough to merely be concerned with getting, for example, the right number (*i.e.* proportion) of senior citizens in the final wave of a panel, or should one also be concerned whether one has the right "mix" of seniors?

This is an empirical question that the present studies cannot answer. Clearly, more research is needed before survey researchers can be more certain whether it is preferable to over-sample at Wave 1 or to "compensate" for attrition through statistical adjustments to subsequent waves of panel data.

Another aspect of the attrition problem is associated with efforts to minimize the loss of those persons whom interviewers are able to re-contact at Wave 2; *i.e.* respondents who refuse or who are "never at home" to complete the Wave 2 interview. This type of loss accounted for 29% of the Study 1 attrition, and 34% in Study 2. What can be done so that interviewers might be more successful at minimizing these losses, other than merely employing traditional interviewer training techniques and making many call-back attempts?

In this age of microcomputers it is quite feasible for interviewers to be given a Wave 1 "profile" of each respondent, so as to be more familiar with the person to be re-interviewed. Care would have to be exercised to avoid creating expectations on the part of interviewers that might bias respondents' Wave 2 answers. We are not suggesting that the interviewer necessarily use this information in verbatim form to identify the Wave 1 respondent; we believe name, gender and age are adequate for that purpose. But, there may be subtle changes in an interviewer's verbal behavior that may lead to increased success at re-interviewing when the interviewer has a more detailed idea of "who" the respondent is. This suggestion must await testing before it can be confidently endorsed, but were it to prove effective without introducing bias into the data, it would be relatively easy to do.

Similarly, introductory statements read by interviewers at Wave 2, could be targeted with special appeals to those demographic groups who appear most likely to refuse at Wave 2: in this case we are referring to the elderly, those with less formal education, those with relatively lower income, and especially those who were rated by Wave 1 interviewers as showing little interest and/or cooperation. Here again, a computer could be programmed to generate special Wave 2 introductory spiels based on Wave 1 data about particular respondents.

These appeals must contain incentives for such persons to participate at Wave 2, as they are often persons with the least intrinsic motivation to participate in surveys. When planning for subsequent waves, surveyors should think of "why" such people would want to cooperate and work such reasoning into the interviewers' introduction for these persons. Such introductions may be lengthy and may even contain some rapport-building questioning. It may even be possible to give the prospective respondent some feedback about Wave 1 findings, without biasing Wave 2 responses. If so, the respondent may regard the re-contact attempt to be more of a "two-way" exchange.

Regardless, computers could be used to generate these special introductions, which in turn would be matched only with those respondents for whom the message is targeted. Again, we have no empirical evidence to cite regarding the efficacy of this suggestion, but we believe it merits consideration and study.

5. CONCLUSION

Our findings suggest that attrition in RDD panels when respondent names are unknown is not of such magnitude as to render the surveying technique invalid or impractical. Due to

its nonreactivity, it would certainly appear to be the preferred approach in two-wave RDD panels in which the researcher has *a priori* reason not to want Wave 1 respondents to know they will be re-contacted. These findings also should provide encouragement for those who are thinking about converting an RDD cross-sectional survey into a panel. We hope that this primarily descriptive paper will encourage other survey methodologists to conduct and report the results of more controlled studies that investigate the nature and magnitude of RDD panel attrition, so that eventually, researchers can more confidently implement strategies to reduce the level of attrition. We suggest that this research should be guided by the observation that reductions in the magnitude of RDD panel attrition appear most likely to occur with well-organized surveying in which each respondent is approached as the individual that he/she is.

ACKNOWLEDGMENTS

The authors would like to thank Professors Fay Cook, Christopher Jencks, Dan Lewis and Dennis Rosenbaum for permission for access to the data sets which were used for the secondary analyses reported in this paper. The authors also appreciate the helpful comments of Professor Peter V. Miller on an earlier version of this manuscript.

REFERENCES

BURGESS, R.D. (1989). Major issues and implications of tracing survey respondents. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons.

CRANO, W.D., and BREWER, M.B. (1973). *Principles of Research in Social Psychology*. New York: McGraw-Hill.

CRIDER, D. M., WILLETS, F.K., and BEALER, R.C. (1971). Tracking respondents in longitudinal surveys. *Public Opinion Quarterly*, 35, 613-620.

DROEGE, R.C., and CRAMBERT, A.C. (1965). Follow-up techniques in a large-scale test validation study. *Journal of Applied Psychology*, 49, 253-256.

FREEMAN, D.S., THORTON, A., and CAMBURN, D. (1980). Maintaining response rates in longitudinal studies. *Sociological Methods & Research*, 9, 87-98.

HAUSMAN, J.A., and WISE, D.A. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica*, 47, 455-473.

LAVRAKAS, P.J. (1987). *Telephone Survey Methods: Sampling, Selection and Supervision*. Newbury Park, CA: Sage.

LEHNEN, R.G., and KOCH, G.G. (1974). Analyzing panel data with uncontrolled attrition. *Public Opinion Quarterly*, 38, 40-56.

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

McALLISTER, R.J., GOE, S.J., and BUTLER, E.W. (1973). Tracking respondents in longitudinal surveys: some preliminary considerations. *Public Opinion Quarterly*, 37, 413-416.

SOBOL, M.G. (1959). Panel mortality and panel bias. *Journal of the American Statistical Association*, 54, 52-68.

WINER, R.S. (1983). Attrition bias in econometric models estimated with panel data. *Journal of Marketing Research*, 177-186.

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 153–161
Statistics Canada

An Exact Test for the Presence of Stable Seasonality With Applications

BRAJENDRA C. SUTRADHAR, ESTELA BEE DAGUM
and BINYAM SOLOMON[1]

ABSTRACT

The X-11-ARIMA seasonal adjustment method and the Census X-11 variant use a standard ANOVA-F-test to assess the presence of stable seasonality. This F-test is applied to a series consisting of estimated seasonals plus irregulars (residuals) which may be (and often are) autocorrelated, thus violating the basic assumption of the F-test. This limitation has long been known by producers of seasonally adjusted data and the nominal value of the F statistic has been rarely used as a criterion for seasonal adjustment. Instead, producers of seasonally adjusted data have used rules of thumb, such as, F equal to or greater than 7. This paper introduces an exact test which takes into account autocorrelated residuals following an SMA process of the $(0,q)(0,Q)_s$ type. Comparisons of this modified F-test and the standard ANOVA test of X-11-ARIMA are made for a large number of Canadian socio-economic series.

KEY WORDS: Standard Anova; Autocorrelated residuals; Seasonality.

1. INTRODUCTION

In the analysis of social and economic time series, it is traditional to decompose the observed series into four unobserved components, namely the trend, the cycle, the seasonal variations, and the irregulars.

Socio-economic time series are often presented in seasonally adjusted form so that the underlying short-term trend can be more easily analysed and current socio-economic conditions can be assessed. There are several seasonal adjustment methods available which estimate the seasonal component present in a time series, but the Census X-11 variant (Shiskin, Young and Musgrave 1967) and the X-11-ARIMA method (Dagum 1980) are the most widely applied. To identify the presence of stable seasonality in a time series, the X-11-ARIMA method as well as the Census X-11 variant use the results of the usual F-test in a one-way ANOVA between monthly seasonal variations and the residuals. However, the residuals in this ANOVA are often autocorrelated, so the nominal significance level of the F-test may not be valid. Aware of this limitation, producers of seasonally adjusted data, do not guide themselves by the nominal significance level of the F-test for presence of stable seasonality but by some rule of thumb based on empirical knowledge (see *e.g.* Shiskin and Plewes 1978). In fact, implicit in the X-11-ARIMA test for the presence of 'identifiable seasonality' is that the F-value for stable seasonality should be greater or equal to 7 if moving seasonality is not present.

The testing for stable seasonality (similarly for annual seasonal shifts) can be approached as a test for the significance of certain regression coefficients in a linear model with autocorrelated errors. The traditional Wald test, the likelihood ratio test, and the tests falling within a generalized least squares framework, all run into convergence problems in testing such a linear

[1] Brajendra C. Sutradhar, Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland, A1C 5S7; Estela Bee Dagum, Time Series Research and Analysis Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. Binyam Solomon, Directorate of Social and Economic Analysis, National Defence Headquarters, Ottawa, Ontario, K1A 0K2.

model with highly autocorrelated errors (*cf.* Sutradhar and Bartlett 1990). Pierce (1978) constructed an F-test based on transformed residuals which are approximately white noise. The transformation suggested in Pierce (1978) is equivalent to using the inversion of the error covariance matrix. But, the inverse of the error covariance matrix may not be obtained for highly autocorrelated errors. Recently Sutradhar, MacNeill and Dagum (1991) proposed a modified F-test, within a linear model framework for testing for the presence of stable seasonality. Their modified F-test is derived following Sutradhar, MacNeill and Sahrmann (1987), and the test accounts for the presence of autocorrelation in the residuals. The test does not require any transformation or any inversion of the error covariance matrix.

Exact tests for testing the null hypothesis that the seasonal pattern changes over time against the alternative that the seasonal pattern is constant have been developed by Franzini and Harvey (1983). Unlike Franzini and Harvey, the present approach assumes that the seasonal pattern is stable over time possibly at different levels (due to annual shifts) and then tests for the presence of significant stable seasonality.

In most empirical cases, a seasonal moving average (SMA) error model of the $(0,q)(0,Q)_s$ type is sufficient. In this investigation we simplify the exact test proposed by Sutradhar, MacNeill and Dagum (1991), for such error models. The test is applied to examine for the presence of stable seasonality as well as of annual seasonal shifts in a number of socio-economic series.

The plan of this paper is as follows. Section 2 presents the exact test. Section 3 analyses the results from the application of the modified F-test to a set of socio-economic time series and compares them with the values given by the X-11-ARIMA method. Section 4 gives the conclusions.

2. MODIFIED F-TEST

2.1 Selected Model

Consider a stationary seasonal time series $\{Z_t\}$, given by

$$Z_t = S_t + U_t, \tag{2.1}$$

where Z_t is the observed series at time t, S_t is the seasonal component, and U_t the irregulars. If the time series contains a trend, which is most likely, it is assumed that a suitable detrending technique will yield the model (2.1). In the latter case, the detrended series may be obtained from the original series by taking appropriate differences as in ARIMA modelling (Box and Jenkins 1970) or as is traditionally done by statistical agencies which use the X-11-ARIMA method or Census X-11 variant.

Next, suppose there are k seasons in a year and there are kn observations in a time series of n years. Let $Z\{(i-1)n+j\}$ be the jth $(j = 1, \ldots, n)$ observation under the ith season $(i = 1, \ldots, k)$ which corresponds to Z_t in (2.1). We shall denote in similar manner the (i,j)th components of S_t and U_t, for all $t = 1, \ldots, kn$. Then, the model assumed for S_t is (*cf.* Sutradhar and MacNeill 1989):

$$S((i-1)n+j) = \mu + \alpha_i + \beta_j, \tag{2.2}$$

with $\sum_{i=1}^{k}\alpha_i = 0$, $\sum_{j=1}^{n}\beta_j = 0$.

The α's and β's in (2.2) represent, respectively, the stable seasonality and annual seasonal shifts in the seasonal time series. Thus, when testing for the presence of stable seasonality, we test the hypotheses

$$H_0\colon \alpha_i = 0 \quad \text{vs.} \quad H_1\colon \alpha_i \neq 0 \text{ for at least one } i; \tag{2.3}$$

and when testing for the presence of annual seasonal shifts, we test the hypotheses

$$H_0\colon \beta_j = 0 \quad \text{vs.} \quad H_1\colon \beta_j \neq 0 \text{ for at least one } j. \tag{2.4}$$

Consequently, the rejection of H_0 in (2.3) and (2.4) would indicate that the series contains significant stable seasonality as well as annual seasonal shifts.

Taking into account model (2.2), the model (2.1) can be written as

$$Z^* = X\gamma + U^*, \tag{2.5}$$

where

$$Z^* = [Z(1), \ldots, Z(n), Z(n+1), \ldots, Z(kn)]',$$

$$U^* = [U(1), \ldots, U(n), U(n+1), \ldots, U(kn)]',$$

$$\gamma = [\mu, \alpha_1, \ldots, \alpha_{k-1}, \alpha_k, \beta_1, \ldots, \beta_{n-1}, \beta_n]'$$

and X is the appropriate $kn \times (k + n + 1)$ design matrix.

2.2 Test Statistics

U^* in (2.5) can be represented by seasonal autoregressive moving average (SARMA) stationary process $(p,q)(P,Q)_s$. In most empirical cases we found, however, that a $(0,q)(0,Q)_s$ model is sufficient. Let Σ^* denote the $kn \times kn$ covariance matrix of U^*. Naturally, Σ^* will contain $\theta \equiv (\theta_1, \ldots, \theta_q)$ and $\Theta \equiv (\Theta_1, \ldots, \Theta_Q)$, where θ and Θ's are the parameters associated with the SARMA $(0,q)(0,Q)_s$ process.

For the usual ANOVA model, *viz.*, when the components of U^* are i.i.d. $N(0,\sigma^2)$, one tests the null hypotheses $\beta_j = 0$, and $\alpha_i = 0$ by using the classical F-statistics F_{A1} and F_{A2} respectively, given by

$$F_{A1} = (k-1)Q_1/Q_3, \quad \text{and} \quad F_{A2} = (n-1)Q_2/Q_3,$$

where

$$Q_1 = k \sum_{j=1}^{n} (\bar{Z}_{.j} - \bar{Z}_{..})^2, \quad Q_2 = n \sum_{i=1}^{k} (\bar{Z}_{i.} - \bar{Z}_{..})^2,$$

and

$$Q_3 = \sum_{i=1}^{k} \sum_{j=1}^{n} (Z_{ij} - \bar{Z}_{i.} - \bar{Z}_{.j} + \bar{Z}_{..})^2$$

with

$$\bar{Z}_{i.} = \sum_{j=1}^{n} Z_{ij}/n, \quad \bar{Z}_{.j} = \sum_{i=1}^{k} Z_{ij}/k, \quad \text{and} \quad \bar{Z}_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n} Z_{ij}/kn,$$

Z_{ij} being the jth observation under the ith season. In the present set-up, however, these statistics are inappropriate for testing the above hypotheses. This is because, the expected values of the sums of squares are affected by the dependence among observations. Also, sums of squares are not mutually independent. For the case when U^* in (2.5) follow a SARMA $(0,q)(0,Q)_s$ process, it can be shown that

$$E(Q_1) = k \sum_{j=1}^{n} \beta_j^2 + \sigma^2(n-1)C_1(\theta,\Theta),$$

$$E(Q_2) = n \sum_{i=1}^{k} \alpha_i^2 + \sigma^2(k-1)C_2(\theta,\Theta),$$

and

$$E(Q_3) = \sigma^2(k-1)(n-1)C_3(\theta,\Theta),$$

where, for example, for the SARMA $(0,1)(0,1)_{12}$ process,

$$C_1(\theta,\Theta) = (1+\theta_1^2)(1+\Theta_1^2) - (\theta_1/6)(1+\Theta_1^2)(11-1/n) + (2\Theta_1/n)(1+\theta_1^2)$$
$$+(\theta_1\Theta_1/6)\{1-22/n-(n-2)/n(n-1)\},$$

$$C_2(\theta,\Theta) = (1+\theta_1^2)(1+\Theta_1^2) - 2(1-1/n)\Theta_1(1+\theta_1^2)$$
$$+1/6\{1+(1-1/n)/11\}\theta_1(1+\Theta_1^2) - (4/11)(1-1/n)\theta_1\Theta_1,$$

$$C_3(\theta,\Theta) = (1+\theta_1^2)(1+\Theta_1^2) + (2\Theta_1/n)(1+\theta_1^2) + (\theta_1/6)(1+\Theta_1^2)(1-1/11n)$$
$$-(\theta_1\Theta_1/6n)[n/11 - 2(n-2)/11(n-1) - 2].$$

Consequently, the null hypotheses $\beta_j = 0$, and $\alpha_i = 0$ may be tested by using the modified F-statistics F_{M1} and F_{M2} respectively, given by

$$F_{M1} = d_1(\hat{\theta},\hat{\Theta})F_{A1}, \tag{2.6}$$

$$F_{M2} = d_2(\hat{\theta},\hat{\Theta})F_{A2}, \tag{2.7}$$

(see also Sutradhar, MacNeill and Sahrmann 1987, Sutradhar, MacNeill and Dagum 1991), where $d_1(\theta,\Theta) = C_3(\theta,\Theta)/C_1(\theta,\Theta)$, $d_2(\theta,\Theta) = C_3(\theta,\Theta)/C_2(\theta,\Theta)$. The modified F-statistics F_{M1} and F_{M2} account for autocorrelation of the residuals.

Notice that in the independence case when $\theta = 0, \Theta = 0, C_1(\cdot) = C_2 = (\cdot) = C_3(\cdot) = 1$, which is obvious. In that case the problem reduces to testing the hypotheses by using standard ANOVA F-statistics.

2.3 Computation of p-value

A simulation study (*cf.* Sutradhar and Bartlett 1989, Table IV, p. 1587) indicates that for the cases when k groups are independent, the distribution of the modified F-statistics for the SMA/$(0,q)(0,Q)_s$ process, may be approximated by the usual F-distribution. In general, the F approximation to the modified F-statistic would be inappropriate, in particular when k groups are correlated and n is small.

In this paper we use the well known Satterthwaite (1946) approximation (*cf.* Sutradhar, MacNeill and Dagum 1991) to calculate the p-value, namely, $P_r(F_{M1} \geq f_{M1})$, where f_{M1} is the data based value of F_{M1}. In order to do it, we first compute the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0 = \lambda_{r+1} = \ldots = \lambda_s > \lambda_{s+1} \geq \ldots \geq \lambda_n$ of

$$\Sigma^{*1/2}[d_1(\theta,\Theta)D_1 - f_{M1}(I_{kn} - D_2)]\,\Sigma^{*1/2}, \tag{2.8}$$

where $d_1(\cdot)$ is given in equation (2.6), $D_1 = R(RR')^{-1}R'$, with $R = C(X'X)^{-1\prime}$, $D_2 = X(X'X)^{-1}X'$, C being a suitable matrix obtained by expressing the: $H_0 : \beta_j = 0$ in the form $C\gamma = 0$, where γ is given in model (2.5). In equation (2.8) I_{kn} is the $kn \times kn$ identity matrix. Then the Satterthwaite approximation yields

$$P_r(F_{M1} \geq f_{M1}) = P_r[F_{a,b} \geq bd/ac], \tag{2.9}$$

where $F_{a,b}$ denotes the usual F-ratio with degrees of freedom a and b, with

$$a = \left(\sum_{j=1}^{r} \lambda_j\right)^2 \Big/ \sum_{j=1}^{r} \lambda_j^2, \quad b = \left(\sum_{j=s+1}^{n} \lambda_j\right)^2 \Big/ \sum_{j=s+1}^{n} \lambda_j^2.$$

In equation (2.9),

$$c = \sum_{j=1}^{r} \lambda_j^2 \Big/ \sum_{j=1}^{r} \lambda_j, \quad d = \sum_{j=s+1}^{n} \lambda_j^2 \Big/ \sum_{j=s+1}^{n} |\lambda_j|.$$

Similarly, $P_r(F_{M2} \geq f_{M2})$ may be calculated by using $d_2(\cdot)$ and f_{M2} in place of $d_1(\cdot)$ and f_{M1} respectively in equation (2.9). The construction of D_1 will now depend on a different C matrix which will be obtained by expressing the $H_0 : \alpha_i = 0$ in the form $C\gamma = 0$.

3. APPLICATIONS

3.1 Monthly Series

The modified F statistics F_{M1} and F_{M2} of equations (2.6) and (2.7) were calculated for a set of 26 monthly series obtained from various economic sectors, namely, Imports, Exports, Consumer Prices and Labour. All series cover the period January 1979 till December 1988 inclusive.

Since the modified F-test is not valid when moving seasonality is present (except for annual seasonal shifts), none of the series selected are affected by moving seasonality according to certain preliminary tests available in X-11-ARIMA. (We also looked at the plots of the seasonal-irregular ratios.)

The X-11-ARIMA method was applied to obtain the detrended series $\{Z_t : t = 1, \ldots, 120\}$. Diagnostic checks show that the errors of the detrended series, U_t (see equation 2.1) follow a $(0,1)(0,1)_{12}$ SMA model for each of the monthly series. The estimates $\hat{\theta}_1$ and $\hat{\Theta}_1$ are used to compute the modified F-statistics F_{M1} and F_{M2}.

In testing for the presence of annual seasonal shifts, the p-values for the modified F-test based on the Satterthwaite approximation and on the standard ANOVA F-test generally were found to be different. For both cases, however, the p-values were very large for each of the series indicating that there is no moving seasonality in the form of annual shifts.

Table 1

Diagnostics of Stable Seasonality in Monthly Series

Series	Parameter Estimates θ_1	Θ_1	X-11-ARIMA F-Test[a]	Modified F F_{M2} (p-value in %)	Final Diagnostic[c]
IMPORTS					
1. Fodder and feed	−0.09*	−0.01	3.68	3.43(0.06)	Y
2. Coal related materials	0.02	−0.01	64.40	58.76(0.00)	Y
3. Crude vegetable products	0.02	−0.07*	3.48	2.94(0.27)	Y
4. Wool & man made materials	0.02	0.29*	10.98	20.63(0.00)	Y
5. Precious metals	0.27*	0.01	1.25	1.20(31.10)	N
6. Oils & fats	0.41*	0.01	8.59	8.22(0.00)	Y
7. Non-metal minerals	0.04	0.02	16.50	16.68(0.00)	Y
8. Aircraft engines	0.32*	0.00	2.53[b]	2.36(1.79)	N
9. Other trans. equipments	0.19*	−0.18*	3.48[b]	2.43(1.31)	N
EXPORTS					
10. Wheat	0.04	−0.03	1.89	1.71(8.71)	N
11. Asbestos	0.13*	−0.03	6.83	6.15(0.00)	Y
12. Wood pulp	−0.27	0.20*	6.45	9.61(0.00)	Y
13. Textile fabrics	0.52*	0.13*	12.05	15.06(0.00)	Y
14. Other fabrics	0.04	0.11*	5.03	6.19(0.00)	Y
15. Television & telecommunication	0.12*	0.01	9.26	8.99(0.00)	Y
16. Domestic export pass.	−0.30*	−0.14*	24.50	18.52(0.00)	Y
CPI					
17. Eggs	−0.04	−0.01	6.90	6.50(0.00)	Y
18. Pasta	−0.05*	−0.04	3.69	3.24(0.10)	Y
19. Onions	−0.42*	−0.03	26.90	23.49(0.00)	Y
20. Housing	0.11*	−0.34*	19.02	9.28(0.00)	Y
21. Clothing	0.03	−0.42*	47.42	24.30(0.00)	Y
22. Transport	−0.09*	−0.02	4.21	3.74(0.02)	Y
LABOUR					
23. Sask. employment (25-34)	−0.19*	−0.11*	67.40	52.35(0.00)	Y
24. Sask. not in labour force	0.12*	−0.36*	22.98	12.69(0.00)	Y
25. Ontario unemployment (25-44)	−0.21*	0.07*	31.4	34.23(0.00)	Y
26. Ontario unemployment male & female (20-24)	−0.02	0.19*	24.27	34.78(0.00)	Y

a Critical value is $F(11,99;\ 0.01) = 2.47$.
b X-11-ARIMA and Modified F give conflicting inference.
c Y (Yes) – stable seasonality is significant
N (No) – stable seasonality is not present.
* Significant values at 5% level.

To test for the presence of stable seasonality, we computed the p-values of the modified F-statistic F_{M2} (2.7) by using the Satterthwaite approximation and compared them to those given by the X-11-ARIMA F-test (which is equivalent to the standard ANOVA F_{A2}) for the 26 monthly series. The results are shown in Table 1.

The p-values of the modified F-statistic in Table 1 show that among the nine import series, three series do not have significant stable seasonality at the 1% significance level (the critical value of $F(11,99; 0.01) = 2.47$). Among the seven exports series, only one series, namely Wheat, appears to have no seasonality. All six CPI series have significant stable seasonality and similarly the four Labour series.

The X-11-ARIMA F-test values give same results (either rejection or acceptance of the null hypothesis) as the modified F-test for a large number of series. It seems that for most of the monthly series, under the SMA $(0,1)(0,1)_s$ error structure, the X-11-ARIMA F-test (or equivalently standard ANOVA F-test) is more affected by large negative values of Θ_1, *i.e.* when there is seasonal autocorrelation in the residuals. This can be generalized by looking at the values of $C_3(\theta,\Theta)/C_2(\theta,\Theta)$. By examining when this fraction is greater or less than 1, it may be seen that the direction of the inequality is affected by the signs of θ_1 and the size by the value of Θ_1. Only two series, namely, Imports Aircraft Engines and Imports other transportation Equipments, have standard F-test values which lead to contradictory conclusions with respect to the modified F-test. On the other hand, if we would follow the rule of thumb of $F \geq 7$ to justify seasonal adjustment, then the modified F-test would be in contradiction for eight out of twelve series. We then seasonally adjusted these eight series with the X-11-ARIMA method and found that the quality of the adjustment was acceptable for six out of the eight cases. All series passed the extrapolation ARIMA model automatically chosen for the program, six out of the eight series passed the X-11-ARIMA guidelines criteria for acceptance; and the four series for which the F_{M2} values were relatively small, that is, falling between 3.24 and 3.74 were really strongly affected by trading-day variations. Only Imports Fodder and Feed and Imports Crude Vegetable products gave a seasonally adjusted output that could not be considered reliable.

3.2 Quarterly Series

The X-11-ARIMA method was applied to four quarterly series of the System of National Accounts to obtain the detrended values (z_t, $t = 1, \ldots, 40$). It was found that for all four series U_t follow a $(0,1)(0,1)_4$ model. The computation for the modified F-test is quite similar to the case for monthly series but since the covariance matrix Σ^* is different, the formulas for $C_1(\cdot)$, $C_2(\cdot)$, and $C_3(\cdot)$ in equations (2.6) and (2.7) were adjusted accordingly.

Similar to the monthly series, the p-values for testing the presence of annual shifts based on the F_{M1} test were found very large and thus rejecting this pattern of moving seasonality.

The results of the modified F_{M2} test and the X-11-ARIMA F-test for testing for the presence of stable seasonality in each of the four series, are given in Table 2. The p-value for two series namely, Deposits in other Institutions and Small Mortgages are not significant and in agreement with those obtained from X-11-ARIMA. Thus we conclude that these two series contain significant stable seasonality. For the remaining two quarterly series, the modified F-test and the X-11-ARIMA F-test give conflicting inferences. Contrary to the X-11-ARIMA F-test, the modified F-test yields significant p-values for these two series. Thus we conclude that these two quarterly series, namely, Net Financial Investments and Corporate claims should not be seasonally adjusted.

Table 2

Diagnostics of Stable Seasonality in Quarterly Series

Series	Parameter Estimates		X-11-ARIMA F-Test[a]	Modified F F_{M2} (p-value in %)	Final Diagnostic[c]
	θ_1	Θ_1			
1. Deposits in other institutions	0.53*	0.11*	9.03	9.67(0.04)	Y
2. Net financial investments	0.77*	−0.37*	4.86[b]	2.56(8.16)	N
3. Small mortgages	0.17*	−0.01	6.65	4.88(1.02)	Y
4. Corporate claims	0.77*	−0.31*	7.88[b]	3.58(3.20)	N

a Critical value is $F(3,27; 0.01) = 4.51$.
b X-11-ARIMA and Modified F give conflicting inference.
c Y (Yes) - Stable seasonality is significant.
N (No) - Stable seasonality is not present.
* Significant values at 5% level.

4. CONCLUSIONS

This paper has introduced an exact test for the presence of stable seasonality and annual seasonal shifts based on the modified F-test by Sutradhar, MacNeill and Sahrmann (1987). The new test takes into account the possibility of autocorrelated residuals in the seasonal-irregular ratios of the X-11-ARIMA method. The residuals are assumed to follow a simple Seasonal Moving Average (SMA) model $(0,q)(0,Q)_s$. This test is applied to a set of quarterly and monthly series from the system of National Accounts, Imports, Exports, Consumer Prices and Labour. The residuals from the X-11-ARIMA method are found to follow seasonal moving average models (SMA) where either $\hat{\theta}$ and/or $\hat{\Theta}$ were significant. The exact F-test gives values very different from those of the F-test in X-11-ARIMA (also in the Census X-11 variant) when the autocorrelation of the residuals is of a seasonal character, *i.e.*, whenever $\hat{\Theta}$ is significantly different from zero.

Among the 26 monthly series analysed, only in two cases, the standard F-test values gave conflicting conclusions with respect to the modified F-test. On the other hand, if we would follow the common rule of thumb of $F \geq 7$ to justify seasonal adjustment, then the modified F-test gave contradictory results for eight out of twelve series.

By looking at the seasonal adjustment output of these eight series we found that six can be soundly seasonally adjusted by the X-11-ARIMA method.

Concerning the quarterly series, the modified F-test indicates that there is no stable seasonality in two out of the four series analysed. Furthermore, in one case, the F-test of X-11-ARIMA gives an F value greater than 7 whereas the modified F accepts the null hypothesis.

It has been assumed throughout the paper that moving seasonality may be present in the series only in the form of annual shifts. The present test is not suitable to detect other types of moving seasonal patterns in the series. This raises the necessity of further investigations in this direction.

ACKNOWLEDGEMENTS

We would like to thank an anonymous referee for his valuable comments to an earlier version of this paper.

REFERENCES

BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis, Forecasting and Control.* San Francisco: Holden-Day.

DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method.* Catalogue 12-564E, Statistics Canada.

FRANZINI, L., and HARVEY, A.C. (1983). Testing for deterministic trend and seasonal components in time series models. *Biometrika*, 70, 673-682.

PIERCE, D.A. (1978). Seasonal adjustment when both deterministic and stochastic seasonality are present. In *Seasonal Analysis of Economic Time Series*, (Ed. A. Zellner). Washington, D.C.: U.S. Bureau of the Census, 242-272.

SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.

SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1967). The X-11 Variant of Census Method II: Seasonal Adjustment Program. Technical Paper 15, Bureau of the Census, U.S. Dept. of Commerce.

SHISKIN, J., and PLEWES, T. (1978). Seasonal adjustment of the U.S. unemployment rate. *The Statistician*, 27, 181-202.

SUTRADHAR, B.C., and BARTLETT, R.F. (1989). An approximation to the distribution of the ratio of two general quadratic forms with application to time series valued designs. *Communications in Statistics - Theory and Methods*, 18, 1563-1588.

SUTRADHAR, B.C., and BARTLETT, R.F. (1990). An Exact Large and Small Sample Comparison of Wald's, Likelihood Ratio and Rao's Tests for Testing Linear Regression with Autocorrelated Errors. Technical Report, Department of Mathematics and Statistics, Memorial University of Newfoundland.

SUTRADHAR, B.C., and MacNEILL, I.B. (1989). Two-way analysis of variance for stationary periodic time series. *International Statistical Review*, 57, 169-182.

SUTRADHAR, B.C., MacNEILL, I.B., and DAGUM, E.B. (1991). A Simple Test for Stable Seasonalities. Statistics Canada, Methodology Branch, Working Paper No. TSRA-91-007.

SUTRADHAR, B.C., MacNEILL, I.B., and SAHRMANN, H.F. (1987). Time series valued experimental designs: One-Way Analysis of Variance with Autocorrelated Errors. In *Time Series and Econometric Modelling*, (Eds. I.B. MacNeill and G.J. Umphrey) Dordrecht: Reidel, 113-129.

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 163–181
Statistics Canada

A Theory of Quota Surveys

JEAN-CLAUDE DEVILLE[1]

ABSTRACT

Simple or marginal quota surveys are analyzed using two methods: (1) behaviour modelling (super-population model) and prediction estimation, and (2) sample modelling (simple restricted random sampling) and estimation derived from the sample distribution. In both cases the limitations of the theory used to establish the variance formulas and estimates when measuring totals are described. An extension of the quota method (non-proportional quotas) is also briefly described and analyzed. In some cases, this may provide a very significant improvement in survey precision. The advantages of the quota method are compared with those of random sampling. The latter remains indispensable in the case of large scale surveys within the framework of Official Statistics.

KEY WORDS: Quota surveys; Super-population models; Restricted sampling; Regression estimation.

1. INTRODUCTION

Quota sampling is the method most frequently used in France by private polling institutions. It is easy to implement, inexpensive, and has many practical advantages. However, its disadvantages are also well known: likelihood of bias, no possibility of processing non-responses, and the need for external information in order to set the quotas. In the English literature (Cochran 1977; or Madow *et al.* 1983, for example) quotas have a very bad reputation due to the lack of a reliable theory on which statistical inference can be based. The only "defenders" of the method (Smith 1983, in particular) base their arguments on the principles of inference conditional upon sampling, where the sampling plan may generally be ignored.

This paper proposes a theory of quota surveys based on two types of modelling: population behaviour modelling (which is the approach of Smith or the ideas expressed in Gourieroux 1981), and modelling the method of sample collection, which may correspond to a more realistic idea.

In both cases, variance estimates are obtained by resorting to variations of regression estimators.

The first section of the paper describes the quota method and the results of the survey theory that can be subsequently useful. Parts 2 and 3 develop models for the behaviour of individuals in a population, or of those conducting the survey, which justify the method. The last section examines the problems raised, and attempts to demonstrate how the quota method can be used to add to the traditional probabilistic methods, rather than compete with them.

2. A BRIEF REVIEW OF THE QUOTA METHOD AND SURVEY THEORY

2.1 Cell Quotas; Quotas on the Margins of a Contingency Table – Some Practical Aspects of the Method

At the simplest level, the quota method resembles stratified sampling. The distribution in the population of a discrete characteristic h possessed by N_h individuals ($h = 1$ to H) is known.

[1] Jean-Claude Deville, Institut National de la Statistique et des Études Économiques, 18, Boulevard Adolphe Pinard, 75675, Paris Cedex 14, France.

The sample includes n_h individuals in category h; however, the choice of these individuals is left up to the those conducting the survey. The sampling rate $f_h = n_h/N_h$ may vary from category to category.

In practice, we prefer to control several criteria expressed as $i, j, \ldots, h$ ($i = 1$ to I, $j = 1$ to J, $\ldots$, $h = 1$ to H). Ideally, knowing the $N_{ij\ldots h}$ values of the multiple-entry contingency table allows the use of the previous method to define the number $n_{ij\ldots h}$ of members in the sample depending upon the $f_{ij\ldots h}$ rates. Except in very specific cases (few criteria having few modalities each) this method is unrealistic, because it leads to a search for individuals who are extremely difficult to find.

Thus, it is preferable to use **marginal quotas**, by calibrating the sample so that its distribution in accordance with the first criterion leads to a given $n_{i+\ldots+}$ number of members, and the same is done for the other criteria. The only constraint on these marginal values is that they must be added to the overall sample size n. However, in practice, a single sampling rate f is adopted for each set of quotas: $n_{i+\ldots+} = fN_{i+\ldots+}$, $n_{+j\ldots+} = fN_{+j\ldots+}$ and $n_{++\ldots h} = fN_{++\ldots h}$ with the obvious notations (+ in place of an index indicates the addition of all the modalities in the category represented by the index).

Beyond the obvious collection advantages, this technique is the one most often imposed by the external data on which the quotas are based. These are obtained, for example, from various sources, thus preventing any cross-correlations. Another situation arises when the quotas are established on the basis of a large survey (a labour survey, for example): each distribution is done in accordance with a criterion (age, socio-professional category, *etc.*) that may be considered to be reliable. On the other hand, the cross-correlations are affected by a large random error, and cannot be used to set the quotas.

In practice, the quota method is most often used to complement more traditional methods as the last sampling technique used in a multi-stage stratified survey on a geographic basis (region, size of the agglomerations). Each primary unit is assigned to a survey officer for whom quotas have been set. The survey officer also receives instructions to distribute his sample in order to make data collection as close to random as possible.

2.2 Traditional Survey Theory

We want to measure the total Y of a variable whose value Y_k for individual k is fixed, with no randomness. Only sample s is random, and the law of probability that governs s is known, since it is controlled by the statistician. Thus, we also know the possibility π_k that each individual will appear in s. Without any other information, the natural (unbiased) estimator to be used is the estimator based on inflated values:

$$\hat{Y} = \sum_{k \in s} Y_k/\pi_k = \sum_s d_k Y_k \quad \text{with} \quad d_k = 1/\pi_k.$$

When the π_k are all equal to n/N, the sampling rate, we have:

$$\hat{Y} = N/n \sum_s Y_k = N\bar{y},$$

where $\bar{y}$ is the mean of Y in the sample.

This estimator has a known variance, which is a quadratic form $V(Y_U)$ on the vector of Y_k in the population:

$$\mathrm{Var}(\hat{Y}) = V(Y_U) = \sum_k Y_k(d_k - 1) + \sum_{kl} Y_k Y_l d_k d_l(\pi_{kl} - \pi_k \pi_l), \quad (2.2.1)$$

where π_{kl} is the probability of simultaneously having k and l in s.

Similarly, the variance of $\hat{Y}$ can be estimated by a quadratic form on vector Y_s of the Y_k in the sample:

$$\hat{V}(Y_s) = \sum_{kl \in s} \Delta_{kl} Y_k Y_l,$$

with

$$\Delta_{kl} = (1 - \pi_k)/\pi_k^2 \quad \text{if} \quad k = l$$

$$= (\pi_{kl} - \pi_k \pi_l)/(\pi_{kl} \pi_k \pi_l) \quad \text{if} \quad k \neq l.$$

Depending upon the sampling plans, these expressions take the specific forms found in the manuals (Desabie 1965; Cochran 1977; Wolter 1985).

Any external information can improve the quality of the estimate. This is usually presented in the form of a vector X in which each of the p components is the total of a measurable variable in each of the possible samples. The estimate of Y can thus be improved by using regression estimation:

$$\hat{Y}_{\mathrm{Reg}} = \hat{Y} + (X - \hat{X})'\hat{B},$$

where B is the vector of the coefficients of the regression of the Y_k on the X_k estimated by:

$$\hat{B} = \sum_s (d_k X_k X_k')^{-1} \sum_s d_k X_k Y_k.$$

When the constant is part of the regressors, or if it is a linear combination of the regressors and the sample has equal probabilities, the formula is simplified as follows:

$$\hat{Y}_{\mathrm{Reg}} = X'\hat{B}.$$

The variance of $\hat{Y}_{\mathrm{Reg}}$ is simply expressed by introducing the residuals of the regression $E_k = Y_k - X_k'B$ into the population. We know that we have:

$$\mathrm{Var}(\hat{Y}_{\mathrm{Reg}}) = V(E_U)$$

thus, we introduce in formula (2.2.1) vector E_U of residuals E_k. At the same time, we approximate an estimate of this variance by $\hat{V}(e_s)$, where e_s is the vector of $e_k = Y_k - X_k'\hat{B}$, the estimated residuals of the regression.

Under some sampling plans, these expressions assume particular forms. As a general rule, V and $\hat{V}$ are the positive quadratic forms, and the E_k or e_k quantities smaller than the Y_k; the regression estimator leads to substantial improvements over the inflated values.

A particularly important case that we will use later is one where X is a vector of the total accounting variables (values on the basis of which the quotas are constructed). Typically, the additional information is the vector of dimension $I + (J - 1) + \ldots + (H - 1)$ formed by the quantities: $N_{i++\ldots+}$, $N_{+j+\ldots+}$, $N_{+\ldots+h}$ for $i = 1$ to I, $j = 1$ to $J - 1$, and $h = 1$ to $H - 1$ (keeping only those variables that are linearly independent). Thus, the regressors

are the indicative variables of categories $i(i = 1 \text{ to } I)$, $j(j = 1 \text{ to } J - 1)$, and $h = 1$ to $(H - 1)$. Since the constant is a linear combination of the regressors (it is the sum of the first I of them), the regression estimator takes the form:

$$\hat{Y}_{\text{Reg}} = \sum_i N_{i++\ldots+}\hat{A}_i + \sum_j N_{+j\ldots+}\hat{B}_j + \ldots + \sum_h N_{++\ldots h}\hat{C}_h, \qquad (2.2.2)$$

where $\hat{A}_i$ (for example) indicates belonging to category i.

If we are only working with a single category, the regressors are orthogonal 2 by 2 and we have:

$$\hat{Y}_{\text{Reg}} = \sum_i N_i \hat{\bar{Y}}_i$$

where $\hat{\bar{Y}}_i$ is the estimator of the mean of Y in category i. Thus, i. $\hat{Y}_{\text{Reg}}$ is nothing but the post-stratified estimator.

2.3 Sampling Theories Based on Models

In this approach, we consider that the Y_k are random variables governed by a super-population model. This consists of parameters that we estimate on the basis of the sample. We can then calculate the probability, under the estimated model, of the non-observed values of Y, that is, $\hat{Y}_k$. The prediction estimator is the sum of the observed and predicted values and can be obtained as follows:

$$\hat{Y}_{\text{Pred}} = \sum_s Y_k + \sum_{U-s} \hat{Y}_k.$$

If, for example, in an equal probabilities survey, the model is a regression $Y_k = X_k' \cdot \beta + \epsilon_k$, ϵ_k, when the k values are independent, centred, and of equal variance, and when the constant appears on the regression (or when we have a linear combination of X_k that is constant), we have: $\sum_s Y_k = \sum_s X_k' \beta$; and the prediction estimator and the regression estimator are the same.

We say that $\hat{Y}$ is without bias under the model when, for all s, $\mathcal{E}(\hat{Y} - Y) = 0$ (conditionally upon the sample, the probability and variance under the model are expressed as $\mathcal{E}$ and $\mathcal{V}$). For the prediction estimator, we must only have, for all k, the natural condition $\mathcal{E}\hat{Y}_k = \mathcal{E}Y_k$, in order for this to be true. With the model, we can also evaluate the average quadratic deviation: $\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2$, since we know that the two terms $\hat{Y}_{\text{Pred}}$ and Y are random, and that $\hat{Y}_{\text{Pred}}$ depends upon sample s. The above-mentioned probability is thus conditional upon sample s. This follows a certain probability law already discussed in the previous paragraph. The precision of this estimator can be measured by calculating:

$$\mathcal{V}(\hat{Y}_{\text{Pred}}) = E\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2.$$

If the law of s is such that the Y_k are independent (the so-called non-informative sampling), then this quantity equals:

$$\mathcal{E}(E(\hat{Y}_{\text{Pred}} - Y)^2),$$

where the internal probability is conditional upon Y_k. If $\hat{Y}_{\text{Pred}}$ is equal to $\hat{Y}_{\text{Reg}}$, and we have a condition of independence, we will have:

$$\mathcal{V}(\hat{Y}_{\text{Pred}}) = \mathcal{E}(\text{Var}(\hat{Y}_{\text{Reg}})).$$

2.4 Comments on the Two Approaches Applied to the Quota Method

a) In both cases, the process of estimation will be effective if the variable of interest is well explained by category indicators on which the quotas are roughly based, because the regression adjustment residuals will be small.

b) In a quota survey the "sampling plan" is not known by the statistician. Thus, he cannot make inferences without using a model. The latter may be a population behaviour model ("model" approach) that requires him to assume certain responsibilities regarding the nature of what he observes. This approach will be developed in the second part of this paper. This may also consist of modelling the sampling plan; which means taking responsibility for the operation of the collection process. This approach will be developed in the third section of this paper.

In all cases, the modelling speculation must be mobilized in order to validate a kind of inference. The question is to know whether it is easier and more plausible to model the behaviour of the individuals surveyed, or to model the sample collection process (including the contacts between interviewer and interviewee).

c) In this respect, the hypothesis made in section 2.3 regarding the independence between randomness in the population and randomness in the collection process is **crucial**. If sampling is controlled by the statisticians, this guarantee can be ensured, except for the effect of non-responses. In the case of the quota method, there are no guarantees. Let us assume, for example, that we want to measure incomes Y_k, the probability π_k of finding k in the sample may be very low if Y_k is large. In other words, the fact of belonging to the sample (which is 1 if k is in s, and 0 otherwise) and the residual of the super-population model ϵ_k are negatively correlated. This example illustrates well the main danger of the quota method, which the following theory does not take into account.

3. QUOTA THEORY WITH A SUPER-POPULATION MODEL

3.1 Cell Quotas

There is a single cell category $i = 1$ to I for the known values N_i. The model that can be imagined is as follows:

$$Y_k = m_i + \epsilon_k, \tag{3.1.1}$$

ϵ_k centred independently of variance σ_i^2 where i is the cell to which k belongs.

The Gauss-Markov estimators of m_i are the means observed in the various $\bar{y}_i$ cells. Thus, the prediction estimator is:

$$\hat{Y}_{\text{Pred}} = \sum_i (N_i - n_i)\,\bar{y}_i + \sum_i n_i\bar{y}_i = \sum_i N_i\bar{y}_i. \tag{3.1.2}$$

This has the form of the post-stratified estimator. Moreover:

$$\text{Var}(\hat{Y}_{\text{Pred}} - Y)^2 = \sum_i \sigma_i^2 N_i(N_i - n_i)/n_i. \tag{3.1.3}$$

This quantity does not depend upon sample s, as the latter always includes (with a probability of 1 !) n_i individuals in cell i.

$E\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2$ can be estimated by replacing σ_i^2 by its usual estimator $s_i^2 = (n_i - 1)^{-1} \sum_{k \in s_i} (Y_k - \bar{y}_i)^2$ with s_i being part of s in cell i.

These results are from Gourieroux (1981) and represent, to a certain extent, a justification of the simple quota method.

3.2 Marginal Quotas - "Representative" Case

In this and the following paragraphs, we will restrict ourselves to the case of quotas overlapping 2 criteria i and j. The generalization with more than 2 criteria does not pose any particular problems, but leads to very complex notations that we prefer to avoid (see Appendix).

Thus, the situation is as follows: the values N_{i+} and N_{+j} of the two universe breakdowns are known. The sampling only allows samples of fixed size $n = fN$ including $n_{i+} = fN_{i+}$ individuals for each i, and $n_{+j} = fN_{+j}$ individuals for each j.

We postulate an analysis of variance model in the population, formulated as follows:

If k belongs to cell (i,j):

$$Y_k = \alpha_i + \beta_j + \epsilon_k. \tag{3.2.1}$$

The ϵ_k are centred, independent, and we have $\text{Var}\, \epsilon_k = \sigma_i^2 + \gamma_j^2$.

For reasons of identification of the model, we postulate that $\beta_J = 0$.

This is equivalent to postulating that $Y_k = (\alpha_i + u_{ik}) + (\beta_j + v_{jk})$ where u_{ik} and v_{jk} are independent, and their respective variances are σ_i^2 and τ_j^2.

We estimate α_i and β_j using the ordinary least squares (OLS) method, because we ignore the values of the variance elements; the $\hat{\alpha}_i$ and $\hat{\beta}_j$ are solutions of the system:

$$\begin{aligned} \sum_j n_{ij}\bar{y}_{ij} &= n_{i+}\hat{\alpha}_i + \sum_j n_{ij}\hat{\beta}_j \quad (i = 1 \text{ to } I) \\ \sum_i n_{ij}\bar{y}_{ij} &= n_{+j}\hat{\beta}_j + \sum_i n_{ij}\hat{\alpha}_i \quad (j = 1 \text{ to } J - 1), \end{aligned} \tag{3.2.2}$$

with $\bar{y}_{ij}$ the mean of the Y_k over the s_{ij} part of the sample in cell (i,j). Thus, the prediction estimator can be written as follows:

$$\hat{Y}_{\text{Pred}} = \sum_{ij} (N_{ij} - n_{ij})(\hat{\alpha}_i + \hat{\beta}_j) + \sum_{ij} n_{ij}\bar{y}_{ij}.$$

Result 1: Under model (3.2.1), the prediction estimator using the OLS is $N\bar{y}$. We check that it is unbiased for the model; that is, that $\mathcal{E}(N\bar{y} - Y) = 0$.

Proof: Immediately from (3.2.2), and because of the fact that the quotas are proportional to the numbers in the population.

Result 2: We have:

$$\mathcal{E}(N\bar{y} - Y)^2 = (N^2/n)(1 - f)n^{-1}\left(\sum_i n_{i+}\sigma_i^2 + \sum_j n_{+j}\tau_j^2 \right).$$

This quantity does not depend upon the sample (as it depends only upon the quotas). Thus, to a certain extent, this is a justification for the marginal quotas method.

Proof: With $m_k = \mathcal{E}Y_k$, using the unbiased character of the estimator we have:

$$\mathcal{E}(N\bar{y} - Y)^2 = \mathcal{E}\left((N/n)\sum_s (Y_k - m_k) - \sum_U (Y_l - m_l)\right)^2$$

$$= \mathcal{E}\left((N/n)\sum_s \epsilon_k - \sum_U \epsilon_l\right)^2$$

$$= (N/n)^2 \sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2) - 2(N/n)\sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2) + \sum_{ij} N_{ij}(\sigma_i^2 + \tau_j^2).$$

But

$$\sum_{ij} N_{ij}(\sigma_i^2 + \tau_j^2) = \sum_i N_{i+}\sigma_i^2 + \sum_j N_{+j}\tau_j^2$$

$$= (N/n)\left(\sum_i n_{i+}\sigma_i^2 + \sum_j n_{+j}\tau_j^2\right)$$

from which:

$$\mathcal{E}(N\bar{y} - Y)^2 = (N^2/n)(1 - f)n^{-1}\left(\sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2)\right)$$

$$= (N^2/n)(1 - f)\left(\sum_i p_{i+}\sigma_i^2 + \sum_j p_{+j}\tau_j^2\right)$$

$$\text{with} \quad p_{i+} = N_{i+}/N \quad \text{and} \quad p_{+j} = N_{+j}/N.$$

The estimate of the precision of $E\,(N\bar{y} - Y)^2$ is derived from this. In fact, with this model, s_{ij}^2 has a probability of $\sigma_i^2 + \tau_j^2$. Thus, an unbiased estimator of the precision is obtained by

$$(N/n)^2\,(1 - f)\sum_{ij} n_{ij}s_{ij}^2$$

if all the n_{ij} are equal to or greater than 2.

This estimator is formally identical to the one that we would use in a complete post-stratification on cells (i,j). We can also use $(N/n)^2\,(1 - f)\,\sum_s e_k^2$, where e_k are the estimated residuals of the model.

3.3 What Happens if the Model is False?

3.3.1 An initial way of looking at the question is to put model (3.2.1) into the general model where the mean of Y_k depends upon the pair (i,j). This can be written as follows:

$$Y_k = \alpha_i + \beta_j + \gamma_{ij} + \epsilon_k, \tag{3.3.1.1}$$

with the usual hypotheses for ϵ_k and the terms of interaction γ_{ij} that verify the constraints of identifiability:

$$\sum_j N_{ij}\gamma_{ij} = 0 \quad \text{and} \quad \sum_i N_{ij}\gamma_{ij} = 0. \tag{3.3.1.2}$$

Thus we have:

$$\mathcal{E}(N\bar{y} - Y) = \sum_{ij} (Nn_{ij}/n - N_{ij})\gamma_{ij}, \tag{3.3.1.3}$$

such that the estimator is biased for the model except when $n_{ij} = fN_{ij}$, which has no reason to exist.

This means that the terms of sum (3.3.1.3) may well compensate for each other, since their signs are *a priori* undetermined.

On the other hand, if "good" sampling precautions are taken, $Nn_{ij}/n - N_{ij}$ should usually be close to 0.

It is clear, in any case, that the more suitable the additive model is (small γ_{ij}), and the more the sampling plan approaches randomness, the more likely it is that bias will be reduced.

3.3.2 Another way to view the misrepresentation of the model, which has already been described, is to no longer admit that there is independence between the randomness of the sample and the randomness of the additive model. This means that distinct models should be developed for the $(Y_k, k \in s)$ and $(Y_l, l \notin s)$ vectors. This approach has often been used in the econometric literature, to which the reader is referred. It is clear that risk-taking in regards to the data becomes enormous, and is often incompatible with objective work on the part of the statistician.

3.4 Marginal Quotas with Unequal Rates

In the case of cell quotas, we can arbitrarily set quotas for each cell. Until now, in the case of marginal quotas, we have only examined the case where the quotas were proportional to the size of the population.

In many cases however, we may be tempted to over-represent certain categories. If, for example, we want to study household assets, we may want to set the largest quotas for older households (quotas by age group), on the one hand; and for those where the head is self-employed (quotas by social categories), on the other.

Thus, we formally force the sample to fall within a given size n_{i+} and n_{+j} (however, the sum of n_{i+} is always equal to the sum of n_{+j}).

In this case, always using the OLS as an estimation technique, we can easily find that the total prediction estimator is:

$$\hat{Y}_{\text{Pred}} = \sum_i N_{i+}\hat{\alpha}_i + \sum_j N_{+j}\hat{\beta}_j, \tag{3.4.1}$$

$\hat{\alpha}_i$ and $\hat{\beta}_j$ always verify estimating equations (3.2.2). It is easy to see that this estimator may be expressed as follows:

$$\hat{Y}_{\text{Pred}} = \sum_{ij} (w_i^{(1)} + w_j^{(2)})\, n_{ij}\bar{y}_{ij} = \sum_{ij} \hat{N}_{ij}\bar{y}_{ij}.$$

Thus, the quantities $(w_i^{(1)} + w_j^{(2)})\, n_{ij}$ seem to be estimates of the size of cells (i,j), an idea that will be largely exploited in the following sections.

On the other hand, the variance of this estimator under the model depends upon all the n_{ij}, and this can be demonstrated by a rather cumbersome calculation. The justification of the quota method described above no longer works.

4. MODELS FOR THE SAMPLING PLAN

4.1 A Model Sampling Plan

The idea is one of a simple random sampling constrained by the quotas imposed. The selection algorithm, while totally unrealistic, consists of drawing a series of simple random samples until we find one that verifies the quotas. Thus, each sample that verifies the quotas has the same positive probability of being drawn, the samples that do not verify the quotas have a zero probability of being drawn.

The purpose is to model the fact that the person conducting the survey will correctly follow the dispersion constraints on the survey units assigned to him.

4.2 Cell Quotas

This sampling model is based on an *a priori* stratification. Its practical advantage is that it does not require a sampling frame where the stratification variables are present. It is implemented rigorously in certain cases, for example, in a telephone survey based on a non-informative random list of telephone numbers, and when surveys are carried out only until the quotas are met.

The formulas that provide the estimators, the variances, and the precision estimates are those given in all the manuals. They have a certain similarity with those described in section 3.1 (see Gouriboux 1981).

4.3 The Case of Marginal Quotas: General Estimators

The sampling model is that of simple random sampling constrained by marginal quotas. SRS provides samples with n_{ij} members in the various cells that can be taken as a random vector (in whole values) in R^{IJ}. The quota constraint means that we are limited to a random vector as follows:

$$\sum_j n_{ij} = n_{i+} \ (i = 1 \text{ to } I) \quad \text{and} \quad \sum_i n_{ij} = n_{+j} \ (j = 1 \text{ to } J - 1),$$

that is, one that varies within a sub-space of size $IJ - I - J + 1$. We place ourselves in the case where the overall sampling rate is negligible, and the law of the n_{ij} can be compared to a multinomial law $(n, p_{ij} = N_{ij}/N)$.

Conditional upon n_{ij}, the $\bar{y}_{ij}$ estimate the $\bar{Y}_{ij}$ without bias. The idea is now to construct an estimator of the total of Y by weighting the $\bar{y}_{ij}$ by the estimators of N_{ij}, that is, the p_{ij}. If we choose to maximize the probability, this is proportional to:

$$\prod_{ij} p_{ij}^{n_{ij}}. \tag{4.3.1}$$

Thus, we maximize

$$\sum_{ij} n_{ij} \operatorname{Log} p_{ij} \tag{4.3.2}$$

under the following constraints

$$\sum_j p_{ij} = p_{i+} \ (i = 1 \text{ to } I) \quad \text{and} \quad \sum_i p_{ij} = p_{+j} \ (j = 1 \text{ to } J - 1) \tag{4.3.3}$$

which leads to solving the system for a_i, b_j ($p_{i+} = N_{i+}/N\, p_{+j} = N_{+j}/N$ are known):

$$\sum_j \hat{p}^\circ_{ij}\,(a_i + b_j)^{-1} = p_{i+} \quad (i = 1 \text{ to } I)$$
$$\sum_i \hat{p}^\circ_{ij}\,(a_i + b_j)^{-1} = p_{+j} \quad (j = 1 \text{ to } J - 1;\ b_J = 0), \tag{4.3.4}$$

with $\hat{p}^\circ_{ij} = n_{ij}/n$ frequency in the sample.

The estimators of p_{ij} are thus $\hat{p}^\circ_{ij}\,(a_i + b_j)^{-1}$ and the estimator we are looking for can be written as follows:

$$\hat{Y}_Q = (N/n) \sum_{ij} n_{ij}(a_i + b_j)^{-1}\, y_{ij} = (N/n) \sum_s w_k Y_k, \tag{4.3.5}$$

where $w_k = (a_i + b_j)^{-1}$ is the weight added to Y_k in the case when k appears in cell (i,j). This estimator is asymptotically without bias under the SRS model in U, as are the maximum probability estimators. The quotas do not play an explicit role in (3.3.4), but they affect the values of a_i and b_j.

In the normal case when the marginal quotas are "proportional", with a fixed sampling fraction f, the solution of equations (4.3.4) is evident: $a_i = 1$ for any i, and $b_j = 0$ for any j. The estimator of the total is $N\bar{y}$, as could be expected, and has the same expression as the equal-probability probabilistic sampling.

Comment: The use of maximum probability to estimate the proportions is rather arbitrary. A chi-square criterion (minimize $\sum_{ij} (p_{ij} - \hat{p}^\circ_{ij})^2/\hat{p}^\circ_{ij}$) would make the (4.3.4) system linear.

4.4 Variance of the Estimator and its Estimate

4.4.1 To establish a variance formula we will use the parametrization of variable Y used by J.C. Deville and C.E. Särndal (1990), which we will express in the form of a:

Lemma: For any variable $Y = (Y_k; k \in U)$, we can choose an uniquely defined parametrization

$$Y_k = \bar{Y}_{ij} + R_k \quad \text{if } k \text{ is in cell } (i,j) \quad (k \in U_{ij}) \quad \text{with} \sum_{k \in U_{ij}} R_k = 0,$$

$$\bar{Y}_{ij} = A_i + B_j + E_{ij} \quad \text{with} \quad B_J = 0$$

$$\sum_j N_{ij} E_{ij} = 0 \quad i = 1 \text{ to } I$$

$$\sum_i N_{ij} E_{ij} = 0 \quad j = 1 \text{ to } J - I.$$

In fact, A_i and B_j are numbers that minimize the quantity $\sum_U (Y_k - A_i - B_j)^2$ where, in an equivalent manner $\sum_{ij} N_{ij}\,(\bar{Y}_{ij} - A_i - B_j)^2$.

Thus, we can write:

$$\hat{Y}_Q = (N/n) \sum_{ij} n_{ij} (a_i + b_j)^{-1} (A_i + B_j + E_{ij} + \bar{R}_{ij}) \quad \text{where} \quad \bar{R}_{ij} = \sum_{s_{ij}} R_k/n_{ij}.$$

Taking into account equation 4.3.4 and the lemma:

$$\hat{Y}_Q - Y = \sum_{ij} \hat{N}_{ij} (E_{ij} + \bar{R}_{ij}) \quad \text{with} \quad \hat{N}_{ij} = (N/n)\, n_{ij} (a_i + b_j)^{-1}, \qquad (4.4.1)$$

which is the basic expression for the calculation of the variance.

Conditional upon n_{ij}, the $\hat{N}_{ij}$ are constant, and sub-samples s_{ij} are independent simple random samplings. Thus we have:

$$\text{Cond bias}(\hat{Y}_Q) = \sum_{ij} \hat{N}_{ij} E_{ij} = N \sum_{ij} \hat{p}_{ij} E_{ij}$$

$$\text{Cond Var}(\hat{Y}_Q) = \sum_{ij} \hat{N}_{ij}^2 V_{ij}/n_{ij} \quad \text{where} \quad V_{ij} = (1/N_{ij}) \sum_{U_{ij}} R_k^2.$$

Thus (demonstration in the Appendix) we have:

Result 1:

$$\text{Var}\left(\sum_{ij} \hat{p}_{ij} E_{ij} \right) = 1/n \sum_{ij} p_{ij} E_{ij}^2.$$

Furthermore, the probability of $\hat{p}_{ij}^{\circ}(a_i + b_j)^{-1}$ is (in terms close to $1/n$) $p_{ij}(a_i^{\circ} + b_j^{\circ})^{-1}$ where a_i° and b_j° are the solutions to equations (4.3.4), in which $\hat{p}_{ij}^{\circ}$ are replaced by the exact p_{ij}.

This leads to:

Result 2: The variance of the quota estimator $\hat{Y}_Q$ is given by:

$$\text{Var}(\hat{Y}_Q) = (N^2/n) \sum_{ij} p_{ij}(E_{ij}^2 + (a_i^{\circ} + b_j^{\circ})^{-1} V_{ij}).$$

If the quotas are proportional to the size of the population, we will have:

$$\text{Var}(\hat{Y}_Q) = (N^2/n) \sum_{ij} p_{ij}(E_{ij}^2 + V_{ij}).$$

4.4.2 Estimating the Variance

The conditional variance of $\hat{Y}_Q$ can be estimated by:

$$\sum_{ij} \hat{N}_{ij}^2 s_{ij}^2/n_{ij} = (N^2/n) \sum_{ij} \hat{p}_{ij}(a_i + b_j)^{-1} s_{ij}^2,$$

where s_{ij}^2 is the usual unbiased estimator of V_{ij}. The probability of the square of the conditional bias is $(N^2/n)\sum_{ij} p_{ij} E_{ij}^2$ and is estimated by $(N^2/n)\sum_{ij}\hat{p}_{ij}\hat{E}_{ij}^2$ where $\hat{E}_{ij} = \bar{y}_{ij} - \hat{A}_i - \hat{B}_j$ and $\hat{A}_i$ and $\hat{B}_j$ are the solutions of:

$$\begin{aligned} \sum_j \hat{p}_{ij}(\hat{A}_i + \hat{B}_j) &= \sum_j \hat{p}_{ij}\bar{y}_{ij} \quad (i = 1 \text{ to } I), \\ \sum_j \hat{p}_{ij}(\hat{A}_i + \hat{B}_j) &= \sum_j \hat{p}_{ij}\bar{y}_{ij} \quad (j = 1 \text{ to } J - I) \quad \text{with} \quad B_J = 0. \end{aligned} \tag{4.4.2}$$

In other words, the estimate of E_{ij} is obtained by fitting to the data an additive ANOVA model without interaction, the fitness criterion being that of least squares weighted by $(a_i + b_j)^{-1}$.

Thus, the variance estimator is:

$$\widehat{\mathrm{Var}}(\hat{Y}_Q) = (N^2/n)\sum_{ij}\hat{p}_{ij}(\hat{E}_{ij}^2 + (a_i + b_j)^{-1}s_{ij}^2). \tag{4.4.3}$$

When the quotas are proportional to the population numbers, this expression can be simplified as follows:

$$(N^2/n)\sum_{ij} n_{ij}(\hat{E}_{ij}^2 + s_{ij}^2)/n. \tag{4.4.4}$$

If the n_{ij} are all sufficiently large that $n_{ij}/(n_{ij} - 1) = 1$, the sum of the formula is the sum of the squares of the residuals estimated in the OLS adjustment of the $Y_k = A_i + B_j +$ residual model. Thus, the estimation procedure is simple:

- use the OLS to fit the additive model to the individual data
- create the variable e_k of the estimated residuals
- $\widehat{\mathrm{Var}}(\hat{Y}_Q) = (N^2/n) \cdot (1/n)\sum_s e_k^2$.

This formula is precisely that proposed in paragraph 2, and based on the super-population model. A rather neat situation!

4.4.3 Discussion of the Results

The variance breaks down into two parts: one that can be seen as the probability of the square of the conditional bias; and one as the probability of the conditional variance.

The first term does not depend upon the quotas imposed on the sample, but only upon the quality of the fit of an additive model to the variable of interest. This part of the variance is diminished by choosing quota criteria that can best explain what we want to measure.

The second term, on the other hand, depends upon the remaining variability $(N_{ij}^2 V_{ij}/n_{ij})$ and the number of observations collected in each cell. Since the size of the sample is fixed, we must attempt to make the n_{ij} as close as possible to Neyman's distribution: $n_{ij} \propto N_{ij} V_{ij}^{1/2}$. This may be achieved approximately by overloading quotas n_{i+} and n_{+j}, which correspond to large values of V_{ij}. Thus, in some cases, it is possible to improve the precision of a quota survey considerably.

4.5 Combination of the Quota Method and Stratified or Multi-Stage Samplings

4.5.1 The Case of Stratified Sampling with a Quota in Each Stratum

If the size of the criteria used to set the quotas are known in each stratum, the method described above makes it possible to construct an unbiased estimator, under the hypothesis that sampling functions like an SRS constraint in **each stratum**. If the allocation of quotas is proportional to the size of each stratum, the estimator is the natural estimator of the stratified sampling. If "national" quotas are used with each stratum, a correction should be made by reweighting.

On the other hand, if the size of the quota variables is unknown at the stratum level, it is not possible to correct the estimators to eliminate "structure effects" related to the stratification. Since, furthermore, the purpose of stratification is to construct dissimilar sub-populations, the corrections required will generally be quite large. Thus, the quota method is not recommended (except when the validity of the additive model is quite clear, *cf* part 3).

4.5.2 The Case of Two-Stage Sampling

Let us assume a two-stage sampling (inside a stratum where the sizes of the quota variables are known). If the sizes of the quota variables are known at the level of each primary unit, there are no problems. The theory in section 4.4 makes it possible to obtain an estimator of the total Y in each primary unit, as well as to calculate its variance, and an estimator of the latter. These quantities can then be used to obtain an estimator of Y, as well as an estimator of precision (*cf* Rao 1975). If the sizes of the quota criteria are not known at the level of the primary units, but only at the stratum level, we again have a problem that is impossible to correct. However, there is generally little harm if the PU are relatively similar: the structure of each PU is close to that of the stratum as a whole, and the corrections to be made for each PU are close to those that must be made at the stratum level.

4.5.3 In Conclusion

In conclusion, in the case complex multi-stage stratified sampling, the quota method may be used as the final sampling method if the stratification was carried out effectively by regrouping the similar primary units together, and if quotas derived from the data relative to each stratum are used with each PU.

To the extent that the hypothesis of simple random sampling constrained in each PU may appear to be quite satisfactory, the quota method is justified independently of any super-population model.

5. CONCLUSIONS AND PROBLEMS

5.1 How Should Non-response Be Taken into Account?

As we have already shown, this is the most important limitation in our theory. As far as sampling using the quota method is concerned, we do not have, in principle, any information on members of the population who refuse to respond to the survey, and we find ourselves lacking individual information on the subject of non-respondents. However, the situation is not as desperate as one might think. Let us illustrate this using a very simplified example.

We have carried out a simple quota survey using a sample of ni individuals in category i with a population N_i. An acceptable model of non-response postulates a response probability of r_c if an individual belongs to category c with a population N_c. The (unknown) population

of the intersection between quota category i and class c of the non-response model is expressed as N_i^c. The population likely to respond in category i is thus $N_{ri} = \sum_c N_i^c r_c$. By setting a quota n_i in this category, within the framework of model (4.1), we obtain a probability of inclusion in the sample of $w_i^{-1} = n_i/N_{ri}$. In the sample, we collect n_i^c individuals belonging to the intersection (i,c) between the two categories. This quantity is random, and its probability is $N_i^c r_c w_i^{-1}$. If we attempt to estimate N_i^c, we will solve the estimating equations derived from the following relations:

$$N_i^c = n_i^c w_i r_c^{-1},$$

$$\sum_c N_i^c = N_i,$$

$$\sum_i N_i^c = N^c.$$

Thus, ranking ratio technique makes it possible to obtain estimates of $\hat{r}_c$ and $\hat{w}_i$, and to derive estimators $\hat{N}_i^c = n_i^c \hat{w}_i \hat{r}_c^{-1}$ from the sizes of the intersection (i,c). We can also obtain an estimator of the total of Y:

$$\hat{Y}_{NR} = \sum_{ic} N_i^c \bar{y}_i^c = \sum_{ic} r_c^{-1} w_i n_i^c \bar{y}_i^c,$$

where $\bar{y}_i^c$ is the mean of the Y_k values in the sample in category (i,c). Thus, estimation techniques based on fitting should allow for the honourable processing of non-responses in quota surveys.

5.2 Some Points of Comparison with Probabilistic Surveys

Regardless of how we try to understand it, the quota method demands the formulation of a hypothetical model to fit the data. On the other hand, a probabilistic survey does not, in principle, depend upon any model. In practice, sampling for a probabilistic survey is a model to which the reality of data collection attempts to conform. In fact, we are well aware that, in any probabilistic survey, some compromises of detail must be made with the model (necessary exclusion of certain units, replacement of others after selection but before data collection, *etc*). However, we can say that statistical biases are always much lower in probabilistic selection than when using the quota method. On the other hand, quotas make it possible to use, in the sampling stage, additional information that cannot be mobilized in a probabilistic selection process. As a result, the variance of a quota sampling is similar to that of a regression estimation, and is thus generally smaller than that resulting from a probabilistic survey associated with its estimate of standard inflated values. The choice is between bias due to the model associated with low variance, against lack of bias. Two types of conclusions can be drawn from this approach:

5.2.1 Precision depends mostly upon the size of the sample. On the average, in the case of small samples, probabilistic sampling will produce the worst results; and the bias of a quota survey will be more tolerable than the lack of precision of a probabilistic survey. For large samples, on the other hand, the quota method will have a clear bias that is obviously incompatible with the confidence interval without bias of a probabilistic survey.

Where should the boundary between the two methods be set? It is hard for the theory to be specific. On the other hand, experience in the French institutes may lead to a solution to this question: most national quota surveys are carried out on samples of 1,000 to 2,000 individuals. On the other hand, no national probabilistic survey mobilizes less than 5,000 units. It would seem fair to say that a size of 2,500 to 3,000 surveys is a practical boundary between the two types of surveys.

5.2.2 Official Statistics or Marketing

In a survey, the use of any speculative model represents methodological risk-taking. This may be perfectly reasonable if the users are aware of it, and if they have ratified the speculations leading to the specification of the model. This is typically what happens, at least implicitly, in marketing surveys: an organization, company, administration, or association requests a sampling survey from a polling company. A contract marks the agreement between the two parties respecting the implementation of the survey, its price, the result delivery schedule, and **the methodology used.** In this methodology, models are used to formalize the sampling or behaviour of the population. Thus, from this point of view, the use of the quota method may be quite proper.

Official statisticians, on the other hand, are responsible for generating data that can be used by the entire society; and that can be used, in particular, in the arbitration of disputes between various groups, parties, and social classes. The use of statistical models, particularly econometric models that describe the behaviour of economic agents, may turn out to be very dangerous, partial, or affected by a questionable or disputed economic theory. Official statistics should not tolerate any uncontrollable bias in its products. It should carry out sample surveys using probabilistic methods.

There is no real opposition between quota survey techniques and those using controlled randomness, quite the opposite – they are complementary. As a proof of this, the statistics that are used to construct the quotas are themselves very often derived from large surveys carried out by the National Statistics Services. However, quota survey technicians find it hard to admit that these data are obtained using methods other than traditional, confirmed, and well-founded probabilistic techniques.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to the referee and editor for their help in improving the quality of this paper.

APPENDIX

Demonstration of the Results of Section 4.4

1. Notation and Results

In order to deal with the question in a general way, we will require certain convenient notations. We have Q qualitative variables whose modalities are indicated by using indices from 1 to I_q when $q = 1$ to Q. A "cell" is denoted as c; that is, a series of Q indices where the qth could have a value of 1 to I_q; and q_c is the value of the qth index (q^{th} projection of c); in a finite population U, of size N, U_c is the population of individuals in cell c, when the size of the cell is N_c. The quantity $N_i^{+q} = \sum_{q_c} = i N_c$ is the total of the Q-dimensional contingency table where the cells are represented by c for the ith modality of the qth variable. If we postulate that

$$\bar{Y}_c = \frac{1}{N_c} \sum_{k \in U_c} Y_k.$$

We will obtain the following results:

Result 1: Variable $Y_k (k \in U)$ may be parametrized by the following numbers: $A^q_{q_c}$, E_c and R_k by:

$$\bar{Y}_k = \bar{Y}_c + R_k \quad \text{if} \quad k \in U_c. \quad \text{We have} \quad \sum_{U_c} R_k = 0 \quad \text{for any} \quad c.$$

$$\bar{Y}_c = \sum_{q=1}^{Q} A^q_{q_c} + E_c \quad \text{with} \quad A^q_{I_q} = 0 \quad \text{for} \quad q = 2 \text{ to } Q \quad \text{and}$$

$$\sum_{q_c = i} N_c E_c = 0 \quad \text{for} \quad q = i \text{ to } Q \quad \text{and} \quad i = 1 \text{ to } I_q.$$

These numbers are obtained from the minimization of:

$$\sum_U \left(Y_k - \sum_{q=1}^{Q} A^q_{q_c(k)} \right)^2 = \sum_c N_c \left(\bar{Y}_c - \sum_{q=1}^{Q} A^q_{q_c} \right)^2.$$

Let us assume that we have a sample s. We will use n to denote all quantities in the sample that are similar to whatever we have already indicated in the population.

We assume that s was obtained on the basis of simple random sampling (with or without replacement) in accordance with an equal probability scheme constrained by the totals n_i^{+q} ($q = 1$ to Q, $i = 1$ to I_q), the quotas.

The purpose of this appendix is to demonstrate the following result:

Result 2: The variance of $\sum_c \hat{N}_c E_c$ is approximately equal to $1/n \sum_c N_c E_c^2$ when n, and N/n become arbitrarily large.

The following section will provide a more precise formulation for this result.

2. Sampling Plan and Asymptotic Reduction

Let us consider the following two sampling models SR and AR:

SR: Bernouilli Sampling. Each of the units of N belong to s with a probability f, and the N drawings are independent.

AR: Each unit is drawn a number v_k of times; v_k follows Poisson's law with f parameters. The v_k are independent variables.

A simple random survey without replacement (SRSWOR) of fixed size n is an SR sampling if the total size of the sample is n.

A simple random survey with replacement (SRSWR) of fixed size n is an AR sampling when we have n observations; that is, when $\sum_k v_k = n$.

In the case of SR sampling, the law of the vector n_c is obtained as follows:

$$\Pr(\{n_c\}) = \prod_c \binom{N_c}{n_c} f^{n_c} (1 - f)^{N_c - n_c}.$$

In the case of AR sampling, we have:

$$\Pr(\{n_c\}) = \prod_c \frac{(N_c f)^{n_c}}{n_c!} \exp(-fN_c).$$

In both cases the variables n_c are independent.

In the case of SR sampling constrained by $\sum n_c = n$, the law of the n_c is hypergeometric:

$$\Pr(\{n_c\}) = \prod_c \binom{N_c}{n_c} \binom{N}{n}^{-1}.$$

In the case of the restricted AR sampling, the law is multinomial:

$$\Pr(\{n_c\}) = \prod_c p_c^{n_c}/n_c!.$$

The sampling plan model retained by the quota method described in paragraph 3 corresponds to constraints on these two schemes; which is equivalent to constraints on the SR and AR plans.

If we assume that N tends toward infinity, that f tends towards 0, and that $n^* = fN$ tends toward infinity, then in the two plans, the law of the $u_c = n^{*-1/2} (n_c - fN_c) = n^{*1/2} (p_c^* - p_c)$, with $p_c^* = n_c/n^*$, tends toward a multidimensional normal law with independent u_c, with zero probability and variances equal to p_c.

3. Proportional Sampling

In this case, we have $\hat{N}_c = N/n\, n_c$, so that the quantity for which we want to determine the variance is:

$$\frac{N}{n^{*1/2}} \sum_c u_c E_c,$$

where the vector of the u_c follows a centered normal law with a diagonal covariance matrix $\Delta = \operatorname{diag}(p_c)$, constrained by the relationships expressed by the quotas:

$$\sum_{q_c = i} u_c = 0 \text{ for } q = 1 \text{ to } Q, \quad i = 1 \text{ for } I_q \text{ if } q = 1, \quad i = 1 \text{ for } I_q - 1 \text{ if } q = 2 \text{ for } Q.$$

If we let U represent the vector of the u_c, the relationships can be written as follows:

$$AU = 0,$$

with A matrix with $l = \sum_q I_q - (Q - 1)$ rows and $k = \Pi_q I_q$ columns, where 1 and 0 represent the constraints. This also expresses the fact that U varies in the kernel L of the operator defined by matrix A. The (asymptotic) law of U is thus that of a centered gaussian vector W with a matrix whose covariances equal Δ, when $AW = 0$. Thus, it is a question of evaluating the variance of a scalar product $U'\underline{E}$, where $\underline{E}$ is the vector of the E_c.

It is important to emphasize the following two points:

- The constraints upon the E_c given in result 1 can be expressed on the basis of matrix analysis by $A\Delta\underline{E} = 0$. In other words, $\Delta\underline{E}$ is a vector of $L = \mathrm{Ker}A$, or a vector of $\mathrm{Ker}(A\Delta)$.
- Let P be the projection of $\mathfrak{R}^k$ on L orthogonal in the Δ^{-1} metrics. P verifies the following relations:
  - $\forall x \in L,\ Px = x;\ Im\, P = L$
  - $Py = 0 \Leftrightarrow \forall x \in L,\ x'\Delta^{-1} y = 0;\ \mathrm{Ker}P = \Delta(L^{\perp})$,

where $L^{\perp}$ is the supplementary line orthogonal to L in the natural metrics.

The gaussian vectors PW and $(1 - P)W$ vary in L and $\Delta(L^{\perp})$ respectively; and their sum is equal to W. Moreover, they are independent; in fact, their covariance matrix is $E(PW)((1 - P)W)' = P\Delta(1 - P')$. Thus, P' is the kernel projector $L^{\perp}$ and can be represented as $\Delta(L^{\perp})^{\perp}$. The image of the projector $(1 - P')$ is thus $L^{\perp}$. That of $\Delta(1 - P')$ is $\Delta(L^{\perp})$; that is, the kernel of P, *q.e.d.*

At this point, we have to evaluate the variance of $\sum_c u_c E_c = U'\underline{E}$. Thus, in accordance with the previous statements, we can write $W = U + V$, when U and V are independent. The law of W conditional upon $W \in L$ is none other than the law of W conditional upon $V = 0$.

Moreover, we have:

$$V'\underline{E} = (\Delta^{-1} V)'(\Delta E).$$

Since ΔE is in L, and V varies in $\Delta(L^{\perp})$, the scalar product above is zero. From this, we can deduce that:

$$\mathrm{Var}(U'\underline{E}) = \mathrm{Var}(W'\underline{E}) = \underline{E}'\Delta\underline{E} = \sum_c p_c E_c^2.$$

The asymptotic variance of is thus equal to $N/n^* \sum_c n_c E_c$

$$\frac{N^2}{n} \sum_c p_c E_c^2 = \frac{N}{n} \sum_c N_c E_c^2.$$

4. Sampling using "Non-Proportional" Quotas

Let us complete the preceding asymptotic reduction. Now, the vector $\hat{p}^{\circ}$ of n_c/n^* is constrained by

$$A\hat{p}^{\circ} = Ap + n^{*-1/2} AV_0,$$

where Ap is the vector (1-dimensional) of the "proportional quotas", and V_0 is the only vector (k-dimensional) of $\Delta(L^{\perp})$, so that $A(p + n^{*-1/2} V_0)$; that is, the vector of the quotas imposed.

Thus, as in the previous paragraph, $U. = n^{*1/2} (\hat{p}_\circ - p)$ may be analyzed as a gaussian vector $W = U + V$ conditional upon $V = V_0$. Thus, $EU_0 = V_0$, and the covariances matrix of U_0 is the same as that of U.

Moreover, we go from $\hat{p}^\circ$ to $\hat{p}$ by estimating the maximum resemblance. Under asymptotic gaussian conditions, this consists of minimizing the quadratic form $(\hat{p}^\circ - \hat{p})'\Delta^{-1} (\hat{p}^\circ - \hat{p})$ under constraints $A\hat{p} = Ap$. Since $\hat{p}^\circ$ varies in the related subspace $L + V_0$ that is parallel to L, and minimization is a question of projecting $\hat{p}^\circ$ upon L orthogonally for Δ^{-1}; that is, along $\Delta(L^{\perp})$, it follows that we have $\hat{p} = \hat{p}^\circ - n^{*-1/2} V_0$ under asymptotic conditions. The random vector $\hat{p}$ is thus obtained from $\hat{p}^\circ$, is unbiased, and has the same covariance matrix as $\hat{p}^\circ$, so that $n^{*-1/2} U$.

Finally, we have:

$$E\left(\sum_c \hat{p}_c E_c \right)^2 = E(\hat{p}'\underline{E})^2 = \frac{1}{n^*} \sum_c p_c E_c^2$$

as in the previous case.

REFERENCES

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley & Sons.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley & Sons.

DESABIE, J. (1965). *Théorie et pratique des sondages*. Paris: Dunod.

DEVILLE, J.C., and SÄRNDAL, C.E. (1990). Calibration estimators and generalized raking techniques. Manuscript submitted for publication.

GOURIÉROUX, C. (1981). *Théorie des sondages*. Paris: Economica.

MADOW, W.G., OLKIN, I., and RUBIN, D.B., (Eds.) (1983). *Incomplete Data in Sample Surveys*. New York: Academic Press.

RAO, J.N.K. (1976). Unbiased variance estimation for multistage designs. *Sankhyā*, Series C, 37, 133-139.

SMITH, T.M.F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society*, A, 146, 394-403.

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

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 183–194
Statistics Canada

Sampling Flows of Mobile Human Populations

GRAHAM KALTON[1]

ABSTRACT

Surveys are often conducted of flows of persons, such as: visitors to museums, libraries and parks; voters; shoppers; hospital outpatients; tourists; international travellers; and car occupants. The sample designs for such surveys usually involve sampling in time and space. Methods for sampling flows of human populations are reviewed and illustrated.

KEY WORDS: Mobile populations; Exit polls; Traffic surveys; Time and space sampling; Systematic sampling.

1. INTRODUCTION

Most surveys of human populations are household based, typically with a sample of households selected with a multi-stage sample design, and individuals sampled within the selected households. The household survey is a powerful method for collecting data on a wide range of characteristics about the population, such as social, demographic, economic and health characteristics and the population's opinions and attitudes. The method is, however, not so effective for studying the characteristics of mobile populations. Two types of mobile populations may be distinguished: those who do not reside regularly at a fixed location, such as nomads and the homeless; and members of the general population who belong to the mobile population under study because they are in transit, such as visitors to libraries and parks, voters at polling booths, shoppers, hospital outpatients, travellers, and car occupants. This paper reviews sample design issues for this latter type of mobile population.

Although there are many surveys concerned with flows of mobile human populations, the general sampling literature contains little discussion of the sampling issues involved. The purpose of this paper is to describe the sample designs commonly adopted for surveys of flows of human populations, to discuss some of the special sampling issues faced, and to illustrate the range of applications for such surveys. The next section of the paper reviews the general time and space sample design used for sampling persons in transit and some of the issues involved in employing this design in particular situations. Section 3 then illustrates the application of the design in a range of different settings. Section 4 presents some concluding remarks.

2. SAMPLING IN TIME AND SPACE

It will be useful to consider a specific example in describing the general time and space sample design for sampling flows of human populations. Suppose that a survey of visitors to a summer sculpture exhibition in a city park is to be conducted to find out the visitors' socio-economic characteristics, how they heard about the exhibition, what means of transport they used to get to the park, and perhaps their views of the exhibition. Suppose that the exhibition is held from April 1 to September 30 in the year in question, that it is open from 10 a.m. until 6 p.m. daily, and that there are three sites where visitors enter and leave the exhibition.

[1] Graham Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan 48106-1248, U.S.A.

The sampling frame for a survey of this type is usually taken to be a list of time interval/site primary sampling units (PSUs). This frame is constructed by dividing the time period of the survey into a set of time intervals for each site. A simple construction of PSUs for the current example would be to divide each exhibition day at each site into two time intervals, one from 10 a.m. until 2 p.m. and the other from 2 p.m. until 6 p.m. A more complex construction of PSUs could involve time intervals of different lengths on different days and/or at different sites. Once the PSUs are defined, a two-stage sample design is often employed. At the first stage a sample of PSUs is selected, and at the second stage a sample of visitors is drawn, usually by systematic sampling, in the sampled PSUs.

The actual specification of the sample design for a survey of persons in transit within the two-stage sampling framework depends on features of the mobile population under study and of the survey data collection procedures. A key feature is the nature of the flow of the mobile population. In particular, is there a predictable variability in the rate of flow across PSUs? For instance, is the flow at one site higher than that at another site, or are the flows at some time intervals (say, Saturday afternoons) higher than those at others? Also, is the flow within a PSU a smooth one throughout the time interval or is it uneven, with visitors arriving (or leaving) in sizeable groups? Both these aspects of flow affect the sample design for the survey.

If the flow is fairly uniform across the PSUs, and if the PSU time intervals are the same, then the number of visitors per PSU is approximately constant. In this case, the PSUs may be sampled with equal probabilities, and a constant subsampling fraction can be applied within the selected PSUs to generate an equal probability, or epsem, sample of visits. The PSUs can be classified in two or more dimensions (*e.g.* day of week, time of day, and site), and a carefully balanced sample across these dimensions can be obtained using lattice sampling (Yates 1981; Cochran 1977 and Jessen 1978).

In many cases, the level of flow varies across the PSUs in a manner that is partly predictable. For instance, the attendance at the sculpture exhibition may be known to be generally higher in the later shift each day and at the weekends, and particularly low on Mondays. Thus the PSUs comprise different numbers of visitors, that is, they are PSUs of unequal sizes. The usual procedure for handling PSUs of unequal sizes is to sample them with probabilities proportional to their sizes (PPS), or estimated sizes (PPES). In the current context, the actual PSU sizes are not known in advance, and estimated sizes must therefore be used. Sampling the PSUs with PPES works well provided that reasonable estimates of the sizes can be made. When PSUs are selected by PPES sampling, then the application within the selected PSUs of subsampling fractions that are inversely proportional to the estimated sizes of the PSUs produces an overall epsem sample of visits. In general, an attraction of PPES sampling (with reasonable estimates of size) is that the subsample sizes in the PSUs do not vary greatly from one PSU to another. This feature is of especial value for conducting the fieldwork in surveys of persons in transit. When time/site PSUs are sampled by PPES sampling, lattice sampling cannot be applied for deep stratification. Instead, controlled selection may be employed for this purpose (Goodman and Kish 1950; Hess *et al.* 1975).

An important consideration in any two-stage sample design is the allocation of the sample between first-and second-stage units, that is, how many PSUs to select and how many elements to select per sampled PSU. In the case of surveys of persons in transit, that allocation is strongly affected by the fieldwork procedures to be used and the nature of the flow within the PSUs. The aim of the design is to make full use of the fieldworkers assigned to a sampled PSU while maintaining a probability sample of persons entering (or leaving) the site during the sampled time interval.

Many surveys of persons in transit use self-completion questionnaires, in which case the fieldwork process for the two-stage design described above consists of counting persons as they enter (or leave) the sampled site during the time interval, selecting every kth person for a systematic sample, and asking the selected persons to complete the questionnaire. If the flow is light and evenly spread throughout the time interval, one fieldworker may be able to handle all the tasks involved. When this is so, the sampling interval k can be chosen to give the fieldworker time to perform all the tasks in an unpressured way. If, however, the flow is heavy, either constantly or intermittently, two fieldworkers may be needed, one simply to count entrants (or leavers) and identify sampled persons, and the second to hand out the questionnaires and to instruct respondents on how they should be completed and returned. With this fieldwork arrangement, the sampling interval can be chosen to keep the second fieldworker as fully occupied as possible, while making sure that he or she is able to distribute questionnaires to all (or at least nearly all) of those sampled. Nonresponse can be a major concern with the self-completion mode of data collection. It is often possible to keep nonresponse to an acceptable level when sampled persons complete and return the questionnaire at the site. However, when they are handed the questionnaire with the request to complete it later and return it by mail, the level of nonresponse can be very high and, moreover, there is generally no way of following up the nonrespondents.

When face-to-face interviewing is used for data collection, the fieldwork team for a PSU usually contains one counter and a small team of interviewers. The size of the interviewer team depends on the regularity of the flow and the length of the interview. Since persons in transit are likely to be unwilling to be delayed for long, interviews are necessarily mostly short. Longer interviews may, however, be possible if the sampled persons are in a waiting mode, such as waiting in line or in an airport departure lounge. The choice of sampling interval has to be such that there is always (or nearly always) an interviewer free to interview the next sampled person, and that the interviewers do not spend too much time waiting for the next sampled person to be selected. If the flow is irregular, allowance needs to be made to accommodate the peaks (for instance, the arrival of a coachload of visitors to the sculpture exhibition).

The PPES selection of the PSUs works to equate the subsample size for each sampled PSU. For face-to-face interview surveys, the interviewer load is thus roughly the same for each selected PSU, and hence the same-sized interviewer team can be used for each PSU. A problem occurs, however, when the PPES measure used in selecting the PSU at the first stage is seriously in error. For example, a thunderstorm may substantially reduce the number of visitors to the sculpture exhibition on a particular Saturday afternoon, or an unforeseen holiday may substantially increase the number on another day. In the first case, applying in that PSU a sampling interval inversely proportional to its estimated size will leave the interviewers largely unoccupied, whereas in the second case it will result in a workload that the interviewers cannot handle. A modification that may be adopted in such cases is to change the sampling interval at the start of data collection to one that is more suitable for the flow actually encountered. Since this modification destroys the epsem property of the sample, weights are needed in the survey analysis.

A general limitation to the systematic sampling of visitors at selected PSUs is that if the sampling interval is made long enough to enable interviewers to cope with peak flows, they spend much of their time without work. On the other hand, if the sampling interval is reduced, the interviewers are more fully occupied, but they cannot cope with peak flows. Various methods have been proposed to circumvent these problems (Heady 1985). One procedure is to take a systematic sample of times (say, every 10 minutes) and to select the next visitor to enter after each sampled time. This procedure might have fieldwork attractions, but it does

not produce a probability sample of visitors. Persons arriving in busy periods are less likely to be chosen, as are those who travel in groups, and the walking habits of persons travelling in groups may affect the chances of selection in unknown ways. The sample generated by this procedure is clearly not an epsem sample. An attempt can be made to compensate for the selection bias that operates against visitors arriving in busy times by dividing the time interval for selected PSUs into a set of much shorter intervals, and keeping a log of arrivals in each such interval. Then weighting adjustments can be employed to compensate for the variation in the flow across the shorter intervals.

Another alternative procedure to systematic sampling of visitors is to take the next person to enter (or leave) after the last interview was completed. With this procedure, the first persons to arrive after gaps in the flow, perhaps the leaders of groups, clearly have greater chances of selection. Also interviewers may deliberately speed up or slow down their current interview in order to avoid or to select a particular individual. For these reasons, variants on this procedure that select the nth person after the completed interview, where n might be set at 2, 3, 4 or 5, have been employed. These alternatives to straightforward systematic sampling of visitors make more effective use of interviewers' time, and hence enable larger samples to be obtained for a given fieldwork budget. However, they produce nonprobability samples, with the risk of selection bias that this form of sampling entails. Probability sampling provides the security of objective statistical inference without the need for assumptions about the sample selection process. With nonprobability sampling, assumptions need to be made about the way the sample was generated, a common assumption being that all the elements in the population have an equal chance of selection. Failure of the assumptions can lead to serious bias in the survey estimates.

Visitors may be sampled either as they enter or as they leave a location. If data about the visitors' activities in and opinions of the location are required, then leavers need to be sampled. In other cases, the choice between sampling entrants and leavers may depend on the nature of the flows. It may, for example, be difficult to sample and interview people leaving a theatre because they leave en masse and because they will not want to be delayed. On the other hand, they may be readily sampled and interviewed as they line up to enter the theatre.

In concluding this section, attention should be drawn to the fact that the samples described here are samples of visits not visitors. The standard two-stage design may produce an epsem sample of visits, but this is not the same as an epsem sample of visitors unless each visitor visits the place under study (the sculpture exhibition) only once (or they all visit the same number of times). For most flow surveys, the visit, rather than the visitor, is the appropriate unit of analysis. There are, however, situations where the analytic unit is problematic. Using the visit as the unit of analysis, the researcher might readily accept visits to the sculpture exhibition on two separate days as distinct visits, but might not be willing to treat two entries on the same day (one, perhaps, after leaving briefly for refreshments) as two visits. The use of the visitor as the unit of analysis presents severe problems because of the issue of multiple visits, and the fact that visitors will not be able to report their multiplicities. They may be able to recall past visits reasonably well, but they will usually be unable to forecast future visits accurately.

3. SOME EXAMPLES

This section presents some examples of surveys of flows of human populations in order to indicate the wide range of applications and to illustrate some of the special considerations that arise in particular settings.

3.1 A Survey of Library Use

A survey of the use of the 18 libraries at the University of Michigan was conducted in 1984 (Heeringa 1985). Each sampled person exiting a library was asked whether he or she had used the library's materials and services during that visit. If so, the person was asked to complete a short self-completion questionnaire of seven questions on the materials and services used. Most of the 5,184 respondents completed the questionnaires on the spot and returned them to the survey fieldworkers; others sent them back by campus mail. A response rate of 96% was obtained.

The sample design followed the two-stage time/site sample design described in Section 2. The survey covered the full 1984 calendar year. Each day the libraries were open was divided into 10 two-hour time intervals, starting at 7.30 a.m. and lasting until 3.30 a.m. the next morning, the two-hour interval being chosen on the grounds that it was a suitable shift for the fieldworkers. The PSUs were then defined to be time interval/library combinations. The PSUs were selected by PPES sampling, where the estimated size for a PSU was the estimated number of persons exiting from that library in the specified time period. Rough estimates of these numbers were derived from average daily usage based on November, 1983, turnstile counts where available, and on librarians' estimates where not, and on an assumption that library exit volume was twice as high between 9.30 a.m. and 5.30 p.m. as at other times. The libraries were stratified into four types, and within each stratum controlled selection was employed to give a proportionate distribution of the sample across libraries, days of the week, and time intervals.

For each selected PSU, a systematic sample of persons exiting the library was selected for the survey, with the sampling interval being determined to yield an overall epsem sample of visits. Fieldworkers were provided with a record sheet of integers from 1 up to 430, with the selected numbers marked on them. All they then needed to do was check off a number for each person exiting the library, and select the persons associated with the sample numbers. An advantage of this scheme is that fractional sampling intervals are readily handled. Where the exit volume for a sampled PSU was expected to be low, one fieldworker was assigned to perform both the counting and the contacting of sampled persons. Where the exit volume was high, two fieldworkers were assigned, one to count and one to contact sampled persons. There was also a need for more than one fieldworker for libraries with more than one exit.

3.2 A Survey of Museum Visits

A face-to-face interview survey of visitors leaving the National Air and Space Museum in Washington, D.C. was conducted from mid-July until December, 1988 (Doering and Black 1989). The interview, which took about four to six minutes to complete, collected data on the sampled person's socio-demographic background, place of residence, activities on the visit, exhibits of special interest, reason for visit, the size and type of group if part of a group visit, and mode of transport used. Children under 12 years old and persons working at the museum were excluded from the survey. Data were collected from 5,574 respondents, with a response rate of 86%.

Each day in the survey period was divided into two half-days. Interviewing was conducted on one half-day every second day, alternating between mornings and afternoons. During the summer season, three public exits from the museum were in operation, while later in the year only two of them were open. During the selected half-days, survey data collection was rotated on an hourly basis between the exits that were open.

The fieldwork team for an exit at a sampled hour comprised one or two counters and two interviewers. The lead counter used a mechanical counter and a stop watch to keep track of the number of persons exiting, and to maintain a record that gave the numbers of persons exiting in each 10-minute interval in the hour. The lead counter also identified the persons to be interviewed. The selection of sample persons was made in order to keep the interviewers fully occupied. The lead counter noted when an interviewer had completed an interview and was ready to begin another one, and then chose the fifth person exiting after that time as the next sampled person. The 10-minute flow counts were used in the analysis to develop weights to compensate for the variation in the chance of selection associated with the variable flow of persons across time.

The distinction between the "visit" and the "visitor" is particularly salient for this survey. Persons could, of course, visit the museum on several days throughout the survey period, and also could visit the museum several times on a given day. This latter possibility is particularly likely with the National Air and Space Museum because entry to the Museum is free, and hence there is no incentive to enter only once. Given this situation, it may be appropriate to define multiple entries on one day as a single visit for some types of analysis. For some purposes, this definition could be applied by restricting the analysis to those exiting for the first time on the sampled day.

3.3 Exit Polls

A number of major news organizations conduct polls of voters on election days in the United States (Levy 1983; Mitofsky 1991). Voters are sampled as they leave polling places. Those selected are asked to complete a short and simple self-completion questionnaire, and to deposit the completed questionnaire in a ballot box. A typical questionnaire contains around 25 questions asking how the respondent voted, what the respondent's position is on key issues, what opinions the respondent has on various topics, and what are the respondent's demographic characteristics. Refusal rates for the CBS exit polls have averaged 25% for recent elections (Mitofsky and Waksberg 1989).

The sampling of voters for election polls usually employs a straightforward two-stage sample design. At the first stage a stratified PPES sample of voting precincts is drawn, where the size measure is the number of voters in the precinct. At the second stage a systematic sample of voters leaving the polling place is selected, with a sampling interval chosen to produce an approximately epsem sample of voters within states. Usually only one interviewer is assigned to each selected precinct. The fieldwork is straightforward when a polling place has a single exit, and the interviewer is permitted to get close to it. When there are two or more exits, interviewers alternate between the exits, covering each one for set periods of time. When this applies, the sampling interval has to be modified accordingly. In some states interviewers are not allowed to approach within a certain distance of a polling place, and this can create problems if it results in voters departing in different directions before the interviewer can contact them.

3.4 Ambulatory Medical Care Survey

The U.S. National Ambulatory Medical Care Survey (NAMCS) employs a flow survey design to collect data on visits to physicians' offices for physicians in office practice who direct patient care (Bryant and Shimizu 1988). The NAMCS has been conducted a number of times since it was introduced in 1973. For each survey, data collection has been spread throughout the survey's calendar year in order to provide annual estimates of visit characteristics. Individual sampled physicians have, however, been asked to provide information for a sample of their visits occurring in only one week. The annual coverage is achieved by asking different sampled physicians to report on different weeks of the year.

The sample for the NAMCS is based on a complex three-stage design, which has varied over time. A broad overview of the design will serve for present purposes; for more details, the reader is referred to Bryant and Shimizu (1988). The first stage of the NAMCS sample design is the selection of a stratified PPES sample of areal PSUs, selected with probability proportional to population size. At the second stage, physicians are sampled from lists within the selected PSUs with different sampling intervals from PSU to PSU to take account of the unequal selection probabilities for the PSUs (in the more recent surveys, different specialty classes are sampled at different rates). Sampled physicians are then assigned at random in a balanced way to one of the 52 reporting weeks of the year. Each physician is asked to record information for a systematic sample of his or her patient visits occurring during the sampled week, with the sampling interval being chosen to yield about 30 sampled visits in the week. A sampling interval of 1, 2, 3 or 5 is chosen for a particular physician on the basis of the number of office visits the physician expects during the week, and the number of days he or she expects to see patients. The fieldwork procedures consist of keeping a log of patient arrivals for sampling purposes, and then completing a short 16-item record for each sampled visit.

The NAMCS is a survey of patient visits not patients. As such, it provides useful information about the nature of physicians' work on a visit basis – the frequency of use of diagnostic tests, the therapies provided, and the demographic characteristics of the patients seen. It does not, however, provide estimates on a patient basis, such as treatments and outcomes for patients' episodes of illness.

3.5 Surveys of International Passengers

A number of countries conduct surveys of their international travellers, both those entering and those leaving the country by land, sea or air. This subsection will briefly describe the sample designs for a survey of international air passengers conducted by the United States, for surveys of international air and land travellers conducted by Canada, and for a survey of international air and sea passengers conducted by the U.K.

The United States Travel and Tourism Administration conducts an In-flight Survey of International Air Travelers to survey both foreign travellers to the U.S. and U.S. residents travelling abroad (see, for instance, United States Travel and Tourism Administration 1989). The survey is conducted through the voluntary cooperation of some thirty airlines. A stratified sample of scheduled flights is selected for the third week of each month and all passengers on those flights are included in the sample. Participating airlines are provided with a survey kit of instructions and questionnaires in appropriate languages for each sampled flight. The airline cabin personnel distribute the self-completion questionnaires in boarding areas or in flight to all adult passengers and collect them prior to debarkation. Nonresponse is a serious problem with these surveys. For the 1988 survey of visitors to the United States, one half of the flight kits issued resulted in no returned questionnaires. For flights for which questionnaires were returned, the estimated response rate for non-U.S. residents was 44% and for U.S. residents it was only 20%.

The International Travel Section of Statistics Canada conducts international travel surveys at both airports and landports in Canada. The surveys are undertaken in cooperation with Canada Customs, with customs officers being responsible for distributing the self-completion mail-back questionnaires. The account here is based on the report by the International Travel Section, Statistics Canada (1979). It reflects the survey designs that applied prior to some changes that have recently been made. The sample designs for the landports and airports have been similar, and therefore only the design for the landports will be outlined here.

At one time the sampling scheme at landports for returning Canadian residents who had spent at least one night abroad was to distribute survey questionnaires to every travel party

on every fourth day throughout the year, the days being chosen by systematic sampling. This scheme proved to be unworkable because the customs officers too often failed to apply it correctly. It was therefore replaced by a stint scheme in which a landport was assigned two periods, or stints, for each quarter of the year during which the questionnaires were to be distributed. The stints were expected to last from 6 to 10 days, with successive stints starting about 6½ weeks apart (Gough and Ghangurde 1977). The number of questionnaires sent to a landport for a particular stint was determined from the expected traffic at that port. The customs officers were then instructed to start the distribution of the questionnaires on a given day, and to continue to distribute them until none were left. This sample design is geared to operational limitations resulting from the use of customs officers, for whom the survey is of only secondary concern, as survey fieldworkers. The design has some major drawbacks, but perhaps a more serious concern is a response rate of 20% or less.

The U.S. and Canadian surveys of international travellers both rely on cooperation from other agencies in conducting the fieldwork. This cooperation has notable benefits in costs, but a price is paid in terms of a lack of ability to apply rigorous controls to the fieldwork procedures. The U.K. surveys of air and sea travellers employ more costly face-to-face interviewing procedures.

The 1984 U.K. International Passenger Survey included the three Heathrow terminals, Gatwick and Manchester airports as strata (Griffiths and Elliot 1987). Within each airport, days were divided into mornings and afternoons, and these periods constituted the PSUs. A stratified sample of PSUs was selected, and systematic samples of passengers were chosen in selected PSUs. A sample of PSUs for other airports was also included. Two alternative data collection procedures were used at seaports. At some seaports, interviewers sampled and interviewed passengers at the quayside. At others, the interviewers travelled on the ship, interviewing passengers during the voyage. In the former case, they worked shifts that covered several sailings, and the shift became the PSU. In the latter case, the crossings were the PSUs.

3.6 Surveys at Shopping Centers

Surveys conducted at shopping centers are of two types. One type aims to describe the shoppers' socio-economic characteristics, their areas of residence, and their shopping activities in the center. The other type uses the shopping center as a convenient location to obtain samples of people from the general population of the area.

An example of a survey of the first type is a study that was conducted to examine the impact of the opening of a hypermarket on the outskirts of the city of Southampton, England (Wood 1978). Surveys of shoppers were conducted in four neighboring shopping centers both before and after the hypermarket opened (and also at the hypermarket). At each center, the first step in the survey process was the enumeration of all the retail outlets and their hours of opening. The second step was a counting of departures of groups of shoppers from sampled shops at sampled hours, with counting being conducted for 15 minutes within the hour. The counting operation was carried out over a period of one month. Based on the counts obtained, interviews were allocated between shop types and days of the week, and to specific shops and hours. Interviewers were then instructed to interview the given number of people leaving the shop, interviewing the next person to leave after they had completed the previous interview. The sample is one of shop visits, and shoppers could visit several shops on a particular trip to the shopping center. Respondents were asked about previous visits to shops in the center on this particular trip, and also about the number of extra shops they planned to visit. These data were used to develop weights for analyses of trips.

The second type of shopping center survey uses the selected persons at shopping centers as a convenience sample of the general population. Mall intercept surveys of this type are widely used in market research (Bush and Hair 1985; Gates and Solomon 1982). The procedures are often haphazard, and the samples are potentially biased. The issues involved are reviewed by Sudman (1980), who discusses procedures for sampling shopping centers, locations at selected centers, and time periods to improve the sample designs, and by Blair (1983), Dupont (1987), and Murry *et al.* (1989).

3.7 Road Traffic Surveys

One form of road traffic survey relates to traffic passing through one or more locations. Time and space sample designs can be applied for these surveys in a relatively straightforward manner. Kish *et al.* (1961), for example, describe the sample design for an origin-destination survey of vehicles using the Port of New York Authority's bridges over and tunnels under the Hudson river during 1959. A four-stage stratified PPES sample design was used for this survey. The PSUs were combinations of eight-hour shifts and particular bridges or tunnels. A sample of these PSUs was selected at the first stage, a sample of contiguous toll lanes (locations) was selected at the second stage within selected PSUs, a sample of specific lanes was selected at the third stage within selected locations, and finally a systematic sample of vehicles was selected at selected lanes. Interviewers stayed at one sampled location for four hours, and moved each hour from one traffic lane to another according to a prescribed pattern.

Another type of road traffic survey relates to general traffic on the road. Surveys of occupants of passenger vehicles to study seat belt usage and drivers' blood alcohol concentrations are of this type. A full discussion of the complex design issues involved in such surveys is outside the present scope; instead only a few general observations will be made.

The method of data collection to be employed exerts a strong influence on the sampling procedures for a general traffic survey. Seat belt usage is mostly studied by observational methods, whereas the measurement of blood alcohol concentrations usually involves breathtesting. Shoulder belt usage of front-seat occupants can be observed in moving traffic, but lap belt usage and the seat belt usage of other occupants can be observed only when the vehicle has stopped briefly, for instance at traffic lights. Lack of street lights can preclude observation of seat belt usage at night at some sites. Breathtesting requires the vehicle to be stopped, and this can be done safely only in locations where the stopped vehicle does not hinder the other traffic. Unlike observational surveys, interview surveys that stop vehicles face a significant nonresponse problem.

An ingenious method of studying seat belt usage on interstate highways is described by Wells *et al.* (1990). For this study, an observer sat behind the driver in a passenger van that travelled at a slower speed than the prevailing traffic in the right hand lane of the highway. From that vantage point, the observer noted the shoulder belt usage of front-seat occupants of cars, light trucks, and vans that passed the observer's van in the adjacent lane.

A more usual approach to studying seat belt usage is to take observations at road intersections and freeway exits controlled by traffic lights, and sometimes at shopping centers and parking lots (Ziegler 1983; Bowman and Rounds 1989). O'Day and Wolfe (1984) describe an observational survey of seat belt use in Michigan applying this approach. They sampled a number of areal units, sampled a number of intersections with traffic signals within these areas, sampled days for observations to be taken at these intersections, and sampled five periods of one hour each between 8 a.m. and 8 p.m. for observation on each selected day. Each hour of observation was conducted at a different intersection. The hours were selected by a scheme that alternated one hour working and one hour free, with the observers moving between

intersections in the free hours. Observations of seat belt usage were taken at the selected intersections at the specified times for vehicles that stopped at the traffic lights. When more than one vehicle was stopped, observation began with the second vehicle, because of the bias associated with the first vehicle to stop at a light. In order to obtain more detailed information on the usage of child-restraints, observations were also made on vehicles entering shopping centers and rest areas.

The usual approach to analyzing observational data on seat belt usage is to calculate the proportion wearing seat belts among those observed. Brick and Lago (1988) propose an alternative measure, the proportion of estimated time front-seat occupants are belted in eligible vehicles to the total time in eligible vehicles. For their survey a probability sample of all roadway intersections, whether they had traffic signals or not, was selected. To avoid selection bias, observers were told the site they were to use for observation and the direction of the traffic to be observed in the specified 40-minute interval of observation. The time occupants were on the road was estimated as the length of the road segment leading to the intersection divided by the estimated average speed of the traffic on that segment. This estimated time was used as a weighting factor in the analysis.

The sampling considerations for roadside breathtesting surveys are broadly similar to those for seat belt usage surveys, except that the locations for data collection need to be places where vehicles can be stopped safely. In the 1986 U.S. National Roadside Breathtesting Survey, local police officers cooperated in the survey by flagging down selected drivers and directing them to the survey interviewers (Wolfe 1986). The interviews lasted about 5-6 minutes. When an interviewer finished an interview and the respondent had taken the breath test, the interviewer would signal to the police officer to stop the next passing vehicle. Interviewing was conducted for a period of two hours at each sampled location. A count was made of all the vehicles passing the location in the sampled direction during the period, and the ratio of this count to the number of interviews conducted was used as a weighting factor in the analysis.

4. CONCLUDING REMARKS

As the examples in the previous section illustrate, fieldwork considerations and the economics of data collection play major roles in the choice of sample design for surveys of persons in transit. The length of the time interval used in defining the PSUs may, for instance, be dictated by the length of a suitable workshift for the fieldworkers, and this may result in PSUs with substantial internal variation in the rate of flow. For example, in a survey of passengers arriving at a railway station, a morning interviewer workshift may include a peak flow of early morning commuters and a low rate of flow later on. If it were not for the need to make the PSU time interval conform to the fieldworkers' workshift, it would be preferable to avoid such variation in flow within PSUs since it leads to problems in how to subsample in the selected PSUs.

When the flow of persons within a PSU is uneven, the use of systematic sampling, or any epsem sampling scheme, for selecting persons creates a variable workload over time. If this variability in workload is substantial, there are difficulties in deciding how to staff the PSU for the survey fieldwork, particularly for a face-to-face interview survey. The assignment of sufficient staff to cope with peak flows is uneconomic since interviewers will then often be inactive at off-peak times. Sometimes staffing for somewhat below peak flow may be preferable. This will introduce some nonresponse at times of peak flow because no interviewer is available to conduct an interview with some sampled persons, but it will more fully use the interviewers' time.

The most effective use of the interviewers' time is to assign them to interview the first person to arrive (or leave) after they have completed their current interview. Schemes of this type suffer the disadvantage of not producing probability samples, and hence there is a risk of bias in the survey estimates. Where cost effective probability sampling designs can be devised, they are to be preferred. However, the choice of a sampling scheme in which the first (or second, or third) person is selected after an interviewer becomes free is understandably attractive for face-to-face interview surveys when the flow is very variable and unpredictable. When this kind of scheme is employed, it is useful to take counts of the flow over short intervals of time. These counts may then be used to make weighting adjustments to compensate for the unequal selection probabilities caused by the uneven flow.

ACKNOWLEDGEMENT

I would like to record my thanks to many researchers who generously provided me with information about the flow surveys with which they have been associated.

REFERENCES

BLAIR, E. (1983). Sampling issues in trade area maps drawn from shopper surveys. *Journal of Marketing*, 47, 98-106.

BOWMAN, B.L., and ROUNDS, D.A. (1989). Restraint System Usage in the Traffic Population, 1988 Annual Report. Washington, D.C.: U.S. Department of Transportation.

BRICK, M., and LAGO, J. (1988). The design and implementation of an observational safety belt use survey. *Journal of Safety Research*, 19, 87-98.

BRYANT, E., and SHIMIZU, I. (1988). *Sample Design, Sampling Variance, and Estimation Procedures for the National Ambulatory Medical Care Survey*. Vital and Health Statistics, Series 2, No. 108, Washington D.C.: U.S. Government Printing Office.

BUSH, A.J., and HAIR, J.F. (1985). An assessment of the mall intercept as a data collection method. *Journal of Marketing Research*, 22, 158-67.

COCHRAN, W.G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley.

DOERING, Z.D., and BLACK, K.J. (1989). Visits to the National Air and Space Museum (NASM): Demographic Characteristics. Working Paper 89-1, Institutional Studies, Smithsonian Institution.

DUPONT, T.D. (1987). Do frequent mall shoppers distort mall-intercept survey results? *Journal of Advertising Research*, 45-51.

GATES, R., and SOLOMON, P.J. (1982). Research using the mall intercept: State of the art. *Journal of Advertising Research*, 22, 4, 43-49.

GOODMAN, R., and KISH, L. (1950). Controlled selection – a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

GOUGH, J.H., and GHANGURDE, P.D. (1977). Survey of Canadian residents returning by land. *Survey Methodology*, 3, 215-231.

GRIFFITHS, D., and ELLIOT, D. (1987). Sampling errors on the International Passenger Survey. Unpublished paper, Social Survey Division, U.K. Office of Population Censuses and Surveys, London.

HEADY, P. (1985). Note on some sampling methods for visitor surveys. *Survey Methodology Bulletin*. U.K. Office of Population Censuses and Surveys, 17, 10-17.

HEERINGA, S.G. (1985). The University of Michigan 1984 Library Cost Study: Final Report. Institute for Social Research, University of Michigan.

HESS, I., RIEDEL, D.C., and FITZPATRICK, T.B. (1975). *Probability Sampling of Hospitals and Patients*. Ann Arbor, Michigan: Health Administration Press.

JESSEN, R.J. (1978). *Statistical Survey Techniques*. New York: John Wiley.

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*, 47, 54-67.

MITOFSKY, W.J. (1991). A short history of exit polls. In *Polling in Presidential Election Coverage*. (Eds. P. Lavrakas and J. Holley). Newbury Park, California: Sage.

MITOFSKY, W.J., and WAKSBERG, J. (1989). CBS models for election night estimation. Paper presented at American Statistical Association San Diego Winter Conference.

MURRY, J.P., LASTOVICKA, J.L., and BHALLA, G. (1989). Demographic and lifestyle selection error in mall-intercept data. *Journal of Advertising Research*, 46-52.

O'DAY, J., and WOLFE, A.C. (1984). Seat Belt Observations in Michigan - August/September 1983. Ann Arbor, Michigan: University of Michigan Transportation Research Institute.

SUDMAN, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research*, 17, 423-31.

STATISTICS CANADA (1979). Data Collection and Dissemination Methods for International Travel Statistics in Canada. International Travel Section, Statistics Canada.

UNITED STATES TRAVEL AND TOURISM ADMINISTRATION (1989). In-flight Survey: Overseas and Mexican Visitors to the United States. Survey Period: January-December 1988. Washington, D.C.: United States Travel and Tourism Administration.

WELLS, J.K., WILLIAMS, A.F., and LUND, A.K. (1990). Seat belt use on interstate highways. *American Journal of Public Health*, 80, 741-742.

WOLFE, A.C. (1986). 1986 U.S. National Roadside Breathtesting Survey: Procedures and Results. Ann Arbor, Michigan: Mid-America Research Institute.

WOOD, D. (1978). The Eastleigh Carrefour: a hypermarket and its effects. London: U.K. Department of the Environment.

YATES, F. (1981). *Sampling Methods for Censuses and Surveys* (4th ed.). London: Charles Griffin.

ZIEGLER, P.N. (1983). Guidelines for Conducting a Survey of the Use of Safety Belts and Child Safety Seats. Washington, D.C.: U.S. Department of Transportation.

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 195–210
Statistics Canada

A Sampling and Estimation Methodology for Sub-Annual Business Surveys

M.A. HIDIROGLOU, G.H. CHOUDHRY and P. LAVALLÉE[1]

ABSTRACT

A sample design for the initial selection, sample rotation and updating for sub-annual business surveys is proposed. The sample design is a stratified clustered design, with the stratification being carried out on the basis of industry, geography and size. Sample rotation of the sample units is carried out under time-in and time-out constraints. Updating is with respect to the selection of births (new businesses), removal of deaths (defunct businesses) and implementation of changes in the classification variables used for stratification, *i.e.* industry, geography and size. A number of alternate estimators, including the simple expansion estimator and Mickey's (1959) unbiased ratio-type estimator have been evaluated for this design in an empirical study under various survey conditions. The problem of variance estimation has also been considered using the Taylor linearization method and the jackknife technique.

KEY WORDS: Continuous surveys; Sample updating; Ratio estimator; Variance estimation.

1. INTRODUCTION

The universe for sub-annual business surveys continually changes on account of births, deaths, splits, mergers, amalgamations, and classification changes. The sample design associated with such a universe should have the following characteristics. Firstly, it should result in samples which reflect the changing structure of the population. Secondly, it should distribute response burden by rotating units in and out of the sample. Thirdly, if there are significant changes in the stratification of the universe, it should be possible to redraw a new sample which reflects the stratification and possible changes in sampling fractions. The resulting new sample should have maximum overlap with the previous sample in order to minimize abrupt changes in the estimates and increased costs due to the introduction of new units in the sample. The sample design which has been proposed to satisfy these requirements is that of a simple random sample of randomly formed rotation groups (clusters) within each of the strata. Each rotation group represents either a group of units or a single unit. All units within a selected rotation group are selected in the sample. Rotation of the sample takes place under the constraints that units must stay in the sample for a certain period of time and be kept out of the sample for at least a certain period of time after they have rotated out of the sample.

For given domains of interest, unbiased (or nearly unbiased) estimates are developed along with the associated measures of reliability (coefficients of variation). A desirable property of the estimation is that the estimates of domain totals should add up to the population total when the domains are exhaustive and non-overlapping. This can be ensured by using one set of weights which is independent of the domains.

In section 2, the rotation group sampling design is developed and a number of alternative estimation procedures are described in section 3. In section 4, the results of an empirical study showing the performance of these estimators under various survey conditions are given. Finally, section 5 contains some concluding remarks.

[1] M.A. Hidiroglou, G.H. Choudhry and P. Lavallée, Business Survey Methods Division, Statistics Canada, 11th floor R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.

2. SAMPLING DESIGN

2.1 Stratification and Sample Allocation

The stratification of a business universe is usually based on one or more of the following characteristics: industry, geography, and size. The size measure can be univariate (*e.g.* sales or number of employees) or multivariate (*e.g.* revenue and assets). In our context, the primary strata are cross-classifications of industry and geographic regions for which estimates are required. Within these primary strata, secondary strata are formed using the size measure of the units. The secondary strata are comprised of a completely enumerated "take-all" stratum and a number of strata called 'take-some' strata where sampling occurs. It is necessary to have a take-all stratum on account of the highly skewed nature of the business universe. The take-all stratum boundary can be determined by a method introduced by Hidiroglou (1986). This method finds the optimum boundary between the take-all and the take-some strata within each primary stratum so as to minimize the overall sample size for a given coefficient of variation. The determination of this boundary also takes into account that certain units are to be sampled with certainty irrespective of their size. These pre-specified "take-all" units are units which are to be included in the sample on account of their complex structures. An example of a unit with a complex structure could be one which operates in more than one of the primary strata. The boundaries for the take-some strata are obtained either using the cum $\sqrt{f}$ rule introduced by Dalenius and Hodges (1959) or the cum $\sqrt{x}$ rule given by Hansen *et al.* (1953). Here x is a size variable available for stratification of the units in the population.

The sample sizes for the primary strata are computed so as to satisfy planned levels of precision for certain key estimates. The computation of these sample sizes also takes into account the required allocation scheme of the units to the take-some strata. It is assumed that the information available for computing these sample sizes is well correlated with the planned key variables. Given that the take-all sample units have been taken into account, the remaining sample is allocated to the take-some strata within the primary stratum, proportional to M^q or X^q, where M is the number of units in the take-some stratum and X is the take-some stratum total for the size variable being considered. The power q where $0 \leq q \leq 1$ is chosen according to the required allocation. Letting $q = 1$ results in Neyman allocation, whereas as q approaches zero, the resulting coefficients of variation become more equal amongst the different strata provided that $S_h/\bar{X}_h$ does not vary significantly from stratum to stratum and that the finite population correction factors can be ignored. The advantages of these power allocations are discussed in Bankier (1988). The allocation can be adjusted to achieve the desired minimum sample sizes and/or maximum weights for each secondary stratum.

The reliability criteria (in terms of coefficients of variation) can be associated with the primary strata in one of two ways. Either they can be specified for each primary stratum, or, for a given global (national) coefficient of variation (c.v.), the c.v. at the primary stratum level can be determined so that the c.v.'s for each industry group and geographic region are equal. An iterative procedure is used to determine the desired c.v.'s for each of the primary strata and hence the sample size within each primary stratum, so that the planned c.v.'s at the global and marginal levels are achieved.

2.2 Sampling Scheme

For each stratum, the M population units within that stratum are randomly allocated to a predetermined number P of population rotation groups, so that initially, the number of units in each of any two rotation groups differ by at most one unit. The number of rotation groups

is a function of sampling fractions, and time-in and time-out constraints. It may be noted that in order to achieve unbiasedness, the time-in and time-out constraints may sometimes have to be violated. A simple random (SRS) sample of p rotation groups is selected from the P population rotation groups. The number of rotation groups p to be selected is determined such that p/P is approximately equal to the desired sampling fraction f. The sample consists of all the units in the p selected rotation groups. Rotation of the sample occurs by acquiring an out-of-sample rotation group and dropping an in-sample rotation group. Births are randomly allocated to the P population rotation groups, one at a time, in a systematic fashion. Deaths are removed from the stratum only if they are detected by a source independent of the survey, or if they have been dead for more than a pre-specified period of time. Methods proposed by Kish and Scott (1971) are adopted for sample updating with re-stratification due to population units changing strata. The sample update maximizes the overlap between the current and the new samples. There are obvious advantages to redrawing the sample in this fashion. First, it minimizes the introduction of new units into the sample, resulting in a smoother transition from an operational point of view, and also minimizes cost. Second, discontinuity in the estimates on account of sample redraw is kept to a minimum.

There are other sampling schemes which can be used to select the sample and rotate the units. These include Poisson and collocated sampling. The properties of these schemes have been discussed by Brewer, Early and Joyce (1972), and by Sunter (1977). Poisson sampling as defined by Hajék (1964) allows each unit in the population to be drawn in the sample independently with a given probability of inclusion. Decisions as to whether the unit is selected in the sample or not are made using an independent random draw or Bernoulli trial for each unit. Supposing that the inclusion probability of a given unit i is π_i, and that a random number u_i uniformly distributed over the interval (0,1) is generated, then the i-th unit is selected if $u_i \leq \pi_i$. This probability of inclusion corresponds to the sampling fraction of the stratum that the unit belongs to. Although the advantage of Poisson sampling lies in the simple manner in which sample rotation is exercised, it has certain disadvantages. Its main disadvantage is that the realized sample size is a random variable. This can be serious if the number of units in the stratum is small, possibly resulting in samples of size zero. Early and Brewer (1971) remedied this weakness by using a scheme known as collocated sampling. Collocated sampling is similar to Poisson sampling but reduces the variation in sample size by equispacing, at the cell level, the units over the interval (0,1). Properties of this method are provided in more detail in Brewer, Early and Hanif (1984). Whereas in Poisson sampling, the addition of births and removal of deaths do not affect the random numbers attached to existing units, the use of collocated sampling requires that these random numbers be slightly perturbed, possibly disturbing the rotation scheme by violating the time-in and time-out constraints.

The rotation group sampling scheme has several advantages over the two previously mentioned schemes. For the rotation group sampling scheme, in contrast to the Poisson scheme, the expected number of units on each rotation cycle is almost equal. The removal of dead units on a universal basis may disturb the balance of units amongst the different rotation groups. This can be remedied by periodically redrawing the sample with maximum overlap, keeping the stratification and sampling fractions unchanged. The rotation for the rotation group scheme can be performed without perturbing the units, thereby satisfying the time-in and time-out constraints. This may not necessarily be true with collocated sampling on account of the slight perturbations of the random numbers due to population births and deaths. These effects may become non-trivial over a long period of time. Another advantage of the rotation group scheme over the other two methods is that re-stratification and new sampling fractions can easily be accommodated while maximizing sample overlap.

2.2.1 Determination of the Number of Rotation Groups

Assume that for a given take-some stratum, the number of population units is M and that the desired sampling fraction is f. Let t_{in} be the desired number of occasions a unit should stay in the sample. Let t_{out} be the minimum required number of occasions a unit must stay out of the sample, once it has rotated out of the sample. The required number of population rotation groups "P" and *in*-sample rotation groups "p" are determined as follows. Let, $x = int\,[t_{in}\,(1 - f)/f + 0.5]$ where $int[\cdot]$ denotes the integer portion of the argument. Two conditions arise:

a) If $x \geq t_{out}$, then the number of in-sample rotation groups is $p = t_{in}$ and the number of population rotation groups is $P = t_{in} + x$.

b) If $x < t_{out}$, then the number of in-sample rotation groups is

$$p = int\left[\frac{f}{1-f}t_{out} + 0.5\right]$$

and the number of population rotation groups is $P = p + t_{out}$.

It must be noted that p/P is only approximately equal to f on account of the integer operations.

2.2.2 Allocation of Units to Rotation Groups

Given that at the time of initial selection, there are M population units to be allocated to P population rotation groups, two distinct cases arise with respect to the relative sizes of M and P: $M \geq P$ or $M < P$.

When $M \geq P$, at least one unit can be allocated to each population rotation group. Suppose $M = a\,P + \ell$, where $a > 0$ and $\ell \geq 0$ are integers. In order to equalize the rotation group sizes as much as possible at the time of initial selection and on subsequent occasions, the following procedure is used. A 2 by P matrix is used to assign a rotation sequence to the units that will satisfy the requirements of almost equal rotation group sizes. It is used for initial sample selection and subsequent addition of births. The first "assignment" row is labelled from 1 to P, whereas the second "rotation" row is a randomized order of the first row. The corresponding rotation group numbers in the second row determine which units are in sample at any point in time. The M population units are assigned sequentially to the assignment rotation group numbers $1, 2, \ldots, P$, the P-th unit going to the P-th assignment rotation group number. The $(P + 1)$-th unit is assigned to assignment rotation group number 1 and so on. This eventually results in having the first "ℓ" assignment rotation groups with $(a + 1)$ units and the next $(P - \ell)$ assignment rotation groups with (a) units. The rotation group to which the M-th unit is assigned is termed the last assignment rotation group. This rotation group, which is assigned rotation group number at time of initial selection, is used for assigning future births starting from the next assignment rotation group number, *i.e.* $\ell + 1$.

When $M < P$, the M population units can only be allocated to a subset of M out of the P rotation groups. These rotation groups must be as equispaced as possible to ensure that the expected sample size, $\bar{m} = fM$, will be achieved from one survey occasion to the next. For this case the allocation matrix is 2 by M. The first assignment row is labelled from 1 to M. The second rotation row is a randomization of M "z" numbers where $1 \leq z_i < z_j \leq P$ for $i \neq j$, $i = 1, \ldots, M$ and $j = 1, \ldots, M$. The "z" numbers are created as follows.

i) Find integers s and q such that $P = sM + q$ where $q < M$ and $s \geq 0$.

ii) Generate r_j ($j = 1, \ldots, M$) numbers randomly assuming the values 0 or 1, such that q of them have the value equal to 1 and $M - q$ of them have the value equal to 0.

iii) Select a random integer "b" such that $1 \leq b \leq P$.

iv) Compute $z_1 = (b + r_1 - 1) \bmod P + 1$ and $z_j = (z_{j-1} + s + r_j - 1) \bmod P + 1$ for $j = 2, \ldots, M$.

v) Randomize the "z" numbers. Let the sequence of randomized "z" numbers be $z_{i_1}, z_{i_2}, \ldots, z_{i_M}$.

Now the M population units are assigned sequentially to the M assignment rotation group numbers, thereby picking up their rotation numbers. The last assignment rotation group number is M. Future births will be assigned starting from assignment rotation group number 1.

It is now a simple matter to perform the basic functions of sample selection and updating.

2.2.3 Sample Selection and Updating

At time of initial sample selection, a given stratum will have $N = \min(M,P)$ distinct rotation groups. The units belonging to the initial sample are those whose rotation numbers are included in the closed sampling interval $[1,p]$. When $M \geq P$, the number of in-sample rotation groups n is equal to p. When $M < P$, the number of in-sample rotation groups n is approximately equal to fN on account of the equispacing.

Sample rotation is carried out by shifting the sampling interval by one rotation group at each sampling occasion in a circular fashion. On the t-th occasion, units in the sample are those whose rotation number is contained in the interval defined as

i) $[(t - 1) \bmod P + 1, (t + p - 2) \bmod P + 1]$, if $(t - 1) \bmod P \leq (P - p)$

and

ii) $[1, (p - P) + (t - 1) \bmod P] \cup [(t - 1) \bmod P + 1, P]$, otherwise.

Effectively, rotation occurs by dropping a rotation group from in-sample and acquiring a rotation group from out-of-sample in a modular fashion.

"Births" occur as a result of starting a new business activity, or a change of industrial activity of a unit from out-of-scope to in-scope for the survey. Births are stratified and given an assignment rotation group number within the stratum as follows. Assuming the last assignment rotation group number was ℓ, where $1 \leq \ell \leq P$, the q-th birth will be given the assignment rotation group number $(\ell + q) \bmod P$. The next birth will be given the assignment rotation group number $(\ell + q + 1) \bmod P$. The rotation number is then immediately obtained through the one-to-one correspondence between the assignment and rotation numbers.

"Deaths" occur as a result of the termination of business activity for in-scope units or changes of industrial activity from in-scope to out-of-scope to the survey. Deaths that occur in a take-all stratum are immediately removed from the population and sample. Deaths that are part of a take-some stratum are removed immediately if they are identified as such by a source independent of the survey process. Otherwise, they are removed after a given time period. This time period should be sufficiently long so that most of the population deaths would have been identified. Deaths in the sample and in this latter category which have not yet been removed are assigned a value of zero for estimation purposes. Classification values are also retained as such until they have been identified as changes by a source independent of the survey.

2.2.4 Periodic Resampling

The sampling frame changes continually due not only to births and deaths, but also due to changes of classification variables used in the stratification (*i.e.*, geography, industry and size). These changes in the classification variables are reflected in the estimation process by

use of domain estimation (*i.e.* estimation for sub-populations). That is, the latest classification is assigned to data for tabulation purposes, using the original sampling weight. Over a period of time, changes in classification may be sufficiently important to require the examination of the stratification and subsequent sampling rates. One solution would be to redraw an independent sample, taking into account these changes, but ignoring the current sample. Such an approach has certain disadvantages from an operational point of view. An independent redraw implies that i) the newly sampled units must be initiated into the sample, ii) time-in and time-out constraints can be violated, and iii) the estimates may change substantially. It is therefore desirable to maximize the overlap between the current sample and the new sample. The following methodology provides such a procedure for resampling. It is an adaptation of the Kish and Scott (1971) method, and is based on the property that each rotation group is a simple random sample from the population rotation groups.

At time of resampling, rotation will have occurred at different rates amongst the strata, resulting in sampling intervals with different starting and end points. Hence, assuming that rotation started at time t_1 and that we are currently at time t_2, the number of rotations that have occurred is $r = t_2 - t_1 + 1$. At time t_2, the sampling interval(s) associated with a given stratum currently labelled as k ($k = 1, 2, \ldots, K$) is(are)

$$[(r-1) \bmod P_k + 1, (r + p_k - 2) \bmod P_k + 1] \quad \text{if} \quad (r-1) \bmod P_k \leq (P_k - p_k)$$

and

$$[1, p_k - P_k + (r-1) \bmod P_k] \quad \text{and} \quad [(r-1) \bmod P_k + 1, P_k] \quad \text{otherwise.}$$

The first step associated with the resampling is to relabel the different sampling intervals, which have different starting points, into sampling intervals which have the same starting point. For the k-th stratum, the resulting sampling interval is $[1, p_k]$. Let b denote the starting point of the sampling interval at time t_2 where b is given by $(r-1) \bmod P_k + 1$. All units labelled with rotation number "g" are relabelled as $(g - b + 1)$ if $b \leq g \leq P_k$ and as $P_k - (b - g - 1)$ otherwise.

The second step is to associate with each population unit currently classified to stratum k its new stratum "h". The population units of the new h-th stratum, U_h, can therefore be expressed as the union of K non-overlapping and exhaustive sets U_{hk}, $h = 1, 2, \ldots, L$. Each set U_{hk} is comprised of population units whose new stratification is h and current stratification is k. Some of these sets may be empty.

The third step is to rank, on the 0 to 1 scale, sampling units within each set U_{hk}, taking into account their current rotation numbers. Assume that there are M_{hk} units in the set U_{hk} and that their current rotation numbers are labelled between 1 and P_k. Rank these units from 1 to M_{hk} based on their associated current rotation number. Units which have the lowest rotation numbers are assigned the lowest ranks and units which have the highest rotation numbers are assigned the highest ranks. If there are any ties, these can be broken up randomly by generating uniform random numbers. This results in the units in set U_{hk} to be ranked from 1 to M_{hk}. Next, a unit with rank "i" in set U_{hk}, $1 \leq i \leq M_{hk}$, is assigned a number $r_{hki} = (a_{hk} + i - 1)/M_{hk}$, where a_{hk} is a uniformly generated random number between 0 and 1 for each set U_{hk} within U_h. These numbers represent the current rotation groups transformed to the range 0 and 1. Assume that the new sampling fraction associated with the new stratum is f_h and that the current sampling fraction is f_k. If $f_h \geq f_k$, this implies that all units currently sampled in U_{hk} will stay in the new sample and that units in the closed interval $[0, f_h]$ will be included in the new sample. If $f_h < f_k$, this implies that units must be dropped (rotated out) from the current

sample. The units which must be dropped are those which have the lowest r_{hki} values. These represent the rotation groups which have been in the sample the longest. In order that the units in the new sample be contained in the closed interval $[0, f_h]$ it is necessary to relabel the r_{hki}'s as $r_{hki} - (f_k - f_h)$ if $r_{hki} \geq (f_k - f_h)$ and as $r_{hki} - (f_k - f_h) + 1$ otherwise. Assuming that the population units belonging to the new h-th stratum are ranked based on the ordered r_{hki}'s, define $b_{hi} = i/(M_h + 1), i = 1, 2, \ldots, M_h$. Using the b_{hi}'s, new rotation numbers will be obtained as follows. For a given new stratum h, let N_h be the number of distinct rotation groups. Form N_h disjoint intervals

$$I_u = \begin{cases} [(u-1)/N_h, u/N_h] & \text{for} \quad u=1, \ldots, N_{h-1} \\ [(N_h - 1)/N_h, 1] & \text{for} \quad u = N_h. \end{cases}$$

The union of these intervals is the closed interval $[0,1]$ D_{u_i}. For D_{u_j} the new stratum h, label the new rotation numbers as where $D_1, D_2, \ldots, D_{N_h}$ where $D_{u_i} < D_{u_j}$ for $u_i < u_j$, $u_i = 1, \ldots, N_h$. The i-th unit acquires rotation number D_u if its corresponding b_{hi} value belongs to the interval I_u. Assuming that all the M_h units have been assigned new rotation numbers in this fashion, the units in sample will be those whose rotation number belongs to the interval $[1, p_h]$.

3. WEIGHTING AND ESTIMATION

The simplest estimator which can be used in conjunction with the rotation group design described in Section 2.2 is the simple expansion (or simple domain) estimator. Although this estimator is unconditionally unbiased, it can have a large conditional bias when the rotation group sizes are not balanced. The removal of dead units can cause such an imbalance in the distribution of rotation group sizes. Other estimators which take the auxiliary rotation group size information into account have therefore been considered. These include the separate and combined ratio estimators. A drawback of the separate estimator is that its bias may accumulate in a non-trivial manner across strata. The combined estimator will have negligible bias, but possibly large variances for stratum level estimates. We have therefore evaluated the performance of an unbiased separate ratio estimator due to Mickey (1959). The penalty for achieving unbiasedness is an increase in the variance. The primary objective is to determine which of the above estimators is the most suitable one for the rotation group design. The criteria for choosing the most appropriate estimator will be based on bias and mean squared error. In order to simplify the comparisons, it will be assumed that each sampled unit has valid response data.

As mentioned earlier, the h-th stratum ($h = 1, 2, \ldots, L$) is defined at some given level of industry, geography and size. Estimates are required for domains which can span all the sampling strata or be a subset of these strata. Examples of such domains are aggregations of variables of interest at the sub-provincial level given that the sampling may have occurred at a higher level, *e.g.* province. A desirable feature of the estimates is that the sum of any non-overlapping domain set must always add up to the domain defined as their union. In order to achieve consistency, only one set of weights can be used.

Let y denote the characteristic of interest and y_{hij} be its value for the j-th unit in rotation group (cluster) i of stratum h. Let $\delta_{hij}(d)$ be an indicator variable defined as 1 if the hij-th unit belongs to domain "d", and 0 otherwise. Then, the parameter of interest is the population total $Y(d)$ given by:

$$Y(d) = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} y_{hij}(d),$$

where $y_{hij}(d) = \delta_{hij}(d)\, y_{hij}$.

As described earlier, we have a simple random sample of n_h rotation groups selected without replacement from the N_h rotation groups in the h-th stratum. Let M_{hi} be the number of units in the i-th sampled rotation group within stratum h. Without loss of generality, we can assume that the sampled rotation groups are indexed $i = 1, 2, \ldots, n_h$. Let $y_{hi}(d)$ be the total response of the units belonging to domain "d" from the i-th sampled rotation group within stratum h, *i.e.*

$$y_{hi}(d) = \sum_{j=1}^{M_{hi}} y_{hij}(d),\ i = 1, 2, \ldots, n_h.$$

We will consider a number of alternative estimators for the population parameter $Y(d)$ and their corresponding variance. The estimators considered are of the form,

$$\hat{Y}_h(d) = \sum_{i=1}^{n_h} w_{hi}\, y_{hi}(d),$$

where w_{hi} is the product of the design weight and an adjustment which reflects the estimation procedure used. Estimators of $Y(d)$ are obtained by aggregating over strata, that is,

$$\hat{Y}(d) = \sum_{h=1}^{L} \hat{Y}_h(d).$$

3.1 Estimators of Total

A. Simple Expansion Estimator

Since the probability of selecting a rotation group in the h-th stratum is n_h/N_h, the design weight is $w_{hi} = N_h/n_h$ for $i = 1, 2, \ldots, n_h$, $h = 1, 2, \ldots, L$. The simple expansion estimator is given by

$$\hat{Y}_E(d) = \sum_{h=1}^{L} N_h\, \bar{y}_h(d), \tag{3.1}$$

where

$$\bar{y}_h(d) = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}(d).$$

As mentioned earlier, this estimator is unconditionally unbiased, but it can have a large conditional bias. Moreover, it may not be very efficient because it does not make use of available auxiliary information, such as rotation group sizes. As the variation in the rotation group sizes may increase over time on account of removal of deaths, it may become more and more inefficient.

B. Separate Ratio Estimator

If the correlation between $y_{hi}(d)$ and rotation group sizes M_{hi} is large, efficiency gains can be realized through the separate ratio estimator defined as

$$\hat{Y}_{SR}(d) = \sum_{h=1}^{L} \left(\frac{M_h}{\bar{m}_h}\right) \bar{y}_h(d) \tag{3.2}$$

where

$$\bar{m}_h = n_h^{-1} \sum_{i=1}^{n_h} M_{hi}$$

and

$$M_h = \sum_{i=1}^{N_h} M_{hi}.$$

One major drawback of this estimator is that it is subject to the ratio estimation bias. Consequently, if the bias tends to be positive or negative in the majority of the strata, its accumulated effect can be quite significant when aggregating over the strata.

C. Combined Ratio Estimator

The accumulated effect of aggregation bias can be significantly reduced using a combined version of the ratio estimator. The combined ratio estimator is given by

$$\hat{Y}_{CR}(d) = M \frac{\sum_{h=1}^{L} N_h \bar{y}_h(d)}{\sum_{h=1}^{L} N_h \bar{m}_h}, \tag{3.3}$$

where $M = \sum_{h=1}^{L} M_h$.

D. Unbiased Ratio-type Estimator

The bias problem caused by the ratio estimation can be completely eliminated using the following adjusted ratio-type estimator suggested by Mickey (1959). The Mickey estimator is given by

$$\hat{Y}_{MI}(d) = \sum_{h=1}^{L} \left(\bar{r}_h(d) M_h + (N_h - n_h + 1) \left[\sum_{i=1}^{n_h} y_{hi}(d) - m_h \bar{r}_h(d) \right] \right), \tag{3.4}$$

where

$$\bar{r}_h(d) = \frac{1}{n_h} \sum_{j=1}^{n_h} \bar{r}_h^{(j)}(d); \; \bar{r}_h^{(j)}(d) = \frac{\sum_{i \neq (j)} y_{hi}(d)}{\sum_{i \neq (j)} M_{hi}}; \; m_h = \sum_{i=1}^{n_h} M_{hi}.$$

An undesirable feature of the Mickey estimator is that it can have weights less than one, including negative weights.

For the separate and combined ratio estimators, the variances are estimated using the Taylor linearization method. In the case of Mickey's estimator, a jackknife procedure is used, leaving out one rotation group at a time and re-computing Mickey's estimator for the remaining $(n_h - 1)$ rotation groups in the sample. Denote each jackknifed estimator as for $\hat{Y}_{MI,h}^{(j)}(d)$ for $j = 1, 2, \ldots, n_h$,

where

$$\hat{Y}_{MI,h}^{(j)}(d) \sum_{i \neq (j)} w_{hi}^{(j)} y_{hi}(d)$$

with

$$w_{hi}^{(j)} = [M_h - (m_h - M_{hj})] \, (N_h - n_h + 2) \, b_{hi}^{(j)} + (N_h - n_h + 2)$$

and

$$b_{hi}^{(j)} = (n_h - 1)^{-1} \sum_{i \neq (j)} \frac{1}{(m_h - M_{hj} - M_{hi})}.$$

A jackknife variance estimator of $\hat{Y}_{MI,h}(d)$ is given by

$$v_j(\hat{Y}_{MI,h}(d)) = (1 - f_h) \frac{(n_h - 1)}{n_h} \sum_{j=1}^{n_h} (z_h^{(j)}(d) - \bar{z}_h(d))^2,$$

where $z_h^{(j)}(d) = \hat{Y}_{MI,h}^{(j)}(d)$ and $\bar{z}_h(d) = n_h^{-1} \sum_{j=1}^{n_h} z_h^{(j)}(d)$.

It can be shown that all the estimators are equivalent and unconditionally unbiased when the rotation group sizes M_{hi} are all equal in each stratum h. However once the rotation group sizes (M_{hi}) become unequal, all estimators, except for the simple expansion and the Mickey estimator, are unconditionally biased. For these estimators, the magnitude of their unconditional biases and their efficiency was assessed in a simulation study which is presented next.

4. SIMULATION STUDY

The purpose of this simulation was to determine which of the four estimators of aggregate total $Y(d)$ and the stratum total $Y_h(d)$ would be the most "appropriate" for the sample design described in Section 2. For simplicity, the simulations were confined to a single variable (y), gross business income (GBI). Also, for the purpose of this simulation the domains coincided with strata. Therefore, the symbol "d" used to denote the domain will be omitted.

4.1 Description of the Study

The universe for the simulation study was defined as the set of smaller sized units belonging to the Wholesale Trade sector in the province of Québec for the May 1989 reference period. The size of each unit was based on a GBI derived from payroll deductions using a ratio model. Units whose GBI was below a given threshold were retained, resulting in a population of 10,953 units. The stratification of this population was defined on the basis of Standard Industrial Classification at the 3 digit level. This resulted in 30 strata with a minimum stratum size of 18 units. For each of the 30 strata, 16 rotation groups were formed by randomly assigning the units to the rotation groups as described in Section 2.2.2.

For each stratum h, samples of 4 rotation groups were obtained from the 16 rotation groups using simple random sampling without replacement. From each stratum there were 1,820 possible samples of size 4. Over the 30 strata there were 54,600 (30 strata times 1,820 samples per stratum) possible different estimates for the separate ratio estimation procedure. On the other hand, for the combined ratio estimator, a total of $(1{,}820)^{30}$ different estimates could be produced. For the simple expansion, the separate ratio estimator, and Mickey's estimator, all 54,600 possible samples were drawn. For the combined ratio estimator, 100,000 samples were randomly drawn from the $(1{,}820)^{30}$ possible samples.

4.2 Evaluation Criteria

The evaluation criteria involved bias and mean squared error. These are described next.

For each selected sample k, an estimate $\hat{Y}_h^{(k)}$ was produced for each stratum h and for each of the four estimators. The stratum expectation $E(\hat{Y}_h^{(k)})$ of this estimate was obtained as

$$E(\hat{Y}_h) = \frac{1}{K} \sum_{k=1}^{K} \hat{Y}_h^{(k)},$$

where K is the total number of samples drawn. It should be noted that for estimators (3.1) - (3.2) and (3.4), $E(\hat{Y}_h)$ was in fact the true expectation since all possible samples were drawn. For the combined ratio estimator (3.3), it corresponded to an unbiased estimate of the expectation. The resulting stratum bias was

$$\text{Bias}(\hat{Y}_h) \doteq E(\hat{Y}_h) - Y_h.$$

The total bias, $\text{Bias}(\hat{Y})$, was obtained by summing the stratum bias over all strata.

For estimators (3.1) - (3.2) and (3.4), we have that

$$\text{Var}(\hat{Y}_h) \doteq \frac{1}{K} \sum_{k=1}^{K} (\hat{Y}_h^{(k)} - E(\hat{Y}_h))^2$$

and

$$\text{Var}(\hat{Y}) = \sum_{h=1}^{L} \text{Var}(\hat{Y}_h).$$

For the combined ratio estimator (3.3), we have that

$$\text{Var}(\hat{Y}_h \doteq \frac{1}{K-1} \sum_{k=1}^{K} (\hat{Y}_h^{(k)} - E(\hat{Y}_h))^2$$

and

$$\text{Var}(\hat{Y}) = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{Y}^{(k)} - E(\hat{Y}))^2,$$

where $\hat{Y}^{(k)} = \sum_{h=1}^{L} \hat{Y}_h^{(k)}$ and $E(\hat{Y}) = \sum_{h=1}^{L} (\hat{Y}_h)$.

Finally, the stratum mean squared error, $\text{MSE}(\hat{Y}_h)$, of each estimator was defined as

$$\text{MSE}(\hat{Y}_h) = \text{Var}(\hat{Y}_h) + (\text{Bias}(\hat{Y}_h))^2$$

while the aggregate mean squared error, $\text{MSE}(\hat{Y})$, of each estimator was given by

$$\text{MSE}(\hat{Y}) = \text{Var}(\hat{Y}) + (\text{Bias}(\hat{Y}))^2.$$

Four criteria were used in comparing the relative behaviour of the proposed estimators. The first criterion was absolute relative bias. The stratum average absolute relative bias was computed as

$$\overline{\text{ARB}} = \frac{1}{L} \sum_{h=1}^{L} \left| \text{Bias}(\hat{Y}_h) \right| \Big/ Y_h,$$

while the aggregate absolute relative bias was computed as

$$\text{ARB} = \left| \sum_{h=1}^{L} \text{Bias}(\hat{Y}_h) \right| \Big/ Y,$$

where

$$Y = \sum_{h=1}^{L} Y_h.$$

The second criterion was the ratio of absolute bias to standard error which was called "absolute standard bias". The stratum average absolute standard bias was computed as

$$\overline{\text{ASB}} = \frac{1}{L} \sum_{h=1}^{L} \left| \text{Bias}(\hat{Y}_h) \right| \Big/ \sqrt{\text{Var}(\hat{Y}_h)}$$

while at the aggregate level, it was computed as

$$\text{ASB} = \left| \text{Bias}(\hat{Y}) \right| \Big/ \sqrt{\text{Var}(\hat{Y})}.$$

Following Cochran (1977), a reasonable value for the maximum acceptable bias over the standard error should not exceed 10% . Indeed, since the precision of an estimator is usually measured by its variance and not by its MSE, too large a bias as compared to the standard deviation would give a false impression of the precision of the estimator used.

The third criterion was efficiency, defined as the ratio of the root mean squared error of the estimator under study, $\text{RMSE}(\hat{Y}^{EST})$, to that of the simple expansion estimator $\text{RMSE}(\hat{Y}^{EST})$. The stratum average relative efficiency was computed as

$$\overline{\text{EFF}} = \frac{1}{L} \sum_{h=1}^{L} \left\{ \text{RMSE}\left(\hat{Y}_h^{EXP}\right) \Big/ \text{RMSE}\left(\hat{Y}_h^{EST}\right) \right\},$$

while at the aggregate level, the relative efficiency was computed as

$$\text{EFF} = \text{RMSE}\left(\hat{Y}^{EXP}\right) \Big/ \text{RMSE}\left(\hat{Y}^{EST}\right).$$

Finally, the fourth criterion was to observe the proportion of negative weights.

4.3 Description of the Scenarios

Four different scenarios were considered for the possible configuration of the population of rotation groups for the rotation group sample design described in Section 2.2. The four scenarios provided different combinations of the rotation group size balance (good, poor) and of the correlation between the rotation group sizes M_{hi} and the survey variable y_{hi} (good, scattered). In the context of rotation group balance, "good" means that the rotation groups do not differ much in size, whereas "poor" means that they differ significantly. In the context of correlation, "good" means that the correlation between the survey variable and the rotation group size is quite high throughout the strata, whereas "scattered" means that it varies from low to high amongst the strata.

These scenarios represent possible configurations that will arise as the survey progresses through time. Scenario 1 reflects the survey at time of initial selection: for this case, the balance of rotation group sizes is good, and the correlation between rotation group sizes and the survey variable is good. Scenario 2 reflects the deterioration of the correlation (scattered) between the rotation group size and the survey variable as time progresses, due to dead units accumulating in the population. For this scenario, since the dead units have not been removed from the population, the balance in rotation group sizes is good, but the correlation between the survey

variable and the rotation group size is weakened. Scenario 3 implies that removal of the dead population units may result in imbalance of the rotation group sizes (poor), but strengthening the correlation (good) between rotation group size and the survey variable. Finally scenario 4 represents the worst possible case, which is poor correlation between rotation group size and the survey variable, and poor balance in rotation group sizes.

Scenario 1 was constructed by varying the rotation group sizes and leaving the GBI values y_{hi} unchanged for all the rotation groups. The 16 rotation group sizes were varied by sorting them in ascending order of y_{hi}. Their size was set as follows. For rotation groups 1-4, 5-8, 9-12 and 13-16, the rotation group size was set to $0.22\,M_h/4$, $0.24\,M_h/4$, $0.26\,M_h/4$ and $0.28\,M_h/4$ respectively. The average correlation between the GBI and the rotation group sizes was 0.86, ranging from 0.69 to 0.96 at the individual stratum level. The average coefficient of variation of the rotation group sizes was 9.2%.

For scenario 2, the population units were randomly permuted and assigned systematically to one of 16 rotation groups, using the procedure described in Section 2.2.2. Approximately 20% of the population units were then randomly assigned a y-value of zero to represent a high proportion of dead units. The overall correlation between the GBI and the rotation group sizes was 0.11, ranging from -0.23 to 0.74 at the individual stratum level. The average coefficient of variation of the rotation group sizes was 4.1%.

For scenario 3 the procedure was similar to scenario 1 except that the rotation group sizes differed. For rotation groups 1-4, 5-8, 9-12 and 13-16, the rotation group size was set as $0.05\,M_h/4$, $0.20\,M_h/4$, $0.30\,M_h/4$ and $0.45\,M_h/4$ respectively. The overall correlation between the GBI and the rotation group sizes was 0.87, ranging from 0.70 to 0.96 at the stratum level. The average coefficient of variation of the rotation group sizes was 60.2%.

For scenario 4 a random rotation group size was assigned independently of the GBI values as follows. Suppose that for each stratum h, $a_h = \min\{M_{hi}: i = 1, \ldots, N_h\}$ and $b_h = \max\{M_{hi}: i = 1, \ldots, N_h\}$. For each stratum h, the size M_{hi}^* for rotation group i was set to $r_h\, e_{hi}$ where e_{hi} is uniformly distributed on the interval (a_h, b_h). Here r_h is a scaling factor such that $M_h = \sum_{i=1}^{N_h} M_h^*$. The average correlation was 0, ranging from -0.49 to 0.56 at the stratum level. The average coefficient of variation of the rotation group sizes was 49.2%.

4.4 Discussion of Results

Based on the 4 scenarios described in the previous section, simulations were performed to compute the absolute relative bias (ARB), the absolute standard bias(ASB), the efficiency (EFF), and the proportion of weights less than or equal to 0. Those quantities were computed for each individual stratum and at the aggregate level. The results are given in Tables 1 to 3. Note that all of these results are presented as percentages.

In terms of absolute relative bias (ARB), as shown in Table 1, both the simple expansion and Mickey's estimator have no bias, as expected, neither at the overall nor at the stratum level. The separate ratio estimator displays the most absolute relative bias while the combined ratio estimator displays the least relative bias. For the biased estimators, the absolute relative bias increases as the coefficient of variation of the rotation group sizes increases, and the correlation between the rotation group sizes and the variable of interest decreases.

Turning to absolute standard bias (ASB), as shown in Table 2, the following observations can be made. The separate ratio estimator is unacceptable for most scenarios using this criterion. Its performance worsens as the variation in rotation group sizes increases, and as the correlation between the rotation group sizes and as the variable of interest decreases. The performance of the combined ratio estimator is acceptable, both at the aggregate and stratum level.

Table 1
Percentage Absolute Relative Bias ($\overline{ARB}$)

Scenario	Aggregate Level		Stratum Level	
	Separate Ratio	Combined Ratio	Separate Ratio	Combined Ratio
1	1.27	0.07	1.31	0.11
2	0.02	0.01	0.24	0.05
3	2.88	0.14	3.19	0.29
4	5.51	0.22	5.72	0.30

Table 2
Percentage Absolute Standard Bias ($\overline{ASB}$)

Scenario	Aggregate Level		Stratum Level	
	Separate Ratio	Combined Ratio	Separate Ratio	Separate Ratio
1	13.41	0.76	3.37	0.24
2	0.44	0.26	0.58	0.25
3	45.64	2.13	12.11	0.69
4	43.29	1.96	9.88	0.71

The behaviour of the estimators with respect to relative efficiency (EFF) is provided in Tables 3a and 3b. For Scenario 1, which represents good rotation group balance and good correlation, all the estimators are nearly equivalent, both at the aggregate and the stratum levels. For Scenario 2, which represents well balanced rotation groups and scattered correlation, the same conclusion holds. For Scenario 3, which represents poor rotation group balance and good correlation between the rotation group sizes and the survey variable, the ranking of the estimators at the aggregate level from highest EFF to lowest EFF is: i) the combined ratio, ii) the separate ratio estimator, iii) Mickey's estimator, and iv) the simple expansion estimator. For Scenario 4, which represents the worst in terms of rotation group balance and correlation between the rotation group sizes and the survey variable, the best estimator at both the aggregate and stratum levels is the simple expansion estimator. The combined ratio estimate is the next best choice.

Weights smaller than zero occured for the Mickey estimator in 2% of the cases.

In conclusion, given the above four scenarios, the combined ratio estimator is a reasonable choice for estimation for sub-annual surveys which use the rotation group design. The simple expansion estimator may also be considered on account of its simplicity. However, one should be aware of its poor conditional properties if the rotation group sizes are not balanced.

Table 3a
Percentage Relative Efficiency (EFF) at the Aggregate Level

Scenario	Simple Expansion	Separate Ratio	Combined Ratio	Mickey
1	100.0	108.0	107.9	107.3
2	100.0	100.2	99.8	100.1
3	100.0	148.3	160.3	143.5
4	100.0	74.3	92.3	84.3

Table 3b
Percentage Average Relative Efficiency ($\overline{EFF}$) at the Stratum Level

Scenario	Simple Expansion	Separate Ratio	Combined Ratio	Mickey
1	100.0	109.6	108.6	108.6
2	100.0	100.9	99.5	100.6
3	100.0	183.3	180.2	174.2
4	100.0	80.0	99.4	83.7

5. CONCLUSION

In this paper, we have presented a sample design which can accommodate the necessary requirements for a sub-annual business survey. These requirements have included initial sample selection, sample rotation and updating. Given this rotation group design, a number of estimation procedures have been considered and they have been evaluated via a simulation study. These estimation procedures are equivalent when the rotation group sizes are well balanced within each of the strata. In the case of unbalanced rotation group sizes, the use of the combined ratio estimator which used rotation group sizes as auxiliary information is recommended.

ACKNOWLEDGEMENTS

The authors would like to thank Lyne Guertin for the programming of the simulation study. We would also like to thank the referees and Professor J.N.K. Rao for valuable comments and constructive suggestions.

REFERENCES

BANKIER, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

BREWER, K.R.W., EARLY, L.J., and HANIF, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15-30.

BREWER, K.R.W., EARLY, L.J., and JOYCE S.F. (1974). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 232-239.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Editon). New York: John Wiley.

DALENIUS, T., and HODGES, J.L., Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

EARLY, L.J., and BREWER, K.R.W. (1971). Some estimators for arbitrary probability sampling. Master's thesis.

HAJÉK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley and Sons.

HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.

RAO, J.N.K., and KUZIK, R.A. (1974). Sampling errors in ratio estimation. *Sankhyā*. Series, C, 36, 43-58.

SUNTER, A.B. (1977). Response burden, sample rotation and classification renewal in economic surveys. *International Statistical Review*, 45, 209-222.

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 211–225
Statistics Canada

County Estimates of Wheat Production

ELIZABETH A. STASNY, PREM K. GOEL
and DEBORAH J. RUMSEY[1]

ABSTRACT

Although farm surveys carried out by the USDA are used to estimate crop production at the state and national levels, small area estimates at the county level are more useful for local economic decision making. County estimates are also in demand by companies selling fertilizers, pesticides, crop insurance, and farm equipment. Individual states often conduct their own surveys to provide data for county estimates of farm production. Typically, these state surveys are not carried out using probability sampling methods. An additional complication is that states impose the constraint that the sum of county estimates of crop production for all counties in a state be equal to the USDA estimate for that state. Thus, standard small area estimation procedures are not directly applicable to this problem. In this paper, we consider using regression models for obtaining county estimates of wheat production in Kansas. We describe a simulation study comparing the resulting estimates to those obtained using two standard small area estimators: the synthetic and direct estimators. We also compare several strategies for scaling the initial estimates so that they agree with the USDA estimate of the state production total.

KEY WORDS: Non-probability sample; Regression; Simulation; Small area estimation.

1. INTRODUCTION

County estimates of farm production are more and more in demand by government agencies for use in local economic decision making and by companies selling fertilizers, pesticides, crop insurance, and farm equipment. The United States Department of Agriculture (USDA) is currently implementing a program to standardize and improve county estimates of farm production (Bass *et al.* 1989). County estimation programs in the past have been carried out individually within each state. Because of this there has been little consistency across states in data collection and estimation methods used to produce county estimates. The goal of the USDA program for county estimation is to provide a set of sampling and estimation procedures for the states so that county estimation programs across the United States may yield estimates of comparable quality.

The new USDA county estimation program encompasses every stage of the production of county estimates from the construction of sampling frames through the estimation itself. The research described here is concerned only with the estimation of bushels of wheat produced. We hope, however, that our methods may prove useful in other aspects of the county estimation program, for example in estimating acres planted and for crops other than wheat.

Although the county estimation procedures used in the past varied from state to state, some parts of the procedures were similar. A typical procedure involved obtaining initial estimates from the data available within each county. Then an expert would review the estimates, alter them in light of his personal knowledge of the farms in the sample, weather conditions, and other factors, and then note the implications of the adjustments on the estimated total production for the state. The expert might repeat this process for a number of iterations until the

[1] Elizabeth A. Stasny, Prem K. Goel and Deborah J. Rumsey, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, USA.

estimates within each county seemed reasonable and the resulting state production total agreed with the USDA state estimate. (The USDA state total is estimated based on a large probability sample and is thus thought to be a more accurate estimate than the total based on the county estimation procedure. For this reason, states typically constrain their county estimates to sum to the USDA estimate.)

Written documentation of the current county estimation procedures, as outlined above, typically is not available. Thus the assumptions and methods that the expert uses can not be inspected by others and it is practically impossible to study the procedures or replicate calculations. In addition, one cannot obtain variance estimates or use the procedures of one state in another state. New methods for county estimation must address these problems.

The data that we use in this research were collected in Kansas in 1987, before the new USDA county estimation sampling procedures were in use. Data from 1987 were used because the United States Agricultural Census was taken in that year and we may, therefore, use the Census data in our estimation procedure. Kansas data were chosen for use in this study because the county data collection program in Kansas was one of the more comprehensive programs in the United States. Nevertheless, the data used for county estimation in Kansas, as in most other states, were not collected from a probability sample of farms. Therefore, our estimation procedure must not require a probability sample of wheat farms. Such a procedure may also be useful under the new county estimation program since states still will not be required to choose probability samples of farms.

There is much recent research on small area estimation (see for example Platek *et al.* 1987). Standard small area estimation procedures, however, require known selection probabilities since the inverses of these probabilities are used to weight observations in standard estimators such as synthetic and direct estimators. (See for example Section 2 of Särndal and Hidiroglou 1989 for a discussion of standard small area estimators.)

The methods considered here must be different from the usual small area estimation techniques. First, the sample of farms available to produce county estimates is not typically a probability sample. Second, the county estimates must be constrained to sum to the USDA-produced state totals. Since most state agriculture departments currently do not have large computing facilities, an additional preliminary constraint on the estimation procedure is that computations must be simple enough to be performed on a personal computer. Thus, for our initial efforts, we prefer to avoid computationally intensive estimators such as those described by Fay and Herriot (1979). For these reasons, we consider a computationally simple estimator based on a regression model for producing county estimates of wheat production.

In Section 2 of this paper we describe the Kansas data bases used in this study. Section 3 presents the regression procedure for estimating wheat production while Section 4 describes several methods for scaling those estimates to the USDA state total. In Section 5 estimates from the regression models are obtained and compared to the published county estimates and to estimates produced using the synthetic estimator and the direct estimator. In Section 6 we present the results of a simulation study conducted to compare these same estimators. Section 7 gives conclusions and areas for future research.

2. KANSAS DATA

For the purpose of reporting farm production, all states are divided into nine or ten districts. Kansas is divided into nine districts such that each of the 105 counties in Kansas is completely contained within one of the districts. The locations of the districts and the number of counties within each district are as shown below:

District Number	District Location	Countries in District
1	Northwest	8
2	West Central	9
3	Southwest	14
4	North Central	11
5	Central	11
6	South Central	13
7	Northeast	11
8	East Central	14
9	Southeast	14

Two data bases which are used in the production of Kansas county estimates, the Planted Acres Data Base and the Small Grain Data Base, were available for our use in this research. Most of our work was done with 1987 data but we also verified our results with the 1988 data. The 1987 Planted Acres Data Base contains information on planted acreage for 37,094 farms throughout Kansas. (A farm is defined by USDA to be any place with annual sales of agricultural products of $1,000 or more.) Of these farms, the 22,300 that reported planting some wheat were used in the simulation study described in Section 5. The 1987 Small Grain Data Base contains production information for 5,802 farms which reported planting small grain crops. Of these, the 1,707 that reported planting some wheat were used in our study.

Records on the Planted Acres Data Base are a composite of Kansas farm data from a number of sources collected at a number of times. First a list of names and addresses of farms is created using data collected by county appraisers. This data may be replaced and/or corrected using data from the Quarterly Agricultural Surveys and from Monthly Farm Reports. The Quarterly Agricultural Surveys use stratified systematic samples of approximately 2,600 farms. The response rate is approximately 80%. The Monthly Farm Report is completed by about 3,000 farmers who have agreed to file the reports. The same farmer may complete monthly reports for many years. The most recent data for each item appears in the Planted Acres Data Base and the record for any one farm in any year may contain information from a number of sources.

The 1987 Small Grain Data Base contains information on acres planted, acres harvested, and bushels produced for farms responding to the Quarterly Agricultural Surveys and the Kansas Small Grain Survey. About 6,000 surveys were mailed to a random sample of farms for the 1987 Kansas Small Grain Survey; about 50% of the surveys were completed and returned.

In addition to the potential problem with nonresponse bias in the Small Grain Data Base, there is typically a problem with response bias. The production reported by farmers is often lower than the actual production. The non-standard sample, nonresponse bias, and response bias lead us to develop the county estimation procedure described in the following sections.

3. REGRESSION MODELING

We propose the development of a regression model for use in producing county estimates. The calculations for fitting a multiple regression model can be performed using a number of statistical packages available for personal computers. In addition, our proposed estimator allows for the fact that we do not have a probability sample of farms and will produce county estimates that sum to the desired state total.

The steps in our procedure are as follows:

1) Use multiple regression to model the relationship between farm production and some predictor variables using the non-probability sample of farms.
2) Assume that the regression relationship holds for the entire population of farms in the state, and estimate farm production for all farms in each county.
3) Adjust the estimates of farm production to sum to the USDA state total.

To describe the regression model we need the following notation. For $i = 1, 2, \ldots, I$ ($I = 105$ counties in Kansas) and $j = 1, 2, \ldots, n_i$ let

n_i = number of farms from i^{th} county in sample;

$n = \sum_{i=1}^{I} n_i$ = total sample size;

N_i = total number of farms from i^{th} county in population;

$N = \sum_{i=1}^{I} N_i$ = total number of farms in population;

Y_{ij} = wheat production of jth farm in i^{th} county (in bushels);

$X_{ij} = (1\ X_{ij1}\ X_{ij2} \ldots X_{ijp})$ = vector of p predictors for j^{th} farm in i^{th} county.

It is important, as we will see later, to choose predictor variables for which county totals are known or for which very accurate estimates of the county totals are available. The predictor variables must also include information related to the probability that a farm is included in the sample, such as a measure of the size of a farm. This will allow us to use the regression model to adjust for the fact that the sample is not a probability sample.

We consider regression models of the form

$$Y_{ij} = f(X_{ij} \mid \beta) + \epsilon_{ij},$$

where $\beta = (\beta_0\ \beta_1\ \beta_2 \ldots \beta_p)$ is a vector of parameters and ϵ_{ij} is a random error term with variance σ^2. Let the fitted values, which will be obtained using data from the Small Grain Data Base, be denoted by

$$\hat{Y}_{ij} = f(X_{ij} \mid \hat{\beta}).$$

Then the county total for the ith county may be estimated as follows:

$$\hat{Y}_{i+} = \sum_{j=1}^{N_i} \hat{Y}_{ij} = \sum_{j=1}^{N_i} f(X_{ij} \mid \hat{\beta}),$$

where a "+" in a subscript indicates summation over the corresponding subscript.

For a general form of $f(X_{ij} \mid \beta)$, it would be necessary to know the value of X_{ij} for all farms in the ith county. It is, of course, not possible to have such extensive information. If, however, $f(X_{ij} \mid \beta)$ is a linear function, then we only need to know county totals of the predictor variables. This is the case since, for a linear regression equation,

$$\hat{Y}_{i+} = \sum_{j=1}^{N_i} \hat{Y}_{ij} = \sum_{j=1}^{N_i} [\hat{\beta}_0 + \hat{\beta}_1 X_{ij1} + \hat{\beta}_2 X_{ij2} + \ldots + \hat{\beta}_p X_{ijp}]$$

$$= \hat{\beta}_0 N_i + \hat{\beta}_1 X_{i+1} + \hat{\beta}_2 X_{i+2} + \ldots + \hat{\beta}_p X_{i+p},$$

where X_{i+k} is the total of the kth predictor for the ith county.

The $\hat{Y}_{i+}$ will be reasonable county estimates if the regression model describes the relationship between the predictor variables and production for all farms in each county as well as for the farms in the data base. These county estimates, however, will not necessarily sum to the USDA state total for production. Methods for resolving this problem will be considered in Section 4.

In addition to providing county estimates of farm production, the linear regression model proposed above also permits us to obtain variance estimates. This is easiest to see if we write the county estimates in terms of matrices. Let

X = $n \times (p + 1)$ matrix of actual data with rows being the X_{ij} defined above;

Z = (unknown) $N \times (p + 1)$ matrix of predictor variables for all farms in the state;

$\hat{Y}$ = (unknown) $N \times 1$ vector of estimates of wheat production for all N farms in state;

B_i = $N \times 1$ column vector with elements b_{ij}

$$\text{where } b_{ij} = \begin{cases} 1 \text{ if the } j^{\text{th}} \text{ farm is in the } i^{\text{th}} \text{ county} \\ 0 \text{ otherwise} \end{cases}$$

$A = [B_1\, B_2\, B_3 \ldots B_I]_{N \times I}$.

The estimation procedure described above does not provide $\hat{Y}$ but instead provides a vector of county estimates $\hat{Y}_{i+} = A^T\hat{Y}$, where "$T$" indicates the transpose of a matrix.

The variance for the county estimates is thus

$$\text{Var}(\hat{Y}_{i+}) = \text{Var}(A^T\hat{Y}) = A^T\text{Var}(\hat{Y})A = A^T\text{Var}(Z\hat{\beta})A = \sigma^2 A^T Z(X^T X)^{-1} Z^T A.$$

Although Z itself is unknown, the product $A^T Z$ is a known matrix containing only the numbers of farms in a county, N_i, and the county totals, X_{i+k}, for the predictor variables. Thus, if we use the regression mean square error (mse) as an estimate of σ^2, we may obtain estimates of the variances of the county estimates. Variance estimates for county estimates have not previously been available.

The estimator based on a regression model as described in this section meets the requirements for a computationally simple estimator from a non-probability sample. In the following section we consider methods to adjust the estimates to sum to the USDA state totals for farm production.

4. SCALING ESTIMATES TO SUM TO STATE TOTAL

Let Y be the USDA's estimated total wheat production for Kansas. In general, $\sum_{i=1}^{I} \hat{Y}_{i+} \neq Y$. Thus, we define new estimates

$$\tilde{Y}_{i+} = c_i \hat{Y}_{i+},$$

where the c_i are constants such that $\sum_{i=1}^{I} \tilde{Y}_{i+} = \sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$. An important question is how to choose the c_i. Current methods used for county estimation take $c_i = c$ (at the district level) and thus adjust all estimates by a common proportion. Instead, one could choose the c_i to minimize the sum of the squared differences or relative differences between the $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$. Values of c_i and $\tilde{Y}_{i+}$ for three criterion for choosing c_i are given below.

1) Choose $c_i = c$

If c_i is taken to be a constant, then it is easy to show that

$$c_i = c = Y \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}$$

and

$$\tilde{Y}_{i+} = Y\left(\hat{Y}_{i+} \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}\right).$$

2) Choose c_i to minimize the sum of squared differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$

To choose c_i to minimize the sum of the squared differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$ subject to $\sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$, we must minimize $\sum_{i=1}^{I} (\tilde{Y}_{i+} - \hat{Y}_{i+})^2 = \sum_{i=1}^{I} (c_i \hat{Y}_{i+} - \hat{Y}_{i+})^2$ with respect to c_i using a Lagrange multiplier to impose the desired constraint. Doing this, we find that

$$c_i = 1 + \left[\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \hat{Y}_{i+}^2 \sum_{i=1}^{I} (1/\hat{Y}_{i+})\right]$$

and

$$\tilde{Y}_{i+} = \hat{Y}_{i+} + \left[\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \hat{Y}_{i+} \sum_{i=1}^{I} (1/\hat{Y}_{i+})\right].$$

Note that the scaled estimates, $\tilde{Y}_{i+}$, are obtained by adjusting the original estimates, $\hat{Y}_{i+}$, by adding a factor which is a proportion of the difference between the USDA state total and the sum of the original county estimates. The proportion is based on the harmonic mean of

the original estimates. Although some of these scaled estimates could be negative in theory, this is not considered likely in practice because farmers often underreport the amount of production on their farms. If the total of the original estimates exceeds the USDA state total, then scaled estimates corresponding to counties with small original estimates may be negative.

3) Choose c_i to minimize the sum of squared relative differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$

To choose c_i to minimize the sum of the squared relative differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$ subject to $\sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$, we must minimize $\sum_{i=1}^{I} [(\tilde{Y}_{i+} - \hat{Y}_{i+})/\hat{Y}_{i+}]^2 = \sum_{i=1}^{I} (c_i - 1)^2$ with respect to c_i using a Lagrange multiplier to impose the desired constraint. Doing this, we find that

$$c_i = 1 + \left[\hat{Y}_{i+}\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}^2\right]$$

and

$$\tilde{Y}_{i+} = \hat{Y}_{i+} + \left[\hat{Y}_{i+}^2\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}^2\right].$$

The scaled estimates, $\tilde{Y}_{i+}$, are again obtained by adjusting the original estimates, $\hat{Y}_{i+}$, by adding a factor which is a proportion of the difference between the USDA state total and the sum of the original county estimates. The proportion here is based on the squared values of the original estimates. As in method 2, these scaled estimates may be negative, although it is unlikely in practice.

Note that we have chosen to consider the difference $\tilde{Y}_{i+} - \hat{Y}_{i+}$ relative to $\hat{Y}_{i+}$ rather than to $\tilde{Y}_{i+}$. This choice was made because in the later case the estimator, $\tilde{Y}_{i+}$, does not have a closed-form solution. Thus, to meet the goal of developing computationally simple estimators, we chose to consider the difference $\tilde{Y}_{i+} - \hat{Y}_{i+}$ relative to $\hat{Y}_{i+}$.

In the following section we will consider the effects of these three scaling methods on the county estimates of wheat production.

5. COMPARISON OF ESTIMATES OF WHEAT PRODUCTION

We used a linear regression model, as described in Section 3, to model the relationship between wheat production (measured in total bushels produced) and some predictor variables for farms in the 1987 Small Grain Data Base. The possible predictor variables that we considered included: acres planted in wheat, acres of wheat harvested, a prediction of wheat production based on the 1986 county estimates, acres of irrigated wheat, acres of non-irrigated wheat, indicators of the district in which the farm is located, indicators of region of the state (east, central, west), and interaction terms.

The most important predictor variables for the regression model were acres planted in wheat and some indicator of the location of the farm within the state. The variable based on the previous year's county estimates did not seem to be a useful predictor for the amount of wheat produced on a farm in the current year. Because other possible predictor variables, such as irrigated acres, are not known as accurately at the county level, we decided that acres planted would be the single continuous predictor variable included in the model. Not all district indicators were needed in the regression model; that is, some districts were similar and could

Table 1

Regression Models Fitted to Actual Data

	Fitted Models	R^2	$\sqrt{\text{mse}}$
Model 1	Bushels $= -811 + 32(\text{Pla}) + 3{,}248I_1 + 3{,}088I_2 + 2{,}190I_3 + 2{,}526I_4 + 1{,}241I_5 - 562I_6 + 1{,}047I_7 + 399I_8$	85	5,945
Model 2	Bushels $= -281 + 28(\text{Pla}) + 138I_1 + 1{,}861I_2 + 2{,}328I_3 + 329I_4 - 359I_5 - 334I_6 - 42I_7 + 500I_8 + 11(\text{Pla})I_1 + 5(\text{Pla})I_2 + 3(\text{Pla})I_3 + 11(\text{Pla})I_4 + 9(\text{Pla})I_5 - 0.2(\text{Pla})I_6 + 15(\text{Pla})I_7 - 7(\text{Pla})I_8$	86	5,818

Note: Pla is planted acres, I_i is the indicator variable for the ith district.

have been grouped together. We decided, however, to include all district indicators in the model since groupings of districts might change from year to year or might be different for crops other than wheat.

We chose to focus our study on two possible regression models: Model 1 contained acres planted in wheat and the district indicators while Model 2 contained these same variables and the interaction terms involving acres planted and the indicator variables. The models and measures of their fits are shown in Table 1. Although the root mean squared errors did not differ considerably for the two models, we felt that the difference might be magnified when the models were used to estimate farm production for the entire state. Thus, in the following, we obtain and compare estimates from both models.

To verify that these regression models are not simply a result of some unusual feature in the 1987 Kansas Small Grain Data Base, we used the same set of possible predictor variables and searched for reasonable regression models using the 1988 data. The fits of Models 1 and 2 to the 1988 data are similar to the 1987 fits and no other model appeared to be superior for fitting the 1988 data. The estimates for the parameter corresponding to acres of wheat planted were fairly similar in both 1987 and 1988, but the parameters corresponding to the indicator variables for districts showed considerable change. We believe that the indicator variables for districts are reflecting the effects of weather and different farming practices in different parts of the state. For example, irrigation is more commonly used in western and central Kansas than in eastern Kansas. Although farming practices are not likely to change dramatically from one year to the next, weather conditions may be quite different. Thus, it seems reasonable that the contribution of the district variable in predicting wheat production could change considerably from year to year.

Both models were used to obtain county estimates for all 105 counties in Kansas. In Table 2, the unscaled estimates and their standard errors under both Models 1 and 2 are given for nine counties, one county chosen at random from within each district so that the nine counties are spread over the entire state. An inspection of Table 2 suggests that the estimated standard error for Shawnee county is an anomaly. The variance of a county estimate depends on the number of farms in the county, the total acres planted in wheat in the county, and the number of farms sampled from the district in which the county lies. District 8, in which Shawnee county is located, had relatively few farms in the Small Grain Data Base. The county has a moderate number of farms growing wheat but these farms are small in terms of acres planted. These three factors together result in the rather large standard error for the estimates from Shawnee county.

Table 2
Regression Model Estimates for Nine Counties in Kansas

District	County	Estimated Bushels of Wheat Produced (in thousands of bushels)	
		Model 1 (no interaction terms)	Model 2 (with interaction terms)
1	Decatur	4,944 (180)	4,778 (179)
2	Trego	4,378 (174)	4,229 (188)
3	Hodgeman	4,808 (123)	4,908 (125)
4	Jewell	5,555 (275)	5,550 (269)
5	Marion	5,144 (313)	4,931 (315)
6	Comanche	2,615 (59)	2,480 (63)
7	Leavenworth	231 (53)	262 (61)
8	Shawnee	232 (106)	226 (104)
9	Butler	2,374 (331)	2,272 (338)

Note: Standard errors are given in parentheses below each estimate.

The estimates shown in Table 2 are reasonably similar to the published county estimates (Kansas Agricultural Statistics 1988). While it is encouraging that our estimates are not wildly different from those published by Kansas, there is no theoretical basis for using the Kansas estimates as a standard. Thus, we carried out a simulation study to help us evaluate our estimators. This study is described in the following section.

6. SIMULATION STUDY

6.1 The Estimators to be Compared

In the simulation study, we compared the estimates from our two regression models with those from two standard small area estimators: the synthetic and direct estimators. (See, for example, Section 2 of Särndal and Hidiroglou (1989) for a discussion of standard small area estimators, including the synthetic and direct estimators.) The synthetic estimates are obtained by allocating the state total for wheat production to the counties according to the proportion of total acres planted in wheat within each county. The direct estimates are obtained using only the sampled farms in a county to estimate wheat production for that county.

We expect the synthetic estimates to have a large amount of bias because counties in different parts of the state have different farming practices and different weather conditions, while the synthetic estimator treats each county as if it were representative of the entire state. The synthetic estimates, however, will have relatively small variances because they are obtained using all the data from the entire state.

Since the direct estimate for a county is based only on the sample data within that county, it will have a relatively large variance but it should have smaller bias than the synthetic estimate. At least one farm from a county must appear in the sample to make it possible to obtain an estimate for that county, and at least two farms are needed in the sample to make variance estimation possible. In the 1987 Kansas Small Grain Data Base, three counties had no wheat farms in the sample and three additional counties had only a single farm in the sample. Although we are comparing our regression model estimates to the synthetic and direct estimates, it should be noted that the latter two estimators require that the data be from a probability sample. This requirement is not met by the Kansas data.

Table 3

Numbers of Farms and Production Levels by District and Planted Acres

District		Planted Acres in Farm				
		0-99	100-249	250-499	500-999	≥1,000
1	M_i*	354	638	531	302	85
	m_i*	27	45	51	40	9
	bu/pa*	34.68	37.18	37.76	39.21	38.68
2	M_i	266	550	572	377	161
	m_i	27	49	47	55	33
	bu/pa	35.92	33.62	36.78	39.09	34.85
3	M_i	264	549	610	537	264
	m_i	31	80	76	98	61
	bu/pa	26.93	32.84	35.03	36.79	33.13
4	M_i	956	939	626	271	50
	m_i	62	37	23	21	7
	bu/pa	36.81	36.91	39.70	39.87	39.41
5	M_i	1,236	1,529	912	350	54
	m_i	92	93	51	26	3
	bu/pa	31.79	32.25	31.69	36.85	33.65
6	M_i	1,181	1,427	1,160	793	249
	m_i	96	96	81	55	20
	bu/pa	26.24	26.88	28.78	27.87	26.72
7	M_i	957	242	67	9	3
	m_i	62	5	2	0	0
	bu/pa	33.87	40.81**	40.81**	40.81**	40.81**
8	M_i	1,126	251	52	9	1
	m_i	56	11	2	0	0
	bu/pa	26.02	11.48**	11.48**	11.48**	11.48**
9	M_i	1,122	431	166	59	12
	m_i	47	19	7	3	1
	bu/pa	23.57	23.87	27.63**	27.63**	27.63**

* M_i is the number of farms on the Planted Acres Data Base, m_i is the number of farms in the Production Data Base, and bu/pa is the ratio of bushels produced to acres planted.

** Cells of this district were grouped to obtain bu/pa values.

6.2 The Simulated Population and Samples

We first simulated a population of wheat farms by generating production values for all 22,300 farms reporting acres planted in wheat on the Planted Acres Data Base. Because production rates appear to vary by district and size of farm (see Table 3), we generated bushels-per-planted-acres (bu/pa) from 37 different distributions. These distributions were based on the bu/pa data from the Small Grain Data Base. (Notice that in the eastern districts of Kansas, districts 7, 8 and 9, there were few or no sampled farms in several size-of-farm classifications. Those classifications were grouped as indicated in Table 3 for the purpose of simulating bu/pa values.) Histograms of the observed bu/pa from the Small Grain Data Base were generated by district and five sizes of farm: 0-99, 100-249, 250-499, 500-999, and 1000 or more acres of wheat planted. Since these histograms generally appeared mound-shaped, we chose to use normal distributions to model the distributions of the bu/pa. The means and variances of the normal distributions were taken to be the sample means and variances of bu/pa from the wheat farms in the Small Grain Data Base within the 37 district by size-of-farm classifications.

After the bu/pa values were generated from the appropriate normal distributions for each farm, the bushels of wheat produced were obtained by multiplying the simulated bu/pa by the reported acres planted in wheat for each farm. Ten samples were generated from the resulting simulated population. Since there was no sampling design to follow in creating these samples, we sampled each farm within the district by size-of-farm classifications with probabilities equal to the observed frequencies with which farms on the Planted Acres Data Base appeared in the Small Grain Data Base. That is, farms within classification C, say, were chosen to be in the sample with probability equal to

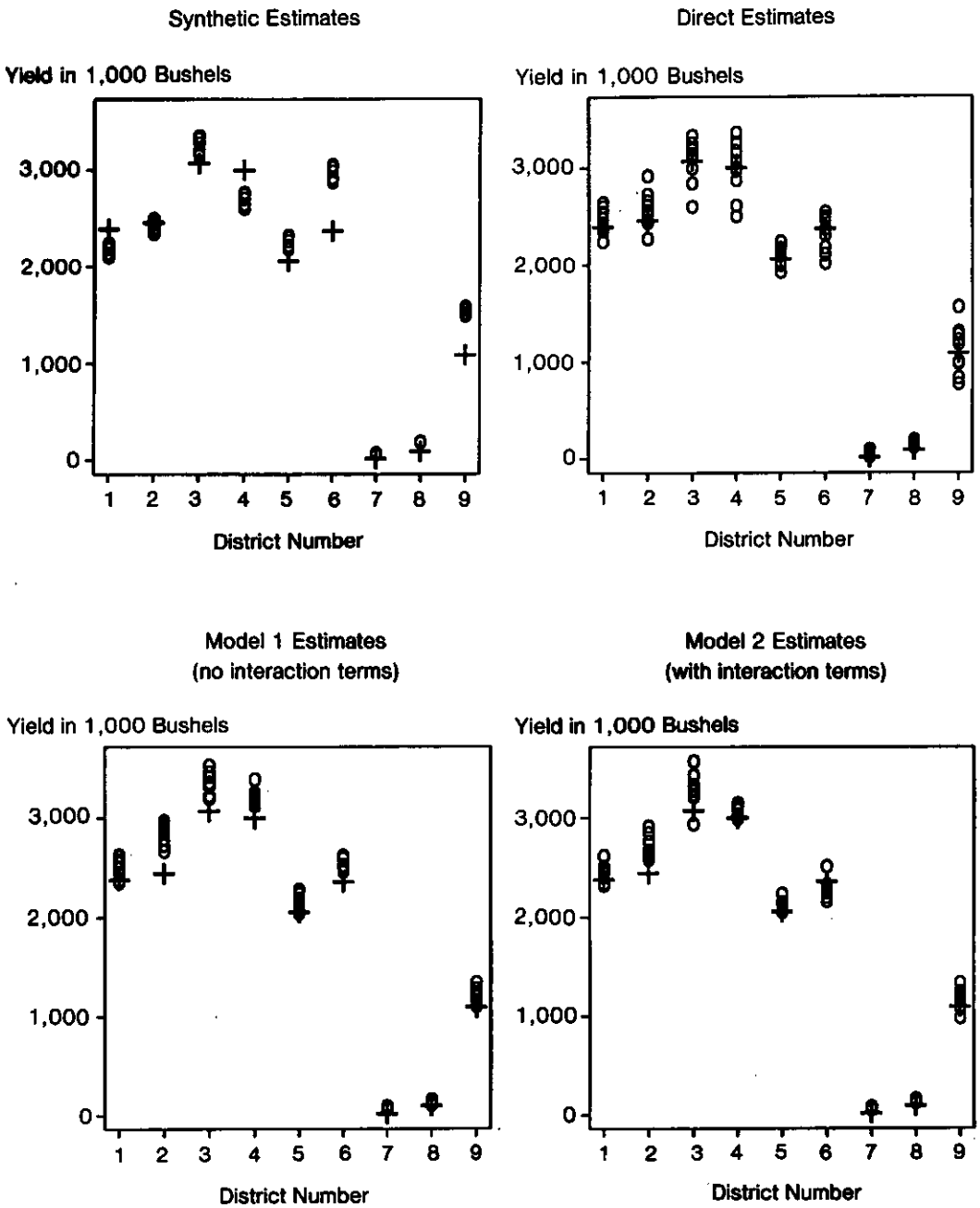
$$\frac{\text{Number of farms in classification } C \text{ in Small Grain Data Base}}{\text{Number of farms in classification } C \text{ in Planted Acres Data Base}}.$$

Our goal in using such a sampling scheme was to make the simulated samples as similar as possible to actual samples even though we do not know what the selection probabilities for the actual samples were.

6.3 Comparison of the Four County Estimators

We used the four county estimators (the two regression, the synthetic, and the direct) to obtain wheat production estimates for all 105 counties from each of the ten simulated samples. The resulting estimates were then compared to the "true" production values obtained for each county from the simulated population. This comparison allows us to evaluate the amounts of bias and variability in the estimates for each county. Figure 1 presents the values of all four estimates from each of the ten samples along with the true production values for the nine-randomly chosen counties, one from each district, which were previously mentioned in Section 5.

As expected, the synthetic estimates exhibit considerable bias. Indeed, only in district 2 does the range of estimates include the true population value. The ranges of the direct estimates are all larger than those of the synthetic estimates but those ranges do include the population values. The ranges of estimates from the regression models appear to be less than those of the direct estimates. For about half of the counties pictured in Figure 1, the estimates from Model 1 appear to exhibit some bias. The estimates under Model 2 seem to exhibit less bias. On the basis of this comparison of estimators we prefer Model 2, the regression model with the interaction terms.



Note: Estimates are for one county chosen at random from within each district.
o = estimate from one of the ten simulated samples,
+ = true value from simulated population.

Figure 1. Comparison of Estimators for Nine Counties

6.4 Comparison of the Scaling Methods

The same four sets of estimates for all counties from the ten sets of simulated samples were next scaled to agree with the state total from the simulated population using the three scaling methods described in Section 4. The resulting scaled estimates were compared to the true county production values for the simulated population. The comparison was made using the mean of the absolute value of relative error which is defined as follows:

$$(1/I) \sum_{i=1}^{I} \left| (\tilde{Y}_{i+} - Y_{i+})/Y_{i+} \right|.$$

Figure 2 shows the values for all ten samples of the mean over the 105 counties of absolute relative error. This error is given for all four estimators under no scaling and under each of the three methods of scaling.

From Figure 2A, we see that the scaling method which minimizes the sum of squared differences produces very poor final estimates; the average of absolute relative differences between the final estimates and the county production values for the simulated population is quite large compared to that of the other scaling methods. This large error results from the fact that the total wheat production in one county may be quite different from that in another county. Since the scaling procedure minimizes the squared differences between the original and the final estimates, a county with a very small original estimate may have a final estimate that is changed considerably relative to the original estimate. These large changes in estimates do not seem warranted; hence we drop this method of scaling from consideration.

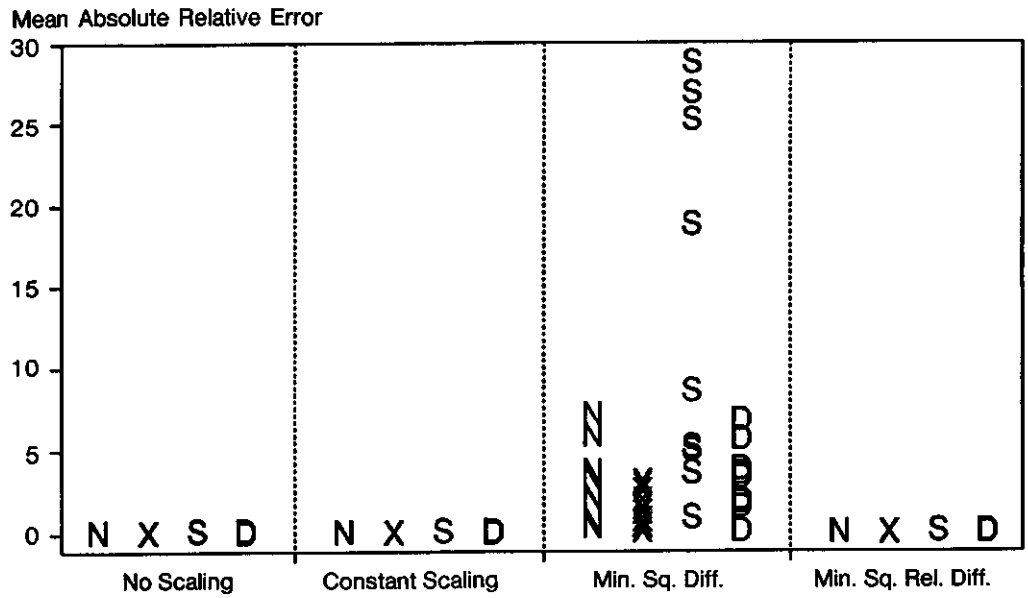
Figure 2B, a refinement of Figure 2A, provides a more detailed comparison of the four estimators under no scaling and under the two remaining scaling methods. We see from this figure that the error is generally smallest for the regression model with the interaction terms. This supports our choice of Model 2 in the previous subsection. In addition, Figure 2B suggests that there is little difference between the original unscaled estimates and the final estimates under either scaling method. In fact, the total of the original county estimates is not far from the simulated population total. Thus, the scaling constants, c_i, are all quite close to one. Since the two methods of scaling produce similar estimates, there is no reason to use the more difficult scaling method; the constant scaling method may be used.

7. CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

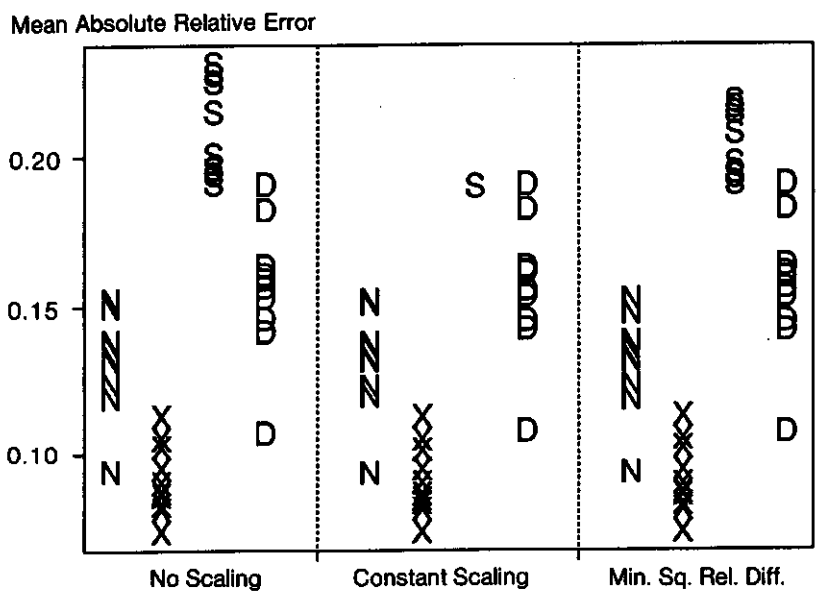
We have shown that a regression model may be used to obtain reasonable county estimates of wheat production. The model we selected used acres planted, district indicators, and interaction terms as predictor variables. The regression model does not require a probability sample of farms and it does permit the estimation of variances of the county estimates. The estimates based on the regression model may be scaled to agree with state total production using a constant scaling factor since the alternative scaling method did not produce markedly different county estimates.

Many areas for future research in county estimation of farm production remain. For example, the county estimates from our simulation study suggested that the inclusion or exclusion of large farms (1,000 or more acres of wheat planted) from the sample for a district could have a large effect on the estimates for counties in that district. This was particularly true for

A.



B.



Note: Data are from the simulated samples.
N = Model 1 Estimator (no interaction terms),
X = Model 2 Estimator (with interaction terms),
S = Synthetic Estimator,
D = Direct Estimator.

Figure 2. Comparison of Estimators and Scaling Methods

districts which had few of these larger farms. Since large farms most likely account for a sizable proportion of farm production, it might be worthwhile to handle large farms separately in a county estimation procedure. States might also consider altering their sampling plans so that the largest farms are included in the samples with certainty.

Additional work is needed to determine whether a regression model similar to that developed for wheat is appropriate for other crops as well. In particular, it would be useful to discover if such models can be used for rare crops where there is much less available data. We should also note that the similarity in the state total and the total of the county estimates, which was observed for the actual data as well as for the simulated samples, may be characteristic of wheat production but not of all crop production. Future research should consider whether other crops require a scaling method other than constant scaling.

We chose to begin our research on the county estimation problem by studying methods of estimating production. An additional problem for future research is the estimation of total acreage planted for various crops. In this research we used 1987 agricultural census data to provide the needed information on numbers of farms and acres planted in wheat within each county. The agricultural census, however, is taken only every five years. In the intermediate years, changes in numbers of farms and acres planted must be estimated from sample data. We expect such changes in census values to be small for major crops like wheat in Kansas, but we anticipate greater difficulty estimating these quantities for less common crops.

Finally, the requirement for a computationally simple estimator, which led us to propose an estimator based on a regression model, may no longer be necessary as state agricultural offices are being linked to a large, national computer system. Thus, in our future research on county estimates of farm production, we plan to consider more computationally intensive small-area estimators.

ACKNOWLEDGEMENTS

This research was supported in part by the United States Department of Agriculture under Cooperative Agreement No. 58-3AEU-9-80040. The authors take sole responsibility for the contents of this paper. The authors wish to thank Gary Keough and Leland Brown at USDA and Ronald Sadler, Melvin Perrott, Eldon Thiessen, and M. E. Johnson at the Kansas Department of Agriculture for their help on this project. We also thank the referee and associate editor for their helpful comments on an earlier version of this paper.

REFERENCES

BASS, J., GUINN, B., KLUGH, B., RUCKMAN, C., THORSON, J., and WALDROP, J. (1989). Report of the Task Group for Review and Recommendations on County Estimates. USDA National Agricultural Statistics Service, Washington, D.C.

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*, 74, 269-277.

KANSAS AGRICULTURAL STATISTICS. (1988). Kansas Farm Facts, prepared by the Statistical Division of the Kansas Department of Agriculture in cooperation with the National Agricultural Statistics Service of the U. S. Department of Agriculture, Topeka, Kansas.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.E., and SINGH M.P. (Eds.) (1987). *Small Area Statistics*. New York: John Wiley & Sons.

SÄRNDAL, C.E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

Survey Methodology, December 1991
Vol. 17, No. 2, pp. 227–240
Statistics Canada

Canada's General Social Survey: Five Years of Experience

D.A. NORRIS and D.G. PATON[1]

ABSTRACT

The Canadian General Social Survey is an annual survey that aims to provide data on the demographic and social characteristics of Canadians. This paper provides an overview of the program, based on the experience of the first five surveys. The objectives of the program, the methodology used, the themes and issues addressed, the program outputs and the plans for the future are all discussed.

KEY WORDS: Social surveys; Telephone surveys; Random digit dialing; Time use surveys; Health surveys.

1. INTRODUCTION

Statistics Canada's social statistics program is concerned with providing information on the demographic and social characteristics and conditions of Canadians. The program's output sustains the development of policy on many critical social issues.

The Census of Population, held every five years, is the cornerstone of the social statistics program, providing benchmark information on the demographic, social, and economic conditions of the population and the basis for future sample surveys of the population. In addition to the Census, activities include on-going surveys and other statistical programs, many based on administrative data sources, in the areas of Health, Education, Culture, Justice, Public Finance, Employment and Unemployment, Income and Expenditures and Demography.

While household surveys have long been an important part of the social statistics program, the regular survey program has historically been directed mainly at labour market and income related issues and there have been no regular ongoing surveys in areas such as health, education, justice or culture. In order to partially fill this data gap Statistics Canada established in 1985 a General Social Survey (GSS) program.

The purpose of this paper is to outline the nature and scope of the GSS program and to describe its evolution over the past five years. Included is a description of the methodology and the content of the five surveys that make up the program. Finally there is a brief discussion of some future directions for the program.

2. GSS PROGRAM OBJECTIVES AND STRUCTURE

The period 1930-1980 witnessed a rapid rise in the number and size of social programs in Canada. Whereas in the early 1930's all government expenditures on social programs accounted for about 10% of GNP, by the early 1980's this expenditure had climbed to about 30%. Along with this rise came an increased demand for and use of data and information to monitor and analyze social trends, and over the years, Statistics Canada expanded its social statistics program to meet growing requirements. Nonetheless, the more extensive use of available data in recent

[1] D.A. Norris and D.G. Paton, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

years revealed major areas of weakness where relevant data were too narrow and restrictive for the effective planning of policy programs, products, and services, or for determining the allocation of resources between competing alternatives.

In the early 1980's, a shortcoming of the social statistics program was that aside from the labour market and income areas, most other social data were derived from administrative records or surveys of institutions. These data sources provided only limited information on the population who came in contact with social institutions and no data on the need for, or impact of, social programs on the general population. Such data can only be obtained through a general population survey.

While a case could have been made for regular, frequent and large scale surveys in a variety of fields (*eg.* health, education, criminal victimization), resources to mount such a large scale program were not available. Instead Statistics Canada initiated a much more modest annual General Social Survey which over five years would cover major topics of importance and which would in the long term serve as a vehicle for monitoring social change. In the short term it could also serve as a vehicle to collect limited data on topics of current social policy interest. The total annual budget for the GSS was originally set at about one million dollars (CAN) and the program was funded by an internal reallocation of Statistics Canada resources derived from efficiency gains in the Labour Force Survey program.

The objectives of the GSS program are two-fold:

- To gather data with a degree of regularity on a broad range of social trends in order to monitor temporal changes in the living conditions and well-being of Canadians; and,
- To provide information on specific social policy issues of current or emerging interest.

To meet these objectives, the GSS program was established with an annual survey cycle. In order to cover the wide range of social issues for which data are required, the GSS program consists of five survey cycles, each covering a different core topic. The collection of data for these topics is thus repeated every five years. The core topics identified for the five cycles are:

1. Health
2. Time Use
3. Personal Risk (accidents and criminal victimizations)
4. Education and Work
5. Family and Friends.

An additional objective in planning content was to include questions that would be useful in deriving indicators of the quality of life, for example, measures of life satisfaction, attitudes, perceptions, or beliefs.

The content of a GSS cycle consists of the following three modules:

- Core content, which is repeated every five years in order to gather information to monitor trends in living conditions and well-being.
- Focus content, which varies from survey to survey and is aimed at the second survey objective of providing information on specific policy issues of particular interest to certain federal departments or policy groups.
- Classification content, which is collected in every cycle and consists of a set of basic demographic and socio-economic variables that enable the delineation of various population groups to facilitate the analysis of core and focus content.

While core and classification content are funded by Statistics Canada, costs associated with focus content are recovered from sponsors.

The target population for the GSS consists of the non-institutionalized population aged 15 and over living in the ten provinces. It was decided that the Labour Force Survey would not

be used as a vehicle for the GSS in order to avoid placing an excessive response burden on LFS respondents and to allow the GSS to use sampling and collection methodologies and sample allocations that differ from those of the LFS. The target sample size for each cycle is 10,000 individuals which was arrived at as a compromise between the competing demands of precision of estimates,budget and length of interview. However, there exists within the GSS program the potential for sponsors to expand the sample for a target population or geographic area. The first survey on health was conducted in late 1985 and the other surveys followed at approximately one year intervals. The fifth cycle on the family was conducted in early 1990 and data collection for Cycle 6 began in January 1991.

The themes and research issues which are covered by each of the surveys are discussed in more detail below. However, before considering these the methodology of the survey is examined in more detail.

3. METHODOLOGY

3.1 Requirements and Constraints

The following are the principal methodological requirements of the GSS: i) it should allow for extensive analysis of the adult Canadian population at a national level and somewhat less detailed analysis at the regional level (this has implications on both the sample size and on the amount of data collected from each respondent); ii) it should have an acceptable cost; iii) it should have a design appropriate for a multipurpose survey; and iv) it should provide public use microdata sets that could be used for analysis by researchers outside Statistics Canada without too much difficulty.

These requirements all interact with the choice of data collection mode, sample design and sample size, but the last two were principally responsible for the choice of sample design, while the sample design and the first two requirements were largely responsible for the choice of data collection mode and sample size.

The last requirement suggests that the sample design be simple as the design information that would be necessary to analyse complex survey data cannot generally be made available on public use files. Requirement iii) suggests that the design not be highly optimized for specific variables.

3.2 Mode of Data Collection

The choice of data collection mode involved balancing a number of competing factors: cost per interview, length of interview, response rate, accuracy of information collected and sample size. The level of detail required in the data collected meant that interviews were expected to last 20 to 30 minutes per respondent. To reduce response burden at the household level and to avoid a cluster effect at the household level it was decided that only one person per household would be interviewed. The principal data collection methods considered for the survey were: self-completed mail-back questionnaire; personal interview; and telephone interview. The high non-response rates experienced with self-completed mail-back questionnaires were felt to be unacceptable (in terms of potential biases) given the heterogeneity of the target population. Personal interviews were felt to offer a number of advantages that would improve the quality of the data collected such as low non-response rates and low item non-response rates, but suffered from the disadvantage of high cost. In addition, many designs used to reduce the cost of personal interviewing have multiple stages of selection and are highly optimized for a few variables. (To not use a design and frame currently used for personal interviewing would have

been unreasonably costly.) These complicated designs make analysis of the resulting datasets difficult and the optimization leads to high design effects for some variables. These high design effects make such designs less appropriate for multipurpose surveys like the GSS. Experience with telephone surveys at Statistics Canada indicated that fairly high response rates could be achieved at reasonable cost. In addition, random digit dialing (RDD) sampling methods allow the efficient selection of samples that are simple random samples or nearly simple random samples.

For these reasons, the GSS has used telephone sampling (RDD) methods and telephone interviewing for most of its sample in all cycles conducted to date. When there has been a need to focus on special target groups its main sample has been supplemented with individuals selected from list frames. In Cycle 1 it was felt that face to face interviews should be used for many of the interviews with elderly respondents.

3.3 Target Population

The target population of the GSS is all persons over the age of 14 permanently living in Canada, with the following two exclusions: i) residents of the Yukon and Northwest Territories, and ii) residents of institutions. This target population is different from that of the Labour Force Survey, which in addition excludes residents of Indian Reserves and full-time members of the Canadian Armed Forces.

3.4 Sampled Population

The sampling methods used for the GSS exclude some members of the target population from the sample. During weighting, these exclusions are implicitly assumed to be similar to the sampled population (missing at random) and the final weights produce estimates for the target population.

When telephone interviewing methods are used, those persons living in households without telephones are excluded from the sample. This affects less than 2% of Canadian households covered by the Labour Force Survey (Statistics Canada 1989, 1990b). This high rate of telephone penetration is not uniform across age and income groups and varies from province to province: 95.4% of households in the province of Prince Edward Island have telephones while 99.2% of those in the province of Ontario do; 99.1% of households with incomes between 20 and 25 thousand dollars have telephones while only 93.9% of those with incomes less than 10 thousand do. Some subpopulations have much lower rates of telephone ownership than the average; for instance, only 86.7% of low income persons under 65 living alone have telephones.

The GSS does not in general accept proxy responses and so individuals who cannot use a telephone (those unable to hear or unable to speak) or who cannot be reached by phone during the survey period or who do not speak either English or French are excluded from the responding population. (For the sixth GSS cycle (on the health of Canadians) it was decided to accept proxy responses in those situations where the selected respondent could not complete the interview due to a health problem.)

When supplementary samples are drawn from lists of households interviewed by the Labour Force Survey (as was done for GSS Cycles 1, 5 and 6), residents of Indian Reserves and full-time members of the Canadian Armed Forces are excluded from these samples. These exclusions represent less than 0.5% of the population over the age of 65. (This is the only age group that has been sampled this way.)

3.5 Stratification

The stratification used by the GSS is determined by estimation requirements, operational requirements, restrictions on the definition of strata imposed by RDD sampling, weighting problems specific to RDD sampling and the special needs of sponsors. Since some estimates are required at the provincial level the GSS strata never cross provincial boundaries. For operational reasons a stratum must be interviewed from a single Regional Office, thus strata never cross Regional Office boundaries. The RDD sampling method used requires that strata be defined as aggregations of telephone exchanges. During weighting, accurate estimates of the sizes of strata are needed, thus the strata (defined on the basis of telephone geography) need to correspond closely to aggregations of units for which accurate population data or estimates were available. These accurate data are available in intercensal years at the Census Metropolitan Area (CMA) level.

The basic stratification based on these requirements starts with the provincial boundaries as stratum boundaries. In Cycles 1 to 5, Saskatchewan and Ontario were each covered by two Regional Offices, so they were both divided in two by a stratum boundary. Further, within each of the areas thus obtained, the CMA's formed a stratum and the non-CMA areas another stratum. In addition, the two largest CMA's, Montreal and Toronto, were each separate strata. For Cycles 1 to 5 this gives us a total of 25 strata: one in Prince Edward Island (there is no CMA in PEI), two in each of Newfoundland, Nova Scotia, New Brunswick, Manitoba, Alberta and British Columbia, three in Quebec, four in Saskatchewan and five in Ontario. For Cycle 6 there were 21 strata: one in Prince Edward Island, two in each of Newfoundland, Nova Scotia, New Brunswick, Manitoba, Saskatchewan, Alberta and British Columbia, and three in Quebec and Ontario.

This is the basic stratification used by the GSS, but modifications to this basic common stratification to accommodate the particular needs of the subject matter or of the sponsors are possible and have been implemented. In Cycle 2 the special interest in language use indicated that separate strata with higher sampling fractions should be used in "contact regions" in which there were thought to be large numbers of both anglophones and francophones. In Cycle 5 the interest of a client in producing estimates for certain sub-provincial regions of Ontario led to the definition of a special stratification.

3.6 Allocation

The target sample size of the GSS is 10,000 completed interviews. This sample has been allocated to provinces in proportion to the square roots of their population sizes. The allocation to strata within provinces has been in proportion to their sizes. The square root allocation is a method of increasing the sample sizes for the smaller provinces (when compared with a proportional allocation) without compromising the precision of Canada level estimates as much as an equal allocation. The method of Kish (1976) for arriving at an allocation that explicitly balances the need for provincial and Canada level precision has been investigated, but the resulting allocations yield little improvement in the precision at the Canada level while changing the allocations to some provinces dramatically and in a way felt to be undesirable.

3.7 Telephone Sampling Method

Except for supplemental samples of the population over 65 selected using lists of households interviewed for the Labour Force Survey, the GSS samples have been selected using random digit dialing methods. Two methods of sample selection have been used, the Waksberg (1978) method and the elimination of non-working banks method. Both methods use information obtained from telephone companies to improve the success rate of reaching households. The choice of methods depends on the level of detail of the information available.

Telephone numbers in Canada are ten digit numbers that can be decomposed into a three digit "Area Code", a three digit "Prefix", and two two digit fields, the first of which we refer to as a bank identifier. Thus within each "Area Code-Prefix" (ACP) there are ten thousand possible numbers and within each "Area Code-Prefix-Bank" (ACP-Bank or simply bank) there are one hundred possible numbers. For example, here is a fictitious telephone number and its components:

216-357-4675

216	Area Code
357	Prefix (exchange)
46	Bank Identifier
75	Number
216-357	ACP
216-357-46	ACP-Bank (bank).

When the only information that is available is a list of ACP's, the GSS uses the Waksberg method of generating the sample. In this method, banks are selected with probability proportional to size, where the size measure is the number of residential telephone numbers in the bank. Within each selected bank a simple random sample of residential numbers is selected. When the sample size is the same in each bank, this method yields an equal probability sample of residential telephone numbers. The sample size within banks used by the GSS has been 6. This method has the advantage of improving the success rate of selecting residential numbers with the disadvantage of producing a clustered sample. For instance, in some rural areas of western Canada only approximately 6% of the numbers generated using lists of ACP's are residential, while the success rate during the second stage of seletion is about 50%. The design effects due to clustering are small for many variables, on the order of 1.0-1.3.

When more detailed information is available that allows the creation of a list of banks containing one or more residential numbers ("working banks") the method which we call the elimination of non-working banks method (ENWB) is used. A simple random sample of numbers within the working banks is selected and non-residential numbers are rejected, yielding a simple random sample of residential numbers. Since the first GSS in 1985, sampling has shifted more and more to the ENWB method as more information has become available from the telephone companies. For Cycle 6 (conducted in 1991) the ENWB method was used for the entire sample.

A system of computer programmes for the Regional Offices of Statistics Canada has been written to implement these two sampling schemes and to monitor the progress of the survey. Within a stratum, the entire sample must be generated using the same sampling method.

After a household has been reached by telephone, a list of the names and ages of all household members is collected and, using this list and a set of random numbers printed for each questionnaire, one person 15 years of age or older in the household is selected to be interviewed. This is the method of Kish (1949).

3.8 Special Samples

Sponsors of the GSS have the opportunity to fund additional interviews. These additional samples can be simple increases in the RDD sample size for one or more strata or they can be drawn from other sampling frames.

In Cycles 2 and 5 the RDD samples in strata of special interest to sponsors were increased.

In Cycles 1, 5 and 6 additional samples of special interest groups were used to supplement the RDD sample. In these cases samples of persons aged 65 and over were selected using lists of households that had recently been part of the LFS sample.

3.9 Response Rates

One disadvantage of telephone surveys is that respondents seem to find it easier to refuse to participate in a telephone survey than in a survey with personal interviews. Telephone soliciting is being used regularly by businesses to sell products and services and everyone has to learn to say no over the phone. In addition, new technologies such as answering machines and special features being added to telephone systems are making it possible and easy for people to screen their incoming calls.

Table 1 gives response rates for the first five cycles of the GSS. The categories "Other Household Non-Response" and "Other Respondent Non-Response" include non-interviews due to language problems, illness, death in the family and absence for the survey period; some of these non-responses are undoubtedly refusals in disguise. In all cycles except Cycle 2 interviews were conducted as soon as possible after contacting the households. In Cycle 2 there was a gap of about a month between the initial contact with the households and the interviewing; there is a component of non-response that can be directly attributed to this time lag. From the table it seems that there may be a trend toward lower response rates over the five cycles.

If we consult Table 2, which presents response rates for individual Regional Offices for Cycles 3 to 6, we see that the situation is not so simple, with many offices (Halifax, Montreal, Winnipeg) showing little change in response rate over these cycles. In fact if we exclude the results obtained in Toronto, the response rate declined only slightly between Cycle 3 and Cycle 5. We have observed that more experienced interviewers tend to be more successful at achieving high response rates. The dramatic change in response rates over three cycles experienced by the Toronto office may in large part be due to the difficulty in hiring and retaining staff in a city that at the time had a booming economy. It is also possible that some of the change is due to a change in the population sampled from the Toronto office.

Preliminary results from eight (January to August 1991) months of data collection for Cycle 6 indicate (see Table 2) that it was possible to reverse the trend to lower response rates. There were a number of changes made between Cycles 5 and 6, the most important ones being a change to monthly data collection and the reassignment of the sample from offices not used for data collection in Cycle 6: St. John's' sample was transfered to Halifax, Toronto's to Sturgeon Falls and Edmonton's to Winnipeg.

During data collection for Cycle 3 it was noted by interviewers that an increasing number of calls were answered by answering machines. This raised the concern that respondents might use these machines to screen their calls, resulting in higher non-response rates. We are not able

Table 1

Response and Non-response Rates (%) by Cycle and Type

Result	Cycle				
	1	2	3	4	5
Household Refusal	6.2	6.2	6.0	7.2	10.3
Other Household Non-Response	4.4	6.8	6.6	6.4	7.2
Respondent Refusal	1.3	2.8	1.3	1.7	2.4
Other Respondent Non-Response	4.8	3.5	3.2	3.9	4.3
Special Cycle 2 Non-Response		1.9			
Response	83.4	78.9	82.9	80.7	75.8

Table 2

Response Rates by Cycle and Regional Office
(results for Cycle 6 are preliminary - indicates offices not conducting interviews)

Regional Office	Cycle			
	3	4	5	6
St. John's	84.1	82.8	90.9	-
Halifax	84.7	84.1	85.9	82
Montreal	83.0	79.6	81.2	82
Sturgeon Falls	76.5	81.1	71.5	71
Toronto	87.0	75.4	63.0	-
Winnipeg	84.3	87.0	84.3	89
Edmonton	83.2	79.4	76.8	-
Vancouver	75.3	80.2	79.6	82
Canada	82.9	80.7	75.8	81
Canada (without Toronto)	82.1	81.8	80.1	-

Table 3

Response and Non-response Rates (%) by Type and Contact with Answering Machines

	Did any calls reach an answering machine?			
	Cycle 4		Cycle 5	
	No	Yes	No	Yes
Household Refusal	7.18	8.17	10.34	9.74
Other Household Non-Response	6.45	5.89	7.19	7.41
Respondent Refusal	5.05	4.55	4.39	3.27
Other Respondent Non-Response	1.68	2.01	2.31	3.15
Responses	79.64	79.38	75.76	76.44
Number of Records	10,981 (93.6%)	747 (6.4%)	16,611 (90.6%)	1,715 (9.4%)

Table 4

Response and Non-response Rates (%) by Type and Type of First Contact

	Was the first contact with an answering machine?			
	Cycle 4		Cycle 5	
	No	Yes	No	Yes
Household Refusal	7.24	7.19	10.46	7.90
Other Household Non-Response	6.46	5.40	7.15	8.06
Respondent Refusal	5.08	3.96	4.38	3.07
Other Respondent Non-Response	1.70	1.80	2.43	1.92
Responses	79.52	81.65	75.58	79.05
Number of Records	11,172 (95.3%)	556 (4.7%)	17,023 (92.9%)	1,303 (7.1%)

to identify those calls that were answered by a machine for cycles 1 to 3, but we are for subsequent cycles and so can analyze to some extent the effect of their use on response rates. Table 3 compares the response rates for those households for which none of the calls were answered by a machine with those for which at least one call was. No important effect of answering machines is indicated by this table; however the increase in contacts with answering machines, from 6.4% to 9.4% of households, is dramatic (Table 3). Table 4 compares the response rates for those households for which the first answered call was answered by a machine with those for which it was not. If any effect of answering machines is indicated by this table it is that response rates are higher for those households with a first contact by answering machine. There appears to be no evidence that the use of answering machines is seriously reducing response rates.

3.10 Data Capture and Processing

The data for all five cycles were captured directly into the mini-computers in Statistics Canada's regional offices. Some simple edits to check the validity of data as captured were made at the time of capture, but these could in most cases be overridden using special functions. Following transmission of the raw data to Ottawa an exhaustive set of edits was applied to find, and correct if possible, invalid or inconsistent responses. When a response was missing, invalid or inconsistent with other responses and the approriate value could not be inferred from other responses on the questionnaire an 'unknown' code was assigned. Exceptions to this rule were three variables needed for weighting purposes: age, sex and number of telephone lines. In cases where these variables were missing the questionnaires themselves were consulted to assist in the imputation of values.

3.11 Weighting

3.11.1 Initial Weights

Both the Waksberg and ENWB methods of selecting RDD samples yield self-weighting samples of residential telephone numbers. The Waksberg method does not provide an estimate of this weight, but for GSS weighting purposes it is sufficient to use an initial weight of one (1) for telephone numbers in those strata where that method is used. In ENWB strata the initial weight is the reciprocal of the probability of selection of the telephone number. This probability is simply:

$$\frac{n_c}{100 \times N_B},$$

where:

n_c is the number of telephone numbers selected, and

N_B is the number of working banks in the frame.

3.11.2 Non-response Adjustment

The initial weight is adjusted for non-response using adjustment "strata" based on telephone geography. These are typically banks in Waksberg method strata and ACP's in ENWB strata. The initial weights are inflated by the following factor:

$$\frac{n_R + n_{NR}}{n_R},$$

where:

n_R is the number of responding households in the non-response "stratum", and

n_{NR} is the corresponding number of non-responding households.

3.11.3 Telephone Adjustment

Since households with more than one telephone line have a higher chance of being selected by an RDD survey, the initial weight adjusted for non-response (a weight for telephone numbers) is further adjusted by dividing by the number of telephone lines for the household to yield a household weight.

3.11.4 Initial Person Weight

Since only one eligible respondent per household is interviewed, the household weight must be adjusted by multiplying by the number of eligible respondents to yield a person weight.

3.11.5 Poststratification

At this point populations projected from the census are used as reference totals in the poststratification of the person weights, first to the stratum population sizes and then to the provincial age-sex populations. (It should be noted that it is only after the first stage of poststratification that the weights in Waksberg strata actually sum to a population estimate. Until this step they differ from a set of weights based on the inverses of the selection probabilities by an unknown constant of proportionality.) These two sets of reference totals are then used as the margins for a raking ratio adjustment to the weights.

4. THEMES AND RESEARCH ISSUES COVERED BY THE GSS

As indicated above, in order to cover a wide range of social issues, the GSS examines a different core topic each year for five years and then the topics are repeated. The core topics were chosen to fill perceived data gaps in the social statistics program. The five core themes are discussed in more detail below.

4.1 Health

The core content of the health cycle is directed at providing a range of measures of health status, including short and long term disability, the prevalence of common chronic conditions, such as high blood pressure or diabetes, and the use of various health care services. In addition, data are collected on life-style such as, smoking, drinking, and physical exercise. When linked to health status, these data provide information on the barriers (*e.g.* smoking, drinking) and bridges (*e.g.* physical exercise) to positive health for various population groups.

For the first GSS health cycle, the add-on focus content was directed at older Canadians and covered social networks, support given and received, as well as participation in a range of social activities. The sample size for the elderly population was also increased to allow for more in-depth analyses.

4.2 Time Use

The GSS time use survey consisted of a "24 hour time budget" generally for the day preceding the interview. Respondents provided information on each primary activity engaged

in during that day, the start time and duration of each activity, and associated information on where the activity took place and who was with the respondent at the time (*e.g.* spouse, children, friends, *etc.*). These data provide information on the frequency with which people participate in activities such as paid work, household work, attending cultural events, watching television, and the time spent on these activities.

The survey provides information on how Canadians allocate their time to activities such as paid work, housework and other non-market work and leisure activities. The data can be used to show constraints that limit a person's choice of the use of time and how these are distributed among different population groups. The inclusion of a battery of questions on satisfaction with various dimensions of life allows such measures to be correlated with patterns of time use for different population groups.

The 1986 GSS time use cycle also included a small module on intergenerational social mobility that allows for the analysis of movement on an occupational or educational hierarchy between the respondent and his or her parents.

The add-on focus content for the time use cycle was a detailed set of questions on language knowledge and use. While focus content is generally expected to be related to and complement core content, there was a demand for much more detailed language data than could be included in the population census. The information collected included data on language use at various stages of life (*e.g.* first learned, during childhood, at school) and in various settings including at home, at work, with friends, watching television, and in dealing with federal agencies. In order to allow a more detailed analysis in bilingual regions of the country, sample size was also increased in these geographic areas.

4.3 Personal Risk

The third GSS cycle was based around the topic of personal risk, including both criminal victimizations and accidents. Traditionally, information on these topics has been derived from administrative sources, such as police statistics and hospital records. However, these data provide very little information about the victim and, in addition, there are many crimes (the GSS estimates more than half) and accidents which are not reported to authorities.

The personal risk survey conducted in early 1988 asked respondents about criminal victimizations and accidents that they had experienced during calendar year 1987. Data were also collected on several life-style measures, such as alcohol consumption and frequency of night outings to allow these to be correlated with criminal victimizations and accidents. For each reported crime or accident incident, data were collected on the nature of the incident, the consequences in terms of activity restriction, medical attention and financial loss. In addition, respondents were asked to report their perceptions of crimes and accidents and about precautions taken to prevent these events.

The add-on focus content for the personal risk cycle was a set of questions on contact with the criminal justice system (*e.g.* police, courts, lawyers) and on the awareness and use of services by victims of crime.

4.4 Education and Work

While the monthly Labour Force Survey and other labour related surveys provide a wealth of information about the labour force, none of the existing surveys provides much information on the social aspects of work or the perceived quality of working life. The GSS cycle on education and work, conducted in early 1989, was designed to partially fill this data gap.

The survey was developed around three main themes that reflect fundamental changes in Canadian society: patterns and trends in work and education; new technologies and human resources; and work in the service economy. The themes reflect a range of issues on which more information is required. For example, the accelerating rate of technological innovation demands detailed knowledge about the utilization of and training for computers. Concerns about the effective utilization of the nation's human capital require a better understanding of the links between the labour force and the educational system. We also must anticipate future demands on educational institutions, and changing relationships between educational attainment and socio-economic outcomes. This round of the GSS also augments existing data sources by providing new information about the elderly population as well as some of the socio-economic implications of the baby boom generation entering middle age.

The survey collected a partial work and education history. It also included information on technology training and the use of computers, and on future plans for education. Subjective information also was sought in the form of a series of questions about satisfaction with retirement and other dimensions of life, as well as a block of questions on attitudes to science and technology.

4.5 Family and Friends

The fifth cycle of the GSS was based around issues related to family and friends and was completed in early 1990. While the Census and other household surveys provide family-based data, changes in family life have resulted in a need for new types of information. One shortcoming of existing data is that generally they are based on a rather narrow concept of the family, in particular a nuclear family of parents and children or perhaps an economic family of related individuals living in the same household. This survey looks at the family in a broader context and collects information on the extent and nature of kinship networks and related questions of patterns of informal help and support among family and friends.

A second major theme of the survey is a result of the trends in marriage, divorce, and the increased frequency of common law unions. Increased numbers of Canadians are living in more than one union during their life time. The impact of such changes on family life and children is substantial and can best be studied by an analysis of marital and family history data. Such data were retrospectively collected in a special Family History Survey conducted in 1984 (Burch 1985). The GSS family cycle incorporates the collection of these data on a regular basis. Specific issues that can be addressed include changing patterns of union formation and dissolution, the situation of single parent families, and home leaving patterns of young adults.

A third but more minor theme of the cycle is concerned with the division of household labour.

5. PROGRAM OUTPUTS

The GSS results are disseminated in a variety of ways. For each survey there are one or more publications that present the results of data analysis with respect to particular social issues and the monitoring of conditions and trends. The results of Cycle 1 are reported in Statistics Canada (1987) and Stone (1988); the results of Cycle 2 are presented in Harvey *et al.* (1991) and Creese *et al.* (1991); and the Cycle 3 results are reported in Sacco and Johnson (1990) and Millar and Adams (1991). Publications containing the results of other cycles are in preparation. The general public are made aware of GSS results through the publication of reports in the media which are often based on articles published in *Canadian Social Trends*, a quarterly Statistics Canada publication that is targeted to a general audience.

A second product is a public use microdata file and associated documentation to enable university and other researchers to carry out their own analysis of the data. These data are also useful for teaching purposes. Microdata files from the first five survey cycles are now available.

In addition to the product outputs, the GSS program has developed a survey capacity. This is not simply a system for data collection and processing, but includes other major components. Content research and development, related data specification, analysis of survey and other relevant data, dissemination of informative results as well as the development and use, where applicable, of improved methods of collection, processing, analysis and dissemination are all components of the evolving survey capacity of the GSS group.

6. FUTURE PLANS

As the GSS program moves into the second round of surveys, attention has shifted from the problems of developing and fielding five new surveys to further building the survey program through partnerships with others. The first round of surveys has had a modest success with obtaining buy-ins of additional sample and/or focus content. Only Cycle 4 had neither focus content nor increased sample size. For the first time, the 1990 survey had provincial participation, with the Ontario government funding an increase in sample size.

A new initiative for the GSS program is an investigation of the potential for expanding the scope of the survey to include interviewing a sub-sample of respondents again in future cycles. In the short term, this could provide an enriched data set by linking content from different cycles. In the longer term, it could serve to provide longitudinal data by interviewing respondents on the same topic five years later. A feasibility study was conducted in 1990 and the possibility of interviewing a sample of respondents from a previous cycle is now offered to interested sponsors.

The GSS will also continue to undertake a range of more general research and development activities. Core content of the first set of cycles will be reviewed and input sought from users as to possible improvements for future cycles. While new and alternative survey designs and approaches will be considered, any potential changes will have to be balanced against the impact on data comparability that is required for the long term goal of monitoring change. In addition, the content from the first round of surveys will be reviewed from the point of view of consistency and integration across GSS survey cycles and between the GSS and the 1991 Census and other household surveys. On-going development of the GSS infrastructure will also continue. Consideration was given to changing to monthly data collection (and monthly data collection was implemented for Cycle 6) and will be given to supplemental collection methods (*e.g.* mail). Attempts are also being made to shift processing of the survey to a micro-computer environment to further improve timeliness. Finally, new procedures, such as computer-assisted telephone interviewing, will be considered as these become available as part of a larger Statistics Canada survey development program.

In summary, the GSS Program during the coming years will focus on building on the firm foundation that has been established during the first round of surveys. The primary objective will continue to be the measurement of social conditions and the gradual development of a time series to monitor trends. In addition, flexibility will be maintained in order to quickly respond to new and emerging social information needs.

REFERENCES

BURCH, T.K. (1985). *Family History Survey: Preliminary Findings*. Catalogue 99-955, Statistics Canada.

CREESE, G., GUPPY, N., and MEISSNER, M. (1991). Ups and downs on the ladder of success. *General Social Survey Analysis Series*. Catalogue 11-612E, No. 5, Statistics Canada.

HARVEY, A.S., MARSHALL, K., and FREDERICK, J.A. (1991). Where does time go? *General Social Survey Analysis Series*. Catalogue 11-612E, No. 4, Statistics Canada.

KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.

KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society* A, 139, 80-95.

MILLAR, W., and ADAMS, O. (1991). Accidents in Canada. *General Social Survey Analysis Series*. Catalogue 11-612E, No. 3, Statistics Canada.

SACCO, V., and JOHNSON, H. (1990). Patterns of criminal victimization in Canada. *General Social Survey Analysis Series*. Catalogue 11-612, No. 2, Statistics Canada.

STATISTICS CANADA (1987). Health and social support, 1985. *General Social Survey Analysis Series*. Catalogue 11-612, No. 1, Statistics, Canada.

STATISTICS CANADA (1989). *Household facilities and equipment*. Catalogue 64-202, Statistics Canada.

STATISTICS CANADA (1990a). *Overview of Special Surveys 1989*, Household Surveys Division, Statistics Canada.

STATISTICS CANADA (1990b). *Household facilities by income and other characteristics*. Catalogue 13-218, Statistics Canada.

STONE, L. (1988). *Family and Friendship Ties Among Canada's Seniors*. Catalogue 89-508, Statistics Canada.

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

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees, sometimes more than once, during 1991:

J. Alho, *University of Joensuu*
J. Armstrong, *Statistics Canada*
Y. Bélanger, *Statistics Canada*
D. Bellhouse, *University of Western Ontario*
K. Bennett, *Statistics Canada*
J.-M. Berthelot, *Statistics Canada*
D.A. Binder, *Statistics Canada*
K. Bollen, *University of North Carolina - Chapel Hill*
P.A. Buesher, *North Carolina Center for Health Statistics*
K.P. Burnham, *U.S. Fish and Wildlife and Colorado State University*
S.J. Butani, *U.S. Bureau of Labor Statistics*
G.H. Choudhry, *Statistics Canada*
M.L. Cohen, *University of Maryland*
C.D. Cowan, *Opinion Research Corporation*
E.B. Dagum, *Statistics Canada*
J.-C. Deville, *INSEE*
D. Dolson, *Statistics Canada*
J.D. Drew, *Statistics Canada*
F.J. Fowler, Jr., *University of Massachusetts*
J.G. Gambino, *Statistics Canada*
J.F. Gentleman, *Statistics Canada*
M.E. Gonzalez, *U.S. Office of Management and Budget*
J.-F. Gosselin, *Statistics Canada*
H. Gough, *Statistics Canada*
A. Gower, *Statistics Canada*
G.B. Gray, *Statistics Canada*
M.A. Hidiroglou, *Statistics Canada*
M.A. Hill, *Systat Inc.*
G.J.C. Hole, *Statistics Canada*
D. Holt, *University of Southampton*
J. Hox, *University of Amsterdam*
B. Hulliger-Domingues, *Swiss Federal Statistical Office*
G. Kalton, *University of Michigan*
B. Lefrançois, *Statistics Canada*
T.P. Liu, *Statistics Canada*
M. March, *Statistics Canada*
S.M. Miller, *U.S. Bureau of Labor Statistics*
I. Munck, *Statistics Sweden*
J.C. Nash, *University of Ottawa*
H.B. Newcombe, *Consultant*
S. Presser, *University of Maryland*
D. B. Radner, *U.S. Social Security Administration*
J.N.K. Rao, *Carleton University*
P.S.R.S. Rao, *University of Rochester*
L.-P. Rivest, *Université Laval*
W. Rodgers, *University of Michigan*
D.B. Rubin, *Harvard University*
K. Rust, *Westat Inc.*
I. Sande, *Bell Communications Research*
C.E. Särndal, *University of Montreal*
A. Satin, *Statistics Canada*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
I. Schiopu-Kratina, *Statistics Canada*
K.P. Srinath, *Statistics Canada*
C.M. Suchindran, *University of North Carolina - Chapel Hill*
S. Sudman, *University of Illinois - Urbana-Champaign*
A. Sunter, *A.B. Sunter Research Design & Analysis, Inc.*
L. Swain, *Statistics Canada*
R.B.P. Verma, *Statistics Canada*
P.R. Voss, *University of Wisconsin - Madison*
J. Waksberg, *Westat Inc.*
G.S. Werking, *U.S. Bureau of Labor Statistics*
W.E. Winkler, *U.S. Bureau of the Census*
K.M. Wolter, *A.C. Nielsen*
A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1991 issues: S. Beauchamp and S. Lineger (Photocomposition), G. Gaulin (Author Services), and M. Haight (Translation Services). Finally we wish to acknowledge M. Kent, C. Larabie and D. Lemire of Social Survey Methods Division, for their support with coordination, typing and copy editing.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 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.