ISSUES INVOLVING THE DESIGN AND ANALYSIS OF BUSINESS SURVEYS

BY

JULIE Z. MOJICA

A Practicum
Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

Department of Statistics
University of Manitoba
Winnipeg, Manitoba

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES
*****
COPYRIGHT PERMISSION PAGE

ISSUES INVOLVING THE DESIGN AND ANALYSIS OF BUSINESS SURVEYS

BY

JULIE Z. MOJICA

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University

of Manitoba in partial fulfillment of the requirements of the degree

of

Master of Science

Table of Contents

List of Tables

List of Figures

Abstract

Business surveys are generally conducted by government institutions to provide data necessary for determining key economic indicators to be used in economic monitoring and in the construction of official statistics. In the conduct of these surveys, collecting agencies normally encounter numerous survey issues related to relevance, accuracy, reliability, clarity, timeliness, ease of interpretation and production at reasonable costs. This Practicum intends to discuss some of the issues encountered namely: creation of business frames, definition of basic units being surveyed, sample design and sample allocation, estimation procedures of parameters, data collection and handling of missing data, data editing and imputation and finally, attainment of high quality survey outputs. Due to the complexity of the organizational structure of business units, the skewed distribution of units in the business population, and the volatility of business units over time, there is a need to address these issues in order to help survey designers in developing and conducting cost-effective business surveys.

Acknowledgements

The author wishes to express her sincere gratitude to the King of Kings and Lord of Lords, Jesus Christ, for making available the most valuable assistance of following persons:

Dr. Brian D. Macpherson, her advisor and chairman of the guidance committee, for his valuable and brilliant suggestions, continuous guidance and support all throughout the preparation of her manuscript. Her committee members, Dr. Smiley W. Cheng of the Department of Statistics, University of Manitoba and Mr. Wilf Falk of the Manitoba Bureau of Statistics, for taking time out of their busy schedules to read and give helpful suggestions and comments on her manuscript.

Mr. David Greenwood of the Manitoba Bureau of Statistics for providing some pertinent materials and related references on the study.

Her husband Joel, her daughter Deborah, her son Gideon, her parents and sisters for their love, care and inspiration. Lastly, friends, for their encouragement and prayers.

Chapter 1

Introduction

Progress has always been the quest of civilization. The desire to improve the quality of life has spurred many to develop new ideas and improve existing ones. This is readily observable in the free world where capitalism abounds. Due to the consuming motivation for profit coupled with the consistent passion of developing something new and advanced, which is intended to make life more convenient, numerous types of businesses have come into existence. The private sector has always been known for its agility in responding to every possible business opportunity, hence the abundance of privately owned enterprises. Businesses form the backbone of the economy of any country.

In contrast, with the exception of some capitalist countries, the government has the duties of regulation, policy making and governance. These duties are being performed for the improvement of the standard of living of its constituents situated in a given geographical location. The performance of their duties, particularly policy making, depends on intelligent decisions and intelligent decisions depend on the availability of high quality information. Information could be gathered effectively and efficiently through the use of business survey methodologies developed specifically in the environment where these methodologies are expected to function. This task is primarily performed by government statistical agencies.

The Manitoba Bureau of Statistics (MBS) is the central statistical agency for the province. Its basic role is to meet the statistical requirements of the Manitoba government departments and crown agencies as well as other clients in the business sector and the general public. Under a Federal – Provincial Agreement, it is the designated interface between the province of Manitoba and Statistics Canada. As an interface between the province and Statistics Canada, MBS transforms basic data into a wide range of "ready-to-use" products designed to meet the information needs of its clients. Recently, MBS saw the need of providing its constituents with more detailed inter-provincial information on some of Statistics Canada's on-going business surveys. MBS intends to have more detailed economic statistics than what Statistics Canada provides. It would be interesting to know greater detail about inter-provincial trade flows of goods and services in Manitoba. Moreover, MBS would also like to know the details of the export and import activities of Manitoba businesses. It is also interested in gathering information on any issues or concerns related to Manitoba businesses which would become a vehicle or basis for policy making of the provincial government offices and/or private agencies. It is under this situation that MBS has expressed its need for data specifically gathered from the various regions and industries of Manitoba for Manitobans. This would only be possible through the development of an Omnibus Business Survey Methodology for Manitoba which would enable MBS to better pursue its mandate.

1.1 Objective

The objective of this practicum is to consider issues concerning survey methodologies for business surveys of any government collecting agencies. These issues may be beneficial to the Manitoba Bureau of Statistics in the development of omnibus business survey methodologies and of model future surveys to be conducted in the province.

1.2 Definition

“Business” as a term has no unique and universal definition. In a general sense, business refers to an economic unit engaged in the production of goods and services (Colledge, 1995). This definition covers commercial and industrial enterprises, departments of government, farms, institutions and non-profit organizations. Statistics Canada defines a business entity as: “an economic transactor having the responsibility and authority to allocate resources in the production of goods and services thereby directing and managing the receipt and disposition of income, the accumulation of property, borrowing and lending, and maintaining complete financial statements accounting for these responsibilities” (Cox and Chinnappa, 1995).

Another broad definition of business refers to organizational entities and hence a business survey refers to the study of the characteristics or attributes of these organizational entities. The unit of interest in business surveys is the organization and its attributes, and not the individuals within the organization. The individuals are merely spokespersons for the organization. There are various types of businesses under this broad definition (Cox and Chinnappa, 1995). They are as follows:

- "Businesses: retail and wholesale stores, manufacturers, construction companies, mining operations, financial institutions, transportation companies, public and private utilities, service providers, etc.
- Farms: crop and livestock operations, agribusiness, vineyards, family farms and ranches, plant nurseries, cooperatives, etc.
- Institutions: schools, prisons, courts, hospitals, local governments, professional and trade associations, etc."

Business surveys are conducted to provide data necessary for determining key economic indicators used in economic monitoring and for constructing official statistics. Cox and Chinnappa (1995) described the distinct characteristics of business surveys as follows:

- The size of businesses does not follow a normal distribution; instead they tend to have very skewed distributions composed of a few large businesses and many small businesses in a given geographical area.

- Business surveys in some cases require quick estimators/indicators that reflect the rapid rate of change in the economy at any given time.

- Business surveys have access to alternative sources of data for businesses (e.g. administrative records) which can be used for data editing and data imputation.

- Business surveys should provide information that is consistent across surveys and must conform to the definitions and requirements of the System of National Accounts (SNA) as in the case of the United States.

- Business survey estimates can be validated using other sources in the future which give reliable aggregated estimates.

As in any sample survey, the planning and execution of business surveys involve the following procedure:

1. A clear statement of the objective.
2. Definition of the target population and the elements belonging to it.
3. Identification of the variables of interest to be measured.
4. Identification of the sampling unit and establishment of a list of the sampling units called the frame.
5. The method of selection of the sampling unit in the frame.
6. Design of the questionnaire.
7. A well-defined measurement plan in the gathering of data.

8. The plans for handling nonresponse.

9. The sample estimation procedures.

10. Data analysis, summarization and reporting of results.

11. The review and evaluation of the present survey in order to improve future surveys.

There are issues associated in each of the above-mentioned topics that are peculiar to business surveys. These issues will be addressed in this Practicum.

Chapter 2

Frame and Business Register

The primary goal of a survey is to gather information about populations in order to satisfy a need for specific information. Population is a group of units defined according to the purpose of the survey. In order to gather information from the predetermined population, there is a need for certain procedures, devices, lists, maps, or other acceptable materials which will serve as a guide. These materials are called frames. This chapter will cover the issues relevant to the establishment of frames and deal with the following topics: target and sampled population, survey frame, sampling unit, statistical unit and the structure of the Manitoba business register.

2.1 Target and Sampled Population

The target population is a collection of elements about which an inference is to be made. For instance, the target population in business surveys is composed of all units of businesses located at a given region at a particular point or period of time. However, it is often not possible to gather an exact or complete aggregate of these units because of its large size. This leads then to the concept of sampled population. The sampled population is composed of units from which the sample is actually selected. Ideally, the sampled population should coincide with the target population but this does not always happen in practice. For this reason, statistical agencies conducting surveys spend a great deal of time and effort to make the sampled population as close as possible to the target population. For example, the target population for an annual survey of wholesale trade could be all business engaged in wholesale activities at any time of the year in Manitoba. In this particular case, the sampled population could be all business involved in wholesale trade which are registered in the province for tax purposes.

2.2 Survey Frame

The population to be sampled should have a finite number of identifiable and distinct units called sampling units. The device used to identify the units in the population is called a survey frame. The frame provides the basis for selection and identification of the units to be included in the sample. In business surveys, the sampling unit is the unit that defines the business entity. Unlike in household or social surveys where the sampling unit is a natural unit such as a person or a household, this is not the case for business surveys. The sampling unit for a business survey is an artificial unit and is usually defined in terms of the type of data to be collected for the survey (Cox and Chinnappa, 1995). Businesses 1) vary in sizes, ranging from small businesses to large businesses, 2) vary in organizational structures (e.g. legal, operational, administrative, accounting, etc) from simple to complex, 3) may engage in one or more activities in order to produce goods and services, 4) may operate in one or more geographic locations, 5) may change in composition due to splits, merging, growth or expansion, or 6) may change in existence as a new business (births) or business that have ceased operations (deaths) or business that have resumed operation after being out of business for some period of time. There are issues associated with the above mentioned characteristics of businesses which make the definition of unit in a business frame complicated and difficult. These issues will be discussed in Section 2.3.

The structure of the frame and the information it contains will determine the type of sampling designs and the estimation procedures that can be used in the survey (Lessler and Kalsbeek, 1992). Colledge (1995) described The information which must be included in a survey frame. They are as follows:

- Identification data. – These are the unique identification characteristics of each unit such as name, address and identifier.
- Classification data. – These are the variables used for the classification of each unit based on certain characteristics such as size (which can be defined

in terms of employment, assets, revenues, profits, etc.), industrial or regional classification, etc.

- Contact data. – These are the information items needed to locate the sample such as contact person, mailing address, telephone number, etc.
- Maintenance and linkage data. – These are the information items related to the changes in the unit or linkages to other surveys.

Collectively, these data are referred to as frame data.

There are two types of survey frames encountered in business surveys, namely, list frame and area frame.

2.2.1 List Frame

List frame is a list of all units and all the associated frame data which identify the sampled population. In the list frame, the units are the business entities themselves which will be discussed in detail later. Examples are 1) an administrative list such as the list of businesses registered for value added taxes (VAT), or a register of employers making payroll deductions or unemployment insurance payments for employees and 2) a commercial list such as a list of businesses registered by Chambers of Commerce or by telephone or electricity utilities. The business survey may use either a single list frame or combination of list frames. An example of an agency which uses a combination of two list frames is the Survey of Employment, Payrolls and Hours (SEPH) of Statistics Canada. It uses the 1) Business Register (BR) maintained by the Business Register Division of Statistics Canada and 2) a list of all payroll deduction accounts maintained by Canada Customs and Revenue Agency (CCRA) (Statistics Canada, 1998a). In business surveys, the administrative list is usually used as the starting point in the construction of the list frame.

The advantages of using a list frame in business surveys are as follows:

- List frames are effective in identifying large and rare units in the sampled population (Bush and House, 1993).
- Efficient sample designs can be developed because of the available information associated with each unit in the list for example, known measures of employment size (Colledge, 1995).
- List frame units are less expensive to develop and can be generated by mail, telephone or electronic medium (Kott and Vogel, 1995).

On the other hand, the disadvantage of using a list frame in business surveys is the incomplete coverage of the population units at any given time because of the rapid changing nature of the business populations. Businesses come and go at any period of time thus causing the frame to become outdated quickly and therefore made incomplete. There is also the issue of obtaining "identifier" information for a business entity such as nature of business, employment, revenue, organizational structure.

2.2.2 Area Frame

The term area frame refers to the collection of non-overlapping geographical areas from which areas are selected. Then within the selected areas, all the associated units are enumerated. In contrast with list frames, the sampling units are area segments and the businesses in the population are then linked to the area segments in the frame to which they belong.

In list frames, units are selected directly from the list using stratified single-stage sample designs but in area frames, the units are often selected using stratified multi-stage designs (Kott and Vogel, 1995). For instance, in the U.S. an area frame is being used by the U.S. Department of Agriculture (USDA) National Agricultural Statistical Service (NASS) to provide estimates of crop acreages, livestock inventories, farm expenditures, farm labor and other similar agricultural items (Bush and House, 1993). The area sampling from the area frame is done in a two stage design where the primary sampling units (PSUs) are parcels of land averaging 6 to 8 square miles in size. The sampling process selects the PSUs and the selected PSU's are divided into segments generally one square mile in size which have natural boundaries and may have irregular shape. The secondary sampling units are the segments selected from each PSUs. Similarly, a two stage design for area sampling is being used by the USDA's Soil Conservation Service for the National Resource Inventory Survey to provide reliable data on land use, erosion, conservation treatment needs and other conservation issues at various nonfederal lands in mainland US, Hawaii, Puerto Rico, and the US Virgin Islands (Bush and House, 1993).

In Canada, an area frame is being used by the Agricultural Division of Statistics Canada in its Farm Establishment Survey to provide agricultural statistics. The design of area samples is being revised from a two stage design to a single stage design. The two stage design used the "Census of Agricultural Enumeration Areas" as the PSUs. The selected PSU's were divided into segments of about 6 to 10 square kilometers using natural boundaries. A second stage sample of segments was then selected, in this case only one per PSU. In the new single stage design, Statistics Canada uses the Universal Transverse Mercator projection to divide Canada into 3 x 2 kilometers rectangles or cells. The boundaries of these cells and the Census of Agricultural Enumeration Areas are overlaid. Cells that overlap the Agricultural Enumeration Areas form the population of area segments. The single stage sample is drawn from this population of area segments.

Area frames have been useful as a base of sampling in rural areas where the farm establishments are the units of interest. It is also useful in surveys that provide basic statistics on agriculture and ecological resources that are related to land area.

The advantages of using an area frame in business surveys are the following:

- Complete coverage of population units can be ensured since current and future units are contained within an area segment in the frame (Kott and Vogel, 1995).
- It has a long life span and need not be updated frequently unless dramatic changes happen in the geographical features of the units within (Bush and House, 1993).

However, the area frame also has some disadvantages. They are as follows:

- Area frames are useful when the units are fairly evenly distributed geographically or when the size (in terms of the number of units) of the area segments are available or can be estimated accurately. This is generally not true for business populations because businesses vary in size and may be sparsely distributed geographically (Kott and Vogel, 1995).
- Sampling design may not be efficient because of the clustering of units in the area segments (Cox and Chinnappa, 1993).
- More costly to develop because names and addresses of businesses are usually not available and need to be enumerated personally (Kott and Vogel, 1995; Bush and House, 1993) .

- Small businesses which are not visible or have no outward sign of business activity such as those operating in private homes are likely to be missed in an area frame and so bias the estimates (Hirschberg and Nisselson, 1993).

2.2.3 Multiple Frames

The concept of the multiple frame survey involves combination of several frames such as an area frame combined with one or more list frames. This type of frame is commonly used in agricultural surveys for example where a list of farms is available but the list is incomplete due to changes of ownership or other such reasons. The area frame is added to provide complete coverage of farm production. The usefulness of multiple frames is based upon the assumption that the units in the target population should belong to at least one of the frames and that for any sampled unit, the frame where it belongs is identifiable (Kott and Vogel, 1995).

However, there are issues associated with multiple frame surveys. They are as follows:

- Identifiability. – There is a need to know whether a sampled unit from any one of the frames belongs to any other frame. Businesses need to be identified by primary name, address and other relevant information. In the case of large businesses, which have complicated structure and/or are situated in more than one location, such information should be clearly indicated. It is important to also note other business names and names of individuals associated with the business in order to avoid duplication of the same sample unit in the population. With so much information needed to uniquely identify the sampled units, additional cost would be required to create a final frame consisting only of unique units.

- Record Linkage. – With multiple frames, the survey designer should decide whether to use each frame separately or combine the frames to form a single list frame for sampling purposes. If they are to be combined, there is a need for some matching and record linkage methodologies to match the units from each of the frames. The matching and record linkage procedures include the specification of the variable identifiers to be used in matching, the criterion or rules for matching, the algorithms to be followed for matching and decision making when inconsistent situations arise. The use of computers facilitates matching, centralizes and speeds-up the processing, provides better quality control, easily reproduces results and reduces manual review of matching and record linkage. The essence of all these efforts is to match the same unit, the business entity, regardless of different names. There are various rules being followed in matching names between any two frames and Winkler (1995) discussed the details of these methodologies. Questionable links should be checked and, if possible, corrected because if not attended to, this would result in large non-sampling error.

- Estimation Difficulty. With multiple frames, the survey process and the estimation procedures become difficult and complicated.

An example of an agency that used a multiple frame survey is the U.S. Bureau of the Census Sample Survey of Retail Stores. It was the first business survey which used the dual frame methodology by combining a single list frame and single area frame but its continued use has been investigated (Kott and Vogel, 1995). On the other hand, in Italy, the use of dual frame sampling design for establishment surveys has been introduced and its usage is being continued (Petrucci and Pratesi, 1993). The preliminary results of the study showed better survey estimates as compared to the use of area frame sampling design .

2.2.4 Survey Frame Errors

Whether involving a single frame or multiple frames, a good survey frame should provide a complete, exhaustive and up-to-date representation of the target populations for any survey, even a business survey. But this is not always possible. Survey frames often have deficiencies or imperfections that can introduce bias into the survey estimates. The common errors encountered in the creation of frames and their associated effects on the outcomes of the surveys (Lessler and Kalsbeek, 1992; Willeboordse, 1998a) are as follows:

- Missing population elements; also referred to as undercoverage, noncoverage or incomplete coverage. This means that some members of the target population are not included in the survey frame. Failure to recognize and correct the missing population elements may cause the totals of the variable of interest to be underestimated and so bias other statistics. This is the most serious problem of such a frame error because it cannot be recognized from the sample or from the frame. An example of such a situation is where self-employed businesses, newly established businesses, or new establishments of existing businesses may be missing from the list frame of all businesses.

- Inclusion of non-population elements or overcoverage. This means that the frame contains elements that are not part of the target population and hence should not be in the frame. If such errors are not corrected, the survey variable totals will be overestimated and again other statistics may be biased as a result. The potential for bias is less compared to undercoverage because these non-population elements can be easily recognized as being such and hence can be eliminated during the survey. An example of this are businesses which no longer exist but have not been removed and are still present in the frame.

- Multiplicity problems or duplication of listings. This arises when the elements of the target population are linked to more than one frame unit. If the duplicate units are not removed, this will affect the inclusion probabilities of sample elements and consequently cause an upward bias in the survey estimates.

- Incorrect auxiliary information. Some examples of auxiliary information are the classification data used for stratification, measures of size, such as employment size or revenue, used for probability proportional to size sampling and characteristics highly correlated with the survey variable of interest which can be used for ratio and regression estimation. Incorrect information causes a decrease in the precision of the survey estimates.

- Incorrect accessing information. This refers to the out-of-date or inaccurate listing of units within the frames. The characteristics of the frame units such as contact person and addresses have not been updated and so the unit cannot be located. The target elements are known to exist but cannot be accessed. The effect of such error is similar to the problem of undercoverage.

The survey designer should be aware of these frame imperfections and should undertake some actions to reduce the discrepancies between the target population and the frame population. As much as possible, the survey frame should contain minimal errors of omission, inclusion and duplication of units and the information about the frame units (e.g. identification numbers, classification variables, addresses, and contact persons, etc.) should be of high quality.

Lessler and Kalsbeek (1992) gave some suggestions for factors to consider when deciding as to which type of frames to use. These include:

- the amount variability or bias that would result from frame errors encountered,

- the cost and time to create a quality frame,

- the frequency of usage of the frame,
- the type of design and estimation procedures applicable to the frame and the cost of carrying them out, and
- the type of information to be collected, the cost and its quality.

But most importantly in the creation, use , maintenance and monitoring of survey frames, the operational and cost constraints of the survey are the primary considerations. When limited funds are available, maintenance efforts must be concentrated on the large businesses to ensure their critical information is indeed correct because their impact on the survey estimates is significant as compared to the small businesses.

2.2.5 Administrative Sources

With budgetary constraints in mind, the creation of a survey frame from administrative sources provides a suitable alternative to the building of a frame from scratch for any type of survey. This is certainly true for business surveys. Administrative sources are lists that are maintained for administrative or commercial purposes. Government agencies, organizations or departments are collecting data related to individual businesses and are maintaining their own registers for specific administrative purposes or for serving other administrations or government offices. With such administrative data in existence, the costs and burden of direct data collection from businesses can be avoided (Colledge, 1995; Statistics Canada, 1998b; Boegh-Nielsen, 1998; Perry, 1993; Harala, 1998; Hostrup-Pedersen, 1993). Administrative sources have been found useful not only in the creation of frames but also in the maintenance of frames. Data from administrative sources can also be used in editing, imputation and weighting of survey data and in the evaluation of statistical outputs for checks or quality control.

One concern regarding the use of existing administrative sources is on the issue of privacy and confidentiality of information in the public domain especially when these records are linked to other sources of data (Statistics Canada, 1998b). It is important that the surveying agency be ready to give explanation and justification for such secondary use. But in the case of Statistics Canada, the Statistics Act gives it the authority to access administrative records for statistical purposes.

There are some issues regarding the use of administrative sources for the creation and maintenance of a business survey frame. These are:

- Administrative units may not correspond to the statistical units because of differing concepts or differing identification numbering system (Boegh-Nielsen, 1998; Perry, 1993).
- The classification data (e.g. industry code) from the administrative source may not be in a suitable form and do not follow the standard industrial classification codes (Colledge, 1995) . The codes may be outdated or be based upon the old SIC coding, for example.

- There is less control on the quality of incoming data from administrative sources because the manner by which information are put together is in the hands of the administrative organization (Statistics Canada, 1998b).
- In terms of timeliness , the availability of data from the administrative source may not coincide with the surveying agency's schedule resulting in an unacceptable delay for the data.
- If there is a need to use more than one administrative source for coverage sufficiency, matching and consistency problems may arise because sources don't share a common identification numbering system (Colledge, 1995; Statistics Canada, 1998b; Boegh-Nielsen, 1998; Perry, 1993). While there are a variety of record linkage techniques available that are automated, fast and inexpensive, doubtful links need to be checked manually which becomes an added cost in operation.

With all these shortcomings, the survey designer must be aware and be prepared to implement appropriate measures to overcome these deficiencies and to safeguard quality in the use of administrative data as best as one can.

The advantages of using administrative sources outweigh the disadvantages. Statistical offices in many countries have used and recommended this approach in creating a business statistics register. Statistics Canada Business Register is using CCRA administrative data as its base source for business accounts, supplemented by updates from the Statistics Canada Nature of Business Survey which collects information on new businesses (Statistics Canada, 2002a). The Inter-Department Business Register (IBDR) of the United Kingdom is using both the records of traders registered for Value Added Tax (VAT) and employers registered with the Inland Revenue for Pay- As-You-Earn (PAYE) tax purposes (Perry, 1993). Statistics New Zealand's register uses the value-added tax data, Statistics Netherlands uses data from the Chamber of Commerce while the U.S. Bureau of Census is using payroll deduction data (Colledge, 1995). The Australian Bureau of Statistics (ABS) Business Register uses data from Australian Taxation Office (Australian Bureau of Statistics, 2001). Denmark's Central Register of Enterprises and Establishments uses data reported to the Ministry of Taxation, Central Customs and Tax Administration (Hostrup-Pedersen, 1993).

Further discussion about business registers will be presented at the end of this chapter.

2.2.6 Recommendations

Having considered the advantages and disadvantages of using list frames, area frames and multiple frames as well as taking into account the issues associated with each of these frames and of using a variety of administrative sources, it is recommended that a list frame be utilized in conducting surveys in the province of Manitoba. A list frame is recommended for the reasons that follows.

- Availability of a Business Register from administrative sources. – Statistics Canada maintains a Business Register which contains a comprehensive and a complete-as-possible listing of statistical establishments (i.e. the smallest unit which defines a business entity) which are categorized by economic region, industry and employment size. This information is available per province. Since Manitoba Bureau of Statistics (MBS) has a data sharing arrangement with Statistics Canada, creation of a list frame containing Manitoba businesses will be easy and quick.

- Access to up-to-date information. – The Business Register is constantly updated by the Business Register Division of Statistics Canada. It keeps a record of businesses which come and go. As a result, information contained therein is always up-to-date. A new file is received every six months by MBS.

- Less expensive to develop. – Since there would be no need to start from scratch in establishing a survey frame due to the availability of the Business Register maintained by Statistics Canada, this may well result in modest or low budgetary requirements.

2.3 Statistical Unit

A sampling unit is the unit on which observations are made and from which data are collected. In social surveys, the sampling unit is usually a person or household. In business surveys, it is often difficult to decide which units within a business, particularly in large businesses, about which and from which data will be obtained (Colledge, 1995). A small business may have only one structure to carry out its legal, administrative and operations functions. A large business may have an organizational structure where the various functions are carried out separately by different sets of units within the business. Each level of the organizational structures may be maintaining their own bookkeeping and accounting system and hence may be capable of reporting a variety of business data.

For statistical purposes, the unit of observation or measurement for which data are obtained or collected is called a statistical unit. Within a business, there could be more than one type of statistical unit depending on the survey objectives and the type of data being collected. Examples of business data are production, commodity, employment, financial, capital expenditures and so forth. The collection scheme for the different types of data may vary. For example, production data such as revenue earned from sales of goods and services, employment, wages and salaries may be available for each physical location of the business. Financial data such as consolidated profit or loss accounts, balance sheets of assets and liabilities could well be known only for the entire business. The different levels and hierarchy in the organizational structure of a business may give rise to different levels and hierarchy in the statistical units as well. The different types of statistical units and their definitions will be discussed in sub-section 2.3.1.

The unit reporting the data may not necessarily be the same as the statistical unit. This happens in cases where a business has some hierarchical structures and records are maintained not at the lowest level but at a higher level of the business. For example a business retailer may be operating at three locations and the owner's bookkeeping office handles the completion of the survey questionnaire for each location separately. The bookkeeping office is the reporting unit. Reporting unit is more of a contact address or person responsible for completing the forms or questionnaire. Reporting unit is also referred to as collecting unit.

The are some issues related to the definition of the statistical unit in a business survey (Nijhowne, 1995; Colledge, 1995). Such definitional issues include:

- Varied sizes of business - A small business engaged in a single industry at a single location can be regarded as a single statistical unit. For large businesses, which are engaged in a range of activities and/or are situated in more than one location, defining of statistical units becomes a problem. One solution suggested to handle this problem is to employ a process called profiling. Profiling involves making a personal contact with a specific business with the intention of getting in-depth information about the organizational structure of the business (Archer, 1995; Pietsch, 1995). This process involves additional costs, time, labor and manpower resources. The positive consequence of its use is an improved sample frame.

Another strategy being used by Statistics Canada to gather detailed information on new businesses is to conduct its Nature of Business Surveys.

- Different types of data may require different collection schemes. This issue has been mentioned earlier.
- Operational, accounting and legal structures. A business may be structured according to production operations, administrative and management, accounting systems, legal structures, geographical locations and so forth. In small businesses, all of the above mentioned structures are considered one and thus can be regarded as a single unit. For large businesses, the operational structure is often different from the legal structure and so the units of the business operational structure differ from the units of the legal structure.
- Multiplicity of activities to produce goods and services. A production process may involve a single process or series of integrated processes (horizontally or vertically). An example of horizontal integration is a business composed of several unrelated divisions of businesses such as transport, hotels and electronic communications. An example of vertical integration is a business composed of several related divisions of businesses such as printing division and bookbinding division.
- Geographical locations at and from which the business operates may vary from single physical location to multi-locations. An activity producing goods

takes place at one or more physical locations. On the other hand, an activity producing services, like a construction company or engineering consultants, may take place from a specific location related to the client and not from the main office of the business.

2.3.1 Standard Statistical Unit

Business surveys are expected to produce data in the form of economic statistics that will describe the behavior and activities of businesses including all transactions undertaken. The data collected from individual businesses need to be integrated across businesses. With the varying sizes, organizational structures, ranges of activities and geographical locations of businesses, data integration will be a problem. Thus, there is a need to have a standardized statistical structure for collecting survey data. In this way, the data to be collected would be consistent, comparable and can be integrated across businesses and even across surveys.

Statistical units play an important role in the production of economic statistics. They serve as building blocks for aggregation of micro-data into sectors such as industry, region or institutional sector. There is no international standardization of statistical units yet because national statistical agencies across countries are still using different definitions, names and concepts to describe different statistical units (Nijhowne, 1995). Some countries like US, Canada and Mexico have been working together for standardization of statistical units. However, national statistical agencies have agreed and recognized at least two distinct standard statistical units associated with the two levels of the business organizational structure, namely 1) the level at which financial decisions are made and 2) the level at which production decisions take place. Therefore, the two main types of data that require two distinct standard units are a) financial data and b) production data (Colledge, 1995).

Before mentioning the appropriate statistical units for each type of data, the various types of statistical units will be discussed first. These are 1) Enterprise, 2) Kind-of-Activity Unit (KAU), 3) Local Kind-of-Activity Unit, 4) Establishment, 5) Location, 6) Ancillary Unit, and 7) Global Enterprise. Nijhowne (1995) showed how these type of units are related hierarchically with each other in a diagram (Figure 1).

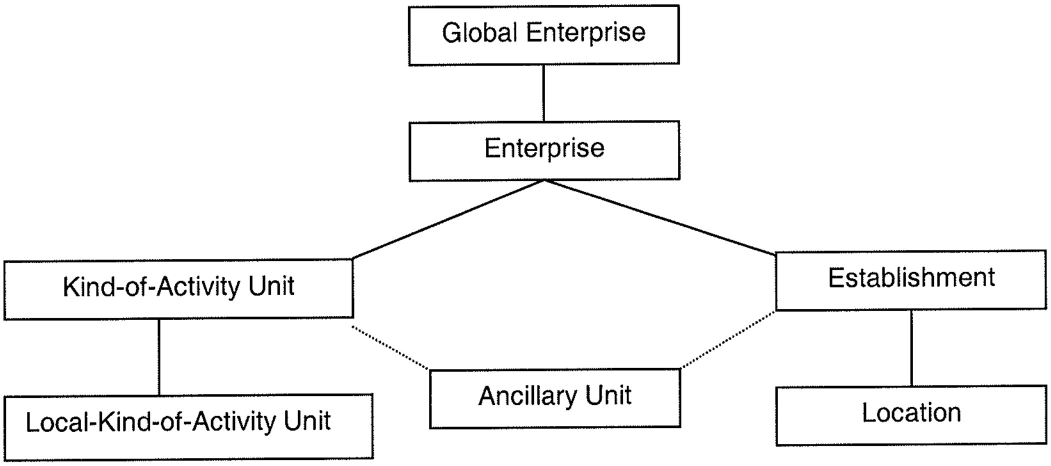

Figure 1. A diagram depicting the logical hierarchical relationship among types of statistical units. Source: Nijhowne, S. (1995). "Defining and Classifying Statistical Units". In *Business Survey Methods*, B.G. Cox et al (eds), ISBN 0-471-59852, p. 54.

- Global Enterprise. Global enterprise is the highest unit in the hierarchy. Another term for it is Enterprise Group (Willeboordse, 1998b). It is made up of enterprises with common ownership and is bound together by legal or financial links. The term global indicates that this unit is not confined to the national boundaries of a country. The data for the global enterprise are derived from enterprises under its ownership or control and thus, a separate unit is usually not needed on the business register (Nijhowne, 1995).

- Enterprise. Enterprise is defined as "the organizational unit of the business that has autonomy with respect to financial and investment decision making as well as authority to allocate resources for the production of goods and services (Nijhowne, 1995)". In the Canada Survey of Employment, Payroll and Hours (SEPH), enterprise is defined as "any business or institution whether incorporated or not; it comprises sole proprietorship, partnership, companies and other forms of organization. An enterprise is considered to be simple if all its establishments operate in the same province/industry classification; otherwise, an enterprise is classified as complex (Statistics Canada, 1998a)." The Annual Survey of Manufactures of Canada defines enterprise as "a company or a family of companies which is a result of common ownership and are controlled or managed by the same interest (Statistics Canada, 1977)". Willeboordse (1998b) defined enterprise as "the smallest combination of legal units that is an organizational unit producing goods or services, which benefits from a certain degree of autonomy in decision making, especially for the allocation of its current resources".

An enterprise can consist of one or more legal units and can own or control one or more KAU's or establishments. Thus, an enterprise can engage in one or more activities at one or more locations (Willeboordse, 1998b). The financial position and status of the entire business are normally determined at the enterprise level. Therefore for financial statistics, enterprise is the recommended standard statistical unit (Nijhowne, 1995).

- KAU or Establishment. The next level in the hierarchy of the statistical structure is the KAU or establishment where the production activities take place and are being managed. Both should carry out a principal economic activity although secondary activities belonging to other classes of industry classification are also allowed (Nijhowne, 1995). The essence of KAU or establishment is to group all units of the enterprise contributing to the performance of an activity at the same level of industry classification (Willeboordse, 1998b). Thus, an enterprise can be divided into two or more KAUs or establishments depending on the degree of differences between or among the production activities and on availability of separate data accounts. The data accounts include the operating revenues, operating costs, commodity inputs and outputs, staff costs, employment, operating surplus and value added. An example given by Willeboordse (1998b) is an enterprise that produces coffee and milk. Obviously, the production processes for these two products are very different. If separate accounts are available, then the enterprise is split into two KAUs.

For production data, the KAU or establishment is the recommended standard statistical unit. The unit of choice depends upon the desired precision on the location of economic activities, completeness of accounting records and level of management responsibility (Nijhowne,1995). KAUs are production units whose activities are carried out at or from one or more locations. Establishment, on the other hand, has its production activities conducted at or from one physical location only. If emphasis is given on the location of the activity, the unit of choice is establishment. However, if location is less of a concern and the emphasis is on the level of the managerial responsibility with respect to production decisions, then the choice of unit is the KAU.

- Ancillary Unit. Ancillary unit refers to the unit within the enterprise which renders support services to the rest of the business. The Canada Survey of Manufactures refers to it as auxiliary unit. Examples are the support services provided by the head office, accounting department, computer department, sales department, warehouse, etc. There are costs associated with these ancillary activities which need to be accounted for. Nijhowne (1995) mentioned two issues about the ancillary unit namely, 1) whether the ancillary activities need to treated as a separate unit and 2) if so, what industrial or geographical classification will they be assigned. The general practice is not to treat these activities as separate units and their operating costs are simply added to the principal activity of a KAU or establishment. But when an ancillary activity is conducted at different geographic location from the principal activities of the enterprise, Nijhowne (1995) suggested to identify ancillary unit as separate unit. An industrial classification for the unit is to be chosen and should be done carefully. Then the assignment of the geographical location code to the unit follows. The ancillary unit may become useful for some purposes.

- Local Kind-of-Activity (LKAU) or Location. LKAU or location refers to a lower level unit under the KAU or establishment. In the case of a KAU where the activities are conducted at or from more than one geographic location or area, each geographic location is considered a LKAU. If the purpose is to compile production statistics for every geographic location, this lower-level-unit is useful. But Nijhowne (1995) pointed out that the need to identify each location as a separate unit may not be necessary if the specific information about geographical locations can be obtained from the KAU or establishment level.

It would be interesting to know how national statistical agencies of different countries define the statistical units for business surveys.

- Statistics Canada (2002a) uses a four level hierarchy of statistical units namely the Enterprise, the Company, the Establishment and the Location. Each type of unit is associated with a particular level of economic data. The enterprise is associated with a complete set of financial statements. The company is defined as "the organizational unit for which income and expenditure accounts and balance sheets are maintained from which operating profit and the rate of return on capital can be derived (Barfoot, 1993)". The data associated with the company level is the operating profit. The establishment is associated with the accounting data related to the production process. Example of production data are costs of principal inputs, revenues, salaries and wages. The location, the lowest in the hierarchy, is the producing unit at a single geographic location. It is used for example by the monthly survey of retail sales and inventories to provide employment data (i.e. the number of employees).

- The US Bureau of the Census Standard Statistical Establishment List (SSEL) maintains three organizational units namely Establishment, Legal Entity and Enterprise. Establishment is the basic building block of the SSEL and the establishments are linked to the Legal entities and to Enterprise (Hanczaryk and Mesenbourg, 1993). The Legal entity is defined as the organizational unit with assigned Employer Identification Number (EIN) by the Internal Revenue Services (IRS) and is used for tax reporting purposes. Here the legal entity is considered as a statistical unit.

- UK identifies two main statistical units in its Inter-Department Business Register (IDBR) namely Enterprise and Local Units (Perry, 1993). Local unit is defined as an individual site operated by the Enterprise. One or more local units comprised an enterprise.

- Australian Bureau of Statistics (ABS) uses the following statistical units: Enterprise Group, Enterprise, Management Unit, Establishment and Location. ABS defines these units as:

  "Enterprise group is the unit covering all the operations in Australia of one or more legal entities under common ownership and/or control. Enterprise is the unit covering all legal entities within an enterprise group. Management unit is the largest type of unit within an enterprise group which controls its productive activities and for which accounts are kept. Establishment is the smallest type of accounting unit within a Management Unit within a State or Territory, which controls its productive activities. Location is a site occupied by an establishment on a relatively permanent basis (Australian Bureau of Statistics, 2001)."

- Central Bureau of Statistics (CBS) Netherlands defines the statistical units as the Enterprise Group, the Enterprise, the Local Unit, the Kind of Activity Unit (KAU) and Local Kind of Activity Unit (LKAU). Except for the Local Unit, these units have been defined previously. Local unit is a sub-part of the enterprise in terms of geographical location, may consist of more than one legal units, may engage in more than one activity but is not an autonomous decision making unit (Willeboordse, 1998b).

2.3.2 Industry Classifications and Geographical Locations

So far, economic statistics are presented on a per statistical unit. The statistical units need to be grouped by similarity of characteristics namely, industry and geographical locations in order to have a useful and meaningful analysis of economic data.

Industrial classification refers to the allocation of the statistical units into homogenous groups based upon similarity of principal activities, input and output structures, production functions, processes and technology, and goods and services produced. Such classification provides data that would be useful in the study of the economic activity or commodity of an industry or groups of industries. The industrial classifications have hierarchical structures and each level would be useful for analytical purposes. The levels are represented by two, three or four digits code depending on classification level, that is primary hierarchy, secondary hierarchy and so forth. The details of these levels will be discussed in a separate section. There is already an established industrial classification system called 1980 Standard Industrial Classification (SIC) System which was adopted by Canada. In 1997, United States, Canada and Mexico developed a new system which is called North American Industry Classification System (NAICS). Other European countries have been revising their SIC systems as well. The focus of the revision is to have international comparability of the industry definitions and to come up with an international or multinational standard of industrial classification.

Geographical classifications refer to the allocation of statistical units based on geographical areas. Areas may be defined by administrative and political boundaries or by some economic concepts such as urban core and labor market areas. The areas also have hierarchical structures such as local, regional and provincial hierarchies. The geographical classification has been found useful for regional analysis of economic activities.

Both industrial and geographical classifications need to be revised periodically to remain relevant and analytically useful. In the case of industrial classification revisions every 10-15 years is generally recommended (Colledge, 1995; Mac Donald, 1995). On the other hand, revision in the geographical classifications depends upon the availability of new information on the changes in the administrative boundaries or labor market areas.

2.3.3 Recommendations

For the Manitoba omnibus business survey, the establishment should be used as the standard statistical unit since Manitoba's economy is already defined in terms of the number of establishments. Establishment is defined as "the smallest unit of a company (whether sole proprietorship, partnership, cooperative, corporation, etc.) that is a separate operating entity capable of reporting all elements of basic industrial statistics (Manitoba Bureau of Statistics, 2002)". This definition of establishment includes 1) a single establishment providing accounting information on a single plant or 2) an establishment operating in more than one locations but the accounting information for all locations are combined into one when reporting income or 3) firms operating on an inter-provincial basis and have available accounting information. The establishments will also be classified based on the major industrial activity and regional areas. The details of the Manitoba Business Structure will be discussed in the next section.

2.4 Manitoba Business Register / Structure

Statistics Canada has been maintaining a comprehensive business register which is primarily intended to become a central source of frames for all business surveys. A number of surveys, (such as the Retail, Wholesale and Manufacturing Surveys and the Survey of Employment, Payroll and Hours (SEPH)), are already using the register. For the Manitoba omnibus business survey, a register of Manitoba businesses can be derived from the Business Register maintained by Statistics Canada.

2.4.1 The Statistics Canada Business Register

The Business Register contains information reflecting the universe of businesses, incorporated and unincorporated, including government departments and institutions, non-profit organizations, religious organizations etc. in Canada (Barfoot, 1993). The register includes business entities meeting at least one of the following criteria: “1) having a workforce for which they submit payroll remittances to CCRA, or 2) having a minimum of $30,000 annual sales revenue, or 3) are incorporated under a federal or provincial act and are active Federal Corporation Tax filers (Manitoba Bureau of Statistics, 2002)".

The purpose of the register is primarily for sample design and survey operations. The business register is intended to provide a reliable and good quality survey frame for a wide range of surveys measuring economic data of various target populations (Mac Donald, 1995). Statistics Canada makes sure that the register's coverage is complete, accurate and up-to-date. Statistics Canada also ensures that consistent classification of statistical units is being achieved.

The register's information comes from Statistics Canada surveys and from administrative data of four CCRA programs, namely the incorporated tax accounts (T2), the goods and services tax accounts (GST), the payroll deduction accounts (PD) and the import/export accounts (Castonguay and Monty, 2000). In the register, each business entity is assigned a unique identifier referred to as Business Number (BN). This identifier was introduced by CCRA in 1994 and was implemented in all of the CCRA programs. The BN which replaced the GST/T2/PD account numbers in CCRA was used to combine and link data from various CCRA programs. The Business Register Division of Statistics Canada then receives the data according to BN identification.

On the Provincial Business Register Extract, each business entity has an associated statistical entity numbers namely 1) Statistical Enterprise Number is the enterprise number to which the statistical establishment record(s) are linked, and 2) Statistical Establishment Number is the unique record identifier for each establishment. The Statistical Enterprise Number identifies a group of statistical establishments with common ownership and control. The Statistical Establishment Number denotes the smallest operating entity capable of reporting all basic industrial statistics (Statistics Canada, 2002b).

The other characteristics included in the business register are the geographic area code, industry code and employment size class.

The geographical classification of a business entity or enterprise is determined using the postal code of the physical address of the business. The classification is changed and updated when the business changes its postal code. The Standard Geographical Classification (SGC) code is composed of Province or Territory code and Census Division (county) code and Census Subdivision (municipality).

The employment size classification is determined based on a model that estimates the number of employees. The model is generated from the data provided by employers to CCRA. The size range instead of exact size of employees is used for size classification. If the model detects a significant change in size , then the business is moved to a different size range and the size classification is changed and updated. Employment size ranges are done both at the enterprise level and at the establishment level. There is also an additional category called “Indeterminate” included in the employment size ranges. The “Indeterminate” category refers to establishments which have a workforce but do not maintain an employee payroll because the people included in the workforce are the business owners, family members, part time or contract workers.

The industry classification code is assigned by Statistics Canada based on the description provided by the business when registering with CCRA or based on the information from the Nature of Business Survey. When new information comes in from Statistics Canada survey programs, then the industrial code is changed and updated. The industrial classification system is discussed in detail in the next section.

The Business Register is continuously updated (Statistics Canada, 2002a). Included in the updating process is the identification of new businesses, businesses that undergo changes in structure (e.g. break-up, split-off, merged, take-over) and businesses that ceased activities. Any changes in name, address and size of employees for businesses are also included in the updates.

2.4.2 Industrial Classification System

Statistics Canada is using the new industrial classification system called 1997 North American Industry Classification System (NAICS) previously mentioned.

The NAICS replaced the Standard Industrial Classification (SIC) system. Statistics Canada developed the SIC in 1948 because of the need of the government to establish a comprehensive system for reporting economic data. The concept, terminology and groupings of industries were introduced. In 1960, standardization of the unit of observation was emphasized and the "establishment" became the smallest unit capable of reporting industrial statistics. In 1970, the industry groupings were revised to reflect changes in the industrial structure of the economy. In 1980, the revision focused on the links to the System of National Accounts (SNA). The 1980 SIC system classified the establishments into a hierarchical structure with four levels namely, the Division, the Main Group, the Industry Group and the Industry Class. SIC is supposed to be revised at 10 year intervals but in 1990, the revision was postponed to take into account the statistical need of the Free Trade Agreement and the need to develop an industrial classification common to Canada, Mexico and the United States for North America Free Trade Agreement (NAFTA) purposes.

NAICS classifies the establishments by type of economic activity. It aims "to facilitate the collection, tabulation, presentation and analysis of data relating to an establishment and to promote uniformity and comparability in the presentation and analysis of statistical data describing the economy (Statistics Canada, 2002c)". NAICS is a collaborative effort of the INEGI of Mexico, Statistics Canada and the United States Office of Management and Budget (OMB). Its mission is directed to making the industrial statistics produced comparable across the three countries of Mexico, Canada and the United States. However, differences in the national and institutional structures, and limitation in time and resources caused the NAICS structure to be not entirely comparable across the three countries, particularly at the individual industry level.

The NAICS structure has a five level hierarchy structure classifying the establishments from the broadest level to the most detailed level. Table 1 shows the five levels of the NAICS structure with the associated coding system and the number of classes.

Table 1. The 1997 NAICS Canada hierarchy structure, coding system and number of classes.

Level	Hierarchy Structure	Coding System	Number of Classes
1	Sector	2 digits	20
2	Sub-sector	3 digits	99
3	Industry Group	4 digits	321
4	Industry	5 digits	734
5	National Industry	6 digits	925

Source: Statistics Canada (2002).

The twenty sectors included in the 1997 NAICS Canada are shown in Table 2.

Table 2. The new NAICS Sectors and the corresponding number of industries.

Sector	Title	No. of Canadian Industries
11	Agriculture, Forestry, Fishing and Hunting	48
21	Mining and Oil and Gas Extraction	29
22	Utilities	10

23	Construction	36
31-33	Manufacturing	259
41	Wholesale Trade	78
44-45	Retail Trade	69
48-49	Transportation, Warehousing and Storage	58
52	Finance and Insurance	46
51	Information and Cultural Industries *	30
53	Real Estate, Rental and Leasing *	21
54	Professional, Scientific and Technical Services *	40
55	Management of Companies and Enterprises *	2
56	Administration and Support, Waste Management and Remediation Services *	34
61	Educational Services	12
62	Health Care and Social Assistance	37
71	Arts, Entertainment and Recreation *	31
72	Accommodation and Food Services	18
81	Other Services (except Public Administration)	38
91	Public Administration	29

*New sectors in the 1997 NAICS and not present in the 1980 SIC. Source: Statistics Canada (2002c).

2.4.3 Recommendations

It is recommended that the Manitoba Business Register adopt all the guidelines of the Statistics Canada Business Register but at a provincial level. The classification codes for the industry, employment size category by enterprise and establishment levels, and geographical area of the Statistics Canada Business Register should be followed. Establishments are assigned to an industry sector based on the 1997 NAICS. For the employment size category, the size ranges are 1-4, 5-9, 10-19, 20-49, 50-99, 100 & over and include an "Indeterminate" category (Manitoba Bureau of Statistics, 2002). In addition to the above, there will be a regional classification. There are eight official regions in Manitoba called Economic Regions. Each region is divided into one or more Census Divisions of which Manitoba has 23. Each Census Division is further divided into Census Subdivisions which may mean a city, town/village, rural municipality, district or Indian Reserve. The eight Economic Regions are listed in Table 3 and geographically shown in Figure 2.

Table 3. The eight economic regions of Manitoba.

Official Regions for Data Collection	Economic Region	Census Divisions in Region
Southeast	10	1,2,12
South Central	20	3,4
Southwest	30	5,6,7,15
North Central	40	8,9,10
Winnipeg	50	11
Interlake	60	13,14,18
Parklands	70	16,17,20
North	80	19,21,22,23

Source: Manitoba Bureau of Statistics (2002).

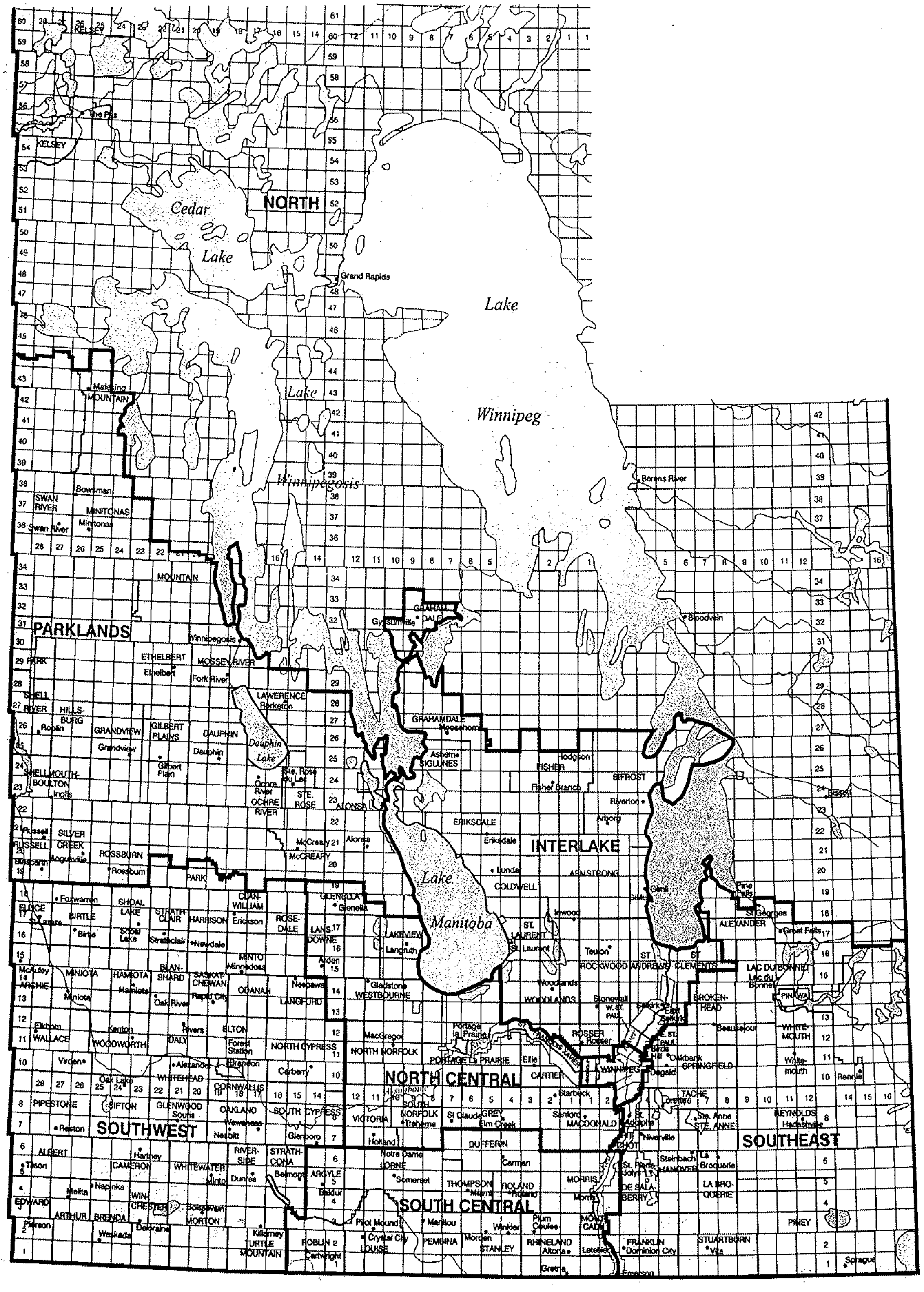

Figure 2. Map showing the boundaries of the eight economic regions of Manitoba. Source : Manitoba Bureau of Statistics.

Chapter 3

Sample Design and Selection

Sample design is a broad concept that describes both 1) the sample selection plan, such as definition of the sampling unit, number of sampling units or sample size, and method of selection of sampling units or sampling design; and 2) the procedures for estimation (Murthy, 1977; Sarndal et al, 1992). Both are important because the precision of the estimates is determined by the sample design. Estimates are the values assigned to unknown population characteristics or parameters. In the design stage, a combination of sampling design and estimation method is selected to produce estimates that will lie as near as possible to the values of the parameters. The choice is usually determined in relation to the required precision and available budget. The survey designer will search for methods that will give estimates of best possible precision under a given budget or of the lowest possible cost for a fixed precision. The choice of sampling design which best fits for a particular survey also depends on supplemental information or auxiliary information on the units in the frame or from other sources. The more auxiliary information that is available, the better the sampling design can be tailored to the survey goals (Koeijers and Hilbink, 1998a). This chapter will discuss only the sample selection plan. Procedures for estimation will be presented in a separate chapter.

The development of sample designs for business surveys is challenging because of the distinctive features of business populations as discussed earlier. The sampling design commonly used by many business surveys is stratified sampling (Sigman and Monsour, 1995; Koeijers and Hilbink, 1998a). In stratified sampling, the population is divided into non-overlapping sub-divisions or sub-populations called strata and then independent probability samples are selected from within each stratum. The sampling and estimation procedures may be the same or different from stratum to stratum depending on the information available. The allocation of sample size to the strata and the demarcation of strata boundaries may be done with reference to the specified sampling variability and survey cost.

Why is stratified sampling being used for business surveys? 1) The first reason is related to the skewness in the typically-used measures of size (e.g. number of employees, revenue or assets) of business populations. For example, about 90% of US establishments contain less than 20 employees and about 98% have less than 100 employees (Chapman, 1993). The use of stratification by size in this situation helps reduce the sampling variability of the characteristic being measured and consequently results in better estimates of the stratum parameters. 2) Since each stratum is treated as a separate population, the selection of samples in each of the strata can be carried out independently. Still related to the skewness in size characteristics in business surveys, the stratum containing the large establishments are usually over sampled or completely enumerated. This is because the characteristics measured from the few large businesses greatly influence the survey estimates (Colledge, 1995). 3) There are administrative sources or lists available which contain the information that can be used to measure size, to define strata and to determine the sampling rates in strata. 4) Lastly, the measurement, response and auxiliary information may differ from one stratum to another and thus lead to different choices of sampling design and estimation procedure across the various strata.

3.1 The Construction of Strata

Sigman and Monsour (1995) described the main objectives for the construction of strata namely: to obtain domain estimates, reduce design variances and reduce survey costs. The construction of strata for stratified sampling involves a choice of the stratification variables to be used, a demarcation of the boundaries between strata and determination of the number of strata.

Stratification variables are the characteristics used for subdividing the population into strata. The units within a stratum should be made as similar as possible to achieve an efficient stratification. Ideally, the best characteristic for stratification is the frequency distribution of the variable(s) under study. However in practice, this situation does not often exist. Thus, the stratification variables usually chosen are those that are strongly related with the variable(s) under study. In the business world, the stratification could be based on one or more characteristics such as industry, geography and measure of size. The measure of size can be expressed in terms of number of employees, sales, revenue or assets.

Stratification variables can be either categorical or continuous. Example of categorical variables are geographical region and industry classification. An example of a continuous variable is employment size but it becomes a categorical variable if employment size classes are used instead (e.g. 1-4, 5-20, 21-49, etc.).

For categorical variables, the boundaries of strata are easily determined. But when the stratification variable is continuous, say Employment Size, the lower and upper endpoint values for each stratum need to be specified. For a fixed number of strata, Dalenuis and Hodges (1959) proposed the cumulative square root of the frequency method (i.e. $cum\sqrt{f}$) to delineate the strata and obtain the best stratum boundaries or endpoints, where f represents the frequencies at each class boundary. Scheaffer, Mendelhall and Ott (1990) illustrated the $cum\sqrt{f}$ rule in the following example. The frequency distribution of yearly sales by \$50,000 increments for 56 firms and the associated $cum\sqrt{f}$ are shown in Table 4. The investigator wants to allocate the firms to L=3 strata.

Table 4. Calculation of stratum boundaries by the $cum\sqrt{f}$ rule.

Income (\$)	$f(y)$	$\sqrt{f(y)}$	$cum\sqrt{f}$
100,000-150,000	11	3.32	3.32
150,001-200,000	14	3.74	7.06
200,001-250,000	9	3.00	10.06
250,001-300,000	4	2.00	12.06
300,001-350,000	5	2.24	14.30
350,001-400,000	8	2.83	17.13
400,001-450,000	3	1.73	18.86
450,001-500,000	2	1.41	20.27
Total	56		

Source: Scheaffer, Mendelhall and Ott (1990). *Elementary Survey Sampling*, p. 128.

The total of $cum\sqrt{f}$ is 20.27 and with three strata, the division points should be set approximately equally spaced between 0 and 20.27, hence at 6.7 and 13.52. Based on Table 4, the nearest points on the actual scale are the following: 7.06 is closest to 6.76 and 14.30 is closest to 13.52. Thus the resulting three strata are:

Stratum	1	2	3
Boundaries	\$100,000 - \$200,000	\$200,001 - \$350,000	\$350,001 - \$500,000
Interval width	7.06	7.24	5.97
Number of Firms	25	18	13

The $cum\sqrt{f}$ method requires that the stratification variable should be uniformly distributed between adjacent endpoints. But this assumption is violated when the distribution of the population is highly skewed as in the case of business populations. To handle the problem of skewed population, the creation of a stratum that will include all the largest units has been suggested (Sigman and Monsour, 1995; Raj, 1968). This stratum is referred to as "certainty" stratum or "self-representing" stratum or "take-all" stratum. Then for the rest of the units, $cum\sqrt{f}$ rule will be applied to demarcate the boundaries of each stratum given a fixed number of strata. These strata are called "non-certainty" or "non-self-representing" or "take-some" strata.

The next concern is the determination of the number of strata to be used. To decide on the number of strata, two things should be considered: 1) the survey cost and 2) the magnitude of gain in efficiency as the number of strata increases (Murthy, 1977; Cochran, 1977). Cochran (1977) fitted a regression model for determining the reduction of variance for varying number of strata and determined that beyond 6 strata, the reduction in variance was negligible. A cost function which shows how the cost depends on number of strata was also fitted and the result obtained showed that an increase in strata beyond 6 would no longer be profitable. But this recommendation is not normally followed in business surveys where for example, one of the stratification variables is geographical region with number of regions exceeding 6. In Canada, there are 10 provinces and 3 territories which may form the strata. In addition, when using other criteria for creating strata in business surveys, there could in fact be hundreds of useful strata.

So far the information needed to stratify is assumed to be present in the frame. However, when such information is absent, a different strategy called a two phase sampling design could be considered. The design involves the selection of a first large sample which is then stratified and a sub-sample is selected from within each stratum. If this design is adopted, there are issues that need to be considered such as the cost of sampling at each phase, the amount of information available at each phase and the gain in precision obtained by the stratification during the first phase (Statistics Canada, 1998b).

3.2 Notations

The suffix h denotes the stratum and i the unit within the stratum in the following notations:

N	Total units in the population
L	Number of strata or subdivisions
N_h	Total number of units in stratum h, h=1,2,... L
n_h	Sample size in stratum h
Y_{hi}	Value of variable y on i^{th} unit within stratum h
$\overline{Y}_h = \frac{\sum_{i}^{N_h} Y_{hi}}{N_h}$	True mean in stratum h

$S_h^2 = \frac{\sum_{i=1}^{N_h}\left(Y_{hi} - \overline{Y}_h\right)^2}{N_h - 1}$ True variance in stratum h

$W_h = \frac{N_h}{N}$ Stratum weight or proportion of the population size attributable to h th stratum

y_{hi} Value of variable y on i th sample unit within stratum h

$\overline{y}_h = \frac{\sum_{i}^{n_h} y_{hi}}{n_h}$ Sample mean in stratum h

$s_h^2 = \frac{\sum_{i}^{n_h}\left(y_{hi} - \overline{y}_h\right)^2}{n_h - 1}$ Sample variance in stratum h

$f_h = \frac{n_h}{N_h}$ Sampling fraction in stratum h

3.3 Sample Allocation within Strata

In each stratum, a sampling design and sample size must be specified. This section discusses the sample size and the method of sample allocation while the next section will discuss the sampling design.

For a skewed population as would be anticipated for businesses, it has been found desirable to take the largest units with certainty and select a sample from the rest (Raj, 1968; Sigman and Monsour, 1995). As mentioned earlier, the certainty stratum is referred to as “take-all stratum” and the non-certainty stratum as “take–some” stratum. Different methods of sample allocation would therefore be applied to the “take-some” strata.

3.3.1 Take-all Stratum

The take-all stratum is constructed by putting all large businesses together and sampling these units with certainty. A subjective way to identify large businesses is to first sort the frame list by a classification variable. The outliers or potential outliers in the list are classified as large businesses. A more objective procedure for constructing the take-all stratum has been proposed by Glasser in 1962 and later on by Hidiroglou in 1986. Both of them used some “approximate cut-off rules” for stratifying the population into take-all and take-some strata (Hidiroglou, 1986). The assumptions are that in the take-all strata, all units are included and in the take-some strata, the sample units are selected using simple random sampling without replacement. Glasser’s method is designed to minimize sampling variance for fixed sample size while Hidiroglou’s method minimizes sample size for fixed sampling variance.

The number of units to be allocated to the take-all stratum depends on the desired precision, the population mean, the population variance and the total number of units in the population. Hidiroglou (1986) mentioned the advantages of using his method when the population is highly skewed. First, for a fixed coefficient of variation, the sample size obtained from this method is lower compared to the sample size obtained when there is no stratification. Secondly, since the largest observations are separated from the rest of the highly skewed distribution, the remaining observations would follow a less skewed distribution. Finally, this type of stratification protects against the overestimation of population characteristics.

3.3.2 Take-some Strata

In business surveys, the methods that could be used for sample allocations for the take-some strata are: 1) Proportional Allocation, 2) Optimal Allocation, 3) Neyman Allocation and 4) X -Optimal Allocation.

- Proportional Allocation. Proportional allocation is the allocation where the sample size for each stratum is made proportional to the stratum's size. This allocation is used when the strata differ in size. If stratum h has N_h units , the sample size in the stratum would be

$$n_h = \frac{N_h}{N} n = W_h n.$$

- Optimal Allocation. This is the scheme which allocates the sample size to minimize the variance of the estimates of the population mean (or total) for a specified cost or to minimize the cost for a specified variance value. Cochran (1977) described the scheme when the cost function is linear. The optimum values of sample size in a stratum, n_h, for the two scenarios are shown below.

i) Minimizing the variance with fixed cost and assuming linear cost function (i.e. $C = C_o + \sum_{h=1}^{L} c_h n_h$), the optimum value of n_h is

$$n_h = \frac{(C - C_o)\sum\left(N_h S_h / \sqrt{c_h}\right)}{\sum\left(N_h S_h \sqrt{c_h}\right)}$$

where C_o is the overhead cost

c_h is the unit cost in the h^{th} stratum

ii) Minimizing the cost with specified variance, the optimum value of n_h is

$$n_h = \frac{\left(\sum W_h S_h \sqrt{c_h}\right)\sum W_h S_h / \sqrt{c_h}}{V + N^{-1}\sum W_h S_h^2}$$

where V = a specified fixed variance .

This optimum scheme allocates a larger sample size to the larger or the more variable strata and a smaller sample size to the expensive and more difficult to sample strata. This allocation requires known stratum variance which is usually not available. Thus, estimates of variance from previous surveys or planned pilot surveys are used to calculate approximations of n_h (Cochran, 1977). In this scheme, cases may also occur where the optimum allocation of n_h is greater than the population size N_h. To handle such case, all units in the strata are included in the sample while the remaining sample size is allocated to the rest of the strata (Cochran, 1977).

This sample allocation is being used in the Current Employment Statistics (CES) Monthly Payroll Survey by the US Bureau of Labor Statistics (Bureau of Labor Statistics, 1997).

- Neyman Allocation. When the cost per unit is the same for all strata, the optimum allocation results in Neyman allocation and the value of the stratum sample size becomes :

$$n_h = n \frac{N_h S_h}{\sum^{L} N_h S_h} \ .$$

- X-Optimal Allocation. Another sample allocation which is found useful in business surveys is the X-optimal allocation. This sample allocation uses an auxiliary variable which is highly correlated with the primary variable. For example, let X be the auxiliary variable with known stratum deviation S_{xh} ; then the X-optimal allocation is obtained as :

$$n_h = n \frac{N_h S_{xh}}{\sum^{L} N_h S_{xh}} \ .$$

In business surveys, the auxiliary variable is often a measure of size. An example of this is a case in a retail trade survey where the auxiliary variable is the number of employees or the total annual sales from a recent census.

3.4 Sample Selection Methods

The procedure by which a sample of units is selected from the population is called the sampling design or sampling method. Each possible sample S_i is assigned a known probability of selection π_i. The selection probabilities can be equal or unequal for all units in the population. The sampling designs commonly used in business surveys will be discussed in this section. In stratified sampling, the "population" could be the sub-population associated with a particular take-some stratum. The sampling design need not be the same in all take-some strata.

3.4.1 Sampling with Equal Probability

- Simple Random Sampling Without Replacement (SRSWOR). This is a sampling design in which n_h distinct units are selected from the N_h units in the population in such a way that every possible combination of n_h units are equally likely to be the sample selected. The n_h selections are independent and without replacing a selected unit back into the population, each unit has the same probability of inclusion in the sample namely, $\frac{n_h}{N_h}$. Simple random samples are usually generated using a table of random numbers or by a computer "random number" generator. Simple random sampling is customarily carried out as a series of draws from the entire population or a given subset of population and each draw results in an element selected for the sample. The US Bureau of Labor Statistics (BLS) used SRSWOR in its Current Employment Statistics (CES) Monthly Payroll Survey (Bureau of Labor Statistics, 1997).

- Sequential SRSWOR. This method of sample selection is similar to SRSWOR except that the units being sampled are determined by random numbers associated with each unit. For example, let N_h be the size of the stratum. For each frame unit i, a random number ε_i is assigned. The random number is drawn independently from a Uniform [0,1] distribution. Then the ε_i of all the frame units are sorted in ascending order. The first n units in the ordered list are taken as the sample. This type of selection of samples is referred as Sequential SRSWOR by Ohlsson (1995).

Alternately, the position of the first element of the sample need not always be the first in the ordered sequence. There is flexibility as far as position of the first and last elements in the sample is concerned, as long as n units are selected.

If more than one survey of the same population is needed to be carried out, simultaneous selection of several non-overlapping random samples can be done if desired. On the other hand if overlapping samples are needed, the degree of overlap desired in the samples can be controlled. The practice of avoiding or minimizing any overlap for more than one survey helps reduce the response burden since the element will not be approached too often and consequently, would help increase the response rate.

There is another scheme where random numbers between 0 and 1 are assigned permanently to each element or unit in the frame. A scheme that is being used by Statistics Sweden in its annual and sub-annual surveys was introduced by Atmer (Ohlsson, 1995). Each unit in the list has an associated permanent random number (PRN). Any new unit or birth is simply added to the frame and assigned a unique random number. Deaths on the other hand are withdrawn from the list frame together with their associated random number.

The scheme is called sequential SRSWOR with PRNs and is used by Sweden SAMU System to coordinate sampling across surveys and over time. SAMU is an acronym for 'SAMordnade Urval inom foretagsstatisiken' or 'coordinated samples' in Swedish (Teikari, 2000; Ohlsson, 1995). Two terms related to coordination are being used in business surveys namely positive coordination and negative coordination. Positive coordination refers to the method of maximizing the overlap between samples of different surveys whereas negative coordination refers to the minimization of the overlap between samples of different surveys (Ohlsson, 1995). The two types of coordination are illustrated in Figures 3 and 4. For the positive coordination, consider a sample of 5 units drawn at time 1 and again at time 2.

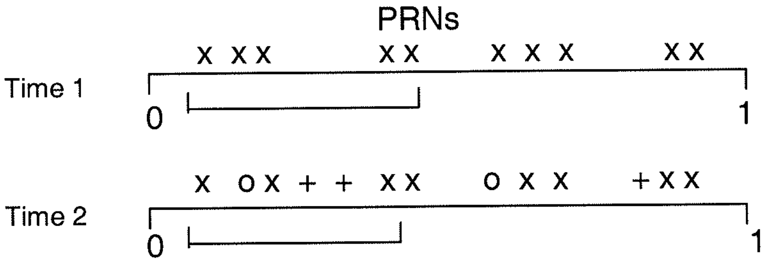

Figure 3. Positive coordination in SAMU system (denote the active units by x, new units by + and closed-down units by o). Source : Ohlsson, E. (1995). "Coordination of Samples Using Permanent Random Numbers". In *Business Survey Methods.* B.G. Cox et al (eds). p.155.

The PRNs associated with the sampled units are expected to be not equally spaced as shown in Figure 3. On each sampling occasion, the PRNs are used to select samples by sequential SRSWOR. To obtain as large an overlap as possible between the two sampling occasions, the same PRN should be used as a starting point. In Figure 3, for the first five units of both sampling occasions, only three units at Time 1 were retained at Time 2, due to the two newly formed units and one close-down unit at Time 2. The chance that the units stay or leave the sample for the next sampling occasion depends on the number of births and deaths that will occur in between the two sampling occasions.

For the case of negative coordination of samples across surveys, the overlap between samples can be reduced by choosing two constants a_1 and a_2 in the interval (0,1). If no overlap is desired between the two samples, allow at least 0.5 interval between the two constants. A sample of n units is selected either to the right starting from the points a_1 and a_2 as shown in Figure 4, or from the left as appropriate.

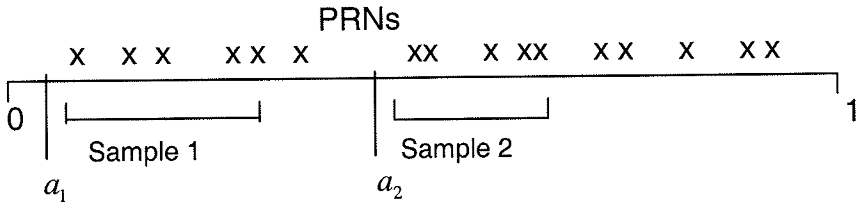

Figure 4. Negative coordination in SAMU system. Source : Ohlsson, E. (1995). "Coordination of Samples Using Permanent Random Numbers". In *Business Survey Methods.* B.G. Cox et al (eds). p. 156.

- Systematic Sampling. Suppose a stratum consists of N_h units that are ordered in some fashion and numbered 1 to N_h. A sample of n_h units is selected by first selecting a unit at random from the first k_h units, where k_h is a positive integer closest to $\frac{N_h}{n_h}$, and the rest of the samples are selected every k_h th unit thereafter until the end of the population list. There are only k_h possible samples, each unit having a probability of $\frac{1}{k_h}$ of being selected. This sampling design is known as systematic sampling. A systematic sample is drawn separately from each stratum with the starting points independently determined. When $N_h = n_h k_h$, the k_h systematic samples result in samples of equal size and such a scheme is called linear systematic sampling. When $N_h \neq n_h k_h$, a procedure called circular systematic sampling is used to overcome the problem of varying sample size (Singh and Chaudhary, 1986; Sarndal et al, 1992). The procedure involves selecting every k_hth unit from a random start, and when the last unit in the list for the stratum is reached, the selection process continues from the start of the list until n_h units are obtained.

Systematic sampling is easy, simple to execute and spread out the sample more evenly over the population. However there are also drawbacks in the use of systematic sampling. The efficiency of systematic sampling greatly depends on the particular ordering of the elements in the population. The elements in the population may be randomly ordered or ordered according to certain characteristic or may follow some definite periodic or cyclic pattern. It is less efficient when the elements in the population have periodic or cyclic pattern. To become more efficient, the arrangement of units should be made in such a way that the units within the systematic sample are heterogeneous to the maximum extent.

Another concern in systematic sampling with a single sample is the non-availability of an unbiased estimate of the variance of the estimator.

However, this drawback has been remedied by taking m_h systematic sub-samples with independent random starts, each sub-sample contains $\frac{n_h}{m_h}$ units to keep the same total sample size within the stratum the same. This method is referred as interpenetrating systematic sampling (Cochran, 1977; Singh and Chaudhary, 1986). But if, there is reason to believe that the units are randomly ordered then systematic sampling can be considered to be SRSWOR. Hence for practical applications, the estimator of the variance for SRSWOR is used to estimate the variance for the systematic sampling (Yamane, 1967).

3.4.2 Sampling with Varying Probability

In business surveys, the stratification by size is sometimes replaced by probability proportional to size (PPS) sampling. Each unit in the frame has an initial probability of selection (p_i) which is equal to some measure of size (X). A major issue for PPS sampling is how to draw the sample. Brewer and Hanif (1983) presented 50 schemes for selecting a sample with unequal probability. Only the methods applied in business surveys are discussed in this section.

- Poisson Sampling and other modifications . Poisson sampling is an unequal probability sampling design. Drawing of a Poisson sample starts with the list of the N units. This sampling method consists of a series of experiments or trials such that one experiment is carried out for each element in the frame list. According to a specified rule, a decision of whether to select or not to select that element is then made (Sarndal et al, 1992; Brewer and Hanif, 1983; Brewer, 2002). The sample consists of all the units which are selected in a series of trials.

Poisson sampling not only makes the drawing of a PPS sample from finite populations easy but also simplifies the rotation of the sample and controls the overlap of different samples for repeated surveys (Ohlsson, 1998).

For the sample selection in Poisson sampling, let the elements in the frame appear in a given order, k=1,2,..., N. Let π_k be a predetermined positive inclusion probability for the k^{th} unit, $0 < \pi_k < 1$. Let ε_1, ..., ε_N be independent random numbers drawn from the Uniform (0,1) distribution, The selection or non selection of the k^{th} element is decided by the following rule: if $\varepsilon_k < \pi_k$, the unit is selected, otherwise not, k=1,2,..., N.

A set of hypothetical data to illustrate selection of sample for Poisson Sampling is shown in Table 5. Let the population consists of N=10 elements. Each element has a measure of size variable X_k and an associated random number ε_k. The inclusion probability π_k is calculated to be directly proportional to the size measure X_k as shown below:

$$\pi_k = np_k = \frac{nX_k}{\sum X_k}.$$

Take the cases where the desired sample size n is 3 and 4.

Table 5. A set of hypothetical data to illustrate selection of sample for Poisson Sampling.

k	X_k	p_k	ε_K	n=3		n=4	
				$\pi_k = np_k$	$\varepsilon_k < \pi_k$	$\pi_k = np_k$	$\varepsilon_k < \pi_k$
1	76	0.067	0.696	0.201		0.267	
2	63	0.055	0.875	0.166		0.222	
3	15	0.013	0.575	0.040		0.053	
4	306	0.269	0.493	0.807	selected	1.077	selected
5	17	0.015	0.826	0.045		0.060	
6	20	0.018	0.857	0.053		0.070	
7	156	0.137	0.732	0.412		0.549	
8	280	0.246	0.183	0.739	selected	0.985	selected
9	150	0.132	0.568	0.396		0.528	
10	54	0.047	0.383	0.142		0.190	
Total	1137						
Number of selected sample units					2		2

Given the above set of random numbers, only 2 sample units passed the rule and were selected when Poisson sampling was used. As the set of random numbers associated with each element varies, the number of sample units to be selected also varies.

The U.S. Bureau of Census uses the Poisson sampling method for its Annual Survey of Manufactures (Ogus and Clark, 1971) and finds it has merit in terms of simple control as to which units are to be included in the sample and which units are not. Poisson sampling was also used in the Swedish Consumer Price Index until 1989 by Statistics Sweden (Ohlsson, 1998).

However the drawback of Poisson sampling is that the sample size is random. Brewer, Early and Hanif (1984) identified the issues involved when the sample size is random, namely: 1) the possible selection of an empty sample resulting in no estimation of population total, 2) larger variance for the estimator even though the samples are not empty, and 3) related to optimum sample allocation such that for a desired sample of size n, a) selecting a larger number of sample units means unnecessary extra effort, however, b) if a smaller number of sample units is selected, it will result in higher mean square error than intended.

A Modified Poisson sampling scheme introduced by Ogus and Clark (1971) solved the issue of selecting an empty sample. The procedure involves drawing an ordinary Poisson sample. If the sample is empty, then a second Poisson sample is drawn with a new set of random numbers and so on until a non-empty sample is obtained.

Collocated sampling is another procedure similar to Poisson sampling which helps reduce the variability of the sample size and selection of an empty sample (Brewer et al, 1984; Ohlsson, 1998). The units in the population are first arranged in random order, assigning unit k the order L_k ($L_k = 1, 2, ..., N$). Then random number ε_k, drawn from the Uniform (0,1) distribution, is assigned independently to each unit in the population. A random variable R_k is defined for each unit as

$$R_k = \frac{L_k - \varepsilon_k}{N} \ .$$

The criterion for selection or inclusion of sample is the same as in Poisson sampling but with R_k replacing ε_k.

Another alteration of Poisson sampling is Sequential Poisson sampling which yields a fixed sample size n (Ohlsson, 1998). The random numbers ε_k associated with the elements in the ordinary Poisson sampling are transformed into the form

$$\xi_k = \frac{\varepsilon_k}{p_k} \ .$$

The units are sorted by ξ_k and then the first n units are selected from the ordered list. Hence, a sample of size n drawn by Sequential Poisson sampling consists of the n units with the smallest transformed random numbers ξ_k. While the sample size can be fixed, the procedure is not a strict

PPS sampling but merely an approximate PPS scheme (Ohlsson, 1998). When Ohlsson (1995) used the method in a simulation study of Swedish CPI data, he found that the inclusion probabilities are quite close to that of those in a strict PPS sampling design and that the mean square error of the estimator used is slightly better than in ordinary Poisson sampling.

Ohlsson (1995) described the use of Poisson, Sequential Poisson and Collocated sampling in combination with PRNs. He found that by using any of these three PPS sampling methods, the coordinating features of samples across surveys and over time are valid if common PRNs are used.

A special case of Poisson sampling which uses permanent random numbers (PRNs) is referred to as Poisson πps sampling (Teikari, 2000). The use of PRN in controlling overlap of samples and the desire to reflect changes in the population such as births and deaths and changes in size measure make the Poisson πps sampling a suitable design for the coordination of surveys.

- Probability Proportional to Size (PPS) sampling without replacement. With fixed sample size, PPS sampling without replacement have inclusion probabilities π_i proportional to a positive size measure X_i defined as

$$\pi_i = \frac{nX_i}{\sum^N X_i} = \frac{nX_i}{X} \qquad \text{where } X = \sum^N X_i \text{, } 0 < \pi_i \leq 1 \text{.}$$

All X_i must be positive and $nX_i \leq X$. The Horvitz-Thompson estimator is the unbiased estimator of the population total Y.

There are various methods of selecting samples for PPS sampling. Examples of these methods are: 1) Brewer's Method, 2) Durbin's Method, 3) Murthy's Method, 4) Midzuno System, 5) Rao- Hartley-Cochran Method, 6) Systematic PPS, 7) Lahiri's Method and 8) Rao-Sampford Scheme. Each method is presented in detail by Brewer and Hanif (1983). Among the procedures, the systematic PPS sampling has the simplest selection procedure but has the disadvantage of no unbiased estimator for the variance. The Murthy sample selection scheme is more complicated than systematic PPS sampling but the variance estimator is simple for n=2 and becomes complicated as n increases. The Rao-Hartley-Cochran selection procedure is slightly more tedious than Murthy's scheme, slightly easier to apply than the Rao-Sampford Procedure but has the advantage of providing a simple estimator of the variance for any n size.

3.5 Recommendations

Stratified sampling and PPS sampling designs are two possible designs that can be used for business surveys. They are found suitable because of the skewed distribution of units in the business population, and the availability of auxiliary information which can be used as stratification variables or as a measure of size for PPS sampling. As a fundamental principle of sample design, the choice for the appropriate sampling design always considers the cost and precision factors (i.e. minimizing cost for a specified level of precision or maximizing precision within the fixed budget). Considering these factors and the other distinctive features of each of the two designs, the Stratified sampling design exhibits more advantages than the PPS sampling design. The stratification of business population in a way that strata are homogenous within themselves presents many advantages; such as 1) large businesses can be segregated from the small or medium businesses and assigned into separate stratum, thereby reducing the variability within strata, 2) a gain in the precision in the estimation of a characteristic of the whole population, 3) flexibility of using different sampling designs in different strata, and 4) convenience in operation. On the other hand, PPS sampling is known for its more complex sample selection procedures, and also complex or nonexistent variance formulas.

Within a stratum, there are again several sampling design options, namely, SRSWOR, Systematic sampling, Poisson Sampling and PPS sampling. The simplest among the methods, in terms of sample allocation, sample selection and estimation procedures is the SRSWOR. In contrast, the most difficult method is PPS sampling. The probability of drawing a sample unit at any given draw is the same for SRSWOR while in PPS sampling, the probability of drawing a sample unit differs from draw to draw. Thus, in PPS sampling, each unit in the population is assigned with an unequal probability of selection depending on the unit size, which makes the process of selecting samples tedious and the estimation formulae complicated. Depending on the available budget, the SRSWOR would be a practical choice.

The suggested stratification variables are the geographic regions and the type of industry. The geographical classification codes are composed of codes for region, census division and census sub-division. The industry classification codes will follow the five level hierarchy structure of NAICS.

With regards to employment size, it can be used either as a stratification variable or a measure of size in PPS sampling, since business size is found to be highly correlated with most of the variables of interest. If SRSWOR is the design chosen for all strata, then employment size must be used as a stratification variable. However , employment size ranges (e.g. 1-4, 5-20, etc.) will be used instead of actual size and the "indeterminate" category will be included. The units then will be made as homogeneous as possible within the size class strata.

The information on stratification variables should be always updated, particularly the industry classification and employment size.

The sample allocation for each of the strata is generally disproportionate over the size classes. It is recommended that the stratum with large business sizes be included with certainty and the rest of the strata should only take some samples using the appropriate sample allocation and sample selection method.

Any of the allocation methods (i.e. Neyman, Optimal, X-optimal) can be used. The most important variable(s) out of the many variables in the survey should be used in the determination and allocation of sample.

Finally, as available resources allow, the efficiency of alternative stratification techniques, sampling methods and allocations should be evaluated by conducting studies using available data from previous censuses, surveys and administrative data.

Chapter 4

Estimation Procedure

The previous chapter introduced the sampling designs commonly used in business surveys and the Stratified SRSWOR design was recommended. The other component of the broad concept of sample design is the procedure of estimation. The implementation of the chosen estimation methodology in the survey process happens after the data have been collected in the field and have been processed (e g. coded, checked and corrected). This chapter will deal with the estimation procedure associated with the recommended design, and also with other designs in order to show the complexities of their estimation formulae. The roles of auxiliary information in improving the survey estimates will be discussed in detail. Other components of the estimation process such as weighting, post-stratification and domain estimation will also be discussed.

4.1 Estimator – Definition and Properties

Estimation is defined as the methodology of computing the value or values of a statistic based on a set of sampled data in order to provide an approximate answer to the value of an unknown population parameter (Khazanie, 1996). A single computed value is called a point estimate of the parameter. An interval or range of numerical values which is believed to contain the unknown parameter together with a degree of confidence is called an interval estimate or confidence interval for the parameter. In business surveys, the point estimators include descriptive statistics such as means, totals, ratios or complicated statistics such as regression coefficients. Along with these estimated parameters are some measures of precision such as 1) the variance which is the average of all squared differences between the estimates and the expectation of the estimator, 2) the mean square error which is the average difference between the estimates and the parameter and 3) the coefficient of variation (%) which is the ratio of the standard deviation of the point estimate and the mean of the point estimate expressed in percentage. The variance measures the accuracy of the estimate or how close the estimate lies to the expectation of the estimator while the mean square error measures the precision of the estimate or how close the estimate is around the parameter. In most surveys, the coefficient of variation (%) is used as the standard measure of precision because of its dimensionless property.

The criteria or properties of a good estimator are unbiasedness, consistency, efficiency and sufficiency (Yamane, 1967; Khazanie, 1996).

- Unbiasedness. An estimator is unbiased when the expected value of estimator is equal to the value of the population parameter to be estimated.

- Efficiency. An unbiased estimator with a variance smaller than that of another unbiased estimator is said to be more efficient than the other estimator. Hence, an estimator which has the smallest variance among all the unbiased estimators of a parameter, if it exists, is said to be the most efficient estimator. Such an estimator is also called the minimum variance estimator. However, for biased estimators, the mean square error would be used instead of variance to determine efficiency.

- Consistency. An estimator is consistent when the estimate approaches the value of the population parameter to be estimated with high probability as the sample size becomes extremely large.

- Sufficiency. An estimator is sufficient if it has utilized all the information contained in the sample for the purpose of estimating a given parameter.

Statistical theory indicates that under simple random sampling, the sample mean is an unbiased and a consistent estimator of the population mean and the sample variance is an unbiased estimator of the population variance assuming the population units are normally distributed.

In business surveys, two approaches for estimation are being used namely 1) the design based inference method and 2) the model-based method.

The design-based inference method follows the classical estimation formula associated with the chosen sampling design. It is important to note that the estimation procedure reflects the sampling design. A poor combination of a sampling design and an estimator generates biased results. The person(s) who is responsible for the sampling design should be in close communication with the person(s) who works on the estimation. All aspects of the sampling design such as stratification, inclusion probabilities of sample, sample size etc. should be appropriately reflected in the estimation procedures. Examples of unbiased estimation procedures for various sampling design applicable to business surveys are presented in section 4.2.

The model-based approach of estimation on the other hand aims to improve the efficiency of the estimates by using auxiliary variables. For instance, an alternative for the Horvitz-Thompson estimator when auxiliary information is available is obtained using generalized regression estimation procedure. Generalized regression estimation is an old technique and has found applications in sampling recently. Sarndal, Swensson and Wretmann (1992) developed the methodology for the model assisted survey sampling using the generalized regression estimators (GREG). More of these will be discussed under the section Auxiliary Variables (section 4.4).

4.2 Estimation Formula for Stratified Sampling Design

The estimation formulae for the recommended design, Stratified SRSWOR along with the other designs, are presented in this section. Refer to Chapter 3, section 3.2 for some notations used therein.

Let the finite population U consists of N units. Then the population is divided into L subdivisions or sub-populations called strata denoted by $U_1,...,U_{h.},..U_L$, such that $\sum_{h=1}^{L} N_h = N$, N_h stratum size, and each unit in the population belongs to one and only one stratum, $U_h = \{k : k$ belongs to stratum h, $h = 1,\dots L\}$. A probability sample S_h according to a design $p_h(\cdot)$ from each stratum is selected such that the selection from each of the strata is independent one to another. Let Y_h be the stratum total and $\overline{Y}_h$ be the stratum mean. The population total can be decomposed as $Y = \sum_{h=1}^{L} Y_h = \sum_{h=1}^{L} N_h \overline{Y}_h$ and the population mean can be decomposed as $\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h$ where $W_h = \frac{N_h}{N}$ denotes the relative size of the stratum U_h.

- The recommended design - Stratified SRSWOR

  The unbiased estimator of the population mean $\overline{Y}$ is

$$\hat{\overline{Y}}_{st} = \overline{y}_{st} = \sum_{h=1}^{L} \frac{N_h \overline{y}_h}{N} = \sum_{h=1}^{L} W_h \overline{y}_h .$$

Note : $N_h \bar{y}_h$ can also be expressed as $\dfrac{N_h}{n_h}\sum_{i=1}^{n_h} y_{ih}$.

Consider the first stratum $(h=1)$, as the take-all stratum with $N_1 = n_1$ and $\dfrac{N_1}{n_1} = 1$. Also consider the strata h=2,3,... L as the take-some strata. The form of the estimator becomes:

$$\hat{\bar{Y}}_{st} = \bar{y}_{st} = \frac{\sum_{i=1}^{N_1} y_{i1} + \sum_{h=2}^{L} N_h \bar{y}_h}{N} .$$

The first term of the estimator corresponds to the take-all stratum and the last term corresponds to the take-some strata.

Variance of the estimate $\bar{y}_{st}$ is

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \ V(\bar{y}_h) = \sum_{h=1}^{L} W_h^2 \ \frac{S_h^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h).$$

An unbiased estimator for variance of $\bar{y}_{st}$ is

$$v(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \ v\,(\bar{y}_h) = \sum_{h=1}^{L} W_h^2 \ \frac{s_h^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} - \frac{\sum_{h=1}^{L} W_h \, s_h^2}{N} .$$

The variance using Neyman or Optimum Allocation is

$$V_o(\bar{y}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{n} - \frac{\sum_{h=1}^{L} W_h \, S_h^2}{N}.$$

The variance using Proportional Allocation is

$$V_p(\bar{y}_{st}) = \frac{\sum_{h=1}^{L} W_h S_h^2}{n} - \frac{\sum_{h=1}^{L} W_h \, S_h^2}{N}.$$

Considering again the case of the take-all stratum $(h = 1)$ and the take-some strata, h=2,3,... L, the variance of the estimator for the take-all stratum is zero, resulting in the form of the variance as:

$$V(\bar{y}_{st}) = \sum_{h=2}^{L} W_h^2 \; V(\bar{y}_h) = \sum_{h=2}^{L} W_h^2 \; \frac{S_h^2}{n_h}\left(\frac{N_h - n_h}{N_h}\right).$$

The above estimation formulae were taken from Cochran (1977), Murthy (1977), and Levellee and Hidiroglou (1988).

- Estimation Formula for Proportion in Stratified SRSWOR

Assume that within stratum h, there are A_h out of the N_h units with a given characteristic such that $P_h = \frac{A_h}{N_h}$. Similarly there are a_h out of n_h units in the sample having the characteristic, so $p_h = \frac{a_h}{n_h}$.

The unbiased estimate of P (the proportion of units in the population with the characteristic), is

$$\hat{P}_{st} = p_{st} = \sum_{h=1}^{L} W_h p_h = \frac{\sum_{h=1}^{L} N_h p_h}{N}.$$

The variance of the estimated proportion p_{st} is

$$V(p_{st}) = \sum_{h=1}^{L} W_h^2 \frac{P_h(1-P_h)}{n_h}\left(\frac{N_h - n_h}{N_h - 1}\right).$$

The unbiased estimator for the variance of p_{st} is

$$v(p_{st}) = \sum_{h=1}^{L} W_h^2 \frac{p_h(1-p_h)}{(n_h - 1)}\left(\frac{N_h - n_h}{N_h}\right).$$

The variance using Optimum Allocation is

$$V_o(p_{st}) = \frac{\left(\sum_{h=1}^{L} W_h \sqrt{P_h(1-P_h)}\right)^2}{n} - \frac{\sum_{h=1}^{L} W_h P_h(1-P_h)}{N}.$$

The variance using Proportional Allocation is

$$V_p(\hat{P}_{st}) = \frac{N-n}{N^2 n}\sum_{h=1}^{L} W_h^2 \frac{P_h(1-P_h)}{(N_h-1)}.$$

The succeeding estimation formulae correspond to sampling designs within stratum which have complex estimation procedure, hence, were not recommended for the omnibus survey. These include the PPSWOR, Poisson and Systematic sampling.

- Case of PPSWOR Sampling in a stratum

The estimator for the stratum total is

$$\hat{Y}_{HT} = \sum_{i}^{n_h} \frac{y_i}{\pi_i}$$

where π_i denotes the inclusion probability of the i^{th} sampled unit.

The variance of the unbiased estimator y_{HT} is

$$V_h\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{N_h} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i,j=1}^{N_h}\sum_{i\neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j$$

where π_{ij} denotes the probability that both i^{th} and j^{th} sampled units are included in the sample.

Another variance form for the estimator derived by Sen-Yates-Grundy for fixed sample size is given by

$$V_h\left(\hat{Y}_{HT}\right) = \sum_{i,j=1}^{N_h}\sum_{j>i} \left(\pi_i \ \pi_j - \pi_{ij}\right) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2.$$

The unbiased estimator for the variance is

$$v_h\left(\hat{Y}_{HT}\right) = \sum_{i=1}^{n_h} \frac{1-\pi_i}{\pi_j} y_i^2 + \sum_{i,j=1}^{n_h} \sum_{i\neq j} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} y_i y_j .$$

The unbiased estimator for the variance suggested by Sen-Yates-Grundy for fixed sample size is

$$v_h\left(\hat{Y}_{HT}\right) = \sum_{i,i=1}^{n_h} \sum_{j>i} \left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 .$$

The above estimation formula were taken from Brewer and Hanif (1983).

- Case of Poisson Sampling in a stratum

Consider each stratum h as the population. Let each unit in the population have a known probability of inclusion denoted by π_i for the i^{th} unit, $i = 1,2,\ldots N_h$ and the sample size denoted by n_h.

The unbiased Horvitz-Thompson estimator of the population total is

$$y'_{PS} = \sum_{i=1}^{n_h} \frac{y_i}{\pi_i}$$

The joint probability inclusion π_{ij} takes the form $\pi_{ij} = \pi_i\pi_j$, hence the variance of the unbiased estimator y'_{PS} is

$$V(y'_{PS}) = \sum_{i=1}^{N_h} (1-\pi_i) \frac{Y_i^2}{\pi_i} .$$

Unbiased estimator for variance of the y'_{PS} is

$$v(y'_{PS}) = \sum_{i=1}^{n_h} (1-\pi_i) \frac{y_i^2}{\pi_i}.$$

Since the sample size is not fixed but random, another estimator y''_{PS}, in a ratio form, is found more efficient than y'_{PS} (Brewer et al, 1984).

$$y''_{PS} = \begin{cases} \frac{y'_{PS}}{m_h} \cdot n_h & \text{if } m_h > 0 \\ 0 & m_h = 0 \end{cases}$$

where m_h denotes the size of the non-empty sample.

The mean square error of y''_{PS} is approximated by

$$V(y''_{PS}) \approx \sum_{i=1}^{N_h} \pi_i (1-\pi_i) \left(\frac{Y_i}{\pi_i} - \frac{Y}{n_h} \right)^2 + P_0 Y^2$$

where $P_0 = \Pr(m_h = 0)$ and $n_h = E(m_h) = \sum_{i=1}^{N_h} \pi_i$.

The conventional estimator of the approximation y''_{PS} is

$$v(y''_{PS}) = \sum_{i=1}^{n_h} (1-\pi_i) \left(\frac{y_i}{\pi_i} - \frac{y''_{PS}}{n_h} \right)^2 + P_0 y''^{\,2}_{PS}.$$

The above estimation formulae were taken from Brewer, Early and Hanif (1984).

- Case of Modified Poisson Sampling in a stratum

The variance for the Horvitz-Thompson estimator using modified Poisson sampling is

$$V(y'_{MPS}) = \sum_{i=1}^{N_h}(1-\pi_i)\frac{Y^2}{\pi_i} - P_0^*\left(Y^2 - \sum_{i=1}^{N_h} Y_i^2\right)$$

where P_0^* is the probability of selecting an empty sample at each draw.

The unbiased estimator of the variance is

$$v(y'_{MPS}) = \sum_{i=1}^{n_h}(1-\pi_i)\frac{y_i^2}{\pi_i^2} - \frac{P_0^*}{1-P_0^*}\left(y'^2_{MPS} - \sum_{i=1}^{n_h}\frac{y_i^2}{\pi_i^2}\right) .$$

Using the ratio estimator , the mean square error approximation is

$$V(y''_{MPS}) \approx \sum_{i=1}^{N_h}\pi_i\left\{1-\left(1-P_0^*\right)\pi_i\right\} - \left(\frac{Y_i}{\pi_i} - \frac{Y}{n_h}\right)^2 .$$

The conventional estimator of the approximation y''_{MPS} is

$$v(y''_{MPS}) = \sum_{i=1}^{n_h}\left\{1-\left(1-P_0^*\right)\pi_i\right\} - \left(\frac{y_i}{\pi_i} - \frac{y''_{MPS}}{n_h}\right)^2 .$$

The source of the estimation formulae for Modified Poisson Sampling was Brewer, Early and Hanif (1984).

- Case of Stratified Systematic Sampling

Let $\bar{y}_{syh}$ denotes the mean of the systematic sample in stratum h.

The estimator of population mean $\bar{Y}$ is

$$\bar{y}_{stsy} == \sum_{h=1}^{L} W_h \bar{y}_{syh} \quad .$$

The variance for the estimator of $\bar{Y}$ is

$$V(\bar{y}_{stsy}) = \sum_{h=1}^{L} W_h^2 \; V(\bar{y}_{syh})$$

where $\quad N_h = n_h k_h$

$$V(\bar{y}_{syh}) = \frac{N_h - 1}{N_h} S_h^2 - \frac{k_h(n_h - 1)}{N_h} S_{wsyh}^2$$

$$S_{wsyh}^2 = \frac{1}{k_h(n_h - 1)} \sum_{i=1}^{k_h} \sum_{j=1}^{n_h} (y_{ijh} - \bar{y}_{i.h})^2 \; .$$

To obtain an unbiased estimate of the error variance, two systematic samples each with a random start and an interval $2k$ should be drawn within each stratum (Cochran, 1977).

4.3 Weighting

The idea of weights is not a new concept for stratified sampling design. For example, in the estimation of the population mean in stratified sampling, the estimates from the individual strata have an associated stratum weight W_h of the form $\frac{N_h}{N}$. The sum of the products of the sample mean of the respective stratum multiplied by stratum weights results in the estimator of the population mean $\overline{Y}$.

In large-scale surveys, when estimates of various parameters are to be computed at the tabulation stage, one of the techniques commonly used to simplify the calculation of estimates in order to save time and reduce work load is to assign certain weights to the sample observations or units. The technique is referred to as weighting. The weights to be used depend on the sampling design and the estimation procedure, and they are usually chosen to make the estimator unbiased (Murthy, 1977). The weights are first calculated and then they are multiplied by the sample values of the various characteristics under study.

Weighting is essential for the following reasons: 1) to expand the sample information to the level of the target population, 2) to cope with missing sampled units including nonresponse, 3) to increase precision of estimates through the use of the auxiliary information and 4) to bring about consistency with data from other sources (Koeijers and Hilbink, 1998b).

In the sample survey, the value of a characteristic for the whole population is estimated on the basis of data gathered from selected samples. Weights are used to inflate or raise the sample observations to get an estimate of the population total (Murthy, 1977) and so weights are also known as multipliers, inflation factor or raising factor. Computationally, weight is equal to the inverse of the inclusion probabilities used by the sampling design. Sometimes this weight is called sampling weight (Statistics Canada, 1998b) or inclusion weight (Koeijers and Hilbink, 1998b) or just basic weight. The application of the sampling weight for the Stratified SRSWOR and PPS sampling designs is shown below.

- For Stratified SRSWOR design, the elements or units in stratum h have weights equal to $\frac{1}{n_h/N_h}$ or $\frac{N_h}{n_h}$ where N_h is the population size and n_h is the sample size in stratum h. Thus, the estimator of the population total Y is given by

$$\hat{Y}_{st} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

where y_{hi} is the value of the i^{th} selected unit within stratum h.

- In the case of the PPS sampling design, the weight generally differs from unit to unit and is equal to $\frac{1}{\pi_i}$ where π_i is the first order inclusion probabilities of the i^{th} unit. The estimator Horvitz-Thompson for the population total is given by $\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ when applied to a probability sample of size n.

Another type of weighting adjustment is related to missing sampled units called "nonresponse". Nonresponse is a non-sampling error and is defined as the failure to obtain a part or complete information from one or more units in the sample. Nonresponse is considered a serious problem because it reduces the efficiency of the estimators by increasing the variance due to reduced sample size and by introducing bias to the estimates (Sarndal et al, 1992; Johnson, 1995; Lessler and Kalsbeek, 1992). Nonresponse will be discussed in detail in Chapter 6.

At the tabulation stage, nonresponse is dealt with by adjusting the sampling weights with some measure of response probability. The idea is to inflate the original sampling weight used in the estimator of the characteristic based on the information from responding units in order to compensate for those values that are lost because of nonresponse. This assumes that the respondents and the non-respondents have similar characteristics and that non-respondents are missing at random (Chaudhuri and Stenger, 1992). The adjusted sample weight is defined as the sampling weight multiplied by the inverse of the response probability (Koeijers and Hilbink, 1998b). But the response probability is unknown and must be estimated in a logical manner. The manner in which the response probability is estimated differentiates the various adjustment methods. Lessler and Kalsbeek (1992) presented in detail three adjustment methods namely the 1) Politz and Simmons adjustment method, 2) the weighting class adjustment and 3) the post--stratification adjustment method.

- The Politz and Simmons' adjustment procedure involves the use of number of days (t_i) during the previous five days that the i^{th} respondent would have been available to be interviewed. The estimated response probability equal to $\frac{(t_i + 1)}{6}$ has the premise that the likelihood of response is tied to the recent availability of the person. This adjustment method is applicable to social surveys rather than business surveys.

- In the weighting class adjustment method, the original sample (which includes both respondents and non-respondents) is divided into H mutually exclusive and non-overlapping groups. Within each group the elements are assumed to have the same response probability while different groups have different response probabilities. These groups are termed in various ways such as adjustment cells (Lessler and Kalsbeek, 1992), adjustment classes or weighting classes (Choudhuri and Stenger, 1992) or response homogeneity group [RHG] (Sarndal et al, 1992). The response rate in each group is obtained. The nonresponse adjustment weight of an element is equal to the inverse of the response rate in the group to which the element belongs. The weighting class adjustment weight is the product of the sampling weight and the nonresponse adjustment weight. The crucial part here is how to form the groups such that a constant response probability within a group can be achieved. This demands survey experience on the part of the survey statistician.

- The post--stratification adjustment method is the same as the weighting class adjustment if the groups coincide with the post- strata. The post-stratification variables should be strongly correlated to the Y variable. The members of the post-stratification adjustment cells are expected to have similar response probabilities with respect to the Y variable. Lessler and Kalsbeek (1992) shows the exact formula for the post-stratification adjustment method.

There are still other weighting adjustments which are being used in many surveys. The Labor Force Survey of Statistics Canada for instance include weight adjustments due to coverage error. For longitudinal studies, two types of weights are provided namely the longitudinal weights which need to be adjusted to take into account the attrition of sample units over time and the cross-sectional weights which are related to the population at each occasion. These two weights differ from one another because of the changing population over time (Statistics Canada, 1998b).

For repeated surveys, the weight adjustment method used involves a ratio imputation approach (Hidiroglou et al, 1995) or adjustment ratio (Chapman et al, 1986). The adjustment ratio is derived from those units that responded in the current period and for which information from the previous period is known. This method is being used in the Monthly Retail Trade Survey by the U.S. Bureau of Census (Chapman et al, 1986).

4.4 Auxiliary Variables

An auxiliary variable is any variable which provides information on individual units in the population. Auxiliary information is found useful both at the design stage and at the estimation stage of a survey. At the design stage, the auxiliary information is used to improve sampling designs and consequently increase the precision of the estimates. These are accomplished when the auxiliary information is used to construct strata in Stratified sampling designs and when it is used in the generation of the inclusion probabilities for PPS sampling designs. At the estimation stage, the auxiliary variables are incorporated in the estimator formula to improve the efficiency of the estimates by reducing their variance. In order for the auxiliary data to be effective it should be highly correlated with the characteristics about which estimates are desired (Singh and Chaudhary, 1986; Sarndal et al, 1992). Auxiliary variable can be incorporated in the 1) ratio method of estimation, 2) regression method of estimation, 3) post-stratification and 4) calibration methods (Hidiroglou et al, 1995).

4.4.1 Ratio Estimators

When the relationship between the study variable y and a single auxiliary variable x is linear and passes through the origin, the ratio estimator provides a precise estimate of the population total or mean. The ratio of the population totals or means of the characteristics y and x is defined as

$$R = \frac{Y}{X} = \frac{\overline{Y}}{\overline{X}}$$

where Y and X are the total of characteristics y and x of the population respectively. An estimator of the ratio R is given by $\hat{R} = \frac{\hat{Y}}{\hat{X}}$, where $\hat{Y}$ and $\hat{X}$ are the unbiased estimators of the population total Y and X respectively. The ratio estimator of total Y is given by $\hat{Y}_R = \hat{R}X = \frac{\hat{Y}}{\hat{X}}X$.

For stratified sampling design, there are two ways of obtaining a ratio estimate of the population total Y, namely 1) separate ratio estimate and 2) combined ratio estimate. Let y_h and x_h denote the sample totals in the stratum h; $\overline{y}_h$ and $\overline{x}_h$ denote the sample means in the stratum h and X_h denotes the h^{th} stratum total of characteristic x.

- Separate ratio estimator. Within a stratum, an estimator of the ratio R is given by $\hat{R}_h = \frac{y_h}{x_h} = \frac{\overline{y}_h}{\overline{x}_h}$ and an estimator of stratum total Y_h is given by $\hat{Y}_{Rh} = \hat{R}_h X_h$. By summing the L estimators, the estimator of the population total Y is defined as

$$\hat{Y}_{RS} = \sum_{h=1}^{L} \hat{R}_h X_h = \sum_{h=1}^{L} \frac{y_h}{x_h} X_h = \sum_{h=1}^{L} \frac{\overline{y}_h}{\overline{x}_h} X_h \ .$$

- Combined ratio estimator. From the strata, the estimate of Y and X are defined as $\hat{Y}_{st} = \sum_{h=1}^{L} N_h \overline{y}_h$ and $\hat{X}_{st} = \sum_{h=1}^{L} N_h \overline{x}_h$. An estimator of the ratio R is given by $\hat{R}_C = \frac{\hat{Y}_{st}}{\hat{X}_{st}}$. The estimator of the population total Y is defined as

  $$\hat{Y}_{RC} = \hat{R}_C X$$

  where X is the overall population total for the characteristic x.

The choice on whether to use a separate ratio estimate or combined ratio estimate depends on the stratum size and the differences among the stratum ratios. If the stratum ratios do not vary much and the stratum size is relatively small, combined ratio estimation is preferred. If the auxiliary totals for each stratum, X_h, are known and the sample size in each stratum is large enough to achieve a valid variance approximation, separate ratio estimation is expected to be more precise than the combined one. When the line of regression of y passes through the origin within each stratum, the separate ratio estimator is more efficient.

When ratio estimation is applied to groups identified after sampling or post-strata, the type of ratio estimator is called the post-stratified ratio estimator. The X_p value, which is the p^{th} post-stratum total of characteristic x, must be known (e.g. either derived from the frame or from a reliable external source) for the post-stratified ratio estimator to be used.

A major concern in the ratio method of estimation is that ratio estimator is biased and its standard error can only be approximately estimated. Much research has been done to reduce or correct the bias. Quenouillie's method, or Jackknife method (Sarndal et al, 1992), reduces the bias in the estimator from order $\frac{1}{n}$ to $\frac{1}{n^2}$ while the method proposed by Hartley and Ross can be used to produce an unbiased ratio-type estimator (Cochran, 1977; Murhty, 1977).

4.4.2 Regression Estimators

When the relationship between the study variable y and the auxiliary variable x is linear but does not go through the origin, it is more precise to estimate the population total or mean by fitting a linear regression equation (Singh and Chaudhary, 1986). The estimator is called a regression estimator. The regression model is used as a means to find an appropriate β (i.e. the regression coefficient of y on x) to be used in the regression estimator formula. Thus the procedure is said to be model assisted and not model dependent.

For Stratified SRSWOR design, as with the ratio estimators, there are two possible regression estimators namely 1) separate regression estimator where regression coefficients are estimated separately for each stratum and 2) combined regression estimator where a pooled regression coefficient for all strata is estimated.

Let y_{hi} and x_{hi} denote the value on the i^{th} sample unit in the stratum h; $\bar{y}_h$ and $\bar{x}_h$ denote the sample means in the stratum h; $\bar{X}_h$ and X denote the hth stratum mean and overall total of characteristic x.

- A separate regression estimator is defined as:

$$\hat{Y}_{SR} = \sum_{h=1}^{L} N_h \left[\bar{y}_h + \hat{\beta}_h \left(\bar{X}_h - \bar{x}_h \right) \right]$$

$$\text{where } \hat{\beta}_h = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}$$

- A combined regression estimator is defined as :

$$\hat{Y}_{CR} = \sum_{h=1}^{L} N_h \bar{y}_h + \hat{\beta} \left(X - \sum_{h=1}^{L} N_h \bar{x}_h \right)$$

$$\text{where } \hat{\beta} = \frac{\sum_{h=1}^{L} w_h \sum_{i=1}^{n_h} (y_{hi} - y_h)(x_{hi} - \bar{x}_h)}{\sum_{h=1}^{L} w_h \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2} \text{ and } w_h = \frac{N_h \left(1 - {}^{n_h}\!/_{N_h} \right)}{n_h (n_h - 1)}.$$

The above formulae were taken from Murthy (1977).

The decision whether to use the separate or combined regression estimator depends on results of the estimated regression coefficients obtained for each of the strata. If the coefficients appear to be the same in all strata, then the combined regression estimator is to be used. If the coefficients differ from stratum to stratum then the separate regression estimator is preferred. Between the two, the separate regression estimator is likely to be more efficient than the combined regression estimator. But when sample size in each stratum is small, the bias in the separate regression estimator is expected to be larger (Singh and Chaudhary, 1986). In such a situation, collapsing the strata could be one alternative in increasing the sample size. The regression estimator is not unbiased, but for large samples it is approximately unbiased (Sarndal et al, 1992). Consequently, the variance and the estimator of variance are also approximations.

The form of the regression estimator has been generalized to maximize the use of auxiliary data (Sarndal et al, 1992; Hidiroglou et al, 1995). In the generalized regression (GREG) estimation procedure, the population can be divided into mutually exclusive sub-populations or groups, with every sample unit belonging to a sub-population. The sub-populations could be the original design strata or a new partition for which one or more auxiliary variables are available. The auxiliary totals for the entire population and sub-populations need to be known and every sample unit in a sub-population must have associated auxiliary data. The number of auxiliary variables may differ across sub-populations and the auxiliary variables being measured need not be the same across sub-populations (Hidiroglou et al,1995).

The GREG estimation procedure consists of fitting a regression model of y on the $x_1, x_2, ... x_k$ auxiliary variables in each of the groups and then estimating the total for each group. To obtain the GREG estimator for the entire population total, the totals over the groups are summed up. The mathematical formulae are shown in details in Hidiroglou, Sarndal and Binder (1995), and Sarndal, Swensson and Wretman (1992).

If there is only one group, only one x auxiliary variable, and the samples are drawn by Stratified SRSWOR design, then the generalized regression estimator is equivalent to the Combined regression estimator. If the groups are the same as the original design strata in stratified sampling design and there is only one x auxiliary variable, then the generalized regression estimator is equivalent to the Separate regression estimator.

Post-stratified regression estimator is another type of estimator applicable when group membership is created after sampling. The GREG estimator of the population total is obtained with the post- strata as the groups.

Statistics Canada has used the regression estimators in the Canadian Survey of Employment, Payrolls and Hours (Hidiroglou et al, 1995). For many years, the Horvitz-Thompson estimator has been used to produce estimates of totals for various characteristics of interest. Recently, a study using auxiliary data in the regression estimator was undertaken. The monthly remittance of payroll deductions from administrative files of Revenue Canada was used as the auxiliary variable. The important variables such as payroll and employment were used as y_1 and y_2 variables. Correlation was found to be high between y_1 and x variables, and between y_2 and x variables. For the larger strata the combined regression estimator was retained. For the smaller strata, both the Horvitz-Thompson estimator and the regression estimator were retained. The regression estimator showed large efficiency gains as compared to the Horvitz-Thompson estimator for the smaller strata.

The efficiency of the regression estimator depends 1) on the chosen auxiliary variables – whether they correlate well with the variable of study, and 2) on the size of sample. Larger sample sizes are more likely to achieve unbiased estimates. Small sample size shall be discussed more under the section Domain (section 4.6).

4.4.3 Calibration Estimators

Calibration method is another way of incorporating auxiliary information into the estimation process. This is done by generating a weighting system with the aid of distance functions and a set of constraints called calibration equations (Deville and Sarndal, 1992). The main goal is to generate new weights, w_k that satisfy the constraints $\sum w_k x_k = X$ where X is the known population auxiliary total. This means that the sample sum of the weighted auxiliary variable values must equal the known population total for that auxiliary variable. These new weights ensure that the sample estimates are consistent with the auxiliary information. The calibrated weights are a modification of the basic sampling weights that were discussed earlier under the Section on Weighting. For every distance function, there is a corresponding set of calibrated weights and a calibration estimator. The use of calibration estimation improves precision of estimates when the resulting calibrated weights are not widely spread out (Statistics Canada, 1998b). When there is a large heterogeneity of weights, the weights need to be bounded. Deville and Sarndal (1992) showed the step-by-step procedure of the calibration approach, the different distance functions as well as the bounding methods. They also discussed the application of the technique in the calibration on the known counts (either cell counts or marginal counts) of a two way frequency table .

The calibration approach was used in the Canadian Monthly Retail Survey. The sampling design is Stratified sampling with the province by industry by size as the strata. The objective is to estimate a characteristic of interest y for a specified province. The auxiliary variables are the known industry totals, $X_{i.}$ and the known size class totals, $X_{.j}$. Following the procedure, the estimate for each of the cell totals X_{ij} was obtained as well as the corresponding estimator for the characteristic of interest y for cell (i,j). The estimator was found better than the Horvitz-Thompson estimator. The CV(%) was reduced from 0.08% to 0.05%.

4.5 Post-stratification

Post-stratification is another technique commonly used in estimation. Post-stratification is a classification of the selected sample into a given number of strata only after selection of the sample. In the same way that stratification is used at the design stage to bring a gain in the precision in the estimation of a characteristic of a population, the same is true in the post-stratification at the estimation stage. The post-strata may be the same as the design strata (in the case of stratified sampling design) or they may be new strata based on variables found suitable for stratification after the sample information is obtained.

The technique of post-stratification consists of 1) dividing the selected sample into a certain number of strata, termed as post-strata, after the sample data are known and 2) estimating $\overline{Y}$ by the weighted means of the estimators of the post-strata means, with the proportion of the units in the post-strata as the weights (Murthy, 1977; Singh and Chaudhary, 1986). The form of the estimator of the $\overline{Y}$ is

$$\hat{\overline{Y}}_{POST} = \sum_{p=1}^{L'} W_p \overline{y}_p$$

where $\overline{y}_p$ is the mean of the n_p sample units falling in the p^{th} post-stratum (p=1,2,... L') and $W_p = \frac{N_p}{N}$ is the proportion of units in the post-strata which is known from other sources.

The number of sample units falling into each post-stratum is a random variable. Thus inferences are then made conditionally given the achieved sample size. In terms of sample size, the post-stratification produces more reliable results when the sample size within a post-stratum is relatively large. Post-stratifying the population too finely is not recommended because the sample sizes within the post-strata may become quite small.

If $\frac{N_p}{N}$ is known and sample size is large (i.e. n_p > 20 in every stratum), post-stratification technique is as precise as stratified random sampling with proportional allocations (Cochran, 1977; Scheaffer et al, 1990).

Post-stratification is a useful technique to improve efficiency of estimators in large-scale surveys. It is an important estimation tool not only in reducing the variance of the estimator but also in reducing the conditional bias of the estimator of a total (Valiant, 1993). Much gain in efficiency is expected when the post-stratum means are quite different from each other (Hidiroglou et al, 1995).

The post-stratification technique is found useful when 1) applied with ratio and regression estimation, 2) obtaining estimates for domains of study by treating the domains as post-strata and 3) correcting or adjusting for sample nonresponse and for coverage problems in sampling frames.

The post-stratification estimation is being used in business surveys by statistical offices in many countries including Statistics Canada, Statistics Netherlands and Australian Bureau of Statistics (ABS). In fact, post-stratification of the sample is a common practice in ABS business surveys.

In Statistics Canada's Survey for Employment, Payroll and Hours (SEPH), post-stratification estimation was applied due to the introduction of a new classification system in October 1990. The post-stratification was based on a comparison of the business register September 1990 size code (based on the 1970 SICs) and the new October 1990 size code (based on the 1980 SICs). After studying and comparing the effect of the changes in size codes on the estimates, the estimates based on the post-stratification were used (Hidiroglou et al, 1995).

4.6 Domain Estimation

In most surveys, estimates are wanted not only for the entire population but also for specific subsets of the population known as domains. Domain estimation is therefore the estimation for the specific subset or sub-populations of interest. The division into domains represents a new type of partitioning of the population into subsets (Sarndal et al, 1992). The elements included in the sub-population are only known after the data have been collected and not before sampling. The number of elements in the sample belonging to a domain is a random variable and may happen to be very small in some cases. A 'small domain' occurs when there are very few observations or none at all within a domain. Domain estimation is done by setting to zero the characteristics of the sampled elements that are found not to belong to the specified domain of interest, otherwise the characteristics take the measured values. The estimators for population mean, totals and proportions can be used for the domain estimations provided the units in the sample not belonging to the sub-population are assumed to have a value of zero for the character of interest.

Aside from the problem of random sample size, there is also a problem of unknown domain size (N_d). In the estimation for the domain mean $\overline{Y}_d$, the issue of unknown N_d has been handled by treating the estimator of $\overline{Y}_d$ as the ratio of two unknown totals. For the variance estimator, an approximate variance estimator conditioned on fixed value of n_d has been suggested. Sarndal, Swensson and Wretman (1992) compared under the simple random design, the scenario where the domains are specified in advance and the sample size controlled, with the scenario that the domains are known after sampling and the sample size uncontrolled. They found that the two variances are roughly the same if the sample size under the controlled conditions is equal to the expected domain sample count under uncontrolled conditions.

To improve the precision of the estimates, the use of auxiliary variables through the ratio estimators, generalized regression estimators, and post-stratified estimators is also applicable for domain estimation. However, in all cases, a large domain size is required to achieve highly efficient estimates.

Domains with very small sample size (or zero) result in estimators which are biased and have high variance. This problem of small domains has been the focus of recent studies. Special techniques like the Synthetic estimator, SY, both for count and ratio (Sarndal et al, 1992), Modified regression estimator, MRE (Hidiroglou and Sarndal, 1985), and Dampened regression estimator, DRE (Sarndal and Hidiroglou, 1989) have been proposed. The SY estimator showed the advantage of reduced variance but can be badly biased in some domains. The MRE estimator gave a small bias in those domains where the SY estimator is greatly biased and in other domains, nearly unbiased and has an advantage of reduced variance compared to the SY estimator. For very small sample size n_d, both SY and MRE estimators may yield estimates outside the range of acceptable values. The DRE estimator effectively removes the likelihood of "wild" estimates while having the advantage of reduced variance (Sarndal and Hidiroglou, 1989). The comparisons of the above estimators were done theoretically and confirmed through a Monte Carlo simulation study based on Canadian business survey data. The application of these methods to real survey data is still under review and study.

Most surveys, including business surveys, use only a simple domain estimator where estimates are required for domains or parts of strata.

Chapter 5

Repeated Surveys

Most business surveys are repeated across time using the same population in order to estimate the same characteristics at different points of time. The information collected on the previous occasion is often utilized to improve the estimator of the current period. Surveys could be conducted either on a monthly, quarterly, sub-annual or annual basis. A dimension of time is therefore added to the choice of sampling design and the method of estimation (Koeijers and Hilbink, 1998a). There are several issues to consider when sampling is done across time. These are the nature of matching in the sample at different occasions, the fraction of sample to be retained, as well as cost, efficiency and administrative convenience (Singh and Chaudhary, 1986).

A key question to be answered is whether to use the same sample on each occasion, to draw a completely new sample, or to use a mixture of the old and the new. The decision to change or retain sampling units depends on the objective of the successive surveys. The objectives may include the following:

1. To estimate the net or gross change in $\overline{Y}$ or Y with the intention of studying the effects of some events acting upon a population,
2. To estimate the average value of $\overline{Y}$ or Y over all occasions,
3. To estimate an average value of $\overline{Y}$ or Y for the most recent occasion.

If the characteristic of the population changes rapidly with time, the interest would be on the estimate of the current occasion (e.g. Objective 3). If the population changes slowly over time, the estimate of the average over all occasions may be adequate (e.g. Objective 2).

The strategies that will meet the objectives mentioned above are shown below. Suppose there are two successive occasions and, using the same sampling design, a sample of size n is selected on each occasion. The measurements on the same unit between two successive occasions most likely will have a positive correlation, ρ.

- For Objective 1, keeping the same sample on both occasions is preferable. Cochran (1977) showed that for occasions 1 and 2, the variance of the estimated change on the same unit is $\left(S_1^2 + S_2^2 - 2\rho S_1 S_2\right)$ while the variance of the estimated change using completely different units is $\left(S_1^2 + S_2^2\right)$ where S_1^2 and S_2^2 are the population variance for occasions 1 and 2 respectively. Since

ρ is likely to be positive, the estimated change from the same unit has a lower variance than that from different units.

- For Objective 2, it is better to draw a completely new sample on each occasion. Again in terms of variance, the estimate for the overall mean for the two occasions has variance equal to $\frac{\left(S_1^2 + S_2^2 + 2\rho S_1 S_2\right)}{4}$ if the same sample is used, and has variance equal to $\frac{\left(S_1^2 + S_2^2\right)}{4}$ if different samples are selected (Cochran, 1977). In this case, changing the sample completely results in lower variance for the estimate.

- For objective 3, either keeping all or changing all sample units on the second occasion is acceptable but replacing the sample partially may be better in some cases. Why this is so will be explained below.

Assume that the sample of size n is selected by Simple Random Sampling and that the population variance S^2 of y_i is the same for both occasions, $i = 1,2$. On the first occasion, y_1, the estimate for $\overline{Y}_1$, has variance $\frac{S^2}{n}$. On the second occasion, m units in the first sample are kept (m indicating matched) and the remaining u units $(u = n - m)$ are removed and replaced (u indicating unmatched) by units not selected on the previous occasion. Then $\overline{y}_2'$, the best estimator of $\overline{Y}_2$, consists of a combination of two independent estimates, one estimate for the unmatched units and another estimate for the matched units. The variance of $\bar{y}_2'$ is given by

$$V(\bar{y}_2') = \frac{S^2\left(n - u\rho^2\right)}{\left(n^2 - u^2\rho^2\right)}$$

where ρ is the correlation between units in population at first and second occasions (Cochran, 1977).

Using the above variance formula, if $u = 0$, meaning both samples are the same or completely matched, then $V\left(\bar{y}_2'\right) = \frac{S^2}{n}$.

If $u = n$, meaning the samples are completely different or there is no matching, then $V\left(\bar{y}_2'\right) = \frac{S^2}{n}$. Thus, for the current estimates, there is the same level of precision whether the same sample is kept or a completely new sample is drawn on every occasion. Now if $\rho = 0$, then $V\left(\bar{y}_2'\right) = \frac{S^2}{n}$ is the same as the cases with no matching or with complete matching. Hence there is no gain in precision whatsoever. However if $\rho = 1$, $V\left(\bar{y}_2'\right) = \frac{S^2}{2n}$, there would be a gain. In fact, the gain in precision over the no matching case is 100% when $\rho = 1$.

However, for other values of u, Cochran (1977) showed that for $0 < \rho < 0.8$, the percent gains in precision over complete matching are modest (i.e. less than or equal 17%). Therefore, when the ρ exceeds 0.8, the alternative of keeping some units and replacing the other units is found better than the two alternatives (Cochran, 1977; Singh and Chaudhary, 1986).

Take note that the 'keeping some and replacing other' strategy assumes that for a fixed sample size, the cost is the same for the replaced and retained units. If extra costs may be involved in drawing and contacting a new sample, cost considerations must not be overlooked.

5.1 Rotation Sampling

The scheme of retaining a portion (p) of the sampling units in one occasion and replacing the other portion (q) with new sampling units in the succeeding occasion is called rotation sampling. Other terms used to refer to it are sample rotation or partial periodic replacement of the sample (Schiopu-Kratina and Srinath, 1991) or rotating panels (Sigman and Monsour, 1995). The size of p depends on the nature and timing of the survey and the population concerned (Brewer et al, 1984). Using larger p values will tend to minimize abrupt changes in the estimates and perhaps reduced costs brought about by the inclusion of new units in the sample (Hidiroglou et al, 1991; Statistics Canada,1998b).

Another reason for employing rotation sampling is related to the burden on the respondents having to answer the same set of questions in the survey more than once. The response burden is distributed when the units are rotated in and out of the sample. The scheme allows the units to stay in the sample for a specific period of time and then be kept out of the sample for at least a certain period of time after they have been rotated out of the sample.

The timing for reporting of data is crucial in rotating panel surveys. In some cases, the data reported for the previous period and for the current period by the same respondent are subject to bias (Sigman and Monsour, 1995). This bias is referred to as differential or early reporting bias. The effect is more of a downward bias in the estimator of the current period than in the estimator of the previous period. It is possible that the respondent may have given a conservative set of responses for the current period because final data for the period are not yet available.

The problem of differential bias can be solved in two ways. The first way is to adjust the reporting time for the survey in order to give ample time for the respondent to get their final data for the current period. The second way is to use benchmarking. Benchmarking is a technique where the data are adjusted, and consequently the estimates, in order to conform to a more accurate survey conducted on a less frequent basis.

An example of the use of benchmarking is the monthly and annual Survey of Manufactures by Statistics Canada (Statistics Canada, 2001). Monthly estimates of shipments are reported for a number of industries in a particular sector. There is also a separate source, the annual survey, which gives the total shipments for the year. Theoretically, the monthly figures are assumed to be equal to the annual figure if summed up. But most often, the two sources of data give different results. If the annual source is considered as binding or exact, then by benchmarking, the monthly figures are revised so that the resulting series adds up to a given annual figure for the same period of time. Here the annual figure is the benchmark.

5.2 Methods of Sample Selection and Rotation

In business surveys, the different methods of sample selection and rotation are applied to the take-some strata and not to a take-all stratum where the same sample is used on every occasion. The methods to be discussed here are rotation group sampling, sampling rotation with PRN and repeated collocated sampling.

5.2.1 Rotation Group Sampling Scheme

Rotation group sampling is the widely used method in the monthly retail and wholesale trade surveys in United States and Canada.

For a stratified sampling design, rotation group sampling is described as a simple random sample design of randomly formed rotation groups (clusters) within each stratum (Srinath and Carpenter, 1995; Hidiroglou et al, 1991). The N_h population units within a stratum are randomly allocated to a pre-specified number of population groups, P. If $\frac{N_h}{P}$ is not an integer, then the P rotation groups will have unequal sizes. Then a simple random sample of p rotation groups is selected from the P population rotation groups. To determine the number of rotation groups p to be selected, the ratio of p to P must be approximately equal to the sampling fraction f_h (i.e. $\frac{p}{P} \cong f_h$) where $f_h = \frac{n_h}{N_h}$. All units within the selected rotation groups are then included in the sample. The rotation of sample units occurs by removing an in-sample rotation group and obtaining an out-of-sample rotation group to replace it.

Srinath and Carpenter (1995) illustrate the rotation group sampling using a hypothetical example. Let the population size be $N = 16$ and the sample size $n = 8$. The sampling fraction $f = 0.5$. Let's suppose that the units will stay in the sample for 4 occasions, setting $p = 4$. The total number of population rotation groups P is computed as $P = 4\left(\frac{1}{f}\right) = 4\left(\frac{1}{0.5}\right) = 8$. The 16 units in the population are allocated to the 8 population rotation groups, each rotation group containing two units .

The following procedure for allocation of units to the groups is used. Number the P rotation groups from 1 to P where the number is referred to as Assign Ordering. Arrange the P integers in a random order (e.g. 5,1,3,6,2,7,8,4) where the random sequence is referred to as the Rotation Ordering. The units in the population are assigned sequentially to the Assign Ordering such that the first unit is assigned to Assign Ordering 1, the second unit to Assign Ordering 2 and so on to the P^{th} unit which is assigned to Assign Ordering P. Then the $(P+1)^{\text{th}}$ unit is assigned again to Assign Ordering 1, the $(P+2)^{\text{th}}$ unit to Assign Ordering 2 and so on. Table 6 shows the assignment of the Assign Ordering, Rotation Ordering and the allocation of 16 units to the 8 rotation groups.

Table 6. Allocation of 16 units to the 8 rotation groups.

Assign Ordering	1	2	3	4	5	6	7	8
Rotation Ordering	5	1	3	6	2	7	8	4
Stratum units	1	2	3	4	5	6	7	8
	9	10	11	12	13	14	15	16

From the table, the rotation group in the first Assign Ordering is rotation group 5 consisting of units 1 and 9, followed by rotation group 1 consisting of units 2 and 10 and so on.

The Rotation Ordering is used for the selection and rotation of units. The rotation groups numbered 1 to p in the Rotation Ordering are included in the sample on the first occasion (i.e. referring to rotation Groups 1 to 4 which consist of units 2,10,5,13,3,11,8,16); on the second occasion, rotation group 1 rotates out of the sample while the $(p+1)^{\text{th}}$ rotation group rotates into the sample (i.e. rotation group 1 is replaced by rotation group 5); and so forth. Table 7 shows the selection and rotation of samples on 8 sampling occasions.

Table 7. Selection and rotation of units in 8 sampling occasions by Rotation Group Sampling.

Occasion	Selected Rotation Groups	Selected Sample Units
1	1,2,3,4	2,10,5,13,3,11,8,16
2	2,3,4,5	5,13,3,11,8,16,1,9
3	3,4,5,6	3,11,8,16,1,9, 4,12
4	4,5,6,7	8,16,1,9, 4,12,6,14
5	5,6,7,8	1,9, 4,12,6,14,7,15
6	1, 6,7,8	2,10, 4,12,6,14,7,15
7	1,2 7,8	2,10,5,13, 6,14,7,15
8	1,2,3 8	2,10,5,13,3,11, 7,15

After the selection and rotation of sample units at the current occasion, the sampling frame needs to be updated in preparation for the next occasion survey. There are three things to be considered in the updating of the frame: 1) the emergence of new business units (births), 2) the units which terminate business activity (deaths) and 3) the changes of classification variables used in stratification (e.g. geography, industry or size).

- Updating of Births. Before the time of selection and rotation of samples , the newly recorded units (births) should be added to the sampling frame. Births within a stratum are assigned in sequence to the rotation groups. Using the previous example, suppose that on the second occasion there are 4 new births (now labelled units 17,18,19,20) in a stratum. In Table 6, the last unit (i.e. unit 16) on the first occasion was assigned to Assign Ordering 8 and Rotation Ordering 4. The first new birth (i.e. unit 17) of the second sampling occasion will be assigned to Assign Ordering 1 and hence Rotation Ordering 5; the next birth (i.e unit 18) to Assign Ordering 2 and hence Rotation Ordering 1; and so forth. Table 8 shows the Assign Ordering, Rotation Ordering and allocation of 4 new births to the rotation groups.

Table 8. Allocation of 4 new births to the 8 rotation groups.

Assign Ordering	1	2	3	4	5	6	7	8
Rotation Ordering	5	1	3	6	2	7	8	4
Stratum units	1	2	3	4	5	6	7	8
	9	10	11	12	13	14	15	16
New births	17	18	19	20				

From the table, at the first Assign Ordering, rotation group 5 will now consists of units 1,9 and 17; followed by rotation group 1 consisting of units 2,10 and 18; and so on. With this scheme, it is important to always note the Assign Ordering and Rotation Ordering to which the last birth is assigned on each sampling occasion, so that continuity of rotation group assignment for subsequent births will be maintained (Srinath and Carpenter, 1995).

The selection and rotation of samples will remain the same as shown in Table 7.

Schiopu-Kratina and Srinath (1991) designed a different birth strategy in the sampling of births for the Statistics Canada's SEPH Survey . The strategy supposed that if B births are added to the current period, then a sample of b births, determined by $b = f_h B$ (where f_h is the sampling fraction in stratum h), is selected for that period. The selected b births are then randomly allocated to the p rotation groups. With the belief that the births may differ from the "old" units, the strategy will ensure that the probability of rotating out of the sample for births is the same as the old units.

- Removal of Deaths. The units that cease business operation (or deaths) must also be removed from the sampling frame if they are detected through an external source that is independent from the survey or if they have been dead for more than a pre-specified period of time (Hidiroglou et al, 1991). When deaths are removed from the sampling frame, the chance of having an imbalance in the number of units per rotation group is possible.

  Deaths identified from the sample survey itself are handled differently. These units are not removed from the frame or sample but associated variables are assigned a zero value at the estimation stage.

- Changes in Classification. The changes in the classification variables of units are implemented in the sampling frame only if they are detected by a source independent of the survey. A simple approach to implement such changes in the frame, as suggested by Srinath and Carpenter (1995) is to treat the units with a new classification variable as births and treat the units with old classification variable as deaths. The guidelines for updating births and deaths will then be followed. But again there is the possibility of having an imbalance in the number of units per rotation group as a result.

5.2.2 Sample Rotation with PRN

The Permanent Random Number (PRN) sampling techniques are being widely used in business surveys by countries like Sweden, Australia, New Zealand, Finland and France. Sample rotation in sampling designs that use permanent random numbers (PRN) can be handled easily. Examples of these designs are sequential SRSWOR, Poisson sampling, sequential Poisson sampling and collocated sampling.

As mentioned in the preceding chapter, the main feature of the technique is the permanent random number associated with every unit in the sampling frame (Ohlsson, 1995). These PRNs are the same on every sampling occasion. New units or births are also assigned with a permanent random number and recorded in the frame. Deaths are withdrawn from the frame together with the associated random number. On each sampling occasion, the PRNs are used to select a new sample. The units that remain (called persistants) for the next occasion depend on the number of births and deaths that have occurred during the previous period. In Sweden, births and deaths comprised less 15% of the list frame (Ohlsson, 1995). Thus, the desired overlap of samples between successive sampling occasions is most frequently met.

Two methods used for the rotation of samples in PRN designs will be discussed namely, 1) the constant shift method, and 2) the random rotation cohort (RRC).

- The constant shift method shifts the starting points of all surveys by a specified distance to the right or to the left. For example, assume that the sample should be rotated every year and the units with selection probabilities of 0.10 or less will stay in the sample for only 5 years, and the shift of starting points is 0.02 to the right each year (Ohlsson, 1995). So initially, a sample is chosen using all those units with PRN ≤ 0.1. That will result in a certain size of sample; albeit a random sample size. On the next occasion, a sample consists of units with PRN between 0.02 and 0.12; the units with PRN < 0.02 are rotated out of sample and replaced by units with PRN between 0.10 and 0.12. On the third occasion, the sample consists of units with PRN from 0.04 to 0.14 and so on. This gives a 20% rotation and the unit can be expected to be out of sample after 5 years. Those units with inclusion probabilities < 0.02 will stay in the sample for only a year.

- The random rotation cohort (RRC) method maintains the starting point of each occasion constant over time but the permanent random numbers change. The technique permanently assigns each of the units to a rotation cohort, of which there are a pre-specified number. The same thing is done to the births when they enter the frame. Using the constant shift example, assuming the sample is to be rotated yearly and there are five rotation cohorts, the PRN's of cohort 1 are shifted 0.10 to the left after the first year. Then for next year, the PRN's of cohort 2 are shifted 0.10 to the left and so on.

This is the method being used by Statistics Sweden in about 15 annual and subannual surveys. Samples of most surveys are drawn by December and a few annual surveys during May. Rotation of samples are done before sampling.

5.2.3 Repeated Collocated Sampling

Repeated collocated sampling is still under study at Statistics Canada and their application of it to real surveys is not yet determined (Srinath and Carpenter, 1995). Nonetheless, the method is worth knowing for future use.

At the first sampling occasion, the units within a stratum are arranged in random order. A random number, ε, drawn from a uniform distribution over the interval [0,1] is assigned to each unit. Then for each i^{th} ordered unit in the population, a sample selection number $SSN(i) = \frac{(i-\varepsilon)}{N}$ is defined. The selected sample are those units whose sample selection numbers fall within the interval $[0, f]$, where f is the desired sampling fraction. Rotation of sample unit is done by shifting the sampling interval by a measure of overlap, a. In the second sampling occasion, the selected samples are those units whose sample selection numbers fall within the interval $[a, a+f]$.

For the births on each sampling occasion, a sample selection number (SSN) is also assigned to the jth ordered unit of the Q new births since the last sampling occasion and defined as $SSN(j) = \frac{(j-\varepsilon)}{Q}$. Then the births with sample selection number falling within the interval $[a, a+f]$ are selected for the next sampling occasion.

The deaths are handled in the same way as in the rotation group sampling procedure. Their removal may cause imbalance in the number of units in the sampled interval.

The changes in the classification variables are handled in the same way as in the rotation group sampling procedure (i.e. treating the units coming into a stratum as births and the units going out of a stratum as deaths). The only difference is that there is an option to either assign a new random number to the units or retain the old random number as these units are shifted to the new strata.

5.3 Estimation Procedures

The estimation procedures for repeated surveys can be categorized into 3 types namely , 1) estimation of mean (or total) on the current occasion, 2) estimation of change between two occasions, and 3) estimation of value of the sum of the means (or total) over two occasions. Each type has its own form of estimators. In rotation sampling, where the second sampling occasion can opt for partial overlap of sample units with the first occasion sample, the information from the first occasion is often used to improve the estimation for the second occasion.

Consider a population of N units. As before, suppose that the samples of size n are selected on two successive occasions by simple random sampling method and the variability of the population remain the same over occasions. On the second occasion, m matched units in the first sample are kept and the remaining u unmatched units $(u = n - m)$ are removed and replaced by units not selected on the previous occasion. Define $m = np$ and $u = nq$ where p is the portion of sample retained out of the sample of size n and q is the portion of sample replaced, and that $p + q = 1$. Define further that

$\bar{y}_{tu}$ = Mean of unmatched portion on occasion t

$\bar{y}_{tm}$ = Mean of matched portion on occasion t

$\bar{y}_{t}$ = Mean of whole sample on occasion t.

The estimators for each of the categories above are as follows:

- Estimation of mean on the current occasion (Cochran, 1977; Singh and Chaudhary, 1986). The estimator for the mean can be derived from 1) a difference estimator based on the matched sample units on both two occasions and/or 2) estimator based on the unmatched sample. When combined, the estimator is referred to as the composite estimator. Both estimates are weighted according to the inverse of their respective variances.

The form of the difference estimator based on the matched sample units is

$$\bar{y}_{2m}{}' = \bar{y}_{2m} + b\left(\bar{y}_1 - \bar{y}_{1m}\right)$$

where b is the regression coefficient of y_2 on y_1 from the matched sample units. The variance of the estimator is

$$V(\bar{y}_{2m}{}') = \frac{S^2\left(1-\rho^2\right)}{m} + \rho^2\frac{S^2}{n} \quad = \frac{1}{W_{2m}}$$

The estimator based on the unmatched sample units is

$$\bar{y}_{2u}{}' = \bar{y}_{2u}$$

and the variance of the estimator is

$$V\left(\bar{y}_{2u}{}'\right) = \frac{S^2}{u} = \frac{1}{W_{2u}} \quad .$$

Thus the combined estimator of $\overline{Y}_2$ for the current occasion is

$$\overline{y}_2' = \varphi_2 \overline{y}_{2u}' + (1-\varphi_2)\overline{y}_{2m}'$$

where $\varphi_2 = \dfrac{W_{2u}}{W_{2u} + W_{2m}}$.

The variance of the combined estimator is

$$V(\overline{y}_2') = \frac{S^2(n - u\rho^2)}{(n^2 - u^2\rho^2)} .$$

This composite estimation method is being used by Statistics Canada in periodic surveys with a large overlap between occasions (Statistics Canada, 1998b). This method is also used by Canadian Labour Force Survey (Statistics Canada, 2002d).

- Estimation of change (Raj, 1968; Singh and Chaudhary, 1986). The unbiased estimator of the change $\overline{Y}_2 - \overline{Y}_1$ is

  $$\overline{d} = p\frac{(\overline{y}_{2m} - \overline{y}_{1m})}{(1-q\rho)} + q\frac{(1-\rho)(\overline{y}_{2u} - \overline{y}_{1u})}{(1-q\rho)} \text{ and}$$

  the variance of the change estimate is

  $$V(\overline{d}) = \frac{2(1-\rho)S^2}{n(1-q\rho)} .$$

- Estimation of the value of the sum of means (or totals) over two occasions (Raj, 1968; Singh and Chaudhary, 1986). The unbiased estimator of the combined means, $\overline{Y}_S = \overline{Y}_1 + \overline{Y}_2$, is

$$\hat{\overline{Y}}_S = \frac{p}{(1+q\rho)}(\overline{y}_{2m} + \overline{y}_{1m}) + \frac{q(1+\rho)}{(1+q\rho)}(\overline{y}_{2u} + \overline{y}_{1u})$$

and the variance of the combined means estimator is

$$V(\hat{\overline{Y}}_S) = \frac{2(1-\rho)S^2}{(1+q\rho)n}.$$

For the estimator of the combined totals, $Y_S = Y_1 + Y_2$, simply multiply the unbiased estimator of the combined means, shown above, by N and the variance of the combined means estimator by N^2 (Sarndal et al, 1992).

The UK Office For National Statistics (ONS) uses different forms of estimators in its repeated monthly business surveys. For instance in ONS Monthly Retail Sales Inquiry Survey, a modified composite estimator is used to estimate the month to month total change . Since the amount of monthly rotation is small and the correlation of responses between successive months is very high (i.e. > 0.95), only the estimator based on matched sample units is considered. Thus the method is referred to as matched pairs estimator (Kokic and Jones, 1997). The method assumes negligible contribution from the unmatched sample units.

There is another estimator being used to estimate the total difference between successive months, called matched pairs estimator with ratio updating (Kokic and

Jones, 1997). The form of the estimator based on the matched sample units is

$$\hat{Y}_{M\,t} = \sum_{h=1}^{L} \hat{Y}_{M\,t\,h}$$

where $\hat{Y}_{M\,t\,h} = \hat{Y}_{M\,,t-i,h} \dfrac{y_{th}}{y_{t-1,h}}$,

y_{th} = total value of variable y at occasion t in stratum h,

$y_{t-1,h}$ = total value of variable y at occasion $t-1$ in stratum h.

In order to apply this estimator at the first sampling occasion, an initial starting estimate must be supplied. The initial estimator suggested is the ordinary ratio estimator of the form

$$\hat{Y}_{R1} = \sum_{h=1}^{L} X_{1h} \frac{y_{1h}}{x_{1h}}$$

where X_{1h} = h^{th} stratum total of auxiliary variable x at first occasion,

y_{1h} and x_{1h} denote the sample totals in the stratum h at first occasion .

Another estimator takes into account the auxiliary information and may produce a more accurate estimate for total differences between consecutive months (Kokic and Jones, 1997). The form of this estimator is

$$\hat{Y}_{D\,\mathrm{t}} = \sum_{h=1}^{L} X_{t-1,h} \frac{d_{t\,\mathrm{h}}}{x_{t-1,h}} + \hat{Y}_{D\,\mathrm{t-1}}$$

where $d_{\mathrm{t\,h}} = \sum_{i=1}^{m} d_{t\,i} = \sum_{i=1}^{m} \left(y_{ti} - y_{t-1,i} \right)$ = sum over the response differences over the matched sample units.

Chapter 6

Nonresponse and Imputation

Nonresponse is a non-sampling error and is defined as the failure to obtain a response or information from one or more units in the sample. Unit nonresponse occurs when units fail to respond to the survey and item nonresponse occurs when units respond to the survey but fail to give information to some questions included in the survey. The occurrence of nonresponse is inevitable in surveys. Its presence in business surveys poses a problem because the nonrespondents may not behave similarly to the respondents in terms of the characteristics of interest. Nonresponse introduces bias to the estimates. Handling of nonresponse or missing data when done inappropriately will result in more potential biases in the estimates. The actions needed to cope with nonresponse are influenced by budget, time, use of data and risk of bias (Statistics Canada, 1998b). Item nonresponse is handled by replacing the missing value with feasible data value using an imputation technique. Unit nonresponse is dealt with by a weighting adjustment approach or by imputation if sufficient auxiliary information from other sources or past occasions is available (Srinath and Carpenter, 1995; Sarndal et al, 1992).

This section will discuss the issues related to nonresponse and imputation from the business surveys perspective. The different imputation methods used by statistical agencies of various countries will be presented as well.

6.1 Nonresponse Issues

Nonresponse causes bias in survey estimates when nonrespondents differ from respondents in the characteristics being measured. In the case of unit nonresponse, the effect is an increase in sampling variance due to a decrease in the sample size.

6.1.1 Response and Nonresponse Rate

Response rate, or nonresponse rate, is defined as a ratio involving counts or weights which represent a given category of response, or nonresponse, in some domain of interest (Hidiroglou et al, 1993). The manner by which the rate is computed varies because of the varying ways of defining the denominator of the formula. Some collecting agencies use the number of contacts for the denominator while others used the number of eligible sample members (Lessler and Kalsbeek, 1992). One of the issues in nonresponse is the need to have a standardized definition of nonresponse (Lessler and Kalsbeek, 1992; Hidiroglou et al, 1993). Is the unit which is included in the sample but turns out to be a non-member of the survey target population (also referred to as out-of-scope sample member) considered to be a nonrespondent?

The reasons given for nonparticipation by the respondents is important to the interpretation of the concept of nonresponse (e.g refusal, noncontact, misunderstanding of the questions, no available data for the particular period, sensitive nature of the data, etc.). These reasons need to be recorded and monitored (Statistics Canada, 1998b; Hidiroglou et al, 1993).

Similarly, the response and nonresponse rate must be reported in a standard way to facilitate comparability between surveys. Statistics Canada has established guidelines for reporting the nonresponse rate applied to both business and social surveys. A framework intended to classify the sampled units in a survey into responding, nonresponding and out-of-scope units was developed by Drew and Gray (Hidiroglou et al, 1993). The schematic diagram is shown in Figure 5.

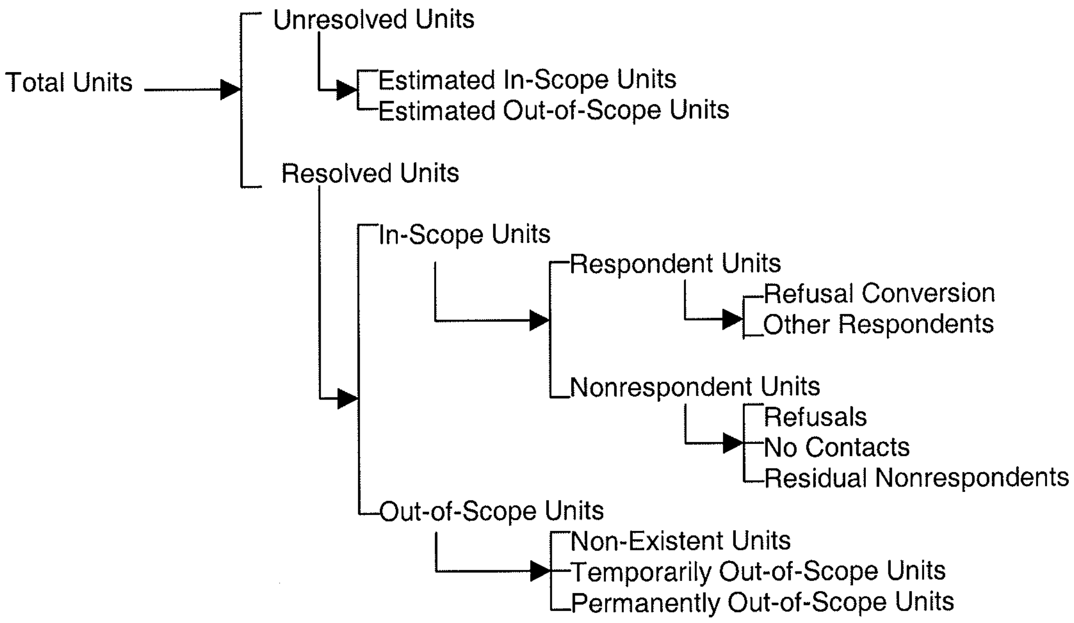

Figure 5. Respondent/nonrespondent components at the data collection phase. Source: Hidiroglou, M.A.; Drew, J.D.; and Gray, G.B. (1993). "A Framework for Measuring and Reducing Nonresponse in Suveys. *Survey Methodology* 19, No 1, p. 82.

The Total Units are units which are thought to belong to the population of interest before the survey process begins. The Total Units is broken down into two main categories; namely, Resolved Units and Unresolved Units. Resolved Units are units whose status is known by the cut-off date of the survey data collection. The Unresolved Units are units whose status have not been determined by the end of data collection period.

The number of Resolved Units is further subdivided into 1) In-Scope Units, units which belong to the survey target population, and 2) Out-of-Scope Units, units not belonging to the survey target population.

- The In-Scope Units category is broken down into 1) Respondent Units, units which have provided the needed information and 2) Nonrespondent units, units which failed to provide the needed information.

- The Out-of-Scope Units category is sub-divided into 3 sub-categories, which may or may not be applicable to a particular survey. These are 1) Non-Existent units, which may include business deaths, 2) Temporarily Out-of-Scope Units, which may include business units which were inactive at the time of survey due to some reasons but might be active later, and 3) Permanently Out-of-Scope Units.

Based on the diagram, the response rate must be calculated as the ratio of Respondent Units to the sum of In-scope Units and Unresolved Units. On the other hand, the nonresponse rate must be calculated as the ratio of the sum of the Nonrespondent Units and Unresolved Units to the sum of In-scope Units and Unresolved Units.

6.1.2 Handling Nonresponse

The best strategy for handling the problem of nonresponse is to keep it from becoming too large right from the start of the survey (Lessler and Kalsbeek, 1992). The survey design, data collection procedures and operations, data follow-up and data editing procedures are factors that can influence response rate. These factors are described by Statistics Canada (1998b) as follows: "the quality of survey frame in terms of coverage and contact information, target population, sampling method, method of data collection (such as mail, telephone, personal interview, computer assisted interview), time of year and length of collection period, response burden imposed (length of interview, difficulty of subject matter, periodicity of the survey), nature of subject matter in terms of sensitivity, length and complexity of the questionnaire, follow-up methodology, expected difficulties in tracing respondents, prior experience with same type of survey, prior experience and demonstrated ability of collection staff, workload of collection staff, established relationships with respondents, the communications strategy, the total budget, the allocation of budget to the various operations, the importance of the survey to users and respondents and respondents incentives."

During the period of data collection, unit nonresponse is dealt with by implementing several strategies namely: 1) preventive measures, 2) nonrespondent substitution, 3) nonrespondent subsampling (Lessler and Kalsbeek, 1992).

- Preventive measures. The preventive methods include high priority mailing to increase the chance of reaching a sample unit; clearer interviewer assignment materials such as correctly spelled full name of the contact person, complete address; follow-up reminders in the form of postcard, letter, or telephone call; and lead letter signed by a credible person to inform, assure and motivate a prospective sample member .

- Nonrespondent substitution. Another way to handle nonresponse is by substitution. To be assured that the sample size will not be less than the intended sample size because of nonresponse , a method of replacing each nonrespondent with another member of the population not included in the original sample is often done. The substitute respondent may be selected at random usually from the same stratum as the nonrespondent, or nonrandom based on a predetermined set of criteria. This method has two advantages, namely, the targeted sample size is achieved and nonresponse bias is reduced but not eliminated. The drawbacks of the substitution method are: the diminishing effort to obtain response from the original sample unit knowing that a replacement is waiting and the potential of overstating the response rate in the calculation, e.g. including substitutes in the numerator but failure to include them in the denominator of the calculated rate.

- Nonrespondent subsampling. The limited amount of time and budgetary constraints associated with many surveys brought the development of the nonrespondent subsampling as a strategy. The procedure involves a random selection of a subsample from the identified initial nonrespondents followed by employing more intensive and costly efforts (e.g. by telephone or in person) to obtain responses from these subsampled units. This method is similar to two-phase sampling.

Statistics Netherlands has undertaken some practical steps to minimize nonresponse in business surveys (Willeboordse, 1998c). These are as follows:

- The timing of the sending out of questionnaires must be close to the date when accounting data are available and should not be more than a week delayed.
- The period of time before the final date of return of questionnaires should not be too long for the respondents to forget. For quarterly surveys, the interval should not be more than two weeks.
- Reminders should first be sent within a week after the final return date.
- Written reminders should be sent first, followed by a telephone reminder since the latter is more expensive.
- Give priority to larger businesses or influential units than to small units.
- Always contact respondents as soon as possible after the return of questionnaires with incomplete and implausible data to let them know that their attitude and behavior is important
- Encourage the respondents to contact the collecting agency for answers to questions and make sure that there is always a qualified staff person to respond to the call.
- Update the mailing addresses.
- Try to make appointments with contact persons.
- Reward or give incentives to the respondents with a copy of the report or publication.
- Train survey staff in handling difficult questions.

- In case of statutory surveys, consider prosecution as the last resort for those respondents consistently refusing to cooperate in the survey.

After having done all preventive measures to minimize nonresponse, there are two ways to compensate for remaining nonresponse at the estimation stage. It could be done by means of weighting adjustment and by imputation (Statistics Canada, 1998b; Sarndal et al, 1992; Lessler and Kalsbeek, 1992). A discussion of weighting adjustment for nonresponse was described earlier in Chapter 4. The imputation technique will be discussed in the next section.

6.2 Imputation Issues

Imputation is a process of replacing missing, invalid or inconsistent responses by feasible data values in order to create a complete data file for analysis (Statistics Canada, 1998b; Kovar and Whitridge, 1995). The advances in technology have made the process of imputation easier to compute, monitor and evaluate as compared to the manual process which is more subjective and requires subject matter experience and knowledge. An automated imputation system yields more objective and reproducible results (Statistics Canada, 1998b; Kovar and Whitridge, 1995). Because of this, care should be made not to abuse or misuse the availability of automatic imputation systems. Some issues when imputing missing data are as follows:

- Imputation on a large-scale basis (or mass imputation) for convenience in operation should be handled with care. Where unit nonresponse arises because of nonresponse as dictated by the design (e.g. two-phase sampling), or because certain questions are not applicable to the respondents for which reason they are not required to report certain characteristics, imputation is not required but weighting adjustment may be needed. However mass imputation is found helpful for cases where large amounts of data are missing; where quick ad hoc estimates are required; and where second-phase samples weights are complicated to compute (Kovar and Whitridge, 1995).

- Data sets with imputed values should not be treated as "clean" observed data. All imputed values must be flagged properly in the data file to distinguish them from the observed values (Statistics Canada, 1998b; Sarndal et al, 1992). The methods of imputation and sources of imputed values must be also clearly identified. This is important if other analysts wish to perform another imputation method on the data file. While imputation results in a complete data matrix and can produce ad hoc estimates of means and totals at various domains of interest, their use for data analysis can be misleading (Statistics Canada, 1998b; Sarndal et al, 1992; Lessler and Kalsbeek, 1992) . This is because the variances and covariances are underestimated. The extent of underestimation can be of the order 2 to 10% in case of 5% nonresponse and as high as 10-50% where there is 30% nonresponse (Kovar and Whitridge, 1995). Several techniques have been proposed to remedy the problem of variance underestimation. Examples of these techniques are the model assisted methods, jackknife method and multiple imputation method (Sarndal et al, 1992).

- The resulting imputation-revised data sets are not verifiable since the true data values are unknown for nonrespondents. However some measures of quality on the imputation process itself or some indications of its success and failures need to be noted.

In spite of its shortcomings, imputation has been found to be a useful and practical tool by data collecting agencies. Estimation of data files with imputed data is simplified and generation of ad hoc tabulations are done quickly and consistently. Imputation is useful in preserving known relationships between variables, in addressing systematic biases, and in reducing nonresponse bias (Kovar and Whitridge, 1995).

6.2.1 Imputation methods

Imputation methods can be broadly classified into two categories; 1) a deterministic category where the imputed values are determined uniquely, or 2) a stochastic category where the imputed values are subject to some degree of randomness (Kalton and Kasprzyk,1986; Kovar and Whitridge,1995).

Examples of deterministic methods are logical imputation, mean imputation, historical imputation, hot deck method, ratio and regression imputation, and nearest-neighbor imputation (Kovar and Whitridge, 1995; Sarndal et al, 1992; Lessler and Kalsbeek, 1992; Statistics Canada, 1998b; Verboon, 1998)

- Logical Imputation or Deductive Imputation refers to the method where the missing value can be substituted with a predicted value attained by logical conclusion. As an example, a missing expense value can be deduced as being zero when the reported income is zero. Deductive Imputation is usually derived based on known and existing functional relationships among variables (Lessler and Kalsbeek, 1992). These relationships can assume different forms depending on the variable. For this reason, Logical or Deductive Imputation is often included as part of data cleaning and editing (Kovar and Whitridge, 1995).

- Mean Imputation is a simple method of assigning the mean of the respondent item to every missing value. The mean value may be the overall mean, the stratum mean, or the post-stratum mean.

  The Mean Imputation is easy to execute and reasonably effective in reducing bias of point estimates for means and totals. However there are two drawbacks to the method. The first drawback is that it destroys the distributions and relationships among variables because of the artificial spike created at the class mean value (Kovar and Whitridge,1995; Lessler and

Kalsbeek, 1992). The other drawback is that the variance estimates are underestimated and the level of confidence in the interval is therefore invalid (Sarndal et al,1992). Mean Imputation is therefore suitable for studies where analysis is limited to simple point estimates without accompanying variances and where there is limited budget for proceeding to complex analysis where variance estimates are required (Lessler and Kalsbeek, 1992). Kovar and Whitridge (1995) recommend choosing this method as a last resort only after all the other methods have failed .

In some cases, Median Imputation instead of Mean Imputation is used to eliminate the effect of outliers (Verboon, 1998).

- Historical Imputation or Cold-deck Imputation replaces the missing data by a reasonable value from other sources such as past surveys or historical data (Lessler and Kalsbeek, 1992; Sarndal et al, 1992; Verboon, 1998). It is most useful in periodic business surveys for variables that tend to be stable over time. When the relationship between previous and current occasions is strong, this method is found most effective (Kovar and Whitridge, 1995).

When the other sources are external to the survey (e.g. administrative sources), the imputation method is called *Exact Match Imputation* (Sarndal et al, 1992). Potential problems associated with this method are related to the credibility of the external source to give quality data value for a match unit, and to the cost and complexity of the unit matching process (Lessler and Kalsbeek, 1992).

- Hot-deck Imputation replaces the missing data by a donor value from respondents in the current survey. There are several variations of Hot-deck procedures such as Sequential Hot-deck Imputation, Hierarchical Hot-deck Imputation and Distance Function Matching or Nearest Neighbor Imputation, Random Overall Imputation, and Random Imputation Within Classes (Sarndal et al, 1992).

Sequential Hot-Deck method replaces the missing data by a donor value from the last responding unit preceding the missing item in the same imputation class as the recipient. The arrangement of units in the data file plays a role in this method. In business surveys, the order of the files is often according to geography or size class. Care must be taken when using sorted files in order to avoid introduction of any systematic pattern, i.e. the donor value is always larger or always smaller than the recipient (Kovar and Whitridge, 1995).

Hot-deck Imputation is appealing because this method uses actual, observed data in imputation as opposed to mean, ratio or regression imputation. Moreover, it can be easily incorporated into existing data validation programs but the system becomes very large making it difficult and costly to maintain. Another drawback of this method is that it often leads to multiple uses of donors which lowers the precision of the estimates (Lessler and Kalsbeek, 1992).

- Nearest Neighbor or Distance Function imputation replaces the missing data by the value from a respondent classified to be "nearest" by some measure of distance defined in terms of known auxiliary variable values. Just like the Hot-deck method, the Nearest Neighbor method uses actual observed data and can use donors repeatedly when there is high nonresponse rate within an imputation class (Kovar and Whitridge, 1995). The distance function matching can be used both for continuous and categorical variables (Lessler and Kalsbeek, 1992).

- Ratio and Regression Imputation replaces the missing value with the ratio or regression predicted value obtained by making use of auxiliary variables. This method is a very effective method in business surveys especially when auxiliary information is sufficiently available among the sampled units and when the auxiliary variable is highly correlated with the variable to be imputed (Kovar and Whitridge, 1995). The method performs well both in cases of random and nonrandom nonresponse. Independent regression variables may be continuous or discrete. The disadvantage of using this method lies in the distributional problems where several units have the same imputed values due to the use of the same values of the independent (auxiliary) variables. Another disadvantage is the added time and effort to develop the model and do the necessary verification or adjustment regularly.

- Multiple imputation uses several different values as candidates to replace for a single missing variable in order to obtain better estimates for variances and covariances. The drawback of this method is the huge amount of work required for the computation of estimates (Sarndal et al, 1992).

When the above methods are modified by adding a degree of randomness on the imputed values, then the methods will be classified under the category of stochastic methods. Examples are the random hot deck method, regression with random residuals and any deterministic method with added random residual (Sarndal, et al, 1992; Lessler and Kalsbeek, 1992).

After identifying all these imputation methods, the decision on which method to choose is still difficult. The statistical effectiveness and the practical implication of their usage are the factors to be considered (Lessler and Kaslbeek, 1992). The use of complex methods (e.g. mutliple imputation applied to hot-deck method) may not be practical when there is no available staff that will interpret the findings or when computer software is not available. Different methods have their own merits depending on the circumstances. A good imputation method reduces the nonresponse bias and produces imputed values which are internally consistent (Statistics Canada,1998b) . As Verboon (1998) remarked "no method is superior to the others in all circumstances".

6.2.2 Imputation Methods Used by Collecting Agencies

Some of the imputation methods used by four collecting agencies namely, 1) Statistics Canada, 2) US Bureau of the Census, 3) UK Office of National Statistics and 4) Australian Bureau of Statistics , will be described in this section.

- Statistics Canada. For periodic business surveys like monthly survey of manufacturing, monthly retail/wholesale trade surveys, Statistics Canada uses a number of imputation methods for nonresponse (i.e. complete or partial) depending on the variables requiring treatment . The methods include 1) historical imputation, 2) ratio imputation, 3) using industry-province cell trends and 4) reference to the related annual survey, e.g. Annual Survey of Manufacturing being the reference for the Monthly Survey of Manufacturing (Statistics Canada, 2001). Then after the imputation, a final verification of the responses that have been imputed is performed.

  For Annual Surveys, Statistics Canada has used the following methods: 1) historical data (i.e. previous year's data), 2) use of other sources like tax data, 3) industry trends and 4) reference from related periodic surveys, e.g. the Monthly Survey of Manufacturing being the reference for the Annual Survey of Manufacturing.

  Statistics Canada has two automated systems which include different algorithms for imputing missing data (for either continuous or categorical type of data). These systems are the Generalized Edit and Imputation System (GEIS) for quantitative and the Numerical Imputation Method (NIM) for categorical data (Statistics Canada, 1998b).

- US Bureau of the Census. The US Bureau of the Census conducts monthly and annual economic surveys on retail/wholesale trade, manufactures, service industries, mineral industries and construction in addition to the associated censuses which are conducted every 5 years (Chapman et al, 1986) .

In its Monthly Retail Trade Survey, for the variable *Retail Sales*, missing values are imputed by the product of the previous month's figure and a *ratio of identicals* (Chapman et al, 1986). The ratio of identicals is defined as the ratio of the weighted sum of the current sales to the weighted sum of the previous month sales for all sample units in the same cell group which has reported sales for both the current and previous months. The cell group is defined by the first 3 digits of the SIC code, by establishment and by sales size class. The weights are the inverse of the selection probability for the responding unit. If a nonrespondent is selected for the first time in the current period, then the previous month sales is imputed from the sales reported in the most recent census, if available. If a nonrespondent is a birth case and was not in the most recent census, then the two months of sales data generally provided when the company was added to the frame will be seasonally adjusted and inflated to an annual-based figure. The obtained value is then used in the imputation as though it were a census sales value.

The historical data may come not only from recent censuses but also from administrative sources like tax form information from Internal Revenue Service (IRS).

The Truck Inventory and Use Survey (TIUS) which is conducted every five years uses a weighting adjustment method to compensate for unit nonresponse.

- UK Office of National Statistics. The UK Office for National Statistics uses more than one method of imputation in a typical short term business survey namely, the Simple Autoregressive Imputation, if data from previous periods are available and Ratio Imputation using auxiliary variable if the nonresponder was selected for the first time (Full, 1999).

Consider a population of N units. Suppose that a sample of size n is selected by SRSWOR. The set of respondents is of size m and the set of nonrespondents is of size $n-m$. The variable of interest is y and the auxiliary variable available to each unit in the population is x. For each nonrespondent, a value $\hat{y}_k$ is imputed.

The steps used in the Simple Autoregressive Imputation method are as follows:

1. For each of the units that responded both in the current and previous period, calculate the ratio of the variable of interest as $\frac{y_{ti}}{y_{t-1,i}}$, i=1,2,..m.

2. Trim the largest 20% of the ratios and the smallest 10% of the ratios.

3. Calculate a trimmed mean of the remaining ratios for the current period-on-period growth, b_1 and the growth for the same period a year previous, b_2 by the formula below.

$$b_1 = \frac{\sum_{i=1}^{m_1} \frac{y_{ti}}{y_{t-1,i}}}{m_1}$$

$$b_2 = \frac{\sum_{i=1}^{m_2} \frac{y_{t-12,i}}{y_{(t-12)-1,i}}}{m_2}$$

   where m_1, m_2 are the number of matched pair ratios remaining after trimming.

4. The two imputation links b_1 and b_2 are weighted and combined together.

$$b = w\ b_1 + (1-w)\ b_2$$

   where w is the weighting value (e.g. default value of w= 0.8).

5. For each nonrespondent unit k, calculate the imputed value for the current period $\hat{y}_{tk}$ as

   $$\hat{y}_{tk} = b\ \mathrm{y}_{t\text{-}1,k}$$

   where $y_{t-1,k}$ is either the observed or imputed value for the previous period.

The Ratio Imputation is used for those nonrespondents which have been selected for the survey the first time and have no response value from the previous year. Auxiliary data are required for each sample unit. The steps are as follows:

1. For each responding unit, i, in the current period, calculate the ratio $\frac{y_{ti}}{x_{ti}}$.

2. Trim the largest 10% of the ratios and the smallest 10% of the ratios.

3. Calculate the trimmed mean of the ratios, b_3

   $$b_3 = \frac{\sum_{i=1}^{m_3} \frac{y_{ti}}{x_{ti}}}{m_3}$$

   where m_3 is the number of responding units after trimming.

4. For each nonrespondent unit k, calculate the imputed value $\hat{y}_{tk}$ for the current period as

   $$\hat{y}_{tk} = b_3\ x_k$$

   where x_k is the associated auxiliary data of the k^{th} nonrespondent unit.

- Australian Bureau of Statistics. The Australian Bureau of Statistics (ABS) categorized the nonresponse in business surveys as partial nonresponse, complete nonresponse and refusal. In all cases, imputation is used to compensate for the nonresponse. ABS uses three imputation methods depending on availability of previous data. These methods are called Beta Imputation, Live Respondent Mean and Hot-deck method (Australian Bureau of Statistics, 2001).

If previous data are available, Beta Imputation is used. The Beta Imputation method estimates the missing value by applying growth rate to the most recent reported data provided the data have been reported in either of the two previous occasions. The Live Respondent Mean method is the same as the mean value imputation previously described. If no information is available, the Hot-deck method is used. For the Hot-deck method, the nonresponding units is assigned the values from the responding units which have similar characteristics.

For the completely enumerated strata or take-all strata, unit nonresponse and refusal are handled by imputing all data items preferably by previously provided data. For the take-some strata, the unit nonresponse and refusal are replaced by values obtained from Live Respondent Mean method, but if previous data are available, imputation based on previous data is preferred.

Chapter 7

Quality Issues

The final output of a series of survey operations is called survey statistics. These are usually point estimates of finite population parameters. Users incorporate these statistics in their decision making or in their research. For this reason, it is important for the collecting agencies to provide the users with the description of the essential features of the survey, the statistical methodologies and the quality of the outputs being released (Sarndal et al, 1992; Griffiths and Linacre, 1995). In this way, the users can evaluate the survey results and judge the data's fitness for their use. The need for survey data quality has been recognized by major statistical agencies in the US, Canada, Australia and European countries. These agencies have begun to issue their own policies of informing users about the different survey operations that led to the survey outputs including sources of errors and some measures of data quality (Sarndal et al, 1992).

This chapter will discuss first the issues in data editing in connection with data quality, the quality assurance of business surveys as a whole and the importance of pilot surveys to quality.

7.1 Data Editing

Data editing is the process of detecting and correcting missing, invalid and inconsistent values resulting from data collection or capture in a survey. Errors in survey data result in invalid estimates, complicate the data processing and analysis, and lower the credibility of the collecting agency (Statistics Canada, 1998b; Sarndal et al, 1992; Granquist, 1995).

Edit rules, or edits, refer to the checks used for identifying missing, invalid and inconsistent values by the use of computer. Edits are classified in many ways by various people. Hidiroglou and Berthelot (1986) classified edits into 1) consistency edits and 2) statistical edits. Consistency edits are checks satisfying specific requirements in the form of inequalities or logical relationships. Consistency edits may verify well-defined variable values such as, values 0 or 1 for a dichotomous variable or ranges of values such as values less than, or greater than, some pre-determined values. In some cases, consistency edits follow the "if-then" type of condition.

Statistical edits on the other hand are checks for outliers obtained by applying statistical tests or procedures. Outliers are extreme values verified as being correct but are very high or very low compared from the values reported by similar units in a stratum (Sarndal et al, 1992). Outliers are expected to occur in very rare cases in the population. But if one or more outliers occur, there are three approaches that could be used at the estimation stage, namely 1) the values of outliers are changed by the Winsorization method, 2) the sampling weights of outliers are reduced, or 3) a robust estimation technique (Lee, 1995). The Winsorization method will be discussed in section 7.1.4.

Van de Pol (1998) distinguished the checks and edits as follows: routing checks, data validation checks and relational checks. Routing checks inspect all questions to determine if those questions which should be answered have been answered. Data validation checks inspect if the data values are permissible and within the valid ranges. In business surveys, the range of values are expected to be wide because of varying sizes of business. Relational checks are either ratio edits or arithmetic checks. Ratio edits use the ratio of two variables which should be within the specified limits. Arithmetic checks specify that the sum of sets of variables should be equal to the total.

Granquist (1995) mentioned another classification of edits as fatal edit, in contrast to query or suspicious edits. Fatal edits identify data that are clearly wrong while query edits identify data which have a high chance of being wrong.

7.1.1 Editing process

The traditional data collection process includes the recording of responses, encoding the responses into a computer, editing, following-up for missing or questionable responses and finally correcting (Sarndal et al, 1992). The editing process is done using computer software with edits specified by the subject-matter analyst. Error messages are reviewed manually and the actions that follow include checking of questionnaire and/or re-contacting of respondent for correction. The advent of computer assisted data entry technology simplifies the process by entering the responses directly into a computer at the time of the interview with the edit checks taking place at the same time. This technique is called Computer Assisted Telephone Interview (CATI). Any inconsistent, erroneous or illogical responses are shown on the computer screen and the interviewer can go back, ask the respondent and resolve the problems immediately. Another technique is the Computer Assisted Personal Interview (CAPI) where the interviewer visits the respondent and uses a lap-top computer to capture the responses. Then at the end of the interview, the data are transferred to a computer at the central office. For both techniques, any further editing is done later. As a result, the time taken for the editing process is shortened.

In business surveys, the mode of data collection is usually by mail because the data collected are mostly quantitative data from accounting records. Follow-up of missing and questionable responses is done through letters, telephone contact or personal visits. The strategies of follow-up vary from one collecting agency to another. In the Australian Bureau of Statistics for instance, business survey follow-up is prioritized for units that contribute significantly to the estimates, for newly selected units and for units that did not respond in the previous survey cycle (Australian Bureau of Statistics, 2001). If follow-up is not possible, the values are simply corrected by either imputation or weighting procedures (Sarndal et al, 1992).

In business surveys, data editing is usually the most expensive activity in the whole survey process, accounting for as much as 40% of the total survey budget. This was reported in the studies undertaken by Statistics Canada in 1994 (Statistics Canada, 1998b). The same results were reported by Statistics Sweden in 1975 and in the United States in 1990 (Granquist, 1995). The high editing costs gives rise to discussion on how to rationally allocate the resources for data processing operations. Collecting agencies have begun to review and evaluate their editing systems and methods to come up with a cost-benefit efficient editing process. Collecting agencies such as Statistics Canada, UK Office of National Statistics, Statistics Netherlands and other agencies recognized the role of editing in providing information about the quality of the survey data and serve as the basis for future improvements in a survey process (Statistics Canada, 1998b; Underwood, 2001; Van de Pol, 1998).

7.1.2 Editing Issues

As an expensive and tedious operation, data editing presents several issues. These are as follows:

- Errors in a data file neither can be nor will be traced completely even with exhaustive data editing. In one way or another there will be many errors which will remain undetected. This should not be much of a concern as long as the important and influential errors are identified and detected (Van de Pol, 1998). Editing in a selective manner, with priority given according to types or severity of errors or according to importance of variables, is found to be efficient and does not have a detrimental effect on data quality. A selective editing technique is being implemented by the Office of National Statistics UK in one of its monthly business survey as a pilot study and as a result the editing effort has been reduced by approximately 35% (Underwood, 2001).

- With available automated systems, the scope and volume of edits and checks can be easily increased. However, with respect to query edits there is a danger of overusing it. The time spent and resources used for editing may not be commensurate with the improvements in data quality. More often, too many data corrections result in a large amount of data changes which can introduce more errors than corrections at the end of the process. Over-editing can also introduce bias into the data (Underwood, 2001). Caution should be undertaken regarding excessive editing.

- There are some errors which may not be identified by edits. Examples are systematic and small errors occurring consistently in periodic surveys despite the strict edits imposed on the data (Statistics Canada, 1998). These errors happen when the survey definitions and that of the business accounting system do not match. In such situations the sensitivity of the edits needs to be improved .

- Follow-ups are added burdens to the respondents and costly to the collecting agencies. The issue of nonresponse has a close link to respondent burden. A reduction in respondent burden will have a positive effect on the response rates (Willeboordse, 1998c). A smart sampling design (e.g. coordinating of samples across surveys), a well-planned questionnaire and trained interviewers at the early stage of the survey are powerful tools to reduce respondent burden and nonresponse (Pierzchala, 1995; Granquist, 1995). In the follow-up of missing and/or questionable data, a huge amount of work is actually wasted since many follow-ups do not lead to a change in the data (Underwood, 2001). By performing selective sampling and follow-up, costs, as well as the respondent burden, are reduced (Underwood, 2001; Pierzchala, 1995; Statistics Canada, 1998b).

- Data editing does not guarantee a corresponding increase in data quality (Statistics Canada, 1998b; Underwood, 2001; Granquist, 1995). The scope of editing should be reduced and efforts should be redirected to error prevention rather than error correction. Granquist (1995) advocated the application of a total quality management approach to data editing.

7.1.3 Editing Systems and Software

At present there are three types of editing systems that have been developed according to functionality namely, 1) interactive integration, 2) Fellegi and Holt system , and 3) top-down (macro- or statistical) editing systems.

- The interactive integration system combines data editing with data entry and interview (Pierzchala, 1995). It also performs model-based imputation, weighting and tabulation tasks. Examples of interactive integration systems are the BLAISE from Statistics Netherlands (Hundepool, 1993) and DC2 from Statistics Canada (Statistics Canada, 1998b).

- The Fellegi and Holt system performs pre-survey rule analysis, interactive edit, automated item deletion and replacement , model-based imputation and donor imputation (Granquist, 1995). Examples of Fellegi and Holt systems are the GEIS from Statistics Canada and SPEER from the US Bureau of the Census.

- A top-down editing system performs interactive edit, graphical or tabular outlier inspection and computes the effect of editing on estimates (Pierzchala, 1995). Examples of top-down editing system are the ARIES from the US Bureau of Labor Statistics (Esposito et al, 1993) and the *gred* system from Statistics New Zealand (Houston and Bruce, 1993).

The automated editing systems do not replace human editing but are tools to make the tasks efficient, quick, consistent and satisfactory. People are still better at recognizing problems that the computer has not been programmed for and in applying subject-matter knowledge to solve the problems (Pierschala, 1995). An effective editing system comprises both human and computer editing which builds data quality.

7.1.4 Editing Procedures by Collecting Agencies

- Statistics Canada. In Statistics Canada's Monthly Survey of Manufacturing, the sampled establishment is contacted by mail or telephone depending on the respondent's preference and then the data capture and preliminary editing are done simultaneously. Businesses whose reports contain errors are followed-up immediately for verification and correction. All the data capture, preliminary edit and follow-up of non-respondents are undertaken at the Statistics Canada regional offices. Statistical edits are performed separately by industry and province classification. Outliers are identified by the magnitude of the deviation from the average response. Follow-ups are done to verify the outliers. Outliers for continuous variables are handled by applying some objective procedures such as outlier-resistant estimators or weight reduction techniques (Statistics Canada, 1998b; Chambers, 1986; Hidiroglou and Srinath, 1981; Lee, 1995). Nonresponse are handled by imputation. Then final verification of the responses after imputation is performed to ensure validity of data.

In Statistics Canada's Annual Survey of Manufactures, the selective editing strategy is being used. Business firms with high impact or deemed influential are dealt with manually while firms with lesser impact are dealt with by the GEIS imputation system (Pierzchala, 1995).

- US Bureau of Labor Statistics (BLS). Editing is done by the State agencies each month regardless of method of data collection. The data are examined to determine if consistent with the data reported in earlier months by the establishment and by the other establishments in the industry. Then the State agencies transfer the data electronically to BLS where further automated editing is done. Any questionable values detected at the editing process are then relayed to the person responsible for the initial collection of data. The person contacts the respondent for clarification and correction.

- UK Office of National Statistics (ONS). The UK Office of National Statistics had a recent review of their current data validation and editing processes and found that these methods are inefficient and antiquated (Underwood, 2001).

A new method called selective editing technique has been used in a pilot study involving one of its monthly business surveys. The survey was the Monthly Inquiry into the Distribution and Services Sectors (MIDSS). The editing resource focused on editing suspect values which were thought to significantly affect the survey estimates without impacting adversely on quality standards. Underwood (2001) described the selective editing method in two steps. The first step involves passing the data through the validation system. The second step involves those data values that fail at the validation system stage where a score is computed as follows:

$$\text{Score} = (\,|\,\text{current value} - \text{previous value}\,|\,)\ \text{x}\ \ (\text{survey weight}).$$

The scores are ordered and then editing is done on those units having scores higher than a pre-determined point or threshold. The results of the selective editing technique were verified with the results of the current editing system. Data quality is deemed to have been maintained if the following relationship holds:

$$\frac{(\text{current system estimate - selective editing estimate})}{\text{standard error}} < 0.1$$

ONS found that for the month of April 2001, the results were consistent in both the current system and the selective editing, with the editing effort reduced by about 35%. The selective editing pilot study was continued in the month of May survey for confirmation. The plan to implement the new system for MIDSS survey was set and ONS has continued undertaking pilot studies on the use of this technique for other business surveys.

- Australian Bureau of Statistics. The Australian Bureau of Statistics used editing in ABS business surveys to correct non-sampling errors (Australian Bureau of Statistics, 2001). Examples of these errors are respondent's misconception of the questions and instruction, interviewer bias, miscoding, unavailable data, incorrect transcriptions, nonresponse and noncontact. The editing process makes use of the respondent's value at the present survey and also on past survey results. In order to reduce the editing load for the survey, only those values which will significantly affect the survey estimates are edited by using imputation and outlier techniques. In handling outliers, ABS uses two methods referred to as: 1) Surprise outlier approach and 2) Winsorizing technique. The first method has been gradually replaced by the second method in most business surveys.

The Surprise outlier approach treats each outlier as if it were the only extreme unit in a stratum. Then the outlier is assigned with weight of one as if it had been selected in a completely enumerated (CE) stratum (or take-all stratum). The weight for units in the outlier's selection stratum (or take-some stratum) is recalculated since the population size and sample size have been changed by the outlier's movement to the CE stratum. The other population units which would have been represented by the outlier are now represented by the average of the other units in the stratum.

The Winsorizing technique uses a pre-determined cutoff value to identify an outlier (Lee, 1995). If a value is greater than the cutoff value, then it is considered an outlier and the value is replaced by the cutoff value plus a small additional amount. The small additional amount is computed by multiplying the difference between the sample value and cutoff by the stratum sampling fraction.

7.1.5 Recommendations

- Data quality should start prior to data collection, during the sampling design and questionnaire development stages, and not in the editing phase of the data collection process.

- Editing efforts should focus on those data problems that significantly affect the survey estimates in order to reduce costs, respondent burden and improve timeliness.

- Evaluation studies on the quality measures of the survey data should be undertaken not only in the current survey but also in other business surveys, in order to provide feedback and serve as basis for future improvements in data processing and other survey processes.

7.2 Survey Quality

Griffiths and Linacre (1995) stated that, “a business survey has quality outputs when its outputs are timely, easy to interpret, relevant to the issues of concern, of known and acceptable reliability and a good value for money spent” . The issue of quality in the survey design philosophies arises within the framework of Total Survey Design, TSD and Total Quality Management, TQM (Depoutot et al, 1998; Statistics Canada, 1998b). For statistical agencies, quality may be viewed in relation to the three elements that comprise TQM namely: 1) knowing and understanding the user’s needs, 2) involving staff in decision-making to meet the needs and 3) reviewing the survey processes continuously for potential redesigning (Statistics Canada, 1998b).

There is no standard definition of quality for data or statistical information but several criteria have evolved to characterize quality as follows: relevance, accuracy, timeliness, accessibility, interpretability and coherence.

- Relevance. Relevance of data or statistical information refers to the value contributed by these data in meeting the user’s needs or the mandate for which they are produced (Depoutot et al, 1998; Statistics Canada, 1998b; Wright, 1983).

- Accuracy. Accuracy of data or of statistical information refers to the closeness of the estimated values derived from the data to the true unknown population values (Cochran, 1977; Yamane,1967). The accuracy of an estimate involves analysis of total error associated with the estimate. In a sample survey, total error is divided into sampling and non-sampling errors. Non-sampling errors include errors of coverage, data response, nonresponse, processing (e.g. coding, data entry, verification, editing, weighting, tabulation) and dissemination (Lessler and Kalsbeek, 1992; Sarndal et al, 1992; Depoutot et al, 1998; Statistics Canada, 1998b).

- Timeliness. Timeliness of information refers to the availability and punctuality of dissemination of the results at a pre-established time. Usually, there is trade-off between timeliness and accuracy in such a way that the statistical information is produced and disseminated quickly but with less reliability, and the other scenario, which is slower but provides more accurate information (Depoutot et al, 1998; Statistics Canada, 1998b).

- Accessibility. Accessibility refers to the availability of information from the providers to the users taking into consideration the suitability and appropriateness of the forms, the media of dissemination and how the information will be made known to users, the ease or convenience of physical access and the affordability of gaining that access (Depoutot et al, 1998; Statistics Canada, 1998b).

- Interpretability. Interpretability of data refers to the ease with which the user is able to understand and use the information because the concepts and methods are adequately and clearly supplied (Statistics Canada, 1998b) .

- Coherence. Coherence of data refers to the degree to which the data originating from a single statistical program and the data brought together across data sets or statistical programs are logically connected and at least not contradictory of each other (Statistics Canada, 1998b). Information from different sources are coherent if they are based on similar definitions, classifications and methodological standards (Depoutot et al, 1998).

The best possible survey would be one that satisfies all these elements of quality. But in practice , these requirements are never met effectively because of errors (controllable and uncontrollable) arising in each stage of the survey process. It is even difficult to quantify quality because of the overlapping and interrelated relationship of these elements with one another. In fact, there has never been any statistical model which has aggregated these elements into a single indicator (Statistics Canada, 1998b). In order to attain an acceptable level of quality, the collecting agencies need to address, manage and balance these elements that constitute quality. Subject-matter knowledge, previous experience, reviews, feedback and consultations are the basis normally used in achieving this balance. All production aspects of the survey from the organization, management, staff, statistical instruments, computer facilities and training process should be efficiently and effectively used (Depoutot et al, 1998).

Each stage in the survey process makes a contribution to the overall quality and errors committed at the start of the process have an effect on subsequent stages.

There should be some measure of performance that could be used to evaluate the quality of each survey process. Gathering of additional information may be necessary through inclusion of special items in the questionnaire; maintenance of different versions of databases corresponding to preliminary data, imputed data and edited data; and documentation of all activities in detail, including the time spent and costs incurred in each stage of the process. The survey manager must be informed of such a measure of quality at every stage of the process so that any adjustment in the process can be done on time (Depoutot et al, 1998; Griffiths and Linacre, 1995).

Two methodological approaches to quality control can be applied in survey operations namely, preventive control and process control (Wright 1983). The survey operations that need to be better controlled in order to improve the quality of the survey (Sarndal et al, 1992; Lessler and Kalsbeek, 1992) are the following:

- Sampling process (includes frames and sample selection)
- Measurement and Field work (includes data collection procedures, questionnaire design, interviewer training)
- Data processing (includes data input , cleaning, editing and imputation)
- Estimation production (includes weighting and estimation procedure)
- Analysis and publications.

Leading national statistical agencies have established their own policy standards or guidelines for quality assurance. The guidelines must include the relevant information users need to know or be aware of such as the concepts and methods underlying the data being published, indicators of data quality or measures of error on each survey operation. In the 1982 Conference of European Statisticians, attended by the national statistical agencies in Canada, UK and other European countries, an agreement was reached as to the minimum information the quality report should contain. Sarndal, Swensson and Wretman (1992) listed the items as follows:

- Basic information on the data source, definitions and classifications
- Data coverage or frame
- Selection and estimation procedures
- Response rates and definition
- Sampling error; definition and formula
- Size of major errors and the relative importance and impact on the statistics
- Information on the significant changes in the procedure that would affect comparability of statistics over time
- Information on the comparability with statistics on the same subject from other sources
- References to the availability of more detailed technical reports.

In 1986, Statistics Canada issued its own policy statement on providing users with some indicators of the quality of data it disseminates and descriptions of concepts, definitions and statistical methodologies (Statistics Canada, 1998b). The data quality descriptions include the coverage, sampling error, response rates, comparability over time, benchmarking and revisions, comparability with other data sources, total variance and its components, nonresponse bias, edit and imputation effect, seasonal adjustment and any other error sources. The adoption of this policy intensifies Statistics Canada's pursuit for high-quality standards in every survey operations and with efforts directed for the systematic control of all operations.

7.3 Pilot Survey

A survey, no matter how well planned, should not be taken immediately to the full field work activity level but a preliminary survey or pilot survey should be undertaken. This is to provide a review of operational issues and cost aspects of data collection and tabulation. The information that is obtained from the pilot survey will be helpful in the conduct of the main survey.

Pilot surveys are useful for :

- assessing issues concerning the sample design and survey frame,
- providing preliminary information on the variability of the characteristics to be studied,

- testing the questionnaires and the related instructions,
- developing suitable procedures for field and tabulation work,
- training field staff,
- testing the schedules drafted for the survey, and
- examining issues of data quality and editing.

Operational difficulties that arise during the pilot survey will cause survey designers to plan, develop and suggest alternative courses of action for implementation in the different phases of the main survey. These efforts will contribute to the efficiency and quality of the survey in terms of costs, time and error control.

Chapter 8

Summary and Recommendations

The distinctive features of business surveys create many issues regarding the implementation of the various survey processes from the frame development, sample design, data collection and processing to estimation and reporting. These issues are gathered from the experiences of the different collecting agencies in the world including Canada, US, Australia, United Kingdom, the Netherlands, Sweden and New Zealand. A knowledge of these issues will facilitate the development of an omnibus business survey which the Manitoba Bureau of Statistics plans to undertake. Good practices can then be carried out while bad practices can be avoided.

8.1 Summary of Issues

The first issue in business surveys relates to the definition of the sampling units because the way by which the units are defined depends upon the data being collected and the organizational structure of units (e.g. geographical locations, operation, legal and administrative structures). These structures can be complex and the units may well be erratic over time due to splits, mergers and growth. Moreover, the units in business surveys tend to have a highly skewed distribution in terms of size; may be involved in single or multiple activities to produce goods and services; may operate in single physical location or multi-locations; and have undergone rapid changes in time due to births and deaths. These characteristics create difficult problems at all steps of the survey process.

Issues involved in the creation of business survey frames are: the choice of type of frames to use (e.g. list frame or area frame or both); the need for standard statistical units; standard industrial and geographical classification systems and rules for handling changes; the advantage and use of business registers derived from administrative sources; and the demand for quality in terms of accuracy of information (e.g. coverage, classification, contact, linkage etc), timeliness and cost.

The skewed distribution of units in the business population, the rapid changes in unit membership and the availability of auxiliary information on the units affect the choice of sampling design; the selection and allocation of samples; and the estimation procedure. Stratified sampling designs are found to be suitable in business surveys. The sample allocation for each of the strata is generally disproportionate over the size classes such that the stratum with large business sizes include all units with certainty while in the remaining strata, random samples are taken using the appropriate sample allocation and selection method. The choice of the sampling design and the estimation procedure is governed by the required level of precision for the survey estimates, available budget and available auxiliary information. The examples of sampling design used in take-some strata are the SRSWOR, Sequential SRSWOR, Poisson Sampling, and PPS Sampling without Replacement. In addition, the choice of sampling design takes into account the issues of coordination of samples for repetitive surveys or across programs of surveys, respondent burden control and data integration.

The estimation procedures in business surveys are handled by either a design-based inference method or a model-based method approach. The sampling design and all its aspects have to be reflected in the estimation process. The availability of auxiliary data is another distinct feature of business surveys. The auxiliary data are used in ratio, regression, post-stratification and calibration estimation to improve the precision of the estimates. Weighting as a necessary component of estimation needs to be done carefully and correctly. The measure of precision of the resulting estimators expressed as variance, mean square estimator or coefficient of variation (%) must reflect both the sampling design and the use of auxiliary data. Domain estimation is common in business surveys. The problem of small sample size resulting in high variance and biased estimates is handled by a model-based approach.

The rotation group sampling and the sampling rotation with PRN are the common methods used to select samples for business surveys repeated across time. Issues involved in repeated business surveys relate to the keeping of constant sampling fractions for the sampled units; to updating of the sampling frame due to births, deaths and change in the classification variables; and early reporting bias.

Nonresponse poses a problem in business surveys because of the highly skewed and quantitative nature of business data. The actions needed to cope with nonresponse are influenced by budget, time, use of data and risk of bias. The survey design, the questionnaire design, the data collection procedure and operations are some of the factors that can influence response rates. Nonresponse is handled by a weighting adjustment approach or by imputation at the estimation stage. The issues related to nonresponse are the need for a standard definition of the response and non-response rate and the redirection of efforts to prevent its occurrence before and during the data collection stage.

Imputation as a method of compensating for nonresponse can be abused, misused or overused because of the availability of automatic imputation systems. The issues related to imputation are its use on a large-scale basis for convenience in operation, treating the data set with imputed values as observed data, the non-verifiability of the imputed value and non-availability of quality measures to indicate its success or failure. However, the method of imputation when done properly is a useful tool in preserving known relationships between variables, in addressing systematic biases and in reducing nonresponse bias.

Data editing as a method for correcting erroneous, inconsistent and missing values does not guarantee a corresponding increase in data quality. Issues regarding data editing are related to the overuse of the edits which may cause an increase in costs, respondent burden and time; the failure of the editing process to detect all kinds of errors and the need for quality measures to evaluate the editing process. The scope of editing should be reduced and efforts should be focused on error prevention rather than error correction.

Lastly, the quality of the survey outputs can be ensured if they are relevant, accurate, reliable, timely, clear, easy to interpret and produced at a reasonable cost. A good working knowledge of the user's needs and the impact of the different survey processes on the quality of the data are necessary in order to attain quality in business surveys. Errors and their sources need to be identified and controlled. There should be some measures of performance to evaluate the quality of each survey process. The survey processes should be continuously reviewed for quality improvement.

8.2 Recommendations

The specific recommendations for the development of the Omnibus Survey of Manitoba Businesses, assuming that the survey is relatively modest in terms of complexity, are the following:

- Survey Frame. A Business Register containing Manitoba businesses derived from Statistics Canada Business Register will be used as the survey frame, to take advantage of the data sharing arrangement of Manitoba Bureau of Statistics with Statistics Canada.

- Sampling Unit. Establishment will be used as the smallest level of statistical unit.

- Sampling Design. A Stratified Sampling design is recommended with large businesses assigned in a separate stratum and all units in that stratum taken with certainty. In the remaining strata, random samples are taken using a SRSWOR design and an appropriate sample allocation (either Proportional, Neyman, Optimal, X-Optimal) for each stratum.

  The stratification variables to be used may include industry classification, geographical classification and employment size category.

The industry classification codes are based on the five level hierarchy structure of 1997 NAICS.

The geographical classification code is composed of Manitoba Economic Regional code, Census Division code and Census Subdivision code.

The employment size code includes an "Indeterminate" category and size ranges of 1-4, 5-9, 10-19, 20-49, 50-99, and 100 & over.

- Estimation Procedures. Either the design-based inference method or model-assisted method can be used. The use of auxiliary variables and post-stratification technique to further improve the efficiency of estimates is recommended. A weighting technique will be used to simplify calculation of estimates.

- Repeated Surveys. The Omnibus Business Survey may be repeated quarterly, bi-annually or annually depending on available budget and survey objective. In this case, the Rotation group sampling method can be used when retaining some sampling units for the next sampling occasion. A composite estimator would then be used to estimate the population mean or total of the most recent occasion.

- Nonresponse Problem. Preventive measures should be undertaken at the planning and during data collection stages. At the estimation stage, unit nonresponse is handled by weighting adjustment and item nonresponse is handled by imputation techniques such as Historical or Cold-deck imputation, or Ratio/Regression imputation.

- Data Editing. Selective editing techniques are recommended with editing priority according to severity of error and importance of variables.

- Pilot Survey. A pilot survey must be undertaken to assess and review the operational issues and cost aspects of data collection and tabulation before proceeding with the full field work activity level.

- Survey Quality. Issues concerning quality management should be addressed in order to minimize or control errors in all stages of the survey process.

Bibliography

Archer, D. (1995). "Maintenance of Business Registers". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 85-100.

Australian Bureau of Statistics (2001). Labour Statistics: Concepts, Sources and Methods. Catalogue 6102. (http://www.abs.gov.au/austats/).

Barfoot, M.T. (1993). "Automatic Generation of Standardized Statistical Structure in the Statistics Canada Business Register". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, NY, pp. 910-915.

Boegh-Nielsen, P. (1998). "Use of Administrative Registers". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands (http://www.forum.europa.eu.int/irc/dsis/bmethods/info/data/new/surveyhbk/).

Brewer, K. (2002). *Combined Survey Sampling Inference.* Oxford University Press Inc., NY.

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities.* Lecture Notes in Statistics 15, Springer Verlag, NY.

Brewer, K.R.W.; Early, L.J. and Hanif, M. (1984). "Poisson, Modified Poisson and Collocated Sampling". *Journal of Statistical Planning and Inference* 10, pp. 15-30.

Bureau of Labor Statistics (1997). "Update on the BLS Sample Design for the Current Employment Statistics Survey". Business Economics, Washington DC. (http://proquest.umi.com/).

Bush, J. and House, C. (1993). "The Area Frame: A Sampling Base for Establishment Surveys". *Proceedings of the International Conference on Establishment Surveys,* Buffalo, NY, pp. 335-344.

Castonguay, E. and Monty, A. (2000). "Recent Developments In Statistics Canada Business Register". *Proceedings of the International Conference on Establishment Surveys II.* (http://www.eia.doe.gov/ices2/).

Chambers, R.L. (1986). "Outlier Robust Finite Population Estimation". *Journal of American Statistical Association* 81, pp. 1063-1069.

Chapman, D.; Bailey, L.; and Kasprzyk, D. (1986). "Non-response Adjustment Procedures at the US Bureau of the Census". *Survey Methodology* 12, No 2, pp. 161-180.

Chapman, D.W. (1993). "Cluster Sampling for Personal Visit Establishment Surveys". *Proceedings of the International Conference on Establishment Surveys,* Buffalo, NY, pp. 645-650.

Chaudhuri, A. and Stenger, H. (1992). *Survey Sampling Theory and Methods.* Marcel Dekker Inc., NY.

Cochran, W.G. (1977). *Sampling Techniques* 3rd ed. John Wiley and Sons Inc.

Colledge, M.J. (1995). "Frame and Business Registers: An Overview". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 21-48.

Cox, B.G. and Chinnappa, B.N. (1995). "Unique Features of Business Surveys". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 1-20.

Dalenius, T. and Hodges, J.L. (1959). "Minimum Variance Stratification". *Journal of American Statistical Association* 54, pp. 88-101.

Depoutot, R.; Hurgen, R. and Ressen, R. (1998). "Quality". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands. (http://www.forum.europa.eu.int/irc/dsis/bmethods/info/data/new/surveyhbk/).

Deville, J.C. and Sarndal, C.E. (1992). "Calibration Estimators in Survey Sampling". *Journal of American Statistical Association* 87, No 418, pp. 376-382.

Esposito, R.; Lin, D. and Tidemann K. (1993). "The ARIES Review System in the BLS Current Employment Statistics Program". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, NY, pp. 843-847.

Full, S. (1999). "Estimating Variance due to Imputation in ONS Business Survey". *Proceedings of International Conference on Survey Non-response,* Portland, OR. (http://www.jpsm.umd.edu/icsn/papers/).

Granquist, L. (1995). "Improving the Traditional Editing Process". In *Business Survey Methods*, B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 385-401.

Griffiths, G. and Linacre, S. (1995). " Quality Assurance in Business Surveys". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 673-690.

Hanczaryk, P.S. and Mesenbourg, T.L. (1993). "Problems and Challenges Associated with Managing a Large Business Register". *Proceedings of the International Conference on Establishment Surveys,* Buffalo, NY, pp. 193-199.

Harala, R. (1998). "Statistical Properties and Quality of Register-based Census Statistics in Finland". *Proceedings Symposium 97 New Directions in Surveys and Censuses.* Catalogue 11-522-XPE, Statistics Canada.

Hidiroglou, M.A. (1986). "The Construction of a Self-Representing Stratum of Large Units in Survey Design". *The American Statistician* 40, No 1, pp. 27-31.

Hidiroglou, M.A. and Berthelot, J.M. (1986). "Statistical Editing and Imputation for Periodic Business Surveys". *Survey Methodology* 12, pp. 173-83.

Hidiroglou, M.A. and Sarndal, C.E. (1985). "An Empirical Study of Some Regression Estimators for Small Domains". *Survey Methodology* 11, pp. 65-77.

Hidiroglou, M.A. and Srinath, K.P. (1981). "Some Estimators of Population Total Containing Large Units". *Journal of American Statistical Association* 47, pp. 663-685.

Hidiroglou, M.A.; Choudhry, G.H.; and Lavallee, P. (1991). "A Sampling and Estimation Methodology for Sub-Annual Business Surveys". *Survey Methodology* 17, No 2, pp. 195-210.

Hidiroglou, M.A.; Drew, J.D.; and Gary, G.B. (1993). "A Framework for Measuring and Reducing Non-response in Surveys". *Survey Methodology* 19, No 1, pp. 81-94.

Hidiroglou, M.A.; Sarndal, C.E.; and Binder, D.A. (1995). "Weighting and Estimation in Business Surveys". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 477-502.

Hirschberg, D.A. and Nisselson, H. (1993). "An Efficient New Dual Frame System Design for Surveys of Small Business". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, NY, pp. 662-667.

Hostrup-Pedersen, S. (1993). "Statistics of Employment in Businesses – Denmark". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, NY, pp. 309-315.

Houston, G. and Bruce, A.G. (1993). "Gred: interactive Graphical Editing of Business Surveys". *Journal of Official Statistics* 9, pp. 81-90.

Hundepool, A. (1993). "Automation in Survey Processing". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, NY, pp. 167-172.

Johnson, A.E. (1995). "Business Survey as a Network Sample". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 219-233.

Kalton, G. and Kasprzyk, D. (1986). "The Treatment of Missing Survey Data". *Survey Methodology* 12, pp. 1-16.

Khazanie, R. (1996). *Statistics in a World of Applications.* 4th ed. Harper Collins College Publisher.

Koeijers, E. and Hilbink, K. (1998a). "Choosing Sampling Design and Estimation Method". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands. (http://www.forum.europa.eu.int /irc/dsis/bmethods/Info/ data/news/ sureyhbk/).

Koeijers, E. and Hilbink, K. (1998b). "Weighting and Reweighting". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands. (http://www.forum.europa.eu.int/irc/dsis/bmethods/info/data/new/surveyhbk/).

Kokic, P. and Jones, T. (1997). "Comparing Estimation Methods for a Monthly Business Inquiry". *Proceedings of Statistics Canada Symposium 97,New Directions in Surveys and Censuses,* pp. 269-277.

Kott, P.S. and Vogel, F.A. (1995). "Multiple Frame Business Survey". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 185-204.

Kovar, J.G. and Whitridge, P.J. (1995). "Imputation of Business Survey Data". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 403-424.

Lavallee, P. and Hidiroglou, M.A. (1988). "On the Stratification of Skewed Population". *Survey Methodology* 1, pp. 33-43.

Lee, H. (1995). "Outliers in Business Surveys". In *Business Survey Methods.* B.G. Cox et al (eds), John Wiley and Sons Inc., pp. 503-526.

Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Errors in Surveys.* John Wiley and Sons Inc.

Mac Donald, B. (1995). "Implementing a Standard Statistical Classification (SIC) System Revision". In *Business Survey Methods.* B.G.Cox et al (eds). John Wiley and Sons Inc., pp. 115-132.

Manitoba Bureau of Statistics (2002). "Manitoba Business Structure". Winnipeg, MB.

Murthy, M.N. (1977). *Sampling Theory and Methods.* Indian Press Private LTD, Calcutta, India.

Nijhowne, S. (1995). "Defining and Classifying Statistical Units". In *Business Survey Methods*. B.G. Cox et al (eds). John Wiley and Sons Inc., pp. 49-64.

Ogus, J.L. and Clark, D.F. (1971). "The Annual Survey of Manufactures: A Report on Methodology". Technical Paper No 24. US Bureau of Census. Washington DC.

Ohlsson, E. (1995). "Coordination of Samples Using Permanent Random Numbers". In *Business Survey Methods*. B.G. Cox et al (eds). John Wiley and Sons Inc., pp. 153-170.

Ohlsson, E. (1998). "Sequential Poisson Sampling". *Journal of Official Statistics* 14, No 2, pp. 149-162.

Perry, J. (1993). "The Development of a Business Survey Frame from Administrative Data". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, NY, pp. 186-192.

Petrucci, A. and Pratesi, M. (1993). "Listing Frames and Maps in Area Sampling Survey on Establishment and Firms". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, NY, pp. 554-559.

Pierzchala, M. (1995). "Editing Systems and Software". In *Business Survey Methods*. B.G. Cox et al (eds). John Wiley and Sons Inc., pp. 425-441.

Pietsch, L. (1995). "Profiling Large Businesses to Define Frame Units". In *Business Survey Methods*. B.G. Cox et al (eds). John Wiley and Sons Inc., pp. 101-114.

Raj, D. (1968). *Sampling Theory*. McGraw-Hill Inc.

Sarndal, C.E. and Hidiroglou, M.A. (1989). "Small Domain Estimation: A Conditional Analysis". *Journal of American Statistical Association* 84, No 450, pp. 266-275.

Sarndal, C.E.; Swensson, B.; and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics, Springer-Verlag New York Inc.

Scheaffer, R.L.; Mendelhall, W.; and Ott, L. (1990). *Elementary Survey Sampling* 4th ed, Duxbury Press, Belmont, CA.

Schiopu-Kratina, I. and Srinath, K.P. (1991). "Sample Rotation and Estimation in the Survey of Employment, Payroll and Hours". *Survey Methodology* 17, No 1, pp. 79-90.

Sigman, R.S. and Monsour, N.J. (1995). "Selecting Samples from List Frames of Registers". In *Business Survey Methods*. B.G. Cox et al (eds). John Wiley and Sons Inc., pp. 133-152.

Singh, D. and Chaudhary, F.S. (1986*). Theory and Analysis of Sample Survey Designs*. John Wiley and Sons Inc., NY.

Srinath, K.P. and Carpenter, R. (1995). "Sampling Methods for Repeated Business Surveys". In *Business Survey Methods*. B.G. Cox et al (eds). John Wiley and Sons Inc., pp. 171-184.

Statistics Canada (1977). "Concepts and Definitions of the Census of Manufactures". Catalogue 31-528, Ottawa, ON.

Statistics Canada (1998a). "Employment, Earnings and Hours". Catalogue 72-002-XPB, Ottawa, ON.

Statistics Canada (1998b). Quality Guidelines 3rd ed. Catalogue 12-539-XIE. Ottawa, ON. (http:www.statcan.ca/statisticalmethods/).

Statistics Canada (2001). Monthly Survey of Manufacturing (MSM). Reference No 2101. Ottawa, ON. (http:www.statcan.ca/statisticalmethods/).

Statistics Canada (2002a). Business Register. Statistical Data Documentation System Reference Number 1105. Ottawa, ON.

Statistics Canada (2002b). Provincial Business Register Extract – Record Layout. Internal Paper.

Statistics Canada (2002c). NAICS Canada 1997. (http://www.stancan.ca/english/subjects/standard /naics2002.html).

Statistics Canada (2002d). Guide to the Labour Force Survey. Catalogue No 71-543-GIE.

Teikari, I. (2000). "Evening Out the Response Burden". *Proceedings of the International Conference on Establishment Surveys II*, Buffalo, NY, pp. 609-618. (http:/webfarm.jrc.it/ETK-NTTS/Papers/final_papers/62pdf).

Underwood, C. (2001). "Implementing Selective Editing in a Monthly Business Survey". *Proceedings of Sixth Government Statistical Service Methodology Conference*. (http://www.statistics.gov.uk/methods_quality/gss_method_conf).

Valliant, R. (1993). "Post Stratification and Conditional Variance Estimation". *Journal of American Statistical Association* 88, No 421, pp. 89-96.

Van de Pol, F. (1998). "Data Editing". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands. (http://www.forum. europa. eu.int/irc/dsis/bmethods/info/data/ new/surveyhbk /).

Verboon, P. (1998). "Imputation". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands. (http://www.forum.europa.eu.int/irc/dsis/bmethods/ info/data/new/surveyhbk/).

Willboordse, A. (1998a). "The Framework of Business Statistics". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands. (http:// www.forum.europa.eu.int/irc/dsis/bmethods/info/data/new/surveyhbk/).

Willboordse, A. (1998b). "The Statistical Business Register". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands. (http://www.forum.ueropa.eu.int/irc/dsis/bmethods/info/data/new/surveyhbk/).

Willboordse, A. (1998c). "Minimizing Non-response". In Handbook on Design and Implementation of Business Surveys. CBS Netherlands. (http://www.forum.europa.eu.int/irc/dsis/bmethods/info/data/new/surveyhbk/).

Winkler, W.E. (1995). "Matching and Record Linkage". In *Business Survey Methods*. BG Cox et al (eds). John Wiley and Sons Inc, pp. 355-384.

Wright, T. (1983). *Statistical Methods and the Improvement of Data Quality*. Academic Press Inc., FL.

Yamane, T. (1967). *Elementary Sampling Theory*. Prentice-Hall Inc., NJ.