Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 16, NUMBER 1
JUNE 1990

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

JUNE 1990

Published under the authority of
the Minister of Industry, Science and Technology

September 1990

Price: Canada: $30.00 a year
United States: US$36.00 a year
Other Countries: US$42.00 a year

Catalogue 12-001, Vol. 16, No. 1

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.

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, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two 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 $30 per year in Canada, US $36 in the United States, and US $42 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 16, Number 1, June 1990

CONTENTS

In This Issue 1

Special Section – History and Emerging Issues in Censuses and Surveys

J.N.K. RAO and D.R. BELLHOUSE
History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis 3
Comments: T.M.F. SMITH 26

S.E. FIENBERG and J.M. TANUR
A Historical Perspective on the Institutional Bases for Survey Research in the United States 31
Comments: R.M. GROVES 47

B.A. BAILAR
Contributions to Statistical Methodology from the U.S. Federal Government 51
Comments: G.J. BRACKSTONE 58

L. KISH
Rolling Samples and Censuses 63
Comments: F. SCHEUREN 72

M.H. HANSEN
Comments on Articles in the Special Section 81

REPLIES: J.N.K. RAO and D.R. BELLHOUSE 87
S.E. FIENBERG and J.M. TANUR 89
B.A. BAILAR 91
L. KISH 93

T. DALENIUS and C.-E. SÄRNDAL
Some Developments of Sampling Techniques and their Use in Official Statistics in Sweden 95

P.S. KOTT
Variance Estimation when a First Phase Area Sample is Restratified 99

D.B. WHITE
Estimation Using Double Sampling and Dual Stratification 105

C. JULIEN and F. MARANDA
Sample Design of the 1988 National Farm Survey 117

CONTENTS – Concluded

D.A. HAY
Does the Method Matter on Sensitive Survey Topics? 131

E.R. LANGLET
Use of Cluster Analysis for Collapsing Imputation Classes 137

Y. BÉLAND and A. THÉBERGE
An Example of the Use of Randomization Tests in Testing the Census Questionnaire 145

P.J. CANTWELL
Variance Formulae for Composite Estimators in Rotation Designs 153

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 1–2
Statistics Canada

In This Issue

In this issue's special section, we take a look back and a look forward. Our contributors to this section are well-known survey statisticians who bring a wealth of experience and knowledge. By looking back with clarity to developments in our field, they enable us to look forward to areas of emerging interest. With one exception, each paper has discussants, with a reply by the authors.

Rao and Bellhouse present an historical perspective on sample survey theory and methods. Beginning with a discussion of some of the earliest developments in the field, they then take us through the design-versus model-based debate, variance estimation methods, analysis of survey data and recent developments in computer software. The paper includes an extensive bibliography. Smith's comments complement the paper, providing a somewhat different perspective, including some thoughts on the position of sample survey theory relative to "mainstream" statistics.

Beginning with a discussion of the role of governments and social researchers in the earliest sample surveys and censuses, Fienberg and Tanur describe the institutional bases for survey research, particularly in the United States. Among the organizations considered are government agencies, statistical associations, polling firms and universities. The authors discuss recent developments including increased telephone interviewing and cognitive aspects of surveys. They end by discussing links among the various sectors which make up the field. In his discussion, Groves also looks at the sectors and states that movement of people among them has been less common than Fienberg and Tanur's examples suggest. He also adds substantially to the list of recent developments.

Whereas Fienberg and Tanur look at government institutions as one component out of several, Bailar focuses on the important role played by the U.S. Bureau of the Census in the development of sample survey methods. She discusses the motivation for, and development of, various methods and approaches including sampling and seasonal adjustment. The paper concludes with a look to the future. Brackstone emphasizes that practical problems gave rise to the advances discussed by Bailar. He also adds several other contributions made by Statistics Canada and other agencies to those mentioned by Bailar. Brackstone also points to the importance of a suitable environment to encourage innovation.

Kish discusses alternatives to current periodic censuses. He rekindles the debate on the feasibility of replacing them by rolling censuses. He discusses the use of administrative data in this context, pointing out the existence of good sources of data in some countries. An important issue is how to cumulate data from rolling samples and censuses. Various alternatives are discussed. In his discussion, Scheuren points out that Kish is, in effect, advocating a major shift in our way of thinking – always a difficult task. While Scheuren feels that pure rolling censuses are likely to be too expensive, variations, along with the use of improved administrative data, should be feasible. Both authors agree that there is much research required for further progress.

We are pleased to have Morris Hansen, who participated in many important developments mentioned by the authors, as a discussant of all the above papers. He adds important historical details and corrects some errors and misconceptions. One item of particular interest is Hansen's discussion of the reluctance to introduce sampling – something which we now tend to take for granted. His insightful comments on individual topics are too numerous and varied to summarize here.

Dalenius and Särndal initially intended to discuss Bailar's paper, but their paper metamorphosed into a history of sampling techniques in Sweden. As such, it serves as a summary and update of Dalenius's 1957 book.

The remaining papers in this issue of **Survey Methodology** deal with a diversity of topics. Kott proposes an unbiased estimator of variance for a two-phase sampling design where both phases are stratified simple random sampling. Such designs are commonly used, especially in agricultural surveys.

Two-phase sampling with stratification at both phases is also the subject of White's paper. An estimator due to Vardeman and Meeden which uses prior information is studied via simulation. Some theoretical results are also given for the case where the prior information is not used.

Julien and Maranda describe the sample design used for the National Farm Survey since 1988. The efficiency of the new design is evaluated by comparing the precision of the survey estimates for 1988 to those for 1987, as well as to the expected precision obtained during the development of the new design.

The results of a study in Saskatchewan are analyzed by Hay to examine the effects on responses of the method of data collection: self-administered questionnaire versus personal interview. Although statistically significant differences are found, they are not of sufficient magnitude to be of practical importance.

Langlet studies the use of cluster analysis to deal with the problem of imputation for item nonresponse. This technique would be especially useful in situations where the number of imputation classes is rather large.

Béland and Théberge use randomization tests to compare two questionnaires which were used to study the questions likely to be asked in the 1991 census. Since tests of this type may not be familiar to many survey methodologists, this paper will serve as a useful introduction.

In his paper, Cantwell derives a simple variance expression for a general composite estimator commonly considered for rotating designs. He deals with both single-level and multi-level rotation plans.

The Editor

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 3–29
Statistics Canada

History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis

J.N.K. RAO and D.R. BELLHOUSE[1]

ABSTRACT

Early developments in sampling theory and methods largely concentrated on efficient sampling designs and associated estimation techniques for population totals or means. More recently, the theoretical foundations of survey based estimation have also been critically examined, and formal frameworks for inference on totals or means have emerged. During the past 10 years or so, rapid progress has also been made in the development of methods for the analysis of survey data that take account of the complexity of the sampling design. The scope of this paper is restricted to an overview and appraisal of some of these developments.

KEY WORDS: Foundations of inference; Analysis of survey data; Computer software.

1. SOME EARLY MILESTONES

The motivation behind much of the work in survey sampling prior to the 1950's or 60's was the desire to obtain reasonably efficient estimates, at a desired cost, of totals, means, or proportions for large, and increasingly complex-structured, finite populations. A discussion of the early work in sampling human populations may be found in several review papers (see *e.g.*, Hansen, Dalenius and Tepping 1985 and Bellhouse 1988).

The history of the mathematical theory of survey sampling has its origins in the late nineteenth century through the work of the Norwegian statistician A.N. Kiaer. Kiaer was the first to promote what was then called 'the representative method', or sampling, over complete enumeration. What Kiaer (1897) meant by representative sampling was that the sample should mirror the parent finite population. This can be achieved in two ways, by randomization or by balanced sampling through purposive selection. Initially, purposive selection was the preferred method of sample selection, but gradually randomization became a strong competitor to balanced sampling for sample selection. By the 1920's random sampling and purposive selection were both widely used as sample selection techniques. The major theoretical developments in both areas which occurred during this era are summarized in Bowley (1926). This summary includes the development of stratified random sampling with proportional allocation and the derivation of formulae to obtain the precision of an estimate from a purposively selected sample.

The equal footing of random sampling and purposive selection gradually changed after the publication of Neyman's (1934) classic paper. Neyman was able to show, both theoretically and with practical examples, why random sampling was preferable to purposive selection for the large-scale sampling problems of the day. With the publication of the 1934 paper, Neyman also opened up new avenues of development for random sample selection techniques. Previously, Bowley and his followers used only sampling designs with equal inclusion

[1] J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6. D.R. Bellhouse, Department of Statistics, University of Western Ontario, London, Ontario, N6A 5B9.

probabilities for every population unit. Their reasoning was that this method of sampling would provide a representative sample of the universe. Neyman (1934) broke out of this sampling straitjacket with his theories of stratified sampling with "optimal" allocation and cluster sampling with ratio estimation. In both situations, "valid" estimates of population totals, means or proportions are obtained without reliance on a representative sample selected through a design with equal inclusion probabilities. Neyman's final contribution to the theory of survey sampling is his introduction of cost functions to find the sample allocation in two phase sampling which minimized the variance subject to a fixed budget (Neyman 1938).

Neyman's fundamental contributions inspired various important extensions of his theory. Among these, we should mention ratio and regression estimation with two-phase sampling (Cochran 1939), determination of "optimal" stratification points and "optimal" allocation with multiple parameters/characters (Dalenius 1957), and sampling on two occasions with partial replacement of units (Jessen 1942) which was subsequently extended by Patterson (1950) and Hansen *et al.* (1953, pp. 470-503) to sampling on more than two occasions (also called rotation sampling). Rotation sampling and associated "composite" estimates are now extensively used to estimate levels and changes from continuing large scale, multi-purpose surveys (*e.g.*, the Current Population Survey (CPS) carried out by the U.S. Bureau of the Census).

Neyman's work also greatly influenced Morris Hansen, William Hurwitz, and their colleagues at the U.S. Bureau of the Census. Inspired by their practical problems in large-scale survey design and by Neyman's approach to sampling theory, Hansen and Hurwitz (1943) developed the theory of sampling with probability proportional to size and with replacement (also called PPS sampling). The effect of this approach to multistage surveys is that it provides approximately equal interviewer work loads which makes the administration of a multistage survey easier. This procedure also leads to significant reductions in the variances of the estimates, by controlling the variability arising from unequal cluster sizes without actually stratifying by size and thus allowing stratification on other variables to reduce variance. The theory of Hansen and Hurwitz was extended by Horvitz and Thompson (1952) and Narain (1951) to unequal probability sampling without replacement. By making the inclusion probabilities of units at each stage proportional to their sizes, the desirable features of the Hansen-Hurwitz method are retained, using the so-called Horvitz-Thompson estimator of a population total. The basic work of Horvitz and Thompson and Narain stimulated many theoretical and applied contributions to unequal probability sampling without replacement. Brewer and Hanif (1983) and Chaudhuri and Vos (1988) have provided comprehensive accounts of these developments.

Madow and Madow (1944) have given the basic theory of systematic sampling, and introduced population models to examine the features of systematic sampling. Cochran (1946) introduced the "superpopulation" approach in which the finite population is regarded as being drawn from an infinite superpopulation having certain properties. The expected (or anticipated) variances under the superpopulation model are then compared to study the relative efficiency of alternative sampling strategies. His 1946 paper stimulated much subsequent research in the use of superpopulation models in the choice of sampling strategies and also for model-dependent or model-assisted inference (see Section 2).

Mahalanobis (1946) developed the technique of interpenetrating subsamples, and used it extensively in large-scale surveys in India for assessing both sampling and non-sampling errors. This technique consists of drawing the sample in the form of two or more independent subsamples according to the same sampling scheme such that each subsample provides a valid estimate of the parameter of interest. By assigning the subsamples to different interviewers

(or interviewer teams), a valid estimate of the total variance can be obtained that takes proper account of the correlated response variance component due to interviewers. Deming (1960) used this method (sometimes called replicated sampling) extensively to obtain simple estimates of variance. It has led to resampling techniques such as the jackknife, balanced repeated replication and the bootstrap for getting variance estimates of complex non-linear statistics (see Section 3).

Yet another milestone in the emergence of ideas and theory surrounding complex surveys is the concept of design effect (DEFF), due to Leslie Kish (see Kish 1965, sec. 8.2). The design effect is defined as the ratio of the actual variance of a statistic under the specified design to the variance which would be achieved under a simple random sample of the same size. The concept of design effect has been found to be especially useful in the presentation and modelling of sampling errors, and also in the analysis of survey data involving clustering and stratification (see Section 4).

2. THEORETICAL FOUNDATIONS

Although Neyman (1934) and others obtained best linear unbiased estimators for simple designs using the standard Gauss-Markov set-up, the development of traditional sampling theory progressed more or less inductively. Estimators (and designs) which appeared reasonable were considered and their relative properties carefully studied by analytical and/or empirical methods, mainly through comparisons of bias and mean square error, and sometimes also using anticipated mean square error or variance under plausible superpopulation models. As noted by Hansen *et al.* (1983), unbiasedness of estimators under a given design was not insisted on since it "often results in much larger mean square errors than necessary". Instead, asymptotic design consistency of estimators was insisted on, at least when aggregate estimates from reasonably large samples are needed, and the mean square errors of selected asymptotically design consistent estimators were compared to arrive at a suitable estimator (and design). Moreover, in large-scale surveys involving a great many statistics, uniform estimation procedures are often insisted on at the expense of variance inflation for some statistics (compared to alternative estimators tailored to each statistic), due to time, cost and other operational constraints.

Despite the usefulness of the traditional approach, the need for a formal framework for inference from survey data was long felt. Realizing this need, several statisticians have made important contributions to the theoretical foundations of inference from survey data, especially during the past 10-20 years. Several review papers (see *e.g.*, Chaudhuri 1988) and two books (Cassel *et al.*, 1977; Chaudhuri and Vos 1988) discuss various aspects of the theoretical foundations.

Most papers on the theoretical foundations of sampling theory have assumed the following somewhat idealistic set-up. A survey population U consists of N distinct elements identified through the labels $j = 1, \ldots, N$. The characteristic of interest y_j (possibly vector-valued) associated with element j can be known **exactly** by observing element j. Thus response or measurement errors are assumed to be absent or ignored if present. The parameter of interest is the population total $Y = y_1 + \ldots + y_N$ or the population mean $\bar{Y} = Y/N$ (if N is known). A sample is a subset s of U and the associated y-values, *i.e.*, $\{(i,y_i), i \in s\}$, selected according to a sampling plan which assigns a known probability $p(s)$ to s such that $p(s) \geq 0$ for all $s \in S$ (the set of all possible s) and $\sum_{s \in S} p(s) = 1$. The selection probability $p(s)$ can depend on known design variables $z = (z_1, \ldots, z_N)'$, such as stratum indicator variables and size measures of clusters, *i.e.*, $p(s) = p(s \mid z)$ where z_j is possibly vector-valued. For

probability sampling, the inclusion probabilities $\pi_j = \sum_{\{s:j\in s\}} p(s)$ are positive, which permits unbiased or consistent estimation of Y in the traditional sense. It is also customary to impose the condition that the joint inclusion probabilities $\pi_{ij} = \sum_{\{s:(i,j)\in s\}} p(s)$ be positive, which permits unbiased or consistent variance estimation in the traditional sense.

The basic problem is to make inferences (estimation, variance estimation and constructing confidence intervals), about the total Y by observing a sample selected according to a specified sampling plan $p(s)$ and also using available supplementary data. This involves essentially three steps: (i) choice of a sampling plan; (ii) choice of an estimator $\hat{Y}$; (iii) choice of a variance estimator and confidence intervals. There are essentially three different approaches to implement these steps: (i) design-based approach, also called probability sampling approach or randomization approach; (ii) model-dependent approach, also called prediction approach or probability speculation approach (Hájek 1981), (iii) a hybrid approach, called model-based approach or model-assisted approach. Developments to date under each of these three approaches are discussed below.

2.1 Design-based Approach

This approach uses probability sampling both for sample selection and for inference from the data. The probability sampling distribution provides valid inferences irrespective of the population y-values, even in complicated situations, in the sense that the pivotal $t = (\hat{Y} - Y)/s(\hat{Y})$ is approximately $N(0,1)$, at least for large samples, where $s(\hat{Y})$ is the standard error of $\hat{Y}$. This approach has been critized on the grounds that such inferences, although assumption-free, refer to repeated sampling from the survey population involving all samples $s \in S$ and the associated probabilities $p(s)$, instead of just the particular s that has been drawn. This criticism can be countered to some extent by using either conditional design-based inference referring to a subset of S that is "relevant" to the particular s or by a model-assisted approach.

Horvitz and Thompson (1952) made a basic contribution to foundational aspects of design-based inference by formulating three classes of linear estimators of Y, and then raising the possibility that the best (minimum variance) estimator among all possible linear unbiased estimators of Y may not exist, even for simple random sampling. Prompted by the Horvitz-Thompson formulation, Godambe (1955) proposed a general class of linear estimators given by $\hat{Y}_b = \sum_{i\in s} b_{si} y_i$, where the weight b_{si} is attached to element i if s is selected and $i \in s$. He proved that no best unbiased estimator of Y could exist in this class, for any sampling plan $p(s)$. Since the criterion of minimum variance had failed, several alternative criteria for the choice of an estimator were proposed. Among these, the admissibility criterion is of some use but is not sufficiently selective in distinguishing between the merits of estimators since too many estimators are admissible. Ghosh (1987) provides an excellent survey of results on admissibility and related criteria in finite population sampling. New criteria that give rise to a unique choice of estimator in the Godambe class for any sampling plan have also been put forth, but the optimality properties established have questionable relevance (see Rao 1971, Rao and Singh 1973). Basu's (1971) well-known "elephants" example demonstrates the futility of two such criteria, *viz.* necessary bestness and hyperadmissibility.

Godambe (1966) obtained the likelihood function from the sample $\{(i,y_i), i \in s\}$ regarding the N-vector $y = (y_1, \ldots, y_N)'$ as the parameter of interest, but it provides no information on $(y_i: i \notin s)$, and hence on the total Y, since the N population units are essentially treated as N separate post strata. A way out of this difficulty is to ignore some of the data to make the sample non-unique and arrive at an informative likelihood function (Hartley and Rao 1968; Royall 1968). Another route is to combine the uninformative likelihood function with exchangeable priors via Bayes theorem to arrive at informative posterior inferences (Ericson 1969).

Conditional inference has attracted considerable attention (and controversy) in classical statistics since Fisher (1925). The choice of a relevant reference set for making conditional inference is not always clear-cut, but in the context of post-stratification it seems sensible to make design-based inferences conditional on the realized strata sample sizes (Durbin 1969). Holt and Smith (1979) provide the most compelling arguments in favour of conditional design-based inference, although their discussion was confined to post-stratification of a simple random sample. Rao (1985) considered a number of real examples involving random sample sizes to illustrate conditional design-based inference and associated difficulties.

Robinson (1987) considered conditional design-based inference from a simple random sample when only the population total X of a concomitant variable x is known. By conditioning on the observed sample mean $\bar{x}$, he showed that the usual ratio estimator $\hat{Y}_r = (\bar{y}/\bar{x})X$ is conditionally biased. He obtained a conditional bias adjusted ratio estimator given by

$$\hat{Y}_r(adj) = \hat{Y}_r + N(r - b)(\bar{x} - \bar{X})\bar{X}/\bar{x}, \tag{2.1}$$

where $r = \bar{y}/\bar{x}$ and b is the sample regression coefficient. He also showed that a customary variance estimator

$$s_c^2(\hat{Y}_r) = N^2(1 - n/N) \sum_{i \in s} (y_i - rx_i)^2/n(n - 1) \tag{2.2}$$

is conditionally biased, while another classical variance estimator

$$s_a^2(\hat{Y}_r) = (\bar{X}/\bar{x})^2 s_c^2(\hat{Y}_r) \tag{2.3}$$

is in fact conditionally unbiased, for large n. Robinson also showed, through a simulation study, that $s_a^2(\hat{Y}_r)$ is very close to the estimator of conditional variance of $\hat{Y}_r(adj)$.

2.2 Model-dependent Approach

A strict model-dependent approach involves purposive sampling, and the model distribution (generated from hypothetical realizations of $\mathbf{y} = (y_1, \ldots, y_N)'$ obeying the model) provides valid inferences referring to the particular sample s that has been drawn.

The model-dependent approach was first proposed by Brewer (1963) and extensively studied by Royall and his co-workers, starting with Royall (1970). It is best illustrated under a simple regression model

$$E_m(y_i) = \beta x_i, \quad i = 1, \ldots, N; \quad \beta > 0, x_i > 0 \tag{2.4}$$

where E_m denotes the model expectation. It is further assumed that the model variance $V_m(y_i) = \sigma_i^2$ where σ_i^2 is known except for a multiplicative constant, and that the model covariance $\text{cov}_m(y_i, y_j) = 0$, $i \neq j$. Royall (1970) showed that the customary design-unbiased estimator, $N\bar{y}$, under simple random sampling is biased under the model given by (2.4), and that $N\bar{y}$ leads to serious underestimation if the observed sample contains mostly units with small sizes, x_i. These results can also be shown under the conditional design-based approach without assuming a model (Rao 1985).

The best linear model unbiased estimator (or prediction estimator) of Y under the model (2.4) is given by

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{\beta} x_i \tag{2.5}$$

which reduces to the usual ratio estimator $\hat{Y}_r$ if $\sigma_i^2 = \sigma^2 x_i$, where $\bar{s} = U - s$ is the set of non-sampled units and $\hat{\beta}$ is the best linear unbiased estimator of β. The uncertainty in $\hat{Y}$ is measured by $E_m(\hat{Y} - Y)^2 = V_m(\hat{Y} - Y)$ which in the case of $\hat{Y}_r$ reduces to

$$V_m(\hat{Y} - Y) = \{X(X - n\bar{x})/(n\bar{x})\}\sigma^2. \tag{2.6}$$

Since (2.6) decreases as $\bar{x}$ increases, the optimal design is a purposive sample consisting of the n units whose x-values are largest, assuming that the population x_i's are known. A model unbiased estimator, $s_m^2(\hat{Y} - Y)$, of $V_m(\hat{Y} - Y)$ is obtained from (2.6) by replacing σ^2 with its weighted least squares estimator $\hat{\sigma}^2$, and the resulting pivotal $t_m = (\hat{Y} - Y)/s_m(\hat{Y} - Y)$ is approximately $N(0,1)$ under the model distribution. These theoretical results are impressive, but such model-dependent strategies could lead to serious biases if the assumed model is not completely correct.

To protect against model misspecifications, Royall and Herson (1973 a,b) considered model deviations consisting of second or higher order polynomial terms in x (say q-th order) or an intercept or both, and demonstrated that a balanced sample for which $\bar{x}^{(j)} = \bar{X}^{(j)}$, $j = 1, \ldots, q$ provides robustness in the sense that $\hat{Y}_r$ remains model unbiased, where $\bar{x}^{(j)} = \sum_{i\epsilon s} x_i^j/n$ and $\bar{X}^{(j)} = \sum_{i\epsilon U} x_i^j/N$. Further, they have shown that stratification on x with optimal allocation and balanced sampling within each stratum together with the separate ratio estimator of Y provides increased efficiency. Purposively chosen balanced samples have a number of difficulties, nevertheless. First, due to lack of rigorous rules in the sample selection one might be tempted to select units whose x_i are close to $\bar{X}$ (in the case of $q = 1$) which can produce an unrepresentative sample if y is positively correlated with x (Yates 1960, p. 40). Second, balancing is sensitive to departures from the polynomial regression model (Madow 1978, p. 320). Balance is required on the alternative model, which may contain higher-order polynomial terms or other variables or both, and the extra variables in the alternative model must be known in advance. Third, balanced sampling is not feasible for surveys with multiple characters of interest since different samples may be required for each variable.

If the extra concomitant variables z in the model are unknown or unmeasured, Royall and Pfeffermann (1982) recommend simple random sampling since it provides "grounds for confidence that the selected sample is not badly unbalanced on z", but more recently Royall and Cumberland (1988) seem to favour some form of restricted randomization: "Many techniques, including restricted randomization, stratification and systematic sampling, can be used to help achieve balanced samples. We are not advocating one scheme over another; . . .". In any case, it appears that most advocates of the model-dependent approach seem to recommend probability sampling in some form, as noted by Smith (1984), and hence the main difference between the probability sampling approach and the model-dependent approach is in the choice of the pivotal involving the estimator $\hat{Y}$ and a measure of its uncertainty.

Despite the above-mentioned limitations, the model-dependent approach is useful for studying the conditional performances of conventional procedures, under different plausible models. For instance, the variance estimator $s_a^2(\hat{Y}_r)$ is consistent with the behaviour of the conditional variance $V_m(\hat{Y}_r - Y)$ under the model (2.4) with $\sigma_i^2 = \sigma^2 x_i$, while $s_c^2(\hat{Y}_r)$ is model-biased (Royall and Eberhardt 1975). The variance estimator $s_a^2(\hat{Y}_r)$ is also robust to deviations from the assumption $\sigma_i^2 = \sigma^2 x_i$.

2.3 Model-assisted Approach

Hansen, Madow and Tepping (1983) illustrated the dangers in using model-dependent strategies even when the model is apparently consistent with the sample data. By introducing

a misspecification to the model (2.4) which is not detectable through tests of significance from samples as large as 400, they showed that the design-based coverage of the confidence intervals derived from the model-dependent pivotal $t_r = (\hat{Y}_r - Y)/s_a(\hat{Y}_r)$ is substantially less than the desired level and that it becomes worse as the sample size increases. The poor performance of t_r was due to the asymptotic inconsistency of the estimator $\hat{Y}_r$ with respect to their stratified random sampling design.

The model-assisted approach considers only asymptotically design consistent estimators $\hat{Y}$ that are also model unbiased under an assumed model. Variance estimators that are consistent for the design variance of $\hat{Y}$ and at the same time model unbiased (at least approximately) for the conditional variance $V_m(\hat{Y} - Y)$ are also constructed. Thus the resulting pivotal leads to valid inferences under an assumed model and at the same time protects against model misspecifications in the sense of providing valid design-based inferences irrespective of the population y-values. However, very little attention has been given to studying conditional design-based properties of model-assisted strategies under model misspecifications.

Godambe (1955) assumed the model (2.4) with $V_m(y_i) = \sigma_i^2$ and $\text{cov}_m(y_i, y_j) = 0, i \neq j$, and obtained a lower bound, $\sum_{i \in U}(1/\pi_i - 1)\sigma_i^2$, to the anticipated variance of any design unbiased linear estimator, $\hat{Y}_b$. He also showed that any fixed sample size plan with $\pi_i = (nx_i)/X$ together with the Horvitz-Thompson estimator, $\hat{Y}_{HT} = \sum_{i \in s} y_i/\pi_i$, attains the lower bound, provided $\sigma_i^2 = \sigma^2 x_i^2$. "Optimal" design unbiased strategies do not exist if $\sigma_i^2 \neq \sigma^2 x_i^2$, and as a result asymptotically optimal strategies were developed by relaxing the restriction to design unbiased estimators and considering asymptotically design-consistent estimators. The generalized regression estimator

$$\hat{Y}_{reg} = \sum_{i \in s} y_i/\pi_i + \hat{\beta}\left(X - \sum_{i \in s} x_i/\pi_i\right) \tag{2.7}$$

for any fixed sample size plan with π_i proportional to σ_i is asymptotically optimal (*i.e.*, the asymptotic anticipated variance attains the lower bound), where $\hat{\beta}$ is a linear model unbiased estimator of β and $E_m E_p(\hat{\beta} - \beta)^2 \to 0$ as $n \to \infty$, where E_p denotes the design expectation (Särndal 1980). In particular, the best model unbiased estimator $\hat{\beta} = (\sum_{i \in s} w_i x_i y_i)/(\sum_{i \in s} w_i x_i^2)$ with $w_i = 1/\sigma_i^2$ may be chosen.

If $\hat{\beta} = (\sum_{i \in s} w_i x_i y_i/\pi_i)/(\sum_{i \in s} w_i x_i^2/\pi_i)$ with $w_i = 1/x_i$ is chosen, then $\hat{Y}_{reg}$ reduces to the simpler form (ratio estimator)

$$\hat{Y}_{reg} = X\hat{\beta} = \sum_{i \in s} g_{si} y_i/\pi_i, \tag{2.8}$$

where $g_{si} = X/(\sum_{i \in s} x_i/\pi_i)$ and g_{si} converges in probability to 1 as $n \to \infty$ (Särndal and Wright 1984). Särndal, Swensson and Wretman (1989) proposed a new variance estimator for estimators $\hat{Y}$ of the form (2.8) which is design consistent and at the same time approximately unbiased for the conditional variance $V_m(\hat{Y} - Y)$. Their variance estimator for $\hat{Y}_{reg}$ is given by

$$s^2(\hat{Y}_{reg}) = \sum_{i<j \in s} (\pi_i \pi_j - \pi_{ij})\pi_{ij}^{-1}(g_{si}\tilde{e}_i - g_{sj}\tilde{e}_j)^2 \tag{2.9}$$

where $\tilde{e}_i = (y_i - \hat{\beta}x_i)/\pi_i$. For simple random sampling, $s^2(\hat{Y}_{reg})$ reduces to $s_a^2(\hat{Y}_r)$, given by (2.3), which was justified under the prediction and conditional randomization approaches. Kott (1987) proposed a ratio adjustment to the conventional Yates-Grundy variance estimator, $s_{YG}^2(\hat{Y})$, of any model unbiased asymptotically design consistent estimator $\hat{Y}$. His variance estimator

$$\tilde{s}^2_{YG}(\hat{Y}) = s^2_{YG}(\hat{Y})[V_m(\hat{Y} - Y)/E_m s^2_{YG}(\hat{Y})] \tag{2.10}$$

is model unbiased and at the same time asymptotically design consistent. However, for estimators of the form (2.8) Särndal *et al.* variance estimator appears simpler since it is obtained simply from the conventional variance estimator $s^2_{YG}(\hat{Y})$ by changing $\tilde{e}_i$ to $g_{si}\tilde{e}_i$.

The conventional regression estimator is obtained by first considering a fixed constant B in place of $\hat{\beta}$ in (2.7), and then substituting a consistent estimator of B_{opt}, the value of B minimizing the design variance. This estimator does not depend on the validity of any model. However, the optimal design variance can be approximately attained in the model- assisted framework by modifying the model (2.4) to $E(y_i) = \beta x_i + \gamma\pi_i$ and then using $(\tilde{\beta},\tilde{\gamma})'$, the weighted regression estimator of $(\beta,\gamma)'$ with weights $w_i = 1/\pi_i^2$. The resulting estimator of Y reduces to (2.7) with $\hat{\beta}$ changed to $\tilde{\beta}$ (Isaki and Fuller 1982; Montanari 1987). Any other choice of $\hat{\beta}$ in (2.7) will give a larger asymptotic design variance.

Little (1983) argued that only models that yield asymptotically design consistent, best linear model unbiased estimators should be used since the latter estimators are optimal if the model is in fact true. One way to accomplish this is by introducing an additional auxiliary variable $u_i = \sigma_i^2(1 - \pi_i)/\pi_i$ into the model (2.4), *i.e.* by using $E(y_i) = \beta x_i + \gamma u_i$ (Särndal and Wright 1984). If we change the model to $E(y_i) = \beta x_i + \gamma\sigma_i^2/\pi_i + \delta\sigma_i^2$ by adding two auxiliary variables σ_i^2/π_i and σ_i^2 to the model (2.4), then we get an asymptotically design consistent, best linear model unbiased estimator of the form $\hat{Y} = \sum_{i\epsilon s} g_{si}y_i/\pi_i$ (Särndal and Wright 1984). The lower bound to asymptotic anticipated variance is also attained if we choose a sampling plan with π_i proportional to σ_i. The above desirable properties, however, are obtained at the expense of a slight increase in the model variance under the original model (2.4).

Godambe and Thompson (1986) employed the theory of estimating functions to derive design consistent estimators through an assumed model. For example, if y_i is expected to be unrelated to π_i for some character y in a multisubject survey, then the "optimal" estimating function gives the Hájek (1971) estimator of $\bar{Y}$:

$$\hat{\bar{Y}}_H = \left(\sum_{i\epsilon s} y_i/\pi_i\right)\Big/\left(\sum_{i\epsilon s} 1/\pi_i\right). \tag{2.11}$$

The superpopulation model here is given by $y_i = \theta + \epsilon_i$, with independent errors ϵ_i, which reflects the situation at hand. The estimator $\hat{\bar{Y}}_H$ avoids the difficulties associated with the Horvitz-Thompson estimator $\hat{Y}_{HT}/N$, as illustrated by the "elephants" example of Basu (1971). The method of estimating functions looks promising, but further work remains to be done on its use in getting "better" estimators or pivotals or both. It is interesting to note that the well-known Fieller method of computing confidence limits for a ratio (Fieller 1932) and the method of Woodruff (1952) for computing confidence limits for medians are essentially equivalent to the method of estimating functions.

The results in Sections 2.2 and 2.3 use models appropriate to unistage sampling. In the case of multistage sampling, the models are more complex due to intra-cluster correlations (Scott and Smith 1969; Montanari 1987). The resulting best linear model unbiased estimators or prediction estimators involve weighted combinations of estimators, where the weights depend on intra-cluster correlations which can be estimated from the sample data. Bellhouse and Rao (1986) investigated the relative efficiency of such estimators, under the repeated sampling framework. Their empirical results suggest that the prediction estimators may not be significantly more efficient than the customary estimator in two-stage sampling with PPS sampling of clusters and simple random sampling within sampled clusters.

If the clusters are regarded as strata and if the strata means are the parameters of interest as in small area estimation, then the prediction estimators of strata means are likely to be significantly more efficient than the customary design-based estimators since the prediction estimators "borrow strength" from all the strata unlike the customary estimators. In the case of two-stage sampling with cluster means as parameters of interest, only a prediction estimator for the nonsampled clusters can be implemented.

3. VARIANCE ESTIMATION AND CONFIDENCE INTERVALS

3.1 Linear Statistics

A substantial part of traditional sampling theory is devoted to the derivation of mean square errors or variances of linear estimators of a total Y, and their estimators. Rao (1979) developed a unified approach for estimators belonging to Godambe's general linear class, $\hat{Y}_b = \sum_{i \in s} b_{is} y_i$, which enables the derivation of mean square error in a straightforward fashion, and also exhibits the necessary form of any non-negative quadratic unbiased estimator of the mean square error. For multistage designs, a general estimator of Y is of the form $\hat{Y}_{bm} = \sum_{i \in s} b_{is} \hat{Y}_i$, where s now denotes a sample of primary sampling units (psu's) and $\hat{Y}_i$ is an unbiased linear estimator of psu total Y_i based on subsampling the psu. Unified variance formulae for multistage designs have been worked out by Raj (1966) and Rao (1975).

Large scale surveys often employ many strata, L, with relatively few psu's n_h, sampled within each stratum h. In fact, it is a common practice to select $n_h = 2$ psu's within each stratum to permit maximum degree of stratification of psu's consistent with the provision of a valid variance estimator. If the psu's are sampled with replacement with probabilities p_{hi} in stratum h, then the estimator of total Y is given by $\hat{Y} = \sum_h \bar{r}_h$, and an unbiased variance estimator is simply obtained as

$$s^2(\hat{Y}) = \sum_h \left\{ \sum_i (r_{hi} - \bar{r}_h)^2 \Big/ [n_h(n_h - 1)] \right\}, \tag{3.1}$$

where $\bar{r}_h = \sum_i r_{hi}/n_h, r_{hi} = \hat{Y}_{hi}/p_{hi}$ and $\hat{Y}_{hi}$ is an unbiased estimator of the i-th psu total in stratum $h (i = 1, \ldots, n_h; h = 1, \ldots, L)$. This stratified design is frequently used in comparing methods for nonlinear statistics (Section 3.2). Because of its simplicity, $s^2(\hat{Y})$ is often used even when the psu's are sampled without replacement. This procedure leads to overestimation of variance, but the relative bias would be small if the first stage sampling fraction is small.

3.2 Non-linear Statistics

Many non-linear, finite population parameters of interest, θ, such as ratio, regression and correlation coefficients, can be expressed as smooth functions, $g(\mathbf{Y})$ of totals $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ of suitably defined variates such that $g(\mathbf{Y}) \propto g_1(Y_1/M, \ldots, Y_{q-1}/M)$, where $Y_q = M$, the population size. The parameter θ is estimated by $g(\hat{\mathbf{Y}}) \propto g_1(\hat{Y}_1/\hat{M}, \ldots, \hat{Y}_{q-1}/\hat{M})$. Such estimators are well-behaved even when the variates attached to the elements t are not related to the inclusion probabilities $\pi_t (t = 1, \ldots, M)$ since $g(\hat{\mathbf{Y}})$ is a function only of the Hájek-type estimators $\hat{\bar{Y}}_j = \hat{Y}_j/\hat{M}$ of the means $\bar{Y}_j$. As an example of $g(\hat{\mathbf{Y}})$, the estimator of a finite population regression coefficient $B = \sum (x_t - \bar{X})(y_t - \bar{Y}) / \sum (x_t - \bar{X})^2$ can be written as

$$\hat{B} = [\hat{Z}/\hat{M} - (\hat{X}/\hat{M})(\hat{Y}/\hat{M})] [\hat{W}/\hat{M} - (\hat{X}/\hat{M})^2]^{-1}, \tag{3.2}$$

where $\hat{X}$, $\hat{Z}$ and $\hat{W}$ are the estimators of the totals X, Z and W of the variates x_t, $z_t = y_t x_t$ and $w_t = x_t^2$ respectively.

Variance estimation methods for non-linear statistics, $g(\hat{\mathbf{Y}})$, include the well-known linearization method and resampling techniques like the jackknife, balanced repeated replication (BRR) and the bootstrap. The linearization method is applicable to general sampling designs, but it involves a separate variance formula for each statistic. On the other hand, resampling methods use a single variance formula for all statistics. The jackknife and BRR, however, are strictly applicable only to those designs in which the psu's are sampled with replacement (or the first-stage sampling fractions are negligible). The bootstrap seems to be more generally applicable, but it is computationally more cumbersome and its properties have not yet been fully examined.

Linearization method

If we denote the variance estimator of $\hat{Y} = \hat{Y}(y_t)$ for a general design as $v(y_t)$, the linearization method provides a variance estimator for a nonlinear statistic $\hat{\theta}$ as $v(z_t)$ for a suitably defined synthetic variable z_t which depends on the form of $\hat{\theta}$. For a general statistic $\hat{\theta} = g(\hat{\mathbf{Y}})$, the variance estimator is given by

$$s_L^2(\hat{\theta}) = v(z_t) \quad \text{with} \quad z_t = \sum_i y_{ti} g_i(\hat{\mathbf{Y}}), \tag{3.3}$$

(Woodruff 1971), where y_{ti} is the value of ith character for tth unit, and $g_i(\hat{\mathbf{Y}})$ is the partial derivative $\partial g(\mathbf{Y})/\partial Y_i$ evaluated at $\mathbf{Y} = \hat{\mathbf{Y}} (i = 1, \ldots, q)$. One drawback of the formula (3.3) is that the evaluation of partial derivatives may be difficult in some cases, although useful approximations to the desired partial derivatives can be obtained using numerical methods (Woodruff and Causey 1976). The variance estimator can also be obtained in many cases, without actually evaluating the partial derivatives g_i, by recasting $\hat{\theta}$ as a ratio-type statistic and using the usual variance formula for a ratio. For example, the sample regression coefficient $\hat{B}$ may be expressed as $\hat{B} = \hat{Y}(z_{1t})/\hat{Y}(z_{2t})$ with $z_{1t} = (y_t - \hat{\bar{Y}})(x_t - \hat{\bar{X}})$ and $z_{2t} = (x_t - \hat{\bar{X}})^2$, so that

$$s_L^2(\hat{B}) = v(z_{1t} - \hat{B} z_{2t})/[\hat{Y}(z_{2t})]^2. \tag{3.4}$$

Similar techniques can be used for other statistics like the multiple regression coefficients (Fuller 1975; Folsom 1974). Binder (1983) extended the scope of linearization method to statistics defined implicitly as the solution of a set of nonlinear equations. His formulation covers finite population parameters derived from generalized linear models which include the linear regression model and the logistic regression model.

Resampling methods

We now turn to resampling methods for the commonly used stratified multistage design of Section 3.1. Letting $\hat{\theta}^{hi}$ be the estimator of θ computed from the sample $\{\mathbf{r}_{hi}\}$ after omitting $\mathbf{r}_{hi} = \hat{\mathbf{Y}}_{hi}/p_{hi}$, a jackknife variance estimator of $\hat{\theta} = g(\sum \bar{\mathbf{r}}_h)$ is given by

$$s_J^2(\hat{\theta}) = \sum_h \{(n_h - 1)/n_h\} \sum_i (\hat{\theta}^{hi} - \hat{\theta})^2. \tag{3.5}$$

Several variations of (3.5) can be obtained; for instance, $\hat{\theta}$ in (3.5) may be replaced by $\hat{\theta}^h = \sum_i \hat{\theta}^{hi}/n_h$.

McCarthy (1969) proposed the BRR method for the important special case of $n_h = 2$. A set of J "balanced" half-samples is formed by deleting one psu in the sample from each stratum. This set may be constructed from Hadamard matrices. The BRR variance estimator is given by

$$s^2_{\text{BRR}}(\hat{\theta}) = \sum_j (\hat{\theta}^{(j)} - \hat{\theta})^2/J, \tag{3.6}$$

where $\hat{\theta}^{(j)}$ is the estimator computed from the j-th half sample. Again, several variations of (3.6) can be obtained. The BRR method has been extended recently to the general case of unequal n_h, using asymmetrical orthogonal arrays (Gupta and Nigam 1987; Wang and Wu 1988).

The bootstrap method for the stratified design involved the following steps (Rao and Wu 1988): (i) Draw a simple random sample $\{\mathbf{r}^*_{hi}\}_{i=1}^{m_h}$ of size m_h with replacement from $\{\mathbf{r}_{hi}\}_{i=1}^{n_h}$, independently for each h. Calculate

$$\bar{\mathbf{r}}_{hi} = \bar{\mathbf{r}}_h + [m_h/(n_h - 1)]^{1/2}(\mathbf{r}^*_{hi} - \bar{\mathbf{r}}_h), \bar{\mathbf{r}}_h = n_h^{-1} \sum_i \mathbf{r}_{hi}$$

and $\tilde{\theta} = g(\sum \bar{\mathbf{r}}_h)$. (ii) Independently replicate step (i) a large number, B, of times and calculate the corresponding estimators $\tilde{\theta}^1, \ldots, \tilde{\theta}^B$. (iii) The bootstrap variance estimator of $\hat{\theta}$ is given by

$$s^2_{\text{BOOT}}(\hat{\theta}) = \sum_b (\tilde{\theta}^b - \hat{\theta})^2/(B - 1). \tag{3.7}$$

Confidence intervals can also be obtained by approximating the distribution of $t = (\hat{\theta} - \theta)/s_J(\hat{\theta})$ by its bootstrap counterpart $\tilde{t} = (\tilde{\theta} - \hat{\theta})/s^*_J(\tilde{\theta})$, where $s^*_J(\tilde{\theta})$ is obtained from $s^2_J(\hat{\theta})$ by jackknifing the particular bootstrap sample $\{\mathbf{r}^*_{hi}\}$. Two-sided $1 - \alpha$ level "bootstrap-t" confidence intervals on θ are then given by

$$\{\hat{\theta} - \tilde{t}_{\text{UP}} s_J(\hat{\theta}), \hat{\theta} - \tilde{t}_{\text{LOW}} s_J(\hat{\theta})\}, \tag{3.8}$$

where $\tilde{t}_{\text{LOW}}$ and $\tilde{t}_{\text{UP}}$ are the lower and upper $\alpha/2$ points of $\tilde{t}$ obtained from the bootstrap histogram of $\tilde{t}^1, \ldots, \tilde{t}^B$. One-sided confidence intervals can also be obtained from the bootstrap histogram. Also, one could use the linearization variance estimator instead of the jackknife variance estimator in constructing the confidence intervals. For confidence intervals we need a much larger number, B, of bootstrap samples than for variance estimation. Regarding the choice of bootstrap sample sizes m_h, the choice $m_h = n_h - 1$ is attractive since it gives $\bar{\mathbf{r}}_{hi} = \mathbf{r}^*_{hi}$.

Comparison of the methods

Theoretical properties of the methods reported in the literature include the following: (1) All the variance estimators reduce to the "standard" one, $s^2(\hat{Y})$ given by (3.1), in the linear case $g(\mathbf{Y}) = Y$. (2) For smooth functions $g(\mathbf{Y})$, all the variance estimators are asymptotically design consistent (Krewski and Rao 1981). The jackknife variance estimator, however, is known to be inconsistent for nonsmooth functions like the quantiles, even in the case of simple random sampling. Hence, caution should be exercised in using jackknife software. (3) If $n_h = 2$ for all h, then the jackknife and linearization variance estimators are asymptotically equal to high order terms for smooth functions $g(\mathbf{Y})$, indicating that the choice between

these methods in this important special case should depend more on other considerations like computational costs (Rao and Wu 1985). Turning to empirical studies, Kish and Frankel (1974) studied the linearization, jackknife and BRR methods, using data from the Current Population Survey and sample designs with $n_h = 2$ clusters from each of $L = 6$, 12 and 30 strata. They evaluated the empirical coverage probability of the $1 - \alpha$ level confidence intervals, $\hat{\theta} \pm t_{\alpha/2} s(\hat{\theta})$, for ratios, regression and correlation coefficients, where $t_{\alpha/2}$ is the upper $\alpha/2$-point of a t-variable with L degrees of freedom and $s^2(\hat{\theta})$ is anyone of the variance estimators. The BRR method performed consistently better, in terms of coverage probability, than the jackknife which in turn was better than the linearization method; the observed differences were small for ratios. The methods performed in the reverse order with regard to stability of variance estimator. Other empirical studies in the literature reported similar results. Regarding the bootstrap, a simulation study by Kovar, Rao and Wu (1988) indicates that the bootstrap t-intervals track the nominal error rate in each tail better than the intervals based on the normal approximation to $t = (\hat{\theta} - \theta)/s(\hat{\theta})$, but the bootstrap variance estimators are less stable than those based on the linearization or the jackknife. The second order equivalence of the latter two variance estimators for the special case $n_h = 2$ is also confirmed.

Computationally simpler methods of variance estimation than the previous methods have also been proposed in the literature, *e.g.*, random group method and partially balanced repeated replication, but these variance estimators do not reduce to the "standard" one in the linear case. Methods of constructing models from which sampling errors can be imputed have also been proposed. Such methods are useful in producing "smoothed" standard errors for estimators for which direct computations have not been made, and also in presenting standard errors in a concise form (*e.g.*, graphs) in published reports.

Wolter's (1985) book gives an excellent introduction to recent developments in variance estimation, and illustrates the methods on data from a variety of large-scale surveys. Recent review papers on variance estimation include Rust (1985) and Rao (1988).

4. ANALYSIS OF SURVEY DATA

Standard methods of data analysis are, in general, based on the assumption of simple random sampling. These methods have also been implemented in standard statistical packages, including SPSS^X, BMDP and SAS. Application of standard methods to survey data without some adjustment for survey design, however, can lead to erroneous inferences, since most such data are obtained from complex sample surveys involving clustering, stratification and unequal probability sampling, and as a result do not satisfy the assumption of simple random sampling. In particular, standard errors of parameter estimates and associated confidence intervals can be seriously understated if the effect of design is ignored in the analysis of data. Similarly, the actual type I error rates of tests of hypotheses can be much bigger than the nominal levels. Standard exploratory data analyses, such as residual analysis to detect model deviations, are also affected. Kish and Frankel (1974) and others drew attention to some of these problems with standard methods and emphasized the need for new methods that take proper account of the complexity of survey data. During the past 10 years or so, rapid progress has been made in developing such methods for the following types of analyses: (a) analysis of multi-way contingency tables; (b) analysis of domain means or domain proportions; (c) linear regression analysis; (d) multivariate analysis including principal component analysis and factor analysis. A brief account of some of these developments is given in this section, and the reader is referred to review articles by Nathan (1988), Rao (1987) and Smith (1984), and a book edited by C.J. Skinner, D. Holt and T.M.F. Smith (1989).

4.1 Analysis of Multi-way Contingency Tables

Chi-squared tests (or likelihood ratio tests) are frequently used for the evaluation and selection of parsimonious models on $\mathbf{p}$, the population cell probabilities, in a multi-way contingency table with T cells. For this purpose, loglinear models are convenient because of their close similarity to analysis of variance in systematically providing test statistics of various hypotheses associated with a multi-way table. Rao and Scott (1984) made a systematic study of the impact of survey design on the standard chi-squared test of goodness-of-fit of a loglinear model, denoted by X^2. They showed that X^2 is asymptotically distributed as a weighted sum, $\sum \delta_i W_i$, of $T - r - 1$ independent χ^2_1 variables W_i, where the weights δ_i are the eigenvalues of a "generalized design effects" matrix and $T - r - 1$ is the degrees of freedom. This general result shows that the survey design can have a substantial impact on the type I error rate of X^2. For instance, under a constant design effects clustering model, $\delta_i = \lambda$ for all i, the actual type I error rate, for nominal level α, is approximately given by $Pr[\chi^2_{T-r-1} > \lambda^{-1}\chi^2_{T-r-1}(\alpha)]$ which increases with the clustering effect, λ.

Rao and Scott (1984,7) obtained simple first-order corrections to X^2 which can be computed from published tables that include estimates of design effects (or standard errors) for cell estimates $\hat{\mathbf{p}}$ and their marginal totals, thus facilitating secondary analyses (see also Fellegi 1980, Gross 1984, and Bedrick 1983). A first-order correction refers $X^2/\hat{\delta}.$ to χ^2_{T-r-1}, where $\hat{\delta}.$ is an estimate of the average design effect $\delta. = \sum \delta_i/(T - r - 1)$ or an estimate of an upper bound on $\delta.$. The corrected test is asymptotically valid in the case of constant design effects clustering, and in general it should perform well when the variability of the δ_i's is small. More accurate, second-order corrections that take account of the variability in the δ_i's can also be obtained by using the Satterthwaite approximation to the weighted sum of independent χ^2 variables (Rao and Scott 1984). These tests, however, require the knowledge of a full estimated covariance matrix of $\hat{\mathbf{p}}$. Alternative methods that take account of the survey design include the Wald statistics based on weighted least squares (Koch, Freeman and Freeman 1975) and the jackknife chi-squared tests (Fay 1985). The latter tests are applicable to survey designs permitting the use of a replication method, such as the jackknife or the BRR. The Wald tests require the full estimated covariance matrix of $\hat{\mathbf{p}}$, whereas the jackknife tests require access to cluster-level estimates.

Fay (1985) and Thomas and Rao (1987) showed that the Wald test which refers to χ^2_{T-r-1}, although asymptotically correct, can become highly unstable as the number of cells in the multi-way table increases and the number of sample clusters decreases, leading to unacceptably high type I error rates compared to the nominal level, α. On the other hand, Fay's jackknife tests and the Rao-Scott corrections performed well under quite general conditions. A simple modification to the Wald test which refers to an F distribution on $T - r - 1$ and $f - T + r + 2$ degrees of freedom performed better than the Wald test in controlling the type I error rate, where f is the degrees of freedom for estimating the covariance matrix of $\hat{\mathbf{p}}$.

4.2 Analysis of Domain Means or Domain Proportions

Analysis of domain (or subpopulation) proportions associated with a binary response variable is of considerable interest to researchers in social and health sciences, and other subject matter areas. Logistic regression models are extensively used for this purpose in conjunction with standard statistical methods for binomial proportions. Rao and Scott (1987) obtained simple first-order corrections to standard chi-squared tests of goodness-of-fit and of nested hypotheses which can be computed from published tables that include estimates of design effects (or standard errors) of domain proportions. Roberts, Rao and Kumar (1987) derived more

accurate second-order corrections to standard tests, but these require access to a full estimated covariance matrix of domain proportions. Diagnostics for detecting outlying domain proportions and influential points in the factor space were developed as well, again taking the sampling design into account.

Koch, Freeman and Freeman (1975) used weighted least squares methods to analyze domain means of a quantitative variable, y, and developed Wald tests of goodness-of-fit of the model and of linear hyptotheses on the model parameters. The performance of Wald tests can be improved, as in Section 4.1, by using an F-modification.

4.3 Linear Regression Analysis

In Section 3.2, we considered design-based inferences on nonlinear, finite population parameters such as the finite population simple regression coefficient B. The pivotal $t = (\hat{B} - B)/s(\hat{B})$ is approximately $N(0,1)$, where $\hat{B}$ is the design-consistent estimator, (3.2), of B, and its standard error, $s(\hat{B})$, can be obtained either through the linearization method as in (3.4) or by using one of the replication methods. This approach readily extends to multiple regression coefficients. The design-weighted estimator $\hat{B}$ or its multiple regression analogue can be obtained by the weighted regression option of standard packages by using the survey weights attached to the sample elements as the weights in the regression. However, the standard error of $\hat{B}$ resulting from this routine remains incorrect.

Some people argue that most users are concerned with inferences on parameters of an appropriate superpopulation model rather than inferences on finite population parameters like B. However, the interest in B can also be justified by considering it as the least squares estimator of the superpopulation parameter β in the model

$$y_i = \alpha + \beta x_i + \epsilon_i \text{ with } E_m(\epsilon_i) = 0, \quad i = 1, \ldots, N. \tag{4.1}$$

If the population size is large, then estimating B is effectively equivalent to estimating β, while if the model (4.1) is misspecified to the extent of making β meaningless, then B may still be of interest as the slope of the least squares line fitted to the N-pairs (y_i, x_i) (Godambe and Thompson 1986).

Scott and Holt (1982) used a model-dependent approach to investigate the effect of two-stage sampling on standard regression analysis. They assumed a regression model of the form (4.1) with equi-correlated error terms ϵ_i within each cluster, as in Fuller (1975). This model also holds for the sample pairs (y_i, x_i), $i \in s$, if the selection probabilities are not related to the dependent variable, as in the case of two-stage random sampling. The results of Scott and Holt indicate that the effect of a positive intra-cluster correlation is to understate the standard errors of parameter estimates, and consequently inflate the type I error rates of customary tests. Wu, Holt and Holmes (1988) made a systematic study of the effect of two-stage sampling on the customary F-statistic, and proposed a correction for the F test for unknown intra-cluster correlation, as an alternative to iterative generalized least squares (GLS) procedure. Both the GLS procedure and the F-correction require known cluster labels which may not be available when the survey data are used for secondary analysis.

If the regression model includes all the design variables z related to the dependent variable, such as stratum indicator variables and size measures of units, and the errors ϵ_i are independent with a constant variance σ^2, then standard regression analysis is valid under the model-dependent approach (Pfefferman and Smith 1985). However, such models may involve too many parameters to be useful. Also, the design variables may not be of intrinsic interest to the user, or may not be available in secondary analysis. In such situations, we are often interested

in models of the form (4.1), where x is not a design variable. The sample pairs (y_i, x_i), $i \in s$ however, may not satisfy the model due to sample selection bias. Nathan and Holt (1980) proposed an adjusted regression approach to take account of selection bias, and compared it with ordinary least squares and the design based approach based on $\hat{B}$ and $s(\hat{B})$. This approach assumes specific relationships between the regression variables and the design variables. Their empirical results indicate that ordinary least squares inferences can be highly unreliable, that the design-based approach is basically reliable except under extreme selection schemes, and that the adjusted regression approach performs well. Pfefferman and Holmes (1985) study the robustness of these procedures to misspecification of relationships between the regression variables, and conclude that the adjusted regression approach is very sensitive to model misspecification. The design-weighted estimator $\hat{B}$ is robust, but a more efficient estimator is obtained by modifying the adjusted regression estimator to be design-consistent for the finite population regression coefficient, B.

4.4 Multivariate Analysis

The methods in Section 4.2 for the analysis of domain means can be extended to the multivariate case of domain mean vectors, but no detailed studies of such extensions have been reported in the literature. The literature on multivariate anlaysis of survey data is largely devoted to the analysis of covariance structures, in particular to principal component analysis and factor analysis. Bebbington and Smith (1977), Tortora (1980) and Skinner, Holmes and Smith (1986) investigated the effect of sample design on standard principal component analysis. Their results indicate that the application of standard methods, without some adjustment for the sample design, can lead to erroneous inferences. In particular, the estimators of eigenvalues and eigenvectors of the covariance matrix, $\sum_y$, can be severely biased for non-self-weighting sample designs. Skinner, Holmes and Smith (1986) proposed maximum likelihood (ML) estimators, under a multivariate normal model, and probability-weighted (or design-based) estimators, to adjust for the effects of the sample design. Their simulation study indicates that both estimators perform well unconditionally, while the probability-weighted estimators exhibit a conditional model bias. The ML estimators, however, may be sensitive to model misspecification. A probability-weighted version of the ML estimators may be more robust, as demonstrated by Pfefferman and Holmes (1985) in the context of the adjusted regression approach (section 4.3). Fuller (1987) derived design-based estimators of the parameters in factor analysis, and the estimated covariance matrix of the estimators. He showed that the estimated variances based on normal theory can seriously underestimate the true variances of the factor estimators.

5. COMPUTER SOFTWARE

Several computer package programs for variance estimation in complex surveys were developed in the mid to late 1970's, often in conjunction with programs for regression analysis of survey data. Wolter (1985, pp. 393-412) reviewed the latest versions of these programs to about 1985. Among the programs listed by Wolter, the ones most commonly used are CLUSTERS (Verma and Pearce 1977), the programs &PSALMS and &REPERR in the OSIRIS IV system (Vinter 1980 and Lepkowski 1982), SUDAAN (Shah 1981a, 1981b, 1982 and Holt 1979), HESBRR (Jones 1983) and SUPER CARP (Hidiroglou, Fuller and Hickman 1980). The programs HESBRR and the OSIRIS IV program &REPERR use balanced repeated replication as the variance estimation technique; the remaining three use the Taylor linearization method.

Cohen, Burt and Jones (1986) evaluated the variance estimation programs for means and ratios, with the exception of CLUSTERS, using a large data set from the National Medical Care Expenditure Survey. They found that the programs SESUDAAN and RATIOEST in the SUDAAN collection were the most efficient in terms of CPU time usage and easier to program than the others.

One major current trend in software development is the development of menu-driven packages on micro-computers. Variance estimation and specialized survey analysis software is no exception to this trend. A notable enhancement to the commonly used variance estimation programs since 1985 is the introduction of PC CARP (Schnell *et al.* 1986 and Schnell *et al.* 1988), available on IBM AT/XT or compatible micro-computers with a math co-processor. This package, like its predecessor SUPER CARP, uses Taylor linearization methods for variance estimation. A second variance estimation package is also available on micro-computers. The package listed as BELLHOUSE in Wolter (1985, p. 399) has been adapted for IBM micros with or without a co-processor by Rylett and Bellhouse (1988) under the program name TREES. This software uses tree structures to mimic the structure of stratified multistage sampling designs and applies tree traversal algorithms, in conjunction with general results on variance estimation in multi-stage sampling (see section 3.1), to the calculation of variance estimates.

A second trend in the computer implementation of survey variance estimation and survey analysis techniques is the integration of survey software with widely used statistical analysis systems. A leader in this trend from the early 1980's is the SUDAAN system, which is comprised of a series of several SAS procedures. Freeman *et al.* (1985) and Hidiroglou and Paton (1987) both used the PROC MATRIX procedure in SAS to obtain survey variance estimates, the former by balanced repeated replication and the latter by Taylor linearization. Mohadjer *et al.* (1986) report the development of a new SAS procedure WESVAR to obtain survey variance estimates by balanced repeated replication.

A variety of packages and computing techniques are available to carry out the analyses of survey data reviewed in Section 4. Among the available specialized packages, the most comprehensive appears to be the PC CARP. The original program, SUPER CARP, was designed to carry out regression analyses developed by Fuller (1975); the PC version retains this option. The current version now contains additional options for categorical data analysis, and inferences on cumulative distribution function and associated quantiles, following methods given by Francisco and Fuller (1986). For categorical data, there is an option for the analysis of two-way contingency tables, based on the Rao-Scott corrections to chi-squared test of independence. The program can also be manipulated to perform factor analyses of survey data.

There are four other specialized packages for the analysis of survey data; between them they cover topics in regression and categorical data analysis. The &REPERR program in OSIRIS IV and the SURREGR procedure in SUDAAN both calculate standard errors of regression coefficients so that regression analyses can be carried out. The programs CPLX, developed by Fay (1982), and RSPLX, also by Fay, handle categorical data analyses of log-linear models for two and multi-way tables. The analysis in CPLX is carried out using jack-knifed chi-square statistics, while RSPLX applies second order Rao-Scott corrections to the usual test statistic.

The four programs for the regression analyses for complex survey data were evaluated by Cohen, Xanthopoulis and Jones (1988). The older version, SUPER CARP, was included in this analysis rather than PC CARP. Similar to the earlier study of Cohen, Burt and Jones (1986) on variance estimation, data from the National Medical Care Expenditure Survey were used. Once again, a program in the SUDAAN suite of programs, SURREGR, was the most

efficient in terms of CPU time usage and easier to program than the others. However, the efficiency of the SUDAAN programs might be balanced by the flexibility of the PC CARP program, depending upon the survey analysis required.

Significant enhancements to SUDAAN are provided in the new SUDAAN system under development (LaVange *et al.* 1989). Variance estimation and data analysis methods not available in SUDAAN are among the many modifications incorporated into the new SUDAAN System.

Running almost parallel to the emerging trend in the calculation of variance estimates, there is pagea move towards incorporating methods for the analysis of complex survey data into standard statistical packages and systems. Following on their variance estimation methods using SAS procedures, Hidiroglou and Paton (1987) describe further SAS procedures to carry out log-linear analyses, with Rao-Scott corrections, of multi-way contingency tables. Likewise, Freeman (1988) notes that he used the SAS procedure PROC MATRIX for both variance estimation and for the analysis of variance of his survey data. Similarly, Mahodjer *et al.* (1986) describe two other new SAS procedures in addition to the variance estimation procedure WESVAR. These are the previously mentioned NASSREG and NASSLOG which carry out weighted least squares regression analyses and logistic regression analyses respectively. Both procedures depend on balanced repeated replication for variance estimation of the model parameters. An alternative approach to using SAS procedures is to use the matrix algebra language GAUSS (Platt 1986). Based on their own experience, Rao and Thomas (1988) favorably report on the use of this language for categorical data analysis in complex surveys.

6. CONCLUDING REMARKS

The early milestones in the development of efficient sampling designs and associated estimation techniques for population totals and means have firmly established sample survey theory and methods as a major discipline in statistics. Subsequent developments in theoretical foundations of sampling theory have provided useful insights into inferential aspects. In particular, the model-assisted approach and the conditional design-based approach appear to be promising since they attempt to fill the "gap" between the traditional approach and the model-dependent approach by retaining the desirable features of both approaches, but more research is needed in this area to handle complex sampling designs. Recent advances in variance estimation and confidence intervals for nonlinear statistics and the associated computer software, are also equally impressive. It is also gratifying that rapid progress has been made in the development of methods for the analysis of survey data that take account of the complexity of the sampling design, and the associated computer software.

We can expect to see important new developments in the next 10 years or so in the areas of variance estimation for nonlinear statistics (especially, nonsmooth functions), analysis of survey data (especially, multivariate analysis), and other topics not covered here (especially, sampling in time and small area estimation).

ACKNOWLEDGEMENTS

The authors would like to thank the editor for helpful comments. This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

BASU, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, (Eds. V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

BEBBINGTON, A.C., and SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis. In *The Analysis of Survey Data* (Eds. C.A. O'Muircheartaigh and C.D. Payne), Vol. 2, New York: Wiley, 175-192.

BEDRICK, E.J. (1983). Adjusted goodness-of-fit tests for survey data. *Biometrika*, 70, 591-595.

BELLHOUSE, D.R. (1988). A brief history of random sampling methods. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 1-14.

BELLHOUSE, D.R., and RAO, J.N.K. (1986). On the efficiency of prediction estimators in two-stage sampling. *Journal of Statistical Planning and Inference*, 13, 269-281.

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

BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv.1, 6-62.

BREWER, K.R.W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

BREWER, K.R.W., and HANIF, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1976). *Foundations of Inference in Survey Sampling*. New York: Wiley.

CHAUDHURI, A. (1988). Optimality of sampling strategies. In *Handbook of Statistics* (Eds. P.R. Krishmaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 47-96.

CHAUDHURI, A., and VOS, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. Amsterdam: North-Holland.

COCHRAN, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.

COCHRAN, W.G. (1946). Relative accuracy of systematic and stratified samples for a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

COHEN, S.B., BURT, V.L., and JONES, G.K. (1986). Efficiencies in variance estimation for complex survey data. *American Statistician*, 40, 157-164.

COHEN, S.B., XANTHOPOULIS, J.A., and JONES, G.K. (1988). An evaluation of statistical software procedures appropriate for the regression analysis of complex survey data. *Journal of Official Statistics*, 4, 17-34.

DALENIUS, T. (1957). *Sampling in Sweden*. Stockholm: Almquist and Wiksell.

DEMING, W.E. (1960). *Sample Design in Business Research*. New York: Wiley.

DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling* (Eds. N.L. Johnson and H. Smith), New York: Wiley-Interscience, 629-651.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, Series B, 31, 195-224.

FAY, R.E. (1982). Contingency tables for complex designs, CPLX. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 44-53.

FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

FELLEGI, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75, 261-268.

FIELLER, E.C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, 24, 428-440.

FISHER, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd 5th edition 1934.

FOLSOM, R.E. (1974). National assessment approach to sampling error estimation, sampling error monograph. National Assessment of Educational Progress, first draft.

FRANCISCO, C.A., and FULLER, W.A. (1986). Estimation of the distribution function with a complex survey. Technical Report, Iowe State University.

FREEMAN, D.H. (1988). Sample survey analysis: analysis of variance and contingency tables. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 415-426.

FREEMAN, D.H., LIVINGSTON, M., LEO, L., and LEAF, P. (1985). A comparison of indirect variance estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 313-316.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.

FULLER, W.A. (1987). Estimators of the factor model for survey data. In *Applied Probability, Statistics and Sampling Theory* (Eds. I.B. MacNeill and G.J. Umphrey), Boston: D. Reidel Publishing Company, 265-284.

GHOSH, M. (1987). On admissibility and uniform admissibility in finite population sampling. In *Applied Probability, Stochastic Processes and Sampling Theory*, (Eds. I.B. MacNeil and G.J. Umphrey), Boston: D. Reidel Publishing Company, 197-213.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, series B, 17, 269-278.

GODAMBE, V.P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society*, series B, 28, 310-328.

GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54,127-138.

GROSS, W.F. (1984). A note on chi-squared tests with survey data. *Journal of the Royal Statistical Society*, series B, 46, 270-272.

GUPTA, V.K., and NIGAM, A.K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika*, 74, 735-742.

HÁJEK, J. (1981). *Sampling From a Finite Population*. New York: Marcel Dekker.

HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sampling Survey Methods and Theory*, Vol. 1. New York: Wiley.

HANSEN, M.H., MADOW, W.G., and TEPPING, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

HANSEN, M.H., DALENIUS, T., and TEPPING, B.J. (1985). The development of sample surveys of finite populations. In *A Celebration of Statistics*: The ISI Centenary Volume (Eds. A.C. Atkinson and S.E. Fienberg), New York: Springer Verlag, 327-354.

HARTLEY, H.O., and RAO, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R. (1980). *SUPERCARP-Sixth Edition*. Survey Section, Ames, Iowa.

HIDIROGLOU, M.A., and PATON, D.J. (1987). Some experiences in computing estimates and their variances using data from complex survey designs. In *Applied Probability, Statistics and Sampling Theory* (Eds. I.B. MacNeill and G.J. Umphrey), Boston: D. Reidel Publishing Company, 285-308.

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

HOLT, M.M. (1979). SURREGR: standard errors of regression coefficients from sampling survey data. Research Triangle Institute, Research Triangle Park, North Carolina.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

ISAKI, C.T. and FULLER, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.

JONES, G.K. (1983). HESBRR (HES variance and crosstabulation program). Version 3, Internal NCHS Report, Hyattsville, Maryland.

KIAER, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistics of Norway.

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

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of Royal Statistical Society*, series B, 36, 1-37.

KOCH, G.G., FREEMAN, D.H., and FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.

KOVAR, J., RAO, J.N.K., and WU, C.F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, Supplement, 25-45.

KOTT, P.S. (1987). Estimating the conditional variance of a design consistent regression estimator. Technical Report.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *Annals Statistics*, 9, 1010-1019.

LAVANGE, L.M., SHAH, B.V., BARNWELL, B.G., and KILLINGER, J.F. (1989). SUDAAN: A comprehensive package for survey data analysis. Technical Report, Research Triangle Institute.

LEPKOWSKI, J.M. (1982). The use of OSIRIS IV to analyse complex sample survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 38-43.

LITTLE, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.

MADOW, W.G. (1978). Comments on papers by Basu and Royall and Cumberland. In *Survey Sampling and Measurement* (Ed. N.K. Namboodiri). New York: Academic Press, 315-322.

MADOW, W.G., and MADOW, L.H. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.

MAHALANOBIS, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-370.

McCARTHY, P.J. (1969). Pseudo-replication: half-samples. *International Statistical Review*, 37, 239-264.

MOHADJER, L., MORGANSTEIN, D., CHU, A., and RHOADS, M. (1986). Estimation and analysis of survey data using SAS procedures WESVAR, NASSREG, and NASSLOG. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 258-263.

MONTANARI, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

NARAIN, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

NATHAN, G. (1988). Inference based on data from complex sample designs. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland.

NATHAN, G., and HOLT, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, series B, 42, 377-386.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, series B, 12, 241-255.

PFEFFERMAN, D., and HOLMES, D.J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, series A, 148, 268-278.

PFEFFERMAN, D., and SMITH, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.

PLATT, W.G. (1986). GAUSS. *American Statistician*, 40, 164-169.

RAJ, D. (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.

RAO, J.N.K. (1971). Some thoughts on the foundations of survey sampling. *Journal of the Indian Society of Agricultural Statistics*, 23, 69-82.

RAO, J.N.K. (1979). On deriving mean square errors and their non-negative unbiased estimators. *Journal of the Indian Statistical Association*, 17, 125-136.

RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.

RAO, J.N.K. (1987). Analysis of categorical data from sample surveys. In *New Perspectives in Theoretical and Applied Statistics* (Eds. M.L. Puri, J.P. Vilaplana and W. Wertz). New York: Wiley, 45-60.

RAO, J.N.K. (1988). Variance estimation in sample surveys. In *Handbook of Statistics* (Eds. P.R. Krishnaiah and C.R. Rao), Vol. 6, Amsterdam: North-Holland, 427-447.

RAO, J.N.K., and SINGH, M.P. (1973). On the choice of estimator in survey sampling. *Australian Journal of Statistics*, 15,95-104.

RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RAO, J.N.K., and SCOTT, A.J. (1987). On simple adjustments to chi-square tests with sample survey data. *Annals of Statistics*, 15, 385-397.

RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association*, 83, 231-241.

RAO, J.N.K., and THOMAS D.R. (1988). The analysis of cross-classified categorical data from sample surveys. *Sociology Methodology*, 18, 213-269.

ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

ROBINSON, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association*, 82, 826-831.

ROYALL, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57,377-387.

ROYALL, R.M., and HERSON, J.H. (1973). Robust estimation in finite populations, I and II. *Journal of the American Statistical Association*, 68, 880-889 and 890-893.

ROYALL, R.M., and EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā*, series C, 37, 43-52.

ROYALL, R.M., and PFEFFERMAN, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika*, 69, 401-410.

ROYALL, R.M., and CUMBERLAND, W.G. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, series B, 50, 118-124.

RUST, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 4, 381-397.

RYLETT, D.T., and BELLHOUSE, D.R. (1988). TREES: a computer program for complex surveys. *Proceedings of the Section on Survey Research Methods*. American Statistical Association,694-697.

SÄRNDAL, C.E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.E., and WRIGHT, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

SÄRNDAL, C.E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted regression technique for estimating the variance of the generalized regression estimator. *Biometrika*, 76, 527-537.

SCHNELL, D., SULLIVAN, G., KENNEDY, W.J., and FULLER, W.A. (1986). PC CARP: Variance estimation for complex surveys. In *Computer Science and Statistics: Proceedings of the 17th Symposium of the Interface* (Ed. D.M. Allen). Amsterdam: North Holland, 125-129.

SCHNELL, D., KENNEDY, W.J., SULLIVAN, G., PARK, H.J., and FULLER, W.A. (1988). Personal computer variance software for complex surveys. *Survey Methodology*, 14, 59-69.

SCOTT, A.J., and SMITH, T.M.F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64, 830-840.

SCOTT, A.J., and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

SHAH, B.V. (1981a). SESUDAAN: Standard errors program for computing of standardized rates from sample survey data. Research Triangle Institute, Research Triangle Park, North Carolina.

SHAH, B.V. (1981b)., RATIOEST: Standard errors program for computing ratio estimates for sample survey data. Research Triangle Institute, Research Triangle Park, North Carolina.

SHAH, B.V. (1982). RTIFREQS: Program to compute weighted frequencies, percentages and their standard errors. Research Triangle Institute, Research Triangle Park, North Carolina.

SKINNER, C.J., HOLMES, D.J., and SMITH, T.M.F. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association*, 81, 789-798.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.

SMITH, T.M.F. (1984). Present position and potential developments: some personal views - sample surveys. *Journal of the Royal Statistical Society*, series A, 147, 208-221.

THOMAS, D.R., and RAO, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

TORTORA, R.D. (1980). The effect of disproportionate stratified design on principal component analysis used for variable elimination. *Proceedings of the Survey Research Section*, American Statistical Association, 746-750.

VERMA, V., and PEARCE, M.(1977). Users manual for CLUSTERS: A sampling program for computation of sampling errors for clustered samples. Technical Report No. 568, World Fertility Survey, U.K.

VINTER, S. (1980). Survey sampling errors with OSIRIS IV. COMPSTAT *1980: Proceedings in Computational Statistics*, Vienna: Physica-Verlag, 72-80.

WANG, J.C., and WU, C.F.J. (1988). An approach to the construction of asymmetrical orthogonal arrays. Technical Report, University of Waterloo.

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

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures, *Journal of the American Statistical Association*, 47, 635-646.

WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

WOODRUFF, R.S., and CAUSEY, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

WU, C.F.J., HOLT, D., and HOLMES, D.J. (1988). The effect of two-stage sampling on the *F* statistic. *Journal of the American Statistical Association*, 83, 150-159.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, Third Edition, London: Griffin.

COMMENT

T.M. FRED SMITH[1]

Sample surveys are one of the most important areas of the application of statistics. The paper by Professors Rao and Bellhouse is an excellent review of the theoretical development of sample surveys and I find it hard to be critical; but in the best traditions of the Royal Statistical Society I shall make the attempt in as constructive and a controversial manner as possible. In any review paper the choice of topics, especially relating to recent work, must be to some extent subjective. This affords a discussant an easy target; criticize the authors for their sins of omission. Also a review must be wide ranging and this allows discussants freedom to ride their own hobby horses over the range. I shall adopt both approaches and my objective in so doing is to identify some additional issues which I believe are important thus widening the review still further.

There is now general agreement about the milestones of our subject. These are associated with the names of Kiaer, Bowley, Neyman, Cochran, Hansen, Hurwitz, Madow, Mahalanobis, Horvitz and Thompson – an international collection dominated, latterly by contributions from the USA. Kiaer and Bowley's work was fundamental because they demonstrated that valid conclusions could be drawn from representative samples of quite small size drawn from large populations with arbitrary values. Representative samples were stratified samples with proportional allocation, and Bowley derived the appropriate theoretical results. Neyman and subsequent authors argued the case for random sampling and developed a comprehensive theory of randomisation inference applicable to most sampling schemes. Durbin (1953) completes the theory with his multi-stage sampling results. Despite the importance of these results sample surveys became a Cinderella subject on the fringes of mainstream statistics, and even today most university departments do not have a sampling statistician on their staff. Why is this?

One reason is that sample survey theory has developed mainly within social science and official government statistics, whereas most statisticians have a training within mathematics and physical science. Although all experimental scientists deal with samples very few seem to recognise this explicitly and those that do, such as geologists and biologists, have developed their own theory of sampling and estimation. In my view it is time to bring together sampling experts from all areas of scientific enquiry to share ideas and experiences and hopefully to establish a global theory of sample surveys.

A second reason is that sample surveys starts with a population which is a real fixed finite population of units. Samples are then drawn from this population according to specific rules. In most scientific enquiries the position is reversed; the population is not well defined and the scientist starts with a sample. One view of the role of the statistician, as enunciated, for example, R.A. Fisher, is to define the hypothetical population from which the sample data can be viewed as a random sample. This approach begs the question whether this hypothetical population has any scientific value. Arguably the sample survey approach of starting with the population has much to commend it.

A third reason is that since the finite population units can take arbitrary values the population cannot be summarized by a few parameters. Notions like sufficiency have little value in sample survey theory, and sample data are usually summarized by a mass of cross-tabulations. The estimation of a large number of cell proportions is the primary aim of sample surveys and the object of inference is usually descriptive rather than explanatory.

A final reason for the separation of sample surveys from mainstream statistics is that the randomisation theory of sample surveys is so complete. It is a closed theory which if accepted

[1] T.M.F. Smith, Department of Mathematics, The University, Southampton, S09 5NH, U.K.

has few remaining problems to be solved. The chief concerns of randomisation researchers since Horvitz and Thompson (1952) provided the general theoretical framework have been the construction of πps sampling schemes with non-zero joint inclusion probabilities, the production of methods and programs for variance estimation and the construction of estimators which employ auxiliary information but can never be generally efficient because of Godambe's result. All of these problems are important, but they are not exciting, they lack the philosophical and mathematical depth to capture the imaginations of young mathematical statisticians.

These reasons are my explanation why sample surveys have been seen in the past as an activity on the fringe of mainstream statistics. The position is changing now and I detect a coming together of the branches of statistics. Much recent work in sample surveys has attempted to integrate surveys into mainstream statistics and many areas of statistics now recognise the importance of selection effects. Has the sample survey Cinderella been invited to the Statisticians's Ball?

In addition to his non-existence theorem Godambe has also shown that within the randomisation framework the likelihood is proportional to the probability of selection, $p(s \mid z)$, where z is the prior information on which the design was based, which for fixed s is a constant. Thus the likelihood is completely uninformative. In the same set-up Basu (1971) showed that the sufficient statistic is $\{(i,y_i):i\epsilon s\}$, namely the complete data tape including the labels. Although these results are also negative, highlighting the distinction between randomisation inference and other forms of inference, they did stimulate interest amongst a wider group of statistician and so had a positive value. My own interest in the theory of sample surveys was stimulated by Ericson (1969), in particular by the way he incorporated the uninformative likelihood into a positive framework via Bayes theorem and exchangeable priors. Ericson's use of exchangeability deserves consideration by all statisticians, not just Bayesians. Is it reasonable, is it even possible, to have a valid theory of predictive inference without some form of exchangeability? If there is no function of the unit values which is exchangeable how can you predict the unobserved values from the sample values? My opinion is that Ericson's work was a milestone in the development of sample survey theory.

The uninformative nature of the randomisation likelihood led some statisticians to question the role of randomisation. Godambe himself refers to "the problem of randomisation" and developed alternative theoretical approaches which required randomisation. Ericson also found a role for randomisation within his exchangeable set-up. He argued that if you employ your prior information, z, to form groups of units which are approximately exchangeable a priori then the use of simple random sampling will guarantee exchangeability. Royall (1970, 1973), however, made the mistake of advocating purposive sampling within his model-based framework. He touched a raw nerve and brought down upon his head the wrath of the randomisation establishment. I thought that Royall had asked some serious questions which deserved an answer and the strength of the reaction surprised me. Why did academic survey samplers and those from government agencies in North America feel so strongly about randomisation? Their colleagues in market research seemed happy with quota samples which could be viewed as a special case of balanced sampling. In Europe many official surveys are based on quota samples. What is so special about official statistics in North America?

I think the answer lies deep in the American political psyche. Thoughtful Americans are democratic in the true sense of that term. They believe in individual freedom and the right to information, they are also deeply suspicious of governments. They recognise the need within a democracy for reliable statistical information. To the official statisticians randomisation is the guarantee of the objective reliability of their data. It is a key source of their professional integrity and any attack on randomisation was seen as potentially dangerous however well

intentioned. I admire this position and it has helped to convince me that randomisation is one of the great contributions of statistics to science.

I have expressed myself with some feeling because I am so unhappy about the present position of official statistics in the U.K.. The tradition in the U.K. is not naturally democratic, we are still a monarchy, we respect authority rather than the individual. This tendency is being exploited and there is now a serious erosion of public confidence in the Government's use of statistics. It has been argued that official statistics in the U.K. are collected to aid the decisions of government, not to help parliament or to inform the electorate. Key series have been stopped, definitions have been changed, information is presented by ministers in ways which are patently false, yet no government statistician can complain publically because of the Official Secrets Act. There is a dangerous public cynicism about statistics and George Orwell's predictions in his novel 1984 may be closer to the truth than we realise. I apologise to the authors for this digression, but I said I would ride some hobby horses, and the issue of the integrity of official statistics is of great importance.

Before leaving randomisation theory I would like to make some comments about repeated surveys and rotation sampling. Again this is an area which the authors have excluded although they did note Patterson (1950) as a milestone paper. Randomisation theory has been developed within the framework of the one-off cross section survey. The extension to repeated surveys is non-trivial for it is difficult to retain the probability structure over time under rotation sampling when the population changes, Fellegi (1963). For the measurement of gross flows, or transition probabilities, the role of the randomisation inclusion probabilities is not clear. The beautiful simplicity of randomisation theory for one-off surveys is destroyed when they are repeated over time. But most important surveys are repeated surveys, especially in the government sector, so what are the implications?

As always the answer is that it depends. If the primary purpose is to produce descriptive statistics of the state of the system at each time period then the surveys can be considered as repetitions of a cross-section survey and each one can be analysed independently. Although composite estimators or time series estimators may be more efficient they should be viewed as secondary estimators rather than primary estimators. If I wanted to use repeated survey data within an econometric model I would prefer to input the cross-section estimates with their known correlation structure rather than complex composite estimates. On the other hand if I wanted the best estimate of the current value of, say, unemployment, for a particular purpose, not for public consumption, then I would use the most efficient procedure available. Similarly if I wanted to explain the change in value of some estimates over time then I would need to go beyond simple randomisation analysis. Thus the problems with randomisation inference for repeated surveys occur mainly for secondary analyses. However, there remains the important issue of which estimates should be reported to the public.

Section 2 of the paper is devoted to work on the theoretical foundations of inference from survey data carried out during the last 30-40 years. The authors have chosen to distinguish three approaches, design-based, model-dependent and model-assisted, the latter being an attempt to find a compromise solution between the other two. Personally I prefer to go for a GUT (Grand Universal Theory) approach integrating both design and models into one framework. The important influences on my thinking in this area, in addition to Ericson, have been Scott (1977) and Rubin (1976). In the GUT approach the survey variables, the sampling mechanism, and any other selection and measurement mechanisms are all introduced explicitly into an overall model. If Y is the $n \times p$ matrix of measured survey variables, z is the prior information, s denotes the sample, $s^* \subset s$ denotes the respondents, then the joint distribution of all these variables is

$$f(Y \mid z;\Theta)g(z;\phi)\mathrm{p}(\mathrm{s} \mid z)q(s^* \mid s,z,Y_s;\eta),$$

where the survey design, represented by $p(s \mid z)$, is of the so-called uninformative type such as random sampling. The design is uninformative because z is assumed known and includes all the usual information on stratification, clustering and measures of size. This general formulation forces statisticians to face up to all their assumptions. Non-response must be modelled explicitly. Measurement errors must be included in the structure of $f(Y \mid z;\Theta)g(z;\phi)$. The decision to use randomisation inference is then an explicit statement that given z the values of Y can be treated as unknown constants; they are arbitrary values about which we have no additional information. A modeller, on the other hand must specify the model to the level needed for inference, for example, by an exchangeable model. Both design-based and model-dependent approaches condition on the same prior information, z, and so both should employ similar, possibly identical, structures. In fact I would rarely expect the point estimators using the different approaches to differ very much in practice. The issue thus becomes that identified by the authors as the choice of a measure of uncertainty. Model-dependent procedures employ conditional variances, strict design-based procedures are unconditional. How to construct conditional design-based inferences is still an open question, but the approach of Robinson (1987) looks promising. The GUT model shows the design-based versus model-based controversy to be what it is, namely a relatively small philosophical dispute within the much bigger framework of total survey analysis.

The failure of both theoretical and practical statisticians to integrate sampling and non-sampling errors into measures of total survey error even after 50 years of intensive research must be noted as one of the failures of this important branch of statistics. But again things are changing and the mood now is no longer merely to report sampling errors and in addition to give vague warnings about the potential size of non-sampling errors but it is to attempt to measure total survey error recognising that some non-sampling biases can far exceed sampling errors.

Section 4 of the paper is devoted to the analysis of survey data, to the analytic rather than descriptive uses of surveys. Here the design-based, model-based dispute pales into insignificance. Analysts must face up to all the classical problems of model choice, estimation and testing, residual analysis and so on, which make up mainstream statistics. Cinderella is at last dancing with the Prince.

My final comments are again personal. If you look at the references at the end of the paper, and if you consider the additional areas which I have discussed, then you will see that Jon Rao has contributed important papers in every area. I think that it was particularly appropriate that he was invited to write this paper. I congratulate both authors on their fine paper.

ADDITIONAL REFERENCES

DURBIN, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society, Series B, 15*, 262-269.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society, Series B, 31*, 195-233.

FELLEGI, I.P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

SCOTT, A.J. (1977). On the problem of randomization in survey sampling. *Sankhyā*, C, 39, 1-9.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 31–50
Statistics Canada

A Historical Perspective on the Institutional Bases for Survey Research in the United States

STEPHEN E. FIENBERG and JUDITH M. TANUR[1]

ABSTRACT

The basic theme of this paper is that the development of survey methods in the technical sense can only be well understood in the context of the development of the institutions through which survey-taking is done. Thus we consider here survey methods *in the large*, in order to better prepare the reader for consideration of more formal methodological developments in sampling theory in the mathematical statistics sense. After a brief introduction, we give a historical overview of the evolution of institutional and contextual factors in Europe and the United States, up through the early part of the twentieth century, concentrating on governmental activities. We then focus on the emergence of institutional bases for survey research in the United States, primarily in the 1930s and 1940s. In a separate section, we take special note of the role of the U.S. Bureau of the Census in the study of non-sampling errors that was initiated in the 1940s and 1950s. Then, we look at three areas of basic change in survey methodology since 1960.

KEY WORDS: Censuses; Cognitive aspects of survey design; Non-sampling errors; Probability sampling; Survey organizations.

1. INTRODUCTION

The development of survey methods in the technical sense can only be well understood in the context of the development of the institutions through which survey-taking is done. The purpose of this paper is to consider survey methods from this broader perspective in order to better prepare the reader for consideration of more formal methodological developments in sampling theory in the mathematical statistics sense that are described in numerous texts on sampling as well as in Rao and Bellhouse (1990). Although our viewpoint and organization is somewhat new, we have relied heavily on secondary sources which provide detailed expositions alternative to ours. Our paper focuses on the American experiences in the development of survey methodology, but it sketches some background of the much broader social science and institutional settings out of which survey methodology grew.

In the next section we present a very brief historical overview of the evolution of this institutional and contextual background, up through the early part of the twentieth century. We see two broad strands – social research and censuses. We begin with a short synopsis of the early history of European social research, turn to a brief overview of census-taking, especially in the context of the United States, and then take up the role of the International Statistical Congresses in the late nineteenth and early twentieth century in establishing the importance of sampling. Even following these congresses, the possible role of probability in sampling was not broadly understood. Further steps required an institutional base.

In section 3, we focus on the emergence of other U.S. institutional bases for survey research in the 1930s and 1940s. In particular, we note that a missing institutional ingredient was provided by the creation of the U.S. statistical agencies at the beginning of the twentieth century.

[1] Stephen E. Fienberg, College of Humanities and Social Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890, U.S.A.; Judith M. Tanur, State University of New York, Stony Brook, U.S.A.

Then a number of factors, including the depression of the 1930s, the development of probability sampling methodology, and a U.S. federal statistical coordinating function came together to launch the modern era of survey methodology in the U.S. We also review market research and the universities as institutional bases. In section 4, we take special note of the role of the U.S. Bureau of the Census in the study of non-sampling errors that was initiated in the 1940s and 1950s. In section 5, we look at some of the basic changes in survey methodology since 1960, focusing on technological advances, the role of longitudinal surveys, and the recent movement to explore cognitive aspects of surveys.

2. A HISTORICAL OVERVIEW OF THE INSTITUTIONAL BACKGROUND FOR MODERN SURVEY METHODS

2.1 Institutional Bases in Early European Social Research

One set of roots of the U.S. tradition of survey methodology and data collection technology is in early European social research (*cf.* Lecuyer and Oberschall 1978, from whose work we have drawn).

In England that tradition can be traced to the seventeenth century. The research, dubbed political arithmetic, was based on administrative records (especially parish records) and personal observation. It was usually carried out by dedicated individuals, such as John Graunt who published his *Natural and Political Observations Made Upon the Bills of Mortality* in 1662. Until the beginning of the eighteenth century, the parish was the unit of local government and administration, so that it was sensible to use the clergy as informants for many inquiries. With the industrial revolution and the rise of cities, this convenient arrangement broke down, necessitating the institution of house-to-house surveys.

By the 1830s statistical societies were formed in England to investigate social problems. They organized committees, which in turn hired agents to go door-to-door to collect data. Although the statistical societies disbanded when the social problems seemed solved, similar procedures were revived towards the end of the nineteenth century when Booth (1889-1891) sent school attendance officers door-to-door to study London's poor.

In France, where the government was more highly centralized, early social research was carried out by the government. District administrators were used as informants to fill out questionnaires on the demographic and economic conditions of their districts. By the mid-eighteenth century what we might consider an early study of the effects of mass communication was carried out in France. Administrators were instructed to spread rumors of increases in taxes and of military conscription and to report on the reactions of the populace.

In the Napoleonic period following the revolution, the French government established a national office responsible for gathering survey-like data on population, social situation, agriculture, and industry and commerce (Bourguet 1988). While this effort was not fully successful, and while it fell short of census methods as we now understand them, it did set in place an institutional structure. During the nineteenth century, France continued the tradition of government responsibility for statistical functions through reporting of data by prefects and in its Bureau de Statistique. The Napoleonic effort also launched a social science data enterprise in France that explicitly rejected the ideas from the theory of probability as it was then known. The French interest in social statistics affected many scientists, such as Laplace and Quetelet (a Belgian who studied in France under Laplace), who in turn contributed in major ways to the art and science of census-taking, attempting to reintroduce ideas from probability, through the use of what we now know as ratio estimation (see Stigler 1986, Chapter 5). After

the revolution of 1830, the Académie des Sciences Morales et Politiques sponsored prize competitions that encouraged statisticians to undertake their own research.

In Germany, the origin of "statistics" (collection of data on the state) was in the universities as early as the end of the seventeenth century. By the early nineteenth century this work was split into three parts, with descriptive political science and historical/quantitative political economy remaining university-based but statisticians collecting data in census bureaus and other government agencies.

In 1872, the Verein for Socialpolitik was founded – part pressure group, part professional organization, part research organization. It drew up questionnaires to be answered by supposedly knowledgeable informants such as landowners, ministers, and notaries. Problems of informants' possibly inaccurate information, haphazardly grouped and imprecise questions, and low response rates dogged these efforts. By the early twentieth century, Levenstein (1912) published what was probably the first large scale attitude and opinion survey, for which he used a snowball sampling technique. At about the same time Max Weber attempted a survey of industrial workers, planning to get some information directly from respondents but finding that the majority did not care to cooperate.

2.2 Censuses: A Prelude to Survey-Taking

Another set of roots of survey methodology is intertwined with the history of methods for census-taking and thus we present a brief overview on censuses and census-taking infrastructures. Many others have observed that the origins of the modern census are found in biblical censuses described in the Old Testament (Madansky 1986) as well as in censuses carried out by the the ancient Egyptians, Greeks, Japanese, Persians, and Romans (Taeuber 1978). The emphasis in the biblical accounts of censuses seemed to be on the results of the enumeration, rather than on how the counting was done, although in several instances we are told about the rapidity of the process. For most practical purposes we can skip from biblical times to the end of the eighteenth century and the initiation of census activities the United States of America, although there is some debate as whether Canada, Sweden, or the United States should be credited with originating the modern census (Willcox 1930).

In the United States, the first census was taken in 1790 (in 1990 the U.S. government will take its bicentennial census) by State officials who were then reimbursed by the Federal government. Then, in the next census of 1800, the enumerators were deputies or assistants to Federal marshals (Duncan and Shelton 1978). It was only with the 1880 census that the central Census Office gained control over field operations and secured the authority to appoint enumerators.

Prior to 1850 the U.S. decennial census considered the family as the unit of interest and reported few data on persons. The change to an individual-focus in census-taking was strongly influenced by the work Lemuel Shattuck, one of the of ASA's founders who had earlier conducted the Boston census of 1845 (Anderson 1988, pp. 36-37), as well as that of Quetelet, who helped to organize the 1846 Belgian census (Willcox 1930).

Progress on the methodology of census-taking continued, as every 10 years, a special operation was mounted to fulfill the constitutional obligation of an enumeration of the U.S. population; however, there was a clear lack of continuity from one census to the next (American Economic Association 1899). It was only after the first 12 censuses had been taken that the Bureau of the Census was created in 1902 as a permanent agency. Over this period there was a steady expansion of the number of censuses of other sorts and the broadening of topics covered in addition to simple enumeration.

2.3 International Statistical Congresses

The move from censuses to sample surveys was slow and laborious. Kruskal and Mosteller (1980) trace some of this movement, especially as it was reflected in the discussions regarding surveys that took place at the meetings of the International Statistical Institute (ISI), and our exposition here owes much to their work. The groundwork for these meetings was laid in the 1850s by Quetelet who helped to organize the first of a series of International Statistical Congresses in 1853. After nine such Congresses from 1853 to 1876, the ISI was founded in 1885. It is interesting to note that there is only one index entry for sample surveys in Stigler's (1986) history of statistics before 1900 – to 1830s work of Quetelet linked to a census method suggested by Laplace – and only two index entries in Porter's (1986) history – one to a 1900 paper by Karl Pearson and the other to the work of Kiaer and the ISI.

As early as the 1895 ISI meeting, Kiaer (1895-1896) argued for a "representative method" or "partial investigation", in which the investigator would first choose districts, cities, *etc.*, and then units (individuals) within those primary choices. The choosing at each level was to be done purposively, with an eye to the inclusion of all types of units. That coverage tenet, together with the large sample sizes recommended at all levels of sampling, was what was judged to make the selection representative.

The idea of less than a complete enumeration was widely opposed, but Kiaer presented arguments for it (with some members agreeing and others disagreeing) at ISI meetings in 1897, 1901, and 1903. Towards the end of this period, the idea of probability sampling entered the discussion, but the topic of the representative method seems absent from the records of the ISI meetings until 1925. By then the record suggests that the representative method was taken for granted, and the discussions centered around how to accomplish representativeness and how to measure the precision of sample-based estimates (Bowley 1926; Jensen 1926). Notions of clustering and stratification were put forward, but purposive sampling was still the method of choice.

It was not until Gini and Galvani made a purposive choice of which returns of an Italian census to preserve and found that districts chosen to represent the country's average on seven variables were, in that sense, unrepresentative on other variables, that purposive sampling was definitively discredited (Gini 1928; Gini and Galvani 1929). Soon thereafter Neyman published his groundbreaking 1934 paper that demonstrated, among other thing, the virtues of probability sampling.

3. THE DEVELOPMENT OF INSTITUTIONAL BASES FOR SURVEY RESEARCH IN THE UNITED STATES

Survey research in the United States grew from a blending of the same three institutional bases that had been influential in Europe – private individuals acting as entrepreneurs in the private sector, universities, and the government. Early social research in this country (before World War I) seems to have followed the earlier British model, being carried out by social workers, public health workers, and reformers. An early university involvement was the hiring by the University of Pennsylvania in 1899 of W.E.B. DuBois to carry out his study of the Philadelphia Negro, conducted as a house-to-house survey. Starting in the 1930's, and especially in the period after World War II, the U.S. experienced a flowering of survey methodology in the three broad institutional bases: market research and polling, universities, and government. But before we describe that flowering we shall take a step backwards and note the establishment of the U.S. government statistical agencies.

Recently, Jean Converse (1987) has written an extremely scholarly and graceful study of the roots and emergence of survey research in the United States, with special focus on market research and polling and on universities. Our exposition on these bases closely follow hers. We have separated out the institutional bases both to reflect a social reality and to structure our exposition. But there is another social reality that we ask the reader to bear in mind; the membranes separating the institutions are permeable. They not only permit the flow of cross-fertilizing ideas and methods in all directions; they also permit a somewhat lesser flow of people, as individuals move from one sector to another over the course of their careers.

3.1 The Establishment of U.S. Statistical Agencies

The establishment of American statistical agencies effectively began in 1863, when the newly created Department of Agriculture released the first crop and livestock report to provide information on Union food supplies during the Civil War. This report was based on data from a purposive sample of 2,000 farmers in 22 states. This agricultural statistical reporting activity has existed in the Department of Agriculture on a continuing basis to the present day and is now known as the National Agricultural Statistical Service. By the late 1920s, correlational and regression methodology was well established in the work of agricultural statisticians (Duncan and Shelton 1978).

In 1884, Congress voted to establish a Bureau of Labor (later renamed the Bureau of Labor Statistics, BLS) to "collect information" on the earnings and the working conditions of "laboring men and women." Under the leadership of Carroll Wright, the first Commissioner, BLS expanded its statistical activities to cover such issues as depressions, strikes and lockouts, women's wages, marriage and divorce, and the domestic liquor trade (Norwood and Early 1984).

With the creation of the Bureau of the Census in 1902, there were three major U.S. agencies in place, each with a mandate to collect national data on a regular basis. During the first three decades of the twentieth century, the role of government statistical agencies expanded considerably and, at the time of the stock market crash of October 29, 1929, data on various facets of economic and social life were available. As late as 1932, however, there were few examples of probability sampling anywhere in the Federal Government (Duncan and Shelton 1978).

Difficult though it is to conceive in a period when we are used to receiving reliable readings on the unemployment rate monthly, there was no comparable survey data resource available in the 1920s and early 1930s. Except for selected monthly non-survey data gathered by BLS from most manufacturing industries and some nonmanufacturing industries, there were no regular national unemployment figures. In the 1920 census the question on unemployment was dropped because of statistical concerns regarding the accuracy of the resulting data. This question was restored to the 1930 census because of the wide-spread concerns regarding the employment situation. The extensive controversy that surrounded the 1930 unemployment data (Van Kleeck 1930) and those from the special January 1931 Unemployment Census was especially acrimonious (Anderson 1988), and played a role in the 1932 presidential election campaign.

3.2 The ASA-SSRC Committee and the Institutionalization of Probability Sampling: An Early Bridge

Thus, at the beginning of the Great Depression of the 1930s in the United States, the federal statistical agencies had difficulty responding to the demand for statistics to monitor the effects of the programs of President Franklin Roosevelt's New Deal. In 1933, Secretary of Labor Frances Perkins asked Stuart A. Rice, the ASA president, to set up an advisory committee on

the programs of BLS. This committee grew into the Committee on Government Statistics and Information Services (COGSIS), sponsored jointly by ASA and the Social Science Research Council (SSRC). Duncan and Shelton (1978) give a detailed account of the activities of COGSIS, and for our discussion here two outcomes are worthy of note.

First, in 1933, COGSIS recommended the creation of a Central Statistics Board (CSB) to help coordinate government statistical activities. With the groundwork laid for a coordinated federal statistical system, COGSIS and CSB proceeded, in early 1934, to arrange for an interagency agreement through which Census would collect basic data on production and labor for BLS.

Second, COGSIS helped to stimulate the use of probability sampling methods in various parts of the Federal government, and it encouraged research on sampling theory, to be done by employees of statistical agencies. For example, to establish a technical basis for unemployment estimates, COGSIS and CSB organized an experimental Trial Census of Unemployment as a Civil Works Administration project in three cities using probability sampling, carried out in late 1933 and early 1934. The positive results from this study and the interagency arrangement mentioned above led in 1940 to the first large-scale, ongoing sample survey on employment and unemployment using probability sampling methods. This survey later became the Current Population Survey.

Another somewhat indirect outcome of the COGSIS emphasis on probability sampling took place at the Department of Agriculture Graduate School where W. Edwards Deming organized a series of lectures in 1937 on sampling and other statistical methods by Jerzy Neyman (1938). These lectures had a profound impact on the further development of sampling theory across the government as well as in universities.

What we see happening in this period is the confluence of a number of factors that served to launch the use and development of sampling methods in the U.S. government statistical agencies. A key prerequisite was the existence of the agencies themselves. A second was the methodological advances in sampling theory as encapsulated in Neyman's landmark 1934 paper. What was required to bring these together was the Great Depression, a new administration hungry for quality data to assess the impact of its social programs, and the joint ASA-SSRC Committee on Government Statistics and Information Services.

3.3 Market Research and Polling

The institutional base of survey methodology in U.S. market research and polling traces its own pre-history to election straw votes collected by newspapers, dating back at least to the beginning of the nineteenth century. Often publicity and circulation boosting were more important than accuracy of prediction. Converse (1987) points out, however, a more serious journalistic base; election polls were taken and published by such reputable magazines as the *Literary Digest* (which had gained a reputation for accuracy before the 1936 fiasco). Then, as now, election forecasting was taken as the acid test of survey validity. A reputation for accuracy in "calling" elections was thought to spill over to a presumption of accuracy in other, less verifiable areas.

There was a parallel tradition in market research, dating back to just before the turn of the century, attempting to measure consumers' product preferences and the effectiveness of advertising. It was seen as only a short step from measuring the opinions of potential consumers about products to measuring the opinions of the general public about other objects, either material or conceptual. By the mid 1930s there were several well established market research firms. Many of them conducted election polls in 1936 and achieved much greater accuracy than did the *Literary Digest*. It was the principals of these firms (*e.g.*, Archibald Crossley, George Gallup, and Elmo Roper) who put polling – election, public opinion, and consumer – on the map in the immediate pre-World War II period.

Data collection technology developed broadly in the market research and polling organizations in this era. Sampling was either by purposively selected groups or by quota. Samples were large, with the size enlarged sequentially until the law of large numbers caused the mean or percentage being estimated to stabilize. Some questionnaires were very informal with the interviewer instructed to bring certain topics into a conversation – what we might now call an unstructured interview. Others were more standardized, but shorter, actual forms. The progression seems to have been that as interviewers became more distanced from the primary investigators – in space, in education, in training, in identification with the research project, and perhaps in their very numerousness – the interview became more standardized.

The same kinds of validity issues that interest survey researchers today surfaced in the period. What should be the balance between open and closed questions? (Practice seems to have favored a combination; the device of the "opinion thermometer" to calibrate answers was first developed by the *Literary Digest* in 1925.) The pollsters tackled the problem of how to ask sensitive questions – about age, income, occupation, and home owning – by providing check lists, functioning much like contemporary visual aids. Experiments in question wording were carried out in the polling houses.

Market research in this early period, as now, of necessity put a premium on the timeliness of results. Then, as now, this tended to create some tension between academics and market researchers, with academics believing commercial workers to be corrupted by money and thus too far from basic science and commercial workers believing academics were overly concerned with the abstract. It is noteworthy, however, that one of the earliest homes of public opinion and market surveys was the Psychological Corporation, an organization of academic psychologists committed to plowing part of their profits back into the research process. The Psychological Corporation carried out its surveys from its Market Surveys Division, organized and run by Henry C. Link.

3.4 The Universities

But the universities were hardly totally above the polling movement. As early as 1911 the Harvard Graduate School of Business established a Bureau of Business Research to carry out consumer research. Such household names of social science as Paul Lazarsfeld, Hadley Cantril, and Rensis Likert moved to university affiliations and attached research institutes. Lazarsfeld came to the United States in 1933 determined to bring the techniques developed in market research to the basic scientific endeavor. He went on to form the Office of Radio Research, later to be called the Bureau of Applied Social Research, at Columbia University. His myriad contributions included the use of panels and a system of causal analysis.

Hadley Cantril was an academic who early on collaborated with Lazarsfeld on research on radio listening. When the two had a falling out, Cantril established the Office of Public Opinion Research at Princeton University. Here studies were carried out to improve data collection techniques. For example, in investigating the effects of question wording, Rugg and Cantril (1944) found that in 1940 – 41 over a six-week period, the percentage of Americans who favored "giving aid [to Great Britain] even at the risk of war" varied between 56% and 78%. At the same time, the percent in favor of "entering the war immediately" ranged from 8% to 22%.

Rensis Likert started out teaching at New York University and with a connection to the surveys of the Psychological Corporation. Moving to business, he carried out a survey of life insurance agents' attitudes, comparing qualitative and quantitative (mostly questionnaire) methods. He then became Director of the Division of Program Surveys at the Department of Agriculture. There he worked to standardize questionnaires. When Likert left the Department of Agriculture after World War II, he brought his group to the University of Michigan to form the Survey Research Center.

4. FROM SAMPLING THEORY TO THE STUDY OF NON-SAMPLING ERROR

As we have seen above, the introduction of probability sampling into government surveys in the mid-1930s came at the time of rapid development in many areas of statistics, and the development of a foundation for experimentation and inference more broadly under the leadership of such statisticians as R.A. Fisher, Walter Shewart, Jerzy Neyman, and Egon Pearson. Among those who worked on the probability-sampling-based trial Census of Unemployment at the Bureau of the Census were Calvert Dedrick, Morris Hansen, Samuel Stouffer, and Frederick Stephan (Anderson 1988; Duncan and Shelton 1978). Hansen was then assigned with a few others to explore the field of sampling for other possible uses at the Bureau, and went on to work on the 1937 sample Unemployment Census. After working on the sample component of the 1940 decennial census (under the direction of Deming), Hansen worked with others (*e.g.*, Jerome Cornfield, Lester Frankel, William Hurwitz and J. Steven Stock) to redesign the unemployment survey based on new ideas on multi-stage probability samples and cluster sampling (Hansen and Hurwitz 1942, 1943). They expanded and applied their approach in various Bureau surveys, often in collaboration and interaction with others, and this effort culminated in 1953 with the publication of a two-volume compendium of theory and methodology (Hansen, Hurwitz and Madow 1953). The recent interview with Hansen (Olkin 1987) and the Duncan and Shelton (1978) volume provide interesting and detailed descriptions of the developments during this period.

Virtually independent and often complementary contributions to sampling theory came via the statistical sampling work in agriculture by P.C. Mahalanobis and students in India and by Frank Yates and William Cochran in England. Cochran's 1939 paper is especially notable because of its use of the analysis of variance in sampling settings and the introduction of superpopulation and modeling approaches to the analysis of survey data (see Fienberg and Tanur 1987, 1988 for related discussion on the design and analysis linkages between sampling and experimentation). In the 1940s, as results from these two separate schools appeared in various statistical journals, we see some convergence of ideas and results.

The 1940s saw a rapid spread of probability sampling methods to other government agencies. It was only after the fiasco of the 1948 presidential pre-election poll predictions (Mosteller *et al.* 1949) that market research firms and others shifted towards probability sampling. Even today many organizations use a version of probability sampling with quotas (Sudman 1987).

Amidst the flurry of activity on the theory and practice of probability sampling during the 1940s, attention was also being focused on issues of nonresponse and other forms of nonsampling error. In a review of work on errors in surveys, Deming (1944) listed 13 factors affecting the ultimate usefulness of surveys (note that most of these are nonsampling errors):

1. variability in response;
2. differences between different kinds and degrees of canvass;
3. bias and variation arising from the interviewer;
4. bias of the auspices;
5. imperfections in the design of the questionnaire and tabulation plans;
6. changes that take place in the universe before tabulations are available;
7. bias arising from nonresponse (including omissions);
8. bias arising from late reports;
9. bias arising from an unrepresentative selection of date for the survey, or of the period covered;

10. bias arising from an unrepresentative selection of respondents;
11. sampling errors and biases;
12. processing errors (coding, editing, calculating, tabulating, tallying, *etc.*);
13. errors in interpretation.

Most of the errors described in this list either had been or would become the focus of research by statisticians at the Bureau of the Census.

A milestone in this effort to understand and model non-response errors was the development of an integrated model for sampling and non-sampling error in censuses and surveys, in connection with planning for and evaluation of the 1950 census (Hansen, Hurwitz, Marks and Mauldin 1951). This analysis-of-variance-like model, or variants of it, has served as the basis of much of the work on non-sampling error over the past 35 years, both inside and outside the Bureau of the Census. An excellent qualitative analysis of the error structure of the Current Population Survey is given in Brooks and Bailar (1978), and reviews of the non-sampling error literature are given by Mosteller (1978) and Fienberg and Tanur (1983). Finally, we note that Groves' (1989) recent book gives an updated approach to a variant of this census model, making a careful distinction between random and fixed components that arise from the various sources of error.

The paper by Bailar (1990) in this issue contains a detailed discussion on non-sampling error from the perspective of the Bureau of the Census.

5. CHANGING DIMENSIONS OF SURVEY METHODOLOGY AFTER 1960

The decades of the 1960s and 1970s saw polls and surveys becoming an all-pervasive fact of American life, beginning with the hard-fought presidential election of 1960 in which both candidates (Kennedy and Nixon) commissioned and relied on private polls of the electorate. Here we focus on three major areas of innovation during recent decades. We refer the reader to other presentations for such important topics as imputation for incomplete data and the ever-present controversies surrounding inferences from survey data (*e.g.*, see Fienberg and Tanur 1983, 1986).

5.1 Mode of Interviewing: The Role of Telephones and Computers in Surveys

The development and diffusion of technology, especially telephones and computers, strongly influenced survey practice in these decades. U.S. telephone coverage, which was estimated to have been only 35% in 1936 and hence contributed to the *Literary Digest*'s problem (Massey 1988), reached 75% by 1960 and 88% in 1970 on its way to around 93% in 1986 (Thornberry and Massey 1988). Thus telephone surveys, often based on random digit dialing (RDD) techniques, became increasingly prevalent and accurate. The movement began among commercial survey researchers, with governmental and academics lagging behind because of their concerns over differential coverage by such variables as income and race (Trewin and Lee 1988) and accompanying fears of lack of "representativeness". Indeed, most government uses of telephone interviewing remain as follow-ups of initial in-person contacts (as in the Current Population Survey which has been using telephone interviewing for households in later months of the survey since 1954). Only recently has there been a marked shift towards the use of RDD for government surveys. Groves and Kahn (1979) provide a review of work on telephone interviewing and, by and large, they document the comparability of survey results through comparisons of data gathered by personal interviews and by telephone.

The advent and proliferation of the computer meant that the tasks of analyses of survey responses could be carried out much more rapidly and broadly than ever before. This led to an increase in the number of surveys carried out under all institutional auspices. In retrospect it seems only natural that computer technology should be combined with telephone technology to produce systems of computer assisted telephone interviewing (CATI). These systems provide automated questionnaires that carry out skip patterns and display the appropriate question on a monitor screen, schedule (and often actually place) calls and callbacks, carry out randomizations, and automate data entry, in addition to other functions. CATI systems were developed by U.S. market research organizations in the early 1970s in part to keep track of respondent characteristics and thus ensure that quotas are precisely and efficiently met (Nicholls 1988). Chilton Research was one of the commercial CATI pioneers, using a CATI system for surveys intended to determine the level of customer satisfaction with services provided by the telephone companies (Nicholls and Groves 1986). Largely independently, university survey organizations began to develop their own CATI systems in the mid-1970s, and introduced them to the larger statistical community with an emphasis on their usefulness for documentation, standardization, and interviewer flexibility. While government agencies exhibited early interest in CATI, they have only recently begun to actually employ systems, sometimes on an experimental basis and often in tandem with other data collection methodologies, as in panel designs where the first interview is carried out in person. At this writing we see the beginnings of a movement to the use of computer-assisted personal interviewing (CAPI), a development made possible by the technological advances that produced truly portable laptop computers.

5.2 Longitudinal Surveys

While panel surveys were conducted in connection with the 1924 and 1940 U.S. presidential election campaigns (Rice 1928; Lazarsfeld *et al.* 1944), interest in over-time survey data did not really become fashionable in social research until the 1960s. This is all the more surprising when we realize that the Current Population Survey has traditionally had a rotating-panel structure and, since 1953, many respondents are interviewed as many as 8 times over a 16 month period. This rotating-panel structure was originally intended to produce estimates of change in aggregate quantities that had smaller variances than those from repeated cross-sections but, in principle, the CPS could have been analyzed in panel form on a regular basis. The fact that the CPS is a survey of sample addresses and not individuals or households is a major obstacle to the use of it as a panel survey (see related comments on the National Crime Survey in Fienberg 1978), but this has not prevented the elaborate use of the CPS to study gross flows in individual employment status (*e.g.*, see Abowd and Zellner 1985 and Stasny 1988).

Not all survey attempts to measure change need be based on longitudinal data; often repeated cross-sections can do at least as well if not better in measuring aggregate change. By the 1970s the Gallup Poll and others had developed a tradition of asking the same questions repeatedly and reporting the results in newspapers. These established time series became incorporated into the burgeoning Social Indicators movement. In 1972 the National Opinion Research Center first fielded the General Social Survey (GSS), funded by the National Science Foundation. GSS was designed by a broadly based group of academics to provide periodic readings on social indicators and to provide an original data set for use by students and academics doing modestly funded research. For purposes of continuity, the designers incorporated into GSS many questions first developed by Gallup and other commercial pollsters, yielding a fruitful cross-institutional collaboration (*e.g.*, see Smith 1975).

The basic idea behind the conduct of longitudinal surveys of panels, however, is to measure changes over time, not by comparing the changes in aggregate quantities, but by focusing on individual change. Such surveys typically focus on changes in status, the duration of activities, and events occurring over time. The rise of interest in longitudinal panel surveys occurred primarily outside the government, and early examples are the Panel Study of Income Dynamics, conducted by the Institute for Social Research at the University of Michigan annually since 1968; the National Longitudinal Surveys of Labor Market Experience, sponsored by the Center for Human Resources Research at Ohio State University beginning in 1966, and currently funded by BLS; and the Longitudinal Retirement History Survey, sponsored by the Social Security Administration from 1969 to 1979. The 1970s saw expanded use of longitudinal panel surveys, especially under government auspices (*e.g.*, see Boruch and Pearson 1988), but the basic survey methodology used often resembled that for traditional cross-sectional surveys. Only in the late 1970s did researchers begin to question the coventional wisdom about longitudinal survey design and analysis and to explore such fundamental issues as the definition of a longitudinal family (for a discussion, see Fienberg and Tanur 1986).

In the 1980s, interest in longitudinal panel surveys expanded and considerable attention was focused on aspects of non-sampling error such as attrition and on issues of data management and analysis. Kalton *et al.* (1989) includes a number of papers on these topics.

5.3 Cognitive Aspects of Surveys

As a result of systematic efforts to improve survey methodology over the past forty years, survey researchers have evolved a highly developed art of questionnaire design and interview procedures to reduce nonsampling errors, such as those described in Deming's list above (*e.g.*, see Payne 1951), and they have carried out many scientific studies to test aspects of that art (*e.g.*, see Sudman and Bradburn 1974, Bradburn and Sudman 1979, and Schuman and Presser 1981). Until recently, however, research on understanding the survey interview situation has been relatively unsystematic. The recent change came, in part, through the recognition that other fields, in particular cognitive psychology, had insights that would assist survey researchers in examining the interview process.

Among non-sampling errors are those occasioned by the cognitive processes that respondents and interviewers are required to exercise in the survey interview situation. Respondents must often recall events and make judgments or estimates, and must always face issues of comprehension of the questions asked – their meaning to respondents as well as their meaning to interviewers. Survey researchers are now beginning to draw on the concepts of cognitive psychology and the expertise of cognitive psychologists to investigate more systematically these issues of non-sampling error. We note especially that the exploration of meaning is not new to the enterprise of survey research. Indeed, Cantril (1944) devotes two chapters to reporting the results of experiments on the meaning and wording of questions. These experiments used many of the same probing and paraphrasing techniques used in today's cognitive laboratory.

This explicit movement to study cognitive aspects of surveys originated in a 1981 conference sponsored by the Bureau of Social Science Research and the Bureau of Justice Statistics that brought together cognitive psychologists and survey researchers to concentrate on the National Crime Survey. A more intensive 1983 conference, sponsored by the Committee on National Statistics (CNSTAT) of the National Research Council, concentrated on the National Health Interview Survey (Jabine *et al.* 1984). From the beginning the movement was, by design, a partnership between people from academia, from research institutes and other academic institutions, and from the government.

A direct outgrowth of the CNSTAT conference was the establishment of a Questionnaire Design Research Laboratory at the U.S. National Center for Health Statistics under the leadership of Monroe Sirken to do pretesting (in parallel with full scale field testing) of major government surveys. It employs government personnel, brings in visiting scholars, and contracts with academics and people in research institutes to carry out its mission. This has been followed by the establishment of similar laboratories at the Bureau of Labor Statistics and the Bureau of the Census. Another outgrowth is the establishment of the Social Science Research Council's Committee on Cognition and Survey Research, which is, itself, both cross disciplinary and cross institutional. The Committee has fostered research in such directions as the interactive process of the survey interview, the uses and pitfalls of retrospective memory, and issues in measuring pain in a survey context. Examples of other outgrowths of this movement are (a) an investigation by the OECD's Working Party on Labor Statistics of cognitive aspects of labor surveys, addressing such issues as the meaning of "looking for work" - a knotty conceptual problem within a culture, and even more problematic across cultures (Schwarz 1987), (b) work at combining the cognitive perspective with statistical work on the embedding of experiments within surveys (Fienberg and Tanur 1989), (c) international conferences on work at the interface of cognition and survey methods (*e.g.*, see Hippler, Schwarz and Sudman 1987).

At the same time that methodological techniques of the cognitive laboratory are being used to shape questionnaire design, findings from the cognitive psychology laboratory are being taken into the field in order to test their generalizability and thus enrich the academic field of cognitive psychology, as well as to ascertain their usefulness for the survey enterprise. Here is yet another instance of interaction between the academic world and the government. For example, a laboratory finding is that people recall visits to health care providers more easily and accurately if they begin with the earliest first (Fathi, Schooler and Loftus 1984). A recent investigation explores whether this advantage holds in the field situation of the pre-test of the NHIS (White and Berk 1987).

The movement to integrate methods from the cognitive sciences into the design of sample surveys is important for several reasons. First, it has brought a renewed scientific base to the problems of questionnaire design. Second, it has opened up the survey domain to the study of selected cognitive phenomena. But most important, it had brought new vigor to the survey enterprise and raised anew issues about the structure and format of the survey interview, going far beyond questionnaire design, that many statisticians thought were resolved in the 1940s and 1950s.

6. COMMENTS

Traditional reviews of the history of survey methods have focused on the role of probability sampling and its refinements, and occasionally on the study of non-sampling errors. Here we have attempted to set this methodological history in the context of the tradition of social science research that evolved over the nineteenth and early twentieth centuries and the institutions, in and outside of government, that facilitated and occasionally directly spawned the methodological developments. This perspective should help remind readers that factors other than the advance of statistical theory have helped to shape the survey domain as we know it today. It should also help them follow the evolution of survey theory and practice as it continues to be shaped by institutional change.

There is an additional facet of institutional shaping of the survey enterprise that we have not addressed heretofore. We wrote above about the permeability of the membranes separating the three sectors: government, market research (the private domain), and the universities and

other academic institutions. We believe that these membranes are becoming even more permeable with the increased presence of a fourth kind of institution, which we shall refer to as a "bridge". We saw earlier how the ASA-SSRC Committee on Government Statistics and Information Services, a bridge between academia and government, prepared the ground for federal statistical coordination. ASA and SSRC continue to provide bridging functions, but other such institutions also exist.

Some vivid examples of other bridges come to mind. For over 40 years the American Association for Public Opinion Research has been bringing together survey practitioners from all sectors in local chapters and in national conferences at which new findings are disseminated and issues of common concern are discussed. The National Science Foundation program on Measurement Methods and Data Improvement (MMDI), under the direction of Murray Aborn, has explicitly seen as part of its mandate the fostering of government/academic collaboration. The mission has been implemented, for example, through the funding of research by academics that both uses and improves government databases (the 1983 seminar on cognitive aspects of survey methodology was sponsored by MMDI) and the funding of an ASA-sponsored fellowship program. That fellowship program places academic researchers for a semester or a year in government statistical agencies to carry out their own research, bring new ideas to the agency, and return to their academic bases with new knowledge and contacts in the federal agencies and new awareness of government data bases and statistical concerns. The National Research Council, an arm of the National Academy of Sciences, maintains a Committee on National Statistics that brings statisticians from academia and the private sector together to interact with representatives of the government agencies. Here, in formal panel studies and informal interaction, individuals come to know one another and common problems are tackled.

While these and other bridges will surely not totally erase the boundaries between the sectors, we see their existence as a positive force for progress in the development of survey methodology. Developments in one sector move more quickly to others across these bridges, but perhaps more important, the bridges facilitate a process whereby problems faced by any sector become legitimate research questions in all sectors.

ACKNOWLEDGEMENTS

The preparation of this paper was supported in part by the National Science Foundation under Grant No. SES-8701606 to Carnegie Mellon University and Grant No. SES-8701816 to the State University of New York at Stony Brook. An earlier version appeared under the title "Some History of Survey Methods and Data Collection Technology," in the *Sesquicentennial Invited Paper Sessions* volume of the 1989 Proceedings of the American Statistical Association, 393-405.

REFERENCES

ABOWD, J.M., and ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics*, 3, 254-283.

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

AMERICAN ECONOMIC ASSOCIATION (1899). The Federal Census. Report of the committee on the Twelfth Census. *Publications of the American Economic Association*, New Series, No. 2, 1-7.

BAILAR, B.A. (1990). Contributions to statistical methodology from the federal government. *Survey Methodology*, 16, 51-61.

BOOTH, C. *et al.* (1889-1891) 1902-1903. *Life and Labours of the People in London*. London: Macmillan.

BORUCH, R.F., and PEARSON, R.W. (1988). Assessing the quality of longitudinal surveys. *Evaluation Review*, 12, 3-58.

BOURGUET, M.-N. (1988). Décrire, Compter, Calculer: The debate over statistics during the Napoleonic Period. In *The Probabilistic Revolution. Volume I, Ideas in History* (Eds. L. Kruger, L.J. Daston and M. Heidelberger). Cambridge: MIT Press.

BOWLEY, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, 6-62.

BRADBURN, N.M., SUDMAN, S., and Associates (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.

BROOKS, C.A., and BAILAR, B.A. (1978). *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Working Paper 3, Office of Federal Statistical Policy and Standards. Washington: U.S. Department of Commerce.

CANTRIL, H. (1944) (1947). *Gauging Public Opinion*, Princeton: Princeton University Press.

COCHRAN, W.G. (1939). The use of the analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.

CONVERSE, J.M. (1987). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.

DEMING, W.E. (1944). On errors in surveys. *American Sociological Review*, 19, 359-369.

DUBOIS, W.E.B. (1899) (1973). *The Philadelphia Negro: A Social Study; Together With a Special Report on Domestic Service by Isabel Eaton*. Millwood, N.Y.: Kraus Reprint.

DUNCAN, J.W., and SHELTON, W.C. (1978). *Revolution in United States Government Statistics, 1926-1976*. U.S. Department of Commerce. Washington: U.S. Government Printing Office.

FATHI, D., SCHOOLER, J., and LOFTUS, E. (1984). Moving survey problems into the cognitive psychology laboratory. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 19-21.

FIENBERG, S.E. (1978). Victimization and the National Crime Survey: Problems of design and analysis. In *Survey Sampling and Measurement* (Ed. K. Namboodiri). New York: Academic.

FIENBERG, S.E., and TANUR, J.M. (1983). Large scale social surveys: Perspectives, problems, and prospects. *Behavioral Science*, 28, 135-153.

FIENBERG, S.E., and TANUR, J.M. (1986). The design and analysis of longitudinal surveys: Controversies and issues of cost and continuity. In *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits* (Eds. R.W. Pearson and R.F. Boruch). New York: Springer-Verlag.

FIENBERG, S.E., and TANUR, J.M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *Canadian Journal of Statistics*, 55, 75-96.

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

FIENBERG, S.E., and TANUR, J.M. (1989). Combining cognitive and statistical approaches to survey design. *Science*, 242, 1017-1022.

GINI, C. (1928). Une application de la méthode représentative aux matériaux du dernier recensement de la population Italienne (1er décembre 1921). *Bulletin of the International Statistical Institute*, 23, 198-215.

GINI, C., and GALVANI, L. (1929). Di una applicazione del metodo rappresentativo all'ultimo censimento Italiano della popolazione (1 dicembre 1921) (In Italian). *Annali di Statistica*, Series 6, 4, 1-107.

GRAUNT, J. (1662) (1939). *Natural and Political Observations Made Upon the Bills of Mortality* (Edited and with an introduction by Walter F. Willcox). Baltimore: Johns Hopkins Press.

GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L. II, and WAKSBERG, J., eds. (1988). *Telephone Survey Methods*. New York: Wiley.

GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

GROVES, R.M., and KAHN, R.L. (1979). *Surveys by Telephone*. New York: Academic Press.

HANSEN, M.H., and HURWITZ, W.N. (1942). Relative efficiencies of various sampling units in population inquiries. *Journal of the American Statistical Association*, 37, 89-94.

HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, New York: Wiley.

HANSEN, M.H., HURWITZ, W.N., MARKS, E.S., and MAULDIN, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190.

HIPPLER, H.-J., SCHWARZ, N., and SUDMAN, S., eds. (1987). *Social Information Processing and Survey Methodology*. New York: Springer-Verlag.

JABINE, T.B., STRAF, M., TANUR, J.M., and TOURANGEAU, R., eds. (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington: National Academy Press.

JENSEN, A. (1926). Report on the representative method in statistics. *Bulletin of the International Statistical Institute*, 22, 359-380.

KALTON, G., KASPRZYK, D., and DUNCAN, G.J., eds. (1989). *Panel Surveys*. New York: Wiley.

KIAER, A.N. (1895-1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin of the International Statistical Institute*, 9, 176-183.

KRUSKAL, W.H., and MOSTELLER, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review*, 48, 169-195.

LAZARSFELD, P.F., BERELSON, B., and GAUDET, H. (1944). *The People's Choice: How the Voter Makes up his Mind in a Presidential Campaign*. New York: Columbia University Press.

LECUYER, B., and OBERSCHALL, A. (1978). Social research, the early history of. In *International Encyclopedia of Statistics* (Eds. W.H. Kruskal and J.M. Tanur). New York: Macmillan and the Free Press.

LEVENSTEIN, A. (1912). *Die Arbeitfrage mit besonderer Berücksichtigung der sozialpsychologischen Seite des modernen Grossbetreibes und der psychopsysischen Einwirkungen auf die Arbeiter*. (In German) Munich: Reinhardt.

MADANSKY, A. (1986). On biblical censuses. *Journal of Official Statistics*, 2, 561-569.

MASSEY, J.T. (1988). An overview of telephone coverage. In *Telephone Survey Methods* (Eds. R.M. Groves *et al.*). New York: Wiley.

MOSTELLER, F. (1978). Errors: 1. Nonsampling errors. In *Intenational Encyclopedia of Statistics* (Eds. W.H. Kruskal and J.M. Tanur). New York: Macmillan and the Free Press.

MOSTELLER, F., HYMAN, H., McCARTHY, P.J., MARKS, E.S., and TRUMAN, D.B. (1949). *The Pre-election Polls of 1948*. Bulletin 60. New York: Social Science Research Council.

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

NEYMAN, J. (1938). *Lectures and Conferences on Mathematical Statistics and Probability*. Washington, DC: Graduate School, U.S. Department of Agriculture.

NICHOLLS, W. L., II (1988). Computer-assisted telephone interviewing: A general introduction. In *Telephone Survey Methods* (Eds. R.M. Groves *et al.*). New York: Wiley.

NICHOLLS, W.L., II, and GROVES, R.M. (1986). The status of computer-assisted telephone interviewing: Part 1 - Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics*, 2, 93-115.

NORWOOD, J.L., and EARLY, J.F. (1984). A century of methodological progress at the U.S. Bureau of Labor Statistics. *Journal of the American Statistical Association*, 79, 748-761.

OLKIN, I. (1987). A conversation with Morris Hansen. *Statistical Science*, 2, 162-179.

PAYNE, S. L. (1951). *The Art of Asking Questions*. Princeton: Princeton University Press.

PORTER, T.M. (1986). *The Rise of Statistical Thinking, 1820-1900*. Princeton: Princeton University Press.

RAO, J.N.K., and BELLHOUSE, D.R. (1990). The history and development of the theoretical foundations of survey based estimation and statistical analysis. *Survey Methodology*, 16, 3-29.

RICE, S. (1928). *Quantitative Methods in Politics*. New York: Knopf.

RUGG, D., and CANTRIL, H. (1944) (1947). The wording of questions. In *Gauging Public Opinion* (Ed. H. Cantril). Princeton: Princeton University Press.

SCHWARZ, N. (1987). Cognitive aspects of labor surveys in a multinational context. Paper prepared for the Working Party on Labor Statistics, OECD. Paris, April 1987.

SCHUMAN, H., and PRESSER, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic.

SMITH, T.W. (1975). Social change and the General Social Survey: An annotated bibliography. *Social Indicators Research*, 2, 9-38.

STASNY, E.A. (1988). Modelling nonignorable nonresponse in categorical panel data with an example in estimating gross labor force flows. *Journal of Business and Economic Statistics*, 6, 207-219.

STIGLER, S.M. (1986). *The History of Statistics. The Measurement of Uncertainty Before 1900.* Cambridge: Harvard University Press.

SUDMAN, S. (1987). *Reducing the Costs of Surveys*. Chicago: Aldine.

SUDMAN, S., and BRADBURN, N.M. (1974). *Response Errors in Surveys: A Review and Synthesis.* Chicago: Aldine.

TAEUBER, C. (1978). Census. In *International Encyclopedia of Statistics* (Eds. W.H. Kruskal and J.M. Tanur). New York: Macmillan and the Free Press.

THORNBERRY, O.T. Jr., and MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. In *Telephone Survey Methods* (Eds. R.M. Groves *et al.*). New York: Wiley.

TREWIN, D., and LEE, G. (1988). International comparisons of telephone coverage. In *Telephone Survey Methods* (Eds. R.M. Groves *et al.*). New York: Wiley.

VAN KLEECK, M. (1930). The Federal Unemployment Census of 1930. *Proceedings of the American Statistical Association*, 189-200.

WHITE, A.A., and BERK, M.L. (1987). Recall strategies in personal interviewing: moving results from the laboratory to the field. *Proceedings of the Social Statistics Section, American Statistical Association*, 66-71.

WILLCOX, W.F. (1930). Census. In *Encyclopaedia of Social Sciences* (Eds. E.R.A. Seligman and A. Johnson). New York: Macmillan.

COMMENT

ROBERT M. GROVES[1]

The writing of histories of the development and use of survey methods signals a certain maturation of the field. Currently, we are seeing the fiftieth anniversary of several important survey innovations - Neyman's breakthrough papers in stratification, the start of the U.S. Current Population Survey, and the greater visibility of election polling. With our attention called to such developments it is natural to review the intervening years, seeking to find some theme for events affecting the field. Professors Fienberg and Tanur have completed such an exercise in their paper.

My comments will review key parts of the work, offering comments as I proceed, and then note some errors of nonobservation, misplaced emphases, and other minor quibbles.

Fienberg and Tanur express the purpose of their paper in two ways "to note that technical developments in surveys can be understood only in the context of institutions within which they occur "(p. 31) or at another point to note that "factors other than the advance of statistical theory have shaped the survey domain" (p. 42). Consistent with this they note:

1. the role of ruling, governing institutions which perceive a need for information on the population's welfare or its reaction to taxation;
2. later, the role of academics in the social sciences in framing central statistical and measurement issues in surveys;
3. the role of mass media use of surveys for election and current events monitoring; and
4. still later, the use of surveys by commercial entities in the market economy.

They document the resolution of controversies in the government sector about use of probability sampling.

Along the way we learn some interesting facts - for example, that for 12 U.S. censuses (120 years) there was no permanent organization for the Census Bureau; that the Department of Agriculture data collection began with need for information about food supplies in the Civil war; that another boost for surveys occurred in the New Deal's creation for government programs. There seems to be a recurring theme here that governments emphasizing services for the welfare of the populace demand more information about their societies than do those pursuing other goals. In addition, we see that governments most sensitive to public opinion demand more measures of that opinion (that reminds this reader of Gallup's early metaphor of the survey as a voting analog).

The focus in the paper on the role that institutions played in the development is convincing only for parts of the review. For example, the institutional focus is appealing in describing Lazarsfeld's evangelical efforts to bring commercial survey and academic inquiry together. The role of the Bureau of Applied Social Research at Columbia University in his partial success at that is enlightening. So too the move of Likert and others from a government agency home (U.S. Department of Agriculture) to academia in order to spread the method to new domains is largely a story of groups of people and organizations which make them effective.

However, the identification of organizations or institutions as the focus can be misunderstood as the stimulus to developments. Nothing I read in the paper changes my opinion that the survey field at its origins attracted broad, creative thinkers. Many were intelligent and charismatic; they led by ideas and mobilized others to work diligently at the definition

[1] Robert M. Groves, The University of Michigan and U.S. Bureau of the Census.

of the new field. Institutions permitted this to happen. They didn't produce the developments. They were homes for the best and brightest.

Within the focus on the institutional, I wished the emphasis of the paper might have been placed more on two related points:

1. Different tasks were more easily accomplished in the different domains. For example, government agencies were by their nature restricted to questions of monitoring social welfare, the commercial, to newsworthy or dollar worthy interests, and the academic to longer term, more basic social issues. Those involved in early developments shaped their agenda to the goals of the organization.

2. Stories of the early days of survey research, as told by those who lived them are filled with the excitement of a new field. I missed in the paper sufficient acknowledgement that the young researchers involved in the work shared an evangelical mission – spreading the gospel of probability sampling, inventing new methods of interviewing because nothing existed. The institutional focus misses the human drama of those days.

Fienberg and Tanur also note "the membranes separating the institutions are extremely permeable". That is, researchers move back and forth between the institutions, contributing to each of them, and transferring knowledge as they move. The evidence the authors cite is the experience of Lazarsfeld addressing basic design issues while conducting radio audience research within an academic setting and of Likert moving from the insurance industry to the Department of Agriculture to the University of Michigan. These moves seem the exception rather than the rule. I have not conducted the appropriate careerline research to demonstrate this, but my impression is that the fences between the sectors have been and remain high and painful to transgress. Further, movement among academic government, commercial is asymmetric. Rarely is there movement from the commercial or government sector to the academic sector (current demands on publication history prevent this). The government-commercial interchange is larger.

The result of this insularity is the development of techniques not shared across the different sectors (edit and imputations schemes, nonresponse reduction techniques). The three sectors to some extent have developed their own language to describe their work (*e.g.*, "stem and banners", "tabs" versus "contingency tables").

The membrane metaphor also fails to observe the large differences in the centrality of surveys to organizations in the three sectors. Academic survey research is not central to any university in the world. It was not central early in the history of the method (viz. the inability of the Likert group to obtain university parking stickers because of their nonfaculty status). Even now it is often viewed as a haven of technicians (several steps below the chemistry laboratory staff) currently on many campuses. In contrast there are government and commercial organizations fully devoted to survey design, collection, and analysis. These have decision-making hierarchies constantly monitoring cost and error structures of surveys without the ongoing debate about the relative worth of the enterprise.

The paper ends with a discussion of three developments since 1960 that are important to understanding surveys. At this point, the institutional context is dropped as the organizing principle of the paper and innovations are the focus. Three developments are highlighted: a) the use of the telephone as a data collection medium and later developments in computer assisted telephone interviewing (CATI); b) the use of longitudinal surveys to study micro-level change over time; and c) the application of cognitive psychological concepts to survey methods.

The authors note the movement of mode of data collection from face to face to phone and development of CATI, but they fail to note that this is largely a US phenomenon in the academic and government sectors (the commercial side had done it years ago). Indeed, it is an example of distinctive methodologies pursued by the three sectors. I share their belief that the merits of longitudinal surveys are increasingly being recognized and note that the 1980's is seeing this spread internationally. The Fienberg and Tanur team was instrumental in launching the U.S. effort the apply cognitive psychological concepts to survey measurement, and we are in their debt for this.

The paper does not make it clear whether the authors believe the CATI, longitudinal surveys, and the effort to "cognitize" survey methodological research are the most important three developments in surveys, but they clearly omit several other important ones. We can all choose our three most important developments since 1960; here are some other candidates:

1. Development of Generalized Statistical Software Packages

This development greatly expanded the number of researchers who could directly pose and answer questions using survey data. In the statistical and social sciences at this writing, it is common for undergraduates to perform analyses of survey data whose complexity would have prevented their being done 25 years ago.

2. Existence of Survey Data Archives

The archiving of survey data on computer media was a further democratizing force in survey analysis. With those developments replication and extension of analysis, a key component of the structure of scientific advance, became trivial. Unfortunately, there were also deleterious effects. Analysts of survey data could do their work in complete ignorance of the survey design, of the interviewer training and supervision guidelines, of nonresponse rates, and of a host of other design features known by those conducting the survey.

3. Growth of Commercial and Nonprofit Industry to do Government Surveys

The U.S. is distinctive in its reliance on academic and commercial groups to conduct surveys on behalf of government agencies. Some of this exists in many Western countries, but to a much smaller degree. This suggests that a cross-cultural strain in the paper might be interesting – to identify unique histories of survey research in various societies.

4. 1960 as Beginning of Widespread Acceptance in Academic Circles of the Social Psychological Model of the Interview

This typically describes survey interviews as "conversations with a purpose" and focuses the researcher's attention on the role of the two actors in the errors produced during measurement.

5. Ubiquity of Surveys

Survey measurement is now a way of life for most large corporations (prior to the breakup of ATT in the U.S. the corporation conducted over 7 million customer satisfaction interviews annually). Surveys are viewed as irreplaceable sources of information about customers, suppliers, and the general society.

6. Nonresponse and the Growing Reluctance of the Population to be Measured

This is certainly a phenomenon of great import to survey researchers in most Western countries. With statistical inference to large populations one of the key virtues of surveys versus other data collection schemes, this issue strikes at the heart of the tool. Again, a cross-national theme to the paper would have highlighted these issues.

We can apply the superpopulation metaphor to any historical account – that is, any series of events (which later we call history) is but one realization of an infinite set of possible series which defines the universe of possible realities. This fits the set of questions that remain unanswered.

1. Why after almost a century hasn't survey research fully evolved into a profession (with specified standards and training criteria)?
2. Why is there so little formal educational structure for survey researchers to get their knowledge base? Why are there departments of communications, operations research, naval architecture but none of survey research (teaching sampling, questionnaire design, data analysis)?
3. Would public education about surveys and statistics (like the ASA/NSF program in quantitative literacy) have made an impact on acceptance of surveys?

We are indebted to the Fienberg/Tanur team for reviewing our collective past. They have helped chronicle the birth and first 50 years or so of what is now an important component in most societies of the world. I do hope that the year 2040 will see the need to ask Fienberg and Tanur to update their paper for that occasion. I hope they will be able to report innovation during those 50 years that made a difference in survey methods.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 51–61
Statistics Canada

Contributions to Statistical Methodology from the U.S. Federal Government

BARBARA A. BAILAR[1]

ABSTRACT

Drawing upon experiences from developments at the U.S. Bureau of the Census, the paper briefly traces some contributions made by practitioners to the theory and application of censuses and surveys. Some guesses about future developments are also given.

KEY WORDS: Sampling; Nonsampling error; Estimation; Confidentiality; Seasonal adjustment.

1. INTRODUCTION

In the United States, the federal government has led the way in the development of statistical methodology in censuses and surveys. I will confine my remarks to examples from the U.S. Bureau of the Census and will discuss four main areas of work – the development of sampling methods, non-sampling error, seasonal adjustment, and the development of methods to protect the confidentiality of respondents, usually called disclosure avoidance techniques. Finally, I will venture to hazard some guesses about future development.

2. SAMPLING

The story of sampling in the U.S. federal government is primarily the story of a remarkable group of people at the Census Bureau, led by Morris Hansen and William Hurwitz. When one considers that the Census Bureau was committed to probability sampling in the early 1940's, one wonders: how could an innovation of this type have occurred so quickly in such a conservative institution? The adoption of innovative methods often takes a very long time and I suspect the Bureau is much slower in adopting and promoting new methodology today. Hansen has given three reasons why he thinks sampling was accepted relatively quickly by the subject-matter divisions of the Bureau. They are: (1) support from the top, (2) conscious development of a team-work approach with the subject-matter divisions, and (3) the development of a corps of sampling experts (later, methods specialists) in the subject-matter divisions who were responsible to the Statistical Research Division (SRD) on technical matters. I think he left out one key ingredient and that is the force and the spirit of the dynamic duo and their cohorts.

In 1936, the Bureau began exploration of sampling and potential applications. Some sampling was already in use, but not probability sampling. There was judgment sampling and sampling of some large establishments. However, there was little or no theory to guide sampling approaches. In 1937, Congress authorized a national voluntary registration of the unemployed and partially employed. A questionnaire was to be delivered by the Post Office to every household. There was some concern that this voluntary registration could have some bias, so an enumerative check census was put in place in a sample of areas. The check census required interviewing all households within a probability sample of postal delivery routes. The mail

[1] Barbara A. Bailar, American Statistical Association, 1429 Duke Street, Alexandria, VA 22314-3402.

carriers did the interviewing and identified and sorted the voluntary mail returns. They then provided separate counts for each postal route, including the sample postal routes. This then gave an independent variable to use in the estimation, one of the earliest demonstrations of ratio estimation. The results of the check census were convincing on the usefulness of sampling. However, the entire effort was remarkable in many ways:

- the effects of nonresponse from a voluntary census were anticipated;
- the use of ratio estimation;
- the speedy results.

Hansen, in an interview in *Statistical Science* (Olkin), reports that the registration took place the week of November 20, 1937; that the household canvas was done during the week of December 4, 1937; and preliminary results became available on New Year's Eve, 1937. I don't think the Census Bureau could beat that record now.

Hansen attributes the success of the 1937 enumerative check census as a demonstration of the use of sampling as key in gaining acceptance within the Bureau. Before then, Bureau staff believed that complete coverage was necessary and that sampling would discredit the Bureau. The success of the study helped gain the acceptance of sampling in the 1940 census, the first census in which some questions were asked of only a sample, not the entire population. Unfortunately, in the last few months, some at Census have dragged out the old chestnut about needing to do the vacant delete check on a 100% basis because a census has less error than a survey. Let's just assume that was a temporary aberration caused by litigation.

A great deal of the theory of sampling was developed in conjunction with the Labor Force Survey. The Works Progress Administration (WPA) sponsored a survey to measure unemployment. In 1942, when the WPA was abolished, the survey was moved to the Census Bureau. The sampling procedures were evaluated and many improvements were made. Several important contributions to sampling theory came from that revision. Some of the sampling principles introduced into the 1942 revision were: enlarged primary sampling units, sampling with probabilities proportionate to a measure of size, and area substratification. These principles were discussed in a 1943 paper by Hansen and Hurwitz in the *Annals of Mathematical Statistics*. Rereading this paper, "On The Theory Of Sampling From Finite Populations," always provides new insights. The article seems to be the first published by federal employees on the topic of sampling of finite populations. Though the concepts had been discussed by others, the extension of theory was new. Also, a hallmark of Hansen and Hurwitz, the results were discussed in a series of practical comparisons highlighting the advantages of the recommended procedures.

Improvements in the Labor Force Survey continued over the years. Composite estimation, using the system of sample rotation to improve the estimates, was introduced. The Current Population Survey, as the Labor Force Survey is now called, has undoubtedly led the way throughout the world in setting the standards for a labor force survey.

Surveys of business establishments presented new sampling problems, also undertaken by the Statistical Research Division. The attitude frequently encountered was that sampling might be all right with relatively homogenous populations such as people but they would not work with highly skewed populations such as businesses. Working with the acknowledged skewness of the population, the sampling group stratified the retail stores by size. The largest stores were necessarily included in the sample, and the smaller businesses were sampled with probability proportionate to a measure of size.

It was also apparent that businesses came into being and died frequently. A static sample would not be able to capture this turnover. Therefore, an area sample to provide estimates

for new stores was incorporated. The Monthly Retail Trade Survey has seen many innovations, but these basic cornerstones remain. The Retail Trade Survey also makes use of composite estimation to provide more precise estimates.

Many other instances of sampling innovations could be mentioned. Many descriptions are given, and the theory and practical applications are described in the book *Sample Survey Methods and Theory* in two volumes, by Hansen, Hurwitz, and Madow (1953). Though the illustrations are seriously outdated, the books still provide more practical sampling applications than any other books I know of. I only regret that they were never updated.

3. NON-SAMPLING ERROR

Another major advance in sample surveys and censuses was to look beyond sampling error to try to control the errors arising from other sources, such as the interviewers, processors, questionnaires, and so forth. Hansen and Hurwitz moved in that direction before the 1950 Census, incorporating many experimental studies in the census designed to estimate the effect of measurement errors in the census. Total survey error became a strong focus at the Census Bureau. The measurement and control of nonsampling errors became a regular feature of Census Bureau work.

An impetus to this nonsampling error work was the recognition that measurement errors could have a much stronger effect on data than sampling errors, especially at larger levels of aggregation. Hansen, Hurwitz and Bershad (1961) developed an integrated model for censuses and surveys that explicitly incorporated sampling error, response error, and bias. The response error component contained what are now known as a simple response variance and a correlated response variance. The simple response variance reflects the basic trial-to-trial variability that arises from differences in respondent reporting, different respondents, different interviewers, and the like. The term has also been generalized to include the variance that arises from trial-to-trial variability in coding. The correlated response variance refers to the variance that arises from a factor that pushes responses into a certain pattern. The most studied factor is that of the interviewer. By having certain expectations or from experience interviewing at a few households, the interviewer can push responses into certain categories. We see wide variability among interviewers working in the same areas on nonresponse rates, on questions about educational attainment, and many other items.

This model was first tested in the 1950 census and was a major factor in the decision to move from an "enumerator census" where an interviewer went to every household, asked the questions, and recorded the answers, to a "mail census", where the questionnaires are sent to every household and householders are asked to fill out the forms and return them by mail. Experiments in the 1960 and 1970 censuses show a large reduction in this variance component when self-enumeration is used (U.S. Bureau of the Census 1968, 1970).

In addition, Hansen and Hurwitz encouraged work on coverage error. The Census Bureau has invested a large amount of time in investigating the effects of coverage error, both in censuses and surveys. After the 1950 census, using a model developed by Ansley Coale at Princeton University, the Census Bureau was able to measure the amount of undercounting in the decennial census at the national level, by age, race, and sex. This method, known as demographic analysis, showed that there was a differential undercount that affected blacks much more severely than whites (Citro and Cohen 1985). In addition, the Census Bureau started development of a post-enumeration survey to learn more about the uncounted population. At first, the Bureau relied on a "do-it-better" approach, but in recent years has turned

to a "do-it-again" approach. This latter emphasis will be used in the 1990 census. Similarly, coverage losses in surveys spurred work on ratio estimation procedures that would dampen the effect. Most Bureau household surveys use those procedures.

The Bureau of the Census now is well known for its work on measurement error. In addition to work on response error and coverage, it has encouraged work on time-in-sample biases that affect the estimates from surveys in which respondents are contacted more than once. The labor force survey, in which respondents are kept in sample four successive months, dropped for eight months, and then contacted for four additional months, has been carefully studied. Bailar (1975) showed the difference between the higher estimates of employment and unemployment for those in sample for the first time and those in sample for later times. These differences affect the levels of employment and unemployment, though probably not the estimates of month-to-month change.

These are only a few examples of the work begun at the Census Bureau on measurement errors. Now work is carried on at all the statistical agencies.

4. SEASONAL ADJUSTMENT

The history of seasonal adjustment in the government began with the efforts of Julius Shiskin when he was at the Census Bureau. He was responsible for introducing computerized seasonal adjustment. Now the X-11 method is used around the world.

According to Julie Shiskin, in the 1950's the Federal agencies were under pressure from the Council of Economic Advisors to produce seasonally adjusted time series. The Census Bureau got the first electronic computer dedicated to data processing, the UNIVAC I, in 1953 and Julie heard a lot about how difficult it was to program from Eli Marks who was in his car pool. It dawned on Julie that the computer could be used for making the seasonal adjustments, so he checked with a computer technician and found that it would take 1 minute to do a 10-year series. Of course, it takes less than that now.

Seasonal adjustment is still somewhat of an art form, since the X-11 program provides so many options, and the analyst can choose among them. However, there was skepticism at the beginning of this computerization about whether a machine could do what a skilled technician could. Julie decided to challenge the Federal Reserve Board. He proposed that they take any series and spend as much time as they wanted adjusting it. Then he would run the same series through the computer. Both series would be plotted and given, without identification of who did the adjustment, to a small, very distinguished group at the Federal Reserve Board who would judge the results. The result was a unanimous decision that the computer method was superior.

The government now seasonally adjusts thousands of time series annually. Model-based methods, because of computer limitations, seemed impractical for many years. Also, new seasonal adjustment factors were developed every year, based on historical experience. For example, a factor to be used in the computation of the seasonally adjusted figures for July would be developed in December of the preceding year. No new data based on more recent events were allowed to influence the adjustment. This made sense when it took several days to prepare punch cards and run the series. But within the last ten years, that method received more criticism and the method of concurrent seasonal adjustment was promoted. The time series staff at the Census Bureau, led by David Findley, did a thorough investigation of the merits of concurrent seasonal adjustment on Census Bureau series, and led the way for the adoption of that method by the Bureau.

The time series staff has also asked some very key questions that are central to seasonal adjustment. First, what kind of standard exists to judge whether or not a series should be seasonally adjusted? Second, given that there are several methods for adjusting time series, how do you evaluate the different methods? In a key paper, Bell and Hillmer (1984) question the need for seasonal adjustment if series can be adequately modeled. They also describe some criteria for evaluating seasonal adjustments. I must be quick to point out that the Census Bureau is not the only government agency that has done ground-breaking work in this area. In fact, one very useful accomplishment of the time series staff at the Census Bureau is to hold regular meetings of interested and involved experts throughout the government. Thus, people at the Federal Reserve Board, Bureau of Labor Statistics, Energy Information Administration, and the Bureau of Economic Analysis, to name only a few, all participate and keep up-to-date on new developments. Estella Dagum at Statistics Canada has led many very successful efforts, including the development of the X-11 ARIMA method.

5. DISCLOSURE AVOIDANCE

Whether or not one agrees with the Census Bureau on its policies about keeping data confidential one must agree that the Bureau has promoted disclosure avoidance techniques to protect data. Disclosure avoidance is an attempt to protect the answers of individual respondents. It has long been a problem in censuses, but is also a problem in surveys, especially surveys that are longitudinal in nature or where records exist that could be linked to the survey results.

Disclosure avoidance problems in the population censuses focus on disclosures that would occur from the publication of very small frequencies. These small numbers lead to the potential identification of single respondents or small groups of respondents. In addition, zeros in cells may also lead to disclosure. Disclosure in frequency tables is usually defined in terms of a threshold rule that states that disclosure occurs if, given any tabulation cell X, one can infer that the number of respondents in X is less than a predetermined threshold value. In 1980 decennial census publications this predetermined threshold value was defined separately for households and persons.

Methods for controlling disclosure in frequency count tables fall into three categories: suppressing all values, perturbing cell values, and replacing numeric cell values by intervals. Cell suppression insures that numeric values are not given and that inferences cannot be derived from manipulation of linear relationships between unpublished and published cell values. Data perturbation means adding or subtracting a small amount from most cell values so that inferences regarding the tabulated values cannot be made with certainty. The third method, replacing point estimates by intervals, is not useful for many data users for cross-classifications.

Cell suppression was the main technique used by the Census Bureau through 1980. Additive restraints along rows and columns of the table generate a series of linear constraints. Once the primary disclosures have been suppressed, mathematical programming is used as a disclosure audit on the table. Though this method was used on an ad hoc basis for years, Cox and his colleagues at the Census Bureau derived the mathematical underpinnings (Causey, Cox and Ernst 1985) and showed how complex cell suppression actually was.

Data perturbation methods, including random rounding, have been developed and used in the United Kingdom, Sweden, and Canada. All of these methods depend on adding or subtracting a small value, sometimes zero, from table cells, with a specified probability.

For data such as sales, value, inventory, and financial information from manufacturing and retail establishments, the Census Bureau is concerned about being able to identify the amount

from respondents. If a competitor reviews a tabulation and subtracts the amount for his firm, the amount for another respondent may be identified. Cell suppression techniques are used. The so-called (n, k)-rule states that X is a disclosure cell if a fixed number of respondents n account for more than a fixed percentage of k of the total cell value. This rule belongs to a class of cell dominance rules, all of which are additive.

Disclosure avoidance work is going on all over the world, primarily in government offices. No doubt this reflects the fact that these offices have serious problems that have been pushed to the fore by the demand for microdata.

6. A LOOK TO THE FUTURE

All four areas presented so far have relied on the development of mathematical models. Sampling, of course, relies on randomization methods, but the control of total survey error led to the formulation of a survey error model, first described by Hansen, Hurwitz, and Bershad (1961). That model and the experiments used to estimate the parameters were the basis for many policy decisions on the conduct of censuses and surveys.

Time series models are used widely around the world, replacing empirical methods such as the X-11. Researchers are now urging that time series methods become integrated with survey estimation methods to produce more accurate results. It will be interesting to observe how or whether this melding will take place.

Another area of active modeling within government agencies is to produce small-area data. Data are often collected for larger areas of aggregation, such as states, and then data needs are expressed for smaller areas, such as counties. Conferences have been held comparing and evaluating different techniques for producing small-area data. The Census Bureau used empirical methods to develop population estimates during the decade. Several models were explored as part of the undercount research at the Census Bureau, and much was learned about the problem.

Ad hoc methods for editing and imputation are now being carefully scrutinized and mathematical models are being developed. We shall undoubtedly see more modeling of this type in the future.

Thus, the future, as I see it, will be a further expansion of models. This is not to denigrate the empirical methods used now. Statisticians have always recognized that theory and practice go hand in hand. Empirical methods that seem to work lead to modeling and theoretical developments that are tempered by practical experience. The government agencies have many fascinating statistical problems that will lead the way, as they have in the past, in certain areas of statistical methodology.

REFERENCES

BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

BELL, W.R., and HILLMER, S.C. (1984). Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics*, 2, 291-317.

CAUSEY, B.E., COX, L.H., and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

CITRO, C.F., and COHEN, M.L. (eds.) (1985). *The Bicentennial Census*. Washington, D.C.: National Academy Press.

DUNCAN, J., and SHELTON, W. (1978). *Revolution in United States Government Statistics.* Washington D.C.: U.S. Government Printing Office.

HANSEN, M.H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2, 180-190.

HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

HANSEN, M.H., HURWITZ, W.N., and BERSHAD, M.A. (1961). Measurement errors in censuses and surveys. *Proceeding of the International Statistical Institute*, 38, 358-374.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W. (1953). *Sample Survey Methods and Theory*, Vols. 1 and 2. New York: John Wiley and Sons.

OLKIN, I. (1987). A Conversation with Morris Hansen. *Statistical Science*, 2, 191-210.

U.S. BUREAU OF THE CENSUS (1968). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: The Effect of Interviewers and Crew Leaders.* Series ER 60 No. 7.

U.S. BUREAU OF THE CENSUS (1979). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1970: Enumerator Variance in the 1970 Census.* PHC(E) No. 13.

COMMENT

G.J. BRACKSTONE[1]

1. Introduction

This paper confirms the significant contributions to statistical methodology made by the Bureau of the Census over the past 50 years. The four examples chosen by Bailar to illustrate these contributions are striking, not only in their intrinsic importance, but also in their variety. These are not variations of a single methodological breakthrough; they are fundamental contributions in four distinct areas. They are perhaps themselves illustrative of the wide variety and challenging nature of methodological problems faced by government statistical agencies – a variety and level of challenge that belie any suggestion that government statistics involves only the routine and the mundane.

Of particular interest in the description of these examples are the insights into the environments in which these developments came about. While the methodological contributions have themselves yielded benefits far beyond the original problems they were designed to address, the processes that led to these original contributions are themselves worthy of attention to identify the circumstances that need to exist to make such breakthroughs possible. I will return to this theme below.

During this same period, the Bureau of the Census was also making significant contributions to the automation of statistical processes. Having pioneered the development of punched card sorting and tabulating equipment in the earlier part of the century, the Bureau of the Census was responsible for the introduction of the first computer into a statistical agency in the 1950s. Subsequently in the 1960s, the Bureau also led the way in the automation of data entry by developing FOSDIC, a device for reading a microfilm copy of a marked questionnaire. Clearly the innovative contributions of the Bureau of the Census permeate many aspects of the work of government statistical agencies.

2. The Diffusion of New Methodology

Each of the contributions to statistical methodology described by Bailar originated with a real practical problem faced by a statistical agency. The need to collect additional data at reasonable cost and with acceptable timeliness motivated the development of sampling methods; the need to improve data quality by understanding, measuring and reducing non-sampling errors led to work in this area; seasonal adjustment developments seem to have been prompted by a need to speed up and standardize a skilled manual procedure; the problem of defining a rational and efficient process for ensuring the confidentiality of individual information in statistical outputs inspired the research on disclosure avoidance. Each of the many other examples that could have been cited share this characteristic of having had a real practical problem as catalyst.

The successful development of statistical methodology to address problems such as these is clearly of direct benefit to the statistical agency involved. But have these contributions had benefits more broadly? Have they added to the body of knowledge and methodology known as Statistics? It will be argued that these developments have had significant and broad benefits to statistical agencies engaged in the production of social and economic data, but that their impact on the subject of Statistics as treated in universities, while growing, has not been as influential as it might have been.

[1] G.J. Brackstone, Assistant Chief Statistician, Statistics Canada, Ottawa, Ontario.

Firstly, consider other government statistical agencies. In most countries the government statistical agency is a unique organization dealing with the problems of running regular large household and business surveys, integrating data from various sources, maintaining and analyzing time series, and making large volumes of data available to the public. (In this respect the United States is an exception in having several major organizations involved in this type of activity in different subject areas.) In most countries, therefore, statistical agencies have to look abroad for experiences similar to their own and for peer discussion and review. The network of interaction between government statistical agencies is extensive among developed countries. Contacts may be bilateral or multilateral. The long-standing and continuing exchange of information and experience between Statistics Canada and the U.S. Bureau of the Census is an example of the former. Statistics Canada has benefitted greatly from being able to adopt, and in some cases extend, statistical methodologies developed at the Bureau of the Census, including all of those described by Bailar; equally, I believe, the Bureau of the Census has benefitted from methodological developments at Statistics Canada.

On the multilateral level, several organizations provide regular fora for the exchange of information between statisticians in government agencies. These include the United Nations and its regional and specialized bodies, the International Statistical Institute, particularly its sections for Survey Statisticians and Official Statistics, and the professional statistical societies of several countries. In addition, both U.S.B.C. and Statistics Canada have instituted annual symposia or research conferences at which new developments and experiences are exchanged. All in all, this mixture of bilateral and multilateral contacts serves well to ensure that contributions to statistical methodology emanating from any agency – and many agencies are making significant contributions – are freely shared and utilized in other agencies.

But what has been the impact of such developments on the statistical profession outside government statistical agencies? Here we will use the specific examples cited by Bailar for illustration, though there are many other areas (some of them listed in Section 4) for which similar arguments would apply. In the case of sampling, the influence on the profession has been far-reaching. The topic of sampling from finite populations is now an established part of many university statistics curricula and is the subject of numerous textbooks. The developments initiated in a government statistical agency have been absorbed and extended by the profession. Indeed, some might argue that they have in some respects been taken far beyond the practical needs of survey- takers. In the case of non-sampling errors, the story is different. These developments have not yet led to a well-established body of theory and methods. That is not to say there have been no developments. On the contrary, there has been a wealth of work. However, much of it has been survey specific. It has improved, one hopes, many individual surveys, documented a great deal of experience, and generated a certain amount of applicable wisdom. But the topic has not yet found a secure niche in statistics curricula. Indeed, the accrued wisdom is often associated with particular areas of application (sociology, demography, *etc.*) rather than with Statistics as a subject.

Seasonal adjustment provides yet another story. With its origins as a rather empirical process used in statistical agencies, it has attracted increasing attention in recent years with attempts to provide it with a sound statistical basis. Bailar refers to some fundamental questions about objectives and yardsticks for seasonal adjustment that are now being addressed. Model based alternatives to the traditional X11 approaches are also being investigated. This is an area of statistical research that has attracted attention among time series experts in universities. Seasonal adjustment techniques clearly have applications well beyond government statistical agencies.

Finally, the most recent example that Bailar describes is disclosure avoidance. This is a problem largely confined to agencies operating under a confidentiality code that prohibits

divulgence of any identifiable individual information. Most of the research in this area is taking place in statistical agencies. The tools being used, however, tend to be from the fields of computer science, numerical analysis and mathematics. This is a relatively new field that has not yet attracted much attention outside government statistical agencies.

These examples show that methodological contributions from government statistical agencies not only solve problems for these agencies but can also lead to significant advances in the field of statistics more generally. Of course, not all such contributions have wide applicability and some may remain confined essentially to statistical agencies. A continuing challenge for government statisticians is to generate interest among other statisticians, particularly those in universities, in research problems arising in government work.

3. An Environment for Innovation

Innovative contributions rarely arise by chance. A suitable environment that allows ideas to develop and research to flourish is required. This is not always easy within an organization whose primary mission is the regular dissemination of data according to pre-determined schedules. Bailar refers to three reasons given by Hansen why sampling was accepted relatively quickly in the Census Bureau. In essence, these same three reasons define prerequisites for an innovative research environment in a statistical agency:

(a) management support in the sense of a willingness to invest in research activity;

(b) co-operative clients in the sense that successful research needs a particular application that represents the initial problem and sets the research schedule – the manager of this program has to be an enthusiastic guinea pig;

(c) competent research staff, not just in terms of expertise in particular areas, but also in terms of the ability to recognize problems susceptible to generalization and solution through statistical methodology.

While these three conditions will help to provide an environment conducive to research, further effort may be required to ensure that research results are in fact used, and used appropriately. This requires persuasiveness and good communication skills on the part of the statistician, as well as adequate institutional support for the new methodology.

4. Other Contributions

Bailar was not trying to be exhaustive in her examples of contributions to statistical methodology. It is worth noting some other areas of statistical methodology in which statistical agencies have made significant contributions. Some of these are mentioned as future topics by Bailar, but pivotal contributions have already been made. The following areas would find a place on a Statistics Canada list.

(a) **Methods for analyzing data from complex surveys** Of great relevance to users of most government statistics, these methods aim to adapt or replace traditional methods of statistical analysis that assume simple random sampling. This is an area of work that has attracted the interest of university researchers who have also made many contributions to the topic.

(b) **Record linkage** This technique is used in deriving statistics from administrative records, in micro-matches to assess quality, and in list frame maintenance. The development of a general theory for record linkage has provided a basis for software to support this activity. Most of the work on this topic has emanated from statistical agencies.

(c) **Editing and imputation** Widely used in many surveys, this technique lacked a sound statistical basis until theory was developed in the 1970s. Since then methodologies and systems have been developed to provide general facilities for performing these functions in a variety of surveys. This topic has generated substantial interest and further work outside statistical offices.

(d) **Small area estimation** In recent years the production of estimates for areas smaller than could be supported by direct estimation from sample surveys has received increasing attention. Statistical offices have developed a variety of methods to address this problem and university researchers have participated actively in this work. To date the utilization of such methods for production purposes has been limited, partly due to lingering concern about the probity of government agencies producing model-based estimates.

(e) **Statistical use of administrative data** As another means of reducing data collection costs, the statistical potential of existing administrative records has been exploited. Such sources present a different array of coverage and data quality problems, from those experienced in surveys. While administrative data may be used alone to produce statistical data, they may more effectively be used in combination with survey or census data in estimation systems that take advantage of the relative strengths of each. Most of this work has taken place in government statistical agencies.

5. Future Areas

In looking to the future, Bailar foresees increased use of models. This is almost certainly correct as statistical agencies strive to extract the maximum information out of existing data and minimize the increasing costs of data collection. In particular, she refers to the melding of time series methods with survey estimation methods, an area now being explored in several statistical agencies. I would add three other domains of activity in which we might look forward to significant developments in the long run, each of them requiring an interaction of statistics with other disciplines.

The first is the application of expert systems to certain activities in government statistical agencies. To use an example already discussed, the choice of the appropriate options or models to use in seasonally adjusting a time series could well lend itself to such an approach. The second area is the use of cognitive methods for understanding and improving the response process. Work in this area is underway at a number of statistical agencies. Drawing on the expertise of psychology, it may provide a basis for enabling statisticians to develop better models of the response process – probably the least well understood component of the survey process. The third area is the development of integrated statistical information systems that combine models of social or economic systems with databases on which the impact of different policy assumptions can be simulated. Such systems serve to facilitate the use of an agency's data for policy analysis, and also help it to recognize data gaps in current programs.

To echo Bailar's conclusion, the problems are fascinating and there are more than enough to go around.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 63–79
Statistics Canada

Rolling Samples and Censuses

LESLIE KISH[1]

ABSTRACT

Rolling censuses combine F nonoverlapping periodic samples of $1/F$ each, so designed that cumulating the F periods yields a complete census of the whole population area with $F/F = 1$. Intermediate cumulations of k samples would yield samples of k/F for more timely uses (annual or quinquennial censuses). Area sampling frames would cover the national territory for naturally mobile populations. These methods may often be preferable to other alternative methods for censuses, also discussed. *Asymmetrical cumulations* are also recommended to counter the problems of small sample cells for area domains (provinces, regions, states) common to most countries and to other population units. *Split-panel-designs* offer another use for cumulating periodic surveys by combining nonoverlapping portions $a - b - c - d -$ with panels p for partial overlaps, $pa - pb - pc - pd -$, for multipurpose designs.

KEY WORDS: Periodic samples; Time sampling; Cumulations; Split-panel designs; Asymmetrical cumulations; Multipurpose designs.

1. INTRODUCTION AND DESCRIPTIONS

Several uses and methods for cumulating data from periodic samples are discussed below. This has been a rather neglected subject, as the literature on periodic and rotating samples has concentrated on the statistics for net changes and for current ("cross section") estimates; not on cumulations. The first concern here is on rolling censuses and samples, and let me attempt a definition of *rolling censuses*: a combined (joint) design of F separate (nonoverlapping) periodic samples, each a probability sample with fraction $f = 1/F$ of the entire population, so designed that the cumulation of the F periods yields a detailed census of the whole population with $f' = F/F = 1$. Intermediate cumulations of $k < F$ periods should yield rolling samples with $f' = k/F$ and with details intermediate between 1 and F periods. We may appreciate that definition by looking at examples and counterexamples. We shall also examine possible variations that would satisfy the definition and conflicting needs that rolling samples can be aimed to meet.

Imagine a weekly national sample, each with *epsem* selection rates of 1/520, and so designed that in 520 weeks they are "rolled over" the entire population and the cumulation yields a complete census of the population averaged over ten years. Each year would yield national and local samples with selection rates of $52/520 = 1/10$. The design would combine weekly national samples into an averaged decennial complete census, and into sample censuses of ten percent each year.

The Health Interview Surveys of the National Center for Health Statistics (1958) cumulate 52 weekly samples of about 1,000 households each. These samples select about $f = 1/80{,}000$ weekly; thus 520/80,000 represents cumulations of nonoverlapping periodic samples over ten years. But they are confined to a set of PSU's for reasons of cost chiefly, but also for better estimates of net change and for current estimates. However, *rolling samples* may better be reserved for samples designed for maximizing (increasing) the spread (representation) of the samples cumulated over national (or broad) populations. The words in the parentheses indicate that rolling samples constitute a special case of the more general *cumulated periodic samples* and that the boundary of the subset need not be precisely clear.

[1] Institute for Social Research, The University of Michigan, Ann Arbor, U.S.A. 48106.

For overlapping between periodic surveys, the requirements for the selection of units of cumulated designs are diametrically opposed to the requirements for the objectives and substantive content of the interviews (the observations, variables). The content of the surveys must be as similar, standardized, identical as possible for the cumulations to be meaningful. Using periodic panels of the same elements for different contents could broaden the scope of surveys, but would not contribute to increasing the sample size for survey statistics. Most periodic surveys collect similar variables, though some may also have other contents attached at times. However, changes of methods, questions, and variables would cause conflicts and problems. Perhaps such changes should be introduced only with extended intervals of "splicing", using both the new and the old methods to study the differences. These problems are fundamentally similar to those faced when measuring differences from periodic surveys, but they seem more novel. I insist (Section 6) that solutions to such problems must be tailored to specific situations.

On the other hand, the cumulation of the same elements (persons, households) does not increase proportionately the sample size (base), and panels of the same elements would not help rolling samples. Many periodic surveys (*e.g.*, labor force surveys of Canada, the USA *etc.*) have partly or largely overlapping fractions of segments (ultimate clusters), and those tend to contribute little toward increasing the sample size. Even in surveys with nonoverlapping segments (like the HIS of the NCHS (1985)), the segments are confined to the same first stage (and second-stage?) units; in these the positive correlations (clustering effects) tend to reduce the "effective" sample sizes for overall statistics. Furthermore, those periodic samples, confined to samples of primary units fail to meet the needs of rolling samples for spreading over the entire (national?) population.

A few more remarks may help to broaden our frame of reference. (1) The discussion often assumes area sampling, but the concept can be generalized to other frames. (2) Equal selection rates for elements are often used, but cumulations may be modified to unequal selection probabilities. (3) The concept may be generalized from regular periodic samples to cumulations over less regular periods. (4) Cumulations over the entire time span (year or ten years) come most readily to mind, but we may envisage systematic sampling of the span; *e.g.*, labor force surveys cover only single weeks of the months over the year.

2. ALTERNATIVE METHODS FOR CENSUSES

Rolling censuses would be expensive, and the reason for such an innovation should include the acknowledged relative weaknesses of the decennial censuses now widely used, and of sample surveys and administrative registers, which are proposed at times as possible alternatives. The chief reason for censuses is the need for detailed information, especially for small areas; and the chief weakness of decennial censuses is their obsolescence between censuses and their great total cost that prevents more frequent censuses. Sample surveys have many advantages for national statistics and for large regions, but they lack geographical and other details. Good registers are rare and they provide few variables beyond a few, bare demographic data.

Decennial censuses of population, housing, agriculture, industry and others, first and foremost, have spread into most countries in the last two centuries, and especially in the last two generations with the help of the United Nations State Statistical Office. In addition to detailed data for small domains, censuses often may obtain better coverage than samples, due to the concentrated publicity and the national "ceremony" connected with censuses; the Chinese census of 1982 is a good example (Kish 1979, 1989). The efforts of the census also yield *lower unit costs* (for short forms) than surveys, but much *higher total costs* than sample surveys, because of much greater size. At 2.6 billions, the 1990 censuses of the USA will cost $10 per

capita or $30 per household. That cost of about half to one hour of the median hourly wage per capita (once in ten years) seems to hold in international comparisons, though the number and complexity of census variables is one of the cost factors. Rolling censuses would probabily be proposed and designed for surveys fairly rich in the numbers and complexity of variables. In Canada 260 weekly samples of 32,000 households would cumulate to the national population. In the USA 520 weekly samples of 160,000 would be needed by decennial cumulations to 80,000,000 households; the CPS surveys have 100,000 with state supplements.

No detailed comparison of decennial censuses with rolling censuses is possible here, but the issue of *timeliness* must be mentioned, because that is the chief issue in the comparison. Up to now the periods for using data from decennial censuses have varied from a start of 1-4 years to 14 year or more. Even with faster computers the start is slower for complex social statistics than for mere head counts; and the obsolescence over the ten intercensal years becomes worse with higher population mobility in our modern civilization. The biases due to obsolescence will be monotonic, if not linear, functions of elapsed time. The sizes of the biases will differ with variables, populations, *etc.*; but they will be present and considerable, I believe; often perhaps greater even than the famous biases due to under coverage (Kish 1981, 1979).

Increasing and rapid obsolescence of decennial census data should chiefly motivate the searches for alternatives, such as in *A Study on the Future of the Census of Population: Alterative Approaches* (Redfern 1987). "A serious weakness of the census is that it occurs relatively infrequently". About a "rolling census" it states: "The merit of this proposal is that ... a much smaller, better trained organization and more experienced staff could be deployed both for the fieldwork and for processing ... the public awareness of the rolling census would not be highly peaked. Whilst that might well lessen the risk of public protest, the reduced publicity would adversely affect the level of coverage achieved ... (The method) would complicate the interpretation of the census results, especially comparisons between areas. Simultaneous national coverage, one of the virtues of the census, would be lost. The idea of a rolling census has not yet been developed and applied".

Most countries will probably still need censuses in 2000 AD. They are being replaced by population registers in the Nordic countries and still need to be introduced in some Third World countries in 1990. They have been stopped by opposition and by obstacles in a few. But most countries need and will have them in 1990. They have been a great and useful invention – like the steam locomotive, and at about the same time. However it is possible that the censuses also may be phased out gradually by some of the alternatives here considered.

Quinquennial annual censuses have been proposed, and quinquennial censuses have been initiated or carried out in a few countries, including Canada and Turkey. But these are not destined for quick acceptance, I suspect. They seem too costly: ten percent samples in two countries had half of the costs of complete censuses. Also they still leave a great deal of obsolescence. On the other hand, much smaller (*e.g.*, 5 or 1 percent) yearly sample censuses would fail to offer enough geographic detail. The one percent "microcensus" of West Germany provides yearly sample data. China had a one percent census in 1987; their yearly samples of 1/2,000 (also about 500,000 people) collect chiefly fertility data only (State Statistical Bureau 1987; Kish 1989). Quinquennial censuses are not frequent enough and yearly censuses would be too costly.

Administrative registers provide a great deal of diverse data in many countries, and they are likely to spread in the future. Excellent *population registers* exist in the Nordic countries of Sweden, Norway, Denmark, and Finland, and perhaps in some other countries of Northern Europe. Their completeness is based on cooperation, motivation (with social incentives), and literacy; in a few cases they are replacing censuses with data from the population registers. In other situations their coverage, quality, and updating are far from adequate. We can expect

future improvements in the quality, spread, and use of population registers but not quickly and not widely. We should not expect them to replace censuses even in developed countries like the USA and Canada, and their use in less developed countries soon is even less likely (Redfern 1989).

Furthermore, even after population registers become adequate in quality and coverage, they will contain and supply only a few, bare demographic variables: head counts, age, sex and little more. Thus, they will fail to meet the demands of modern society for richer sources of statistics. For these the registers will serve only as auxiliary variables.

Synthetic, ratio regression, and raking estimators are being used increasingly for *small area statistics* (Platek *et al.* 1987; Purcell and Kish 1980). Census data are usually obsolete, data from registers inadequate, and sample data lack details for small areas. The weaknesses and strengths of the three methods are complementary, hence combining the advantages of the three methods seems like good strategy. This is the common purpose of the several methods of *small area estimation*: to provide estimates for small areas and for other small domains that are current, accurate, and relevant.

These methods are now being used for local area estimates of population counts for the intercensal years, in order to compensate for the obsolescence of the decennial censuses,thus sometimes called *postcensal estimates*. They also have other uses in increasing numbers, *e.g.*, they have been proposed to compensate for undercount biases. However,those methods have all combined censuses with sample surveys and registers. Therefore,they should not yet be considered as alternatives to censuses. Nevertheless, we may raise the question whether rolling censuses would perform better or worse overall than decennial censuses in those combinations. The answer is uncertain, but I believe that the balance of variance components would favor rolling censuses in most cases. However, theoretical as well as empirical investigations will be needed to decide this question as well as several others here.

Partially overlapping samples from multipurpose designs must be considered because they exist in many countries for several purposes and they absorb some of the funds available for national statistics. These multipurpose surveys often provide labor force statistics and other valuable data. They vary in parameters between countries but they also have several basic features in common with those of the USA and Canada. They are periodic samples with overlaps that are constant and for fixed periods (but all three parameters differ between countries). They use area segments for bases, but not panels of households (movers are not followed). The overlaps are usually large and these are generally justified with references to reductions of variances from positive correlations in the overlaps. But an even greater advantage of overlaps may be the lower costs of interviewing in later calls, especially where telephone calls follow first calls on foot. These "rotation designs" have dominated practice and literature and they represent an important innovation (by H.D. Patterson 1950 and R.J. Jessen 1942). They are designed for measuring net changes and current (level) statistics, but not for cumulations. However, the variances (per household) would not be greatly increased for overlaps of even a small fraction (< 0.3), when compared to the large overlap (> 0.7) commonly used. This is particularly true for many variables like being unemployed, which have low correlations between periods. Furthermore the overlaps could be changed in other ways (Section 5). Therefore it is possible that these surveys could be combined with the cumulations needed for rolling samples and censuses.

3. CUMULATIONS OVER TIME AND SPACE

Changes in populations and in their variables are often recognized as of three kinds: "secular" trends, which are more or less smooth and monotonic, like "growth"; periodic and "cyclical", such as seasonal fluctuations; and irregular variations which are difficult to describe

and often treated as "random". Designs for cumulating, averaging, and sampling over temporal variations face psychological obstacles that differ from our acceptance of designs for variations over spatial variations. Spatial variations can be large and sometimes accountable, but more often irregular. However, we have learned to accept samples, averages, and cumulations over them in population (national) aggregates and averages.

The psychological blocks still facing rolling samples and censuses may be countered with both theoretical and pragmatic arguments. The theoretical and philosophical arguments are hinted at above and in later discussions of alternatives (Kish 1987, 6.1B). The pragmatic and empirical arguments may be buttressed with several types of uses we recognize as common and successful. The same periodic samples for obtaining current data and for measuring changes can also be used for aggregates needed for spatial and domain details. Furthermore, by averaging (over a year or longer) the temporal variations (seasonal or cyclical or erratic) are smoothed over in the moving averages.

Retrospective data. "Children ever born" to women who completed fertility over the entire fertile span of 30 years may represent an extreme for retrospective spans; but other individual interview data aggregated over life spans include serious diseases, education, *etc.* Interviews aggregated over yearly spans include farm production, work history, income, home and auto purchases. Of course, all these data have imperfections, which differ across variables, respondents, methods, *etc.* But even cumulations over a week or over a day (such as purchases of bread or cigarettes) have errors. *Multiround surveys* are used for cumulating short term data; for example, births during the past month have been cumulated from 12 monthly samples over the year.

Cumulating rare elements from periodic surveys has often been used to deal with these difficult and expensive problems. The topic has been dealt with and illustrated in publications on rare items (Kish 1965 11.4; Kalton and Anderson 1986). *Statistics for small domains* may also benefit from cumulations, and single years of birth may exemplify such small domains, which consist of "crossclasses". But geographical and administrative units are "proper domains"; for these the periodic samples are not adequate, because those domains need the designs of rolling samples or censuses.

Cumulations from periodic samples. The Health Interview Survey (NCHS 1958), described above, may be the best known example with yearly cumulations of weekly samples of about 1,000 households from nonoverlapping area segments. It is designed for *multipurpose* objectives (like most periodic surveys) including cumulations for some rare diseases, but also estimates of current levels and net changes. It provides some estimates for larger domains, as well as national estimates for the common diseases. To convert it into a rolling sample, by increasing the spread of the yearly samples, would increase field costs, especially in that portion (about 30 percent only) where the PSU's are counties (not self-representing).

A traffic survey provides an interesting example of cumulations, because the population is very mobile within the sampling frame of sampling units of locations x hours (Kish, Lovejoy and Rackow 1961). The general concept is applicable to nomads and other mobile populations. It may also serve less mobile general populations over a longer period, such as the decennial spread.

The earliest cumulation I found is for a sample of California in 1952 (Mooney 1956). "The samples were selected in such a manner that they resulted in a uniform overall sampling rate of 1 in 385. For purposes of enumeration, the sample was divided into 52 equal subsamples, and a different subsample was enumerated during each week of the survey year. Consequently, each week's enumeration was based on a sample of 1 in 20,020". For smaller states (populations) and/or larger samples one may imagine weekly samples of 1/520, and complete rolling samples in the 520 weeks of the decennial census period. It is likely that such rolling samples have been designed for smaller populations.

The above examples refer to nonoverlapping periodic samples. Cumulations from partially overlapping samples have been used, but with the "effective sample sizes" reduced by the amount of the overlap (Ericksen 1974). Furthermore, this paper concerns cumulations of individual cases, but periodic or repeated surveys may also be used for *combining statistics* from them (Kish 1987, 6.6) as in "meta-analysis".

4. ASYMMETRICAL CUMULATIONS

This term denotes a proposed method of cumulation for problems that arise because "natural" subpopulations generally vary greatly in size. For example, I have been faced within the past few years with ranges of 50 or even 100 to 1 among the provinces (or states) of Canada, USA, Australia and China; and those ranges of relative sizes are similar for the provinces of most countries. Those inequalities arise because administrative units tend to be created roughly equal in areas, but spread over lands with highly unequal population densities. They also exist for districts, counties, *etc.* within most provinces. They also arise for other social units and social organizations, like firms, hospitals, universities. But not for all: military units, census enumeration districts and elementary schools are created roughly equal.

For many other frequency distributions rough equalities of classes are created with traditionally accepted cumulations over roughly logarithmic scales; *e.g.*, income, city size, *etc.* are often tabulated in classes like 10-25, 25-50, 50-100, 100-250, 250-500, 500-1,000, 1,000-2,500, *etc.* This shows a sensible method of cumulation that creates roughly equal cells on a roughly logarithmic scale, and they are traditionally accepted and understood, although highly asymmetrical.

Note also that cells in tables for sample data are *generally cumulated over both space and time*. For example, monthly surveys of labor force often show labor force statistics cumulated over the month (or over a week as a "sample" of the month), and also over the provinces (from a sample of sampling units). Quarterly and yearly statistics show further cumulations, as do the national statistics. The spans of cumulations must balance three parameters of restraints: the span of the reference period that may be relatively flexible; the domains of subpopulations, which may be more rigid, like provinces; and the sample size expressed in sampling units and variance components. Other variables, such as cost factors and "required precisions", tend to be expressed through the basic three parameters of cell size.

Decennial censuses of the population counts represent extremes by emphasizing locational detail: persons are placed in homes as of the reference date (April 1 in the USA). But yearly and longer cumulations are possible for income, *etc.* Time gets sacrificed in obsolescence, and sample sizes and costs in complete coverage. At the other extreme are monthly sample surveys for labor force and health variables, and myriad other variables, where the emphasis is placed on timelines and reduced costs, but at great sacrifice of spatial detail.

Population inequalities between provinces impose severe restraints on timeliness and sample sizes. Often higher sampling rate are introduced for the smaller provinces, but such "optimal" selection rates bring disadvantages in increased variances both overall and for cross-provincial "crossclasses" (age, sex, *etc.*) (Kish 1988, Section 5; Trewin 1987). Thus those mildly unequal rates fail to solve conflicts in provincial sizes of 50:1 or 100:1.

Because of those conflicts the tables for monthly surveys commonly present cells for small provinces with inadequately small sample sizes. Two alternative procedures have been advanced and practiced for such small cells. A. Release the same data for small cells as for large cells, and let the reader (user, consumer) beware, *caveat emptor*, with perhaps warnings posted

to appendixes to sampling errors. B. Don't release, but suppress small cells, leaving them blank, after applying some declared curtailing limits. Readers may be directed to other released publications, based on cumulated data (quarterly, annual).

Asymmetrical cumulation proposes a compromise between symmetrical releases (A) and asymmetrical suppression (B).

C. Asymmetrical cumulation proposes to release for small cells the *specified* cumulations of periodic data. These cumulations may be flexible: for example, quarterly for small cells and yearly for very small cells, instead of the monthly data for large cells. The readers may be notified (with * or italics or other signs); thus they may choose either C (cumulation) or B (disregard).

AC. This procedure would allow readers to choose either A or B or C by publishing *both* the current monthly data A and the cumulated C data.

Procedures B and C have the disadvantage that the cells do not sum to the marginals. But AC like A do sum to the marginals. Some iterative method could overcome these disadvantages of B and C.

5. MULTIPURPOSE SPLIT PANEL DESIGNS (SPD)

In order to find adequate funds for rolling samples and censuses it is desirable to consider how they could be combined with the periodic surveys now being funded and conducted in many countries. These are either monthly or quarterly surveys (sometimes yearly or weekly). They are typically partially overlapping samples designed for improved estimates for current level and net changes. However they are not designed either for cumulated rolling samples, or for panel studies based in the overlaps. I proposed SPD as the design for providing data for all those four purposes; and also for some fringe benefits (Kish 1987, 6.5).

a. Combining two separate periodic samples forms the basis of SPD: to add a panel p to a parallel series of nonoverlapping samples $a - b - c - d$ *etc.*, with the combination then denoted as $pa - pb - pc - pd$ *etc.* The panel p provides individual (micro) changes and the nonoverlaps can be cumulated into larger samples and rolling samples. The combined samples provide the partial overlaps best for current estimates and for net changes; thus they can replace the usual rotating samples. This combined use is a main feature of SPD,together with the provision of a flexible and potentially large sample of nonoverlapping portion for use in cumulating samples.

b. The designs for p and for $a - b - c$ can be separate and distinct, each "optimized" for its own objective. But they must also be combined for joint estimates of net changes and current levels; and for that purpose the populations covered and the measurements used must be similar enough for the combination.

c. SPD has considerable advantages because its overlaps exist for *all* periods, whereas they are rigidly fixed in classical rotation designs. This advantage is clear and important for net changes because it exists for all desired comparisons. But it also exists for current levels, because the correlations may differ among variables.

d. Including proper panels p of elements necessary for measuring individual (micro or gross) changes would be a great advantage for SPD over partial overlaps now used. However, the other features can be satisfied with overlaps p' of area segments as at present. Furthermore a modest and slow rotation can be built into the design of either the panel p or the overlap p', so as to retain most of the gains from covariances and from panel information. Perhaps some alternation may be introduced to reduce panel fatigue or deterioration. Several surveys have used *both* the overlap p' *and* panel p by following as many movers as possible. Most

households belong to both samples. The extra cost for the panel depends on the proportion of movers and their cost (Kish 1987, 6.2, 6.4).

e. The advantages and problems of panel interviewing pose difficult problems, with a large and varied literature and conflicting results (Kish 1987, Sections 6.4, 6.5). The number and spacing of reinterviews that are possible, desirable, and reliable need to be established.

SPD has an advantage in separating the panel p whose cumulated data may be checked against the nonoverlaps for "panel biases", and perhaps even for adjustments of biases when those are measured adequately.

Another useful modification may be to recruit sampling units into the panel by different ("optimal") selection rates on the basis of their being "screened" in the nonoverlaps.

f. The size of $a - b - c - d$ need not always be the same; this flexibility of SPD, which differs from the rigidity of rotating designs, may be used for needed sample enlargements or for cost retrenchments. Such changes would raise weighting problems (solvable) for cumulations.

g. The relative size of the panel p against the nonoverlap $a - b - c - d$ portions depends on feasibilities and costs and needs study (Section 6). For individual changes we need larger p, but for cumulations larger $a - b - c - d$. The larger p portions now common may be favored by lower field costs for telephone reinterviews.

Lower values of p than are now common are good enough for current levels and for net changes with weighted estimates; the optima are insensitive and p between 1/4 and 1/2 are all nearly best; lower p may also be used where the emphasis lies in nonoverlaps $a - b - c - d$ for cumulations.

6. CONCLUSIONS AND QUESTIONS

Cumulated samples provide the bases for four new methods proposed here: rolling samples, rolling censuses, asymmetrical cumulations, and split panel designs. Rolling samples have been designed, but the other three still await practical applications. Meanwhile we should welcome methodological developments that would outline the parameters of feasibility.

However, the chief tasks for these methods must be found in the details of specific situations rather than in theoretical generalities. The factors of costs, variances, biases, feasibilities, and public acceptance for novel procedures must be worked out specifically for each situation. We can do no more than raise a few questions as examples, in addition to those raised implicitly or explicitly in the preceding sections.

1. For rolling samples and censuses what kinds of moving averages may prove most useful? For national aggregates the latest month (or quarter or year) may receive the full weight. But for small local areas the data may be cumulated over ten years; with equal or with increasing weights? Are "shrinking" (Stein-James) estimators useful?

2. How to deal in the aggregates with changes in the population, in methods, in variables?

3. For asymmetrical cumulation similar questions arise. Should the latest monthly estimates (A) be printed together with the cumulated (C)? Methods are needed to make the cells and the marginals consistent.

4. For the split panel design, how large should the overlap (p) be? Can it be a panel or merely overlapping segments? Or must we, can we, have both? How does it depend on the correlations for diverse variables? How do we balance the four chief purposes of periodic surveys?

There will be other interesting questions but this essay must come to an end before they do.

REFERENCES

ERICKSEN, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association*, 69, 867-75.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304.

KALTON, G., and ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society* (A),149, 149-52.

KISH L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

KISH L. (1979). Samples and censuses. *International Statistical Review*, 47, 99-109.

KISH L. (1981). *Using Cumulated Rolling Samples*. Washington: Congressional Research Office, 80-528-0.

KISH L. (1987). *Statistical Designs for Research*. New York: John Wiley and Sons.

KISH L. (1988). Multipurpose sample designs. *Survey Methodology*, 14, 19-32.

KISH L. (1989). Developing statistics in China. *Journal of Official Statistics*, 5, 157-69.

KISH, L., LOVEJOY, W., and RACKOW, P. (1961). A multi-state probability sample for traffic surveys. *Proceedings of the Section on Social Statistics, American Statistical Association*, 227-230.

MOONEY, H.W. (1956). *Methodology in Two California Health Surveys, San Jose (1952) and Statewide (1954-55)*. U.S. Public Health Monograph No. 70.

NATIONAL CENTER FOR HEALTH STATISTICS (1958). *Statistical Designs of the Health Household Interview Survey*. Washington: Public Health Series, 584-A2, 15-18.

PATTERSON, H.O. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B, 16, 140-149.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.E., and SINGH, M.P. (1987). *Small Area Statistics*. New York: Wiley-Interscience.

PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas. *International Statistical Review*, 48, 3-18.

REDFERN, P. (1987). *A Study of the Future of the Census of Population: Alternative Approaches*. Luxembourg: Statistical Office of European Commission No. ISBN 92-825-7429-6.

REDFERN, P. (1989). Population registers: some administrative and statistical pros and cons. *Journal of the Royal Statistical Society*, A, 1-41.

STATE STATISTICAL BUREAU (1987). *The 1987 Nationwide One-percent Population Sample Survey*. Beijing: State Statistical Bureau.

TREWIN, D. (1987). Estimation of trends and time series models from continuing surveys. *Bulletin of the International Statistical Institute*, 46.

UNITED NATIONS (1981). *Principles and Recommendations for Population and Housing Censuses*. Series M No. 67 (E.80.XVII.8).

REDFERN, P. (1989). Population registers: Some administrative and statistical pros and cons. *Journal of the Royal Statistical Society*, A, 152, 1-41.

COMMENT

FRITZ SCHEUREN[1]

The statistical literature has neglected the idea of cumulative samples. Leslie Kish, in several previous papers and in the present one, has tried to rectify matters. Ever forward-looking and practical, he makes a persuasive and compelling case for more work on the design and analysis issues raised by cumulation.

His writing is so down-to-earth that readers may miss the fact that Kish is not just advocating a few minor additions to the already large supply of survey designs and estimation methods. He asks us to look very hard at the topology of the space/time/content trade-offs in surveys – especially in censuses. In fact, Kish seems to be advocating what might be called a "paradigm shift" in census-taking, at least in developed countries like Canada and the U.S.

The word "paradigm" deserves some elaboration (Barker 1988). A paradigm is a way of thinking and then doing, a pattern of belief and behavior, a way of seeing reality and using that sense to accomplish something. Paradigms are common – the way we get to work would be a humble example. Conventional census-taking, under this definition, could be characterized as a major scientific and technical paradigm.

As long as our paradigms work well for us, we tend not to change them. Occasionally, however, paradigms break down and have to be replaced. The bridge goes out and we need to find another route to work. As Kuhn pointed out in his seminal book on the structure of scientific revolutions, paradigms break down in science, as well (Kuhn 1970). Perhaps the most famous example of this is the revolution in the thinking of astronomers that occurred when the Ptolemic earth-centered view of the universe was replaced by the Copernican view of an earth that revolved, with the other planets, around the sun.

Kish, in his paper, argues that major problems exist with the conventional census-taking paradigm. He then goes on to consider two possible alternatives: rolling censuses and administrative registers. My objective here will be to round out and occasionally balance Kish's presentation of these topics.

Conventional Census-Taking

Conventional censuses, like those in Canada and the U.S., continue to do many things very well. Indeed, at present, we have no adequate substitute for them; nonetheless, Kish's point of view on the need for at least some change seems compelling. Rising costs are a big factor. There have been many improvements in census-taking in this century; still, in both Canada and the U.S., total costs and even costs per person have risen significantly:

- The 1990 decennial census in the U.S. is budgeted at about $10 (U.S.) per person. Even adjusting for inflation, this is a four-fold increase over what the per capita expenses were in 1960. Item content differences between the two censuses are small and essentially not a factor in explaining the difference. Both the 1960 and 1990 Census, for example, asked only 7 population questions of everyone (U.S. Bureau of the Census 1989). The Census long-form sample in 1960 contained 35 questions and was to be completed by 25% of the population. For 1990, the Census long-form sample was given to 16% of U.S. households and had 33 questions.

[1] Fritz Scheuren, Director, Statistics of Income Division, Internal Revenue Service. The opinions expressed here are those of the author and do not necessarily represent the position of the Internal Revenue Service.

- The situation in Canada is similar with regard to the costs of census-taking. For example, the 1991 Canadian Census is budgeted at about $9.50 (CAN) per person. Like the U.S. Census, there are again just 7, albeit somewhat different, population items that are asked of everyone. Like the 1990 U.S. Census, questions on housing are included for everyone (2 in Canada and 7 in the U.S.). In Canada, a 20% long-form sample will be employed in 1991. The Canadian long-form questionnaire has 45 items for 1991. The 1961 census in Canada was quite different from that planned for 1991 and, thus, meaningful cost comparisons are hard to make. Nonetheless, looking back 30 years in Canada, the same long-term trend in census-taking costs seems to exist; however, per capita costs have been roughly the same – even declining slightly – in the last two or three censuses.

The U.S. Census Bureau has looked at the growing cost of conventional census-taking and concluded that a major change may be needed (Browne 1989). Labor costs have grown appreciably in recent decades in both Canada and the U.S. Technological improvements have not been great enough to offset these costs, though some, like TIGER (Topographically Integrated Geographic Encoding and Referencing) and CATI (Computer-Assisted Telephone Interviewing), offer promise. Greater attention in the U.S. to improved population coverage is another important factor (Anderson 1990). The degree of public cooperation in the census also seems to be dropping, at least as reflected by the poorer than anticipated mail response rate for the 1990 U.S. census. (It should be noted that, in Canada, public cooperation has fluctuated, with no clear tendency.)

Increasing cost is not the only major problem facing conventional census-taking. Perhaps of even greater importance, as Kish notes, is the growing rate of obsolescence of the information collected. The combination of rising costs and growing information obsolescence has had the effect of reducing the benefit/cost ratio for conventional censuses steadily and dramatically.

To obtain more frequent small area data, some countries have introduced quinquennial censuses. For example, in Canada this was first done nationally in 1956. Budget problems led to the 1986 Canadian Census being cancelled and then reinstated. Indeed, it is unclear whether there will be a Canadian Census in 1996. While a quinquennial census was also legislated in the U.S., funds were never made available.

Rolling Censuses

As Kish rightly observes, conventional census-taking, of necessity, must sacrifice both timeliness and item content (on a 100% basis) to achieve complete spatial detail and high population coverage.

One of the alternatives that Kish asks us to look at is a "rolling census." His proposal envisions the sampling of a country over a decade in such a way that every area is eventually covered. In its purest form, space and time become a single dimension and content remains fixed, such that, at decade's end, we have obtained cumulative information on the entire country for a given set of items.

The chief advantage of a rolling census is that it can avoid the problem of information obsolescence at national and major subnational levels. For small geographic areas, though, there would, of course, still be only one observation per decade. Unlike a conventional census, comparisons among small geographic areas would be very difficult to interpret because the data are being collected at different points in time (Fellegi 1981).

For a rolling census or survey, unit costs could be higher, as Kish notes, than in a more conventional enumeration (indeed, *ceteris paribus*, maybe even higher than the cost of existing survey efforts). In an age of fixed or declining resources, therefore, it might not be possible

to do a complete "enumeration" each decade, even if content were significantly scaled back. Rolling samples would seem to have their greatest attractiveness not as a replacement for conventional censuses, but, say, as part of a strategy to link together census-taking with ongoing surveys and local area population estimates for the intercensal years (Herriot, Bateman and McCarthy 1989).

Both the United States and Canada employ monthly surveys to estimate the national (and some subnational) labor force characteristics. The Canadian Labor Force Survey (LFS) of 64,500 households covers 0.67% of the total Canadian population each month. "Given the rotation pattern in effect for the LFS, the 0.67% sample per month rolls up into a 6.7% sample of unique households over a 5-year period" (Drew 1989). In the Canadian context, at least, Kish's proposal may be feasible. A sample survey vehicle could be designed, with some reduction in the month-to-month household overlap, which could achieve many of the benefits he has stated for a rolling sample, while also meeting the information needs currently met by ongoing household surveys (Drew 1989). This sample would not replace the 100% census count data, itself, but, might be a *partial* substitute for Canada's 20% long-form census sample.

Because the United States has a population about 10 times larger than Canada, the tradeoffs involving rolling samples and overall country coverage are not as favorable as they are in Canada. The U.S. Current Population Survey (CPS), for instance, at about 60,000 households, covers only .06% of the total U.S. population monthly. Even if cumulated over a *whole* decade (but, with no change in its rotation pattern), the CPS would cover just roughly 1% of all U.S. households. This does not compare well in size to the overall 16% long-form sample being conducted as part of the 1990 U.S. Census.

To bring the rolling sample population coverage nearer to the 1990 U.S. decennial sample, major changes in the CPS rotation pattern, like those Kish asks us to look at, would be needed. Other U.S. Census Bureau surveys might also have to be redesigned if the objective were to achieve even a partial substitute. Despite these changes, moreover, the resulting decade-long sample would still be only a small percent of the total U.S. population – perhaps, at best, in the 2% to 3% range, assuming resources and other requirements remained essentially fixed.

In both Canada and the U.S., the likely higher unit costs of a rolling sample may need to be addressed by changes in survey procedures: how area segments are listed (Royce and Drew, 1988); how first contact with households is made, *etc.* Where is it written, for example, that a personal interview contact is needed before using other modes of collection?

It will be no mean challenge to keep *effective* sample sizes equal for the major level and change components now obtained from ongoing surveys (*e.g.*, Tegels and Cahoon 1982). Some compromise may be needed, moreover, in the extent to which the basic content of the current long-form Census samples can be included. Despite these challenges, or perhaps because of them, Kish's ideas on rolling samples deserve continued serious attention and should be the focus of extensive practical experimentation.

Administrative Registers

With the flowering of scientific sample survey methods in the 1940's (Bailar, 1990), the use of administrative records for statistical purposes became relatively less important in many national statistics programs. By the early 1980's, however, at least in the developed countries, the pendulum had begun to swing back. Kish recognizes this trend and rightly quotes Philip Redfern, who has been the major chronicler of this phenomenum internationally (Redfern 1987). While the Danes seem to have gone the farthest (Jensen 1983 and 1987), major efforts have been made in Canada (*e.g.*, Statistics Canada 1990) and even some in the U.S. (*e.g.*, Alvey and Kilss 1990).

A good summary of most of the key barriers to the greater use of administrative registers for census-taking is found in Redfern (1989), including the extensive discussion published with that paper. Perception barriers by the citizens (*e.g.*, in Germany) are mentioned as problems. Psychological barriers by the national statistical service may, however, be of equal or even greater importance. Major scientific "paradigm shifts" generally have this problem (Kuhn 1970). Certainly, this seemed to be part of the reason for the reception given to the proposal (made by me in 1980) to explore the feasibility of making administrative records an integral part of the U.S. Census of Population. While a sketch of such a proposal was eventually given at the 1982 American Statistical Association meetings (Alvey and Scheuren 1982), it seems, with a few fairly limited exceptions (*e.g.*, Irwin 1984, Citro and Cohen 1985), that serious interest at the Census Bureau has been notably lacking.

Suffice it to say that in the U.S. very little of the needed research has been undertaken. This is true, despite continuing efforts to give the proposal prominence (Jabine and Scheuren 1985 and 1987) and to get it discussed widely (Butz 1985). Sadly, therefore, I have to agree that Kish is probably right that in the United States, at least for the year 2,000, "... we should not expect [administrative registers] to replace censuses."

The 1990 U.S. decennial census could have been used as a proving (or disproving) ground for some of the needed research into administrative record alternatives. Why that didn't happen is a matter that can only be speculated about. A contributing factor, quite possibly, is a case of "paradigm paralysis" (Barker 1988). The literally decades-long controversy about whether to adjust census "counts" seems to have locked the U.S. Bureau of the Census into what some, at least, would call an increasingly sterile intellectual position (Fienberg 1990). The viewpoint that they have adopted makes it very hard for them to see any alternative, like a (partial) administrative record approach, that starts out with the notion that adjustments would be required.

The situation is different in Canada. Since the late 1970's, Statistics Canada has assembled many of the building blocks needed to conduct an administrative record census (*e.g.*, Drew 1989; Podoluk 1987; Verma and Raby 1989). While much remains to be done, such a change could even happen as early as 1996. For example, the coverage of the Canadian tax return system, alone, is quite high and growing. In 1987, for instance, it has been estimated that the coverage was about 94% – *i.e.*, about 3% less than the 96.8% coverage achieved in the 1986 Canadian Census. By 1991, tax return coverage, alone, should be up to about 97% or better, with overall administrative record coverage still higher and likely to grow further in the 1990's.

Kish expresses concern that administrative registers, even after they become adequate in quality and coverage, will "supply only a few, bare demographic variables: head counts, age, sex and little more." An immediate observation concerning his remark is that conventional censuses do *little* more than this, themselves, at least for the 100% items. It is also evident that, while the variables on administrative records are not the same as those collected in a traditional census, there may *already* be more available than Kish realizes (*e.g.*, Meyer 1990; Alvey and Scheuren 1982).

More important even than any current item content comparison is the need to emphasize that the proposal to use administrative registers in census-taking does not envision that administrative records have to be used as they are. *Administrative records will need to be changed.* In my personal opinion, limited optimism about achieving needed changes is justified. However, without a doubt, it is too much to expect of administrative records that they will be able to capture exactly the same concepts now measured in censuses and surveys. Additionally, there almost certainly will need to be special efforts, using existing census-taking techniques, to separately enumerate certain groups. The efforts in the 1990 U.S. Census to count the homeless would be one such example.

Censuses and administrative records each have inherent limitations. Unavoidable conceptual differences will be a major barrier to any shift from one medium to another. Administrative feasibility is another issue; however, some hard-to-duplicate census concepts (*e.g.*, households) may not be as important to the measurement process as formerly.

Shifts in methodology (from a conventional census to administrative records) for some uses would potentially be accompanied by a parallel shift in the underlying concepts measured. Some concepts may alter or expand in meaning, including our ability to measure them (*e.g.*, families). We also must ascertain the extent to which respondents answer survey questions the same way they fill out administrative forms that may have real direct impact in their lives.

In recent years, traditional survey methodology has been enhanced by new tools from the field of cognitive psychology. These cognitive research tools could be used to understand any conceptual differences between the meaning of terms when they are used in surveys or drawn from administrative records. We may not have what we think we have anyway (Bates and DeMaio 1989). In any case, there is already an extensive body of cognitive research that can be drawn on (*e.g.*, Dippo 1987; Fienberg and Tanur 1989; Jobe and Mingay 1990).

Kish is close to the mark when he goes on to say that administrative registers "will fail to meet the demands of modern society for richer sources of statistics." Such demands, of course, appear to be insatiable. Even if they were not, administrative records will never have the flexibility and responsiveness of surveys. Registers, however, (including partial ones like those that exist in the U.S.) when linked to survey data, can be extremely important as auxiliary variables in making improved direct national survey – and even subnational survey – estimates. The U.S. Census Bureau's Survey of Income and Program Participation research on the use of Internal Revenue Service data for improving the precision of national survey estimates is a good recent example (Huggins and Fay 1988). Indirect (*e.g.*, synthetic) estimates for small areas would still be needed for variables not on the administrative registers (Platek, Rao, Särndal, and Singh 1987). The registers, though, might provide a source of valuable symptomatic indicators.

Concluding Observations

The case Kish makes for considering a "paradigm shift" in census-taking seems compelling, at least in developed countries like Canada and the U.S. The rolling census alternative he proposes is probably too expensive to fully implement as a complete substitute for a census. Rolling samples do offer real promise, however, if they can be integrated into the current ongoing survey operations of Canadian and U.S. national statistical programs. Such samples could provide a needed link in addressing small area estimation needs that might otherwise not be met. Less promising, but still possible, is their use as a (partial) substitute for the census long-form samples.

Kish may be unduly pessimistic about administrative registers. The Canadian situation, however, differs from the United States:

- In Canada, it is already within the realm of feasibility to combine rolling samples with administrative records as an alternative to conventional census-taking. This is not to say that enormous practical challenges don't remain. The 100% count portion of the Canadian census, though, could be done with administrative records as a starting point, augmented by a large-scale survey to measure and potentially adjust for undercoverage. The Canadian 20% census long-form sample might be, at least partially, replaced by a rolling sample. The content of the Census long-form is considerably richer than that of household surveys, but the content differences could be made up through additional questions "piggybacking" the on-going surveys at regular intervals. Coverage issues surrounding the use of administrative records could also be addressed directly with rolling samples, especially to calibrate for changes in administrative records between censuses.

- In the United States, the U.S. Census Bureau has begun to look at alternatives other than conventional census-taking (Bounpane 1988). Unfortunately, the research needed to look at an administrative register alternative has barely begun. Whether the Census Bureau will find a better approach than the use of administrative records and rolling samples remains to be seen (Browne 1989). Whatever other alternatives they study, however, the use of administrative registers as a partial replacement for the conventional 100% counts definitely needs to be considered. A preliminary research agenda updating earlier ideas is given in Scheuren, Alvey and Kilss 1990.

Kish is right in saying that, with the radical proposals he (and I) are discussing, the answer is uncertain. Like him, I believe that "the balance of variance components" favors a change from conventional census-taking in most cases. "However, theoretical as well as empirical investigations will be needed to decide matters."

In a change as big as the one proposed here, the "balance" that needs to be struck goes, of course, well beyond looking at variance (and bias) components. Kish recognizes this in numerous ways in his paper. One issue that needs to be emphasized more, though, is that some aspects, at least, of the paradigm shifts being considered could go to the heart of the social contract that exists between national statistical agencies and the people that those agencies have a mission to serve. For instance, in the U.S. Constitution, there is a requirement that an "enumeration" of the population take place every ten years. Would the use of administrative records or rolling censuses fit within this "Constitutional paradigm?" Perhaps the starting place is to adopt a broader definition of "enumeration."

Another example where social contract issues arise is the extent to which the greater use of existing (or expanded) administrative data for statistical purposes might be seen as an unwelcome increase in the intrusiveness of the State into the private lives of its citizens (Grace, 1989). As legitimate as concerns about "intrusiveness" might be, though, there is no evidence in a North American context, at least, that they pose an insurmountable barrier. On the contrary, there have been virtually no adverse public reactions to past U.S additions to administrative records for statistical purposes (*e.g.*, of residential address information in 1972, 1974 and 1980 tax returns). To my knowledge the issue, so far, has not come up directly yet in Canada, at least at the Federal level.

In summary, to make changes of the types being discussed by Kish, there is, as he points out, the need for a lot more scientific research. Studying the implementation technologies will be an even bigger job. Finally, the issues go beyond our profession and may well be settled in other arenas. Wherever they are decided, it is incumbent on us, as statisticians, to frame the debate in terms of feasible options. Kish has taken us a long way down that path and is to be greatly congratulated.

REFERENCES

ALVEY, W. and KILSS, B. (eds.) (1990). *Statistics of Income and Related Administrative Record Research*. U.S. Department of the Treasury, Internal Revenue Service. See also Kilss, Beth and Alvey, Wendy (eds.) (1984). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, vols. 1 and 2, U.S. Department of the Treasury, Internal Revenue Service.

ALVEY, W. and SCHEUREN, F. (1982). Background for an Administrative Record Census. *Proceedings of the Section on Social Statistics, American Statistical Association*, 137-146.

ANDERSON, M. (1990). 'According to their respective numbers . . .' for the twenty-first time. *Chance*, 3, 12-18.

BAILAR, B. (1990). Contributions to Statistical Methodology from the Federal Government. *Survey Methodology*, 16.

BARKER, J.A. (1988). *Discovering the Future: The Business of Paradigms.*

BATES, N.A., and DEMAIO, T.A. (1989). Using Cognitive Research Methods to Improve the Design of the Decennial Census Form. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 267-285.

BOUNPANE, P. (1988). A Sample Census: A valid alternative to a complete count census? 46th Session of the International Statistical Institute.

BROWNE, D.L. (1989). U.S. Bureau of the Census: Facing the future labor shortage. *Asian and Pacific Population Forum*, 3, 4.

BUTZ, W. (1985). Comment: The Future of Administrative Records in the Census Bureau's Demographic Activities. *Journal of Business and Economic Statistics*, 393-395.

CITRO, C., and COHEN, M.L. (eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990.* National Academy Press, Washington, DC.

DIPPO, C. (1987). A Review of Statistical Research at the U.S. Bureau of Labor Statistics. *Journal of Official Statistics*, 3, 289-297.

DREW, J. D. (1989). Address Register Development and its possible future role in Integration of Census, Survey and Administrative Data. A paper presented at the U.S. Bureau of the Census/Statistics Canada Interchange. (Unpublished).

FELLEGI, I.P. (1981). Discussion of a paper by Leslie Kish entitled "Population Counts from Cumulated Samples." *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau. An Analysis, Review and Response*, Congressional Research Service, the Library of Congress.

FIENBERG, S. (1990). An Adjusted Census in 1990? An Interim Report. *Chance*, 3, 19-21.

FIENBERG, S., and TANUR J. (1989). Combining Cognitive and Statistical Approaches to Survey Design. *Science*, 243, 1017-1022.

GRACE, J.W. (1989). The Use of Administrative Records for Social Research. Statistics Canada Workshop, December 12, 1989, Ottawa, Ontario.

HAMMOND, R.B. (1990). The 1990 Decennial Census: An Overview. *Conference Proceedings, Advanced Computing for the Social Sciences*, sponsored by the Oak Ridge National Laboratory and the U.S. Bureau of the Census, April 10-12, 1990, Williamsburg, Virginia.

HERRIOT, R., BATEMAN, D.V., and McCARTHY, W. F. (1989). The Decade Census Program - New Approach for Meeting the Nation's Needs for Sub-National Data. To appear in *American Statistical Association Proceedings, Social Statistics Section.*

HUGGINS, V., and FAY, R. (1988). Use of Administrative Data in SIPP Longitudinal Estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.

IRWIN, R. (1984). Feasibility of an Administrative Records Census in 1990. Special report on the use of administrative records, committee on the use of administrative records in the 1990 Census, unpublished Census Bureau report.

JABINE, T.B., and SCHEUREN, F. (1985). Goals for Statistical Uses of Administrative Records: The Next Ten Years. *Journal of Business and Economic Statistics*, 380-391.

JABINE, T.B. and SCHEUREN, F. (1987). Statistical Uses of Administrative Records in the United States: Where Are We and Where Are We Going? *Proceedings of an International Symposium on Statistical Uses of Administrative Data*, J.W. Coombs and M.P. Singh (eds.), Statistics Canada, December 1988, Ottawa, 43-72.

JENSEN, P. (1983). Towards a Register-Based Statistical System - Some Danish Experiences. *Statistical Journal of the United Nations Economic Commission for Europe*, 1, 341-365.

JENSEN, P. (1987). The Quality of Administrative Data from a Statistical Point of View: Some Danish Experience and Consideration. *Proceedings of an International Symposium on Statistical Uses of Administrative Data*, J.W. Coombs and M.P. Singh (eds.) Statistics Canada, Ottawa.

JOBE, J.B., and MINGAY, D.J. (1990). Cognition and Survey Measurement: History and Overview. *Applied Cognitive Psychology*, in press.

KUHN, T.S. (1970). *The Structure of Scientific Revolutions*. Second Edition, Enlarged, The University of Chicago Press, Chicago.

MEYER, B. (1990). The Tax System: Comparisons of Demographic, Labour Force and Income Results for Individuals and Families. Small Area and Administrative Data Division, Statistics Canada.

PLATEK, R., RAO, J.N.K., SÄRNDAL, E.E., and SINGH, M.P. (1987). *Small Area Statistics*, New York: Wiley-Interscience.

PODOLUK, J. (1987). Administrative Data as Alternative Sources to Census Data. *Proceedings of an International Symposium on Statistical Uses of Administrative Data*; J.W. Coombs and M.P. Singh (eds.), Statistics Canada, December 1988, Ottawa, 273-290.

REDFERN, P. (1987). A Study of the Future of the Census of Population: Alternative Approaches Eurostat Theme 3 Series C. Luxembourg: Office for Official Publications of the European Communities.

REDFERN, P. (1989). Population Registers: Some Administrative and Statistical Pros and Cons. *The Journal of the Royal Statistical Society*, Series A (Statistics in Society), 152, 1-41.

ROYCE, D., and DREW, J.D. (1988). Address Register Research: Current Status and Future Plans. 1991 Research and Testing Project, 1991 Census, Statistics Canada, Ottawa.

SCHEUREN, F., ALVEY, W., and KILSS, B. (1990). Paradigm Shifts: Administrative Records and Census-Taking.

STATISTICS CANADA (1990). Research papers and reports. Bibliography, Small area and administrative data division, Ottawa, Ontario. (unpublished).

TEGELS, R., and CAHOON, L. S. (1982). The Redesign of the Current Population Survey: The investigation into alternate rotation plans. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

U.S. BUREAU OF THE CENSUS (1989). *200 Years of U.S. Census Taking: Population and Housing Questions, 1790-1990*. Superintendent of Documents, U.S. Government Printing Office, Washington, DC.

VERMA, R.B.P., and RABY, R. (1989). The Use of Administrative Records for Estimating Population in Canada. *Survey Methodology*, 15, 261-270.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 81–86
Statistics Canada

Comments on Articles in the Special Section

MORRIS H. HANSEN[1]

These are excellent papers that I enjoyed reading. Three of these papers focus primarily on historical and current developments and to some extent looking to the future. The paper by Kish is focused on and is an effort to influence some important future developments. I will attempt to add a little clarification from my own personal history and point of view on the historical summaries, and a little perspective, again from my personal point of view, on Kish's proposal for rolling censuses to replace the more traditional censuses.

Rao and Bellhouse have given a compact but useful survey of sampling development. Their summary begins, after a few preliminaries, at about the time that I first began to participate in censuses and sample surveys, and their improvement.

Their survey is done about as well as can be accomplished in such a compact summary, without elaborating on details. However, I would like to provide a slightly different view than they present on the development of sampling with probabilities proportionate to size or to measures of size (PPS). They accurately indicate that we (Hansen and Hurwitz) developed the theory for PPS sampling with replacement as an approximation. We were unsuccessful in solving the problem of variance estimation with varying probabilities when sampling without replacement that was soon solved by Horvitz-Thompson and others. However, with possibly rare exceptions, we never proposed the use of or used sampling with replacement. In practice, we did PPS sampling without replacement, usually either by choosing two or more units from a stratum by a systematic sampling procedure with the units arranged in a random or systematic sequence, or by choosing one unit per stratum. Units that would have had high probabilities of selection were selected with certainty. We prepared estimates of aggregates and functions of these by weighting by the reciprocals of the probabilities, exactly as in what has come to be referred to as the Horvitz-Thompson estimator. The variance estimators resulted in moderate overestimates because they assumed sampling with replacement as a simplification. Ordinarily, we have not regarded moderate overestimates of variance as a serious concern. The ultimate cluster variance estimator was often used. This is a very simple approximate variance estimator that involves weighting (if subsampling has been used) within the first stage units up to the first stage unit level, and then computing the variance between such first-stage unit estimates (see Hansen, Hurwitz, and Madow, p. 257). Horvitz and Thompson provided the initial breakthrough in variance estimation when sampling more than one unit per stratum with varying probabilities.

Sampling with PPS had the advantages that Rao and Bellhouse briefly describe. In addition, its use was a great convenience in multistage sampling, with probabilities proportionate to measures of size at each stage up to the final. The probabilities at the final stage were often set to achieve uniform overall probabilities of selection of the elementary units.

I add one other comment on their paper with respect to jackknife variance estimation. They indicate that the jackknife variance estimators are known to be inconsistent for nonsmooth functions like quantiles, even in the case of simple random sampling. They might have said, especially in the case of simple random sampling of the elements that are the units of analysis. We have recently demonstrated empirically that variances of medians and (in this case)

[1] Morris H. Hansen, Westat, 1650 Research Boulevard, Rockville, MD, 20850, U.S.A.

of 10th and 90th percentiles can be well estimated with the usual ultimate cluster jackknife variance estimation procedure with multistage sampling in which two or more first-stage units or combinations of them are identified in a stratum (one dropped and the other doubled, to form a replicate). We hypothesize that jackknife worked well in these applications because each ultimate cluster associated with a first-stage unit contains a substantial number of elementary units in the sample. We anticipate that it would work equally well, although we have not demonstrated it, when the jackknife replicates are formed by another procedure often followed, in which a simple random (or stratified random) sample is divided into m simple random subsamples (or stratified random subsamples utilizing the same strata to the extent feasible), and dropping one subsample at a time.

Fienberg and Tanur have presented an interesting perspective on the influence of the institutional setting in which survey research has developed. I agree with their view that an improved understanding of the development of survey methods is achieved by an understanding of the institutions through which survey research and surveys are done. At least those survey developments in which I have participated have arisen largely out of the institutional setting, and the need and opportunity to solve problems that occurred in accomplishing programs of the institution. Again, I have comments on some of the details in the developments in which I was a participant.

Fienberg and Tanur properly indicate that the design of what is now known as the Current Population Survey or CPS (earlier known as the Labour Force Survey) had a key role in the evolution of sampling theory and its application that has influenced other developments. However, they incorrectly suggest that its principal origins were in the experimental Trial Census of Unemployment carried out in late 1933 and early 1934 as a Civil Works Administration (CWA) project in three cities. There is some confusion in their paper of the 1933-34 CWA trial census with the 1937 "Enumerative Check Census" that accompanied the 1937 "Unemployment Census". It was the latter that, as they mention, Dedrick, Hansen, Stouffer, and Stephan jointly worked on, and that was the progenitor of the CPS. The 1937 Unemployment Census was a national registration done through the Post Office. The Enumerative Check Census was taken by mail carriers in a national probability sample of postal routes – they took a complete census of each postal route in the sample. New concepts for measuring labour force and unemployment were developed and applied in it based on behavior in a prior week. It was also a first application of nationwide area probability sampling. Its purpose was to evaluate the 1937 national registration of the unemployed (as discussed in the accompanying paper by Barbara Bailar). That sample survey taught us much, and was the seed for the monthly Labor Force Survey, later to become the Current Population Survey. Again, I was an active participant. Bailar desbribes it well. Stock, Frankel, and Webb and others at the Work Projects Administration (WPA) also had a role in the design of the national registration and of the Enumerative Check Census. Those were the days of dire unemployment, and the need for a continuing measure was obvious and urgent.

With this experience Stock, Frankel, and Webb, along with their colleagues at WPA perceived the opportunity and need for a continuing survey. They initiated a monthly unemployment and labor force survey, introducing some imaginative concepts in survey design (but also some problems that needed later correction). The monthly survey was just getting well established when Pearl Harbor and U.S. entry into World War II occurred, and the needs for information were radically changed. Labor shortage rather than high unemployment became the problem. The WPA was no longer needed and was abolished, and the survey was transferred to the Bureau of the Census to become a labor force survey to measure especially war-time implications of labor force participation and employment. When the survey was transferred to the Bureau of the Census we perceived some problems in the original design and developed

solutions to them, which led to the introduction, among other things, of PPS sampling and other design innovations. These developments for the labor force survey (now the CPS with a much broader role) have had a substantial impact on sample methodology, and more important, on meeting the needs of the nation for up-to-date information, not only on labor force but on many other subjects - demographic, social, and economic.

Feinberg and Tanur might also have emphasized the remarkable consequences of bringing together census-taking and sampling, along with computerization and automated reading of position marks on census questionnaires. In modern censuses in the United States, beginning with the 1960 Census, the questionnaires used for collecting information from all households are relatively brief in content. The principal content of the censuses is now obtained through samples taken simultaneously with and as part of the census, and, of course, on an exceedingly large scale in order to produce useful data for perhaps 40,000 small areas. A related development was the introduction in the 1960 Census of self-enumeration methods. The decision to introduce self-enumeration was guided by the application of the response error model to which Feinberg and Tanur refer, and by associated research and experiments on response errors, and especially on the correlated response errors associated with the work of enumerators. These innovations were guided by large-scale experiments that were done prior to and as part of the 1950 Census and in later censuses as well as in separate experiments. Another contribution was FOSDIC (Film Optical Sensing Device for Input to Computers), a device for reading position marks designed by the Bureau of Standards at the Census Bureau's request, in response to Census Bureau needs to replace the massive key-punching effort in a census. A consequence of the innovations that were introduced was more timely results and generally more accurate censuses, as well as lower costs. The opportunities for progress arose in view of the problems of large-scale census taking, and how they might be solved with the application of sampling and self-enumeration, along with the remarkable advances made possible by the development and application of electronic computers and FOSDIC, in which the Census Bureau was a pioneer.

In the late 1930's, some of the top Census Bureau staff, as well as members of Congress, were reluctant to see sampling introduced into the work of the Census Bureau. Complete enumeration had been the tradition. The use of probability sampling in the 1937 enumerative check census associated with the national unemployment registration was an important factor in achieving the acceptance of sampling as a methodology appropriate to the Bureau of the Census, again as more fully told in the accompanying paper by Bailar. The 1940 population census was a pioneering effort in the application of sampling in the collection of supplemental items of information in a census. In this effort Deming and I worked as colleagues. I was working with Calvert Dedrick, and Deming with Philip Hauser, with effective consultation and advice from Fred Stephan, and we all worked as a team in developing this important milestone in the application of sampling.

I have little in the way of comments to add to the paper by Barbara Bailar. As the paper indicates, I was an active participant along with Bill Hurwitz and our colleagues, in the developments she describes so well. I do have a minor correction. Feinberg and Tanur correctly identify the 1951 paper on response error models by Hansen, Hurwitz, Marks, and Mauldin as the original publication on the model, which Bailar credits to a later (1960) paper by Hansen, Hurwitz, and Bershad. The later paper elaborated those results, and included empirical data from the application of the model in large-scale randomization experiments involving the random assignment of enumerators in the 1950 Census. Analysis of these results as summarized in the 1960 paper showed the substantial and striking impact on small area census statistics of correlated errors within the work of interviewers. Earlier memoranda containing the results reported in that paper, and associated studies, were the principal vehicles that led

to the use of self-enumeration as the procedure for collecting the principal content items in the 1960 Census. They also led to transferring the collection of much of the information to a large sample instead of a complete census, with substantial cost reduction implications, improved timing, and generally improved quality. Bailar's paper provides an excellent summary description.

I should note, in this connection, the remarkable contribution to these developments that came from Bill (William N.) Hurwitz. He and I worked as a team that was far more effective than the sum of our individual contributions. In addition, I cannot give enough credit to our colleagues that we recruited and helped to stimulate and to some extent train, and who became the backbone of developments in the Census Bureau in the application of sampling, quality control, and operational research methods to the successful design and conduct of samples and censuses in wide ranging subject areas. Leaders among these colleagues included Max Bershad, Joseph Daly, Leon Gilford, William Madow, Eli Marks, Harold Nisselson, Jack Ogus, Leon Pritzker, Joseph Steinberg, Benjamin Tepping, Joe Waksberg, Ralph Woodruff, and others. I often get much of the credit, but without Bill Hurwitz, especially, and our colleagues, it could not have occurred.

I should mention that we benefited greatly, also, from the participation and advice from a panel of statistical consultants, with Bill Cochran (William G. Cochran) as chairman, over the years from 1955 until I left the Bureau in 1968. Other principal members included Fred Stephan (Frederick F. Stephan) and Bill Madow (William G. Madow) for the full time period, and Ivan Fellegi from Statistics Canada, H.O. Hartley, and others for part of the time. All were exceedingly able. However, we did not look to them as experts whose advice would simply be sought and generally followed. Instead, we operated on an interactive basis. We discussed specific issues or problems as well as all phases of total survey design for a particular survey, experiment or census. We received much useful advice; they also learned from us.

The paper by Leslie Kish moves the emphasis from historical background and recent and current advances to proposals for taking censuses of the future, through the introduction of what he calls rolling censuses. He also describes rolling samples in various forms.

Each of the kinds of rolling samples that he discusses, with and without overlapping panels are, as he indicates, in use for various purposes at the present time, and his discussion of these does not propose anything new. I suppose he introduces them for generality and as a means of suggesting their potential relationship to a rolling census.

The particular rolling census he describes is a weekly sample, with the total population of housing units at each point of time subdivided into 520 subsamples, one to be covered each week over a 10-year period. Thus, the entire population of housing units would be covered in a decade except for new additions of housing units in samples that had already been covered earlier in the decade. If the procedure were continued over time, then at any point in time the aggregate of the 520 samples for the prior ten years would provide average census results, representing the average situation over the prior 10-year period. It is an interesting and imaginative proposal. However, there are also problems.

He suggests a rolling census without any overlap in the coverage in successive weeks or other periods, except after the full decade when it starts all over again. Such an approach would provide a large national cross section sample each week, as well as average or aggregate results for each month, each year, and for other periods. However, without any overlap in the samples, it will be a relatively crude instrument for measuring changes occurring in small areas from week to week, from month to month, or even from year to year. Overlapping samples might be introduced, as he indicates, but would add greatly to costs. Of course, changes can be measured with the proposed rolling samples, but without partially overlapping samples the result would be large sampling errors of estimates of change for small areas. Providing data

for small areas is a primary purpose of the Decennial Census. I believe that reliably measuring such changes may be as important as providing aggregate measures for points in time. While Kish recognizes this, he seems to dismiss it.

Undercoverage of the population would likely be a particularly serious problem with a rolling census. Because of the general recognition by the public of the need for censuses, along with the intense publicity that is feasible for a census, the completeness of coverage of the censuses has traditionally been much greater than that in even the best sample surveys (although coverage still remains a problem in the censuses). The problem of net undercoverage in sample surveys is quite general – even including the Current Population Survey in the U.S. which is often taken as a model. Public interest with continuing weekly publicity for a rolling census could not conceivably be maintained.

Another issue in my judgment is the likely high cost of such a system. Kish recognizes this, also, and then seems to dismiss it. While I have not seen any cost estimates, I would not be surprised that over a decade the rolling census would cost substantially more than the cost of taking complete censuses quinquennially, plus the cost of relatively large-scale monthly samples to provide measures of change and information on various subjects for states and large areas within most states. Moreover, I anticipate that quinquennial censuses would be easier to interpret and more useful by providing measures for small areas at points in time, or for short intervals of time, rather than providing average measures over periods up to ten years.

The Census Bureau, influenced, in part, by Kish's earlier recommendations for such a rolling census, and the desire to spread the workloads has come up with some proposed alternatives for consideration for taking a brief decennial census along with rotating censuses. They consider some alternative approaches to rotating censuses of whole states over a decade. It is an innovative proposal intended to spread the workload while avoiding the high cost of a rolling census such as described by Kish.

I am one who believes that a quinquennial census, along with ongoing large-scale current surveys, are well worth a substantial cost. However, I believe that if a rolling census were adopted, as proposed by Kish, overlapping samples should be used. A rolling census, even without overlapping samples, may cost considerably more than the cost of the current census program extended to include a quinquennial census. I question if it is worth the added cost, or that it has advantages over a quinquennial census plus substantial intercensal samples. I anticipate that the rolling census approach would yield less useful information than quinquennial censuses for most purposes because it would provide complete census counts only for averages over a 10-year period. Quinquennial censuses, along with sufficiently large current samples to provide relatively up-to-date information for large areas, along with other procedures for providing data for state, county, and perhaps also small area population estimates, seem to have advantages from a cost-benefit point of view.

Kish is to be commended for his efforts to solve some of the census problems by a radical new approach. However, to me, the rolling census does not appear to be the answer. Perhaps more effective utilization of administrative records can provide results that hold more promise, again along with current samples and a decennial, or, hopefully, quinquennial censuses. Perhaps the remarkable new computerized mapping and coding system (known as TIGER) developed by the Census Bureau for the 1990 Census holds much promise for improving census-taking, and for current sample surveys. In addition, incorporating the TIGER geographic coding into the major administration records systems might make them more accessible for population estimates and for other uses. Up-to-date maintenance of TIGER, along with a currently maintained address register, are hopefully to be included in the Census Bureau's future plans.

REFERENCES

HANSEN, M., HURWITZ, W., and MADOW, W. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons.

HANSEN, M. (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, 2, 180-190.

OLKIN, I. (1987). A Conversation with Morris Hansen. *Statistical Science*, 2, 162-179.

DUNCAN, J.W., AND SHELTON, W.C. (1978). *Revolution in United States Government Statistics, 1926-1976*. U.S. Government Printing Office, Washington, D.C.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 87–88
Statistics Canada

Reply

J.N.K. RAO and D.R. BELLHOUSE

We thank the discussants, Hansen and Smith, for their useful comments.

Hansen provided important observations on the development of PPS sampling. He is correct in saying that Hansen and Hurwitz (1943) did not propose the use of sampling with replacement and that only for variance estimation they assumed sampling with replacement. Incidentally, Murthy (1967, p. 184) notes that Mahalanobis (1938) has referred to PPS sampling and the associated unbiased estimator of a total in the context of sampling plots for a crop survey.

Hansen also made some interesting observations on the use of delete-1 cluster jackknife variance estimator for nonsmooth functions like quantiles. It is now well-known that the delete-1 jackknife variance estimator of a quantile is inconsistent under simple random sampling. Empirical results in Kovar, Rao and Wu (1988) indicate that it is also inconsistent under stratified simple random sampling. It is also likely inconsistent under stratified cluster sampling if the subsamples from the clusters are small or if the intra-cluster correlations are significant. In Hansen's application the subsamples from the clusters are quite large and the intra-cluster correlations very small. In this case, the delete-1 cluster jackknife variance estimator may be well-behaved in view of Shao and Wu's (1989) result that the delete-d jackknife variance estimator, under simple random sampling, is consistent, provided $n^{1/2}/d \to 0$ and $n-d \to \infty$ as the sample size $n \to \infty$.

The method of dividing a simple random sample into m subsamples, each of size d say, and dropping one subsample at a time, as suggested by Hansen, is similar to Shao and Wu's delete-d jackknife except that they consider all $\binom{n}{d}$ subsamples in constructing the variance estimator. However, the delete-d jackknife variance estimator is likely to be more stable. Shao and Wu also consider balanced subsampling requiring only b subsets of size $n-d$, where b ($\geq n$) is the number of blocks in a balanced incomplete block design.

Smith provided some important observations on the foundational aspects of sample survey theory, in particular, on the importance of Ericson's (1969) work on Bayesian estimation of a total under exchangeable priors. In this connection, we note that equivalent results for the posterior mean and the posterior variance, under simple random sampling, were also obtained by Hartley and Rao (1968). A. Scott pointed out the similarity of the two approaches in his discussion of Ericson's paper. However, an advantage of the Hartley-Rao approach is that the inferences depend on the sample design, unlike Ericson's approach. Their approach also yields useful classical inferences. Rao and Ghangurde (1972) extended the Hartley-Rao results to stratified random samping, double samping with unknown strata sizes, the Hansen-Hurwitz method for handling nonresponse, and two-stage random sampling.

The GUT approach for inference, proposed by Smith looks very promising. We agree with Smith that the point estimators using the different approaches rarely differ very much in practice, and that the issue essentially reduces to the choice of a measure of uncertainty, as noted in our paper.

We also agree with Smith on the importance of measuring total survey error from ongoing surveys.

ADDITIONAL REFERENCES

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

MAHALANOBIS, P.C. (1938). *Statistical report on the experimental crop census*, 1937. Indian Central Jute Committee.

RAO, J.N.K., and GHANGURDE, P.D. (1972). Bayesian optimization in sampling finite populations. *Journal of the American Statistical Association*, 67, 439-443.

SHAO, J., and WU, C.F.J. (1989). A general theory for jackknife variance estimations. *Annals of Statistics*, 17, 1176-1197.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 89–90
Statistics Canada

Reply

STEPHEN E. FIENBERG and JUDITH M. TANUR

We are grateful to Bob Groves and Morris Hansen for their insightful comments and to the editor of *Survey Methodology* for the opportunity to update our thinking in 1990 rather than waiting for 2040. Groves and Hansen make several important points; we shall attempt to react to them in turn.

We very much like Groves' summary to the effect that governments emphasizing service for the welfare of the populace demand more information about their services than do those pursuing other goals. Consistent with this thesis is the fact that the most substantial new national survey launched in the United States during the 1980s, a decade not noted for an emphasis by the federal government on expanding welfare services, was the Survey of Income and Program Participation, one of whose primary purposes has been to monitor the impact of government welfare programs on income and assets. Moreover, as the countries of Eastern Europe democratize and turn to the West for assistance in upgrading their statistical systems, including the development of infrastructures for the conduct of large scale surveys, we see additional support for such a thesis. Thus it seems to us that Groves shares our belief that the institutional bases for survey research shape the content and direction of such surveys. Whether they provided homes or incubators for the best and the brightest seems to us akin to the nature/nurture debate – more a framework for discussion than an either-or choice. Indeed, we agree with Groves that the purposes of the various sectors shaped their choice of tasks, at least in part. In line with his urging of a cross national perspective, however, we note that institutional roles differ across countries. For example, there has been a widely-held view in the United States that the Federal government should not be in the business of collecting survey data on subjective phenomena (*e.g.*, see Turner and Martin 1984, 31-39) – a quite different stance has been taken by the British government, especially in connection with its annual report, *Social Trends* (Turner and Martin 1984, p.4).

Groves suggests that the membranes between sectors (academic, commercial, and governmental) are less permeable than we suggest. Neither we nor he have collected systematic empirical evidence on this question, but we point again to our concept of bridging institutions which bring together representatives of the various sectors, for the interchange of ideas if not personnel. And we hasten to point out that Groves' own recent appointment to the position of Associate Director of the U.S. Bureau of the Census, as well as Hansen's movement from that position into the commercial domain back in 1968, indicate the value, if not the ease, of membrane crossing.

Groves indirectly speculates that we choose to focus on technological advances, longitudinal surveys, and cognitive aspects of surveys because these are our areas of interest and experience, and he suggests several other developments that are worthy of consideration. Of course he is correct in suggesting that we have focussed on the developments that fit with our interests, but surely technological advances as a topic subsumes Groves' first two additional areas of importance: (1) development of generalized statistical software packages and (2) existence of survey data archives. We wonder, however, if the technological advances we both note, coupled with the ubiquity of surveys that we also both note, do not have negative as well as positive consequences. For example, the complex analyses of survey data by undergraduates (or indeed

any beginners) using statistical software packages often show neither an understanding of the data being analyzed nor the appropriateness of the packaged statistical methods used.

The ubiquity of surveys is a consequence not only of the demand for information but also of the relative ease with which surveys can be carried out and the data analyzed given current technology. (And we believe that the availability of survey data for reanalysis will only increase with the advent and adoption of new storage technologies such as CD-ROM and optical disks). Such ease is a mixed blessing. As Groves notes, the 1980s have seen a growing problem of nonresponse in the United States, a pattern that manifested itself earlier and (so far) more seriously in Europe. We do not need to postulate a growing trend toward demands for privacy to explain this decline in response rates, though such a trend may well exist. We need only look at the major nonresponse problems currently being encountered in the conduct of the U.S. 1990 decennial census, in both the mail-out-mail-back and in the door-to-door phases, to see evidence to support the contention that respondents are merely getting tired of being surveyed so frequently.

Further, as Groves points out, survey research has not been central to the self-image of academe, because survey research has not fully evolved into a separate identifiable discipline, with specified standards and training criteria. Since there are no departments of survey research on university campuses, almost anyone who cares can mount a survey or carry out analyses of survey data. While some people do these tasks well, others do them poorly thereby giving the whole survey enterprise a bad name. Thus, if we are to present the optimistic report on the state of the survey enterprise in 2040 that Groves envisages, it seems to us that the innovations in education and training that neither he nor we are currently able to chronicle will have to become institutionalized.

We are especially pleased to have Hansen's embellishment on our brief account of the development of the survey enterprise in the U.S. government in the 1930s and 1940s. His comments supply some of the human drama that Groves says is lacking in our institutional focus.

Hansen also expands on our account of the link between censuses and sampling and the introduction of self-enumeration into U.S, censustaking, that was guided by the study of response errors. The major decline in completion rates for self-enumeration in the 1990 decennial census suggests the need to reexamine the implications of the various components in the Hansen-Hurwitz-Marks-Mauldin model for non-sampling errors. In addition we note that as part of the 1990 census, the Bureau of the Census will mount a new Post-Enumeration Survey (PES) of 150,000 households whose results will be used to evaluate census coverage. The technological advances in computerized data management and in computer-based matching of files between the PES and the census were essential ingredients to the launching of this major new government survey and its planned use to measure both under-and over-coverage of the household-based population.

ADDITIONAL REFERENCES

TURNER, C.F., and MARTIN, E. (eds.) (1984). *Surveying Subjective Phenomena. Vol. 1.* New York: Russell Sage Foundation.

Survey Methodology, June 1990
Vol. 16, No. 1, p. 91
Statistics Canada

Reply

BARBARA BAILAR

The comments from the discussants describe even more contributions of the Federal Government to the world of statistics. I am very grateful to Gordon Brackstone and Morris Hansen for mentioning these additional topics. The topic I omitted that may have had the biggest impact on statistics as well as other quantitative fields was the development of the computer for data processing and data analysis purposes. Again, the team of Hansen and Hurwitz were the prime movers, urging and funding the development of UNIVAC I and then bringing it into the Census.

Morris Hansen describes the remarkable team at Census who worked with him and Bill Hurwitz on so many topics. I feel very fortunate that I began my career at the Census Bureau when these people were there and that I was able to work with most of them for many years. It is rare that one gets that kind of apprenticeship.

Gordon Brackstone questions whether the statistical methodology developed by the Census Bureau had a benefit to the wider world of statistics. Certainly, given the amount of interaction among government statistical offices, the Bureau of the Census has influenced government statistical operations in other countries. Brackstone finds the impact of the Census Bureau development on university statistics departments rather mixed. He may be correct as far as course offerings are concerned, but I believe the ASA-NSF-Census Fellowship program and the Agriculture Fellowship program have had a big impact. More university professors and graduate students are aware of and working on non-sampling error, disclosure avoidance, and time series problems. The recent addition of Fellowship programs at the Bureau of Labour Statistics and the National Center for Education Statistics have also highlighted these research areas. The NSF now receives many proposals based on research started at one of the government agencies.

The main problem now is to make sure that research results are used. Many government programs are slow to accept new methodology because change is disruptive. Yet, to make sure that methods are improving, change is necessary.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 93–94
Statistics Canada

Reply

LESLIE KISH

In his fine discussion Fritz Scheuren complements our comparisons of alternative census methods by advocating administrative registers for the USA. I support his expert plea to study what these methods could offer as additions, as complements to the decennial censuses. They are coming to many countries and we would like to know where, when, and how? It is even likely that they will not only complement, but even replace decennial censuses soon in some places. When in the USA? I don't know; we were comparatively slow and late in adopting a successful registry of births and deaths. And even now their reporting is rather slow.

Rolling samples could be designed for quick reporting, and timeliness is only one of the advantages of rolling samples. Thus it is biased to compare rolling censuses with traditional censuses, both as regards costs and benefits, only on the basis of the single output for which decennial censuses are designed. It would take detailed, technical investigations to compare the factors of costs, coverage, timeliness, content, *etc.* of rolling versus decennial censuses in the USA. But 10 to 15 million dollars monthly can go far. The issue of adequate censuses is most salient in 1990 in the USA and elsewhere, but the other uses of samples should not be forgotten, as we plan for the last decade of our twentieth century.

My contribution aims mainly to advance the *diverse* advantages of cumulations from periodic samples, which have been neglected in favor of the other benefits that can be obtained from the growing numbers of periodic surveys. Rolling censuses may become someday one of those benefits, and rolling samples have been used already – though not often enough, I believe. Asymmetrical cumulations may exist rarely and obscurely, and the split-panel designs that I propose, not at all.

Furthermore my scope is not merely national (the USA), nor even continental (North America): it is intercontinental and international. For example, registers have come to the Nordic countries and they may come to Canada before the USA. Rolling censuses pose a much smaller expansion of the Labour Force survey in Canada because it is one-tenth the size of the USA, as Fritz and I both show. But some other country may well use them before either.

Not only international, rolling samples and cumulations are also aimed to be inter-disciplinary, not only for making population counts. Good many of the other needs of statistical offices – and *of other institutions for data collections*! – would be better served by a trained "permanent" staff than by a hurriedly hired huge army whose training time roughly equals their brief employment.

Scheuren is most complimentary when he calls rolling censuses a new paradigm. It is true that, as all new paradigms, they meet three big mental blocks when I present cumulations and rolling samples: a) averaging of variable data instead of an arbitrary date like April 1, of the decennial year; b) accepting some of the mobility of human populations instead of fixing them to unique sites; c) rolling samples to replace fixed primary sampling areas. So it may seem paradoxical when Morris Hansen notes that my "discussion of these does not propose anything new." Hansen may have encountered all of these proposals, and perhaps dismissed some of them. Personally I have described rolling samples since at least 1961 and proposed rolling censuses since 1965. But I also found that for many people they come as new ideas, and often as strange new ideas.

Finally let me only add two important origins in the '40's for sampling, although for me personalities and priorities are only minor aspects of the history of any science. Iowa State at Ames should be mentioned, where, under George Snedecor and Henry Wallace, Bill Cochran started in the spring of 1939 the first course of sampling and turned out pioneer MA's, then PhD's in sampling. Then Henry Wallace (again) in the US Dept. of Agriculture started the Division of Program Surveys, hired me in 1941 and Steve Stock in 1942 for the first national samples in Washington in 1942, followed by the 1943 sample at the USBC. Stock, Frankel and Webb (from the WPA samples) began the second sampling course in fall 1939 at the USDA graduate School, which became famous and productive under Hansen, Hurwitz and their Census staff. Among influential courses there I shall testify especially to those of Deming, the major figure at the school. The teaching and learning of samples in the forties was done mostly at Ames and in the USDA, as well as at the USBC.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 95–98
Statistics Canada

Some Developments of Sampling Techniques and their Use in Official Statistics in Sweden

TORE DALENIUS and CARL-ERIK SÄRNDAL[1]

In this paper we present some important features of the history of sample surveys in Sweden, and we comment on related developments of sampling techniques (methods and theory) in official statistics. The account is organized into three periods as follows: (i) before 1900; (ii) 1900-1950; and (iii) after 1950. The emphasis is on the third period.

I. THE PERIOD BEFORE 1900

1. A summary view. As described in Dalenius (1957), there was a noticeable resistance against sample surveys in traditional fields of official statistics, especially among statisticians in leading positions. Sample surveys were considered justified primarily in cases where circumstances did not admit *total* surveys. In other fields there were, however, signs of appreciation, as illustrated in the next section.

2. Two classic illustrations. In the 1820's, the area of meadowland in Sweden was estimated using the following technique. For each county separately, the ratio of meadow acreage to arable land was computed for a sample of farms. This ratio was then applied to the total arable land acreage of the county, for which a separate estimate was available. And in 1830, the proposal was made by an official in a forestry board to estimate the volume of timber in a forest by means of a "strip survey method".

II. THE PERIOD 1900-1950

3. The main features. The potential of sample surveys in official statistics was slowly being understood. To the extent that sample surveys were used during this period, the design typically called for systematic sampling, whenever this was operationally feasible. In many applications, the sampling fraction was 1/10 or 1/5. In the 1940's, a major factor favouring total surveys was the war-time economy with its regulations and rationing. This influence, which lasted roughly until the end of that decade, was however counteracted by the introduction of Gallup polls into Sweden and especially by the spectacular accuracy of the Gallup Institute's forecast of the 1944 election. In particular, these trends were followed with interest by official statisticians.

4. The 1911 Forest Survey in Värmland. The essential feature of the design was that the volume of timber was measured on sample plots along 10 meter wide strips covering the area of Värmland. It is worth noting that the "representative characteristics" of the survey were analysed by means of probability theory.

[1] Tore Dalenius, Brown University, Carl-Erik Särndal, Université de Montréal.
The circumstances did not permit the authors to discuss the contents of this paper with representatives of Statistics Sweden.

5. The 1911 Housing Survey in Göteborg. This survey was carried out by the municipal statistical office in Göteborg. The selection of the sample of apartments was based on an urn scheme. Each building in Göteborg was represented by a slip with identification data. The slips were thoroughly mixed in an urn and a 20% sample of slips was selected. The motive behind the scheme was to avoid that the survey be criticized for using a biased sample. The urn scheme was described by the person in charge of the survey as the only method "which can be called representative".

6. The 1935-36 Partial Population Census. This sample census used an elaborate scheme of controlled selection. The results from this census played a decisive role in an intense debate in Sweden concerning a "population crisis" which was feared as a result of low birth rates at the time.

III. THE PERIOD AFTER 1950

7. The beginnings of a new era. The greatly improved international communications after the end of World War II contributed to making the statistical community in Sweden aware of the recent advancements in sample survey theory, methods, and applications in the United States and India, to mention two of the leading countries. The new developments were studied and discussed, for example, at the conference of the Scandinavian statisticians in Helsinki in 1949. Statisticians were proud to be able to "talk sample survey methods"; to be sure, in some cases this ability was limited to knowledge of certain technical terms, notably "stratification". Mention should also be made of the influence exercised by the United Nations and affiliated agencies such as the Food and Agriculture Organization. In the following we give some examples of sample surveys and related developments of methods and theory. For cases dating to the early 1950's, details are found in Dalenius (1957).

8. The 1950 sample inventory of acreages and livestock. In the 1930's, sample surveys were used to estimate acreages of various crops and animal stocks. These surveys were referred to as "representative counts". They were based on nonprobability selection of farms. The aim, which however was not achieved, was to select 1/10 of the farms in each of several size-groups into which the farms had been divided. In the 1940's, these surveys were carried out on a total basis. A decision was made for the 1950 survey to return to sampling. The design that was suggested and largely implemented for the 1950 survey represented a partial break with the classical tradition of selecting every tenth unit. While the total sample size was fixed by the government authorities to be 1/10 of the total number of farms in the target population, the new design called for stratifying the farms by size groups based on acreage and using minimum variance allocation, which implied a selection of relatively speaking more large farms than small farms. It is interesting to note that the government authorities responsible for assessing the design felt it necessary to consult the U.N. Subcommission on Statistical Sampling about the appropriateness of the drastic deviation from the "every tenth unit rule". The Subcommission wholeheartedly endorsed the design. Consequently it was accepted in principle. The design provided considerable opportunity for research. In fact, three contributions to the theory of stratified sampling emerged, namely, (i) how best to divide a population into L strata; (ii) the best choice of the number of strata; and (iii) sample allocation to the strata for estimation of several parameters. The suggested design also called for addressing the problem of "measurement errors" in the acreage, and a special calibration survey was proposed. However, the authorities rejected this proposal.

9. Yield estimation. During World War II, the yield of various crops was estimated using data collected by "eye estimates" of the yield per unit area. By 1950, it was realized that this data collection method could be seriously biased. In the beginning of the 1950's, time was ripe for considering a different approach, namely, crop estimation based on harvesting sample plots, referred to as "objective crop estimation". Accordingly, a pilot study was carried out to test the use of this approach. The outcome of the test was convincing. From then on, the "objective" method has been used. As part of the pilot survey design, a scheme was developed for without replacement selection of $n = 2$ farms from a stratum with probabilities proportional to size, as discussed in Dalenius (1953). The scheme called for dividing each stratum at random into two parts, and selecting one farm from each part.

10. Developments relating to nonsampling errors. In the early 1950's, the problem of nonresponse received considerable attention in Sweden as in other countries. Surveys with 20-30% nonresponse were not unusual. This generated a vivid and sometimes heated debate in the statistical community about the distortion of the estimates. For a while, the statisticians seemed to have the problem under their control. The public concern about invasion of privacy has lately changed this picture; nonresponse has again become a serious problem. In the last 15 years, several contributions were made in the area of control of nonsampling errors. The problem of "evasive answer bias", to use the term introduced by S. Warner in connection with randomized response, was addressed in Swensson (1976). And Lyberg (1981) successully tackled the problem of controlling the coding operation in a population census or in a survey with interviews.

11. Respondent burden. In recent years there has been a growing concern about respondent burden and its negative effects on response rates. For example, the target population in many business surveys is the same, rather limited population. The problem can be alleviated by special sample selection techniques. The SAMU system for business surveys at Statistics Sweden permits "negative coordination" of samples, in the sense that samples without overlap can be selected with the technique known as JALES. To each unit in the sampling frame, a uniformly distributed random number is attached. This number stays with the unit, and is used in the selection of samples over time.

12. Modeling in combination with traditional probability sampling principles. Since the 1950's, the methodology for surveys had closely followed the strong probability sampling tradition established by Neyman and by Hansen and his co-workers in the United States. However, sometimes modeling is necessary in surveys when the traditional probability sampling theory is not sufficient. Since the 1970's the use of modeling in surveys has been explored. The book Foundations of Inference in Survey Sampling by Cassel, Särndal and Wretman (1977) exposed the new trends. Also, a number of papers by these and other Swedish authors showed how models may assist in inference from surveys. In recent years, methodologists at Statistics Sweden have shown unusual openness to incorporating modeling in the making of survey estimates. An early example where design-based and model-based ideas were combined is the "Öresund survey" for measuring traffic flow between Sweden and Denmark. The design is discussed in Cassel (1978). Some surveys are now designed with the aid of modeling assumptions, as in the work force survey described in Lundström (1987) and in an ongoing project of restructuring of the business survey sector.

13. Safeguarding privacy in surveys. In the last two decades, the general public has become increasingly concerned about invasion of privacy in connection with surveys, including population censuses, carried out by Statistics Sweden. As a result, there has been a trend

towards increasing nonresponse rates in some surveys. Several measures have been taken to deal with the problem: (i) Statistics Sweden has adopted the Ethical Declaration of the International Statistical Institute (1986); a translation of that declaration was distributed to all employees; (ii) In 1987, Statistics Sweden held an international conference which focused on policy issues (as distinguished from "techniques"); the discussions at the conference are summarized in Statistics Sweden (1987); (iii) Statistics Sweden has promoted the development of new safeguards for privacy in its surveys and has taken active steps to apply them. A review is given in Dalenius (1988). Of special interest are papers by Block and Olsson (1976), who describe a measure for the identifying power of quasi-identifiers, and Cassel (1976), who discussed probability-based disclosure.

14. Specific events. The increasing appreciation of sample surveys since around 1950 led to the creation of the Survey Research Center at Statistics Sweden in 1953. A similar interpretation may be given to the establishment of a professorship in "statistics, especially official statistics" at the University of Stockholm in 1965. Also, professorships in survey methodology were recently created at Statistics Sweden.

REFERENCES

BLOCK, H., and OLSSON, L. (1976). Bakvägsidentifiering. (Backwards identification) *Statistisk Tidskrift*, 1976, 135-144.

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

CASSEL, C.M. (1978). Probability based disclosure. In Dalenius, T., and Klevmarken, A. (eds.), *Proceedings of a symposium on personal integrity and the need for data in the social sciences*. Swedish Council for Social Science Research, Stockholm, 189-193.

CASSEL, C.M. (1978). On errors in the predictions with logit models. Technical report, Statistics Sweden.

DALENIUS, T. (1953). Något om metoder för objektiva skördeberäkningar. (About methods for objective crop estimation.) *Kungliga Lantbruksakademiens Tidskrift*, 92, 99-118.

DALENIUS, T. (1957). *Sampling in Sweden. Contributions to the Methods and Theory of Sample Survey Practice*. Stockholm: Almqvist and Wiksell.

DALENIUS, T. (1988). Controlling Invasion of Privacy in Surveys. Statistics Sweden.

INTERNATIONAL STATISTICAL INSTITUTE (1986). Declaration of Professional Ethics. *International Statistical Review*, 54, 227-242.

LUNDSTRÖM, S. (1987). Utveckling av estimatorer för skattning av antal förvärvsarbetande i olika arbetstidsklasser inom små redovisningsgrupper. R&D Report, U/STM 40, Statistics Sweden.

LYBERG, L. (1981). Control of the coding operation in statistical investigations. Urval no. 13, Statistics Sweden.

STATISTICS SWEDEN (1987). Statistics and Privacy: Future Access to Data for Official Statistics - Cooperation or Distrust? Statistics Sweden.

SWENSSON, B. (1977). Survey measurement of sensitive attributes. Ph.D. Thesis, Department of Statistics, University of Stockholm.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 99–103
Statistics Canada

Variance Estimation when a First Phase Area Sample is Restratified

PHILLIP S. KOTT[1]

ABSTRACT

This paper proposes an unbiased variance estimation formula for a two-phase sampling design used in many agricultural surveys. In this design, geographically defined primary sampling units (PSUs) are first selected via stratified simple random sampling; then secondary sampling units within sampled PSUs are restratified based on their characteristics and subsampled in a second phase of stratified simple random sampling.

KEY WORDS: Two-phase sample; Primary sampling unit; Secondary sampling unit; Unbiased.

1. INTRODUCTION

Suppose we have a sample of geographically defined primary sampling units (PSUs) drawn from a stratified area frame. Each sampled PSU contains a number of secondary sampling units (SSUs) which are restratified based on their characteristics. Subsamples of the SSUs are then drawn within each new stratum. To avoid confusion, only the original area strata will hereafter be referred to as strata; the new strata based on SSU characteristics will be referred to as *domains*. Stratified simple random sampling (srs) without replacement is performed at both phases of the sampling design.

This article derives an unbiased variance formula for the estimation strategy described above which is used in many agricultural surveys (for example, see Kott and Johnston 1988) but is not restricted to such surveys. The formula is a generalization of a suggestion by Cochran and Huddleston (1969, 1970), who assumed unstratified srs in the first sampling phase. It is also a special case of a variance formula in Särndal and Swensson (1987). The Särndal and Swensson formula (their equation (4.4)) depends on the calculation of a joint inclusion probability for each pair of subsampled SSUs. This proves cumbersome for the particular application under study because there are six distinct situations which need to be considered (depending on whether or not the two SSUs come from the same PSU, stratum, and/or domain). The derivation presented here follows a different line of reasoning entirely.

2. PRELIMINARIES

Suppose we start with an area survey consisting of n_h (out of N_h) PSUs from each of H strata. The SSUs within sampled PSUs are then restratified into D domains. Within domain d, m_d (out of M_d) SSUs are subsampled. Both phases of the sampling design are stratified srs without replacement.

Let us concentrate on estimating the total for a particular item of interest. To this end, let

[1] Phillip S. Kott, Senior Mathematical Statistician, Survey Research Branch, National Agricultural Statistics Service, USDA, S-4801, Washington, DC 20250, USA.

S^1 = denote the set of all SSUs within a PSU selected in the first phase of sampling whether these SSUs are in the subsample or not,

S_{hj} = denote the set of subsampled SSUs in PSU j of stratum h,

$S_{h\cdot}$ = denote the set of all subsampled SSUs in stratum h,

R_d = denote the set of all subsampled SSUs in domain d,

x_i = denote the value of interest for SSU i,

e_i = $(N_h/n_h)(M_d/m_d)x_i$ (assuming $i \in S_{h\cdot} \cap R_d$) be the "fully expanded" value of interest for SSU i,

$$e_{dhj} = \sum_{i \in S_{hj} \cap R_d} e_i,$$

$$e_{dh\cdot} = \sum_{i \in S_{h\cdot} \cap R_d} e_i,$$

$$e_{d\cdot\cdot} = \sum_{i \in R_d} e_i,$$

$$e_{\cdot hj} = \sum_{i \in S_{hj}} e_i, \text{ and}$$

$$e_{\cdot h\cdot} = \sum_{i \in S_{h\cdot}} e_i.$$

Note that when S_{hj} is empty, e_{dhj} and $e_{\cdot hj}$ are zero. Likewise when $S_{h\cdot}$ is empty, $e_{dh\cdot}$ and $e_{\cdot h\cdot}$ are zero, and when R_d is empty e_{dhj}, $e_{dh\cdot}$, and $e_{d\cdot\cdot}$ are zero.

An unbiased estimator for X, the sum of x_i values across all SSUs in the population, is

$$\hat{X} = \sum_{d=1}^{D} \sum_{i \in R_d} e_i. \tag{1}$$

To see this, observe that $\tilde{X} = \sum_{i \in S^1} (N_D/n_D)x_i$ is an unbiased estimator of X with respect to the first phase of sampling, while $\hat{X}$ is an unbiased estimator of $\tilde{X}$ with respect to the second sampling phase. Mathematically, $E_1(\tilde{X}) = X$ and $E_2(\hat{X}) = \tilde{X}$, which implies $E(\hat{X}) = E_1E_2(\hat{X}) = X$.

3. VARIANCE OF $\hat{X}$

From any of a number of textbooks on sampling theory (*e.g.*, Cochran 1977, p. 276), we know that the variance of a two-phase estimator like $\hat{X}$ is

$$\text{var}(\hat{X}) = \text{var}_1[E_2(\hat{X})] + E_1[\text{var}_2(\hat{X})], \tag{2}$$

where E_k and var_k denote, respectively, expectation and variance with respect to the k^{th} phase of sampling.

The first term in equation (2) is often called the first phase variance because it equals the variance that would be obtained if every SSU within a sampled PSU were part of the subsample. The second term in (2) is often called the second phase variance. It is easier to estimate than the first phase variance and we will attack it first. The problem with first phase variance estimation is that total value of interest for a PSU in the first phase sample can only be estimated using the subsample. As is well known, putting an estimated PSU total in place of a real total in the usual one-phase variance formula biases the resulting estimator.

3.1 Second Phase Variance Estimation

An unbiased estimator of $\text{var}_2(\hat{X})$ given *any* original sample is automatically an unbiased estimator of $E_1[\text{var}_2(\hat{X})]$. To see this, suppose that v_2 is an unbiased estimator of $\text{var}_2(\hat{X})$ given any sample. Since $E_2[v_2 - \text{var}_2(\hat{X})] = 0$ for *every possible* S^1, the first phase expectation of $E_2[v_2 - \text{var}_2(\hat{X})]$ must also be zero. Consequently, $E(v_2) = E_1E_2(v_2) = E_1[\text{var}_2(\hat{X})]$.

Now given our particular S^1,

$$\hat{\text{var}}_2 = \sum_{d=1}^{D} (1 - m_d/M_d)[m_d/(m_d - 1)]\left[\left\{\sum_{i\in R_d} e_i^2\right\} - e_{d\cdot\cdot}{}^2/m_d\right] \tag{3}$$

is the conventional unbiased estimator for $\text{var}_2(\hat{X})$. Moreover, equation (3) would hold whatever first phase sample obtained. As a result, $\hat{\text{var}}_2$ is also an unbiased estimator for $E_1[\text{var}_2(\hat{X})]$.

3.2 First Phase Variance Estimation

Consider a PSU j within stratum h. The value $e_{\cdot hj}$ is an unbiased estimator of (N_h/n_h) times the total value among all SSUs in PSU j whether in the current subsample or not. Consequently, $E_2(e_{\cdot hj})$ is exactly equal to (N_h/n_h) times the total value among all SSUs in PSU j. With this in mind, the following would be an unbiased estimator of the first phase variance of $\hat{X}$:

$$\hat{\text{var}}_1[E_2(\hat{X})] =$$

$$\sum_{h=1}^{H} (1 - n_h/N_h)[n_h/(n_h - 1)]\left[\sum_{j=1}^{n_h} \{E_2(e_{\cdot hj})\}^2 - \{E_2(e_{\cdot h\cdot})\}^2/n_h\right]. \tag{4}$$

Taken as is, equation (4) is of little use since it supposes we know what the $\{E_2(e_{\cdot hj})\}^2$ and $\{E_2(e_{\cdot h\cdot})\}^2$ are. Nevertheless, it does suggest that $\text{var}_1[E_2(\hat{X})]$ would be estimated in an unbiased manner if one could find unbiased estimators for the $\{E_2(e_{\cdot hj})\}^2$ and $\{E_2(e_{\cdot h\cdot})\}^2$ to plug into (4).

Observe first that $e_{\cdot hj}{}^2$ and $e_{\cdot h\cdot}{}^2$ are *not* unbiased estimators of $\{E_2(e_{\cdot hj})\}^2$ and $\{E_2(e_{\cdot h\cdot})\}^2$. In fact,

$$E_2(e_{\cdot hj}{}^2) = \{E_2(e_{\cdot hj})\}^2 + \text{var}_2(e_{\cdot hj}),$$

while (5)

$$E_2(e_{\cdot h\cdot}{}^2) = \{E_2(e_{\cdot h\cdot})\}^2 + \text{var}_2(e_{\cdot h\cdot}).$$

These equations hint towards alternative estimators for $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$. If v_{2hj} and v_{2h}, say, were unbiased estimators of $\text{var}_2(e_{.hj})$ and $\text{var}_2(e_{.h.})$, respectively, then ${e_{.hj}}^2 - v_{2hj}$ would be an unbiased estimator of $\{E_2(e_{.hj})\}^2$, while ${e_{.h.}}^2 - v_{2hj}$ would be an unbiased estimator of $\{E_2(e_{.h.})\}^2$.

From Cochran (1977, p. 143, eq. (5A.68)), one can see that

$$\hat{\text{var}}_{2hj} = \sum_{d=1}^{D} (1 - m_d/M_d)[m_d/(m_d - 1)]\left[\left\{\sum_{i \in S_{hj} \cap R_d} {e_i}^2\right\} - {e_{dhj}}^2/m_d\right]$$

and (6)

$$\hat{\text{var}}_{2h} = \sum_{h=1}^{H} (1 - m_d/M_d)[m_d/(m_d - 1)]\left[\left\{\sum_{i \in S_{h.} \cap R_d} {e_i}^2\right\} - {e_{dh.}}^2/m_d\right]$$

are, respectively, unbiased estimators of $\text{var}_2(e_{.hj})$ and $\text{var}_2(e_{.h.})$.

3.3 Putting It All Together

Observe that combining equations (3) and (6) can yield (after some manipulation) this estimator for the second phase variance of $\hat{X}$:

$$\hat{\text{var}}_2 = \sum_{h=1}^{H} [n_h/(n_h - 1)] \sum_{j=1}^{n_h} \hat{\text{var}}_{2hj} - \hat{\text{var}}_{2h}/(n_h - 1) +$$

$$\sum_{d=1}^{D} \left\{(1 - m_d/M_d)[1/(m_d - 1)] \cdot \right.$$

$$\left. \left(\sum_{h=1}^{H} [n_h/(n_h - 1)]\left[\left\{ \sum_{j=1}^{n_h} {e_{dhj}}^2 \right\} - {e_{dh.}}^2/n_h\right] - {e_{d..}}^2 \right) \right\}. \quad (7)$$

By plugging ${e_{.hj}}^2 - \hat{\text{var}}_{2hj}$ and ${e_{.h.}}^2 - \hat{\text{var}}_{2h}$ respectively into $\{E_2(e_{.hj})\}^2$ and $\{E_2(e_{.h.})\}^2$ of equation (4), we have the following estimator for the first phase variance of $\hat{X}$:

$$\hat{\text{var}}_1[E_2(\hat{X})] = \sum_{h=1}^{H} (1 - n_h/N_h)[n_h/(n_h - 1)] \cdot$$

$$\left[\left\{ \sum_{j=1}^{n_h} {e_{.hj}}^2 - \hat{\text{var}}_{2hj} \right\} - \{({e_{.h.}}^2 - \hat{\text{var}}_{2h})\}/n_h\right].$$

This can then be added to (7) to yield the following estimator for the variance of $\hat{X}$ in (1):

$$\hat{\text{var}} = A + B + C, \quad (8)$$

where

$$A = \sum_{h=1}^{H} [n_h/(n_h - 1)] \left[\left\{ \sum_{j=1}^{n_h} e_{\cdot hj}{}^2 \right\} - e_{\cdot h\cdot}{}^2/n_h \right],$$

$$B = \sum_{d=1}^{D} \left\{ (1 - m_d/M_d)\ [1/(m_d - 1)] \cdot \left(\sum_{h=1}^{H} [n_h/(n_h - 1)] \left[\left\{ \sum_{j=1}^{n_h} e_{dhj}{}^2 \right\} - e_{dh\cdot}{}^2/n_h \right] - e_{d\cdot\cdot}{}^2 \right) \right\},$$

$$C = - \sum_{h=1}^{H} f_h n_h/(n_h - 1) \left[\sum_{j=1}^{n_h} \{e_{\cdot hj}{}^2 - \hat{\text{var}}_{2hj}\} - \{e_{\cdot h\cdot}{}^2 - \hat{\text{var}}_{2h}\}/n_h \right],$$

$f_h = n_h/N_h$ is the first phase sampling fraction in stratum h, and $\hat{\text{var}}_{2hj}$ and $\hat{\text{var}}_{2h}$ are defined by equation (6).

Observe that if all the first phase sampling fractions are very small, then the contribution of C to (8) can be ignored. In any event dropping C would at worst give vâr an upward bias, since $E(C) \leq 0$.

Observe further that vâr would collapse to A if – in addition to C being ignorably small – the sampling design had been conventional two-stage sampling; that is, if each domain had been contained within one of the originally sampled PSU's so that $y_{d\cdot\cdot} = y_{dhj} = y_{dh\cdot}$ and $B = 0$. This should not be surprising, since A is the standard variance estimator in two stage sampling when the first stage is srs with replacement (Cochran 1977, p. 307). Ignorable first stage sampling fractions blur the distinction between srs with and without replacement.

The right hand side of (8) can, in principle, be negative. This is because B is often negative (since $y_{d\cdot\cdot} \geq y_{dh\cdot} \geq y_{dhj}$), while A can theoretically be as small as zero. Kott and Johnston (1988) applied a formula similar to (6) to data from a US Department of Agriculture survey. In the 41 cases they examined the absolute value of B was always less than 7% of A.

One final note. Since $B \leq 0$ and $E(C) \leq 0$, using A alone provides a conservative, unambiguously nonnegative, estimate for $\text{var}(\hat{X})$.

REFERENCES

COCHRAN, R., and HUDDLESTON, H. (1969). Unbiased estimates for stratified subsample designs. U.S. Department of Agriculture, Statistical Reporting Service.

COCHRAN, R., and HUDDLESTON, H. (1970). Unbiased estimates for stratified subsample design. *Proceedings of the Section on Social Statistics, American Statistical Association*, 265-267.

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

KOTT, P.S., and JOHNSTON, R. (1988). Estimating the non-overlap variance component for multiple frame agricultural surveys. RAD Staff Report No. SRB-NERS-8805, U.S. Department of Agriculture, National Agricultural Statistics Service.

SÄRNDAL, C.E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 105–116
Statistics Canada

Estimation Using Double Sampling and Dual Stratification

DONALD B. WHITE[1]

ABSTRACT

The problem considered is that of estimation of the total of a finite population which is stratified at two levels: a deeper level which has low intrastratum variability but is not known until the first phase of sampling, and a known pre-stratification which is relatively effective, unit by unit, in predicting the deeper post-stratification. As an important example, the post-stratification may define two groups corresponding to responders and non-responders in the situation of two-phase sampling for non-response. The estimators of Vardeman and Meeden (1984) are employed in a variety of situations where different types of prior information are assumed. In a general case, the standard error relative to that of the usual methods is studied via simulation. In the situation where no prior information is available and where proportional sampling is employed, the estimator is unbiased and its variance is approximated. Here, the variance is always lower than that of the usual double sampling for stratification. Also, without prior information, but with non-proportional sampling, using a slight modification of the second phase sampling plan, an unbiased estimator is found along with its variance, an unbiased estimator of its variance, and an optimal allocation scheme for the two phases of sampling. Finally, applications of these methods are discussed.

KEY WORDS: Two-phase sampling; Prior information; Variance estimation; Optimal allocation; Non-response.

1. INTRODUCTION

Various stratified sampling designs employ various types of prior information. For example, the usual stratification model assumes full prior knowledge of individual stratum memberships. Post-stratification is useful when there is global information on stratum sizes but no information on individuals. Double sampling for stratification, on the other hand, assumes no prior information on strata. Further, some knowledge of the population values is necessary, for example, for the allocation of sampling resources among strata (see, for example, Cochran 1977, pp. 96-99 and 331-332).

The rigid assumptions inherent to these sampling designs and population models often are not satisfied due to the discrepancy between the population under study and the (possibly dated) prior information. Seeking to appropriately handle this discrepancy, Vardeman and Meeden (1984) have introduced a pair of estimators which combine information on stratum memberships, stratum sizes, and stratum averages with analogous information gained from the current sample. Their two estimators apply to two essentially different situations. The first is where the prior information is global only, *i.e.*, only on stratum sizes and averages. The second estimator applies where there is also partial information on individual stratum memberships. Here, the population is stratified according to various factors, some of which are known and some of which, though not known, may be inexpensive to determine on a first phase of sampling.

[1] Donald B. White, Department of Statistics, State University of New York at Buffalo, 249 Farber Hall, Buffalo, New York 14214.

As an example, consider the use of sampling to determine the spread of an infectious disease. If detection of infection is expensive, then stratification, according to risk categories, is desirable to reduce the second phase sample size. Factors determining risk categories may include gender, age, place of residence, ethnicity, health habits, and contact with potential carriers. As some of these factors are not known prior to sampling, the model of Vardeman and Meeden can be employed since the true risk categories can be predicted by the known factors.

Another example is two-phase sampling for non-response. Extending the method of Hansen and Hurwitz (1946), we have a population which is divided into two post-strata, *i.e.*, responders and non-responders. The methods discussed here apply when there is some prior information which classifies units into pre-strata which are then used to predict whether or not the unit will be in the group of responders.

The notion of employment of prior information in two-phase designs is not without precedent in the sampling literature. As an example, Han (1973) has used prior information on an auxiliary regression variable (to be measured in a first phase sample) to construct a simple hypothesis (say H_0) regarding the mean of that variable. The first phase sample measurements are then used to test H_0. If H_0 is accepted, the value specified by H_0 is used in the estimator; if it is rejected, the sample average is used.

A discussion of the use of the first estimator of Vardeman and Meeden (global information only) can be found in White (1987). There, optimal choices of the weighting constants for prior information relative to the information contained in the current sample were determined. Here, the situation considered is where prior information is also available on individual stratum memberships. After introducing the necessary notation in Section 2, we explore a simulated example in Section 3. In Section 4, in two different sampling situations, unbiased estimators are analyzed in terms of variance, unbiased estimation of the variance, and optimal allocation of sampling resources. In Section 5, applications of these techniques are discussed.

2. THE POPULATION MODEL AND SAMPLING SCHEME

We now present the population model and the proposed sampling design. We begin with a finite population P of units labelled 1, 2, ..., N with associated unknown values $y_1, y_2, \ldots, y_N$. Denote the population total by $\tau = \sum_{i=1}^{N} y_i$. For $1 \leq i \leq N$, unit i also possesses an unknown post-stratum membership j_i, $1 \leq j_i \leq J$, and a known pre-stratum membership k_i, $1 \leq k_i \leq K$.

A variety of population quantities require a specialized notation. Such quantities include sizes of groups, group averages and group variances. Subscripts will identify the group involved: no subscript implies reference to the entire population, "$k\cdot$" refers to pre-stratum k, $1 \leq k \leq K$, "$\cdot j$" refers to post-stratum j, $1 \leq j \leq J$, and the subscript "kj" refers to the intersection of pre-stratum k with post-stratum j. The base symbols N, $\bar{Y}$ and S^2 refer to number of elements, y-average, and finite population variance, respectively. Also, we let P, $P_{k\cdot}$, $P_{\cdot j}$ and P_{kj} denote the subsets of P corresponding to the four categories given above. For example, we have

$$S_{k\cdot}^2 = \frac{1}{N_{k\cdot} - 1} \sum_{i \in P_{k\cdot}} (y_i - \bar{Y}_{k\cdot})^2.$$

Also, we can write

$$\tau = \sum_j N_{.j}\bar{Y}_{.j}. \tag{1}$$

We finally let $W_{kj} = N_{kj}/N_{k.}$, *i.e.*, W_{kj} is the proportion of units in pre-stratum k which fall into true stratum j.

We now discuss the sampling technique. In the first phase of sampling, a stratified simple random sample without replacement s' is selected, with $n'_{k.}$ units (first phase sampling fraction denoted by $f'_{k.} = n'_{k.}/N_{k.}$) selected from pre-stratum k. Samples from different pre-strata are independent. For these $n' = \sum_k n'_{k.}$ units, post-strata, j_i, are observed. Following the notational pattern given above, we let n'_{kj} denote the number of units in s' sampled from pre-stratum k which happen to fall in post-stratum j. Also, $n'_{.j} = \sum_k n'_{kj}$ is the total number of units in s' which fall in post-stratum j. This set of units is denoted by $s'_{.j}$. These quantities are observed, while quantities involving y-values, such as $\bar{y}'$ and $s^{2'}$ (with all four types of subscripts), remain unobserved. Here, and in the following, the average of any empty collection is taken as zero, and, if the size of a group is one or zero, we take its variance s^2 to be zero. We note that for $1 \leq k \leq K$, the random vectors $(n'_{k1}, \ldots, n'_{kJ})$ are independent with each possessing a multivariate hypergeometric distribution.

For the second phase of sampling, we partition s' into $\cup_{j=1}^{J} s'_{.j}$, *i.e.*, by post-stratification. For each j, let $v_j(\cdot)$ denote a known function on and into the non-negative integers with $v_j(0) = 0$ and $1 \leq v_j(x) \leq x$ if $x \geq 1$. The second phase sample s is also stratified, but now is a subsample of s' and stratified according to the post-stratification. The sample from $s'_{.j}$ is denoted $s_{.j}$ and is of size $n_{.j} \equiv v_j(n'_{.j})$. Here, y-values are observed, yielding quantities such as $\bar{y}_{.j}$ and $s^2_{.j}$, the y-average and finite population variance of the units in the phase two sample and stratum j.

The estimates of τ given by Vardeman and Meeden include the option of inclusion of prior guesses for the relative stratum sizes within each pre-stratum and for the stratum averages. Thus, we have prior guesses for the values W_{kj} and $\bar{Y}_{.j}$ which are given by Π_{kj} and $\mu_{.j}$, respectively. In the estimator introduced below, these guesses are given weighting constants which reflect the confidence in the guess relative to the confidence in the corresponding information yielded by the current sample. For each k, the confidence value allotted to the collection $(\Pi_{k1}, \ldots, \Pi_{kJ})$ is denoted $\tilde{M}_{k.} \in [0,\infty]$ and for each j, the confidence value given to $\mu_{.j}$ is denoted $M_{.j} \in [0,\infty]$. In the current sample, the collection $(W_{k1}, \ldots, W_{kJ})$ is estimated by $(n'_{k1}/n'_{k.}, \ldots, n'_{kJ}/n'_{k.})$ and is based on a simple random sample of size $n'_{k.}$. Thus, the confidence in Π_{kj}, say, as opposed to $n'_{kj}/n'_{k.}$, is reflected by the size of $\tilde{M}_{k.}$ versus that of $n'_{k.}$. Similarly, in the current sample, $\bar{Y}_{.j}$ is estimated by $\bar{y}_{.j}$ and is based on a sample of size $n_{.j}$; thus, the relative confidence in the prior guess and the current estimate is reflected by the relative sizes of $M_{.j}$ and $n_{.j}$. Any confidence weight for prior information equal to zero corresponds to no use of the prior information, and, as in the use of stratum sizes in the usual post stratification model, a value of infinity implies no use of the corresponding information in the current sample.

Using the prior guesses, current estimates and confidence weights, we estimate W_{kj} and $\bar{Y}_{.j}$ by $\hat{\Pi}_{kj} = (\tilde{M}_{k.}\Pi_{kj} + n'_{kj})/(\tilde{M}_{k.} + n'_{k.})$ and $\hat{\mu}_{.j} = (M_{.j}\,\mu_{.j} + n_{.j}\bar{y}_{.j})/(M_{.j} + n_{.j})$, respectively. Finally, an estimate $\hat{\tau}$ of the population total τ is constructed by replacing in the formula (1) for τ any unobserved quantity by its estimate given above. Thus, we employ

$$\hat{\tau} = \sum_{j=1}^{J} \left\{ n_{.j}\bar{y}_{.j} + (n'_{.j} - n_{.j})\hat{\mu}_{.j} + \sum_{k=1}^{K} (N_{k.} - n'_{k.})\hat{\Pi}_{kj}\hat{\mu}_{.j} \right\}. \tag{2}$$

Computation of the bias and variance of $\hat{\tau}$ in the general case is left open by Vardeman and Meeden. The case $K = 1$ and $M_{\cdot j} = 0, 1 \leq j \leq J$, has been studied in White (1987). Before proceeding to a result in a more complex situation, we first explore the results of a simulation on a hypothetical population.

3. A MONTE CARLO STUDY

Here we present a specific population and sampling scheme which is modelled after the introductory example regarding estimation of the spread of an infectious disease. For a population of 10,000 individuals who are susceptible, the disease is assumed to be more prevalent among the 5,000 who live in the western section of the area considered. Since this is a known characteristic, the population is partitioned according to the east-west boundary into $K = 2$ pre-strata. Next, we assume that certain easily obtained additional information enables the sampler to categorize the individual as low, medium, or high risk for becoming infected. See Table 1 for the details of the construction of the population.

For estimation of the total number infected ($\tau = 2302$), we assume no prior knowledge of the stratum proportions $\bar{Y}_{\cdot 1}$, $\bar{Y}_{\cdot 2}$, and $\bar{Y}_{\cdot 3}$ and thus take $M_{\cdot 1} = M_{\cdot 2} = M_{\cdot 3} = 0$. There remain four major ingredients to the estimation process: 1) the prior guesses $\{\Pi_{kj}: k = 1, 2, j = 1, 2, 3\}$ for the distribution of individuals from pre-strata to post-strata, 2) the weighting constants $\tilde{M}_1$ and $\tilde{M}_2$ given to these prior guesses, 3) the first phase sample design and outcome, and 4) the second phase sample design and outcome. These are detailed in the following.

First, in White (1987) it was found for the $K = 1$ case that an effective choice of weighting constants was to select M equal to the sample size on which the previous information was based. Following that notion, we allowed, for each simulation, the collection $\{\Pi_{kj}\}$ to select itself through a preliminary sample of size m (either 500 or 2500) from each pre-stratum. That is, Π_{kj} is taken to be the proportion of the m individuals from pre-stratum k falling in post-stratum j.

Second, for each run, the weighting constants were taken as $\tilde{M}_{1\cdot} = \tilde{M}_{2\cdot} = M$ for all $M \in \{0, 100, 200, 300, \ldots, 10{,}000, \infty\}$. Recall that $M = \infty$ corresponds to the situation of the usual post-stratification where no use is made of the current sample to estimate group sizes.

Third, the first phase sample is stratified according to pre-strata with sampling fractions $f'_{k\cdot}$ taken to be $f'_{1\cdot} = f'_{2\cdot} = f$, $f \in \{.10, .20, .30, .40, .50\}$. Recall that in this phase of sampling, only post-stratification is observed. This information is, presumably, inexpensive to obtain.

Table 1

Number Infected/Group Size for the Pre-strata and Post-strata Combinations

Location of Residence	Risk Group j	Low 1	Medium 2	High 3	Total
East ($k = 1$)		40/4000	80/800	100/200	220/5000
West ($k = 2$)		2/200	80/800	2000/4000	2082/5000
Total		42/4200	160/1600	2100/4200	2302/10000

On the other hand, sampling a unit in phase 2, where the presence of infection is determined, is assumed to be rather expensive. The individuals selected are a subsample of the phase one sample, stratified according to post-strata. The sampling fractions in various strata are again taken as equal ($v_j(n'_j) = [c_j n'_j]$ for n'_j large enough, and $c_1 = c_2 = c_3 = c$) and so that different simulations can be compared, c is selected so that the fraction of the entire population which appears in the phase 2 sample remains constant at .10.

Now, the following process is repeated $R = 50{,}000$ times: obtain a preliminary sample of size m from which prior guesses Π_{kj} for W_{kj} are constructed. Next, a sample, stratified according to pre-strata with sampling fractions f, is obtained. Only post-stratification is observed. Then, a subsample, stratified according to post-strata with sampling fractions c, is obtained and units in this sample are classified as infected or not infected. Finally, on each run, $\hat{\tau}$ is obtained for each value of M considered. The standard error of $\hat{\tau}$ is estimated using the R simulated values of $\hat{\tau}$. Recall, however, that in a real-life application, the standard error of an estimate will depend on the particular values of Π_{kj} used; here, these values are different on each run and thus the estimated standard error should be viewed as a long run average for a mixture of distributions of $\hat{\tau}$, mixed according to the distribution of the Π_{kj} based on the preliminary sample.

The simulations were performed on an IBM3031 computer. For this example, where $y_i \in \{0,1\}$ for all i, all random quantities are functions of independent hypergeometric or multivariate hypergeometric variables. Using the fact that the conditional distribution of a univariate marginal of a multivariate hypergeometric distribution given any subcollection of the other coordinates is itself hypergeometric, all random quantities were simulated using the IMSL 92DP hypergeometric simulation subroutine GGHPR. For the first combination of m and f (500 and .10), the simulation process was repeated five times to check internal consistency.

Tables 2 and 3 summarize pertinent characteristics of the variation of the simulated SE($\hat{\tau}$) as a function of M for the five repeated simulations (Table 2), and the simulations for various values of f and m (Table 3). Table 2 gives only highlights which demonstrate internal consistency and confirm that the number of repetitions is chosen large enough. Note that M_0 denotes the value of M for which SE($\hat{\tau}$) is minimized. In Table 3, also given is a comparison with the better of the possible usual techniques (regular two-phase or stratified according to pre-strata) relative to the ideal where the true strata are regarded as known. The standard error of an estimator based on stratified sampling using pre-strata only is 113.27, and for stratified according to true strata, it is 105.47. Thus, letting the estimator in regular two-phase sampling be denoted by $\hat{\tau}_2$ and realizing that SE($\hat{\tau}_2$) depends upon f and c, the values appearing in the columns headed Percent Relative Reduction in SE($\hat{\tau}$) are 100 [min(SE($\hat{\tau}_2$), 113.27)] − SE($\hat{\tau}$)/[min(SE($\hat{\tau}_2$), 113.27) − 105.47].

Table 2

Key Features of the Repeated Runs with $m = 500$, $f = .10$ and $c = 1.0$

Run #	M_0	SE($\hat{\tau}$)			
		$M = 0$	$M = m$	$M = M_0$	$M = \infty$
1	600	113.55	109.67	109.62	112.00
2	700	113.42	109.50	109.45	111.80
3	700	113.92	109.86	109.78	112.00
4	600	113.61	109.71	109.66	112.07
5	600	113.56	109.74	109.70	112.17

Table 3

Key Features of SE($\hat{\tau}$) as a Function of M

					SE($\hat{\tau}$)				Percent Relative Reduction in SE($\hat{\tau}$)		
m	f'	c	SE($\hat{\tau}_2$)	M_0	$M = M_0$	$M = 0$	$M = m$	$M = \infty$	$M = 0$	$M = m$	$M = \infty$
500	.10	1.00	126.29	600	109.62	113.55	109.67	112.00	−3.6	46.2	16.3
500	.20	.50	115.19	600	107.95	109.02	107.97	110.72	54.5	67.9	32.7
500	.30	.33	111.80	600	107.87	108.25	107.87	110.38	56.1	62.1	22.4
500	.40	.25	109.22	750	106.51	106.76	106.52	108.29	65.6	72.0	24.8
500	.50	.20	107.98	700	106.17	106.28	106.18	107.55	67.7	71.7	17.1
2500	.10	1.00	126.29	*	≤ 106.20	113.33	106.42	106.20	−0.8	87.8	90.6
2500	.20	.50	115.19	*	≤ 105.76	108.67	106.02	105.76	59.0	92.9	96.3
2500	.30	.33	111.80	*	≤ 106.63	108.18	106.87	106.63	57.2	77.9	81.7
2500	.40	.25	109.22	*	≤ 105.77	106.59	105.94	105.77	70.1	87.5	92.0
2500	.50	.20	107.98	*	≤ 105.81	106.34	105.96	105.81	65.3	80.5	86.5

* -- > 10,000

A variety of important results can be discerned from Table 3. First is that for $m = 500$, M_0 is very close to, although always slightly larger than, m. This is the result predicted by the $K = 1$ situation from White (1987). For $m = 2500$, though in every case $M_0 > 10{,}000$, one discovers that SE($\hat{\tau}$) at $M = m$ is very close to the minimum at $M = M_0$.

Second is that at $M = m$, the percent relative reduction in SE($\hat{\tau}$) ranges from a minimum of 46% to over 90%. Also, at $M = 0$, corresponding to the situation of dual stratification with no prior information on any population characteristic, the percent relative reduction in SE($\hat{\tau}$) is always over 50% except in the case of the smallest first phase sampling fraction, $f = .10$. In that case, when prior information is not available and the first phase sample size is small, one is better off to use the pre-strata and ignore the true stratification. On the other hand, if one does have a set of prior guesses available for the collection of W_{kj}, but is uncertain of what weights to attach to these values, one could use the usual post-stratification notion of using weight $M = \infty$. If the prior information is good, as in our case $m = 2500$, then the percent relative reduction in SE($\hat{\tau}$) is always over 80%. Even if the prior information is only moderately accurate, as in the case $m = 500$, the reduction in standard error is between 16% and 33%.

In summary, if one is able to identify a weighting constant applicable to prior information on the distributions of units among strata, then a substantial reduction in standard error can be obtained using these methods. Even if one cannot identify such a constant or does not have applicable prior information, one can still decrease standard error using dual stratification by taking $M = 0$ if the prior information on W_{kj} is either poor or non-existent, or $M = \infty$ with accurate prior information. In particular, it thus turns out that the case $M = 0$ is important. This case is examined in detail in the next section.

4. BIAS, STANDARD ERROR, AND OPTIMAL ALLOCATION WITH NO PRIOR INFORMATION

When no prior information is available, we set $M_j = 0$ and $\tilde{M}_k = 0$ for each $1 \leq j \leq J$ and $1 \leq k \leq K$. In this section, we at first also assume that sampling in both phases is proportional to the size of the group from which the sample is drawn, that is, for each

k, $n'_{k\cdot} = fN_{k\cdot}$ (*i.e.*, $f'_{k\cdot} = f$, all k) and for each j, $n_{\cdot j} = cn'_{\cdot j}$ (*i.e.*, $v_j(x) = cx$, all j). This, of course, immediately introduces an approximation (referred to in what follows as approximation A1), since the resulting sample sizes are not necessarily integers. However, in reasonably large populations, and for reasonably large sampling fractions f and c, this approximation has little impact on the derivations that follow.

In this situation, $\hat{\mu}_{\cdot j}$ reduces to $\bar{y}_{\cdot j}$ and $\hat{\Pi}_{kj}$ reduces to $n'_{kj}/n'_{k\cdot}$ and, thus, we have $\hat{\tau} = 1/f \sum_{j=1}^{J} n'_{\cdot j}\bar{y}_{\cdot j}$. The derivations of the expectation and variance of $\hat{\tau}$ are summarized in the appendix. The key features are two conditioning arguments: first, we condition on s' since the second phase sample is a function of s' and, second, because of the multivariate hypergeometric nature of the phase one sample, we condition on the values n'_{kj}, the sizes of the various pre-stratum and post-stratum combinations in the first phase sample.

In the appendix, we show first that $\hat{\tau}$ in this case is unbiased (aside from approximation A1) and that an approximation of its variance is given by

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f} \sum_{k} N_{k\cdot} S^2_{k\cdot} + \frac{1-c}{fc} \sum_{j} N_{\cdot j} S^2_{\cdot j}. \tag{3}$$

As discussed in the appendix in more detail, formula (3) 1) gives answers close to the simulated values, 2) is based on approximations whose error is small for large populations and reasonably large samples, and 3) reduces to the exact formula in all three of the standard situations. In addition, it is easy to show that the variance given by (3) is always smaller than that of the situation of regular two phase sampling.

Now as in any stratification model, there is a question of optimal design. The problem addressed here is that of minimum variance given a fixed cost. To this end, we let $T_1 = \sum_k N_{k\cdot} S^2_{k\cdot}$ and $T_2 = \sum_j N_{\cdot j} S^2_{\cdot j}$. We assume, for the design question at hand, that these are known. In reality, of course, only guesses are available. Next, we let D denote the total budget, d_0, the start-up cost, d_1, the cost per unit in the phase one sample, and d_2, the cost per unit in the phase two sample. Letting D_a denote the number of dollars available for sampling per population unit, we have

$$D_a = \frac{D - d_0}{N} = f(d_1 + cd_2). \tag{4}$$

With f and c subject to constraint (4), we seek to minimize (3), var($\hat{\tau}$), now given by

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f} T_1 + \frac{1-c}{fc} T_2. \tag{5}$$

The solution is easily found to be given by

$$c = \left[\frac{d_1 T_2}{d_2(T_1 - T_2)}\right]^{1/2} \tag{6}$$

with f found using (4). If $T_1 \leq T_2$, we automatically take $c = 1$ since then the pre-stratification is more effective than the post-stratification.

In the case of non-proportional sampling, the estimator given is biased and calculations of the bias and standard error in this more general situation are prohibitive. However, a slight modification of the second phase sampling design along with the associated change in the estimator $\hat{\tau}$ yields an estimator which is unbiased. Following a description of the required modification, we compute the variance and an unbiased estimator of the variance and we find an optimal method of allocating sampling resources to the various pre- and post-strata.

The modification to the sampling plan is to leave the second phase sample within pre-strata rather than pooling within post-strata across pre-strata. Thus, given n'_{kj} units appearing in $s' \cap P_{kj}$, we have a function $v_{kj}(\cdot)$ (like $v_{\cdot j}(\cdot)$ in Section 2) which defines a sample size $n_{kj} = v_{kj}(n'_{kj}) = c_{kj} n'_{kj}$ to be taken by simple random sampling from $s' \cap P_{kj}$. Based upon this sample, we obtain the quantities $\bar{y}_{kj}$ and s^2_{kj} which were defined in Section 2. The estimator is now $\hat{\tau} = \sum_k 1/f'_{k\cdot} \sum_j n'_{kj} \bar{y}_{kj}$.

Now, since samples (and thus estimators) are independent between pre-strata, $\hat{\tau}$ is the sum of independent estimators of the K pre-stratum totals, where each estimator is based on a regular double sampling scheme. Thus, the results of Rao (1973) apply to each pre-stratum and we first observe that $\hat{\tau}$ is unbiased because its summands are unbiased estimators of their respective pre-stratum totals. Second, using Rao's results, we have

$$\operatorname{var}(\hat{\tau}) = \sum_k \frac{1}{f_{k\cdot}} \left[(N_{k\cdot} - n'_{k\cdot}) S^2_{kj} + \sum_j N_{kj} S^2_{kj} (1/c_{kj} - 1) \right]. \tag{7}$$

Also, an unbiased estimator of $\operatorname{var}(\hat{\tau})$ is given by

$$\hat{\operatorname{var}}(\hat{\tau}) = \sum_k N_{k\cdot} \left[(N_{k\cdot} - 1) \sum_j \left(\frac{n'_{kj} - 1}{n'_{k\cdot} - 1} - \frac{n'_{kj} - 1}{n'_{k\cdot} - 1} \right) \frac{n'_{kj} s^2_{kj}}{n'_{k\cdot} n_{kj}} \right.$$

$$\left. + \frac{N_{k\cdot} - n'_{k\cdot}}{N_{k\cdot}(n'_{k\cdot} - 1)} \sum_j \frac{n'_{kj}}{n'_{k}} \left(\bar{y}_{kj} - \sum_{j'} \frac{n'_{kj'}}{n'_{k}} \bar{y}_{kj'} \right)^2 \right]. \tag{8}$$

We note at this point that in the case of proportional sampling considered earlier in this section, we have proposed two different estimators for τ, one based on a pooled second phase sample, the other unpooled. In both cases, the estimator was found to be unbiased, and, also, reduction of formula (7) to the case where $f'_{k\cdot} = f$ for all k and where $c_{kj} = c$ for all k and all j yields formula (3), *i.e.*, the approximate variance for the pooled second phase sampling estimator.

Finally, again following the results in Rao, we derive an optimal allocation of sampling resources. Say that D dollars are available for the two phases of sampling, where sampling a unit in phase 1 from $P_{k\cdot}$ costs $d'_{k\cdot}$ dollars and sampling a unit in phase 2 from $P_{\cdot j}$ costs $d_{\cdot j}$ dollars. Given these costs, we wish to find the values of $f'_{k\cdot}$ and c_{kj} which minimize the variance of $\hat{\tau}$. Using the Cauchy inequality for the phase 2 sample in each pre-stratum, we observe that no matter what the value of $f'_{k\cdot}$, the sampling fraction from post-stratum j is given by

$$c_{kj} = S_{kj} \left(\frac{d'_{k\cdot}}{d_{\cdot j} (S^2_{k\cdot} - \sum_j W_{kj} S^2_{kj})} \right)^{1/2}. \tag{9}$$

Now, the effective expected cost (over both phases of sampling) for each unit sampled in phase 1 and in pre-stratum k is given by

$$d_{k\cdot}^{(e)} = d'_{k\cdot} + \sum_j W_{kj}\, c_{kj}\, d_j . \tag{10}$$

When viewed in this way, for cost considerations, the first phase of sampling can be seen as a regular stratified sample with (effective) cost of a unit sampled in $P_{k\cdot}$ given by (10). Thus, Cochran (1977,p.97) provides the required formulation of the first phase allocation:

$$\frac{n'_{k\cdot}}{n'} = \frac{N_{k\cdot}\, S_{k\cdot}/\sqrt{d_{k\cdot}^{(e)}}}{\sum_{k'} N_{k'\cdot}\, S_{k'\cdot}/\sqrt{d_{k'\cdot}^{(e)}}} \tag{11}$$

where

$$n' = \sum_k n'_{k\cdot} = D \sum_k \frac{N_{k\cdot}\, S_{k\cdot}/\sqrt{d_{k\cdot}^{(e)}}}{\sum_{k'} N_{k'\cdot}\, S_{k'\cdot}/\sqrt{d_{k'\cdot}^{(e)}}} . \tag{12}$$

Following the modifications suggested by Rao, one can handle the situation where one or more of the c_{kj} turn out to be greater than one. One can also modify the results in the usual way to minimize sampling cost in the case of pre-determined variance.

5. APPLICATIONS

One can employ the method of dual stratification presented here at two levels. At one level, double sampling with pre-strata can be employed with no use of prior information on stratum sizes or stratum averages. At a more complex level, if one has in hand prior information on the number of units in each stratum coming from each pre-stratum, and if the sampler has a level of confidence for this information, then a further reduction in standard error can be obtained by employing this prior information.

This two phase sampling and estimation technique could be used in the proposed nation-wide survey to determine the extent of spread of the HTLV-III (Acquired Immune Deficiency Syndrome) virus. The extended incubation period, estimated to be on the average 4.5 years (Lui *et al.* 1986), makes the survey approach imperative, yet there are psychosocial and financial factors which make such a survey extremely difficult to carry out. Thus, methods which assist in reducing sample size while maintaining accuracy must be pursued.

Allen (1984) provides data which suggests a partition of the American population according to a variety of factors which can be used to define risk categories. Known factors, which could be used to define pre-strata, include age, gender, presence of certain diseases, nationality, immigration status, and geographical location. Unknown factors, which could be determined via interview, include sexual preference and drug use. Data on the prevalence of HTLV-III within various subgroups can be both 1) incorporated into the overall estimate of prevalence and 2) used to determine sampling allocations. Such data is available, for example, for blood donors (Kuritsky *et al.* 1986), military results (Redfield and Burke 1987), intravenous drug

abusers in Queens, New York (Robert-Guroff *et al.* 1986) and male homosexuals in Greenwich Village (Casareale *et al.* 1984/5). Though this prior information can be used to reduce cost and increase accuracy, confidentiality and sensitivity/specificity of the HTLV-III test remain as significant obstacles which must be addressed carefully before such a study will provide meaningful results.

ACKNOWLEDGEMENT

The author would like to express his appreciation to the reviewer for helpful comments in the area of non-proportional sampling.

APPENDIX

Derivation of Expectation and Variance With No Prior Information and Proportional Sampling

Using the notation given in Section 2, we proceed first with the derivation of $E(\hat{\tau})$. The conditional expectation given s' is $E(\hat{\tau} \mid s') = 1/f \sum_j n'_{\cdot j} \bar{y}'_{\cdot j}$. Then, writing $n'_{\cdot j} \bar{y}'_{\cdot j}$ as $\sum_k n'_{kj} \bar{y}'_{kj}$, we find $E(\hat{\tau}) = E(E(\hat{\tau} \mid s')) = 1/f \sum_j \sum_k E\left(n'_{kj} E(\bar{y}'_{kj} \mid n'_{kj})\right) = 1/f \sum_j \sum_k E(n'_{kj}) \bar{Y}_{kj} = \tau$ since n'_{kj} is hypergeometric with sampling fraction f and N_{kj} units in pre-stratum k and post-stratum j. Thus, $\hat{\tau}$ is, in this case, unbiased (ignoring approximation A1).

Computation of the variance is along the same lines, yet much more technically detailed. Only certain elements of the computation will be presented and particular emphasis will be placed on the points in the derivation where approximations are made. First, some computation using the two phases of conditioning discussed above, yields

$$\operatorname{var}(E(\hat{\tau} \mid s')) = \frac{1-f}{f} \sum_k N_{k\cdot} S^2_{k\cdot}. \tag{13}$$

We next obtain

$$\operatorname{var}(\hat{\tau} \mid s') = \frac{1-c}{f^2 c} \sum_j \frac{n'_{\cdot j}}{n'_{\cdot j} - 1} \cdot \left[\sum_k (n'_{kj} - 1) s'^2_{kj} + \sum_k n'_{kj} (\bar{y}'_{kj} - \bar{y}'_{\cdot j})^2 \right]. \tag{14}$$

Our second and third approximations are to approximate $n'_{\cdot j}/(n'_{\cdot j} - 1)$ by one (A2) and $(n'_{kj} - 1)$ by n'_{kj} (A3) in equation (14). We now require the expectation of the first term in (14) and find

$$E\left[\frac{1-c}{f^2 c} \sum_j \sum_k n'_{kj} s'^2_{kj}\right] \approx \frac{1-c}{fc} \sum_j \sum_k N_{kj} S^2_{kj}. \tag{15}$$

In (15), one further approximation (A4) is necessary; we ignore the possibility of $n'_{kj} \leq 1$ for any k,j. We also require the expectation of the second term in (14). The exact formula turns out to be

$$\frac{1-c}{fc}\sum_j\left\{\sum_k N_{kj}(\bar{Y}_{kj}-\bar{Y}_{.j})^2+a_1\sum_k S_{kj}^2-a_2\right\} \tag{16}$$

where $a_1 = 1 - f - E[n'_{kj}(1 - n'_{kj}/N_{kj})/n'_{.j}]$ and $a_2 = E[(\sum_k n'_{kj}(\bar{Y}_{kj} - \bar{Y}_{.j}))^2/n'_{.j}]$. We note first that $|a_1| \leq 1$ and thus when combined with N_{kj} in (15), it can be ignored (approximation A5). Also, if in a_2 $n'_{.j}$ is approximated (A6) by its expectation, $fN_{.j}$, since $E[\sum_k n'_{kj}(\bar{Y}_{kj} - \bar{Y}_{.j})] = 0$, we have

$$a_2 \approx \frac{1}{fN_{.j}}\operatorname{var}\left(\sum_k n'_{kj}(\bar{Y}_{kj}-\bar{Y}_{.j})\right) \approx (1-f)\sum_k\frac{N_{kj}}{N_{.j}}(1-W_{kj})(\bar{Y}_{kj}-\bar{Y}_{.j})^2$$

where we have finally approximated $(N_{k.} - 1)$ by $N_{k.}$ (A7) in computing the variance of the hypergeometric variable n'_{kj}. When compared to the similar term with coefficient N_{kj} in (16), we discover that a_2 itself is approximately negligible. Finally, once again ignoring differences between N_{kj} and $(N_{kj} - 1)$ or between $N_{.j}$ and $(N_{.j} - 1)$ (approximation A8), (15) and (16) can be combined to yield

$$E\big(\operatorname{var}(\hat{\tau} \mid s')\big) \approx \frac{1-c}{fc}\sum_j\frac{N_{.j}}{N_{.j}-1}\sum_k\left[(N_{kj}-1)S_{kj}^2+N_{kj}(\bar{Y}_{kj}-\bar{Y}_{.j})^2\right]$$

$$= \frac{1-c}{fc}\sum_j N_{.j}S_{.j}^2. \tag{17}$$

Combining (13) and (17), we finally obtain

$$\operatorname{var}(\hat{\tau}) \approx \frac{1-f}{f}\sum_k N_{k.}S_{k.}^2 + \frac{1-c}{fc}\sum_j N_{.j}S_{.j}^2. \tag{18}$$

The validity of this approximation rests on three facts. First, when (18) is evaluated in the five examples for which simulated data exists, the results compare very favorably. The approximated standard error given by (12) is 113.25, 108.97, 108.09, 106.77, and 106.32 for $f' = .10$, .20, .30, .40, and .50, respectively. These values are nearly equal to those in Table 3 and the column giving $SE(\hat{\tau})$ and $M = 0$ with m equal to 500 or 2500. Second, the error introduced by each approximation made was analyzed and found, with the possible exception of approximation A6, to be negligible in the case of relatively large population and sample sizes. Even in the case of A6, the law of large numbers indicates that $n'_{.j}$ will be well approximated by its expectation if the sample sizes are reasonably large. Finally, as described in the following, this approximation formula reduces to the exact formula in all three standard situations. First, this situation reduces to the usual stratified sampling according to pre-strata when we take $J = K$, $P_{.j} = P_{k.}$ for $j = k$, and $c = 1$. Here, formula (18) reduces to $\operatorname{var}(\hat{\tau}) \approx (1 - f)/f \sum_k N_{k.} S_{k.}^2$ which is well known to be the exact formula. Also, the estimation scheme described reduces to the usual two phase sampling for stratification when we take $K = 1$ and (18) again reduces to the exact formula (see Cochran 1977, p. 329). Similarly, we obtain the situation of regular stratified sampling by post-strata if we take $f = 1$ (here, K and the pre-stratification become irrelevant), and formula (18) again reduces to the exact value.

REFERENCES

ALLEN, J.R. (1984). Epidemiology of the Acquired Immunodeficiency Syndrome (AIDS) in the United States. *Seminars in Oncology*, 11, 4-11.

CASAREALE, D. *et al.* (1984/5). Prevalence of AIDS-associated retrovirus and antibodies among male homosexuals at risk for AIDS in Greenwich Village. *AIDS Research*, 1, 407-421.

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

HAN, C. (1973). Double sampling with partial information on auxiliary variables. *Journal of the American Statistical Association*, 68, 914-918.

HANSEN, M.H., and HURWITZ, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

KURITSKY, J.N. *et al.* (1986). Results of nationwide screening of blood and plasma for antibodies to HTLV-III. *Transfusion*, 26, 205-207.

LUI, K. *et al.* (1986). A model based approach for estimating the mean incubation period of transfusion-associated acquired immuno-deficiency symdrome. *Proceedings of the National Academy of Sciences*, 83, 3051-3055.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

REDFIELD, R.R., and BURKE, D.S. (1987). Shadow on the land: the epidemiology of HIV infection. *Viral Immunology*, 1, 69-81.

ROBERT-GUROFF, M. (1986). Prevalence of antibodies to HTLV-I, -II, and -III in intravenous drug abusers from an AIDS endemic region. *Journal of the American Medical Association*, 255, 3133-3137.

VARDEMAN, S., and MEEDEN, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.

WHITE, D. (1987). Mean squared error of estimators using two stage sampling for stratification and prior information. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 117–129
Statistics Canada

Sample Design of the 1988 National Farm Survey

C. JULIEN and F. MARANDA[1]

ABSTRACT

The National Farm Survey is a sample survey which produces annual estimates on a variety of subjects related to agriculture in Canada. The 1988 survey was conducted using a new sample design. This design involved multiple sampling frames and multivariate sampling techniques different from those of the previous design. This article first describes the strategy and methods used to develop the new sample design, then gives details on factors affecting the precision of the estimates. Finally, the performance of the new design is assessed using the 1988 survey results.

KEY WORDS: Multi-purpose sampling; Multiple frame; Area frame; Multivariate stratification.

1. INTRODUCTION

The National Farm Survey (NFS) is a probability-based sample survey focussing on several subjects related to agriculture in Canada. It is conducted annually in June and July in all provinces except Newfoundland, where a separate survey is carried out.

The previous NFS sample design, dating from 1983, was based on the results of the 1981 Census of Agriculture. A description of it may be found in Ingram and Davidson (1983). However, since 1981 the farm population has changed significantly, reducing the effectiveness of this design. Furthermore, the requirements of the survey have changed somewhat over the years, resulting in the need to update the samples.

A new sample design was therefore developed based on the results of the 1986 Census of Agriculture, and became operational in the summer of 1988.

2. OBJECTIVES OF THE SURVEY

The primary objective of the survey is to provide timely, reliable estimates of levels and annual trends for over 100 agriculture variables. Essentially, these variables may be divided into three categories: cropland areas for the current year; livestock numbers on July 1; and receipts and operating expenses for the previous calendar year. In terms of reliability, the objective of the survey is to obtain coefficients of variation (CV) below 5% at the provincial level for the major parameters.

Survey data are normally summarized to the provincial level. However, primarily for analysis purposes, results for sub-provincial regions are also produced using domain estimation methods.

Another important objective of the survey is to obtain a master sample from which subsamples are chosen for use in other farm surveys conducted by Statistics Canada.

[1] C. Julien is a methodologist with the Census Data Quality and Analysis Section, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6; F. Maranda is chief of the Agriculture Survey Methods Section, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

3. TARGET POPULATION AND SURVEY POPULATION

The target population includes all farms in the provinces surveyed which received $250 or more from the sale of agricultural products during the 12 months preceding the survey. Also included are farms which do not meet the $250 criterion at the time of the survey, but which expect to earn at least this sum during the 12 months following the survey. Such farms, which either began operating just prior to the survey or are temporarily inactive, are relatively few in number.

The survey population, or the group from which the sample is selected, excludes farms operated by institutions as well as those located on Indian reserves or settlements. The terms institution, Indian reserve and Indian settlement are defined in Statistics Canada (1987, pp. 115-117, 145, 152). The cost-benefit ratio associated with collecting data on these types of farms is very high. Because of this, they are excluded in order to enable more efficient use of the resources available for the survey. The contribution of such exclusions to national agricultural production is small and is estimated using adjustment factors which are based on Census data.

4. SAMPLING FRAMES AND THEIR USE

In theory, the survey population is divided into two groups, the first of which includes the farms enumerated in the Census and the second all other farms. These include the undercoverage from the Census and so-called new farms, that is, those which began operating after the Census.

The first group is covered all or in part, depending on the province, by one or two list frames created from the list of census farms. To complement the list frames and ensure complete coverage of the survey population, an area frame, created from the agricultural enumeration areas (EAs), is used. An enumeration area is the geographical region enumerated by a census representative. Furthermore, an EA is said to be agricultural if it contains at least one census farm. An area frame is needed to compensate for the shortcomings of the list frames, particularly their difficulty to identify new farms.

The estimation requirements of the survey and the characteristics of agriculture in Canada vary by region. To better account for these variations, the territory covered by the survey is divided into three regions and a different sample design is used in each one. The three regions involved are: the Prairie provinces and the Peace River district in British Columbia; Quebec and Ontario; and, finally, the Maritime provinces and the rest of British Columbia. The first of these regions is called the Canadian Wheat Board (CWB) region, since the entire region comes under the jurisdiction of this organization.

The total sample size in each of the three regions is essentially based on the overall budget available for data collection. Within each region, sample allocation among the various provinces and, where applicable, among the various frames, depends on several factors. The primary ones are the square root rule applied to the size of the survey population, historical allocations in the survey, and the results of various analyses centred on the expected precision of the estimates.

4.1 The Canadian Wheat Board Region

In this part of Canada, two list frames and one area frame are used in each province.

The first list frame (L1 list) essentially includes the large and medium-sized census farms in relation to key crop, livestock and expense variables. This list is obtained using an iterative process which consists in establishing a threshold for each key variable and including in the

list all farms that exceed at least one of these thresholds. Each threshold is adjusted separately upward or downward so that the L1 list, once completed, includes approximately 35% of the survey population's farms and accounts for 50% to 90% of the total agricultural activity, depending on the key variable in question. These percentages are used because experience has shown that the resulting list is composed of farms which, individually, are more stable over time than the rest of the farms in the survey population. This stability leads to the creation of strata which remain homogeneous over the years, which is a factor in maintaining the efficiency of the sample design.

In each province, the L1 list is then stratified within sub-provincial regions based on nine key variables. A sample of farms is selected and used to obtain data on crops and livestock. Because data on expenses are more difficult and costly to collect, only a sub-sample, called the core sample, is used to obtain this information.

The second list frame (L2 list) includes all census farms with more than 20 acres of cropland which were not included in the L1 list. The L2 list is stratified within crop districts based on a single key variable, namely, cropland area at the time of the Census. The L2 list is used to complement the L1 list for preliminary crop data. These data must be collected within very tight deadlines which, for operational reasons, cannot be met using the area frame.

The area frame includes all agricultural enumeration areas, except those on Indian reserves and in the so-called marginal agricultural regions, that is regions with little agricultural activity. Marginal regions are found mostly in the northern parts of the provinces and in urban fringes. The few census farms located in marginal regions are added to the L1 list, since it is the only list used to collect data on all survey variables.

The area frame is stratified using the same sub-provincial regions and key variables as the L1 list. It ultimately produces a sample of segments which are delineated on topographic maps. The identity of the farmers operating land in one of these segments is obtained through on-site enumeration. Manual matching of names and addresses then enables detection of segment farms overlapping one of the list frames. This detection is essential because each time the area frame is used to complement a list frame, only those segment farms that do not overlap the list in question are used, thus ensuring that the list and area frames represent mutually exclusive domains.

Complete information is required on all segment farms except those overlapping the L1 list, as the data for this list are obtained from the sample selected from it.

4.2 Quebec and Ontario

In each of these provinces, a single list frame, called L1, and an area frame are used.

The list frame is composed of all census farms in the survey population. The methodology used in sampling from this list is similar to that used for the CWB region L1 list, apart from two differences. First, incorporated farms, or farms founded as business corporations, are separated from the other farms, and strata are created independently within the two groups. This preliminary separation is performed because only incorporated farms are required to report their expenses in the survey, since the expenses of the non-incorporated farms are obtained from Revenue Canada tax records. It should be noted that the confidentiality of these records is completely protected under the Statistics Act. Second, sub-sampling for expenses is unnecessary because less than 25% of the farms in the survey population are incorporated.

The area frame and its sample design have not been modified following the last Census, due to a lack of resources. Only the marginal regions were updated, resulting in their enlargement.

4.3 Maritime Provinces and the Rest of British Columbia

In each province of this region, the sample design includes only one list frame, again called L1, which is made up of all census farms in the survey population. Given that a list frame tends to deteriorate with time and that there is no area frame to supplement it, it becomes more difficult to completely cover the survey population. However, because of the relatively small number of farms, under 30 000 in these provinces, more complex procedures were implemented to keep the list up-to-date. Notably, farms which were missed in the Census or which began operating following it may be detected through these procedures. Thus, for all practical purposes, the list frame is considered to ensure full coverage of the survey population.

In each province of this region, the list is stratified and a sample of farms is selected using the same approach as in Quebec and Ontario. All the estimates required are produced from this sample.

5. LIST SAMPLING TECHNIQUES

Samples are taken from the list frames using a one stage, statified sample design where the farms constitute the sampling units. The strategy and methods used to develop this design are essentially the same, regardless of the province and list involved. However, the combination of methods and key variables used may vary from case to case.

The first step consists in identifying the farms with distinct characteristics and in automatically including them in the sample. There are essentially two kinds of these so-called self-representative or take-all farms. The first group includes those with a unique operating structure such as community pastures and multiholding corporations, while the second group contains the farms which clearly stand out from the majority because of their very large contributions to key crop, livestock and expense variables. Due to the skewness (to the right) of the distributions involved, complete enumeration of these farms is an efficient way to reduce sampling variance.

Farms with very large contributions are identified through an intuitively-based rule which produced good results in the previous sample design. This rule, called the sigma-gap rule, is applied separately to each key variable using all farms having a non-zero value for the variable in question. Farms with a sufficiently high contribution to one of the key variables, as determined by this rule, are said to be take-all.

The sigma-gap rule, as adapted to the survey, functions as follows. Given a univariate distribution of points x_i, $i = 1, 2, \ldots, N$, $x_i > 0$ for all i, and given σ as its standard deviation, the points are arranged in increasing order $x_1 \leq x_2 \leq \ldots \leq x_N$; for the half of the distribution to the right of the median, the distance between each successive pair of points $d_i = x_i - x_{i-1}$ is determined; given i_o, the smallest i for which $d_i \geq \sigma$, all points $i \geq i_o$ correspond to take-all farms. If $d_i < \sigma$ for all i, no point in this distribution distinguishes itself sufficiently from the others to be declared a take-all farm.

The second step consists in dividing the rest of the farms in the list into take-some strata. In most cases, the strata are formed within sub-provincial regions according to nine key variables representing the usual three categories: crops, livestock and operating expenses. The number of variables in each category is one, six and two respectively.

The underlying principle to the stratification is as follows. Each farm is characterized by nine variables, and neighbouring farms, defined in terms of Euclidian distance, are grouped together. Two multivariate clustering algorithms are used for this purpose. These algorithms are called FASTCLUS and CLUSTER, since they are available in the procedures of the same name in the SAS statistical analysis software package (version 5).

The FASTCLUS algorithm divides a set of observations into a predetermined number of mutually exclusive clusters. First, the algorithm chooses observations which serve as initial cluster seeds. Each observation is then assigned to the nearest seed, and once this is completed, the cluster seeds are updated by the means of the clusters thus formed. The process is repeated until the changes in the seeds become minimal. The FASTCLUS algorithm is based on work by Hartigan (1975) and MacQueen (1967).

The CLUSTER algorithm groups a set of observations into mutually exclusive clusters in a hierarchical structure. Initially, each observation forms a cluster in itself. Based on a technique inspired by Ward (1963), the two most similar clusters are combined into one, which subsequently replaces them. The process is repeated until only one cluster remains. Massart and Kaufman (1983) provide an introduction to this type of classification. Thus, the set of observations is broken down into as many partitions as there were observations to begin with, and each partition corresponds to a stratification.

These algorithms are used successively as follows. FASTCLUS is used first to group the farms into 250 clusters, which are then progressively combined to form the strata using CLUSTER. Initial classification is performed with FASTCLUS, since using CLUSTER directly with a high number of records would require excessive computer time.

Each of the three categories of variables must contribute equally to strata formation. To ensure this, the initial stratification variables are transformed so that the sum of the transformed variables in each category has a mean 0 and a predetermined variance, usually 1. The crop category with its single variable may be standardized in the usual manner by subtracting its mean and dividing by its standard deviation. In each of the other two categories, two successive transformations are performed independently. Given X_i, the initial variables of a given category C, a principal components analysis was performed to obtain transformed variables Y_i. These new variables, with mean μ_i and variance σ_i^2, are linear combinations of the former ones and mutually independent. The Y_i are then standardized to obtain final stratification variables Z_i as follows:

$$Z_i = \frac{Y_i - \mu_i}{\left(\sum_{i \in C} \sigma_i^2\right)^{1/2}}.$$

Thus, the mean and variance for $\sum_{i \in C} Z_i$ are 0 and 1 respectively.

An empirical approach is used to determine the number of strata. Several stratifications and allocations are performed by varying the number of strata. Then, the coefficient of variation curve is drawn as a function of the number of strata for all key variables and many others. These curves generally resemble Figure 1. Stratification gains are considered to have been virtually fully attained at the point where the majority of curves are practically horizontal. The number of strata chosen is a compromise between this point and the desire to avoid forming too many strata so as to attenuate the effects of incorrect initial classification and stratum jumpers over time, two major causes of outliers or influential observations.

Sample allocation is multivariate and is generally carried out using the same key variables used for stratification. The allocation algorithm consists in minimizing a linear combination of the square of the coefficients of variation of the key variables, within the constraint of a fixed total sample size. Given c_i, coefficient of variation for a key variable, $a_i > 0$ as constant and n_o total sample size, $\sum a_i c_i^2 = f(n)$ must be minimized within the constraint $n = n_o$. The algorithm used is described in Bethel (1986). Adjustments are then made to obtain a minimum sample size of 4 and a maximum weighting factor of 50 in each stratum.

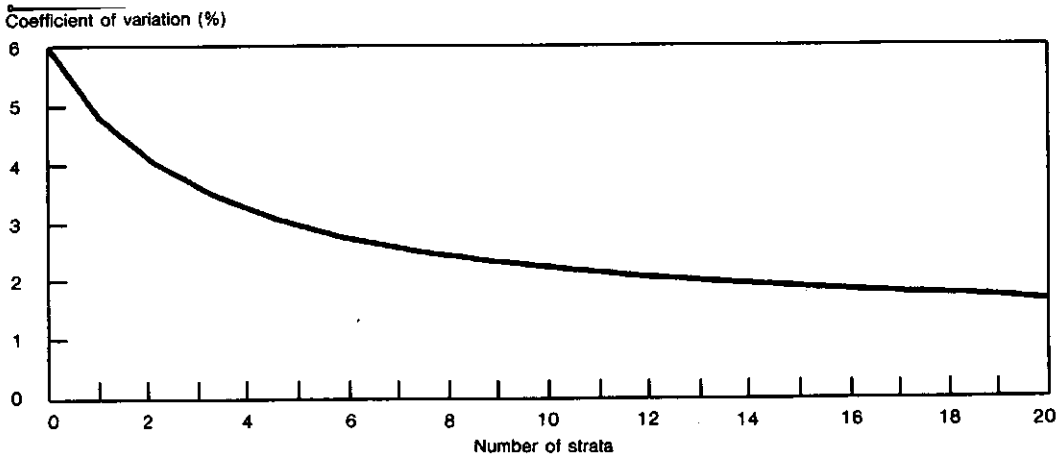

Figure 1. General Curve of the Coefficient of Variation as a Function of the Number of Strata

Finally, once allocation has been completed, the farms are sorted within each stratum by sub-provincial region and total operating expenses and a sample is selected using circular systematic sampling. For the L1 list in the CWB region, the complete sample is chosen first; the core sub-sample is then selected from it using circular systematic sampling.

6. AREA SAMPLING TECHNIQUES

Area samples are selected according to a two-stage stratified sample design. The Census enumeration areas and segments represent the primary and secondary sampling units respectively.

Given that the area sample design has not been modified for Quebec and Ontario, the following paragraphs apply only to the CWB region.

The first step consists in measuring the agricultural activity in each of the frame's EAs by summarizing to the EA level the data for the census farms not included on the L1 list. Excluding the L1 list farms from the summarization process produces EA distributions which accurately reflect the characteristics of small farms. Subsequent use of these distributions enables an area sample complementing the L1 list with respect to small farms to be selected with greater efficiency.

Once the summarization process has been completed, each EA is treated as a farm for sampling purposes. The EA selection strategy and methods are very similar to those applied

to the CWB region L1 list. First, take-all EAs are determined using the sigma-gap rule. The remaining EAs are then allocated to take-some strata within sub-provincial regions using the CLUSTER multivariate clustering algorithm. Preliminary classification with FASTCLUS is unnecessary in this case due to the relatively low number of EAs, never more than 3000 per province, to be processed. Furthermore, the usual standardizations suffice for transforming the key variables. A principal components analysis was not used because the area frame's contribution to provincial estimates does not justify such an approach.

Allocation to strata is performed with the same algorithm used for the list, and the minimum sample size is again established at 4. The sample size is then divided by four in each stratum, and four separate replicates are selected using circular systematic sampling. Replicates facilitate variance calculation, as a single secondary unit is often chosen per primary unit.

Once the EAs have been selected, their boundaries are traced on topographic maps and they are divided into segments of approximately 7.5 km^2 (3 mi^2). Natural boundaries such as roads and rivers are used as much as possible to facilitate the work of field interviewers. Simple random sampling without replacement of the segments is performed at a minimum rate of 1 out of 30 in each selected EA. There are, however, some exceptions to the rule: additional segments are taken so that the overall weighting factor does not exceed 180; a minimum of two segments are selected in each EA belonging to the strata subjected to first-stage complete enumeration; and, finally, when the same EA appears in more than one replicate, measures are taken to avoid selecting the same segment more than once. Nevertheless, these exceptions are rare.

7. RESULTS OF THE SAMPLE DESIGN

Table 1 contains the results of the list frame sample design. The following items are included: the number of farms in the list (N); the number of strata (H); the number of farms in the sample (n); and, finally, the number of farms in the core sub-sample (n-core) in those provinces where it applies.

Table 1

Results of the List Frame Sample Design

Province	L1 List				L2 List		
	N	H	n	n-core	N	H	n
P.E.I.	2,830	26	451				
N.S.	4,273	35	550				
N.B.	3,544	39	498				
Quebec	41,380	80	6,096				
Ontario	72,598	78	8,401				
Manitoba	6,712	48	1,364	490	18,058	29	2,267
Saskatchewan	15,668	48	3,625	1,106	45,798	41	4,573
Alberta	13,928	63	2,981	909	38,504	25	2,973
B.C. (Peace)[a]	494	25	190	190	1,187	6	170
B.C. (rest)[b]	17,042	41	1,999				
Total	178,469	479	26,155	2,695	103,547	101	9,983

[a] Peace River district in British Columbia.

[b] British Columbia minus the Peace River district.

Table 2 contains the results of the area sample design in those provinces where such a design is used. The following items are indicated: the number of EAs in the frame (N); the number of strata (H); the total number of EAs sampled (n); the number of EAs sampled where each EA is counted only once when it appears in more than one replicate (n-once); and, finally, the number of segments chosen (m).

8. FACTORS AFFECTING THE PRECISION OF THE ESTIMATES

To better appreciate the results obtained from the 1988 survey, three factors affecting the reliability of the estimates must be discussed. These factors are the sample size, the treatment of the total non-response and the estimation methodology.

First, the sample size for the L1 list in the CWB region was reduced by 10% in relation to that of the corresponding list used in the previous sample design. This reduction was prompted mainly by the desire to lower costs.

Second, the methodology used to treat total non-response was modified in 1988. Previously, when a farm failed to respond to the survey, its data were imputed using the data from another farm in the same stratum. These imputed data enabled the sample to be completed to its original size. However, in 1988, the cases of total non-response were not imputed; instead only the respondent sample was used and the weighting factors adjusted upward. The actual sample is therefore reduced in relation to the former method.

In the 1988 survey, the total non-response rate varied between 2% and 13%, depending on the province. The national rate was 10%. Non-response rates are presented in detail in Table 3.

Table 2
Results of the Area Sample Design

Province	N	H	n	n-once	m
Quebec	2,065	43	191	182	230
Ontario	2,687	49	195	185	259
Manitoba	794	21	277	264	305
Saskatchewan	1,496	26	328	308	477
Alberta	1,623	32	328	319	434
B.C. (Peace)[a]	54	7	36	32	58
Total	8,719	178	1,355	1,290	1,763

[a] Peace River district in British Columbia.

Table 3
Total Non-response Rate (%) by Province

Province	Refusals	No Contact	Total
P.E.I.	0.00	3.55	3.55
N.S.	0.00	2.18	2.18
N.B.	0.00	1.61	1.61
Quebec	1.71	6.56	8.27
Ontario	2.27	11.11	13.38
Manitoba	3.45	4.03	7.48
Saskatchewan	4.06	6.46	10.52
Alberta	2.68	7.95	10.63
B.C.	1.78	10.28	12.06
Total	2.32	8.11	10.43

The last factor to be discussed is the estimation methodology. The usual estimators corresponding to a stratified simple random sample are used for list frames. For area frames, an estimator described in Wolter (1986 pp. 19-26) and corresponding to a sample design with independent replicates is used. Provincial estimates are obtained by adding the contribution of the list and area frames since, as previously mentioned, these two frames are independent and represent mutually exclusive domains. Details on the estimation methodology are found in Lynch (1988).

9. ASSESSING THE PERFORMANCE OF THE NEW DESIGN

To assess the performance of the new design, the precision of the estimates obtained in 1988 is compared first to that of the 1987 survey, then to the precision anticipated during the development of the sample design.

9.1 The 1988 and 1987 Surveys Compared

Two opposite tendencies are in effect in a comparison of the precision of the estimate in the 1988 and 1987 surveys. The 1988 estimates should be more precise because the 1987 sample design was already four years old. However, the two sample size reduction factors described in section 8 would indicate less precise estimates for 1988.

Precision is compared using the coefficient of variation of the provincial estimates obtained by combining the L1 list and area frames. The estimates used are those for several key variables whose coefficient of variation in 1987 did not exceed 20%.

The precision of 234 estimates is compared in the charts in Figure 2, where each square represents the CV achieved in 1987 on the x-axis and achieved in 1988 on the y-axis for a given estimate. The frequency (as a percentage) of the key variables located within each zone delineated by the straight lines $Y = X/2$, $Y = X$ and $Y = 2X$ is also presented.

Nearly 60% of crop estimates were more precise in 1988 than in 1987. The majority of those that were less precise were so to a small degree only. Close to 95% of livestock estimates were more precise in 1988 than the previous year; in fact, 32% of the estimates were even twice as precise. Finally, over 60% of operating expense estimates were more precise in 1988. Some of the 1987 estimates were a good deal less precise, and 7% were even two times less precise. The latter are from Quebec and Ontario, where data on operating expenses are collected from incorporated farms only. Further more, the legal status of a farm in these provinces is difficult to identify, both in the Census and the survey.

Despite the reduction in the effective sample due to total non-response and cutbacks during the sample design development stage, the 1988 survey generally provided more precise estimates for each category of variables.

9.2 Precision Obtained Versus Precision Anticipated

The precision obtained is expected to be inferior to the precision anticipated for two reasons. First, when the weighting factors are adjusted to account for the total non-response, the variance increases slightly. Second, the data used to create the sampling frame were taken from the 1986 Census of Agriculture. These data are subject to error and the sampling frame deteriorates with changes in agricultural activity.

Precision is compared using the coefficient of variation of L1 list frame provincial estimates only. These estimates are for several key variables whose anticipated CV did not exceed 20%.

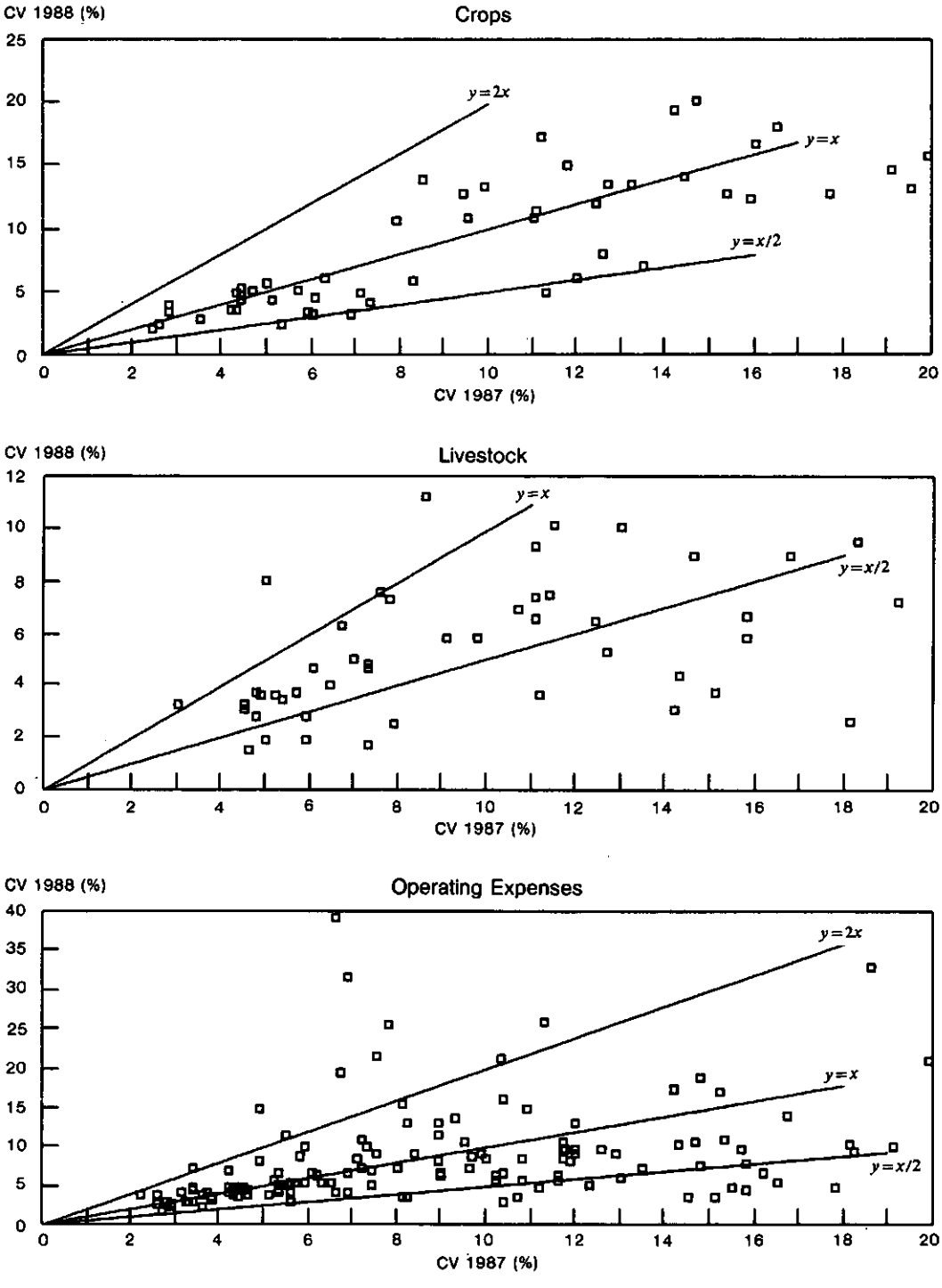

Figure 2. Comparaison of the Precision of Key Variable Estimates in the 1987 and 1988 Surveys by Category of Questions.

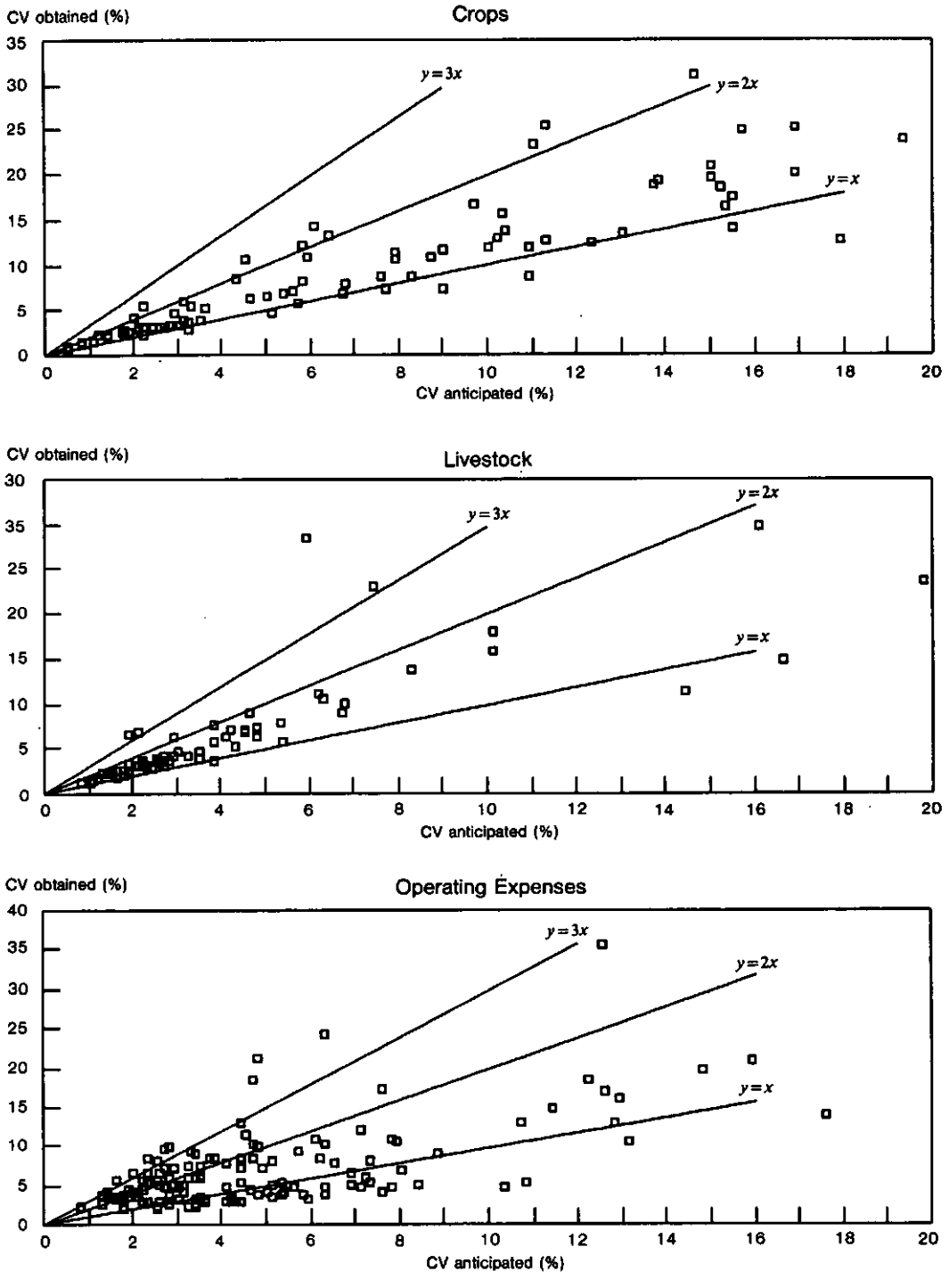

Figure 3. Comparison of the Precision of Key Variable Estimates Obtained in the 1988 Survey and the Precision Anticipated during Development of the Sample Design.

A comparison of the precision of 288 estimates is presented in chart form in Figure 3. In these charts, each square represents the anticipated CV on the x-axis and the obtained CV in 1988 on the y-axis for a given estimate. The frequency (as a percentage) of the key variables located within each zone delineated by the straight lines $Y = X$, $Y = 2X$ and $Y = 3X$ is shown in the charts.

For the crop and livestock categories, approximately 90% of the estimates are sufficiently precise, given the non-response rate, as most of the key variables are located closer to straight line $Y = X$ than to straight line $Y = 2X$. Two tendencies can be seen for the operating expense estimates. Surprisingly, the CV obtained is lower than the anticipated CV in 28% of the cases, the vast majority of which are found in the CWB region. However, 31% of all estimates are more than two times less precise than anticipated. These cases are found in Quebec and Ontario for the reasons given in section 9.1.

Finally, a complementary study was conducted in which the precision obtained was compared to the anticipated precision based on the size of the sample actually observed. This study revealed that the frequency of estimates at least two times less precise than anticipated dropped from 12% to 5% for crops, from 9% to 5% for livestock and from 31% to 7% for operating expenses.

These studies show that in general the precision obtained is acceptable and differs from the anticipated precision mainly because of the treatment for total non-response. This indicates that the sample design is therefore sound and the L1 list frame is adequate. On the other hand, less precise estimates were obtained for operating expenses due to a problem in identifying incorporated farms in Quebec and Ontario in the Census and in the survey. Finally, the list frame, which was two years old at the time of the survey, was observed to have deteriorated somewhat due mostly to bankruptcies and farm sales.

10. CONCLUSION

In general, survey results were substantially improved following implementation of the new sample design. Moreover, the reduction in sample sizes led to cost savings and a considerable reduction in the response burden on the farmers surveyed. Difficulties remain, however, especially regarding the operating expense variables for incorporated farms in Quebec and Ontario. Further studies to resolve these difficulties are being envisaged.

ACKNOWLEDGEMENTS

The authors would like to thank the editor of the journal and the referees, whose valuable comments helped to improve this article.

REFERENCES

BETHEL, J. (1986). An optimum allocation algorithm for multivariate surveys. Technical report of the United States Department of Agriculture, Statistical Reporting Service, Statistical Research Division, No. SF and SRB-89.

GERMAIN, M.-F., DOLSON, D., and MARANDA, F. (1989). Redesign of the 1988 National Farm Survey. Internal working document, Business Survey Methods Division, Agriculture Section, Statistics Canada.

HARTIGAN, J.A. (1975). *Clustering Algorithms*. New York: John Wiley and Sons.

INGRAM, S., and DAVIDSON, G. (1983). Methods used in designing the National Farm Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 220-225.

LYNCH, J. (1988). Cas spéciaux d'estimation dans l'enquête nationale des fermes de 1988. Internal working document, Business Survey Methods Division, Agriculture Section, Statistics Canada.

MacQUEEN, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability*, 1, 281-297.

MASSART, D.L., and KAUFMAN, L. (1983). *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. New York: John Wiley and Sons.

SAS Institute Inc. (1985). *SAS User's Guide: Statistics*, Version 5 Edition. Cary, NC: SAS institute.

STATISTICS CANADA (1987). 1986 Census Dictionary. Catalogue 99-101E, Statistics Canada.

WARD, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

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

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 131–136
Statistics Canada

Does the Method Matter on Sensitive Survey Topics?

DAVID A. HAY[1]

ABSTRACT

The effects of utilizing a self-administered questionnaire or a personal interview procedure on the responses of an adolescent sample on their alcohol consumption and related behaviors are examined. The results are generally supportive of previous studies on the relationship between the method of data collection and the distribution of responses with sensitive or non-normative content. Although of significance in a statistical sense, many of the differences are not of sufficient magnitude to be considered significant in a substantive sense.

KEY WORDS: Data collection; Personal interview; Self-administered questionnaire; Response errors; Alcohol consumption.

1. INTRODUCTION

To "questionnaire" or to interview that is the question to be answered by researchers in the design and conduct of sample surveys on delicate or sensitive topics. The decision on whether to utilize personal or telephone interviews or a variant of the self-administered questionnaires, or a combination there of, is a critical decision that survey researchers have to make in attempting to optimize the quality of the resultant data.

Encompassed by the more general problems of reliability and validity associated with self-reports of attitudes, behaviour and other phenomena of interest to survey practitioners, is the question regarding the relative merits of the interview and self-administered formats in minimizing or reducing non-sampling biases or errors. In other words, then would different results be obtained from the utilization of different modes of data collection (Smith 1975)?

As far back as 1959, Selltiz *et al.* (1959) stated that most questionnaires and interviews were utilized without evidence of their relative merits. More recently, this position has been re-emphasized by Knudsen *et al.* (1967), Alwin (1977) and Newton *et al.* (1982) who maintain that the selection of the survey mode to be utilized is based on convenience, relative costs and other practical considerations rather than on their methodological adequacy and potential response effects. The planning of survey research, Newton *et al.* emphasize should be determined by what is reliably known about the relationship between methods of administration and response patterns, rather than just on the issues of relative costs, respondent motivation and other similar considerations.

Some studies which have compared personal interviews with more anonymous formats such as self-administered questionnaires or telephone interviews have found minimal and/or statistically non-significant differences in the responses to a variety of topics including those of a private or sensitive nature (DeLameter and MacCorquodale 1975; Gibson and Hawkins 1968; Krohn *et al.* 1974; McDonagh and Rosenblum 1965; Metzner and Mann 1952, Newton *et al.* 1982 and Sykes and Collins 1987.) Other researchers have observed that more candid, self-revelatory and informative responses are more likely to be made by questionnaire and telephone respondents than personal interviewees on topics concerning deviant, sensitive

[1] David A. Hay, Associate Professor, University of Saskatchewan, Saskatoon, Saskatchewan, S7N 0W0.

or embarrassing behaviours and attitudes. (Cannell and Fowler 1963; Ellis 1947; Hubbard *et al.* 1976; Knudsen *et al.* 1967; Siemiatycki 1979; Whitehead and Smart 1972 and Wiseman 1972).

The conclusions of the latter studies were generally based on the untested assumption that the increased reporting of deviant, threatening or embarrassing information was more accurate (Blair *et al.* 1977). This point was also emphasized by Schuman (1980) who stated that frequently no external validation data were obtained, but the researchers "assumed that the more such behaviour was reported, the more accurate the reports – a plausible but not air-tight assumption for most of the topics they dealt with."

The present note is concerned with a further comparison of the relationship between personal interviews and self-administered questionnaires and responses obtained from an adolescent population on a "threatening" or deviant topic, namely alcohol consumption. The results being reported are based on a secondary analysis of data from a study of alcohol-related attitudes and behaviors from a sample of teenagers in a Western Canadian province completed in 1977-78 (Hetherington *et al.* 1978 and 1979).

The study which utilized both personal interviews and self-administered questionnaires provides a unique opportunity to compare the potential effects of the mode of data collection on the resultant data. This type of comparison of interest to survey practitioners is generally not possible in the majority of surveys which tend to rely on one method of data collection.

A stratified random sample of 1502 students in grades 6 to 12 was selected from three school regions in the Province of concern. The total sample of students was randomly assigned by grade to either the self-administered questionnaire or to the personal interview procedure. Approximately one half of the students from each grade 6 to 12 were thus allocated to one of the procedures. The number of students assigned to be interviewed was 752 with 750 students being assigned to the questionnaire data collection.

The questionnaire was group administered by a trained researcher in a room made available at each school for that purpose. The interviews were conducted by fifteen interviewers specifically trained for the study.

The survey instrument which consisted of 75 questions was identical in content for both the interview and questionnaire data collection procedures. The majority of the questions were closed ended and required an average of 20 minutes for completion in both types of administration.

2. RESULTS AND DISCUSSION

A comparison of the personal interview and self-administered questionnaire respondents on a number of personal and familial characteristics was conducted to determine if the two groups differed in respects other than the method of data collection. The results indicated that the two groups did not differ by more than could be attributed to chance on variables such as sex, age, grade of enrollment, parent's educational and occupational backgrounds and religious affiliation. A statistically significant difference was observed on the variable of ethnicity with a higher percentage of Canadian identities reported by the interview respondents.

With the exception of ethnic background, the subsequent analysis was, therefore, based on the assumption that the interview and questionnaire respondents were equivalent on a number of variables that could potentially confound the comparison of obtained responses to the two procedures.

Table 1

Frequency Distribution and Z Probabilities on Selected Questions for Interview and Questionnaire Results

Variable	Interview (n = 752)	Questionnaire (n = 750)	Two-tailed Z Probability
Ever drink	62.63	73.73	.000
Ever used cigarettes	29.78	37.60	.001

2.1 Variable Distribution

A comparison of the mean responses or frequency distributions for the interview and questionnaire respondents on a number of questions with non-normative or illegal content lent general support to previous research on similar issues. The questions of primary concern are those related to the consumption of alcohol which are viewed as possessing a considerable degree of threat or deviant content for the population under consideration, the majority (99.8%) of whom were under the legal drinking age at the time of the study.

The frequency distributions in Table 1 indicated that a significantly higher percentage of the questionnaire respondents reported ever having more than a sip or taste of an alcoholic beverage. Similar statistically significant differentials were observed between the interview and questionnaire respondents on reported smoking.

For those respondents reporting that they had consumed a drink of alcohol, the mean drinking levels and average age at first drink shown in Table 2 were also suggestive that the questionnaire respondents are more likely to report on deviant behaviour than were their interview contemporaries. The significantly higher average drinking levels for the questionnaire respondents reflects their reporting higher amounts and frequencies of alcohol consumption. The significantly higher average age at first drink for the interviewees indicates their reporting taking their first substantial drink at an older age than did the questionnaire respondents.

Significant differentials between the interview and questionnaire respondents were also observed on the reporting of parental drinking and on the importance of religion in the home questions. The mean values for these three questions indicated that the questionnaire respondents reported higher drinking levels for their parents than did the interviewees and that religion was perceived as being less important in the homes of the questionnaire respondents. While not possessing the same degree of self revelation or threat to the respondent per se, the differentials were viewed as suggestive of an attempt on the part of the interviewees to portray a more favourable or socially acceptable image about their family life.

However, the greater importance of religion in the home reported by the interviewees was not carried through in their self-descriptions of the importance of religion. The statistical equivalence of the means values on the importance of religion to self indicated that the interview respondents were no more likely to report that religion was important to self than were the questionnaire respondents. The two groups of respondents were also equally likely to report on the drinking habits of friends or peers.

The response patterns on other questions possessing somewhat different aspects of ego-involvement or image favourability did not generally support the potential operation of a social desirability effect as was evident for the alcohol related behaviours. As indicated in Table 2, the questionnaire respondents reported receiving significantly higher school grades, had higher educational aspirations in terms of their future educational plans and reported more positive self images on 4 of the 7 self-esteem items and on the composite self-esteem index. Contrary

Table 2

Means, Standard Deviations and "*t*" probabilities on Selected Questions for Interview and Questionnaire Respondents

Variable[1]	Interview (n = 752)		Questionnaire (n = 750)		Two-tailed "*t*" Probabilities
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	
	Alcohol and Related Behaviour				
Drinking level	2.31	2.92	2.76	3.05	.003
Age at first drink[a]	3.93	1.32	3.64	1.39	.001
Father drinks	1.82	0.62	1.90	0.58	.011
Mother drinks	1.70	0.50	1.75	0.51	.025
Friends drink	1.92	0.57	1.94	0.56	.481
	Educational Variables				
Grades received	4.37	1.49	4.58	1.46	.008
Educational plans	3.02	1.24	3.25	1.24	.001
	Religious Variables				
Importance of religion in the home	3.37	1.16	3.15	1.22	.000
Importance of religion to student	3.22	1.12	3.13	1.18	.130
	Self-Esteem Indices				
Item 1	2.98	0.60	3.12	0.60	.000
Item 2	2.96	0.49	3.08	0.54	.000
Item 3	3.14	0.55	3.27	0.61	.000
Item 4	2.98	0.51	2.05	0.57	.033
Item 5	3.10	0.63	3.01	0.75	.017
Item 6	2.93	0.56	2.97	0.59	.207
Item 7	3.07	0.54	3.12	0.60	.132
Composite	21.17	2.39	21.65	2.85	.001

[a] - Mean value calculated on grouped data.

[1] Variable Codes: Drinking level; composite index of frequency and volume of alcohol consumed 0 = abstainer to 9 = frequent consumer of large amount of alcohol.
Age at first drink: 1 = 6 years or less; 2 = 7-8 years; 3 = 9-10 years; 4 = 11-12 years; 5 = 13-14 years; 6 = 15-16 years; and 7 = ≥ 17 years.
Father, mother and friends drink: 1 = never drinks; 2 = drinks sometimes; 3 = drinks a lot.
Grades received: 1 = mostly D's and F's; 2 = Mostly C's and D's; 3 = mostly C's; 4 = mostly B's and C's; 5 = mostly B's; 6 = mostly A's and B's and 7 = mostly A's.
Educational plans: 1 = will not finish grade 12; 2 = will finish grade 12 only; 3 = will take technical training; 4 = will attend university and 5 = will go to graduate or professional school.
Self-esteem items and index: 1 = strongly disagree; 2 = disagree; 3 = agree and 4 = strongly agree. The additive index for the 7 items ranged from 7 to 28.

to the expectation that the interviewees would attempt to portray a more favourable image, these results tended to indicate that they were more modest in the reporting of school grades received, in their educational aspirations and in their self perceptions. However, the greater anonymity and potential freedom afforded the questionnaire respondents to more willingly report on their alcohol related behaviors may also have resulted in a similar perceived freedom to aggrandize their own merits in relation to these questions on school grades, educational plans and their self conceptions.

However, the presence of a significant distributional response bias between the interview-questionnaire data collections is evident only in the statistical sense of the term. The statistically significant mean value differences on the questions of concern ranged from 0.05 to a maximum of 0.48 on the composite self-esteem index. Given the potential presence of other errors of measurement, the interview-questionnaire response differentials obtained in the present study are not of sufficient magnitude to be considered as indicative of a response bias effect of substantive or practical importance.

Due to the unavailability of reliable information on the actual drinking habits of the students and their parents, the school grades and other responses under consideration, it was not possible to conduct an evaluation of the relative accuracy of the interview and questionnaire responses. As a result it is not possible to indicate the relative superiority of either the self-administered mode or the personal interview for the question responses under consideration. Both types of responses may be subject to an under- or over-reporting bias of an indeterminant direction and/or magnitude.

The results of this note are in general agreement with Bradburn and Sudman (1979) who indicate that no consistent relationship appears to exist between the method of survey administration and the over-reporting of socially desirable behaviour or the under-reporting of socially undesirable behaviors and attitudes. As a result Bradburn and Sudman (1979) and Locander *et al.* (1976) suggest that no data collection procedure is clearly superior for all types of threatening or other questions of concern to survey practitioners.

ACKNOWLEDGEMENT

The author is grateful to S. Parvez Wakil for his critical comments on the original version of this paper and to the anonymous referees and M. P. Singh for their very helpful comments.

The initial study was supported by Health and Welfare Canada Non-Medical Use of Drugs Directorate (Grant #1213-7-10) with additional support from the Applied Research Unit, Psychiatric Research Division, Saskatchewan Department of Health. The study's grantees are also acknowledged for their generosity in allowing the use of the data for this paper.

REFERENCES

ALWIN, D.F. (1977). Making errors in surveys: an overview. *Sociological Methods and Research*, 6, 131-151.

BLAIR, E., SUDMAN, S., BRADBURN, N.M., and STOCKING, C. (1977). How to ask questions about drinking and sex: response effects in measuring consumer behavior. *Journal of Marketing Research*, 14, 316-321.

BRADBURN, N.M., and SUDMAN, S. (1979). *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.

CANNELL, C.F., and FOWLER, F.J. (1963). Comparison of a self-enumerative procedure and a personal interview: a validity study. *Public Opinion Quarterly*, 27, 250-264.

DeLAMETER, J., and MacCORQUODALE, P. (1975). The effects of interview schedule on reported sexual behavior. *Sociological Methods and Research*, 4, 215-236.

ELLIS, A. (1947). Questionnaire versus interview methods in the study of human love relationships. *American Sociological Review*, 12, 541-553.

GIBSON, F.W., and HAWKINS, B.W. (1968). Interview versus questionnaires. *American Behavioral Scientist*, 12, 9-11.

HERZOG, A.R., RODGERS, W., and KULKA, R.A. (1983). Interviewing older adults: a comparison of telephone and face-to-face modalities. *Public Opinion Quarterly*, 47, 405-418.

HETHERINGTON, R.W., DICKINSON, J., CIPYWNYK, D., and HAY, D.A. (1978). Drinking behavior among Saskatchewan adolescents. *Canadian Journal of Public Health*, 69, 315-324.

HETHERINGTON, R.W., DICKINSON, J., CIPYWNYK, D., and HAY, D.A. (1979). Attitudes and knowledge about alcohol among Saskatchewan adolescents. *Canadian Journal of Public Health*, 70, 247-259.

HUBBARD, R.L., ECKERMAN, W.C., and RACHAL, J.V. (1976). Methods of validating self-reports of drug use: a critical review. *Proceeding of the Social Statistics Section, American Statistical Association, Part I*, 406-409.

KNUDSEN, D., HALLOWELL, D., and IRISH, D.P. (1967). Response differences to questions on sexual standards: an interview-questionnaire comparison. *Public Opinion Quarterly*, 21, 290-297.

KROHN, M., WALDO, G.P., and CHIRICOS, T.G. (1974). Self-reported delinquency: a comparison of structured interviews and self-administered checklists. *The Journal of Criminal Law and Criminology*, 65, 545-553.

LOCANDER, W., SUDMAN, S., and BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. *Journal of American Statistical Association*, 71, 269-275.

McDONAGH, E.C., and ROSENBLUM, A.L. (1965). A comparison of mailed questionnaires and subsequent structured interviews. *Public Opinion Quarterly*, 29, 131-136.

METZNER, H., and MANN, F. (1952). A limited comparison of two methods of data collection: the fixed alternative questionnaire and the open-ended interview. *American Sociological Review*, 17, 486-491.

NEWTON, R.R., PRENSKY, D., and SHUESSLER, K. (1982). Form effect in the measurement of feeling states. *Social Science Research*, 11, 301-317.

SCHUMAN, H. (1980). Review of improving interview method and questionnaire design. *Social Forces*, 59, 325-326.

SELLTIZ, C., JAHODA, M., DEUTSCH, M., and COOK, S.W. (1959). *Research Methods in Social Relations* (Revised). New York: Holt, Rinehart and Winston.

SIEMIATYCKI, J. (1979) A comparison of mail, telephone, and home interview strategies for household health surveys. *American Journal of Public Health*, 69, 238-245.

SMITH, H.W. (1975). *Strategies of Social Research*. Englewood Cliffs: Prentice Hall.

SYKES, W.M., and COLLINS, M. (1987). Comparing telephone and face-to-face interviewing in the United Kingdom. *Survey Methodology*, 13, 15-28.

WHITEHEAD, P.C., and SMART, R.G. (1972). Validity and reliability of self-reported drug use. *Canadian Journal of Crimnology and Corrections*, 14, 83-89.

WISEMAN, F. (1972). Methodological bias in public opinion polls. *Public Opinion Quarterly*, 36, 105-108.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 137–143
Statistics Canada

Use of Cluster Analysis for Collapsing Imputation Classes

E.R. LANGLET[1]

ABSTRACT

The problem of collapsing the imputation classes defined by a large number of cross-classifications of auxiliary variables is considered. A solution based on cluster analysis to reduce the number of levels of auxiliary variables to a reasonably small number of imputation classes is proposed. The motivation and solution of this general problem are illustrated by the imputation of age in the Hospital Morbidity System where auxiliary variables are sex and diagnosis.

KEY WORDS: Item nonresponse; Auxiliary variables; Imputation matrix; Donors; Disjoint techniques; Hierarchical techniques; Cluster seeds.

1. STATEMENT OF THE PROBLEM

In surveys, the problem of item nonresponse occurs when some but not all information is collected for a sample unit or when some information is deleted because it fails to satisfy edit constraints. In many surveys, this problem is handled by random imputation within classes, a common form of hot deck imputation method. For this type of imputation, a respondent is chosen at random within an imputation class defined by one or more auxiliary variables and the respondent's value is assigned to the nonrespondent.

The problem considered in this paper can be defined as follows. The classifications of the respondents according to certain auxiliary variables form a multi-dimensional imputation matrix where the number of imputation classes equals the number of cross-classification cells defined by the auxiliary variables. If the number of imputation classes is very large, few or no donors may be available in several classes. In addition, manipulation of this large matrix could be very cumbersome computationally. These problems can be alleviated by collapsing the cells of the matrix either by grouping the cells themselves, or the rows, columns or along some other dimension (or combination of dimensions) so that the resulting groups will be homogeneous with respect to the variables requiring imputation. We propose to use cluster analysis to achieve the desired level of collapsing. For this purpose, the values of the variables of interest from donors (or respondents) for each imputation class can be used to assign numerical scores to each class. In this paper, measures based on empirical distribution function for respondent data are used to quantify imputation classes. Cluster analysis can then be used to group the cells of the matrix according to these numerical scores. It will be shown that cluster analysis is appropriate for the problem under consideration. Related useful references concerning the application of cluster analysis to stratify primary sampling units are Drew, Bélanger and Foy (1985), Judkins and Singh (1981) and other references contained therein.

The above mentioned problem arose in the context of age imputation in the Hospital Morbidity System (HMS). This system uses the auxiliary variables sex and diagnosis as the basis for imputing the age. The number of imputation classes were over 5,000 for each sex. A solution based on the technique of cluster analysis was proposed in order to collapse the levels of

[1] E.R. Langlet, Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

the diagnosis variable to 40 groups of related diagnoses. In section 2, a brief review of the commonly used cluster analysis techniques is presented. Use of cluster analysis for the problem of collapsing imputation classes is illustrated for the example of imputation of age for the HMS data in section 3 including the relative performance of the proposed method with respect to the current method. Both methods utilize a hot deck approach but the proposed method redefines the imputation classes using cluster analysis. Some concluding remarks including possible generalizations of the method are given in section 4.

2. CLUSTER ANALYSIS TECHNIQUES: A BRIEF REVIEW

The problem of classifying a given number of entities described by a number of quantitative variables into groups such that entities within the same groups or clusters will be similar to each other and dissimilar to entities in different groups is considered in this section. A good review of clustering techniques is given by Everitt (1980) mainly based on the work of Cormack (1971). Most clustering techniques can be classified into two groups, namely 'hierarchical techniques' and 'disjoint techniques', the latter one also known as 'optimization techniques'. These two groups of techniques will be described below. Some other methods, are density techniques where clusters are formed by searching for regions containing dense concentrations of entities. This is based on the fact that if entities are described as points in a metric space, there should be parts of the space in which the points are very dense, separated by parts of low density. Another class of techniques is called clumping techniques in which the clusters can overlap. In certain fields such as language studies, for example, classification must permit an overlap between the classes because words tend to have several meanings, and if they are classified by their meanings they may belong in several places.

Hierarchical techniques can be subdivided into 'fusion techniques' and 'divisive techniques'. In fusion methods, each entity begins in a cluster by itself. At each step, the two closest clusters are fused to form a new cluster until only one cluster containing all the observations is left. In divisive techniques, all entities are first grouped into one cluster. Then, at each step, groups of the entities are successively broken down into finer partitions until each entity constitutes a cluster by itself. Hierarchical techniques differ with respects to the definition of the distance measure between observations or groups of observations. An advantage of hierarchical techniques is that a single run can produce results for one cluster to as many as you like by stopping the fusion or division process at the desired level of the hierarchy. Obviously, hierarchical techniques can be used for only small data sets since there are $n(n-1)/2$ possibilities to fuse two entities in a group of n entities and $2^{n-1}-1$ possibilities to break a group of n entities in two groups.

In contrast to hierarchical techniques where observations belong to a series of clusters depending on the level of the hierarchy, disjoint techniques divide observations into a number of clusters (generally predetermined) such that each observation belongs to one and only one cluster. They also differ from hierarchical techniques in that they admit relocation of the observations so that a poor initial partition can be corrected at a later stage. Disjoint techniques are clearly more appropriate than hierarchical techniques to handle large data sets. Disjoint techniques are also called optimization techniques because they seek for a partition of the data which optimizes some predefined criterion. Various disjoint techniques differ in the way the methods obtain an initial partition and in the clustering criterion they try to optimize. Usually, disjoint techniques start by selecting a set of points called cluster seeds as a first guess of the means of the clusters. A number of procedures have been suggested for choosing these points

(Anderberg 1973). Once the cluster seeds have been selected, the entities are then assigned to the closest cluster seeds (usually, the Euclidean distance is used). Estimates of the cluster means might be updated after each allocation (MacQueen 1967) or after all entities have been allocated (Ball and Hall 1967). Once an initial partition has been found (which is equivalent to finding a set of cluster seeds and to allocating each entity to the closest cluster seed), a search is made for entities whose re-allocation to some other group will improve the clustering criterion. This procedure is repeated until no further move of a single entity improves the clustering criterion. A local optimum is then reached. This is what Anderberg (1973) calls 'nearest centroid sorting'. In general, there is no way to know whether a global optimum has been reached.

3. APPLICATION: FORMING IMPUTATION CLASSES FOR THE HMS

3.1 Background

The Hospital Morbidity System (Statistics Canada 1987) consists of a count of inpatient cases, discharged during the year from general and allied special hospitals in Canada except Yukon and Northwest Territories. Each record of the system contains at least one diagnosis code, the age and sex of the patient, the length of stay, *etc.* The first valid diagnosis on the record is called the tabulating diagnosis and is the diagnosis on which tabulations are based in the publications. This diagnosis can be seen as the main cause for which the patient is hospitalized and is coded according to the 9th Edition of the International Classification of Diseases (World Health Organization 1977) which contains more than 5,000 diagnoses.

The age imputation problem in the HMS is currently treated by a hot deck method. In this imputation problem to predict the age of the patient y, two auxiliary variables are used, namely the tabulating diagnosis d which is always present on the record and the sex of the patient s. The sex itself needs to be imputed first if it is missing according to the observed male/female proportions of d over previous years. Classification of the patients according to d and s forms an imputation matrix with the number of imputation classes larger than 5000×2. In order to reduce the dimension of the imputation matrix, diagnoses were regrouped or collapsed, based on the age distribution of each diagnosis. Let F_d denote the age distribution in the population of the patients with tabulating diagnosis d. Then, diagnoses A and B would be collapsed together if F_A is close to F_B. Estimates of F_d from available data can be used for this purpose. It should be noted that the sex variable was not used in defining imputation classes (see section 4 for details on how it could be used) although it was used in the imputation scheme. By not using the sex variable for defining imputation classes, the number of imputation classes of the imputation matrix is reduced by half.

In order to motivate the proposed method for collapsing imputation classes, we will first describe the current method and its limitations. The collapsed groups were created by comparing manually (using histograms) the shapes of the empirical age frequency distributions, $\hat{F}_d$ of all diagnosis codes corresponding to 1974 HMS data. Thirty six groups were obtained and a 37th group was created for those diagnoses for which less than 200 observations were available. The number of groups was determined a posteriori arbitrarily. The main deficiency of the current method comes from the fact that no statistical criterion was used to group diagnoses which makes the method labour intensive and somewhat subjective. These groups were obtained by simply comparing histograms. An evaluation of the current imputation method indicated that the resulting groups of diagnoses were, in a few cases, not homogeneous with respect to $\hat{F}_d$ and consequently needed to be updated.

3.2 Proposed Method

The proposed method can be briefly described as follows. We shall consider the case when only one quantitative variable needs to be imputed. Extension to cases where more than one variable requires imputation is discussed in section 4. Let's denote by y the variable to be imputed and by F_i the distribution of variable y in class i. Note that the classes are defined by the cross-classification of one or more auxiliary variables which are suitably categorized if necessary. The first step is to find an appropriate set of parameters to represent F_i in each class, for example, the first three or four moments of the F_i's or the percentiles. The next step is to estimate these parameters from the respondent data. Finally, a suitable technique of cluster analysis on the set of estimated parameters can be used to condense the number of classes such that classes grouped together will be similar with respect to the parameters representing the F_i's.

A justification for the choice of the proposed method in the context of the age imputation for the Hospital Morbidity System (HMS) will now be presented. First, consider some possible alternative strategies to the collapsing problem. One strategy for this problem might be similar to the original method that was used for 1974 data, that is, to group diagnoses according to the distributions $\hat{F}_d$ but using a statistical criterion for grouping instead of manually comparing histograms. Data would be cross-classified by tabulating diagnoses, sex and a number of age groups, say 10. Two diagnoses would be grouped together if the proportion of cases in each of these ten age groups, $p_1, \ldots, p_{10}$ were judged to be close to each other according to some criterion such as the Euclidean distance or a chi-square measure. Note that the use of a chi-square measure would cause serious computational burden since no commonly available cluster analysis program uses this distance measure. This would imply the calculation of the chi-square distance for all possible pairs of diagnoses. Another possible strategy would be to first use data reduction techniques such as principal components to reduce the dimension of age groups and then decide whether two diagnoses are close based on principal component scores. An obvious disadvantage to all these methods is the number of observations required to obtain a reliable estimate of the categorical age distribution for each diagnosis.

In view of the above problem, we decided to use the first two or three moments to approximately describe F_d. We started with three - the mean m_d, the standard deviation s_d and the skewness coefficient b_d. However, it was found by means of principal component analysis that it was not necessary to include b_d. The approach then is to collapse diagnoses according to the sample mean, m_d, and the sample standard deviation s_d. Cluster analysis can be used to provide a suitable statistical technique for this purpose. An obvious advantage with this approach over other strategies based on the categorical distribution of age is that a reliable estimation of two moments requires much fewer observations than the estimation of the proportion of cases over several age groups. In section 4, implementation of this approach is described for the problem of age imputation.

3.3 Procedure Steps in the Implementation of the Proposed Method for HMS Data

There are four steps in implementing the proposed collapsing method based on cluster analysis for the age imputation problem for HMS data.

Step I: Selection of a clustering method

Before selecting a clustering method, it should be noted that our goal is primarily to partition the diagnoses into homogeneous groups without trying to uncover 'natural' or 'real' clusters. This is called 'data dissection' in the literature (Everitt 1980). Another important consideration is the availability of a well tested clustering program using an efficient

clustering method.The determinant consideration for the selection of a clustering method was the number of observations in our data set which resulted in the selection of a disjoint technique rather than a hierarchical technique.

Taking into consideration the above points, the disjoint clustering technique used in the FASTCLUS procedure of SAS (1985) was chosen to do the analysis. This procedure performs a disjoint cluster analysis based on the usual Euclidean distances computed from a given set of quantitative variables. The FASTCLUS procedure combines an effective method for finding initial clusters (or initial clusters can be given by the user) with a standard iterative algorithm for minimizing the sum of squared distances from the cluster means. FASTCLUS was directly inspired by Hartigan's leader algorithm (1975) and MacQueen's k-means algorithm (1967). A set of cluster seeds is first selected as a guess of the means of the clusters. Each observation is assigned to the nearest cluster seed to form temporary clusters. The cluster seeds are replaced by the means of the temporary clusters each time an observation is assigned (this is an option chosen for our application). After each pass through the data set, the observations are assigned to the nearest cluster seed until the changes in the cluster seeds become small or null (chosen to be null for our application). The final clusters are formed by assigning each observation to the nearest cluster seed.

Step II: Estimation of parameters

Two years of HMS data from 82–83 and 83–84 fiscal years were gathered to get estimates m_d and s_d for each diagnosis d. These estimates were the usual weighted estimates over the two year period. Each diagnosis is represented by two variables, m_d and s_d. The problem is now reduced to finding an appropriate partition of the diagnoses according to m_d and s_d. Three special groups of diagnoses judged as outliers were removed. These three special groups will form the first three rows of the imputation matrix (the columns are defined by the sex variable). A catch-all category was created in the last row of the imputation matrix for those diagnoses with, say, fewer than ten observations available over the two years of data and not included in the three special groups. The choice for the upper bound of ten observations was made arbitrarily. Cluster analysis can then be used to group the remaining diagnoses not included in the three special groups with at least ten observations available.

Step III: Determination of the number of clusters

The determination of the number of clusters was dictated by operational constraints since the imputation module of the program doing the imputation will accept a maximum number of rows not larger than 40. Since there are already three rows for special diagnoses and one row for diagnoses with fewer than ten observations, the maximum number of other rows that would not affect the program is then 36. A small empirical study calculating the R^2 coefficient for different numbers of clusters indicated that the R^2 coefficient was already above 98% for 36 clusters, suggesting that 36 clusters was acceptable. Note that even with 15 clusters, the R^2 could be made as high as 95%. The definition of the R^2 coefficient is given in section 3.4.

Step IV: FASTCLUS implementation

First, an initial partition of the observations into 36 groups was chosen (equivalent to choosing a set of 36 cluster seeds). Better results were obtained by selecting an initial set of cluster seeds than by letting FASTCLUS find initial cluster seeds. Note that different initial cluster seeds and different orders of the input data set will yield different results

due to the fact that the method produces only locally optimal partitions. To select cluster seeds, diagnoses were divided into nine groups of roughly the same size according to m_d and four groups of roughly the same size according to s_d. This procedure produced 36 homogeneous groups of diagnoses of approximately the same size. The means of the two variables m_d and s_d in each group were taken as initial cluster seeds. Several other variations were tried and the procedure giving the largest R^2 was chosen.

Second, since m_d and s_d were based on very different numbers of observations for different diagnoses, it was judged preferable to perform a weighted cluster analysis, the weights being the number of observations available for each diagnosis. Note that, in this case, FASTCLUS would minimize the weighted within cluster sum of squares instead of an unweighted within-cluster sum of squares.

3.4 Relative Performance of the Proposed Method

One way to compare the current and proposed method for collapsing imputation classes is to use the R^2 coefficient pooled over all variables (in our case, it would be the mean and the standard deviation). The pooled R^2 coefficient is the proportion of the total variance explained by the between cluster pooled sum of squares (which should be as large as possible). Each pooled sum of squares is defined as $(SSQ_m + SSQ_s)/2$ where SSQ_m and SSQ_s are the sums of squares of the mean and the standard deviation respectively. The R^2 coefficients obtained from FASTCLUS were 0.993 for m_d and 0.929 for s_d for a pooled R^2 value of 0.986. The current classification of diagnoses into groups would yield an R^2 of 0.735 for m_d and 0.466 for s_d producing a pooled R^2 value of 0.705. Thus, in terms of R^2, results indicated that the groups of diagnoses formed using cluster analysis were much more homogeneous with respect to the variable being imputed than in the case where classes were formed by the earlier method.

4. CONCLUDING REMARKS

A methodology based on cluster analysis for collapsing the imputation classes of an imputation matrix defined by the cross-classification of several auxiliary variables was proposed. This methodology was applied to the imputation of age for the Hospital Morbidity System where diagnosis and sex were used as auxiliary variables.

It should be noted that in this specific application, only one variable, namely the diagnosis, was used to collapse the original imputation classes. The variable sex is, however, used later in the imputation scheme so that a recipient will be matched to a donor of the same sex. In a generalization of the proposed method, one may consider using the two variables, sex and diagnosis, in the collapsing process. For this purpose one might also impose some constraints that male and female cases of the same diagnosis belong to the same row in the final imputation matrix. Alternatively, one could produce two final imputation matrices, one for each sex. In either one of these alternatives, the number of initial imputation classes would clearly be much higher and hence the collapsing problem more complex. In this situation, it is more likely for many classes to have a small number of donors and therefore many of the imputation classes would have to be assigned to the catch all category. This, however, may not be desirable in practice. This problem can be simplified if one could make the assumption that, for most diagnoses, the male and female age distributions are similar to each other. There is some evidence based on significance tests that this is not an unreasonable assumption. In the HMS example considered, it was decided to group diagnoses based on estimates of μ_d and σ_d from the data pooled over sex.

It should also be noted that the choice of mean and standard deviation of age distribution to assign numerical scores to each imputation class was not investigated. Other choices might be percentiles or some other parameters of the age distribution. Clearly, the results of using cluster analysis for collapsing purpose would depend on the choice of the above scores.

Finally, generalization of the proposed method to the case where $k \geq 1$ variables need to be imputed and where $p \geq 2$ auxiliary variables are available follows in a straightforward manner from the simpler case considered in this paper.

ACKNOWLEDGEMENTS

This work was presented at the annual meeting of the "Association canadienne-française pour l'avancement des sciences" in May 1988. I thank Avi Singh for his helpful comments which have led to improvements in this paper. I would like to thank Cyril Nair of the Health Division and his staff for their support especially concerning the production of the computer files required to complete this work.

REFERENCES

ANDERBERG, M.R. (1973). *Cluster Analysis for Application*. New York: Academic Press.

BALL, G.H., and HALL, D.J. (1970). Some implications of interactive graphic computer systems for data analysis and statistics. *Technometrics*, 12, 17-31.

CORMACK, R.M. (1971) A review of classification. *Journal of the Royal Statistical Society,* Series A, 134, 321-367.

DREW, J.D., BÉLANGER, Y., and FOY, P. (1985). Stratification in the Canadian Labour Force Survey. *Survey Methodology,* 11, 95-110.

EVERITT, B.S. (1980). *Cluster Analysis*. Second Edition, London: Heineman Education Books Ltd.

JUDKINS, D.R., and SINGH, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-284.

HARTIGAN J.A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.

MacQUEEN J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings 5th Berkeley Symposium* 1, 281-297.

SAS INSTITUTE Inc. (1985). *SAS User's Guide: Statistics*, Version 5.

STATISTICS CANADA (1986). *Hospital Morbidity 1981-82, 1982-83*. Catalogue No. 82-206, Statistics Canada, Ottawa.

WORLD HEALTH ORGANIZATION (1977). *International Classification of Diseases*. 1975 Revision, Volume 1, Geneva.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 145–151
Statistics Canada

An Example of the Use of Randomization Tests in Testing the Census Questionnaire

YVES BÉLAND and ALAIN THÉBERGE

ABSTRACT

Modular Test 2 was a survey conducted by Statistics Canada that used two different questionnaires. Its purpose was to assist in the making of the 1991 census questionnaire. The sample used for the survey was not a probability sample. This article briefly describes the survey methodology, and the use of randomization tests to compare the two questionnaires.

KEY WORDS: Randomization tests; Non-probability sample; Experimental design.

1. INTRODUCTION

Statistical tests could be classified into two groups, randomization tests and classical tests. A classical test, is based on a comparison of the observed value of a statistic with the distribution, under the null hypothesis, of the values of this statistic for the set of samples that could have been selected. To conduct this kind of test, the probability of selecting any given sample must be known; therefore probability sampling using a known design is required. A randomization test is based on a comparison of the observed value of a statistic with the distribution, under the null hypothesis, of the values of this statistic for all possible permutations of the data. This was the method used by Fisher to compare two seed samples (1935), and Edgington (1987) also discusses various aspects of this method. "Treatments" are required to define the permutations in a randomization test, and the probability of obtaining a given permutation must also be known. Which unit will be given which treatment must be decided randomly; that is, the experimental design must incorporate randomization.

In an organization like Statistics Canada, classical tests are generally used because most of the sample surveys done by Statistics Canada use probability sampling, and also because there are no treatments in these surveys. This article describes how randomization tests were used in a survey that was an exception to the rule.

In Section 2, the methodology used in the modular tests is described briefly. Section 3 describes using simple examples the procedure used in a randomization test. Section 4 describes how randomization tests were applied to Modular Test 2.

2. MODULAR TESTS

As part of the planning for the 1991 census, two modular tests were carried out to test questions likely to be asked in the census. The purpose of these surveys was to ensure that each question whether new or just reformulated was easy to understand. We refer to the tests as "modular" because they were independent surveys that tested different sections of the census questionnaire.

[1] Yves Béland, Social Survey Methods Division; Alain Théberge, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6

Modular Test 1 was carried out in November, 1987 in order to revise newly-formulated questions dealing with population coverage, marital status, fertility, volunteer work, and nuptiality. This first survey used neither classical nor randomization tests.

Modular Test 2, carried out in January, 1988, was designed principally to measure the reaction of ethnic groups to questions on language, ethnic origin, religion, citizenship, and mobility. In Modular Test 2, a two-stage sampling plan was used to select about 3,500 households taken from within the metropolitan areas of Halifax, Québec, Montréal, Toronto, Winnipeg, and Vancouver. To reduce costs and to make data collection easier, and to get a sample that contained people of diverse ethnic origins, a non-probability method was used to select the sample. The questionnaire used in Modular Test 2 came in two versions. The differences are described in Section 4. The households in the sample were given either version 1 or version 2 on a random basis.

Randomization tests were used to allow us to statistically test hypotheses pertaining to Modular Test 2. Randomizations tests can be used to compare two treatments applied to units in samples which may not be probability samples.

3. RANDOMIZATION TESTS

The procedure for doing a randomization test will now be described. First, the value of a statistic is calculated for the observed data. Next, the value of the same statistic is calculated for the other permutations of the data that are possible with the experimental design used. H_0 is rejected if the value of the statistic for the observed data is extreme in relation to the values obtained under H_0 for the set of permutations.

For example, suppose there are four households. Household 1 has three persons, households 2 and 3 have two, and household 4 has one. These households may have been chosen arbitrarily, but a household whose members will receive treatment Y is chosen at random. Members of the three other households will receive treatment X. Suppose that household 4 is selected for treatment Y. For household 1, the treatment succeeds for two of the three members, for households 2 and 3, for one of two members, and for household 4, it fails for the sole member. Our null hypothesis states that the results are independent of the treatment used. To measure the impact of treatment X compared to treatment Y, the statistic S, giving the average number of successes for treatment X minus the average number of successes for treatment Y is calculated. Here $S = (2 + 1 + 1)/(3 + 2 + 2) - 0/1 = 4/7$. To find out whether this value is significant, the values for S obtained by permuting the observations are given in Table 1. Each observation in Table 1 shows the number of members in the household after the vertical bar, and the number of successes before the vertical bar. If a right-tailed test is used, H_0 is rejected when $\alpha \geq 3/12 = .25$, because three of the twelve permutations yield an S value greater than or equal to 4/7, the observed value.

Rather than permuting the observations, we could have permuted the treatments. Table 2 gives the results when this is done. Because only one of the four permutations yields a value for S greater than or equal to 4/7 for a right- tailed test, we again reject H_0 *if* $\alpha \geq 1/4 = .25$. It is not a coincidence if the results are the same. Note n_{ki}, the number of units that receive treatment k ($k = 1, \ldots, K$) and for which the result r_i ($i = 1, \ldots, I$) is observed; $n_{k.} = \sum_i n_{ki}$ the number of units that receive treatment k, $n_{.i} = \sum_k n_{ki}$, the number of units for which the result r_i is observed; and $n_{..} = \sum_k \sum_i n_{ki}$, the total number of units. The number, N_t, of permutations of the treatments is given by

$$N_t = n_{..}! \Big/ \prod_k (n_{k.}!). \tag{1}$$

Table 1
Values of the Statistics S for each Permutation of the Observations

Treatment	Permutations											
X	2 | 3	1 | 2	1 | 2	2 | 3	2 | 3	1 | 2	1 | 2	0 | 1	0 | 1	1 | 2	1 | 2	0 | 1
X	1 | 2	2 | 3	1 | 2	1 | 2	0 | 1	2 | 3	0 | 1	1 | 2	2 | 3	1 | 2	0 | 1	1 | 2
X	1 | 2	1 | 2	2 | 3	0 | 1	1 | 2	0 | 1	2 | 3	2 | 3	1 | 2	0 | 1	1 | 2	1 | 2
Y	0 | 1	0 | 1	0 | 1	1 | 2	1 | 2	1 | 2	1 | 2	1 | 2	1 | 2	2 | 3	2 | 3	2 | 3
S	4/7	4/7	4/7	0	0	0	0	0	0	−4/15	−4/15	−4/15

Table 2
Values of the Statistics S for each Permutation of the Treatments

Observation	Permutations			
2 | 3	X	X	X	Y
1 | 2	X	X	Y	X
1 | 2	X	Y	X	X
0 | 1	Y	X	X	X
S	4/7	0	0	−4/15

Of these N_t permutations, there are N_t^* for which n_{ki} units are associated with treatment k and the result r_i ($k = 1, 2, \ldots, K$; $i = 1, 2, \ldots, I$), where

$$N_t^* = \prod_i \left(n_{.i}! \Big/ \prod_k (n_{ki}!) \right). \tag{2}$$

In addition, there are N_o permutations of the observations where

$$N_o = n_{..}! \Big/ \prod_i (n_{.i}!). \tag{3}$$

Of these N_o permutations, there are N_o^* for which n_{ki} units are associated with treatment k and the result r_i ($k = 1, 2, \ldots, K$; $i = 1, 2, \ldots, I$), where

$$N_o^* = \prod_k \left(n_{k.}! \Big/ \prod_i (n_{ki}!) \right). \tag{4}$$

Because $N_o^*/N_o = N_t^*/N_t$, the tests are equivalent. To reduce the number of calculations, it is preferable to permute the treatments if $N_t < N_o$, and to permute the observations if $N_t > N_o$. Dwass (1957) suggests that when there are a large number of permutations, a sample of permutations can be taken, and the observed value of the statistic can be compared to the set of values for the sample. If all of the permutations are not considered, the level of the test is not affected, only its power is.

If the permutations are sampled, the rule given above can still be applied, not to reduce the number of calculations, but to minimize the loss of power due to sampling. For example, Dwass shows that for a one-tailed test at the 0.05 level, the loss of power for a sample of 999 permutations is no more than 5.5%. Bradley (1968) notes that when the power of randomization and classical tests are compared, the results depend on to what extent the requirements of the classical tests have been met.

Because of the way in which randomization tests are constructed, the inference applies only to the effect of treatment on units in the sample, and not to the entire population. Classical tests, however, are based on a random sample drawn from a population that rarely matches the population of interest. In the present case for example, the population of interest is the Canadian population on Census Day, June 4, 1991. So for both types of tests, non-statistical arguments must be used to generalize inferences to the population of interest.

4. THE USE OF RANDOMIZATION TESTS IN MODULAR TEST 2

As mentioned above, there are two questionnaire versions for Modular Test 2, versions X and Y. Questions on ethnic identity and ethnic origin differ in the two versions. "CANADIAN" is a response category in version X that the respondent can select to answer the questions on ethnic identity and origin. In version Y, those who want to respond "CANADIAN" must write it out in full after selecting the category, "OTHER."

We wanted to know whether questions on ethnic identity and origin in version X of the test questionnaire got more or got less multiple responses than these questions in version Y. By a multiple response we mean any response in which more than one category has been chosen. We also wanted to find out what bearing the type of questionnaire had on multiplicity (number of response categories selected by the respondent), and on the selection of certain response categories (such as "FRENCH") for these questions. The types of questionnaire constitute the treatments. Because the sample for each region had its peculiarities, the randomization tests were done separately for each of the metropolitan areas from which the sample was taken.

First of all, we generated at random a sample of 999 permutations of the questionnaire versions. A permutation is generated as follows: For any given region, let N_x and N_y represent the number of X and Y questionnaires respectively. Using Bebbington's algorithm (1975), from the $N_x + N_y$ households take a simple random sample of N_x households. Household members in this sample are then assigned version X of the questionnaire. This process is repeated 999 times. Next, calculate for a given question the proportion of respondents who gave a multiple response for version X and for version Y. These proportions are denoted P_x and P_y.

Next, for each of the 999 permutations of the questionnaire versions, as well as for the initial observed sample, we calculated the statistic $S = P_x - P_y$. In this way we obtained 1,000 values for S, which we ranked in increasing order. If more than one statistic had the same value, we generated a random number between 0 and 1 and used it to determine the order of statistics of the same value. We used the variable RANKP_{x-y} to represent the rank of an observed S statistic.

Let μ_x and μ_y represent the expected proportion of respondents who gave a multiple response for version X and version Y respectively. For all regions excluding Halifax we tested:

$$H_0: \mu_x = \mu_y$$

versus

$$H_1: \mu_x > \mu_y.$$

For Halifax, the counter-hypothesis H_1: $\mu_x < \mu_y$ was used because more multiple responses were expected for version Y of the questionnaire. Because "CANADIAN" was not an available response category on version Y of the questionnaire and because the majority of households selected in this region were made up of people of British origin (that is, English, Scottish, or Irish), members of households that received version Y marked one or more of these categories. Members of households that received version X had the option of marking only the "CANADIAN" category.

The critical level, $\hat{\alpha}$, is calculated as follows: for the Halifax region, given that H_0 is rejected if the proportion of respondents who gave a multiple response in version X is significantly lower than the proportion observed for Y, the critical level is $\text{RANKP}_{x-y}/1000$. For all the other regions, given that H_0 is rejected if the proportion for X is significantly higher than the proportion observed for Y, the critical level is $(1001 - \text{RANKP}_{x-y})/1000$. The results are shown in Table 3.

Randomization tests were also used to test multiplicity (the number of response categories selected by the respondent) for questions on ethnic identity and origin in each of the regions, but this time ratios (R_x, R_y) are used, instead of proportions (P_x, P_y). Ratio R_x is the average number of response categories selected by respondent for a question in version X of the questionnaire, and ratio R_y is the average number of response categories selected in version Y. The rest of the method is the same except that instead of RANKP_{x-y}, RANKR_{x-y} is used, and the statistic S is defined as $R_x - R_y$. However, because there is greater variability for the values of the statistic S in the tests for multiplicity, a sample of 1,999 permutations was generated instead of 999.

Let F and G represent the distribution functions of the number of response categories selected in version X and version Y respectively. For all the regions excluding Halifax, we test the hypothesis

$$H_0: F = G$$

versus

$$H_1: F(z) \leq G(z) \text{ for all } z \text{ and } F \neq G.$$

If H_0 is rejected, the number of response categories selected for an X questionnaire is said to be stochastically larger than the number of response categories selected for a Y questionnaire. For Halifax, the counter-hypothesis used is H_1: $F(z) \geq G(z)$, for all z and $F \neq G$. The results are shown in Table 3. In the Québec region, the value of R_y is less than 1 for each question. This is because most respondents in this region chose only one response category, and some respondents did not answer one or other of the questions.

Finally, versions X and Y for Modular Test 2 were compared for some regions as to the number of respondents who identified themselves as being of French, Italian, or British origin. By "BRITISH", we mean that at least one of the categories "IRISH," "SCOTTISH," or "ENGLISH" was chosen. For example, if a test was done on the proportion of people selecting "FRENCH", μ_x and μ_y were defined as the expected proportion of questionnaires where the response "FRENCH" would be chosen in versions X and Y of the questionnaire. In all regions, we tested

$$H_0: \quad \mu_x = \mu_y$$

versus

$$H_1: \quad \mu_x < \mu_y.$$

The randomization tests were done using 999 permutations. The results are shown in Table 4.

Table 3
Critical Levels for the Rate of Multiple Responses and Multiplicity

		Multiple Response			Multiplicity		
Question	Region	P_x	P_y	$\hat{\alpha}$	R_x	R_y	$\hat{\alpha}$
ORIGIN	HALIFAX	0.435	0.536	0.087	1.617	1.914	0.062
ORIGIN	QUÉBEC	0.154	0.043	0.001	1.143	0.986	0.001
ORIGIN	MONTRÉAL	0.185	0.194	0.612	1.141	1.152	0.585
ORIGIN	TORONTO	0.127	0.122	0.393	1.124	1.125	0.495
ORIGIN	WINNIPEG	0.293	0.307	0.622	1.439	1.398	0.345
ORIGIN	VANCOUVER	0.285	0.296	0.621	1.440	1.392	0.280
IDENTITY	HALIFAX	0.220	0.335	0.035	1.244	1.502	0.029
IDENTITY	QUÉBEC	0.140	0.016	0.001	1.131	0.959	0.001
IDENTITY	MONTRÉAL	0.159	0.125	0.063	1.075	1.044	0.186
IDENTITY	TORONTO	0.186	0.120	0.001	1.154	1.075	0.005
IDENTITY	WINNIPEG	0.224	0.195	0.248	1.253	1.208	0.298
IDENTITY	VANCOUVER	0.186	0.183	0.457	1.182	1.137	0.202

Table 4
Critical Levels for Selected Variables

Question	Variable	Region	P_x	P_y	$\hat{\alpha}$
ORIGIN	FRENCH	QUÉBEC	0.127	0.897	0.001
ORIGIN	FRENCH	MONTRÉAL	0.038	0.210	0.001
ORIGIN	BRITISH	HALIFAX	0.321	0.837	0.001
ORIGIN	BRITISH	MONTRÉAL	0.034	0.092	0.002
ORIGIN	BRITISH	TORONTO	0.085	0.135	0.003
ORIGIN	BRITISH	WINNIPEG	0.167	0.234	0.054
ORIGIN	BRITISH	VANCOUVER	0.267	0.325	0.065
IDENTITY	FRENCH	QUÉBEC	0.138	0.899	0.001
IDENTITY	BRITISH	HALIFAX	0.153	0.828	0.001
IDENTITY	BRITISH	MONTRÉAL	0.022	0.117	0.001
IDENTITY	BRITISH	TORONTO	0.050	0.215	0.001
IDENTITY	BRITISH	WINNIPEG	0.074	0.276	0.001
IDENTITY	BRITISH	VANCOUVER	0.104	0.325	0.001
IDENTITY	ITALIAN	TORONTO	0.412	0.463	0.060

5. CONCLUSION

The results for tests on the rate of multiple responses are similar to those on multiplicity, which is not surprising. When you compare the critical levels for the question on ethnic origin to the critical levels for the question on ethnic identity, it is seen that the differences between the two versions of the questionnaire affect the responses to the question on ethnic identity the most.

Our main reason for using randomization tests was that the sample for Modular Test 2 was a non-probability sample. However, there are also other cases where randomization tests are appropriate. For example, to do a "Student's" *t* test for means equality the hypothesis of normality is required, and it must also be assumed that the variances are equal. These assumptions are not needed for a randomization test. It should be kept in mind that the results of a randomization test apply to the sample, and not necessarily to the entire population, unless a simple random sample is used.

REFERENCES

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

BRADLEY, J.V. (1968). *Distribution-free Statistical Tests*. Englewood Cliffs: Prentice-Hall.

DWASS, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28, 181-187.

EDGINGTON, E.S. (1987). *Randomization Tests*, (2nd ed.). New York: Marcel Dekker.

FISHER, R.A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Survey Methodology, June 1990
Vol. 16, No. 1, pp. 153–163
Statistics Canada

Variance Formulae for Composite Estimators in Rotation Designs

PATRICK J. CANTWELL[1]

ABSTRACT

In many government surveys, respondents are interviewed a set number of times during the life of the survey, a practice referred to as a rotation design or repeated sampling. Often composite estimation – where data from the current and earlier periods of time are combined – is used to measure the level of a characteristic of interest. As other authors have observed, composite estimation can be used in a rotation design to decrease the variance of estimators of change in level. In this paper, simple expressions are derived for the variance of a general class of composite estimators for level, change in level, and average level over time. Considered first are "one-level" rotation designs, where only the current month is referenced in the interview. Results are developed for any sampling pattern of m interviews over a period of M months. Subsequently, "multi-level" plans are addressed. In each month one of p different groups is interviewed. Respondents then answer questions referring to the previous p months. Results from the several sections apply to a wide range of government surveys.

KEY WORDS: Repeated sampling in surveys; Balanced designs; Month-to-month change; Yearly average.

1. INTRODUCTION

Rotation designs of various types are used in many major household surveys. The Current Population Survey (CPS) is conducted by the U.S. Bureau of the Census for the U.S. Bureau of Labor Statistics. Statistics Canada operates the Labour Force Survey (LFS). Both surveys yield estimates of labor force characteristics, including unemployment. In each survey, households are interviewed a number of times before leaving the sample. In the CPS, each household is "rotated in" for interviews in four consecutive months, rotated out of the sample for eight months, and finally back in for four more months. In the LFS, a participating household responds for six consecutive months and does not return.

A survey with a rotation design lies somewhere between a fixed panel survey, where participants remain in sample indefinitely, and a survey using independent samples, where respondents are interviewed once and retired from sample. The total overlap of a fixed panel from one time period to the next can minimize the variance of estimators of change when measurements are positively correlated across periods. Also, certain costs are incurred only the first time a unit is placed in sample. However, response burden on the members of a fixed panel can be excessive. Using a rotation design is an attempt to realize variance or cost reductions without overly burdening sample participants. In the CPS and the LFS, there are sample overlaps of 75% and 83%, respectively, from one month to the next. For more on these topics, see Woodruff (1963), Rao and Graham (1964), or Wolter (1979).

Some estimators used with rotation designs are composite in nature. In order to take advantage of repeated sampling, they combine rotation group estimates obtained for the current month with those from prior months into a final estimator.

[1] Patrick J. Cantwell, Mathematical Statistician, Statistical Research Division, Bureau of the Census, Washington, DC 20233, USA. This paper reports research undertaken by a member of the Census Bureau's staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

While the variance of composite estimators can be decreased by selecting the combination wisely, calculating this variance may become more complex because of the correlation patterns involved among the repeated groups. For general rotation plans, subject to specific restrictions, simple formulae are presented in this paper for the variance of estimators of level and change. The derivations are applied to an important and quite general class of estimators called the generalized composite estimator (Breau and Ernst 1983).

These formulae can be of use if the correlations between estimates from the same rotation group one or more time periods apart can be estimated and are suffiently large to render composite estimation worthwhile. In continuing government surveys, past sample data will typically enable the estimation of these correlations. Characteristics involving household income and labor force usually exhibit moderately high correlations. For others, such as the incidence of crime, however, correlations across time periods may not be large enough to realize the benefits of composite estimation. Of the surveys mentioned in this paper, only CPS currently uses a composite estimator.

In the developments which follow, two types of surveys are treated separately. In surveys such as CPS and LFS, participants supply information only for the current month. Such surveys are called "one-level" surveys. On the other hand, the U.S. Census Bureau conducts the Survey of Income and Program Participation (SIPP) to acquire data on income level, sources of income, program participation, and other items. During each interview, respondents in the SIPP refer back to the previous four months. A different group is then interviewed the following month. The SIPP design is consequently called "multi-level." The level of a survey was used by Wolter (1979) to indicate the number of periods for which information is solicited in one interview.

Another distinction is made between these two types of surveys. Let the term "design gap" indicate a period of time between interviews which is never referenced in any interview. While the LFS contains no design gaps, CPS includes one of eight months. For the sections pertaining to one-level designs, the results and derivations apply *regardless of the pattern of interviews and design gaps*. Therefore, the formulae are relevant not only to the current design of CPS and LFS, but also to other designs under consideration.

For reasons discussed later, designs gaps are generally not a feature of multi-level rotation plans in practice. The SIPP is no exception. Accordingly, the multi-level plans addressed in this paper do not include design gaps.

One-level designs are treated in Sections 2 and 3. In Section 2, the generalized composite estimator is defined. Notation, definitions and covariance assumptions are introduced. The main results – Theorems 1 through 3 – are given in Section 3. Variances of estimators of level and change in level are stated. The formulae are determined for single time periods (such as months) and combinations (such as quarters or years). They apply to one-level designs with any pattern of interviews and design gaps. When seeking the optimal rotation plan and composite estimator, the user must determine how best to combine variance reductions/increases for the resulting estimators of level, "month-to-month" change, and average over many periods.

In Section 4, these results are extended from one-level to certain multi-level designs, which include the SIPP. Subject to minor restrictions – in particular, the exclusion of design gaps in the sampling scheme – theorems similar to those in Section 3 are stated. Because the derivations are analogous to those for one-level plans, the results are not proved.

2. ONE-LEVEL DESIGNS: NOTATION AND DEFINITIONS

Although rotation schemes can assume infinitely many forms, the discussion in Sections 2 and 3 is restricted to one type. At each period of time, a new rotation group enters the sample,

and follows the same pattern of interviews and design gaps as every preceding group. In addition, responses refer only to the current period of time, whether or not the participants were in sample in the previous period. This design is called a balanced one-level rotation plan. The design is "balanced" because the number of groups in sample at any time is equal to the total number of time periods any one group is included in the sample.

The scheme used in the LFS satisfies these restrictions. Each month a new group enters, and remains in the sample for five more months. The CPS as it currently operates follows these guidelines in a 4-8-4 plan. Before July 1953, however, CPS used an unbalanced design where five rotation groups entered, one each in consecutive months. In the sixth month, *no new group entered*. The process then continued in the same manner, with groups exiting after six months in sample.

One problem with the CPS design before 1953 is the introduction of month-in-sample bias, often referred to as rotation group bias. Of greater concern here is the changing pattern of rotation group appearances. The variance of a composite estimate depends on when each participating group appeared in sample before, and the covariance structure for identical groups in different months. If the pattern of appearances changes from month to month, the variance formula of the estimator also changes. Under a balanced design with stationary covariance structure, general derivations are possible.

Throughout this paper, the word "month" refers to the period of time in which interviews are done, partly for brevity, but also because most government surveys use the month to divide the life of the survey. However, the results in this section and the next apply to any period of time, provided the rotation plan is balanced and one-level.

Some notation and vector definitions are now introduced. Suppose that every rotation group is in sample for a total of m interviews over a period of M months. That is, it is out of sample for $M - m$ months after first entering and before exiting. The balanced design ensures that m groups are in sample during any month.

The set T_0 is defined as follows. Consider any rotation group. Let T_0 index the set of "months" when this group is *not in sample*, labeling as month one the month this group is first interviewed, and stopping at month M. Because the design is balanced, the composition of T_0 does not depend on which group is selected. Note that, if respondents are interviewed in m consecutive months, *i.e.*, there are no design gaps, then m and M are the same, and T_0 is empty.

Next, given a set of m values $w_1, \ldots, w_m$, it is possible to define the $M \times 1$ vector w as follows. Define the ith component of w to be 0 if $i \in T_0$. This step fills $M - m$ positions in w. Then the values $w_1, \ldots, w_m$ are inserted in order into the remaining m components, starting with the first. The resulting w is called a vector "in design form." For example, in a 4-8-4 rotation plan, $T_0 = \{5, 6, \ldots, 12\}$, and $w^T = (w_1, w_2, w_3, w_4, 0, 0, 0, 0, 0, 0, 0, 0, w_5, w_6, w_7, w_8)$.

It is useful to introduce the $M \times M$ matrix R as: $R_{ii} = 1$ if $i \notin T_0$, and 0 if $i \in T_0$; and $R_{ij} = 0$ if $i \neq j$. It is clear that R is a diagonal matrix where diag(R) is a set of 1's "in design form," R_{11} and R_{MM} are 1, and $\sum_{i=1}^{M} R_{ii} = m$.

Observe that, for any $M \times p$ matrix V, RV is the same as V, but with 0's across each row i such that i is in T_0. In other words, premultiplication by R "removes" (turns to 0) the rows of V indexed by T_0. If the columns of V are already in design form, then $RV = V$. Similarly, for any $p \times M$ matrix U, postmultiplication by R "removes" the columns of U which are indexed by T_0. If the rows of U are already in design form, then $UR = U$.

Let L be the $M \times M$ matrix with 1's on the subdiagonal, and 0's elsewhere. Formally, $L_{ij} = 1$, if $i - j = 1$, and 0, otherwise. For any $M \times 1$ vector written as $w^T = (w_1, \ldots, w_M)$, the product Lw becomes $(0, w_1, w_2, \ldots, w_{M-1})^T$, and $w^T L$ is $(w_2, w_3, \ldots, w_M, 0)$.

Turning to the data, let $x_{h,i}$ denote the estimate of "monthly" level for some characteristic to be measured from the rotation group which is in sample for the ith time in month h, where $i = 1, \ldots, m$. Breau and Ernst (1983) defined the generalized composite estimator (GCE) of level recursively as follows. For monthly level, let:

$$y_h = \sum_{i=1}^{m} a_i x_{h,i} - k \sum_{i=1}^{m} b_i x_{h-1,i} + k y_{h-1}, \tag{1}$$

where $0 \leq k < 1$, and the a_i's and b_i's may take any values, including negative ones, subject to $\sum_{i=1}^{m} a_i = 1$ and $\sum_{i=1}^{m} b_i = 1$. The "current composite" and AK composite estimators used in CPS are special cases of the GCE. For information on these, see Hanson (1978), Huang and Ernst (1981), and Kumar and Lee (1983).

The GCE is more restrictive than a general linear estimator which combines $x_{h,i}$ values from the current period with those from many prior months (see Gurney and Daly 1965). However, the GCE has been shown to perform almost as well (Breau and Ernst 1983). It has the advantage that only data from two months – the current month and the preceding one – need to be stored. Although y_h incorporates earlier data, it is summarized through y_{h-1}.

To facilitate variance computations, (1) is expressed in vector form. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be $M \times 1$ vectors in design form comprising, respectively, the sets of constants $a_1, \ldots, a_m$ and $b_1, \ldots, b_m$. Similarly, for any h, the observations $x_{h,1}, \ldots, x_{h,m}$ make up $\boldsymbol{x}_h$, also an $M \times 1$ vector in design form. Then

$$y_h = \boldsymbol{a}^T \boldsymbol{x}_h - k \boldsymbol{b}^T \boldsymbol{x}_{h-1} + k y_{h-1}. \tag{1a}$$

The data are assumed to exhibit a stationary covariance structure:

(i) $\mathrm{Var}(x_{h,i}) = \sigma^2$ for all h and i;

(ii) $\mathrm{Cov}(x_{h,i}, x_{h,j}) = 0$ for $i \neq j$, *i.e.*, different rotation groups in the same month are uncorrelated; and

(iii) $\mathrm{Cov}(x_{h,i}, x_{s,j}) = \rho_{|h-s|}\sigma^2$, if the two x's refer to the same rotation group $|\ h - s\ |$ months apart; or 0, otherwise. Take ρ_0 to be 1. (2)

From the first two parts of (2), it is clear that $\mathrm{Var}(\boldsymbol{x}_h) = \sigma^2 \boldsymbol{R}$, for all h. Part three implies that $\mathrm{Cov}(\boldsymbol{x}_h, \boldsymbol{x}_{h-1}) = \sigma^2 \rho_1 \boldsymbol{R}\boldsymbol{L}\boldsymbol{R}$. This follows because (a) the matrix $\boldsymbol{L}$, with 1's on the subdiagonal, "represents" the one month lag between the $\boldsymbol{x}_h$ and $\boldsymbol{x}_{h-1}$ values, and (b) premultiplying (postmultiplying) by $\boldsymbol{R}$ inserts 0's corresponding to 0's in $\boldsymbol{x}_h$ ($\boldsymbol{x}_{h-1}$) (months not in sample).

It is readily seen that $(\boldsymbol{L}^r)_{ij} = 1$ if $i - j = r \geq 0$ and $1 \leq j, i \leq M$; take $\boldsymbol{L}^0$ to be the identity matrix. The same development as above gives $\mathrm{Cov}(\boldsymbol{x}_h, \boldsymbol{x}_{h-2}) = \sigma^2 \rho_2 \boldsymbol{R}\boldsymbol{L}^2\boldsymbol{R}$. In general,

$$\mathrm{Cov}(\boldsymbol{x}_h, \boldsymbol{x}_{h-r}) = \sigma^2 \rho_r \boldsymbol{R}\boldsymbol{L}^r\boldsymbol{R}, \text{ for } r = 0, 1, 2, \ldots, \text{ and all } h. \tag{3}$$

For $r \geq M$, $\boldsymbol{L}^r = 0$, and $\mathrm{Cov}(\boldsymbol{x}_h, \boldsymbol{x}_{h-r}) = 0$.

For the theorems which follow, define the $M \times M$ matrix $\boldsymbol{Q}$ by: $\boldsymbol{Q}_{ij} = k^{i-j}\rho_{i-j}$, if $1 \leq j < i \leq M$, and 0, otherwise. Finally, let $\boldsymbol{I}$ be the $M \times M$ identity matrix.

3. ONE-LEVEL DESIGNS: THEOREMS AND PROOFS

Three theorems are now stated and proved.

Theorem 1. If the GCE of level is defined as in (1), and the covariance structure as expressed in (2) holds, then

$$\text{Var}(y_h) = \sigma^2\{a^Ta + k^2b^T(b - 2a) + 2(a - k^2b)^TQ(a - b)\}/(1 - k^2). \quad (4)$$

Notice that when one uses an unweighted average of the estimates from the m rotation groups of the current month, $k = 0$, $Q = 0$, and $a_i = 1/m$, for $i = 1, \ldots, m$. Then $\text{Var}(y_h) = \sigma^2/m$, as expected.

Proof of theorem 1. Substitution into (1a) recursively leads to

$$y_h = a^Tx_h + (a - b)^T \sum_{i=1}^{\infty} k^i x_{h-i}. \quad (5)$$

From (3), the variance of this sum is

$$\begin{aligned}
\text{Var}(y_h) &= a^T\sigma^2Ra + (a - b)^T \sum_{i=1}^{\infty} k^{2i}\sigma^2R(a - b) \\
&\quad + 2a^T \sum_{i=1}^{\infty} k^i\sigma^2\rho_iRL^iR(a - b) \\
&\quad + 2(a - b)^T \sum_{i=1}^{\infty} \sum_{j=i+1}^{\infty} k^{i+j}\sigma^2\rho_{j-i}RL^{j-i}R(a - b) \\
&= \sigma^2\Big\{a^TRa + (a - b)^TR(a - b)k^2/(1 - k^2) \\
&\quad + 2a^TR\Big(\sum_{i=1}^{\infty} k^i\rho_iL^i\Big)R(a - b) \\
&\quad + 2(a - b)^TR\Big(\sum_{i=1}^{\infty} k^{2i}\Big[\sum_{j=i+1}^{\infty} k^{j-i}\rho_{j-i}L^{j-i}\Big]\Big)R(a - b)\Big\}. \quad (6)
\end{aligned}$$

Because a and $a - b$ are vectors in design form, $a^TR = a^T$, $(a - b)^TR = (a - b)^T$, and $R(a - b) = (a - b)$. The sum $\sum_{i=1}^{\infty} k^i\rho_iL^i$ is seen to be the matrix Q: its ijth entry is $k^{i-j}\rho_{i-j}$, if $1 \le j < i \le M$, and 0, otherwise. A change of variables will show that the sum in brackets is also Q. Expression (6) can be rewritten as:

$$\begin{aligned}
&\sigma^2\{a^Ta + (a - b)^T(a - b)k^2/(1 - k^2) + 2a^TQ(a - b) \\
&\quad + 2(a - b)^TQ(a - b)k^2(1 - k^2)\}.
\end{aligned}$$

Simple rearrangement of these terms produces the result in (4).

Theorem 2. Let $y_h - y_{h-1}$ be the GCE estimator of "month-to-month" change. Then $\mathrm{Var}(y_h - y_{h-1})$ is

(i) $2\sigma^2 a^T(I - \rho_1 L)a$, if $k = 0$, and

(ii) $\sigma^2(a^Ta + k^2b^Tb - 2k\rho_1 a^TLb)/k - (1 - k)^2\mathrm{Var}(y_h)/k$, if $0 < k < 1$.

Proof of theorem 2:

(i) If $k = 0$, $y_h = a^Tx_h$. From (3), the variance of $a^Tx_h - a^Tx_{h-1}$ is

$$2a^T\sigma^2Ra - 2a^T\sigma^2\rho_1 RLRa = 2\sigma^2 a^T(I - \rho_1 L)a.$$

(ii) If $0 < k < 1$, define W_h as $a^Tx_h - kb^Tx_{h-1}$. From prior results, it is quickly seen that

$$\mathrm{Var}(W_h) = \sigma^2\{a^Ta + k^2b^Tb - 2k\rho_1 a^TLb\}. \tag{7}$$

From (1a), $y_h = W_h + ky_{h-1}$. Then

$$\mathrm{Var}(y_h) = \mathrm{Var}(W_h) + k^2\mathrm{Var}(y_{h-1}) + 2k\mathrm{Cov}(W_h, y_{h-1}); \tag{8}$$

the covariance term can be isolated for later use. Finally, $y_h - y_{h-1} = W_h - (1 - k)y_{h-1}$. When computing the variance of this difference, substitution from (8) and (7) produces the desired result.

Often of primary importance are the average level over a certain length of time (*e.g.*, a quarter or a year), the difference in these averages from one "year" to the next, or the difference in "monthly" level for two months a year apart. Denote by $S_{h,t}$ the sum of the GCE's for the last t months:

$$S_{h,t} = y_h + y_{h-1} + \ldots + y_{h-t+1}, \quad t \geq 1. \tag{9}$$

Commonly used values of t include three, four and twelve. It is left to the reader to divide $S_{h,t}$ by t if an average desired rather than a sum.

Theorem 3:

(a) The expressions $S_{h,t}$, $S_{h,t} - S_{h-t,t}$, and $y_h - y_{h-t}$ can be written as $\sum_{i=0}^{\infty} v_i^T x_{h-1}$, where

(i) for $S_{h,t}$, $v_i =$:

$$a + [(k - k^{i+1})/(1 - k)](a - b), \quad \text{for} \quad i = 0, 1, \ldots, t - 1,$$
$$[k^{i-t}(k - k^{t+1})/(1 - k)](a - b), \quad \text{for} \quad i = t, t + 1, t + 2, \ldots;$$

(ii) for $S_{h,t} - S_{h-t,t}$, $v_i =$:

$$a + [(k - k^{i+1})/(1 - k)](a - b), \quad \text{for} \quad i = 0, 1, \ldots, t - 1,$$
$$[(2k^{i-t+1} - k - k^{i+1})/(1 - k)](a - b) - a, \quad \text{for} \quad i = t, t + 1, \ldots, 2t - 1,$$
$$-[k^{i-2t+1}(1 - k^t)^2/(1 - k)](a - b), \quad \text{for} \quad i = 2t, 2t + 1, \ldots;$$

(iii) for $y_h - y_{h-t}$, $v_0 = a$, $v_t = k^t(a - b) - a$, and $v_i =$:

$$k^i(a - b), \quad \text{for} \quad i = 1, 2, \ldots, t - 1,$$
$$-k^{i-t}(1 - k^t)(a - b), \quad \text{for} \quad i = t + 1, t + 2, \ldots; \tag{10}$$

(b) For the sets of vectors $v_0, v_1, v_2, \ldots$ defined in (a),

$$\text{Var}\left(\sum_{i=0}^{\infty} v_i^T x_{h-i} \right) = \sigma^2 \left\{ \sum_{i=0}^{\infty} v_i^T v_i + 2 \sum_{i=0}^{\infty} v_i^T \sum_{n=1}^{M-1} \rho_n L^n v_{i+n} \right\}; \tag{11}$$

the sums in (11) converge.

Proof of theorem 3. For (a), successive inclusion of terms y_h through y_{h-t+1}, and the application of (5) to y_{h-t} yield

$$\begin{aligned} S_{h,t} = {} & a^T(x_h + x_{h-1} + \ldots + x_{h-t+1}) + k(a-b)^T x_{h-1} \\ & + (k + k^2)(a-b)^T x_{h-2} + \ldots \\ & + (k + k^2 + \ldots + k^{t-1})(a-b)^T x_{h-t+1} \\ & + (k + k^2 + \ldots + k^t)(a-b)^T \sum_{j=t}^{\infty} k^{j-t} x_{h-j}. \end{aligned} \tag{12}$$

The three sets of v_i's are then determined from (12) and (5).

The proof of (b) is similar to that of Theorem 1, once it is seen that the v_i's defined in (a), being linear combinationsi of a and $a - b$, are in design form. To prove convergence, note that, for all three sets of v_i's in (a), v_i is proportional to $k^i(a-b)$ for i sufficiently large. There exists a constant $\lambda > 0$ such that, for $i \geq 2t$ and each component j, $| v_{ij} | \leq k^i \lambda$. Recalling that $| \rho_i | \leq 1$, and that each row of L^n has at most one nonzero element (equal to 1), the finite sum in (11) is seen to be an $M \times 1$ vector, each of whose components is bounded above in absolute value by $k^i(M-1)\lambda$. Convergence of the double summation then follows geometrically in k^{2i}.

4. EXTENSION TO MULTI-LEVEL DESIGNS

Although the results developed in Sections 2 and 3 apply to all balanced one-level rotation plans, it was observed that many surveys operate under multi-level designs. For example, in the Survey of Income and Program Participation (SIPP), one of four rotation groups is interviewed each month, and respondents supply information about the previous four months. Although the design is always subject to change, the first rotation group is interviewed in February, June, October, February, *etc.*, for a total of eight interviews. A second group is interviewed in March, July, *etc.* The remaining two groups follow the same sampling pattern, beginning in April and May. A SIPP panel is the set of four concurrent rotation groups covering about two and one-half years. Each year, a new panel is introduced. For example, the 1986 panel ran from 1986 through 1988, while the 1987 panel spanned 1987-89. Data from different panels are not combined, even though they may cover a common year or two. For further details on the SIPP design, see Nelson, McMillen and Kasprzyk (1984).

When one-level designs were addressed, a rotation group was allowed to assume any pattern of interviews and design gaps - intermediate months which are never referenced - provided the design was balanced. In a multi-level plan, however, design gaps can create problems with recall. Looking back several months, a respondent may find it difficult to assign an event to

the correct period of time. Design gaps can only add to the confusion. For this reason, and because multi-level surveys which incorporate design gaps are rare in practice, this section considers only designs where (i) the sample comprises p rotation groups, (ii) groups are interviewed every pth "month" in an alternating sequence, and (iii) the period of reference is the previous p months.

Many multi-level surveys, for example, the National Crime Survey, sponsored by the U.S. Bureau of Justice Statistics, have a more intricate rotational pattern than that covered here. As expected, variance formulae applied to composite estimators would tend to be more complex.

The interview of a rotation group will refer to the collective gathering of information in the assigned month from all sample units in that group. For a particular characteristic which is to be estimated, let $x_{h,i}$ denote the estimate of "monthly" level for month h from the group which is interviewed in month $h + i$, where $i = 1, \ldots, p$. The index i measures recall time – the amount of time between the month of reference and the interview. Table 1 depicts the estimates $x_{h,i}$ for a four-group four-level design. In the diagram solid lines separate estimates which are obtained in different interviews. These boundaries between the reference periods of consecutive interviews are called "seams" in the SIPP.

Table 1
Layout of Estimates in a Longitudinal 4-Level Design

MONTH ↓	ROTATION GROUPS → 1	2	3	4
1	$x_{1,4}$			
2	$x_{2,3}$	$x_{2,4}$		
3	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	
4	$x_{4,1}$	$x_{4,2}$	$x_{4,3}$	$x_{4,4}$
5	$x_{5,4}$	$x_{5,1}$	$x_{5,2}$	$x_{5,3}$
6	$x_{6,3}$	$x_{6,4}$	$x_{6,1}$	$x_{6,2}$
7	$x_{7,2}$	$x_{7,3}$	$x_{7,4}$	$x_{7,1}$
8	$x_{8,1}$	$x_{8,2}$	$x_{8,3}$	$x_{8,4}$
9	$x_{9,4}$	$x_{9,1}$	$x_{9,2}$	$x_{9,3}$
10	$x_{10,3}$	$x_{10,4}$	$x_{10,1}$	$x_{10,2}$
11	$x_{11,2}$	$x_{11,3}$	$x_{11,4}$	$x_{11,1}$
12	$x_{12,1}$	$x_{12,2}$	$x_{12,3}$	$x_{12,4}$
13	$x_{13,4}$	$x_{13,1}$	$x_{13,2}$	$x_{13,3}$
14	$x_{14,3}$	$x_{14,4}$	$x_{14,1}$	$x_{14,2}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Note: $x_{h,i}$ denotes the estimate of "monthly" level for month h from the group which is interviewed in month $h + i$. Interviewing begins in month 5. Solid horizontal lines (seams) separate estimates which are obtained in different interviews.

Let the vector x_h, defined as $(x_{h,1}, x_{h,2}, \ldots, x_{h,p})^T$, comprise the p estimates for month h obtained from the p groups *in different interviews*. Note that $x_{h,p}, x_{h+1,p-1}, \ldots, x_{h+p-1,1}$ are estimates for p different months obtained *from one group in a single interview* (in month $h + p$).

As in Sections 2 and 3, the generalized composite estimator for monthly level is defined as

$$y_h = \sum a_i x_{h,i} - k \sum b_i x_{h-1,i} + k y_{h-1}, \tag{13}$$

where the summations now range from 1 to p. Defining a and b as $(a_1, \ldots, a_p)^T$ and $(b_1, \ldots, b_p)^T$, respectively, the GCE can again be written as

$$y_h = a^T x_h - k b^T x_{h-1} + k y_{h-1}.$$

The covariance structure of the monthly rotation group estimates is assumed to be stationary in time. Under this multi-level design, however, the length of time between the target month h and the corresponding interview in month $h + i$ may affect the variability of the response, $x_{h,i}$. For $i = 1, \ldots, p$, let d_i^2 represent the response variability as a function of the amount of time between the reference month and the interview. The following covariance structure is postulated:

(i) $\mathrm{Var}(x_{h,i}) = d_i^2 \sigma^2$ for all h and i, where $d_i > 0$;

(ii) $\mathrm{Cov}(x_{h,i}, x_{h,j}) = 0$ for $i \neq j$; and

(iii) For $r \geq 0$: $\mathrm{Cov}(x_{h,i}, x_{h-r,j}) = \rho_{r,i} d_i d_j \sigma^2$, if the two x's refer to the same group r months apart; or 0, otherwise. Take $\rho_{0,i}$ to be 1 for all i. (14)

It may well be that $d_1 \leq d_2 \leq \ldots \leq d_p$, if response variability increases with recall time. The subscript r in the correlation coefficient $\rho_{r,i}$ is the amount of time between the months referenced by estimates $x_{h,i}$ and $x_{h-r,j}$. The subscript i indicates that the estimate for month h is obtained from an interview i months later. For specified values of h, r and i, there is only one value j, $1 \leq j \leq p$, for which the estimates $x_{h,i}$ and $x_{h-r,j}$ refer to the same panel and $\mathrm{Cov}(x_{h,i}, x_{h-r,j})$ is nonzero. (This value is $j = \mathrm{mod}_p(i + r - 1) + 1$, where $\mathrm{mod}_p(n)$ is the value of the integer n, modulo p.) Otherwise, the covariance is 0. In some cases, it may be appropriate to replace $\rho_{r,1}, \ldots, \rho_{r,p}$ with a commom ρ_r.

No assumptions are made about bias. In addition to the effect of recall on variances of group estimates as postulated in (14), a bias related to recall time might also be incurred. Another source - time-in-sample bias - can result according to the number of times a respondent has been interviewed (Bailar 1975). Although these biases need not be measured to derive the variance formulae given in this section, they might constitute a nontrivial component of mean squared error.

Define the $p \times p$ matrices D, P_r and J as follows. Let D and P_r, for $r \geq 0$, be diagonal matrices with $d_1, \ldots, d_p$ and $\rho_{r,1}, \ldots, \rho_{r,p}$, respectively, along the diagonal. Define J as: $J_{i,i+1} = 1$ for $i = 1, 2, \ldots, p - 1$; $J_{p1} = 1$; and $J_{ij} = 0$, otherwise. The powers of J form a cycle with $J^p = I$, where I is the $p \times p$ identity matrix. An argument similar to that in Section 2 leads to $\mathrm{Var}(x_h) = \sigma^2 D^2$ for all h, and, in general, $\mathrm{Cov}(x_h, x_{h-r}) = \sigma^2 D P_r J^r D$, for $r = 0, 1, 2, \ldots$, and all h.

Finally, define the matrix Z as $\sum_{n=1}^{\infty} k^n P_n J^n$. For general p, i, and j, it can be shown that the ijth cell Z_{ij} is an infinite sum of terms:

$$Z_{ij} = \sum_{m=0}^{\infty} k^u \rho_{u,i}, \quad \text{where} \quad u = pm + 1 + \text{mod}_p(p - i + j - 1).$$

Because the ρ values represent correlation coefficients, it follows easily that Z is finite.

Analogous to theorems 1, 2, and 3 proven earlier are theorems 4, 5, and 6 presented below. The former three allow any pattern of design gaps, but apply only to one-level designs. Theorems 4, 5, and 6 do not permit designs gaps.

The proofs of the theorems are similar to those in Section 3 and are not repeated. All results apply to the limiting case where rotation groups have been in sample long enough to eliminate the effect of phasing in the sample. If the $\rho_{r,i}$'s decrease rapidly with r, or if k is relatively small, the "steady-state" arrives within a couple of interviews.

Theorem 4. If the GCE of level is defined as in (13), and the covariance structure of (14) holds, then

$$\text{Var}(y_h) = \sigma^2\{a^T D^2 a + k^2 b^T D^2 (b - 2a) + 2(a - k^2 b)^T DZD(a - b)\}/(1 - k^2).$$

Theorem 5. Let $y_h - y_{h-1}$ be the GCE estimator of "month-to-month" change. Then $\text{Var}(y_h - y_{h-1})$ is

(i) $2\sigma^2 a^T D(I - P_1 J)Da$, if $k = 0$, and

(ii) $\sigma^2(a^T D^2 a + k^2 b^T D^2 b - 2ka^T DP_1 JDb)/k - (1 - k)^2 \text{Var}(y_h)/k$, if $0 < k < 1$.

Theorem 6. Define $S_{h,t}$ as in (9), the sum of the GCE's for the last t periods. Then $S_{h,t}$, $S_{h,t} - S_{h-t,t}$, and $y_h - y_{h-t}$ can again be written as $\sum_{i=0}^{\infty} v_i^T x_{h-i}$, where the vectors v_0, v_1, v_2, ... are found in (10). For these sets of vectors,

$$\text{Var}\left(\sum_{i=0}^{\infty} v_i^T x_{h-i}\right) = \sigma^2\left\{\sum_{i=0}^{\infty} v_i^T D^2 v_i + 2\sum_{i=0}^{\infty} v_i^T \sum_{n=1}^{\infty} DP_n J^n D\, v_{i+n}\right\}; \tag{16}$$

the sums in (16) converge.

ACKNOWLEDGMENTS

The author wishes to thank Lynn Weidman and Larry Ernst for checking the text and proofs. Lynn graciously read through several drafts, and offered many helpful suggestions to improve the presentation of the paper. The comments and suggestions of an associate editor and three referees are also noted and greatly appreciated.

REFERENCES

BAILAR, B. A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

BREAU, P., and ERNST, L. R. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

GURNEY, M., and DALY, J. F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 242-257.

HANSON, R. H. (1978). The Current Population Survey: Design and Methodology. Technical Paper 40, U.S. Bureau of the Census, Washington, D.C.

HUANG, E. T., and ERNST, L. R. (1981). Comparison of an alternative estimator to the current composite estimator in CPS. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 303-308.

KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 403-408.

NELSON, D., McMILLEN, D., and KASPRZYK, D. (1984). An Overview of the Survey of Income and Program Participation. SIPP Working Paper Series, No. 8401, U.S. Bureau of the Census, Washington, D.C.

RAO, J. N. K., and GRAHAM, J. E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

WOLTER, K. M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

WOODRUFF, R. S. (1963). The use of rotating samples in the Census Bureau's monthly surveys. *Journal of the American Statistical Association*, 58, 454-467.

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.