Statistics Canada — Statistique Canada

SURVEY METHODOLOGY

TECHNIQUES D'ENQUÊTE

June - 1978 - Juin

VOLUME 4

NUMBER 1 - NUMÉRO 1

A Journal produced by
Statistical Services Field
Statistics Canada

Publié par Le Secteur des
Services Statistiques,
Statistique Canada

C O R R I G E N D A

Table of Contents:
should read Vol. 4, No. 1

Table des Matières:
pour Vol. 4, No. 2
lire Vol. 4, No. 1

The page numbered 20 should
follow page 17
(correct order is 17-20-18-
19-21 ...)

La page numérotée 20
devrait suivre la page 17
(l'ordre correct est donc
17-20-18-19-21 ...)

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1978 Vol. 4 No. 1

A Journal produced by Statistical Services Field, Statistics Canada

Publié par le secteur des services statistiques, Statistique Canada

C O N T E N T S

The Utilization of Administrative Records for Statistical Purposes
L.E. ROWEBOTTOM 1

The Effect of a Two-Stage Sample Design On Tests Of Independence
J. COWAN and D.A. BINDER 16

Approximate Tests Of Independence And Goodness Of Fit Based On Stratified Multi-Stage Samples
I.P. FELLEGI 29

A Survey Design System For The Measurement Of Truck Cargo Flows In Peru
A. SATIN and R. RYAN 57

A Comparison Of Correlated Response Variance Estimates Obtained In the 1961, 1971 and 1976 Censuses
K.P. KRÓTKI and C.J. HILL 87

A Study Of Refusal Rates To The Physical Measures Component Of The Canada Health Survey
B.N. CHINNAPPA and B. WILLS 100

An Estimate Of The Efficiency Of Raking Ratio Estimators Under Simple Random Sampling
M.D. BANKIER 115

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1978 Vol. 4 No. 1

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. 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 the Department. Copies of papers in either Official Language will be made available upon request.

Politique de la rédaction:

La revue Techniques d'enquête veut donner aux personnes qu'intéressent les aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration des méthodes d'enquête: les problèmes de conception causé par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada. On pourra se procurer sur demande des exemplaires d'un article dans l'une ou l'autre langue officielle.

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1978 Vol. 4 No. 1

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Household Surveys Development Division, Statistics Canada, 10th Floor, Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested. Authors of articles for this journal are free to have their articles published in other statistical journals.

Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division de l'élaboration d'enquêtes ménages, Statistique Canada, 10^{e} étage, Edifice Coats, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi. Les auteurs des articles publiés dans cette revue sont libres de les faire paraître dans d'autres revues statistiques.

THE UTILIZATION OF ADMINISTRATIVE RECORDS FOR STATISTICAL PURPOSES[1]

L.E. Rowebottom[2]

This paper advances the case that administrative records are a powerful source of statistics and in support of this conclusion provides an overview of the extensive utilization in Canada of administrative records for statistical purposes. The paper discusses recent developments and the changing environment which are seen as major determinants of both the creation of administrative data bases as well as their utilization. The capabilities of the computer, combined with the extensive demand for statistics and the limited financial resources available to meet that demand, are seen as combining to lead to more extensive use of administrative records. A variety of problems associated with the use of administrative records is specified and the development of strategies to meet these problems and permit utilization of administrative records is described. Recent developments in Canada intended to support the use of administrative records are indicated.

1. INTRODUCTION

This paper, on the utilization of administrative records in Canada for statistical purposes, comprises a survey of:

- a history of such utilization;
- reasons why administrative records are a powerful source of statistics;
- recent important developments and the changed environment within which such utilization must take place;
- problems accounting for the relatively slow pace of developments;
- prerequisites of effective and extensive utilization;
- ways being pursued to more fully utilize administrative records.

---

[1] Adapted from a paper prepared for the XIII Session of the Committee on Improvement of National Statistics (COINS) of the Inter-American Statistical Institute, November 1977.

[2] L.E. Rowebottom, Assistant Chief Statistician, Institutions and Agriculture Statistics Field, Statistics Canada.

Administrative records are defined for purposes of this paper as those generated as a result of carrying out a wide range of functions such as registering, insuring, educating, taxing, and other benefits or penalties so that decisions which affect individuals may be taken by departments, agencies, institutions and other organizations of government. Administrative records contain information not only about individuals, e.g., admission to a hospital, but also about organizations and institutions of government, e.g., amount spent on construction or about businesses, e.g., gross business income, employment, etc. In many respects, they are the public sector equivalent of the accounting records of businesses which form the basis of various economic statistics. In much the same way, administrative records are created for non-statistical purposes but may be used for statistical purposes.

2. STATISTICAL USES OF ADMINISTRATIVE RECORDS IN CANADA

Many countries have long and effective records of utilizing administrative files for purposes of compiling both economic and social statistics. Thus, in Canada, registrations of birth, death, and marriages have long formed the basis of vital statistics, and custom invoices have similarly constituted the basis of external trade statistics. Examples of other administrative files and the statistics derived from them are as follows. This listing is not exhaustive but makes clear that substantial portions of the national statistical system are in large part dependent upon administrative records.

The Administrative Records	The Statistics Derived From Them
Hospital admission and discharge records	Statistics on hospitalized illness, cause of hospitalization, days of hospital care, by age and sex. They comprise the only national measure of morbidity.

The Administrative Records	The Statistics Derived From Them
Student enrolment	Numbers of students by age, sex and grade, as well as level of education attained and "drop-outs" before graduation. For university students, measures of enrolment by discipline.
Personnel records of teachers	Salaries of teachers by qualification, years of experience, discipline, and other characteristics.
Accounting records of boards, commissions, departments and agencies of government	For federal, provincial and municipal governments, statistics of income by source, expenditure by object and function, assets and debts.
Offenses reported to police and persons charged by police	Statistics of crime by type of crime, by age and sex of persons charged, for Canada, the provinces and major municipalities.
Personal income tax returns	Statistics of income by geographic area, occupation, age sex, level of income, marital status and number of dependents. Statistics concerning businesses of which the individual may be a partner or the proprietor.
Corporate tax returns	Statistics of incorporated businesses.
Payroll Deduction Accounts	Statistics of payroll and employment.

A characteristic of many of the files referred to above is that they are generated by the provinces as a result of their responsibility for the administration of such important areas as health, justice, education, the registration of vital events and a wide variety of other functions. Provinces and municipalities also account for a large portion of the total income and expenditure of government. Given such division of responsibilities and a corresponding creation by the provinces of important administrative records, Statistics Canada as the national statistical agency has devoted a great deal of effort over many years to the co-ordination of administrative records so that they would be comparable from province to province and with the federal systems, and usable to

produce significant national statistics. Careful painstaking and continuing effort to obtain and maintain agreement on definitions, classifications, codes, and wording of questionnaires and administrative forms, has been the dominating activity required to derive good statistics from administrative records. Progress has been rewarding although slow. In many areas substantial progress has been made, in others, such as welfare, much remains to be done.

3. WHY USE ADMINISTRATIVE RECORDS FOR STATISTICS

It should not be taken for granted that there is a widespread agreement that administrative records provide significant sources of data for the derivation of useful statistics. Before further discussing the problems and possibilities of using administrative records for statistical purposes, it is important to describe why they should be so used. Skepticism is not uncommon and as a basis for giving priority to their exploitation it is important to provide a wider basis of understanding to both those who hold and those who use administrative files.

For a number of administrative files, the principal historic reason for using them, and still the overwhelming reason, is that they provide almost the sole source of essential statistics. Examples are import and export statistics based on customs declarations, and mortality and demographic statistics based on birth, marriage and death registration. Alternative sources of trade statistics, for example, could be developed but at inordinate cost in resources and response burden, as well as reductions in the statistics available. Even where administrative records do not compromise the sole source, they are frequently powerful complements to other data sources. Thus, income statistics derived from tax returns and those derived from household surveys complement each other and the statistical system is enriched by the use of both sources. For example, tax returns provide geographic detail not otherwise available on an annual basis and survey data provides more socio-demographic information about those in receipt of various levels of income. Similarly,

personal tax returns can supply financial data relating to unincorporated businesses which complements data from business surveys and can often reduce significantly the number of respondents needed for those surveys.

Another important reason for the use of administrative records is that they have been generated for other purposes and can be utilized for statistical purposes at marginal cost. Of course, even the marginal cost is not zero and can be quite high in absolute terms; and the usefulness of statistics derived from administrative records must justify the cost of their production.

A combination of these reasons - the uniqueness of the source and benefits which justify marginal costs - account for the historic derivation of many statistics from administrative records. In the areas of health, education and crime, it has been possible to study both the institutions and the population involved with them, e.g., patients and students, by utilizing the administrative records of the institutions.

At the same time, an important limitation of such files for social statistical purposes is that they relate to particular populations and not to the population at large, and it is this limitation which has given emphasis to the development of household surveys to measure such characteristics of the whole population as level of education attained and health status. However, as with the income statistics referred to above, statistics of education and health, for example, derived from the two sources - administrative records and household surveys - complement each other and one can seldom be replaced with the other. For example, information about the delivery of health care based on hospital records is greatly enriched, not replaced, by information about the health/illness conditions of the whole population derived from household surveys.

In the case of business statistics, the entire population of corporations, employers and unincorporated businesses are covered by the corporate, Payroll Deduction Account, and personal taxation systems. The bureau has access to these files for statistical purposes and they form the basis for the Business Register - a major tool in the integration of business survey-taking activity.

As an example of the way in which different data sources complement each other, household sample surveys can be taken frequently and can explore social issues in depth, but they can seldom provide information for small local areas where many public decisions are made. These characteristics of household surveys highlight a contrasting characteristic of some important administrative data files, namely, the fact that they are continuously updated and also relate to very large populations. Files which have such characteristics have great potential for the production of small area data, as well as longitudinal data derived from cohorts of individuals for lengthy periods. Both of these kinds of statistics are expensive to obtain from surveys. Some of the administrative records which relate to large portions of the population and are continuously maintained are those derived from the administration of unemployment insurance, family allowances, the collection of personal income taxes and health insurance.

Finally, the use of such files for statistical purposes does not involve significant additional risk related to privacy or confidentiality <u>if</u> the records are handled appropriately so that the statistical <u>output</u> contains no identifiable information. However, their use to produce statistics may not be perceived as involving minimal risk. In fact, an important task confronting statisticians is to demonstrate and persuade that the risk is so low, and that files are so carefully safeguarded, that the small risk which is inevitably involved is more than justified by the benefits derived.

4. RECENT DEVELOPMENTS AND THE CHANGING ENVIRONMENT

While the historic use of administrative records provides a perspective from which to view current developments, it does not provide an adequate understanding of recent developments which have radically changed the environment within which administrative records are used (or not used) for statistical purposes. Probably the most important of these developments are computers which have eliminated the monopoly, or near monopoly, on data processing capability formerly held by large statistical offices. Effective ability to process data is now available to many agencies of government, computer stored data bases now provide administrators with ready access to their records for management purposes, and the development of data base technology involving random access to computer stored data is rapidly extending that ability further than even the tremendous improvements which sequential batch computer processing made possible. The process of creating, storing and retrieving administrative records is usually referred to as a management information system but the product of such systems is typically statistics. The essential points are that where there was historically only one or a few producers of statistics derived from administrative records there can now be many, and that senior managers of administrative programs want access to information derived from their data bases.

A second change of major consequence has been the recent development by governments of large socio-economic programs such as unemployment insurance and old age pensions which have generated large data bases typically covering significant proportions of the population. These new data files, along with long established files, processed by the new computer technology, have resulted in the statistical system rather suddenly acquiring new producers of significant statistics in various subject-matter areas and a new major need to co-ordinate the data in this enlarged system.

While this paper has described these developments as "new" they have, of course, been underway for some while and were recognized and articulated at least four years ago (see [1] and [2]). Since that time there have been large increases in the number of machine-readable data files.

We have entered a still further phase in the utilization of administrative records for statistical purposes reflecting the following additional changes in the environment:

- financial constraint on the part of all levels of government which are seeking to reduce their commitments;
- reflecting such constraints, a concern about the costs of statistics;
- frequent complaints on the part of the business community, politicians and the media about the burden of completing statistical questionnaires;
- widespread concern about privacy and confidentiality;
- elements (known or assumed) of duplication within the statistical system, including duplication between administrative and other data bases;
- a conviction on the part of both suppliers of data and users of resource;
- record linkage technology which is now relatively advanced;
- an insistent demand by sophisticated users for access to micro-data bases reflecting the development of new analytical tools;
- statisticians' concern about the quality and comparability of the numbers being generated by the many more players who are now part of the system;
- in certain cases the simple fact that sufficient data to satisfy the demands for statistics cannot be obtained by traditional survey-taking methods without massive increases in expenditure and respondent burden;

- machinery for protection of confidentiality and security of data files is greatly improved;
- developments in methodology which enable us to use financial data from business financial statements to generate pseudo-survey data usable at the macro-level, when processed in conjunction with data derived from small sample surveys.

All of the above have led to the conviction that not using administrative records to the full is a luxury which can no longer be afforded, and that priority (one amongst several) must go to their fuller effective use as an important part of the statistical base.

5. PROBLEMS OF USING ADMINISTRATIVE RECORDS

There are a number of prerequisites for such effective use but before turning to them it is helpful to examine some of the problems associated with the use of administrative records which explain why progress has not been more rapid. Some problems referred to earlier are worth repeating in the following listing:

- administrative records are frequently held in multiple jurisdictions, for example in the provinces as already described;
- notwithstanding earlier comments about the importance of computers some of the technologies are still relatively new and not yet familiar to those responsible for data handling, e.g., computer utilization, data base technology, record linkage, micro-data bases;
- concern about possible negative public and political reaction has undoubtedly inhibited development, and properly so. Given the widespread concern about "big brother" the use of administrative records could be imperilled if statisticians were to move too quickly ahead of public and political opinion;

- the problems of standards, comparability, consistency, coverage, and similar problems are intrinsically complex; for example, administrative records are subject to changes in procedures which may seriously affect comparability through time;
- in contrast to surveys of households and businesses, administrative records produced by program departments of government are not under the control of statisticians, and their use as a data source requires collaboration of administrators well beyond that required of respondents to most surveys;
- responsibility for the processes involved in using administrative files for statistical purposes is diffused, frequently overlapping between the central statistical office and other agencies;
- there are frequently uncertainties as to who has access to data files, under what authority, and with what rights and obligations;
- there is a fairly widespread perception amongst statisticians that their business is surveys, and that while the use of administrative records is not to be inhibited, neither is it a priority task to be supported with the dedication of significant resources responsible for researching, investigating, co-ordinating, and producing statistics derived from administrative records;
- frequently administrative records do not fit logically into one subject-matter area or one sphere of observation, and the old proverb "everybody's business is nobody's business" applies.

6. PREREQUISITES FOR UTILIZATION

The above represents a formidable array of problems. In addition, the following are considered to be almost prerequisites for extensive exploitation:

i) Co-ordination in terms of standards, concepts, definitions, classifications, to ensure comparability of files containing data relating to similar subject-matter and variables. However, even more important than the co-ordination of such technical matters is the development of a spirit of co-operation between administrators and statisticians, based on mutual understanding of each other's problems, objectives and contributions.

ii) The existence of a central authority to develop such standards in consultation with those acquiring and holding the administrative records, and to work towards implementation of the standards, either vested with legal authority or being persuasive as it largely is in the case of Statistics Canada.

iii) Some organization which will give sustained support to the use of administrative records as a priority.

iv) An organization which develops and maintains an inventory of administrative records, including good documentation of machine-readable (and other) data files. The organization responsible for compiling such an inventory, and disseminating information about its content, should have authority to acquire required documentation about each file maintained in the jurisdiction involved, e.g., the departments and agencies of the central government.

v) The central statistical agency should have a legal right of access to administrative records of program departments and agencies, probably with some neutral agency monitoring the effectiveness of the relationships between statisticians and administrators. This is now the case, for taxation records.

vi) A central statistical agency which does have right of access must operate under strong confidentiality laws, practices, and procedures. These are universally agreed upon as essential. In terms of modern technology they are also required to enable linkage of data files (from both administrative and survey sources); to develop and implement policies and procedures for the dissemination of micro-data files containing only unidentifiable records; and to constrain the unnecessary production of data bases containing the combined data from many separate systems.

vii) The critical accumulations of technical skills and computer capacity required to acquire, edit, link and process administrative records, as well as evaluate, analyze and disseminate the statistics derived from them.

viii) A willingness on the part of administrators, based on enlightened self-interest, to utilize their records in collaboration with statisticians for the purpose of publishing statistics based on them, or to make them available to statisticians for such purposes; and also to perhaps modify their own procedures somewhat for a recognized common good. Statisticians also need to equip themselves with the expertise to participate in the design of systems of administrative data files in such a way as to assist administrators and at the same time meet their own needs.

7. RECENT DEVELOPMENTS

Some of these prerequisites have existed in Canada for some while - most importantly the relatively strong central statistical agency with legal right of access to administrative records; laws and practices protecting confidentiality; much of the required skill and technology and co-operation between statisticians and administrators. (It is worth noting that the "right of access" referred to above is only exercised in collaboration with federal, provincial and municipal departments and agencies, and is not enforced by either confrontation or reference to the courts).

More recently a number of additional steps have been taken to improve the co-ordination of all statistical activities within the federal government and a number of them relate to the above prerequisites. A committee of senior government officials on federal statistical activities, chaired by the Chief Statistician, has been created. This committee is concerned with questions of access, priority-setting and planning for the statistical system. Under its aegis a federal statistical clearinghouse has been set up to prepare an inventory of machine-readable files, to promote their use, to facilitate assessment and cost evaluation, and to improve processing. It is intended that the clearinghouse continuously collect documentation about federal data files, prepare and distribute catalogues and disseminate information about classifications, codes and data handling.

Also, the federal government has passed human rights legislation within the past few months which is intended to co-ordinate the collection, retention, use and storage of information by government institutions. The legislation entitles individuals to ascertain what records concerning them are held within the control of a federal government institution and used for administrative purposes. To this end it calls for the annual publication of the name or other identification of each federal information bank and the types of records contained in it. The clearinghouse operation referred to above will be used to provide such a list. The individual's right to access to records about himself or herself relates only to files used for administrative purposes and not to files used for statistical purposes. However, the legislation does impact upon the statistical system, since it requires the responsible cabinet minister (who may or may not be the minister responsible for Statistics Canada) to cause to be kept under review the utilization of existing information banks and to approve proposals for the creation of new information banks or the substantial modification of existing ones. For a more extensive description of this legislation and its impact see [3] (Part 3, Privacy: Legal Basis in Canada). Statistics Canada is expected to participate in the review of proposals for new and substantially modified information banks, in much the same way that it has been responsible in recent years for reviewing new survey proposals for purposes of avoiding duplication, extending the use of existing survey data and co-ordinating survey activities of the federal government.

As previously indicated all the provinces produce extensive administrative files, and most of them have provincial statistical bureaus supported by legislation similar to that of Statistics Canada. A Consultative Council on Statistical Policy comprising directors of provincial statistical bureaus (or others where such bureaus do not exist) which is also chaired by the Chief Statistician, meets annually to review programs which involve the provinces as both users and producers of

statistics. This council recently created a federal-provincial committee on administrative records to initiate the development of inventories of administrative files held in the provinces, similar to that being undertaken by the federal clearinghouse referred to earlier in this paper.

In Statistics Canada an office of co-ordination of administrative records has recently been created as part of a restructuring of Statistics Canada. The co-ordinating activities to be carried by the office are now being developed.

Statistics Canada is also giving priority to the use of administrative records by assigning budgetary priority to such activities even in a period of severe financial constraint. How this budgetary priority can be made manifest also remains to be determined.

Statistics Canada has also developed a policy for the release of micro-data files which ensures that the observations contained in the files cannot be identified with any particular individual and a number of such files are now released routinely. Taxation-derived data, except where the release is provided for explicitly under existing legislation (e.g., CALURA) are, of course, fully confidential under the Income Tax Act. Finally, a limited number of files have been linked, e.g., cancer and death files, and a policy regarding record linkage is being developed by the office of co-ordination referred to above.

These recent developments provide important additions to a long established base which has permitted the utilization of administrative records as has been described in this paper. It is anticipated that, taken together, they will make possible new and much more extensive exploitation of the very substantial statistical potential that lies within administrative records.

RESUME

Cet article suggère que les dossiers administratifs sont une source puissante de statistiques. Afin d'appuyer cette conclusion on fournit une description générale de l'exploitation au Canada des dossiers administratifs à des fins statistiques. L'article traite de certains développements récents et de l'environnement changeant qui sont considérés comme les principaux déterminants de la création et de l'exploitation des bases de données administratifs. On estime que les capacités de l'ordinateur, avec la demande extensive pour les statistiques et les ressources financières limitées disponibles pour satisfaire cette demande, mènent conjointement vers une exploitation plus étendue des dossiers administratifs. On spécifie une variété de problèmes qui sont associés avec l'utilisation des dossiers administratifs et on décrit le développement de stratégies qui répondraient à ces problèmes et qui permettraient l'exploitation des dossiers administratifs. On indique certains développements récents au Canada qui devraient appuyer l'utilisation des dossiers administratifs.

REFERENCES

[1] Ohlsson, I., "Statistical Aspects of Computerized Administrative Information Systems", paper prepared by the National Central Bureau of Statistics, Sweden for the Conference of European Statisticians, June 1973 - Restricted CES/R.1, 8 May 1973, United Nations Economic and Social Council.

[2] Podoluk, J.R., "Administrative Data as a Source of Statistics", paper prepared by Statistics Canada for the Eighth Conference of Commonwealth Statisticians, December, 1975.

[3] Worton, D.A., "Confidentiality of Statistical Information", paper prepared by Statistics Canada for the XIII Session of COINS of the Inter-American Statistical Institute, November 1977.

THE EFFECT OF A TWO-STAGE SAMPLE DESIGN ON TESTS OF INDEPENDENCE IN A 2 by 2 TABLE

J. Cowan and D.A. Binder[1]

When a two-stage sample is used to collect data, the correlations between the sampled units make the χ^2 test of independence invalid. Use of the ordinary χ^2 tables generally results in a test which is greater than the desired level of significance. The effect of the sample design comes from two main areas: the sample size within PSU's and the degree to which the characteristics are independent within each PSU. The effect of the sample size within PSU's is greatest when there is no independence within each PSU, and diminishes as the degree of independence increases.

1. INTRODUCTION

Classical statistical inference has been developed through the years under the assumption of independent observations. In recent years, attention has turned to attempts to develop data analysis techniques for complex sampling procedures, especially in the area of social surveys (Kish & Frankel [5]). The Canada Health Survey has set up a data analysis group to monitor new developments in this field, to attempt to adapt existing techniques for its own use, and to look into new areas. This paper deals with one of these new areas.

One of the basic statistical tools currently in popular use is contingency table analysis for testing the independence of two or more characteristics in a population. One of the key assumptions in developing the distributional theory is the independence of the observations. This leads to the multinomial distribution and, asymptotically, the chi-squared test. If the assumption of independence is violated, as in a complex sample survey, this theory loses its validity. Several studies have been done on the analysis of contingency tables

---

[1] J. Cowan and D.A. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada.

when correlations are present between observations, notably by Cohen [4], Altham [1], and Nathan [6]. Shuster and Downing [7] have derived a test statistic applicable for any sampling scheme.

The question has been raised as to how much the sampling scheme affects the inference. For example, is it possible to assume independent observations without distorting the true situation too much. The answer will depend, of course, on the sampling scheme and what the analysts consider "too much". A study is currently being done which empirically investigates the effect that a two-stage sample has on inferences about independence of two characteristics in a population, although the theory developed can be applied to any self-weighting sampling scheme.

2. DISTRIBUTION OF TEST STATISTICS

Denote the usual Pearson chi-square statistics by P:

$$P = \sum_{i,j} \frac{(n\hat{\Pi}_{ij} - n\hat{\Pi}_{i.}\hat{\Pi}_{.j})^2}{n\hat{\Pi}_{i.}\hat{\Pi}_{.j}},$$ where the $\hat{\Pi}$ are the MLE's of Π in the case of i.i.d. sampling.

Next, approximate P by taking its Taylor series expansion around the point $\{\Pi\colon \Pi_{ij} = \Pi_{i.}\Pi_{.j}\}$, where the null hypothesis of independence is true. It can easily be shown that if $\hat{\Pi}$ is unbiased for Π, the constant and linear terms are zero. To evaluate the quadratic term, we find that

$$\left.\frac{\delta P}{\delta\hat{\Pi}_{pq}\,\delta\hat{\Pi}_{rs}}\right|_{\hat{\Pi}=\Pi} = \begin{cases} 2n\left(2 - \frac{1}{\Pi_{p.}} - \frac{1}{\Pi_{.q}} + \frac{1}{\Pi_{pq}}\right) & \text{if } p=r,\ q=s \\ 2n\left(2 - \frac{1}{\Pi_{p.}}\right) & \text{if } p=r,\ q\neq s \\ 2n\left(2 - \frac{1}{\Pi_{.q}}\right) & \text{if } p\neq r,\ q=s \\ 2n\ (2) & \text{if } p\neq r,\ q\neq s \end{cases}$$

generally not at the desired level. Since the asymptotic distribution of P can be calculated for large m, the actual level of the test can be empirically investigated by generating random variables with this distribution and calculating the proportion which are greater than the critical value. With independent observations, this proportion will be α on the average.

3.1 Method

The first step in calculating the distribution of P is to calculate A, the matrix of the quadratic form which approximates the statistic. This is done by specifying the parameter Π, the population proportions falling into each category. The next step in calculating the distribution of P is to derive V, the matrix of variances and covariances of $\sqrt{mh}\ \hat{\Pi}$ for this particular sample design. It was decided to take M as infinite, as this would be a fairly accurate approximation to the Canada Health Survey, and the calculations are simplified. The derivation of V is shown in Appendix A. It is necessary to specify the distribution of ${}_i\Pi = ({}_i\Pi_1, {}_i\Pi_2, {}_i\Pi_3, {}_i\Pi_4)$ over all PSU's in order to calculate V. For simplicity the Dirichlet distribution was used. Appendix B gives the properties of this distribution.

The calculation of the eigenvalues of AV was done by a routine due to Sparks and Todd [8], and generation of normal random variables was done by a routine due to Marsaglia and Bray (see [2]). For fixed values of the parameters, the approximation to P was generated 10,000 times and the proportions which were greater than the .10, .05, .01, and .001 level critical values were recorded. Many combinations of the parameters were tried and the results are given in the next section.

3.2 Results

In all cases, three of the four eigenvalues are negligible and can be taken as zero, so the distribution of the Taylor series approximation to P reduces to that of a chi-squared random variable with one degree

of freedom multiplied by the largest eigenvalue (λ_1) of the matrix AV. This eigenvalue determines the level of the test. If it is greater than 1, the null hypothesis will incorrectly be rejected more often than 100 α percent of the time, and if less than 1, the level drops below α. The effects of the parameters are described below.

Appendix C gives the results of some of the simulations. Although this is far from a complete listing of the possible combinations of parameters, it still gives some idea how the level of significance is changed by the two-stage design.

a. Effect of Π (population proportions)

If all other parameters are fixed, a change in the population proportions falling into each of the four categories results in a negligible change in λ_1. This means that the level of the test does not depend on Π. Individually, the A, V and AV matrices are changed as Π changes, but the matrix AV will have the same eigenvalues.

b. Effect of h, H (within-PSU sample and population sizes)

For fixed h, variation in H gives limited variation in λ_1 but for fixed H, variation in h results in a great deal of variation in λ_1 (see Appendix C). Table 1 displays the largest eigenvalues for various values of H and h. It can be seen that an increase in the PSU population size for fixed number of SSU's within each PSU results in a slight increase in the significance level. This is caused by an increase in the covariance between the elements of $\hat{\Pi}$ as H increases. The table also shows that the effect of the sample size within PSU's is far greater than that of the PSU population size, and that the sampling fraction within PSU's by itself is not informative. This is due to the clustering effect of taking more SSU's within PSU's.

Let $A = \frac{1}{2n}\left(\frac{\delta P}{\delta \hat{\Pi}_{pq}\ \delta \hat{\Pi}_{rs}}\bigg|_{\hat{\Pi} = \Pi}\right)$;

then P is approximated by:

$$P \doteq n(\hat{\Pi} - \Pi)'\ A(\hat{\Pi} - \Pi) = Q\ .$$

Since $\hat{\Pi}$ is assumed to be asymptotically normal, the distribution of Q is the weighted sum of independent central chi-squared random variables, each on one degree of freedom, where the weights $(\lambda_1, \lambda_2, \ldots)$ are the eigenvalues of the matrix AV (see Box [3], Theorem 2.1). Since the asymptotic distribution of Q is known, it is possible to perform an empirical study of the effect that the sample design has on inference by generating values of Q and calculating the proportion of these values which are greater than the α-level critical point of the usual χ^2 - test. This should give some idea how the probability of a Type I error is changed by the sampling scheme.

3. SIMPLE TWO STAGE DESIGNS

The distribution derived for the usual test of independence is used to examine the effect that a two-stage sample has on Type I errors in testing the independence of two characteristics, each with two categories. The population is divided into M primary sampling units (PSU), from which m are drawn. Each PSU contains H units from which h secondary sampling units (SSU) are chosen. Each sampled unit is then classified according to the two characteristics, so a 2 by 2 table can be constructed displaying the proportions of the sample population falling into each of the four categories, and P can be calculated. If the observations had been independent, this statistic asymptotically would have a chi-squared distribution with one degree of freedom. The null hypothesis of independence of the characteristics in the population could be tested by comparing the observed value of the statistic with the χ^2 tables. Since the observations are not independent, use of the α - level critical point from the χ^2 tables results in a test which is

TABLE 1:

H	h	λ_1
1000	50	9.13
500	50	9.08
300	50	9.03
100	50	8.75
1000	40	7.47
800	40	7.46
100	40	7.17
1000	30	5.81
400	30	5.77
100	30	5.59
1000	20	4.15
750	20	4.15
500	20	4.13
100	20	4.01
700	15	3.32
400	15	3.30
100	15	3.22
1000	10	2.49
500	10	2.49
250	10	2.47
100	10	2.42

λ_1 = largest eigenvalue of AV when $\theta. = \sum_{i=1}^{4} \theta_i = 5$

c. Effect of $\theta.$ (parameter of the distribution of $_i\Pi$)

Small values of $\theta.(= \sum_{i=1}^{4} \theta_i)$ give large values of λ_1, and as $\theta.$ increases to infinity, λ_1 decreases to $(1-\frac{h}{H})$. A typical example is for the case H = 300, h = 50 (Table C4): $1-\frac{h}{H} = 0.833$.

$\theta.$	λ_1
5	9.030
25	2.727
125	1.226
625	0.915
3125	0.852
15625	0.839
78125	0.837

This result shows that the more independent the characteristics are within strata, the less the size of the test. Notice that if all units within strata are sampled (h = H), then independence

within strata gives a value of 0 for the approximation to P, as it should. Another result is that for small $\theta_.$, variation in h causes wide variation in λ_1, but for large $\theta_.$, the variation in λ_1 becomes a great deal smaller.

4. FUTURE DIRECTIONS

The results here are very restrictive because the model is simple. The two-stage model was chosen to be as close to multinomial sampling as possible. However, because the usual chi-squared behaves so poorly here, we would expect things would get worse in a more realistic setting.

One interesting point is that all the above results yielded only one dominant eigenvalue. Is this true for more realistic settings (e.g. unequal sized PSU's, unequal probability samples)? If so then we may be able to estimate the largest eigenvalue and derive a statistic whose distribution is much closer to chi-squared. Also the use of the Dirichlet distribution is only one of many possible distributions that could be considered.

APPENDIX A: Derivation of V

Let $u_i = 1$ if PSU i, is in the sample $i = 1, \ldots, \infty$
$= 0$ otherwise .

Let $x_{ijk} = 1$ if the j^{th} unit in the i^{th} PSU belongs to category k
$= 0$ otherwise.

Then $\bar{X}_{i.k}$ is the proportion of units in PSU i that belong to category k,

and $\hat{\bar{X}}_{i.k}$ is the proportion of sampled SSU's in PSU i that belong to category k.

Let $S_{ik\ell} = \frac{1}{H-1} \sum_{j=1}^{H} (x_{ijk} - \bar{X}_{i.k})(x_{ij\ell} - \bar{X}_{i.\ell})$.

Now, $y_k = \sum_{i=1}^{\infty} u_i \, h \, \hat{\bar{X}}_{i.k}$ = total number of sampled units that belong to category k

and $E(y_k) = E_{\underset{\sim}{u}}\left(E(y_k|u)\right) = h \, E_{\underset{\sim}{u}}\left(\sum_{i=1}^{\infty} u_i \, \bar{X}_{i.k}\right) = hm \, E(\bar{X}_{i.k})$.

$$\text{Also, } \operatorname{Cov}(y_k, y_\ell) = E_{\underset{\sim}{u}}\left(\operatorname{Cov}(y_k, y_\ell|\underset{\sim}{u})\right) + \operatorname{Cov}_{\underset{\sim}{u}}\left[E(y_k|\underset{\sim}{u}), E(y_\ell|\underset{\sim}{u})\right]$$

$$= mh(1 - \tfrac{h}{H}) \, E(S_{ik\ell}) + h^2 m\left[E(\bar{X}_{i.k} \, \bar{X}_{i.\ell}) - E(\bar{X}_{i.k}) \, E(\bar{X}_{i.\ell})\right],$$

and since $E(S_{ik\ell}) = \frac{H}{H-1} E\{\delta_{k\ell} \bar{X}_{i.k} - \bar{X}_{i.k} \bar{X}_{i.\ell}\}$ ($\delta_{k\ell} = 1$ if $k = \ell$
$= 0$ otherwise)

we have:
$$\operatorname{Cov}\left(\frac{y_k}{\sqrt{mh}}, \frac{y_\ell}{\sqrt{mh}}\right) = \frac{H-h}{H-1} E\{\delta_{k\ell} \bar{X}_{i.k} - \bar{X}_{i.k} \bar{X}_{i.\ell}\}$$
$$+ h\left[E(\bar{X}_{i.k} \, \bar{X}_{i.\ell}) - E(\bar{X}_{i.k}) \, E(\bar{X}_{i.\ell})\right].$$

APPENDIX B: Distribution of ${}_i\Pi$

Within PSU i, we know that the sum of the proportions in each category is 1, or $\sum_{j=1}^{4} {}_i\Pi_j = 1$ for $i = 1, 2, \ldots \infty$. We also know that the average value over all strata of the proportion in category j is Π_j, or $\lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} {}_i\Pi_j = \Pi_j$ for $j = 1, 2, 3, 4$. One simple way to accomplish this is to let ${}_i\Pi$ follow a multivariate analogue of the Beta distribution, called the Dirichlet distribution.

If the random vector $(y_1, \ldots, y_k)$ follows a Dirichlet distribution with parameters $\theta_1, \ldots, \theta_k$, then: $\sum_{i=1}^{k} y_i = 1$; $y_1 \geq 0$, $i = 1, \ldots, k$,

the density is given by

$$\frac{\Gamma\left(\sum_{i=1}^{k} \theta_i\right)}{\prod_{i=1}^{k} \Gamma(\theta_i)} \prod_{i=1}^{k} y_i^{\theta_i - 1},$$

$$E(y_i) = \frac{\theta_i}{\Sigma\theta_j}, \quad E(y_i^2) = \frac{\theta_i(\theta_i+1)}{(\Sigma\theta_j)(\Sigma\theta_j+1)},$$

$$E(y_i y_k) = \frac{\theta_i\, \theta_k}{(\Sigma\theta_j)(\Sigma\theta_j+1)}.$$

To apply this distribution to the calculation of V, we specify Π_k, $k = 1, 2, 3, 4$ and equate $\frac{\theta_k}{\Sigma\theta_j}$ to Π_k. This determines $(\theta_1, \theta_2, \theta_3, \theta_4)$ to within a constant multiple. Various values for $\theta. = \sum_{i=1}^{4} \theta_i$ will then determine various degrees of independence of the two characteristics within PSU's. As $\theta.$ increases, $\mathrm{Var}({}_i\Pi_j)$ decreases to zero for all j so that the distribution of the proportions of units falling into each of the categories is identical from PSU to PSU. Since the average value over all PSU's is Π_k, the within PSU distribution must be identical to the population proportion, and since independence holds in the population, it necessarily holds within PSU.

APPENDIX C: Proportion of times the approximation to P is greater than the critical point at the given significance level

C1. # UNITS/PSU = 400
SSU /PSU = 30

max eigenvalue :	5.77277	2.04550	1.15805	.97376	.93662	.92918	.92769
θ. sums to →	5	25	125	625	3125	15625	78125
↓ significance level							
.10	.491	.250	.130	.094	.089	.088	.088
.05	.412	.168	.070	.046	.042	.043	.042
.01	.282	.071	.017	.008	.007	.008	.006
.001	.171	.021	.002	.0008	.0007	.0006	.0006

C2. # UNITS/PSU = 400
SSU /PSU = 15

max eigenvalue :	3.30409	1.50472	1.07630	.98733	.96940	.96581	.96509
θ. sums to →	5	25	125	625	3125	15625	78125
↓ significance level							
.10	.364	.175	.110	.098	.093	.093	.093
.05	.283	.108	.057	.049	.047	.044	.045
.01	.157	.035	.012	.009	.009	.009	.008
.001	.068	.007	.002	.0009	.0008	.0007	.0009

C3. # UNITS/PSU = 500
SSU /PSU = 10

max eigenvalue :	2.48497	1.32881	1.05353	.99637	.98485	.98254	.98208
θ. sums to →	5	25	125	625	3125	15625	78125
↓ significane level							
.10	.300	.155	.108	.097	.095	.095	.094
.05	.215	.088	.055	.049	.048	.048	.047
.01	.102	.025	.012	.009	.009	.009	.009
.001	.037	.004	.0012	.0009	.0009	.0009	.0008

C4. # UNITS/PSU = 300
SSU /PSU = 50

max eigenvalue :	9.03010	2.72704	1.22631	.91466	.85185	.83927	.83675
θ. sums to →	5	25	125	625	3125	15625	78125
↓ significance level							
.10	.587	.320	.136	.084	.073	.068	.069
.05	.516	.236	.077	.041	.033	.030	.031
.01	.395	.119	.019	.007	.005	.005	.005
.001	.274	.046	.0025	.0004	.0004	.0002	.0004

C5. # UNITS/PSU = 250
SSU /PSU = 10

max eigenvalue :	2.46988	1.31140	1.03557	.97829	.96674	.96443	.96397
θ. sums to →	5	25	125	625	3125	15625	78125
↓ significance level							
.10	.293	.150	.104	.096	.095	.092	.092
.05	.212	.086	.053	.047	.045	.045	.045
.01	.099	.025	.011	.009	.008	.008	.008
.001	.035	.004	.0013	.0010	.0007	.0007	.0008

C6. # UNITS/PSU = 500
SSU /PSU = 20

max eigenvalue :	4.13494	1.69416	1.11302	.99234	.96801	.96314	.96217
θ. sums to →	5	25	125	625	3125	15625	78125
↓ significance level							
.10	.415	.204	.117	.098	.095	.092	.092
.05	.333	.131	.062	.048	.045	.046	.044
.01	.204	.048	.014	.009	.008	.008	.008
.001	.105	.012	.0020	.0011	.0008	.0008	.0009

C7. # UNITS/PSU = 750
SSU /PSU = 20

max eigenvalue :	4.14553	1.70638	1.12563	1.00502	.98072	.97585	.97487
θ. sums to →	5	25	125	625	3125	15625	78125
↓ significance level							
.10	.417	.211	.120	.100	.093	.095	.093
.05	.337	.138	.064	.051	.048	.047	.047
.01	.207	.051	.017	.010	.009	.008	.008
.001	.108	.012	.002	.0012	.0010	.0009	.0007

C8. # UNITS/PSU = 500
SSU /PSU = 50

max eigenvalue :	9.08484	2.79020	1.29147	.98023	.91751	.90494	.90243
θ. sums to →	5	25	125	625	3125	15625	78125
↓ significance level							
.10	.582	.325	.148	.095	.083	.082	.083
.05	.513	.244	.086	.047	.040	.037	.038
.01	.393	.125	.023	.010	.008	.007	.007
.001	.275	.050	.004	.0009	.0007	.0006	.0005

RESUME

Quand on utilise un échantillon à deux degrés pour rassembler des données, les corrélations entre les unités échantillonnées rendent le test d'indépendance χ^2 invalide. Si on utilise les tables ordinaires de χ^2, on obtient généralement un test qui est plus grand que le seuil significatif voulu. L'effet du plan d'échantillonnage provient de deux facteurs principaux: la taille de l'échantillon dans les UPE et le degré d'indépendance des caractéristiques dans chaque UPE. L'effet de la taille de l'échantillon dans les UPE est à son maximum quand il n'y a pas d'indépendance dans chaque UPE, et diminue à mesure que le degré d'indépendance augmente.

REFERENCES

[1] Altham, P.M.E. (1976). "Discrete variable analysis for individuals grouped into families", Biometrika, 63, 2, 263-269.

[2] Atkinson, A.C. and Pearce, M.C. (1976). "The computer generation of beta, gamma, and normal random variables", JRSS, A, 139, 431-461.

[3] Box, G.E.P. (1954). "Some theorems on quadratic forms applied in the study of analysis of variance problems I", AMS, 25, 2, 290-302.

[4] Cohen, J.E. (1976). "The distribution of the chi-squared statistic under clustered sampling from contingency tables", JASA 71, 665-670.

[5] Kish, L. and Frankel, M.R. (1974). "Inference from complex samples", JRSS(B), 36, 1, 1-37.

[6] Nathan, G. (1975). "Tests of independence in contingency tables from stratified proportional samples", Sankhya (C), 37, 77-87.

[7] Shuster, J.J. and Downing, D.J. (1976). "Two-way contingency tables for complex sampling schemes", Biometrika, 63, 271-276.

[8] Sparks, D.N. and Todd, A.D. (1973). "Latent roots and vectors of a symmetric matrix", Applied Statistics, 22, 2, 260-265.

APPROXIMATE TESTS OF INDEPENDENCE AND GOODNESS OF FIT BASED ON STRATIFIED MULTI-STAGE SAMPLES

I.P. Fellegi[1]

The impact on linear statistics of the sample design used in obtaining survey data is the subject of much of sampling literature. Recently, more attention has been paid to the design's impact on non-linear statistics; the major factor inhibiting these investigations has been the problem of estimating at least the first two moments of such statistics. The present article examines the problem of estimating the variances of non-linear statistics from complex samples, in the light of existing literature. The behaviour of the chi-square statistic computed from a complex sample to test hypotheses of goodness of fit or independence is studied. Alternative tests are developed and their properties studied in simulation experiments.

1. INTRODUCTION

The impact of the actual sample design used in obtaining data from a given survey has been recognized and studied by a number of authors. Its impact on linear statistics (e.g. population means and totals) has, of course, been the main subject of a large part of sampling literature. In the last ten years, or so, increasing attention has been paid to this impact as it affects non-linear statistics -- regression coefficients, correlations, multiple and partial correlations, etc. Among numerous related papers, the landmark contribution of Kish and Frankel [4] must be mentioned.

The major limiting factor inhibiting the investigation of the impact of the actual sample design on non-linear statistics has been the problem of estimating at least the first two moments of such statistics. It is well known that even from complex stratified multi-stage cluster samples, if the design is self-weighting (i.e. the inclusion probability of each unit in the population is the same), approximately unbiased

---

[1] I.P. Fellegi, Assistant Chief Statistician, Statistical Services Field, Statistics Canada.

estimates can be obtained of the population variances and covariances of the variables which were collected. Moreover, these estimates are formally identical to those derived under the assumption of simple random sampling. Therefore, consistent estimates are available for those non-linear statistics which can be constructed as functions of the estimated population variances and covariances. The estimation of the variance of such non-linear statistics has, however, been a major obstacle -- certainly variance estimators assuming simple random sampling can be quite misleading, as Kish and Frankel [4] have shown.

For purposes of the present paper the most important development enabling the estimation of variances of non-linear statistics from complex samples is the paper by McCarthy [5]. McCarthy's method of variance estimation, known as balanced repeated replication (BRR), is predicated only on the availability of two primary sampling units being selected in each stratum independently (or at least with a correlation between them which is negligible) and a within primary sampling unit sample design which is independent (although not necessarily identical) as between the two psu's of each stratum. The procedure boils down to forming two overall half-samples by combining one of the two psu's from each stratum. Any overall statistic which can be estimated from the complete stratum can also be estimated from each of the two half-samples. If there are L strata, there are 2^L different ways of forming half-samples and each of these is called a replicate. The following points have been made by McCarthy or subsequent authors:

a) If $\hat{T}^{(k)}$ and $\tilde{T}^{(k)}$ are statistics based on the two half-samples of the k-th replicate, $\bar{T}$ is the corresponding estimate made from the full sample (we will use throughout the paper the symbols ^ and ~ to refer to estimates derived from half-samples and - to estimates derived from the complete sample), then

$$(\hat{T}^{(k)} + \tilde{T}^{(k)})\ /2 = \bar{T} \qquad \text{for all } k \tag{1}$$

if $\bar{T}$ is a linear statistic. However, even for non-linear statistics (1) is found to apply approximately.

b) The variance of $\bar{T}$ can be estimated by each of the following expressions:

$$\left.\begin{aligned} v_1^k &= (\hat{T}^{(k)} - \bar{T})^2 \\ v_2^k &= (\tilde{T}^{(k)} - \bar{T})^2 \\ v_3^k &= (\hat{T}^{(k)} - \tilde{T}^{(k)})^2/4 \end{aligned}\right\} \quad \text{for all } k.$$

In the case of linear statistics the three expressions are identical and provide an unbiased estimate of $\mathrm{Var}(\bar{T})$. In the case of non-linear statistics the three estimates are observed to be very close.

c) In the case of linear statistics, there is a way of creating K replicates ($L \leq K \leq L+3$), called balanced repeated replication (BRR), in such a fashion that these K out of the 2^L possible replicates capture all the available L degrees of freedom for estimating the variance of $\bar{T}$. In this case

$$\frac{1}{K} \sum_{k=1}^{K} v_i^{(k)} \qquad i=1,\ 2,\ 3$$

provides an unbiased variance estimate of $\bar{T}$ with L degrees of freedom. The same phenomenon is conjectured to hold approximately for non-linear statistics.

The main contribution of the present paper is to call attention to the fact that the chi square statistic, when computed from a complex sample to test a hypothesis of goodness of fit or of independence in a contingency table, behaves in a way which is fundamentally different from that of the many common descriptive statistics investigated by Kish and Frankel (e.g. regression coefficients, multiple and partial correlations) -- not only the dispersion of the statistic is altered (with the

mean being more or less unchanged) but the distribution is both shifted and the dispersion affected in a more or less predictable way. Alternative tests are also developed and their behaviour is studied in simulation experiments.

Consider, first of all, the chi square statistic as a test of goodness of fit. First, assume a simple random sample. Let there be m categories and denote the number of observations in the sample of n units which falls into the i-th category by

$$\bar{T}_i \qquad\qquad i=1, \ldots, m$$

where $\sum_i \bar{T}_i = n.$

To test the null hypothesis H_o, that

$$E(\bar{T}_i/n) = P_i \qquad\qquad i=1, \ldots, m,$$

the statistic s is computed

$$s = \sum_{i=1}^{m} \frac{(\bar{T}_i - nP_i)^2}{nP_i} \tag{2}$$

and, as is well known, it is distributed asymptotically under H_o as chi square with m-1 degrees of freedom.

Now assume that the estimates $\bar{T}_i$ arise from a complex self-weighting design. The variance of $\bar{T}_i$ is modified. Kish calls the quantity below the design effect

$$\mathrm{Deff}_i = \mathrm{Var}(\bar{T}_i)/nP_i(1-P_i) \quad .$$

The value of Deff depends on the nature of the sample design and the variables being measured, and in well-designed surveys it ranges typically between 1 and 3, although values as high as 6 have been reported and the most common values appear to be between 1.4 and 2. If all the values Deff_i above are equal, then s does not have the chi square distribution. It is conjectured (and supported by empirical

investigations of the last section) that the distribution might be close to Deff times the χ^2 distribution. At least the mean of s can easily be shown to be multiplied by Deff. More generally, if Deff_i are not equal, then s is conjectured to have the distribution of chi square multiplied by a weighted average of the Deff_i. The impact of Deff (or its applicable weighted average) is profound on the achieved significance levels (under the null hypothesis) as the table below indicates.

Table 1

Achieved significance levels (as opposed to a nominal significance level of 0.05) for different values of Deff and different degrees of freedom

df \ Deff	1.1	1.2	1.4	1.6	1.8	2.0	2.5	3	4	5	6
1	0.062	0.074	0.098	0.121	0.144	0.166	0.215	0.258	0.327	0.381	0.424
2	0.066	0.082	0.118	0.154	0.189	0.224	0.302	0.368	0.473	0.549	0.607
3	0.069	0.089	0.134	0.181	0.227	0.272	0.373	0.457	0.582	0.668	0.729
4	0.071	0.095	0.148	0.204	0.261	0.315	0.434	0.531	0.668	0.755	0.812
5	0.073	0.100	0.161	0.227	0.292	0.354	.491	0.595	0.736	0.819	0.870
6	0.076	.105	.174	.248	.321	.391	.539	.650	.790	.866	.910
7	.077	.110	.186	.268	.349	.425	.584	.698	.833	.902	.938
8	.079	.115	.197	.287	.376	.458	.625	.739	.868	.928	.958
9	.081	.119	.209	.306	.401	.489	.661	.775	.896	.947	.971
10	.083	.123	.219	.324	.426	.518	.695	.807	.918	.961	.980
15	.091	.142	.270	.408	.534	.641	.820	.910	.975	.992	.997
20	.097	.160	.317	.481	.624	.735	.874	.959	.993	.999	1.000
25	.103	.177	.361	.546	.697	.805	.940	.981	.998	1.000	1.000
30	.109	.193	.402	.604	.757	.858	.955	.992	.999	1.000	1.000

A quick perusal of the table should indicate that the use of the standard chi square as a test of goodness of fit can be drastically misleading. It will be seen in subsequent sections of this paper (particularly the simulation results) that much the same holds for chi square tests of independence in contingency tables. In fact, the achieved significance levels are particularly misleading for higher degrees of freedom and, not surprisingly, for higher values of Deff. But even with the degrees of freedom only as large as 4 (or more) and with Deff as large as or larger than 1.6, the significance tests are practically useless: the null hypothesis would be rejected with a probability of 0.2, rapidly rising to .5 or more with larger values of Deff and/or larger degrees of freedom. It should be kept in mind that large degrees of freedom arise quite commonly in the case of contingency tables: e.g. a 4x5 table gives rise to 12 degrees of freedom, a 5x6 table to 20.

One final introductory note on the chi square test is necessary. Conceptually, one can at least contemplate using the test in the case of self-weighting designs (although, as seen above, the results can be most misleading). However, in a not negligible proportion of sample designs actually used in practice the inclusion probability of all units in the population is not equal. In this case the unweighted sample frequencies will not provide unbiased estimates of the corresponding population statistics, thus even if H_o holds in the population as a whole, one would expect that a chi square test based on the unweighted sample frequencies would be rejected far more often than the nominal significance level (whether used as a test of independence or of goodness of fit). However, the chi square statistic (2) does not lend itself to weighting at all, because the numerator increases with the square of any weight whereas the denominator increases linearly.

2. OBSERVATIONS ON RELATED WORK IN THE LITERATURE

Cohen [3] investigates a very special case of the general problem of testing goodness of fit from complex samples. He assumes a simple random sample of n clusters consisting of two units each. This sample of 2n units is classified into r cells. In the model studied by Cohen, if p_i is the probability of unit 1 of a cluster being in cell i, then the probability of the two units being in cells i and j respectively is

$$p_{ij} = (1-a)\ p_i p_j \qquad \text{if } i \neq j$$

$$p_{ii} = p_i\ [a + (1-a)p_i]$$

for values of a between 0 and 1.

Under this model it can easily be shown that

$$\mathrm{Var}(\bar{T}_i) = (1+a)\ 2n\ p_i\ (1-p_i) \qquad i=1,\ \ldots,\ m\ .$$

So the design effect Deff is equal to 1+a for all i. Cohen shows that the statistic

$$s/1+a$$

is, in fact, distributed as chi square (under H_o) with m-1 degrees of freedom -- as is conjectured in the introduction more generally.

The most sustained work on tests of independence from complex samples has been carried out by Nathan ([6], [7], [8], [9], [10]). He also reviews the work of several other authors, such as Bhapkar and Koch [1] and Chapman [2].

Consider the usual contingency table: there are r rows, c columns, overall sample size n, P_{ij} are the estimated proportions of frequencies in the (i,j)-th cell (i=1, ..., r; j=1, ..., c). The statistic

$$t = \sum_{i,j} \frac{(n\bar{P}_{ij} - n\bar{P}_{i.}\ \bar{P}_{.j})^2}{n\bar{P}_{i.}\ \bar{P}_{.j}} \tag{3}$$

has approximately the chi square distribution under the null hypothesis <u>if</u> n is based on a simple random sample and is sufficiently large. The quantities $\bar{P}_{.j}$ and $\bar{P}_{i.}$ are obtained by summation over the missing subscripts. Under the null hypothesis the expression

$$\bar{P}_{ij} - \bar{P}_{i.}\,\bar{P}_{.j} \tag{4}$$

has zero expected value but generally an expected value different from zero if the null hypothesis does not hold. The zero expected value of (4) is the result of a number of variance and covariance terms in $E(\bar{P}_{i.}\,\bar{P}_{.j})$ cancelling. In effect, under simple random sampling and the null hypothesis

$$\begin{aligned} E(\bar{P}_{i.}\,\bar{P}_{.j}) &= P_{i.}\,P_{.j} + \mathrm{Var}\,(\bar{P}_{ij}) \\ &\qquad + \sum_{(k,k')\neq(i,j)} \mathrm{Cov}(\bar{P}_{kj},\bar{P}_{ik'}) \\ &= P_{ij} + \frac{1}{n} P_{ij}(1-P_{ij}) - \frac{1}{n} \sum_{(k,k')\ (i,j)} P_{kj}\,P_{ik'} \\ &= E(\bar{P}_{ij}). \end{aligned} \tag{5}$$

In complex surveys the variances and covariances in the penultimate line of (5) would each have to be multiplied by their respective design effect multipliers and therefore may not cancel out -- thus the expected value of $\bar{P}_{ij}-\bar{P}_{i.}\,\bar{P}_{.j}$ may not be equal to zero even under H_o.

The work of both Nathan and Bhapkar and Koch starts out with the construction of an expression involving estimates of P_{ij}, $P_{i.}$ and $P_{.j}$ which has zero expected value under the null hypothesis, even in the case of complex samples. For this purpose they both resort to balanced repeated replication and make use of the fact that the two half-samples of any replicate are uncorrelated under the assumptions outlined in the introduction. Thus if $\hat{P}^{(k)}_{ij}$, $\hat{P}^{(k)}_{i.}$, $\hat{P}^{(k)}_{.j}$ are estimated, under a complex sample design, from the first half-sample of the k-th replicate and $\tilde{P}^{(k)}_{ij}$, $\tilde{P}^{(k)}_{i.}$, $\tilde{P}^{(k)}_{.j}$ are the corresponding quantities estimated from the

second half-sample, Nathan's test is based on the expression

$$\bar{U}_{ij}^{(k)}(N) = \hat{P}_{ij}^{(k)} + \tilde{P}_{ij}^{(k)} - \hat{P}_{i.}^{(k)}\tilde{P}_{.j}^{(k)} - \tilde{P}_{i.}^{(k)}\hat{P}_{.j}^{(k)} \tag{6}$$

and Bhapkar and Koch's is based on

$$\bar{U}_{ij}^{(k)}(B) = \hat{P}_{ij}^{(k)}\tilde{P}_{rc}^{(k)} - \hat{P}_{ic}^{(k)}\tilde{P}_{rj}^{(k)} \tag{7}$$

Both (6) and (7) have zero expected values under the null hypothesis. Chapman's test is based on

$$\bar{U}_{ij}^{(k)}(C) = \hat{P}_{ij}^{(k)} - \tilde{P}_{i.}^{(k)}\tilde{P}_{.j}^{(k)} \tag{8}$$

and it does not necessarily have a zero expected value even under H_o.

Now if an estimate $\hat{V}$ can be constructed for the covariance matrix for the (r-1)x(c-1) linearly independent quantities among the rxc U_{ij} values, and if U is the corresponding vector of these values, then

$$\bar{U}' \; (\hat{V})^{-1} \; \bar{U} \tag{9}$$

would, for large enough n, and apart from a suitable constant multiplier, be either distributed approximately as F or a χ^2 -- depending on whether $\hat{V}$ is estimated from a large enough number of degrees of freedom. Note that (9) overcomes the problem of weighting in the case of disproportionate sampling -- its effect would, so to speak, automatically be reflected in $\hat{V}$. The problem, however, is to estimate $\hat{V}$.

In the case of simple random sampling each cell of the covariance matrix of (4) is readily estimated approximately as

$$\hat{v}_{ij,fg} = \frac{1}{n}\bar{P}_{i.}\bar{P}_{.j}\,(1-\bar{P}_{i.})(1-\bar{P}_{.j}); \qquad (i,j)=(f,g)$$

$$= -\frac{1}{n}\bar{P}_{i.}\bar{P}_{.j}\,\bar{P}_{f.}\bar{P}_{.g} \qquad ; \qquad (i,j)\neq(f,g)$$

and the resulting estimate of $\hat{V}$ is based on a large number of degrees of freedom so (9) would be distributed as chi square. In the case of

complex samples the analogous estimation cannot be carried out without some very strong simplifying assumptions.

Alternatively, one may observe that, even in the case of complex samples, the vectors $\bar{U}^{(k)}$ in (6)-(8) are identically distributed and hope to derive estimates of variances and covariances from

$$\sum_{k=1}^{K} (\bar{U}_{ij}^{(k)} - \bar{U}_{ij})(\bar{U}_{fg}^{(k)} - \bar{U}_{fg}) \tag{10}$$

Now if the K replicate values of $\bar{U}_{ij}^{(k)}$ were independent, the expression (10) above, divided by K-1, would provide an unbiased estimate of $v_{ij,fg}$. However, far from being independent, they are very highly correlated. In the case of $\bar{U}^{(k)}(N)$ the correlations are very close to one -- not too surprisingly, since Kish and Frankel noted that for all replicates the sum of two analogous non-linear statistics, computed respectively from the two half-samples, is very nearly the same for all K replicates and is identically the same in the case of linear statistics. In order to correct (10) for the correlations involved, one would have to estimate these correlations and that, in turn, again requires strong simplifying assumptions. Moreover, when the correlations are close to one the numerical behaviour of the estimates is very bad.

Thus whichever of the two methods of estimating the covariance matrix is attempted, strong simplifying assumptions are needed in the case of complex samples. Nathan [9] is forced to make the assumption, among others, that for each stratum h there is a number n_h which depends only on the number of final units selected in stratum h in each of the two primary sampling units (psu), and if $\hat{P}_{ijha}$ is the estimate of the proportions in cell (i,j) derived from psu a (a=1,2) of stratum h, then $n_h \hat{P}_{ijha}$ has approximately the multinomial distribution with parameters n_h and P_{ijh}. However, this assumption implies that the expected value of an estimate $\hat{P}_{ijha}$ derived from any selected psu, conditional on that psu being in the sample, depends on the stratum only and <u>not</u> on

the selected psu. Thus the total between-psu component of variance is assumed away. But, for example, in the case of two stage stratified sampling, with simple random sampling within each of the psu's, this assumes away all the within-stratum design effects. Other assumptions of Nathan, less important to his development, assume away the effects of stratification and disproportionate sampling in different strata as well.

In light of the comments above, it is not too surprising to find that the test statistic proposed by Nathan behaves very badly with respect to its achieved significance levels. The simulation results reported in his paper [9] are flawed, as pointed out by the author in his own subsequent paper, Nathan [10]. The results reported in [10] refer to stratified cluster sampling with a self-weighting design, so the traditional chi square test can be applied and serves as a measure of comparison. The achieved significance levels of his test statistic under H_o are .038, .144 and .190 for the nominal significance levels of .01, .05 and .10 respectively. However, these are almost identical to the achieved significance levels of the traditional chi square test: the latter differ from those of Nathan's test by at most 0.002. It may be noted for interest that if one assumed that the statistic t of (3) is distributed as chi square multiplied by a factor of 1.4, the achieved significance level of this hypothetical variable at the nominal levels of .01, .05 and .10 would be .037, .118 and .193 respectively -- quite remarkably close to the values reported by Nathan. One might conjecture that Deff in his example was about 1.4.

Finally, a few comments will be made relating to the special case of stratified simple random sampling with proportional allocation.

Nathan in [10] claims* that in this case the traditional chi square statistic is asymptotically distributed under H_o as chi square with

* In a private communication with the author, Nathan identified the error in his proof.

$(r-1)\times(c-1)$ degrees of freedom. This is not generally true, as a simple counter-example proves. Indeed, suppose that there are exactly $r\times c$ strata, each of equal size. Suppose also that corresponding to each cell of the contingency table there is one and only one stratum which contributes to that cell and, conversely, every unit in a stratum is classified to exactly one cell of the contingency table. Now, proportional allocation is equivalent to selecting the same number of units in each stratum, say d, the total sample size being $n=rcd$. Then every cell of the contingency table will contain exactly d entries with probability equal to one. Thus the statistic t in (3) will be equal to zero with probability equal to one -- no matter how large n becomes. Parenthetically, one may observe that the design effect in this case is equal to zero.

In the relatively simple case of stratified simple random sampling with proportional allocation to strata, it is easy to prove under H_o for the test of goodness of fit, s

$$E(s) = m-1 - \sum_{i=1}^{m} \frac{1}{nP_i} \sum_h n_h \, (P_{ih}-P_i)^2$$

where P_{ih} is the proportion of units in stratum h belonging to category i and n_h is the sample size in stratum h. Since the expected value of s under simple random sampling is m-1, a <u>reduction</u> is observed in the expected value of s -- as indeed in such cases Deff is known to be less than or equal to 1, the extent of the reduction increasing, roughly speaking, with the between stratum differences in the proportions P_{ih}. In the case of the example of the previous paragraph $E(s) = 0$ which, given that $s\geq 0$, is only possible if $s=0$ with probability equal to one.

3. TWO ALTERNATIVE TESTS

The desirable feature of the two half-samples of a replicate is that they have the same distribution and are uncorrelated. This can either

be used to construct, in the case of complex samples, a quantity like Nathan's or Bhapkar and Koch's U(N) or U(B) in (6) or (7) whose expectation is zero under H_o, or it can be used to estimate the variance of linear or non-linear statistic -- but two half-samples cannot be used for both purposes. A simple way out would be available if more than two psu's were selected per stratum, but this is so rarely the case that it is hardly worth considering.

Given the difficulties of variance estimation when the half-samples are used for another purpose, it is a natural motivation to go back to the question: how far away from zero is actually the expected value under H_o of the quantity

$$\bar{U}_{ij} = \bar{P}_{ij} - \bar{P}_{i.} \bar{P}_{.j}$$

when it is based on the whole sample or, in fact, on one of the two half-samples of a replicate?

Several observations will be made on this question.

a) The expected value of $\bar{U}_{ij}$ is zero under simple random sampling or, slightly more generally, if the Deff of all the variances and covariances in (5) is equal to one. Actually, an even more general sufficient condition is that all the Deff's are equal. While this cannot be assumed to be the case generally, it is often the case that Deff's from the same survey for a wide variety of variables are within a quite narrow range of one another.

b) Under the most general conditions,

$$\sum_i \bar{U}_{ij} = 0 \qquad j=1, 2, \ldots, c$$

and

$$\sum_j \bar{U}_{ij} = 0 \qquad \text{for all } i=1, 2, \ldots, r$$

so that the expectations of $\bar{U}_{ij}$ are subject to r+c-1 linear constraints. Whereas this does not exclude the possibility that $E(\bar{U}_{ij})$ may be quite large in absolute value, it most certainly ensures that they are not all of the same sign.

c) Most important of all, it can easily be shown that so long as all Deff values are bounded for all variances and covariances considered (i.e. |maximum Deff| $\leq$ B for some B in whatever way $n\to\infty$),

$$E(\bar{U}_{ij}^2) = \mathrm{Var}(\bar{U}_{ij}) + 0\left(\frac{1}{n^2}\right) \tag{11}$$

where the left hand side is of $0\left(\frac{1}{n}\right)$.

Indeed,

$$E(\bar{U}_{ij}) = -\ \mathrm{Cov}\ (\bar{P}_{i.}, \bar{P}_{.j}). \tag{12}$$

The right hand side of (12) is obviously $0\left(\frac{1}{n}\right)$ in the case of simple random sampling, but in the case of complex designs it will only be multiplied by the appropriate Deff. Now there are no reported values in the survey literature of Deff exceeding 10, in fact values above 3 or 4 are exceedingly rare -- for the simple reason that the survey would never be allowed if it was that inefficiently designed. At any rate, so long as maximum|Deff | $\leq$ B as $n\to\infty$ over any class of designs subject only to the independent selection of two psu's per stratum and the availability of unbiased stratum-level estimates of linear statistics from each of the two psu's,

$$|E(\bar{U}_{ij})| \leq \frac{B}{n} P_{i.} P_{.j}\ . \tag{13}$$

Further,

$$E(\bar{U}_{ij}^2) = \mathrm{Var}\ \bar{U}_{ij} + [E(\bar{U}_{ij})]^2\ , \tag{14}$$

and (11) follows immediately from (13) and (14).

Since, however, under the same conditions

$$\text{Var } \bar{U}_{ij} = 0 \left(\frac{1}{n}\right)$$

it follows that for sufficiently large n one may well be able to treat $\bar{U}_{ij}$ as if it did have a zero expected value. Typically, in complex surveys n is quite large: at least of the order 1000-2000 and very often, in the case of large national surveys, tens of thousands.

That being the case, one can construct U_{ij} from each of the two half-samples of the k-th replicate. Consider the vector of U_{ij} values corresponding to the (r-1)x(c-1) upper left hand corner of a contingency table, constructed from one of the two half samples of the k-th replicate

$$\hat{U}^{(k)} = \hat{U}^{(k)}_{1,1}, \ \hat{U}^{(k)}_{1,2,\ldots,} \ \hat{U}^{(k)}_{1,c-1}, \ \hat{U}^{(k)}_{2,1,\ldots,} \ \hat{U}^{(k)}_{r-1,c-1}$$

and similarly $\tilde{U}^{k}$ constructed analogously from the second half sample of the same replicate

$$\hat{V} = \frac{1}{4K} \sum_{k=1}^{K} \hat{U}^{(k)} \ (\tilde{U}^{(k)})' \tag{15}$$

provides an approximately unbiased estimate of the variance of

$$\bar{U} = \sum_{k=1}^{K} (\hat{U}^{(k)} + \tilde{U}^{(k)})/2K \tag{16}$$

where K is the number of orthogonal replicates.

Therefore, under the null hypothesis, the statistic

$$\bar{U}' \ (\hat{V})^{-1} \ \bar{U} \tag{17}$$

is approximately distributed as Hotelling's T^2, or multiplied by an appropriate constant, as F. From standard textbooks this constant is easily seen as (L-m)/m(L-1) where L is the number of strata and m=(r-1)x(c-1). It follows that

$$t' = \frac{L-m}{m(L-1)} \bar{U}' \ (\hat{V})^{-1} \ \bar{U} \tag{18}$$

is approximately distributed as $F(m,L-m)$. It is easy to see that (18) also provides a test of goodness of fit: the vector U has to be replaced with the vector $\bar{P}-P^{o}$, where $\bar{P}$ is the vector of observed proportions, P^{o} the proportions under H_o, $\hat{V}$ is the covariance matrix of $\bar{P}$ estimated through BRR and $m=rc-1$.

The second test is more heuristically constructed than the first. Consider first the test of goodness of fit

$$\sum_i \frac{(n\bar{P}_i-n\,P_i^o)^2}{nP_i^o} = \sum_i \frac{n(\bar{P}_i-P_i^o)^2}{P_i^o} . \tag{19}$$

As discussed above, when H_o holds the <u>expected</u> value of the numerator of each term in (19) under the given design is the appropriate Deff times its expected value under simple random sampling. Assume that not only the expected value of the numerator but its <u>distribution</u> was also equal to that obtained by multiplying by Deff the corresponding statistic under simple random sampling. This would then suggest that dividing each term by the estimated Deff of the numerator would restore the distribution (under H_o) to chi square.

In effect, by dividing each term in the numerator by its corresponding Deff, the multiplier n becomes what is known as the <u>effective sample size</u>

$$n_i' = \frac{n}{\text{deff}_i}$$

so that the statistic

$$\sum_i \frac{n_i'\ (\bar{P}_i-P_i^o)^2}{P_i^o}$$

is distributed as chi square.

Since

$$\frac{1}{2}\left(\hat{P}_i^{(k)} + \tilde{P}_i^{(k)}\right) = \bar{P}_i \qquad \text{for all } k$$

and the variance of the expression above can be estimated as

$$a_i = \frac{1}{4K} \sum_{k=1}^{K} \left(\hat{P}_i^{(k)} - \tilde{P}_i^{(k)}\right)^2 \tag{20}$$

and since the variance of $\bar{P}_i$ under simple random sampling (under H_o) is estimated as

$$\frac{1}{n} \bar{P}_i \ (1-\bar{P}_i) \tag{21}$$

it follows that (20) divided by (21) provides an estimator of $Deff_i$. In fact,

$$b_i = a_i/\bar{P}_i \ (1-\bar{P}_i)$$

is an estimator of the inverse of the effective sample size, n_i'. Thus

$$\sum_i \frac{(\bar{P}_i - P_i^o)^2}{b_i \ P_i^o}$$

might be distributed approximately as chi square. In simulation studies the statistic above tended to be too large. However, by using the average of the b_i values, good results were obtained. So the second statistic proposed as a test of goodness of fit and evaluated through simulation studies is

$$t'' = \frac{1}{\bar{b}} \sum_i \frac{(\bar{P}_i - P_i^o)^2}{P_i^o} \tag{22}$$

where

$$\bar{b} = \frac{1}{r} \sum_{i=1}^{r} b_i. \tag{23}$$

Note that (22) can be computed whether or not disproportionate sampling among the strata has been used.

The test (22) can readily be generalized to obtain a test of independence. Let a_{ij}, b_{ij} and $\bar{b}$ be defined, respectively, as

$$a_{ij} = \frac{1}{4K} \sum_{k=1}^{K} (\hat{P}_{ij}^{(k)} - \tilde{P}_{ij}^{(k)})^2$$

$$b_{ij} = a_{ij}/\bar{P}_{ij} \ (1-\bar{P}_{ij})$$

$$\bar{b} = \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} b_{ij} \tag{24}$$

then

$$t'' = \frac{1}{\bar{b}} \sum_{i,j} \frac{(\bar{P}_{ij}-\bar{P}_{i.}\,\bar{P}_{.j})^2}{\bar{P}_{i.}\,\bar{P}_{.j}} \tag{25}$$

Note that (25) differs from Pearson's chi square test simply by the replacement of the actual sample size by the average effective sample size under complex designs. Note also that (25) can be computed whether or not proportionate allocation among the strata has been used. Also, since $\bar{P}_{ij}$ are linear statistics, their variances (hence their Deff's) can be estimated through traditional methods, i.e. without BRR. This makes the calculation of t'' quite easy: apart from a package to compute the traditional chi square statistic, only an efficient variance estimation program is needed.

4. EMPIRICAL RESULTS

The results of seven simulated examples are presented in this concluding section. In every instance the simulated sample design is stratified, two primary sampling units are selected with equal probabilities and replacement, and the sampling within the psu's is simple random, also

with replacement. Except for the with replacement sampling of psu's (which may well enough be approximated in practice if there are a large number of psu's, say 20, and no more than two of them are selected), the remaining simplifications in the simulations were imposed by the need to keep the programs simple -- as opposed to theoretical restrictions. All examples, except example 7, are based on 500 simulations, example 7 on 250. In all simulations, except example 6, the contingency tables are based on 2 rows and 3 columns. In examples 1 and 6 proportional allocation to the strata was used, in the others the sampling rates differed in the proportions 1:2:3. The total sample size in all examples was 1200.

The features of the examples are summarized in Table 2 below while the behaviour of the unweighted chi square test values is shown in Table 3.

Table 2

Summary of features of seven examples

	No. of rows	No. of columns	No. of strata	Relative sampling rates	Range of P_{ij} (H_o)	Range of Deff	No. of simulations
Example 1	2	3	6	1:1:1	1/6	1.98	500
Example 2	2	3	6	1:2:3	.161-.172	1.49-2.07	500
Example 3	2	3	6	1:2:3	.158-.173	1.57-2.55	500
Example 4	2	3	6	1:2:3	.158-.175	1.62-3.03	500
Example 5	2	3	6	1:2:3	.156-.178	1.66-3.51	500
Example 6	3	4	12	1:1:1	1/12	1.89	500
Example 7	2	3	30	1:2:3	.150-.180	2.07-3.09	250

Table 3

Observed significance level (under H_o) of unweighted sample counts in Pearson's chi square test

Nominal level	.01	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	Avg. test ÷ d. of f.	Avg. Deff
Example 1	.090	.198	.290	.420	.536	.604	.696	.770	.826	.886	.938	1.87	1.98
Example 2	.100	.212	.290	.426	.542	.596	.656	.728	.778	.868	.936	1.91	1.76
Example 3	.150	.294	.388	.502	.566	.636	.702	.772	.846	.914	.962	2.45	2.10
Example 4	.190	.330	.416	.528	.594	.666	.734	.810	.870	.918	.958	2.75	2.29
Example 5	.232	.362	.436	.556	.632	.688	.740	.820	.880	.922	.964	3.08	2.48
Example 6	.174	.350	.452	.606	.706	.782	.856	.896	.938	.960	.978	1.90	1.89
Example 7	.212	.324	.416	.532	.616	.684	.776	.860	.900	.940	.976	2.81	2.60

Looking at Table 3, one's immediate observation is the rising level of nominal significances from examples 1 to 5 with the rising level of average Deff. However, in examples 2 to 5 another consideration is also important: the unweighted counts do not have an expected value which is consistent with H_o: i.e. H_o is valid over all strata but not in each stratum, hence the unweighted counts are inconsistent with H_o in the case of disproportional allocation. The simulation model used was such that, going from examples 2 to 5, not only the range and average of Deff values but also the within stratum departures from H_o increased. This explains the reason why within these examples the average test value divided by the degrees of freedom (which for a valid chi square test, of course, ought to be 1), rises faster than the average of the Deff values.

It is interesting to note the entries in examples 1 and 6 (for which proportional allocation was used) at the nominal significance level .05: .198 and .350 respectively. Compare these with the corresponding

proportions "predicted" by Table 1: for Deff = 1.87 and 2 degrees of freedom it is .201; for Deff = 1.90 and 6 degrees of freedom it is .356. The agreement is, indeed, very close. Also very close for these two examples is the average value of the test statistic divided by the degrees of freedom: 1.87 and 1.98, and 1.90 and 1.89 respectively.

Finally, it should be emphasized that, as predicted, the classical chi square test provides totally misleading results in the presence of Deff's which are moderately large, even in the case of proportional allocation -- and more so otherwise.

Table 4 shows the observed significance level (under H_o) of the test t', i.e. the F test.

Table 4

Observed significance level (under H_o) of the test t' (F test)

Nominal level	.01	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	Avg. test ÷ Exp. value	Chi square (deciles)
Example 1	.008	.038	.068	.152	.240	.346	.452	.554	.664	.768	.870	0.81	12.60
Example 2	.012	.052	.080	.178	.276	.374	.482	.570	.670	.776	.888	0.96	4.72
Example 3	.016	.054	.102	.188	.280	.384	.474	.578	.688	.804	.906	1.04	3.96
Example 4	.014	.058	.108	.184	.270	.384	.478	.560	.676	.790	.894	0.99	8.38
Example 5	.018	.058	.104	.184	.278	.388	.484	.564	.658	.786	.906	1.03	11.12
Example 6	.018	.072	.124	.222	.336	.430	.500	.604	.708	.786	.872	0.83	16.04
Example 7	.004	.040	.112	.228	.304	.404	.500	.612	.696	.788	.900	1.00	4.00

The one before the last column contains entries obtained by dividing the average of our test statistic by its theoretical expected value (if d is

the degrees of freedom for the denominator of the F test, this value is $d \div d-2$). The last column is the chi square test statistic for goodness of fit applied to the decile values of the table above. The critical value of chi square with 9 degrees of freedom at the 5% level is 16.92, thus at least this test is consistent with the hypothesis that t' is distributed as F. The nominal significance levels, particularly at the .01 and .05 levels, which are usually of greatest interest, behave very well -- their average over the seven examples is .013 and .053 respectively.

Notwithstanding the non-significant values of the chi square goodness of fit test, particularly the first five examples show what appear to be consistent departures from the nominal levels in the range .3 - .8. It appears, however, that this is not due to the approximation whereby (13) is assumed to be zero (i.e. $E(\bar{U}_{ij}) = 0$). Indeed, this approximation holds exactly if all the Deff's in (5) are equal -- as noted before. This is the case in examples 1 and 6. Yet, in many requests, they appear to follow the F test worst (although still well enough at the .01 and .05 levels). Thus it would appear that the problem, if it can be called that, is due to the normal approximation, as opposed to the approximation (13). Note that n = 1200 -- small by standards of survey sampling.

Next the behaviour of the test t'' is shown, still under H_o.

Table 5

Observed significance level (under H_o) of the test t''
(adjusted chi square)

Nominal level	.01	.05	.1	.2	.3	.4	.5	.6	.7	.8	.9	Avg. test ÷ d. of f.	Chi square (deciles)
Example 1	.034	.068	.104	.216	.306	.390	.496	.600	.698	.782	.892	1.10	4.96
Example 2	.028	.092	.146	.248	.340	.408	.494	.612	.680	.790	.898	1.16	17.40
Example 3	.040	.102	.164	.250	.324	.416	.510	.600	.696	.790	.906	1.23	27.56
Example 4	.042	.100	.158	.244	.326	.416	.510	.600	.676	.796	.902	1.25	25.68
Example 5	.048	.112	.156	.250	.324	.420	.506	.600	.670	.772	.896	1.28	27.96
Example 6	.032	.072	.126	.238	.326	.412	.502	.610	.696	.816	.912	1.07	9.64
Example 7	.008	.068	.124	.216	.296	.400	.504	.596	.684	.780	.908	1.02	10.88

Clearly, t'' has a distribution which in four of the seven examples is significantly different from chi square at the 5% level of significance. However, in exploring the results in Table 5 somewhat further, the following might be observed.

a) Most notably, if the alternative is between using the unadjusted chi square test (Table 3) or using the simple adjustment which leads to t'', clearly t'' is very much closer to chi square. While the statistics corresponding in Table 3 to the last column of Table 5 were not computed, this much is clear to the naked eye.

b) The extent to which the distribution of t'' departs from chi square seems to follow closely the extent to which the expected value of t'' departs from the expected value of the corresponding chi square distribution (penultimate column of Table 5). However, t'' was constructed in such a manner that its expected value would asymptotically

be equal to that of chi square. The convergence, however, depends on the number of available replications from which to compute $\bar{b}$ of (25) -- not on n. In fact, we are faced with the usual bias of a ratio estimate whose magnitude heavily depends on the variance of the denominator, i.e. of $\bar{b}$, which in turn depends on the number of replicates or, since BRR need not be used, on the number of strata. The good behaviour of t'' in examples 6 and 7, where the number of strata is 12 and 30 respectively, is consistent with this line of reasoning. Thus one might expect t'' to behave acceptably well for the survey designs encountered in practice, where the number of strata is usually quite large.

c) While t' behaves consistently better than t'', one has to note that t' is not applicable if the degrees of freedom is greater than or equal to the number of strata. This may well occur in multi-level consistency tables.

Finally, Table 6 is a tentative attempt to compare the power of t' and t'' under an alternative hypothesis H_1. Since H_1 could not be consistently chosen across the examples, the comparison of the results in Table 6 should be restricted to comparing t' and t'' within the same example.

Table 6

Observed proportion of times t' and t'' exceed their respective significance levels, under H_1

Nominal level		.01	.05	.10
Example 1	t'	.026	.178	.340
	t''	.250	.418	.510
Example 2	t'	.060	.232	.406
	t''	.282	.498	.632
Example 3	t'	.062	.214	.394
	t''	.242	.438	.568
Example 4	t'	.066	.212	.380
	t''	.218	.408	.552
Example 5	t'	.062	.206	.382
	t''	.220	.378	.524
Example 6	t'	.160	.480	.686
	t''	.714	.866	.920
Example 7	t'	.148	.332	.476
	t''	.136	.316	.452

No authoritative conclusions can be drawn from the above, primarily because of the fact that t'' is biased upward under H_o -- at least for the first six examples. In fact, for this reason a comparison of the two test statistics at their respective nominal level is misleading -- it would make t'' appear to have considerably more power than it actually does. Analyzing Table 6 together with Tables 4 and 5 is more realistic. By and large t' at the .10 nominal level would appear to be more comparable with t'' at the .05 nominal level. Even so, t'' would appear to have somewhat more power than t', except for example 7. This is more or less what one could expect, since t', based on the F test, will asymptotically be distributed as chi square if the degrees

of freedom for the denominator is large, i.e. if the number of strata is large compared to the number of subclasses.

In summary, the unadjusted chi square is subject to intolerable biases under complex designs even with moderate Deff's; t' appears to be distributed as expected even for values of n as small as 1200; t'' behaves incomparably better than the unadjusted chi square test but still appears to have higher than nominal significance levels particularly when the number of strata is small; t'' is much easier to calculate than t'; t'' appears to have greater power than t' unless the number of strata is large compared to the number of subclasses.

RESUME

Une grande partie de la littérature sur l'échantillonnage se concentre sur l'effet que le plan d'échantillonnage utilisé pour rassembler des données dans une enquête porte sur les statistiques linéaires. Récemment, on a considéré davantage l'effet du plan d'échantillonnage sur les statistiques non-linéaires. Le facteur le plus important qui empêche ces recherches a été le problème de l'estimation d'au moins les deux premiers moments de ces statistiques. Le présent article étudie le problème de l'estimation des variances des statistiques non-linéaires des échantillons complexes, en considérant la littérature existante. On étudie les attributs de la statistique chi-carré calculée à partir d'un échantillon complexe pour tester des hypothèses de la qualité de l'ajustement ou d'indépendance. On développe des tests alternatifs et on étudie leurs attributs en faisant des expériences simulées.

REFERENCES

[1] BHAPKAR, V.P. and KOCH, G.G., "On the Hypothesis of 'no interaction' in Contingency Tables", Biometrics 24(3), pp. 567-594, September 1968.

[2] CHAPMAN, D.W., "An Approximate Test of Independence Based on Replications of Complex Sample Survey Design", Unpublished master's thesis, Cornell University, 1966.

[3] COHEN, J.E., "The Distribution of the Chi-squared Statistic Under Clustered Sampling from Contingency Tables", Journal of the American Statistical Association, Volume 71, pp. 665-669, September 1976.

[4] KISH, L. and FRANKEL, M.R., "Inference from Complex Samples", Journal of the Royal Statistical Society, Series B, Vol. 36, pp. 1-37, 1974.

[5] McCARTHY, P.J., "Replication - an Approach to the Analysis of Data from Complex Surveys", National Centre for Health Statistics, Vital and Health Statistics, Series 2, No. 14, Washington, D.C., 1966.

[6] NATHAN, G., "Tests of Independence in Contingency Tables from Stratified Samples", In N.L. Johnson and H. Smith, eds., New Developments in Survey Sampling, Wiley, New York, pp. 578-600, 1969.

[7] NATHAN, G., "A Simulation Comparison of Tests for Independence in Stratified Cluster Sampling", Bulletin of the International Statistical Institute, Vol. 44, Book 2, pp. 289-295, 1971.

[8] NATHAN, G., "On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples", Journal of the American Statistical Association, Vol. 67, pp. 917-920, 1972.

[9] NATHAN, G., "Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples", National Centre for Health Statistics, Vital and Health Statistics, Series 2, No. 53, Washington, D.C., 1973.

[10] NATHAN, G., "Tests of Independence in Contingency Tables from Stratified Proportional Samples", Sankhya, Vol. 37, Series C, Part 1, pp. 77-87, 1975.

A SURVEY DESIGN SYSTEM FOR THE MEASUREMENT OF TRUCK CARGO FLOWS IN PERU[1]

A. Satin and R. Ryan[2]

This paper describes a survey design established to measure truck commodity flows in Peru. The article addresses the conceptual and operational features of the survey design as well as describing its elements and implementation techniques in the context of a pilot project. Finally, the paper illustrates how the results of this pilot might be used to design and implement a full-scale national survey.

1. INTRODUCTION

This article documents elements of a proposed survey design system which will allow for the collection of data (product-origin-destination, i.e. P-O-D) on cargo movements by truck in Peru. The basic design was established during a three-week field trip to the Office of Sectorial Statistics within the Ministry of Transport and Communications in Lima, Peru. It was decided that the design would be tested in a pilot study during the summer of 1978. If successful, the design would serve as a potential base for the conduct of a full-scale, continuous national roadside survey. The techniques and procedures may also be adapted to similar P-O-D studies conducted at the regional level. Background material used in the development of the system is included in the papers [1], [2], [3], [4] and [5].

2. DATA SOURCES AND REQUIREMENTS

With respect to the development of a design to collect P-O-D information, three potential data sources were reviewed by the authors:

---

[1] This project was commissioned by the Canadian Transportation Commission as part of an overall technical assistance program for the Canadian International Development Agency in the transportation field.

[2] A. Satin, Household Surveys Development Division and R. Ryan, Special Surveys Co-ordination Division, Statistics Canada.

(a) the consignor - consignee (the shipper)
(b) the carrier company
(c) the driver.

Under alternative (a), one would be required to locate and obtain data from the shipper. There are many shippers without known addresses, especially farmers in the Sierra, many of whom sell their produce (even in the field before harvesting) to a middleman-agent-carrier; the middleman transports and markets the goods and the original producer may not know how much he had to ship or where it was going. For these reasons (a) is considered an unlikely source of reliable P-O-D information.

The carrier company (alternative (b)) is also viewed as an unlikely source. A high proportion of the carriers (more than 50% and perhaps as much as 85%) are one vehicle-one driver operations. The location of many of these small operators is unknown so they would be unavailable for mail or personal contact. In addition, it is felt that the records kept by these individuals are poor (i.e. virtually non-existent), so their movements and cargo over any but the immediately preceding period would be uncertain.

The remaining data source under alternative (c), namely the truck driver, is viewed for now as the best alternative. It is understood that they are an unco-operative group as a whole, viewing police and government as an undesirable intrusion into their private affairs. They are suspected of going to some lengths to avoid such questions as may be required to collect the data for a survey.

The advantage of the truck driver as the data source is that certain data is automatically available, namely the commodity, origin-destination and the vehicle description. This approach has the additional advantage that it has been used before, and its risks are known and, to some extent, the drivers are used to it.

One of the frequently mentioned "solutions" to the problems of data collection is the "Planica Unica" or waybill document (mandatory) used for all shipments with extra copies automatically provided for control and statistical purposes. It is apparently in successful use at this time in Colombia. Occasional planning for its use in Peru goes back at least to 1970. In this type of system, the shipper and carrier provide all data needed, prior to moving the goods, and copies of the documents are collected at various control points in the course of a trip.

The introduction of such a system of documents is likely to be an involved process, time consuming and costly. It is estimated that about four years were necessary before statistics started to flow from the Colombian model.

Such a system does have obvious advantages, but in light of the long lead time, and the lack of success in reaching agreement on it in Peru in the past, it would be more appropriate, for the present, to look for another solution.

Finally, the use of models might be viewed as having the potential to satisfy data requirements. The basic drawback here is the lack of reliable statistics (consumption of products, production of products) to render such an approach workable at this time.

In summary, it was concluded that, for the present, the only viable alternative for collecting product-origin-destination (P-O-D) statistics is by means of a roadside survey, the details of which are described in the following sections of this report.

With respect to data requirements, the most basic information required on a regular basis and likely to be available from a truck driver is:

(1) Commodity(ies): The code system suggested is the European Classification. Another system under consideration is the U.N. Commodity Classification.

(2) Origin(s) Destination(s): A place name, associated with the specific area in the country, to be coded ultimately to the provincial (regional) level.

Only major flows will be measured as the measurement of minor flows would require that stations be placed along all road segments, which is not practical.

(3) Quantity: In metric tons, cubic metres, or units (where other measures are unavailable) to be converted to weight as appropriate.

(4) Vehicle: Licence plate, number of axles, type of body.

The basic unit of the study is a commodity trip. Thus, for each loading and/or unloading of cargo, a separate commodity origin-destination data set would be prepared. The vehicle trip would be defined by the point of first loading of cargo to the point of ultimate emptying of the vehicle, for the commodities on the truck at the survey point. Cargo collected and discharged prior to the interview is not to be considered, nor is cargo to be picked up and delivered later in the vehicle trip.

3. THE PILOT STUDY

In designing the frame for the pilot, two major areas of consideration rested with the determination of (1) the geographical or spatial coverage of the pilot, and (2) the temporal coverage associated with the sampling methodology. It was determined that both the pilot and the extension for the future would involve the measurement of road freight traffic for major flows in the country. The measurement of all flows would not be financially or operationally feasible or practical.

In terms of spatial coverage, it was important that the locations selected for the pilot and the associated survey points (control stations) simulate as much as possible the conditions that might be experienced in the rest of the country. In this way, all methodologies associated with the pilot could be realistically tested and evaluated in terms of the possibilities of extending to a national survey. The criteria used to select the area for the pilot were determined to be:

(1) high volume of traffic,
(2) a diversity of product mix,
(3) testing data collection methods under extremes in climatic conditions, and
(4) an area where external data sources can be used for validation.

Peruvian officials listed three areas for consideration; namely, (1) the Ancash area in the north, (2) the Lima-La Orolla route in the central area, and (3) the area around Pisco in the south.

The three areas were discussed in light of the selection criteria and it was determined that the region of Ancash around Chimbote would-best meet the criteria and so was chosen for the pilot to be run during the course of a full trimester. A secondary objective of the pilot was to furnish data for the region, if possible. This would depend upon the extent to which the sample design, field procedures, etc. were to be modified and improved during the course of the pilot.

Five control points were established to measure the major flows in the region.

Prior to discussing the basic design, it should be noted that the following constraints had to be incorporated into the design and estimation procedures. The field staff (brigade) consisting of 4-5 men at each control station could work a maximum of 12 hours per day, 20 days per month with a minimum rest period of 6 hours between interview periods.

In order that traffic flows not be unduly disrupted during interview periods, a systematic sampling scheme could not be adopted (particularly in heavy traffic flows). The field procedure which is viable to implement can be briefly described as stopping the first truck, obtaining a complete interview, then stopping the next truck which passes the station when the interview is completed. During a selected interview period, a second member of the brigade (classifier) independently records the movements of <u>all</u> trucks passing the station according to truck type.

4. SAMPLE DESIGN (TEMPORAL)

4.1 Introduction

At each control station an independent sample is constructed. The design can be described as a stratified 3-stage replicated probability sample of trucks which pass a control station in a calendar year (trimester in the pilot).

4.2 Stratification

Stratification in a time frame is a process of classifying time units into certain collections called strata. An advantage of stratified sampling is the possible increase in efficiency per unit cost in estimating the population characteristics. In the context of a future national survey, stratification provides the flexibility of redesigning the sample of a specified stratum or groups of strata, without affecting the design in the remaining strata.

It is important that the variables selected for stratification meet certain requirements if the stratification is to result in the increase of design efficiency. The requirements for the P-O-D survey were:

(i) The stratification variables selected should be highly correlated with road cargo movements. Accounting for major sources of variation in cargo flows should result in the creation of relatively homogeneous strata which increases design efficiency.

(ii) For the purpose of the national survey, the variables selected should result in strata which maintain their homogeneity for the life of the survey.

(iii) The design is to be constructed so as to allow for the production of data at the annual level (national survey) as well as for each of the four trimesters. This constraint must also be accommodated in the stratification.

Sources of temporal variation in cargo flows are dependent specifically on the location of the interview stations (control points). For the purposes of the pilot, the following sources of temporal variation were determined on the basis of existing data, a field trip to several control stations, and subject matter knowledge of O.S.E. officials.

(a) time of day - morning, afternoon, evening, night

(b) day type - workday, non-workday

workday - (1) Mondays, Fridays

(2) Tuesdays, Wednesdays, Thursdays

non-workday - (3) Saturdays, Sundays, Holidays

(c) months

(d) trimesters.

The stratification variables selected were trimester (4) x months (3) x day type (3). Time of day is incorporated in the design through the sampling of time periods and ratio estimation.

In total, there are 9 strata for each trimester and hence 36 for the calendar year.

4.3 Sample Size

In view of financial constraints and the lack of reliable historical data, it was decided to select a maximum of 20 days per month for the 3 months of the trimester. The data resulting from the pilot will serve as input towards the determination of sample size requirements for the national survey. The sample size for the national survey will be based upon (1) data reliability requirements, (2) the sample design and estimation procedures used, (3) the systems implemented for geographical and commodity classifications and (4) financial constraints.

4.4 Sample Allocation

The following sample allocation strategies were discussed in detail:

(i) Neyman allocation,
(ii) X-proportional allocation (proportional to traffic volume), and
(iii) proportional to size (no. of time units in strata).

The Neyman allocation approach was eliminated due to the fact that (a) no historical reliable information base exists (i.e. no variance statistics), (b) the allocation could vary considerably with the commodity being measured and (c) the flow patterns could change over time which would adversely affect the design efficiency.

X-proportional allocation could be considered when reliable information on traffic flows becomes available. Traffic flow information was not considered sufficiently reliable to use this approach in the pilot. Further, although this approach may improve the reliability of many commodity estimates, the measurement of those commodity movements which are negatively correlated with total traffic flow would be adversely affected. In addition, it would be preferable to have truck traffic rather than total traffic (trucks and other vehicles) flows if such information were available. Again, if the flow pattern were unstable over the life of the survey, the design would require updating or possible redesign to accommodate such changes.

The strategy chosen for the pilot survey was proportional to size (i.e. proportional to the number of days in each stratum). This approach is conservative and relies least on reliable historical information. When the pilot is completed, a decision will be made as to whether or not to choose a compromise between this allocation and X-proportional for the national survey in view of the considerations outlined above.

4.5 Sampling Stages and Sample Selection Methods

Having determined the sample sizes (days) and allocation strategy, the first stage of sample selection is carried out. In each stratum an independent sample of days (20/month) is selected systematically. To facilitate variance estimation, in view of the complex survey design, the sample of days is systematically selected (without replacement) within each of two replicates for each stratum.

The second stage of selection involves the random selection of one six-hour period (morning, afternoon, evening or night) within each of the selected days. The selection of six hour periods in the pilot was such as to satisfy the field constraint requiring a minimum 6 hour rest period between interview periods.

Having selected the days and the time period within each of the days, the final stage of selection involves the selection of trucks within the selected time periods.

4.6 The Problem of Direction

To handle the problem of direction, it was decided to conduct interviews in each direction for 3 hours within the six hour time period. (In the development of a national survey, light flows may be accommodated with interviewing in both directions simultaneously without unduly disrupting traffic flows or statistical efficiency).

4.7 Truck Classification

During the six hour interview period, a second member of the field brigade records the truck traffic by type. This is required for the purpose of estimation (expansion of the sample to the population). To accommodate the interview methodology, this classification information should be recorded within sufficiently small time periods for statistical accuracy. For the pilot, classification is to be carried out for each 1-hour period instead of the entire 6-hour periods. With sufficiently small time periods, the sampling method approaches systematic sampling.

The time of interview is recorded by the interviewer on each questionnaire, for the purpose of estimation. It can also be noted that should the estimates for a major flow not be sufficiently reliable, two options are available to adjust the design, (1) adjust sample allocation, (2) increase the sample size (i.e. a second brigade at the same control station).

5. THE FORMS

5.1 Forms Design

The two major forms are: (1) Sample Control Document and (2) interview schedule.

(1) The Sample Control Document (SCD) contains all the information required to weight the sample to the population. A separate SCD is completed, corresponding to each 6-hour interview period. All the necessary control information is inserted before field operations begin. Such control information refers to location, date, time period, stratum, replicate, direction and sub-weight (see weighting). The SCD is then transferred to the brigade member responsible for truck classification. For each hour, he inscribes the total number of trucks by type which pass by the control station.

(2) The interview form is to be completed by the brigade member responsible for interviewing "Cammioneros". The control information outlined above is transferred onto the batched questionnaires before the field operations begin. The questionnaire presently allows for multi-commodity shipments. The total number of commodities is recorded and up to three are enumerated with "major" origin and "major" destination. The time of interview is recorded on the questionnaire for estimation purposes.

The following points should also be noted:

(i) Truck types and a commodity classification scheme have yet to be finalized. It was indicated that the greater the detail of the commodity classification, the lower will be, in general, the reliability of commodity estimates. As well, there will be an increase in the complexity of field operations, coding, quality control, and data processing procedures.

Since the commodities carried by different types of trucks vary by type and weight, defining major truck types for the purpose of estimation will improve the statistical efficiency of the design.

(ii) It is recognized that there is some information loss, particularly with agricultural products, for which it is probable that some trucks pick up the same commodity at different points along the route and drop off a commodity at several destinations along the route. To lessen the complexity of the field data collection, coding, quality controls and data processing, only one origin and one destination will be required.

(iii) A standardized system for recording weight as well as a standardized geographical code structure must be developed and clearly workable at the field level and by the data processing staff.

5.2 Batching, Colour Coding

It is suggested that the forms be colour coded corresponding to the interview stations to increase control of the survey documents.

The questionnaires are to be bound and enclosed with the SCD corresponding to each six hour period selected in the sample. Upon their completion, the forms are to be bound in a similar fashion and then forwarded.

5.3 Clerical Operations

The forms are to be checked for their completeness and accuracy upon their return from the field. Instructions need to be prepared for the staff to check the data and resolve discrepancies (see section 10).

After the data has been checked and corrected, the effective sample size (i.e. usable questionnaires) is determined on the basis of the time of interview which has been recorded on the questionnaires. This number is entered onto the SCD beside the corresponding truck type total which has been recorded by the truck classification officer in the field.

The clerical staff then checks the SCD for completeness, performs the necessary collapsing (when the effective sample take is zero) and, finally, adjusts the sub-weight according to eight pre-determined categories within each trimester defined by time period (4) and direction (2).

Having adjusted the sub-weight, the final weight (see weighting) is determined by the clerical staff and inserted onto the completed, checked questionnaires. The questionnaires are then ready for data entry.

6. TRAINING OF FIELD STAFF

It had been recognized that the success of the statistical system fundamentally rests with the reliability of the data collection. The importance of centralized or regional training programs - training sessions, interviewers' manuals, procedures manuals were discussed in the context of the data collection exercise. Provision has been made in the launching of the training program for the implementation of all these aspects.

7. ESTIMATION

7.1 Introduction

In a probability sample, the sample design itself determines the weights which may be used to produce unbiased estimates. Each record may be weighted by the inverse of the probability of selecting the unit to which the record refers.

The file to be created for tabulation purposes contains one record per commodity for each selected truck in the sample. Instead of physically duplicating the sample records, an overall weighting factor is entered in each record. For example, if the total weight of copper shipped by truck between a given origin and destination is required, this is done by sorting the records referring to those trucks in the sample which carry copper between the origin and destination and summing the product of the sample weights by the copper weights entered on these records.

Since objective information concerning the time periods and directions is available for the universe, the reliability of estimates can be improved by utilizing such auxiliary information. Ratio estimation is one of the most prevalent techniques of utilizing relevant information.

Upon comparing the estimates derived from the survey with those obtained externally, the estimates from outside sources are divided by the sample estimates for each classification and the weights of the records in each classification are adjusted by multiplying the weights by this factor. After the adjustment of the weights, the estimated aggregates will now agree with the estimate from the independent source for each classification. Ratio estimation is quite simple as compared to other methods of using external information and at the same time results in increased efficiency. The choice of external information is, however, very crucial to the procedure as it leads to higher efficiency only if such information is highly correlated with the characteristics of interest in the survey.

7.2 Weighting

In the cargo freight study, the final weight attached to each record is derived in five steps and is the product of five factors. These are referred to as the day weight, time period weight, direction weight, correction factor for non-response, and truck weight.

(i) Day Weight

This weight corresponds to the inverse sampling ratio of days selected for each replicate within each stratum. Since each replicate corresponds to a 1/2 sample, the weight is divided by 2.

(ii) Time Period Weight

Since one of four time periods corresponding to either 0:01-6:00; 6:01-12:00; 12:01-18:00; 18:01-24:00 is selected for each sampled day, this weight is four in all cases.

(iii) Direction Weight

Since classification and interviewing in the pilot is carried out in one direction only for each of three consecutive 1-hour periods, this weight is two in all cases. For classification and interviewing in 2 directions, the weight is 1.

The product of the above three factors will be referred to as the sub-weight. This sub-weight is entered into the SCD before the field operations commence.

(iv) Correction Factor for Non-Response

Since some SCDs will be lost (unusable, etc.) or since in some cases interviewing trucks was not possible, the information corresponding to the sample period will be lost unless the weights for other sample periods are adjusted to compensate for this. The sub-weights for completed SCDs are adjusted within eight classes defined by time period (4) and direction (2) based on the assumption that the volume and commodity characteristics of trucks that have been successfully interviewed represent the volume and commodity characteristics of trucks that should but were not interviewed within the above classes. However, if this assumption is not true, the estimates will be biased and the bias will be large with a high rate of non-response. The exact magnitude of bias introduced by the adjustment for non-response is impossible to calculate. Consequently, such an adjustment should be viewed as a last resort and every effort should be made to reduce it in the field.

The sub-weights multiplied by the correction factor provides the corrected sub-weight.

(v) Truck Weight

After the questionnaires have been checked and corrected, the effective sample take is recorded on the SCD according to truck type beside the total count of trucks. The ratio provides the sampling fraction for trucks and the inverse provides the truck weight.

The final weight corresponding to an interviewed truck corresponds to the product of the above five factors. This weight is calculated manually and is attached to each commodity of a sampled truck.

The estimate of total weight of a given commodity moved between a given origin and destination in a time period is obtained by first sorting all records referring to this commodity in the time period (trimester or year) having the corresponding origin and destination codes, and then aggregating the product of the final weight and the commodity weight.

A ratio adjustment was considered using as an auxiliary source, information supplied by traffic counters. The idea was dropped for three reasons, namely:

(1) traffic counters record number of axles not vehicles,
(2) procedures for incorporating the information were complex operationally and hence quite expensive, and
(3) a validation of the physical reliability of the counters would be necessary before any serious attempt to use the information is made.

7.3 Description of the Estimation

The estimation procedure outlined above can be stated algebraically and the following notation can be used for that purpose.

h - stratum ($h = 1,2, \dots 36$)
i - replicate ($i = 1,2$)
j - day ($j = 1,2,\dots 20$)
k - time period ($k = 1,2,3,4$)
ℓ - direction ($\ell = 1,2$)
m - hour ($m = 1,2, \dots 6$)
r - type of truck ($r = 1,2,3,4$)
o - interviewed truck
t - trimester ($t = 1,2,3,4$)

The final weight for an interviewed truck as represented by $W_{hijk\ell mro}$ can be expressed as the produce of five weighting factors as follows:

$$W_{hijk\ell mro} = W_{hij} \, W_{hijk} \, W_{hijk\ell} \, W_{k\ell t} \, W_{hijk\ell mr}$$

Each of the component weighting factors may be expressed as follows:

(i) $W_{hij} = \dfrac{N_h}{2n_{hi}}$

N_h : Number of days in stratum h

n_{hi}: Number of days selected from stratum h in replicate i

(ii) $W_{hijk} \equiv 4$ (one of four time periods is selected)

(iii) $W_{hijk} \equiv 2$ (selection of one direction as in the pilot; heavy flows in the national survey)

$\equiv 1$ (two directions for light flows)

(iv) $W_{k\ell t} = \dfrac{N_{k\ell t}}{\hat{N}_{k\ell t}}$

$N_{k\ell t}$: Number of time periods of type k corresponding to direction ℓ in the t^{th} trimester in the universe

$$\hat{N}_{k\ell t} = \sum_{h \varepsilon t} \sum_{i} \sum_{j} W_{hij} \, W_{hijk} \, W_{hijk\ell}$$

i.e. $\hat{N}_{k\ell t}$ is the sum of the sub-weights of usable SCDs referring to the k^{th} time period type corresponding to direction ℓ in the t^{th} trimester.

(v) $W_{hijk\ell mr} = \dfrac{N_{hijk\ell mr}}{n_{hijk\ell mr}}$

$N_{hijk\ell mr}$ = Total number of trucks of type r passing the control station during the m^{th} hour corresponding to direction ℓ in the k^{th} time period type of the j^{th} selected day belonging to replicate i in the h^{th} stratum as recorded by the truck classification officer.

$n_{hijk\ell mr}$ = Corresponding number of interviewed trucks.

The final weight is replicated for each commodity of a sampled truck.

7.4 Pooling of Hours and Truck Types in Case of Zero Observations

Let $T_{k\ell mr}$ refer to the total truck traffic of truck type r in the m^{th} hour in the ℓ^{th} direction within time period type k and $t_{k\ell mr}$ to the corresponding number of interviews.

If it happens that, for some r and m, $T_{k\ell mr} = 0$, there is no problem; the truck weight is defined to be 0 and simply does not appear in the estimate.

If, however, $T_{k\ell mr} \neq 0$ and $t_{k\ell mr} = 0$, the above procedure does not apply since a portion of the traffic would be omitted from all estimates. It is,therefore, necessary to develop a strategy for pooling over hours and/or truck types to account for zero observations.

The procedure to be followed in the P-O-D survey involves the collapsing of a given truck type over adjacent hours. If, for example, $t_{k\ell 11} = 0$ and $t_{k\ell 21} \neq 0$, the first two hours are collapsed for truck type one to form the following quantity necessary for the calculation of truck weights: $\frac{T_{k\ell m*1}}{t_{k\ell m*1}}$ where m* refers to a pooled two hour time segment.

If pooling over all the hours in one direction still results in zero observations, pooling is then carried out between adjacent truck types. Finally, should zero observations remain after such pooling, the number of interviews within a direction in a given time period is zero and the corresponding period is treated as non-response. This component of non-response is handled by an adjustment to the sub-weights discussed in the estimation section.

8. SPECIAL CONSIDERATIONS

8.1 Empty Trucks

The code value corresponding to 'empty' is inserted onto the questionnaire. The commodity carried is 'no commodity' and no distinction is made for the purpose of weighting. Empty trucks are then handled through domain estimation.

8.2 Duplication

Since more than one control point may be stationed along a given route between a defined origin and destination, the problem of duplication arises. The problem of duplication stems from the fact that a truck has a non-zero probability of being stopped at each station along the route. The problem cannot be overcome by sampling each station at different times or rejecting duplicate trucks in the sample. The following four methods are proposed to resolve this problem.

1) Select the control station in advance which will provide the P-O-D estimate. This option is easy to apply. The principal drawback rests with the fact that a station is chosen arbitrarily and that information from other control stations is lost.

2) Select the control station for which the P-O-D estimate is based on the largest sample size. The drawback here again is that information from other control stations is lost and that record counts must be determined for each estimate to select the station.

3) Create a simple average of the estimates from the control stations along an O-D route. The advantage here is that all information is used. The drawbacks are that each station's contribution to the estimate is the same (i.e. equal weight) regardless of the sample size and that programming is required to construct such estimates.

4) Create a weighted average of the estimates from the control stations along an O-D route on the basis of relative sample size. This method is the one which is the most statistically sound. The advantage again is that all information is used in a manner which minimizes the variance of P-O-D estimates. The drawback stems from the fact that record counts must be determined for all estimates and extra programming to construct such estimates is necessary.

8.3 Alternative Routes from an Origin to a Destination

There are some (not many) alternative routes from a major origin to a major destination in Peru. One such road network has been included in the pilot test. Estimates of P-O-D obtained from alternative routes must be aggregated. After aggregation, the situation reduces to that corresponding to one route.

9. VARIANCE ESTIMATION

On the basis of sample data, estimates of the sampling variability of P-O-D estimates can be calculated.

The methodology adopted for estimating variances in the pilot test is pseudo-replication. For the purpose of variance estimation, the two replicates selected from within each stratum are assumed to have been selected independently. In fact, the units for each of the two replicates were selected without replacement. In addition, the adjustment for non-response is carried out for both replicates together rather than independently for each replicate. This adjustment is carried out at the

trimestrial level rather than the stratum level. The replicates are, therefore, somewhat correlated but this has been assumed negligible in the pilot. A more precise formulation for the construction of variance estimates requires more complex programming. Such a system may be developed for the national survey as explained in a subsequent section of this report. The advantage of the present formulation is that it should be relatively easy to program and should provide useful approximations with respect to data reliability.

Letting $\hat{X}_t$ represent a given P-O-D estimate for trimester t, the variance estimate denoted by $\hat{V}(\hat{X}_t)$ may be determined as follows:

$$\hat{V}(\hat{X}_t) = \sum_{\substack{h=1 \\ h\varepsilon t}}^{9} (\hat{X}_{h1} - \hat{X}_{h2})^2$$

where X_{h1} = half sample estimate corresponding to replicate 1 in stratum h
X_{h2} = half sample estimate corresponding to replicate 2 in stratum h

and $\sum_{\substack{h=1 \\ h\varepsilon t}}^{9}$ refers to summation over all strata in trimester t.

The variance estimate of the corresponding annual P-O-D estimate $\hat{X}_A$ denoted by $\hat{V}(\hat{X}_A)$ may be determined as follows:

$$\hat{V}(\hat{X}_A) = \sum_{t=1}^{4} \hat{V}(\hat{X}_t)$$

where $\sum_{t=1}^{4}$ refers to the summation over the 4 trimesters.

10. EDITING AND IMPUTATION

10.1 Introduction

The purpose of editing and imputation is to identify and correct invalid entries or codes, to reconcile conflicting data and as far as possible to fill in missing values in information fields of records with partial

non-response. This procedure is distinct from weighting adjustments for complete non-response which do not take the form of changes within 'individual' records but may change an entire record's relative weight in the estimates.

Editing and imputation must be done according to a specified set of steps and decision rules which are based mainly on external knowledge and on logical rules.

All the editing and imputation to be carried out on the pilot responses are to be done at the level of the individual record, bypassing the necessity of a "hot-deck" frequency imputation approach.

The following sections will illustrate the general approach to edit and imputation since exact specifications will depend upon the final questionnaire content.

10.2 Coding and Transcription Errors

All editing will be carried out manually. Invalid codes may appear in single fields of identification. Correction must be made using other available identification information. Alternatively, several fields may individually contain allowable codes, but taken in combination, indicate a non-existent case. For example, it is not enough simply to ensure that 'Location' is one of the allowable codes; it must correspond to a selected day and time period combination. This sort of conflict is easy to specify by means of a list of valid code combinations.

10.3 Edit/Imputation Specifications

All edits must be specified before manual editing can be carried out. The first step involves the identification of all allowable responses. This is handled by means of a coded list of control stations, origins, destinations, commodities, weight classification, etc. For actual commodity weight, any non-negative integer, for example, may be acceptable

but usually a limit must be placed. Failures may essentially take one of three forms: a single field containing an invalid code which must be corrected; a single field containing an invalid blank; or two or more fields in conflict, that is, each having a legal code but forming an invalid combination. For variables which appear on the SCD and questionnaire, one would normally demand agreement on information appearing on these records. In case of disagreement, the problem is to decide which record contains the correct codes.

Further, there are some logical edits. Consider, for example, the situation where the total number of commodities recorded is less than the number of commodities listed.

Finally, one may impose "reasonableness" edits. Could a two-axle truck carry 10 commodities, 3 of which are listed and whose total weight is 1,000 tons? Subject matter knowledge will determine how many of this type of edit will be worth imposing. Since the editing is to be done manually, the complexity of the edit structure should be kept to a minimum.

Having identified those records which fail one or more edit rules, the problem is how to correct them. The first task is to separate questionnaires into three groups: (a) records which pass all edit rules, (b) records which require some correction and (c) unusable returns. The first two groups will provide the data for further processing. Rules need to be established concerning the manner in which edit failures are corrected for group (b).

Single field correction is to be resolved by referring to the SCD for the same information or deterministically on the basis of a defined set of imputation actions. Most difficult to resolve is the situation involving multiple field edits. In some cases, one may just change the minimum number of fields required to obtain a valid combination. In other cases, one field or a relationship between fields may be determined

to be of over-riding importance. In any event, all imputation rules must be developed which can be easily and consistently applied manually.

11. COMPUTER PROCESSING AND TABULATION

From the resulting clean and weighted set of questionnaires, a clean file can be created subject to errors involved in the data entry which can be minimized by means of key-edit, etc. Construction of any desired estimate is then relatively straightforward: the estimate for any O-D is found by just summing the product of the final weights and the commodity weights for the period under consideration. Tabulation is a matter of deciding what estimates to produce and arranging a suitable output format.

12. CONSIDERATIONS FOR THE SURVEY DESIGN OF A NATIONAL CONTINUOUS P-O-D SURVEY

Outlined below are recommendations concerning an analysis of the pilot survey data and an evaluation of the survey operations to serve as input towards the development of a national or regional P-O-D survey.

12.1 Concepts and Definitions

The adequacy of the concepts and definitions used in the pilot to obtain origin, destination, single and multi-commodity shipments should be assessed in terms of statistical and operational effectiveness.

In particular, the manner in which the problem associated with several origins/destinations corresponding to the same commodity is handled in the pilot, should be evaluated. Further, decisions must be made on the basis of the pilot as to the maximum number of commodities that can be reliably obtained during the course of a roadside interview.

12.2 Classification Systems

The commodity, truck, weights and measures, and geographical classification systems adopted in the pilot, should be evaluated in terms of their ability to satisfy data requirements, and their operational and statistical efficiency. Revised systems can then be implemented at the regional and national levels.

12.3 Field Operations

The methods and procedures used with respect to hiring and training field brigades in the pilot should be evaluated and possibly revised. In particular, the field procedures should be reviewed with respect to their ability to handle heavy traffic flows and surveying under adverse climatic conditions. It should be possible to determine how many interviews can be successfully handled by a brigade as a function of traffic volume. Such information can be used to monitor the performance of brigades in the national survey. Systematic programs for field observation should, as well, be an integral part of a national data collection exercise.

12.4 Sample Design

(a) Stratification

On the basis of an analysis of major commodity estimates, the efficiency of the stratification variables selected for the pilot can be assessed. Those variables which are found to be important can be incorporated into the revised sample design.

(b) Sample Size/Sample Allocation

Analysis of the variance estimates for selected commodity estimates should guide decisions with respect to adjustments of the sample size and/or allocation. For example, heavy flows might be handled by two brigades instead of one.

The allocation strategy which should be considered in the context of a national survey, is a compromise allocation falling between proportional to size of strata and proportional to estimated traffic volume of strata.

(c) Selection of Time Periods

If large differences in traffic flow are found between the four periods of the day, the periods within selected days may be selected with probability proportional to estimated traffic flows. Such a selection procedure should be considered only if traffic flow information is judged to be sufficiently accurate and reasonably stable over time.

The truck weight (fourth factor in the overall weight) is very large if a truck is selected during a heavy traffic flow time period. The selection of time periods by PPS would result in smaller time period weights (second factor in the overall weight) corresponding to such periods of heavy traffic. The overall weights for trucks would then be more even resulting in more stable estimates.

The selection of time periods was carried out in the pilot to satisfy field operating constraints. It is suggested that the selection of time periods be carried out on a random basis for each selected day and subsequently adjusted to satisfy field constraints.

12.5 Estimation

It is recommended that the adjustment of the sub-weights for non-response be carried out at the stratum level within each replicate rather than at the trimestrial level over both replicates. This procedure, of course, adds to the complexity of the operation and might be considered when experience has been gained and the development of automated procedures becomes feasible and practical.

The assumption underlying the adjustment for non-response is that information obtained for certain time periods represent other time periods for which information is not available. The assumption is more valid if the correction is carried out at the stratum level.

For the purpose of variance estimation outlined below, the replicates constructed for variance estimation purposes will be correlated unless the non-response adjustment is carried out independently for each replicate. The variance estimates would otherwise be subject to bias, the magnitude of which could be high if each of the replicates have very different response rates.

12.6 Variance Estimation

To account for the ratio weight adjustment by time period type and direction to improve the statistical accuracy of P-O-D estimates, more precise variance estimates than those used in the pilot can be constructed on the basis of the following formulation.

Letting $\hat{X}_t$ and $\hat{X}_A$ refer to the trimestrial and annual P-O-D estimates of commodity weight X and $\hat{V}(\hat{X}_t)$ and $\hat{V}(\hat{X}_A)$ to their variance estimates, the following is calculated for each stratum.

$$\Delta_h = \hat{X}_{h1} - \hat{X}_{h2} - \sum_{k=1}^{4} \sum_{l=1}^{2} \frac{\hat{X}_{tk\ell}}{T_{tk\ell}} (\hat{T}_{h1k\ell} - \hat{T}_{h2k\ell})$$

where $\hat{X}_{hi}$: refers to the half sample estimate of the total weight to commodity X derived from the i^{th} replicate in stratum h $(i=1,2)$

$\hat{X}_{tk\ell}$: refers to the total weight of commodity X in trimester t corresponding to the k^{th} time period type in the ℓ^{th} direction

$T_{tk\ell}$: refers to the total number of time periods of type k in the ℓ^{th} direction in trimester t

$\hat{T}_{hik\ell}$: refers to the half sample estimate of the number of periods of type k in the ℓ^{th} direction obtained from the i^{th} replicate in stratum h (i = 1,2).

The variance estimate corresponding to $\hat{X}_t$ is calculated as follows:

$$\hat{V}(\hat{X}_t) = \sum_{h=1}^{9} \Delta_h^2$$

The corresponding variance estimate for the annual estimate $\hat{X}_A$ is calculated as follows:

$$\hat{V}(\hat{X}_A) = \sum_{t=1}^{4} \hat{V}(\hat{X}_t).$$

12.7 Data Handling

The problems encountered in the manual data processing cycle: coding, clean up of forms, weighting, etc. should be evaluated in terms of cost, timeliness and accuracy. Certain procedures (e.g. part of the edit/imputation or weighting) may be more effectively and efficiently handled through the development and implementation of automated packages.

12.8 Time and Cost Analysis

The time and cost of all phases of the pilot survey should be recorded and later evaluated to estimate such parameters in the proposed national survey. This will serve as input into all phases of the development and implementation of the national survey.

ACKNOWLEDGEMENTS

The research presented in this paper was carried out under the general direction of Dr. K.W. Studnicki-Gizbert, Executive Director of the Research Branch of the CTC. This project is part of an overall technical assistance program in the transportation field which has been commissioned by the Canadian International Development Agency in Peru.

The authors wish to acknowledge the helpful suggestions of Messrs M. Nargundkar, H. Gough, and Miss J. Forgie of Statistics Canada and the support and direction of Miss M. Fleming (CTC) and Mr. R. Platek, Statistics Canada and Messrs A. Gemmell and D. Napier, the CTC representatives in Peru for the background research and their support to us while in Lima.

RESUME

Cet article décrit un plan d'enquête qui a été élaboré pour mesurer le flot de marchandises transportées par camion au Pérou. L'article considère les traits conceptuels et opérationnels du plan d'enquête, et en décrit les éléments et les techniques d'exécution dans le contexte d'un projet pilote. Enfin, on démontre comment on pourrait utiliser les résultats de ce projet pilote pour élaborer et exécuter une enquête nationale à grande échelle.

REFERENCES

[1] Gough, J.H., Ghangurde, P.D. (1976), "An Alternative Method of Surveying International Travellers At Frontier Points - Methodology Report", Statistics Canada.

[2] Platek, R. Singh, M.P. (1976), "Methodology of Canadian Labour Force Survey", Household Surveys Development Division, Statistics Canada.

[3] "For-Hire Trucking Survey (1976)", Transportation and Communications Division, Statistics Canada.

[4] Koop, J.C. (1960), "On Theoretical Questions Underlying the Technique of Replicated or Interpenetrating Samples", Institute of Statistics, North Carolina State College.

[5] Gemmell, A., Napier, D. (1977), "Background Document On Inter-Urban Cargo Statistics (Road)", Research Branch, Canadian Transportation Commission.

A COMPARISON OF CORRELATED RESPONSE VARIANCE ESTIMATES OBTAINED IN THE 1961, 1971 AND 1976 CENSUSES

K.P. Krótki and C.J. Hill[1]

The total variance of a survey estimate incorporates sampling variance, simple response variance and correlated response variance. The last component reflects the part of the total variance due to a common influence on a group of respondents. In the Canadian census, self-enumeration was adopted as the standard method of enumeration in the 1971 Census. One factor in favor of introducing this method was evidence, from the 1961 Census, that correlated response variance made an important contribution to the total variance of census estimates. Based on a study conducted using interpenetration of interviewers, this article compares correlated response variances from the 1961, 1971 and 1976 Censuses. The empirical results demonstrate that although the self-enumeration adopted in the 1971 Census did not completely remove the correlated response variance, this approach has considerably reduced the magnitude of this component of variance for almost all the characteristics examined.

1. INTRODUCTION

The total variance of a survey estimate incorporates both sampling and response variance. To see this formally one can decompose the total mean squared error into total variance and bias. In turn, the total variance can be decomposed into sampling and response variance. Finally, the response variance can be expressed as the sum of the simple response variance and the correlated response variance. The first component measures trial-to-trial variability of the response of a given respondent. It is the part of the response variance that is produced by tendencies of individual respondents to commit response errors independently of any other respondents. The correlated response variance, on the other hand, reflects the part of the total variance due to a common influence on a group of respondents.

[1] K.P. Krótki and C.J. Hill, Census Survey Methods Division, Statistics Canada.

The decomposition of the total variance is laid out in detail in the seminal work by Hansen, Hurwitz and Bershad [9]. This is often referred to as the Census model ([1], [2], [3]). The decomposition of the mean squared error can be expressed algebraically as:

$$M.S.E.(\bar{x}) = \frac{1}{n}\sigma_r^2 + \frac{n-1}{n}\rho_r\sigma_r^2 + \frac{N-n}{N-1}\frac{\sigma_s^2}{n} + \frac{2(n-1)}{n}\sigma_{rs} + B^2$$

where $\bar{x}$ is the mean variable X, n is the sample size, N is the population size, σ_s^2 is the sampling variance, σ_r^2 is the simple response variance, $\rho_r\sigma_r^2$ is the correlated response variance, σ_{rs} is the sampling-response covariance and B is the bias. Similar results can be obtained through a linear model approach (see [4]) in which the observed value is expressed as a linear combination of the true value, a term reflecting the constant bias of each enumerator and an independent random component. An overall review of total variance can be found in [12].

It has been shown empirically that in the case of canvasser enumeration the correlated response variance is the dominant component of total response variance ([5], p. 1035). In fact, in the case of a Census, for data collected from 100% of the population, there is no sampling variance and the correlated response variance becomes the dominant component of the total variance. This component is due to an effect on the respondents that causes respondents in the same interviewer assignment area to commit similar errors. It is postulated that this homogeneity of errors is due to the effect of the interviewers on the respondents.

In the Canadian census, self-enumeration was adopted as the standard method of enumeration in the 1971 Census. One factor in favor of introducing this method was evidence, from the 1961 Census, that correlated response variance made an important contribution to the total variance of census estimates [5]. In principle, self-enumeration would totally remove this component of variance. In practice, even while using this method, there remains some enumerator-respondent contact. This being the case, it is to be expected that self-enumeration will reduce rather

than completely remove all correlated response variance. This article gives evidence to show that this is indeed the case by making a comparison between the estimates obtained for the 1961 Census with those obtained in 1971 and 1976.

2. THE MODEL AND EXPERIMENTAL DESIGN

Whereas sampling variance may be calculated straightforwardly from the sample elements, the calculation of correlated response variance involves an experimental design to provide multiple observations on each response. Replication of the survey is one way to achieve this. However, problems of contamination (lack of independence between the first and second interview) make this approach subject to criticism. The design used in the 1961, 1971 and 1976 Censuses to measure total variance and correlated response variance is based on interpenetration of interviewers and respondents. The use of interpenetration in this fashion is due to Mahalanobis (1949). The theory and experimental designs for investigating response errors and the particular methods applied are discussed in detail elsewhere ([5], [8]). A brief summary is provided here.

An interpenetrated design model was set up for the 1961, 1971 and 1976 Censuses by selecting samples of pairs of adjacent enumeration areas (EAs) in which both enumerators shared the same supervisor and then partitioning the EAs at random into 2 approximately equal parts. Each enumerator of the pair is then responsible for the completion of the paired assignment. In 1961 the study was effected in 96 EAs in the Cornwall area, whereas in 1971 and 1976 a stratified random sample of 376 pairs of EAs from across Canada was used. Only a small number of canvasser EAs in the North, collective EAs and collectives (institutions such as hospitals, prisons, etc.) within EAs was excluded from the scope of this study.

An additional feature of the 1961 study was the use of re-enumeration. Re-enumeration allows the estimation of additional parameters including

the estimation of simple response variance, but introduces considerably increased costs. A drawback of re-enumeration, however, is that it is impossible (in a response error study) to ensure independence between repeated measurements due to the effect of recall. For this reason and the evidence that under canvasser enumeration the simple response variance contributes less to the total variance than does the correlated response variance ([5], p. 1035), the 1971 and 1976 studies were limited to interpenetration.

3. FORMULAE

A detailed derivation of the formulae for total and correlated response variance is given in [5] for the case in which both interpenetration and replication are applied. Assuming certain factors negligible, a somewhat shorter derivation applicable to the interpenetrated design of 1971 and 1976 is presented in [3]. The basic developments of this work are presented here.

The following notation will be used in the formulae.

P	is the number of EAs in Canada
k	is the subscript denoting the EA
h	is the subscript denoting the household within an EA
$k(i)$	denotes the ith half of EA k
N_k, n_k	denotes the total number of households in the population and in the sample respectively for the kth EA
n_{ki}	is the number of households in the sample for the ith half of the kth EA
X_{kh}	denotes the observed characteristic for household h in EA k
σ_k^2	is a measure of response variance
$\rho_k\sigma_k{}^2$	is a measure of correlated response variance
S_{xk}^2	is the sampling variance.

The estimate of the population total for a Census sample characteristic can be written as

$$\hat{X} = \sum_{k=1}^{P} \frac{N_k}{n_k} \sum_{h \varepsilon S_k} x_{kh}$$ where S_k is the set of sample households in EA k,

and its total variance is then given by

$$V(\hat{X}) = \sum_{k=1}^{P} [N_k^2 \frac{\sigma_k^2}{n_k} [1+(n_k-1)\rho_k] + (1- \frac{n_k)}{N_k} \frac{S_{xk}^2}{n_k}].$$

From the experimental design two estimators can be obtained. The first is the between enumerator variance, C_k, that is a measure of variance for EA k.

$$C_k = \frac{1}{2} [\bar{x}_{k(1)} - \bar{x}_{k(2)}]^2$$

$$\text{and} \quad E(C_k) = \frac{2\sigma_k^2}{n_k} [(1+(\frac{n_k}{2} - 1)\rho_k] + \frac{2S_{xk}^2}{n_k} .$$

The second is D_k, a within enumerator variance for EA k.

$$D_k = \frac{\sum_{i=1}^{2} \sum_{h \varepsilon S_{ki}} (x_{kh}-\bar{x}_{k(i)})^2}{n_{k1} + n_{k2} - 2}$$

$$\text{and} \quad E(D_k) = \frac{2}{n_k} [\sigma_k^2 (1-\rho_k) + S_{xk}^2].$$

Finally, it can be shown that, with a certain bias,

$E(C_k-D_k) = \rho_k\sigma_k^2$ a measure of correlated response variance for EA k.

In order to make the 1971 and 1976 results comparable to the 1961 results, a weighted average of the $\rho_k \sigma_k^2$ over all EAs in the project was calculated. Weights reflected the size of the EA in terms of number of people.

It should be pointed out that for several reasons the results presented in this paper do not correspond to the published total variance results for 1971 and 1976 publications. One reason is that the publications contain estimates of total variance which differ from the estimates of correlated response variance. Secondly, the publications present results inflated to the Canada level. The formulae used to calculate these total variance results at national and sub-national levels can be found in [3]. Finally, for purposes of publication, for any EA in which $C_k - D_k$ was negative this quantity was set to zero. Looking at the results, for several characteristics there is a sufficient number of negative $C_k - D_k$ values to make the overall average value of the estimate of the correlated response variance negative. A discussion of this problem can be found in [13] in which a statistical explanation is given for negative estimates of the correlated response variance.

4. LIMITATIONS IN COMPARING 1961, 1971 AND 1976 CORRELATED RESPONSE VARIANCE

The major limitations in comparing the 1961, 1971 and 1976 correlated response variance estimates are the differences in scope and design of the projects. As has been indicated above the 1961 study included both replication and interpenetration in its design but was only applied in the Cornwall area, whereas the 1971 and 1976 studies, which only included interpenetration, were applied to a Canada wide sample. These differences reflect different objectives for the studies. The 1961 study attempted a detailed investigation of the factors contributing to total variance. The 1971 study accepted as given the 1961 result that the correlated response variance dominates the simple response variance. Thus, rather than provide more detail on the components of total variance,

the 1971 study sought to give a reliable measure of variance at a Canada wide level to accompany the published Census estimates.

The difference in the domain of study, i.e. Cornwall as opposed to Canada may not be of importance for the majority of variables. There is no reason to believe that the enumerators in Cornwall were more or less inclined to influence responses than anywhere else in Canada. However, in certain crucial respects Cornwall is atypical of the whole of Canada. It is a boundary area between French speaking and English speaking regions containing a substantial proportion of both language groups. Clearly if 100% of the persons in an enumeration area have the same characteristics the likelihood of there being response errors for any particular characteristic is very low, whereas in a heterogeneous area errors are to be expected, if for no other reason than that there exists a proportion of persons for whom the 'true' response is ambiguous. In effect, therefore, for variance estimates of either English or French, the comparison is between an area with potentially high correlated response errors and all of Canada that includes both areas of this nature and other areas where response errors will be low.

There are two additional limitations to be considered in making the comparisons.

(1) The 1961 Census asked all questions of the entire population, whereas in 1971 some questions were asked only of a 1/3 sample. Comparisons are therefore sometimes between 100% questions and at other times between 100% questions and sample questions. It is not clear whether or not this difference is critical.

(2) The wording of questions has often changed from one Census to the next, indeed in some cases there is a slight change in the concept covered by the questions.

5. THE RESULTS

The results are given in two tables. The first provides 1961, 1971 and 1976 results for those characteristics that were included in [5]. The second table is an extension of Table 1 in that it gives 1961, 1971 and 1976 comparisons for characteristics not published in [5].

Two estimates of correlated response variance for 1961 are available. The first is calculated as $\delta_{21}\,\sigma_{r1}^{2}$, using the notation from [5]. In fact, the estimate of this quantity involves several other terms (see formulae 39 and 41 in [5]) which are assumed to be negligible. Furthermore, the estimate of this version of the correlated response variance is based on both the interpenetrated and replicated aspects of the 1961 survey. Thus this estimate is not directly comparable to the estimate used in 1971 and 1976. The estimate from [5] that is comparable to the estimate used in 1971 and 1976 is $\frac{1}{n}(C_1 - F_1)$ where n is the average enumerator assignment size, C_1 is the between enumerator variance for the first survey and F_1 is the within enumerator variance for the first survey ([5], p. 1033). This estimate is based only on the interpenetrated part of the study. It is this estimate that is used in comparing 1961, 1971 and 1976 results.

Comparing 1961 results with those from 1971 and 1976 in Table 1, it is evident that the estimate of correlated response variance is in general considerably reduced. All the characteristics with positive values in 1961 give lower estimates in 1971 and 1976 including two that are negative. The only estimate that is not lower in 1971 is for Age = 5 which in any case is scarcely different from zero. The most substantial difference occurs for Ethnic Group = French.

Table 2 comparing the unpublished estimates from 1961 with the 1971 and 1976 figures gives essentially the same results. Official language spoken = French Only gives a higher estimate, but all the other results are either lower or negligible. The main reservation in interpreting these estimates is the presence of negative results that give numerically large values.

The possible explanations for these results are that (1) errors are introduced which actually reduce total variance and (2) the variance of the correlated response variance is high particularly for those variables that were sample variables in 1971 and 1976. The variance of the correlated response variance may be further inflated by the fact that these variables are clustered and only apply to persons over 15.

In some cases (e.g. Mother Tongue = English), a decline of the correlated response variance can be observed across all three time points. Calculation of variances for more characteristics in which the results are not negative could shed more light on the prevalence of this situation. It is also interesting to note that for all three Censuses the values for Mother Tongue = English are larger than those for Mother Tongue = French. Little can be said about the implications of this result until more is known about the languages of both interviewers and respondents.

6. CONCLUSION

The comparison of the 1961, 1971 and 1976 correlated response variance estimates gives empirical evidence of a reduction in the variance between 1961 and 1971. This reduction can presumably be attributed to the change from canvasser to self-enumeration collection of data. The reduction of variance is seen to persist through to the 1976 Census. In fact, in some cases, the variance is even further reduced.

These findings are, however, qualified by a number of considerations concerning the variance and accuracy of the estimators and the problems of preparing results derived under different circumstances. Research concerning the variance and accuracy of the estimators is now being carried out on two fronts. First, a new method of calculating the response variance [6] is being investigated with data from the 1971 and 1976 Censuses.

Some preliminary results are already available ([9], [11]). Second, the old estimator is being studied in depth to shed further light on its theoretical and empirical properties.

RESUME

La variance totale d'un estimateur dans une enquête comprend la variance due à l'échantillonage, la variance due aux réponses simples et la variance due aux réponses correlées. Ce dernier composant reflète la partie de la variance totale causée par une influence commune sur un groupe de répondants. Dans le cas du recensement canadien, on a adopté l'auto-énumération comme méthode générale d'énumération pour le recensement de 1971. Un facteur en faveur de l'introduction de cette méthode était l'évidence, dans le recensement de 1961, que la variance due aux réponses correlées apportait une contribution importante à la variance totale des estimations du recensement. Cet article, basé sur une étude faite en utilisant l'interpénétration des interviewers, compare les variances dues aux réponses correlées des recensements de 1961, 1971 et 1976. Les resultats démontrent que, bien que la méthode d'auto-énumération adoptée pour le recensement de 1971 n'ait pas enlevé complètement la variance due aux réponses correlées, cette approche a considérablement réduit l'importante de cette composante de la variance pour presque toutes les caractéristiques examinées.

REFERENCES

[1] Bailar, B. and Dalenius, T. (1969), "Estimating the Response Variance Components of the U.S. Bureau of the Census' Survey Model", Sankhya B 31 (parts 3 & 4):341-60, December.

[2] Bailey, L., Moore, T.F. and Bailar, B. (1978), "An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample", Journal of the American Statistical Association, 73(361):16-23, March.

[3] Brackstone, G.J. and Hill, C.J. (1976), "The Estimation of Total Variance in the 1976 Census", Survey Methodology 2(2):195-208, December.

[4] Dodds, D.J. and Smith, T.M.F. (1973), "Estimation of Correlated Response Variance Under a Linear Additive Model", presented at the IASS Conference, August.

[5] Fellegi, I.P. (1964), "Response Variance and its Estimation", Journal of the American Statistical Association, 59(308):1016-41, December.

[6] Fellegi, I.P. (1974), "An Improved Method of Estimating the Correlated Response Variance", Journal of the American Statistical Association, 69(346):496-501, June.

[7] Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961), "Measurement Errors in Censuses and Surveys", Bulletin of the ISI, 38(2):359-74.

[8] Hill, C.J. (1976), "1971 Census Evaluation Programme, 1971 Response Variance Project, Methodology Report and Results", internal report, Census Survey Methods Division, Statistics Canada.

[9] MacLeod, A.D. (1978), "Investigation of Fellegi's 'Improved' Method of Estimating Correlated Response Variance", internal report, Census Survey Methods Division, Statistics Canada, April.

[10] Mahalanobis, P.C. (1946), "Recent Experiments in Statistical Sampling in the Indian Statistical Institute", Journal of the Royal Statistical Society, 109:325-70.

[11] Maurice, C., (1978), "Un nouvel estimateur de la variance corrélée de réponse", internal report, Census Survey Methods Division, Statistics Canada, August 1977, revised April.

[12] Nisselson, H. and Bailar, B.A. (1976), "Measurement, Analysis and Reporting of Non-sampling Errors in Surveys", Proceedings of the 9th International Biometric Conference, invited paper, Vol. 2:301-22, Boston, 22-27 August.

[13] United States Bureau of the Census (1968), "Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Effects of Interviewers and Crew Leaders", Series ER60, no. 7, U.S. Govt. Printing Office, Washington, D.C.

Table 1

A Comparison of 1961, 1971 and 1976 Correlated Response Variance Estimates for Characteristics Published in Fellegi (1964)

	1961 estimate x 10^4 comparable to 1971	1971 estimate x 10^4	1976 estimate x 10^4(4)
Sex: Male	13.0	0.2	- 3.4
Age: 5	- 0.9	0.4	- 0.6
Ethnic Group: French	1688.0	151.1	-
Highest Grade of School Attended: High School, Grade 5 or Univ. (2)	23.3	9.8	-
Persons Looking For Work Last Week	40.8	5.6	-
Persons Who Usually Work 40 Hours a Week (3)	83.1	48.9	-
Industry: Manufacturing	6.7	- 27.0	-
Industry: Trade	18.5	- 3.4	-

NOTES: (1) Age, Sex and Ethnic Group were 100% variables in all three Censuses. The other variables were 100% in 1961 but sample variables in 1971 and 1976

(2) 1961 Wording. In 1971 and 1976 the estimate is for Grade 12, 13 or University.

(3) 1961 Wording. In 1971 the estimate is for Persons Who Usually Work 40-44 Hours a Week.

(4) Entries marked with - are unavailable for 1976.

Table 2

A Comparison of 1961, 1971 and 1976 Correlated Response Variance Estimates for Characteristics not Published in Fellegi (1964)

Characteristic	1961 estimate x 10^4 comparable to 1971	1971 estimate x 10^4	1976 estimate x 10^4(1)
Official Language Spoken English only	129.2	112.5	-
Official Language Spoken French only	193.8	681.5	-
Mother Tongue English	228.2	12.0	9.8
Mother Tongue French	97.2	- 0.7	2.2
EDUCATION (Highest Grade)			
Elementary Only	93.6	- 196.0	-
High School (Grade 1 or 2)	77.0	- 52.9	-
High School (Grade 3 or 4)	6.1	- 69.3	-
AGE			
4	- 0.6	- 0.4	- 0.2
5	- 0.9	0.4	- 0.6
6	0.8	- 0.3	- 0.7
64	0.2	- 0.1	- 0.7
65	0.5	- 0.2	- 0.6
66	- 0.5	0.2	- 0.5
OTHER CHARACTERISTICS			
Relation to Head = Son	16.1	7.4	0.7
Marital Status = Married	1.6	0.8	- 1.4

(1) Entries marked with - are unavailable.

A STUDY OF REFUSAL RATES TO THE PHYSICAL MEASURES COMPONENT OF THE CANADA HEALTH SURVEY

B.N. Chinnappa and B. Wills[1]

This article presents the findings of an experimental design set up to study the variation in refusal rates to the different modules of the physical measures component of the Canada Health Survey. The study indicated that interviewer teams have a significant impact on refusal rates. Also, a large proportion of the refusals was due to total family refusals rather than individual refusals within responding families.

1. INTRODUCTION

The broad objectives of the Canada Health Survey are to provide reliable information on the health status of Canadians. The survey consists of two components:

a) an <u>interview component</u> which covers self-perceived and self-reported data collected by trained interviewers, and data given by respondents in self-administered questionnaires;

b) a <u>physical measures component</u> where the observations and tests are taken with the help of qualified nurses.

The three modules comprising the physical measures component are:

A - anthropometric measurements (height, weight, blood pressure and arm skinfold thickness)

B - blood tests

F - fitness test

There was concern over public reaction to collection of the physical measures data in a field survey. The greatest concern was that in requesting blood samples from respondents for module B, reaction would be such that response to the entire physical measures component (and eventually to the Canada Health Survey) would be jeopardized.

---

[1] B.N. Chinnappa and B. Wills, Institutional and Agriculture Survey Methods Division, Statistics Canada.

A pilot survey was mounted in Peterborough, Ontario, in the last week of July and the first two weeks of August 1976 to test the acceptability of each of the three modules of the physical measures component to the general public. The primary objective of the test was to examine the variation in response rates (or conversely in refusal rates) to the modules of the physical measures component and to study the causes of such variation. For this purpose, the test was set up as a factorial split-plot design controlling for the effects of two factors which it was felt could affect the refusals rates — interviewing teams and family types.

Although the test sample covered both rural and urban families, because of problems with the availability of suitable sampling frames, the main experimental design to test refusal rates to modules A, B and F was confined to urban families. Single member families and families containing only aged persons were also excluded owing to inadequate numbers of such cases in the frame, and to logistical difficulties.

A sample of urban families with only aged persons was used for a design involving modules A and B only, since aged persons (defined as those aged 65 and above) were not eligible for the fitness test. In rural areas an experimental design controlling for different factors was not possible and all that was attempted was a feasibility study of the field and laboratory procedures of the physical measures component. Details of the Peterborough test are given in [1] and [2].

This article describes the main experimental design set up for urban families (excluding single member families and those with only aged persons) and the analysis of the refusal rates observed in the experiment.

2. SAMPLE DESIGN

Peterborough city, as defined by the 1971 Census, consisting of 103 census enumeration areas (EA's) was the target population.

A two-stage sample of families was selected where EA's were the first stage sampling units. A random systematic sample of 8 EA's was selected from the 103 in Peterborough city after excluding uninhabited EA's, collective EA's (which were large institutions and hotels etc.) and those EA's covered by the current Labour Force Survey. The EA's were arranged in increasing order of average income per household (1971 Census data) before selection, in an attempt to represent families with different economic status in the sample.

The list of households in the sampled EA's was obtained from the 1975 city street directory for Peterborough and those households with telephones were contacted by telephone to build the sampling frame for the second stage sample. This telephone listing operation provided information on the composition of the households by families which was used to classify them into types for the experimental design. (Details on the success of the telephone enquiry in contacting households, types of families, etc. are given in Table A.1 of the Appendix.)

Cost considerations suggested a field staff of 12 interviewer teams, each consisting of an interviewer and a nurse. It was expected that a team would be able to complete 1 family on the average per working day, working in the evenings, so that a total sample size of about 200 families was feasible in the three week test. (As it turned out, many of the interviews were completed during the day itself which resulted in the assignments being completed earlier than expected.)

Families other than single member families and those with aged persons only were classified into two types for the experimental design:

TYPE 1: adults only
TYPE 2: mixed families (adults with aged and/or children).

It was felt that the presence of aged and/or children might affect the response rates to the different modules.

From each of family types 1 and 2, 80 families were selected at random for the experimental design. A random sample of 32 households from among those that could not be contacted by telephone during the listing operation was also included for the test and those families among them that belonged to types 1 or 2 were added to the appropriate sample.

The total sample consisted of all the individuals in these 192 families. On contact some of these families were not eligible for the enquiry since they were single member families or had only aged persons. The effective sample size was therefore reduced to 177 families and 540 persons in these families.

3. EXPERIMENTAL DESIGN

The main objective of the test was to study the effects of modules B and F and their interaction on refusal rates. Module A was a standard part of the Physical Measures component to be applied as a 'control' to all the sampled persons. A factorial design therefore appeared suitable with combinations of two levels (presence and absence) of each of the 'factors' B and F giving rise to four treatments to be tested—modules A, AF, AB and ABF. A split-plot design was set up to improve the precision of the comparisons between the treatments and to allow an adequate workload of 16 families to each of the interviewer teams. Family types were used as blocks, each block consisting of about 96 families (80 that were sampled within the type and those among the additional sample of 32 families that belonged to the type). Within each type, random subsamples of 8 families were used as plots (so as to best represent each block) and a plot was assigned at random to each of the 12 interviewing teams which were the 'whole plot treatments'. Random subsamples of 2 families within each plot were used as sub-plots and these were assigned at random to the four treatments.

As explained earlier, the family 'types' as determined by the listing operation needed correction. Consequently the total sample size reduced to 177 families and the sample sizes in the different split-plots were unequal.

Table 1 shows the layout of the design indicating the number of sample families allotted to each interviewer team x family type x module combination cell.

TABLE 1: Distribution of sample families by interviewer-team, family type and module

Interviewer Team	Family Type								Total
	1 adults only				2 mixed				
	Module Combination								
	ABF	AF	AB	A	ABF	AF	AB	A	
1	2	2	1	1	2	1	2	2	13
2	2	2	2	3	2	1	2	0	14
3	2	1	1	1	1	2	3	3	14
4	2	3	3	1	2	1	1	3	16
5	1	2	1	1	2	1	2	3	13
6	1	3	1	2	2	1	2	2	14
7	1	2	1	2	2	3	3	2	16
8	2	2	3	2	2	1	1	2	15
9	2	2	1	2	2	2	3	2	16
10	2	3	2	2	2	1	2	1	15
11	2	2	1	2	2	2	3	2	16
12	3	1	1	1	1	3	3	2	15
Sub-total	22	25	18	20	22	19	27	24	177
Total	85				92				177

Any family that was away (e.g. on vacation) and could not be interviewed during the test period was replaced by a family belonging to the same type from a reserve list to retain the sample size for the experimental design. 16.4% of the sample families needed such replacement. Replacement was not done for the cell with no family (module A, family type 2, interviewer team 2 in table 1) because that was not discovered until after the completion of the field work.

Table 2 shows the distribution of the number of eligible persons in the sample families in the different cells of the split-plot design. This number excluded infants (who could not stand) and members who were temporarily absent during the field test period.

TABLE 2: Distribution of no. of persons in sample families by interviewer-team, family type and module

Interviewer Team	Family Type								Total
	1 adults only				2 mixed				
	Module Combination								
	ABF	AF	AB	A	ABF	AF	AB	A	
1	5	5	2	2	6	5	7	3	35
2	4	6	4	6	9	6	7	0	42
3	5	5	2	2	4	8	8	12	46
4	5	9	7	3	8	2	4	11	49
5	3	5	4	2	8	3	7	15	47
6	3	9	2	4	6	4	8	6	42
7	2	6	2	6	7	11	9	8	51
8	6	5	8	6	8	3	2	5	43
9	5	7	3	4	6	8	11	6	50
10	4	6	5	5	8	7	11	4	50
11	5	4	2	7	5	6	6	8	43
12	4	3	2	4	3	11	10	5	42
Sub-total	51	70	43	51	78	74	90	83	540
Total	215				325				540

4. DATA COLLECTED IN THE EXPERIMENT

For each sample family a short interview component was administered as a preface to the physical measures component and to collect basic information on the family composition. That was followed by data collected for all the members in the family for the combination of modules assigned to it as per the experimental design. A family and a person were the two types of units considered for the study of refusals to the modules and the analysis of variance was done only for persons because of the small numbers of families in each cell.

5. RESULTS AND ANALYSIS

There were three types of refusals to the physical measures component, each of which was handled differently for the response analysis.

a) <u>Entire family refusals to both interview and physical measures components:</u>

In this case, the entire family refused and no data was obtained on the family during the survey. The only data available for such families was the number of persons in them obtained from the telephone listing operation if they were contacted in that operation. The number of refusals was then set equal to the number of persons in the family. For those family refusals that were not contacted by telephone (5 families), family type and size was imputed by randomly selecting a contacted survey family and imputing its type and size to the refusal family.

b) <u>Family refusals to physical measures component only:</u>

For those families that refused only the physical measures component, data on family size was obtained from the interview part of the survey and the number of refusals for that family was set equal to its family size.

c) Individual refusals to the physical measures component within participating families.

Table 3 shows the distribution of families and persons within these families that were classified as refusals of type (a), (b), or (c). Refusals of type (a) were most common, refusals of type (b) were infrequent, while individual refusals to the physical measures component within participating families were rare.

The majority of person refusals was due to the refusal at the doorstep to the interviewer on first contact.

TABLE 3: Distribution of families and persons within these families by type of refusal

	Families		Persons	
	Number	% of total	Number	% of total
Type (a) refusals of whole families to interview and physical measures components ..	20	11.3	58	10.7
Type (b) refusals of whole families to physical measures component only	9	5.1	26	4.8
Type (c) refusals of individuals within participating families to physical measures component	-	-	20	3.7
Total refusals to physical measures	29	16.4	104	19.3
Total number in sample	177	100.0	540	100.0

The person refusal rate, p, in each cell of the experimental design layout was calculated as $\frac{r}{n}$ where r = total no. of persons who had refused the physical measures among the families in that cell and n = no. of eligible persons in the families in the cell.

Because of the fact that p has a binomial distribution and n varies from cell to cell, the transformed variable $y = \text{Sin}^{-1}\sqrt{c + (1-2c)p}$ was used for the analysis of variance where $c = \frac{1}{4\bar{n}}$ and $\bar{n}$ is the harmonic mean of the n's. (Refer [3]).

The following table gives the analysis of variance for y for the experimental design adopted. The formulae used are described in Appendix B. The value of p = 0 was imputed for the cell with the missing observation since it belonged to Module A and 17 of the 24 module A cells had p = 0.

TABLE 4: Analysis of variance for refusals to the physical measures

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F
Between family types	1	0.1324	0.1324	3.0839
Between interviewers	11	1.4323	0.1302	3.0338*
Main plot error	11	0.4721	0.0429	-
Main effect - blood	1	0.0968	0.0968	1.0563
Main effect - fitness ...	1	0.0136	0.0136	0.1484
Interaction - blood x fitness ...	1	0.1606	0.1606	1.7525
Interaction - module x team	33	4.2851	0.1299	1.4174
Residual error	36	3.2980	0.0916	
Total	95	9.8908		

FOOTNOTE:

	*5% significance level	**1% significance level
$F_{1,11}$	4.84	9.65
$F_{11,11}$	2.82	4.46
$F_{1,36}$	4.11	7.39
$F_{33,36}$	1.76	2.23

The F test showed that the variation among the interviewer teams was the only component to achieve significance at the 5% level. None were significant at the 1% level.

Table 5 gives the refusal rates to physical measures, by persons for each of the module combinations, family types 1 and 2 and interviewer teams 1 to 12.

TABLE 5: Refusal rates to physical measures by person

Total refusal rate among all sample persons		- 19.3%
By module combinations:	ABF	- 20.9%
	AB	- 21.5%
	AF	- 22.6%
	A	- 11.9%
By household types:		
Type 1: multiple member adult hhlds.		- 23.7%
Type 2: mixed hhlds. (adults with aged and/or children)		- 16.3%
By interviewer teams: Team No.	1	- 31.4%
	2	- 42.9%
	3	- 10.9%
	4	- 18.4%
	5	- 19.1%
	6	- 9.5%
	7	- 19.6%
	8	- 34.9%
	9	- 10.0%
	10	- 8.0%
	11	- 23.3%
	12	- 9.5%

Although the analysis of variance showed that there was no significant variation among refusal rates to the modules, the above table shows that the overall refusal rate for A only was half of that for the other modules. The refusal rates between interviewer teams varied from as low as 8.0% to 42.9%.

6. CONCLUSIONS

The following conclusions are based on the findings of the Peterborough test presented above. It must be emphasized that this was a specially mounted first test on the general public at one point of time, confined to one small English speaking area. The test has a high publicity profile and was undertaken under survey conditions that did not and could not quite simulate those that would prevail in the final survey. It must be noted also that the experiment was set up to compare response rates between modules and not to obtain estimates of response or refusal rates. As such, the main findings presented above

and conclusion in (a) below are valid outcomes of this test. However estimates of refusal rates that occurred in the test and that are incidental to its main findings have been presented above and commented upon in (b) below since they are indicative of the order of magnitude of refusal rates that might be met in the final survey and of the causes for such refusals. Given the scope and purposes of the test, it is recognized that generalizations are difficult and simple extrapolations could be misleading.

a) The analysis of variance of refusal rates for urban families other than single member and aged families showed that interviewer teams rather than modules, family types or interactions between them, was the significant factor affecting refusal rates to the physical measures component.

   Initial fears that the blood test module would increase refusal rates appear to be unfounded on the basis of this test, although it should be noted that the refusal rate to the A module was half that to the other modules.

b) The overall refusal rate to the physical measures at the family level was high at 16.4% of all sample families. The refusal rate at the person level was even higher at 19.3%. A large proportion of these refusals (15.5% out of the 19.3%) was accounted for by total family refusals to the interview or physical measures component, and these were basically refusals on behalf of the whole family of the person first contacted by the interviewer. To add to this, 16.4% of the households could not be contacted during the survey and were replaced for the experimental design.

Also, it should be noted that this study was conducted essentially among households that were initially contacted by telephone so that the 8% refusals to the telephone and 21% non-contacts by telephone were excluded from the study. (See Table A.1 of the Appendix.)

Debriefing of 7 of the complete refusal families showed that these families had decided that they would not participate even before the interviewer had contacted them. The general comment was that they did not wish to participate in a government survey—it was a waste of time and money. 5 of the refusals to physical measures were debriefed. It appeared that usually the 'too old' were not interested and that the main reasons for other refusals were 'too busy' or 'other household members would not be interested'. (Refer [2])

In a survey situation, the total refusals at the household or family level would be decreased by attempts to call back on such households to persuade them to participate in the survey—attempts that were not made in this test. Also, it is difficult to judge the net result of the positive and negative effects that the high publicity profile (initial telephone contact, letter to the household and the various publicity programmes on the news media) had on response rates. On the other hand, in a survey situation, replacements for non-contacted families would not be allowed. This would inflate the total non-response rates although it would be mitigated to some extent by the lack of pressure to make early contacts and fix appointments for the physical measures that operated in this test.

The test suggests the need for strengthening door-step diplomacy tactics to persuade the first person contacted to co-operate in the survey and the need for steps to be taken while hiring and training interviewers to ensure a standard interviewer approach to respondents that would lower refusal rates.

The formulae for sums of squares in the ANOVA table are as follows:

Source of Variation	d.f.	Sum of Squares
Between hhld types	1	$\sum_i \frac{y_{i..}^2}{48} - CF$
Between interviewer teams (main plots)	11	$\sum_j \frac{y_{.j.}^2}{8} - CF$
Main plot error	11	$\sum_{ij} \frac{y_{ij.}^2}{4} - \sum_i \frac{y_{i..}^2}{48} - \sum_j \frac{y_{.j.}^2}{8} + CF$
Main effect B	1	$\sum_{B=0}^{1} \frac{y_{...(B.)}^2}{48} - CF$
Main effect F	1	$\sum_{F=0}^{1} \frac{y_{...(.F)}^2}{48} - CF$
Interaction B x F	1	$\sum_{B=0}^{1} \sum_{F=0}^{1} \frac{y_{...(BF)}^2}{24} - \sum_{B=0}^{1} \frac{y_{...(B.)}^2}{48} - \sum_{F=0}^{1} \frac{y_{...(.F)}}{48} + CF$
Interaction: interviewer team x module (main-plot x sub-plot)	33	$\sum_{jk} \frac{y_{.jk}^2}{2} - \sum_j \frac{y_{.j.}^2}{8} - \sum_k \frac{y_{..k}^2}{24} + CF$
Sub-plot error	36	obtained by subtraction
TOTAL	95	$\sum_{ijk} y_{ijk}^2 - CF$

(where $CF = \frac{y_{...}^2}{96}$)

ACKNOWLEDGEMENTS

The authors thank the referee for several helpful comments and suggestions.

RESUME

Cet article présente les conclusions tirées d'un plan expérimental qui a été élaboré pour étudier la variation des taux de refus aux différents modules de la composante des mesures de bien-être physique de l'Enquête Santé Canada. L'étude a indiqué que les équipes d'interviewers ont un effet significatif sur les taux de refus. De plus, une grande proportion des refus étaient des refus de toute la famille plutôt que des refus individuels dans des familles répondantes.

REFERENCES

[1] Canada Health Survey - Experimental Design and Analysis for the Peterborough Field Test of Physical Measures (Test 3) dated June 29, 1976.

[2] Canada Health Survey - Peterborough Test - Preliminary Report, August 26, 1976.

[3] Chanter, D.O., Modification of the Angular Transformation, Applied Statistics, 1975, 24, No. 3.

APPENDIX A

TABLE A.1: Results of the Telephone Listing Operation

Total no. of households listed in the 8 sample EA's ...	1,028	
Households with telephones	920	(100%)
Households which had moved or which had disconnected phones	115	(12%)
Non-contacted households (after at least 3 call-backs).	201	(21%)
Refusals	72	(8%)
Effective number of households contacted	532	(59%)

APPENDIX B

Formulae for Analysis of Variance of Refusal Rates

Let the proportion of refusals in each cell of the experimental design be $p = \frac{r}{n}$, where n is the no. of sampled persons in that cell and r is the number of refusals. The transformation used for analysis is

$$y = \mathrm{Sin}^{-1}\sqrt{c + (1-2c)\frac{r}{n}}$$

where c $\frac{1}{4\bar{n}}$ and $\bar{n}$ is the harmonic mean of the n's, where n varies. This is necessary to allow the analysis of variance and usual tests of significance. (Refer [3])

Let y_{ijk} be the transformed variable for the k^{th} module in the i^{th} household type and j^{th} team; and $y_{ijk}(BF)$ the corresponding value for different levels of B and F (B = 0 if absent, 1 if present; F = 0 if absent, 1 if present).

AN ESTIMATE OF THE EFFICIENCY OF RAKING RATIO ESTIMATORS UNDER SIMPLE RANDOM SAMPLING

M.D. Bankier[1]

Raking ratio estimators give estimates of the population values of characteristics examined on a sample basis utilizing the row and column totals of a contingency table of characteristics examined on a 100% basis. In this paper, the asymptotic variance of the maximum likelihood estimator of a sample characteristic subject to the marginal constraints of the above contingency table is derived. From this, we are able to compute the loss in efficiency of the raking ratio estimators relative to the maximum likelihood estimator in an empirical study.

1. INTRODUCTION

Raking ratio estimators given estimates of the population values of characteristics examined on a sample basis utilizing the row and column totals of a contingency table of characteristics examined on a 100% basis. Arora and Brackstone [1] described the use of RRE in the 1971 and 1976 Canadian Censuses of Population and Housing. They also derived formulae for the asymptotic variance of RRE under simple random sampling without replacement (s.r.s.w.o.r.) and presented an empirical study. Rao [3] found the asymptotic variance-covariance (V-C) matrix of the maximum likelihood estimators (MLE) for those variables examined on a 100% basis where the estimators were subject to the marginal constraints of the above contingency table. He restricted himself to the special case of estimators of frequency counts.

In this paper, Rao's results are generalized to variables examined on a sample basis. The results are derived under the assumption of simple random sampling with replacement (s.r.s.w.r.) since assuming sampling without replacement makes the problem much more complex. Because the

[1] M.D. Bankier, Census Survey Methods Division, Statistics Canada.

MLE are asymptotically most efficient, we can compute the loss in efficiency of the RRE relative to the MLE. This has been done in an empirical study presented in Section 4.

Variables examined on a 100% basis and variables examined on a sample basis will be known as 2A-categories and 2B-categories respectively.

2. DERIVATION OF THE ASYMPTOTIC VARIANCE OF THE MLE

Suppose there are t 2B-categories and let

$$N_{ijk} \qquad \begin{array}{l} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots, s \\ k = 1, 2, \ldots, t \end{array}$$

be the number of individuals in the population that belong to the kth 2B-category and fall in the (i,j)th cell of the 2A-category cross-classification table. Define

$$\pi_{ijk} = \frac{N_{ijk}}{N}$$

so that

$$\sum_j \sum_k \pi_{ijk} = \pi_{i..} \qquad i = 1, \ldots, r-1$$

$$\sum_i \sum_k \pi_{ijk} = \pi_{.j.} \qquad j = 1, \ldots, s-1 \tag{2.1}$$

$$\sum_i \sum_j \sum_k \pi_{ijk} = 1$$

where $\pi_{i..} = \frac{N_{i..}}{N}$ and $\pi_{.j.} = \frac{N_{.j.}}{N}$ are the known marginal proportions.

In this section the asymptotic variance of the MLE π^*_{ijk} of π_{ijk} is found subject to the constraints (2.1).

Let us assume that we have taken an s.r.s.w.r. Then the likelihood of the sample frequencies n_{ijk}'s is given by

$$L \propto \prod_{i,j,k} (\pi_{ijk})^{n_{ijk}} . \tag{2.2}$$

We maximize

$$\ln L = \text{constant} + \sum_i \sum_j \sum_k n_{ijk} \ln \pi_{ijk} \tag{2.3}$$

subject to the constraints (2.1) to find the MLE π^*_{ijk}. To do this we would have to solve a system of non-linear equations iteratively.

Silvey [4] has given a general method for finding the asymptotic V-C matrix of the MLE. Let

$$\underset{\sim}{B}(\underset{\sim}{\pi}) = \frac{1}{n} \left(E \left(- \frac{\partial^2 \ln L}{(\partial \pi_{ijk})(\partial \pi_{i'j'k'})}\right)\right) \tag{2.4}$$

denote the (rst) x (rst) information matrix. Let $\underset{\sim}{H}(\underset{\sim}{\pi})$ be the (rst) x (r + s - 1) matrix of derivatives of (2.1) with respect to the π_{ijk}'s. The derivatives are in the order that the π_{ijk}'s fall in $\underset{\sim}{\pi}$ where

$$\underset{\sim}{\pi}' = (\underset{\sim}{\pi}'_{1..} | \underset{\sim}{\pi}'_{2..} | \cdots | \underset{\sim}{\pi}'_{r-1,..} | \underset{\sim}{\pi}'_{.s.} | \underset{\sim}{\pi}'_{r..} | \underset{\sim}{\pi}'_{rs.}) \tag{2.5}$$

$$\pi'_{i..} = (\underset{\sim}{\pi}'_{i1.} | \underset{\sim}{\pi}'_{i2.} | \cdots | \underset{\sim}{\pi}'_{i,s-1,.}) \tag{2.6}$$

$$\underset{\sim}{\pi}'_{.j.} = (\underset{\sim}{\pi}'_{1j.}|\underset{\sim}{\pi}'_{2j.}|\cdots|\underset{\sim}{\pi}'_{r-1,j.}) \tag{2.7}$$

and

$$\underset{\sim}{\pi}'_{ij.} = (\pi_{ij1}, \pi_{ij2}, \ldots, \pi_{ijt}). \tag{2.8}$$

The asymptotic V-C matrix of the π^*_{ijk} is given by $\frac{1}{n}\underset{\sim}{D}(\underset{\sim}{\pi})$ where

$$\left[\begin{array}{c|c} \underset{\sim}{B}(\underset{\sim}{\pi}) & \underset{\sim}{H}(\underset{\sim}{\pi}) \\ \hline \underset{\sim}{H}'(\underset{\sim}{\pi}) & \underset{\sim}{0}_{(r+s-1),(r+s-1)} \end{array}\right]^{-1} = \left[\begin{array}{c|c} \underset{\sim}{D}(\underset{\sim}{\pi}) & \underset{\sim}{Q}(\underset{\sim}{\pi}) \\ \hline \underset{\sim}{Q}'(\underset{\sim}{\pi}) & \underset{\sim}{R}(\underset{\sim}{\pi}) \end{array}\right] \tag{2.9}$$

and $\underset{\sim}{0}_{(r+s-1),(r+s-1)}$ is a matrix of zeroes of order $(r+s-1) \times (r+s-1)$.

Using the formula for the inverse of a partitioned matrix, we find

$$\begin{aligned} \underset{\sim}{D}(\underset{\sim}{\pi}) &= \underset{\sim}{B}^{-1}(\underset{\sim}{\pi}) - \underset{\sim}{B}^{-1}(\underset{\sim}{\pi})\ \underset{\sim}{H}\ (\underset{\sim}{\pi})\ [\underset{\sim}{H}'(\underset{\sim}{\pi})\ \underset{\sim}{B}^{-1}(\underset{\sim}{\pi})\ \underset{\sim}{H}(\underset{\sim}{\pi})\ ]^{-1}\ \underset{\sim}{H}'(\underset{\sim}{\pi})\underset{\sim}{B}^{-1}(\underset{\sim}{\pi}) \\ &= \underset{\sim}{B}^{-1}(\underset{\sim}{\pi}) - \underset{\sim}{A}\ \underset{\sim}{F}^{-1}\ \underset{\sim}{A}' \end{aligned} \tag{2.10}$$

where

$$\underset{\sim}{A} = \underset{\sim}{B}^{-1}(\underset{\sim}{\pi})\underset{\sim}{H}(\underset{\sim}{\pi}) \text{ and } \underset{\sim}{F} = \underset{\sim}{H}'(\underset{\sim}{\pi})\ \underset{\sim}{B}^{-1}(\underset{\sim}{\pi})\ \underset{\sim}{H}(\underset{\sim}{\pi}).$$

Define

$$\underset{\sim}{c}' = \underbrace{[(\underset{\sim}{e}_t^k)'|(\underset{\sim}{e}_t^k)'|\ldots|(\underset{\sim}{e}_t^k)']}_{\text{rs submatrices}} \tag{2.11}$$

where

$$(\underset{\sim}{e}_t^k)' = \overbrace{[0, \ldots, 0, 1, 0, \ldots, 0]}^{\text{t columns}} \tag{2.12}$$

and the one is in the kth column.

It can be shown that the asymptotic variance of $\pi^*_{..k}$ is

$$V(\pi^*_{..k}) = \sum_i \sum_j \sum_{i'} \sum_{j'} C(\pi^*_{ijk}, \pi^*_{i'j'k})$$

$$= \frac{\underset{\sim}{c}'\,(\underset{\sim}{B}^{-1}(\underset{\sim}{\pi}) - \underset{\sim}{A}\underset{\sim}{F}^{-1}\,\underset{\sim}{A}')\,\underset{\sim}{c}}{n}$$

$$= \frac{1}{n}\,(\pi_{..k} - \underset{\sim}{a}'\,\underset{\sim}{F}^{-1}\,\underset{\sim}{a}) \tag{2.13}$$

where $\underset{\sim}{a}' = (\pi_{1.k}, \pi_{2.k}, \ldots, \pi_{r-1,.k}, \pi_{.1k}, \pi_{.2k}, \ldots, \pi_{.,s-1,k}, \pi_{..k})$ (2.14)

$$\underset{\sim}{F} = \left[\begin{array}{c|c} \underset{\sim}{F}_{11} & \underset{\sim}{f}_{12} \\ \hline \underset{\sim}{f}'_{12} & 1 \end{array}\right] \tag{2.15}$$

$$\underset{\sim}{f}'_{12} = (\pi_{1..},\pi_{2..},\ldots,\pi_{r-1,..},\pi_{.1.},\pi_{.2.},\ldots,\pi_{.,s-1,.}) \tag{2.16}$$

$$\underset{\sim}{F}_{11} = \left[\begin{array}{c|c} \underset{\sim}{E}_1 & \underset{\sim}{E}_2 \\ \hline \underset{\sim}{E}'_2 & \underset{\sim}{E}_3 \end{array}\right] \tag{2.17}$$

$$\underset{\sim}{E}_1 = \text{diag}\ (\pi_{1..},\pi_{2..},\ldots,\pi_{r-1,..}) \tag{2.18}$$

$$\underset{\sim}{E}_3 = \text{diag}\ (\pi_{.1.},\pi_{.2.},\ldots,\pi_{.,s-1,.}) \tag{2.19}$$

and

$$\underset{\sim}{E}_2 = \begin{bmatrix} \pi_{11.} & \pi_{12.} & \cdots & \pi_{1,s-1,.} \\ \pi_{21.} & \pi_{22.} & \cdots & \pi_{2,s-1,.} \\ \vdots & & & \vdots \\ \pi_{r-1,1.} & \pi_{r-1,2.} & \cdots & \pi_{r-1,s-1.} \end{bmatrix} . \tag{2.20}$$

Rao [3] demonstrated that the inversion of $\underset{\sim}{F}$ can be reduced to the inversion of $\underset{\sim}{E}_3 - \underset{\sim}{E}'_2\ \underset{\sim}{E}_1^{-1}\ \underset{\sim}{E}_2$ which is a (s-1) x (s-1) matrix.

3. AN EMPIRICAL STUDY

This empirical study uses the same data and categories as the paper by Arora and Brackstone [1]. The data came from one Electoral District (ED) that contained 15 Weighting Areas (WAs) and the data were gathered in the 1974 Canadian Test Census.

In the previous section, the asymptotic variance of the MLE $\pi^*_{..k}$ of $\pi_{..k}$ was derived under s.r.s.w.r. To allow efficiency comparisons with the RRE, the large-sample variances of the RRE without the f.p.c. are used. The estimate SE^*_0 of the asymptotic standard error of $N\pi^*_{..k}$ is calculated using zero iteration raking ratio estimates $\hat{\pi}^{(0)}_{ijk} = \frac{n_{ijk}}{n}$ for the π_{ijk}'s in (2.13). Because the zero iteration estimator of π_{ijk} is unbiased, it is felt that it is better to use it rather than the fourth iteration raking ratio estimate $\hat{\pi}^{(4)}_{ijk}$. SE^*_0 expressed as a percentage of SE_p (the estimated pth iteration RRE standard error under s.r.s.w.r.) will be denoted by RE^0_p. The table at the end of this paper gives at the ED level SE_p, RE^0_p (p=0, 1, 2, 3, 4) and SE^*_0. The categories examined are:

Class A(2B-Categories):	A1 Households with Employed Heads
	A2 Households with Unemployed Heads
	A3 Households with Heads Not in Labour Force
	A4 Household with Heads Not Moved in 5 Years
	A5 Households with Heads Moved in Last 5 Years in Same Municipality
	A6 Highest Grade of Head is 1 to 10
	A7 Heads with Bachelor Degree or Higher
Class B(2A-Categories)	B1 Households with 3 or 4 Persons
	B2 Age of Head is Less Than 25
	B3 Age Of Head is 25 to 34
	B4 Head Who is Widowed, Divorced, or Separated

Class C(2A-Categories): C1 Households with 2 or Fewer Persons
C2 Age of Head is 65 or More
C3 Owned Dwellings
C4 Rented Apartments

4. ANALYSIS OF THE RESULTS

For all 2B-categories, it can be seen that $RE_2^0 \geq 98.6\%$ and $RE_4^0 \geq 99.8\%$. These results indicate that the fourth iteration RRE is almost as efficient as the MLE and that the RRE do not gain much in efficiency after the second iteration.

The results are similar for the 2A-categories with the exception of C3 and C4. For C3, the RRE at the third iteration and the MLE both have zero standard errors while for C4, $RE_3^0 = 99.9\%$. Thus for the 2A-categories there is near equality between the standard errors of the RRE at either the third or fourth iteration and the MLE.

5. CONCLUSIONS

The empirical study, which assumes s.r.s.w.r. and that n is large, indicates that the RRE by the second iteration is almost as efficient as the MLE for most categories and that generally only small gains in efficiency are made at the third and fourth iteration.

6. ACKNOWLEDGEMENTS

This problem was suggested to me by Professor J.N.K. Rao for my Master's Thesis [2]. I would like to thank him for his time spent and for his advice. I would also like to express my appreciation to H.R. Arora and G.J. Brackstone who had many discussions with me about their work.

ESTIMATED EFFICIENCY OF THE RRE VERSUS THE MLE

UNDER S.R.S.W.R. AT THE ED LEVEL

Category	Iteration 0		Iteration 1		Iteration 2		Iteration 3		Iteration 4		MLE
	SE_0	RE_0^0	SE_1	RE_1^0	SE_2	RE_2^0	SE_3	RE_3^0	SE_4	RE_4^0	SE_0^*
A1	102.9	77.4	101.7	78.3	79.8	99.8	79.9	99.8	79.7	99.9	79.7
A2	39.7	98.6	39.5	99.0	39.2	99.9	39.2	99.9	39.2	99.9	39.1
A3	99.1	75.1	98.2	75.8	74.6	99.8	74.7	99.7	74.5	99.9	74.5
A4	116.0	83.3	104.4	92.6	98.0	98.6	96.8	99.8	96.8	99.8	96.6
A5	103.6	94.2	100.8	96.9	98.0	99.6	97.8	99.9	97.8	99.9	97.7
A6	115.1	94.9	113.9	95.9	109.4	99.8	109.4	99.8	109.3	99.9	109.2
A7	59.6	97.3	59.2	97.9	58.1	99.8	58.1	99.9	58.0	99.9	58.0
B1	114.4	84.1	113.2	85.1	96.4	99.9	96.6	99.6	96.4	99.9	96.3
B2	61.1	83.9	57.7	88.8	51.5	99.5	51.5	99.6	51.3	99.9	51.3
B3	98.0	51.1	96.0	52.1	50.2	99.7	51.1	98.0	50.1	99.9	50.1
B4	87.8	67.4	86.3	68.6	59.2	99.9	59.6	99.3	59.2	99.9	59.2
C1	112.5	40.4	106.7	42.7	45.6	99.8	47.7	95.4	45.5	99.9	45.5
C2	82.8	25.0	81.7	25.4	20.8	99.7	21.4	96.8	20.7	99.9	20.7
C3	109.4	0.0	0.0	-	41.9	0.0	0.0	-	11.2	0.0	0.0
C4	80.1	17.6	14.7	96.4	34.7	40.7	14.2	99.9	17.0	83.4	14.1

RESUME

Les estimateurs d'échantillon en formation donnent des estimations de la valeur, dans la population, des caractéristiques qui ont été étudiées à partir d'un échantillon, en utilisant les totaux des rangées et des colonnes d'un tableau de contingence des caractéristiques qui ont été étudiées pour toutes les unités de la population. Dans cet article, on donne la variance asymptotique de l'estimateur du maximum de vraisemblance d'une caractéristique échantillonné, soumise aux contraintes marginales dudit tableau de contingence. A partir de cette variance on peut calculer, dans une étude empirique, la diminution de l'efficacité des estimateurs d'échantillon en formation relatifs à l'estimateur du maximum de vraisemblance.

REFERENCES

[1] Arora, H.R. and Brackstone, G.J. (1977), "An Investigation of the Properties of Raking Ratio Estimators With Simple Random Sampling", Survey Methodology, Vol. 3, No. 1, 62-83.

[2] Bankier, M.D. (1977), "Properties of Raking Ratio Estimators", Master's Thesis, Carleton University.

[3] Rao, J.N.K. (1974), "Raking Ratio Estimator - III", Technical Report, Statistics Canada.

[4] Silvey, S.D. (1959(, "The Lagrangian Multiplier Test", Ann. Math. Statist., 30, 389-407.

SURV

June/juin 1977 No. 1

A Journal produced by s Canada.

Publié par le secteur ue Canada.

Adjustment for Non-Res
R. PLAT 1

Some Estimators of Popu
Simple Random Samples
M.A. HIDIROGLOU and K.P. SRINATH 25

Redesign of the Sour Cherry, Peach and Grape
Objective Yield Surveys in the Niagara Peninsula
G. DAVIDSON 38

An Investigation of the Properties of Raking Ratio
Estimators: I With Simple Random Sampling
H.R. ARORA and G.J. BRACKSTONE 62

The Response Incentives Experiment in the Canadian
Labour Force Survey
A.R. GOWER 84

Estimation of Variance in Multi-Stage Sampling
K.P. SRINATH and M.A. HIDIROGLOU 104

Yukon Labour Force Survey - Methodology of a Feasibility Study
D. ROYCE 114

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/décembre 1977 Vol. 3 No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

CONTENTS

Confidentiality of Statistical Information
D.A. WORTON 127

Synthetic Estimation In Periodic Household Surveys
P.D. GHANGURDE and M.P. SINGH 152

L'enquête sur la profession des salaries (E.P.S.)
M.A. HIDIROGLOU 182

Some Factors Affecting Non-Response
R. PLATEK 191

Survey of Canadian Residents Returning By Land
J.H. GOUGH and P.D. GHANGURDE 215

An Investigation of the Properties of Raking
Ratio Estimators: II With Cluster Sampling
H.R. ARORA and G.J. BRACKSTONE 232