

## TECHNIQUES D'ENQUÊTE

June - 1979-Juin

## VOLUME 5

NUMBER 1 - NUMÉRO 1

A Journal produced by Methodology Staff Statistics Canada

Préparé par les méthodologistes de Statistique Canada

## SURVEY METHODOLOGY/TECHNIQUES D'ENQUETE

June/juin 1979 Vol. 5 ..... No. 1
A Journal produced by Methodology Staff, Statistics Canada.
Preparé par les méthodologistes de Statistique Canada.
CONTENTS
Management of Information: Future Trends PETER G. KIRKHAM ..... 1
Non-Response In the Canadian Labour Force Survey
A.R. GOWER ..... 29
An Empirical Investigation of An Improved Method of Measuring Correlated Response Variance A. MACLEOD and K.P. KROTKI ..... 59
Selecting A Sample Of Size $n$ With PPSWOR
From A Finite Population
G.H. CHOUDHRY ..... 79
On The Inclusion Of Large Units In Simple Random Sampling M.A. HIDIROGLOU ..... 96
Unbiased Estimation Of Proportions Under Sequential Sampling M.D. BANKIER ..... 116

A Journal produced by Methodology Staff, Statistics Canada.
Preparé par les méthodologistes de Statistique Canada.

| Editorial Board/ | R. Platek - Chairman/Président |
| :--- | :--- |
| Comité de rédaction: | M.P. Singh - Editor/Rédacteur en chef <br> P.F. Timmons |
| Assistant Editor/ <br> Rédacteur adjoint | J.H. Gough |

## 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, nonsampling 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 de 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 or l'autre langue officielle.

No. 1

A Journal produced by Methodology Staff, Statistics Canada.
Preparé par les méthodologistes de 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, Census and Household Survey Methods Division, Statistics Canada, 6th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, KIA OT6. 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 des méthodes d'enquêtes ménages et du recensement, Statistique Canada, $6^{\mathrm{e}}$ étage, Edifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, KlA OT6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi. Les auteurs des articles publiés dans cette revue sont libre de les faire paraître dans d'autres revues statistiques.

MANAGEMENT OF INFORMATION: FUTURE TRENDS ${ }^{1}$

Peter G. Kirkham ${ }^{2}$


#### Abstract

This paper discusses the management of information within the context of the information industry and indicates some likely future trends related thereto. The information industry itself is first briefly described. Then the process used in producing information, the organizational structure required for such production, and the legislation relating to the information industry are discussed in turn. Finally, some approaches to solving the problems of the future are suggested.


## 1. INTRODUCTION

The purpose of this paper is to discuss the management of information and indicate some likely future trends relating to such management. Either of two obvious directions could have been chosen for the discussion: the management of information as it related to the internal management of organizations, or the management of information within the context of the information industry. The nature of the discussions relating to each of the above possible choices are not mutually exclusive.

I have chosen the latter perspective. It is broader in scope, bears directly on the question of public administration, and has some essential lessons for the former that might not be apparent if the focus of the discussion were changed.

[^0]
## 2. THE INFORMATION INDUSTRY

In a study of the information industry published in 1959 [13], Professor Fritz Machlup noted that it was not useful to distinguish between the meanings of the terms "knowledge" and "information". Hence, he accepted the definition that "information is knowledge communicated by others or obtained by personal study and investigation', or alternatively, "knowledge of a special event, situation or the like" and that, in his study, the two words were sufficiently synonymous that it was not necessary to use the redundant phrase "knowledge and information ${ }^{11}$. For the purposes of this paper, we take the same position. In addition, Professor Machlup took pains to indicate that the industry was not well defined and that statistics relating to the various components were fragmentary, incomplete, and often not relatable.

More recently, a Japanese study [21] defined the "information society", albeit primarily in computer terms, to be comprised of the information industry (which includes data supply and information processing services), data processing related industries, and the knowledge industry (which includes mass communications, education, research and development, lawyers and accountants). Figure 1 (taken from Appendix ll of [16]) more clearly identifies the scope of the Japanese definition.

[^1]
I T\&กDİ

The lack of clarity with respect to the scope and dimension of the industry, and the absence of any reliable statistical profile of its activities, is a handicap to our more complete understanding ${ }^{1}$. Nevertheless, the more advanced industrial societies are heading towards what have been termed information societies, that is, post-industrial societies in which processing and handling of information constitute the leading sectors of their economies (see, for example, [5] or [20]).

Referring to Figure 1 , it becomes immediately apparent that a very vast diversity of firms, organizations and activities are summarized under the umbrella of this information industry. The difficulty that this immediately presents is that there is no unified organization with respect to the industry, nor the availability of an appropriate forum wherein the views, needs and plans of the various components of the industry can be presented and discussed in a manner that would foster some uniformity in terms of standards and development. It has been observed that there is a need for a conceptualization of the industry and the design and promotion of coordination of the industry and the design and promotion of co-ordination mechanisms to foster its orderly development. Further, there is a requirement to review existing legislation as it relates to the many facets of this industry in order to ensure that each piece of legislation fits one with the other and, in total, is in appropriate legislation structure within which this industry should operate. See, for example, [18], p. viii, where it is stated that "A key question is how to structure the policymaking process so that the country can begin to develop a national information policy that is comprehensive to the implications of the Information Age'.

1
Kettle [10] provides an estimate of the size of the Canadian information industry.

This article will describe the developments and plans that are being fostered within Statistics Canada and relate these developments to the many issues that are implied in the above discussion. I hope that such an institutional case study will make the more general discussion of these topics in the literature more concrete and meaningful.

The remainder of this paper is divided into 3 basic parts, the production process used in producing information, the organizational structure required for such production, and the legislation relating to the information industry itself.

## 3. THE PRODUCTION PROCESS

In the past, the typical product of a statistical agency was a fixed set of statistical outputs, normally produced in the form of a hard copy publication, deriving from a specific survey or census. The product originated with a request by one or more users to supply a specified set of information. This information request was translated into a specific collection vehicle, usually a survey, and an organizational entity accompanied by budgetary resources was structured to carry out the total activity. Invariably, as a consequence, the statistical product, its associated survey, and the organizational entity performing the activity, became a fixed reality. The accompanying argument for the necessity for 'continuity" in statistical series soon assured this total activity a degree of permanence in the statistical fabric.

Consequently, when new demands for statistical information were received, these demands had to be met either by directing the requestor to a less than perfect existing substitute in the form of a printed publication or incrementally adding another survey, with its attendant organizational unit, to the statistical program.

While there was a continual effort to redirect existing resources in the light of perceived shifting priorities, the changes to existing programs were usually limited by the above-noted rigidities.

The whole question of priority setting within statistical agencies is exacerbated by the nature of the product. Machlup [13] succinctly summaries the dilemma:

> "There are several insurmountable obstacles in a statistical analysis of the knowledge industry. In the first place, there are no physical outputs. Indeed, for most parts of the production of knowledge no possible measure of output can be conceived that would be logically separate from a measure of input; and those relatively rare kinds of knowledge for which independent indices of output could be concocted cannot in any meaningful way be compared, let alone aggregated, with other kinds of knowledge. In addition, most of the services of the knowledge industry are not sold in the market but instead are distributed below cost or without charge, the cost being paid for in part or in full by government (as in the case of public schools), by philanthropists (as in the case of some private schools), and by commercial advertisers (as in the case of newspapers, magazines, radio and television). Hence, we lack the valuations which for most other industries the consumer puts on the product by paying a price for it. There are no "total sales" and no "selling prices".

Because of the non-measurability of the product, the consequent lack of productivity data, and the absence of market prices, one cannot even state with assurance that an increase in the expenditures for knowledge, relative to Gross National Product, will result in more knowledge being provided to society. Even in the few exceptional instances where we do have consumption expenditures in the form of purchases of products at market prices, the heterogeneity of the product makes a quantification of output most difficult.

Thus, while there have been periodic calls to apply the principle of cost/ benefit analysis to the priority-setting problem little progress has been made in this area to date and the prospects for any near-term breakthroughs are not great. [19] has a good discussion of the problems and summarizes some of the latest views on this matter.

The implied method of management of information within statistical agencies, explicitly or implicitly identified in the foregoing characterization of their activities have, in the past 25 years, been taking place in an external (and internal) environment that has itself been undergoing a change of revolutionary proportions.

Nanus [17] has described five information revolutions in the history of mankind: language, printing, mass media including TV/radio, computers, and community information facilities, the last representing the whole complex of computer hardware, software databanks, and communication systems operating in an environment where a wide array of information services is available to public and private users, including individuals, on-line in their own environments. The radical changes began with the "fourth information revolution', the introduction of computers. Certainly within the statistical sector of the information industry, the advent of the computer was initially foreseen as the "potential saviour" of the production problems inherent in the manual processes which were then required within mass data processing. On the other hand, many of the industry participants did not anticipate the very radical shift in user behaviour and demand patterns that would simultaneously emerge with the introduction of the computer.

As we all know, the consequences were just the converse. On the production side, the shortage of trained personnel plus the continuing difficulties with computer-software development did not allow the agencies to resolve their production problems as originally envisaged. At the same time,
these difficulties were compounded by the explosion of demand for information occasioned by user access to the computer in conjunction with the influx into the public service and other user organizations of university trained personnel knowledgeable about the computer and associated analytical research techniques.

These developments have led to strenuous efforts on the part of information suppliers to meet the level of demand articulated. In retrospect, we have now reached the stage where decision makers can be inundated with copious quantities of information relating to practically any decision or issue they wish to address. The problem is that for the decision maker the range of choice of data is so extensive that it becomes difficult to make a rational selection of the data/information most appropriate to the decision-making situation.

This situation is exacerbated by the fact, that the existing information systems are largely what Churchman ([1]], p. 33) has termed "suggestive information systems"l. Such "suggestive" systems do not make strong assumptions about how the whole system, to which the information in question relates, ought to work. In the absence of "decisive information systems", where strong assumptions about the whole system are made and courses of action more clearly indicated, there is likely to be a built-in level of frustration for users with existing information systems. This appears to have manifested itself in part in further user data demands of a very specific nature for very specific purposes, with the clear intent to make the information more directly applicable and decisive in the decision-making context. This non-substitutability of data supply has brought about large and growing demand for more and more data in micro-form which are, or appear to be, more specific and, hence, more decisive.

In addition, where data choices had to be made for specific decisions, the plethora of possible choices, even from often admittedly inadequate data, and the inability to consider rationally all of the available data, led to ad hoc, uneven choices in the information that did get used.

In summary, there is now a growing recognition that it is not more data that are needed but rather a better synthesis of existing data, a synthesis or "packing-down" which would lead to easier, more relevant and informative choices relating to the decision-making situation; alternatively, a synthesis that comes closer to the characteristics of a "decisive information system" in contrast to a "suggestive information system".

Another offshoot of this current situation is the growing demand for more "custom-tailored" information, especially in micro-data form. Such custom tailoring becomes extremely difficult when all of the existing production processes are geared to producing fixed statistical products, often in hard copy form.

The recognition of the above situation by Statistics Canada has led to a fundamental re-thinking of the manner in which we produce our information. The thrust internal to the agency is now to specify, from a systems analysis point of view, the individual functions that must be performed in the statistical process. Figure 2 (taken from [20]) describes such a systems analysis of the statistical production process. While Figure 2 and the subsequent discussion is in the context of the statistical system, I believe appropriate modifications could easily generalize this discussion to deal with the broader problem of management of other forms of information.

It may be noted that, in Figure 2, the systematic analysis of functions to be performed in any statistical process highlights several important requirements for any such statistical system.


First, there must be a function, labelled 1.1 in figure 2, to analyse user requirements and effectively determine what is needed. This is not always a simple task because users often do not have a good idea of precisely what they want, and, even if they do, such a description may not be in a form (standard terms, quality, frequency, scope of population) that is easily translatable into terms used in the statistical system.

Second, there must be a data clearing-house function which makes it possible to search an index listing of the detailed information to be found in existing data bases, using standard terms and definitions, to determine whether the request can be met out of an existing inventory, whether close, acceptable substitutes exist, or whether a new collection activity must be initiated to satisfy the need. If it is determined that the latter alternative is the only course of action, such a request must be considered in the light of the other competing requests and the resources available to meet such requests. In figure 2, this latter function is shown as 1.2.

Naturally, if the data, or close, acceptable substitues are available, they must be in a form readily available for access and dissemination. This dissemination function is identified as function 1.5 .

Several other functions are included in figure 2. The survey or census. design and the actual collection of the data is shown as function 1.3, and the analysis and interpretation of the data collected is function 1.4. An important intervening point is represented by function 1.6 , the maintenance of micro and meta data bases. These data bases result from the collection process and represent the clean edited micro data series obtained in the collection process plus the meta data which describe the characteristics of each micro data series, including name, definitions, scope, frequency, measures of quality, etc. It is these micro and meta data bases that represent the true raw material for analysis and publication of information and which are, collectively, the data capital resource about which statistical agencies in the future must be crucially concerned. Only through the conscious production of such micro and meta data, under conditions of rigid
quality control, can the necessary raw material required for customtailored products and meaningful analysis be generated.

The old concept of producing a predetermined finished product must be supplanted by the idea of manufacturing common sub-components that can be assembled in many ways in order to produce multiple end-products. Such an approach not only promises great efficiencies in the direct costs of operations but also a potential significant reduction in the indirect costs inherent in the response burden of data collection.

The bottom half of figure 2, representing functions 2.1, 2.2, and 2.3, relate to the planning process and the development, maintenance and promulgation of standard concepts and classifications, tools and practices.

Traditionally, all of the functions just described were handled by each survey group within statistical agencies. Naturally, over time and between groups, the functions were not always handled in an even, uniform manner, and the priorities attached to each function varied considerably. In other instances, with this survey-by-survey approach, it was not possible to handle other functions at all.

## 4. THE ORGANIZATIONAL ARRANGEMENTS

The foregoing discussion has focused on the process of production and has not addressed the question of how such a process is to be organized. There are two aspects to the organization of such a statistical system in Canada, namely the organization internal to Statistics Canada and the organization of all other components in the system.

Internal to Statistics Canada, the functions explicitly identified in figure 2 are being organized on the basis of several basic principles. First, the concept "sphere of observation" has been coined. Four specific examples of spheres of observation are the business sector, the household sector, the agriculture sector, and the public institutions sector, the last providing statistics on such services as justice, education, health and government.

The concept "sphere of observation" connotes two aspects. First, it is a distinct phenomenon about which information is desired in its own right, not only for the purposes of understanding the entity itself, but because programs relating specifically to the entity exist and policy analysis and decision-making revolve around such entities. Second, each sphere of observation represents a "respondent universe" from which information is sought and for which the information to be collected must be integrable and reasonably balanced in terms of coverage.

While one may wish to formulate issues and undertake analysis in terms of a single sphere of observation, issues for which information is sought increasingly span more than one sphere of observation. Thus, the concept, "sphere of observation", is usefully viewed as a construct focused more on collection and the issue of integrability of data, rather than the concern for analysis, although the latter is not ignored in this model.

In the present organization of the bureau, organizational units have already been assembled into "spheres of observation", and considerable progress has also been made in allocating, within the latter, organizational responsibility for the operations and analysis functions respectively.

The operations group(s) will be in charge of function 1.3, namely survey design and data collection, relating to a particular sphere of observation. This grouping will design surveys and collect data according to pre-determined specifications and will be responsible for producing the data, on the basis of clearly defined quality control standards, up to the point of providing the clean micro-data and meta data for the data bases identified in function 1.6.

The data content and analysis group(s) will be responsible for data analysis, interface with the user community, analysis of data user needs and provision of specifications for such data needs, for purposes of survey work. The data analysis groups will be organized according to appropriate user groupings and will be expected to know enough about the relevant area of user concern
so that the user can obtain the data and analysis desired but also to permit the analyst to act as a true consultant to the user.

Such analyst groupings may be organized around issues of concern that span more than one "sphere of observation". For example, labour market analysis must draw upon information collected from virtually all of the "spheres of observation". In order to be able to do this, the compatibility and integrability of the data must be established at the specification stage in the survey design process, and all the relevant data relating to the labour market contained in the micro and meta data bases identified in function 1.6 must be accessible to the group.

The third component, the program control group, which does not yet have explicit organizational recognition, will be responsible for function 1.2 and for ensuring that adequate inputs are provided for function 2. The program control group is charged with the responsibility of controlling the complete "information collection program relating to its sphere of observation', ensuring the necessary integrability of the data to be collected, the adherence to standards and production quality, and introducing and ensuring that a proper balance is maintained in the collection of various components of information from that "sphere of observation".

Three other points should be noted. First, the organization above is based on the principle that interfaces across organizational boundaries should be as simple as possible and, if complicated interfaces are required, they should be kept internal to an organization where the problems are better understood and the human contacts are presumably better established. Thus, in the collection phase, the interface between a given respondent community and the collectors of data is simplified; at the same time, an attempt is made to organize the analytical component so that it best conforms to well defined user groupings and interests, with the object of facilitating "one-stop" shopping on their part.

Second, it was noted earlier that there was a continuing escalating demand for information, particularly in micro-data forms, but that this demand was accompanied by an increasing frustration on the part of the user who could not obtain the information most relevant to the problem, while at the same time being overwhelmed by the supply of existing data, making it virtually impossible to use it all in the decision-making process; further any selection process was ad-hoc and uneven. This issue is dealt with organizationally in several ways. First, in function 1.1 where requirements are analyzed, it should be possible to introduce some aid to data selection. Second, the clearing house function documenting what is available will help in this selection process.

Finally, the micro-data bases, combined with the analytical groups, will permit much more "custom work" to be done for users, and their expertise in modelling and the use of other analytical techniques, combined with data specifically designed to be integratable and comparable, should allow the process of "packing-down" the data into fewer signals, each with more informational content, to be achieved. This latter development will also simplify the selection process.

Further, the analytical function, particularly as it related to modelling and other techniques, embodies the kinds of assumptions, in addition to the raw data, that translate the data base, in Professor Churchman's words, from a "suggestive" data base to a "decisive" data base.

The foregoing discussion of the production process and the internal organization of Statistics Canada has emphasized process and underplayed the specific forms of product to be produced. However, it should be perfectly clear to all concerned that in spite of the sweeping, allinclusive mandate given to Statistics Canada in the Statistics Act, it is virtually impossible to expect the agency to meet all of the needs of the total potential user clientele.

The current thinking internal to Statistics Canada is that the functions of the agency need to be explicitly identified and operationally defined. These functions should at least include:
(a) providing the necessary statistical capacity (statistical vehicles) capable of carrying out the level of information production required of the agency;
(b) to generate and maintain an agency data base(s) of clean micro and meta data files essential for the production of a set of "national statistics" and its associated publication program;
(c) the capability to respond to special user requests which draw upon, in an essential way, the information/data contained in the agency data base(s);
(d) provide sufficient but limited statistical capacity, especially in the area of survey design and collection, to satisfy special user ad-hoc requests for information/data not contained in the existing agency data base(s);
(e) help foster, coordinate, provide and maintain the necessary analytical capacity and data clearing house facilities required to provide an efficient information/data search facility for users;
(f) through intellectual leadership and coordination help conceptualize, design, and foster the construction of a total systematic or holistic approach to the management of information within the context of the information industry in Canada.

Several issues arise from this partial listing of functions. First, Statistics Canada cannot do everything in the way of meeting the informational needs of the user community. Emphasis is given above to the satisfaction of national needs with the proviso that where special needs can be met from the same micro data base, capacity will be available to handle this demand. Such a statement leaves open the question of what level of capacity will be provided and what cost recovery or pricing policies would apply.

In addition, there is a provision in this suggested scheme to do some collection, albeit limited, of data for special user groups utilizing our existing statistical vehicles. In other words, the proposal would be to have a regular core of national statistics from which national needs are met and, in addition, have a limited "custom-product" capacity to serve special user requests.

Clearly such a policy would signal the need for alternative statistical capacity, whether it be for collection or analysis, located outside of Statistics Canada. Such statistical capacity could reside with other levels of government or be located in the private sector.

To the extent that some or all of the statistical collection vehicles of Statistics Canada could be geographically located in the regions or provinces of the country, arrangements could likely be developed whereby such capacity could be utilized and shared with other levels of government, thereby in part meeting their needs for more specific local products. And presumably to some extent it could, in special circumstances, also be available to service private sector requirements.

In such cooperative or shared ventures, the intent would clearly be, where feasible, to employ concepts, standards, definitions, and survey design procedures which would permit the data collected to be utilized in more than one application and by more than one user group.

Furthermore, the intent is clear from the above suggestions that such arrangements would allow the promotion and fostering of an information industry that is privatized to the greatest degree possible and/or desirable and in which no single party has an unduly dominant role ${ }^{l}$. There are clearly basic informational requirements that must be performed by a government agency. But similarly there are specific functions as well as products that are probably better handled privately. The above conceptual approach to the statistical process and the accompanying organizational arrangements would facilitate the fostering of this private involvement wherever possible within the context of an overall conceptual, organizational approach to the total requirements.

Specifically, various functions in figure 2 can be organizationally and geographically distributed throughout the industry with, in some cases, groups of functions being performed solely by one organization, and in other cases, a single function being carried out by a large number of organizations.

If this conceptualized approach is followed and all parties adhere to certain common practices, relating to stardards, concepts, definitions, survey design and subsequent processing, the resulting micro and meta data bases acquire a potential usefulness for all users within the system.

1
Diebold [3] has observed that "No matter how we re-organize things to produce management efficiency, nothing is going to keep increasingly unproductive, labour-intensive public services in important areas of life from declining .... Most of our really important activities - education, public transport, medical service distribution, the running of cities - don't rely on advanced technology and therefore decline in quality and their costs spiral up.

The real problem here is not to try to force technology and modern management through the system ... but to see if we can't find ways to create the kind of demand that stimulates innovation and high productivity in the private sector'.

The suggested approach to statistical information management discussed in this paper allows such potential developments as prescribed above by Diebold to take place to the extent that it is felt to be feasible and desirable.

Such data bases, irrespective of organizational or geographic location, if interconnected by a computer network, create the potential for a truly useful reservoir of information for the decision-maker both inside and outside of government. For example, it should be possible for a user, utilizing a high-level computer language, to request information from any node in such a network. The request should automatically generate a search of the data clearing house mechanism to locate the information desired, the required data bases should be accessed automatically, the appropriate data massaged into the form desired, and transmitted to the user at the point where the initial request was made. On the dissemination side, CANSIM, the Statistics Canada computerized data base is an embryonic version of a very primitive mode. The realization of a fuller model is dependent upon the development of the organizational arrangements referred to, plus the appropriate legislative structure to support such a statistical information system design.
5. LEGISLATIVE POLICIES FOR THE INFORMATION INDUSTRY

The technology required for the realization of the previous scenario is available today. The literature provides an extensive discussion of this technology and since, in my view, this aspect will not be one of the constraining factors in future trends, it is not treated further here.

The fundamental determinants of future developments are threefold: the social attitudes of the Canadian people towards information and its management within our society, the legislation required to embody these public attitudes and views into practice, and the "people problems" inherent in any organizational and cooperative ventures necessary for the implementation of such practices.

The societal issues relate to privacy, confidentiality, compulsion versus voluntarism in response, and the question of freedom of information. The resolution of these issues in turn will determine the outcome of such subsidiary issues as the degree of sharing of information, who has access
to information, who must supply information, the use to which various sets of information can and will be put, the use of common identifiers, questions of data-linkage, the level of response burden acceptable in the system, who will control the system, how it will be controlled, and the forms of checks and balances necessitated by any particular organizational form.

Specific legislation in Canada relating to such issues includes the recently enacted human rights legislation, the proposed freedom of information legislation, statistic acts in general and the Federal Statistics Act in particular. Other legislation, less of ten identified with these issues but nevertheless relevant, include copyright laws as they relate to photocopying, data banks, computer software protection, information exchange and third-party record keeping; in fact the whole concept of public and private property. Trade secrets, patent laws, the impact of information sharing on the publication industry and standardization of computer hardware technology are additional concerns. In summary, the level of public understanding of such fundamental issues as privacy and confidentiality and the final determination of the preferred public position on such issues, reflected in such laws as the now existing human rights legislation and the proposed freedom of information legislation will structure the range of organizational choices within the context of the previous scenario.

In order to ensure that the optimum choice is made from the point of view of the public, substantial public discussion and debate of the underlying issues must take place and the legislative framework within which the information industry needs to operate must openly and fairly reflect the public's opinion on these matters.

Substantive work designed to address these issues has been proceeding in a number of countries. (For example see [18], [20], [1], [15]). However, a current dilemma in many countries is captured in the remarks of the Conference Board (see [16], p. 6):

'Within the [US] Federal Government, policy responsibility relating to the information industry is still distributed and fragmented among different agencies, precluding the formulation of comprehensive and coherent "alternative" policies that extend beyond the needs of a particular agency'.

Such a dilemma is further exacerbated in a country like Canada by the fact that the concept of an information industry is still ill-defined. For example, Mauerhoff ([15], p. 35), has recently observed "in Canada the concept of an information industry is still fragmented. There is a publication industry, a broadcast industry, a computer industry, a telecommunications industry, but no attempt has been made to include them in a single framework with information as the common denominator'. Further, few writers have given explicit recognition to the role of the statistical system in such an industry. One notable exception is Kettle ([10], p. 40), who explicitly identified Statistics Canada as part of the "Professional Services" sub-component of the Canadian information industry. However, not only will the statistical system of the future represent an increasingly effective source of information in its own right within the industry, on the basis of data obtained from surveys, censuses, and increasingly from administrative data bases, but the data clearing house function of the statistical system identified in figure 2, will also play a crucial role in any such industry.

The data clearing house function of the future must provide a mechanism whereby the numeric and qualitative information within statistical agencies can be cross-referenced and associated with the textual material in books, articles, periodicals, and scholarly works located in libraries. While, in the first stage of development of such clearing houses, perhaps only an index listing of information will be available, ultimately there will be a requirement for an "intelligent search and selection" requirement to be built in. In the, absence of such devices, the deluge of information will render the system grossly inefficient at best and inoperable at worst. (See [1], P. 30).

The foregoing discussion has highlighted the need for an overall view of the legislative structure of the information industry. What has not been emphasized is the equally important need to ensure that the organizational arrangements take account of the need for the services of this industry to be as wide open as possible and available to society as a whole.

The respondent, who supplies the information in the system, has often been in the past an individual or small organization who has not had either the resources or the capacity to benefit from the information within the system. These respondents have been the "information-poor", but the information that they have supplied has of ten been used by those who do access the information system, "the information-rich', in ways that of ten have been interpreted by the small organizations as detrimental to their interests.

Unless the system can be developed in an open manner, in which information is available to all in an accepted, equitable fashion and in which, through the provision of facilities and training; the "information-poor" are given equitable treatment, the source of the information to be put into the system could well dry up.

Finally, the foregoing discussion has highlighted the diversity of functions, activities, and organizations within the information industry on the one hand, and the need for a common approach to concepts, definitions, terminology, organization, conduct and behaviour on the other. Since so many different organizations with different levels of authority and mandates must come together in this industry, one must ask what the possibility of achieving some kind of co-ordinated approaches really are.

Arrow ([2], pp. 68-70) has summarized the dilemma especially in those instances where the price system does not work well':

'lauthority is needed to achieve a co-ordination of the activities of the members of the organization ... An organization whose members have identical interests and identical information will be one in which sponteneous consensus would be efficient ... When either interests or information differ among the members of the organization, the costs of achieving consensus rise, and hence the value of consensus as a mode of organizational decisionmaking declines relative to that of authority'.

These remarks suggest that in the absence of co-ordinated legislation designed to promote the orderly development of this industry and the provision of intellectual leadership in the conceptualization and design of the overall organizational approach, the orderly, uniform development of the industry will be difficult to achieve. If, on the other hand, such legislation and conceptualization can be structured to promote individual initiatives that are both in self-interest of the individual organization within the industry and in the collective interest of all the other participants, the proper development of the industry will be assured.

[^2]
## 6. CONCLUSIONS

The full realization of major organizational trends and developments required within the information industry must await the partial or complete outcome of public debate on the issues involved and the related legislative changes.

In the meantime, the national statistical agency is attempting to restructure its internal activities along the functional lines outlined in figure 2 , define its mandate in operational terms, and compile a set of medium term plans consistent with such a mandate. Emphasis is being given to the process to be managed with a clear attempt to structure the substantive work content in such a way that more meaningful, multipurpose uses, better suited to user needs, will become a statistical reality.

The functional approach to the production of information and the conceptual approach to its organization offers a very robust model; robust in the sense that, from the point of view of the statistical agency, just about any position taken by society on the issues of privacy, confidentiality and the other areas of concern can be accommodated within the model. Naturally, the degree of privatization of functions and the degree of sharing of information is heavily dependent on the particular position taken, but the general approach described can, it is hoped, accommodate a wide range of possible public positions in these matters. I also feel that the management of non-statistical information can be usefully approached on the basis of the concepts discussed.

In addition, in keeping with the expressed need for openness in the system, in order to ensure better public understanding and acceptance of our role in the management of information, efforts are being made to develop mechanisms that will effectively meet these perceived needs of openness, fairness and equity within the system.
-

Recognition is also given to the problem of escalating quantities of information. Clearly the steps outlined here to deal with this problem fall short of providing a full solution. The whole area of decisionmaking and the development of theories and approaches to decision-making have been completely omitted from this discussion.

In part, the suggestions by Churchman ([11], p. 33) with respect to a "decisive data base" go part way to meet this concern. However, people like Marschak [14] clearly see the need to include, in a much more fundamental way, the issue of approaches to decision-making in any discussion and design of a rational information system.

This does not necessarily mean that the system must be geared to produce only sophisticated, technical information relating to highly scientific subjects. As Kochen has observed ([1], p. 195):
> "Though that is important, it does not compare with the seriousness of the day-to-day problems faced by community leaders and individuals as non-specialists, and with which information systems could perhaps be expected to help. Much of the knowledge and understanding they need or get is not scientific'.

But, in the final analysis, the crucial issue remains: "the most serious shortcomings ... are to be found in the management and organization of the services engaged in the collection, processing and redistribution of information' ([1], p. 28). The suggestions put forward in this paper address many of those issued directly and offer a promising direction of development in resolving them.

L'auteur examine la gestion de l'information dans le contexte de l'industrie de l'information et présente quelques tendances futures possibles s'y rattachant. L'industrie elle-même est brièvement décrite au début. L'auteur examine ensuite successivement la production de l'information, la structure organisationnelle nécessaire à cette production et la législation qui s'y rapporte. Enfin, l'auteur propose quelques solutions aux problèmes de l'avenir. $^{\prime}$.aver

## REFERENCES

[1] Anderla, George, "The Future of Information: A Challenge For Governments and Society', The OECD Observer, No. 63, April 1973, pp. 27-32.
[2] Arrow, Kenneth J., The Limits of Organization, W.W. Norton and Company, Inc., New York, 1974.
[3] Diebold, John, 'Why Things Don't Work Anymore', Newsweek, July 18, 1977.
[4] Diener, Richard A.V., "Value: Data Information", Knowledge, American Society for Information Science, Collection of Papers presented at 6th Mid-Year Meeting, May, 1977, Syracuse University.
[5] Drucker, Peter, The Age of Discontinuity, New York, Harper and Ron, 1969.
[6] Henry, Nicholas, "Public Policies for Information Management: An Overview', Information Processing and Management, Vol. 11, pp. 67-74, 1975.
[7] "Knowledge Management: A New Concern for Public Administration", Public Administration Review, May/June 1974, pp. 189-196.
[8] "Bureaucracy, Technology, and Knowledge Management", Public Administration Review, November-December, 1975, pp. 572-578.
[9] Hillman, Donald, J., "Model for the on line management of knowledge transfer'', On-Line Review, 1977, Vol. 1, No. I, pp. 23-30.
[10] Kettle, John, "The Knowledge Industry", Executive Magazine, November 1974, pp. 39-42.
[11] Manfred Kochen, ed., Information For Action, From Knowledge to Wisdom, Academic Press, New York, 1975.
[12] Ludlow, Hope T., "Information Technology:The Revolution Reconsidered", pp. 52-63, The Conference Board Record, June, 1972.
[13] Machlup, Fritz, The Production and Distribution of Knowledge In the United States.
[14] Marschak, Jacob, 'Economics of Inquiry, Communicating, Deciding", American Economic Review, May, 1968, pp. 1-18.
[15] Mauerhoff, George, R., "An Information Industry for Canada", Business Quarterly, Summer, 1977, pp. 35-41.
[16] Morris, Dubow, S., Jr., ed., Information Technology, Initiatives for Today Decisions That Cannot Wait, The Conference Board Inc., New York, 1972.
[17] Nanus, Bert, "Profiles of the Future, Managing the Fifth Information Revolution, pp. 5-14, Business Horizons, April, 1972.
[18] National Information Policy, Report to the President of the United States, prepared by the Domestic Council Committee on the Right of Privacy, Nelson A. Rockefeller, Chairman, Washington, D.C., 1976.
[19] Setting Statistical Priorities, Report of the Panel on Methodology for Statistical Priorities of the Committee on National Statistics, National Academy of Sciences, Washington, D.C., 1976.
[20] Systems Analysis of An "Ideal Statistical System of the 1980's", Statistics Canada paper, prepared by Dr. I.P. Fellegi, for the Conference of European Statistician's Seminar On Statistical Services in Ten Years Time, Washington, March, 1977.
[21] The Japanese Computer Usage Development Institute, The Information Society: A Year 2000 Japanese Nation Goal, 1970, as reported in [12].

# NON-RESPONSE IN THE CANADIAN LABOUR FORCE SURVEY 

A.R. Gower ${ }^{1}$


#### Abstract

This paper includes a description of interviewer techniques and procedures used to minimize nonresponse, an outline of methods used to monitor and control non-response, and a discussion of how nonrespondents are treated in the data processing and estimation stages of the Canadian Labour Force Survey. Recent non-response rates as well as data on the characteristics of non-respondents are also given. It is concluded that a yearly non-response rate of approximately 5 percent is probably the best that can be achieved in the Labour Force Survey.


## 1. INTRODUCTION

The Labour Force Survey (LFS) is carried out as a monthly probability sample of dwellings. Households within the selected dwellings are interviewed once a month for six consecutive months. After six months these dwellings are replaced by another group of dwellings in such a way that every month one-sixth of the sample is replaced or, in other words, rotated [1].

In one particular week (called survey week) each month about 62,000 dwellings throughout Canada are contacted by approximately 1,100 interviewers. Information is collected by the interviewers on the demographic characteristics and labour force activities of the civilian, noninstitutional population 15 years of age and over who are members of households belonging to these dwellings.

[^3]For various reasons interviewers are not able to obtain an interview at every selected dwelling. These non-interviews occur from the following sources:
(a) household non-response - including reasons such as refusal, no one at home and temporarily absent;
(b) vacant dwellings - including unoccupied dwellings, vacant seasonal dwellings, dwellings under construction, and dwellings occupied by persons not eligible to be interviewed; and
(c) 'non-existent' dwellings - including dwellings which were demolished, moved or listed in error.

More precise definitions of the reasons for non-interviews are given in Appendix A, together with a definition of the non-response rate.

It is essential that household non-response be kept at a minimum level since a high non-response rate leads to a high sampling variability for labour force population estimates and may increase the mean square error as a result of non-response bias. Although adjustments for nonresponse are made in the data processing and estimation stages of the survey, it is very important that every effort be made in the field to interview all households.

Vacant and "non-existent" dwellings do not contribute any bias to the sample but do result in a higher sampling variance because of a smaller household count. It should be noted that vacant dwellings are visited every month in order to interview any which may have become occupied by persons eligible to be interviewed. After "non-existent" dwellings have been detected, they are not visited again and are excluded from the sample.

The focus of this paper is on household non-response in the LFS with emphasis on the methodology which has resulted in non-response rates of approximately 5 percent. This very low non-response rate has been
achieved through effective interviewing techniques and procedures which are described in Section 2 as well as regular monitoring and controls which are described in Section 3. A summary of recent nonresponse rates is given in Section 4 . The treatment of non-respondents in the data processing and estimation stages of the survey is discussed in Section 5, while demographic and labour force characteristics of non-respondents are presented in Section 6. Concluding remarks are given in Section 7.

## 2. INTERVIEWING TECHNIQUES AND PROCEDURES TO CONTROL NON-RESPONSE

### 2.1 Assignment Planning

Interviewers are instructed to enumerate all dwellings in their assignments while keeping expenditure and travel to a minimum. Their assignments must be completed before the end of the six-day survey week. Interviewers work only on the LFS during this time and, although supplementary surveys are occasionally carried out simultaneously with the LFS, priority is given to completing the LFS within the specified time limits.

Generally speaking an interviewer's workload depends on the type of area enumerated and whether telephone interviews are permitted. Usually rural assignments consist of approximately 40 to 50 dwellings. Urban assignments are larger: 70 to 80 dwellings in the case of telephone assignments. It is suggested that interviewers make up to three or four calls to every dwelling and at least one call before the fourth day of survey week. Interviewers usually complete first month interviews and nonresponses from the previous month as early as possible during survey week. During subsequent visits interviewers attempt to contact households at the "best time to call" (determined at the time of the first contact with a household). If unsuccessful, they make calls at different times on different days.

Interviewers are encouraged to plan their schedules and routes in order to make efficient use of time. For instance, areas close to the interviewer's own home are visited first, so that call-backs can be made on the way home. In telephone assignments interviewers telephone frequently at different times and follow-up unsuccessful attempts to telephone with a personal visit. Interviewers determine from relatives, neighbours or superintendents when a household is most likely to be at home so that call-backs can be scheduled accordingly. If the household is not at home on the first visit, the interviewer leaves a brochure describing the survey with a note requesting an appointment [2].

### 2.2 First Contact

Through improved training methods and policies such as "doorstep diplomacy"
[3] interviewers are becoming more knowledgeable on how to gain respondent cooperation and conduct interviews effectively. During training interviewing techniques such as appearance, introduction, asking the questions, handing delicate situations and ending the interview are emphasized. Very important in this training is the interviewer's introduction and presentation of the survey.

It is very important that the respondent be informed about the nature and purpose of the survey. For this reason, in cases where a mailing address is known, an introductory letter together with a brochure describing the LFS is sent to the household prior to the first interview. When the interviewer visits the respondent already knows something about the survey; this helps the interviewer with her introduction. If mailing is not possible, then the interviewer presents the letter at the time of her first visit. In all cases, interviewers must ensure that every household has material explaining the survey.

Every interviewer must carry a Statistics Canada identification card and present it to the respondent at the beginning of the interview. This helps gain the respondent's confidence and ensures that the respondent clearly
understands who the interviewer is and whom she represents. The interviewer gives a short explanation of the survey, and assures the respondent that the information which is being collected is confidential. In subsequent interviews the interviewer tries to contact the person who was interviewed last month (particularly important in the case of telephone interviews), but if this is not possible then another responsible member of the household is interviewed.

### 2.3 Proxy Interviews

Because of time and cost constraints it is virtually impossible to obtain non-proxy responses from every individual. For this reason proxy interviews are accepted in the LFS. Generally only one member of a household is interviewed, and this member responds on behalf of all other members. Occasionally, separate interviews are required for household members such as roomers or boarders. On the average it has been found that proxy interviews account for approximately 50 percent of all respondents [4]. Furthermore, because proxy interviews are accepted, it is possible to obtain complete responses for all household members in virtually 100 percent of all responding households. In fact, there are less than 0.2 percent of all households where interviews (proxy or non-proxy) are obtained for some, but not all, members of the household. The acceptance of suitable proxy respondents is, therefore, an effective means of reducing non-response.

### 2.4 Refusal Follow-Ups

Refusal households are followed up whenever feasible. In most cases this involves a personal visit by a senior interviewer or a regional office representative. In areas where this type of follow-up is not possible, a letter may be sent. Households are usually provided with additional information about the survey and how the data will be used. The importance of the survey and the co-operation of the respondent are emphasized. The result is that many of these refusal households can be successfully
interviewed the following month. In the case of households which cannot be persuaded to respond, interviewers are told not to visit them again unless there is a complete change in household composition.

### 2.5 Post Survey Week Follow-Ups

During the summer months when the number of temporarily.absent and no one at home non-responses are higher than usual, follow-up interviews are carried out during the week following survey week. Although this post survey week follow-up program is carried out only in July and August on a regular basis, it may occasionally be used in other months if nonresponse rates are higher than usual.

In the post survey week follow-up program interviews are carried out on the Monday and Tuesday of the week following survey week by a team consisting of at least one regional office staff member and a few interviewers who make post survey follow-ups to all types of non-response households except refusals and those which are not suitable for followup action due to location, respondent svailability or other special circumstances. The team carries out the post survey week follow-ups either by telephone from the regional office or by personal visit. When a respondent is contacted, the follow-up interviewer identifies herself, explains why she is calling or visiting, and refers to the name of the original interviewer. In the case of a first month visit the follow-up interviewer completes the interview in the usual way and tells the respondent that a different interviewer (i.e. the regular interviewer) will be contacting the household next month. It is very important to explain to the respondent the week (two weeks ago!) to which all survey questions refer and to ensure that the respondent understands this time frame. The interviewer then is very careful to ask all questions referring to reference week.

Indications are that the post survey week follow-up program has resulted in a decrease in non-response during the months of July and August (refer
to section 4 of this paper). Data collected for one regional office showed that the total non-response rate was reduced by approximately 2.0 and that 25 percent of the non-responses during survey week were successfully followed-up. Many of these households required only one telephone call to complete an interview, while others required up to three calls. Households contacted four times or more usually remained non-responses. Interviewers found that their most productive time appeared to be Monday afternoon and early Monday evening, and that by later Monday evening and Tuesday they were making many calls to a household without success [5].

### 2.6 Telephone Interviewing

The telephone interviewing procedure which is used in the LFS involves a combination of personal visits and telephone calls and is carried out only in large urban areas. Interviewers must conduct all first month interviews in person, and telephone interviews can only be carried out in subsequent months if the respondent agrees to be interviewed by telephone. Experience has shown that approximately 75 percent of households are interviewed by telephone in assignments where the telephone interviewing procedure is allowed.

Although the primary reason for using telephone interviewing in the LFS is to reduce enumeration costs, telephone interviewing seems worthwhile from the point of view of non-response since (a) it allows interviewing to be completed on time regardless of weather conditions, (b) it is especially ideal for single persons, small family households and apartment dwellers who are difficult to find at home and who can often be reached only during the evening (it is easier for an interviewer to phone at night than to make a personal visit), (c) it allows interviewing to be conducted more readily at the convenience of the respondent (if one time is not suitable, then another can be easily arranged), and (d) it has the potential of reducing non-response by allowing more opportunity and time for callbacks.

However, the telephone interviewing procedure does not appear to have a direct impact on the reduction of non-response rates. This was indicated by the results of a telephone interviewing experiment carried out during 1972 and 1973 [6]. The experiment showed that respondents who agree to be interviewed by telephone are very unlikely to be nonrespondents during subsequent interviews. Those households which do not agree to telephone interviewing or which cannot be telephoned (for reasons such as no telephone available, party line telephone, unlisted telephone number, complete change in household composition, language or hearing problems, etc.) can be expected to contribute nearly 50 percent of all non-responses, even though this group of households accounts for only 10 to 20 percent of all households in telephone assignments. Therefore, if more effort is directed towards these households, then presumably the non-response rate can be reduced. According to the present interviewing procedures, interviewers are instructed to complete as many telephone interviews as possible during the first day or two of survey week and to try to contact rotate-in households and non-responses from last month as early as possible during survey week. In this way, least priority is given to contacting the households which cannot be interviewed by telephone for reasons other than first month interviews. This suggests that non-response may be reduced among those households which can be interviewed by telephone, but that more non-response may be occurring among those households which are not telephoned.

### 2.7 Response Incentives

Respondent participation in the LFS is compulsory under the Statistics Act, but no monetary incentive or any other form of incentive is offered. However, a response incentives experiment was carried out in the LFS during 1975 and 1976 in order to determine the effectiveness of a response incentive on improving respondent relations and interviewer performance [7]. The response incentive used in the experiment was the "Canada Handbook", an annual Statistics Canada publication. In half of the LFS sample interviewers gave "Canada Handbooks" to all households being visited for the
first time, while all other households received no response incentive.

Households which received the "Canada Handbook" had a marginally lower refusal rate than households which did not receive it. Indications were that the distribution of the "Canada Handbook" had very little effect on converting a refusal to a response at the time of the interviewer's first contact with a household, but respondents who received the "Canada Handbook" were less likely to refuse at some later time than respondents who were not given a response incentive. Although the majority of interviewers indicated that they feel response incentives such as the "Canada Handbook' are useful in establishing a good rapport with respondents, most interviewers believe that materials such as the introductory letter and identification card are actually more effective than gifts.

The response incentives experiment showed that there is real need to provide the respondent with more information on the purposes of the survey and the uses of the data. It is very important, therefore, that interviewers be equipped to provide this information since they have the main responsibility in gaining the co-operation of the respondent. Interviewers can acquire this knowledge through training which emphasizes the purposes and importance of the survey and by having support material available such as an introductory letter or explanatory brochure which illustrates these points.

## 3. MONITORING NON-RESPONSE

For dwellings where no contact can be made interviewers identify the precise reason for the non-interview and record this reason on the Household Record Docket. Interviewers also complete a non-interview report

[^4]explaining the circumstances for no interview as fully as possible. Following the interviewer's coding of the non-interview the regional office decides what action should be taken the next month. This action is then pre-printed on next month's Household Record Docket indicating whether the interviewer should attempt to interview the household again or whether she should not conduct an interview unless there is a complete change in household composition.

Information on the Household Record Dockets is transmitted from the regional offices to Head Office where a non-response file is created which summarizes relevant information related to the response status of every selected dwelling in the LFS sample. This non-response file enables a series of comprehensive reports on non-response to be produced every month within five or six days following the end of survey week. The reports include detailed breakdowns of sample sizes, nonresponse rates and vacancy rates for interviewers, regional offices, provinces as well as every type of area (SRU, NSRU and special areas) ${ }^{3}$ They help Head Office personnel identify areas where non-response is a problem and, if necessary, regional offices can be contacted to take remedial steps to reduce non-response rates in these areas. Using the non-response file various other analyses of non-response rates can be made including breakdowns for specified sampling units, apartment and nonapartment samples, rotation groups as well as telephone and personal visit assignments. This type of information is very useful when investigating the behaviour of non-resfonse rates over a period of time and isolating where high non-response rates may be occurring.

The performance of interviewers is continually monitored and reviewed.

```
\({ }^{3}\) SRU's (self-representing units) are areas whose population exceeds
    15,000 or whose unique characteristics demand their establishment as
    SRU's. NSRU (non-self representing units) are those areas outside SRU's
    and are comprised of rural areas and small urban centres. Special areas
    include military establishments, remote areas, and institutions such as
    hospitals, schools and hotels.
```

Regional offices have a month'ly report on non-response rates at the interviewer level. This report can be produced in each region as soon as survey week has been completed, and it allows supervisors to take immediate action in specific circumstances when interviewers have unusually high non-response rates. The objective is to remedy these situations in time for the next survey and to maintain non-response rates at a satisfactory level.

In the LFS there are also regular programs of observation and re-interview. The observation program [8] is carried out for the purpose of evaluating and improving the performance of interviewers. Every month about one-tenth of the interviewers are selected for observation including interviewers scheduled for systematic observation as well as new interviewers and interviewers whose performances suggest they need observation. The observer, usually a senior interviewer, accompanies the interviewer into the field during survey week and evaluates the interviewer in nearly all aspects of her work. This observation provides an excellent opportunity to train or re-train interviewers in the use of proper interviewing procedures. Among specific areas evaluated are the interviewer's knowledge of non-interview procedures and her ability to minimize the number of non-responses in her assignment.

The re-interview program [9] is conducted in the week immediately following survey week by a senior interviewer or supervisor. Households which were enumerated during survey week are contacted again, and re-interviews proceed with the re-interviewer repeating the same questions previously asked by the interviewer. Any observed differences between the two sets of responses are attributable to several sources including the respondent, the interviewer, the re-interviewer, shortcomings in the instructions or training provided to interviewers, and the wording and sequence of the questions. Although the primary purpose of the re-interview program is to measure response errors, it also allows the opportunity to check the quality of an interviewer's work including her handling of non-interviews. Checking with the respondent, for example, indicates whether or not the interviewer varied the time of her calls and made several call-backs. The re-interview program, therefore, is a complement to the observation
program since it helps identify interviewer weaknesses and needs for further training. On the basis of re-interview findings a special observation or re-training may be recommended.

## 4. NON-RESPONSE RATES

In the LFS it has been possible to maintain low non-response rates through effective training programs, regular monitoring and controls as well as interviewing techniques and procedures which help reduce non-response. With this approach it seems that the total non-response rate cannot be brought lower than 4.0 in any one month and that, due to seasonal variations, a realistic objective is an annul average of 5.0. In 1977 and 1978 the average non-response rate was 5.4 each year. During this time the refusal rate remained at approximately 1.3 to 1.5 : these refusals are likely the "hard core" refusals, and it may be very difficult, if not impossible, to convert any more of this group to interviews. An optimum value for the no one at home rate appears to be 1.0 , given the constraints of cost and the length of the survey week. Over the last two years this component averaged 1.5 with only slight fluctuations occurring from month to month. The temporarily absent rate fluctuated considerably from month to month each year but was usually at least 1.5 in any one month. Other reasons for non-response contributed about 0.5 to the total non-response rate.

Graphs 4.1 and 4.2 show the total non-response, refusal, no one at home and temporarily absent rates over the four year period from 1975 to 1978. Highlights of these graphs include the following points.
(a) The level of non-response generally decreased over the four years (when corresponciing months are compared).
(b) The total non-response rate showed a tendency to decrease through the months of January to April each year (with the exception of April 1976 and January to April 1978). It increased from April to May, usually decreased a little during June, and increased from June to
Non-Response Rates in the LFS


July reaching a peak in July. Once the non-response rate had peaked in July, it decreased rapidly during August and September. The non-response rate remained reasonably stable during October, November and December, while an increase occurred from December to January each year.
(c) The total non-response rate peaked each year during the July survey due to a very high temporarily absent rate. However, the total non-response and temporarily absent rates were substantially lower in July 1977 and 1978 than in July 1975 and 1976. This was probably due to the implementation of the post survey week follow-up program during the July and August surveys in 1977 and 1978.
(d) In addition to the June, July and August surveys the temporarily absent rate had a tendency to increase during the winter months of February and March. This probablyreflects the fact that more people are taking winter vacations.
(e) Improvements occurred in the no one at home rate over the four years, and it is now approximately 1.5.
(f) Except for the first few months of 1975 the refusal rate remained fairly stable. The increases which occurred in April 1976 and April 1978 and continued into the May survey those two years apparently were the result of the Survey of Consumer Finances which was conducted as a supplement to the LFS during April. Over the last two years the refusal rate has remained at approximately 1.3 to 1.5 with the exceptions already noted, compared to 2.5 in May 1976.

Table 4.1 summarizes non-response rates according to the number of times (one to six) that households were enumerated. The rates shown on the table represent averages over twelve rotation groups ${ }^{4}$ which entered the survey for the first time from July 1977 to June 1978 and remained in the survey for a period

[^5]of six months (rotating out from December 1977 to November 1978). The results are considered typical for any consecutive twelve month period.

## TABLE 4.1

Non-Response Rates (\%) According to Tenure of Households in the LFS (Averaged over 12 rotation groups, entering survey from July 1977 to June 1978)

| Number of months in survey | Non-response rates (\%) |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | ```Total non- response``` | Refusal | No one at home | $\begin{aligned} & \text { Temporarily } \\ & \text { absent } \end{aligned}$ |
| 1 | 8.04 | 1.43 | 2.96 | 2.94 |
| 2 | 5.09 | 1.21 | 1.44 | 1.99 |
| 3 | 4.71 | 1.32 | 1.10 | 1.90 |
| 4 | 4.65 | 1.46 | 1.09 | 1.79 |
| 5 | 4.62 | 1.51 | 0.99 | 1.77 |
| 6 | 4.45 | 1.52 | 0.78 | 1.73 |

On the basis of the results shown on Table 4.1 the following comments can be made:
(a) The total non-response rate was highest during the first month, presumably because interviewers had more difficulty finding people at home having not yet determined the best time to call. The rate then decreased sharply in the second month and continued to decrease through the third and fourth months. A very slight decrease occurred from the fourth month to the fifth month, while a larger decrease occurred in the sixth month.
(b) The refusal rate decreased in the second month, increased gradually through the third, fourth and fifth months and levelled off in the
sixth month. It should be noted that refusal rates tend to be cumulative since a refusal one month will likely remain a refusal the next month. In this way any sharp increase in the refusal rate for one month can be expected to result in higher refusal rates during subsequent months, with the rate gradually decreasing as the respondents who refused rotate out of the sample. The decrease in the refusal rate observed in the second month was concentrated within SRU's where the rate decreased from 1.9 to 1.4 , while in NSRU's the refusal rate remained steady at about 1.0 during the first two months. The decrease in the SRU refusal rates reflects the success of refusal follow-ups in these areas. This decrease was not observed in NSRU's, probably because no refusal follow-ups are carried out due to the remote nature of NSRU's.
(c) The no one at home rate decreased sharply from the first month to the second month by roughly 50 percent. It continued to decrease from the second month to the third month but decreased very gradually through the fourth and fifth months. A larger decrease then occurred in the sixth month. The behaviour of the no one at home rate over the six month tenure of households in the survey was most probably due to the fact that the longer a household is in the survey the more familiar the interviewer becomes with knowing when the respondent is most likely to be at home.
(d) The temporarily absent rate decreased through all six months, particularly from the first to second month. It is difficult to explain this phenomenon since the temporarily absent rate should not be expected to depend on how long a household remains in the survey. One can hypothesize that interviewers may have confused no one at home and temporarily absent types of non-response.

Whereas the total non-response rate at the Canada level averaged 5.4 during 1977 and 1978, non-response varied from region to region due to many reasons such as geography, respondent characteristics and attitudes in each area, weather conditions and regional office procedures. Table 4.2 illustrates this variation for 1978.

## TABLE 4.2

Non-Response Rates (\%) by Regional Office
(Monthly average:1978)

| Regional office | Number of households | Total nonresponse | Refusal | No one at $\qquad$ home | Temporarily absent |
| :---: | :---: | :---: | :---: | :---: | :---: |
| St. John's | 3,024 | 4.7 | 0.8 | 1.2 | 2.1 |
| Halifax | 9,632 | 5.5 | 1.3 | 1.7 | 1.8 |
| Montreal | 7,865 | 4.9 | 1.3 | 1.7 | 1.7 |
| Ottawa | 2,729 | 5.1 | 1.4 | 1.4 | 1.9 |
| Toronto | 8,428 | 6.3 | 1.8 | 1.7 | 2.3 |
| Winnipeg | 9,242 | 4.9 | 1.4 | 1.3 | 2.1 |
| Edmonton | 9,302 | 4.9 | 1.4 | 1.2 | 2.0 |
| Vancouver | 6,132 | 6.4 | 2.0 | 1.5 | 2.3 |
| Canada | 55,354 | 5.4 | 1.4 | 1.5 | 2.0 |

Non-response rates also depend upon the type of area being enumerated. Depending upon the distribution of areas within a region, this can affect the level of non-response in a regional office. At the Canada level nonresponse rates according to type of area are shown on Table 4.3.

These results show that the total non-response rate was higher in SRU's than in NSRU's. Averaged over twelve months temporarily absent rates were the same in both areas. However, temporarily absent rates were higher (by as much as 30 percent) in SRU's than in NSRU's during the months of May, June, July, August and September, while NSRU temporarily absent rates were higher (by as much as 35 percent) during the remaining months of the year. This phenomenon may have been due to the fact that people in rural areas
move to larger centres during the winter and that families living in cities usually take summer vacations. Although the temporarily absent rates averaged over the year were the same for SRU's and NSRU's, the no one at home rate in SRU's was almost 25 percent higher than the corresponding rate in NSRU's. The refusal rate was approximately 40 percent higher in SRU's than in NSRU's.

TABLE 4.3

## Non-Response Rates (\%) by Type of Area <br> (Monthly average: 1978)

| Type of Area | Approximate proportion of sample | Total response | Refusal | No one at home | Temporarily absent |
| :---: | :---: | :---: | :---: | :---: | :---: |
| NSRU | 0.48 | 5.0 | 1.2 | 1.3 | 2.0 |
| - urban5 | 0.18 | 5.3 | 1.0 | 1.4 | 2.4 |
| - rural ${ }^{5}$ | 0.30 | 4.9 | 1.3 | 1.3 | 1.8 |
| SRU | 0.51 | 5.7 | 1.7 | 1.6 | 2.0 |
| - built-yp ${ }^{6}$ | 0.37 | 5.5 | 1.6 | 1.5 | 2.0 |
|  | 0.10 | 4.8 | 1.6 | 1.2 | 1.8 |
| - apartment ${ }^{7}$ | 0.04 | 9.8 | 2.7 | 3.6 | 3.0 |
| Special Areas | S. 0.01 | 6.4 | 0.4 | 0.9 | 2.8 |

[^6]Special areas, on the other hand, had a higher total non-response rate than either SRU's or NSRU's, mainly due to the temporarily absent and "other reasons" components of non-response. These relatively high nonresponse rates likely resulted from the remote nature and composition (including hospitals, schools, and hotels) of many special areas. The attention on non-response is usually directed toward SRU's and NSRU's since these areas account for 99 percent of the LFS sample, while special areas contribute only $l$ percent of the sample.

Within NSRU's the total non-response rate was higher in the urban portion due to higher temporarily absent rates among NSRU urban households. The no one at home rates in the urban and rural portions were roughly the same, but the refusal rates were 30 percent higher in NSRU rural areas than in NSRU urban areas.

Within SRU's built-up areas had a higher total non-response rate than fringe areas due to higher no one at home and temporarily absent components. Thus, it appears that people living in the core areas of cities tend to be more difficult to contact than people living in the fringe areas; the differences, however, were not large.

SRU apartments had a higher total non-response rate than any other area shown on Table 4.3. In fact, the total non-response rate in the SRU apartment sample was almost twice the rate in the SRU non-apartment sample (consisting of both built-up and fringe areas). The refusal, no one at home and temporarily absent components were also highest among apartments.

The no one at home rate was almost three times higher in the apartment sample than in the non-apartment sample. This large difference may be due to the different lifestyles of apartment and non-apartment dwellers. Apartment households usually consist of single persons or very small families who tend to be more mobile and difficult to find at home, while non-apartment households are more likely to contain larger families with children. Another problem with apartments is that interviewers often find
it difficult to gain entrance into apartment buildings.

Whereas the temporarily absent rate was usually twice as high for apartments as non-apartments, the difference was less noticeable during July and August than in the other ten months. This probably resulted from the fact that it is easier for single persons and families without children to take their vacations during the fall, winter and spring than it is for families with school-age children.

The refusal rate was almost always higher in the apartment sample than in the non-apartment sample, although the difference in the refusal rates between the two samples was not as great as the differences observed for the no one at home and temporarily absent rates. Recent results, however, indicate that the gap in the level of refusal rates between apartments and non-apartments is gradually widening, to the extent that the refusal rate in the apartment sample is now almost double the corresponding rate in the non-apartment sample.

The significance of examining non-response rates according to breakdowns such as SRU and NSRU is that this approach helps establish relationships among the various types of area in terms of the behaviour of non-response rates. For instance, the overall non-response rate is always expected to be higher in SRU's than in NSRU's, and any deviation from this relationship is considered unusual. The same holds true for the no one at home and refusal rates in SRU's and NSRU's. Another example is the two-to-one ratio of the total non-response rate in the apartment sample to the corresponding rate in the non-apartment sample. If non-response rates ever increase beyond average or expected levels, then knowledge of these relationships is very useful for the purpose of analyzing the situation and taking remedial action.

The average non-response rate at the Canada level was 5.4 in 1978. As already indicated, fluctuations occurred from month to month, from region to region and from one type of area to another type of area. It would also be expected, of course, that non-response rates varied amoung interviewers.

Many interviewers, in fact, achieved 100 percent response rates, while a few interviewers did no better than 75 percent. It is interesting to look at the distribution of interviewers according to the level of their total non-response rates as shown on Table 4.4.

## TABLE 4.4

Distribution of Interviewers According to Their Total Non-Response Rates
(Monthly average: 1978)

Total non-response rate (\%)

Number of
interviewers
$\qquad$
$0.0 \quad 159$
0.1 to 5.0
5.1 to 10.0
10.1 to 15.0
15.1 to 20.0
over 20.0
TOTAL

434
333
98
29
10
$\overline{1063}$

Percentage of total interviewers
$\qquad$
15.0
40.8
31.3
9.2
2.7
1.0
100.0

The data on Table 4.4 are based on all interviewers who enumerated assignments with at least 20 households and represent an average over the twelve months of 1978. The table indicates that 56 percent of interviewers achieved non-response rates of 5.0 or better.

The 13 percent of interviewers with non-response rates higher than 10.0 accounted for 31 percent or almost one-third of all non-responses. Therefore, it is clear that the majority of interviewers have been very successful in maintaining excellent non-response rates, and a large percentage of the total number of non-responses have actually resulted from a small group of interviewers. It is also interesting to observe that 59 percent
of interviewers did not record any "no one at home" non-responses and that 54 percent encountered no refusals. Furthermore, 71 percent and 70 percent of interviewers achieved no one at home rates and refusal rates respectively of 2.0 or better. These percentages are certainly impressive and reflect the success of the on-going training programs, monitoring and controls as well as interviewing techniques and procedures which are all aimed at maximizing response levels in the LFS.

## 5. TREATMENT OF NON-RESPONDENTS IN the data processing and estimatION STAGES

A very important consideration is how non-respondents are treated in the data processing and estimation stages of the LFS. In order to reflect the entire population more adequately it is necessary to impute missing information for non-respondents or to adjust the weights of the interviewed portion of the sample. Imputation and adjustment procedures are carried out in the following way [1,10].

In the case of refusal, no one at home and temporarily absent non-responses which responded in the previous month, records for these households are copied from the previous month with suitable transformations applied to certain items (for example, if a person in the previous month had been looking for a job for 6 weeks, then this information would be updated to 10 weeks for the current month). For estimation purposes these households are treated in the same way as responding households. Records are not carried forward for more than one consecutive month. In 1978, on the average, this method of imputation was carried out for approximately 31 percent of all non-respondents. This percentage ranged from 22 percent in December to 45 percent in July. Most other months about 30 percent of records for non-respondents were copied from the previous month. Throughout the year an average of 11 percent of refusal records, 33 percent of no one at home records and 52 percent of temporarily absent records were carried forward.

The remaining non-respondents for whom records cannot be carried forward from the previous month include non-responses due to road or weather conditions, circumstances within the household, no interviewer available, "no shows", rotate-in households which are non-responses, and households which have been non-responses for more than one month. The adjustment procedure used to compensate for these non-respondents increases the weight of the interviewed households when sample observations are inflated to produce the labour force estimates. Specifically, the weight of an interviewed household is increased by a balancing factor which is determined in the following manner. In NSRU's every sampled primary sampling unit (PSU) is divided into two balancing units (a rural part and an urban part), while every sub-unit in SRU's is treated as a balancing unit. For each balancing unit the balancing factor is calculated by dividing (a) the number of households which should have been interviewed by (b) the number of households which were interviewed or whose records were carried forward from the previous month. In special areas the balancing unit is the stratum and the balancing factors are calculated in the same way as they are in SRU's.

## 6. CHARACTERISTICS OF NON-RESPONDENTS

Although imputation and adjustment procedures are carried out in the data processing and estimation stages of the LFS in order to compensate for non-response and minimize non-response bias, very little is actually known regarding the composition and characteristics of non-response households. For non-respondents whose records cannot be carried forward from the previous month, the assumption is made that their charcteristics are the same as those of respondents. Recent studies, however, indicate that nonresponse households differ from interviewed households in characteristics such as household size and labour force status [11].

Table 6.1 summarizes demographic and labour force characteristics of responding and non-responding households in the LFS during the period from

January to December 1978. Results were obtaịned for 70 percent of all non-response households, 33 percent of refusal households, 79 percent of no one at home households and 88 percent of temporarily absent households using information collected for these households in months when they responded. The characteristics were averaged over all interviewed households and non-respondents for whom data were available. They were also averaged over the refusal, no one at home and temporarily absent components of non-response. All averages were weighted by using the inverse sampling rate according to province and type of area. Only civilian household members 15 years of age and over were included in the tabulations; members of the armed forces and individuals under 15 years old were excluded. Changes in household composition from one month to another (estimated to affect less than 2 percent of non-respondents) were not taken into consideration.

The results shown on Table 6.1 provide very interesting information on the influence and character of non-respondents in the LFS. Differences are indicated in the demographic and labour force charcateristics between respondents and non-respondents, especially in the case of no one at home and temporarily absent non-respondents, while refusal households appear to be very similar to interviewed households.

Because nearly 90 percent of temporarily absent households and 80 percent of no one at home households are represented in the tabulations, the results seem to be a very good characterization of these non-respondents. Compared to respondents, no one at home non-respondents belonged to smaller households, were younger, had higher unemployment rates and had substantially higher participation rates. It would seem, therefore, that their lifestyle made it difficult for interviewers to find them at home during survey week.

While the characteristics of interviewed households were very stable and only small variations occurred from month to month in the characteristics of no one at home households, the characteristics of temporarily absent households varied considerably throughout the year. Temporarily absent
households, like no one at home households, were much smaller than interviewed households. Further, temporarily absent non-respondents had lower particpation rates and were older than both respondents and other types of non-respondents. During the summer months when more temporarily absent non-responses were encountered because families with school-age children were on vacation, it is not surprising that temporarily absent non-residents belonged to larger households, were younger, had lower unemployment rates and had higher participation rates than at other times of the year.

TABLE 6.1
Demographic and Labour Force Characteristics of Interviewed and Non-Respondent Households
(Monthly average: 1978)

| Characteristic | Interviewed <br> households | Total non- <br> response | Refusal | No one <br> (1) Size of Household |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |

(4) Number of persons 'not in labour force ${ }^{11}$
0.83
0.66
0.79
0.42
0.77
(5) Unemployment Rate (\%)
8.3
8.9
9.5
9.7
8.5
(6) Participation rate
(\%)
62.5
63.6
$63.7 \quad 74.2$
55.8
(7) Age of head of household (years)
45.3
44.6
$45.7 \quad 38.6$
48.4
(8) Age of household members (years)
39.4
41.1
$39.9 \quad 36.3$
45.2

Note that the sum of characteristics (2), (3) and (4) equals characteristic (1).

The results for refusal households indicate that their demographic and labour force characteristics were very similar to those of respondents. However, these results should be interpreted with caution since only onethird of all refusals are represented. The tabulations only include refusals who responded at least once during the six months while they were in the LFS, while refusals who were non-respondents every month were not included.

## 7. CONCLUSION

Indications are that non-respondents differ from respondents in terms of characteristics such as household size and labour force status. It is clear, therefore, that the presence of non-response in the LFS leads to a bias in the final survey estimates and that this non-response bias can be expected to increase with higher non-response rates. Although imputation and adjustment procedures are carried out in the data processing and estimation stages of the survey to compensate for the non-response which does occur, the bias due to non-response is never completely eliminated. Consequently, it is very important that every effort must always be made in the field to maximize response levels.

In the LFS every reasonable effort is made to minimize non-response, and during 1977 and 1978 an average non-response rate of 5.4 was achieved. interviewer training programs and procedures emphasize how to gain the respondent's co-operation and conduct interviews in the most effective manner. Regular monitoring and controls help maintain non-response rates at satisfactory levels in all areas and ensure that interviewers are following correct and efficient procedures. In addition, programs such as refusal and post survey week follow-ups have been introduced to reduce non-response, and experimental studies directed toward the non-response problem are undertaken from time to time. With these procedures and programs it seems that the non-response rate cannot be brought lower than 4.0 in any one month and, due to seasonal fluctuations, a realistic objective is a yearly average of 5.0. This 5.0 non-response rate includes
a refusal rate of approximately 1.5 , a no one at home rate of 1.5 , a temporarily absent rate between 1.5 and 2.0 , and up to 0.5 for other reasons.

RESUME

Cet article comprend une description des techniques de l'interviewer et des méthodes utilisées pour minimiser la non-réponse, une brève description des méthodes de surveillance et de contrôle de la non-réponse et un examen de la façon dont on s'occupe des non-répondants lors du traitement et de l'estimation des données de l'enquête canadienne sur la population active. Il donne également les taux de non-réponse récents ainsi que des données sur les caractéristiques des non-répondants. L'auteur conclut qu'un taux de non-réponse annuel de 5\% approximativement est probablement le meilleur que l'on puisse réaliser dans le cadre de l'enquête sur la population active.

## REFERENCES

[1] "Methodology of the Canadian Labour Force Survey, (1976)", Statistics Canada, Catalogue 71-526 Occasional, October 1977.
[2] 'Labour Force Survey Interviewer's Manual', LFS-100, Statistics Canada, 1979.
[3] "Doorstep Diplomacy", a Statistics Canada booklet prepared by Field Division, 1976.
[4] Gower, A., 'LFS Proxy Rates', a Statistics Canada internal memorandum, 6 July 1978.
[5] Gower, A., "Post Survey Week Follow-Up", a Statistics Canada internal memorandum, 15 August 1978.
[6] Muirhead, R.C., Gower, A.R. and Newton, F.T., "The Telephone Experiment in the Canadian Labour Force Survey', Survey Methodology, Vol. 1, No. 2., December 1975, pp. 158-180.
[7] Gower, A.R., "The Response Incentives Experiment in the Canadian Labour Force Survey", Survey Methodology, Vol. 3, No. I, June 1977, pp. 84-103.
[8] 'Labour Force Survey Observation Instructions', LFS-480, Statistics Canada, November 1977.
[9] 'Labour Force Survey Re-interview Instructions", LFS-430, Statistics Canada, July 1979.
[10] Ashraf, A. and Macredie, I., "Edit and Imputation in the Labour Force Survey', presented at the Annual Meeting of the American Statistical Association, San Diego, California, U.S.A., August 1978.
[1I] Gower, A., 'Characteristics of Non-Respondents in the LFS'", a Statistics Canada internal memorandum, 18 April 1979.

## APPENDIX A

## Non-Interview Classifications

1. Non-response households are classified according to the following types:
(a) refusal - a responsible household member refused to provide any information about the household;
(b) no one at home when the interviewer called - the occupancy of the dwelling could not be contacted after several attempts, or someone was inside the dwelling but no one answered the door;
(c) temporarily absent - the household was absent for the entire survey week;
(d) no interviewer due to curcumstances within the household such as sickness, death, language problems or other unusual sitllations.
(e) no interview due to road or wather conditions;
(f) no interviewer was available; and
(g) survey forms arrived too late for processing (usually called "no shows"' ).
II. Vacant (or vacant-type) dwellings are classified according to:
(a) vacant dwellings - includes unoccupied dwellings, newly constructed dwellings ready for occupancy, and vacant trailer stalls in commercial trailer parks;
(b) vacant seasonal dwellings - includes seasonal dwellings such as summer cottages, ski chalets, and fishing or hunting lodges which were not occupied when the interviewer visited;
(c) dwellings under construction - includes any unoccupied dwelling which had a roof but was not ready for occupancy;and
(d) dwellings occupied by persons not eligible to be interviewed includes dwellings where all household members were (i) fulltime members of the Canadian Armed Forces, (ii) embassy, consular or armed forces personnel (including families) of foreign

> countries, (iii) residents of foreign countries on vacation or business in Canada, (iv) occupants of a seasonal dwelling during survey week with their usual place of residence elsewhere, or (v) inmates of religious or penal institutions, hospitals or nursing homes.
111. "Non-existent" dwellings include dwellings which were demolished, converted into business premises, moved (such as a mobile home), abandoned (unfit for habitation), or listed in error.

Definition of Non-Response Rates

The (total)non-response rate is defined as the total number of non-response households expressed as a percentage of the total number of sampled households (including both interviews and non-responses).

The refusal rate is defined as the number of refusal households expressed as a percentage of the total number of sampled households. The definitions are similar for the no one at home rate and the temporarily absent rate.

AN EMPIRICAL INVESTIGATION OF AN IMPROVED METHOD OF MEASURING CORRELATED RESPONSE VARIANCE ${ }^{1}$
A. MacLeod and K.P. Krótki ${ }^{2}$


#### Abstract

Two methods for estimating the correlated response variance of a survey estimator are studied by way of both theoretical comparison and empirical investigation. The variance of these estimators is discussed and the effects of outliers examined. Finally, an improved estimator is developed and evaluated.


## 1. INTRODUCTION

The correlated response variance (CRV) is one component of the overall or total variance of a survey estimator. Its definition is based upon considering a survey process as theoretically repeatable so that a series of independent trials of the survey exists. Response deviation is defined as the difference between an individual's response in the actual survey trial and the expected value of that individual's response, with the expectation taken over all conceptual survey trials. The CRV is then defined as a measure of the variability in response caused by the correlation of these response deviations for any pair of individuals within the given trial of the survey process. It has been shown empirically that, in a setting where enumerators help the respondents fill out the questionnaire, the CRV is an important component of total variance [3].

[^7]r

To estimate the CRV for various variables, two methods were employed. The first of these, hereafter referred to as the old method, is described in [2]. It is the method upon which Statistics Canada based its published estimates of total variance in the 1971 and 1976 Censuses. The other method, hereafter identified as the new method, is defined and derived in [4].

This paper investigates and compares these two methods, both in terms of the estimators themselves (Section 2) and the corresponding estimates that are yielded (Section 3). These estimates will be provided and discussed for data collected on both a $100 \%$ and sample basis. In addition, the variance of both estimators is presented in Section 4 and Section 5 discusses the effects of outliers. Section 6 consists of the development of an improved estimator with lower variance than either the old or new estimators. The paper concludes with Section 7 which contains summary comments, an overall evaluation of the improved method and some research ideas.

## 2. A THEORETICAL COMPARISON OF THE ESTIMATORS

Since a survey process cannot in practice be repeated a number of independent times, a technique of interpenetration of enumerators and enumeration areas (EAs) ${ }^{1}$ is used for both CRV estimators. In normal Census practice, one enumerator is assigned to handle one EA. With interpenetration, however, neighbouring EAs are paired and within these pairs the households in each EA are randomly split into two equal groups. One enumerator is then randomly assigned to one of these two groups in each EA of the pair while the second enumerator is given the other group from each EA of the pair.

[^8]For the o!d estimator, only a sample of interpenetrated EA pairs is needed. The new estimator, on the other hand, requires not only the sample of interpenetrated EA pairs but also a sample of non-interpenetrated EA pairs. In theory, all of the pairs to which the interpenetration technique is not applied could be used, instead of just a sample of such pairs. However, practical budgetary constraints dictate the use of a sample.

The correlated response variance for a Census estimator $X$ based on sample datal is given by the following expression:

$$
\begin{equation*}
\operatorname{CRV}(X)=\sum_{k=1}^{2 P} \quad N_{k}^{2} \frac{\left(n_{k}-1\right)}{n_{k}} \rho_{k} \sigma_{k}^{2} \tag{2.1}
\end{equation*}
$$

where $k$ is the EA index denoting the EA,
$P$ is the number of $E A$ pairs in Canada;
$N_{k}$ is the total number of households in the $k$ th EA and
$n_{k}$ is the number of sample households in the $k$ th EA ( $n_{k}$ is approximately equal to $N_{n} / 3$ ).

The entity $\rho_{k} \sigma_{k}^{2}$ is the key term in (2.1). It is defined as the average value of $E\left(x_{k h}-x_{k h}\right)\left(x_{k h}{ }^{-}-x_{k h} h^{\prime}\right)$, for pairs of households $\left(h, h^{\prime}, h \neq h^{\prime}\right)$ in the $k$ th EA that are enumerated by the same enumerator. $x_{k h}$ is the actual Census response and $X_{k h}$ is the mean of all conceptual responses so that ( $x_{k h}-x_{k h}$ ) is the response deviation in the hth household of the $k$ th EA. The expectation is taken over all conceptual responses, including enumerator assignments, for a given household. For $100 \%$ data, $n_{k}=N_{k}$ and so (2.1) reduces to

$$
\begin{equation*}
\operatorname{CRV}(x)=\sum_{k=1}^{2 P} N_{k}\left(N_{k}-1\right) \rho_{k} \sigma_{k}^{2} \tag{2.2}
\end{equation*}
$$

[^9]Dealing first with sample characteristics, the old estimator is given by:

$$
\begin{equation*}
\hat{C R V}(x)=\frac{P}{P_{1}} \sum_{k=1}^{2 p_{1}} N_{k}^{2} \frac{\left(n_{k}-1\right)}{n_{k}}\left(C_{k}-D_{k}\right) \tag{2.3}
\end{equation*}
$$

where $p_{l}$ is the number of interpenetrated $E A$ pairs in the sample, $C_{k}$ is the between enumerator sum of squares and $D_{k}$ is the within enumerator sum of squares. Expressions for $C_{k}$ and $D_{k}$ are given in [2]. Under certain conditions, $\left(C_{k}-D_{k}\right)$ is an unbiased estimator of $\rho_{k} \sigma_{k}^{2}$.

For $100 \%$ data, the expressions for $C_{k}$ and $D_{k}$ are modified slightly. The CRV estimator in this situation is given by

$$
\begin{equation*}
\hat{C R V}(X)=\frac{P_{1}}{P_{1}} \sum_{k=1}^{2 P_{1}} N_{k}\left(N_{k}-1\right)\left(C_{k}-D_{k}\right) . \tag{2.4}
\end{equation*}
$$

It is shown in [2] that under certain conditions both estimators are unbiased for correlated response variance; i.e., (2.3) is an unbiased estimator of (2.1) and (2.4) is an unbiased estimator of (2.2).

A weighted average of the $\left(C_{k}-D_{k}\right)$ expression over the interpenetrated EAs has been used as an estimator of CRV and total variance. For example, Statistics Canada employed such a procedure for the 1971 and 1976 Censuses. The methodology and results are discussed in [5] and [6]. The U.S. Bureau of the Census also used a weighted combination of $\left(C_{k}-D_{k}\right)$ 's based on 1960 data, as discussed in [7]. Bailey, Moore and Bailar [1] calculated averages of expressions equivalent to $\left(C_{k}-D_{k}\right)$ over all pairs of assignment areas in a national crime survey.
-

In the nev: method, the non-interpenetrated and interpenetrated samples are considered separately at the EA pair level. For each pair m, the quantity $A_{m}=\left(t_{m 1} t_{m 2}\right)^{2}$ is determined where $t_{m j}$ is the number of units in some category in the $j$ th EA of the mth pair $(j=1,2)$. The $A_{m}$ are next averaged separately for the non-interpenetrated pairs and for the interpenetrated pairs. Finally, a weighted difference of these two resulting averages is obtained. This difference is shown by Fellegi [4] to be a biased estimator of $\rho_{m} \sigma_{m}^{2}$, which is the average value of $E\left(x_{k h}-x_{k h}\right)\left(x_{k^{\prime} h^{\prime}}-x_{k \prime h^{\prime}}\right)$, ( $k=1,2$; if $k=k^{\prime}$, then $h \neq h^{\prime}$ ). The expectation is over all conceptual responses for a given household. The average is taken over all pairs of households in the mth EA pair which were interviewed by the same enumerator and over both enumerators in the EA pair. Fellegi [4] anticipates that the bias of this estimator is small.

Since estimates based on formulae (2.3) and (2.4) were already available, it was decided to obtain a form of the new estimator with the same expected value as the estimators in (2.3) and (2.4). This was somewhat complicated since the old estimator is based on calculations within individual EAs while the new estimator is based upon calculations within pairs of EAs. The problem was solved by looking at the expected values of the old and new estimators and by assuming that $\rho_{k} \sigma_{k}^{2}=\rho_{k}, \sigma_{k}^{2}=\rho_{m} \sigma_{m}^{2}$, where $k$ and $k^{\prime}$ are the two EAs of any pair m. For $100 \%$ data, (2.2) can be rewritten as:

$$
\operatorname{CRV}(x)=\sum_{m=1}^{P}\left[N_{m 1}\left(N_{m 1}-1\right)+N_{m 2}\left(N_{m 2}-1\right)\right] \rho_{m} \sigma_{m}^{2}
$$

where $m$ is the index denoting the pair under consideration and $N_{m i}$ is the total number of private households in the ith EA ( $1=1$ or 2) of the mth pair. For sample data, (2.l) can be similarly rewritten.

For sample data, the revised new estimator was found to be:

$$
\begin{align*}
& \hat{C R V}(X)=2 P\left\{\sum_{m=p_{1}}^{P} W_{m} \frac{\left\{\frac{N_{m 1}}{n_{m 1}} t_{m 1}-\frac{N_{m 2}}{n_{m 2}} t_{m 2}\right\}^{2}}{P-P_{1}}-\sum_{m=1}^{P} W_{m} \frac{\left.\frac{N_{m 1}}{n_{m 1}} t_{m 1}-\frac{N_{m 2}}{n_{m 2}} t_{m 2}\right\}^{2}}{p_{1}}\right\}  \tag{2.5}\\
& \text { where } \quad W_{m}=\frac{N_{m 1}^{2}\left(n_{m 1}-1\right)}{N_{m}^{2} n_{m l}}+\frac{N_{m 2}^{2}\left(n_{m 2}-1\right)}{N_{m}^{2} n_{m 2}}
\end{align*}
$$

For $100 \%$ data, the revised new estimator is:

$$
\begin{align*}
\hat{C R V}(X) & =2 P \frac{1}{P-P_{1}} \sum_{m=P, 1}^{P} \frac{N_{m}^{2}\left(N_{m-1}\right)-2 N_{m 1} N_{m 2}}{N_{m}^{2}}\left(t_{m 1}^{-t_{m 2}}\right)^{2} \\
& -\frac{1}{P_{1}} \sum_{m=1}^{P_{1}} \frac{N_{m}^{2}\left(N_{m-1}\right)-2 N_{m 1} N_{m 2}}{N_{m}^{2}}\left(t_{m 1}-t_{m 2}\right)^{2} . \tag{2.6}
\end{align*}
$$

For algebraic convenience, it is assumed in both (2.5) and (2.6), without loss of generality, that the first $P$, EA pairs are interpenetrated. All future uses of the term "new method" or "new estimator" shall refer to (2.5) for sample data and to (2.6) for $100 \%$ data.

## 3. AN EMPIRICAL COMPARISON OF THE ESTIMATORS

The old and new estimates are based on sample sizes of 375 interpenetrated EA pairs and 564 non-interpenetrated EA pairs. The samples were obtained through a two-stage process. In the first stage, 188 Census Commissioner Districts (CCDS) ${ }^{1}$ were selected by PPS (probability proportional to size) sampling, where the measure of size was the number of EAs per CCD. These CCDs were chosen independently from within sixteen strata formed by crossclassifying eight geographical regions of Canada with two types of enumeration methods (pick-up and mail-back). In the second stage, within each selected CCD, all EAs were first paired. Then two pairs were randomly selected for the interpenetrated sample, followed by three additional pairs for the noninterpenetrated sample. Selection of EAs within CCDs was done without replacement.

In order to be paired, two EAs had to be contiguous and of the same enumeration type and had to possess similar linguistic, agricultural, and density characteristics. In addition, each EA of a non-interpenetrated pair had to be a "full-load" EA. That is, it was assigned to an enumerator who was assigned no other EAs. These pairing criteria tend to make the two EAs in a pair alike so the assumption in Section 2 that $\rho_{k} \sigma_{k}^{2}=\rho_{k^{\prime}} \sigma_{k}^{2}=\rho_{m} \sigma_{m}^{2}$ for each pair $m$ is supported.

Using the old estimator, the correlated response variance was calculated for a large number of categories based on $100 \%$ data (age,sex, mother tongue, marital status) and sample data (education, labour force, mobility). A similar selection was made for use with the new estimator. The selection of variables and categories was made to cover a wide variety of results and subject-matter interests. Resulting CRV estimates for both estimators are given for $100 \%$ data in Table 1 and for sample data in Table 2. The corresponding coefficients of variation are also given. The coefficient of variation is defined as the square root of the absolute value of the CRV estimate for that category divided by the number of people in Canada belonging to the category.

1 A CCD on the average has 21 EAs.

For both $100 \%$ and sample data, the majority of old and new CRV estimates are negative. There are some patterns to this negativity. For example, almost all of the mobility status estimates are positive whereas most of the labour force estimates are negative. However, some interesting departures from these patterns can be observed. For example, all of the new mother tongue estimates listed are negative whereas only two of the five old estimates are negative. On the other hand, three of the four old estimates for age categories are negative but all of the new estimates for age are positive.

The estimates examined under the old method tend to show an inverse relationship between the coefficient of variation and the size of the category. This same tendency also exists to a lesser degree when looking at the new method results but there are more noticeable exceptions. For example, "mother tongue: French" and "mobility status: mover" have both fairly large populations and coefficients of variation while "unemployed" and "attending school full time" have both fairly small populations and coefficients of variation.

Table 1
Estimates of Correlated Response Variance Using the 01d and New Estimators With Corresponding Coefficients of Variation for Selected

Categories Based on 100\% Data

| Category | Old <br> CRV <br> Estimate | Coefficient <br> of <br> Variation | New <br> CRV <br> Estimate | Coefficient <br> of <br> Variation |
| :--- | ---: | :--- | ---: | :--- |
| Marital Status: <br> Married | $-322,041$ | $5.35 \times 10^{-5}$ | $25,951,715$ | $4.81 \times 10^{-4}$ |
| Mother Tongue: <br> English | $2,529,933$ | $1.13 \times 10^{-4}$ | $-49,979,989$ | $5.01 \times 10^{-4}$ |
| $\quad$ French | 841,832 | $1.56 \times 10^{-4}$ | $-51,555,178$ | $1.22 \times 10^{-3}$ |
| $\quad$ German | $-677,412$ | $1.73 \times 10^{-3}$ | $-2,466,910$ | $3.30 \times 10^{-3}$ |
| $\quad$ Italian | 206,108 | $9.34 \times 10^{-4}$ | $-7,485,919$ | $5.65 \times 10^{-3}$ |
| $\quad$ Ukrainian | $-336,429$ | $2.06 \times 10^{-3}$ | $-957,338$ | $3.47 \times 10^{-3}$ |
| Age: |  |  |  |  |
| 5 | $-97,258$ | $8.29 \times 10^{-4}$ | 93,087 | $8.11 \times 10^{-4}$ |
| $20-24$ | $-110,764$ | $1.56 \times 10^{-4}$ | $2,900,633$ | $7.98 \times 10^{-4}$ |
| $25-29$ | 325,555 | $2.86 \times 10^{-4}$ | $3,175,864$ | $8.94 \times 10^{-4}$ |
| 35 | $-147,409$ | $1.35 \times 10^{-3}$ | 15,733 | $4.41 \times 10^{-4}$ |
| Sex: |  |  |  |  |
| Male | $-255,958$ | $4.45 \times 10^{-5}$ | $-7,608,908$ | $2.43 \times 10^{-4}$ |

Table 2
Estimates of Correlated Response Variance Using the 01d and New Estimators With Corresponding Coefficients of Variation for Selected Categories Based On Sample Data

| Category | Old <br> CRV <br> Estimate | Coefficient of Variation | New CRV Estimate | ```Coefflcient of Variation``` |
| :---: | :---: | :---: | :---: | :---: |
| Highest Degree Received: High School rertificate | 2,344,591 | $4.56 \times 10^{-4}$ | 11,293,155 | $1.00 \times 10^{-3}$ |
| Eachelors | -899,953 | $1.20 \times 10^{-3}$ | 1,058,321 | $1.30 \times 10^{-3}$ |
| Masters | -1,928,039 | $8.24 \times 10^{-4}$ | -147,189 | $2.28 \times 10^{-3}$ |
| Non-university Cert. | -1,230,405 | $6.05 \times 10^{-4}$ | 485,724 | $3.80 \times 10^{-4}$. |
| Highest Grades Completed: <br> Less Than Grade 5 | 932,373 | $1.12 \times 10^{-3}$ | 1,398,493 | $1.37 \times 10^{-3}$ |
| Grades 5-8 | 1,648,121 | $3.60 \times 10^{-4}$ | -4,066,516 | $5.66 \times 10^{-4}$ |
| Grades 9-10 | 1,502,907 | $3.68 \times 10^{-4}$ | -3,689,625 | $5.76 \times 10^{-4}$ |
| Not Attending School | 4,197,176 | $1.43 \times 10^{-4}$ | 10,793,519 | $2.29 \times 10^{-4}$ |
| Attending School Full-Time | 495,245 | $4.20 \times 10^{-4}$ | 618,369 | $4.69 \times 10^{-4}$ |
| Unemployed On Temporary Layoff | $-2,171,495$ $-5,263,069$ | $2.11 \times 10^{-3}$ $2.24 \times 10^{-2}$ | $-756,508$ $-88,831$ | $1.24 \times 10^{-3}$ $2.91 \times 10^{-3}$ |
| Waiting to Start New Job | -2,761,460 | $1.39 \times 10^{-2}$ | $-88,831$ $-33,105$ | $1.53 \times 10^{-3}$ |
| Looked for Work | -2,275,514 | $3.15 \times 10^{-3}$ | -609,951 | $1.63 \times 10^{-3}$ |
| Employed | 3,357,282 | $1.92 \times 10^{-4}$ | 41,506,918 | $6.74 \times 10^{-4}$ |
| Not in Labour Force | 4,241,919 | $3.01 \times 10^{-4}$ | -966,135 | $1.46 \times 10^{-4}$ |
| ```Mobility Status (in last 5 years):``` |  |  |  |  |
| Non-Mover | 6,204,320 | $2.28 \times 10^{-4}$ | -122,571,179 | $1.01 \times 10^{-3}$ |
| Mover | 3,093,314 | $1.71 \times 10^{-4}$ | 218,080,472 | $1.43 \times 10^{-3}$ |
| Non-Migrant ${ }^{\text {a }}$ | 1,567,852 | $2.51 \times 10^{-4}$ | 26,889,868 | $1.04 \times 10^{-3}$ |

a. A non-migrant is a perison living in the same municipality as five years ago.

The old coefficients of variation also tend to be smaller in magnitude than the new ones. There are exceptions to this trend, especially when looking at sample data; the most notable ones deal with unemployment. Nevertheless, when all of the sample estimates in Table 2 are considered, the old coefficients are smaller than the new ones more often than not. One possible reason for this general phenomenon is that there is an unmeasured effect confounded with all new CRV estimates. The quantity $\left(t_{m 1}-t_{m 2}\right)^{2}$ in the new estimator is partially and undesirably affected by the fact that the numbers of households in the first and second EAs of pair $m$ are not equal to each other. However, with the old estimator, since each enumerator has exactly the same number of households to enumerate in each EA, there is no such undesired extraneous effect on $C_{k}$ and $D_{k}$ and hence on the overall estimate.

It is reported in [7] that relatively high response variances existed in the 1960 U.S. Census for data pertaining to low education levels and unemployment (among other subject-matter areas). An examination of Table 2 tends to support this view, insofar as the coefficients of variation for these categories are among the highest displayed in either Table 1 or 2. However, it should be noted that there are many other estimates which have coefficients of variation of a similar magnitude, such as mother tongue (other than English and French) and many of the 'highest degree received" and "mobility status" categories.

## 4. THE VARIANCE OF THE CRV ESTIMATORS

The next step in this study was to attempt to estimate the variance of the CRV estimators so that some measure of reliability could be attached to the CRV estimates. The jacknife technique was used. The sixteen strata described in Section 3 served as the strata for the estimation of the variance. Ten of the strata contained an odd number of CCDs necessitating the random deletion of one in each. In all, fifty pairs of EAs were deleted from the sample (twenty interpenetrated pairs and thirty non-interpenetrated pairs).

Estimated variances of the old and new CRV estimators are given for selected categories based on $100 \%$ and sample data in Tables 3 and 4, respectively. These tables also include the coefficient of variation which is defined as the square root of the variance of the CRV estimator divided by the absolute value of the CRV estimate.

The tables show that the variance estimates are almost all higher for the new CRV estimator than for the old one, with most of the exceptions occurring for estimates related to unemployment. However, a comparison of old and new coefficients of variation does not seem to reveal any clear patterns. Even for codes of a single variable, there is no obvious pattern for old coefficients. It is conceded that most of the labour force categories have larger new coefficients than old coefficients and that most of the mobility status categories have smaller new coefficients than old coefficients but there are still exceptions in both these cases. The most important observation then, to come from the estimates of the coefficient of variation, is their large magnitude. With many of them greater than one, there is a strong evidence of high variance in both the old and new CRV estimates.

Table 3
Variance Estimates (With Coefficients of Variation) of the 01d and New CRV Estimators for 100\% Data

| Category | Variance of 01d CRV | Coefficient of Variation (old) | Variance of New CRV | Coefficient of Variation (New) |
| :---: | :---: | :---: | :---: | :---: |
| Marital Status: Married | $4.355 \times 10^{11}$ | 2.049 | $5.424 \times 10^{14}$ | 0.897 |
| Mother Tongue: <br> English <br> French German Italian Ukrainian | $5.861 \times 1012$ $7.119 \times 1011$ $4.852 \times 1012$ $2.997 \times 1012$ $1.107 \times 10$ | 0.957 1.002 1.028 8.399 0.989 | $1.612 \times 1016$ $4.806 \times 1015$ $1.206 \times 1014$ $5.626 \times 1011$ $6.987 \times 10$ | 2.540 1.345 1.408 3.169 0.873 |
| ```Age:None``` | $1.361 \times 109$ $8.922 \times 10$ $1.482 \times 10$ $4.995 \times 10$ | 0.379 2.697 1.182 0.152 | $7.818 \times 10^{9} 12$ $6.716 \times 10^{12}$ $8.043 \times 10^{12}$ $2.365 \times 10^{9}$ | 0.950 0.893 0.893 3.091 |
| Sex: <br> Male | $3.267 \times 10^{11}$ | 2.233 | $2.256 \times 10^{15}$ | 6.242 |

Table 4
Variance Estimates (with Coefficients of Variation) of the Old and New CRV Estimators for Sample Data

| Category | Veriance of old CRV | Coefficient of Variation (OId) | Variance of New CRV | Coefficient of Variatio (New) |
| :---: | :---: | :---: | :---: | :---: |
| Highest Degree Received: High School Certificate | $1.088 \times 10^{12}$ | 0.445 | $2.473 \times 10^{13}$ | 0.440 |
| Bachelors | $4.388 \times 10^{11}$ | 0.736 | $5.801 \times 10^{11}$ | 0.720 |
| Masters | $8.694 \times 10^{11}$ | 0.484 | $9.277 \times 10^{9}$ | 0.654 |
| Non University Cert. | $2.424 \times 10^{11}$ | 0.400 | $7.545 \times 10^{12}$ | 5.655 |
| Highest Grade Completed: |  |  |  |  |
| Less Than Grade 5 | $3.535 \times 10^{12}$ | 2.017 | $4.193 \times 10^{12}$ | 1.464 |
| Grades 5-8 | $3.623 \times 10^{12}$ | 1.155 | $3.608 \times 10^{13}$ | 1.477 |
| Grades 9-10 | $1.834 \times 10^{12}$ | 0.901 | $4.435 \times 10^{13}$ | 1.805 |
| Not Attending School | $1.013 \times 10^{13}$ | 0.758 | $2.968 \times 10^{15}$ | 5.047 |
| Attending School Full Time | $5.706 \times 10^{11}$ | 1.525 | $4.688 \times 10^{12}$ | 3.501 |
| Unemployed: | $4.147 \times 10^{11}$ | 0.297 | $4.515 \times 10^{11}$ | 0.888 |
| On Temporary Layoff | $1.747 \times 10^{13}$ | 0.794 | $1.629 \times 10^{10}$ | 1.437 |
| Waiting to Start New Job | $8.143 \times 10^{11}$ | 0.327 | $1.764 \times 10^{9}$ | 1.269 |
| Looked for Work | $2.620 \times 10^{11}$ | 0.225 | $2.139 \times 10^{11}$ | 0.758 |
| Employed | $7.000 \times 10^{12}$ | 0.788 | $7.739 \times 10^{14}$ | 0.670 |
| Not In Labour Force | $1.357 \times 10^{13}$ | 0.868 | $2.544 \times 10^{14}$ | 16.012 |
| Mobility Status (in last |  |  |  |  |
| 5 years): <br> Non-Mover | $4.309 \times 10^{13}$ | 1.058 | $8.912 \times 10^{14}$ | 0.400 |
| Mover | $9.542 \times 10^{13}$ | 3.158 | $2.110 \times 10^{13}$ | 0.402 |
| Non-Migrant | $7.176 \times 10^{12}$ | 1.709 | $2.512 \times 10^{14}$ | 0.589 |

## 5. the effects of outliers on the crv estimates

In order to pinpoint some of the causes of the large coefficients of variation, a study was made of the effects of outliers on both the old and new CRV estimates. There was initially a strong suspicion that a few outliers were heavily influencing some of the new CRV estimates. The first problem in such a study was to develop a systematic criterion for defining an outlier. Whereas some outliers were rather obvious, other potential ones were notso evident. This meant that some decision rule was required.

For the new method, the values of $\left(t_{\mathrm{ml}}{ }^{-t_{m 2}}\right)$ were examined for the 564 noninterpenetrated pairs and for the 375 interpenetrated pairs (separately). Any pair for which ( $t_{m 1}{ }^{-t_{m 2}}$ ) was more than five standard deviations from the mean was declared an outlying pair.

For the old method, the quantities $C_{k}$ and $D_{k}$ were examined for each of the 750 interpenetrated EAs for given characteristics. It was observed that the values of $C_{k}$ and $D_{k}$ were more dispersed about their means than were the ( $\left.t_{m 1} t^{-t}\right)_{2}$ ) values. The criterion developed for the old method was that, if either $C_{k}$ or $D_{k}$ was more than eight standard deviations from the mean, then the EA $k$ would be called an outlying EA (for the particular characteristic).

The next stage was to recalculate the CRV estimates for some categories with the outliers deleted (and the sample size accordingly reduced). Looking at the old estimator first, an average of about four EAs was removed for each category but there was little observed change in the CRV estimates (Table 5). With the outliers gone, the CRVs were slightly closer to zero but the order (power of ten) of the estimates was the same with or without outliers for virtually all of the categories. The old estimator, therefore, did not appear to be very sensitive to outliers and so further investigations were omitted.

The removal of outlying $E A$ pairs did have a more noticeable effect on the CRV estimates. On the average, there were about three non-interpenetrated outlying pairs and about two interpenetrated outlying pairs per category, although these figures were slightly higher for mother tongue categories and slightly lower for age categories. Thus, perhaps not too surprisingly, the CRVs did not change very much for the age categories but for mother tongue, the tendency of the CRVs to approach zero was quite pronounced, as can be seen in Table 5 .

Another method for dealing with outliers with the new estimator was to replace in each outlying pair the value of $\left(t_{m l} t_{m 2}\right)$ by the appropriate mean, depending on whether the pair was non-interpenetrated or interpenetrated. This approach was also attempted here and gave very similar results to the approach of removing the outliers completely, as can be seen by comparing columns (ii) and (iii) of Table 5.

The variance of the re-estimated CRVs was determined for the categories in Table 5 using the new estimator. For age 6 and $4-6$, the variance actually increased slightly with the outlying pairs removed, but the other categories had lower variance with the outliers removed. These variance decreases were most noticeable for the smaller mother tongue groups.

In conclusion, although removing outliers had more effect on the new CRV estimator than on the old one in reducing both the CRV estimates and their variances, the effects were still not significant. With the exception of the small mother tongue groups, the new CRVs remained farther from zero than their old counterparts and the variance of the new CRVs remained larger than those of the old CRVs.

Table 5
CRV and Var(CRV) Estimates With and Without Outliers Using the New Estimator

| Categories | (i) | (ii) | (iii) | (iv) | (v) |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | CRV | CRV | CRV | $\operatorname{Var}$ (CRV) | $\operatorname{Var}$ (CRV) |
|  | Including | Deleting | Replacing | Including | Deleting |
|  | Outliers | Outliers | Outliers | Outliers | Outliers |

Mother Tongue:

English
French
German
Italian
Ukrainian
Age: 6
4-6
20-24
25-29
$-49,979,989-38,256,600$
$-38,371,439$
$1.612 \times 10^{16}$
$0.469 \times 10^{16}$
$-51,555,178-3,928,355$
$-3,067,913$
$4.806 \times 10^{15}$
$3.312 \times 10^{15}$
$198,1521.206 \times 10^{13}$
$0.018 \times 10^{13}$
202,452
541,320
$5.626 \times 10^{14}$
$0.023 \times 10^{14}$
32,289
$6.987 \times 10^{11}$
$1.666 \times 10^{11}$
$\begin{array}{ll}-2,466,910 & 202,452 \\ -7,485,919 & 555,478\end{array}$
-957,338 29,849
$-14,3654.264 \times 10^{9}$
$5.734 \times 10^{9}$
$-156,133 \quad 7.301 \times 10^{11}$
$8.940 \times 10^{11}$
184,275 -136,416
2,031,966
$6.716 \times 10^{12}$
$4.190 \times 10^{12}$
3,175,864
2,290,961
2,292,835
$8.043 \times 10^{12}$
$2.928 \times 10^{12}$

## 6. AN IMPROVED ESTIMATOR

It was suggested in Fellegi [4, p. 500] that a weighted combination of the old and new estimators would provide an improved estimator. Such an estimator was calculated with weights inversely proportional to the variances of the estimators. The results are presented in Table 6. The improved estimate was calculated only for categories based on sample data. The results are presented in two parts according to whether the old CRV was negative or positive. The table presents the improved CRV, its variance and the degree of improvement using the variance of the old CRV as a base. For the first seven categories both the old and improved CRVs are negative and the interpretation of the last column is not obvious. Nevertheless, it can be seen that in three of the seven cases major reduction has occurred and, in three of the remaining four cases, substantial reduction has occurred. For the categories with positive old CRVs, the mean reduction is $15.4 \%$ and for individual categories the reduction rate ranges from a low of $0.3 \%$ to a high of $81.9 \%$. In conclusion, based on the categories chosen and the data studied, it appears that substantial improvements in the reliability of the estimators can be achieved by resorting to the improved method.

Table 6
Estimates Using the Improved Method

| Category | Improved CRV | Variance of Improved CRV | Variance Decrease Over 0ld CRV (\%) |
| :---: | :---: | :---: | :---: |
| Highest degree received |  |  |  |
| Bachelors | -56,602 | $2.498 \times 10^{11}$ | 43.1 |
| Masters | -165,990 | $0.092 \times 10^{11}$ | 98.9 |
| Non-university certificate | -1,176,987 | $2.348 \times 10^{11}$ | 3.1 |
| Labour Force Status |  |  |  |
| Unemployed | -1,494,059 | $2.162 \times 10^{11}$ | 47.9 |
| On Temporary Layoff | -93,651 | $0.002 \times 10^{13}$ | 99.9 |
| Waiting to Start New Job | -39,002 | $0.017 \times 10^{11}$ | 99.8 |
| Looked for Work | -1,358,562 | $1.178 \times 10^{11}$ | 55.0 |
| Highest degree received High School certificate | 2,721,694 | $1.042 \times 10^{12}$ | 4.2 |
| Highest grade completed |  |  |  |
| Less than grade 5 | 1,145,589 | $1.918 \times 10^{12}$ | 45.7 |
| Grades 5-8 | 1,126,646 | $3.292 \times 10^{12}$ | 9.1 |
| Grades 9-10 | 1,296,708 | $1.761 \times 10^{12}$ | 4.0 |
| School Attendance |  |  |  |
| Not Attending School | 4,219,613 | $1.010 \times 10^{13}$ | 0.3 |
| Attending School Full Time | 508,604 | $5.087 \times 10^{11}$ | 10.6 |
| Labour Force Status |  |  |  |
| Employed | 3,699,256 | $6.937 \times 10^{12}$ | 0.9 |
| Not in the Labour Force | 3,978,183 | $1.288 \times 10^{13}$ | 5.1 |
| Mobility Status (in last 5 years) |  |  |  |
| Non-mover | 2,476,082 | $4.110 \times 10^{13}$ | 4.6 |
| Mover | 9,917,528 | $1.728 \times 10^{12}$ | 81.9 |
| Non-migrant ${ }^{\text {a }}$ | 2,271,132 | $7.176 \times 10^{12}$ | 2.7 |

a. A non-migrant is a person living in the same municipality as five years ago.

## 7. CONCLUSION

It has been shown how an improved estimator can be derived as the weighted combination of the old and new estimators. The estimator is improved in the sense that its variance is lower than that of either of the two original estimators. For certain variables the reduction in variance and consequent increase in stability is quite substantial. There are no extra field costs in determining the improved estimator. However, it should be pointed out that the improved estimator requires the noninterpenetrated EAs (be they a sample or the entire population) to be paired. It is our experience that this pairing is a time-consuming and tedious exercise especially if one insists on the same pairing criteria that were imposed on the interpenetrated EAs. In summary, the extra costs of the improved method are minimal and the stability of the improved estimator is considerably superior.

The importance of increase in stability should not be underestimated. One of the problems involved in calculating and disseminating estimates of correlated response variance is that the estimates are small and have large variance. This poses the problem of how to present to users of census data estimates of correlated response variance in a meaningful way. Developments which produce estimators with improved stability should be pursued.

The investigation of outliers in this paper was one such attempt to derive a more stable estimator. The fact that the situation changed very little when outliers were removed is disappointing, for, a priori, the outliers appeared to be a strong cause of instability of the estimates. Future research might investigate this problem more thoroughly. It has been suggested that one source of instability is the varying sizes of the EAs. To compensate for this phenomenon, a new model could be developed in which the mean for an EA, rather than the total count, would be used as the basic unit of analysis. This average could use the number of either persons or households in an EA as the base. Some theoretical development
of the formulae is called for here since the basic unit of observation would become the ratio of two random variables. The current model has a complexity which is ignored in the estimators. This complexity refers to the sample design which is not a simple random sample but involves stratification and more importantly, clustering of EAs in CCDs. Future work might consider incorporating these complexities into the calculation of CRV. Finally, this empirical investigation should be replicated on other data sets for a variety of variables in order to provide a more solid foundation on which to base evaluation of the improved method.

RESUME

Deux méthodes d'estimation de la variance de réponse correlée d'un estimateur d'enquête sont examinées à partir d'une comparaison théorique et d'une étude empirique. On examine ensuite la variance de ces estimateurs et les effets des observations détachées. Enfin, un estimateur amélioré est défini et évalué.

## REFERENCES

[1] Bailey, L., Moore, T.F., and Bailar, B.A., "An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample'', JASA, Vol. 73, 1978, pp. 16-23.
[2] Brackstone, G.J. and Hill, C.J., "The Estimation of Total Variance in the 1976 Census", Survey Methodology, Vol. 2, No. 2, 1976, pp. 195-208.
[3] Fellegi, I.P., "Response Variance and Its Estimation'", JASA, Vol. 59, 1964, pp. 1016-1041.
[4] Fellegi, I.P., ''An Improved Method of Estimating Correlated Response Variance' ${ }^{1}$, JASA, Vol. 69, 1974, pp. 496-501.
[5] Krótki, K.P. and Hill, C.J., "Estimation of Correlated Response Variance'", ASA, Proceedings of the Section on Survey Research Methods, 1978, pp. 609-14.
[6] Krótki, I.P., "1976 Census Parametric Evaluation: Measurement of Total Variance For the 1976 Census - Methodology and Results Report'", final report no. 18, Census Survey Methods Division, Statistics Canada, November 1978.
[7] U.S. Bureau of the Census. Evaluation and Research Program of the U.S. Censuses of Population and Housing (1960): Effects of Interviewers and Crew Leaders. Series ER 60, Number 7, Washington, D.C., 1968.

SELECTING A SAMPLE OF SIZE n WITH PPSWOR FROM A FINITE POPULATION

G.H. Choudhryl


#### Abstract

Let $U=\{1,2, \ldots . i, \ldots, N\}$ be a finite population of N identifiable units. A known "size measure" $\mathrm{x}_{\mathrm{i}}$ is associated with unit $i$; $i=1,2, \ldots, N$. A sampling procedure for selecting a sample of size $n(2<n<N)$ with probability proportional to size (PPS) and without replacement (WOR) from the population is proposed. With this method, the inclusion probability is proportional to size (IPPS) for each unit in the population.


## 1. INTRODUCTION

Yates and Grundy [4] have considered a selection procedure for selecting $n$ units with PPSWOR where at the first draw one unit is selected with probability proportional to size, and at the second draw one unit is selected with probability proportional to the size from the remaining units and so on. But with this procedure the overall probability of including a unit in the sample is not proportional to its size. Fellegi [1] has proposed a method whereby the probability of selecting a unit is proportional to its size at each of the $n$ successive draws. This is achieved by determining $n$ sets of selection probabilities called "Working Probabilities". Thus the inclusion probability is proportional to size for each of the $N$ units in the population. This method, however, requires cumbersome evaluation of "Working Probabilities" at each draw except the first one.

[^10]In the present method, selection is made by Yates and Grundy [4] scheme at the first ( $n-1$ ) draws and a set of 'Working Probabilities' is determined for selecting a unit at the nth draw, so that the overall probability of including a unit in the sample becomes proportional to the size for each unit in the population. Empirical results show that the efficiency of this method is the same as that of Fellegi's [1] method.
2. SAMPLING PROCEDURE

Define the "Normalized Sizes" $p_{i}$ proportional to $x_{i} ; i=1,2, \ldots, N$ such that $\sum_{i=1} p_{i}=1$, i.e.

$$
\begin{equation*}
p_{i}=\frac{x_{i}}{\sum_{i=1}^{N} x_{i}} \quad i=1,2, \ldots, N . \tag{2.1}
\end{equation*}
$$

Let $\Pi_{i}$ denote the probability that unit $i$ is in the sample; then it can be shown that

$$
\begin{equation*}
\sum_{i=1}^{N} \Pi_{i}=n \tag{2.2}
\end{equation*}
$$

It is required that the inclusion probability $\pi_{i}$ be proportional to $p_{i}$; $\mathrm{i}=1,2, \ldots, \mathrm{~N}$. This condition along with (2.1) and (2.2) imply

$$
\begin{equation*}
\pi_{i}=n p_{i} \quad i=1,2, \ldots, N \tag{2.3}
\end{equation*}
$$

At the first draw one unit is selected with probability proportional to size, and at the second draw one unit is selected with probability proportional to the size from the remaining ones and so on up to $(n-1)$ th draw. The probability of $i$, th unit being selected at the first draw is $p_{i_{1}}$, the conditional probability of $i_{2}$ th unit being selected at the second draw (given that $i$, th unit was already selected at the first draw) is equal to

etc. The conditional probability of $i_{n-1}$ th unit being selected at $(n-1)$ th draw (given that the units $i_{1}, i_{2}, \ldots, i_{n-2}$ were previously selected) is equal to

$$
\frac{p_{i_{n-1}}}{1-p_{i_{1}}-p_{i_{2}} \cdots-p_{i_{n-2}}}
$$

Let $q_{i} ; \mathbf{i}=1,2, \ldots, N$ be the set of "Working Probabilities" for selection at the nth draw, then the conditional probability of $i_{n}$ th unit being selected at the $n$th draw (given that $i_{1}, i_{2}, \ldots i_{n-1}$ were previously selected) is equal to


Then the overall probability, $\delta_{i}(k)$, of selecting $i$ th unit at the $k$ th draw is

$$
\begin{align*}
& \delta_{i}(k)=\sum_{(k-1, i)} p_{i_{1}} \times \frac{p_{i_{2}}}{1-p_{i_{1}}} \times \frac{p_{i_{3}}}{1-p_{i_{1}}-p_{i_{2}}} \ldots \times \frac{p_{i_{k-1}}}{1-p_{i_{1}}-p_{i_{2}}-\cdots-p_{i_{k-2}}} \times \\
& \frac{p_{i}}{1-p_{i_{1}}{ }^{-p_{i_{2}}} \cdots^{-p_{i_{k-1}}}} \quad k=1,2, \ldots, n-1 \tag{2.4}
\end{align*}
$$

and

$$
\begin{align*}
\delta_{i}(n)= & \sum_{(n-1, i)} p_{i_{1}} \times \frac{p_{i_{2}}}{1-p_{i_{1}}} \times \frac{p_{i_{3}}}{1-p_{i_{1}}-p_{i_{2}}} \times \ldots \times \frac{p_{i_{n-1}}}{1-p_{i_{1}}-p_{i_{2}} \cdots-p_{i_{n-2}}}
\end{align*} \frac{q_{i_{1}}}{1-q_{i_{1}}{ }^{-q_{i_{2}} \cdots{ }^{-q_{i_{n-1}}}}} \times
$$

where $\Sigma$ (as in Fellegi's [1] paper) denotes the summation over all possible ordered $(k-1)$ tuples of $\left(i_{1}, i_{2}, \ldots, i_{k-1}\right)$ such that $i_{1}, i_{2}, \ldots, i_{k-1}$ are different integers between 1 and $N$, and none of them is equal to $i$. Then $\Pi_{i}$, the probability that the $i$ th unit is in the sample, is given by

$$
\begin{aligned}
\pi_{i} & =\sum_{k=1}^{n} \delta_{i}(k) \\
& =\sum_{k=1}^{n-1} \delta_{i}(k)+\delta_{i}(n)
\end{aligned}
$$

where $\delta_{i}(k)$ for $k=1,2, \ldots, n-1$ is given by (2.4) and $\delta_{i}(n)$ is given by (2.5). In the expression for $\delta_{i}(n), q_{i} ; i=1,2, \ldots, N$ must be determined so that the condition $\Pi_{i}=n p_{i} ; i=1,2, \ldots N$ is satisfied, i.e.

$$
\begin{aligned}
n p_{i}= & \sum_{k=1}^{n-1} \delta_{i}(k)+\sum_{(n-1, i)} p_{i_{1}} \times \frac{p_{i_{2}}}{1-p_{i_{1}}} \times \ldots \times \frac{p_{i_{n-1}}}{1-p_{i_{1}}-p_{i_{2}} \cdots-p_{i_{n-2}}} \times \\
& \frac{q_{i}}{1-q_{i_{1}}-q_{i_{2}} \cdots-q_{i_{n-1}}}
\end{aligned}
$$

$$
\begin{gather*}
q_{i}=\frac{n p_{i}-\sum_{k=1}^{n-1} \delta_{i}(k)}{(n-1, i)} \frac{p_{i_{1}} p_{i_{2}} \cdots p_{i_{n-1}}}{\left(1-p_{i_{1}}\right)\left(1-p_{i_{2}}\right) \cdots\left(1-p_{i_{1}} p_{i_{2}} \cdots-p_{i_{n-2}}\right)\left(1-q_{i_{1}}-q_{i_{2}} \cdots-q_{i_{n-1}}\right)} \\
i=1,2, \ldots N .
\end{gather*}
$$

The set of 'Working Probabilities' $q_{i} ; i=1,2, \ldots N$ can be obtained by solving the set of simultaneous non-linear equations given in (2.6), by iterative procedure where the initial value for $q_{i}$ can be taken as $p_{i} ; i=1,2, \ldots, N$.

For $n=2$, the method is the same as the one given by Fellegi [1], but for $n \geq 3, \pi_{i}$ can be made equal to $n p_{i}$ for all $i$ by evaluating only one set of Working Probabilities instead of ( $n-1$ ) sets of Working Probabilities as in Fellegi's method.

## 3. CALCULATION OF $\pi_{i j}$

The joint probability of including both the units $i$ and $j$ in the sample; $\Pi_{i j}, i=1,2, \ldots N-1 ; j=i+1, \ldots N$ can be calculated as follows:

Let $\delta_{i j}(k, \ell)$ denote the probability that the unit $i$ was selected at $k$ th draw and the unit $j$ was selected at the $\ell$ th draw, where $k<\ell$. Then $\delta_{i j}(k, l)$ is given by

$$
\begin{aligned}
\delta_{i j}(k, \ell)= & \underset{(\ell-2, i, j)}{\Sigma} p_{i_{1}} \times \frac{p_{i_{1}}}{1-p_{i_{1}}} \times \frac{p_{i_{3}}}{1-p_{i_{1}}{ }^{-p_{i_{2}}}} \times \ldots \times \frac{p_{i_{k-1}}}{1-p_{i_{1}}-p_{i_{2}} \cdots-p_{i_{k-2}}} \times \\
& \frac{p_{i_{i}}^{1-p_{i_{1}}{ }^{-p_{i_{2}}} \cdots-p_{i_{k-1}}} \times \frac{p_{i_{k+1}}}{1-p_{i_{1}}{ }^{-p_{i_{2}} \cdots-p_{i_{k-1}}-p_{i_{i}}}} \times}{}
\end{aligned}
$$

$$
\begin{aligned}
& \ldots \times \frac{p_{j}}{1-p_{i_{1}}{ }^{-p_{i_{2}}} \cdots^{-p_{i_{k-1}}}{ }^{-p_{i}-p_{i_{k+1}}} \cdots^{-p_{i_{\ell-1}}}} . \\
& \mathrm{k}=1,2, \ldots, \mathrm{n}-2 \\
& \ell=k+1, \ldots, n-1
\end{aligned}
$$

and $\delta_{i j}(k, n)$ is given by

$$
\begin{aligned}
& \delta_{i j}(k, n)=\sum_{(n-2 ; i, j)} p_{i_{1}} \times \frac{p_{i_{2}}}{1-p_{i_{1}}} \times \frac{p_{i_{3}}}{1-p_{i_{1}}-p_{i_{2}}} \times \ldots \times \frac{p_{i_{k-1}}}{1-p_{i_{1}}-p_{i_{2}} \ldots-p_{i_{k-2}}} \times \\
& \frac{p_{i}}{1-p_{i_{1}}{ }^{-p_{i_{2}}} \cdots{ }^{-p_{i_{k-1}}}} \times \frac{p_{i_{k+1}}}{1-p_{i_{1}}{ }^{-p_{i_{2}}}{ }^{\cdots-p_{i_{k-1}}-p_{i}}} \times \\
& \cdots \times \frac{p_{i_{n-1}}}{1-p_{i_{1}}{ }^{-p_{i_{2}} \cdots-p_{i_{k-1}}-p_{i}-p_{i_{k+1}} \cdots-p_{i_{n-2}}}} \times \\
& \ldots \times \frac{q_{j}}{1-q_{i_{1}}{ }^{-q_{i_{2}}} \cdots{ }^{-q_{i_{k-1}}}{ }^{-q_{i}-q_{i_{k+1}} \cdots-q_{i_{n-1}}}} \\
& k=1,2, \ldots n-1 \text {, }
\end{aligned}
$$

where $\underset{(\ell-2 ; i, j)}{\Sigma}$ denotes the summation over all possible ordered ( $\ell-2)$-tuples of ( $i_{1}, i_{2}, \ldots i_{k-1}, i_{k+1}, \ldots i_{\ell-1}$ ) such that $i_{1}, i_{2}, \ldots i_{k-1}, i_{k+1}$, $\ldots, i_{\ell-1}$ are different integers between 1 and $N$ and none of them is
equal tu $i$ or $j$. Then $\Pi_{i j}$, the probability that the units $i$ and $j$ are both in the sample is given by

$$
\begin{aligned}
\pi_{i j}= & \sum_{k=1}^{n-2} \sum_{\ell=k+1}^{n-1}\left[\delta_{i j}(k, \ell)+\delta_{j i}(k, \ell)\right]+\sum_{k=1}^{n-1}\left[\delta_{i j}(k, n)+\delta_{j i}(k, n)\right] \\
= & \sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n}\left[\delta_{i j}(k, \ell)+\delta_{j i}(k, \ell)\right] \\
& i=1,2, \ldots, N-1 \\
& j=i+1, \ldots, N .
\end{aligned}
$$

## 4. ROTATING SAMPLE

Suppose that in a stratum we want to conduct the survey in $m_{0}$ first stage units (f.s.u.'s) for some specified period of time. This period could be fixed pre-specified or may occur as and when one or more f.s.u.'s get exhausted. In order to accommodate such a rotation scheme, we initially select $n=\sum_{t=0}^{T} m_{t} \quad$ f.s.u.'s where $m_{0}$ is as defined above, and $m_{t}$ are the number of f.s.u.'s needed for rotation at the time period $t ; t=1,2,3, \ldots, T$. At time $t=0$, take a simple random sample of $m_{0}$ out of $n$ f.s.u.'s and for the purpose of rotation at time period $t$, a simple random sample of $m_{t}$ units is selected from the remaining $n-\left(m_{0}+m_{1}+\ldots+m_{t-1}\right)$ out of the $n$ initially selected units. Since the original probability of selecting unit $i$ is $\Pi_{i}=n p_{i}$, and at any given time the conditional probability of a unit being selected (given that it was originally selected in the first stage of sampling) is equal to $m_{0} / n$, therefore the unconditional probability, $\pi_{i}^{\prime}$, that the unit $i$ is in the final sample is

$$
\begin{aligned}
\Pi_{i}^{\prime} & =\Pi_{i} \times \frac{m_{0}}{n_{m_{0}}} \\
& =n p_{i} \times \frac{m_{0}}{n} \\
& =m_{0} p_{i} \quad i=1,2, \ldots, N
\end{aligned}
$$

as required.

Similarly, the unconditional probability, $\Pi_{i j}^{\prime}$, that the unit $i$ and $j$ are both in the sample is given by

$$
\begin{aligned}
\Pi_{i j}^{\prime}= & \frac{m_{0}\left(m_{0}-1\right)}{n(n-i)} \pi_{i j} \\
& i=1,2, \ldots, N-1 \\
& j=i+1, i+2, \ldots, N .
\end{aligned}
$$

where $\pi_{i j}$, is given by (3.1).
In Fellegi's [l] scheme, since the probability of selecting a unit is proportional to the size at each of the successive draws, therefore for a rotating sample, additional f.s.u.'s are selected at the time of rotation.

## 5. ESTIMATOR FOR THE POPULATION TOTAL AND ITS VARIANCE

Let $s=\left\{i_{1}, i_{2}, \ldots, i_{n}\right\}$ denote the $n$ sampled units and $y_{i}$ be the value of study variable $y$ for unit $i$ in the population; $i=1,2, \ldots$, $N$. The unknowr population total $Y=\sum_{i=1}^{N} y_{i}$ is to be estimated from the observations $y_{i}$ for $i \varepsilon s$. Horvitz and Thompson [3] estimator for the population total $Y$ is

$$
\begin{equation*}
\hat{Y}=\frac{1}{n} \sum_{i \varepsilon_{s}} \frac{y_{i}}{P_{i}} \tag{5.1}
\end{equation*}
$$

and the variance of $\hat{Y}$ as given by Yates and Grundy [4] is

$$
\begin{equation*}
V(\hat{y})=\frac{1}{n^{2}} \sum_{i<j} \sum\left(\Pi_{i} \Pi_{j}-\pi_{i j}\right)\left(\frac{y_{i}}{p_{i}}-\frac{y_{j}}{p_{j}}\right)^{2} \tag{5.2}
\end{equation*}
$$

where $\Pi_{i}$ is the probability that the unit $i$ is in the sample and $\Pi_{i j}$ is the probability that both the units $i$ and $j$ are in the sample.

An unbiased estimator of $V(\hat{Y})$ is

$$
\begin{equation*}
\hat{V}=\frac{1}{n^{2}} \underset{(i, j \varepsilon s)}{\sum \sum}\left(\frac{\Pi_{i} \Pi_{j}-\Pi_{i j}}{\Pi_{i j}}\right)\left(\frac{y_{i}}{p_{i}}-\frac{y_{j}}{p_{j}}\right)^{2} \tag{5.3}
\end{equation*}
$$

In the following section, the results of an empirical study using data from Fellegi [1] and Gray [2] have been presented. The $\pi_{i j}$ values have been tabulated for Fellegi's [1] method and the proposed method for samples of size 3 and 4 . The non-negativity of the variance estimator can be checked from the tabulated $\Pi_{i j}$ values, i.e. $\pi_{i j}<\pi_{i} \Pi_{j}$ for all $(i, j)$ pairs in the population. Variances of $\hat{Y}$ and variances of $\hat{V}$ have been computed for the two methods for samples of size 3 and 4 using the two sets of data, i.e. Fellegi [1] and Gray [2].

## 6. EMPIRICAL RESULTS

### 6.1 Example 1 Data from Fellegi [1].

The population consists of six primary sampling units. The $p_{i}$ and $y_{i}$ values are given in Table (6.1.1).

Table (6.1.1): $p_{i}$ and $y_{i}$ Values for Example (1).

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $P_{i}$ | 0.10 | 0.14 | 0.17 | 0.18 | 0.19 | 0.22 |
| $y_{i}$ | 0.60 | 0.98 | 1.53 | 2.16 | 2.85 | 4.18 |
|  | $\sum_{i=1}^{N} \quad y_{i}=12.30$ |  |  |  |  |  |

The 'Working probabilities" for selecting a sample of size 3 are given in Table (6.1.2) for the two schemes. The "Number of iterations" column is the number of iterations it took to obtain the convergence* to the solution. Note that $p_{i}(k) ; i=1,2, \ldots, N$ are the 'Working probabilities" at the kth draw; $k=1,2, \ldots, n$ for Fellegi's [1] scheme, where $p_{i}(1)=p_{i} ; i=1,2, \ldots, N$. Further $q_{i} ; i=1,2$, $\ldots, N$ are the "Working probabilities" at the nth draw for the proposed scheme. Recall that $p_{i} ; i=1,2, \ldots, N$ are the 'Working probabilities' at each of the first $n-1$ draws for the proposed scheme.

[^11]Table (6.1.2): 'Working Probabilities" for Selecting 3 units.

| No. of <br> iterations | $\mathbf{i}$ | 1 | 2 | 3 | 4 | 5 | 6 |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| - | $p_{i}$ | 0.100000 | 0.140000 | 0.170000 | 0.180000 | 0.190000 | 0.220000 |
| - | $p_{i}(1)$ | 0.100000 | 0.140000 | 0.170000 | 0.180000 | 0.190000 | 0.220000 |
| 6 | $p_{i}(2)$ | 0.090190 | 0.132410 | 0.167577 | 0.180157 | 0.193252 | 0.236414 |
| 9 | $p_{i}(3)$ | 0.076367 | 0.119154 | 0.160684 | 0.177404 | 0.196167 | 0.270224 |
| 9 | $q_{i}$ | 0.068868 | 0.113368 | 0.158820 | 0.177494 | 0.198613 | 0.282837 |

From the above table we notice that for Fellegi's [1] scheme, the 'Working probabilities' for units 1, 2, and 3, which are the three smallest units in the population, decrease during successive draws, whereas for units 4, 5, and 6, which are the three largest units in the population, the 'Working probabilities' increase during successive draws. Since for the proposed scheme, the 'Working probabilities" at the first $n-1$ draws remain unchanged; therefore at the nth draw (3rd draw in this case) the 'Working probabilities' for units 1, 2, and 3 i.e. the three smallest units, are smaller than the corresponding 'Working probabilities" for Fellegi's [1] scheme, and for units 4, 5, and 6 i.e. three largest units, the 'Working probabilities" are larger than the corresponding 'Working probabilities' for Fellegi's [1] scheme.

The following table exhibits the values of $\pi_{i j}$ for the two schemes for sample size 3. The values above the main diagonal correspond to Fellegi's [1] scheme and those under the main diagonal correspond to the proposed scheme.

Table (6.1.3): $\pi_{i j}$ Values for Sample size 3.

| i j | 1 | 2 | 3 | 4 | 5 | 6 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | x | 0.086165 | 0.109898 | 0.118557 | 0.127675 | 0.157705 |
| 2 | 0.086163 | x | 0.161733 | 0.174315 | 0.187513 | 0.230274 |
| 3 | 0.109902 | 0.161742 | x | 0.221121 | 0.237550 | 0.289698 |
| 4 | 0.118557 | 0.174316 | 0.221111 | x | 0.255473 | 0.310534 |
| 5 | 0.127674 | 0.187510 | 0.237547 | 0.255479 | $x$ | 0.331790 |
| 6 | 0.157704 | 0.230269 | 0.289699 | 0.310538 | 0.331791 | $x$ |

From the above table it is seen that the $\Pi_{i j}$ values for the two schemes do not differ up to 4 decimals, and since the variance is a function of $\pi_{i j}$ values, therefore, the two schemes will be equally efficient as seen from the following table.

Table (6.1.4): Variance of $\hat{\gamma}$ and Variance of $\hat{V}$ for the Two Schemes for Sample Size 3.

| Selection Scheme | $V(\hat{Y})$ | $V(\hat{V})$ |
| :--- | :---: | :---: |
| Fellegi's Scheme | 3.8258 | 4.6166 |
| Proposed Scheme | 3.8259 | 4.6171 |

Similarly for a sample of size 4 , the following tables give the 'Working Probabilities', the $\pi_{i j}$ values, and variance of $\hat{Y}$ and variance of $\hat{V}$ for the two schemes.

Table (6.1.5): 'Working Probabilities" for Selecting 4 units.

| No. of <br> iterations | $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| - | $p_{i}$ | 0.100000 | 0.140000 | 0.170000 | 0.180000 | 0.190000 | 0.220000 |
| - | $p_{i}(1)$ | 0.100000 | 0.140000 | 0.170000 | 0.180000 | 0.190000 | 0.220000 |
| 6 | $p_{i}(2)$ | 0.090190 | 0.132410 | 0.167577 | 0.180157 | 0.193252 | 0.236414 |
| 9 | $p_{i}(3)$ | 0.076367 | 0.119154 | 0.160684 | 0.177404 | 0.196167 | 0.270224 |
| 16 | $p_{i}(4)$ | 0.051667 | 0.086649 | 0.130509 | 0.153692 | 0.184616 | 0.392867 |
| 17 | $q_{i}$ | 0.033017 | 0.070222 | 0.121849 | 0.150012 | 0.187892 | 0.437008 |

Table (6.1.6): $\pi_{i j}$ Values for Sample size 4.

| $i j$ | 1 | 2 | 3 | 4 | 5 | 6 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $x$ | 0.166245 | 0.216290 | 0.235994 | 0.256805 | 0.324666 |
| 2 | 0.167197 | $x$ | 0.320554 | 0.348706 | 0.377519 | 0.466976 |
| 3 | 0.216261 | 0.320123 | $x$ | 0.445971 | 0.478668 | 0.578517 |
| 4 | 0.235761 | 0.348406 | 0.446103 | $x$ | 0.513248 | 0.616081 |
| 5 | 0.256432 | 0.377327 | 0.478869 | 0.513522 | $x$ | 0.653760 |
| 6 | 0.324349 | 0.466948 | 0.578645 | 0.616208 | 0.653850 | $x$ |

Table (6.1.7): Variance of $\hat{y}$ and Variance of $\hat{V}$ for the Two Schemes for Sample Size 4.

| Selection Scheme | $V(\hat{Y})$ | $V(\hat{V})$ |
| :--- | :---: | :---: |
| Fellegi's Scheme | 1.5323 | 0.4672 |
| Proposed Scheme | 1.5269 | 0.4553 |

6.2 Example 2: Data from Gray [2].

The population in this example is a stratum in Nova Scotia in LFS with dummy characteristics. The stratum consists of ten primary sampling units. The $p_{i}$ and $y_{i}$ values are given in Table (6.2.1).

Table (6.2.1): $p_{i}$ and $y_{i}$ Values for Example (2).

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $p_{i}$ | 0.0957 | 0.1043 | 0.1043 | 0.1006 | 0.0896 | 0.0881 | 0.0986 | 0.1055 | 0.1149 | 0.0984 |
| $y_{i}$ | 10.06 | 10.35 | 10.38 | 9.57 | 9.30 | 8.96 | 10.00 | 10.50 | 11.33 | 9.55 |

$$
y=\sum_{i=1}^{N} y_{i}=100.00
$$

As in example (1), the 'Working probabilities', the $\pi_{i j}$ values, and variance of $\hat{Y}$ and variance of $\hat{V}$ for the two schemes for samples of size 3 and 4 were computed from the data in table (6.2.1) above. The behaviour of the "Working probabilities" was similar to those in example (1), and the $\Pi_{i j}$ values for the two schemes were identical to 5 decimals both for samples of size 3 and 4 . Due to space tables of "Working probabilities" and those of $\Pi_{i j}$ values are not given. In the following table "Number of iterations" required to obtain the "Working probabilities", and variance of $\hat{y}$ and variance of $\hat{V}$ for samples of size 3 and 4 are given.

Table (6.2.2) "Number of iterations" to obtain "Working Probabilities" and Variance of $\hat{Y}$ and Variance of $\hat{V}$ for the two Schemes for samples of size 3 and 4 .

| Selection Scheme | Sample Size | No. of iterations <br> at draw |  |  | $V(\hat{Y})$ | $V(\hat{V})$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Fellegi's Scheme | 3 | - | 5 | 6 |  |  |
| Proposed Scheme | 3 | - | - | 6 | 2.0509 | 2.7418 |
| Fellegi's Scheme | 4 | - | 5 | 6 | 7 | 1.3287 |
| Proposed Scheme | 4 | - | - | - | 7 | 1.3287 |

For the two numerical examples in this study, it is observed that the $\Pi_{i j}$ values for the two selection schemes, ie., Fellegi's [1] scheme and the proposed scheme are almost identical. Although it seems that the underlying design for the two selection schemes is the same,
-
choice between the two should be made on operational convenience. Since Fellegi's scheme requires the evaluation of 'Working Probabilities" at each draw except the first one, whereas the scheme proposed in this paper requires the evaluation of "Working Probabilities" at the last draw only, this results in considerable reduction in computing.

## ACKNOWLEDGEMENT

The author wishes to thank Mr. G.B. Gray and Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, and the referee for helpful comments.

RESUME

Soit $U=\{1,2,3, \ldots, i, \ldots, N\}$ une population finie de $N$ unités indentifiables. Une "mesure de la taille" connue $x_{i}$ est associée à l'unité $i, i=1,2, \ldots, N$. L'auteur propose une méthode d'échantillonnage pour choisir une taille d'échantillon $n(2<n<N)$ dont la probabilité est proportionnelle à la taille et sans remise. De cette façon la probabilité d'inclusion est proportionnelle à la taille pour chaque unité de la population.

## REFERENCES

[1] Fellegi, I.P., "Sampling With Varying Probabilities Without Replacement: Rotating and Non-Rotating Samples', Journal of the American Statistical Association, Vol. 58 (1963), pp 183-201.
[2] Gray, G.B., "Variance Components and Variance Function", Proceedings of 'Statistics 71' Canada, pp 119-26.
[3] Horvitz, D.G. and Thompson, D.J., "A generalization of Sampling Without Replacement from a Finite Universe' ${ }^{\prime \prime}$, Journal of the American Statistical Association, Vol. 47 (1952), pp 663-85.
[4] Yates, F. and Grundy, P.M., "Selection Without Replacement from Within Strata With Probability Proportional to Size'", Journal of the Royal Statistical Society, Series B, Vol. 15 (1953), pp 235-61.

# ON THE INCLUSION OF LARGE UNITS <br> IN SIMPLE RANDOM SAMPLING ${ }^{1}$ 

## M.A. Hidiroglou ${ }^{2}$


#### Abstract

Approximate cutoff rules for stratifying a population into a take-all and take-some universe have been given by Dalenius (1950) and Glasser (1962). They expressed the cutoff value (that value which delineates the boundary of the take-all and take-some) as a function of the mean, the sampling weight and the population variance. Their cutoff values were derived on the assumption that a single random sample of size $n$ was to be drawn without replacement from the population of size $N$.

In the present context, exact and approximate cutoff rules have been worked out for a similar situation. Rather than providing the sample size of the sample, the precision (coefficient of variation) is given. Note that in many sampling situations, the sampler is given a set of objectives in terms of reliability and not sample size. The result is particularly useful for determining the take-all - take-some boundary for samples drawn from a known population. The procedure is also extended to ratio estimation.


## I. INTRODUCTION

The stratification of a population is a technique often used in survey sampling. This technique may produce gains in precision in estimating characteristics of the population. The problem considered in this paper is the stratification of a population into two strata, the take-all and take-some strata. The take-all stratum contains some of the largest units in the population while the take-some stratum contains the remaining units. The take-all stratum units are surveyed entirely while a simple random sample is drawn from the take-some stratum. This type of stratification is

[^12]particularly useful for populations whose distribution exhibits a marked positive skewness, with a few large units and many small units. Failure to recognize that such highly skewed populations should be stratified in the above manner may result in over-estimation of the population characteristics. This last point has been studied by Hidiroglou and Srinath (1977).

Approximate cutoff rules for stratifying a population into take-all and take-some universes have been given by Dalenius (1950) and Glasser (1962). Glasser (1962) expressed the cutoff value (that value which delineates the boundary of the take-all and take-some subuniverses) as a function of the mean, the sampling weight and the population variance. Their cutoff values were derived on the assumption that a prespecified sample size $n$ was to be drawn without replacement from a population of size $N$. In the present context, exact and approximate cutoff rules have been worked out for a similar situation. Rather than providing the sample size, the desired level of precision $c$ (coefficient of variation) of the estimates is given. Note that in many sampling situations, the sampler is given a set of objectives in terms of reliability of the estimates.

Singh and Singh (1975) have adapted Glasser's result to the case where pps with replacement sampling was used. The number of units $n$ to be drawn from the population is prespecified. They delineate the take-all and take-some universes by providing a lower and upper bound for the boundary point. These bounds are strictly a function of the measures of size used to compute the selection probabilities. The reason that the bounds do not involve any values concerning the study variable is that the study variable is expressed as a function of the measures of size through a regression model. In the present paper, the results are also extended to the case of the ratio estimator. Provided that a linear model exists between the study variable and the known auxiliary variable, a cutoff point can be constructed. This cutoff point will strictly be a function of the auxiliary variable.

## 2. THE SAMPLING PROCEDURE AND METHOD OF ESTIMATION

Consider a finite population $\phi_{N}$ consisting of $N$ units labelled $y_{1}, y_{2}$, $\ldots, y_{N}$. Define ordered statistics $y_{(N)}, y_{(N-1)}, \cdots, y_{(1)}$ where $y_{(N)} \leq y_{(N-1)} \leq \cdots \leq y_{(1)}$.

Let a simple random sample of size $n(\ell)$ be selected. Note that $n(\ell)$ is no longer a fixed sample size. Rather, it is a variable which depends on the number of take-all units $\ell$ to be included in the sample. Assume that the desired level of precision for the estimated total is given as c. The total $Y$ may be written as:

$$
\begin{equation*}
y=\sum_{i=1}^{\ell} y_{(i)}+\sum_{i=\ell+1}^{N} y_{(i)} . \tag{2.1}
\end{equation*}
$$

Given that $\ell$ units are take-all and $n(\ell)-\ell$ units are take-some, an estimator of the total $Y$ would be:

$$
\begin{equation*}
\hat{y}=\sum_{i=1}^{\ell} y_{(i)}+\frac{N-\ell}{n(l)-\ell} \sum_{i=1}^{n(\ell)-\ell} z_{i} \tag{2.2}
\end{equation*}
$$

where $y_{(N)} \leq z_{i} \leq y_{(\ell+1)}$ for $i=1,2, \ldots, n(\ell)-\ell$.
The variance of $\hat{\gamma}$ is:

$$
\begin{equation*}
V(\hat{Y})=\frac{(N-\ell)\{N-n(\ell)\}}{n(\ell)-\ell} s_{N-\ell}^{2} \tag{2.3}
\end{equation*}
$$

where

$$
\begin{aligned}
& S_{N-\ell}^{2}=\frac{1}{N-\ell-1} \sum_{i=\ell+1}^{N}\left(y(i)-\mu_{N-\ell}\right)^{2} \\
& \mu_{N-\ell}=\frac{1}{N-\ell} \sum_{i=\ell+1}^{N} y_{(i)}^{N} .
\end{aligned}
$$

In terms of reliability $c, V(\hat{Y})$ may be re-expressed as $V(\hat{Y})=c^{2} y^{2}$. Substituting $V(\hat{Y})=c^{2} Y^{2}$ into (2.3) and solving for $n(\ell)$ :

$$
\begin{equation*}
n(\ell)=\ell+\frac{(N-\ell)^{2} S_{N-\ell}^{2}}{c^{2} Y^{2}+(N-\ell) S_{N-\ell}^{2}} \tag{2.4}
\end{equation*}
$$

## 3. THE OPTIMUM POINT

The objective is to find the optimum value of $y$ which minimizes the sample size $n(\ell)$ for the given level of precision $c$. A necessary condition for the optimum point is that (2.4) with $\ell=m$ shall not exceed (2.4) with $\ell=m-1$ or $\ell=m+1$. This means that the optimum value of $y\left(y^{*}\right)$ is found whenever $n(m-1) \geq n(m)$ and $n(m) \leq n(m+1)$. This condition can be made more flexible if we introduce a real number $b$ into the inequalities, that is,

$$
\begin{equation*}
n(m-1) \geq n(m)+b-1 \tag{3.1}
\end{equation*}
$$

and

$$
n(m) \leq n(m+1)+b-1,
$$

where $b$ can be used to control the number of units to include in the take-all stratum.

Stopping rule (3.1) is the exact one for finding the optimal cutoff for a given b. To express (3.1) in terms of the optimal cutoff neighbouring values $y_{(m)}$ and $y_{(m+1)}$, we need the following two relations.

$$
\begin{equation*}
(N-m) S_{v+1}^{2}=(N-m-1) S_{v}^{2}+\frac{N-m}{N-m+1}\left(y(m)-u_{v}\right)^{2} \tag{3,2}
\end{equation*}
$$

and

$$
(N-m-2) S_{v-1}^{2}=(N-m-1) S_{v}^{2}-\frac{N-m}{N-m-1}\left(y(m+1)-\mu_{\nu}\right)^{2}
$$

where

$$
s_{v+k}^{2}=\frac{1}{N-m+k-1} \sum_{i=m-k+1}^{N}\left(y_{(i)}-\mu_{v+k}\right)^{2}
$$

and

$$
\mu_{v+k}=\frac{1}{N-m+k} \sum_{i=m-k+1}^{N} y_{(i)}
$$

for $k=-1,0,1$ and $v=N-m$.

Substituting (2.4), (3.2) into (3.1) one can show that:

$$
\begin{equation*}
\left(y_{(m)}-\mu_{v}\right)^{2} \geq\left\{\frac{\left[b N-n_{m}-(b-1) m\right](N-m)}{\left(n_{m}-m\right)\left(N-n_{m}-b+1\right)}+\frac{1}{N-m}\right\} \quad s_{v}^{2} \tag{3.3}
\end{equation*}
$$

and

$$
\left(y_{(m+1)}-1_{v}\right)^{2} \leq\left\{\frac{\left[b N-n_{m}-(b-1) m\right](N-m-2)}{\left(n_{m}-m\right)\left(N-n_{m}+b-1\right)}+\frac{1}{N-m}\right\} \quad s_{v}^{2} .
$$

The compromise for (3.3) if $m$ is the optimum number of units to include with certainty is

$$
\begin{align*}
\left(y *-\mu_{v}\right)^{2}= & \left\{\frac{b(N-m-1)}{n_{m}-m}+\frac{1}{2}\left[\frac{(b-1)(N-m)}{N-n_{m}-b+1}+\frac{(b-1)(N-m-2)}{N-n_{m}+b-1}\right]\right. \\
& \left.+\frac{1}{2} \frac{b(b-1)}{\left(n_{m}-m\right)}\left[\frac{N-m}{N-n_{m}^{-b+1}}-\frac{N-m-2}{N-n_{m}+b-1}\right]\right\} . \tag{3.4}
\end{align*}
$$

Note that if $m$ is the optimum number of units to be included in the sample with certainty, then $y_{(m+1)}<y^{*} \leq y_{(m)}$. Also, equation (3.4) is one solution of the system of inequalities given by (3.3). While (3.4) is a necessary condition for an optimum, it is not necessarily sufficient. More than one solution may exist, in which case the one that minimizes $n(\ell)$ for given $b$ would be chosen. As Glasser (1962) points out, while it may not pay to include with certainty a given unit by itself, it may pay to include it with several other units.

Noting that

$$
\begin{equation*}
n_{m}=m+\frac{(N-m)^{2} s_{v}^{2}}{c^{2} Y^{2}+(N-m) s_{v}^{2}} \tag{3.5}
\end{equation*}
$$

and

$$
\begin{equation*}
N-n_{m}=\frac{(N-m) c^{2} Y^{2}}{c^{2} Y^{2}+(N-m) S_{V}^{2}} \tag{3.6}
\end{equation*}
$$

substitution of (3.5) and (3.6) into (3.4) yields

$$
\begin{equation*}
\left(y \div-\mu_{v}\right)^{2} \equiv \frac{b c^{2} y^{2}}{N-m}+(2 b-1) s_{v}^{2}+\frac{(N-m)(b-1) s_{v}^{4}}{c^{2} y^{2}}, \tag{3.7}
\end{equation*}
$$

provided that $\mathrm{b}-\mathrm{l}$ is very much smaller in magnitude than $N C Y^{2} /\left(c r^{2}+N S_{\nu}^{2}\right)$.

An upper limit for $y^{*}$ can be obtained in terms of the population variance $S_{N}^{2}$, population size $N$ and mean $\mu_{N}$ by using the following inequalities:

$$
\begin{align*}
& N\left(y *-\mu_{N}\right) \leq(N-m)\left(y *-\mu_{v}\right),  \tag{3.8}\\
& (N-m-1) S_{v}^{2} \leq(N-1) S_{N}^{2}=\frac{m N\left(y *-\mu_{N}\right)^{2}}{N-m} \tag{3.9}
\end{align*}
$$

where it is true that for $m>0, y^{*}<\mu_{m}, \mu_{m}$ being the mean of the $m$ largest units in the population. Substituting inequalities (3.8) and (3.9) into (3.7), we obtain after some simplification the approximate cutoff rule

$$
\begin{equation*}
y *<\mu_{N}+\left[\frac{b c^{2} y^{2}}{N}+S_{N}^{2}\left\{(2 b-1)+\frac{N(b-1) S_{N}^{2}}{c^{2} y^{2}}\right\}\right]^{1 / 2} . \tag{3.10}
\end{equation*}
$$

This inequality depends only on the population size, the coefficient of variation $c, b, \mu_{N}$ and $S_{N}$. This approximation will be good only when $m$ is relatively small compared to $N$. The more extreme and the more variable the large units, the less well the limit approximates the exact solution. Although the computer programming and time involved in obtaining the exact cutoff point is quite minimal, it is nevertheless instructive to characterize the bound in terms of known population values.

Approximation (3.10) reveals one point about b's effect on the boundary point. If $b_{2}>b_{1}$, then the boundary point associated with $b_{2}$ will be higher than the one associated with $b_{1}$. Note that the converse also follows. The choice of $b$ is user dependent. Under various situations, the number of units in the take-all stratum may be varied. For instance, in business surveys, a possible determining factor affecting the cutoff rule could be the portion of the population that the take-all units represent in terms of the study variable. In this case, the user would probably take $b \leq 1$. Another factor could be response burden. The user would most likely introduce a rotation scheme which would permit some of the large units to rotate in and out of the sample. For this case, fewer units would be included in the take-all by choosing $b \geq 2$.

## 4. APPLICATION TO RATIO ESTIMATION

Suppose that the population $\phi_{\mathrm{N}}$ consists of N two dimensional vectors labelled $\left(y_{1}, x_{1}\right),\left(y_{2}, x_{2}\right), \ldots,\left(y_{N}, x_{N}\right)$. Define ordered statistics $x_{(N)}, x_{(N-1)}, \ldots, x_{(1)}$, where $x_{(N)} \leq x_{(N-1)} \leq \ldots \leq x_{(1)}$. Let the corresponding $y$ values be $y_{(N)}, y_{(N-1)}, \ldots, y_{(1)}$, where $y_{(i)}$ is the value associated with $x_{(i)}$ in the two dimensional vector $\left(y_{(i)}, x_{(i)}\right)$. If a sample of size $\Pi(\ell)$ is selected without replacement using simple random sampling, $\ell$ being the number of take-all units, an estimator for the total Y would be:

$$
\begin{aligned}
& \hat{Y}_{R}=\sum_{i=1}^{\ell} y_{(i)}+\frac{\bar{y}_{n}(\ell)-\ell}{\bar{x}_{n}(\ell)-\ell} x_{N-\ell} \\
& \bar{y}_{n(\ell)-\ell}=\frac{1}{n(\ell)-\ell} \sum_{i=1}^{n(\ell)-\ell} z_{i} \\
& \bar{x}_{n(\ell)-\ell}=\frac{1}{n(\ell)-\ell} \sum_{i=1}^{n(\ell)-\ell} u_{i} \\
& x_{N-\ell}=\sum_{i=\ell+1}^{N} \quad x_{(i)},
\end{aligned}
$$

${ }_{(N)} \leq u_{i} \leq x_{(\ell+1)}$ and $y_{(N)} \leq z_{i} \leq y_{(\ell+1)}$ for $i=1,2, \ldots, n(\ell)-\ell$. The variance of the total given by (4.1) is:

$$
V\left(\hat{Y}_{R}\right)=\frac{(N-\ell)\{N-n(\ell)\}}{n(\ell)-\ell} \frac{1}{N-\ell-1} \sum_{i=\ell+1}^{N}\left(y_{(i)}-R_{N-\ell} x_{(i)}\right)^{2}
$$

where $R_{N-\ell}=Y_{N-\ell} / X_{N-\ell}, Y_{N-\ell}=\sum_{i=\ell+1}^{N} y_{(i)}$, is the ratio of the $y$
population total to the x population total. Note that the optimal value of $y\left(y^{*}\right)$, will be found whenever for some $r, n(r-1) \geq n(r)$ and $n(r) \geq n(r+1)$, where

$$
\begin{gather*}
n(\ell)=\ell+\frac{(N-\ell)^{2} S^{2} R \cdot N-\ell}{c^{2} Y^{2}+(N-\ell) S_{R \cdot N-\ell}^{2}},  \tag{4.2}\\
S_{R \cdot N-\ell}^{2}=\frac{1}{N-\ell-1} \sum_{i=\ell+1}^{N}\left(Y(i)-R_{N-\ell} x_{(i)}\right)^{2}
\end{gather*}
$$

Note that expression (4.2) involves the study variable y. Consider the finite population $\phi_{N}$ as a sample from a superpopulation where the following model holds:

$$
y_{i}=\beta x_{i}+e_{i}
$$

where

$$
\begin{align*}
& E\left(e_{i} \mid x_{i}\right)=0, \\
& E\left(e_{i}, e_{j} \mid x_{i}, x_{j}\right)=0,  \tag{4.3}\\
& E\left(e_{i}^{2} \mid x_{i}\right)=\alpha x_{i} .
\end{align*}
$$

for $\alpha \geq 0$ and $i=1,2, \ldots, N$. Under this model, $n(l)$ can be expressed as a function of the $x$ 's. That is,

$$
\begin{equation*}
n(i)=\ell+\frac{(N-\ell)^{2} a\left\{x_{N-\ell}^{2}-\sum_{i=\ell+1}^{N} x_{(i)}^{2}\right\} /(N-\ell-1)}{c^{2} g x_{N}+(N-\ell) a\left\{x_{N-\ell}^{2}-\sum_{i=\ell+1}^{N} x_{(i)}^{2}\right\} /(N-\ell-1)} . \tag{4.4}
\end{equation*}
$$

where $g=\alpha x+\beta^{2} x^{2}$. The above relationship is easily derived by noting that under model (4.3),

$$
E Y^{2}=\alpha X+\beta^{2} x^{2}
$$

and

$$
E S_{R \cdot N-\ell}^{2}=\frac{\alpha}{N-\ell-1}\left(X_{N-\ell}-\frac{1}{X_{N-\ell}} \sum_{i=\ell+1}^{N} x_{(i)}^{2}\right)
$$

When the optimal $y$ has been found, it can be shown that $n(r-1) \geq n(r)$ is equivalent to

$$
\begin{align*}
& \left\{(N-r+1) \propto c^{2} g\left[X_{\rho}^{2}+(N-r) s_{X \cdot \rho}^{2}\right]-c^{4} g^{2} X_{\rho}\right\} \times(r)  \tag{4.5}\\
& \geq c^{4} g^{2} x_{\rho}^{2}+\frac{N-r+1}{N-r} \propto c^{2} g x_{\rho}^{3}-(N-r+1) \propto c^{2} g s_{X \cdot \rho}^{2} X_{\rho}
\end{align*}
$$

where

$$
\begin{aligned}
& x_{\rho+k}=\sum_{i=r-k+1}^{N} x_{(i)}, \\
& s_{x \cdot \rho+k}^{2}=\frac{1}{N-r-1}\left[\sum_{i=r-k+1}^{N} x_{(i)}^{2}-\frac{x_{\rho+k}^{2}}{N-r+k}\right],
\end{aligned}
$$

$$
\text { for } \rho=N-r \quad \text { and } \quad k=-1,0,1
$$

Note that use of the following two relations is needed to derive (4.5):

$$
x_{\rho+1}=x_{\rho}+x_{(r)}
$$

and

$$
x_{\rho+1}^{2}-\sum_{i=r}^{N} x_{(i)}^{2}=x_{\rho}^{2}-\sum_{i=r+1}^{N} x_{(i)}^{2}+2 x_{(r)} x_{\rho} .
$$

Similarly, using

$$
x_{\rho-1}=x_{\rho}-x_{(r)}
$$

and

$$
\left.x_{\rho-1}^{2}-\sum_{i=r+2}^{N} x_{(i)}^{2}=x_{\rho}^{2}-\sum_{i=r+1}^{N} x_{(i)}^{2}+2 x_{(r+1)}^{2}-2 x_{(r+1}\right) x_{\rho},
$$

it can be shown that $n(r) \leq n(r+1)$ is equivalent to

$$
\left\{-c^{4} g^{2} x_{\rho}-\frac{N-r-1}{N-r-2}(N-r-2) \propto c^{2} g x_{\rho}^{2}-(N-r)(N-r-1) \propto c^{2} g s_{x \cdot \rho}^{2}\right.
$$

$$
+2 \frac{(N-r)(N-r-1)}{N-r-2} \propto c^{2} g x_{\rho} \times(r+1)^{3 \times}(r+1)
$$

$$
\begin{equation*}
\leq-c^{4} g^{2} x_{\rho}^{2}-\alpha c^{2} g x_{\rho}^{3} \frac{N-r-1}{N-r-2}+\frac{(N-r)(N-r-1)}{N-r-2} \alpha c^{2} g x_{\rho} s_{x \cdot \rho}^{2} \tag{4.6}
\end{equation*}
$$

If $r$ is the optimum number of large units to include with certainty, then $x(r) \geq x^{*} \geq x_{(r+1)}$ and this is satisfied whenever the following quadratic equation holds,

$$
\begin{align*}
& \left\{(N-r) \alpha c^{2} g s_{X \cdot \rho}^{2}-\frac{N-r}{N-r-2} \alpha c^{2} g x_{\rho}^{2}-c^{4} g^{2} x_{\rho}^{2}\right. \\
& \left.+\frac{(N-r)(N-r-1)}{N-r-2} \propto c^{2} g x_{\rho} x^{*}\right\} x^{*} \\
& =-\frac{\alpha c^{2} g x_{\rho}^{3}}{(N-r)(N-r-2)}+\frac{\alpha c^{2} g x_{\rho} S_{X \cdot \rho}^{2}}{(N-r-2)} \tag{4.7}
\end{align*}
$$

Note that the above expression depends on knowledge of $\alpha$ and $\beta$. Thus, any estimate of $\alpha$ and $\beta$ values obtained from some earlier surveys or other sources may be utilized to great advantage.

## 5. SOME PRACTICAL ILLUSTRATIONS

The use of the procedure given in section 2 presumes that the population from which the sample is to be drawn, is to be a good proxy for the target population. An example where such a procedure may be used is the following. All the values associated with the units of a business universe are known at time $t_{1}$. A sample is drawn from this universe at time $t_{1}+k_{1}$, $k_{1} \geq 0$, and to be used as a basis for inference to the universe characteristics from $t_{1}+k_{1}$ to $t_{1}+k_{2}$ where $k_{2}>k_{1}$. In this instance, the universe at time $t_{1}$, $\phi\left(t_{1}\right)$, may be different from the universe at time $t_{2}, \phi\left(t_{2}\right), t_{2}>t_{1}$. However, if it can be assumed that the cutoff value computed at time $t_{1}$ is not too different from the one that would be computed at time $t_{2}$, then partitioning of the population $\phi\left(t_{1}\right)$ will still yield gains.

The data used to illustrate the results given in section 3 is from the 1976 Food and Beverage Annual Survey. This survey is essentially a census of all eating and drinking establishments covered by the Merchandising and Services Division of Statistics Canada. Establishments covered in this survey includes all known businesses with establishments classified to the Standard Industrial Classification code 886(1970). The Standard Industrial Classification code is broken down further into seven kinds of businesses that range from licenced restaurants to beverage rooms, bars and night clubs. Data for this survey is presently being published at a subprovincial by kind of business cross-classification. The example takes a situation where the business universe is known at time $t_{1}$ (the 1976 Food and Beverage Restaurant Survey) and a sample is to be drawn at time $\mathrm{t}_{1}{ }^{+\mathrm{k}_{1}}$ (the projected Monthly, Tavern, Caterers and Restaurant Survey).

The cutoff rule that is illustrated is the one given by (3.1) with b chosen equal to 1 and 2 respectively. Four subprovincial by kind of business strata have been chosen to provide the examples. They are respectively: Beverage Rooms, Bars and Night Clubs in Newfoundland (stratum 1), Beverage Rooms, Bars and Night Clubs in the non-metropolitan areas of New Brunswick (stratum 2), Licenced Restaurants in HalifaxDartmouth (stratum 3), and Beverage Rooms, Bars, Night Clubs in the non-metropolitan areas of Quebec (stratum 4). Some of the statistical characteristics for those strata are given in Table 5.1. These are the minimum, maximum and mean sales for each of the strata. The standard deviation, $S_{N}$, is also provided with the associated population size. Note that these statistics imply that the associated frequency distributions are positively skewed. This remark is supported by Fig. 5.1 through Fig. 5.4.

Table 5.l: Statistical Characteristics for the Strata of Interest

| Stratum | Minimum | Maximum | Mean | Standard Dev. | $N$ |
| :---: | ---: | ---: | :---: | :---: | :---: |
| 1 | 3,000 | 476,141 | 139,380 | 67,800 | 170 |
| 2 | 4,000 | 463,000 | 181,930 | 90,160 | 61 |
| 3 | 15,045 | $1,223,360$ | 350,250 | 263,830 | 63 |
| 4 | 3,345 | 885,333 | 132,770 | 72,520 | 632 |

For each of the strata in question and given the coefficient of variation desired, we provide the number of units to be included in the take-all substratum, the exact and approximate cutoff and the sample that would have been selected had no take-all substratum been formed. This information is displayed in Table 5.2. Note that the approximate cutoff point is given by inequality (3.10) and the exact cutoff point by equation (3.3) with $b=1$.

Table 5.2: Information Concerning the Take-all Procedure Given by Inequalities (3.1) for $b=1$

| Stratum | $c$ | Exact Cutoff | Approximate <br> Cutoff | $m$ | $n(m)$ | $n(0)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 0.1116 | 353,351 | 353,230 | 3 | 15 | 17 |
| 2 | 0.1063 | 339,071 | 357,840 | 4 | 13 | 16 |
| 3 | 0.1359 | 806,999 | 811,060 | 6 | 10 | 21 |
| 4 | 0.1209 | 598,192 | 542,450 | 3 | 12 | 20 |

In the above table, $m$ is equal to the number of units to be included in the sample with certainty and $n(m)$ is the corresponding overall sample size required to achieve the desired reliability. Note that $n(m)<n(0)$ for all strata considered, where $n(0)$ is the sample size with no take-all units. Hence, if take-all units are to be found, the overall sample size will be smaller than that of the sample with no take-all units. Note that the approximate cutoff given by (3.10) is quite close to the exact cutoff given by (3.3). Results for $b=2$ provided in Table 5.3 highlight the effect of $b$ on the boundary points.

Table 5.3: Information Concerning the Take-all Procedure Given by Inequalities (3.1) for $b=2$

| Stratum | c | Exact Cutoff | Approximate <br> Cutoff | p | $\mathrm{n}(\mathrm{p})$ | $\mathrm{n}(0)$ |
| :---: | :---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 0.1116 | 476,141 | 450,133 | 1 | 16 | 17 |
| 2 | 0.1063 | 462,303 | 451,950 | 1 | 15 | 16 |
| 3 | 0.1359 | $1,223,360$ | $1,077,048$ | 1 | 19 | 21 |
| 4 | 0.1209 | 832,991 | 717,265 | 2 | 15 | 20 |

Note that $p$ stands for the number of units in the take-all stratum and $n(p)$ is the corresponding overall sample size. Again, the approximate cutoff given by (3.10) is quite close to the exact cutoff given by (3.3) with $\mathrm{b}=2$. The exact bound with $\mathrm{b}=2$ tends to yield fewer take-all units than the exact bound with $b=1$. The same conclusion is reached if the approximate bound is used.


Fig. 5.3. Frequency distribution of 63 units in stratum 3


Fig. 5.4. Frequency distribution of 632 units in stratum 4


Fig. 5.1. Frequency distribution of 170 units in stratum 1


Fig. 5.2. Frequency distribution of 61 units in stratum 2

## 6. COMCLUSION

It is desirable to stratify highly skewed populations on the basis of the size of the units. The approach suggested in the present paper is to put a certain number of large units into a take-all stratum and sample those with certainty. The remaining units, those attached to the take-some stratum, are sampled at an appropriate rate. The number of units to include, with certainty depends on the desired level of precision $c$, the scalar $b$, the population mean $\mu_{N}$, the population variance $S_{N}^{2}$ and the number of units $N$ when criteria (3.1) is used. Note that the sampler may vary the number of units in the take-all stratum by varying b. The approximate stopping rule (3.10) may be used as an initial estimate for the corresponding exact cutoff given by (3.1) provided that the necessary information on the population of interest is available.

There are several advantages in stratifying a highly skewed population for the given method. For a fixed level of reliability, the overall sample size associated with this procedure will invariably be lower than the sample size associated with no stratification. Cochran (1963, p. 38-39) points out that for frequency distributions that are not reasonably close to normality, it is risky to use the normal approximation as a basis for constructing confidence intervals. By separating some of the largest observations from highly skewed distributions, confidence intervals are essentially based on populations which are less skewed. This last point should encourage the sampler in having more confidence in using the normal approximation. Finally, this type of stratification guards against overestimation of population characteristics when highly skewed distributions are sampled.

Glasser (1962) has pointed out that the definition one should assign to large units depends on the method of sampling the remainder of the population and the method of estimation. For the case of ratio estimation, the quadratic equation given by (4.7) should be solved sequentially to obtain the cutoff point $\times *$.

RESUME


#### Abstract

Dalenius (1950) et Glasser (1962) ont énoncé des règles approximatives de partage pour la stratification d'une population en un univers à tirage complet et un univers à tirage partiel. Ils ont exprimé la valeur de partage (qui marque la frontière entre les deux types d'univers) en fonction de la moyenne, du poids de l'échantillonnage et de la variance de la population. Leurs valeurs de partage ont été calculées à partir de l'hypothèse d'un échantillon aléatoire unique de taille $n$ tiré sans remise d'une population de taille $N$.

Ici, l'auteur a élaboré des règles de partage exactes et approximatives pour une situation semblable. Au lieu d'avoir la taille de l'échantillon, on dispose de la précision (coefficient de variation). Il est à noter que dans de nombreux cas d'échantillonnage le chercheur a un ensemble d'objectifs exprimés en fonction de la fiabilité et non de la taille de l'échantillon. Le résultat est particulièrement utile lorsqu'il s'agit de déterminer la limite de partage pour des échantillons tirés d'une population connue. Cette méthode est également utilisée dans le cas de l'estimation par quotient.


## REFERENCES

[1] Cochran, W.G. (1963), Sampling Techniques, Wiley, 2nd ed., New York.
[2] Dalenius, T. (1950), "The Problem of Optimum Stratification in a Special Type of Design", Skandinavisk Aktuariediskrift, p. 61-70.
[3] Glasser, G.J. (1962), "On the Complete Coverage of Large Units in a Statistical Study', Review of the International Statistical Institute, Vol. 30, p. 28-32.
[4] Hidiroglou, M.A. and Srinath, K.P. (1977), "Some Estimators of Population Totals from Simple Random Samples Containing Large Units'. Proceedings of the American Statistical Society, Chicago, Social Statistics Section, p. 903-908.
[5] Singh, D. and Singh, R. (1975), "On the Inclusion of Large Units in the Sample', Journal $c i$ the Indian Society of Agricultural Statistics, p. 24-31.

# UNBIASED ESTIMATION OF PROPORTIONS UNDER SEQUENTIAL SAMPLING 

M.D. Bankier ${ }^{\text {l }}$


#### Abstract

Under a sequential sampling plan, the proportion defective in the sample is generally a biased estimator of the population value. In this paper, an unbiased estimator is given. Also, an unbiased estimator of its variance is derived. These results are applied to an estimation problem from the 1976 Canadian Census.


## 1. INTRODUCTION

In a quality control (Q.C.) operation, a sample of $m$ units from some lot is examined. The number $x$ of defective units in the sample determines if the lot is accepted or rejected. Frequently, a sequential sampling plan is used. A definition of such plans is given in Section 2. For these sampling plans, the number of units sampled is a random variable and $\frac{x}{m}$ is usually a biased estimator of the proportion defective in the lot. In Section 3, an estimator $p(m, x)$ is presented which is unbiased under both sequential sampling with replacement (w.r.) and without replacement (w.o.r.). This estimator was first proposed by Girshick, Mosteller, and Savage [3J under sampling w.r. A result given by DeGroot [1] is also discussed. it states for which sampling plans $p(m, x)$ is the uniformly minimum variance unbiased (UMVU) estimator under sampling w.r. from an infinite population. In Section 4, unbiased estimators of $V(p(m, x))$ are derived under both sampling w.r. and w.o.r. These estimators are constructed using an approach taken under sampling w.r. in [3]. In Section 5, the results are applied to an estimation problem from the 1976 Canadian Census.

[^13]
## 2. DEFINITION OF A SAMPLING PLAN

Let ( $m, x$ ) represent the event that $m$ units have been sampled and of these, $x$ are defective. A sampling plan is defined as a function $S$ such that if

$$
\begin{array}{ll}
S((m, x)) & =1 \\
\text { If } & \text { then another unit is sampled and inspected. } \\
S((m, x))=0 & \text { then no more units are sampled. } \tag{2.2}
\end{array}
$$

The event $(m, x)$ is called a boundary point if $S((m, x))=0$ and $\operatorname{Pr}((m, x))>0$. The set of all boundary points of a sampling plan will be labelled $B$. This set contains the possible outcomes of the sampling plan. A sampling plan is said to be closed if

$$
\begin{equation*}
\left(m^{\Sigma}, x\right) \varepsilon_{B} \operatorname{Pr}((m, x))=1 \tag{2.3}
\end{equation*}
$$

A bounded sampling plan is one where there exists a positive integer $b$ such that

$$
\begin{equation*}
\operatorname{Pr}(m \leq b)=1 \tag{2.4}
\end{equation*}
$$

The size of a sampling plan is the smallest b for which (2.4) holds. Only closed and bounded sampling plans will be considered in this paper. Single, double, multiple and sequential bounded sampling plans used in quality control all satisfy the above criteria.

## 3. AN UNBIASED ESTIMATOR OF THE PROPORTION DEFECTIVE

Let

$$
x_{i}=\left\{\begin{array}{l}
1 \text { if the ith unit sampled is defective }  \tag{3.1}\\
0 \text { otherwise. }
\end{array}\right.
$$

A path to ( $m, x$ ) will be defined as a sequence of points

$$
\begin{array}{ll} 
 \tag{3.2}\\
\left(j, \sum_{i=1}^{j} x_{i}\right) & j=1, \ldots, m \\
& \sum_{i=1}^{m} x_{i}=x
\end{array}
$$

such that

$$
\begin{equation*}
s\left(\left(j, \sum_{i=1}^{j} x_{i}\right)\right)=1 \text { for } j=1, \ldots, m-1 . \tag{3.3}
\end{equation*}
$$

Assume the units are sampled w.r. with $P$ the proportion defective. Under these assumptions, it can be seen that

$$
\begin{equation*}
\operatorname{Pr}((m, x))=c(m, x) p^{x}(1-p)^{m-x} \tag{3.4}
\end{equation*}
$$

where

$$
\begin{equation*}
c(m, x)=\text { the number of distinct paths to }(m, x) . \tag{3.5}
\end{equation*}
$$

Let

$$
\begin{equation*}
\bar{y}_{i}=\frac{\text { number of defective units in first } i}{i} \tag{3.6}
\end{equation*}
$$

The sample mean $\bar{y}_{m}=\frac{X}{m}$ is generally a biased estimator of $P$ except when the sample size $m$ is fixed. However, because at least one unit is always sampled, $\bar{y}_{1}$ is unbiased. The estimator

$$
\begin{align*}
p(m, x) & =E\left(\bar{y}_{1} \mid(m, x)\right) \\
& =\frac{\operatorname{Pr}\left(\bar{y}_{1}=1 \text { and }(m, x)\right)}{\operatorname{Pr}((m, x))} \\
& =\frac{d(m, x) P^{x}(1-P)^{m-x}}{c(m, x) P^{x}(1-P)^{m-x}} \\
& =\frac{d(m, x)}{c(m, x)} \tag{3.7}
\end{align*}
$$

is also an unbiased estimator of $P$ with the same or a smaller variance than $\bar{y}_{1}$ by the Rao-Blackwell Theorem. In the above expression

$$
\begin{align*}
d(m, x)= & \text { the number of distinct paths to }(m, x) \text { that }  \tag{3.8}\\
& \text { pass through the point }(1,1)
\end{align*}
$$

Under sampling w.o.r. from a finite population, expression (3.7) still holds. The $P^{x}(1-P)^{m-x}$ is replaced by

$$
G(m, x ; M, x)=\left\{\begin{array}{lc}
\binom{x}{x} & \binom{M-x}{m-x}  \tag{3.9}\\
\hline\binom{m}{x} & \binom{M}{m} \\
0 & \text { for } \underset{\substack{m \leq M \\
x \leq x \\
x \leq M}}{x \leq m}= \\
\text { otherwise }
\end{array}\right.
$$

where $M$ is the number of units in the population and $X$ is the number of defectives. In this case, $P=X / M$.

A sampling plan is complete if the only estimator $f$ such that $E(f)=0$ for all $r$, is one defined by $f(m, x)=0$ for all $(m, x) \varepsilon B$. In DeGroot [1], it is shown that under sampling w.r. a sampling plan of size $b$ is complete if and only if the boundary $B$ contains exactly $b+l$ points. If the sampling plan is complete and the population size is infinite, then by the Lehmann-Scheffe Uniqueness Theorem, (Roussas [4], p.216), $p(m, x)$ is the UMVU estimator of $P$.

## 4. AN UNBIASED ESTIMATOR OF $V(p(m, x))$

For any closed, complete and bounded sampling plan with size $b \geq 2$, it is possible to construct an unbiased estimator $v(p(m, x))$ of $V(p(m, x))$ under both sampling w.r. and w.o.r. Expressions for $v(p(m, x))$ under sampling w.o.r. from a finite population are given below for three types of boundary point sets $B$. These expressions reduce to those for sampling w.r. if $\frac{M-1}{M}$ is replaced by 1 and $\frac{1}{M}$ is replaced by 0 .

Case (1): It is assumed that (2,2) $\varepsilon$ B. Then

$$
v(p(m, x))=\left\{\begin{array}{l}
0 \text { for }(m, x)=(2,2)  \tag{4.1}\\
p(m, x)\left(p(m, x)-\frac{1}{M}\right) \text { otherwise for }(m, x) \varepsilon B .
\end{array}\right.
$$

Case (2): It is assumed that $(2,2) \notin B$ but $(1,1) \in B$. Then

$$
v(p(m, x))=\left\{\begin{array}{l}
0 \text { for }(m, x)=(1,1)  \tag{4.2}\\
\frac{M-1}{M} \frac{e(m, x)}{c(m, x)} \text { otherwise for }(m, x) \in B
\end{array}\right.
$$

where $e(m, x)=$ the number of distinct paths going through ( 1,1 ) and ( $m, x$ ) under the sampling plan with boundary point set $B^{*}$. $B^{*}$ is identical to $B$ except the point $(1,1)$ has been removed and the point $(2,2)$ has been added.

Case (3): It is assumed that $(2,2) \notin B$ and $(1,1) \notin B$. Then

$$
\begin{equation*}
v(p(m, x))=p(m, x)\left(p(m, x)-\frac{1}{M}\right)-\frac{M-1}{M} \frac{f(m, x)}{c(m, x)} \tag{4.4}
\end{equation*}
$$

where $f(m, x)=$ the number of distinct paths going through $(2,2)$ and ( $m, x$ ) under the sampling plan with boundary (4.5) point set $B$.

Proof: This is given for case lonly. The proofs for the other two cases are similar.

$$
\begin{align*}
V(p(m, x)) & =E\left(p^{2}(m, x)\right)-p^{2} \\
& =\sum_{(m, x) \in B^{2}}^{\sum} p^{2}(m, x) \operatorname{Pr}((m, x))-p^{2} \tag{4.6}
\end{align*}
$$

It can be seen that $d(2,2)=1$ and $c(2,2)=1$. Thus, under sampling with replacement

$$
\begin{equation*}
p^{2}(2,2) \operatorname{Pr}((2,2))-p^{2}=p^{2}-p^{2}=0 . \tag{4.7}
\end{equation*}
$$

Thus

$$
\begin{equation*}
V(p(m, x))=\sum_{\substack{(m, x) \varepsilon B \\(m, x) \neq(2,2)}}^{\sum} p^{2}(m, x) \operatorname{Pr}((m, x)) . \tag{4.8}
\end{equation*}
$$

Therefore

$$
v(p(m, x))=\left\{\begin{array}{l}
0 \text { for }(m, x)=(2,2)  \tag{4.9}\\
p^{2}(m, x) \text { otherwise for }(m, x) \in B
\end{array}\right.
$$

is an unbiased estimator of $V(p(m, x))$ under sampling w.r. Under sampling w.o.r., eq. (4.7) does not reduce to 0 . In fact

$$
\begin{align*}
& p^{2}(2,2) \operatorname{Pr}((2,2))-P^{2} \\
& =G(2,2 ; M, X)-P^{2} \\
& =-\frac{P(1-P)}{M-1} . \tag{4.10}
\end{align*}
$$

This implies that the estimator (4.9) under sampling w.o.r. has a bias of $\frac{P(1-P)}{M-1}$. However if $(2,1) \varepsilon B$ then
is an unbiased estimator of $P(1-P) \frac{M}{M-1}$ since

$$
d(m, x)=\left\{\begin{array}{l}
1 \text { for }(m, x)=(2,1) \text { or }(2,2)  \tag{4.12}\\
0 \text { otherwise for }(m, x) \in B
\end{array}\right.
$$

and

$$
\begin{equation*}
G(2,1 ; M, X)=P(1-P) \frac{M}{M-1} . \tag{4.13}
\end{equation*}
$$

If $(2,1) \notin B$, then let $B^{\prime}$ be the set of boundary points $B$ with $(2,1)$ added. Also let

$$
\begin{align*}
k(m, x)= & \text { the number of distinct paths to }(m, x) \text { under }  \tag{4.14}\\
& \text { sampling plan with boundary point set } B^{\prime}
\end{align*}
$$

and

$$
g(m, x)=\left\{\begin{array}{l}
0 \text { if }(m, x)=(2,1)  \tag{4.15}\\
\text { the number of paths going through }(2,1) \text { and } \\
(m, x) \text { under the sampling plan with boundary } \\
\text { point set } B \text { if }(m, x) \neq(2,1) .
\end{array}\right.
$$

It is obvious that

$$
\begin{equation*}
k(m, x)=c(m, x)-g(m, x) . \tag{4.16}
\end{equation*}
$$

Al so,

$$
\begin{align*}
& \sum_{(m, x) \varepsilon B^{\prime}} k(m, x) G(m, x ; M, X)=1  \tag{4.17}\\
& k(2,1) G(2,1 ; M, x)+\sum_{(m, x) \varepsilon B}^{\Sigma} k(m, x) G(m, x ; M, x)=1,  \tag{4.18}\\
& k(2,1) G(2,1 ; M, x)+\sum_{(m, x) \varepsilon B}^{\Sigma} c(m, x) G(m, x ; M, x) \\
& -\underset{(m, x) \varepsilon B}{\sum} g(m, x) G(m, x ; M, X)=1 \tag{4.19}
\end{align*}
$$

and

$$
\begin{equation*}
\sum_{(m, x) \in B} g(m, x) G(m, x ; M, X)=k(2,1) G(2,1 ; M, X) \tag{4.20}
\end{equation*}
$$

Now if $(1,0) \varepsilon B$, then $k(2,1)=1$ and

$$
g(m, x)=\left\{\begin{array}{l}
0 \text { for }(m, x)=(2,2)  \tag{4.21}\\
d(m, x) \text { otherwise for }(m, x) \varepsilon B
\end{array}\right.
$$

If $(1,0) \notin B$ then $k(2,1)=2$ and

$$
g(m, x)=\left\{\begin{array}{l}
0 \text { for }(m, x)=(2,2)  \tag{4.22}\\
2 d(m, x) \text { otherwise for }(m, x) \in B
\end{array}\right.
$$

Thus from eq. (4.21) and (4.22), eq. (4.20) can be rewritten

$$
\begin{equation*}
\underset{(m, x) \neq \xi_{1}}{(m, 2)} d(m, x) G(m, x ; M, x)=G(2,1 ; M, x) \tag{4.23}
\end{equation*}
$$

Thus $q(m, x)$ given in eq. (4.1l) is also an uniased estimator of $G(2,1 ; M, X)$ if $(2,1) \notin B$. Therefore, an unbiased estimator of $v(p(m, x))$ under sampling w.o.r. from a finite population is given by eq. (4.1).

## 5. ESTIMATING THE PROPORTION OF DEFECTIVE QUESTIONNAIRES IN THE 1976 CENSUS

In the 1976 Canadian Census, Quality Control Technicians were used to examine and reject enumeration areas (EAs) where the data, on a sample basis, was of poor quality. A sample of population and housirig questionnaires (Forms 2B) were sampled w.o.r. one at a time. If a sampled Form 2B did not meet certain quality standards, it was rejected. After examining a Form $2 B$, the technician looked at Table 1 to determine whether to accept the EA, reject the EA or continue sampling Forms 2B. For example, after 17 households had been examined, the EA was accepted, rejected or sampling continued if 2,4 or 3 households respectively had been rejected. A minimum of two Forms $2 B$ and a maximum of 24 Forms $2 B$ were sampled. A sequential sampling plan requires a smaller sample on the average than a fixed sample size plan to distinguish between good and bad quality EAs with reasonable accuracy (see Duncan [2], p. 178).

For future planning purposes, an estimate for Canada was needed of the proportion of defective Forms $2 B$ (those that did not meet the quality standards) at the beginning of the Q.C. operation. A stratified sample of EAs was picked with proportional allocation. The EAs were stratified by regional office and EA methodology. Information recorded from the QC forms included the number of Forms 2B sampled and the number rejected. The proportion of defective Forms 2B on the Canadian level was estimated by

$$
\begin{equation*}
\hat{p}=\frac{\sum_{i} \frac{N_{i}}{n_{i}} \sum_{j=1}^{n_{i}} M_{i j} p_{i j}}{\hat{M}} \tag{5.1}
\end{equation*}
$$

where
$N_{i}=$ the number of EAs in the ith stratum,
$n_{i}=$ the number of EAs in the $i$ th stratum sample,
$M_{i j}=$ the number of forms $2 B$ in the $j$ th EA sampled from the ith stratum,
$p_{i j}=$ the estimator for the proportion of defective Forms $2 B$ in the $j$ th EA sampled from the ith stratum.

Table 1: Decision Table For the 1976 Census Form 2B Sequential Sampling Plan

| Number of Forms 2B Sampled | Accept EA if the Following Number of Forms 2B are Rejected. | Reject EA if the <br> Following Number of <br> Forms 2B are Rejected |
| :---: | :---: | :---: |
| 2 | * | 2 |
| 3 | * | 2 |
| 4 | * | 2 |
| 5 | * | 2 |
| 6 | * | 2 |
| 7 | 0 | 3 |
| 8 | 0 | 3 |
| 9 | 0 | 3 |
| 10 | 0 | 3 |
| 11 | 0 | 3 |
| 12 | 1 | 4 |
| 13 | 1 | 4 |
| 14 | 1 | 4 |
| 15 | 1 | 4 |
| 16 | 1 | 4 |
| 17 | 2 | 4 |
| 18 | 2 | 5 |
| 19 | 2 | 5 |
| 20 | 2 | 5 |
| 21 | 2 | 5 |
| 22 | 2 | 5 |
| 23 | 3 | 5 |
| 24 | 4 | 5 |

* indicates that the EA cannot be accepted with this sample size
and

$$
\begin{equation*}
\hat{M}=\sum_{i} \frac{N_{i}}{n_{i}} \sum_{j=1}^{n_{i}} M_{i j} \tag{5.6}
\end{equation*}
$$

Estimators of the variance and bias of $p$ are given by

$$
\begin{equation*}
v(\hat{p})=\frac{1}{\hat{M}^{2}} \sum_{i}\left(s\left(r_{i j}, r_{i j}\right)+\left(\frac{N_{i}}{n_{i}}\right)^{2} \sum_{j=1}^{n_{i}} M_{i j} v\left(p_{i j}\right)\right) \tag{5.7}
\end{equation*}
$$

and

$$
\begin{equation*}
b(\hat{p})=-\frac{1}{\hat{M}^{2}} \sum_{i} s\left(M_{i j}, r_{i j}\right) \tag{5.8}
\end{equation*}
$$

where

$$
\begin{equation*}
r_{i j}=M_{i j}\left(p_{i j}-\hat{p}\right) \tag{5.9}
\end{equation*}
$$

and
$s\left(w_{i j}, v_{i j}\right)=N_{i}^{2}\left(\frac{1}{n_{i}}-\frac{1}{N_{i}}\right) \frac{1}{n_{i}-1}\left(\sum_{j=1}^{n} w_{i j} v_{i j}-\frac{1}{n_{i}}\left(\sum_{j=1}^{n} w_{i j}\right)\left(\sum_{j=1}^{n} v_{i j}\right)\right)$.

The standard formulae were used to linearize the ratio estimator $\hat{p}$ in $v(p)$ and $b(\hat{p})$. Conditional expectations were applied in the derivations because of the two-stage sampling. Sample quantities were substituted for population values where necessary.

To calculate $p_{i j}$ required finding $c(m, x)$ and $d(m, x)$ in expression (3.7). This was done in Figures 1 and 2. The number in a cell in Figure 1 , for example, is the value of $c(m, x)$ for that $m$ and $x$. Cells with solid lines drawn around them are points where an EA could be accepted or rejected in Table l. The calculations began in Figure 1 by placing the number 1 in the first column. The number in a cell to the right was calculated by adding together the numbers in any cells to the left with arrows pointing into that
Figure 1: Calculating $c(m, x)$ for the Table 1 Sampling Plan

cell. Figure 1 shows that this sampling plan has 25 boundary points. The point $(8,0)$, for example, is not a boundary point since $\operatorname{Pr}((8,0))=0$. This indicates (see section 3 ) that if the questionnaires had been sampled w.r. from an infinite population, then $p(m, x)$ would have been the UMVU estimator of $P$. The unbiased estimator of $V\left(p_{i j}\right)$ is given by eq. (4.1).

The results below are based on a sample of 1199 EAs:

$$
\begin{gather*}
\hat{p}=2.63 \times 10^{-2},  \tag{5.11}\\
b(\hat{p})=2.82 \times 10^{-8},  \tag{5.12}\\
\sqrt{v(\hat{p})}=3.22 \times 10^{-3},  \tag{5.13}\\
\frac{\sqrt{v(\hat{p})}}{\hat{p}} \times 100 \%=12.2 \% \tag{5.14}
\end{gather*}
$$

It can be seen that the estimate of the Canadian proportion defective has a reasonably small coefficient of variation and a very small bias.

## ACKNOWLEDGEMENTS

I would like to thank my supervisor R. Burgess for his advice and also $G$. Brackstone for his suggestion that $p(m, x)$ be investigated.

Dans un plan de sondage séquentiel, la proportion défectueuse de l'échantillon est en général un estimateur biaisé de la valeur de la population. L'auteur de l'article propose un estimateur sans biais, dont un estimateur sans biais de la variance est également défini. Les résultats sont appliqués à un problème d'estimation tiré du recensement de 1976.

## REFERENCES

[1] DeGroot, M.H. (1959), "Unbiased Sequential Estimation for Binomial Populations'", Ann. Math. Stat., 30, 80-101.
[2] Duncan, A.J. (1965), "Quality Control and Industrial Statistics', Third Edition, Richard D. Irwin, Inc.
[3] Girshick, M.A., Mosteller, F. and Savage, L.J. (1946), "Unbiased Estimates for Certain Binomial Sampling Problems With Applications' ${ }^{\prime \prime}$ Ann. Math. Stat., 17, 13-23.
[4] Roussas, G.G. (1973), "A First Course in Mathematical Statistics", Addison-Wesley Publishing Company.
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
CONTENTS
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. KROTKI 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 ..... 115December/décembre 1978
The Evolution of a National Statistical Agency J. SPEAR ..... 125
Non-Response and Imputation
R. PLATEK and G.B. Gray ..... 144
The Application Of A Systematic Method Of Automatic Edit And Imputation To The 1976 Canadian Census of Population And Housing C.J. Hill ..... 178
Large Scale Imputation Of Survey Data M.J. COLLEDGE, J.H. JOHNSON, R. PARE, and I.G. SANDE ..... 203
Some Methods For Updating Sample Survey Frames
And Their Effects On Estimation
J.D. DREW, G.H. CHOUDHRY, and G.B. GRAY ..... 225
Alternative Estimators In PPS Sampling M.P. SINGH ..... 264


[^0]:    ${ }^{1}$ Adapted from a paper presented to the Conference of the Institute of Public Administration of Canada, September 7, 1977.

    2 Peter G. Kirkham, Chief Statistician of Canada, Statistics Canada.

[^1]:    ${ }^{1}$ Professor Machlup provides a detailed discussion of the definitional problems relating to such terms as "information" and "knowledge". Drucker [5], Kochen [11] and Ludlow [12] provide alternative definitions. Diener [4] in part distinguishes between "information" and "knowledge" by the type of questions that can be answered by each; "information" can be used to answer such questions as What?, Where?, When? and Who? while the answers to such questions as How?, and Why? require "Knowledge". The arguments in this paper are not contingent upon a precise definition of these terms so the subject is not pursued further here. However, a more complete and detailed exposition of the subject would demand careful consideration of the meaning of these terms.

[^2]:    ${ }^{1}$ Note that Professor Machlup identified the absence of an effective pricing system as a major characteristic of the knowledge industry. [13], p. 8.

[^3]:    ${ }^{1}$ A.R.Gower, Census and Household Surveys Methods Division, Statistics Canada.

[^4]:    2 The Household Record Docket provides a record of all persons (i.e. household members) found in a selected dwelling for the period that the dwelling is in the Labour Force Survey and contains information which helps interviewers plan their assignments and conduct interviews.

[^5]:    ${ }^{4}$ The LFS sample consists of six rotation groups, each of approximately equal size. Every selected dwelling belongs to one of these rotation groups and remains in the survey for six consecutive months. In any one month approximately one-sixth of the sample rotates out and is replaced by dwellings rotating into the sample for the first time (for example, a dwelling which rotates into the survey in January is enumerated each month from January to June and is replaced by another dwelling in July).

[^6]:    5 Every primary sampling unit $i n$ an NSRU is divided into an urban and a rural portion.
    6
    SRU's are stratified into sub-units, and sub-units are classified as "built-up" or "fringe" on the basis of their potential for future growth. Generally speaking, SRU fringe households belong to the fringe or suburban areas.
    In seventeen large cities across Canada there is a separate frame of apartments having at least five storeys and thirty or more units.

[^7]:    ${ }^{1}$ Adapted from a paper presented at the Annual Meeting of the American Statistical Association, Section on Survey Research Methods, 13-16 August, 1979, Washington, D.C., U.S.A.

    2 A. MacLeod and K.P. Krótki, Census and Household Surveys Methods Division, Statistics Canada.

[^8]:    1 The EA is a spatial unit composed of a cluster of geographically contiguous households and assigned to one enumerator. In 1976, Canada was divided into 35,154 EAs.

[^9]:    1 In the 1976 Census, basic demographic data and mother tongue were collected on a $100 \%$ basis. Data on migration, education and labour force activity were based on a $1 / 3$ sample.

[^10]:    1
    G.H. Choudhry, Census and Household Survey Methods Division, Statistics Canada.

[^11]:    * The iteration procedure was terminated when the change in the value of each of the elements of the probability vector was less than or equal to $1.0 \mathrm{E}-8$ in magnitude.

[^12]:    ${ }^{1}$ Presented at the Annual Meeting of the American Statistical Association, August 13-16, 1979, Washington, D.C., U.S.A.
    2 M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada.

[^13]:    ${ }^{\text {IM.D. Bankier, Census and Household Survey Methods Division, }}$ Statistics Canada.

