C-3

1000

SURVEY SAMPLING:

A NON-MATHEMATICAL GUIDE

By A. Satin and W. Shastry

AUUG 3888382

Statistics Canada

Federal Statistical Activities Secretariat and
Census and Household Survey Methods Division

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

April 1983
8-3200-502

12-X-504E

ISBN 0-660-11224-8

Ottawa

Version française de cette publication
disponible sur demande.

PREFACE

For some time now, Statistics Canada has been offering a series of how-to seminars on the subject of sample surveys. These seminars, which generally run two to three days, have been given in various cities across Canada and are based on this manual. The manual is a basic guide on survey sampling for those with little or no formal training in mathematical statistics. Although this is the first non-mathematical guide on sample design to be published by Statistics Canada, it is unique for other reasons as well.

The manual explains the fundamental concepts and methods of survey sampling by way of examples and illustrations rather than by theory and algebraic expressions. The mainstay of the textual illustrations is a survey on farm expenditures. In brief, the purpose of the survey is to determine what it costs to run a farm, particularly a farm that grows wheat and other grain-growing crops.

It should be pointed out that Statistics Canada actually does conduct such a survey. The Farm Expenditure Survey (FES) as it is called, was first conducted on an annual basis throughout the prairies and the Peace River District of British Columbia. Through the years, the FES has evolved into a multipurpose enumerative survey. It has proven to be invaluable as a means of obtaining annual estimates on selected crops, livestock and financial items.[1]

Many surveys could have been used as the central example for this manual. The FES was chosen simply because it incorporates many of the sample design techniques illustrated in this manual. But, while the farm survey used in the manual is modelled after the FES, it is by no means intended to be a replica of it.

The manual itself is developed and organized in a way which concentrates on the many aspects of survey sampling. Chapter 1 provides a brief look at the history of survey sampling within the context of Statistics Canada. Subsequently, it deals with the basic differences between a census and a sample survey. It also outlines the importance of survey objectives and addresses itself to the way in which these objectives are transformed into a set of data requirements.

Chapter 2 deals with some specific elements of sample design such as the population, the frame and survey units.

In Chapter 3, the rudiments of sampling and non-sampling error are discussed as a preliminary introduction to Chapters 4, 6 and 7.

There are two major areas of study in survey sampling: probability and non-probability sampling. Chapter 4 deals with the first area; namely probability sampling. The manual describes what these sampling methods are, the general circumstances in which they are used, as well as their advantages and disadvantages. The sampling techniques described cover simple random sampling, systematic sampling, stratified sampling, unequal probability sampling and area sampling. There is also a brief explanation of replicated and multiphase sampling followed by a summary of all the schemes described.

Chapter 5 deals with the second sampling method; namely non-probability sampling. Here, discussion ranges from the simplest sampling scheme (the man-on-the-street-interview) to the merits and demerits of quota sampling. This chapter also includes a look at the advantages and disadvantages of non-probability sampling.

Although the manual is primarily concerned with sampling, it also briefly addresses the problem of estimation in Chapter 6. This is followed by a look at sample size determination as well as at some special problems in sampling and estimation. The final chapter attempts to put sampling into its proper perspective in terms of the practical considerations in the conduct of a survey.

Finally, the manual concludes with an Appendix of algebraic expressions intended for those who might be interested in how sample estimates are computed and how their precision is assessed. Included here are examples of sample size determination and estimation based upon a survey of post-secondary institutions.

Where it has been considered appropriate or valuable, various references to the Appendix have been made throughout the text. In addition to the Appendix, there is a

1. The purpose of the FES is to provide information needed to administer the Western Grain Stabilization Act (WGSA). (This act was passed in 1976 by the Canadian parliament to stabilize cash flows of grain producers). At the same time FES data serves to improve estimates of net agriculture income for National Accounts purposes.

Bibliography which should prove useful for those wishing a more technical treatment of the subject.

Since sampling is a science rooted in mathematics, it is recognized that this guide cannot possibly be a substitute for the weightier texts that exist on the subject. It is however, sure to prove valuable as a basic guide on survey sampling. Yet, ultimately the success of any publication rests with its readership. We are confident that this manual will find a wide and enthusiastic audience.

Information Services
Statistics Canada
1982

ACKNOWLEDGEMENTS

The close association between the seminar and the manual makes it very difficult to assign the appropriate credits to all those who have contributed.

We would like to acknowledge those who have taught the course with us, in particular, D. Royce and H. Hofmann since their close involvement has been invaluable in modifying and improving not only the seminar but the manual as well. Several other friends and associates provided valuable assistance in reviewing the manual among whom special mention should be made of N. Chinnappa, H. Gough, F. Hardy, M. Hidiroglou, G. Mitchell, M. Nargundkar, R. Platek, G. Shane, M. Sheridan, L. Swain, T. Thompson and E. Wilhelm. We would also like to express our appreciation to Jonina Wood of the Information Division for her work on editing and refining the manual.

Finally, we would like to aknowledge the Federal Statistical Activities Secretariat of Statistics Canada which commissioned this work in response to the many requests for a non-mathematical guide on survey sampling.

Alvin Satin
Wilma Shastry

Census and Household Survey Methods
Division

TABLE OF CONTENTS

CHAPTER 1

AN INTRODUCTION TO SURVEY SAMPLING

Chapter 1

An Introduction to Survey Sampling

1.1 A Brief History of Sampling at Statistics Canada

It was only recently that survey sampling was officially recognized. Up until fifty years ago, even the theory was still in its infancy. Sampling had yet to evolve to the point where organizations and people would accept it as an efficient and inexpensive way to collect information.

The change came in 1943. At the time, Canada was still in the midst of World War II. Yet, at Statistics Canada, (then the Dominion Bureau of Statistics) the demand for information was unprecedented.

In fact, never before in the history of the country had the need for statistics been such a pressing one. The cost-of-living index now became a key figure. The creation of war departments meant a greater need for employment figures. Industry was expanding.

The post-war years were to bring no respite. If anything, they created additional needs. Since the government was assuming more responsibility for employment and social security, policies such as unemployment insurance, old age pensions and taxation had to be planned. At the same time, Canada's membership in the newly-created United Nations meant a greater need for international statistics, far beyond anything called for in the past.

By late 1943, the Bureau was ready to recognize that survey sampling was an essential scientific technique: a technique proven to be effective, quick and above all cheaper than the enumeration methods the Bureau had historically used. Official recognition came as the Dominion Statistician called for a complete reorganization of the Bureau. One of the most significant results of the new order was the establishment of a Sampling Organization.

The Sampling Organization heralded the beginning of numerous surveys in conjunction with the Bureau's Quarterly Survey of the Labour Force. By 1948, the Statistics Act had been amended "to authorize the collection of statistics by means of sampling".

In 1952, in a document prepared for a United Nations seminar held in Ottawa, the Bureau noted that:

> *(sampling) permits surveys to be made much more quickly and with a fraction of the staff required for complete enumeration, yet it can yield results well within the margin of error necessary for practical purposes; indeed when properly applied, this method is frequently capable of furnishing data of a higher quality than can be obtained by ordinary enumeration.*[1]

In the years since 1952, many changes have taken place. Statistics Canada now has the responsibility for producing current economic accounts of production and trade in all commodities, and indicators such as the unemployment rate and the consumer price index, all of which involve survey sampling.

Statistics Canada is, by federal law, the statistical agency for the entire nation. Under the Statistics Act, 1971, the mandate of the Bureau is to:

> *"collect, compile, analyze, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and condition of the people."*

As a result of this legal mandate, economic indicators are developed almost exclusively at Statistics Canada, based on numerous surveys which are taken on a monthly, quarterly and annual basis.

Today, sample surveys provide information on everything from the amount of sugar an average Canadian consumes in a year to the number of hours he or she might spend playing tennis or jogging. It provides information on unemployment figures, on import and export figures; the list is endless. When the Bureau noted, in 1952, that sampling could produce "data of a higher quality" than complete enumeration, it set the stage for a new national industry of statistics.

1.2 Defining a Survey

The following definition of a survey is one which serves the purpose of this manual and one that we feel is suitable.

Essentially, a *survey* involves the collection of information about characteristics of interest from some or all units of a population using well-defined concepts, methods and procedures, and the compilation of such information into a useful summary form.

1. Dominion Bureau of Statistics: *History, Function, Organization* (Ottawa: Queen's Printer, 1952). p.11.

Surveys are carried out for either one of two purposes: descriptive or analytical.

The main purpose of a *descriptive survey* is to estimate certain characteristics or attributes of a population. Examples of this might be the average income of farmers, or the number, proportion or percentage of unemployed persons in the labour force. A descriptive survey could undertake to determine how much money certain industries spend on research and development.

Analytical surveys are generally concerned with testing statistical hypotheses or exploring relationships among the characteristics of a population. Their main purpose is to explain rather than to describe. An example of an analytical survey would be one that determines whether the proportion of families who own their own homes increases or decreases following the introduction of a government housing program. An analytical survey could also determine whether there is any relationship between frequency of health-related conditions to area of residence, diet and age.

For the purposes of this manual, discussion will centre largely on surveys with a descriptive purpose, since they are, by far, the more common of the two. In fact, even for analytical purposes, descriptive analysis is often the first step.

1.3 Sample vs. Census Surveys

One of the first questions that many people ask when beginning the study of surveys is: "Wouldn't it be better to carry out a complete count rather than take a sample?"

The question is an excellent one and one that even statisticians must address. In fact, whenever a survey is considered to be the best way to gather information, statisticians are always faced with the alternatives of carrying out either a census or a sample survey.

But before that question can be answered, the difference between the two kinds of surveys must first be clarified.[2]

A *census survey*, then, refers to the collection of information about characteristics of interest *from all units in a population*.

A *sample survey* refers to the collection of information about characteristics of interest *from only a part of the population*.

So why take a sample rather than a census survey? The answer briefly, is that sampling permits the survey taker to reduce costs. At the same time, as the following subsections will show, survey takers do not have to hire as many people, they can get to the information faster and in many instances achieve even greater accuracy than they would have through a census survey.

Cost

In most cases the main justification for a sample over a census is to reduce cost. It is possible, with a relatively small sample, to obtain results that reasonably approximate the actual characteristics of a large population. Suppose, for example, that information was needed on all the people in Canada who are over 15 years of age. Rather than collect information from all 17 million Canadians, over the age of 15, a survey of a small percentage of them (possibly as few as 1,000 or 2,000 depending on information requirements), might provide sufficiently adequate results.

At the same time, one can readily see that the cost of obtaining information through a sample would be a lot less than obtaining it through a census.

Timeliness

One of the advantages of sampling is that it permits investigators to move quickly. It is often the case that survey results are required shortly after the need for the information has been identified. For example, if one wants to conduct a survey to measure public awareness of a media ad campaign, it is necessary to conduct a survey shortly after the campaign is undertaken.

Since sampling requires a smaller scale of operations, it reduces the data collection and processing time, while allowing for greater design time as well as more complex processing programs.

Accuracy

Since survey results are subject to error, sample surveys can often be more accurate than their census counterparts.[3] In case of face-to-face or telephone interview surveys, for example, higher levels of accuracy can be achieved through more selective recruiting of interviewers, more extensive training programs, a closer supervision of the personnel involved and a more efficient monitoring of the fieldwork.

The smaller scale of operations associated with a sample survey also allows for more extensive follow-ups of nonrespondents and for a higher level of quality control for such data processing activities as coding and data capture.[4]

Specialized Needs

Sometimes, conducting a sample survey is the only option open to the investigator. Consider for example, cases where information of a technical nature requires highly trained personnel and specialized equipment. It would be difficult and expensive to consider a census in such cases.

2. Whenever a survey is mentioned, the reference may apply to either a sample or a census survey.

3. The sources of errors in surveys and the concepts of accuracy and precision are explained in Chapter 3.

4. Quality control in the context of coding and data capture is explained in Chapter 9.

The Canada Health Survey is a prime example. This survey was developed by the Department of Health and Welfare and Statistics Canada to assess the general fitness of Canadians. It requires the participation of nurses and the use of medical equipment. As part of the process, the nurses, who in this case act as the interviewers, are sent to selected households to conduct a series of physical tests. These tests include measurements of height and weight. At the same time, blood samples are taken and people are requested to run up and down stairs to measure blood pressure and heart rate. An instrument called a Harpenden caliper is used to obtain information on body fat.

In such a survey, it is impossible to hire untrained personnel, arm them with calipers and syringes and charts, and expect them to bring back valid results. No less impossible would be to undertake a census. Even if there were enough funds available for such a costly enterprise, it is dubious that sufficient numbers of nurses could be summoned to the cause!

Reduced Respondent Burden

Respondent burden has become a serious issue in survey taking in view of increasing demands for timely and accurate information. With sample surveys, information is sought from only a part of the population, so fewer people are inconvenienced.

Yet, if such a strong case can be made for sampling, the question now becomes: "Why not always go for the sample and forget the complete count?" The answer, in brief, is that there are occasions where the nature of the information makes a census not only desirable, but essential.

Consider the case where a very small population is under study. Here it would be appropriate to use a census since no substantial savings in time or cost would be realized through sampling.

Also, a census can be a necessity where detailed information on the characteristics of a population is required. Such a case could arise where information is to be disaggregated or broken down over very small geographical areas or into very detailed classifications.

A census survey may also be necessary to provide benchmark information to efficiently design a sample survey. The sample design in the Labour Force Survey, for example, is based upon information from the Census of Population and Housing.

1.4 Survey Objectives

The first task in planning a survey is to specify the objectives as thoroughly as possible. It is not enough to indicate that the purpose is to provide information on, say, "housing conditions of the poor". Such a nebulous statement may well serve as a broad description of the survey, but ultimately it must be broken down into more specific language.

What, for example, is meant by the phrase "housing conditions"? Does it refer to the type of dwelling, its age and/or location? What precisely is meant by "poor"? Is poverty measured in terms of debts or salary or both?

The key to the exercise, at this stage of the survey, is to come up with clearly defined concepts and terms. Once the basic objectives have been broken down and defined, the researcher can then proceed to develop operational definitions.

Operational definitions indicate who or what is to be observed and what is to be measured. In the case of the "poor" the definition might include all families whose gross income is below a specified level. The terms family and income must then be defined. What is considered a family? What constitutes income? From what geographical region will the selected families come: region, province, city? Over what period is income to be measured? The answers to all these questions depend on the ultimate use to which the data is put.

Once operational definitions are developed, the researcher can specify the data requirements of the survey and decide upon a level of error that is acceptable in the survey results.

Finally, the statement of objectives should indicate the purpose of the survey, the areas to be covered, the kinds of results expected, the users as well as the uses of the data, and the level of accuracy which is desired.

Clearly, the original phrase "housing conditions of the poor" contains few of these points. It may well serve as the departure point, in a casually descriptive conversation, but it really has little place in the formal design of a sample survey.

CHAPTER 2

ELEMENTS OF A SAMPLE PLAN

Chapter 2

Elements of a Sample Plan

Once the survey objectives have been clarified and the data requirements have been established, the sample plan can be developed. The sample plan is an integral part of the overall survey plan. Basically, it consists of three elements: sample design, estimation procedures, and estimates of precision.

The sample design refers to what a sample consists of and how the sample is to be obtained. Estimation methods indicate how estimates of the population characteristics are to be constructed from the sample. Procedures to measure sampling error establish how the precision of these estimates is to be determined.[1]

The first component of the sample plan is the sample design. Measures used to estimate sampling error are presented in Chapter 3 and subsequently, in the Appendix. Estimation methods are explained in Chapter 6.

2.1 Sample Design

The sample design is really a set of specifications which describe target and survey population, the frame, the survey units, the size of the sample and the sample selection methods.

In this section, attention will focus on the first three of these elements. Sample size and estimation methods will be discussed following the explanations of the various kinds of sample selection methods.

2.1.1 Population

The *population* is the aggregate or collection of units to which the survey results apply. In this sense, it refers not only to people but can be a collection of households, schools, hospitals, farms, businesses, vehicles, etc.

Once the population has been determined, the units which compose it are described in terms of their age, size or any other features that clearly identify them. At this point, the geographic boundaries of the population must also be outlined. Details such as coverage by municipal, provincial, regional or national boundaries must be worked out. In addition, a time or reference period must be decided.

To illustrate these points, consider a survey which has, as its objective, the task of determining doctors' salaries. In this case, the units of the population would refer to the doctors. The spatial location of these "units" might be Toronto, or any major city, or even an entire province. The time period of interest might be for those doctors practising in 1980. Again, that time period could be extended or shortened depending on the purpose of the survey.

Yet, such an offhand example should in no way imply that this stage of the sample design is an easy one. In fact, defining a population is generally anything but straightforward. Often the target population (the population for which information is required) and the survey population (the population actually covered) differ for practical reasons, even though ideally, they should be the same. Sometimes, it may be necessary to impose geographic limitations to exclude certain parts of the target population which may be too difficult or costly to access (e.g. isolated areas in Northern Canada).

Where it is necessary to change geographic boundaries, the changes should be explicitly stated. It might also be the case that the survey concepts and/or methods used are inappropriate for certain parts of the population.

To illustrate this last point, consider a sample survey of Ontario post-secondary graduates. The objective is to determine if indeed the graduates have found jobs and if so just what kinds of jobs they have found. In this case, the survey population might exclude graduates of such specialized institutions as religious seminaries, military schools or business colleges, since graduates of these institutions would be reasonably assured of securing employment in their respective fields. It might include only those individuals who graduated from universities, community colleges and hospital schools of nursing during 1980.

Or, consider a survey of the capital expenditures of Canadian manufacturing firms. For the purpose of a sample survey, the population might be narrowed to include only those manufacturing firms with 25 or more employees, operating in this country in the year 1980.

Another survey proposition might be to see how Canadians are getting along with the metric system. In this case, the survey population might cover all civilians 15 years of age

1. Sampling error is the error attributed to the fact that only a part of the population is being surveyed. To assess this error, measures such as standard deviation, coefficient of variation and variance can be used. Explanations of these measures are outlined in Chapter 3. Their corresponding mathematical expressions are provided in the Appendices.

and over from the 10 provinces of Canada. It might exclude, however, the Northwest Territories and the Yukon, as well as those persons living on Indian Reserves and Crown lands, inmates of institutions and members of the armed forces. (It may seem as if the target population has now been considerably reduced, but, in fact, the exclusions account for about 2% of the population of Canada.)

In all these examples, there are definite gaps between the target and the survey populations. The researcher should be aware of any gaps that are created between the target and survey populations and understand that the conclusions must be limited to the survey population only.

2.1.2 Frame

Once the population has been defined, in particular, the target and the survey population, the next step is to establish a means of access to it.

The *frame* will provide this means of access. In its simplest form the frame is a list of elements covering the survey population. It can be in the form of a *physical list* such as a computer printout, magnetic tape, telephone book or set of cards. Or it may be a conceptual list of, for example, all those vehicles which enter a provincial park between the hours of 9:00 a.m. and 5:00 p.m. during the month of July. Such a frame is called a *list frame*.

An *area frame* can be considered as a special kind of *list frame* where the elements now correspond to a geographical area. In the Census of Population and Housing, the frame consists of areal units called Enumeration Areas (EAs). An EA is the geographical area canvassed by one interviewer in the Census of Population and Housing. These EAs are readily identifiable on maps and in the field. Collectively, they comprise an area frame.

Getting to the survey population may also be accomplished through a *hierarchy of frames*. In this case, the units which comprise a frame at one level of the hierarchy are potentially divisible into units which comprise a frame at the next level in the hierarchy. Consider a household survey, where a sample of geographical areas such as EAs might be selected from a list of all such areal units. Subsequently, a sample of dwellings could be selected from a complete list of all dwellings within each selected EA.[2]

To further illustrate a frame, consider again, the survey of post-secondary graduates. Here the frame might consist of university and college files which list all students who have graduated in a particular year.

Within Statistics Canada, the business register and the farm register are important frames for carrying out a wide range of business and agricultural surveys.

Yet, sometimes a single frame is not enough to adequately cover the survey population. In such a case, *multiple frames* may be used. These frames may cover different parts of the survey population or they may even overlap. If one were to undertake a survey of the salaries of all Canadians, an area frame could be used in conjunction with a list frame of, say, high income earners. In the case of overlapping frames statistical techniques have been recently developed to resolve overlap.[3]

However, for the time being, the discussion will centre on list frames and will return later to area frames. It should also be noted, at this time, that a frame is required for both census and sample surveys.

Finding an Adequate Sampling Frame

One of the first problems in sample design is finding a suitable frame for the survey population.

The frame plays a central role in the design of a survey. It determines how well a population is covered, affects the method of enumeration and influences the efficiency with which a sample can be designed.

Thus, in the case of a mail or telephone survey, it is essential that the units of the frame have accurate addresses or telephone numbers.

It is also desirable that the units of the frame contain auxiliary information so that an efficient sample plan can be developed. In the case of a business or agricultural survey, it might be desirable for the frame to contain information on the size (number of employees, acreage) and/or location of the units.[4]

In the case of business surveys requiring financial information, it is also essential to know where financial statements are kept, the units to which they apply, not to mention the location of the people who can provide access to such often classified information.

There are many cases where frames already exist. When it comes to specific populations such as hospitals with cancer

2. Hierarchy of frames is used in conjunction with multi-stage sampling. For a fuller explanation, see Chapter 4. Further, for simplification purposes, a hierarchy of frames will also be referred to as an area frame within the context of this manual.

3. These techniques are described in *H.O. Hartley, Multiple Frame Surveys* (Proc. Stat. Sect. Amer. Stat. Assoc., 1962) pp. 203-206, also *H.O. Hartley, Multiple Frame Methodology and Selected Application* (Sankhya C 361974) pp. 99-118.

4. The manner in which such information can be used to efficiently design a sample is explained in Chapter 4.

treatment facilities, business establishments engaged in foreign trade, persons receiving pension or job disability payments, administrative files are usually available. Often these can be adapted for use as sampling frames.

Yet, frames do not always lend themselves perfectly to the survey at hand. To illustrate the problems which can arise, consider the following example:

Suppose that the population of interest consists of the members of an association, such as the Association of Professional Engineers of Ontario (APEO).

According to APEO rules, each applicant must be a resident of Ontario and have a recognized degree in any discipline of engineering along with a minimun of two years experience. The registration must also be accompanied by an annual membership fee of $60.

Suppose now that the objective of the survey is to determine the average income earned by members of this association in a given year. In this case, the frame is, very simply, a list of the members.

However, there are a number of problems that may arise with respect to the suitability of such a list. In the first place, there is the issue of undercoverage, where not all elements that should be included in the population are on the list or frame. The frame may not contain the names of all eligible members. There may be newly qualified engineers who have not yet applied. Or there may be some who have applied, but who have not yet been registered due to delays in processing.

Secondly, there is the problem of overcoverage where elements that are listed on the frame are not *bona fide* members of the population. A list can fall out-of-date very quickly. Some members may have moved to a different province; some may have died; or there may be members still listed who are no longer practising engineers.

Finally, there is the problem of duplication. Duplication often arises when the frame is made up of a combination of lists which have overlapping memberships. It can also arise when several lists are merged. If the APEO survey was national in scope, the frame might have been established by merging the lists from all provinces. Again, the rules state that a member of the APEO can keep a non-resident membership even if he or she decides to move to another province. Since some members of the Association might be registered in more than one province, their names may appear several times on the combined list.

These are only some of the problems associated with frames. Depending on the circumstances, one might decide to (i) discard the list and use or create another or (ii) use the existing list and ignore its defects or (iii) adjust the list through updating or linking it with other files or (iv) use multiple frames.

Since the frame provides the means of accessing or getting to a population, its quality is of crucial importance. Potential frames should be carefully evaluated early in the planning stage to assess the extent of the defects and their potential impact on the results of the survey.

2.1.3 Survey Units

For the purpose of sample selection, the population should be divisible into a finite number of distinct, non-overlapping and identifiable units called *sampling units*, so that each member of the population belongs to only one sampling unit.

Naturally, the type of sampling unit depends on the nature of the study. A dwelling may be considered as the sampling unit in a family expenditure survey; a farm or plot in a crop survey; or a business establishment[5] in an employee payroll survey.

Sampling units may or may not correspond to the *units of analysis*. An example where the sampling unit may be different from the unit of analysis is the case of a household survey. The units selected may be dwellings, whereas the units of analysis would be people or families. In addition, two other kinds of units: *respondent units* and *units of reference* can also come into play.

The respondent unit is that unit which provides the information. The unit of reference is that unit about which information is obtained from the respondent. In many cases, these units are identical, but, on occasion, they can be different.

To illustrate the various units, consider the following example:

The objective is to determine the research and development activities of large Canadian businesses. In such a survey, the sampling unit could be the company. The unit of analysis might be those companies with research and development

5. In Statistics Canada a business establishment is broadly defined as the smallest unit for which a set of separate financial records are kept.

activities exceeding $50,000 in 1980. The respondent unit might be the Head of Research and Development within each company. Finally, the unit of reference could refer to the division(s) within each company engaged in research and development activities.

Now that the sample design specifications of the sample plan have been partially fulfilled, the investigator is ready to determine the size of the sample and the way in which it will be selected. However, before going on to these steps, it is necessary to have some understanding of those factors which affect accuracy. The next chapter provides a brief explanation of sources of error which affect the accuracy of survey results. This will in turn provide the criteria by which one can compare the various sampling schemes. It will also provide a basis for determining the sample size of a survey.

CHAPTER 3

SAMPLING AND NON-SAMPLING ERRORS

Chapter 3

Sampling and Non-Sampling Errors

In the lexicon of survey terms, *accuracy* refers to the difference between a survey result and the true value of a characteristic of the population. *Precision*, on the other hand, refers to the difference between a sample estimate and the result that would be obtained from 100% enumeration.

In surveys there are two basic types of error which arise: *sampling error* and *non-sampling error*.

3.1 Sampling Error

Sampling error is the error attributed to studying a fraction of the population rather than carrying out a census under the same general conditions.

The extent of this error depends on a multitude of factors. For example, the size of the sampling error generally diminishes as the size of the sample increases. However, size is not the only consideration. The nature of such aspects as the variability of the characteristic of interest in the population, the sample design and the estimation method will also have an impact on the extent of the error. Through the development of an efficient sample plan, where proper use is made of available information in developing the sample design and estimation procedure, the sampling error can be reduced.

3.2 Non-Sampling Errors

Non-sampling errors, on the other hand, are present in both sample surveys and censuses. They can arise during the course of virtually all survey activities such as a result of errors in the frame, or difficulties in establishing precise operational definitions. Sometimes, respondents may not be willing to provide correct information, or if they are, they may interpret the questions in different ways. Or, in the later stages of the survey, there may be mistakes in the processing operations. All of these situations contribute to *non-sampling errors*.[1]

Since both sampling and non-sampling errors affect the accuracy of sample results, surveys are designed to minimize their levels to the extent possible.

3.3 Measurement of Sampling Error

Since it is an unavoidable fact that sample results are subject to sampling error, users must be given some indication of just what that error will be. Ideally, the way to assess it would be to measure the difference between the results of a sample estimate and a census. The "Catch 22" of determining sampling error in this way, is that survey sampling was devised to avoid a census. Therefore, taking a census to determine the sampling error of the survey estimate would be defeating the cause!

Since it is seldom possible to measure this difference directly, the approach used is to determine the extent to which sample estimates based upon different possible samples of the same size and the same design differ from one another. In this way, one estimates the sampling error on the assumption that it is possible to draw repeated samples, using the same procedure.

Guides to the *precision* (reliability) of sample results or potential size of sampling errors are provided through *sampling variance*, (defined on the basis of differences in the sample estimates observed in all possible samples), or the *standard error* (square root of the sampling variance) of the estimates.

A relative measure of precision, which is frequently used in sample surveys, relates the standard error of an estimate to its size. Such a measure is called the *coefficient of variation*. This measure is very useful in comparing the precision of different sample estimates, where their sizes or the scale of sample estimates differ from one another.[2]

Realistically speaking, one does not, of course, draw all possible samples to calculate the variance or the standard error of an estimate. However, if *probability sampling methods* (discussed in Chapter 4) are used, the sample

1. For some large scale surveys such as the Census of Population and Housing, and the Labour Force Survey, special studies have been designed to measure some major components of non-sampling error.

2. For the algebraic expressions of variance, standard error and coefficient of variation, See Appendix A, Section 1.2.

estimates and their associated measures of sampling error can be determined on the basis of a single sample.

Estimates are often presented in terms of what is called a confidence interval, to express precision in a meaningful way. A confidence interval constitutes a statement on the level of confidence that the true value for the population lies within a specified range of values.

A 95% confidence interval can be described as follows:

> If sampling is repeated indefinitely each sample leading to a new confidence interval, then in 95% of the samples the interval will cover the true population value.[3]

To further illustrate how measures of precision are related to one another, consider a sample survey on Canadian smoking habits. Let us suppose that the proportion of persons aged 15 and over in Canada who smoke is estimated to be .40, with an estimated variance of .0004, then

i) the estimated standard error is $\sqrt{.0004} = .02$

ii) the estimated coefficient of variation is

$$\frac{.02}{.40} = .05$$

iii) and a 95% confidence interval is between .36 and .44. By this, one means that, with 95% confidence, between 36% and 44% of the target population is made up of people who smoke.[4]

In comparing different probability sampling schemes, reference will be made to a statistical term called *efficiency*. A particular sampling scheme is said to be more 'efficient' than another if, for a fixed sample size, the sampling variance of survey estimates for the first scheme is less than that for the second. Often comparisons of efficiency are made with simple random sampling (see Chapter 4, Section 4.1) as a basic scheme using the ratio of their variances. This is referred to as a design effect.

3. G.W. Snedecor, W.G. Cochran, Statistical Methods, (Iowa University Press, 1967) p.8.

4. Assuming the so-called normal distribution holds, the range of values is determined to be approximately two standard errors above and below the estimate.

CHAPTER 4

PROBABILITY SAMPLING

Chapter 4

Probability Sampling

There are basically two types of sampling methods: *probability sampling* and *non-probability sampling*.

Probability sampling involves the selection of units from a population, based on the principle of *randomization*. It is further characterized by the fact that every unit of the population has a calculable probability of being selected in the sample. For a probability sample a theoretical basis is established for the process of extending the sample results back to the population. In addition, for well-designed probability samples, sampling error tends to be smaller than that of non-probability samples and can be measured.

Selecting probability samples invariably involves the use of random numbers which are available in published tables or can be generated by computer algorithms. Random numbers are usually created by a mechanism, which, when repeated a large number of times, ensures approximately equal frequencies for the digits from 0 to 9 and also correct frequencies for various combinations of these digits. A random number is then used to select one or more sampling units.

There are various types of probability sampling schemes. A sample design uses a combination of one or more of these schemes and techniques. In the following section, descriptions of these schemes are discussed with constant references to a survey on farm expenditures which, as described in the Preface is to determine what it costs to run a farm.

4.1 Simple Random Sampling

Simple random sampling (SRS) is a basic probability selection scheme in which a predetermined number of units from a population list is selected so that each unit on that list has an equal chance of being included in the sample. SRS also makes the selection of every possible combination of the desired number of units equally likely. In this way, each sample has, by definition, an equal chance of being selected. For the probability sampling schemes discussed here, units are drawn one at a time in successive draws.

Sampling may be done with or without replacement. Sampling 'with replacement' allows for a unit to be selected on more than one draw. Sampling 'without replacement' means that once a unit has been selected, it cannot be selected again.

Simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR) are practically identical if the sample size is a very small fraction of the population size. This is because the possibility that the same unit will appear more than once in the sample is small. Generally, sampling 'without replacement' yields more precise results and is operationally more convenient. For the purpose of the present discussion, all references will be to sampling without replacement, unless otherwise indicated.

As a means of illustrating the technique of simple random sampling, consider, then, the farm survey. The object will be to obtain estimates of just what it costs to run a farm in the prairies in a given year.

The survey population for the study will be all farms in Alberta, Saskatchewan and Manitoba which received $250 or more from the sale of agricultural products in 1981.

It will be assumed that a suitable list of such farms is available or can be created from existing sources. Such a list will serve as the sampling frame.

Now, suppose that the population list contains $N = 153{,}000$ farms of which a sample of size $n = 9{,}000$ is needed. The next step is to decide how to select those 9,000 farms.

Selection of the sample can be undertaken by using a table of random numbers (see Table 1, p. 18). In the selection process, the first step involves selecting a six-digit number (six since this is the number of digits in 153,000). One can now begin the selection of a number anywhere in the table and then proceed in any direction. If the decision is made to proceed down the column, the first 9,000 six-digit numbers that do not exceed 153,000 will be selected.

Suppose row 01 and columns 85 to 90 are selected as the starting point. Proceeding down these columns, the respective numbers are 18968, 25668, 98403, 74494, 144147, etc. The selection is continued until 9,000 different numbers are obtained. The result is a sample that consists of farms that carry these numbers in the listing of the population. (It should be noted that since the method under discussion is SRSWOR, any number which appears more than once must be subsequently ignored.)

Although the use of the random number tables has been explained above in the context of manual selection, practically speaking, such a long list of farms would likely be in the form of a computer file and the sample would be generated by means of a computer program.

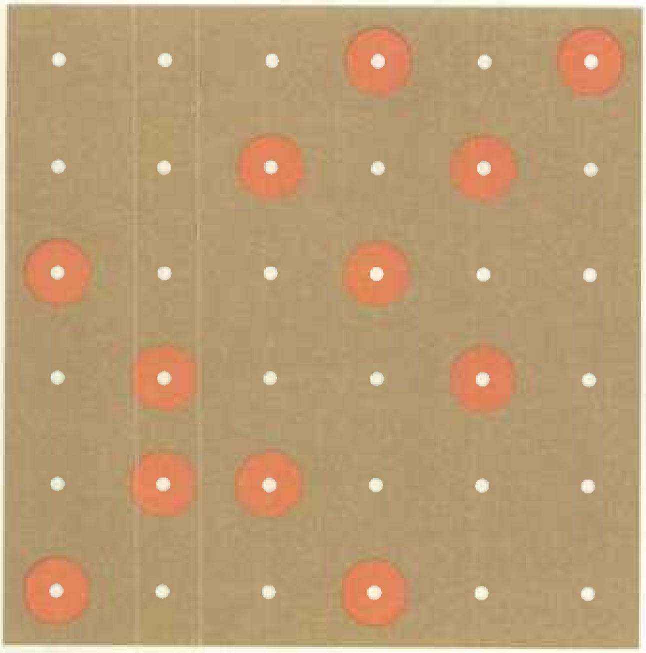
Table 1

Table of Random Numbers

	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
00	59391	58030	52098	82718	87024	82848	04190	96574	90464	29065
01	99567	76364	77204	04615	27062	96621	43918	01896	83991	51141
02	10363	97518	51400	25670	98342	61891	27101	37855	06235	33316
03	86859	19558	64432	16706	99612	59798	32803	67708	15297	28612
04	11258	24591	36863	55368	31721	94335	34936	02566	80972	08188
05	95068	88628	35911	14530	33020	80428	39936	31855	34334	64865
06	54463	47237	73800	91017	36239	71824	83671	39892	60518	37092
07	16874	62677	57412	13215	31389	62233	80827	73917	82802	84420
08	92494	63157	76593	91316	03505	72389	96363	52887	01087	66091
09	15669	56689	35682	40844	53256	81872	35213	09840	34471	74441
10	99116	75486	84989	23476	52967	67104	39495	39100	17217	74073
11	15696	10703	65178	90637	63110	17622	53988	71087	84148	11670
12	97720	15369	51269	69620	03388	13699	33423	67453	43269	56720
13	11666	13841	71681	98000	35979	39719	81899	07449	47985	46967
14	71628	73130	78783	75691	41632	09847	61547	18707	85489	69944
15	40501	51089	99943	91843	41995	88931	73631	69361	05375	15417
16	22518	55576	98215	82068	10798	82611	36584	67466	69377	40054
17	75112	30485	62173	02132	14878	92879	22281	16783	86352	00077
18	08327	02671	98191	84342	90813	49268	95441	15496	20168	09271
19	60251	45548	02146	05597	48228	81366	34598	72856	66762	17002
20	57430	82270	10421	00540	43648	75888	66049	21511	47676	33444
21	73528	39559	34434	88596	54086	71693	43132	14414	79949	85193
22	25991	65959	70769	64721	86413	33475	42740	06175	82758	66248
23	78388	16638	09134	59980	63806	48472	39318	35434	24057	74739
24	12477	09965	96657	57994	59439	76330	24596	77515	09577	91871
25	83266	32883	42451	15579	38155	29793	40914	65990	16255	17777
26	76970	80876	10237	39515	79152	74798	39357	09054	73579	02359
27	37074	65198	44785	68624	98336	84481	97610	78735	46703	98265
28	83712	06514	30101	78295	54656	85417	43189	60048	72781	72606
29	20287	56862	69727	94443	64936	08366	27227	05158	50326	59566
30	74261	32592	86538	27041	65172	85532	07571	80609	39285	65340
31	64081	49863	08478	96001	18888	14810	70545	89755	59064	07210
32	05617	75818	47750	67814	29575	10526	66192	44464	27058	40467
33	26793	74951	95466	74307	13330	42664	85515	20632	05497	33625
34	65988	72850	48737	54719	52056	01596	03845	35067	03134	70322
35	27366	42271	44300	73399	21105	03280	73457	43093	05192	48657
36	56760	10909	98147	34736	33863	95256	12731	66598	50771	83665
37	72880	43338	93643	58904	59543	23943	11231	83268	65938	81581
38	77888	38100	03062	58103	47961	83841	25878	23746	55903	44115
39	28440	07819	21580	51459	47971	29882	13990	29226	23608	15873
40	63525	94441	77033	12147	51054	49955	58312	76923	96071	05813
41	47606	93410	16359	89033	89696	47231	64498	31776	05383	39902
42	52669	45030	96279	14709	52372	87832	02735	50803	72744	88208
43	16738	60159	07425	62369	07515	82721	37875	71153	21315	00132
44	59348	11695	45751	15865	74739	05572	32688	20271	65128	14551
45	12900	71775	29845	60774	94924	21810	38636	33717	67598	82521
46	75086	23537	49939	33595	13484	97588	28617	17979	70749	35234
47	99495	51434	29181	09993	38190	42553	68922	52125	91077	40197
48	26075	31671	45386	36583	93459	48599	52022	41330	60651	91321
49	13636	93596	23377	51133	95126	61496	42474	45141	46660	42338

Source: Rand Corporation (1955) table of random numbers.

Figure 1:
Simple Random Sample (Illustrated)



4.2 Systematic Sampling

The use of simple random sampling can be a long and tedious process if both the sample and the population are large, and particularly if the sample is selected manually. In such instances, *systematic* sampling is a more commonly used selection procedure.

Systematic sampling involves selecting units from a list using a selection interval (K), so that every Kth element on the list, following a random start selected between 1 and K, is included in the sample. If the population size N is an exact multiple of the desired sample size n, then $K = \frac{N}{n}$. Systematic sampling, therefore, requires a sampling interval and a random start.

Consider the farm survey. Suppose that n = 9,000 farms are to be selected using systematic sampling from N = 153,000 farms. The selection interval here is $K = \frac{N}{n} = \frac{153,000}{9,000} = 17$.

If the random start number generated between 1 and 17 is 4, then the units in the sample would correspond to the farms numbered 4, 21, 38, 55, 72, 89, etc. Once the sampling interval is determined, the random selection of the starting point determines the whole sample. In this case, there are 17 such possible samples that can be chosen.

Where N is not a multiple of n, the easiest solution is to use the whole number just below or above K as the interval. This usually results in a sample that is slightly larger or smaller than the initial sample required. For practical purposes, it may be easier to round down in computing the interval K, so that the sample is larger, and then perform systematic deletions. As another alternative, the sample can be selected by a technique referred to as *circular* systematic sampling. This method consists of choosing a random start between 1 and N and thereafter, every Kth unit in a cyclical manner, until a sample of n units is obtained, where K is the integer nearest to $\frac{N}{n}$.

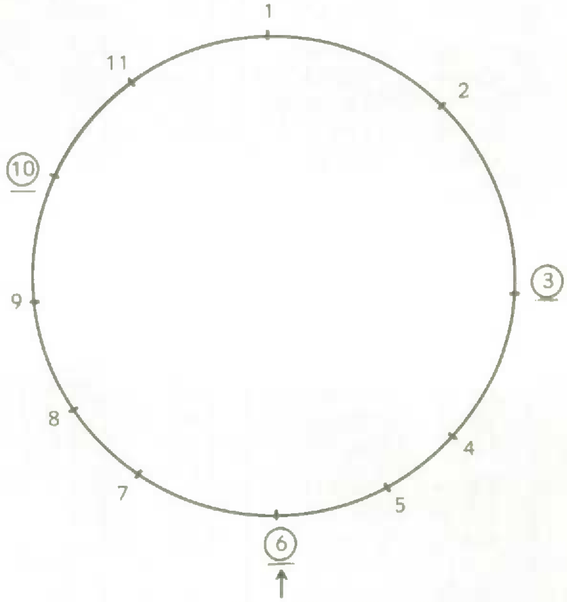
Figure 2:
Systematic Sample Selection Techniques

Example: N = 11, n = 3, K = 4.

linear systematic sample random start (2) selected units (2, 6, 10)

1 (2) 3 4 5 (6) 7 8 9 (10) 11

circular systematic sample random start (6) selected units (6, 10, 3)



Systematic sampling has two advantages over simple random sampling. First, the sample is easier to draw since only one random number is required. Secondly, it tends to distribute the sample over the listed population in a more even way. Higher precision is often associated with systematic sampling, especially if the arrangement of the units in the list is related to the characteristic of interest.

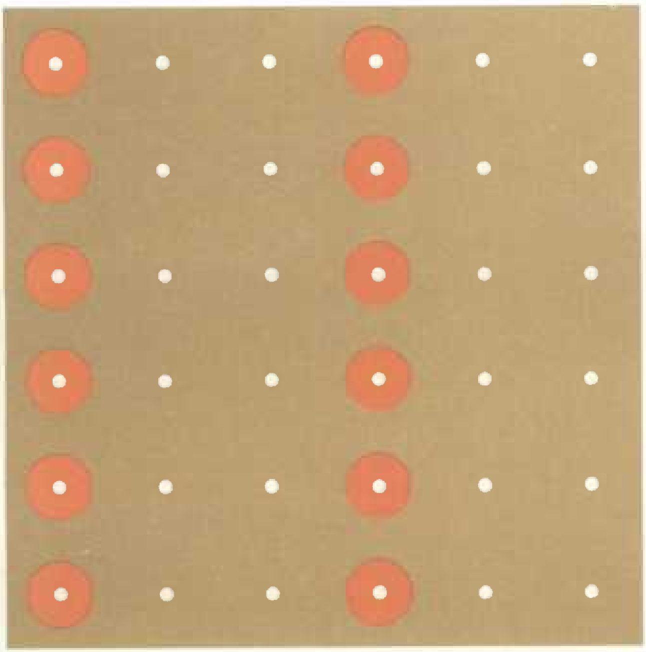
In the farm survey, farms might be listed in geographic order, that is from the south-east to the north-west of a province. If the geographic location of farms is related to farm expenditures, systematic sampling can be more efficient than simple random sampling because of its tendency to scatter the sample throughout the province in a more even way.

There is, however, a disadvantage with the systematic technique. Usually it is referred to as *periodicity*. It is difficult to conceptualize the problem of periodicity. It occurs if or when the list is arranged in a cyclical way. This cyclical pattern in turn may coincide with the sampling interval in such a way as to yield samples that are not representative of the population.

To appreciate the nature of the problem, consider the following example.

If the goal of a survey is to estimate the number of vehicles entering a provincial park over a period of a month, a sample of days might be selected and the total number of vehicles entering the park during selected days observed. If days are arranged in order, then a sampling interval of seven, for example, will consistently yield the same day of the week. However, traffic varies from day to day. Thus, counting vehicles that drive through on a Monday may give the researcher an altogether different result than counting vehicles which arrive on a Saturday.

Figure 3:
Systematic Sample (Illustrated)



The fact is, if the elements of a list are believed to be arranged in a cyclical manner, then simple random sampling may be the appropriate alternative.

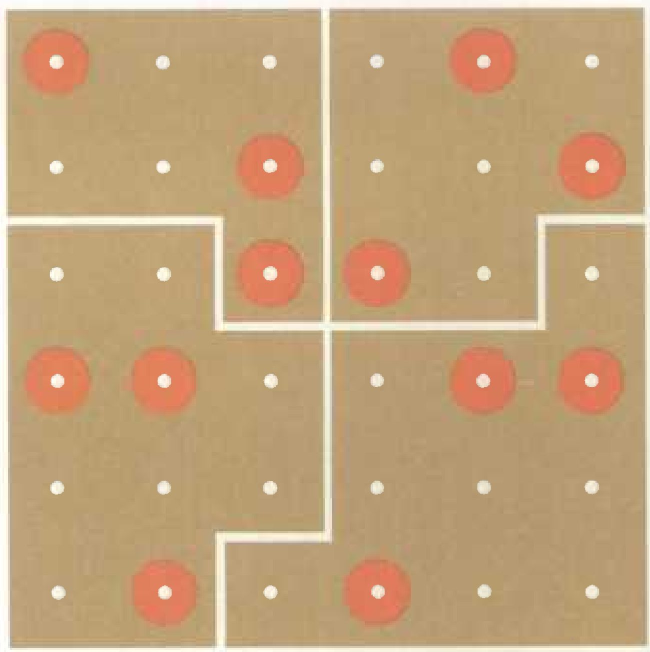
4.3 Stratified Sampling

In the case of simple random sampling, the selection of the sample is left entirely to chance. All that is required to select a sample is a population list and the use of random numbers. No use is made of any relevant information which might be available for members of the population. *Stratified sampling* is a technique which uses such information in order to increase efficiency. Stratified sampling involves the division or stratification of a population into relatively homogeneous groups called *strata* and the selection of samples independently in each of those strata.

To understand how the use of relevant information reduces sampling variance, consider a survey of cigarette smokers. The goal is to determine the proportion of smokers in the population. It is clear that the sample selected must properly represent both men and women, especially if it is true that more men light up than do women! The proportion of smokers is also known to vary considerably between age and occupational groups. Therefore, it would be desirable to select a sample which properly represents each of these groups. By using relevant information about a population, stratified sampling reduces the possible samples which can be selected to those which provide a better representation of the population.

Stratified random sampling, in particular, involves dividing the population into strata, and then selecting simple random samples from each of the strata. Stratification variables may be geographic (region, province, rural/urban) or non-geographic (income, age, sex, number of business employees, etc.) It should be kept in mind that stratification is limited only to those items of information which are available on the frame.

Figure 4:
Stratified Random Sample (Illustrated)



4.3.1 Advantages of Stratified Sampling

Basically, stratified sampling attempts to restrict the possible samples to those which are 'less extreme' by ensuring that all parts of the population are represented in the sample. It follows that the more homogeneous the groups, the greater the precision of the sample estimate.

Sometimes, separate stratum estimates are required at the stratum level. In household surveys, for example, estimates may be required by province, income group, occupation, age group, urban size group, etc. In business surveys, estimates are often required by Standard Industrial Classifications and Standard Geographical Classifications.[1] Stratified sampling ensures that each sub-population is adequately represented in the overall sample.

Stratified sampling is administratively convenient. It can enable a survey organization to control the distribution of fieldwork among its regional offices. Also, for large complex surveys, stratified sampling can facilitate sample design work by enabling such work to be carried out within operationally manageable units.

Sometimes, different parts of the population may call for different sampling procedures. In the study of human populations, a different procedure may be used to sample persons in sparsely populated rural areas than that used for the more densely populated urban areas. There may also be differences in the lists available for different parts of the population.

In the farm expenditure survey, it is desirable to stratify the population of farms on the basis of information which is highly related to the cost of running a farm. If it is the case that farm expenses vary with the size of farm and that expenses vary with the type of farming activity, one may consider *size* (acreage) and *type of farming* (crops, livestock, other, etc.) as appropriate stratification variables. If separate estimates are required by province, the list of farms may be further stratified by province.

However, in order to group the population of farms by province, by size classes and by type of farming, such information must be available for all farms on the sampling frame. Assuming these strata can be formed, one may then proceed to select random samples of farms from each of the strata. (See Table 2, p. 21).

4.3.2 Sample Allocation

An important consideration in stratified sampling is the way in which the total sample size is allocated to each of the strata. This may be done on a proportionate or disproportionate basis.

With *Proportional Allocation*, the sample allocated to each stratum is proportional to the total number of units in the stratum. That is, the sampling fraction in each stratum is made equal to the overall sampling fraction of the population.

In the farm survey, the first stratum in Alberta contains

$$\frac{20,000}{50,000} = \frac{2}{5}$$

of all farms in Alberta.

If a sample size of say 3,000 farms is to be proportionately allocated to the strata in Alberta, then $(\frac{2}{5} \times 3,000) = 1,200$ farms would be allocated to the first stratum. The sample allocation to the strata in Alberta is given in Table 3 on page 22.[2]

Table 2:

Example of Sample Allocation to Strata

Province: Manitoba **Strata**

Type of Farming	Size (Acreage) Class			Total
	0-250	251-500	501 +	
Crops	25,000	7,000	2,000	34,000
Livestock	7,000	5,000	3,000	15,000
Other	4,000	3,000	4,000	11,000
Total	36,000	15,000	9,000	60,000

Province: Saskatchewan

Type of Farming	Size (Acreage) Class			Total
	0-250	251-500	501 +	
Crops	18,000	9,000	1,000	28,000
Livestock	4,000	2,000	2,000	8,000
Other	4,000	2,000	1,000	7,000
Total	26,000	13,000	4,000	43,000

Province: Alberta

Type of Farming	Size (Acreage) Class			Total
	0-250	251-500	501 +	
Crops	20,000	8,000	5,000	33,000
Livestock	6,000	3,000	2,000	11,000
Other	4,000	1,500	500	6,000
Total	30,000	12,500	7,500	50,000

1. Standard Industrial Classification is a scheme developed by Statistics Canada which classifies industries on the basis of their principal activities. This scheme consists of 11 divisions which include all branches of economic activity. Statistics Canada has also developed the Standard Geographical Classification, a system for identification and coding of geographical areas. This system employs unique code numbers which reflect municipal boundaries and has a list of place names related to these units. The objective of the system is to make available a standard set of geographical units which can be used by Statistics Canada and others to facilitate the comparison of statistics for particular areas.

2. The data used here is for illustration only.

Table 3:

Example of Proportional Allocation

Province: Alberta

Type of Farming	Size (Acreage) Class			Total
	0-250	251-500	501 +	
Crops	1,200	480	300	1,980
Livestock	360	180	120	660
Other	240	90	30	360
Total	1,800	750	450	3,000

If the variability among units differs substantially from stratum to stratum or the cost per interview differs between strata, a disproportionate allocation scheme known as *Optimum Allocation* may be considered.

Optimum Allocation is intended to increase the sampling rates in those strata in which the variances are relatively larger, and to decrease the sampling rates in those strata in which the cost of an interview is relatively greater. However, in order to use this scheme, accurate information on stratum variances and interview costs is required.

Where the consideration of differential interview costs among strata is ignored, this allocation scheme reduces to what is referred to as *Neyman Allocation*.

Another disproportionate allocation scheme is called *X-Proportional Allocation*. This scheme is not concerned with the number of elements, but rather measures of their size (such as farm acreage). This scheme can be used when sampling units differ in size and the size is highly correlated with the characteristics of interest. Instead of allocating a sample proportionally to the total number of units in a stratum as in the case of proportional allocation, the sample is allocated proportionally to the total size of all units in a stratum.[3]

Disproportionate allocation schemes may also be used when the strata themselves are of principal interest. Suppose that estimates of farm expenditures are required for each of the three prairie provinces. In this case, it would be important to ensure that each province be allocated a sufficient part of the overall sample to enable such estimates to have sufficient reliability. In the farm survey, each province has therefore, been allocated 3,000 units out of the total sample size of 9,000.

So far, only simple random sampling has been discussed in the context of stratification. But, since stratification is a technique for structuring the population, it can be used with any of the sampling techniques that will be discussed.

4.4 Unequal Probability Sampling

In the case of simple random and systematic sampling, units are selected with equal probability. Probability sampling methods require that every unit in the population has a calculable probability of selection. Thus, there is no restriction that units be selected with equal probability.

If sampling units vary in size and these sizes are known, such information may be used in the sample selection process to increase efficiency. A sampling technique in which size measures are considered in selecting the sample is referred to as sampling with *probability proportional to size* (PPS).

Suppose now, that the cost of running a farm is directly related to its size. The effect of using size information in the selection of the sample can be explained by the following example. Here the survey on farm expenditures has been simplified to the selection of farms from a smaller population or stratum.

Farm	Acreage	Farm Expenditures ($)
1	50	26,000
2	1,000	470,000
3	125	63,800
4	300	145,000
5	500	230,000
6	25	12,500
	2,000	947,300

If one farm is selected with equal probability to represent the population of six farms without taking its size into account, the selection of a small farm (say Farm #1) would tend to yield an underestimate of total farm expenditures (26,000 x 6 = 156,000) while the selection of a large farm (say Farm #5) would tend to yield an overestimate (230,000 x 6 = 1,380,000).

The estimate would, however, be unbiased as the following table demonstrates.[4]

3. For an algebraic description of the various allocation schemes, see Appendix A, Section 1.3.

4. An estimator is unbiased if the values associated with all possible samples will average to the population value. Strictly speaking, an estimator is a mathematical rule. An estimate on the other hand is a specific value assumed by an estimator for a given sample. For simplicity, only the term estimate will be used.

Farm	Acreage	Farm Expenditures	Estimate of Total Expenditure ($)
1	50	26,000 x 6 =	156,000
2	1,000	470,000 x 6 =	2,820,000
3	125	63,800 x 6 =	382,800
4	300	145,000 x 6 =	870,000
5	500	230,000 x 6 =	1,380,000
6	25	12,500 x 6 =	75,000
	2,000		5,683,800

$$(\text{average} = \frac{5{,}683{,}800}{6} = 947{,}300)$$

On the other hand, if Farm #2 were selected, one might reason that since it accounts for $\frac{1{,}000}{2{,}000}$ = 50% of the total acreage of all farms, it might also account for 50% of the total expenditure. Under such an assumption, the estimate of total farm expenditures would be 470,000 x $\frac{2{,}000}{1{,}000}$ = 940,000. Estimates associated with each of the other six farms is given below.

Farm	Acreage	Farm Expenditures	Estimate of Total Expenditure ($)
1	50	26,000 x (2000/50) =	1,040,000
2	1,000	470,000 x (2000/1000) =	940,000
3	125	63,800 x (2000/125) =	1,020,800
4	300	145,000 x (2000/300) =	966,667
5	500	230,000 x (2000/500) =	920,000
6	25	12,500 x (2000/25) =	1,000,000
	2,000		(average) 981,245

In the table immediately above, one can easily see that the estimates associated with each farm, tend, on an average, to be closer to the actual farm expenditure (= 947,300) than the estimates in the previous table which ignore farm size. One might also note that if farm expenditures were exactly proportional to farm acreage then any farm could be selected and used to estimate the total farm expenditures of the population.

Although the estimates in the last column of the table above tend to be close to the actual population value, the estimate is biased, since the average over all possible samples does not equal the population value (1,040,000 + 940,000 + 1,020,800 + 966,667 + 920,000 + 1,000,000) / 6 = 981,245 ≠ 947,300). In this table one can see that there is a tendency on the average to overestimate the actual value.

By varying the probabilities with which a unit is selected according to its size (that is, selecting the units with probability proportional to size) one can ensure that the estimate is unbiased. These probabilities of selection appear in the following table.

Farm	Estimates of Total Expenditures ($)	Probability of Selection
2,000		
1	1,040,000	$\frac{50}{2{,}000}$
2	940,000	$\frac{1{,}000}{2{,}000}$
3	1,020,800	$\frac{125}{2{,}000}$
4	966,667	$\frac{300}{2{,}000}$
5	920,000	$\frac{500}{2{,}000}$
6	1,000,000	$\frac{25}{2{,}000}$
		2,000

In this case, the average value of the estimate over all possible samples =

$$(1{,}040{,}000 \times \frac{50}{2{,}000}) + (940.000 \times \frac{1{,}000}{2{,}000}) + (1{,}020{,}800 \times \frac{125}{2{,}000})$$
$$+ (967{,}667 \times \frac{300}{2{,}000}) + (920{,}000 \times \frac{500}{2{,}000}) + (1{,}000{,}000 \times \frac{25}{2{,}000})$$
$$= 947{,}300$$

Here, the number of acres is referred to as a *size measure* of the farms. In selecting a sample of farms with PPS, Farm #2 has a higher probability of being in the sample since it is larger than any other farm. Similarly, Farm #5 has a higher probability of selection than Farms #1, #3, #4 or #6.

Generally, sampling with probability proportional to size should be considered when accurate measures of size are available for sampling units and where there is a strong correlation between size and the characteristics of interest.

There are a number of procedures to select units with probability proportional to size. Two of these methods are explained in the following pages but, neither can be undertaken without the development of a crucial first step: the cumulation of size measures for each of the farms. The following table contains a *range* of numbers to facilitate the mechanical selection of the farms.

Table 4

Framework For PPS Selection

Farm	Acreage	Cumulative Size	Range	Farm Expenditure
1	50	50	1-50	26000
2	1000	1050	51-1050	470000
3	125	1175	1051-1175	63800
4	300	1475	1176-1475	145000
5	500	1975	1476-1975	230000
6	25	2000	1976-2000	12500

4.4.1 Probability Proportional to Size – Random Method (PPS-Random)

With PPS random sampling, a random number between 1 and 2000 is selected. Subsequently, the farm corresponding to the range in which the random number falls is selected. If, for example, the random number 100 is drawn, then the farm associated with the range 51-1050 (Table 4) is selected. In this case Farm #2 would be selected. Since a random number between 1 and 2000 will fall in the range of 51-1050 with probability $\frac{1000}{2000} = \frac{1}{2}$, the probability that Farm #2 will be selected is seen to be proportional to its size. With this method, more than one unit can be selected either with or without replacement. In the case where more than one unit is selected, *without* replacement, complications arise both in attempting to keep probabilities directly proportional to size and in estimating the sampling variances of survey estimates.

The problems are even more serious when more than two or three units are selected with PPS without replacement, and in fact, is the subject of considerable research.[5]

A second widely-used method of PPS sampling without replacement is PPS-Systematic.

4.4.2 Probability Proportional to Size – Systematic Method (PPS-Systematic)

PPS systematic, ensures that the probabilities of selection are proportional to size and is fairly easy to apply. Suppose then, that two units are to be selected from Table 5. The sampling interval $K = \frac{\text{cumulative size}}{\text{sample size}} = \frac{2000}{2} = 1000$ is first determined. A random start number between 1 and 1000 is then drawn. Subsequent numbers are determined by repeatedly adding the sampling interval to the accumulated total. The farms corresponding to the ranges in which the random numbers fall are then selected in the sample. For example, if the random number drawn between 1 and 1000 is 69, then in a sample of size 2, the next number is 1069. Since 69 falls in the range 51-1050 and 1069 falls in the range 1051-1175, farms 2 and 3 are selected in the sample.

A third widely-used method combines elements from the two discussed here. In the *randomized systematic method with probability proportional to size*, the order of the sampling units is first randomized and a systematic sample is then selected with PPS.

Yet, these methods do pose certain problems. For example, if the size of any unit is greater than the interval, it may just be selected more than once. This problem can only be overcome by placing such large units into separate strata and sampling them independently. A second problem deals with the difficulty of estimating sampling variances.[6]

For the most part, the problem with PPS sampling centres on the size measures of the sampling units. In order for PPS to be efficient, size measures must be accurate. Often, however, size measures go out of date quickly. Business surveys provide a prime example of this changeability where the 'size' (e.g. number of employees) can change substantially over time. Considering this, it is often preferable to stratify by broad size classes for which class membership is fairly stable.

Up until now, the discussion on *size measure* has referred to the largeness of a unit (i.e. acres). Later on, it will be seen that a size measure can also refer to the number of elements contained within sampling units.

List Samples and Area Samples

So far, the presentation of probability sampling schemes has been concerned with 'list samples' where the sample is selected from a complete list of units of the survey population.

If such a list of units for the survey population is not available or is inadequate (e.g. out-of-date), a list may have to be constructed for the survey. In the event that the cost and time of creating a list of all units in the survey population is prohibitive, attention is directed to methods in which the creation of lists may be confined to a limited number of geographical areas.

Further, a sample selected from a geographically dispersed population is likely to be very widely dispersed as well. This may not pose a problem in the case of mail surveys but is of great concern for surveys which require on-site personal interviews or observation. To reduce travel costs, one can consider techniques in which the resulting sample may be concentrated within a number of geographical areas.

5. Much of this research is contained in the writings of Horvitz and Thompson (1952), Yates and Grundy (1953), Rao, Hartley and Cochran (1962) and Fellegi (1963). (For full references, see Bibliography.)

6. These problems are discussed in W.S. Connor (1966), M.A. Hidiroglou and G.B. Gray (1981) and G.B. Gray (1971). (For full references, see Bibliography.)

In the absence of a suitable sampling frame and/or when one is faced with a widely dispersed population where travel cost may be a key factor, area sampling techniques may be used.

'*Area Sampling*' involves the selection of geographical areas. The requirements for an up-to-date list is confined to areas in which the sample itself is concentrated. The probability sampling schemes used in area sampling are cluster and multi-stage sampling.

4.5 Cluster Sampling

In order to use cluster sampling, a population has to be structured in terms of a hierarchy. Consider people who live in a city. Their homes form the base of that hierarchy. Those homes in turn make up city blocks and ultimately the city blocks make up the city.

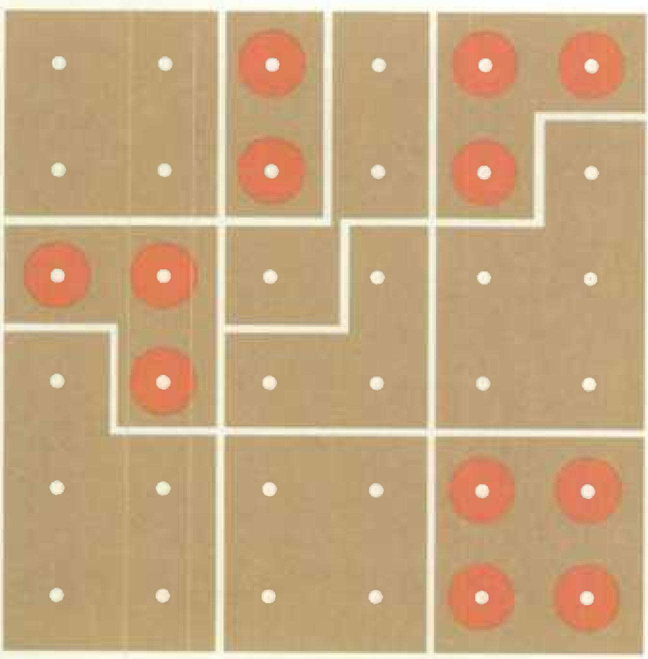
The process of sampling city blocks or dwellings in order to select a sample of the people who live within them is called *cluster sampling*.

There are many advantages to cluster sampling.

Sometimes situations arise where a list of units for the survey population (e.g. people) is not available or is out-of-date. However, lists of city blocks are available or can be easily obtained from such sources as the Census of Population and Housing. An advantage here is that the lists of people required is necessary only for *selected* clusters.

Another advantage of cluster sampling is that it concentrates a sample into compact groups. This reduces costs associated with travel between units as well as the necessary supervision of fieldwork and the callbacks of non-respondents.

Figure 5:
Cluster Sample (Illustrated)



To further illustrate the technique, consider an area frame composed of Census Enumeration Areas (EAs). In the farm survey, one might consider only those EAs which contain the headquarters of at least one farm. A sample of these EAs might be selected and then all the farms within the selected EAs could be listed and selected for the interview.

One of the disadvantages of cluster sampling is that it usually yields less efficient estimates than a simple random sample. The problem is that neighbouring elements tend to be more alike.[7] The farm survey provides an excellent example since neighbouring farms tend to have the same kind of soil and terrain; characteristics they do not share with distant farms.

The problem extends itself to human populations, as well. People living within a district of a city are more likely to share similar social and economic characteristics as a group than the population of a city as a whole. The same relation holds true for individual families. People who share a roof are likely to have more characteristics in common than members of the population at large.

Occasionally, when the relationship among the members of clusters is such that they are more dissimilar with respect to the characteristics of interest than the survey population in general, cluster sampling is more efficient than simple random sampling. In the Labour Force Survey, for example, labour force participation is negatively correlated among individuals in households (occupied dwellings). The use of households as clusters in this case is more efficient than selecting an equivalent number of persons taking one person per household.

In fact, the more heterogeneous the clusters are within themselves, the more efficient the cluster sample is likely to be. This is just the opposite of what is required in the stratification of a population, where the strata should be as homogeneous as possible. But, if there is a superficial resemblance of clusters to strata, it is because a cluster, like a stratum is a grouping of members of the population. Where the two differ is in the methods used to select the sample. In stratified sampling, each stratum is sampled independently. In cluster sampling, a sample of clusters is selected. Cluster sampling is applied to groups of population members where each group is considered a single unit in the selection process.

For maximum precision in cluster sampling, it is therefore desirable that the units which comprise clusters vary as much as possible. Since, it has already been seen that units within a cluster tend to be similar, it is better to survey a large number of small clusters than a small number of large clusters.

7. This problem is treated algebraically in Section 1.5.3 of Appendix A.

Consider again the farm survey. While the EAs are the smallest geographical units for which lists and information is readily available, it might be desirable to select units that are smaller. This might be done by breaking each EA into segments, or smaller size clusters. These segments could then be selected rather than the EAs.

However, one of the disadvantages is that *all EAs* would have to be subdivided into segments to construct the frame. Such an operation would be costly and time consuming.

Also, breaking EAs into segments would likely mean a greater geographic spread in their location, and again, this would increase the travel costs involved.

To avoid excessive loss of efficiency associated with the selection of large clusters such as EAs, and to avoid excessive travel costs, and complexities in the construction of the frame associated with selection of smaller clusters (segments), one might consider a scheme which involves selecting the required number of segments from a *predetermined* number of selected EAs as in the case of multi-stage sampling.

4.6 Multi-Stage Sampling

Multi-stage sampling refers to a process of selecting a sample in two or more successive stages. It involves a hierarchy of different types of units. Each 'first-stage' unit is potentially divisible into 'second-stage' units and so on.

At each stage of sampling, a good sampling frame is required. Initially, it consists of a list of all the first-stage units. A specified number of first-stage units is then selected from this frame.

For the second stage of selection, a frame is required for the second-stage units within the larger units *which have already been selected* at the first stage. In fact, one of the advantages of multi-stage sampling is that units selected at one stage provide the frame for the next stage of sampling.

The sampling units at the first stage are called primary sampling units, while the sampling units at the second stage are called secondary sampling units. Primary sampling units may be EAs, groups of EAs, census subdivisions or even city blocks.[8]

In the farm survey, one can select Enumeration Areas on the basis of a list (possibly stratified) of such EAs. Within selected EAs, a list of smaller areas called segments may then be prepared for each EA (a segment being one of the equal divisions of an enumeration area). Such lists would provide the frame for the second stage of selection.

Next, farms from these selected segments may be listed in the field, and a sample of farms selected from that list. Such a process provides an example of a three-stage sample design.

8. The data source which often serves as the basis for the creation and selection of such higher stage units is the Census of Population and Housing.

The overall probability of selecting a farm in the sample is now calculated as the product of the probabilities of selection of the sampling units at each stage.

In other words, the probability, P, that a farm is selected is contingent on:

a) selecting the EA in which the farm is located from a list of EAs, (P_1),
b) selecting the *segment* in which the farm is located from a list of segments within the selected EA, (P_2) and,
c) selecting the farm from a list of farms within the selected segment, (P_3), such that

$$P = P_1 \times P_2 \times P_3$$

(Remember, as long as units at each stage are selected with a probability which can be calculated, the resulting sample is a probability sample.)

To further illustrate the concept of a multi-stage sample, consider the case of a personal interview survey designed to measure public opinion on a new government service to be offered in a city.

The units at the first stage might be city blocks or groups of city blocks. The second stage might consist of dwellings selected from a list of dwellings prepared for selected city blocks. Units at the third stage would then be persons selected from a list of persons within the selected dwellings.

Techniques may vary depending on whether the survey is in Trois-Rivières or Toronto. Since Toronto is such a large city, city blocks might be initially stratified on the basis of administrative districts within the city.

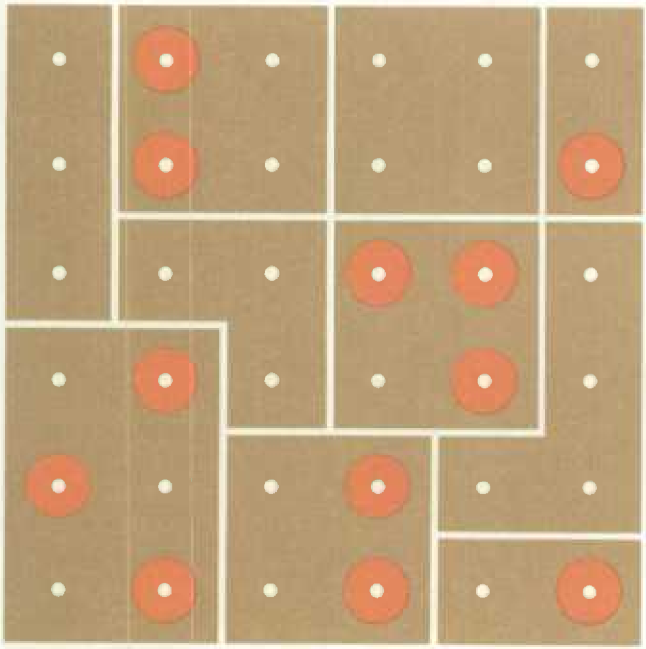
Since frames for higher stage units are generally more stable than those for the lower stages, the latter are often based on current field counts and listings.

In the discussion on cluster sampling, it was pointed out that it is best to keep the size of clusters as small as possible. This principle does not, however, extend to multi-stage sampling. In this case, the use of large units as primary sampling units with further subsampling is often more efficient than the use of smaller clusters without sub-sampling. However, as in cluster sampling, it is desirable that higher stage units be as heterogeneous as possible.

Sampling units may be selected with equal or unequal probability at any stage in a multi-stage sample design. If accurate size measures are available for higher stage units such as EAs or segments, the units may be selected with probability proportional to size. (An advantage of PPS selection is further discussed in Chapter 6.)

So far, cluster sampling and multi-stage sampling have been discussed in the context of area sampling. It should be emphasized, however, that these techniques do have other applications. One such application, for example, might be to measure the characteristics of travellers (origin/destination, expenditures) passing through border points by car over a certain period of time. It might be necessary to confine the sample of selected travellers to a limited number of time periods within the reference period to reduce interview costs.

Figure 6:
Multi-Stage Sample (Illustrated)



4.7 Multi-Phase Sampling

A multi-phase design is one in which some information is collected from a large preliminary sample and additional information is collected from subsamples of the entire sample, either at the same time or at a later time. In the case of only one subsample, the technique is called two-phase or double sampling. *Multi-phase sampling* is generally used:

1) as a means of increasing the efficiency of a sample;
2) for surveys which have different levels of data requirements involving considerably different costs of collection and/or respondent burden.

The farm expenditure survey will be used to illustrate the first context. If the type of farm (livestock, crops, etc.) is highly correlated with farm expenditure it would be desirable to stratify the population by type of farm prior to sample selection. Such an approach is of course precluded if information on the type of farm is not available for farms listed in the sampling frame. To increase the efficiency of a sample by exploiting the relationship between the type of farm and farm expenditure, a large preliminary sample might be selected to obtain only information on the type of farm. This large sample might be stratified and the actual detailed information on farm expenditure then collected from a random subsample selected from the initial sample. In effect, this situation may be considered to be a special case of *post stratification*. Post stratification refers to the stratification of a sample rather than the entire population and is used in situations where information which would be useful for stratification is unavailable for the survey population.

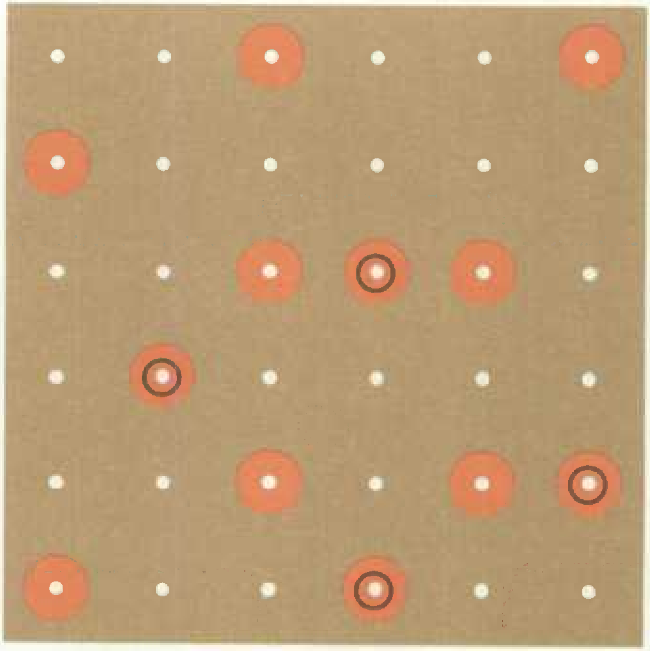
The Canada Health Survey and the Census of Population and Housing provide good illustrations of the second context of multi-phase sampling. In the Canada Health Survey there were two levels of data requirements: (a) items dealing with general living habits which included factors such as smoking, nutrition, exercise and incidence of various diseases, and (b) physical measures which required the taking of blood samples, blood pressure readings, step tests, etc. Registered nurses and specialized equipment were required to administer the physical measures component of the survey. This information was considerably more expensive to collect and subjected persons to greater response burden than the first component of the survey. Hence, it was collected on a *subsample* (2nd phase sample) of households selected for the first level of data. In the Census of Population and Housing a short form containing a set of basic questions (including age, sex, marital status, family size) is administered to the entire population. A longer form, which includes all the questions on the short form, contains further items on, industry/occupation, housing and employment. The long form is administered to a random sample of the population.

In some cases, subsamples of the entire population are not randomly selected but defined to satisfy certain criteria. These are referred to as *screened samples*.

Screened samples may be used where one is interested in, say, only those families whose incomes fall below a certain level; or in the farm survey, it might be used if one is interested only in those farms which produce wheat. A sample may have to be screened, if one cannot pre-identify the particular part of the population of interest in the frame.

It should be remembered that for multi-phase sampling, one is concerned with the *same type* of sampling unit at each phase; whereas for multi- stage sampling, different types of units are sampled at different stages of sampling.

Figure 7:
Multi-Phase Sample (Illustrated)



4.8 Replicated Sampling

Replicated sampling involves the selection of a number of independent samples from a population rather than a single sample. Instead of one overall sample, a number of smaller samples, of roughly equal size, called replicates, is independently selected, each based upon the same sample design.

Thus, if a stratified three-stage sample design were employed for the farm expenditure survey, each replicate would be based upon the same stratified three-stage design.

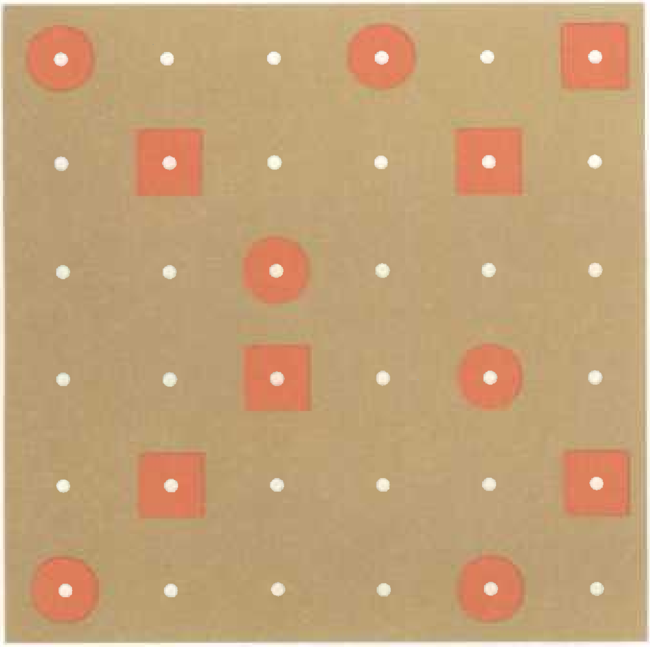
The principal reason for replicated sampling is to facilitate the calculation of standard errors of survey estimates. While it is generally possible to calculate standard errors of estimates based on probability samples, such calculations can be exceedingly difficult depending on the complexity of the sample plan. The problem is that mathematical expressions for standard errors are difficult to derive and tedious and costly to program. In particular, in the case of systematic sampling, variance estimates cannot be calculated directly unless assumptions are made about the arrangement of units in the list.

It has already been explained that measures of sampling error are determined by examining the extent to which sample estimates, based upon all possible samples of the same size and same design, differ from one another. Replicated sampling simulates or copies this concept. Instead of drawing all possible samples, it allows a reasonable number of smaller samples to be selected using identical methods. For example, instead of selecting one sample of size 10,000, one might draw 10 independent samples of size 1,000.

Further, the reliability of estimates of standard error increases with the number of replicates used in the sample design. Yet, there are drawbacks to the approach. Since it is not practical to use very many replicates, a disadvantage of this scheme is that estimates of standard errors tend to be less precise than if they were based directly on the statistical expressions which incorporate sample design features such as multi-stage, stratification, etc. Sample estimates based on replicated samples also can be less precise if a sampling unit is allowed to be selected in more than one replicate.[9]

Replicated sampling might also be used in situations where preliminary results are needed quickly. Such preliminary results might be based upon the processing and analysis of a single replicate.

Figure 8:
Replicated Sample (Illustrated)



SUMMARY

In the discussion on probability sampling methods, one may think of two distinct steps in the sample design.

The first step involves a process of *structuring* the frame prior to any selection. The second step involves the process of *selecting* the units from the frame after it has been structured.

Structuring in turn refers to the possibility of stratification of units and/or clustering of units into larger groups. Such larger groups of units may already exist through the Census (EAs) or may have to be prepared for the survey from maps or other sources. Finally, there is the selection of units. This is carried out using either a probability or a non-probability sampling scheme. In the case of probability sampling, units may be selected with equal probability (SRS, systematic) or unequal probability (PPS-random, PPS-systematic) and with or without replacement.

The development of any sample design involves decisions concerning the number and type of stratification variables and formation of strata, the size of the sample, how it is to be allocated to the strata, and how the sample is to be selected within each stratum.

Such decisions are contingent on a careful consideration of cost, reliability and operational suitability.

Although a probability sample is quite often referred to as a "Scientific Sample", the use of this term, often wrongly, lends credence to the results of a sample survey. While a probability sample does have a well-founded theoretical basis, it can nevertheless yield poor results. This can happen in a number of ways. The sample design itself may not be an efficient one. The concepts, operational definitions, procedures and control mechanisms may not be adequate for the survey. The response rate may be very low. Indeed, whether a poorly designed probability sample survey is in fact better than a well-controlled non-probability sample survey is a moot point.

9. More technical information on replicated sampling can be found in the writings of McCarthy (1966) and Keyfitz (1957). (See Bibliography for complete references.)

CHAPTER 5

NON-PROBABILITY SAMPLING

Chapter 5

Non-Probability Sampling

Although *non-probability* sampling methods are generally less statistically accurate than probability sampling methods they are generally cheaper and more convenient. However, the fact that there is no way to measure the precision of non-probability samples makes most statisticians reluctant to use them.

The difference between the two sampling methods has to do with a basic assumption about the nature of the universal population under study. Non-probability sampling depends on assumptions of an even or random distribution of characteristics in the population. Probability sampling methods, on the other hand, do not need such an assumption about the structure of the population. Randomization is already a part of the selection process.

Since with non-probability sampling, elements are chosen in an arbitrary manner, there is no way of estimating the probability of any one element being included in the sample. Moreover, there is no assurance that every element has a chance of being included. This makes it impossible to estimate either the sampling variability or identify the possible biases involved.

Since reliability cannot be measured, the only way to address the quality of survey data is to compare some results of the survey with available information about the population. Even here, there is still no assurance that the estimates will have an acceptable level of error. However, despite such drawbacks, non-probability sampling can still be a useful tool, as the following examples illustrate.

5.1 Haphazard Sampling

At some point in life, everyone uses haphazard sampling techniques. At a cocktail party, one might try 'sampling' several drinks to find out which one tastes best. Television reporters often go after so-called man-in-the-street interviews to find out what the general public's views are on an issue in the news. In both these cases, little conscious planning goes into the selection of the sample.

However, despite the fact that useful applications of the technique are limited, haphazard samples can and do provide useful results when the population examined is homogeneous.

Consider, for example, the problem of determining the concentration of a chemical in a lake or the sugar level of the blood.

On the assumption that the lake or circulating blood is well mixed, any sample would give very similar information. This technique might also be applied to investigate the range of people's attitudes and opinions on certain topics of interest.

5.2 Sampling of Volunteers

In cases of certain psychological or medical experiments it would not be practical to enlist people randomly selected from the population. In such cases, the sample will consist of people who have volunteered their services, knowing that the process will be lengthy or demanding, and perhaps even unpleasant.

It should be recognized though, that the difference between these people and the general population may introduce large biases. In the case of attitude surveys, for example, volunteers have often been found to have favourable or at least neutral attitudes whereas the general population tend to hold a wider range of attitudes on topics of interest.

5.3 Judgment Sampling

As the term implies, *judgment sampling* calls for a selection of units on the basis of certain judgments concerning the make-up of a population. Such an approach is often used in exploratory studies such as pilot tests, pretests of survey questionnaires and focus groups. It is also frequently used in experimental settings in which the subjects (mice, crops, people) of an experiment reflect the investigator's judgment about the population.

A major advantage of judgment sampling is the reduced cost and time involved in acquiring the sample. However, since there are often disagreements between different investigators on the way to choose representative units, sample selection can be severely biased. In addition, there is often a tendency to eliminate 'extreme' units found in the population in attempting to select 'typical units'. The result can lead to a distorted picture of the underlying characteristics of the population.

5.4 Quota Sampling

Quota sampling is widely used in opinion and market research surveys. Sometimes, it is called a *purposive* type of sample since interviewers are instructed to obtain the required number of interviews in each group defined by such variables as geographical area, age, sex and possibly other demographic variables. These quotas are often determined to be roughly proportional to the corresponding size of the population represented by the group.

To illustrate this technique, suppose that a national opinion survey is to be based on a quota sample. Just as in the case of a stratified random sample, the first step might be to stratify the population by region or province, city or county. In fact, the two methods do not necessarily differ from one another when it comes to selecting areas such as EAs or groups of EAs within the strata.

It is common, though not necessary, for quota samples to employ random selection procedures at the initial stage of selection in exactly the same way as probability samples. The essential difference lies in the selection of sampling units in the final stages of the process. With a probability sample, units are based on up-to-date lists and a sample is selected according to a random process. With quota sampling, each interviewer is given an assignment of interviews with instructions specifying how many of them are to be with men and how many with women, how many with people in various age groups and so forth.

In this way, quotas are calculated from available data such that, for the population under study, the sexes, age groups and social classes are represented in the sample in the right proportions.

Arguments for Quota Sampling

It has already been noted that a quota sample is generally less expensive than a probability sample.

In addition to the cost factor, it is also generally easier to administer a quota sample, since tasks of listing, random selection and follow-up of non-respondents can be avoided.

Quota sampling can be a handy approach where information is urgently needed. In fact, in order to reduce recall errors, it is sometimes the only option.

Quota sampling can be carried out independently of the existence of sampling frames, particularly in the final stages of selection. Indeed, it may be the only practical method of sampling a population for which no suitable frame is available.

Arguments Against Quota Sampling

On the other hand, quota sampling does not permit sampling errors, to be estimated. Also within each quota, interviewers may fail to secure a representative sample of respondents.

Given that all the quotas are correctly filled, who is to say that the selection within groups has been such that a representative sample is assured? For example in a quota survey of a group of people who are aged 65 or over, interviewers may restrict interviews to those who are around 65 or 66, neglecting those in their seventies, eighties or nineties.

In defence of the method, however, it is often claimed that interviewers do have instructions and constraints imposed on them so as to guard against selection biases, but this cannot always be assured.

One of the givens in sample surveys is non-response. Since follow-ups of non-respondents due to refusal or non-contact are generally avoided, sample results could be biased. The extent of that bias will naturally depend on the level of non-response and the differences in the characteristics of those who do respond and those who do not.

SUMMARY

By and large, the problems posed by non-probability sampling methods tend to outweigh the benefits. However, despite the statistical weaknesses of the methods described, non-probability sampling does have advantages in certain situations particularly in exploratory studies.

CHAPTER 6

ESTIMATION METHODS

Chapter 6

Estimation Methods

Once the selection of the sample has been carried out and once the interviewers are back in the office with the results of their fieldwork, the task still remains to relate the sample back to the population. This process is referred to as estimation. Essentially, estimation methods are used to draw conclusions about the population based on the information which has been gathered in the sample.

Not unexpectedly, there is a direct association between probability sampling methods and estimation procedures. In fact, the sample design itself determines the so-called weights or expansion factors which are used to produce the estimates. (An example is provided in Appendix D.)

6.1 Sampling Weights

At the basis of estimation procedures is the *sampling weight* of a unit. The sampling weight for a unit corresponds to the inverse of the probability of selection of the unit in the sample. More simply put, it indicates the number of units in the population that are represented by a unit in the sample.

Consider the farm survey. In the case of a simple random sample of 9,000 farms selected from a population of 153,000 farms, the probability of selection of all sample farms is the same, or $\frac{9{,}000}{153{,}000} = \frac{1}{17}$. In this case the sampling weight is 17 or the inverse of the probability of selection. Thus, since every farm represents 17 other farms in the population, if each one is reproduced 17 times, one will have expanded or weighted the sample of 9,000 'back-up' to the population of 153,000.

Suppose, for example, that an estimate of the total number of units in the population possessing a certain characteristic is required. In the case of the farm survey, this might be all the farms which produce wheat. In this case, one would obtain the estimate simply by counting the number of units which have this characteristic.

However, instead of actually reproducing the information, one can physically attach a weight to records of each of the farms in the sample. In this way, estimates can be produced in an operationally convenient way.[1]

Suppose now, that one sets out to estimate:

(a) the number of farms whose expenditures exceeded $10,000;
(b) the average farm expenditure in the population; and
(c) the average farm expenditure for those farms which produce wheat.

In example (a), farm records are sorted out according to whether or not the presence of the characteristic in question (in this case $10,000 or more) is indicated and then the weights of the corresponding farm records are aggregated or added up. The total of these weights indicates the number of farms in the general population whose expenditures exceed $10,000. In example (b), the product of the weight and the farm expenditure for each farm record is calculated and then aggregated over all farm records. This value is then divided by the total number of farms in the population. Finally, in example (c) once the records for the farms that produce wheat are ferreted out, the product of the weight and the farm expenditure for each farm in this group is calculated and then aggregated over all farm records. This value is then divided by the sum of the weights of records (for the wheat farms) to obtain the average for the group.

Sometimes, sample designs are *self-weighting*. This occurs when sampling weights are the same for all units in the sample. Such designs are time-saving and operationally convenient, particularly in large samples. For self-weighting designs, it is not necessary to actually attach a sampling weight to each record. In fact, the sampling weight can be ignored altogether in the production of statistics such as proportions and averages. The production of totals simply requires the sample total to be inflated by the sampling weight (inverse sampling ratio).

There are a number of points to be considered in the development of a self-weighting design. In the case of single-stage sample designs, units must be selected according to an equal probability selection scheme. In the case of stratified sample designs, it is necessary to allocate the sample size proportionally to the size of the strata in order to keep the sampling ratio for the strata the same as that for the overall population.

1. Such estimates are referred to as Horvitz-Thompson estimates. The Horvitz-Thompson estimator was developed in 1952. It estimates population totals when sampling is without replacement, from a finite population and when unequal probabilities of selection are used. The estimator is unbiased, linear and can be used with a variety of basic sample designs. Reference D.G. Horvitz, and D.J. Thompson, (1952).

Finally, in the case of a multi-stage design (in which the overall sampling rate of the population is fixed in advance) units must be selected with probability proportional to size at all stages except the final one. The units at the last stage are then selected with equal probability. PPS selection techniques are often used in multi-stage designs since they lead to a self-weighting sample with control of the sampling rate of the population, (See Appendix B).

In some cases, it may not be desirable to make an entire sample self-weighting. In the case of a national survey, for example, the sample size required to produce sufficiently reliable estimates for smaller regions (provinces) may be larger than that resulting from proportional allocation of the national sample to the regions (provinces).[2]

The use of *sampling weights* to produce estimates of population characteristics has been the topic of discussion thus far. Now, the focus will be on a brief overview of some techniques for handling the problem of information which may be completely or partially unavailable for some units in the sample.

6.2 Non-Response

Non-response refers to a situation where information from sampling units is unavailable for one reason or another. All surveys suffer from this problem. Generally, the extent of the non-response is directly contingent on the subject matter as well as the methods of data collection and data processing. With respect to field operations, for example, the use of skilled interviewers and adequate follow-up procedures all have a direct impact on the quality of information which is collected and the resulting level of non-response.

Yet, regardless of how successful the field and data processing operations are, there is a point beyond which non-response cannot be further reduced at reasonable cost within reasonable time.

There are numerous methods for dealing with complete or partial (item) non-response.[3] The suitability of any of the methods is dependent upon the type of survey and the nature of the non-response. Some of the methods are, adjustment of sampling weights of respondents, similar record substitution, sub-sampling of non-respondents and collecting data using more effective data collection methods or imputation.[4] In the following chapter, the effect of non-response is further considered in the context of determining sample size.

6.3 Use of Auxiliary Information

An estimation procedure may also be designed to incorporate external independent sources of information (if available) to increase the reliability of sample estimates.

The Statistics Canada's provincial population projections of the number of persons classified by age and sex group is an external data source which can be used in a national or provincial survey of the general population. Such projections might be used in surveys where the characteristics measured are highly correlated with age and sex.

Ratio estimation is a method often used in surveys for incorporating such relevant information. This method incorporates auxiliary information through weight adjustments. When producing estimates for many characteristics, it is operationally convenient to attach a permanent weight to each record as allowed for by this method and use these weights for the production of all survey estimates.

Ratio estimation works in the following way. The weights of the records in each classification of the population (eg age-sex groups) are adjusted by a multiplying factor. This factor is the ratio of the external data value and the sample estimate for each classification. When the weights are adjusted in this way, the estimate agrees with the external value for each classification.

This method requires (1) accurate external sources of information concerning the population and (2) collection of corresponding information for the sample. It is important that the external data source pertain to the same population and be based upon comparable concepts, definitions, reference periods, etc. as that of the survey. In the farm expenditure example, accurate information might be available elsewhere on total farm sales for the population. To the extent that total farm sales and expenditures are correlated, and provided that information on farm sales is available for the sample, the reliability of estimates of farm expenditure may be improved through ratio estimation.

As has been earlier indicated, relevant information about a population can be used in a number of ways to increase the efficiency of sample estimates. In the case of stratification, for example, one is looking for relevant information by which the population may be divided into different groups prior to sample selection. The information used for stratification purposes must be available for all units of a frame. At the time of sample selection, one might use information on the size measures of units which, again, must be available for all units of the frame. At the time of estimation, one might use relevant information available from the sample in combination with information available from external sources to improve efficiency.

2. Refer to section 7.2 Factors Which Affect Precision.

3. Information on these methods can be found in papers by Fellegi I.P. and Holt D. and by Platek R. and Gray G.B. (See bibliography for complete references.)

4. Imputation is a procedure of completing a response by using values from one or more records on the same file or from external sources. (e.g., historical data on respondents, administrative sources, etc.)

CHAPTER 7

DETERMINING THE SAMPLE SIZE

Chapter 7

Determining the Sample Size

One of the first considerations in the planning of a sample survey is the size of the sample. Since every survey is different, there can be no hard and fast rules for determining size.

Generally, the factors which decide the scale of the survey operations have to do with cost, time, operational constraints and the desired precision of the results. Once these points have been appraised and individually assessed, the investigators are in a better position to decide the size of the sample.

7.1 Desired Precision of Sample Estimates

One of the major considerations in deciding sample size has to do with the level of error that one deems tolerable and acceptable.

It has already been explained that measures of sampling error such as standard error or coefficient of variation are frequently used to indicate the precision of sample estimates. Since it is desirable to have high levels of precision, it is also desirable to have large sample sizes, since the larger the sample, the more precise estimates will be.

The sample size can be determined by specifying the precision required for each major finding to be produced from the survey, and the level of disaggregation to which the precision must apply.

Often estimates are required not only on a global basis, but for sub-populations as well. Such sub-populations might be defined in terms of age/sex groups or geographic areas. The sample size falling into each sub-population should be large enough to enable estimates to be produced at specified levels of precision (see Chapter 8). Sometimes, it will simply cost too much to take the size of sample required to achieve a certain level of precision. In this case, decisions must be made on whether to relax precision levels, reduce data requirements, increase the budget, or find other areas of the survey where cost cutting can be carried out.

7.2 Factors Which Affect Precision

In any decision related to the precision expected of the sample survey, a number of factors must be taken into account. Such elements as population size, variability of characteristics in the population and the sample plan itself will all affect the precision of the estimates. Consequently, all these factors are identified in the statistical formulae which ultimately relate sample size to the desired level of precision. In the following sub-sections, these factors are considered individually.

7.2.1 Size of the Population

Contrary to popular belief, the size of a sample does not increase in proportion to the size of the population. In fact the population size plays only a moderate role as far as medium-sized populations are concerned and an almost non-existent role as far as large populations are concerned.

Consider, for example, a simple random sample of 500 from a population of 200,000. Those 500 units will provide, for most practical purposes,the same precision as a simple random sample of 500 from a population of 10,000.

For very small populations, the relationship is more direct, and often more substantial proportions of the population must be surveyed in order to achieve the desired precision. In some cases, it is more prudent to consider taking a census rather than a sample.

7.2.2 Variability of Characteristics in the Population

Since the magnitude of differences between members of a population with respect to characteristics of interest is not generally known in advance, it must often be approximated, on the basis of previous surveys or pilot test results.

In general, the greater the difference between population units, the larger the sample size required to achieve specific levels of reliability.

7.2.3 Sample Plan

Many surveys involve moderately complex or very complex sampling and estimation procedures. A more complex design such as stratified multi-stage sampling with ratio estimation can often lead to higher variance in resulting estimates than might a simple random sample design. If, then, the same degree of precision is desired, it is necessary to inflate the sample size to take account of the fact that simple random sampling is not being used. This is often done by the use of a factor known as a 'design effect' in the calculation of sample size. Design effect refers to the ratio of the variance of the estimate for a particular design to the variance of the estimate for a simple random sample of the same size. The value of the design effect depends upon the sample plan, as well as the characteristics being measured. It can be estimated from similar past surveys, pilot surveys or using conservative judgment.

7.2.4 Non-Response

Non-response can occur for many reasons. Sometimes, members of a population being surveyed may not be available. Sometimes, they may refuse to answer questionnaires or take part in interviews. It is rare, indeed, when a 100% response rate is achieved. If non-response is not taken into account, the effective number of units in the sample will be smaller than expected. Consequently the precision of the estimates produced will also be lowered.

To overcome this, the sample size is sometimes inflated at the design stage to account for an anticipated rate of non-response. While this procedure is effective in reducing the variance, it does not reduce the bias resulting from the non-response. In fact, the magnitude of the bias is a function of the size of the non-response and the difference in characteristics between respondents and non-respondents. Since, however, there is a point beyond which non-response cannot be further reduced without an unreasonable expenditure of time and money, compensation should be considered at the time of estimation (e.g. adjustment of sampling weight of respondents).

Unfortunately, it is not often possible to know in advance what the non-response rate will be. This is especially true of surveys that are breaking new ground. In some instances, the response rate can be estimated with the help of a pilot survey or from past experience with similar surveys.

7.3 Cost and Time

It is a rare world in which considerations of time and cost are not paramount and most survey takers are not exempt from such restrictions. It almost goes without saying that the time and cost involved have a very definite effect on the size of a sample.

In many studies, funds are allocated and time deadlines set even before the specifics of the study have been decided. It may turn out that the sample size required to implement a survey is larger than existing funds can accommodate. In this case, if more money cannot be found, obviously the sample size must be reduced, thus lowering the precision of the estimates. The same is true for time considerations. If the time allowed is simply not sufficient, the size and scale of the sample may have to be limited to accommodate the deadlines.

7.4 Operational Constraints

Since surveys require properly trained field staff, coding and editing staff, as well as processing facilities, any limitations on these resources will mean that the size of the sample must be reduced.

In practice, the sample size is evaluated in terms of data requirements, precision, cost, time and operational feasibility. Such an exercise often results in a re-examination and possible modification of the original objectives, data requirements, levels of precision and elements of the survey plan. The survey designer in the process of interrelating such factors will generally attempt to develop a number of feasible design options for consideration and choose the one that best meets all these often conflicting requirements and constraints.

CHAPTER 8

SPECIAL CONSIDERATIONS ON SAMPLING AND ESTIMATION

Chapter 8

Special Considerations on Sampling and Estimation

The basic elements of a sample plan have been covered in previous sections. Some important topics in sample design and estimation will be explained next.

8.1 Domain Estimation

Mention has already been made of sub-populations in survey sampling. These sub-populations, which can be characterized on the basis of age and sex or industry and occupation, are sometimes referred to as *domains*. It is important to have a large enough sample in each domain to produce estimates with sufficient reliability.

It may also be desirable to turn the domains into separate strata whenever possible. Since sampling is carried out independently for each stratum (domain), the sample size can be controlled for each domain. In fact, disproportionate stratified sampling is frequently used to ensure an adequate representation of domains. With human populations, for example, estimates of characteristics relating to small domains of the population (such as those with income exceeding $100,000) might require that the sampling fractions for such groups be increased to achieve adequate levels of precision.

If it is not possible to turn domains into separate strata, a uniform sampling fraction must be applied to several domains. Smaller domains, therefore, receive a smaller proportion of the total sample than do larger domains. The sampling fraction required to estimate a characteristic with sufficient reliability in all the domains could be very large particularly if small domains are to be adequately represented.

8.2 Problem of Large Units

It is frequently the case that a few large cities (in surveys of human populations) or a few large units (in institutional and business surveys) can exert a great and often troublesome influence in sampling and estimation.

The proper representation of large cities for example, can be ensured by making such cities separate strata and sampling each of them separately.

Large business establishments and institutions often have extremely large values for characteristics typically measured in a business or institutional survey. If special provisions are not made in the sample plan for such units, estimates of characteristics would tend to be too large if such units are overrepresented (or too small if such units are underrepresented in the sample). To overcome this problem, business establishments (or institutions) could be stratified according to size. The largest units may be selected on a 100% basis, while smaller units may be sampled.

8.3 Multi-Purpose Surveys

In most cases, survey objectives call for the measurement of many characteristics. In the survey on farm expenditures for example, one might want to determine much more than the overall costs of running a farm. One could also be interested in the cost of farm machinery, wages, loans, pesticides, seeds, etc.

It is, however, a fact that design decisions are often compromises made to satisfy data requirements.

For example, to address the problem of sample size determination, sample sizes required for major items of interest may have to be examined carefully in light of the precision allowed or tolerated.

It might happen that the largest of the values meets the requirements of all the major items but is not feasible in terms of available time or money. It may also happen that the measurement of certain items calls for the sampling plans that are radically different from one another. It may be necessary then to re-examine the objectives and data requirements with a view to dropping certain characteristics or accepting lower levels of precision.

To accommodate the measurement of several items within one survey plan, it might be necessary to make compromises in many areas of the survey design. The method of data collection (telephone, personal interview, mail-out/mail-back) chosen may be suitable for measurement of some characteristics but not suitable for others. The survey design must be made to properly balance statistical efficiency, time, cost and other operational constraints.

8.4 One-time vs. Continuing Surveys

One-time surveys differ from *periodic* or *continuing surveys* in many ways. The aim of periodic or continuing surveys is often to study trends or changes in the characteristics of interest over a period of time.

With panel studies, data are collected from the same sample on several occasions. Such studies nearly always measure changes in the characteristics of a population with greater precision than do a series of independent samples of the same size.

Administratively, panels have the advantage. Overhead costs of survey development and sample selection can be spread over many surveys and this in turn cuts the cost of the fieldwork. The main problems associated with panels are changes in the size and structure of the population which over time are not reflected in the sample, sample mortality, conditioning effects over a long period of time, and respondent burden. One design which is intermediate between independent samples on successive occasions and the panel sample method takes the form of partial replacement of the sample over time.

One such example is the Canadian Labour Force Survey (LFS), the largest continuing survey conducted by Statistics Canada. The LFS was established in 1945. It was originally designed to produce quarterly estimates, but since 1952, has been conducted on a monthly basis. The survey employs a rotation design in which households are included in the sample for six consecutive months. Every month 1/6 of the sample is replaced by new households.

This design offers the advantage of measuring monthly changes with greater precision, less cost and with less disruption to the field operations than would be the case if independent samples were used, and also reduces the problem of respondent burden associated with panel studies. To reflect changes in the size and structure of the population and data requirements the Labour Force Survey undergoes periodic redesigns, usually in the wake of the decennial census.

Decisions made in the sample design of periodic or continuing surveys should take into account the possibility of deterioration in design efficiency over time. Designers may elect, for example, to use stratification variables that are more stable, avoiding those that may be more efficient in the short run but which change rapidly over time.

In Chapter 4, it was indicated that the selection of units with probability proportional to size required that the size measures of units be accurate for the selection method to be efficient. In the case of continuing multi-stage surveys, it was also pointed out that higher-stage units tend to be subject to less change over time than lower-stage units. For multi-stage surveys which use PPS selection methods in the design, a gradual deterioration of the design efficiency can be offset by ensuring that the selection of lower-stage units be based on current field counts and listings and by using PPS selection methods which facilitate periodic updating of higher-stage units.[1] The Rao-Hartley-Cochran random group method for example, facilitates the process of updating.

For single-stage continuing surveys, size measures refer to the largeness or magnitude of units. If such measures are subject to frequent change, it might be preferable to take size into account through stratification by broad size groups and/or in the estimation procedure rather than using these measures in the sample selection procedure.

Another feature of a periodic or continuing survey is that, in general, a great deal of information is available which is useful for design purposes. The adequacy of various features of the sample design such as the appropriateness of stratification variables and boundaries, method of sample allocation, size of units at various stages of a multi-stage design, etc., may be studied over time with a view to increasing design efficiency. Often, information required to efficiently design a one-time survey is very limited.

In the design of a continuing survey, provisions must be made to accommodate such events as births, deaths and changes in size measures. The sampling and estimation methods used in continuing surveys should incorporate these changes in a statistically efficient way with as little disruption as possible to the ongoing survey operations.

1. N. Keyfitz, "Sampling with probabilities proportional to size: adjustment for changes in the probabilities", Journal of the American Statistical Association, Vol. 46 (1951) pp. 105-109. (See also J.D. Drew, G.H. Choudhry, G.B. Gray, "Some methods for updating sample survey frames and their effects on estimation" (unpublished Statistics Canada in-house paper, 1978.))

CHAPTER 9

PERSPECTIVES ON SAMPLING: BEYOND THE SAMPLE DESIGN

Chapter 9

Perspectives on Sampling: Beyond the Sample Design

So far, the concern has been primarily on matters concerning the development of a sample plan. For those who have been involved in survey taking, one thing should be apparent. There is far more to the design of a sample survey than understanding the differences between a cluster or multi-stage sample or the relative merits of probability and non-probability sampling. While it is essential to understand and appropriately apply the various techniques, it is also important to have a perspective on the role of sampling in the design of a survey. Such is the purpose of the present chapter.

To understand where a sample plan fits into an overall survey, it is first useful to outline the general steps of a survey and some of the important functions they entail. The major stages of a survey can thus be listed as initial feasibility assessment, development, implementation, data analysis and evaluation.

9.1 Feasibility Assessment

It should be recognized that surveys, whether they be based upon censuses or samples, are not the only and often not the most suitable means for satisfying information requirements. The use of available information including administrative sources, controlled experiments, and statistical models are other approaches to be considered depending, of course, on the specific nature of the research problem.

The *feasibility assessment* attempts to clarify the nature of the research problem and to indicate whether a survey is indeed the best way to meet the needs of the problem. It may also include descriptions of one or more feasible approaches. More specifically, the feasibility assessment starts out by defining and clarifying survey objectives, defining and operationalizing survey concepts and establishing a set of data requirements to be met by the survey.

9.1.1 Conceptualization

There is an amusing illustration which goes a long way toward summing up some of the difficulties encountered during the conceptualization process. The illustration,[1] which was developed by Ackoff and Pritzer, focuses on a survey designed to determine the number of chairs people have in their living rooms. One of the first tasks the survey designers face, of course, is to define what is meant by a chair. Yet, once the chair has been defined in terms of, say, its height and weight and shape, the problem of conceptualization has still not been altogether solved. What about the chairs that are built into the wall, or those chairs that are without legs?

Although the illustration does tend to exaggerate the difficulty of the conceptualization process, it does indicate an important point. If the concepts of a survey study are not properly defined, the utility of the survey can be seriously jeopardized.

There is generally considerable difficulty in arriving at a conceptualization of the subject under study. The Ackoff and Pritzer illustration considers a fairly obvious example, but even in something that might first appear as self-evident as the survey on farm expenditures, it must first be determined what exactly is meant by a farm. After all, some people may call their country residences farms, yet the only livestock they have are cats and dogs and the only crops they grow are corn and potatoes.

Once the business of conceptualization has been accomplished, the feasibility assessment can then proceed with the development of one or more possible methodological frameworks for the survey. Included within such frameworks would be sample design and estimation methods, data collection methods, a data processing strategy and a data analysis and evaluation framework. The assessment would further indicate the specific data requirements which could or could not be met, the error levels to be expected, and the impact of these on meeting survey

1. Herbert H. Hyman, *The Major Types of Surveys*, ed. by Bernard Berelson and Morris Janowitz, from the Reader in Public Opinion and Communication (The Free Press, New York, 1966), p. 624.

objectives. The assessment would also attempt to provide an estimate of time and cost associated with the options considered. Finally a feasibility study would indicate the need, if any, for testing of new methods or procedures within the broad methodological framework established for the conduct of the survey.

9.2 Survey Development

Having established a broad methodological framework, detailed work on the various elements of a survey can now be carried out in what is referred to as the development stage. One of the major activities at this time is the design of the questionnaire. The problems faced here are how to best word and arrange questions in a manner consistent with the method of enumeration so as to yield the information required. The intention, of course, is to obtain information with a minimum level of error and minimal inconvenience to those participating in the survey and in a form suitable for subsequent processing. While there are well established principles for questionnaire design[2], crafting a good questionnaire essentially remains an art requiring ingenuity and experience on the part of the survey designer. If the data requirements are not properly transformed into a structured data collection instrument of high quality, a 'good' sample will yield 'bad' results.

Indeed, it is at this stage where any required pretests or pilot studies are carried out to assess, for example, the adequacy of the questionnaire or suitability of a sampling frame. All field materials (interviewer training and instruction manuals, sample control documents) are prepared for the data collection stage. Sample selection and estimation procedures within the sample plan are finalized in the form of specifications. Specifications for coding[3], data capture[4], systems analysis and computer program development are all carried out to set the stage for data processing. Quality assurance and control programs are also developed at this stage.

9.2.1 Quality Assurance and Control Programs

In a perfect world in which it were possible to select a perfect sample and design a perfect questionnaire, it would as well be possible to have perfect interviewers gather perfect information from perfect respondents. No mistakes would be made in the recording of information nor in its conversion into a form which could be processed by computer. In such a perfect world, of course, there would be no need for a survey since all information would be known!

Experience with even the most straightforward survey problem quickly shatters such an illusion of survey taking and replaces it with the well known law: 'whatever can go wrong will go wrong'. As indicated in earlier chapters, there are a host of problems which, if not anticipated and controlled, can introduce errors to the point of rendering survey results useless. Quality assurance programs such as interviewer training, spot checking of data collection and outputs of other major survey activities, provisions for follow-up of non-respondents, editing of information and testing of computer programs are required to minimize non-sampling errors introduced at various stages of a survey. Statistical quality control programs ensure error levels, introduced as a result of a survey operation, are controlled to within specified levels, with minimum inspection[5].

At Statistics Canada, a procedure known as acceptance sampling[6] is used in quality control operations. This approach controls the error levels resulting from survey operations such as coding and data capture.

In order to minimize and control the errors which can be introduced at various stages of a survey, it is a good practice to devote a part of the overall survey budget to quality assurance and control programs.

2. A.N. Oppenheim, *Questionnaire Design and Attitude Measurement* London: Basic Books New York. 1966. S.L. Payne, *The Art of Asking Questions* Princeton N.J., Princeton University Press 1951.

3. Coding can be described as the process of assigning a numerical value to items of information to enable such information to be processed by computer. Coding systems widely used by Statistics Canada are Standard Industrial Classification, Standard Geographical Classification and the Standard Occupational Classification.

4. Data capture involves converting the information into a computer readable form for data processing.

5. For further information on quality control, see also Harold F. Dodge and Harry G. Romig, *Sampling Inspection Tables Single and Double Sampling*. New York: John Wiley and Sons Inc., 1959.

6. Acceptance sampling involves dividing work into units called batches, selecting and checking a sample from each batch and accepting or rejecting the batch depending on the extent of errors encountered in the sample. The remainder of rejected batches is completely inspected.

9.3 Survey Implementation

Having made sure that all systems are in place the count-down begins 10, 9,.....1, 0 (i.e., funding request is approved) – button is pressed and the survey is launched.

All questionnaires and survey control forms and manuals are printed. Interviewers are trained, the sample is selected and information is collected, all in a manner established during the development stage. Following these activities information is coded (generally manually), captured in computer readable form and processed according to specifications. Processing activities include the editing[7] of survey data (followed by the application of corrective actions), and the application of weights to survey records in compliance with estimation specifications. The result is a well-structured and 'clean' data set from which it is possible to produce required tabulations of survey results.

9.4 Analysis and Evaluation

After the production and analysis of descriptive statistics, any required data analysis is carried out in compliance with survey objectives. Data analysis basically involves the application of statistical methods to data sets to test statistical hypotheses, explore relationships among characteristics of interest, and carry out time series studies, etc.

As an aid to data users, it is good practice to provide indicators to evaluate data quality. For example, measures such as the standard error or coefficient of variation should accompany major survey results to give the users an indication of the extent of sampling error. Other indicators such as non-response rates, error rates encountered during the implementation of major survey activities such as coding also provide the data user with some idea of the data quality in order that information be used in an appropriate way.

It is also good practice to evaluate the efficiency and cost of major survey operations, particularly in the case of continuing surveys, so that improvements in their design and implementation can be made over time.

Summary

Despite the complexities and costs associated with survey studies, the point is that sample surveys answer pressing and urgent needs. The power of the survey, which first emerged in the 17th century as a form of crude state counting has become a science of information gathering, indispensable to the efficient operation of both government and industry. At Statistics Canada, an information base is provided for many areas both in the economic and social spheres. Indeed, the samples used to track these spheres have become a vital tool in the planning, decision-making and program evaluation that government, business, social agencies and other organizations undertake. It should therefore be clear that any effort to underscore some of the inherent complexities of survey sampling is only to underscore the importance of this tool in a world increasingly in need of detailed and accurate information.

7. Editing is the process of checking survey data for obvious or suspected errors and can involve manual interface.

BIBLIOGRAPHY

Babbie, E.R. *Survey Research Methods*. Belmont, California: Wadsworth Publishing Company Inc., 1973.

Cochran, W.G. *Sampling Techniques*. New York: John Wiley and Sons, Inc., 1963.

Connor, W.S. "An exact formula for the probability that two specified sample units will occur in a sample drawn with unequal probabilities and without replacement", *Journal of the American Statistical Association*, Volume 61 (1966), pp. 384-390.

Deming, W.E. *Sample Design in Business Research*. New York: John Wiley and Sons, Inc., 1960.

Dodge, Harold F.; Romig G. Harry. *Sampling Inspection Tables Single and Double Sampling*. New York: John Wiley and Sons, Inc., 1959.

Dominion Bureau of Statistics. *History, Function, Organization*.
Ottawa: Queen's Printer, 1952.

Fellegi, I.P. "Sampling with Varying Probabilities Without Replacement Rotating and Non-Rotating Samples". *Journal of the American Statistical Association*, Volume 58 (1963), pp. 183-201.

Fellegi, I.P. and Holt, D. "A Systematic Approach to Automatic Edit and Imputation", *Journal of the American Statistical Association*, Volume 71 (1976), pp. 17-35.

Ferber, R. ed. *Handbook of Marketing Research*. New York: McGraw Hill, 1974.

Gray, G.B. "Joint probabilities of selection of units in systematic samples," Proc. Amer. Stat. Association, (1971) pp. 271-276.

Hansen, M.H.; Horvitz, N.N.; Madow, W.G. *Sample Survey Methods and Theory*. Volume 1 of 2. New York: John Wiley and Sons, Inc., 1953.

Hartley, H.O. "Multiple Frame Surveys" Proc. Amer. Stat. Association, (1962) pp. 203-206.

Hartley, H.O. "Multiple Frame Methodology and Selected Application" *Sankhya, C36* (1974) pp. 99-118.

Hidiroglou, M.A.; Gray G.B. "Construction of joint probabilities of selection for systematic p.p.s. sampling," *Applied Statistics*, Volume 29, No. 1, (1980), pp. 107-109.

Horvitz, D.G.; Thompson, D.J. "A Generalization of Sampling Without Replacement from a Finite Universe". *Journal of the American Statistical Association*, Volume 47 (1952), pp. 663-685.

Hyman, Herbert H. "The Major Types of Surveys." *Reader in Public Opinion and Communication*. Edited by Bernard Berelson and Morris Janowitz. New York: The Free Press, 1966.

Keyfitz, N. "Estimates of Sampling Variance Where Two Units are Selected From Each Stratum". *Journal of the American Statistical Association*, Volume 52 (December, 1957), pp. 503-510.

Keyfitz, N. "Sampling with probabilities proportional to size: adjustment for changes in the probabilities", *Journal of the American Statistical Association*, Volume 46 (1951), pp. 105-109.

Kempthorne, O.; Folks, L. *Probability, Statistics and Data Analysis*. Ames, Iowa: Iowa State University Press, 1971.

Kish, L. *Survey Sampling*. New York: John Wiley and Sons Inc., 1965.

Moser, C.A.; Kalton, G. *Survey Methods in Social Investigation*. New York: Basic Book Inc., Publishers, 1972.

Murthy, M.N. "*Sampling Theory and Methods*." Calcutta Statistical Publishing Society, 1967.

Platek, R. and Gray, G.B. "Non-Response and Imputation". *Survey Methodology Journal*, Volume 4 No. 2 (1978), pp. 144-177.

Raj, D. *Sampling Theory*. New York: McGraw-Hill, 1968.

Rao, J.N.K.; Hartley, H.O.; Cochran, W.G. "On a Simple Procedure of Unequal Probability Sampling Without Replacement" *Journal of Royal Statistical Society*. Series B, Volume 27: (1962), pp. 482-490.

Slonim, Morris James. *Sampling in a Nutshell*. New York: Simon and Shuster, 1960.

Snedecor, G.E.; Cochran, W.G. *Statistical Methods*. Iowa: Iowa State University Press, 1967.

Stuart, Alan. *Basic Ideas of Scientific Sampling*. 2nd ed. London: Charles Griffin and Co. Ltd., 1976.

Sukhatme, P.V.; Sukhatme, B.V. *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State University Press, 1970.

Wheeler, Michael. Lies, Damn Lies and Statistics; *The Manipulation of Public Opinion in America*. Liverwright: New York, 1967.

Wordsnorth Communication Services Ltd., "1981 Census Communication Program Appraisal". Winnipeg, 1981.

Yates, F. *Sampling Methods for Censuses and Surveys*. London: Charles Griffin and Company, London, 1960.

Yates, F.; Grundy, P.M. "Selection Without Replacement from Within Strata with Probability Proportional to Size". *Journal of the Royal Statistical Society*, Series B. Volume 15 (1953), pp. 235-261.

APPENDICES

The appendices contain algebraic expressions designed to serve as a background to the material presented in this manual. References made to estimates apply to 'simple' estimates, not the more complex ratio or regression estimates mentioned in Chapter 6 which are beyond the scope of this manual.

Appendix A

Notation and Algebraic Expressions[1]

1. Notation

N: Number of units in the population

n: Number of units in the sample

$f = \frac{n}{N}$: Sampling fraction of the population

$1 - f$: Finite population correction factor

y_i: Value of characteristic 'y' for the i^{th} unit

X_i: Measure of size of i^{th} unit

$Y = \sum_{i=1}^{N} y_i = y_1 + y_2 + \ldots + y_N$: Total value of characteristic y in the population

$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$: Mean value of characteristic y in the population

$N_y = \sum_{i=1}^{N} y_i$: Number of units in the population which have attribute y

: $y_i = 1$ if unit i has attribute y

: $= 0$ if unit i does not have attribute y

$P = \frac{N_y}{N}$: Proportion of units in the population which have attribute y

$Q = 1 - P$: Proportion of units in the population which do not have attribute y

$S^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{N - 1}$: Variance of characteristic y in the population

$N_h, n_h, f_h, Y_h, \bar{Y}_h, N_{yh}, P_h, Q_h, S^2_h$: above quantities for stratum 'h'

$X_h: \sum_{i=1}^{N_h} X_i$: size of stratum h

1. The algebraic expressions used in these appendices are taken from W.G. Cochran, *Sampling Techniques*. (See Bibliography)

1.1 Sample Estimates of Mean, Total and Proportion for a Simple Random Sample

Estimate of Mean $\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{y_1 + y_2 + \ldots + y_n}{n}$

Estimate of Total $N\bar{y} = \frac{N}{n} \sum_{i=1}^{n} y_i$

Estimate of Proportion $p = \frac{n_y}{n}$

Where n_y is the number of units in the sample which have attribute y.

1.2 Variance, Standard Error, Coefficient of Variation and Confidence Interval for Estimates in Appendix A, 1.1

Table A-1: Algebraic Expression for Variance, Standard Error, Coefficient of Variation and Confidence Interval for a Simple Random Sample

Estimate	Variance	Standard Error	Coefficient of Variation	Confidence Interval (1-∝) %
$\bar{y}$	$\frac{(1-f)S^2}{n}$	$\sqrt{1-f}\,\frac{S}{\sqrt{n}}$	$\sqrt{1-f}\,\frac{S}{\sqrt{n}\,\bar{y}}$	$\bar{y} \pm t_{\alpha/2}\sqrt{1-f}\,\frac{S}{\sqrt{n}}$
$N\bar{y}$	$\frac{N^2(1-f)S^2}{n}$	$N\sqrt{1-f}\,\frac{S}{\sqrt{n}}$	$\sqrt{1-f}\,\frac{S}{\sqrt{n}\,\bar{y}}$	$N\bar{y} \pm t_{\alpha/2} N\sqrt{1-f}\,\frac{S}{\sqrt{n}}$
p	$\frac{(1-f)PQ}{n-1}$	$\sqrt{\frac{(1-f)PQ}{n-1}}$	$\sqrt{\frac{(1-f)Q}{P(n-1)}}$	$p \pm t_{\alpha/2}\sqrt{\frac{(1-f)PQ}{n-1}}$

Notes: (1) Estimates of the variance, standard error, etc., are obtained by replacing S^2, $\bar{Y}$, P and Q by the sample estimates s^2, $\bar{y}$, p, q, respectively where

* The factor $\frac{N}{N-1}$ is assumed to be close to 1 and has therefore been omitted.

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$$

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

$$p = \frac{n_h}{n}$$

$$q = 1 - p$$

(2) It is assumed in the confidence interval given above that the sample sizes are large enough for the so-called normal distribution to apply. The values $t_{\alpha/2}$ may be found in normal distribution tables. A 95% confidence interval, for example, is obtained by setting $\alpha = .05$ ($t_{\alpha/2} = t_{.025} = 1.96$).

1.3 Sample Estimates of Mean, Total and Proportion for a Stratified Sample and Variance of Estimates

Estimate of Mean $\sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h$ variance $\sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 V(\bar{y}_h)$

Estimate of Total $\sum_{h=1}^{L} N_h \bar{y}_h$ variance $\sum_{h=1}^{L} N_h^2 V(\bar{y}_h)$

Estimate of Proportion $\sum_{h=1}^{L} \frac{N_h}{N} P_h$ variance $\sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 V(P_h)$

Notes:

(1) $V(\bar{y}_h)$, $V(p_h)$ denote the variance of sample estimates of the mean or proportion for stratum h.

(2) In the case of stratified random sampling

$$V(\bar{y}_h) = (1 - f_h)\frac{S_h^2}{n_h}$$

$$V(p_h) = (1 - f_h)\frac{P_h Q_h}{n_h - 1}$$

(Estimates of variance may be obtained by replacing S_h^2, P_h, Q_h by sample estimates s_h^2, p_h, and q_h in stratum h.)

Allocation of Sample to Strata

Proportional Allocation: $n_h = \frac{nN_h}{N}$

X-Proportional Allocation: $n_h = \frac{nX_h}{\sum_{h=1}^{L} X_h}$

Neyman Allocation: $n_h = \frac{nN_h S_h}{\sum_{h=1}^{L} N_h S_h}$

Optimum Allocation: $n_h = \frac{nN_h S_h/\sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h/\sqrt{C_h}}$

1.4 Horvitz-Thompson Estimates

Estimates expressed as the weighted sum of values of individual units in the sample where the sampling weights are the inverse of the selection probabilities are called Horvitz-Thompson estimates. Such estimates, generally used in sample surveys, can be expressed as follows:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}$$

where π_i: probability of selecting unit i in the sample.

Notes:

(1) An estimate of the mean (proportion) can be obtained by dividing $\hat{Y}_{HT}$ by N.

(2) In the case of simple random or systematic sampling, $\pi_i = \frac{n}{N}$ and the H – T estimates reduce to those given in Appendix 1.1.

(3) In the case of stratified sampling, π_{ih} refers to the probability of selecting unit i in stratum h. The H – T estimates reduce to those given in Appendix 1.3 (e.g., for stratified random sampling $\pi_{ih} = \frac{n_h}{N_h}$).

(4) In the case of PPS sampling, $\pi_i = n\frac{X_i}{X}$ where X_i refers to the size of unit i and $X = \sum_{i=1}^{N} X_i$.

(5) In the case of cluster sampling, π_i denotes the probability of selecting the i^{th} *cluster* in the sample.

(6) In the case of multi-stage sampling, π_i is the product of the probabilities of selection of the sampling units at each stage of sampling as explained in Chapter 4.6.

(7) Expressions for the variance of H – T estimates, sample estimates of the variance and their properties are presented in numerous texts listed in the Bibliography.

1.5 Precision of Stratified Random, Cluster, and Multi-Stage Sampling Relative to Simple Random Sampling

1.5.1 Stratified Random Sampling

The precision of stratified random sampling depends upon the allocation of the overall sample to the strata. The variance of a mean for a stratified random sample is given below for a) Proportional Allocation (denoted by V_{prop}) and for b) Neyman Allocation (denoted by V_{ney}). This is compared to the corresponding variance for a simple random sample (denoted by V_{srs}).

Assumption: The sampling fraction for each stratum is negligible.

a) $$V_{srs} - V_{prop} = \frac{\sum_{h=1}^{L} N_h(\bar{y}_h - \bar{y})^2}{nN}$$

b) $$V_{srs} - V_{ney} = \frac{\sum_{h=1}^{L} N_h(\bar{y}_h - \bar{y})^2}{nN} + \frac{\sum_{h=1}^{L} N_h(S_h - \bar{S})^2}{nN}$$

Observations

1) The difference in variances in a) is attributed to differences in the stratum means. The larger the differences between the stratum means, the greater is the precision of stratified random sampling with proportional allocation relative to simple random sampling.

2) The differences in the variances in b) are attributed to the differences noted above (first term on the right hand side of the equation) as well as to the differences between the stratum standard deviations. The larger such differences are, the greater will be the precision of stratified random sampling with neyman allocation relative to simple random sampling.

1.5.2 Cluster Sampling[2]

To examine the precision of a cluster sample relative to a simple random sample, it will be assumed that a population of MN units consists of M clusters each of size N from which a simple random sample of m clusters is selected.

The sampling variance for a mean for such a cluster sample denoted by V_c is given by $V_c = V_{srs} [1 + (N-1)\varrho]$

Where V_{srs}: Variance of a simple random sample of mN units

ϱ : intra-class correlation coefficient for clusters.[3]

Observations

1) The variance of a cluster sample relative to a simple random sample depends upon the cluster size (N) and the intra-class correlation coefficient (ϱ).

2) If N = 1, (each cluster contains only 1 unit) then cluster sampling is equivalent to simple random sampling.

3) If $\varrho = 0$, then each cluster is as heterogeneous as the general population and $V_c = V_{srs}$.

4) If $\varrho > 0$, then $V_c > V_{srs}$, implying that for such a situation cluster sampling is less precise than simple random sampling.

5) If $\varrho < 0$, then $V_c < V_{srs}$ and, in this case, simple random sampling is less precise than cluster sampling.

2. C.A. Moser and G. Kalton, *Survey Methods in Social Investigation*. New York: Basic Books Inc., Publishers, 1972. pp. 79-110.

3. C.A. Moser and G. Kalton, *Survey Methods in Social Investigation*. New York: Basic Books Inc., Publishers, 1972. pp. 79-110.

1.5.3 Multi-Stage Sampling[4]

A 2-stage design will be used to examine the precision of a multi-stage sample relative to a simple random sample. To compare precision, the following assumptions are required for simplification:

1) m Primary Sampling Units (PSU's) are selected randomly with replacement from M PSU's each of size N.
2) Simple random samples of size n are selected from each PSU selected in the first stage.

The sampling variance of a mean for such a 2-stage sample is denoted by V_{mult} and is given by:

$$V_{mult} = V_{srs}\,[1 + (n-1)\varrho]$$

where

V_{srs}: variance of a simple random sample of mn units.

ϱ: intra-class correlation coefficient for PSU's.

Observations

1) The variance of a 2-stage sample relative to a simple random sample depends upon the sample size selected within PSU's (n) and the intra-class correlation coefficient (ϱ).
2) 2-Stage sampling is more efficient than cluster sampling, when $\varrho > 0$ and $n < N$. This follows from the fact that to achieve the same overall sample size, the sample is "spread" over more PSU's in a 2-stage design than clusters in a 1-stage design.

4. C.A. Moser and G. Kalton, *Survey Methods in Social Investigation*. New York: Basic Books Inc., Publishers, 1972. pp. 79-110.

Appendix B

Sampling with Probability Proportional to Size in the Context of a Self-Weighting Multi-Stage Sample Design

An operational advantage of PPS sampling in the context of multi-stage sampling is that a self-weighting design can be achieved while controlling the size of the sample. This will now be illustrated for a 2-stage design. Suppose a population of 800 units is grouped into 5 primary sampling units.

Example:

PSU	SIZE (N_i)
1	120
2	200
3	150
4	50
5	280
	N = 800

The overall probability of selecting a unit in the sample is the product of the probabilities of selection of the sampling units at each stage. Thus, for a 2-stage design, this probability is given by $P = P_1 \times P_2$ where P_1 is the probability of selection of units in the sample at the first stage and P_2 is the probability of selection of units in the sample at the second stage within selected first stage units.

In order for the design to be self-weighting, it is necessary that P be the same for all units selected in the sample. Suppose $P = 1/20$ in the example above. Suppose also that 2 units (say PSU's 1 and 5) are selected with equal probability (i.e., $P_1 = 2/5$). The probability (P_2) of selection of units at the second stage *within* selected PSU's 1 and 5 must be 1/8 in order that the overall probability (P) be 1/20 (= 2/5 × 1/8). There would be 15 = (1/8 × 120) units selected from PSU 1 while 35 = (1/8 × 280) would be selected from PSU 5 for a total of 50 units. Thus, the number of units selected in the sample depends upon the size of the PSU's selected at the first stage.

If, on the other hand, PSU's were selected with PPS, then the probability of selecting PSU 1 would be $2 \times \frac{120}{800}$ $(= 2\frac{N_1}{N})$ and PSU 5 would be $2 \times \frac{280}{800}$ $(= 2\frac{N_5}{N})$. In order that the overall probability P be 1/20, the probability of selection of units at the second stage would be 20/120 (= 1/6) for PSU 1 and 20/280 (= 1/14) for PSU 5. Thus, for units selected with equal probability *within* PSU 1, the overall probability is $\frac{240}{800} \times \frac{20}{120} = \frac{1}{20}$. Similarly, for units selected with equal probability within PSU 5 the overall probability is $\frac{560}{800} \times \frac{20}{280} = \frac{1}{20}$. Twenty units are required from each selected PSU (regardless of its size) for a total sample size of 40 units. Since the total sample size does not depend on which PSU's are selected at the first stage, the overall sample size can be controlled.[1] Also, since the same number of units are selected from each PSU regardless of size, interviewer assignment sizes can be readily controlled.

Thus, the sample size for a multi-stage self-weighting design can be controlled provided the units at all stages except the last stage are selected with probability proportional to size and units at the last stage are selected with equal probability.

1. It has been assumed above that the size measures of the selected PSU's upon which the selection is based correspond to actual up-to-date field counts for these PSU's. In practice, there would likely be some change in the size measures. Thus, if the actual size of PSU's 1 and 5 were 144 and 266, the second stage sampling fractions of 1/6 and 1/14 would yield 24 and 19 units.

Appendix C

Example of Sample Size Determination

The example describes how the sample size was determined for the Survey of 1976 Graduates of Post-Secondary Programs. The object of this survey was to study post-secondary graduates from 1976 in an effort to find out the kinds of jobs they had, if indeed they had found jobs. More formally stated, the specific objectives of the survey were to:

a) measure the current employment status of graduates;
b) identify the types of jobs they had;
c) total time they had spent jobless since graduation;
d) measure the difficulties they had finding work;
e) measure their annual salaries;
f) identify their future educational intentions.

The *population* for this survey consisted of all individuals enrolled in post-secondary institutions in Canada (except Quebec) who had completed course requirements for graduation in 1976.

The post-secondary institutions under consideration included universities, community colleges and hospital schools of nursing.

The *frame* was made up from files of the various institutions which listed all individuals enrolled in the academic year for 1976.

It was obvious that due to the detail and extent of the survey objectives, the *sample size* would have to be quite large. In fact, it was calculated to be 45,533, which was approximately 40% of the target population.

Data was collected by telephone. It was felt that, under the circumstances, it would be appropriate to take a stratified random sample.

To see how this sample size was determined, each of the factors affecting the sample size as described in Chapter 7.1 will be assessed.

Desired Precision of Sample Estimates

The estimates produced from the survey were to be within 15% of the true value at a 95% level of confidence.

That is, the true value of a characteristic X (where X is a mean, proportion or total) would belong in the interval from $\hat{X} - .15\hat{X}$ to $\hat{X} + .15\hat{X}$ (Where $\hat{X}$ is a sample estimate of X) with 95% level of confidence.

This degree of precision may equivalently be stated as follows:

1) the *standard error*[1] should not exceed .075X
2) the *variance* should not exceed $(.075X)^2$
3) the *coefficient* of variation should not exceed .075

Since this survey was being carried out for the first time, good estimates of X of S^2 were not available on the characteristics of interest. If a similar survey had been undertaken in the past, historical data might have been used in the formula to arrive at a sample size. A pilot study was at first considered but was not judged to be appropriate for this study primarily because of the timeliness of the final results and cost considerations. Also, since a pilot study would necessarily be limited in scope, estimates of X and S^2 for a number of different characteristics would tend to be imprecise.

Since most of the important estimates were proportions of graduates having characteristics of interest, the precision statements were specified in terms of proportions.[2] The variance of an estimated proportion for a simple random sample given in table A-1 on page 55, was used to algebraically express the precision as follows:

$$V(p) = B(1 - \frac{n}{N})\, P\, \frac{(1-P)}{n-1} = (.075P)^2$$

Since in this survey a stratified random sample was selected B should be less or equal to 1.[3] In fact if the stratification were efficient B should have a value less than one. A conservative value of B = 1 was used in the formula since no historical information was available to estimate B.

1 Assuming the normal distribution holds, the range of values is determined to be approximately two standard errors above and below the estimate in the case of a 95% confidence interval.

2 In the case of proportions, X = P and $S^2 = P(1 - P)$.

3 B is the ratio of the variance of an estimate for the sample design actually used to the variance of the estimate for a simple random sample.

Solving the formula for n gives the following result.[4]

$$n = \frac{N(1 - P)}{N(.075)^2 P + (1 - P)}$$

The reader may verify that as the value of P decreases the value of n increases. The smallest value of P for which the required precision was to apply was P = .05.

Thus setting P = .05 and N = 101,118 in the expression above, the following result is obtained

$$n = 3268$$

Because the population size N is so large the finite population correction factor (f.p.c.) $(1 - \frac{n}{N})$ had a small effect on the sample size.

In fact, if the f.p.c. had been ignored the result would have been

$$n_0 = 3378$$

Finally, the anticipated rate of non-response (1 − r) is arrived at in the following way.

The overall rate of complete non-response experienced in the 1974 Ontario Graduate Survey was .27. It is felt that, although the method of enumeration was different – telephone interview versus mailout with telephone reminder in 1974 – this rate would not change appreciably. Indeed, an examination of the 1974 non-response figures shows that the proportion of non-response attributable to the actual interviewing is quite minimal. The greatest proportion resulted from the tracing operation, a procedure which is not expected to change.

This anticipated rate of non-response was used to arrive at a final sample size value in the following way

$$n = \frac{3268}{(1 - .27)} = 4477$$

The reader may note that the above sample size is only 9.8% of the size mentioned earlier. Why was the sample size for the survey so much larger? What has not been considered in the formulation of the problem?

The answer lies in the fact that the sponsor of the Graduate Survey required not only global estimates but separate estimates for each province and within each province for each of five different sub-populations. The extent of the data requirements is presented next.

Desired Levels of Disaggregation

The statement of desired precision applies to the overall population of graduates, not to any particular sub-group(s). Thus it is true that if the estimated proportion of, for example, graduates who are found to be unemployed at the time of the survey is say .05, the variance will be $(.075P)^2$ = .000014063. When estimating the proportion of graduates who are unemployed in PEI or Ontario or among those having a masters degree, this precision level no longer applies since only a part of the entire population is considered. The different parts of the population for which estimates are to be produced are called domains. As has been indicated on page 88, it is important to have a large enough sample in each domain to produce estimates with desired precision.

For each province, the five sub-populations of interest for which separate estimates were required were identified. The estimates produced for each of these domains were to achieve the desired level of precision. The five sub-populations of interest were:

a) individuals who, in 1976, completed the requirements for graduation from a one or two-year program in a community college or hospital school of nursing (C_1);
b) individuals who, in 1976, completed the requirements for graduation for a three or four-year program in a community college or hospital school of nursing (C_2);
c) individuals who, in 1976, completed the requirements for graduation from a university with an undergraduate degree (UB);
d) individuals who, in 1976, completed the requirements for graduation from a university with a masters degree (UM);
e) individuals who, in 1976, completed the requirements for graduation from a university with a doctorate degree (UD).

These five classifications were referred to as *level of qualification categories.*

The population sizes for each of the domains and the anticipated rates of non-response applicable to these domains are given in the following tables.

4 To facilitate the calculation of sample size (n-1) has been replaced by N. Because the sample size is large the effect is negligible.

Domain Sizes for Levels of Certification by Province

Province	C1	C2	UB	UM	UD	Total
Newfoundland	244	305	1,483	118	7	2,157
Prince Edward Island	155	103	282	–	–	540
Nova Scotia	757	130	4,345	360	46	5,638
New Brunswick	503	85	2,740	179	24	3,531
Ontario	12,408	4,378	35,506	5,806	879	58,977
Manitoba	1,025	203	4,298	409	66	6,001
Saskatchewan	695	232	2,930	231	26	4,164
Alberta	2,478	1,327	6,114	754	185	10,858
British Columbia	1,955	970	5,295	871	161	9,252
TOTAL	20,220	7,733	63,043	8,728	1,394	101,118

The anticipated non response rate for each domain was taken from the 1974 Ontario Graduate Survey. The rates given below according to level of qualification were applied to all the provinces.

Rates of Complete Non-Response Obtained in 1974 OGS

C1: $1 - r = .2712$

C2: $1 - r = .2705$

UB: $1 - r = .2587$

UM: $1 - r = .3724$

UD: $1 - r = .3724$

The statistical formulation given on pages 60-61 was applied to each of the domains of interest. It is interesting to note that while the overall finite population correction factor (f.p.c.) is very close to one, thereby having a small effect on the sample size, the same situation does not apply to the domains. The f.p.c.'s had a much greater effect particularly in the case of very small domains. The sample sizes which were calculated are given in the following tables. For comparison purposes a variance of $(.10P)^2$ has been considered.

Sample Sizes for $V(p) = (.075P)^2$, $P = .05$ and $V(p) = (.10P)^2$, $P = .05$

	$P = .05$ $V(p) = (.075P)^2$ $n_o = 3378$						$P = .05$ $V(p) = (.10P)^2$ $n_o = 1900$					
	C1	C2	UB	UM	UD	Total	C1	C2	UB	UM	UD	Total
Nfld	244	305	1390	118	7	2064	244	305	1124	118	7	1798
P.E.I.	155	103	282	–	–	540	155	103	282	–	–	540
N.S.	757	130	2564	360	46	3857	742	130	1782	360	46	3060
N.B.	503	85	2041	179	24	2832	503	85	1513	179	24	2304
Ont.	3643	2614	4155	3402	879	14693	2267	1822	2432	2282	879	9682
Man.	1025	203	2550	409	66	4253	913	203	1778	409	66	3369
Sask.	695	232	2117	231	26	3301	695	232	1556	231	26	2740
Alta.	1960	1307	2935	754	185	7141	1475	1072	1957	754	185	5443
B.C.	1699	970	2781	871	161	6482	1322	880	1886	871	161	5120
Total	10681	5949	20815	6234	1394	45163	8316	4832	14310	5204	1394	34056

Sampling Fractions for $V(p) = (.075P)^2$, $P = .05$ and $V(p) = (.10P)^2$, $P = .05$

	$P = .05$ $V(p) = (.075P)^2$ $n_o = 3378$						$P = .05$ $V(p) = (.10P)^2$ $n_o = 1900$					
	C1	C2	UB	UM	UD	Total	C1	C2	UB	UM	UD	Total
Nfld	1.00	1.00	.94	1.00	1.00	.96	1.00	1.00	.76	1.00	1.00	.83
P.E.I.	1.00	1.00	1.00	–	–	1.00	1.00	1.00	1.00	–	–	1.00
N.S.	1.00	1.00	.59	1.00	1.00	.68	.98	1.00	.41	1.00	1.00	.57
N.B.	1.00	1.00	.74	1.00	1.00	.80	1.00	1.00	.55	1.00	1.00	.65
Ont.	.29	.60	.12	.59	1.00	.25	.18	.42	.07	.39	1.00	.16
Man.	1.00	1.00	.59	1.00	1.00	.71	.89	1.00	.41	1.00	1.00	.56
Sask.	1.00	1.00	.72	1.00	1.00	.80	1.00	1.00	.52	1.00	1.00	.66
Alta.	.79	.99	.48	1.00	1.00	.66	.60	.81	.32	1.00	1.00	.50
B.C.	.87	1.00	.52	1.00	1.00	.70	.68	.91	.36	1.00	1.00	.55
Total	.53	.77	.33	.72	1.00	.45	.41	.62	.23	.60	1.00	.34

Strata as Domains of Interest

Since information on the level of qualification and province was available on the frame it was possible to turn the domains into separate strata. This made it possible to select the required sample sizes within each stratum (domain) since they were sampled independently.

Had it not been possible to stratify by level of qualification, for example, a much larger overall sample size would have been required to ensure that *at least* the sizes specified on page 62, would fall into each of the domains. In order to select all 7 Doctoral Graduates in Newfoundland, for example, a census would have been required.

Also, since it was felt that the characteristics of the graduate labour force is highly correlated with specific fields of study, the primary strata (domains) were further stratified according to major fields of study to form the sub-strata of the sample design. The sample size for each primary stratum was allocated on a proportional basis to each of the sub-strata. Subsequently a simple random sample of graduates was selected from each sub-stratum.

Other Considerations

Operational Considerations

It was felt that if the sampling fraction was .85 or greater for any domain, a complete census of the domain was warranted. This would save time and expense by avoiding the sample allocation and selection stages. The sample sizes and corresponding sampling fractions appearing below have been adjusted for such cases. Again for comparison purposes a variance of $(.10P)^2$ has been considered.

Sample Sizes for $V(p) = (.075P)^2$, $P = .05$ and $V(p) = (.10P)^2$, $P = .05$

	$P = .05$, $V(p) = (.075P)^2$, $n_o = 3378$						$P = .05$, $V(p) = (.10P)^2$ $n_o = 1900$					
	C1	C2	UB	UM	UD	Total	C1	C2	UB	UM	UD	Total
Nfld.	244	305	1483	118	7	2157	244	305	1123	118	7	1797
P.E.I.	155	103	282	–	–	540	155	103	383	–	–	540
N.S.	757	130	2561	360	46	3854	757	130	1777	360	46	3070
N.B.	503	85	2038	179	24	2829	503	85	1514	179	24	2305
Ont.	3643	2614	4155	3042	879	14693	2267	1822	2429	2281	879	9678
Man.	1025	203	2547	409	66	4250	1025	203	1775	409	66	3478
Sask.	695	232	2133	231	26	3317	695	232	1562	231	26	2746
Alta.	1961	1327	2931	754	185	7158	1473	1072	1952	754	185	5436
B.C.	1955	970	2778	871	161	6735	1323	970	1884	871	161	5209
Total	10938	5969	20908	6324	1394	45533	8442	4992	14298	5203	1394	34259

Sampling Fractions for $V(p) = (.075P)^2$, $P = .05$ and $V(p) = (.10P)^2$, $P = .05$

	P = .05, V(p) = (.075P)², n_0 = 3378						P = .05, V(p) = (.10P)² n_0 = 1900					
	C1	C2	UB	UM	UD	Total	C1	C2	UB	UM	UD	Total
Nfld.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.76	1.00	1.00	.83
P.E.I.	1.00	1.00	1.00	–	–	1.00	1.00	1.00	1.00	–	–	1.00
N.S.	1.00	1.00	.59	1.00	1.00	.68	1.00	1.00	.41	1.00	1.00	.54
N.B.	1.00	1.00	.74	1.00	1.00	.80	1.00	1.00	.55	1.00	1.00	.65
Ont.	.29	.60	.12	.59	1.00	.25	.18	.42	.07	.39	1.00	.16
Man.	1.00	1.00	.59	1.00	1.00	.71	1.00	1.00	.41	1.00	1.00	.58
Sask.	1.00	1.00	.72	1.00	1.00	.80	1.00	1.00	.52	1.00	1.00	.66
Alta.	.79	1.00	.48	1.00	1.00	.66	.59	.81	.32	1.00	1.00	.50
B.C.	1.00	1.00	.52	1.00	1.00	.73	.68	1.00	.36	1.00	1.00	.56
Total	.54	.77	.33	.72	1.00	.45	.42	.64	.23	.60	1.00	.34

Because of the small domain sizes involved, a complete census of graduates from community college programs is required in all provinces except Ontario, and the 1 and 2-year program of Alberta. In addition, a complete census of university undergraduate degree graduates is required in Newfoundland and Prince Edward Island, a complete census of university master degree graduates is required in all provinces except Ontario and a census of university doctorate degree graduates is required everywhere.

The sample size was judged to be operationally feasible in view of the fact that the large scale survey taking capacity exists at Statistics Canada.

Timeliness and Cost Considerations

The sample size was evaluated in terms of balancing survey costs and timeliness of results with survey benefits. It was decided to proceed with the 45,533 sample size option considering the known and potential users and uses of the survey data. The survey data was released in January 1979 at an overall survey cost of $550,000.

Conclusions

The reader may be somewhat surprised by the large sample size required in this survey. The need for such a large sample size is mainly attributed to the requirement for highly disaggregated estimates having high levels of precision. Such a situation applies to many surveys carried out by Statistics Canada including the Labour Force Survey and the Census of Population and Housing mentioned frequently in the manual.

Appendix D

Example of Weighting

The example summarizes the weighting procedures for the Survey of 1976 Graduates of Post-Secondary Programs. This example should be read in conjunction with a general description of the survey given in Appendix C.

1.1 Weighting Procedure

A weight is entered onto each record of the survey file which is used to generate cross tabulations. Tabulations of estimated population counts (proportions, percentages) are obtained by aggregating the weights of records which indicate the presence of specified characteristics. Two factors must be taken into account in calculating weights for units selected from each stratum. These are:

(i) the basic selection probability for each unit selected in stratum;

(ii) the nonresponse factor applied to each responding unit as a result of complete nonresponse to the survey

The combination of these two factors results in the following weight being applied to records in stratum h (h = 1, 2, ... L)

$$W_h = \frac{N_h}{n_h} \times \frac{n_h}{n_h^1} = \frac{N_h}{n_h^1}$$

where N_h = number of units in stratum h

n_h = number of units selected in stratum h

n_h^1 = number of responding units in stratum h

1.2 Production of Population Estimates

Provincial estimates ($\hat{Y}_p$) of the total number of persons with a certain characteristic (e.g., total number of persons who have a Masters degree in engineering and who were found to be employed at the time of the survey) is given by

$$\hat{Y}_p = \sum_{h \varepsilon p} W_h \sum_{i \varepsilon h} Y_{hi}$$

where Y_{hi} = if person i in stratum h has characteristic (e.g., Masters degree in engineering and employed)

0 otherwise

Operationally, estimates of totals are formed by sorting records according to whether or not the presence of characteristic is indicated and summing the corresponding weights. Estimates of proportions (percentages) of persons having a certain characteristic (e.g., proportion of Masters degree graduates who are employed at the time of the survey) are obtained by dividing the estimated totals (e.g., employed Masters degree graduates) by the sum of the weights of records in the domain of interest (e.g., Masters degree graduates).

Subject Index