Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 10, NUMBER 2
DECEMBER 1984

Canada

Statistics Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

December 1984

Published under the authority of
the Minister of Supply and
Services Canada

April 1985
8-3200-501

Price: Canada, $10.00, $20.00 a year
Other Countries, $11.50, $23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 10, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 10, Number 2, December 1984

CONTENTS

SURVEY METHODOLOGY

A Journal of Statistics Canada

EDITORIAL POLICY

The Survey Methodology Journal will publish 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, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, smoothing and extrapolation methods, demographic studies, data integration and analysis and related computer systems development and applications. The emphasis will be on the development and evaluation of specific methodologies as applied to actual 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, Census and Household 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.

To Mr. Paul Francis Timmons

in recognition of his contributions as a founding member

of the Editorial Board until his retirement.

Survey Methodology, December 1984
Vol. 10, No. 2, pp. 127–140
Statistics Canada

Post '81 Censal Redesign of the Canadian[1] Labour Force Survey

M.P. SINGH, J.D. DREW, and G.H. CHOUDHRY[1]

ABSTRACT

Following each decennial population census, the Canadian Labour Force Survey (CLFS) has undergone a sample redesign to reflect changes in population characteristics and to respond to changes in information needs. The current redesign program which culminated with introduction of a new sample at the beginning of 1985 included extensive research into improved sample design, data collection and estimation methodologies, highlights of which are described.

KEY WORDS: Continuous survey; Multi-stage sample design; Stratification; Sample reallocation; Telephone interviewing; Raking ratio estimation.

1. INTRODUCTION

The Canadian Labour Force Survey (LFS), the largest monthly household survey conducted by Statistics Canada, has been redesigned in the past following each decennial census. As a part of 1981 post censal redesign, an intensive program of research as outlined in an earlier paper (Singh and Drew 1981a) was undertaken in the areas of sampling, estimation and data collection methodologies. As the reliability of labour market data at the national and provincial levels was of sufficiently high standard, the major emphases in this redesign were on improving the reliability of subprovincial data and on making the survey more cost efficient. Towards the latter, the main thrusts were on increased automation of various steps involved in sampling, greater use of Census data in place of independently obtained information in updating the sample, and increased telephone interviewing as a regular feature of the survey. As for the improvement in the subprovincial data, alternative sampling and estimation methods were investigated leading to refinements in the earlier methods, coupled with reallocation of the sample within provinces.

This paper presents an overview of the findings of various theoretical and empirical investigations and field tests undertaken during the redesign program. Sections 2 and 3 provide the background information on objectives and the old design, while Sections 4, 5 and 6 highlight the findings of investigations leading to changes in sampling and data collection methodologies. Section 7 deals with estimation issues, and sample reallocations are discussed in Section 8. Implications of the changes made in the redesigned sample on other associated surveys are outlined in Section 9, and finally in Section 10 the benefits from the major improvements in the redesigned sample are briefly recounted, along with some mention of future research plans.

2. OBJECTIVE SETTING

A fundamental step in the redesign of a recurring survey is the re-establishment of survey objectives. For the LFS, this involved reassessment not only of the survey's principal role as a provider of current labour market information, but also of its use as a central vehicle within Statistics Canada for conducting household surveys (Singh and Drew 1981b).

---

[1] Presented at ASA meetings, Section on Survey Research Methods, Philadelphia, August 1984.

[2] M.P. Singh, J.D. Drew, and G.H. Choudhry, Census and Household Survey Methods Division, Methodology Branch, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

At the early stages of the redesign program it was decided that while primary orientation towards the LFS should be maintained, efforts should also be made to enhance flexibility of the vehicle for general applications. In this light, several changes are being adopted that will benefit not only the LFS but its associated surveys as well. Requirements for the labour market data are noted below, while highlights of the changes resulting in the improvements for associated surveys are given in Section 9.

Objectives relating to provision of labour market data were established in consultation with the statistical focal points within each of Canada's ten provinces, and with key federal user departments, through annual Federal/Provincial Conferences on Labour Statistics and bilateral follow-ups. In general, these consultations revealed satisfaction with data reliability for provincial and national data, but a strong desire for improved subprovincial data. Specific data reliability objectives adopted for the redesigned sample are as follows:

(i) for Canada and each of the ten provinces, no reduction in current reliability for monthly estimates of level and estimates of month-to-month change in total employment and unemployment.

(ii) for the 24 Census Metropolitan Areas as defined by the 1981 Census, monthly estimates of unemployed with coefficients of variations (CV's) of 20% or less.

(iii) for 66 subprovincial Economic Regions agreed to in consultation with the provinces, monthly estimates of unemployed with CV's of 25% or less.

(iv) for cities with population of 60,000 or more in Quebec and Ontario, and 25,000 or more in the remaining provinces, quarterly estimates of unemployed with cv's of 25% or less.

Attainment of these objectives necessitated a reallocation of sample within the provinces that entailed a shift of sample from larger cities and Economic Regions to smaller ones. This, coupled with the desire to reduce the cost of the LFS as discussed in Section 8, created high expectations from the research projects for identifying more cost efficient strategies for data collection and production of statistics from the LFS. In the following sections, these issues are addressed for the two main parts of the survey namely, Self-Representing (SR) and Non Self-Representing (NSR) Areas.

3. OLD LFS DESIGN

A complete description of the old LFS design is given by Platek and Singh (1976). Salient features are noted in this section to provide a context for discussions in later sections.

The Self-Representing Units (SRUs) in the old design corresponded to those cities which, when designing the survey, were sufficiently large to yield an expected sample of 20 dwellings, the minimum thought acceptable for an interviewer. Minimum SRU sizes ranged from a population of 10,000 in the Atlantic Region to 25,000 in Quebec and Ontario.

Within large SRUs, deep geographic stratification was carried out by grouping together contiguous Census Tracts — geostatistical areas with populations in the 3,000 - 5,000 range, whose stability from one Census to the next makes them a convenient operational unit — without any regard to optimality of characteristics. Primary sampling units, called clusters, corresponding approximately to city blocks, were delineated on the basis of field counts obtained in 1973. A two stage sample of clusters and dwellings was selected following the Rao, Hartley and Cochran (1962) pps random group method. In addition to the area frame, an open-ended list frame of apartments was maintained in the larger cities.

A major advantage of the selection method for the area frame lies in its flexibility for changes in the sample size (Singh and Drew 1977), and for sample updating (Platek and Singh 1977, Drew, Choudhry, and Gray 1978). Sample updating is necessary in SRUs because over time the design counts used in the pps selection become out of date, leading to higher sampling variances for survey estimates. Since sampling is done independently in each random group, a Keyfitz (1951) sample update can be carried out, under which revised probabilities of selection based on recent dwelling counts can be incorporated, while maximizing retention of already selected units and avoiding overlap of selected dwellings between the pre- and post-update-samples. Regular updating of high growth SR areas occured from 1978 until the beginning of the redesign period in 1982, during which time almost half of the frame was updated. While the intensity of updating was not sufficient to reduce survey design effects to levels experienced during the first 4 years of the survey, it was sufficient to arrest further deterioration which had been averaging 7 - 8% per year for unemployed.

The Non-Self Representing (NSR) units are the areas outside the SRUs comprised of rural areas and smaller urban centers. In NSR areas, stratification based on industry classifications was carried out within Economic Regions, subject to the restriction that strata should be contiguous land areas. Within strata, Primary Sampling Units (PSUs) were delineated such that each PSU represented its stratum to the extent possible with respect to the ratio of rural to urban population and important LF characteristics. While the rural portions of a PSU were comprised of collections of contiguous rural EAs, urban portions could not always be made contiguous to the rural due to the restrictions placed on maintaining the rural to urban population ratio. Frequently larger urban centres had to be shared amongst several PSUs within the stratum.

At the time of the 1973 Redesign, two PSUs were selected per stratum using the randomized pps systematic method (Hartley and Rao 1962). In 1977, the LFS sample size was increased from 33,000 to 55,000 households per month, with the additional sample being allocated so as to improve data reliability at the province level. Thus the smaller provinces received larger proportionate sample size increases. The increase was achieved in NSR areas by selecting 1 - 4 additional PSUs per stratum (Gray 1975).

Within selected PSUs, urban and rural portions were sampled independently. In the urban portion of the selected PSU a two stage sample of clusters and dwellings was selected, whereas in the rural portion of the PSU a three stage design was used with secondaries (which are Census EAs or combinations), clusters (identifiable land areas having up to 20 dwellings) and dwellings as the sampling units. Except for the last stage, randomized pps systematic sampling was used in the selection.

4. REDESIGN OF THE SELF REPRESENTING AREAS

The criterion for cities to qualify as SRUs in the new design was raised to a minimum sample of 50 dwellings, since analysis of cost data indicated significantly higher unit costs for smaller SR assignments. The composition of the SR universe remained largely unaffected, however, due to the off-setting influences of the increase in the LFS sample size from 33,000 to 55,000 households during the late 1970's, and due to the reallocation of the redesigned sample.

For reasons noted earlier, the basic design in the SR areas remained the same and the main thrust for these areas was to update the size measures without resorting to a costly independent field count as used in the last redesign. In order to exploit the data collected during the 1981 census for this purpose different approaches were used in the block-faced cities (larger cities where data were available at block face level) and non-block-faced cities. The choices of updating method and sampling units by the two approaches are given below, whereas the two level stratification adopted for SR areas is given in Section 5.

Block-Faced Cities

The availability of Census data at the block face level in the built-up portions of the larger cities was the key factor in deciding to completely redesign the sample in these cities, which account for ⅔ of the SR frame. The redesign of these cities also provided the opportunity to introduce improved stratification as described in the next section.

For block-face portions of the cities, Census blocks were adopted as clusters (i.e., PSUs). Variance components under a two stage RHC random group design were studied for different choices of clusters — Census Blocks or EAs — by simulating the LFS design using 1976 Census data for the SRUs of Halifax and Saskatoon (Choudhry, Drew, Lee 1984). Study results, for the case of up-to-date size measures, showed little difference between the EA and block in terms of sampling variances. Hence, the decision in favour of blocks was made on the basis of operational considerations. The blocks provided a ready made frame (with splitting or combining needed in only 5 – 10% of cases), permitting a highly automated design with very low redesign costs. Importantly as well, data for future Censuses will be retrievable for the geostatistically stable blocks (but not for EAs which as operational units change at each Census), permitting low cost quinquennial updating of the sample. The built-up portions comprise 86% of these cities.

The EA was adopted in the non-block faced portions of the cities. Whereas the study results considered only the up-to-date case, it was felt that the EA being a larger unit than the block would be more robust to the highly clustered growth which can occur in the non-built-up portions of cities. Also adoption of the EA in these areas permitted very low redesign costs, since roughly 80% of these areas fell in Quebec and Ontario where due to the low sampling rates very little splitting was needed.

The variance study results, combined with cost results from the Time and Cost Study (Lemaître 1983), showed variances per unit cost to be quite flat in the range of 2 – 8 selected dwellings per cluster. Hence it was decided to retain the density of 4 – 5 dwellings per cluster used in the old design in strata where clusters were blocks, but to increase the density to 6 – 8 dwellings in the case of the EAs, due to their larger sizes.

Non-Block-Faced Cities

Since over 70% of the non-block-faced SRUs were either new or had changed boundaries, and since most of those remaining had not been updated since the 1973 redesign, it was decided to completely redesign these cities. Clusters were taken as individual or combined Census blocks in the built-up portions of the cities, with the dwelling counts obtained directly from visitation records and maps completed by Census enumerators. In the non-built-up portions, EAs or split EAs were taken as clusters, with field counts sometimes being required to do the splitting.

The use of Census visitation records, while more costly than procedures followed in the block-faced areas, nevertheless resulted in significant cost savings over the procedure followed in the old design of obtaining independent field counts.

5. STRATIFICATION

5.1 Algorithm and Choice of Stratification Variables

A modified version of a non-hierarchical algorithm due to Friedman and Rubin (1967) was adopted for stratification purposes in both SR and NSR areas, on the strength of findings reported by Judkins and Singh (1981), and Kostanich, Judkins, Singh and Schautz (1981) who evaluated several stratification algorithms for use in the U.S. Bureau of the Census' Current Population Survey. New features incorporated into the algorithm were a capacity to form geographically contiguous and/or compact strata, and the option to form either homogeneous clusters (i.e., strata) or heterogeneous clusters (i.e., primary sampling units within NSR strata). A detailed description of the method, and results of empirical evaluation studies are given by Foy, Bélanger, Drew and Joncas (1984). An overview is presented below.

The algorithm starts with a random partitioning of units into a specified number of strata. An iteration consists of examining in turn each stratification unit and moving it to any stratum which will reduce a weighted multivariate within stratum sum of squares, while continuing to satisfy constraints on strata sizes. The algorithm converges at a local optimum when moving any one of the units would increase the within strata sum of squares. Based on the findings of Judkins and Singh (1981), local optima were improved upon by the use of a moderately large number of random starts (i.e., 30).

For the contiguity option, a matrix is inputted specifying for each unit, all others contiguous to it. Contiguous initial strata respecting the size constraints are then built up from units chosen as random starting points. During the optimization step, an extra condition is imposed on transfer of units that contiguity be maintained. To achieve compactness, population centroids (longitude and latitude) are added as variables in the weighted sum of squares to be minimized.

For both NSR and SR areas, a multivariate stratification has been carried out using 1981 Census data for up to 17 stratification variables. Population variables include: total employed; employment income; persons with secondary education; population 15+, 15 - 24 and 55+; and labour force in agriculture, forestry-fishing, mining, manufacturing, construction, transport, and services. Dwelling related variables include: total dwellings, dwellings rented, one person households, and two persons households.

Any industry variables accounting for less than 2% of the labour force of the area being stratified were dropped and other variables were scaled to be equally important in the optimization process. Unemployed was not included as a stratification variable due to its instability. Study findings showed that strata formed without unemployed when evaluated at the next Census were more efficient not only for other characteristics, but for unemployed as well. The inclusion of the neighborhood type variables, on the other hand, did result in improved efficiency for unemployed.

5.2 Two Level Stratification in SR Areas

In the larger SRUs with sample sizes of 300 or more households, two levels of stratification were adopted. Primary strata, with sample yields of 150 - 170 households from the area and apartment samples combined, are comprised of collections of geographically contiguous Census Tracts. As such, primary strata are designed to correspond to two interviewer assignments. Three to four non-geographic secondary areal strata each yielding six or a multiple of six sampled clusters are formed within primary strata, with Census Tracts as stratification units, and with optimization based on the 1981 Census characteristics of non-apartment dwellers.

Apartments are sampled separately, in the form of a pps systematic sample from an open-ended list, which generally comprises a single stratum for the entire SRU. Sorting of the apartments existing at the time of design by primary strata yielded an implicit geographic stratification to the apartment sample.

In the smaller block-faced SRUs which warranted neither separate apartment sample nor geographic primary strata, optimal non-geographic areal strata were formed directly. In the non-block-faced cities, with considerably less scope for stratification, simple geographic strata were opted for.

The two levels of stratification in the larger SRUs had appeal on both operational and technical grounds. The relaxing of geographic constraints over those existing in the old design permitted greater optimality to be achieved, while the retention of contiguity at a higher level will provide a suitable unit for sample updating purposes later in the decade, and will facilitate the planning of interviewer assignments. Also, in the old design, SR strata were likely to be covered entirely by a single interviewer, and hence the variance estimates did not reflect the correlated response variance component of total variance. To the extent that within strata, interviewer assignments are geographic and secondary strata are non-geographic, an interpenetration of strata and interviewer assignments will be achieved in the new design without incurring any additional data collection costs, resulting in this component being better reflected in the variance estimates.

Table 1 presents study results for two SRUs — Ottawa and Quebec City — comparing efficiencies of the geographic strata used in the old design with those of optimal two-level strata, formed using 1971 Census data. Percent reductions in the first stage variance due to stratification, calculated at the time of the 1981 Census, indicate largest improvements under the optimal stratification for income and rented dwellings. The only marginal gains for other characteristics including employed and unemployed point to the strength and robustness of the simple, but deep, geographic stratification in the old design.

Table 1

% Reduction in First Stage Variance Due to Stratification
Old vs New Methods

Variable	Stratification Method old	Stratification Method new	Variable	Stratification Method old	Stratification Method new
total employed	9.1	12.6	agriculture[1]	5.9	3.9
employment income	18.1	30.4	forestry/fishing[1]	3.1	2.4
secondary education	39.4	42.1	mining[1]	4.8	3.0
population 15+	9.2	12.6	manufacturing	23.5	23.1
population 15 - 24	12.9	17.6	construction	11.9	11.2
population 55+	25.3	29.7	transport	4.2	6.4
total dwellings	28.5	33.1	services	14.5	19.8
dwellings rented	20.9	28.8	unemployed[1]	7.1	9.7
1 person households	33.7	38.4			
2 person households	27.5	29.6			

[1] characteristics not used in optimization for new method

In the NSR areas, the same clustering algorithm was used within each Economic Region to form either rural, or mixed urban and rural strata, depending on the design adopted, as discussed in Section 6.3. Also the adaption of the clustering algorithm for use in PSU formation is described in Section 6.5.

6. DESIGN CONSIDERATIONS IN NSR AREAS

6.1 Extension of Telephone Interviewing

Telephone interviewing for months 2-6 in the sample was introduced during the early 1970's in Self Representing Areas, primarily to reduce cost. However in NSR areas, all interviewing continued to be done in person due to concern over the high instance of party lines vis-à-vis the confidentiality of the data being collected. Nevertheless, in recognition that not only immediate cost benefits from telephoning were at issue, but so also were the longer term potential benefits from the use of new technologies such as Random Digit Dialing and Computer Assisted Telephone Interviewing (CATI), it was decided to test the feasibility of extending telephone interviewing to NSR areas.

A first field test was restricted to urban areas having over 80% private lines. The test was conducted on a portion of the actual LFS sample, with the principal objective of assessing the data quality implications of telephone interviewing. To facilitate this analysis, interviewer assignments were split between the telephone and personal procedure.

This test ran from January 1982 to June 1983 with a gradual phase-in to ensure no adverse impact on the ongoing survey. Principal findings were: lower non-response rates for the telephone sample (3.4% versus 4.3% for the control sample); a high instance of households with telephones (96% for all provinces but one); a low instance (1%) of households not agreeing to telephone interviewing; and no detectable differences in estimates for labour force characteristics.

A second test carried out in the rural areas had comparable findings. Based on the positive findings from both tests, the decision was taken to introduce telephone interviewing across the board in NSR areas during the remainder of 1983 and early 1984.

The decision to extend telephone interviewing had the following principal implications on the design of the NSR sample:

(i) Increase in assignment sizes: In the old design, NSR assignment sizes averaged 50 dwellings. Evidence that per unit costs were lower for larger assignments (Lemaître 1984), and the reduction in travelling under telephoning, both supported increasing the design yield per NSR PSU to 55 - 60 dwellings.

(ii) Level of assignment of rotation numbers: Unlike the old design, in the new design, all dwellings within secondaries will receive the same rotation number, which will cut down on the number of visits to the secondaries in month 2 - 6 in the sample.

6.2 Elimination of Stage of Sampling in Rural Areas

In the old design, the rural sample within PSUs was selected in three stages: secondaries (Census Enumeration Areas), clusters, and dwellings. The clusters corresponded to identifiable land areas containing up to 20 dwellings, which were delineated on the basis of field counts obtained whenever a new secondary entered the sample. Within secondaries, generally 5 - 6 clusters, with 3 - 4 dwellings per cluster, were selected.

The rural cluster stage was identified early in the redesign program as a possible candidate for elimination, on the grounds that (i) the sample variance would be reduced due to having one less stage of samplings, and (ii) the lead time required to introduce new secondaries into the sample could be shortened from 13 months to 7 months.

A field study was carried out on a sample of secondaries entering the LFS sample, in order to assess the feasibility of maintaining good quality dwelling lists for entire rural EAs, and to examine costs under such a procedure, with positive results on both counts. The variance implications of eliminating the cluster stage were also studied. Using 1971 Census data to simulate both the old and alternative design, components of variance were obtained for the Horwitz-Thompson estimator without ratio estimation. The percent reduction in total variance under the alternative design was found to range from 20 - 25% for major labour force characteristics (Choudhry, Lee, and Drew 1984).

On the basis of these findings, an early decision was taken to eliminate the rural cluster stage of sampling, and attention was turned to more global aspects of the NSR design.

6.3 Design with Urban/Rural Stratification

The old design featured implicit urban/rural stratification. PSUs were formed to have approximately the same ratio of urban to rural population as the stratum, and within selected PSUs the urban and rural portions were sampled independently. A premise underlying the design was that the PSU should correspond to an interviewer's assignment. However, in practice this correspondence was weakened since in order to attain the desired urban/rural ratio, frequently the urban and rural portions of PSUs were not contiguous.

As an alternative to the old design, D_0, (with the rural cluster stage eliminated), a design, D_1, featuring explicit urban/rural stratification was studied. Like D_0, the alternative design D_1 consisted of 3 stages of sampling in both urban and rural areas. In urban strata, the stages were: PSUs (consisting of individual or nearby urban centers), clusters, and dwellings. In rural strata, the stages were: PSUs, (consisting of collections of nearby rural EAs), secondaries (EAs), and dwellings. Under D_1, both urban PSUs and rural PSUs were designed independently to yield samples corresponding to interviewer assignments.

The two design alternatives were evaluated, from the point of view of variance and cost (Choudhry, Drew, Lee 1984). In the variance study both designs were simulated for the case of 2 PSUs per stratum using design counts based on the 1971 Census, and study variables based on 1976 Census data.

In terms of costs, a simple model was developed for D_0, the old design under telephone interviewing, and components were estimated using results from a detailed Time and Cost Study (Lemaître 1983). Relative costs for the travel components between designs D_0 and D_1 were estimated by means of a simulation study, in which average dispersion of the sample under the two designs was obtained up to the second stage of sampling using the population centroids of the EAs.

Findings were that the design D_1 was 1.09 times as cost efficient as D_0, and that from the combined perspective of cost and variance, D_1 outperformed D_0 with overall efficiencies of 1.25 for employed and 1.05 for unemployed.

Based on these findings, design D_1 was adopted in 70% of Economic Regions with sufficient urban and rural population to yield separate strata. In the remaining Economic Regions, with the exception of Prince Edward Island, design D_0 was adopted (see Section 6.6).

6.4 Number of PSUs Selected Per Stratum

In the LFS design, since the sample yield per PSU is fixed, the number of PSUs selected per stratum also determines the number of strata. In over two thirds of cases, the urban, rural or combined strata within ERs yielded only enough sample for 2 or 3 PSUs. Further stratification in these cases was ruled out on the grounds that there should be at least 2 PSUs per stratum to permit unbiased estimation of variance.

The remaining ERs were stratified to the point of 2 or 3 PSU's per stratum. Estimated first stage variance reductions over the situation under the old design of from 3 - 6 selected PSU's per stratum were up to 14% for employed (Choudhry, Lee, and Drew 1984). The stratification was carried out using the clustering algorithm described in Section 5.

6.5 Use of Clustering Algorithm in Formation of NSR PSUs

In both the old and new LFS, stratification is carried out prior to formation of NSR Primary Sampling Units. PSUs are delineated within the stratum to be as similar as possible with respect to stratification variables, while being as geographically compact as possible. PSU delineation which was carried out using the clustering algorithm noted earlier, required minimization of geographic and maximization of the non-geographic variables.

6.6 Two Stage Design for Prince Edward Island

For Canada's smallest province, Prince Edward Island, sampling rates are 4% in order to produce monthly LF estimates with required levels of data reliability. In view of the high sampling rates, a less clustered design consisting of a two stage sample of EAs and dwellings, with deep geographic stratification was adopted. It was found to have marginally higher costs than D_0, however from the overall perspective of cost and variance, it came out well ahead with efficiencies of 2.21 and 1.11 for employed and unemployed relative to D_0 (Choudhry, Lee, and Drew 1984).

7. ESTIMATION

7.1 Final Stage Ratio Estimation

In the old LFS, a final stage ratio estimation was carried out by detailed province/age/sex cells. With the development within Statistics Canada of improved and more timely subprovincial population estimates (Verma, Basavarajappa, and Bender 1982), an intermediate ratio estimation step was studied in which survey estimates of population 15+ for subprovincial areas are ratio adjusted to external estimates prior to the usual final ratio estimation. Findings were that the procedure, while not impacting on the variances of provincial level data, resulted in variance reductions for sub-provincial areas ranging from close to 70% for employed to 7% for unemployed (Earwaker and Bélanger 1981). In practice a raking ratio procedure in which the two ratio estimation steps are iterated until both marginal controls are satisfied was adopted, beginning in 1983.

7.2 Improved Estimates for Household and Family Units

Paul and Lawes (1982) used LFS longitudinal data files, which link households over the six months in the sample, to demonstrate that non-response rates are higher amongst households with fewer members. For the old LFS, non-response adjustment consisted of re-weighting at local area levels. This was done without regard to household size, hence the resulting estimates of households and families by size had 1 - 3% biases. Another problem related to the inconsistency of family and individual based statistics (Macredie 1983). When demographic estimates of families by size, currently under development by Statistics Canada's Demography Division, become available, it is intended to incorporate them as an extra dimension in the final stage raking ratio estimation procedure, to address both problems.

As an interim measure, the use of LFS longitudinal data is being studied as a mean to derive household size distributions based on both respondents and non-respondents, prior to the final raking ratio estimation (Ghangurde 1984).

7.3 Small Area Estimation

Demand for Labour Force estimates for small areas (domains) such as Federal Electoral Districts (FEDs) and Census Divisions (CDs), both of which number over 250 units across Canada, has increased in recent years. Since it was not possible to respect the boundaries of such areas in the design of the survey, various alternative small area estimation methodologies were evaluated. A sample dependent estimator was proposed as a combination of post-stratified and synthetic estimators, which relies entirely on the post-stratified estimator whenever the sample size in the domain is sufficient according to certain criteria, and which otherwise introduces a synthetic component whose relative weight depends on the deficiency of the sample in the domain. Based on study findings, it was recommended that the sample dependent approach be developed as a means of providing annual or multi-year average estimates for areas such as FEDs and CDs (Drew, Singh, and Choudhry 1982). Implementation and further research and developmental work is proceeding under Statistics Canada's Small Area Data Program.

7.4 Variance Estimation

The methodology for variance estimation for the redesigned sample will continue to be based on Keyfitz's (1957) method, although it will be further modified to the case of a two step final stage ratio estimation, i.e., to a single iteration in the raking ratio estimation procedure. As subsequent iterations exert only a very small influence on estimates, they are being ignored in variance estimation. Some further refinements of the current variance estimation procedure are under study, such as adopting clusters as replicates in SRUs, as opposed to the current practice of grouping clusters into two psuedo-replicates.

It should be noted that variance estimators given by Rao, Hartley, and Cochran (1962), and by Rao (1975) were evaluated as alternatives to the current method in SRUs, where the RHC design is followed (Choudhry, Lee, and Sida 1984). The current method and the alternatives were studied both with and without ratio adjustment. The current method without ratio adjustment was found to overestimate the variance for certain characteristics (e.g., 20% for employed), however with ratio estimation, biases were negligible. Estimated biases were also negligible for the alternatives. The principal advantage of the alternatives was that they were more stable. The current method was retained however, due to its simplicity and also because of the complications in estimating variances of change or averages under the alternative methods.

7.5 Composite Estimation

In the LFS, moderate to high month-to-month correlations exist for most characteristics due to the 5/6'th common sample. Different composite estimators were studied by Kumar and Lee (1983), which take advantage of these correlations by use of data from previous samples to improve the current month's estimates. Their studies focussed on a class of AK composite estimators studied recently by Huang and Ernst (1981) and others in the context of the U.S. Bureau of the Census' Current Population Survey.

Findings under the assumption that the ratio estimator is unbiased, were that from the perspective of mean square error, a compromise choice of the A and K weights yielded up to 5% gains for monthly estimates of level for unemployed and employed, and from 5% - 16% gains for corresponding month-to-month change estimates. A decision on implementation of composite estimation was delayed pending further studies on the impact on composite estimators of any changes in rotation group bias, stemming from modifications in non-response adjustment and ratio estimation procedures, and pending closer examination of its operational implications.

7.6 New Rounding and Release Policy

In the old LFS estimates of level were rounded to thousands and released if greater than 4 thousand. This policy was applied uniformly in all provinces for all estimates, with the intent that released data should have a coefficient of variation of 33.3% or less.

More rigorous, provincially based rounding and release criteria were developed for the redesigned sample to satisfy the conditions that the CV of unrounded estimates should be 33.3% or less, and that the rounding error should not exceed 20% of the standard error of the unrounded estimate. Findings were that release criteria could be dropped to 2 – 3 thousand, for all provinces except Quebec and Ontario, and that estimates for subprovincial areas should de rounded to hundreds instead of thousands (Kumar 1982).

8. SAMPLE REALLOCATION

Specific data reliability objectives of the redesign having particular emphasis on better data at subprovincial level are noted in Section 2. In addition to the general improvements in the data reliability levels through the refinements in the methods and procedures, it became necessary to consider reallocation of the sample within provinces to meet the reliability levels noted in objectives (ii), (iii) and (iv). Sample size increases were needed in 13 out of 66 Economic Regions, 6 out of 24 CMAs and 27 out of 42 non-CMA cities. An average 28% reduction in the CV's for unemployed was obtained for these cases. In addition, for the 30 ERs with old CV's in the range of 15 – 25%, the reallocations achieved an average 12% reduction in CV's. As a rule of thumb, under the redesigned sample, monthly data for ERs and CMAs and quarterly data for other cities will be based on minimum monthly sample sizes of 300 and 120 households per month respectively.

It is worth noting that the redesign objectives did not directly consider two important uses of LFS data by federal government departments. These are the use of 3 month moving average unemployment rates for subprovincial Unemployment Insurance (UI) Regions in determining the regionally variable number of weeks worked to qualify for UI benefits, and the use of 3 year average data for 180 – 200 areas consisting of individual or combined Census Divisions, for use in allocating federal funds to assist new industrial initiatives. However, the re-distributions of the sample will indirectly benefit both of these applications.

In determining sample size requirements to meet the objectives, average unemployment rates for the period 1980 – 82 were used, in view of medium term forecasts for sustained high unemployment during the 80's.

A general implication of these reallocations was the movement of a significant proportion of the sample from larger CMA's and Economic Regions to smaller ones. This had an adverse impact on the provincial and national estimates due to the departure from the usual proportional allocations. This decrease in reliability at higher levels was however more than compensated by the general increase in the reliability achieved through the structural improvements in the methods and procedures as a result of research investigations.

A study was also carried out using the cost-variance model suggested by Fellegi, Platek and Gray (1967) to arrive at optimum sampling rates in the NSR and SR areas. This resulted in a shift of sample from NSR to SR areas . This was most pronounced in Quebec and Ontario where the proportion of SR sample increased from .60 to .72, as compared with .78 of the frame, yielding gains for provincial estimates of unemployed equivalent to a 5% variance reduction, assuming a fixed sample size. In addition, this optimization helped in achieving objectives (ii) and (iv), and it benefitted the Survey of Consumer Finances and Rent Survey.

It is worth noting that the structural improvements in the design achieved through factors such as improved stratification, reduction in a stage of sampling in the NSR areas, incorporation of subprovincial controls in the estimation procedures, and more refined sample allocations resulted in better than current reliability levels for national and provincial estimates while meeting the objectives for the subprovincial data. This opened up the possibility of reducing the overall sample of the LFS in order to release funds for the collection of data on certain other socio economic issues from time to time. The size of the redesigned sample was therefore fixed as 51,500 households per month down from 55,000. This overall reduction of 6 - 7% was achieved through a uniform reduction in all provinces with the exception of Prince Edward Island. In addition, per unit data collection costs will be reduced due to increased telephone interviewing.

9. IMPLICATION OF CHANGES ON LFS ASSOCIATED SURVEYS

Most of the household surveys conducted by Statistics Canada take advantage of the investment the LFS represents in terms of sample frame and design, data collection vehicle and processing systems to obtain data more quickly, at less cost and greater reliability than would be possible through independent surveys. The design and operations of these surveys are integrated with those of the LFS to varying degrees.

Most common are supplementary surveys consisting of additional questions to LFS respondent, which incur only incremental costs of the extra questions. Surveys which due to the sensitivity of the subject matter due to the length of the questionnaire, might impact on the LFS are not done as supplements. Typically such surveys have been conducted by LFS interviewers on a separate sample of households selected in the same areas as the LFS. Less closely integrated with the LFS are surveys which select different areas from the LFS, but which benefit from use of the LFS sample design and from control of overlap with the LFS sample.

As noted earlier the main orientation of the redesign program was towards the LFS, but efforts were also made to enhance flexibility of the vehicle for general applications. In this light changes being adopted for the LFS that will benefit its associated surveys are briefly highlighted below:

The sample reallocation resulting in a shift of sample from NSR to SR areas will be more robust for general applications and in particular will improve estimation of income and rent changes from the SCF and Rent Survey. Also, the adoption of a 300 household minimum sample size for CMAs will benefit these surveys for which CMA estimates are produced.

The general multi-variate stratification using 15 variables adopted (in both NSR and SR areas) will also represent an improvement for non-LFS applications over the current industry specific or simple geographic stratification.

Three changes will specifically benefit surveys using different sets of households: (i) the elimination of a stage of sampling in rural areas will considerably shorten the lead time to 7 months from 13 months in the old design, (ii) PSUs will be more geographically compact due to the adoption of explicit rural and urban strata, which will benefit smaller surveys where greater correspondence between PSUs and interviewers assignments is needed, and (iii) the flexibility introduced through the refinements in the sample stabilization program will allow selection of subsamples of virtually any size at the national, provincial or subprovincial levels for surveys using the LFS vehicle.

Finally the modification in the method of estimation introduced in the redesign, in the use of a household size distribution in the ratio adjustment as an interim measure (with eventual incorporation of demographic estimates of families by size as an additional dimension in the final staged raking ratio estimation procedure) will improve the consistency amongst family and individual related labour market statistics, and will improve family statistics on expenditure and income.

10. SUMMARY OF CHANGES AND FUTURE RESEARCH PLANS

Most of the research investigations for the post-1981 Censal redesign of the LFS have been completed, with the implementation of the new sampling design underway and certain aspects of research in estimation methodology still to continue. It is worth noting that many of the investigations carried out during this program have confirmed the soundness of the past methods and procedures used in the LFS such as those of sampling two PSUs per stratum in the NSR areas, use of the RHC method and existing density factor (4 - 5 households per cluster) in the SR areas and continuation of the six month rotation pattern. However, several investigations have as well lead to improvements both in the redesign process and the new survey design; primarily due to change in emphasis on the data reliability objectives (as noted in Section 2), availability of better and easily accessible information and technological advances.

Improvements in the redesign process included the use of 1981 Census data in place of independently obtained field counts for updating the SR sample, the reduction of the clustering operation, and automation of stratification and PSU formation. Also cost savings will result from phasing-in much of the redesigned sample in an "on-line" fashion. Under this approach, the new sample will be introduced and the old sample will be dropped one rotation group at a time over a 6 month period, as compared with the traditional method of keeping the old sample at full strength for a 3 - 4 month period while building up the new sample (Mayda, Drew and Lindeyer 1984). Process cost savings as compared with the previous redesign are estimated at $1.8 million (in 1983/84 dollars).

Principal improvements in the cost efficiency of the LFS survey design include the extension of telephone interviewing to months 2 - 6 in the sample in NSR areas, the adoption of an NSR design featuring explicit urban/rural stratification, elimination of a stage of sampling in rural areas, the general purpose stratification in both SR and NSR areas, the use of subprovincial population controls in the estimation procedure, and more refined sample allocation procedures. These improvements were sufficient to permit gains in the reliability of subprovincial data, while retaining the status quo for provincial level reliabilities, and while decreasing the overall sample size by 6 - 7%. Reliability gains averaged 14% for coefficients of variation of unemployed for the half of the Economic Regions and CMAs with poorest reliabilities under the old design, with for the most part, little or no change in remaining areas. Subprovincial gains for estimates of employed will be even greater. On the cost side, the sample size decrease, coupled with reduced costs due to the extension of telephone interviewing will result in estimated cost savings of $.7 million per year (1983/84 dollars).

Following the completion of the sample redesign a principal focus of design related research and development for the LFS in coming years will be on investigation of a dual frame methodology, underwhich a portion of the sample would be converted to a telephone frame, using Random Digit Dialing (RDD) techniques. A multi-year program of RDD testing including research into implications of higher non-response to the RDD telephone sample, of research into dual frame estimation methodologies and of study of centralization versus decentralization of telephone interviewing is currently in the planning stages, as part of a newly established telephone survey development program (Hofmann, Drew, Catlin and Mayda 1984). Another design related initiative will be aimed at developing cost efficient means of intercensally updating the area sample in SR areas.

ACKNOWLEDGEMENTS

The authors are grateful to the LFS Sample Redesign Committee for its support and guidance throughout the course of the redesign program, most notably to I.P. Fellegi, D.B. Petrie, G.J. Brackstone, R. Platek, I. Macredie, M. Levine, M. Brochu and F. Mayda. Special thanks are also due to all members of the methodology team engaged in both the research and implementation phases of the project, whose collective efforts have made possible the improvements described herein. The referee's helpful comments are also appreciated.

REFERENCES

CHOUDHRY, G.H., LEE, H. and DREW, J.D. (1984). Cost Variance Optimizations for the Canadian Labour Force Survey. Internal Technical Report (in preparation), Census and Household Survey Methods Division, Statistics Canada.

CHOUDHRY, G.H., LEE, H. and SIDA, R. (1984). Variance Estimation for the Redesigned Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

DREW, J.D., CHOUDHRY G.H. and GRAY G.B. (1978). Some Methods for Updating Sample Survey Frames and Their Effects on Estimation. *Survey Methodology,* 4, 225-263.

DREW, J.D., SINGH, M.P. and CHOUDHRY, G.H. (1982). Evaluation of Small Area Estimations Techniques for the Canadian Labour Force Survey. *Survey Methodology,* 8, 17-47.

EARWAKER, S. and BÉLANGER, Y. (1981). Ratio Estimation at the Subprovincial Level for the Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

FELLEGI, I.P., GRAY, G.B. and PLATEK, R. (1967). The New Design of the Canadian Labour Force Survey. *Journal of the American Statistical Association,* 62, 421-453.

FOY, P., BÉLANGER, Y., DREW, J.D. and JONCAS, M. (1984). Multivariate Clustering Algorithm for Stratifications and its Application to the Canadian Labour Force Survey. Technical Report (in preparation), Census and Household Survey Methods Division, Statistics Canada.

FRIEDMAN, H.P. and RUBIN, J. (1967). On some Invariant Criteria for Grouping data. *Journal of the American Statistical Association,* 62, 1159-1178.

GHANGURDE, P.D. (1984). Evaluation of LFS Non-Response Adjustment in Household Size Cells. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

GRAY, G.B. (1973). On Increasing the Sample Size (No. of PSUs). Technical Report, Census and Household Survey Methods Division, Statistics Canada.

HARTLEY, H.O. and RAO, J.N.K. (1962). Sampling with Unequal Probabilities without Replacement. *Annals of Mathematical Statistics,* 33, 350-374.

HOFMANN, H., DREW, D., CATLIN, G., and MAYDA, F. (1984). A Proposal for a Telephone Survey Development Program. Internal Statistics Canada Report.

HUANG, E. and ERNST, L. (1981). Comparison of an Alternate Estimator to the Current Composite Estimator in CPS. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 303-308.

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 Meetings,* 274-284.

KEYFITZ, N. (1951). Sampling with Probabilities Proportional to Size: Adjustment for Changes in the Probabilities. *Journal of the American Statistical Association,* 46, 105-109.

KEYFITZ, N. (1957). Estimates of Sampling Variance where two Units are Selected from each Stratum. *Journal of the American Statistical Association,* 52, 503-510.

KOSTANICH, D., JUDKIN, D., SINGH, R. and SCHANTZ, M. (1981). Modification of Friedman - Rubin's Clustering Algorithm for Use in Stratified PPS Sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings.*

KUMAR, S. (1982). Investigation of the Labour Force Survey Rounding and Release Criteria. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

KUMAR, S. and LEE, H. (1983). Evaluation of Composite Estimation for the Canadian Labour Force Survey. *Survey Methodology,* 9, 178-201.

LEMAITRE, G. (1983). Some Results from the Time and Cost Study. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

MACREDIE, I. (1983). Family Oriented Measures of Employment and Unemployment. Presented at OECD Working Party on Employment and Unemployment Statistics.

MAYDA, F., DREW, D., and LINDEYER, J. (1984). Phase-in of the Redesigned Labour Force Survey Sample. Technical Report, (in preparation), Census and Household Survey Methods Division, Statistics Canada.

PAUL, E.C., and LAWES, M. (1982). Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey. *Survey Methodology,* 8, 48-85.

PLATEK, R. and SINGH, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.

PLATEK, R. and SINGH, M.P. (1977). A Strategy for Updating Continuous Surveys. *Metrika,* 25, 1-7.

RAO, J.N.K. (1975). Unbiased Variance Estimation for Multi Stage Designs. *Sankhya,* Series C, 37, 133-139.

RAO, J.N.K., HARTLEY, H.O. and COCHRAN, W.G. (1962). On a Simple Procedure of Unequal Probability Sampling without Replacement. *Journal of the Royal Statistical Society, Series B,* 24, 482-490.

SINGH, M.P. and DREW, J.D. (1977). Sample Expansion in Self Representing Units of the Canadian Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

SINGH, M.P., and DREW, J.D. (1981a). Research Plans for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association Meetings.*

SINGH, M.P. and DREW, J.D. (1981b). Redesigning Continuous Surveys in a Changing Environment. *Survey Methodology,* 7, 44-73.

VERMA, R.B.P., BASAVARAJAPPA, K.G. and BENDER, R.K. (1983). The Regression Estimates of Population for Sub-Provincial Areas in Canada. *Survey Methodology,* 9, 219-240.

Survey Methodology, December 1984
Vol. 10, No. 2, pp. 141–156
Statistics Canada

Analysis of Categorical Data from Surveys with Complex Designs: Some Canadian Experiences[1]

D.A. Binder, M. Gratton, M.A. Hidiroglou, S. Kumar and J.N.K. Rao[2]

ABSTRACT

Goodness of fit tests, tests for independence in a two-way contingency table, log-linear models and logistic regression models are investigated in the context of samples which are obtained from complex survey designs. Suggested approximations to the null distributions are reviewed and some examples from the Canada Health Survey and Canadian Labour Force Survey are given. Software implementation for using these methods is briefly discussed.

KEYWORDS: χ^2 statistic; Wald Statistics; Goodness of fit; Independence in two-way tables; Log-linear and logistic regression model.

1. INTRODUCTION

A sketch of the historical development of modern categorical data analysis has been given in the excellent review paper by Imrey, Koch and Stokes (1981). These techniques, applied in the context of random samples derived as independent selections from a common distribution function, are not directly applicable to survey samples collected using complex survey designs.

Koch *et al* (1975), Shuster and Downing (1976), developed asymptotically valid methods, based on the Wald statistic that take the survey design into account, but requiring access to the micro-data file or at least the full estimated covariance matrix of cell estimates. Cohen (1976) and Altham (1976) proposed a simple model for clustering and showed that the generalized Wald statistic for goodness of fit is a multiple of χ^2, when the model holds. Brier (1978) considered a similar model, but studied general hypotheses on cell probabilities, and proved that a multiple of the corresponding Pearson statistic is asymptotically distributed as a χ^2 random variable, when the model holds. Fellegi (1980) deflated the χ^2 using a correction factor based on the mean of the estimated design effects. Fay (1985) developed jackknife χ^2 and G^2 statistics, also taking the design into account, but requiring the cell estimated at the primary sampling unit level. Rao and Scott (1981) developed a correction to χ^2 (or G^2) based on the Satterthwaite to approximation to the asymptotic distribution of χ^2, requiring the full estimated covariance matrix.

In this paper, we discuss the problems of fitting models and testing hypotheses with categorical data resulting from complex designs. For data collected using complex designs, some adjustments to the classical methods described by Imrey, Koch and Stokes (1981) are necessary in order to make valid inferences. If the published tables are provided along with the cell and marginal design

[1] This paper is a revised and expanded version of that presented at the Seminar on Recent Developments in the Analysis of Large Scale Data Sets sponsored by Statistical Office of European Communities, November 16-18, 1983, Luxemburg.

[2] D.A. Binder, Institutional & Agriculture Survey Methods Division, M. Gratton, EDP Planning and Support Division, M.A Hidiroglou, Business Survey Methods Division, S. Kumar, Census & Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, and J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, CANADA.

effects, some of the approximations to the null distributions of our test statistics can be obtained, without having access to the complete micro-data file. On the other hand, for applications where the complete micro-data file is available, alternative approaches will be described.

For illustrative purposes, Section 2 begins with the standard goodness of fit problem. This discussion is then extended in Section 3 to tests of independence in a two-way contingency table. This leads to a general discussion of log-linear models in Section 4. Logistic regression models are described in Section 5. In Section 6 we summarize the existing situation with respect to software development for these methods at Statistics Canada. In Section 7, we discuss the appropriateness of these methods. Numerical examples are taken primarily from the Canada Health Survey. An application from the Canadian Labour Survey is given in Section 5.

2. GOODNESS OF FIT

2.1 Multinomial Sampling

Suppose we select n independent and identically distributed observations $Y_1, \ldots, Y_n$ from a discrete distribution with k categories, where $\Pr(Y = i) = \pi_i$; $\Sigma_{i=1}^{k} \pi_i = 1$. We observe the random vector $\underset{\sim}{n} = (n_1, \ldots, n_{k-1})^T$, which has a multinomial distribution. Our estimate of $\pi = (\pi_1, \ldots, \pi_{k-1})^T$ is given by $\underset{\sim}{p} = \underset{\sim}{n}/n$. This estimate is unbiased and has covariance matrix given by $\underset{\sim}{V}\{\underset{\sim}{p}\} = (\underset{\sim}{D}_\pi - \underset{\sim}{\pi}\underset{\sim}{\pi}^T)/n = \underset{\sim}{P}/n$, where $\underset{\sim}{D}_\pi = \operatorname{diag}\{\pi_1, \ldots, \pi_{k-1}\}$. Note that $\underset{\sim}{P}^{-1} = \underset{\sim}{D}_\pi^{-1} + (\underset{\sim}{1}\underset{\sim}{1}^T/\pi_k)$. Asymptotically, $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{P})$. For a given $\underset{\sim}{\pi}_o$, the goodness of fit problem is to test the hypothesis.

$$H_o\colon \underset{\sim}{\pi} = \underset{\sim}{\pi}_o,$$

against the alternative

$$H_1\colon \underset{\sim}{\pi} \neq \underset{\sim}{\pi}_o. \tag{2.1}$$

Letting $\underset{\sim}{P}_o$ represent $\underset{\sim}{P}$ evaluated at $\underset{\sim}{\pi}_o$, the Wald statistic for this test is

$$\begin{aligned} W_1 &= n(\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \underset{\sim}{P}_o^{-1} (\underset{\sim}{p} - \underset{\sim}{\pi}_o) \\ &= n \sum_{i=1}^{K} \{(p_i - \pi_{io})^2/\pi_{io}\}, \end{aligned}$$

which is the familiar Pearson chisquare test. Under H_o this is asymptotically χ^2_{k-1}. The likelihood ratio test for this problem is given by

$$LR_1 = 2n \sum_{i=1}^{k} p_i \log(p_i/\pi_{io}).$$

Since $2p_i \log(p_i/\pi_{io})$ is asymptotically equivalent to $2(p_i - \pi_{io}) + (p_i - \pi_{io})^2/\pi_{io}$ under H_o, we see that the likelihood ratio test is asymptotically equivalent to the Pearson chisquare statistic under H_o.

Another possible test for this hypothesis is derived by defining the vector of logs, $\underset{\sim}{\mu}_o = \log \underset{\sim}{\pi}_o$ and $\underset{\sim}{\hat{\mu}} = \log \underset{\sim}{p}$. Now under the null hypotheses $\underset{\sim}{\hat{\mu}} - \underset{\sim}{\mu}_o$ is asymptotically equivalent to $\underset{\sim}{D}_{\pi o}^{-1}(\underset{\sim}{p} - \underset{\sim}{\pi}_o)$. Therefore, $n^{1/2}(\underset{\sim}{\hat{\mu}} - \underset{\sim}{\mu}_o) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{D}_{\pi_o}^{-1} - \underset{\sim}{1}\underset{\sim}{1}^T)$ under H_o and the Wald statistic is

$$\begin{aligned} W_2 &= (\underset{\sim}{\hat{\mu}} - \underset{\sim}{\mu}_o)^T \left[\underset{\sim}{D}_{\pi o} + (\underset{\sim}{\pi}_o \underset{\sim}{\pi}_o^T/\pi_{ko})\right] (\underset{\sim}{\hat{\mu}} - \underset{\sim}{\mu}_o) \\ &\doteq \sum_{i=1}^{k} \pi_{io}(\hat{\mu}_i - \mu_{io})^2, \end{aligned}$$

where $\mu_{ko} = \log \pi_{ko}$ and $\hat{\mu}_k = \log p_k$.

This approximation is obtained by noting that under H_o

$$\pi_{ko}(\hat{\mu}_k - \mu_{ko}) \doteq p_k - \pi_{ko}$$
$$= -\ (\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \underset{\sim}{1}\ -\ (\underset{\sim}{\mu} - \underset{\sim}{\mu}_o)^T \underset{\sim}{\pi}_o\ .$$

Note that W_2 is also asymptotically equivalent to the Pearson chisquare test under H_o.

2.2 Other Sampling Schemes

These results for W_1, W_2 and LR_1 are well-known. The question of interest to us here is the implication of the more general assumption that $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{V})$, where $\underset{\sim}{V}$ is not necessarily equal to $\underset{\sim}{P}$. Here $\underset{\sim}{p}$ is a survey estimate of $\underset{\sim}{\pi}$ and may depend on sampling weights and other adjustment factors. This situation often arises in sampling under a complex sample design. We assume that $\hat{\underset{\sim}{V}}$ is a consistent estimate of $\underset{\sim}{V}$. There are two approaches which we shall consider here. The first is to construct the appropriate Wald statistic for the given sample design. This would be

$$W_3 = n(\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \hat{\underset{\sim}{V}}^{-1} (\underset{\sim}{p} - \underset{\sim}{\pi}_o),$$

where the rank of $\hat{\underset{\sim}{V}}$ is $k-1$ so that W_3 is asymptotically χ^2_{k-1} under H_o.

An alternative approach would be to use W_1, W_2 or LR_1 directly as a test statistic. Now from multivariate normal theory, we know that the distribution of $n(\underset{\sim}{p} - \underset{\sim}{\pi}_o)^T \underset{\sim}{P}_o^{-1}(\underset{\sim}{p} - \underset{\sim}{\pi}_o)$ is that of $\Sigma\ \delta_i Z_i^2$, where $\{Z_i^2\}$ are independent χ_1^2 random variable and $\underset{\sim}{\delta} = (\delta_1, \ldots, \delta_{k-1})^T$ are the eigenvalues of $\underset{\sim}{P}_o^{-1}\underset{\sim}{V}$; see Johnson and Kotz (1970, pg. 150). This result was shown by Rao and Scott (1981), who call the δ_i's generalized design effects. We note that for $k = 2$, we have $\delta = n\sigma_p^2/\{\pi_o(1 - \pi_o)\}$, where $\sigma_p^2 = V\{p\}$. This is the usual design effect for p under H_o.

2.3 Approximations

Now, in general, the distribution function for linear combinations of χ_1^2 random variables is cumbersome, although their moments are easily obtained. Rao and Scott (1981) have suggested two approximations to obtain the significance levels. The first is to approximate the distribution as being proportional to a χ^2_{k-1} random variable, the proportionality constant being determined by equating the mean of the approximating distribution to that of the theoretical distribution. This results in the approximation

$$\sum_{i=1}^{k-1} \delta_i Z_i^2 \doteq \{\sum_{i=1}^{k-1} \delta_i/(k-1)\}\chi^2_{k-1} \tag{2.2}$$

Now,

$$\Sigma\ \delta_i = \mathrm{tr}(\underset{\sim}{P}_o^{-1}\underset{\sim}{V})$$
$$= \sum_{i=1}^{k} v_{ii}/\pi_{io}$$
$$= \sum_{i=1}^{k} d_i(1 - \pi_{io}),$$

which depends only on the cell design effects $\{d_i\}$, where v_{ii} is the $i-th$ diagonal element of $\hat{\underset{\sim}{V}}$ and $d_i = v_{ii}/[\pi_{io}(1 - \pi_{io})]$. This approximation is particularly convenient when the full covariance matrix is not known, but the cell design effects are given. This is often the case for official published data.

Table 1

Age Distribution Among Those Consuming 1-6 Drinks Per Week.
Census Age Distribution for Canada (1978-9)

	Age							
Census	15-19	20-24	25-34	35-44	45-54	55-64	65+	Total
Distribution	.133	.127	.218	.152	.140	.115	.115	1.000
Distribution of those consuming 1-6 drinks/week	.117	.150	.264	.175	.148	.093	.053	1.000
Design Effect	1.4	1.2	2.2	1.1	0.6	1.1	1.0	

Example 1

For the Canada Health Survey (1978-9), a stratified multi-stage household survey, data was derived for the age distribution among those consuming one to six drinks per week, based on a sample of 5,204 persons, aged 15 years and over. A description of the survey may be found in "The Health of Canadians" (Statistics Canada Catalogue No. 82-538).

The data, taken from Hidiroglou and Rao (1981), are presented in Table 1. The raw value for W_1 is 298. This is reduced to 248 by taking the approximation given by (2.2). For these data, the post-stratification adjustments for age and sex lead to small design effects.

A second approximation to the distribution of $\Sigma\, \delta_i\, Z_i^2$, suggested by Rao and Scott (1981), is the Satterthwaite (1946) approximation: $\Sigma\, \delta_i\, Z_i^2 \approx a\chi_\nu^2$. To obtain a and ν, it is necessary to compute

$$\Sigma\, \delta_i^2 = \operatorname{tr}\{(\underline{P}_o^{-1}\hat{\underline{V}})^2\}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{k} \nu_{ij}^2/(\pi_{io}\pi_{jo}).$$

However, this depends on all the terms of the matrix $\hat{\underline{V}}$. The important point, though, is that some adjustment to the multinomial test statistic is necessary to obtain the appropriate significance level.

An alternative approximation, suggested by Fellegi (1980), is to divide the statistic $n(p-\underline{\pi}_o)^T\, \underline{P}_o^{-1}(p-\underline{\pi}_o)$ by the average design effect, $\bar{d}$, instead of the weighted average given in (2.2). The effect of this on the data in Table 1 is that the adjusted chisquare value is 243, which is comparable to Rao and Scott's (1981) approximation.

3. TESTS OF INDEPENDENCE IN A TWO-WAY TABLE

3.1 Multinomial Sampling

We now suppose that the categories of the multinomial distribution can be cross-classified into an $r \times c$ table, where for the bivariate observation (Y_1, Y_2) we have $\Pr(Y_1 = i, Y_2 = j) = \pi_{ij}$; $\Sigma_{i=1}^r\, \Sigma_{j=1}^c\, \pi_{ij} = 1$. We denote $\pi_{i+} = \Sigma_{j=1}^c\, \pi_{ij}$ and $\pi_{+j} = \Sigma_{i=1}^r\, \pi_{ij}$. We denote $\underline{\pi} = (\pi_{11}, \ldots, \pi_{1c}, \ldots, \pi_{r1}, \ldots, \pi_{rc})^T$, $\underline{\pi}_R = (\pi_{1+}, \ldots, \pi_{(r-1)+})^T$, $\underline{\pi}_C = (\pi_{+1}, \ldots, \pi_{+(c-1)})^T$, $\underline{P}_R = \underline{D}_{\pi_R} - \underline{\pi}_R\underline{\pi}_R^T$. $\underline{P}_C = \underline{D}_{\pi_C} - \underline{\pi}_C\underline{\pi}_C^T$. We observe the random vector $\underline{n}$ from the multinomial distribution, where $E\{\underline{n}\} = n\underline{\pi}$. We let $\underline{p} = \underline{n}/n$, $p_{i+} = \sum_j p_{ij}$, and $p_{+j} = \sum_i p_{ij}$.

We wish to test the hypothesis of independence

$$H_o\colon \pi_{ij} - \pi_{i+}\pi_{+j} = 0 \text{ for } 1 \leq i \leq r-1;\ 1 \leq j \leq c-1,$$

against the alternative

$$H_1\colon \pi_{ij} - \pi_{i+}\pi_{+j} \neq 0 \text{ for some } (i, j).$$

If we construct $h_{ij} = p_{ij} - p_{i+}p_{+j}$, for $1 \leq i \leq r-1$ and $1 \leq j \leq c-1$, then under multinomial sampling under H_o, the asymptotic covariance matrix for $\underset{\sim}{h} = (h_{11}, \ldots, h_{1,c-1}, \ldots, h_{r-1,c-1})^T$ is $\underset{\sim}{P}_R \otimes \underset{\sim}{P}_C$, where $\otimes$ denotes the direct matrix product operation. Hence, the Wald statistic under H_o becomes

$$\begin{aligned} W_4 &= \underset{\sim}{h}^T(\hat{\underset{\sim}{P}}_C^{-1} \otimes \hat{\underset{\sim}{P}}_R^{-1})\underset{\sim}{h} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} (p_{ij} - p_{i+}p_{+j})^2/(p_{i+}p_{+j}), \end{aligned}$$

the familiar chisquare test with $(r-1)$ $(c-1)$ degrees of freedom.

Another test, which is asymptotically equivalent to W_4 under H_o, is the likelihood ratio test given by

$$LR_2 = 2n\left[\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log p_{ij} - \sum_{i=1}^{r} p_{i+}\log p_{i+} - \sum_{j=1}^{c} p_{+j}\log p_{+j}\right].$$

An alternative approach for this problem, which is a special case of the methods described in Grizzle, Starmer and Koch (1969) is to consider a Wald statistic based on

$$\{ f_{ij} = \log p_{ij} - \log p_{i+} - \log p_{+j};\ \text{for } 1 \leq i \leq r-1, \text{ and } 1 \leq j \leq c-1 \}.$$

The asymptotic covariance matrix for $\underset{\sim}{f} = (f_{11}, \ldots, f_{1,c-1}, \ldots, f_{r-1,c-1})^T$ is $(\underset{\sim}{D}_{\pi_R}^{-1} - \underset{\sim}{1}\,\underset{\sim}{1}^T) \otimes (\underset{\sim}{D}_{\pi_C}^{-1} - \underset{\sim}{1}\,\underset{\sim}{1}^T)$. Therefore the Wald statistic becomes

$$W_5 = \underset{\sim}{f}^T\left[\left(\hat{\underset{\sim}{D}}_{\pi_R} + \frac{\underset{\sim}{\pi}_R\underset{\sim}{\pi}_R^T}{p_{r+}}\right) \otimes \left(\hat{\underset{\sim}{D}}_{\pi_C} + \frac{\underset{\sim}{\pi}_C\underset{\sim}{\pi}_C^T}{p_{+c}}\right)\right]\underset{\sim}{f}$$

Now under H_o we note that f_{ij} is asymptotically equivalent to

$$\frac{p_{ij}}{\pi_{i+}\pi_{+j}} - \frac{p_{i+}}{\pi_{i+}} - \frac{p_{+j}}{\pi_{+j}} + 1,$$

so that $\sum_{i=1}^{r} \pi_{i+} f_{ij} \doteq \sum_{j=1}^{c} \pi_{+j} f_{ij} \doteq 0$. Using this approximation W_5 becomes

$$W_5' = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{i+}p_{+j}f_{ij}^2\,.$$

It should be noted that under H_o, the statistics W_4, LR_2 and W_5 are all asymptotically equivalent to

$$\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(p_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} - \sum_{i=1}^{r} \frac{(p_{i+} - \pi_{i+})^2}{\pi_{i+}} - \sum_{j=1}^{c} \frac{(p_{+j} - \pi_{+j})^2}{\pi_{+j}} \tag{3.1}$$

This result will prove useful in Section 3.3.

3.2 Other Sampling Schemes

Now, relaxing the assumption that $\underset{\sim}{n}$ is multinomial, we assume instead that $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{V})$ where $\underset{\sim}{p}$ is a survey estimate which may depend on sampling weights and other adjustment factors. For this case, Shuster and Downing (1976) and Fellegi (1980) suggest that we construct the Wald statistic based on $\{h_{ij} = p_{ij} - p_{i+}p_{+j}\}$. If we let $\underset{\sim}{J}_a$ be the $(a-1) \times a$ matrix given by

$$\underset{\sim}{J}_a = [\, \underset{\sim}{I} \mid \underset{\sim}{0} \,] \tag{3.2}$$

$$\text{and let } \hat{\underset{\sim}{H}} = (\underset{\sim}{J}_r - \underset{\sim}{\pi}_R \underset{\sim}{1}^T) \otimes (\underset{\sim}{J}_c - \underset{\sim}{\pi}_C \underset{\sim}{1}^T) - (\underset{\sim}{\pi}_R \underset{\sim}{1}^T \otimes \underset{\sim}{\pi}_C 1^T)$$

then the Wald statistics is

$$W_6 = \underset{\sim}{h}^T(\hat{\underset{\sim}{H}} \hat{\underset{\sim}{V}} \hat{\underset{\sim}{H}}^T)^{-1} \underset{\sim}{h},$$

which under H_0 is asymptotically $\chi^2_{(r-1)(c-1)}$.

Alternatively, we could construct a Wald statistic based on $\{f_{ij} = \log p_{ij} - \log p_{i+} - \log p_{+j}\}$. This is a special case of the log-linear model approach to be discussed in Section 4. We define $(r - 1) \times r$ and $(c - 1) \times c$ matrices as follows:

$$\hat{\underset{\sim}{E}}_R = \left[\hat{\underset{\sim}{D}}{}^{-1}_{\pi_R} \mid \underset{\sim}{0}\right], \; \hat{\underset{\sim}{E}}_C = \left[\hat{\underset{\sim}{D}}{}^{-1}_{\pi_C} \mid \underset{\sim}{0}\right].$$

We let $\hat{\underset{\sim}{F}} = (\hat{\underset{\sim}{E}}_R - \hat{\underset{\sim}{E}}_R \underset{\sim}{1}\underset{\sim}{1}^T) \otimes (\hat{\underset{\sim}{E}}_C - \hat{\underset{\sim}{E}}_C \underset{\sim}{1}\underset{\sim}{1}^T) - (\hat{\underset{\sim}{E}}_R \underset{\sim}{1}\underset{\sim}{1}^T \otimes \hat{\underset{\sim}{E}}_C \underset{\sim}{1}\underset{\sim}{1}^T)$.

The appropriate Wald statistics is

$$W_7 = \underset{\sim}{f}^T(\hat{\underset{\sim}{F}} \hat{\underset{\sim}{V}} \hat{\underset{\sim}{F}}^T)^{-1} \underset{\sim}{f}.$$

Now, analogously to the goodness of fit problem in Section 2, Rao and Scott (1981) have considered null distributions of the test statistics based on W_4, LR_2 and W_5, which are all asymptotically equivalent to the null distribution of (3.1). We see, therefore, that the null distribution is the same as $\Sigma_{i=1}^{(r-1)(c-1)} \delta_i Z_i^2$, where $\{Z_i^2\}$ are independent χ_1^2 and the δ_i's are the eigenvalues of

$$(\underset{\sim}{P}_R^{-1} \otimes \underset{\sim}{P}_C^{-1})(\underset{\sim}{H} \underset{\sim}{V} \underset{\sim}{H}^T).$$

Cowan and Binder (1978) investigated the properties of the eigenvalue from a simple two-stage self-weighting design for a 2×2 table. They found that the eigenvalue increases as the degree of independence of the cell proportions within the primary sampling units decreased.

3.3 Approximations

An approximation for the distribution of $\Sigma\, \delta_i Z_i^2$ is

$$\Sigma\, \delta_i Z_i^2 \approx \frac{\Sigma\, \delta_i}{(r-1)(c-1)} \chi^2_{(r-1)(c-1)},$$

as in (2.2). Since the statistic is asymptotically equivalent to(3.1) under H_o, by computing the mean of (3.1) we obtain

$$\Sigma\, \delta_i = \sum_{i=1}^{r} \sum_{j=1}^{c} d_{ij} (1 - \pi_{i+}\pi_{+j}) - \sum_{i=1}^{r} d_i^{(r)} (1 - \pi_{i+}) - \sum_{j=1}^{c} d_j^{(c)}(1 - \pi_{\pm j}),$$

where d_{ij} is the cell design effect; $d_i^{(r)}$ and $d_j^{(c)}$ are the row and column margin design effects, respectively. This particularly simple expression was obtained by Rao and Scott (1983). Fellegi (1980) suggested an alternative approximation as:

$$(\sum_{i=1}^{r} \sum_{j=1}^{c} d_{ij}/rc)\, \chi^2_{(r-1)(c-1)}$$

Example 2

In Table 2, we give a 4×2 table from the Canada Health Survey, which cross-classifies drug use (four categories; 0, 1, 2, 3+ drug classes in a 2-day period) and sex (male, female). Here n = 31,668.

The raw value for W_4 is 774. Rao and Scott's (1981) adjustment reduces this to 437. Fellegi's (1980) adjustment reduces this to 327. The Wald statistics, W_6, is 538. Hidiroglou and Rao (1981) found that the Rao and Scott (1981) approximation performs quite well relative to the Satterthwaite (1946) approximation which is based on the complete covariance matrix.

LOG-LINEAR MODELS

4.1 Multinomial Sampling

We now extend the results of the previous section to more general cross-classifications of the multinomial distribution. The standard results for these models are given in Bishop, Fienberg

Table 2

Variety of Drugs Taken by Sex for Canada (1978-79)

		Number of Drug Varieties				
Sex		0	1	2	3±	Total
Male	Proportion	0.293	0.134	0.048	0.021	0.496
	Design Effect	1.56	3.37	1.15	1.38	0.00*
Female	Proportion	0.228	0.159	0.072	0.045	0.504
	Design Effect	3.59	3.13	2.85	1.96	0.00*
Total	Proportion	0.521	0.293	0.120	0.066	1.000
	Design Effect	6.03	6.46	1.65	2.57	

* Because of age-sex post-stratification, these design effects are zero.

and Holland (1975) and Fienberg (1980). We have $\underset{\sim}{\pi} = (\pi_1,\ldots,\pi_k)^T$ is a vector of cell proportions; $\Sigma_{i=1}^{k} \pi_k = 1$. We observe $\underset{\sim}{n} = (n_1,\ldots,n_k)^T$, the counts in each cell from a random sample, so that $\underset{\sim}{n}$ has a multinomial distribution ($\Sigma\, n_i = n$). We let $\underset{\sim}{p} = \underset{\sim}{n}/n$ and define

$$\underset{\sim}{\mu} = \log \underset{\sim}{\pi}.$$

The log-linear model assumes that for a parameter vector $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_t)^T$, we have

$$\underset{\sim}{\mu}(\underset{\sim}{\theta}) = u(\underset{\sim}{\theta})\underset{\sim}{1} + \underset{\sim}{X}\underset{\sim}{\theta},$$

where $\underset{\sim}{X}$ is a known $k \times t$ matrix of full rank and $\underset{\sim}{X}^T\underset{\sim}{1} = \underset{\sim}{0}$. Note that $t \leq k-1$. If $t = k-1$, we have the saturated model.

The maximum likelihood estimate for $\underset{\sim}{\theta}$ is given by solving

$$\underset{\sim}{X}^T(\underset{\sim}{p} - \hat{\underset{\sim}{\pi}}) = \underset{\sim}{0}, \tag{4.1}$$

where $\hat{\underset{\sim}{\pi}} = \underset{\sim}{\pi}(\hat{\underset{\sim}{\theta}})$. Now, asymptotically we have

$$\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi} \doteq \underset{\sim}{P}\underset{\sim}{X}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}),$$

where $\underset{\sim}{P} = \underset{\sim}{D}_\pi - \underset{\sim}{\pi}\underset{\sim}{\pi}^T$. From (4.1), we then obtain

$$\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} \doteq (\underset{\sim}{X}^T\underset{\sim}{P}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T(\underset{\sim}{p} - \underset{\sim}{\pi})$$

and

$$\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi} \doteq \underset{\sim}{P}\underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{P}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T(\underset{\sim}{p} - \underset{\sim}{\pi}).$$

Since $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{\pi}) \to N(\underset{\sim}{0}, \underset{\sim}{P})$ we obtain

$$n^{1/2}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \to N[\underset{\sim}{0}, (\underset{\sim}{X}^T\underset{\sim}{P}\underset{\sim}{X})^{-1}]$$

$$n^{1/2}(\underset{\sim}{\pi} - \underset{\sim}{\pi}) \to N[\underset{\sim}{0}, \underset{\sim}{P}\underset{\sim}{X}(\underset{\sim}{X}^T\underset{\sim}{P}\underset{\sim}{X})^{-1}\underset{\sim}{X}^T\underset{\sim}{P}].$$

Suppose now that the linear expression $\underset{\sim}{X}\underset{\sim}{\theta}$ can be decomposed as $\underset{\sim}{X}_1\underset{\sim}{\theta}_1 + \underset{\sim}{X}_2\underset{\sim}{\theta}_2$ where $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ are full rank, $\underset{\sim}{X}_1$ is $k \times r$, $\underset{\sim}{X}_2$ is $k \times s$, $\underset{\sim}{\theta}_1$ is $r \times 1$ and $\underset{\sim}{\theta}_2$ is $s \times 1$, where $r + s = t$.

We consider the problem of testing

$$H_o\colon \underset{\sim}{\theta}_2 = \underset{\sim}{0},$$

against the alternative

$$H_1\colon \underset{\sim}{\theta}_2 \neq \underset{\sim}{0}.$$

We use $\underset{\sim}{\theta}_1$, $\underset{\sim}{\theta}_2$, $\underset{\sim}{\pi}$, etc. to denote the estimates under the full model H_1. Alternatively, we let $\hat{\underset{\sim}{\theta}}_1$, $\hat{\underset{\sim}{\pi}}$, to denote estimates under H_o.

Now,

$$n^{1/2}(\underset{\sim}{\theta}_2 - \underset{\sim}{\theta}_2) \to N[\underset{\sim}{0}, (\tilde{\underset{\sim}{X}}_2^T\underset{\sim}{P}\tilde{\underset{\sim}{X}}_2)^{-1}]$$

where

$$\tilde{\underset{\sim}{X}}_2 = [\underset{\sim}{I} - \underset{\sim}{X}_1(\underset{\sim}{X}_1^T\underset{\sim}{P}\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1^T\underset{\sim}{P}]\underset{\sim}{X}_2 \tag{4.2}$$

so that the Wald statistic is

$$W_8 = n\hat{\hat{\theta}}_2^T \tilde{X}_2^T \hat{P} \tilde{X}_2 \theta_2.$$

Under H_o, this is asymptotically equivalent to the Pearson chisquare statistic

$$n\,(\hat{\pi} - \hat{\hat{\pi}})^T \hat{D}_\pi^{-1} (\hat{\pi} - \hat{\hat{\pi}}),$$

or the likelihood ratio test

$$LR_3 = 2n \sum_{i=1}^{k} p_i \log(\hat{\pi}_i / \hat{\hat{\pi}}_i).$$

Under H_o, these statistics are asymptotically χ_s^2.

4.2 Other Sampling Schemes

We still assume that the cell proportions, π, satisfy $\mu = \log \pi = u(\theta_1, \theta_2)1 + X_1\theta_1 + X_2\theta_2$ but we now have $n^{1/2}(p - \pi) \rightarrow N(0, V)$, where p is a survey estimate.

Rao and Scott (1983) suggest the following Wald statistic for testing $\theta_2 = 0$. We let C be any $k \times s$ matrix with $C^T X_1 = 0$, $C^T 1 = 0$ and $C^T X_2$ nonsingular. For example if $X_1^T X_2 = 0$ then $C = X_2$ is convenient. Now the hypothesis is equivalent to $C^T \mu = 0$. We have

$$C^T(\hat{\mu} - \mu) \doteq C^T D_\pi^{-1}(\hat{\pi} - \pi)$$

$$\doteq C^T X (X^T P X)^{-1} X^T (p - \pi),$$

where π is obtained from (4.1), based on the survey estimate, p.

We therefore have the Wald statistics

$$W_9 = n\hat{\mu}^T C [C^T X (X^T \hat{P} X)^{-1} (X^T \hat{V} X)(X^T \hat{P} X)^{-1} X^T C]^{-1} C^T \hat{\mu}.$$

Similar results were also given in Binder (1983). If under H_1, the model is saturated $(r + s = k - 1)$, then $p = \pi$ and we obtain

$$W_9 = n\hat{\mu}^T C [C^T \hat{D}_\pi^{-1} \hat{V} \hat{D}_\pi^{-1} C]^{-1} C^T \hat{\mu}.$$

Rao and Scott (1984) show that if we use $\hat{P}$ instead of $\hat{V}$ in W_9 then these are asymptotically equivalent to the likelihood ratio or Pearson χ^2 test statistics. They also show that the likelihood ratio test statistics is distributed as $\Sigma_{i=1}^{s} \delta_i Z_i^2$ under H_o, where $\{Z_i^2\}$ are independent χ_1^2 and $\{\delta_i\}$ are the eigenvalues of

$$(\tilde{X}_2^T P \tilde{X}_2)^{-1} (\tilde{X}_2^T V \tilde{X}_2), \tag{4.3}$$

for $\tilde{X}_2$ defined in (4.2).

4.3 Approximations

As before, we approximate the null distribution

$$\sum_{i=1}^{s} \delta_i Z_i^2 \approx \left(\frac{\Sigma\, \delta_i}{s}\right) \chi_s^2.$$

This involves computing the trace of (4.3). Rao and Scott (1984) show that if the model admits explicit solutions for both $\hat{\pi}$ and $\hat{\hat{\pi}}$, then the approximation depends on the matrix V only through cell design effects and marginal design effects. This observation is particularly convenient when ony the estimated design effects for the cell proportions and margins are available, as is often the case for published tables.

Example 3

Hidiroglou and Rao (1983) considered all direct estimates from the three-way table: Drug use (5 categories: 0, 1, 2, 3, 4+ drug classes in a 2 day period) × Age (4 categories; 0-14, 15-44, 45-64, 65+) × Sex (male, female), taken from the Canada health Survey. We give the results for testing whether Age and Sex are independent in each drug category (n = 31,668). This is equivalent to the hypothesis

$$H_0: \pi_{ijk} = \pi_{ij+}\frac{\pi_{i+k}}{\pi_{i++}}.$$

Using Bishop, Fienberg and Holland's (1975) notation, where $\log \pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$, the null hypothesis is equivalent to

$$H_0: u_{23(jk)} = u_{123(ijk)} = 0 \quad \text{for all } (i, j, k).$$

The raw chisquare value is 23 based on 15 degrees of freedom. The average eigenvalue is 1.39, so that the approximation reduces the chisquare value to 16. Whereas the unadjusted chisquare value would lead the analyst to reject the hypothesis at the 10% level, the approximation indicates that h_o cannot be rejected even at the 30% level.

5. LOGISTIC REGRESSION MODELS

5.1 Multinomial Sampling

We now consider a logistic regression model for the conditional distribution of a binary response variable y given the vector $\underset{\sim}{x}$ of independent variables. In particular, this conditional distribution is

$$\Pr(y_i \mid x_i) = \pi(\underset{\sim}{x}_i)^{y_i}\,[1 - \pi(\underset{\sim}{x}_i)]^{1-y_i},$$

where $y_i \in \{0, 1\}$.

For the logistic regression model, we have

$$\log\left\{\frac{\pi(\underset{\sim}{x}_i)}{1 - \pi(\underset{\sim}{x}_i)}\right\} = \underset{\sim}{x}_i^T\,\underset{\sim}{\theta},$$

where $\underset{\sim}{\theta}$ is an unknown vector of parameters.

We note that if $\underset{\sim}{x}_i$ is a categorical vector of 0's and 1's, this is a special case of a log-linear model as described in Section 4. Here we allow $\underset{\sim}{x}_i$ to be arbitrary. The extension to the case of k-categories for the y-variable is straight-forward, it is also possible to generalize the model to

$$\log\left\{\frac{\pi(\underset{\sim}{x}_i)}{1 - \pi(\underset{\sim}{x}_i)}\right\} = f(\underset{\sim}{x}_i^T\,\underset{\sim}{\theta}),$$

for a known function $f(\cdot)$, but we do not discuss this here.

Now, the maximum likelihood estimate for $\underset{\sim}{\theta}$ is given by

$$\underset{\sim}{X}^T(\underset{\sim}{y} - \hat{\underset{\sim}{\pi}}) = \underset{\sim}{0}$$

where $\underset{\sim}{y} = (y_1, \ldots, y_n)^T$, $\hat{\underset{\sim}{\pi}} = [\hat{\pi}(\underset{\sim}{x}_1), \ldots, \hat{\pi}(\underset{\sim}{x}_n)]^T$ and $\underset{\sim}{X} = [\underset{\sim}{x}_1 \ldots \underset{\sim}{x}_n]^T$.

Under suitable regularity conditions, we have

$$n^{1/2}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \rightarrow N\,[\underset{\sim}{0},\ n\,(\underset{\sim}{X}^T\underset{\sim}{\Delta}\underset{\sim}{X})^{-1}], \text{ where } \underset{\sim}{\Delta} = \underset{\sim}{D}_\pi(\underset{\sim}{I} - \underset{\sim}{D}_\pi).$$

If we have $\underset{\sim}{X}\underset{\sim}{\theta} = \underset{\sim}{X}_1 \underset{\sim}{\theta}_1 + \underset{\sim}{X}_2 \underset{\sim}{\theta}_2$ and consider testing the hypothesis

$$H_0\text{: } \underset{\sim}{\theta}_2 = \underset{\sim}{0}$$
$$H_1\text{: } \underset{\sim}{\theta}_2 \neq \underset{\sim}{0},$$

we obtain the Wald statistic

$$W_{10} = n\, \hat{\underset{\sim}{\theta}}_2^T (\tilde{\underset{\sim}{X}}_2^T \underset{\sim}{\Lambda} \tilde{\underset{\sim}{X}}_2)\, \hat{\underset{\sim}{\theta}}_2$$

where

$$\tilde{\underset{\sim}{X}}_2 = [\underset{\sim}{I} - \underset{\sim}{X}_1(\underset{\sim}{X}_1^T \hat{\underset{\sim}{\Lambda}} \underset{\sim}{X}_1)^{-1} \underset{\sim}{X}_1^T \hat{\underset{\sim}{\Lambda}}] \underset{\sim}{X}_2.$$

The likelihood ratio test here is

$$LR_4 = 2 \sum_{i=1}^{n} \left[y_i \log \left(\frac{\hat{\pi}_i}{\hat{\hat{\pi}}_i} \right) + (1 - y_i) \log \left\{ \frac{(1 - \hat{\pi}_i)}{(1 - \hat{\hat{\pi}}_i)} \right\} \right]$$

which is asymptotically equivalent to W_{10} under H_0.

5.2 Other Sampling Schemes

Suppose now that $n^{-1/2} \underset{\sim}{X}^T(\underset{\sim}{y} - \underset{\sim}{\pi}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{V})$ and that $\hat{\underset{\sim}{V}}$ is a consistent estimator of $\underset{\sim}{V}$. Here $\underset{\sim}{y}$ is not necessarily a vector of 0's and 1's, but may in fact depend on the sampling weights and other adjustment factors. Estimating $\underset{\sim}{V}$ is usually possible since $\underset{\sim}{X}^T(\underset{\sim}{y} - \underset{\sim}{\pi})$ is the sum of random observations and most sample designs admit a consistent estimator of the sum of (not necessarily independent) observations. To estimate $\underset{\sim}{V}$ we use $\hat{\underset{\sim}{\pi}}$ instead of $\underset{\sim}{\pi}$ in the estimate. Since asymptotically

$$(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \doteq (\underset{\sim}{X}^T \underset{\sim}{\Lambda} \underset{\sim}{X})^{-1} \underset{\sim}{X}^T (\underset{\sim}{y} - \underset{\sim}{\pi}),$$

we have that

$$n^{1/2}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}) \rightarrow N[\underset{\sim}{0},\, n^2(\underset{\sim}{X}^T \underset{\sim}{\Lambda} \underset{\sim}{X})^{-1} \underset{\sim}{V} (\underset{\sim}{X}^T \underset{\sim}{\Lambda} \underset{\sim}{X})^T];$$

see Binder (1983) for a detailed justification of this result. Now, a Wald statistic may be constructed from the estimated covariance matrix for $\hat{\underset{\sim}{\theta}}_2$.

Table 3

Logistic Regression Model for Explaining Use of Physician Services

Variable	Type	d.f.	Wald Statistic
Age	Categorical	4	19.232
Sex	Categorical	1	12.494
Age-Sex Interactions	Categorical	4	36.001
Family Income	Categorical	5	14.642
Occupation	Categorical	3	8.614
Occupation-Sex Interactions	Categorical	3	11.501
Marital Status	Categorical	3	45.752
Medical History	Categorical	2	36.700
Number of Health Problems	Quantitative	1	81.554
Drug Use	Categorical	2	272.175
Number of Accidents	Quantitative	2	106.372
Number of Disability Days	Quantitative	2	29.052
Community Size	Categorical	2	11.751
Provincial Physician – Population Ratios	Quantitative	1	0.540

Example 4

A logistic regression model was fit on 20,726 respondents from the Canada health survey to explain use or non-use of physician services over a 12-month period. In total it was estimated that 77% of the population visited a physician at least once. The results are summarized in Table 3. For more complete details, see Binder (1983). The logistic model seemed to fit the data very well.

5.3 Qualitative Explanatory Variables

The theory of this section was obtained by G. Roberts in an unpublished manuscript (Carleton University). Here the explanatory variables are all qualitative. We label the domains, $\{1, \ldots, I\}$. We let p_i be the survey estimate of the i-th domain proportion and $\hat{N}_i$ is the estimate of the size of the i-th domain, N_i. Under the model, the expected proportion in the i-th domain is f_i, where

$$\log \{f_i/(l-f_i)\} = \underset{\sim}{a}_i^T\theta,$$

for $\underset{\sim}{a}_i$ known and $\underset{\sim}{\theta}$ an unknown parameter. We define $\underset{\sim}{A} = [\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_I]^T$ and let $\underset{\sim}{D}_{\hat{N}} = \text{diag}\{\hat{N}_i, \ldots, \hat{N}_I\}$.

Under the model, the survey estimator of $\underset{\sim}{f} = (f_1, \ldots, f_I)^T$ is given by $\hat{\underset{\sim}{f}}$, the solution to

$$\underset{\sim}{A}^T\underset{\sim}{D}_{\hat{N}}(p-\hat{\underset{\sim}{f}}) = \underset{\sim}{0}. \tag{5.1}$$

Since asymptotically

$$\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta} \doteq (\underset{\sim}{A}^T\Delta\underset{\sim}{A})^{-1}\underset{\sim}{A}^T\underset{\sim}{D}_N(\underset{\sim}{p}-\underset{\sim}{f}),$$

where $\underset{\sim}{\Delta} = \text{diag}\{N_1f_1(1-f_1), \ldots, N_If_I(1-f_I)\}$, we have

$$n^{1/2}(\underset{\sim}{\theta} - \underset{\sim}{\theta}) \rightarrow N[\underset{\sim}{0}, (\underset{\sim}{A}^T\underset{\sim}{\Delta}\underset{\sim}{A})^{-1}\underset{\sim}{A}^T\underset{\sim}{D}_N\underset{\sim}{V}_P\underset{\sim}{D}_N\underset{\sim}{A}(\underset{\sim}{A}^T\Delta\underset{\sim}{A})^{-1}]$$

whenever $n^{1/2}(\underset{\sim}{p} - \underset{\sim}{f}) \rightarrow N(\underset{\sim}{0}, \underset{\sim}{V}_P)$.

Under independent binomial sampling, the covariance matrix reduces to $(N/n)(\underset{\sim}{A}^T\underset{\sim}{\Delta}\underset{\sim}{A})^{-1}$, where n is the sample size.

The likelihood ratio test for testing goodness of fit is

$$LR_5 = 2(n/\hat{N})\sum_{i=1}^{I}\hat{N}_i[p_i\log(p_i/\hat{f}_i) + (1-p_i)\log\{(1-p_i)/(1-\hat{f}_i)\}],$$

where n is the sample size and $\hat{N} = \Sigma\hat{N}_i$. Under H_o this is asymptotically equivalent to

$$W_{11} = (n/\hat{N})\sum_{i=1}^{I}\hat{N}_i(p_i - \hat{f}_i)^2/[f_i(1 - f_i)].$$

In general, the distribution of LR_5 will be that of $\Sigma\,\delta_i Z_i^2$, where $\{Z_i\}$ are independent χ_1^2, and $\{\delta_i\}$ are the eigenvalues of $\hat{N}^{-1}\underset{\sim}{D}_{\hat{N}}[\underset{\sim}{\Delta}^{-1} - \underset{\sim}{A}(\underset{\sim}{A}^T\underset{\sim}{\Delta}\underset{\sim}{A})^{-1}\underset{\sim}{A}^T]\underset{\sim}{D}_{\hat{N}}\underset{\sim}{V}_P\underset{\sim}{D}_{\hat{N}}[\underset{\sim}{\Delta}^{-1} - \underset{\sim}{A}(\underset{\sim}{A}^T\underset{\sim}{\Delta}\underset{\sim}{A})^{-1}\underset{\sim}{A}^T]\underset{\sim}{\Delta}\underset{\sim}{D}_{\hat{N}}^{-1}$. By taking the expectation of W_{11}, and approximating

$$W_{11} \approx \frac{\Sigma\,\delta_i}{I - s}\chi^2_{I-s}$$

where $s = \text{rank}(\underset{\sim}{A})$, we obtain

$$\Sigma\,\delta_i = (n/\hat{N})\sum_{i=1}^{I}\hat{N}_i v_{ii}^{(r)}/\{f_i(1-f_i)\}$$

where $v_{ii}^{(r)} = V\{p_i - \hat{f}_i\}$. The $\{v_{ii}^{(r)}\}$ may be computed using the relationship $\underset{\sim}{p} - \hat{\underset{\sim}{f}} \doteq [\underset{\sim}{I} - \text{diag}\,\{f_i(1-f_i)\}\,\underset{\sim}{A}\,(\underset{\sim}{A}^T\underset{\sim}{\Delta}\underset{\sim}{A})^{-1}\underset{\sim}{A}^T\underset{\sim}{D}_N]\,(\underset{\sim}{p} - \underset{\sim}{f})$.

Example 5

The data from the October 1980 Canadian Labour Force Survey was used to fit logistic (logit) models for the probability of being employed. The sample consisted of males aged 15-64 who were in the labour force and not full time students. A logit model, quadratic in age and in education, was fitted. Age-group levels were formed by dividing the interval [15, 64] into ten groups with the jth age-group being the interval $[10 + 5j, 14 + 5j]$, $j = 1, 2, \ldots, 10$. The midpoint of each age-group was used as the value of the age for all persons in that age-group. Six levels of education were formed by assigning to each person a value based on the median years of schooling. Age by education classification led to the formation of 60 cells.

Let $\pi_i = \Pr\{\text{an individual in the ith cell is employed}\}$, $i = 1, 2, \ldots, 60$. We assume that $0 < \pi_i < 1$. Hence $1 - \pi_i$ represents the probability that the individual in the ith cell is unemployed. The model, considered for fit, was

$$\ln \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + \beta_3 d_i + \beta_4 d_i^2, \qquad (1)$$

$$i = 1, 2, \ldots, 60$$

where a_i and d_i are the age and education variable values for the individuals in the ith cell.

Using the survey estimates p_i of π_i, the values of Pearson's statistic W_{11} and the likelihood ratio statistic LR_5 were computed as $W_{11} = 98.94$ and $LR_5 = 101.20$. The upper 5% point of the chi-square distribution, with 55 degrees of freedom, is 73.31. Using these values of W_{11} or LR_5 we would reject the model l. These values of W_{11} or LR_5, however, are appropriate only if the sample was a random sample.

The estimate average eigenvalue, $\Sigma\, \delta_i/55$, for testing goodness of fit for this data is 1.88. This would reduce W_{11} to 52.63 and LR_5 to 53.83. Hence, with this adjustment, we find that the data are consistent with the model (1).

The use of the Wald statistic, $(\underset{\sim}{p} - \hat{\underset{\sim}{f}})^T\,[\hat{\underset{\sim}{V}}^{(r)}]^-\,(\underset{\sim}{p} - \hat{\underset{\sim}{f}})$, for testing the goodness of fit was also considered. Here we use the g-inverse of $\hat{\underset{\sim}{V}}^{(r)}$ since the matrix is singular. Some perturbation to the estimates of p_i, when $p_i = 1$, was necessary for computing the Wald statistic. It was found that the Wald statistic was unstable for our problem. Minor perturbations in the estimates of p led to considerable change in the value of the Wald statistic. Also the value of the Wald statistic is very large here due to instability in the estimated covariance matrix involved in its calculation. The Wald statistic is at least 30 times larger than our adjusted Chi-squared values.

6. SOFTWARE CONSIDERATIONS

Advancement of computer technology has made data collection, storage and retrieval operations easy and efficient. Powerful generalized software systems, such as TPL, STATPAK and ESTIMATION SYSTEM, have been used to produce cell estimates and some of their variances fairly easily to users and analysts. As well a number of commercially available packages such as BMDP, SPSS and SAS are powerful analytic tools in certain contexts. However, the ability to perform analysis such as those described in this paper are limited. For example, in situations involving hypothesis testing or statistical inference, these packages assume that the data to be analyzed come from surveys with simple random samples.

At present, an integrated software package, similar to the ones mentioned above, but designed for analyses of the type of data discussed in this paper, is not available. As a result, the researcher requiring a quick solution to his problem is usually forced to use existing statistical packages which may not be appropriate.

The alternatives are

- use existing packages with modifications
- use existing stand-alone software
- write customized programs
- use combinations of the above.

For the analyses given in this paper, modifications to the MINI CARP program (Hidiroglou, Fuller and Hickman; 1980) were incorporated to obtain the results in Examples, 1, 2 and 3. For Example 4, a combination of PL/1 and SAS programs were developed. The analysis of the Labour Force Survey data (Example 5) used a combination of customized programs and SAS.

For the above alternatives, some practical drawbacks have been experienced. they include:

(a) If an existing package is to be modified, intimate knowledge of the package is often required;
(b) Identical information may have to be duplicated on separate data files, as these alternatives are not integrable like generalized systems;
(c) Compared to an integrated "user-friendly" package, these alternatives lack elegance and operational efficiency as software;
(d) Comprehensive documentation is not generally available for specially written programs limiting the availability of software.

Work is now ongoing to develop SAS based procedures for performing many of these analyses, Our ultimate goal is similar to that proposed by Shah (1981); namely, the development of an integrated software package for survey data analysis. This is a goal worth striving for, if we are to avoid the frustrations now being experienced by researchers who are faced with either developing their own software or using existing software which could lead to erroneous results and conslusions.

7. DISCUSSION

We have examined a number of problems which arise when fitting models to categorical data which have been collected under complex sampling designs. The basic approach has been to derive the appropriate Wald statistic for the fitted model or to use the test statistic which is motivated from multinomial-type sampling designs and find a suitable approximation to its null distribution.

We have not addressed the issue as to whether one should really be taking a model-based or design-based approach to begin with. Instead, we have concentrated on design-based inferences.

To put this issue into focus, let us reconsider the test of independence in a two-way contingency table. The question of independence arises if we are interested in whether knowing the value of variable Y_1 affects our knowledge about variable Y_2. If it does not, for all the individuals in the population, then we say the variables are independent. However, if we also know the value of Y_3, it may turn out that Y_1 and Y_2 are no longer independent. This is particularly important when Y_3 is a design variable (such as geographic stratum). Since design variables are usually known for all sampled individuals, we have one of two options: (a) we can say that the question of independence is no longer relevant, or (b) we can marginalize out Y_3, and say that we are only interested in Y_1 and Y_2, unconditionally. Assuming that we take approach (b), the results of this paper seem appropriate. In some cases it may be possible to test if Y_1 and Y_2 are conditionally independent given Y_3.

There is a further difficulty, however. Suppose we are interested in the cell proportions π_{ij} from a finite population of size N. If we were to take a census from this population, it is highly unlikely that we would obtain $\pi_{ij} = \pi_{i+} \pi_{+j}$ exactly. The best that we could hope for is that some measure of association such as $N \Sigma\Sigma \, (\pi_{ij} - \pi_{i+}\pi_{+j})^2/\pi_{i+}\pi_{+j}$ is small. Note that even under a super-population model of exact independence, we would not expect this measure of association to be zero. Perhaps, we should instead be testing hypotheses such as

$$H_0\text{: Measure of Association} \leq C$$

$$H_1\text{: Measure of Association} > C.$$

Further research is needed in this area. However, for practical circumstances where the sampling fraction is not large, the methods given in this paper are suitable.

REFERENCES

ALTHAM, P.A.E. (1976). Discrete vriable analysis for idividuals grouped into families. *Biometrika,* 63, 263-269.

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

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

BRIER, S.S. (1978). Discrete Data Models with Random Effects. Technical Report, University of Minnesota, School of Statistics.

COHEN, J.E. (1976). The distribution of the chi-squared statistic under cluster sampling from contingency tables. *Journal of the American Statistical Association,* 71, 665-670.

COWAN, J. and BINDER, D.A. (1978). The effect of a two-stage sample design on tests of independence in a 2 by 2 table. *Survey Methodology,* 4, 16-28.

FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of 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 American Statistical Assocociation,* 71, 665-670.

FIENBERG, S.E. (1980). *The Analysis of Cross Classified Data,* (2nd ed.). Cambridge, Massachusetts: MIT Press.

GRIZZEL, J.E., STARMER, C.F. and KOCH, G.G. (1969). Analysis of categorical data by linear models. *Biometrics,* 25, 489-504.

HIDIROGLOU, M.A., FULLER, W.A. and HICKMAN, R.D. (1980). MINICARP: A program for estimating simple descriptive statistics and their variances for multi-stage stratified designs. Iowa State University: Ames, Iowa.

HIDIROGLOU, M.A. and RAO, J.N.K. (1981). Chisquare tests for the analysis of categorical data from the Canada Health Survey. Paper presented at the International Statistical Institute Meetings, Buenos Aires, 1981.

HIDIROGLOU, M.A. and RAO, J.N.K. (1983). Chi-square tests for the analysis of three-way contingency tables from the Canada Health Survey. Technical Report, Statistics Canada.

IMREY, P.B., KOCH, G.G. and STOKES (1981). Categorical data analysis: Some reflections on the log linear model and logistic regression. part I: Historical and Methodological Ovewview. *International Statistical Review,* 49, 265-283.

JOHNSON,N.L. and KOTZ, S. (1970). *Continuous Univariate Distributions.* Boston: Houghton Mofflin.

KOCH, G.G., FREEMAN, D.H. JR., and FREEMAN, J.L. (1975). Strategies in the Multivariate Analysis of Data from Complex Surveys. *International Statistical Review,* 43, 59-78.

RAO, J.N.K. and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-square tests for goodness of fit and independence in two-way tables. *Journal of American Statistical Association,* 76, 221-30.

RAO, J.N.K. and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics,* 12, 48-60.

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

SHAH, B.V. (1981). Development of survey data analysis software. Research Triangle Institute, Research Triangle Park, North Carolina.

SHUSTER, J.J. and DOWNING, D.J. (1976). Two-way contingency tables for complex sampling schemes. *Biometrika,* 63, 271-276.

Survey Methodology, December 1984
Vol. 10, No. 2, pp. 157–164
Statistics Canada

Estimating Economic Cycles in Semi-Annual Series

PIERRE A. CHOLETTE[1]

ABSTRACT

This paper presents a moving average which estimates the trend-cycle while eliminating seasonality from semi-annual series (observed twice yearly). The proposed average retains the power of all cycles which last three years or more; 90% of those of two years; and 55% of cycles of one year and a half. By comparison, the *two by two* moving average retains the power of respectively 75%, 50% and 25% of the same cycles.

KEY WORDS: Moving averages; Economic cycles; Spectral analysis.

1. INTRODUCTION

In some cases, semi-annual series exist for which there are no corresponding monthly data. In such instances, one cannot derive the seasonally adjusted semi-annual series from the monthly seasonally adjusted values. In addition, to our knowledge, there are no seasonal adjustment methods for semi-annual series.

This paper presents a moving average which eliminates seasonality and estimates the trend-cycle of semi-annual series. The approach of quadratic minimization used originates with Whittaker (1923) and was further developped by Leser (1961 and 1963), Cholette (1980), Schlicht (1981) and others.

The average derived has five terms and comprises a set of central weights for the semestres (half-years) at the centre of series; and two sets of end weights, for the two first and last semestres. Consequently there is no loss of estimates at the ends of series, as with the *two by four* moving average (used for quarterly series) for instance.

The spectral properties of the central weights prove to be superior to those of the *two by two* moving average, which first comes to mind as a way of processing semi-annual series. The properties of the end weights will also be examined.

2. ILLUSTRATION OF THE AVERAGE

Figure 1 shows the observed semi-annual original series z_t (dashed line) along with the trend-cycle c_t (solid line) estimated by the semi-annual cyclical average presented in this paper. As expected, the trend-cycle behaves smoothly and displays short run cycles, namely a three-year cycle extending from the second semestre (half-year) of 1977 to the first of 1980. An estimate is available for each observation, including the two first and last observations. The trend-cycle produced by the *two by two* moving average (dotted curve) on the other hand does not yield any estimate for the first and the last semestres. Furthermore, the two by two does not reach as deeply into the cyclical peaks and troughs compared to the proposed average.

[1] Pierre A. Cholette, Time Series Research and Analysis, Statistics Canada, 25th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

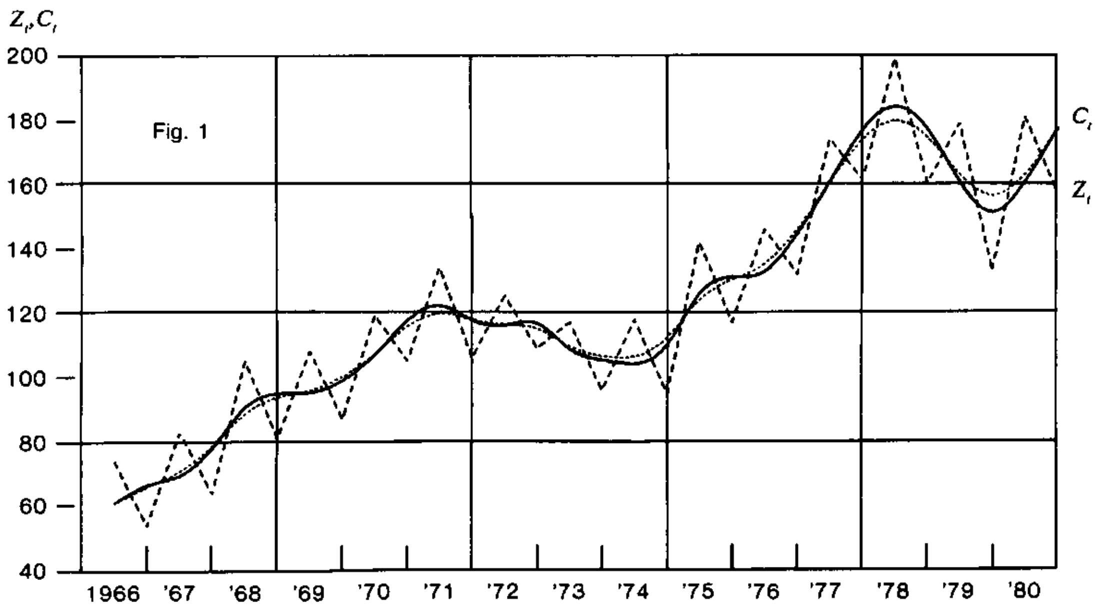

Figure 1. Semi-annual seasonal series (----) and its trend cycle (——) estimated by the proposed semi-annual cyclical moving average and by the *two by two* moving average (····)

3. WEIGHTS OF THE AVERAGE

Table 1-A displays the exact values of the weights of the semi-annual cyclical moving average used. The first row gives the *modified* central weights pertaining to estimates 3 to 28 (in Figure 1); the second, the end weights pertaining to the second-last estimate; and, the third, to the last estimate. Table 1-B shows the central weights derived according to the methodology of Section 6. We judged however these should be replaced by the *modified* central weights of Table 1-A for reasons to be explained.

Table 1-A

Exact weights of the proposed semi-annual cyclical average

Modified Central weights	−0.1000	0.2500	0.7000	0.2500	−0.1000
Second-last Set of weights	0.0625	−0.2500	0.3750	0.7500	0.0625
Last Set of weights	−0.0625	0.2500	−0.3750	0.2500	0.9375

Table 1-B

Unmodified central weights

−0.0625	0.2500	0.6250	0.2500	−0.0625

4. SPECTRAL ANALYSIS OF THE AVERAGE

The curves of Figures 2 and 3 represent the gain functions of the set of weights analysed. The gain value on the ordinate indicates the percentage of amplitude of the sinusoidal waves preserved, that is *passed* to the estimates, by the weights. The frequency of these waves is shown on the abscissa and varies from 0 to 0.500. Frequency 0.500 corresponds to the annual wave of two semestres (1/.50), that is to stable seasonality. Moving seasonality is accounted for by a few neighbouring quasi annual frequencies: 0.467 and 0.483.

Frequency 0.333 corresponds to a three (1/.33) semestre wave; frequency 0.250, four semestres; 0.200, to five semestres; 0.167, six semestre, etc. Frequencies associated with waves of three semestres or more (left of 0.333 in figures 2 and 3) pertain to the trend-cycle of series and constitute the target frequencies of the estimator.

The frequencies between 0.333 and 0.467 exclusively are associated with fluctuations of periodicity less than one and a half years and superior to the quasi-annual seasonal frequencies. They pertain to the irregular component of series. An ideal cyclical average should eliminate 100% of these irregular frequencies, 100% of the seasonal and quasi-seasonal frequencies and preserve only the cyclical frequencies from 0 to 0.333 inclusively.

a) Analysis of the central weights

The solid curve of Figure 2 shows that the modified central weights of the semi-annual cyclical average preserves 100% of all waves of five semestres (2 years) and more: everywhere left of frequency 0.200 the curve is above 100%. By comparison, the *two by two* moving average, represented by the dotted curve, only preservs 65% of the 5-semestre waves and 93% of the 10-semestres waves. Furthermore, the modified central weights pass 55% of 3-semestre waves and 90% of 4-semestre (2-year) waves; against 25% and 50% respectively for the two by two.

Both sets of weights completely eliminate stable seasonality, with gain valued at zero for the seasonal frequency 0.500; and nearly all the moving seasonality. However, the two by two eliminates slightly more of the irregular frequencies than the modified central weights. When choosing between the two averages, one then faces the following trade-off: to let the estimates contain more cyclical movements but also more irregularity or less cyclical movements and less irregularity. When a series is known to contain more cyclical movements (especially faster movements) than irregularity, the modifed central weights of the proposed average are certainly preferable to the two by two.

The dashed curve of Figure 2 represents the gain of the unmodified central weights of the semi-annual cyclical average, as obtained from Section 6. At the cyclical frequencies, its performance proves superior to that of the *two by two*; but, inferior to that of the modified central weights. For instance, the latter reproduces 101% of the 5-semestre waves (frequency 0.200) against 88% for the unmodified. The amplification of 5% (gain of 105%), at the 6-semestre (frequency 0.167) wave with the modified central weights, seems to us preferable to a comparable reduction of 6% (gain of 94%) with the unmodified set of weights. Indeed the analyst stands a better chance to detect an amplified signal in a series than a reduced signal.

b) Analysis of the end weights

Ideally, the gains of the end weights should be identical to the gain of the central weights. In such a case, end and central weights would have the same effect on the processed series (except for possible phase-shifts).

The gain of the weights for the second-last estimate (dotted line of Figure 3) is quite similar to the gain of the modified central weights (solid curve). Note that the former is more similar to the latter than to the set of unmodified central weights of Figure 2. This is the reason why we modified the weights.

The weights for the second-last estimate preserve the cyclical frequencies and eliminate stable seasonnality. However, they preserve 11 and 21% of the moving seasonality frequencies 0.467 and 0.483; and, even more of the noise frequencies. From the view point of the gain, the weights of the second-last estimate should yield less reliable estimates than the modified central weights.

Fig. 2

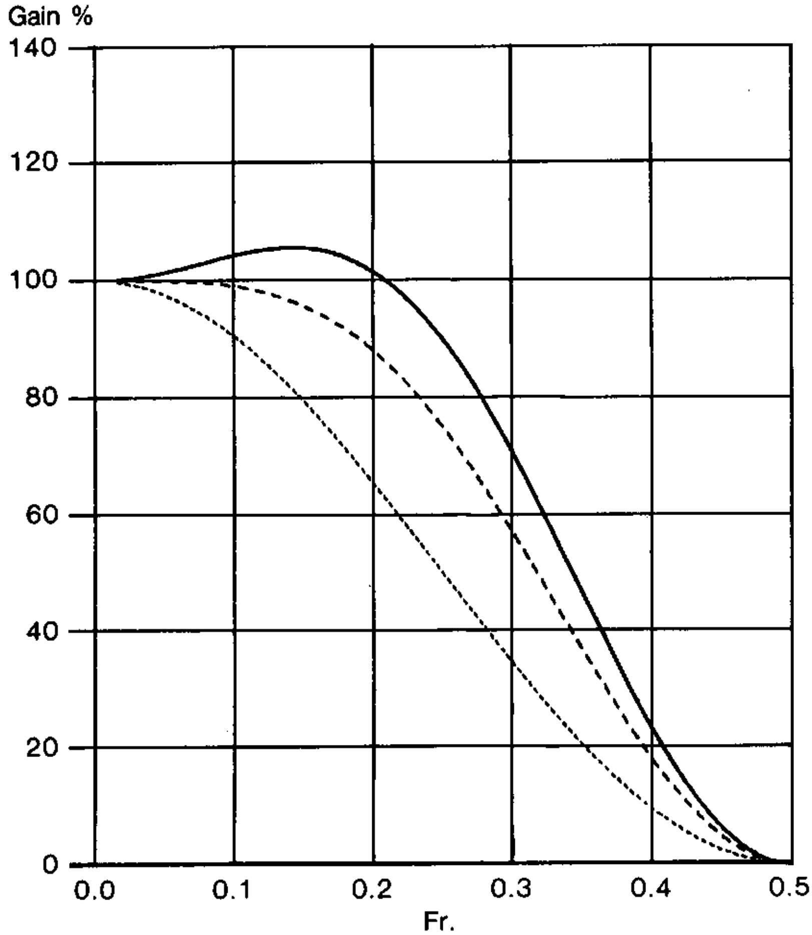

Figure 2. Gain functions of the modified (——) and non-modified (----) central weights of the proposed semi-annual cyclical average and of the *two by two* average (····)

The situation gets worse for the set of weights for the last estimate (broken line in Figure 3). Here, a strong amplification of some noise and fast cyclical frequencies is observed (gains reaching up to 137%). Caution should then be exrcised in interpreting the estimate yielded by these weights. One should perhaps disregard the last estimate completely, for series which are reputed to be irregular (containing those magnified frequencies).

As seen in Table 1, the end sets of weights are not symmetric. Consequently, they cause phase-shifts, which are compiled in number of semestres in Table 2 for certain selected frequencies. At the target cyclical frequencies, a small phase-shift is observed for the second-last weights. In this case, a cyclical wave of five semestres will be delayed by 0.09 semestres in the estimates; one of four semestres, by 0.16 semestres; and one of three semestres, by 0.28 semestres; etc.

The phase-shift reaches its maximum at the fundamental seasonal frequency (0.500). This does not matter however, since the frequency is totally eliminated by the weights. It does matter a little for the moving seasonality frequencies 0.467 and 0.483, since they are not completely eliminated.

Fig. 3

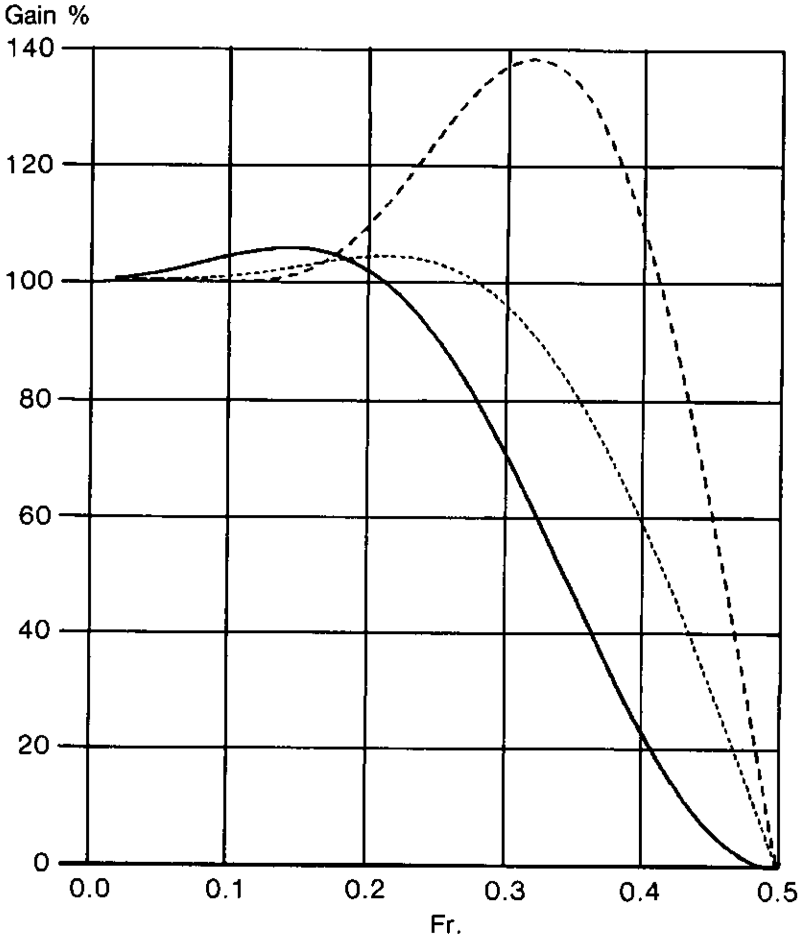

Figure 3. Gain functions of the modified central weights (——) and of the weights for the second-last (····) and the last (----) estimates

Table 2

Phase-shifts observed in number of semestres for the sets of end weights at certain selected frequencies

	second-last	last
cyclical frequencies:		
0.100 (10 semestres)	0.01	0.01
0.167 (6 semestres)	0.05	0.05
0.200 (5 semestres)	0.09	0.05
0.250 (4 semestres)	0.16	0.00
0.333 (3 semestres)	0.28	0.17
seasonal frequencies		
0.467	0.46	0.45
0.483	0.48	0.47
0.500 (2 semestres)	0.50	0.50

5. GRAPHICAL ANALYSIS OF END ESTIMATES

Figure 4 displays the preliminary estimates derived using the two sets of end weights for years 1968 to 1980, accompanied by the corresponding central final available estimates. Figure 4 a) shows the end estimates falling in the second semestre; and 4 a), in the first semestre. (One single plot would have been too crowded.)

If the central estimates are considered as true (or at least more reliable, the end estimates are seen to cause five false signals: in 1968 (arrow in fig. 4 b), in 1972 (4 a), in 1974 (4 b), in 1975 (4 a) and in 1976 (4 b). A false signal is said to occur here when the end estimates show a change in the direction of the trend-cycle and when that change is later contradicted by the central final estimates (becoming available with new observations). These false signals tend to appear when the series slows down in one direction and resumes the movement in the same direction. When there is a strong change of direction like in 1978, this does not seem to occur.

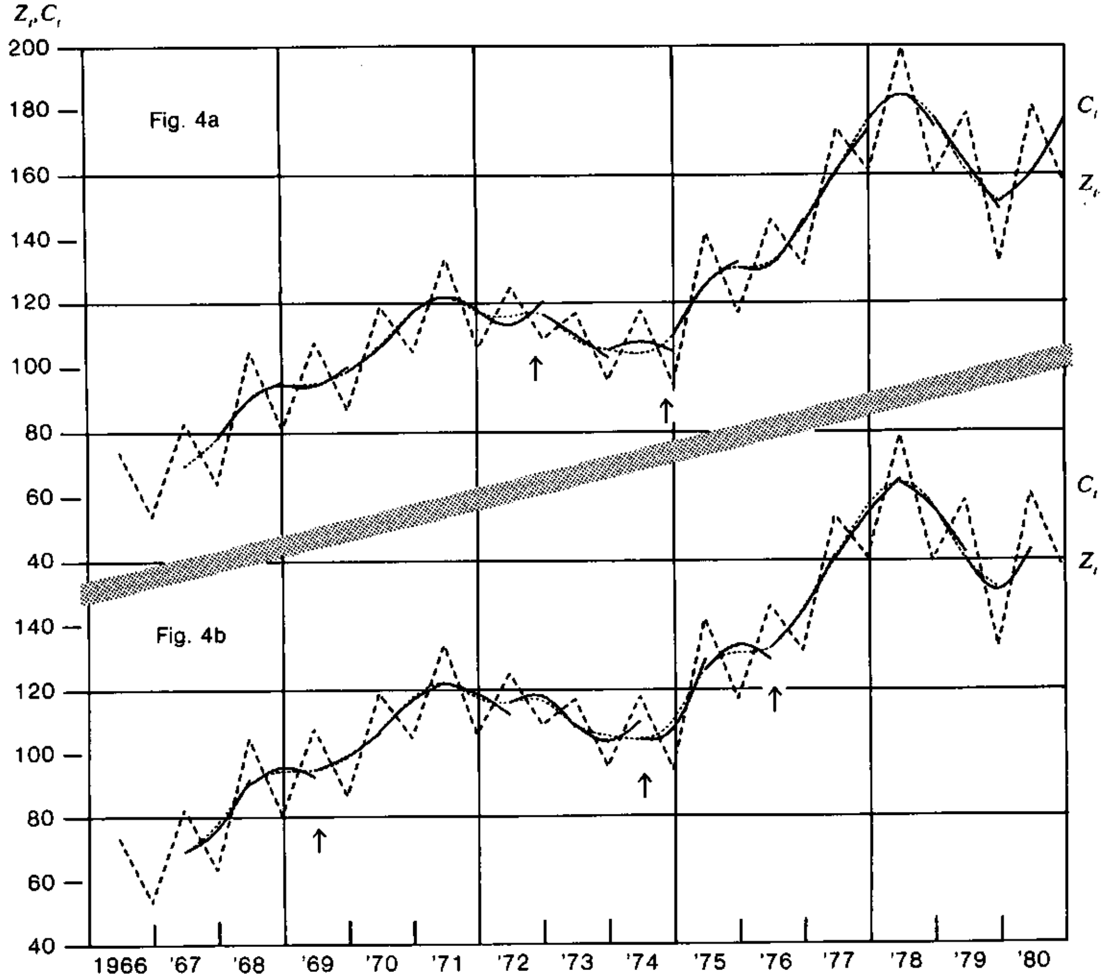

Figure 4. Semi-annual seasonal series (----); preliminary estimates of its trend-cycle by the ends weights (——) of the proposed semi-annual cyclical moving average a) for the second semestres and b) for the first semestres; final estimates (····) by the central weights of the average.

If the estimates derived by the last set of weights were omitted, many false signals would *disappear*. However, the estimates sould become less timely. This illustrates the statistician's dilemma between the timelyness and the reliability of estimates under any estimation method. (In practise, a serious analyst would wait for at least one confirmation of a signal before *believing* it).

Apart from the five false signals mentionned, the preliminary estimates display a movement which is very similar and sometimes undistinguishable from that of the final estimates.

5. CALCULATION OF THE WEIGHTS OF THE SEMI-ANNUAL CYCLICAL AVERAGE

The observed series z_t comprises the trend-cycle c_t to be estimated and a seasonal-irregular residual $s_t + e_t$ $(= z_t - c_t)$:

$$z_t = c_t + (s_t + e_t), \; t = 1, \ldots, 5. \tag{1}$$

Following the approach of Leser (1961 and 1963) and of Cholette (1980), the desired trend-cycle minimizes the quadratic sum of fourth differences (first term of (2)). On the five-semestre estimation interval, the component as much as possible approximates a time polynomial of the third degree. This specification allows for a full economic cycle with its four phases of expansion, turning-point, recession and recovery over the interval.

The seasonal-irregular residual $(z_t - c_t)$ minimizes the quadratic sum of first seasonal differences taken on corresponding semestres (second term of (2)). This specification means that the seasonal-irregular residual of one semestre should resemble that of the same semestre in the neighbouring year as much as possible.

Furthermore, the seasonal-irregular residuals minimize the quadratic sum of their sums on two consecutive semestres (third term of (2)). This criterion indicates that the seasonality of two neighbouring semestres should cancel out and that the irregularity should not affect the level of the desired trend-cycle.

The three criteria specified for the components combine into the following objective function:

$$\begin{aligned} f(c) &= \sum_{t=5}^{5} (c_t - 4c_{t-1} + 6c_{t-2} - 4c_{t-3} + c_{t-4})^2 \\ &+ \sum_{t=3}^{5} \{(z_t - c_t) - (z_{t-2})\}^2 + \sum_{t=2}^{5} \{(z_t - c_t) + (z_{t-1} - c_{t-1})\}^2 \end{aligned} \tag{2}$$

Equation (3) can be rewritten in linear algebra:

$$\begin{aligned} f(C) &= C'A'AC + (Z - C)'B'B(Z - C) + (Z - C)'F'F(Z - C) \\ &= C'HC + (C - Z)'G(C - Z) \end{aligned} \tag{3}$$

where A, B and C respectively stand for the matrix operators of quadruple differences, first seasonal differences and annual sums defined as follows:

$$A = [\,1 \;\; -4 \;\; 6 \;\; -4 \;\; 1\,], \quad B = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The normal equations associated with (3) read

$$dF/dC = 2HC + 2G(C - Z) = 0 \tag{4}$$

and imply solution:

$$C = (H + G)^{-1}GZ = WZ \tag{5}$$

The third central row of matrix W contain the non-modified central weights of the semi-annual cyclical average ot Table 1-B; and the fourth and fifth row, the end weights of rows 2 and 3 of Table 1-A.

7. HISTORY OF MOVING AVERAGES BY QUADRATIC MINIMIZATION

This approach of quadratic minimization originates with Whittaker (1923). Leser (1961, 63) showed how quadratic minimization could be applied to develop cyclical moving averages. Cholette (1980) proposed substitutes for the two by twelve and the two by four moving averages. These substitutes were incorporated into the Dagum (1980) seasonal adjustment programme as optional.

The semi-annual cyclical average presented in this paper could also be incorporated in a seasonal adjustment method of the X-11 type. This would allow seasonally adjusting semi-annual series and the calculation of the seasonal factors by means of the seasonal moving averages usually applied for monthly and quarterly series.

8. CONCLUSION

This paper presented a 5-term moving average which eliminates seasonality from semi-annual time series. The estimator reproduces the economic cycles more exactly than the two by two moving average. The two by two also has the disadvantage of not providing any estimate for the first and last semestres of the series.

REFERENCES

AKAIKE, H. and ISHIGURO, M. (1980). BAYSEA, a Bayesian Seasonal Adjustment Program. *The Institute of Statistical Mathematics*, Computer Science Monograph No. 13, Tokyo.

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

CHOLETTE, P.A. (1980). A Comparison of Various Trend-Cycle Estimators. *in Time Series Analysis*, (O.D. Anderson and M.R. Perryman Eds.), 77-87.

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

KOOPMANS, L.H. (1974). *The Spectral Analysis of Time Series.* Academic Press.

LAROQUE, G. (1977). Analyse d'une mthode de dsaisonnalisation: le programme X-11 du U.S. Bureau of the Census, version trimestrielle. *Annales de l'I.N.S.E.E*, 28, 105-127.

LESER, C.E.V. (1961). A Simple Method of Trend Construction. *Journal of the Royal Statistical Society* Ser. B, 23, 91-107.

LESER, C.E.V. (1963). Estimation of Quasi-Linear Trend and Seasonal Variation. Journal of the American Statistical Association, 58, 1033-1043.

MACAULY, F.R. (1931). *The Smoothing of Time Series.* National Bureau of Economic Research, Washington.

PHLIPS, L. BLOMME, R. (1973). *Analyse chronologique*, (Vander Eds.). Louvain, 334.

SCHLICHT, E. (1981). A Seasonal Adjustment Principle and a Seasonal Adjustment Method Derived from this Principle. Journal of the American Statistical Association, 76, 374-378.

SHISKIN, J., YOUNG, A.H, and MUSGRAVE, J.C. (1967). *The X-11 Variant of Census the Method II Seasonal Adjustment Program.* Technical Paper No. 15, U.S. Bureau of the Census.

WHITAKKER, E. (1923). On a New Method of Graduation. *Proceedings of the Edinburg Mathematical Society*, 41, 63-75.

Survey Methodology, December 1984
Vol. 10, No. 2, pp. 165–179
Statistics Canada

The Use of Matching in the Evaluation of Non-Sampling Errors in the 1981 Canadian Census of Agriculture

J. COULTER[1]

ABSTRACT

This paper discusses the use of matching between files of comparable data in the evaluation of non-sampling error. As an example of the technique, the data quality evaluation of the 1981 Canadian Census of Agriculture is described and some results presented.

KEY WORDS: Non-sampling error; Coverage; Response error; Matching; Record Linkage; Census of Agriculture.

1. INTRODUCTION

As the use of probability sampling in data collection has evolved, the evaluation and control of sampling errors has been a constant concern. Extensive research has been devoted to the design of sampling schemes which would reduce sampling error and facilitate its measurement. In many situations, however, major portions of the survey error arise not from sampling, but from the effects of other components of the data collection operation. In censuses particularly, in which data are obtained through 100 percent enumeration of the population of interest, sampling error is nonexistent. Instead survey error is due entirely to the influences of respondents, interviewers, coders, keyers, and others during the collection, capture, and processing stages of the survey operation. As the impact of these non-sampling errors on data quality has become more fully understood, the development of techniques to control and measure them has gained in importance.

2.. MODELS FOR SURVEY ERROR

Early papers on total survey error, such as that by Deming (1944), outlined the potential sources of error and discussed the need to consider their varying effects when planning data collection operations. As the study of survey error developed, general models were propose by Hansen et al. (1951), Sukhatme and Seth (1952), Hansen, Hurwitz, and Bershad (1961), and others to describe the components of sampling and non-sampling error. Studies were conducted on the correlations between errors which result from influences such as interviewers or coders, and methods were developed for measuring their effects. Fellegi (1964) presented a detailed model which included correlations between numerous error sources.

Other models have followed which consider both single and correlated non-sampling errors and propose methods for evaluating them. Some examples include the U.S. Bureau of the Census survey error model described by Nisselson and Bailar (1976), the discussion of measure-

[1] J. Coulter, Census Operations Division, Statistics Canada, 2nd Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

ment errors by Cochran (1977), and the model of survey error presented by Andersen et al. (1979) which was based on an earlier model by Kish (1965). In a recent paper Hartley (1981) described a model with terms for interviewer, coder, and respondent errors, and proposed a sample design to facilitate estimation of these errors.

Throughout the literature the components of non-sampling error have been categorized in a variety of ways. In this paper we will divide non-sampling error into two elements, coverage error and response error. A coverage error will be defined to have occurred when a unit which satisfies the definition of the universe of interest is missed or counted more than once, or a unit not belonging to the desired universe is included. Coverage errors cause the invalid inclusion or exclusion of all data for the incorrectly enumerated unit. As a result they may influence the estimates for any or all data items.

Response errors will be defined as affecting the values of individual items within the data for units which have been correctly included in the enumeration. They may have arisen at the initial collection of the data or during subsequent processing stages; potential sources include misinterpretation of a question by the respondent, total or partial non-response, the influence of the interviewer, and data capture or coding error.

3. EVALUATING NON-SAMPLING ERROR

In most survey error models, error is defined as the difference between the true value and the collected value for the particular data item. Thus in order to evaluate non-sampling error, one would in theory propose to compare the data collected by the survey with the true values for the item of interest. In practice, however, the true values are seldom known, if in fact they even exist. Instead the survey data must be compared to estimates from an alternate source which is believed to provide the closest available approximations to the true values.

In determining the data source most suitable to represent the unknown true values, a number of factors must be considered. The alternate data must be collected independently from those of the survey of interest. Optimally, the definitions and concepts employed in the collection of the data, and the reference periods to which the data applied, would be equivalent for the two sources. The universe covered by the alternate data would be the same as that of the survey, or comparable subuniverses would be identifiable. The purpose and methods of collection of the alternate data, and of any subsequent processing or updating stages, would be fully understood. Perhaps most important of all, the data would be of high enough quality to act as a standard against which the survey data could be compared.

In practice, of course, all of these conditions are seldom satisfied by a single alternate data source. In some cases one source may be best for a particular subpopulation while another is preferable for the remainder of the universe. Adjustments may be possible to remove the effects of differences in reference dates or definitions of variables between the two data sources. For the majority of cases, however, even the best estimator of the true values will involve major failures of some of these conditions, and their influence on the comparison may not be measurable or even identifiable.

Approximations to the true values may be obtained from a variety of sources. Estimates from one census or survey may be compared to those from another. Examples include the comparisons of the Current Population Survey and the U.S. Census of Population, the Labour Force Survey and the Canadian Census of Population, and the Agriculture Enumerative Survey and the Canadian Census of Agriculture (Statistics Canada 1979). Demographic projections have also been employed to approximate the true values, such as in the evaluations of the Labour Force Survey and the Canadian Census of Population described by Fellegi (1973).

Administrative data are also being used more and more in evaluation studies as respondent burden beccomes an increasing concern. Estimates from income tax, family allowance, motor vehicle licence, agricultural marketing board, and other files may provide approximations to the required true values. Income tax files, for example, are currently being studied at Statistics Canada for use in the collection and evaluation of farm income data.

As these few examples imply, a wide range of alternate data sources have been used as standards of comparison for survey estimates. However, while macro-level comparisons provide indications of the total error in the survey results, they cannot identify or measure the components of the error. The effects of coverage and response errors cannot be identified. Hence additional methods are required which facilitate more detailed investigations into the total error observed.

One technique, which can provide the analyst with a wealth of information on errors and their sources, links the survey results at the individual record level to a file of comparable data. Using the matched records, one-to-one comparisons can be made between the values reported for the survey and those from the alternate data source, which are assumed to represent the true values. Cross-classifications over other data items provide insight into the characteristics of units displaying certain types of inconsistencies. As well, the study of records which could not be matched can indicate areas of potential over or undercoverage by the survey.

In this paper we will consider the use of matching in the evaluation of survey non-sampling error. The strengths and weaknesses of the technique, and the types of studies which it makes possible, will be discussed. As an example of the use of matching, the data quality evaluation of the 1981 Canadian Census of Agriculture will be outlined and some results presented.

4. THE USE OF MATCHING

Study of the literature on the evaluation of total survey error reveals many studies which have made use of matching. As with the macro-level comparisons, a wide variety of comparable data sources have been employed. Post enumeration surveys, aimed at collecting data of a higher quality and in more detail than the original survey, have been conducted in a number of situations. Examples include reinterview by appraisers in a study on the reporting of the market value of homes (Kish and Lansing 1954), the post enumeration survey used to evaluate the U.S. Census of Agriculture (U.S. Bureau of the Census 1982), the Labour Force Survey reinterview program (Tremblay, Singh, and Clavel 1976), and the Vacancy Check Operation of the Canadian Census of Population and Housing (Statistics Canada 1980).

Independently-collected censuses and surveys have also been linked in order to evaluate data quality. Matches have been performed, for example, between the Labour Force Survey and the Canadian Census of Population (Krotki 1980), the Current Population Survey and the U.S. Census of Population (U.S. Bureau of the Census 1964), and the Agriculture Enumerative Survey and the Canadian Census of Agriculture (Statistics Canada 1979).

With the concern for reducing respondent burden, administrative data have been used increasingly in evaluation studies. Records of doctors and hospitals have been matched to health survey results by Andersen et al. (1979) and Horvitz (1981). Immigration and birth records have been employed in the Reverse Record Check Content Study of the Canadian Census of Population (Krotki 1980), and tax records of the IRS have been used to study response errors in the U.S. Census of Population (U.S. Bureau of the Census 1970). Other examples of linkage with administrative data include the evaluation of agricultural survey results by Faulkenberry and Tortora (1981) and the reporting of sensitive topics by Marquis, Marquis, and Polich (1981).

In general, the quality of an evaluation study based on a record linkage operation depends on two major factors:

1) the quality of the data on the alternate source to which the survey is matched, and
2) the uniqueness of the identifiers used for the match, and the accuracy of the match technique itself.

Firstly, by definition of the study objectives, the alternate data are to act as approximations to the true values. Random errors in the data which tend to cancel one another out over all records do not noticeably affect macro-level comparisons. However such errors can have a serious effect on studies conducted at the individual record level.

Some data bases chosen to act as standards for comparison may be assumed to be free from error. Birth records, for example, provide very accurate data on place of birth and age. Other data sets may be known to contain certain response or coverage errors, but if the errors are measurable they can be taken into account in the analysis, and the assumption of no error remains valid. In many cases, however, the comparable data are subject to errors which cannot be completely identified or measured. In these situations, the best approximation to the true values is provided by the data set which is least affected by error. Data which have been collected using more accurate methods or better trained staff than the survey of interest may be assumed to contain less error, and hence can provide a reasonable basis for comparison.

The second major factor affecting the quality of the study is a function of the match operation itself. In some situations, each member of the population will have been assigned a unique identification number which has been accurately stored on both files. Linking the records would then be a straightforward process of matching on these unique identifiers. At the other end of the scale, the best available identifiers may be non-unique characteristics, such as name, which are prone to the introduction of error during data collection or capture.

The linkage algorithm itself can also have an impact on the quality of the match, particularly when the keys or identifiers are less than perfect. The algorithm may tend to allow invalid matches between records with similar keys, or may prevent valid matches when the keys differ due to minor errors or omissions. The extent to which such errors occur can have a significant effect on the composition of the files of matched and unmatched records.

Other factors which affect the comparability of the two sets of data, as for macro-level studies, include differences in collection date and method, concepts and definitions, and reference period. Due to the greater detail of investigation and cross-classification over related variables which is involved in micro-levels studies, such differences can have a much greater impact on the analysis than for the macro-level comparisons.

In order to study the use of matching in the analysis of non-sampling error, we will now consider the example of the data quality evaluation of the 1981 Canadian Census of Agriculture. For this study, independently-collected agricultural data for macro and micro-level comparisons were provided by the Agriculture Enumerative Survey (AES) and the Farm Enumerative Survey (FES), annual probability surveys conducted by Statistics Canada.

5. COMPARING THE CENSUS AND SURVEYS

The Canadian Census of Agriculture was conducted on June 3, 1981, sharing field operations with the quinquennial Census of Population and Housing. Data were to be collected for every census farm in Canada, defined as any farm, ranch or other agricultural operation which received $250 or more from the sale of agricultural products during the twelve months prior to census day, or which had the potential to produce that value in the next twelve months. During drop-off of the population and housing questionnaire, the census representative was to ask at each household whether any member operated a farm or other holding which satisfied the above definition. If so, a Census of Agriculture form was left to be completed by the operator.

In order to improve coverage, results of the 1976 Census and subsequent agricultural surveys were used to identify farms which were major producers of one or more specified agricultural commodities. The census representatives then had to account for each of these "specified farms" located in the area to be enumerated.

The questionnaire, delivered prior to June 3 to the operator of each census farm, was to be completed by self-enumeration on census day. Items covered in the census included crops, livestock, land use, sales, expenses, and other areas of interest to the public and private sectors. (Further details on the methodology and content of the 1981 Census of Agriculture may be obtained from the publication Statistics Canada (1982).)

The Agriculture Enumerative Survey (AES) and Farm Enumerative Survey (FES) together covered the majority of Canada's agricultural land. The FES enumerated the Prairie provinces of Manitoba, Saskatchewan, and Alberta plus the Peace River district of British Columbia, and the AES covered the remainder of British Columbia and the provinces of Prince Edward Island, Nova Scotia, New Brunswick, Quebec, and Ontario. The survey universe consisted of agricultural holdings which satisfied the census farm definition described above. However it excluded types of organization which were of marginal economic influence, such as institutional farms, and areas which contain little or no agricultural activity, such as urban cores. In order to provide comparable universes for the evaluation, the census file was adjusted by removing operations of these types. The deletions consisted of only 2.8 percent of the farms and 1.8 percent of the total farm area from the complete census file.

The probability surveys collected data on the same major agricultural variables as the census, such as crops, livestock, land use, and operating expenses, and used similar concepts and definitions. Some differences existed in wording and format, and in the instructions on what to include or exclude, for particular questions. As will be indicated in the discussion of the results, the effect of these inconsistencies had to be taken into consideration when comparing data from the two sources.

The AES and FES were conducted on July 1, 1981, approximately one month after the June 3 census date. Some data, such as farm expenses for the previous year, were expected to be relatively unaffected by the difference in reference data. However, other items were more likely to change between June 3 and July 1. The effect was expected to be particularly significant for livestock items, due to the constant fluctuations in inventories caused by birth, deaths, purchases, transfers, etc. As a result, operators responding to the survey were asked to indicate the changes in numbers of cattle and pigs between June 3 and July 1. Evaluation indicated that, while the data obtained were of some use in reconciling the differences due to reference date, they were subject to high non-response and questionable accuracy. Hence, the comparison of these and any other variables which would tend to be influenced by date of response had to take into consideration the difference in reference dates between the census and survey.

The samples for the AES and FES were selected from an area frame of agricultural enumeration areas, supplemented by a list frame of farms which were major producers of certain important commodities. Data collection was performed by trained enumerators during a personal interview with the operator of each selected farm. Following the necessary processing stages, an estimation procedure was applied to scale the counts up to the level of estimates for the entire population of interest. (Further details on the sample design are found in Statistics Canada (1984) and Phillips (1978).)

The survey estimates were subject to the same types of non-sampling errors as those from the census. However, due to the concentration on a smaller number of holdings, and the improved control of operations which was thus possible, it was expected that these types of errors would have a lesser impact on the surveys. Hence the surveys provided acceptable approximations to the true data values. On the other hand, the survey estimates were affected by sampling error, which had to be taken into account when making comparisons with macro-level estimates obtained from the census.

5.1 Macro-level Comparisons

Prior to the evaluation using the matched file, estimates from the complete census and survey data files were studied. Since the two vehicles covered comparable universes, these macro-level comparisons for provinces and regions provided initial indications of census coverage. By comparing census point estimates with survey 95 percent confidence intervals for totals of livestock, crop acreages, and other items, areas of potential over or underestimation were identified. Further investigation of the macro-level differences was then initiated to determine if they were confined to particular categories of the items of interest. The macro-level studies, in addition, provided the experience and familiarity with the two sets of data which were required for the detailed analysis which followed.

As an example of the results of the macro-level comparisons, Table 1 presents estimates for Canada for the number of farms, total farm area, and land use. A significant difference between census and survey estimates was observed for total farm area in Canada, yet the size and direction of differences varried greatly among the component land use categories. Census estimates for classes of improved land differed from the survey estimate by as much as 25.5 ± 7.7 percent for other improved land to as little as -3.4 ± 2.2 percent for cropland. The

Table 1

Comparison of Census and AES-FES Estimates for Number of Farms, Area and Land Use (in thousands of acres), 1981, Canada[a]

Item	Census Estimate[b,c]	Survey Estimate[c]	Percent Difference[d]
Total number of farms	309,410*[e]	319,476	− 3.2 ± 2.6
Total area of farms	159,866*	175,543	− 8.9 ± 2.4
Improved land	112,390	114,610	− 1.9 ± 2.3
Cropland	75,532*	78,211	− 3.4 ± 2.2
Improved pasture	10,523*	9,460	11.2 ± 7.3
Summerfallow	23,827*	24,939	− 4.5 ± 3.7
Other improved land	2,509*	1,999	25.5 ± 7.7
Unimproved land	47,477*	60,933	−22.1 ± 4.3
Woodland	8,211*	17,751	−53.7 ± 3.9
Other unimproved land	39,265*	43,182	− 9.1 ± 6.5

[a] Excluding Newfoundland,Yukon and Northwest Territories.

[b] Excluding specified marginal areas and farms not belonging to the survey universe.

[c] Census and survey totals may not equal the sum of the components due to rounding. Survey estimates for Canada are based on a sample of 18,327 farms.

[d] Percent Difference = $\frac{\text{(Census Estimate - Survey Estimate)}}{\text{Survey Estimate}} \times 100$; the percent difference may not be consistant with the totals represented due to rounding. The indicated confidence interval, resulting from the sampling error in the survey, is equal to

$$\pm\ 2 \times \text{(survey coefficient of variation)} \times \frac{\text{census estimate}}{\text{survey estimate}}.$$

[e] An asterisk, identifying a significant difference between estimates, is indicated when the census estimate lies outside the survey 95 percent confidence interval.

major discrepancies in land area, however, were concentrated in the categories of unimproved land, particularly woodland. Further analysis into the reporting of woodland, which was prompted by these results, is discussed in section 5.5.

Macro-level comparisons also included the study of estimated frequency distributions prepared from the census and survey files. Distributions of the estimated number of farms over variables such as type of organization, land area, area of cropland, and sales were compared. Differences in the distributions identified possible over or undercounting of farms with particular characteristics.

Table 2 presents the census and survey frequency distributions by type of organization for the estimated number of farms in Canada. No significant differences were observed between the estimates for individual or family farms or corporations. However further study was initiated into the coverage of partnerships on the basis of the discrepancies noted for this category.

The limitation of the macro-level comparisons for evaluation of coverage was the inability to separate the effects of response errors from the effects of coverage errors. For example, the differences between census and survey estimates for improved land categories, shown in Table 1, seemed to exhibit too much variation in direction and magnitude to be the result of coverage errors alone. The discrepancies for woodland might also have been caused by factors other than coverage. Perhaps differences in field procedures or questionnaire format had resulted in inconsistencies between the census and surveys in the inclusion or exclusion of land of questionable agricultural value, or the classification of certain categories of land use. The micro-level match provided the needed mechanism for investigating these types of issues.

Table 2

Comparison of Census and AES-FES Estimates for Number of Farms by Type of Organization, 1981, Canada[a]

Type of Organization	Census Estimate[b]	Survey Estimate[c]	Percent Difference[d]
Total number of farms	309,410* [e]	319,476	− 3.2 ± 2.6
Individual or family farm	268,199	267,396	0.3 ± 3.0
Partnership			
- with a written agreement	11,160*	15,908	−29.8 ± 16.7
- with no written agreement	17,646*	22,855	−22.8 ± 10.8
Corporation	11,744	12,160	− 3.4 ± 10.4
Other type of organization	661*	1,142	−42.1 ± 13.0

For footnotes, see Table 1.

5.2 The Micro-level Match

Past experience with other agricultural censuses and surveys has indicated that even the most careful attention to quality cannot entirely prevent response errors. Despite all attempts to provide clear, unambiguous questions, problems such as differing interpretations of certain agricultural terms across regions of Canada, or a lack of consensus on the appropriate classification for certain types of land use, influence the data collected. Misinterpretation is particularly common for items which are of marginal economic or agricultural value,or which do not apply to most respondents. The micro or record level match with the AES-FES files provided the means to evaluate the impact of response errors on the 1981 Census of Agriculture.

The match between the Census of Agriculture and the AES-FES was based on the operator name, address, telephone number, and postal code for each holding. The link was performed in thirteen stages, each requiring a match on a different combination of the identifiers or their components. At each stage of the procedure survey records which has not yet been matched were identified, and the census file was searched by computer to locate the corresponding records. For each survey holding, the specified matching variables or keys were compared character by character with those of the census records which had not yet been linked. A match was identified if all characters of the matching variables were equal. At the Canada level, a computer match rate of 75.7 percent was achieved for the 18,327 survey records.

It was inevitable that, for a certain number of survey records, no census farm would be identified by the computer. Discrepancies in spelling of names and addresses, which had arisen during collection or capture of the census or survey data, prevented links in many cases. For example, J. Smith might have been reported on the census as opposed to J. Smyth on the survey, James Smith as opposed to Jim Smith, or St. Catherines rather than St. Catharines. Partnerships or corporations for which one operator had responded on the census but a different partner or manager had been interviewed by the survey could not be matched by a computer link on operators. Similarly, records for holdings which had changed operators between the census and survey collection dates could not be linked by computer.

In order to improve the match rate, and eliminate the possible biases in the matched file which might have resulted from those operations which could not be linked by computer, a manual resolution process was initiated. Using additional data from the questionnaires, such as corporate or farm name, addresses and names of partners, land description of the holding, and comments, clerical staff attempted to identify the corresponding census farm for each unmatched survey record. Of the 4,459 unmatched survey farms which remained following the computer link, 3,228 were matched during the manual resolution process. At its completion, 93.3 percent of the total 18,327 AES and FES records for Canada had been linked to census operations.

With further input of time and resources, it may have been possible to link some of the remaining 6.7 percent of the survey records to the census data base. However, in many cases the needed identifiers has not been collected on either the census or survey, and would have required investigation of administrative records or contact with the operators themselves. It was felt that the possible benefits were not sufficient to warrant the expenditures required, and no further manual resolution was attempted.

The studies which were facilitated by the record linkage can be grouped into two main types, those based on the unmatched survey records and those using the matched census-survey pair.

5.3 Studies of the Unmatched Records

In order to study the characteristics of census undercoverage, the unmatched records were assumed to be representative of the farms which should have been enumerated by the census but were missed. It was known that the unmatched records overestimated the number of missed farms, due to certain conditions of the data sources and the matching algorithm. For example, it was probable that some records on the survey file had been covered in the census, but could not be matched by either computer or manual means due to missing or invalid name or address data.

Because of the resulting potential for overestimation of the land and commodities missed by the census, one had to proceed with caution in using the estimates produced from the unmatched records.Nonetheless the Canada level estimates were most valuable as initial indicators of the characteristics of the farms which were underenumerated by the census.

In the first stage of the study, sample expansion factors were applied to the unmatched records to produce commodity estimates for the "missed farms". These were then compared to the commodity estimates for the entire survey universe, and the fraction of the total estimate which was accounted for by the "missed farms" was calculated. This fraction was then compared with the fraction which the missed farms comprised of the total estimated number of farms. For example, Table 3 shows that the unmatched file contained only 4.4 percent of the estimated total farm area and 3.9 percent of the cropland, whereas it was responsible for almost 9.7 percent of the total estimate of farms. These results provided an initial indication that the "missed farms" were not representative of the complete universe, but were smaller than average in terms of land area and other characteristics. This implied that the extent of undercoverage could not be measured by the number of farms missed alone. Instead, the characteristics of the missed farms over particular commodities had to be considered.

To provide further insight into the characteristics of undercoverage, frequency distributions of the estimated number of missed farms were prepared over classes of land area, sales, livestock, and other commodities. Comparison with similar frequency distributions for the entire survey universe indicated that the missed farms had a higher proportion of holdings with small acreages and low sales. It can be seen from Table 4, for example, that 42.3 percent of the estimated missed farms reported less than 70 acres of total farm area, as compared with 15.8 percent of the complete survey population. Sales of less than $1,200 were reported by an estimated 27.7 percent of the missed farms, but only 7.3 percent of the survey universe.

The frequency distributions were also compared by considering the ratio of the unmatched estimate to the estimate for the entire survey universe, that is, the fraction of the total survey estimate accounted for by the unmatched farms. As shown in Table 4, the unmatched file contained 36.5 percent of the estimated farms with less than 10 acres of land, the smallest size range presented as comapred with only 4.3 percent of those in the largest range of 760 acres or more. Similarly 36.8 percent of the estimated farms with sales less than $1,200 were obtained

Table 3

Comparison of AES-FES Estimates for Total Farms and Unmatched Farms, Area and Land Use (in thousands of acres), 1981, Canada[a]

Item	Estimate from Total AES-FES File[b]	Estimate from Unmatched AES-FES Records[c]	Percent of the Total Estimate Accounted for by the Unmatched Farms
Number of Farms	319,476	30,975	9.7
Total Farm Area	175,543	7,768	4.4
Total Improved Land	114,610	4,502	3.9
Cropland	78,211	2,792	3.6
Improved Pasture	9,460	603	6.4
Summerfallow	24,939	1,004	4.0
Other Improved Land	1,999	104	5.2
Total Unimproved Land	60,933	3,266	5.4
Woodland	17,751	1,325	7.5
Other Unimproved Land	43,182	1,941	4.5

[a] Excluding Newfoundland, Yukon and Northwest Territories.
[b] Survey estimates are based on a sample of 18,327 farms.
[c] The unmatched file contained 1,231 farms.

Table 4

Percentage Distribution of AES-FES Estimates for Total Farms and Unmatched Farms by Total Farm Area and Total Value of Agricultural Products Sold During 1980, 1981, Canada[a]

Item	AES-FES Estimate of Total Number of Farms[b]	AES-FES Estimate of Number of Unmatched Farms[c]	Percent of the Total Estimate Accounted for by the Unmatched Farms
	Cumulative Percent	Cumulative Percent	Percent
Total Farm Area			
Under 10 acres	3.5	13.2	36.5
10 - 69 acres	15.8	42.3	22.9
70 - 399 acres	62.8	85.0	8.8
400 - 759 acres	78.4	90.5	3.4
760 acres and over	100.0	100.0	4.3
Total Value of Agricultural Products Sold			
Under $1,199	7.3	27.7	36.8
$ 1,200 - $ 2,499	12.3	41.2	26.5
2,500 - 9,999	29.6	65.4	13.5
10,000 - 49,999	67.8	88.3	5.8
50,000 and over	100.0	100.0	3.5

[a] Excluding Newfoundland, Yukon and Northwest Territories.
[b] Survey estimates are based on a sample of 18,327 farms.
[c] The unmatched file contained 1,231 farms.

from the unmatched file, compared to 3.5 percent of those with sales of $50,000 or more.

In summary, the results of the study of unmatched records were able to provide concrete evidence that the holdings missed by the census tended to be smaller than average in terms of agricultural production and value, a theory which had been widely held but not proven.

5.4 Studies Using the Matched File

The matched file was composed of agricultural holdings for which census and survey records could be linked by the computer and manual processes described. Since these census and survey values were assumed to have been collected from the same set of holdings, the effects of coverage differences were removed. In addition, the influence of imputation was lessened by excluding records for which census or survey data had been entirely imputed due to non-response. Hence the nature and extent of potential response differences between the two data collection vehicles could be studied. Prior to discussing the results of the matched record studies, however, a number of limitations which existed within the matched file, and which influenced the evaluation process, should be described.

Although every effort was taken to lessen the chance of spurious matches, it is possible that a small number of survey farms may have been linked to the wrong census holding due to similarities in name or address. In the case of extremely large agricultural operations which made a significant contribution to provincial commodity totals or land areas, linkage to the wrong census operation could noticeably skew the results. A detailed study of a sample of matched records, undertaken to determine the quality of the computer link, had not been completed at the time of the evaluation. As a result, the potential influence of spurious matches had to be considered when studying results from the matched file.

The second limitation on the matched file analysis affected the comparison of total counts from the census and survey data. The matched records consisted of a subset of the non-self-weighting sample of the AES and FES, since they contained only the survey farms which could be linked to census records.It would have been preferable to apply sample expansion factors to produce weighted estimates for the matched file. However, the expansion factors for the area frame were calculated using reported land use values from the survey, and it was as yet undetermined whether the factors were valid when applied to census data from the matched file. Census-related expansion factors could not be calculated using the census data, since one component of the factor, the farm area inside the selected segment of land, was collected only by the survey. As a result of this uncertainty regarding the application of survey expansion factors to census matched data, it was decided to restrict the analysis to unweighted census and survey totals from the matched file. (Study of the use of weighted estimates from the matched file was underway at the time of writing, but no conclusions had yet been reached.)

Despite the limitations of a non-self-weighting sample and the possible existence of spurious matches, the matched file proved to be a valuable evaluation tool. When matched totals identified discrepancies between census and survey values, further detailed investigations were undertaken into the possible causes of the observed differences.

As an example of the use of the unweighted matched totals, Table 5 presents counts for total farm area and categories of land use at the Canada level. The results indicate that less land was reported on the census than the survey for all land use items except improved pasture and other improved land. Relative differences between census and survey totals were smallest for items such as cropland, which are of major economic value and hence are clearly defined and seldom misunderstood by farm operators. Items of more marginal agricultural and economic value, however, tended to display greater discrepancies. The largest relative differences were observed for the category of woodland. In order to demonstrate some of the detailed evaluation techniques made possible by the matched file, further results of the study of data on woodland will be discussed.

Table 5

Comparison of Census and AES-FES Totals for Matched Farms, Land Use (thousand of acres), 1981, Canada[a]

Item	Census Total[b]	Survey Total[b]	Percent Difference[c]
Total area of farms	13,059	14,091	−7.3
Improved land	8,798	8,801	--
Cropland	6,046	6,167	−1.9
Improved pasture	804	682	18.0
Summerfallow	1,777	1,816	−2.1
Other improved land	170	137	24.3
Unimproved land	4,261	5,291	−19.5
Woodland	523	1,102	−52.5
Other unimproved land	3,737	4,189	−10.8

[a] Excluding Newfoundland, Yukon and Northwest Territories.

[b] Records for which census or survey data were entirely imputed have been excluded leaving 16,388 matched farms. Census and survey totals may not equal the sum of the components due to rounding.

[c] Percent Difference $= \frac{\text{Census total} - \text{Survey total}}{\text{Survey Total}} \times 100$

5.5 Detailed Comparisons of the Matched Records

While the matched totals provided a measure of the overall biases in reporting, they masked detailed information on where and why the differences occurred. For instance, was the difference in woodland caused by large discrepancies in only a handful of holdings, or was it consistent across all records? Did the response differences vary in magnitude or direction across types of operations or regions of the country? By comparing census and survey responses at the individual record level, the characteristics of the reporting differences were studied in detail.

Table 6 presents an example of the type of investigation facilitated by the matched evaluation file. Only those holdings for which a non-zero value of woodland was reported on either the census or survey are included. The operations are classified by the size and direction of the difference between the census and survey values for the item, and cross-classified by the amount of woodland reported on the census.

The table indicates that less woodland was reported on the census than on the survey for the majority of holdings. Differences tended to be small, clustered in the 1 to 50 acres and 51 to 150 acres ranges even for operations with large amounts of woodland on the census. Of note also were the 3,177 holdings, or 34.3 percent of the universe of interest, for which the census value of woodland was zero but the survey value was greater than zero. This is in contrast to the 807 holdings, or 8.7 percent of the universe, in which the opposite case of zero acres of woodland on the survey but greater than zero acres on the census was observed. Examination of these results suggested that the census and surveys may not have been obtaining measures of the same quantity, and prompted further study into their collection methodologies and questionnaire formats.

On the census questionnaire the respondent was asked to report the area of woodland, with further instructions in the census representative's manual indicating that only land "with seedlings or trees which had or would have value as timber, fuelwood, or Chirstmas trees" be included. In contrast the AES and FES interviewers instructed respondents to "include woodlots, cut-over land, etc." with no additional instructions that the land be of present or future commercial value. As well, woodland was requested twice on the surveys, once immediately after the reporting of total area, and later as a component of land use, and thus received greater emphasis.

Table 6

Comparison of Census and AES-FES Responses for Matched Records, Difference in Woodland by Census Value of Woodland, Canada[a], 1981

Difference: Census Woodland - Survey Woodland (acres)	Census Value of Woodland (acres)								Percent of Reporting Farms
	0	1 to 2	3 to 9	10 to 69	70 to 239	240 to 399	400 or more	Total	
	Number of Reporting Farms[b]								
less than −500	241	0	3	16	16	6	11	293	3.2
−500 to −251	248	0	4	24	21	12	7	316	3.4
−250 to −151	268	4	1	27	44	9	4	357	3.9
−150 to −51	715	5	30	167	145	24	19	1,105	11.9
−50 to −1	1,705	59	277	1,053	370	57	28	3,549	38.3
0	–	26	165	481	151	21	21	865	9.3
1 to 50	–	55	229	1,288	492	53	30	2,147	23.2
51 to 150	–	–	–	50	294	45	32	421	4.5
151 to 250	–	–	–	–	65	25	17	107	1.2
251 to 500	–	–	–	–	–	33	42	75	0.8
greater than 500	–	–	–	–	–	–	36	36	0.4
Total	3,177	149	709	3,106	1,598	285	247	9,271	100.0

[a] Excluding Newfoundland, Yukon, and Northwest Territories.
[b] Including all operations which reported woodland on either the census or survey. Excluding records for which either the census or survey data were totally imputed due to non-response.

As a result of these differences, it was believed that certain areas of woodland of questionable commercial value, which were reported on the survey, may have been excluded from the census. As well, some areas may have been reported on the census under different categories of land use, such as other unimproved land. Study continues into these and other hypotheses, using the matched file to investigate possible causes of the observed response differences.

When summary tabulations from the matched file failed to suggest causes for observed biases, the study of individual records which displayed large discrepancies between census and survey values for the items of interest was often informative. By comparing census and survey responses for other related items, it was sometimes possible to identify misclassification between categories or other causes of reporting differences.

Table 7 shows a number of variables for a record with one of the largest differences between census and survey values for total farm area. In this case, the discrepancy in total area was

Table 7

Census and Survey Recorded Values for a Particular Matched Record

	Total Farm Area (acres)	Land Use (acres)						Total Cattle and Calves
		Cropland	Improved Pasture	Summerfallow	Other Improved Land	Woodland	Other Unimproved Land	
Census Values	2,640	1,035	0	1,240	15	0	350	920
Survey Values	17,000	1,010	0	970	20	0	15,000	815

concentrated in the category of other unimproved land; only minor differences existed between responses for cropland, summerfallow, and other improved land. It appears that most of the land which was reported as other unimproved on the survey was excluded from the census response. Similar studies of other individual records identified other potential discrepancies in the reporting of unimproved land. Theories developed from these studies of individual cases were then tested on the entire file to determine if they might apply in general.

6. CONCLUSION

The link between the 1981 census and survey files was a powerful tool for the evaluation of census coverage and response errors. Errors were known to exist in the survey data which were being used as approximations to the true values, and limitations of the linkage operation were known to have caused spurious matches and unmatches. However, the matched file was still a valuable source of data for investigation of quality concerns. Studies based on the record level matches broadened the initial results obtained from the macro-comparison of census and survey estimates. In addition they brought individual problems into focus by allowing detailed investigation of particular aspects.

The evaluation produced valuable results on census undercoverage. Studies based on the unmatched survey records showed that the holdings missed by the census were, in general, smaller than average in terms of total land area, livestock, and value of agricultural products sold. Thus concrete evidence was provided to support the widely-held theory that the census tended to miss holdings of marginal economic and agricultural value.

Studies of response differences for land use identified categories in which discrepancies were concentrated, tendencies for confusion between certain classes, and variations in differences among regions of the country. Possible revisions to questionnaire format and wording, or collection methods, have been considered as a result of the study. Some of the variations are known to have been caused by difficulties in defining certain land use categories, due to the lack of clarity in the concepts themselves. Problems such as these, that result from confusion in the minds of the respondents as to which land should be reported under which category of usage, may never be completely solved. However, the recognition of the existence of a problem, and the study of the characteristics of its occurrence and its effect on the data, are very valuable contributions to future planning.

The 1981 data quality evaluation had far-reaching effects for the census. In response to its main goal, the study identified quality concerns in the 1981 census data. A publication (Statistics Canada 1984) was prepared to provide users with an indication of data quality, and to advise them with respect to particular problems which had a noticeable impact on the data. Looking further ahead, the evaluation has served as input to the planning of 1986 census procedures. By identifying items for which coverage or response errors occurred in 1981, the study has provided a list of areas requiring further consideration of collection and processing methods.

The impact of the data quality evaluation was not restricted to the Census of Agriculture alone. The comparison of census and survey responses also identified problem areas in the other data collection vehicles. Improvements to the National Farm Survey, the annual probability survey which has replaced the AES and FES, may result from the census study.

A further benefit of the study is the knowledge gained on the use of record linkage for evaluation purposes. The experience in matching data at the individual record level, using both computer and manual means, could provide valuable input to other linkage projects. In particular, knowledge of the problems encountered, their causes, characteristics, and possible solutions, could result in improved procedures for other studies.

In summary, the 1981 Census of Agriculture Data Quality Evaluation project has provided further evidence of the power of matching in the study of non-sampling error. The investigations using matched and unmatched records, and macro and micro-level comparisons, have produced measures of quality of the 1981 census data, and identified items for which errors have impacted significantly upon the results. As an extension of the original project mandate, input was provided to the planning of the 1986 and subsequent censuses, by indicating areas in which further research into possible changes in methodology was required. The evaluation also identified potential problems in the surveys used for comparison, thereby contributing to the planning of future vehicles for the collection of agricultural data. Finally, the study provided valuable experience and insight into the application of record linkage techniques for data quality evaluation.

ACKNOWLEDGEMENTS

The author would like to thank D. Royce, M.R. Dibbs, B.N. Chinnappa, K. Thatcher, the reviewer, and the other members of Statistics Canada whose support and valuable comments were most appreciated during the writing of this paper.

REFERENCES

ANDERSEN, R., KASPER, J., FRANKEL, M.R., and associates (1979). *Total Survey Error*. San Francisco: Jossey-Bass Publishers.

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

DEMING,W.E. (1944). On Errors in Surveys. *American Sociological Review*, 9, 359-369.

FAULKENBERRY, D., and TORTORA, R.D. (1981). Non-sampling Errors in an Agriculture Survey. *1981 Proceedings of the Section on Survey Research Methods, of the American Statistical Association*, 493-495.

FELLEGI, I.P. (1964). Response Variance and its Estimation. *Journal of the American Statistical Association*, 59, 1016-1041.

FELLEGI, I.P. (1973). The Evaluation of the Accuracy of Survey Results: Some Canadian Experiences. *International Statistical Review*, 41, 1-14.

GOSSELIN, J.-F., CHINNAPPA, B.N., GHANGURDE, P.D., and TOURIGNY, J. (1978). *A Compendium of Methods of Error Evaluation in Censuses and Surveys* (Catalogue 13-564). Ottawa, Canada: Statistics Canada.

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.

HANSEN, M.H., HURWITZ, W.N., and BERSHAD, M. (1961). Measurement Errors in Censuses and Surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

HARTLEY, H.O. (1981). Estimation and Design for Non-sampling Errors of Surveys. In *Current Topics in Survey Sampling*, ed. D. Krewski, R. Platek, and J.N.K. Rao. New York: Academic Press.

HORVITZ, D.G. (1981). Response Error Research Issues in Health Surveys. *1981 Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 397-399.

KIBLER, W.E. (1978). Controlling Non-sampling Errors in Surveys. Summary Report of the 29th Federal-Provincial Committee on Agricultural Statistics, Statistics Canada..

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

KISH, L., and LANSING, J.B. (1954). Response Errors in Estimating the Value of Homes. *Journal of the American Statistical Association*, 49, 520-538.

KROTKI, K. (1980). *Response Error in the 1976 Census of Population and Housing*. Working Paper, Ottawa, Canada: Minister of Supply and Services Canada.

MARQUIS, K.H., MARQUIS, M.S., and POLICH, J.M. (1981). Survey Responses to Sensitive Topics. *1981 Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 339-341.

NISSELSON, H., and BAILAR, B.A. (1976). Measurement, Analysis, and Reporting of Nonsampling Errors in Surveys. *Proceedings of the 9th International Biometric Conference*, 2, 201-322.

PHILLIPS, J. (1978). 1979 Farm Expenditure Survey Design and Estimation Procedures. Working Paper, Institutional and Agriculture Survey Methods Division, Statistics Canada.

STATISTICS CANADA (1979). *1976 Census of Canada - Agriculture - Evaluation of Data Quality* (Catalogue 96-872). Ottawa, Canada: Minister of Supply and Services Canada.

STATISTICS CANADA (1980). *1976 Census of Canada - Quality of Data - Series I: Sources of Error - Coverage* (Catalogue 99-840). Ottawa, Canada: Minister of Supply and Services Canada.

STATISTICS CANADA (1982). *1981 Census of Canada - Agriculture* (Catalogue 96-901). Ottawa, Canada: Minister of Supply and Services Canada.

STATISTICS CANADA (1984). *1981 Census of Canada - Agriculture - Evaluation of Data Quality* (Catalogue 96-918). Ottawa, Canada: Minister of Supply and Services Canada.

SUKHATME, P.V., and SETH, G.R. (1952). Non-sampling Errors in Surveys. *Journal of the Indian Society of Agriculture Statistics*, 4, 5-41.

TREMBLAY, V., SINGH, M.P., and CLAVEL, L. (1976). Methodology of the Labour Force Survey Re-interview Program. *Survey Methodology Journal*, 2, 43-62.

U.S. BUREAU OF THE CENSUS (1964). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics as Measured by the CPS - Census Match*. Series ER60, No. 5, Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1970). *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Record Check Study of the Accuracy of Income Reporting*. Series EDR60, No. 8, Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1982). *1978 Census of Agriculture Volume 5 Special Reports - Part 3 Coverage Evaluation*, (AC78-SR-3). Washington, D.C.: U.S. Government Printing Office.

Survey Methodology, December 1984
Vol. 10, No. 2, pp. 181-185
Statistics Canada

Unbiased Estimation of Domain Parameters in Sampling without Replacement

ARIJIT CHAUDHURI AND RAHUL MUKERJEE[1]

ABSTRACT

A finite population of size N is supposed to contain M (unknown) units of a specified category A (say) constituting a domain with mean μ. A procedure which involves drawing units using simple random sampling without replacement till a preassigned number of members of the domain is reached is proposed. An unbiased estimator of μ is also derived. This is seen to be superior to the corresponding possibly biased estimator based on a comparable SRSWOR scheme with a fixed number of draws. The proposed scheme is also shown to admit unbiased estimators of M and the domain total T.

KEY WORDS: Domain estimation; Simple random sampling without replacement.

1. INTRODUCTION

In large scale sample surveys, utilization of available resources and consideration for efficiency often demand realization in a sample of adequate representation from a specified category (A, say) of members with required characteristics. For example, clients and users of survey data may insist on estimates from a sample with a specified (m, say) number: (1) of farmers (i) using a particular fertilizer, (ii) employing a particular irrigation and cultivation technique and (iii) ready to respond truthfully to queries made; (2) of manufacturers using iron and steel with a specific purpose; (3) of household members with a requisite academic qualification, etc. While designing a sampling plan for the purpose, in spite of careful efforts, it is often possible that 'frames' may not be accurately constructed. The faulty list may be supposed to include N units which are well in excess over the M genuine units of the required A-category. Hence arises a problem of sampling to yield estimators for the mean, (and also total and size), of the domain of A-members. A solution to this problem is attempted below using 'inverse' SRSWOR scheme. Inverse sampling plans with replacement are, however, available in the literature (vide Haldane 1945, Sampford 1962 among others) for estimating the proportion $f = M/N$ of domain elements. Domain estimators for μ are also given by Rao (1975) but they are ratio estimators and are not unbiased. The proposed inverse SRSWOR scheme is seen to admit an unbiased estimator of μ which is more efficient than the corresponding possibly biased estimator based on a comparable SRSWOR scheme with the fixed number of draws.

2. A METHOD OF SAMPLING AND ESTIMATION

The population $I_N = (1, ..., j, ..., N)$ is supposed to consist of N units labelled $1, ..., j, ..., N$ and valued $y_1, ..., y_j, ..., y_N$. Of them some M (unknown) units possess certain exclusive features to constitute a class or domain, say A. In practice, some idea about M is usually available and let the parameter space for M be $\mathcal{M} = \{r, r+1, ..., R\}$, where $r(\geq 1)$ and $R(\leq N)$ are known. In almost all real life situations r will be much greater than 1 and R much less than N.

[1] Arijit Chaudhuri and Rahul Mukerjee, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India.

Writing $X_1, ..., X_i, ..., X_M$ as the y_j values for the M units of A, estimators are required for $\mu = (\Sigma_1^M X_i)/M$, and perhaps also for M and $T = \Sigma_1^M X_i$, from a sample containing a preassigned number, say m ($\leq r$), of units of A. The expressions for the variances of these estimators, presented later as functions of m, may be employed for an appropriate choice of m. For convenience, we shall write $X_{M+1} = \ldots = X_N = 0$ for the 'non-A' units of I_N.

Let units be chosen in successive draws by SRSWOR till exactly m units of A are realized. The number of draws, u, is then a random variable with a probability distribution $P_M(\cdot)$ (say, depending on the unknown parameter M) which is given by

$$P_M(u=n) = \frac{\binom{M}{m-1}\binom{D}{n-m}}{\binom{N}{n-1}} \cdot \frac{M-m+1}{N-n+1} = g_{Mn} \text{ (say) } (m \leq n \leq D+m), \tag{2.1}$$

where $D = N - M$. To avoid trivialities, hereafter, we shall make the reasonable assumption that m $\geq$ 2. Then the following results hold for the above inverse sampling scheme.

Lemma 2.1. Every parametric function $f(M)$ is unbiasedly estimable.

Proof. Let $h(u)$, if available be an unbiased estimator (UE) of $f(M)$. Then

$$f(M) = \sum_{n=m}^{D+m} h(n)\, g_{Mn},\; r \leq M \leq R. \tag{2.2}$$

If the above system of $R - r + 1$ equations in $N - r + 1$ unknowns $h(m), ..., h(N - r + m)$ be written in matrix notation, then the fact that $g_{Mn} > 0$ ($m \leq n \leq D + m, r \leq M \leq R$) implies that the resulting coefficient matrix is of full row rank. This guarantees the existence of a solution and completes the proof.

Remark. In particular, if $R = N$ then the number of equations in (2.1) equals the number of unknowns. As such the coefficient matrix becomes nonsingular and every parametric function $f(M)$ becomes uniquely unbiasedly estimable.

Corollary 2.1. A UE of M based on u is $\hat{M}(u) = N(m - 1)/(u - 1) = \hat{M}$ (say).

Proof. First observe that the assumption $m \geq 2$ ensures that $u > 1$ with probability 1 (whatever be M) so that $\hat{M}(u)$ is well defined. Now

$$E\left(\frac{1}{u-1}\right) = \sum_{n=m}^{D+m} \frac{1}{n-1} g_{Mn}$$

$$= \frac{M!\, D!}{(M-m)!\,(m-1)!\, N!} \sum_{n=m}^{D+m} \frac{(n-2)!\,(N-n)!}{(n-m)!\,(D-n+m)}$$

$$= \frac{M!\, D!}{(M-m)!\,(m-1)!\, N!} \cdot \frac{(M-m)!\,(m-2)!\,(N-1)!}{(M-1)!\, D!} = \frac{M}{N(m-1)},\; \forall\, M \in \mathcal{M}.$$

Hence the result.

Remark. The relation (2.1) and Lemma 2.1 may be employed to find $V_M(\hat{M})$ and a UE of this variance. The resulting algebraic expression, although straightforward to evaluate numerically in any practical situation, are somewhat involved and will not be presented here.

In the following, $S^2 = (M - 1)^{-1} \Sigma_1^M (x_i - \mu)^2$, $q(u)$ and $l(u)$ are any UE's for M^{-1} and M^2 respectively (available by (2.2) above) Σ' denotes summation over the A-units included in the sample, $\bar{x} = m^{-1}\Sigma' X_i$, $Z = m^{-1}\Sigma' X_i^2$ and $s^2 = (m - 1)^{-1} \Sigma'(X_i - \bar{x})^2$.

Theorem 2.1. A UE of μ is $\bar{x}$ with $V_M(\bar{x}) = S^2(1/m - 1/M)$. A UE of $V_M(\bar{x})$ is given by $\nu(\bar{x}) = s^2(m^{-1} - q(u))$.

Proof. Easy and hence omitted.

Theorem 2.2. (i) A UE of T is $\hat{T} = \hat{M}\bar{x}$ with

$$V_M(\hat{T}) = S^2(1/m - 1/M)\ E_M(\hat{M}^2) + \mu^2 V_M(\hat{M}).$$

$$\text{(ii) } \nu(\hat{T}) = \hat{T}^2 - [\ell(u)(Z - s^2) + \hat{M}s^2] \text{ is a UE of } V_M(\hat{T}).$$

Proof. The proof of (i) is easy and hence omitted. To prove (ii) note that

$$E\ [\ \{\ell(u)(Z - s^2) + \hat{M}s^2\}\ |\, u\ ]$$

$$= \ell(u)(M^{-1}\sum_1^M X_i^2 - S^2) + \hat{M}S^2 = \ell(u)(\mu^2 - M^{-1}S^2) + \hat{M}S^2.$$

Hence

$$E_M\nu(\hat{T}) = E_M(\hat{T}^2) - [M^2(\mu^2 - M^{-1}S^2) + MS^2] = E_M(\hat{T}^2) - T^2 = V_M(\hat{T}).$$

3. COMPARISON WITH SRSWOR WITH A FIXED NUMBER OF DRAWS

In this section, first it will be shown that if one insists on unbiased estimation of μ then our strategy will be superior to the one based on SRSWOR with a fixed number of draws. Secondly, this superiority will be demonstrated even when biased estimators are allowed.

Let d be a fixed (somehow) number of draws in SRSWOR sampling, $\hat{s}$ a sample so drawn, $\hat{s} \cap A$ the set of A-units in $\hat{s}$ and C the cardinality of $\hat{s} \cap A$. We will use, for this scheme also previous notations P_M, E_M, V_M to imply phenomena relevant here. Then for such a sampling we have:

Theorem 3.1. μ admits a UE if and only if $d \geq N - r + 1$.

Proof. Let $d \geq N - r + 1$. Then $P_M\,[c = 0] = 0$, $\forall\, M \in \mathcal{M}$ and $\hat{\mu} = c^{-1}\Sigma' X_i$ is a UE of μ.

To prove the necessity it will be enough to show that if $d = N - r$, then μ does not admit a UE. For this the following notations will be used. let $j_1, \ldots, j_d$ be d distinct increasingly ordered units out of $1, \ldots, N$, constituting the elements of $\hat{s}$ and such that some k of them $(0 \leq k \leq d)$, say $i_1, \ldots, i_k$ (increasingly ordered) belong to A. Then we write $\hat{s} = (j_1, \ldots, j_d)$, $\hat{s}' = (i_1, \ldots, i_k) = \hat{s} \cap A$ (so that $k = 0 \Rightarrow \hat{s}' = \Phi$ and $k = d \Rightarrow \hat{s}' = \hat{s}$) and $X(\hat{s}') = (X_{i_1}, \ldots, X_{i_k})$, a sequence of X_i values for the units in $\hat{s}'$. Then if there exists a UE for μ, say t, we may write $t = t(X(\hat{s}')|\hat{s})$ such that

$$E_M(t) = \mu,\ \forall\, X_1, \ldots, X_M,\ \forall\, M \in \mathcal{M}. \tag{3.1}$$

For $0 \leq k \leq d$, let $t_k = \Sigma_k t(X(\hat{s}')|\hat{s})$, Σ_k being sum over all samples with exactly k A-units. Clearly t_o is free from X_i's.

If $d = N - r$, then $\mathcal{M} = \{N - d, N - d + 1, \ldots, N\}$. Suppose $M = N - d + j$ $(0 \leq j \leq d)$. Then the A-units may be chosen in $\binom{N}{M} = \binom{N}{d-j}$ ways. Accordingly $\binom{N}{d-j}$ equations are involved in (3.1). Summing over the number of ways of choosing the A-units, (3.1) yields

$$\sum_{w=j}^{d} a_{jw} t_w = \frac{\binom{N}{d}\binom{N-1}{d-j} T}{N - d + j}\ , \tag{3.2}$$

where, for $0 \leq j \leq w \leq d$, $a_{jw} = \binom{N-d}{w-j}$ if $N - d \geq w - j$; 0, otherwise. From (3.2) the solutions for the t_w's may be obtained as

$$t_w = \binom{N}{d}\binom{d}{w}\frac{T}{N}, \; 0 \leq w \leq d, \tag{3.3}$$

and the validity of (3.3) follows from the fact that

$$\sum_{w=j}^{d} \binom{N-d}{w-j}\binom{d}{w} = \binom{N}{d-j}.$$

In particular, (3.3) yields $t_o = N^{-1}\binom{N}{d}\,T$. But then t_o is not free from the X_i's implying a contradiction, proving the necessity and completing the proof.

Thus with a fixed size (d) SRSWOR scheme, for unbiased estimation of μ we need $d \geq N - r + 1$ which may become too large (especially if r is small) making the scheme operationally inconvenient. Even if $d \geq N - r + 1$, the fixed size SRSWOR scheme together with the UE $\hat{\mu} = c^{-1}\Sigma' X_i$ can be seen to be less efficient than the strategy described in the preceding section when compared at equal level of cost of inspection.

To elaborate, suppose $d \geq N - r + 1$ and note that

$$V_M(\hat{\mu}) = S^2\left[E_M(1/c) - 1/M\right]. \tag{3.4}$$

For our inverse sampling scheme, by (2.1) the expected number of draws is given by $m(N + 1)/(M + 1)$ and, to make our scheme comparable to a fixed size (d) scheme, this should equal d i.e. one should have $m = d(M + 1)/(N + 1)$, in which case Theorem 2.1 yields

$$V_M(\bar{x}) = S^2\left[\frac{N + 1}{d(M + 1)} - \frac{1}{M}\right]. \tag{3.5}$$

Since

$$E_M(c^{-1}) > [E_M(c)]^{-1} = \frac{N}{dM} > \frac{N + 1}{d(M + 1)},$$

it follows that (3.4) is greater than (3.5), proving our assertion.

It is also interesting to compare our strategy with the fixed size scheme when a possibly biased estimator of μ is allowed in the latter. In fixed size (d) SRSWOR scheme, consider the usual (ratio) estimator of μ given by [vide e.g. Rao (1975)]

$$\begin{aligned} \mu^* &= c^{-1}\Sigma' X_i \qquad &&\text{if } c > 0 \\ &= 0 &&\text{if } c = 0 \end{aligned}$$

The bias in μ^* equals $-\mu\, P_M(c = 0)$ (observe that if $d \geq N - r + 1$, then $P_M(c = 0) = 0$, $\forall\, M \in \mathcal{M}$ and μ^* reduces to the UE $\hat{\mu}$ defined earlier) and it can be shown that

$$MSE_M(\mu^*) = S^2 \sum_{a \geq 1} (1/a - 1/M)\, P_M(c = a) + \mu^2 P_M(c = 0). \tag{3.6}$$

A straightforward analytic comparison between (3.5) and (3.6) is difficult but as numerical examples including the two cited below suggest, in most practical situations (3.5) will be smaller than (3.6), indicating the superiority of our strategy even when a possibly biased estimator is allowed in the fixed size scheme.

Example 3.1. The following data relate to the aggregate percentage of marks of all the students who passed the Bachelor of Statistics Examination of the Indian Statistical Institute (ISI) during the last five academic years ended 1984[1].

---

[1] The data are obtained from the office of the ISI Dean of Studies to whom the authors are grateful for granting an access to them.

68	80	80	72	87	71	55	75	85	52	82
76	73	54	57	51	56	48	73	54	76	69
87	81	68	74	58	56	71	66	69	81	59
65	83	79	72	50	44	65	61	57	50	73
85	87	64	70	48	58	61	53	56	62	61
74	62	56	62	58	58	66	70	80	74	80

Suppose it is desired to estimate from a sample the mean score of those students who obtained a first class (i.e. sixty percent or above). Then $N = 66$, $M = 44$, $\mu = 73.1818$, $S^2 = 61.6871$. For a fixed size SRSWOR scheme with $d = 10$, (3.6) equals 9.5967. The comparable m in our inverse sampling strategy is $d(M + 1)/(N + 1) = 6.72$ and with this $= 6, 7$, (3.5) equals 8.8792, 7.4105 and the resulting gains in efficiency, compared to the fixed size scheme, are 8.08 and 29.50 percent respectively.

Example 3.2. As a somewhat less traditional example, consider the problem of estimating the mean of the prime numbers among the first sixty natural numbers. The $N = 60$, $M = 18$, $\mu = 24.5$, $S^2 = 350.1471$. For a fixed size SRSWOR scheme with $d = 7$, the value of (3.6) is 205.4654. The comparable inverse sampling strategy requires $m = d(M + 1)/(N + 1) = 2.18$ and with this $= 2$, (3.5) equals 155.6209, indicating a gain in efficiency by 32.03 percent.

4. CONCLUDING REMARKS

In this paper we have considered the estimation problem for a single domain. In large scale surveys estimators are often required for several domains in which case the present procedure may be modified as follows.

Let there be t domains, the domain sizes M_k being unknown, having respective parameter spaces $\mathcal{M}_k = \{r_k, r_{k+1}, ..., R_k\}$, where r_k and R_k are known $(1 \leq k \leq t)$. Let μ_k and S_k^2 and denote the population mean and variance of the study variate in the kth domain. The sampling scheme may be inverse generalized hypergeometric, i.e. inverse SRSWOR may be continued till at least $m_1, m_2 ..., m_t$ $(m_k \leq r_k$ for each $k)$ units of the 1st, 2nd, ..., tth domains are realized. For each k, clearly the number of units, say ξ_k, in the sample from the kth domain is now a random variable, with $P_M(\xi_k \geq m_k) = 1$ (where $m = (M_1, ..., M_t)'$, P_M, E_M the corresponding probability and expectation operator), since even when the quota for kth domain is filled up, sampling may have to be continued to fill up those for the other domains thus possibly including in the sample some additional units from the kth domain. The mean, $\bar{x}_k$, of the units in the sample from the kth domain is a UE of μ_k with a variance $S^2 [E_M(1/\xi_k) - 1/M_k]$, $1 \leq k \leq t$.

In this set-up also numerical investigations (records omitted since they seem uninteresting in the present context) suggest that the inverse sampling strategy will be more efficient than the one based on fixed size SRSWOR when compared at the same level of cost. For multidomain situations, however, the detailed algebraic expressions become somewhat involved so that an analytic comparison along the line of the preceding section becomes difficult and hence not reported here.

ACKNOWLEDGEMENT

The authors are thankful to the referee for his highly constructive suggestions that helped improvement on an earlier draft.

REFERENCES

Haldane, J.B.S. (1945). On a method of estimating frequencies. *Biometrika*, 33, 222-225.

Rao, J.N.K. (1975). Analytical studies of sample survey data. *Survey Methodology*, 1, 1-76.

Sampford, M.R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika*, 49, 27-40.

Survey Methodology, December 1984
Vol. 10, No. 2, pp. 187–197
Statistics Canada

A Methodology for Surveying Disabled Persons using a Supplement to the Labour Force Survey[1]

D. DOLSON, P. GILES, and J.-P. MORIN[2]

ABSTRACT

In response to a need for data on disabled persons in Canada, Statistics Canada undertook a program to create a disability database. This includes using supplements to the Canadian Labour Force Survey in the Fall of 1983 and the Spring of 1984, as well as including questions on the 1986 Census of Population. A general discussion of the background and content of the survey is presented. A comparison of screening methodologies conducted by Statistics Canada in November 1982 and January 1983 is presented and the results are compared.

KEY WORDS: Disability; Screening; Activities of Daily Living.

1. INTRODUCTION

On May 23, 1980 the Canadian government created the Special Parliamentary Committee on the Disabled and the Handicapped. In February 1981 (the International Year of the Disabled) this committee published its report, entitled "Obstacles" [4]. Recommendation 113 of the "Obstacles" report reads in part:

"That the Federal Government directs Statistics Canada to give a high priority to the development and implementation of a long-term strategy which will generate comprehensive data on disabled persons in Canada, using population-based surveys and program data."

The government, wishing to respond positively to the recommendations contained in the report, thus requested Statistics Canada to undertake a survey of disabled persons.

This paper focuses on disability surveys conducted as supplements to the Canadian Labour Force Survey (LFS) in October 1983 and June 1984 and on tests which were done in November 1982 and January 1983.

2. DEFINITIONS

Definitions developed by the World Health Organization (W.H.O.), given in McWhinnie (1980), were employed by the Special Parliamentary Committee. These definitions arise out of a model which focuses on the consequence of disease, and addresses the following illness-related phenomena.

$$\begin{tikzcd} \text{DISEASE} \arrow[r] & \text{IMPAIRMENT} \arrow[r] \arrow[rr, bend right] & \text{DISABILITY} \arrow[r] & \text{HANDICAP} \end{tikzcd}$$

[1] This paper is a combined version of the two papers entitled "A Methodology for Surveying Disabled Persons Using a Supplement to the Canadian Labour Force Survey" by P. Giles and D. Dolson, and "The Canadian Experience with Screening for Disabled Persons in a Household Survey" by P. Giles, D. Dolson and J.-P. Morin. These papers were presented at the 1983 ASA meetings in Toronto.

[2] P. Giles, Business Survey Methods Division, D. Dolson and J.-P. Morin Institutional & Agriculture Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

As given in World Health Organization (1980), the definitions of these terms are as follows.

Impairment: In the context of the health experience, it is any loss or abnormality of psychological, physiological or anatomical structure or function.

It is characterized by losses or abnormalities that may be temporary or permanent, and that include the existence of an anomaly, defect, or a loss in a limb, organ, tissue, or other structure of the body, including the systems of mental function. Impairment represents the exteriorization of a pathological state, and in principle reflects disturbances at the level of the organ.

Disability: It is any restriction or lack of ability (resulting from an impairment) to perform an activity in the manner or in the range considered normal for a human being.

It is characterized by excesses or deficiencies of customarily expected activity, and may be temporary or permanent, reversible or irreversible, and progressive or regressive. Disabilities may arise as a direct consequence of impairment or as a response by the individual, particularly psychologically, to a physical, sensory, or other impairment. Disability represents the objectification of an impairment, and as such, it reflects disturbances at the level of the person.

Handicap: It is a disadvantage for a given individual, resulting from an impairment or disability, that limits or prevents the fulfillment of a role that is normal (depending on age, sex, and social and cultural factors) for that individual.

Handicap is concerned with the value attached to an individual's situation or experience when it departs from the norm. It is characterized by a discordance between the individual's performance or status and the expectations of the individual himself or the particular group of which he is a member. Handicap thus represents the socialization of the impairment or disability, and as such reflects the consequences for the individual (cultural, social, economic and environmental) that stem from the presence of impairment and disability.

To explain these definitions more clearly, consider an example:

DISEASE	Arthritis
↓	
IMPAIRMENT	Inflammation of the joints in hands and fingers
↓	
DISABILITY	Inability to grasp and hold
↓	
HANDICAP	Limited in the kind or amount of work which can be done

3. TARGET POPULATION

Ideally the target population would include all persons in Canada who are disabled according to the above definition and subject to constraints on severity and duration. Severity can be regarded in terms of the person; i.e., how severely is a person disabled, or in terms of disability; i.e., how severe is the specific disability. In defining a target population, a measurement of severity must be included, if only implicitly. The duration of disability must be explicitly addressed. To capture all disabilities, including those arising out of acute illnesses of relatively limited duration, would identify a large percentage of people and run contrary to the spirit of the "Obstacles" recommendation. Nevertheless, to limit the population to the permanently disabled is avoiding the issue and ignoring the needs of the long-term but not chronically disabled.

As explained in the next section, the data were collected through the use of the supplementary capacity of the LFS. In addition to the normal constraints of the LFS, one significant limitation was imposed by the use of the LFS. The target population of the disability survey was not to include the "mentally ill". For example, the target population excluded illness such as amnesia, neuroses and phobias but included impairments of intelligence such as mental retardation or dyslexia. It was felt that asking for this information could be very sensitive in nature and negative reactions could compromise the primary objectives of the LFS.

Thus the target population for the disability survey tests includes all persons having one or more physical (nonbehavioural) disabilities, or knowledge acquisition or other educational disabilities (arising from impairments in intelligence, attention, psychomotor functions and language), whose duration has been or is expected to be at least six months. It also includes individuals suffering from diseases of a chronic and degenerative nature and which have a high probability of producing impairments which are physically disabling. In addition the normal constraints of the LFS are in effect which precludes individuals in institutions.

4. DATA COLLECTION

The difficulty is in translating the definitions into a set of questions which identify persons of interest from a set of persons in the general population. This leads to setting an objective to collect information on those who have a high probability of being disabled by any user's definition, and at the same time, keeping the number of people surveyed within reasonable limits.

The first option studied was to include questions on the Census of Population. However the diverse data requirements would have required ten to thirty additional questions, which was clearly not possible. The second option considered was to include a limited number of questions on the Census of Population, which would identify the disabled. In order to meet the data requirements, a follow-up survey would be required. In fact, this option has been approved and disability questions will be included on the 1986 Census questionnaire. This will provide detailed estimates for small areas. However, results from the follow-up survey will not be available before 1988 or 1989. Given that the current demand for data was high as well as the fact that this demand was mainly for national baseline estimates, the Census/follow-up option was considered inadequate by itself.

For meeting current data requirements, two alternatives were considered. The first possibility would be to mount on a continuous or periodic basis a household survey similar to the Canada Health Survey, to provide a profile of the disabled and handicapped. Such a survey could use survey methods designed particularly for the collection of disability data. However, resource constraints made this proposal not feasible. The second alternative was to use the supplementary capacity of the LFS. For reasons of expediency and cost, this was the method chosen. This has the secondary advantage that the resulting disability data on the individual can be directly linked to their labour force data collected by the LFS.

The data collection for the disability survey was conducted in two stages. First, all persons in all households in the LFS sample, except the one-sixth of the sample which is in its first month of the survey, underwent a "screening" process. Persons of potential interest were identified by means of a "screening" questionnaire. The mode of data collection was the same as for the LFS interview. However, the interviewers were asked to obtain non-proxy interviews as often as possible, even if it meant calling back at a later time. This "screened in" population was then asked another set of questions in a follow-up survey. All of these interviews took place about a week after the screening interview. They were all personal interviews and non-proxy responses only were accepted. This second set of questions was designed to collect the data identified as being desirable by a consultation process with the users.

The schedule for the survey was as follows. Three proposed screening questionnaires were tested in November 1982 and January 1983. More details on these surveys will be given later. Based on the results of these surveys, one screening questionnaire was developed. Two "full" surveys with a screen and a more detailed questionnaire as described above were conducted in October 1983 and in June 1984.

5. APPROACHES TO SCREENING - OTHER SURVEYS

The first step in constructing a set of screening questions was to investigate experiences encountered by other groups that had previously conducted disability surveys.

The one approach that has been used in many surveys is the Activities of Daily Living (ADL) approach. The Activities of Daily Living are a set of activities which any person is required to perform during the course of their regular living pattern. Although there is no generally recognized "best" set of activities that should be used, the set developed in 1978 by the Organization for Economic and Co-operative Development (OECD) and noted in McWhinnie (1980) has been used by surveys in several countries; see Klaukka (1981), Mizrahi and Mizrahi (1981), Raymond, Christe and Clemence (1981), Van Sonsbeek (1981), Wilson and McNeil (1981).

Since a person's ability to perform an ADL may depend on their use of a physical aid, such as an artifical limb, the use of a list of physical aids for screening could be appropriate.

Another approach for screening is that of major activity limitation. If a person is limited in his/her major activity (i.e., work, school, home) that person is probably experiencing some disability. This approach has been used in the United States in a pretest for a disability survey (1980) and in the annual Health Interview Survey, and in Canada in the Canada Health Survey (1978-79).

A list of chronic conditions could be useful for screening since persons with chronic conditions are in the target population but may be missed by ADL's or activity limitation if the person has intermittent difficulty.

Finally a person could be asked a single self-perception question such as "Do you have any physical disabilities or handicaps?".

6. TEST OF SCREENING MECHANISMS

The three Statistics Canada screening tests used combinations of these approaches. Also persons aged 15 or over were administered a different questionnaire than those under 15 years of age. No suitable set of ADL's has been compiled for children. In fact, most disability surveys that have been previously conducted have excluded children. Here, we will consider only persons aged 15 years and older.

In the November 1983 Labour Force Survey, each respondent was asked "Does ... now have any disability or handicap which has lasted or is expected to last six months or more?" This was called Test 1. Persons screened in by this question were those responding "yes".

The other approaches to screening were tested using two different questionnaires each administered in the January 1983 Labour Force Survey. These questionnaires were called Test 2 and Test 3.

Test 2 included the following sections: a list of special aids, a list of ADL's, and the activity limitation question "Are there any (other) conditions or health problems that now prevent or limit . . . when carrying out his/her normal daily activities at a job, in school, or in the home? Please report only difficulties which are expected to last more than six months". Persons who reported using at least one of the special aids or having trouble doing at least one of the ADL's or who answered "yes" to the above question were screened in.

Test 3 included a list of ADL's, a list of chronic conditions, and the following two work disability questions: "Is ... limited in the kind or amount of work he/she can do at his/her job or business because of a long-term physical condition or health problem?" (asked only to employed persons) and "Is ... prevented or limited in the kind or amount of work he/she could do at any job or business because of a long-term physical condition or health problem?". Persons who reported having trouble doing at least one of the ADL's or who had at least one of the chronic conditions or who replied "yes" to either of the above two questions were screened in.

Lists of the special aids, the chronic conditions, and the ADL's used in these tests are given in the appendix. It should be noted that the ADL's used in Test 2 and Test 3 were identical although slightly modified from the OECD list. In addition, each test takes a different approach to the ADL list. Test 2 permits the use of aids to perform the activities, while Test 3 does not.

These proposed screening methods do not permit an assessment of whether or not the target population is being correctly identified, unless they are used on a control population. This has not been done for the Canadian survey.

Test 1 was administered to all persons in households in the November 1982 Labour Force Survey. In January 1983, two rotation groups were used for each of Test 2 and Test 3. These rotation groups were chosen so that each in-sample household had also been in-sample in November 1982. This facilitated the comparison of the results of Test 1 with those of Test 2 and Test 3.

6.1 Major Findings of the Pilot Tests

The major goal in the analysis of the tests was to determine the set of questions that would be most effective in screening in those persons who belong to the target population. Another important factor was to determine an effective screen that would also not unduly increase respondent burden or cost. In addition, sources of non-sampling errors that became evident during the analysis were noted so that, for example, survey procedures or questionnaire design could be changed appropriately. The following discussion presents the major findings of the analysis.

The sample size for Test 1 was about 115,000 persons. For each of Test 2 and Test 3 the sample size was almost 38,000 persons. For all three tests the samples were about 49% male and 51% female.

Table 1 shows by sex the percentage of the sample screened in by Test 1, Test 2, Test 3 and by each section of Test 2 and Test 3. Most notable in this table is that Test 1 screened in only 5.6% of the sample as compared to about 16% of the sample for either of the ADL questions. Given that functional limitation as measured by the activities of daily living is a key indicator of disability, this shows that the single question asked in Test 1 is not effective for screening in the entire target population.

Table 1

Percent Screened in by Each Section of
Each Questionnaire, by Sex

Section	Percent Screened in	
	Male	Female
Test 1:	5.5	5.7
Test 2: ADL	14.7	16.2
Test 2: Aids	3.2	2.9
Test 2: Activity Limitation	5.9	6.1
Test 2:	18.3	19.4
Test 3: ADL	15.4	16.7
Test 3: Work Disability	13.0	13.1
Test 3: Chronic Conditions	25.6	27.4
Test 3:	29.8	31.2

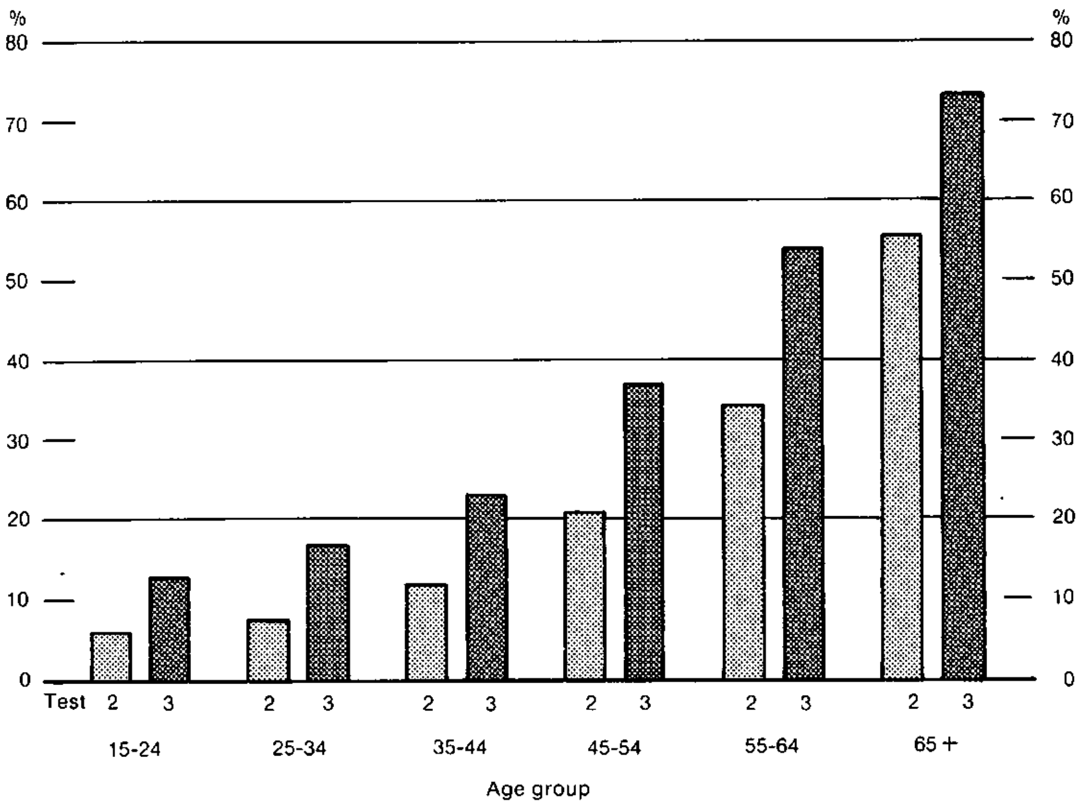

Figure 1. Percent Screened in by Test 2 and Test 3

Percentages of the sample screened in by Test 2 and by Test 3 are shown by age in Figure 1. For both tests, the probability of being screened in is highly related to age. The probability of being screened in is an increasing function of age. Although the degree varies, the same relationship holds for each of the Sections of Test 2 and Test 3, for each of the seventeen ADL's and for most of the chronic conditions (multiple sclerosis, epilepsy, cerebral palsy, cystic fibrosis and muscular dystrophy are exceptions).

Since functional limitation is a good indicator of disability the question for the analysis was how to include the ADL's on the screen rather than whether to include them. The "with special aid approach" (Test 2) selected 15.5% of the sample while the "without special aid approach" (Test 3) selected 16.0% of the sample. The difference between these figures is not statistically significant. For both respondents and interviewers the "with special aid" concept seemed to be more natural and more easily understood and was therefore chosen for the finalized screen.

Since the complete set of activities can be used to obtain a measure of degree of disability, all seventeen ADL's were retained for the finalized screen.

The special aids section in Test 2 screened in 3.0% of the sample. Of these, 84.2% were screened in by the ADL section. Thus although this type of data is of interest for the disability database, the aids section is not an efficient screening mechanism, especially in combination with the ADL section. Consequently, questions on aids are not included on the finalized screen.

In addition to activities of daily living, major activity limitation is also an important aspect of disability. The Test 3 questionnaire addressed this by the two questions noted earlier in this section. These questions, however, considered it only from the point of view of work disability. Of the persons in the Test 3 sample, 13.1% were screened in by these questions (6.8% of employed persons, 6.6% of unemployed persons and 22.7% of persons not in the labour force).

Although there were no obvious problems with the data from the work disability questions, there were some operational difficulties. In particular, the questions sometimes seemed irrelevant to retired persons. Thus, the finalized major activity limitation question was adapted in order to better suit persons not in the labour force.

The chronic conditions section in Test 3 screened in 23.9% of the Test 3 sample. Of these, 37.9% were not otherwise screened in. There were two main problems with this section. First, the question was difficult to answer for respondents who were not sure of the nature of their particular condition(s). Another difficulty is that the data sought on the follow-up questionnaire, are generally more pertinent to persons who are currently disabled. Thus persons screened in by a chronic condition, but not currently having a functional limitation or a major activity limitation would be interviewed for the follow-up and probably provide little useful data.

Given these problems, chronic conditions are not used as screening criteria on the finalized screen. One exception to this is mental handicap. This one condition is retained as a screening item since there may be persons with mental handicaps who are not screened in by the ADL's, or even by the major activity limitation question.

The project team felt that there would likely be differences in proxy and non-proxy responses related to any particular person. To increase non-proxy response, interviewers were instructed that whenever possible the questionnaire was to be completed by interviewing the individual to whom it applied. If a knowledgeable household member insisted upon responding for other household members, then this response was to be accepted; although the practice was to be avoided.

The level of proxy response obtained is considered to be fairly low (20.7% for females, 32.2% for males). Even after accounting for age-sex differences, it was found that proxy respondents were slightly less likely to be screened in than non-proxy respondents. Two reasons can be suggested as to why the probabilities of selection differ. First, persons who are unavailable and for whom proxy responses were provided may be less likely to be disabled. Second, proxy respondents may be less likely to state that a person has trouble doing an ADL or a major activity than the person himself/herself.

7. DATA REQUIREMENTS

As a result of a solicitation of data requirements from users, 173 responses were received which identified 588 issues of data needs. The following eleven areas were identified.

Issue	Number of users requesting data
1. Nature of impairment	123
2. Demographic characteristics	95
3. Employment	85
4. Assistance	77
5. Education	50
6. Accommodation	45
7. Economic Characteristics	41
8. Transportation	29
9. Social activities	26
10. Health	9
11. Communication	8

The nature of impairment/disability/handicap is basic to the survey. Considerable detail is collected, including cause of disability. Most users of the data are interested in focusing on the impairment or disability groupings separately. Demographic characteristics are always important data as they allow the user to identify sectors of the population falling into different categories.

It can be easily understood that employment data about the disabled would be an important issue, as employment is a key component to the independent living of a disabled person. A great deal of employment data are already collected by the LFS. The follow-up survey will collect data related to employment limitations experienced as a result of the disability. In addition to the analysis of the data for the disabled population, these data will permit comparisons of the labour market characteristics of the disabled with those of the Canadian population as a whole.

Three aspects of assistance are considered: technical aids and skills, employment related assistance, and education related assistance. In all three areas, need for aids or assistance was deemed more important than was use. Under technical aids and special skills, interest is greatest for those aids and skills which are most prevalent, or for which special services or facilities must be provided. The aids would be grouped under hearing, speaking, seeing and mobility. Employment related assistance refers to the impact of aids on the ability of the disabled to work.

The LFS already determines the highest level of education achieved by each respondent. In addition, the follow-up survey will collect data on current educational activity and the impact of disability on current and past education.

The LFS collects information on the dwellings of the respondents. Additional accommodation data will be collected on special architectural/structural features, both inside and outside the home and other buildings.

Economic characteristics will be considered in the following areas: personal income including financial assistance received due to disability, sources of financial assistance, and special expenses incurred as a result of the disability.

Transportation data will be collected on three types of travel: travel to work or school, other local travel and long distance travel. Details on each area will identify the modes of transportation used, frequency of use and problems encountered due to the disability.

Although some interest was expressed by users in data on social/leisure activities, health and communication, no data will be collected for these issues by the present survey. For the first and third of these issues it was felt that reliable and useful data could not be collected in this survey. Questions related to health are also not included because of the already substantial response burden imposed by issues of higher priority.

8. RELIABILITY OF ESTIMATES FROM THE DISABILITY SURVEY

When determining the content of a questionnaire, consideration must be given to the reliability of estimates produced for the various data items. It is useless to collect data which will not be reliable enough to publish, even if the data requirement has a high priority. The reliability of an estimate is tied directly to the sample size. For this survey, the number of persons receiving a screening questionnaire is fixed. Therefore the reliability of the estimates produced will depend on the number of disabled falling into the sample and the prevalence rates of each characteristic of interest. Based on population projections from the 1981 Census of Population and certain assumptions it is possible to estimate minimum prevalance rates required to produce an estimate which is "reliable enough" to publish. An estimate whose coefficient of variation is less than or equal to 16.5% is considered releasable without qualification by LFS. Table 2 displays the expected minimum releaseable estimates for the disability survey. Estimates of this size or higher will have coefficients of variation of less than 16.5%, subject to the validity of the following assumptions.

(1) All LFS sampled households are administered the screening questionnaire except the one-sixth of the sample which is in its first month of the survey,

(2) 2.95 persons per household on average,

(3) 5% LFS non-response rate,

(4) 5% disability survey non-response rate,

(5) Design effect of 2.5 (this accounts for the fact that a simple random sample design was not used),

(6) 19% of total adult population and 8% of total child population (aged less than 15) are screened in.

To explain the table in more detail, consider, for example the province of Newfoundland. An estimate of 9,000 persons possessing a particular characteristic will have a coefficient of variation less than 16.5% and is publishable whereas an estimate of 7,000 will have a coefficient of variation greater than 16.5% and is not publishable. An estimate of 8,000 is approximately 1.4% of the population of Newfoundland. Given the assumptions about percentage disabled in the population, an estimate of 8,000 is approximately 10.1% of the adult disabled population and 59.1% of the child disabled population of Newfoundland.

The design effects observed from Test 2, Test 3 and the October 1983 Disability Survey for number of persons screened in were about 1.5. This suggests that design effects for number of screened in persons with specified characteristics are probably also much less than 2.5.

In the October 1983 Disability Survey 12.9% of adults and 4.8% of children in the sample were screened in.

Table 2

The expected minimum releaseable estimates

Province/Region	Min P[a]	Min X[b]	Min D[c]	
			Adults	Children
Atlantic	0.4	7,500	2.3	16.2
NFLD	1.4	8,000	10.1	59.1
PEI	2.9	3,500	20.0	>100
NS	1.1	8,500	6.9	54.5
NB	1.0	7,000	6.7	49.5
Quebec	0.5	30,500	3.3	28.2
Ontario	0.4	32,500	2.6	22.0
Prairies	0.3	10,000	1.7	12.4
MAN	0.9	9,000	7.0	47.6
SASK	0.8	7,000	5.0	38.0
ALTA	0.6	13,000	4.1	29.9
British Columbia	0.7	17,500	4.4	38.2
Canada	0.1	18,000	0.6	4.2

[a] Min P = minimum estimable percentage of the total population,

[b] Min X = minimum estimable total,

[c] Min D = minimum estimable percentage of disabled adults or children.

APPENDIX

Special Aids

Does . . . now use
- a wheelchair?
- crutches or other walking aids?
- any kind of brace excluding braces for teeth?
- medically prescribed orthopedic shoes?
- artifical limb(s)?
- a hearing aid?
- a guide dog?
- a white cane?
- any other kind of special aid?

Activities of Daily Living

Does . . . now have any trouble
- walking 400 metres without resting (about 3 city blocks)?
- walking up and down a flight of stairs?
- carrying an object of 5 kg. 10 metres (e.g. carrying a 12 lb. bag of groceries 30 ft.)?
- moving from one room to another?
- standing for long periods of time (e.g. more than 20 minutes)?
- when standing, bending down and picking up an object from the floor (e.g. a shoe)?
- dressing and undressing himself/herself?
- getting in and out of bed?
- cutting own toenails?
- using fingers to grasp or handle?
- reading?
- cutting own food?
- reading ordinary newsprint (with glasses if normally worn)?
- seeing clearly the face of someone from 4 metres (e.g. across a room) (with glasses if normally worn)?
- hearing what is said in a normal conversation with one other person?
- hearing what is said in a normal conversation with at least two other persons?
- speaking and being understood?

Chronic Conditions

Which, if any, of these long term conditions or health problems does . . . presently have?
- heart disease
- kidney disease
- lung disease
- cancer
- diabetes
- epilepsy
- cerebral palsy
- multiple sclerosis
- cystic fibrosis
- muscular dystrophy
- paralysis of any kind
- arthritis or rheumatism of a serious nature
- high blood pressure
- hearing trouble (uncorrected by aid)
- vision trouble (uncorrected by aid)
- mental handicap
- any missing limb(s) including finger(s) and toe(s)
- any other long-term condition or health problem (please specify)

REFERENCES

STATISTICS CANADA. (1983). Data Content for the Statistics Canada Survey of the Disabled. *Disability Database Development Project,* Health Division, technical report.

CARTER, R.G., GILES, P.D., and SHERIDAN, M.J. (1982). Description and Rationale for the Screen Tests for the January 1983 Disability Survey. *Disability Database Development Project,* Health Division, Statistics Canada.

GRABOWIECKI, F. (1982). Discussion of the Target Population for the Disability Survey. *Disability Database Development Project,* Health Division, Statistics Canada.

HOUSE OF COMMONS. (1981). Obstacles. *Report of the Special Committee on the Disabled and Handicapped.* Ottawa.

MORIN, J.-P. (1983). Enquête sur les handicapés compte rendu des requêtes des utilisateurs. *Disability Database Development Project,* Health Division, Statistics Canada.

WORLD HEALTH ORGANIZATION. (1980). International Classification of Impairments, Disabilities and Handicaps. Geneva, Switzerland.

McWHINNIE, J.R. (1980). Disability Indicators for Measuring Well-being. *OECD Social Indicators Programme Technical Report Series,* Paris, France.

McDOWELL, I. (1981). An Examination of the OECD Survey Questions in a Canadian Study. *Revue d'épidémiologie et de santé publique,* 29, 421-429.

KLAUKKA, T. (1981). Application of the OECD Disability Questions in Finland. *Revue d'épidémiologie et de santé publique,* 29, 431-439.

MIZRAHI, A. and MIZRAHI, A. (1981). Évaluation de l'état de santé de personnes agées en France, à l'aide de plusieurs indicateurs, dont les questions de l'OCDE. *Revue d'épidémiologie et de santé publique,* 29, 441-450.

RAYMOND, L., CHRISTE, E., CLEMENCE, A. (1981). Vers l'établissement d'un score global d'incapacité fonctionelle sur la base des questions de l'OCDE, d'apres une enquête en Suisse. *Revue d'épidémiologie et de santé publique,* 29, 451-459.

VAN SONSBEEK, J.L.A. (1981). Applications aux Pays-Bas des questions de l'OCDE relatives à l'incapacité. *Revue d'épidémiologie et de santé publique,* 29, 461-468.

McWHINNIE, J.R. (1981). Disability Assessment in Population Surveys: Results of the OECD Common Development Effort. *Revue d'épidémiologie et de santé publique,* 29, 413-419.

WILSON, R.W., McNEIL, J.M. (1981). Preliminary Analysis of OECD Disability on the Pretest of the post Census Disability Survey. *Revue d'épidémiologie et de santé publique,* 29, 469-475.

'Logistic Regression Analysis of Labour Force Survey Data' by S. Kumar and J.N.K. Rao, Survey Methodology (1984), 10, 62-81.

In the formula (26) for the projection matrix $M = I - H$, the matrix $\hat{V}_b$ should have been given as

$$\text{diag } [\hat{p}_1(1-\hat{p}_1)\ (nw_1),\ \ldots,\ \hat{p}_I(1-\hat{p}_I)\ (nw_I)]$$

instead of diag $[\hat{p}_1(1-\hat{p}_1)/(nw_1), \ldots, \hat{p}_I(1-\hat{p}_I)/(nw_I)]$. Because of this error, the figures 3 and 4, based on the diagonal elements m_{ii} and h_{ii} of M and H respectively, are incorrect. The corrected figures are given here as figures 3* and 4* which indicate that cells numbered 2, 3 and 55 warrant further examination. It may be noted that the diagnostics for assessing the impact of extreme points on the fit (see figures 5-10, pp. 79-81) have also identified cells 2 and 3 as extreme points, excepting figure 8. The overall observation in the paper that cells 2, 3 and 7 may be possible candidates for deletion is unaffected by the above error.

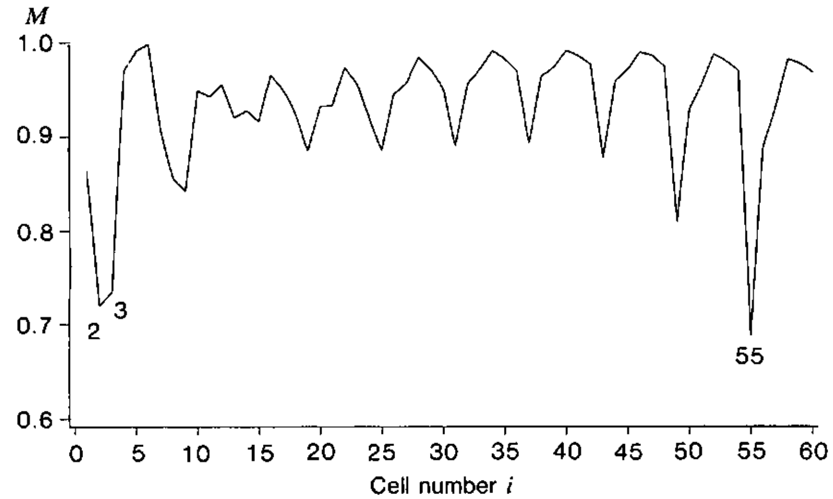

Figure 3*. Index plot of m_{ii}.

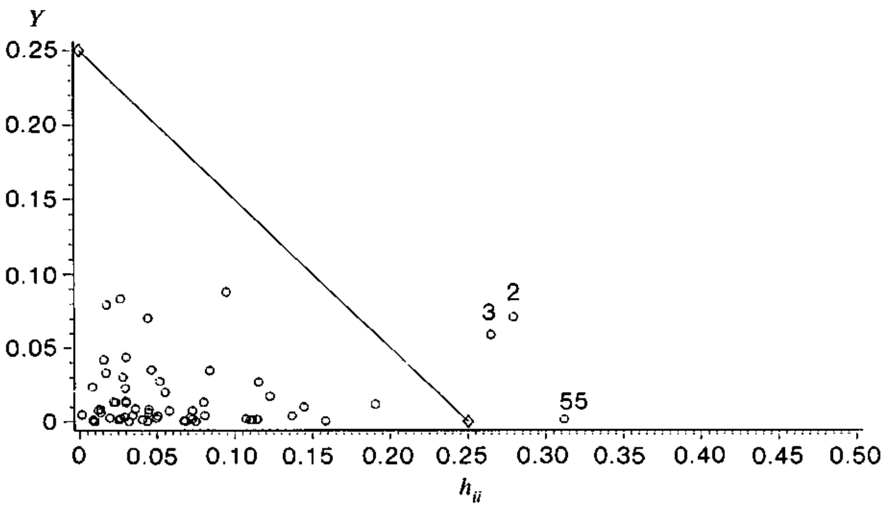

Figure 4*. Scatter plot of χ_i^2/χ^2 vs. h_{ii}.

ACKNOWLEDGEMENTS

The Survey Methodology Journal wishes to thank P. Pariseau and C. Paslawski for their patient preparation of the manuscripts and to D. Lemire for his service as liaison with the translation and the production services.

Acknowledgement is also due to H. Gough, I. Trottier, M. Brodeur and G. Lemaître for their help in proofreading of French version of papers.

Finally, the Journal wishes to thank the following persons who have served as referees during the past year.

M. Bankier	M. Lawes
G. Brackstone	D. Paton
N. Chinnappa	C. Patrick
D. Drew	D. Rhoades
J. Gambino	K.P. Srinath
J. Kovar	R. Verma

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.