Tutkimuksia Forskningsrapporter Research Reports 232

Ismo Teikari

Poisson mixture sampling in controlling the distribution of response burden in longitudinal and cross section business surveys

Tilastokeskus
Statistikcentralen
Statistics Finland

TUTKIMUKSIA FORSKNINGSRAPPORTER RESEARCH REPORTS 232

Ismo Teikari

POISSON MIXTURE SAMPLING IN CONTROLLING THE DISTRIBUTION OF RESPONSE BURDEN IN LONGITUDINAL AND CROSS SECTION BUSINESS SURVEYS

Tilastokeskus
Statistikcentralen
Statistics Finland

Cover
Maija Sohlman

Layout
Seija Töyräänvuori

ISSN 0355–2071
ISBN 951–727–873–X

"Published also as A–189,
Helsinki School of Economics
and Business Administration,
ISBN 951–791–611–6, ISSN 1237–556X"

Hakapaino Oy, Helsinki 2001

PREFACE

Users of survey data have insatiable demand for detail, which is only limited by cost and the response burden. Especially in business surveys the response burden is an important aspect since one business may fall in surveys many times in a given time interval. This also raises another question. How to even out this burden as fairly as possible? Poisson sampling, with Bernoulli sampling and Poisson πps sampling as special cases, has been found in the early seventies to have good properties as regards sample co-ordination. Using permanent random numbers it is possible to rotate some units out and some units in the sample in each draw. However, at the very beginning when I started this study I found one drawback in Poisson sampling. It bypasses some small units in each rotation round so that it is not fair to all small units. I found that by combining two special cases it is possible to fill at least some of this bypassed part. Thanks to Prof. Carl-Erik Särndal from the University of Montreal we have now the mathematical model and algorithm for this sampling scheme, named Poisson Mixture (PoMix) sampling. In this context I must also thank Hannu Kröger, who then worked at Statistics Finland, for preparing the co-ordination system with me and for discussing much of this problem. Hannu Kröger has done the programming work for Monte Carlo simulation studies in Chapter 12.

Prof. Särndal became interested in my problem and, thanks to the Finnish sampling project, I got the possibility to visit Canada in March 1996 to begin co-operation with Prof. Särndal. Our project team of three found in simulation studies that the PoMix sampling, combined with some parts of Bernoulli sampling, gave a smaller variance than the Poisson πps sampling when auxiliary information was used at the estimation stage. This was very important and we welcomed the result, which encouraged us to approach another drawback of the Poisson sampling, that is the random sample size. Using the method of Esbjörn Ohlsson and Prof. Bengt Rosen we prepared an order PoMix sampling which gives fixed size sampling and some bias but effectively the same as the random sample size PoMix sampling.

I thank the referees, Prof. Erkki Pahkinen and Prof. Osmo Kolehmainen, for their valuable comments on my work. I also want to thank Prof. Antti Kanto from the Helsinki School of Economics, and Dos. Seppo Laaksonen and Prof. Risto Lehtonen from Statistics Finland, with whom I could discuss my thesis and who encouraged and stimulated me at the beginning of and during the work, especially when I prepared the comprehensive introduction part concerning survey practice and theory of sample co-ordination. I am also obliged to Statistics Finland for its financial and other support for my work.

Last, but not least, I want to warmly thank my family, and especially my wife Raija, for their understanding both towards my work and towards the time preparing this thesis has taken from our time together.

Helsinki, March 2000

Ismo Teikari

CONTENTS

1.
INTRODUCTION

A response burden arises from the need for statistical information about finite populations. The demand for this information for the purposes of business surveys has grown rapidly in recent times. There are two reasons why governments, organizations and citizens need information: to plan national policies and to monitor the effects of implementing these policies. To create the necessary conditions for business operations, data are needed; but the gathering of such data creates a response burden for those businesses which are included in the survey. Reducing this response burden requires a proper survey program. *A survey* is an operation involving the collection of data to be used for statistical purposes, and *a survey programme* is a set of surveys conducted by a single statistical organization or commercial organization (Colledge 1995).

In planning a survey programme, the response burden should always be considered. The request for survey data should be matched to bookkeeping practices so that replying to the questionnaire will not take up too much time. The questions must be comprehensible and the number of questions should not be greater than is needed. To reduce the response burden on businesses, administrative data should be used as much as possible. As is well known, the burden imposed on businesses by governments and other organizations can be divided into an administrative and a response burden. The Small Business Institute at the Turku School of Economics and Business Administration, for example, when studying the problems of small, young businesses (Malinen 1994) on the basis of a sample population of 300 businesses with a maximum of 199 employees, found that they spent a reported average of 290 hours a year administrative tasks, which were of the following kinds:

1. The most expensive and time-consuming tasks were concerned with taxation
2. Extremely expensive task included membership fees of Trade Unions, employee benefits etc.
3. Utterly useless tasks comprised inquiries from statistical agencies, reports submitted to various authorities, inspections, public health matters etc.
4. The easiest and least expensive tasks were the issuing of testimonials, working agreements, notices etc.

There is also a considerable overlap in the demand for reports.

Completing questionnaires is not the most expensive aspect; that is the maintenance of an information system for inquiries. Consequently the response burden is greatest in the case of small businesses, because their accounting systems are often incomplete.

The response burden caused by a statistical agency or other authorities should be reduced by using existing administrative data as far as possible and by designing the survey policy so that it is easy to reply to questionnaires. This is not enough, however. It is also important that the existing response burden should be distributed as evenly as possible. To show that random sampling without any control over the response burden is unfair, Cox and Chinappa (1995) considered, as an example, a population stratum sampled at a rate of 25 percent each year, so that each unit would be expected to be sampled once over a 4-year period. If Simple random sampling is used, however, we have the following results:

- 31.6 percent will not be sampled at all
- 42.2 percent will be sampled once
- 21.1 percent will be sampled twice
- 4.7 percent will be sampled three times
- 0.4 percent will be sampled four times.

Random samples drawn from a single population can thus cause a problem of uneven response burden. If, however, a single frame is used for numerous surveys of a business population, it is possible to control the distribution of the response burden by suitable statistical methods. This is discussed in more detail in Chapter 7. The aim of this work as a whole is to present suitable statistical methods for controlling the distribution of the response burden.

We have so far used the term "business" without a definition. In everyday speech, the word business does not have a clear-cut meaning, but Kogan (1970) gives the following definition:

> "Business is any gainful occupation in which profit is the goal and in which there is risk of loss."

In this thesis the term "business" is used to refer to a legal entity in the real world, but in reference to the units listed in the sampling frame, the more exact and internationally harmonized statistical term "enterprise" is used. In business surveys an enterprise is identical to an institutional unit. It is useful to make a distinction between the real, observable world and what corresponds to it in administrative files and in statistical business registers. Thus the population element referred to in this study is a business but its counterpart in the frame is an enterprise. The enterprise is not a suitable unit for use in all business surveys, however, and thus it is often necessary to use more homogeneous units for regional or product surveys. Units in business surveys are discussed more detail in Chapter 3.

A survey is often understood as meaning only a sample survey. In its widest meaning, however, a survey also includes a census, i.e. a complete enumeration. A census of population can be defined as "the total process of collecting , compiling, evaluating, analyzing and publishing demographic, economic and social data pertaining to the whole population at a specified time". A census could also be defined as a special type of survey in which a whole population is surveyed instead of a subset. The distribution of the response burden does not have to be controlled in the case of a census, as all the units are studied in any case. The term survey is therefore understood in

its narrower meaning in the present connection and includes only sample surveys.

Instead of controlling the distribution of the response burden, it is sometimes useful to use the term "coordination". Two terms are widely used in the case of business surveys: positive coordination and negative coordination. Positive coordination means maximization of the overlap between two successive samples, whereas negative coordination means minimization of the overlap between two successive or parallel samples. Positive coordination is necessary in panel studies.

The coordination of business surveys is very important for statistical agencies that send out numerous questionnaires to businesses every year. In Statistics Finland, as in many other statistical agencies, the most suitable place for the coordination of business surveys is the Business Register, which is a list of enterprise units with their contact information and the most important auxiliary variables. The Business Register also provides a framework of concepts, economic measures, definition of units etc. Reality cannot be observed or measured without applying an observation framework and set of concepts. A more detailed discussion of the Business Register is given in Chapter 3.

One of the first coordination systems for business surveys was introduced by Johan Atmer and Lars-Erik Sjöberg in Sweden in the 1970s. The system is called SAMU (SAMordnade Urval inom företagsstatistiken) and the method used in it is called JALES, an acronym derived from the names of its inventors. The principal aim of the SAMU-system was negative coordination based on Sequential Simple random sampling. This method resembles the sequential sampling methods presented by Fan et al. (1962) which are based on random numbers which are not associated permanently with the units. The JALES method, on the other hand, is based on the use of Permanent Random Numbers (PRNs). A PRN is a Unif(0,1) random number attached to unit k at its creation and remaining with it during its entire life.

At the same time Brewer et al. (1972) in Australia introduced a method which was also based on the use of PRNs but with positive coordination as its principal aim. The method of coordination in this system was based on Bernoulli sampling and Poisson πps sampling schemes. The authors speak only about Poisson sampling, as the term Bernoulli sampling was not used until the book of Särndal et al. (1992) was published. Poisson sampling was introduced by Hajek in 1964, but surprisingly, it had been used in the USA before that. Bernoulli sampling is a special case of Poisson sampling with equal inclusion probabilities.

Although there has been much discussion over the randomness of random numbers, this question does not lie within the scope of the present work. We shall assume throughout that a program exists which assigns proper random numbers.

The coordination system used in Canada is based on the rotation group method, which does not involve the use of PRN's. The history of business sample coordination is presented in more detail in Chapter 8 and a description of existing coordination systems in Chapter 9.

While business surveys have many features in common with social surveys, they also have many special features, one of which is the rapidly changing structure of the sampling frame. New enterprises and establishments enter the frame continuously, and at the same time numerous old ones die. Some persistent businesses change their industrial class, some expand and some decline. The study of these phenomena, called business demography, is discussed in Chapter 4.

One of the most important features of a business survey population is its inevitably very skewed distribution, with substantial numbers of small businesses and just a few large ones. Simple random sampling (SRS) is hardly ever a good procedure for surveying businesses, because it draws on small and large enterprises in the same proportions. This leaves a great number of small businesses that do not contribute much to the estimates. Thus Probability Proportional-to-Size (PPS) samples and samples stratified by size -measures are the most suitable for business surveys. Survey and sampling designs and estimation methods in general are briefly discussed in Chapters 5 and 6. Schemes suitable for sampling coordination are presented in more detail in Chapter 8, which is a review of the history of coordination in business surveys.

At the beginning of this study, Poisson sampling was observed to have the property that, when rotating successive samples, it bypassed some small units in every rotation round. Poisson sampling is a highly suitable sampling design for the coordination of surveys, because it is very simple and makes it possible to use Permanent Random Numbers (PRN). It has two shortcomings, however: the bypassing of some small enterprises, as mentioned, and the size of the random sample, as described in Chapter 5. I do not consider this latter shortcoming very serious and will concentrate on the first one.

Studies of Small and Medium-sized Enterprises (SME) are very important at the present time, because politicians expect this group to have the greatest effect on employment. At the same time, interest in longitudinal studies has increased. Longitudinal data sets have to take into account the fact that individual businesses are heterogeneous. Time series and cross-sectional studies do not control this heterogeneity, and they therefore run the risk of obtaining biased results. Longitudinal data give more information, more variability, less collinearity among variables, higher degrees of freedom and greater efficiency, and are also better able to cope with the dynamics of adjustment. Moreover, many variables can be measured more accurately at the micro level, and biases resulting from aggregation over firms or individuals are eliminated (Baltagi 1996).

Bypassing the group of small enterprises weakens the possibilities for carrying out the studies mentioned in the previous paragraph, and at the same time the bypassing property makes the Poisson sampling scheme unsuitable for cases where all units should be updated within a certain period. The Poisson Mixture (PoMix) sampling scheme, which greatly alleviates this problem and in some cases even eliminates it entirely, will be presented in Chapter 10.

PoMix is a family of sampling schemes with two extremes: marked by Bernoulli sampling and Poisson πps sampling. PoMix sampling is based on the use of Permanent Random Numbers and on a constant shift in the rotation of successive samples. By changing the Bernoulli part and the width of the constant shift it is possible to control the area of bypassed units.

It was surprising to find in the Monte Carlo simulation results presented in Chapters 11 and 12 that PoMix sampling was more efficient than traditional Poisson πps sampling with some Bernoulli widths. The more skewed the distribution of population, the more effective PoMix sampling is. Thus PoMix sampling is very suitable for a business population.

It should be remembered that business surveys have human respondents and face the same problems as social surveys. There are some problems that will not be taken up for discussion here, e.g. those arising from data collection, such as response errors, missing data etc. It will be assumed throughout this work that a Business Register is available as a sampling frame for business surveys. This implies that direct element sampling is always possible, which means that a sampling frame exists that identifies every element in the population.

2.
SURVEY DESIGN

We set out from the words of Deming (1960):

> "Before there is any thought of survey or experiment someone must have a problem which is associated with some subject field. To find a rational solution for a problem, it requires a statement of aims and a criterion for evaluating the effectiveness of the system of solution".

Sometimes the chance of selecting the wrong solution to a problem can be lessened by the use of statistical information not now at hand. Whether statistical information is useful or not depends on the information and the problem. If the information is useful, the statistician can reformulate the problem in statistical terms, i.e. translate the need for information into a plan by which to acquire estimates of totals or proportions that will be relevant to the problem. This plan may be called "a survey design".

Surveys can be carried out in different ways. Two major classes are census surveys and sampling surveys. A sample can be defined as any fraction of all elements in a universe. However, sampling is not only a procedure for selecting a part of a whole, but is, as Deming (1960) suggests, a scientific method of investigation and inference, in which demonstrable reliability is required. This means that a sampling design must not only produce estimates but it must also provide demonstrable measures of the reliability of those estimates.

A census survey involves complete coverage, and can thus be regarded as a special case of a sample survey – based on a 100% sample. Official statistical agencies in almost all countries nowadays perform at least some census surveys. They are costly, but free of sampling errors. In addition, they provide frames for use in both social surveys and business surveys of the sampling kind. A census can also be filled in using administrative data as a source of information. In the widest sense of the word, the gathering of data by administrative sources must also be regarded as a type of survey.

Assuming that a need for information exists and that a survey seems to be the best way to meet that need, the next step is to decide what kind of survey seems to offer the best way of obtaining answers to the problem.

In general, the purpose of a survey is not to use data to make decisions about individual elements, but to obtain summary statistics covering a population or specific subgroups in a population. However, due to dissimilarities between and within subgroups, information based on microdata in addition to ordinary macrodata will be needed. Firms within the same industry do not use same production process, produce identical products and face identical costs. Numerous studies have proved that they differ dramatically even within the same geographical area and within the same

four-digit industry classes as defined by the SIC (Standard Industrial Classification). Heterogeneity is observed across time as well as across units. It is therefore obvious that different kinds of information are needed. Duncan and Kalton (1987) present eight kinds of information needs:

1. Estimates of characteristics, activity, behaviour, or attitudes for one point in time
2. Estimates of net change between two or more time periods
3. Estimates of gross change between two or more time periods
4. Estimates of trends based on several time periods
5. Estimates of duration, transitions, or frequency of occurrence for specific kinds of events and specific groups of people
6. Estimates of characteristics based on cumulative data over time
7. Estimates of rare events based on cumulative data over time
8. Estimates of relationships among characteristics

To satisfy these information needs, Bailar (1989) distinguishes five types of survey:

1. Single time surveys are designed to produce estimates of characteristics, activity, behaviour or attitudes for a single point in time. Estimates of duration and transition are possible if there are questions covering these topics. Often some of the information is the same as is collected with other surveys, so that it may be possible to estimate net changes.
2. Repeated surveys with no overlap between units investigate a given topic at regularly scheduled time points. In addition to estimates for a single point of time, these surveys give estimates for net changes and trends.
3. Repeated surveys with partial overlap are surveys where units are included a number of times and then rotated out of the survey. The main reason for the overlap is variance reduction. This approach allows gross changes to be estimated.
4. Longitudinal surveys with no rotation are designed to monitor a particular group of units over time. They allow characteristics to be traced for longer periods of time based on cumulative data. The main purpose is to estimate gross changes.
5. Longitudinal surveys with rotation are designed to monitor a particular group of units for a specified period of time, to introduce new sample units at specified intervals, to create longitudinal records for each observation unit and to include longitudinal analysis. Estimation of gross changes for a specified period is possible. This method also allows rare events to be identified in cumulative data.

Table 2.1

Survey types by kind of estimates that can be produced. X means always possible, (X) means sometimes possible.

Kind of Estimate	Type of Survey				
	Single Time	Repeated, no overlap	Repeated, partial overlap	Longitu-dinal, no rotation	Longitu-dinal, with rotation
For one point in time	X	X	X	X	X
Durations, transitions, frequency of occurrence	(X)	X	X	X	X
Relationships among characteristics	X	X	X	X	X
Net change	(X)	X	X	X	X
Trends		X	X	X	X
Rare events - cumulated		X	X		X
Gross change			(X)	X	X
Characteristics for longer time periods based on cumulated data				X	X

The above five types of survey and eight kinds of information needs have been tabulated by means of a modified version of the technique used by Bailar (1989). The table helps the statistician to decide what kind of survey to select. After this has been done a decision must be made regarding the set of elements (population) and the possible subpopulations (domains of study), including the characteristics of the population which are of interest. It is assumed in this thesis that a frame exists which makes it possible to reach the population units and which also includes the main population characteristics. I will return to this problem in the next two chapters. After the steps mentioned above we have some information on the distinctions between the frame and the population. The next steps are to design the sampling procedure and a way of collecting the data. The last of these problems lies outside the scope of this study.

A brief summary of Chapter 2

When it is decided to carry out a survey it must first be determined what kind of information is needed and what kind of survey can best elicit answers to the problems of interest. Expenditure considerations usually determine whether a census or sampling survey is used. The information needs tell us what kinds of estimates have to be produced, and these determine what type of survey should be carried out. The type of survey, in the last resort, determines what kind of control over the distribution of the response burden it is possible and reasonable to use. If a census is used, no such control is needed. Once the type of survey has been determined, the next step is to decide what are the population entities and to seek a frame for the population of interest.

3.
SAMPLING FRAMES FOR BUSINESS SURVEYS

This study concentrates on direct element sampling schemes. The term is used by Särndal et al. (1993), for example, to denote selection from a frame that directly identifies the individual elements of the population of interest. Thus direct element sampling means sampling based on a frame that contains a list of population elements, whereas the elements in indirect element sampling are units which are sets of elements in the first sampling stage, this being based on a frame that contains a list of sets, often called clusters. Since these primary sampling elements are often geographical areas, the frame is called an area frame.

Frame quality, which is a critical point for a successful survey, can be evaluated through the relations that exist between the target population and the frame population. Särndal et al. (1993) present eight necessary properties for a frame in direct element sampling:

1. The units in the frame have been identified
2. All units can be found, if selected in the sample.
3. The frame is organized in a systematic fashion
4. The frame contains a vector of auxiliary information for each unit
5. The frame specifies the domain to which each unit belongs.
6. Every element in the population of interest is present only once in the frame.
7. No element not in the population of interest is present in the frame.
8. Every element in the population of interest is present in the frame.

In practice, a frame is never perfect , mostly due to the continually changing structure of the business population. For example, there may be new birthed or recently dead businesses not recorded in the frame at the moment of sampling. Some frame units of interest may not be elements of the population and, conversely, some population elements may not be present in the frame. These frame imperfections are called undercoverage and overcoverage.

Figure 3.1 Undercoverage and overcoverage in a frame.

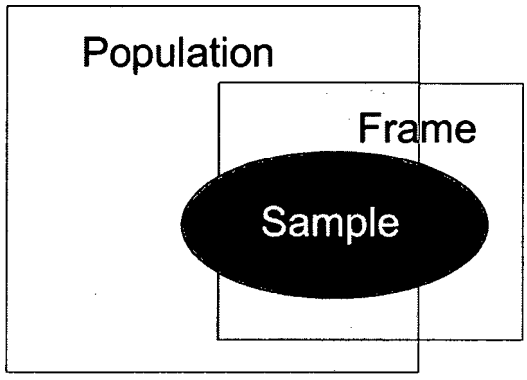


Figure 3.1 shows that a sample may include some frame units not contained in the population. Conversely, some units of interest are not attainable because they are not included in the frame.

Many other imperfections may also be found in a frame. Two examples are: 1. some population elements have links to more that one frame unit; 2. frame information may be incorrect, or not detailed and current enough to allow access to target population elements.

There are numerous definitions of a frame in the sampling literature. Two examples are quoted here, from Lessler (1982) and Colledge (1995) . Lessler says:

> "The materials or devices which delimit, identify and allow access to the elements of the target population. In a sample survey, the units of the frame are the units to which the probability sampling scheme is applied. The frame also includes any auxiliary information that is used for (1) special sampling technique, such as stratification and probability to proportional sampling, or for (2) special estimation techniques such as ratio or regression estimation".

The definition proposed by Colledge is as follows:

> "The survey frame is defined as a set of units comprising the sampled population with identification, classification, contract, maintenance, and linkage data for each unit".

Both definitions show the importance of the frame for delimiting the population of interest and for assembling the identification data, contact data and auxiliary data necessary for sampling. Identification data may include an alphanumeric identifier, name, address etc. Contact data are items required in order to locate units in samples, such as mailing address and telephone number. The frame should also include the maintenance and linkage data needed for repeated surveys or programmes comprising several surveys, and also classification data, including stratification variables and size variables for probability to proportional sampling.

Thus the frame is a list of units which are associated with population elements in which economic activities connected with the producing of goods

and services take place These elements are transformed. in the frame into units which are suitable for statistical purposes.

In the real world, economic activities are performed in entities having an operational and legal structure of their own (UN, Statistical Office 1990). The operating structure reflects the way a business makes decisions about its use of resources and the production of goods and services. This structure may divide an enterprise into two or more entities in which economic activity takes place. This activity can be principal, secondary or ancillary, where a principal activity is a process producing more than the other processes in the entity, either principal products or by-products. Secondary activities produce secondary products or by-products.

Principal and secondary activities cannot be carried out without the support of ancillary activities such as bookkeeping, transportation, storage, purchasing, sales promotion, cleaning, repair and maintenance, and security (UN Statistical Office 1990). As the production process is generally not viable without the support of ancillary activities, these latter should be allocated over all the production activities that they support. Sometimes they are organized into a central ancillary entity, however, and then it is expedient to use supplementary tabulation for ancillary entities in a survey.

The legal structure of a population entity provides the basis for its ownership. This structure forms a legal entity, which may be a corporation (possibly a public corporation), a joint stock company, a co-operative society, an incorporated non-profit association, a partnership, an individual proprietorship, or some other form of association. In some cases legal structures contain more than one legal entity. This entity, called a group of enterprises, is often created in a process of concentration (mergers, take-overs) or deconcentration (split-offs, break-ups). Concentration and deconcentration as demographic events are described in more detailed in the next chapter. Strategy and control are centralized into a relatively small top management group functioning in the manner of a management consulting firm within the group of enterprises.

Most economic entities are small enterprises whose economic activities are generally located in one place. In this case it is easy to collect information both geographically and on a detailed activity level. Economic activity in large and complex enterprises takes place in units, which are grouped into hierarchical structures for management, administrative and decision-making purposes, and this means that there is often a lack of geographically or industrially detailed information at the enterprise level. These enterprises must be divided into smaller parts, The most useful ones, whose definition is internationally standardized, being establishments, local units and activity units. For homogeneous use in international statistics, the U.N. Statistical Office recommends that reference should be made internationally to economic activity units. The third revision of the International Standard Classification of All Economic Activities published by the UN Statistical Commission in 1990 (ISIC rev.3) includes a definition of statistical units. The importance of such definitions is described as follows:

" The statistical unit serves as a tool to measure in an unduplicated and exhaustive fashion several aspects of the economy. In general, the utility of using standard classifications of activities, institutional sectors and geographical regions is weakened if they are applied to sets of transactors which are not defined in a standard way".

A basic statistical unit in business statistics is an institutional unit defined as a transactor in the system. This unit is also supposed to be capable of engaging in a transaction on its own behalf and in its own right. This unit can be a household or a legal, social or economic entity. In the case of business statistics, an institutional unit is normally a legal entity which owns or manages the property of the organizations, enters into contracts, receives incomes and maintains an independent, complete set of accounting records, including profit-and-loss accounts and a balance sheet. In most cases an institutional unit is a single legal entity whose existence is recognized in law independently of the persons or institution owning it. Thus a household cannot be a legal entity, and it follows that the boundary between households and enterprises is sometimes difficult to specify.

In some cases a legal entity is not independent enough to be an institutional unit. In this case a suitable statistical unit may be composed or two or more legal entities. This basic unit is called an enterprise, which is defined as an institutional unit, or the smallest combination of institutional units, that encloses and directly or indirectly controls all the functions necessary fro carrying out its production activities.

The smallest combination of institutional units means that there exist no legal entities created and owned by one or more other legal entities only for reasons of tax shelter or liability. Such units may not be able to survive without the rest of the corporation. It will then be convenient to integrate the activities of these entities into the corporation so that there exists only one enterprise.

Most enterprises are single-establishment enterprises, being homogeneous in terms of both the location of the unit and its activities. On the other hand, in multi-establishment enterprises different activities are often carried out in different locations. It follows that an enterprise is an unsuitable unit for either activity classification or geographical classification. The Statistical Commission of the UN has recommended use of the activity unit, the local unit or the establishment in such cases.

The activity unit is an enterprise or a part of an enterprise engaging in one kind of economic activity without being restricted to the geographical area in which that activity is carried out. It falls under unitary ownership or control from the outside and carries out only one activity. It can be heterogeneous in its location, however. The use of this unit allows the statistician to compile statistics that are homogeneous with regard to economic activities. When statistics are to be provided for individual geographical areas which are smaller than a country and no further breakdown according to economic activity is necessary, it will nevertheless be more appropriate to use the local unit for statistical purposes, as this covers all economic activities carried out by an enterprise at one location.

The statistical unit is often required to be homogeneous with regard to both its location and its activity, And in this case the ideal unit is the establishment.

Most economic statistics are significant only when the breakdown occurs by activity. A consistent classification across surveys and over time is then needed. The standard industrial classification of all economic activities is an integral part of the frames called Business Registers. The ISIC (International Standard Industrial Classification of all Economic Activities) serves as the standard for this purpose. Its two-digit or lower level is a hierarchical system of categories coded with Arabic numerals and arranged on a decimal system. At the one-digit level, which does not belong to any particular hierarchical level, the letters A - Q are used. This makes it possible to use more than ten tabulation categories.

The Statistical Commission of the UN recommends using the ISIC with such modifications as may be necessary to meet national requirements, without disturbing the framework of the classification. The ISIC has links to the Central Product Classification (CPC), the central instrument for classifying goods and services, and the Statistical International Trade Classification (SITC), the classification for transportable goods in international trade statistics.

The council regulation on the statistical classification of economic activities in the European Community recommends a classification called NACE, which is identical to the ISIC at the two-digit level.

Business survey statisticians are often interested in links between survey units over time as recorded in the Business Register. The treatment of changes is linked to what is recorded in the Business Registers at any point in time. In other words, whatever events in the outside world are deemed relevant, their consequences for entries in the statistical business register should be described between units. As the Business Register is in general based on the files of the tax authorities, the events recorded are administrative events of a kind that is not suitable for statistical purposes in all cases. This is described more detail in the next chapter.

The benefit of having an administrative source for the Business Register is that it provides an inexpensive starting point for the register and a source of continuous information for its maintenance. An administrative business list cannot satisfy all the register's data requirements, however. Its coverage may be inadequate, or its data items may be insufficient for classification and contact. Moreover, the legislation governing the administrative process places limits on the use of administrative data, so that we have a distinction between the statistical world and the register world. It follows that business registers also use supplementary information from other sources (Colledge M. J., 1995).

The problems raised by large, multi-establishment enterprises are usually different from those of small, single-establisment enterprises. The problems arising from small businesses are the following. A large number of units must be maintained, high volumes of data on the formation and closure of businesses must be updated, and accurate classification data is laborious to

handle. Furthermore, bookkeeping systems in small enterprises are often incomplete, which results in the respondent burden becoming high if the register data has to be complemented by surveys. Fortunately, administrative and statistical units most often coincide in small enterprises, so that use can often be made of administrative data.

Problems arising from large units concern the difficulty of defining suitable statistical units and reporting units and the updating of organizational changes. The tax authorities are generally interested only in the accounts of the whole enterprise, which means that data on individual establishments must be collected by direct inquiries, called register surveys. Some countries use a process referred to as profiling to determine the statistical structure of large, complex businesses. Profiling means direct contacts with large businesses aimed at defining their structures (Pietch L. 1995).

The good thing is that the response burden is not a major problem among large enterprises, as they usually have good bookkeeping systems. On the other hand, the study of Tuominen (1999) shows that larger enterprises have fairly high response burdens, due to the complex structure of their operating and bookkeeping systems.

A brief summary of Chapter 4

The quality of the frame is a critical part of a business survey, as it must include and define the units and classifications needed for the survey. Units must be suitable for statistical purposes and coincide with real units in the population. These purposes depend on whether information is needed about the geographical level or the activity level. The Statistical Commission of the UN recommends the use of units defined according to the third revision of the International Standard Classification of all Economic Activities (1990).

A frame is never perfect in practice, and there is always some undercoverage and overcoverage in frames due to births of new businesses and the deaths of old ones. In addition, there may be some errors in the contact data, making communication with the businesses impossible.
A frame is often in linked to a coordinating system, which may be needed to update data used for coordination purposes, such as permanent random numbers.

4.
BUSINESS DEMOGRAPHY

Some changes that can occur in the business population were discussed in the previous chapter. A new entry in the administrative register does not necessarily imply the birth of an enterprise in the real world. Some questions arise regarding this problem. How should the birth of an enterprise be recognized? When should identification numbers be changed for statistical units? How should the register track units?

Business Registers should identify changes in the business population in an appropriate way, and the types of changes should be identified and classified. Struits and Willeboords (1996) present a basic classification of changes (see below)

Table 4.1.
Basic Classification of changes in business population. x:y means x units before and y units after the event)

Change class	Number of units involved before:after the event	Identity continued
1. Change of Characteristic	1:1	Yes
2. Change of Existence		
2.1 Birth	0:1	No
2.2 Death	1:0	No
3. Change of Structure		
3.1 Concentration		
3.1.1. Merger	x:1	No
3.1.2. Takeover	x:1	Yes
3.2 Deconcentration		
3.2.1. Break-Up	1:y	No
3.2.2 Split-Off	1:y	Yes
3.3. Restructuring	x:y	Yes or No

A change of characteristic means that the identity of the enterprise has not changed but its main activity or size class or some other important character has changed. The number of units before and after this change is the same and the identity of the enterprise continues.

Changes of existence involve units which are not related to any unit of the population to which they could be compared, before or after change. In the case of birth there is no related enterprise before the event, and in the case of death there is no related enterprise after the event. There is no continuity of identity in either of these cases, of course.

Changes of structure involve more than one unit before or after the change, which can therefore be a process of either concentration or deconcentration. **Concentration** means that two or more units combine to

form one. The unit emerging from this concentration may or may not be essentially the same as one of the units before the event. In the case of take-over one unit continues its identity whereas other units lose theirs, While in the case of merger all the units lose their identity and a new enterprise appears after the event. During **deconcentration** a unit either breaks up without any part retaining the identity of the original unit or one or more units split off from another, which retains its identity. This case is the reverse of concentration. **Restructuring** involves more complex changes in structure.

All of these changes except changes of characteristics entail the administrative birth or death of an enterprise, in addition to which a case of birth in an administrative sense can occur when an enterprise changes its legal form.

These changes can be identified in the administrative register By using the continuity or change of telephone numbers or addresses, for example, to separate real cases of birth and death from purely administrative ones. In Finland the tax authorities gives an ID number to each enterprise when it in appears in the register, and this number disappears in the case of closure and in some cases of a change in structure. The Business Register, which investigates the numbers of new enterprises which appear in its lists and the locations of their establishments, gives an establishment ID-number to each, which is independent of the enterprise number and in general does not change in response to future demographic events. It is thus possible to classify some demographic events using the enterprise number and establishment number together, making use of available information on the continuity or existence of the new enterprise before the change (Laaksonen, Teikari 1998).

It is possible to use the movement of employees from one enterprise to another to improve the accuracy of the classification. Tuija Mustaniemi (1997) studied real instances of the formation of enterprises in the business register as a proportion of administrative ones, constructing for the purpose of this analysis a longitudinal worker-establishment data set covering the years 1988-1992. The data included variables from the Business Register, the Regional Employment statistics and the Statistics of Bankruptcies. For a real birth she used three criteria: 1) a new enterprise must start its activities by creating a new establishment, i.e. a new enterprise creates its own factors of production, 2) a new enterprise is not allowed to share a high proportion of employees with any other enterprise that has closed down or remained active, because this usually indicates an administrative birth that is not real, and 3) a new enterprise must be economically active. Using these three criteria the analysis showed only 54 percent of all openings of enterprises in the retail trade and 63 percent in manufacturing to be classifiable as real births.

Enterprise demography has attracted a great deal of attention recently due to its importance for political decisions. In the process of sample coordination, administrative birth or death may mean that we miss a unit which exists in the real world or we may send a questionnaire out to it in succeeding surveys simply because the its ID number has changed. It would therefore be desirable that the history of an enterprise should be contained in the sample frame.

A brief summary of Chapter 4

Changes in business population cause lead to overcoverage and undercoverage in a frame, And can also cause some problems in sample coordination. We can send a questionnaire to an enterprise which has changed its ID-number and was in the same inquiry last time with another ID-number. Business demography is essential in order to reduce these difficulties.

5.
SAMPLING DESIGN

Sampling does not give information on every unit of the population, and thus it results in sampling errors, but it is preferred because it is cheaper and easier to carry out than a census. When sampling is designed carefully the resulting data can be sufficiently precise, with moderate unbiasedness and variance, and for the purpose at hand. It may sometimes be as precise as a census, or even more precise, due to the fewer non-sampling errors. To be sufficiently precise the sample must be representative. Neyman (1934) described representative sampling in the following manner:

> "I should use these words with regard to the method of sampling and to the method of estimation, if they make possible an estimate of the accuracy of the results obtained in the sense of the new forms of the problem of estimation, irrespectedly of the unknown properties of the population studied."

By accuracy, Neyman meant small bias.

In the first third of the twentieth century there were two versions of representative methods of sampling, one based on purposive selection and the other on random selection. Following Neyman's work in the 1930's, sampling based on random selection has predominated, because purposive sampling does not give the inclusion probabilities for units, so that it is impossible to measure the precision of the sample. Purposive sampling is a method that selects sample units according to the judgement of the researcher, who assumes the result to be representative. For that reason the method is called also judgement sampling. An example of purposive sampling is the farm surveys carried out by the Australian Bureau of Agricultural and Resources Economics, as described in more detail by Bardsley and Chambers (1984).

Even though the purposive method was very popular for business sampling until the 1980's, this study will concentrate on randomized sampling methods. As it is assumed that there exists a sampling frame, the analysis is restricted to direct element sampling techniques (see later chapters). This means that a two-stage sampling design, two-phase sampling design and clustered sampling design are all bypassed and that two basic types of randomized sampling method are taken for consideration, namely the draw-sequential and list-sequential sampling schemes.

The draw-sequential sampling scheme involves drawing randomly units from the entire population and including every drawn unit in the sample, and can be carried out either with or without replacement of every unit in the frame after each draw. Typical sampling schemes are Simple random sampling With Replacement (SRSWR) and Simple random sampling without replacement (SRS). SRS can also be performed sequentially, as we will see in

later chapters. Ohlsson (1993) calls this Sequential Simple random sampling.

The list-sequential sampling scheme consists of successive experiments (unit by unit) for frame units; not necessarily for each frame unit. Each experiment results in either selection or non-selection of the unit in the sample. Bernoulli sampling is a typical form of list-sequential sampling, where each frame unit is given a random number from a uniform distribution *Unif(0,1)*. The unit is then included in the sample if its random number is greater than the sampling fraction.

The sampling design determines the statistical properties of the resulting estimates (sampling distribution, expected values, variance), while the sampling distribution determines the inclusion probabilities that are necessary for calculating point and interval estimates. The sampling distribution *p(.)* gives a probability for each sample in the set of all possible samples. Inclusion of a given element in the sample *(s)* can be indicated by the random variable I_k, defined as

$$I_k = \begin{cases} 1, \text{ if } k \in s \\ 0 \text{ otherwise} \end{cases},$$

which is known as the sample membership indicator (Särndal et al.1992).

The probability of an element k being included in a sample is obtained from the given design *p(.)* as follows:

$$\pi_k = p(k \in s) = p(I_k = 1) = \sum_{k \in s} p(s),$$

where $k \in s$ indicates that the sum is calculated over those samples s which contain k.

If the frame includes auxiliary information which is closely correlated with a variable of interest, it is possible to use sampling designs that effectively exploit this auxiliary information. This is extremely important for business surveys due to the skewness of distribution of business populations. The Stratified Sampling and unequal probability sampling schemes are possibilities in this situation.

The inclusion probability in equal probability sampling schemes is equal to the sampling fraction for each sampling unit. Examples are Bernoulli Sampling, and Simple random sampling without replacement (SRS).

Bernoulli sampling is based on a binomial probability model called a Bernoulli model. There is one trial for each of the N units, and these trials are statistically independent. Each trial is carried out with a random number and results in either success or failure. Each trial has a probability of success π and a probability of failure $(1-\pi)$. Thus there are only two possible outcomes for each trial and their probabilities remain the same throughout the trials. The event "n trials result in k successes and n-k failures" can come about in as many ways as k letters can be distributed among n places. The event thus contains $\binom{n}{k}$ points and each point has the probability $p^k q^{n-k}$.

In Bernoulli sampling the sample membership indicators $I_1,...,I_N$ are independent and identically distributed random variables. The inclusion probability π is a constant, and each I_k follows the Bernoulli distribution

$$\begin{cases} p(I_k = 1) = \pi \\ p(I_k = 0) = 1 - \pi \end{cases}$$

Denote the realized sample size as n_s, then the sampling design BE is expressed by

$$p(s) = \pi^{n_s}(1-\pi)^{N-n_s}.$$

The binomial probability model also gives the answer to another question. The probability of obtaining exactly n_s successes is given by

$$\binom{N}{n_S}\pi^{n_S}(1-\pi)^{N-n_S}.$$

This means that the sample size n_s is a random variable, distributed binomially with a mean

$$E_{BE}(n_S) = N\pi$$

and variance

$$V_{BE}(n_S) = N\pi(1-\pi).$$

In the design *BE* the estimator π for the population total and its variance takes the forms

$$\hat{Y} = \frac{1}{\pi}\sum_s y_k$$

and

$$V_{BE}(\hat{Y}) = (\frac{1}{\pi} - 1)\sum_U {y_k}^2.$$

An unbiased variance estimator is

$$\hat{V}_{BE}(\hat{Y}) = \frac{1}{\pi}(\frac{1}{\pi} - 1)\sum_s {y_k}^2.$$

As the sample size n_s varies in the sampling procedure, Brewer et al. (1984) recommend use of the ratio estimator

$$\hat{Y}_r = N\frac{\sum_s y_k}{n_s} = N\bar{y}_S ,$$

which is called the expanded mean estimator.

The variance estimator takes the form

$$\hat{V}_{BE}(N\bar{y}_s) = N^2\frac{1-\pi}{n_s}(1-\frac{1}{n_s})S^2{}_{yU},$$

where $S^2{}_{yU}$ refers to population variance of y; the variable of interest.

Simple random sampling without replacement (SRS) is based on the assumption of a normal distribution of all possible samples. Every sample *s* of fixed size *n* receives the same probability of being selected, and we have

$$p(s) = \begin{cases} 1/\binom{N}{n_S}, \text{if s is of size n} \\ 0 \text{ otherwise} \end{cases}.$$

Auxiliary information can be introduced by stratifying the population using a size variable which is closely correlated with the variable of interest. The inclusion probabilities are then identical within each stratum and different between the strata. The sampling design is then Stratified simple random sampling (STRSRS). Stratification is treated in more detail in section 8.1.

A Stratified Simple random sampling where the sample fraction differs from stratum to stratum, can be regarded as equal probability sampling as long as we think of each stratum as a separate population. If our study variable *y* is approximately proportional to a known auxiliary variable *x* we can incorporate this auxiliary information by selecting each element *k* with a probability proportional to the size measure x_k. It follows that each unit in the frame has its own inclusion probability. The use of unequal probabilities was first suggested by Hansen and Hurwitz (1943). Their method was based on the replacement of units in the frame after each draw. Sampling with replacement is less efficient than sampling without replacement, but it has four advantages (Brewer and Hanif 1983):

1. Selection of the sample is simple
2. The method can be used for any predetermined (but not necessary distinct) number of units in the sample
3. The unbiased estimator of variance is simple.
4. It is also comparatively easy to obtain unbiased estimators of total variance and components of variance in multistage designs.

However, to avoid the possibility of units being selected more than once, Horwitz and Thompson (1952) presented a general theory of sampling with unequal probabilities without replacement based on the use of an unbiased estimator called an HT estimator

$$\hat{Y} = \sum_s \frac{y_k}{\pi_k},$$

where π_k incorporates the auxiliary information of x_k . In this paper the sampling with unequal probabilities without replacement is called πps sampling.

Brewer and Hanif (1983) review 50 procedures which have been suggested for πps sampling without replacement, and show that the solution requires that the sample either 1) changes some selection probabilities, or 2) allows some variation in sample size. An ideal sample selection method should satisfy the following requirements:

1. Units are selected with probabilities π_k proportional to the size measure x_k
2. The probability, π_{kl}, of joint selection of the units indexed by k and l should be positive for all k, l;
3. Covariance exists between the sample membership indicators $\pi_k\pi_l - \pi_{kl} \geq 0$, for all $k, l\ (k \neq l)$
4. π_{kl} is known or can be calculated

Much attention has also been devoted to the desire for fixed sample size, which means that the first requirement would take the following form

1'. Exactly n units are selected without replacement and with probabilities π_k proportional to size x_k

Sunter (1986) has given some solutions to this problem, but his algorithms are not suitable for the coordination of samples because this is impossible based on PRNs (negative or positive). Poisson sampling, which makes this control possible, is presented in section 8.3 in the context of the history of sample coordination. We will now briefly go into the principles of sample estimation.

A brief summary of Chapter 5

A probability sampling design gives us a device for generalizing a sampled population to the whole population. Sampling coordination using PRNs needs a list-sequential sampling scheme such as Bernoulli sampling, Sequential SRS or Poisson sampling. Bernoulli sampling is a special case of Poisson sampling, that is, equiprobable Poisson sampling. Due to the highly skewed distribution of business populations, stratification is needed for equiprobable sampling schemes. Bernoulli sampling is described in this chapter, and sequential SRS and Poisson sampling are described in more detail in later chapters.

6.
ESTIMATION

When using a sample we aim to obtain point estimates for certain parameters of interest. These parameters may be totals, ratios of totals, domain means, regression coefficients, medians, etc.

Let $\hat{\theta}$ be an estimator of the parameter θ. The Mean Square Error (MSE) is a measure of the accuracy of $\hat{\theta}$,

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + \left[B(\hat{\theta}) \right]^2 ,$$

which depends both on the bias $B(\hat{\theta})$ and the variance $V(\hat{\theta})$ of the estimate $\hat{\theta}$. Bias is an important character which measures the distance between the true value of parameter θ and the expected value $E(\hat{\theta})$ of the estimator $\hat{\theta}$,

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta .$$

Although unbiasedness is an important quality in an estimator, its importance must not be exaggerated. Särndal et al. (1992), for example, present two cases where it is not reasonable to seek an exactly unbiased estimator.

1. Many parameters have a structure that makes it difficult to find an unbiased estimator.
2. An estimator with bias can often have a smaller variance and mean square error *(MSE)* than an unbiased estimator.

Unbiasedness is a characteristic of the *HT*-estimator, which sometimes gives very large variance. By accepting some bias we often achieve a smaller *MSE*. We cannot accept a very large bias, however. It must be small in relation to the standard error. This is important for confidence interval to be valid. A measure for this quality is the bias ratio

$$BR(\hat{\theta}) = \frac{B(\hat{\theta})}{\left[V(\hat{\theta}) \right]^{1/2}} .$$

Following Särndal et al.(1992), we can show that as long as $BR(\hat{\theta})$ is small, a calculated confidence interval will not be greatly in error despite a non-zero

bias. As the above authors have shown, we can see this by making approximation

$$Z = \frac{\hat{\theta} - E(\hat{\theta})}{V(\hat{\theta})^{1/2}} \approx \mathrm{N}(0,1). \qquad (6.1)$$

The coverage probability is then

$$P_{COV} = P\left\{\hat{\theta} - z_{1-\alpha/2}\left[V(\hat{\theta})\right]^{1/2} < \theta < \hat{\theta} + z_{1-\alpha/2}\left[V(\hat{\theta})^{1/2}\right]\right\}$$
$$= P(-z_{1-\alpha/2} - BR(\hat{\theta}) < Z < z_{1-\alpha/2} - BR(\hat{\theta})).$$

P_{cov} equals the desired confidence level $1-\alpha$ only if $BR(\theta)=0$. The effect of the bias ratio on the coverage probability P_{cov} may be ignored, however, if its absolute value is small, as can be seen in the table below.

Table 6.1. Probability P_{cov} as a function of the bias ratio $BR(\hat{\theta})$

$\lvert BR(\hat{\theta})\rvert$	$\mathbf{P}_{\mathbf{cov}}$
0.00	0.9500
0.05	0.9497
0.10	0.9489
0.30	0.9396
0.50	0.9210
1.00	0.8300

If a biased estimator is used, the variance should be replaced by MSE, so that the interval for θ will be

$$\hat{\theta} \pm z_{1-\alpha/2}\left[MSE(\hat{\theta})\right]^{1/2} \qquad (6.2)$$

The value of the probability P_{cov} can be expressed in terms of the bias ratio $BR(\hat{\theta})$. Assuming that $\hat{\theta}$ is normally distributed and the coverage probability $1-\alpha = 0.95$, the coverage probability lies between 0.9489 and 0.95 if $\left|BR(\hat{\theta})\right| \le 0.1$.

According to Hidiroglou et al. (1995) we now consider a point estimator of the population total Y based on auxiliary information about P subgroups of the population U, called model groups and denoted $U_p, \mathrm{p} = 1,...,\mathrm{P}$. Let s_p be the part of the whole sample s that lies in model group U_p. The estimator is given by

$$\hat{Y} = \sum_{p=1}^{P} \sum_{s_p} w_k y_k \,, \qquad (6.3)$$

where $w_k = a_k g_k$. The weight w_k is thus composed of two weights: the sampling weight $a_k = 1/\pi_k$, which is the inverse of the original inclusion probability, and the g-weight

$$g_k = 1 + (\mathbf{X}_p - \hat{\mathbf{X}}_{p\pi})' \left(\sum_{Sp} \frac{a_k \mathbf{x}_{pk} \mathbf{x}'_{pk}}{c_k} \right)^{-1} \frac{\mathbf{x}_{pk}}{c_k}, \qquad (6.4)$$

which incorporates the auxiliary information associated with the particular model groups U_p, $p=1,...,P$ used in the estimation. p is the index of a model group for which one or more auxiliary variable total is known, $\mathbf{x}_{pk}$ is an auxiliary variable vector, and $\mathbf{X}_p = \sum_{U_p} \mathbf{x}_{pk}$. The known constants c_k are determined by the variance structure of the assumed underlying regression model

$$y_k = \boldsymbol{x}'_{pk} \boldsymbol{\beta}_p + \varepsilon_k, \qquad (6.5)$$

where $\boldsymbol{\beta}_p$ is estimated from the sample.

This regression structure also gives the estimated regression residuals $e_k = y_k - x'_{pk} \hat{\beta}_p$, which are needed for the variance estimator given by

$$\hat{V}(\hat{Y}) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{g_k e_k}{\pi_k} \frac{g_l e_l}{\pi_l}. \qquad (6.6)$$

In the case of equal probability sampling designs, (6.6) reduces to simpler forms. For Simple random sampling, for example, it takes the form

$$\hat{V}(\hat{Y}) = N^2 \frac{1 - n/N}{n} \sum_{s} \frac{(g_k e_k)^2}{n-1}.$$

If we do not use auxiliary data, this means that $p=1$ (the entire population is the only model group) and $g_k=1$, because $x_k=0$ for all k, and this leads to the well known HT estimator

$$\hat{Y} = \sum_{s} a_k y_k \,, \qquad (6.7)$$

which uses the auxiliary information only in the design stage (or not at all). HT estimator is unbiased, as is well known, and its variance estimators takes the form

$$\hat{V}(\hat{Y}) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

The Horwitz-Thompson estimator can be improved using auxiliary data, totals for which may be known for the entire population or for specified subpopulations. By incorporating this information into our estimation process through the g-weight, we can improve the estimates obtained using the ratio, post-statification, regression or ranking methods.

Another way to incorporate auxiliary data into the estimates is that known as the calibration approach. This method is described in the paper of Deville and Särndal (1992) , for example.

In business surveys, where we are operating with highly skew distributions, the data are often divided into two strata: the take-all stratum U^{TA} and the rest of the population U^R.

$$Y = \sum_{U^{TA}} y_k + \sum_{U^R} y_k \,. \qquad (6.8)$$

Three estimators for the population total, which are used later in this work, are presented below.

In the following expressions $a_k = 1/\pi_k$ denotes the sampling weights of unit k:

The Horwitz-Thompson (HT) estimator (6.7) takes the form

$$\hat{Y}^{HT} = \sum_{s} a_k y_k = \sum_{U^{TA}} y_k + \sum_{s^R} a_k y_k \,. \qquad (6.9)$$

This estimator does not use auxiliary information in the estimation stage, but the remaining two do. They are not unbiased, but in general give smaller MSE's.

The separate ratio estimator takes the form

$$\hat{Y}^{\mathrm{SRAT}} = \sum_{U^{TA}} y_k + X^R \hat{b}^{SRAT} = \sum_{U^{TA}} y_k + (\sum_{U^R} x_k) \frac{\sum_{S^R} a_k y_k}{\sum_{S^R} a_k x_k} \,, \quad (6.10)$$

and the GREG estimator takes the form

$$\hat{Y}^{\mathrm{GREG}} = \sum_{U^{TA}} y_k + \sum_{S^R} a_k y_k + \hat{b}^{GREG} (\sum_{U^R} x_k - \sum_{S^R} a_k x_k) \,, \quad (6.11)$$

where

$$\hat{b}^{GREG} = \frac{\sum_{S^R} a_k(a_k - 1)y_k x_k}{\sum_{S^R} a_k(a_k - 1)x_k^2}.$$

The slope calculation in $\hat{Y}^{GREG}$ uses the weighting $a_k(a_k - 1)$. Note that $\hat{Y}^{GREG}$ and $\hat{Y}^{SRAT}$ are both members of the GREG family of estimators given by (6.11). By equating these estimators with (6.4), we find the *g-weights* implied by each of them. These weights are required for variance estimation. It will be shown later in Chapter 11 that $\hat{Y}^{GREG}$ and $\hat{Y}^{SRAT}$, which use auxiliary information at the estimation stage, will improve the HT estimator $\hat{Y}^{HT}$ even if the auxiliary information is also used at the sampling stage. Part of the reason why the HT estimator has a comparatively large variance is that the randomness of the sample size penalizes it but not the GREG estimators.

A brief summary of chapter 6

When we estimate totals, ratios of totals, domain means regression coefficients etc, using the sample it is not enough to have point estimates only. We must know how accurate our estimates are. The measure of this accuracy is a mean square error (MSE) which consists of the bias and the variance of the estimate. Even if the unbiasedness is an important quality in an estimator its importance must not be exaggerated. Unbiased estimator often gives large variance and accepting some bias the variance and MSE often can be reduced.

Incorporating the auxiliary information into the weights gives small bias, but reduces often considerably the variances of estimates. This chapter presents three estimates from which one, which is well known HT-estimator, do not use auxiliary information while two others, which are separate ratio estimator and GREG estimator, use auxiliary information in the estimation level. These three estimators are used later in Monte Carlo tests for PoMix sampling.

7.
RESPONSE BURDEN

Since time is money, business owners generally resent spending time on replying to surveys. Brewer et al. (1972) refer to this as sample fatigue, which is an important problem in repeated mail and field surveys unless preventative measures are available. They regard this sample fatigue as an effect which tends to make survey responses different from what they might otherwise have been. A respondent may, for example, refuse to answer yet another voluntary questionnaire if no rotation method is used. On the other hand, repeated surveys are needed because economic analysis usually requires measures of change over time, which can be done with greater precision if there is a substantial overlap between successive samples.

Users of survey data have an insatiable demand for detail, which is only limited by cost and the response burden. Thus requests for survey data should be matched to bookkeeping practices. It is unreasonable to request more detail than can be extracted from a business's accounts.

The largest businesses in the economy are likely to be selected as respondents for most surveys for which they qualify, and their burden can be minimized in this respect only if the statistical agency ensures that the statistical units included in survey frames correctly represent the business's organization and are compatible with structural units for which the data can be reported with least effort. This presupposes a process of setting up statistical reporting arrangements for a business. This is called profiling. At a minimum, profiling involves personal contact with a large business to gain insight into its legal and organizational structures. Good questionnaires impose a low response burden and remain friendly to both the respondent and the interviewer.

Questionnaires from statistical agencies are only a part of the whole administrative burden imposed on businesses. For example, an inquiry into administrative practices in small and medium size enterprises carried out by the Business Research Centre of Turku School of Economics and Business Administration (1994) showed that businesses spend about 290 hours a year and about 40.000 FMK on administrative practices, of which about 5 percent is devoted to inquiries made by statistical agencies. The inquiry concerned aspects of 46 administrative practices, of which the 10 most unnecessary ones and the 10 most expensive ones in the respondents' opinions are presented in the table below.

Table 7.1

The ten most unnecessary and most expensive administrative practices in the opinion of respondents to an inquiry prepared by the Business Research Centre of Turku School of Economics.

Place	The most unnecessary	The most expensive
1.	Collection and payment of trade union fees	Reports for patents and trademarks
2.	Inquires by Statistics Finland	Documents for balance sheets
3.	Contracts regarding trade union fees	Tax returns
4.	Reports to unions	Claims for the costs of occupational health care and reports
5.	Occupational health reports	Collection and payment of trade union fees
6.	Dealings with the provincial administration	Inspections
7.	Dealings with the local authorities	Reports to unions
8.	Unemployment insurance fees	Unemployment insurance fees
9.	Announcements for employment agencies	Dealings with the provincial administration
10.	Claims for the costs of occupational health care and reports	Inquires by Statistics Finland

The public authorities in Finland have paid some attention to this administrative burden, and Statistics Finland has investigated the response burden that it places on businesses. In some countries statistical agencies have indexed and measured this response burden. Terhi Tuominen (1999) sent out a supplementary questionnaire with five business surveys in spring 1998 in order to index and investigate the response burden induced by these questionnaires.

The response burden can be relieved by using administrative data as much as possible, by profiling and by good survey programme designing. The information is still needed, however, so that some response burden must exist. Three basic features of a useful definition of this response burden are listed by Sunter (1977):

1. Each survey questionnaire (j=1,2,...,n) is assessed for its *response 'load'*. This would be expressed conveniently as a money equivalent to the time and effort required to complete the questionnaire and determined by one, or some combination of, the following methods: (i) careful assessment through simulation, role-playing and interviews, of the average time taken; (ii) measurement of the actual time taken, or costs incurred, by respondents in a pilot survey; (iii) negotiation between the agency and business representatives.
2. Each business (k=1,2, ... ,N) is assessed for its "*response obligation*", a measure which reflects the agencies' assessment of its reasonable relative share of the total burden. The assessment might be a function of the size of the business, for example, and the particular function used might be subject to negotiation with business representatives.
3. The response burden allocation system seeks to ensure that the assessed response obligation of each business is exceeded only rarely by its actual response burden, the latter being, of course, the sum of the response loads of all the survey questionnaires it is obliged to answer within some accounting period.

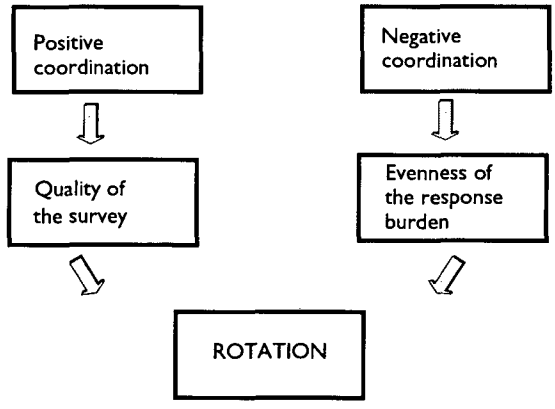
Let β_j denote the response load imposed by the j^{th} survey in a survey program, and π_{kj} the probability of inclusion of the k^{th} business in the j^{th} survey. From the three features presented above, it is now possible to derive an equation for the expected response burden, denoted by RB_k for the business k

$$RB_k = \sum_{j=1}^{M} \pi_{kj} \beta_j , \qquad (7.1)$$

which should not exceed the response obligation of this business.

The formula (7.1) shows that there are two factors that create a response burden: the response load β_j, which could be minimized by the use of administrative data, profiling and a proper survey programme, and π_{kj}, which makes the actual response burden a random variable. This means that the response burden may be distributed unevenly among businesses, many of which suffer in this system which is fair only in that the burden occurs by chance. Thus minimizing the response burden is not enough; steps should also be taken to see that it is distributed as evenly as possible. It is possible to do this by coordinating samples. Coordination can have one of two contrasting aims. Negative coordination aims to achieve the smallest overlap between samples, and so to avoid the situation in which a respondent is burdened with two successive questionnaires, while positive coordination aims to improve the quality of the survey by including successive responses from as many units as possible. Fortunately there is a compromise: the rotation of samples. We will return to the subject of rotation in the next chapter.

Figure 7.1. Coordination of samples



A brief summary of chapter 7

Users of survey data have an insatiable demand for detail. This imposes a burden on respondents which will cause sample fatigue if the burden is not controlled by the statistical agency. A model presented by Sunter includes three main factors: the response load, which express the effort required to complete the questionnaire; the response obligation, which reflects the agency's assessment of it's a given business's reasonable relative share of the total burden the response load and the inclusion probability. The response load and the inclusion probability together form an expected response burden, which must not exceed the response obligation. The inclusion probability makes the response burden a random variable, which means that it is not evenly distributed among businesses. It is possible to make this distribution more even by means of coordination, and the rest of this thesis will be concentrated on the theory of coordination as it applies to business samples.

8.
HISTORY AND THEORY OF SAMPLE COORDINATION IN BUSINESS SURVEYS

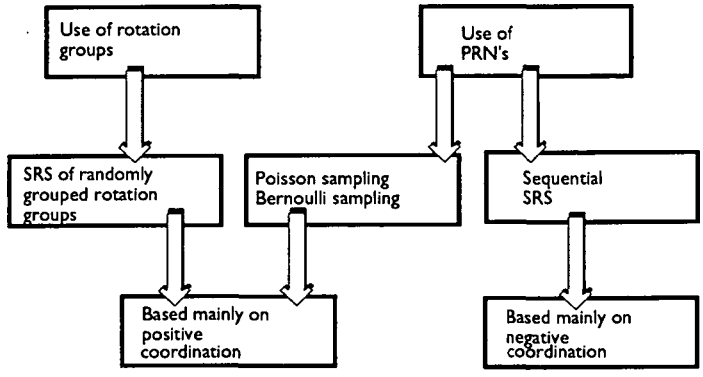
Three types of method for business sample co-ordination will be distinguished below, each of which tries to take into account the following four important features of business surveys:

1. In order to avoid sample fatigue and response burden, the sample units should be changed as often as possible.
2. The structure of the business population is rapidly changing.
3. There are significant changes in strata and inclusion probabilities between successive samples.
4. The distribution of the business population according to the size of the units is highly skewed.

Each of the three methods has its own philosophy, history and application. One of them, Simple random sampling based on the use of randomly formed rotation groups is described in sections 8.1 and 8.2 and its application in 9.1. The other two methods, Poisson (or stratified Bernoulli) sampling and Stratified Sequential Simple random sampling, are based on the use of permanent Random Numbers (PRN). These methods are described in sections 8.3 and 8.4 and their application in 9.1 and 9.2.

These methods also represent two distinct aims of co-ordination, each system apparently having been constructed with a certain primary use in view. The main criterion was initially use for negative coordination or for positive coordination. Even though both criteria are met to a certain extent in each system, we can say that the aim of coordination based on Sequential SRS is mainly negative co-ordination, which means minimizing the overlap between contemporaneous or successive samples, while that in the other two methods is mainly positive coordination, which means maximizing the overlap between successive samples. This division will be used in Chapter 9, where a brief description is given of some existing coordination systems.

Figure 8.1. Grouping the sampling methods with respect to coordination



8.1 Sample Coordination based on Simple Random Sampling in Randomly Formed rotation Groups

The Canadian system of sample coordination is based on rotation groups determined within each strata on criteria of the kind of business activity, geography and size. All the units in a selected rotation group are included in the sample. The primary strata are kind of activity and geography, while secondary strata are formed using some suitable measure of the sizes of units within these primary strata. According to Hidiroglou et al. (1991), this system must have three characteristics. Firstly it should result in samples which reflect the changing structure of the population, secondly it should distribute the response burden by rotating units in and out of the sample, and thirdly, if there are significant changes in the stratification of the universe, it should be possible to redraw a new sample which reflects the changing structure of the population. The third characteristic can be divided into two parts. Firstly, due to the highly skewed nature of the distribution of the business population it is necessary to perform a stratification by size and to divide the population into two parts: a take-all part and a take-some part. An improvement to the theory of setting the cut off points between these parts will be examined in section 8.1.1. Secondly, due to changes in stratification between two dates of sampling, some methods need to be improved to correct for the results of these changes. The improvement of these methods will be examined in section 8.1.2.

8.1.1 Cut-off points in size variables

Because of the highly skewed nature of the business population the secondary strata comprise a certainty part and a part in which the rest of the population is stratified by size. If the distribution of population is extremely skewed the relatively small number of large units will contribute a dominant portion of the total being measured and it could be uneconomical to include any small units in the sample.

The decision when forming the strata is sometimes obvious in view of other decisions made in advance, i.e.the population may be divided into small, medium-sized and large enterprises according to some common principle. If there is no decision in advance, however, the cut-off points should be decided upon by some suitable method.

One of the earliest is the cut-off sampling method introduced by Hansen et al. (1953) for a case where a very small proportion of the units contribute almost the entire aggregate value of the characteristics. This is precisely the case of a population with a highly skewed distribution. This method is usable when the trend in a given auxiliary variable between two dates is similar to that in the larger units and smaller units. In the authors' terminology, a cut-off method is a method for making an estimate that applies to a whole population using only the stratum of larger units and neglecting the other strata, including those composed of small units.

It is essential to this method that there should be data available which include information on a measure of size x_i on two dates i=1,2: where date 1 is a past date and provides an estimate for the whole population, and date 2 is the sampling date. The largest n units on date 2 are selected on the basis of the measure of size on date 1, and the ratio of the total on date 2 to the historical total of the selected large units on date 1 is then computed to obtain a measure of the relationship between the population totals for the two dates. An estimate for the total applying to the current universe on date 2 may then be obtained by multiplying the population total on the past date 1 by the ratio derived for the large establishments. The procedure is as follows

Let $x^{tot}{}_{hi}$ be the sample total for stratum h, h=1,2 on date i, i=1,2 , and let

$$r_h = \frac{x^{tot}{}_{h_1}}{x^{tot}{}_{h_2}}$$

be the ratio of sample totals on dates 1 and 2 for stratum h.

Accordingly, let $X^{tot}{}_{hi}$ be the population total for stratum h, h=1,2 at time i, i=1,2. Then

$$R_h = \frac{X^{tot}{}_{h_1}}{X^{tot}{}_{h_2}} \qquad (8.1.1.1)$$

is the ratio of the population totals in time periods 1 and 2 for stratum h.

Hansen et al. (1953) considered an estimate for the weighted average of the ratio of sampling totals in two strata

$$r = w_1 r_1 + w_2 r_2, \qquad (8.1.1.2)$$

where w_1 and w_2 are the weights for strata 1 and 2 and $w_1+w_2=1$.

After some calculation, they obtained formulae for the optimum values of w_1 and w_2 :

$$w_1 = \frac{S_2^2 - S_1 S_2 + \frac{n}{(X_{12}+X_{22})}(\bar{X}_{12}S_1^2 - \bar{X}_{22}S_2^2)}{(S_2 - S_1)^2}$$

and

$$w_2 = \frac{S_1^2 - S_1 S_2 + \frac{n}{(X_{12}+X_{22})}(\bar{X}_{22}S_2^2 - \bar{X}_{12}S_1^2)}{(S_2 - S_1)^2},$$

where S_1^2 and S_2^2 are approximations for the variances of the ratios R_1 and R_2 (8.1.1.1) in respective samples.

If units are stratified so the large ones form the first stratum and the small ones the second stratum, and if S_1 and S_2 are of roughly the same order of magnitude, then $\bar{X}_{12}S_1$ will be considerably larger than $\bar{X}_{22}S_2$, and as this difference widens, the optimum value of w_1 will increase and w_2 will decrease. When this difference becomes large enough, the optimum values will be approximately $w_1=1$, $w_2=0$ and $n_1=n$.

This method described by Hansen et al. can be used if the trend in a given variable between the two dates is similar for the larger and smaller units. Later, Hidiroglou (1986) introduced an improved method in which the smaller units are sampled, so that similarity between the dates is not necessary.

Before Hidiroglou (1986) introduced his rules, the cut-off method that was widely used was that introduced by Dalenius (1952) and improved by Glasser (1962). Their objective was to minimize the variance of a given auxiliary variable with the given sample size.

Glasser presented his cut-off value as a function of the mean, the sampling weight and the population variance. The parameters he used can be summarized as follows.

	All units	Large units	Small units
Number	N	t	N-t
Mean	$\overline{Y}$	$\overline{Y}_t$	$\overline{Y}_{N-t}$
Sample mean	$\overline{y}$	$\overline{y}_t$	$\overline{y}_{N-t}$
Variance	S^2	S_t^2	S^2_{N-t}
Variance estimator	$\hat{S}^2$	$\hat{S}^2_t$	$\hat{S}^2_{N-t}$

An unbiased estimator for the population mean is

$$\hat{\overline{Y}} = \frac{N-t}{N}\overline{y}_{N-t} + \frac{t}{N}\overline{y}_t ,$$

The variance of the mean estimator takes the form

$$S^2(\hat{\overline{Y}}) = \left(\frac{N-t}{N}\right)^2 \frac{S^2(Y_{N-t})}{n-t}\frac{N-n}{N-t-1} \qquad (8.1.1.3)$$

where $S^2(Y_{N-t})$ is the population variance for small units.
Let $y_1, y_2, \ldots, y_N$ be arranged in ascending order, and let y_m be the value exceeded or equated by only the m largest values. Then

$$y_1 \le y_{N-m} \le (y_m = y^*) \le y_{N-m+1} \le y_N . \qquad (8.1.1.4)$$

The optimum value of Y^* is the value that minimizes (8.1.1.3). The necessary conditions for the optimum point are that

$$S^2(\hat{\overline{Y}}_{(t=m)}) \le S^2(\hat{\overline{Y}}_{(t=m+1)}) \text{ and } S^2(\hat{\overline{Y}}_{(t=m)}) \le S^2(\hat{\overline{Y}}_{(t=m-1)}) .$$

The authors show that, if m is the optimum number of extremes to be included with certainty, then (8.1.1.4) is satisfied if

$$Y^* = \overline{Y}_{N-m} + \sqrt{\frac{N-m}{n-m}}\, S^2(\hat{\overline{Y}}_{N-m}),$$

This is a necessary but not inevitably a sufficient condition for the optimum, as there may exist more than one solution. Glasser gave the following rule for the upper limit:

$$Y^* = \overline{Y} + \sqrt{N/n}\, S^2(\hat{\overline{Y}}) .$$

As the general rule is that it is better to set the cut-off point too high than too

low, this upper limit can provide a good approximation for the optimal cut-off point.

The rules of Dalenius and Glasser predominated in this field until Hidiroglou (1986) presented his cut-off rules for a desired level of precision of estimation. This method can be seen as a generalization of the method of Hansen et al.. Where the objective of Dalenius and Glasser was to minimize the sampling variance for a fixed sample size, that of Hidiroglou was to minimize the sample size for a fixed sampling variance, using the auxiliary variable y as the measure of the size of the units. The sample size is not specified in advance in this method. Hidiroglou expressed the total of a finite ordered population of N units as the sum of the take-all population and the rest of the population N^R. Let

$$U = U^R + U^{TA},$$

where U^{TA} is a population of larger units which are included in the sample with a probability of one and U^R is the rest of population, which are included with a probability of less than one. Let

$$\hat{Y} = \sum_{k=1}^{N-t} y_k + \sum_{k=N-t+1}^{N} y_k = Y^R + Y^{TA},$$

where $y_1 < y_2 < \ldots < y_N$.

There are t large units in the take-all population and $(N\text{-}t)$ small units in the rest of the population.

The sample size of the take-all population t and the overall sample size $n(t)$ are determined so that t units are selected with an inclusion probability of one and the small units are selected from U^R using SRS. The estimator of population total Y is then

$$\hat{Y} = \frac{N-t}{n(t)-1} \sum_{k=1}^{n(t)-1} y_k + \sum_{k=N-t+1}^{N} y_k .$$

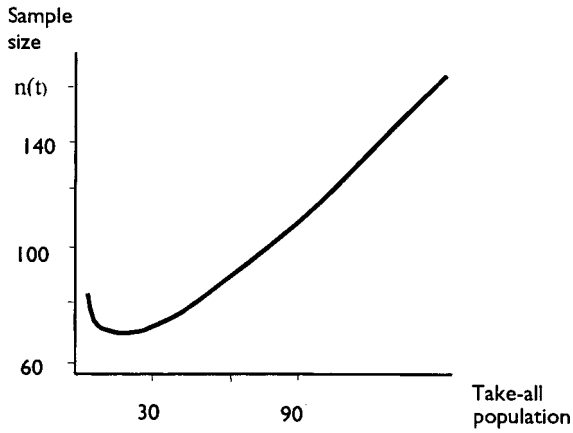
We must now fix the desired level of precision $c = \sqrt{V(\hat{Y})}/\hat{Y}$ for the estimated total, which is the desired coefficient of variation. Once we have calculated the variance for $\hat{Y}$ and substituted $V(\hat{Y}) = c^2Y^2$, we can solve the equation for $n(t)$, which is the overall sample size obtained by adding the required take-some sample to the number of take-all units.

$$n(t) = t + \frac{(N-t)S^2_{(N-t)}}{c^2Y^2 + (N-t)S^2_{(N-t)}} .$$

For c, Y and N as fixed, there exists a minimum for $n(t)$ which is the minimum sample size when stratifying the universe into a take-all stratum and a take-

some stratum. It also gives the optimal size of the take-all part. From the figure (8.1.1.1) below we can see that when the distribution of population is highly skew we can achieve a good total estimate with a small sample size by including most units in the take-all part. In this example the optimum overall sample size *n(t)* is composed of 20 take-all units and 70 representing the rest of the population

Figure 8.1.1.1 Example of sample size versus size of the take-all population with an overall sample size of *n(t)*



Lavallee and Hidiroglou (1988) improved this method further by splitting U^R into two or more strata, the boundaries of which were determined using the *N* and *Y*-proportional power allocation proposed by Bankier (1988) for the rest of the population.

8.1.2 Overlap within strata in successive samples

Another problem in sample coordination based on the Simple random sampling in randomly formed rotation groups is the overlapping or non-overlapping of successive samples. As this method is based on positive coordination, successive surveys are intended to overlap as much as possible. Jumps between strata are then a problem. Nathan Keyfitz (1951) was one of the first to prepare a solution to this problem. As maximization of the overlap between successive samples is the same problem as positive coordination between samples, we can say that Keyfiz was one of the pioneers in the area of sample coordination. A short description will be given below of the Keyfiz procedure, which was prepared for the labour force survey of Statistics Canada.

A sample of one unit is drawn from each stratum h at the time τ using the PPS procedure, and a sample with maximum overlap should be drawn at time $\tau+1$. An unbiased procedure would be to make a new selection within each stratum using PPS for the newly obtained measure of size, but this does not

represent an attempt at maximum overlap.

The first step in the Keyfitz procedure is to check whether the inclusion probability of the unit originally selected has increased or decreased. If it has increased, nothing need be done, but if it has decreased, a change must be introduced. Suppose that a stratum consists of the units 1, 2, 3 and 4. The original inclusion probabilities of these units are P_1, P_2, P_3, P_4 and the new probabilities p_1, p_2, p_3, p_4. Suppose further that $p_1 > P_1$, $p_2 > P_2$, $p_3 < P_3$ and $p_4 < P_4$. The required probabilities of change are $(P_3 - p_3)/P_3$ and $(P_4 - p_4)/P_4$. Let this required probability be 0.07. We then choose a two-digit random number. If this random number lies between 01 and 07 we do not change the original unit included in sample, but if it does not lie in this interval, the unit is dropped from the sample and a new unit (1 or 2) is chosen. The probability of our drawing unit 1 is $(p_1 - P_1)/(p_1 - P_1 + p_2 - P_2)$. Thus, if we obtain a random number which is not greater than this probability we draw unit 1 and otherwise we draw unit 2. A generalization for this special case can be presented as follows. Assume that there are $D + I$ units in the new stratum, where D undergoes decreases in inclusion probability and I undergoes increases.

$P_d < p_d$ and $P_I \geq p_I$, where d=1,2,..,D and I=D+1,D+2,..,D+I

1. If $P_I \geq p_I$, we retain unit k in the sample. Thus it has a new probability P_I of being selected.
2. If $P_d \leq p_d$, there is a probability P_d / p_d that we will retain unit k in the sample. The compound probability of the original draw and this retention is

$$p_d \frac{P_d}{p_d} = P_d . \qquad (8.1.2.1)$$

3. There is a probability $1 - P_d / p_d$ that unit k will be dropped. If this is the case, the probability of our selecting a new unit is

$$\frac{P_I - p_I}{\sum_S (P_I - p_I)} . \qquad (8.1.2.2)$$

The total probability of selecting the k^{th} increasing unit is the sum of the disjoint probabilities for the initial draw and the new draw:

$$p_I + (\sum_I P_I - \sum_I p_I)\frac{P_I - p_I}{\sum_I (P_I - p_I)} = P_I .$$

It follows from (8.1.2.1) and (8.1.2.2) that the desired new probabilities are obtained for both increasing and decreasing units.

Kish and Scott (1971) made some extensions to the Keyfitz method. Important changes in probabilities may be confined to a small proportion of units, those for which P_k/p_k differs greatly from one. Let increases of less than P_k/p_k = 1.1 be deemed insignificant. Then the procedure consists of four steps. Denote the previous probability by p_k and the new probability by P_k, and let D and I be defined as earlier.

1. Calculate the provisional probabilities P_k.
2. A unit for which $P_k/p_k \geq 1.1$ has a new probability P_k of remaining in the sample. Calculate the sum $\sum_I (P_k - p_k)$ of these increases in the stratum.
3. Take enough decreasing units from among those with the smallest values of P_k/p_k to balance the sum of increases exactly, so that $\sum_d (P_d - p_d) + \sum_I (P_I - p_I) = 0$. Assign the probabilities $P_k^* \approx P_k$ to these decreasing units with marginal adjustments to balance the decreases exactly against the increases.
4. For all other units, reassign the old probabilities, $p_k = P_k$.
5. Some units which grow too much are removed from the new stratum and moved elsewhere These units are assigned inclusion probabilities of $p_I = 0$ in their eventual stratum and $P_d = 0$ in their initial stratum.

Kish and Scott also introduced principles for designating the set of initial selection rules in a case where a new stratum is composed of units from several strata.

The method proposed by Hidiroglou and Lavallee is adapted to create the strata, and the procedure proposed by Kish and Scott is used for sample updating when forming the rotation groups for SRS in the randomly formed rotation groups in the Canadian method. The procedure is described in the next chapter.

8.2. Sample Coordination based on Sequential Simple Random Sampling

Simple random sampling was originally designed to be draw-sequential sampling, but list-sequential selection is usually more convenient for a large population of units stored sequentially. Especially if we want to coordinate the units included in different samples, the random numbers procedure requires the list-sequential technique.

One outstanding work in this context was the article of Fan et al. (1962),

who argued that sequential techniques eliminate the need to sort the data according to classification categories prior to sample selection. Computer selection programs based on sequential techniques entail considerable savings in time, because they can utilize data as already arranged, without any need for sorting. All that is required is that the data can be identified as belonging to the sample or not. This method became even more important later, when coordination systems were being prepared for use in business surveys.

One of the proposals made by these authors was the binomial distribution approach, which means that an expected number of n units is selected at random from N population elements. Each unit is selected independently, with a probability π which is equal to the sampling fraction n/N. Each unit is inspected sequentially and the process terminates when all N elements have been inspected.

The procedure is as follows. Each unit k, k=1,2,...,N is assigned a random number r_k from a set of numbers distributed uniformly in the interval (0,1). For each k we test whether $r_k < \pi$. If r_k satisfies this inequality, the unit k is accepted into the sample, otherwise it is rejected.

From the independence of the random numbers r_k it immediately follows that each unit is accepted in the sample independently with a probability π. Thus observations on a binomial distribution with parameters π and N are considered in this procedure.

Another sequential sampling method presented by Fan et al. is more important for the developments in the coordination of business samples. The authors call it the Conditional Distribution Function approach. Each unit is inspected sequentially until n items have been included in the sample. A unit is accepted according to a conditional distribution function that also takes into account the number of units already selected. The procedure is as follows.

Each unit k, k=1,2,...,N is assigned a random number r_k from a set of numbers distributed uniformly in the interval (0,1). The *k:th* unit is accepted in the sample if

$$\frac{n-u}{N-k+1} > r_k, \qquad (8.2.2)$$

where u denotes the number of units already accepted. If (8.2.2) is not satisfied, the *k:th* unit is rejected. The process terminates when $u = n$, which means that n is not random as it was in the Binomial Distribution approach, i.e. Bernoulli Sampling.

The sequential selection rule must clearly provide samples which are equivalent to those obtained by the usual non-sequential methods. This means that the procedure for the Conditional Distribution approach must give a sample which has the same properties as with Simple Randon Sampling. The proof of this can be found in Fan et al. (1962) and Sunter (1977).

In the Conditional Distribution Function approach the population size N is assumed to be known. A "reservoir method" was presented by Cassel (1967)

for the case where N is unknown. Instead of testing every unit, this method searches for the n smallest random numbers. The reservoir method gains its name from the fact that memory space has to be reserved in the computer for the time of sample selection, and this space, which is empty when the procedure starts, is called a reservoir. The procedure consists of six stages:

1. The first n units are read and saved in the reservoir. Each unit is assigned a random number r_k.
2. The units are sorted according their random numbers
3. The unit with the running number $i > n$ is read and assigned a random number.
4. When there are no units left behind to be read into the reservoir, the program stops.
5. If $r_k > r_n$, go to stage 3.
6. If $r_k < r_n$, put the unit k in the reservoir and take the unit n out of the reservoir. k is the unit whose random number is r_k and n is a unit whose random number is r_n.

Later, in 1969, Johan Atmer and Lars-Eric Sjöberg constructed their JALES method, which draws the n units with the smallest random numbers. The method is simple to describe, but in order to prove that it obeys the SRS principle exactly without replacement, it will have to be described in the manner of Atmer et al. (1975).

Let the set $E_{00} = \{e_{00,1} \cdots e_{00,k} \cdots e_{00,N}\}$ be ordered according to some ascending argument of e, and let a permutation $E_0 = (e_{0,1} \ldots e_{0,k} \ldots e_{0,N})$ exist in E_{00}.

The random numbers r_k which are uniformly distributed in the interval (0,1) are ordered according to the same argument of e. Now insert

r_1 into $e_{0,1}$

.....

r_k into $e_{0,k}$

....

r_N into $e_{0,N}$.

When we arrange E_{00} in ascending order of r_k, we obtain the permutation

$$E' = (e'_1 \ldots e'_k \ldots e'_N).$$

All permutations have equal probabilities, and E_{00} can be ordered in the set $N!$. The probability of the permutation E' occurring is then $1/N!$.

Furthermore, let the ordered sets U_k include n ordered elements of N. These n elements can be ordered into *(n!(N-n)!)* permutations of E_{00}.The probability of U_k is then

$$\frac{n!(N-n)!}{N!},$$

which is the probability of an SRS with n elements of population size N. This means that when we order N units according their random numbers in ascending order, any sequence of the n units will constitute an SRS. It is possible to use this property to coordinate the overlap between longitudinal and cross-sectional sample surveys. This important property is one reason why it is used as a base structure in the SAMU system in Statistics Sweden. We will return to these properties in the next chapter.

8.3. History of Sample Co-ordination based on Poisson Sampling

Bernoulli sampling and Sequential Simple random sampling are both equal probability sampling schemes. We will now look at the history of the second approach to controlling the response burden entailed in business samples. This is based on Poisson sampling.

In 1960 Jaroslaw Hajek established necessary and sufficient conditions for the asymptotic normality of estimates based on Simple random sampling from a finite population without replacement. The solution was obtained by approximating Simple random sampling by means of "Poisson sampling". However this was not the first appearance of Poisson sampling. As a matter of fact the annual report on methodology contained in the annual survey of manufactures (1971) states the following:

> "In the selection of new sample panels for the **1959** and 1965 ASM (Annual Survey of Manufactures), each unit was sampled independently of the selection or non-selection of every unit, or combination of units, with the probability of selection varying from unit to unit. For this reason, the ASM sampling procedure is termed **Poisson sampling**, ..."

But let us return to Hajek. In 1964 he used the same method for deriving asymptotic normality conditions for a special kind of sampling with varying probabilities.

There is a particular kind of probability sampling that contains some free parameters which may be controlled by the statistician. These parameters may be related in some way to the size of the units. Hajek assumes that the parameters $\alpha_1,..., \alpha_N$ are non-negative numbers and that

$$\sum_{k\in U}\alpha_k = 1 .$$

Let U be a population consisting of N identifiable units, and let s be a sample which is a subset of U. Rejective sampling of size n can be defined by

$$p_{RE}(s\,|\,n,\alpha_1,...,\alpha_N) = \begin{cases} \chi(n,\alpha_1,...,\alpha_N)\prod_{k\in s}\alpha_k \\ 0 \text{ otherwise} \end{cases}, \qquad (8.3.1)$$

where $\chi(n,\alpha_1,...,\alpha_N)$ is chosen so that $\sum p_{RE}(s|n,\alpha_1,...,\alpha_N)=1$, with s running through all subsets of size n.

Rejective sampling is a realization of n independent draws with fixed probabilities generally varying from unit to unit, given the condition that if the units are not distinct the sample is rejected.

Poisson sampling is defined by

$$p_{PO}(s|n,\alpha_1,\alpha_2,...,\alpha_N) = \prod_{k\in s} n\alpha_k \prod_{l\in U-s}(1-n\alpha_l), \qquad (8.3.2)$$

where $\alpha_k < 1/n$.

Now let $p_1, p_2, ..., p_N$ be the size measures, which are fixed numbers such that $0 < p_k < 1$. We can now define rejective sampling in an equivalent manner to (8.3.1.) as

$$p_{RE} = \begin{cases} C\prod_{k\in s} p_k \prod_{k\in U-s}(1-p_k), \text{ if } s \text{ contains } n \text{ units}, \\ 0 \text{ otherwise} \end{cases}$$

where c is a constant, and Poisson πps sampling as

$$p(s) = \prod_{k\in s} p_k \prod_{k\in U-s}(1-p_k) . \qquad (8.3.3)$$

It can be seen from these equations that rejective sampling may be defined as conditional Poisson sampling. Hajek (1964, 1981) has also shown that rejective sampling may be regarded as sampling with replacement of size n given the condition that the number of distinct units equals n. Conditional Poisson sampling may be regarded as Poisson sampling given the condition that the sample size equals n. If the probabilities $p_1,...,p_n$ are chosen so that $n = \sum_{k\in U} p_k$, the sample size of the rejective sampling will equal that of Poisson sampling.

Poisson sampling and sampling with replacement may both be decomposed into independent sub-experiments, and for this reason they both possess simple variance formulae and simple conditions for an asymptotic normal (Hajek 1964, 1981).

Poisson sampling as defined by Hajék allows each unit in the population to have a given probability of inclusion in the sample. We can also say that all non-rejective samples drawn independently unit by unit as Bernoulli trials with inclusion probabilities p_k are called Poisson samples. There are two special cases of Poisson sampling. The first is the case, where p_k is constant for all k, for example $p_k = n/N$ for all k, where n is the expected sample size. This is called Bernoulli sampling. Another special case is Poisson πps sampling, which is defined as a Poisson sampling design such that

$$\pi_k = \frac{nx_k}{\sum_U x_k}, \qquad (8.3.4)$$

that is, π_k is directly proportional to the size measure x_k.

Why (8.3.4) is the right size measure and why Poisson sampling with unequal inclusion probabilities is more efficient than equal probability sampling schemes require some explanation. Minimizing the variance (8.3.5) introduced below for a fixed expected sample size $n = \sum_U \pi_k$ is equivalent to minimizing the product

$$(\sum_U \frac{y_k}{\pi_k}^2)(\sum_U \pi_k) \text{ (Särndal et al. 1992).}$$

By the Cauchy-Schwarz inequality we obtain

$$(\sum_U \frac{y_k}{\pi_k}^2)(\sum_U \pi_k) \geq (\sum_U y_k)^2$$

and the equality holds only if $\frac{y_k}{\pi_k} = a$ is a constant. Assuming that $y_k > 0$ for all k, we have $\pi_k = \frac{y_k}{a}$.

Finally since $n = \sum_U \pi_k$, we obtain the optimal inclusion probability

$$\pi_k = \frac{ny_k}{\sum_U y_k}.$$

Unfortunately y_k is in general unknown, but if we have an auxiliary variable

X which is closely correlated with Y, we can let p_k be proportional to the known x_k and obtain (8.3.4)

$$\pi_k = \frac{nx_k}{\sum_U x_k}.$$

Poisson sampling that obeys (8.3.4) is referred to below as Poisson πps sampling.

The fact that the random numbers attached to units can be kept permanent means that it is easy to cope with the birth and death of businesses. A random number is attached to a unit on its first appearance in the frame and remains with it until it disappears.

The unbiased Horvitz-Thompson estimator for the population total Y in Poisson πps sampling takes the usual form

$$\hat{Y}^{\pi ps} = \sum_s \frac{y_k}{\pi_k}.$$

As the random numbers r_k are independent, the joint probability of inclusion takes the form $\pi_{kl} = \pi_k \pi_l$. This means that the variance in the HT estimator takes the simple form

$$V(\hat{Y}^{\pi ps}) = \sum_U (1-\pi_k)\frac{y_k^2}{\pi_k}, \qquad (8.3.5)$$

and the unbiased variance estimator the form

$$V(\hat{Y}^{\pi ps}) = \sum_s (1-\pi_k)\frac{y_k^2}{\pi^2_k}.$$

As the sample size is random, it is possible for the realization of sample size to be zero. Brewer et al. (1984) have shown that a ratio estimate that takes the form

$$\hat{Y}^{\pi ps} = \begin{cases} \hat{Y}/n_s, & \text{if } n_s > 0, \text{ where n is the expectation of } n_s \\ 0, & \text{otherwise} \end{cases}$$

is more efficient than the Horvitz-Thompson estimator. The variance for this ratio estimator can be approximated by

$$V(\hat{Y}^{\pi ps}) = \sum_U \pi_k (1-\pi_k)\left(\frac{Y_k}{\pi_k} - \bar{Y}\right)^2 + P_0 \sum_U \bar{Y}^2,$$

where P_0 is the probability of an empty sample.

There are some solutions available for correcting the randomness of Poisson sampling. One of these is the modified Poisson πps sampling introduced by Ogus and Clark (1971), referred to as πpsMO.

In this procedure an ordinary Poisson πps sample is drawn first, but if the sample is empty, a second Poisson sample is drawn, and so on repeatedly until a non-empty sample is achieved. Let π_k be the probability of the unit k being included in the sample. The inclusion probability in each draw, considering the possibility of an empty sample, is $\pi_k(1-p_0^*)$, where P_o denotes the probability of drawing an empty sample, and the second order inclusion probabilities are $\pi_k\pi_l(1-p_0^*)$ for $k \neq l$.

Brewer et al. (1984) give the variance of the HT estimator for $\hat{Y}^{\pi psMO}$ as

$$V(\hat{Y}^{\pi psMO}) = \sum_U (1-\pi k)\frac{y_k^{\,2}}{\pi_k} - p_0^*\left[(\sum_U y_k)^2 - \sum_U y_k^{\,2}\right],$$

and the unbiased variance estimator as

$$\hat{V}(\hat{Y}^{\pi psMO}) = \sum_s (1-\pi k)\frac{Y_k^{\,2}}{\pi k^2} - \frac{p_0^*}{1-p_0^*}(\hat{Y}^{\pi psMO} - \sum_s \frac{Y_k^{\,2}}{\pi k^2}.$$

A more stable estimator can be obtained by multiplying this expression by n/n_s, where $n=E(n_s)$ is the expected sample size.

The difference $V(\hat{Y}^{\pi ps}) - V(\hat{Y}^{\pi psMO})$ is

$p_0^*\left[(\sum_U y_k)^2 - \sum_U y_k^{\,2}\right] > 0$, *if* $p_0^* > 0$, which means that

$V(\hat{Y}^{\pi psMO}) < V(\hat{Y}^{\pi ps})$.

Despite this property, the only advantage of modified Poisson sampling is the non-empty sample. If the sample selected is much smaller (or larger) than the expected sample size, the benefit of modified Poisson sampling is marginal.

Brewer et al. (1972) introduced a procedure which modifies random numbers r_k so that they are uniformly spaced over the interval (0,1). This happens by allocating them according to the following technique:

$$\phi_k = \frac{L_k + r_k - 1}{N}.$$

A random ordering of L_k ($L_k = 1,2,...,N$) is chosen with equal probabilities, and a random variable r_k is selected from a set distributed uniformly in the interval (0,1). A new random number ϕ_k is then calculated for each unit.

The estimation formulae for collocated sampling, that use random number modified like this, is identical to that of Poisson sampling, but the variance formulae are somewhat more complex:

$$V(\hat{Y}^{\pi psCO}) = \sum_U (1-\pi_k)\frac{y_k^{\,2}}{\pi_k} + 2\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l},$$

and the unbiased estimator is given by

$$\hat{V}(\hat{Y}^{\pi psCO}) = \sum_s (1-\pi_k)\frac{y_k^{\,2}}{\pi_k^{\,2}} + 2\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

To see how a collocated sample affects the variation in sample size, we must compare the sample size variances in the two cases. The variance in the sample size for Poisson sampling is

$$V(n_s^{\pi ps}) = n - \sum_U \pi_k^{\,2},$$

and that for a collocated sample is

$$V(n_s^{\pi psCO}) = n - \sum_U \pi_k^{\,2} + \frac{n^2 - \sum_U \pi_k^{\,2}}{(N-1)} - 2\sum_U \frac{(N-1)\pi_k}{N-1}.$$

It is easy to show that $V(n_s^{\pi ps}) > V(n_s^{\pi psCO})$. Collocated sampling thus reduces the variance in sampling size and the probability of drawing an empty sample, but some variability remain.

The method that gives an exactly permanent sample size using unequal sampling probabilities without replacement is **Sequential Poisson sampling,** which is a generalization of sequential Simple random sampling to the case of Poisson πps sampling (Ohlsson E. 1990, 1995, 1998). As we have seen, the unit in Poisson πps sampling is included in the sample if

$$r_k \le \frac{nx_k}{\sum_U x_k}, \qquad (8.3.6)$$

where r_k is the permanent random number given to the unit k, n is the expected sample size and x_k is the size variable for the unit k.
To obtain the same sample size, which is exactly n, Ohlsson introduced a normed random number

$$\eta_k = \frac{r_k}{x_k}. \qquad (8.3.7)$$

According to (8.3.7), the unit k is then included in the sample if

$$\eta_k \leq \frac{n}{\sum_U x_k}. \qquad (8.3.8)$$

The right-hand side of (8.3.8) is constant, and the adjustment of n_s to n adjusts this constant. Adjusting $E(n_s)$ until we obtain a sample size n^0 is equivalent to including the n^0 units with the smallest normed random numbers η_k in the sample.

We can also present this by introducing

$$p_k = \frac{x_k}{\sum_U x_k}. \qquad (8.3.9)$$

We can give an alternative expression for (8.3.6) as follows:

$$r_k \leq np_k$$

Let ξ_k be

$$\xi_k = \frac{r_k}{Np_k} \qquad (8.3.10)$$

We get

$$\xi_k \leq \frac{n}{N}.$$

We draw exactly n units from the beginning of the sampling line. We can do this by sorting the file in ascending order of ξ_k and drawing the first n units from the beginning. Ohlsson calls this procedure Sequential Poisson Sampling. Some units are moved to a take-all stratum, which does not alter the sample but must be properly handled in the estimation process.

Because of the sorting, the procedure is neither list-sequential nor draw-sequential in the sense of Särndal et al. (1992).

Sequential Poisson sampling is not a πps sampling. Nevertheless, because of its close relation to Poisson sampling, as can be seen above, it is approximately πps sampling. The simulation studies of Ohlsson (1995) support this conclusion. It then follows that

$$\hat{Y}^{SEpps} = \frac{1}{n}\sum_s \frac{y_k}{\pi_k}$$

is approximately normally distributed, with mean $\overline{Y}$ and variance σ^2.
In the equal probability case Sequential Poisson Sampling is simply Ordered Simple random sampling.

One problem with Sequential Poisson Sampling is that no closed expression can be given for the first-order and second-order inclusion probabilities. Hence the standard theory for unbiased estimators cannot be used and we must rely on approximations.

The family of fixed sample size order πps sampling designs introduced by Rosen (1996a,b) includes two important cases of sampling coordination, Sequential Poisson Sampling as presented above and Pareto sampling. The family is defined as follows.

Let there be associated with each unit k, k=1,2,...,N, a probability distribution F_k in the interval [0,∞). Order sampling with sample size n and the order distribution $\boldsymbol{F} = (F_1,F_2,...,F_N)$, denoted by $OS(n,\boldsymbol{F})$, is carried out by introducing the independent random variables $Q_1,Q_2,...,Q_N$, called ranking variables, with distributions $F_1,F_2,....,F_N$. The units with the n smallest Q-values constitute the sample.

Let $H(t)$ be a probability distribution of density $h(t)$, and let $\theta=(\theta_1, \theta_2, ... , \theta_N)$ be real numbers. An $OS(n,\boldsymbol{F})$ scheme is said to have fixed order distribution shape $H(t)$ and intensities θ, if either of the following two equivalent conditions is met:

(i) The ranking variables $q_1,q_2,...,q_N$ are of the type $q_k=z_k/\theta_k$, k=1,2,...,N, where $z_1,z_2,..,z_N$ are independent, identically distributed random variables with a common distribution H.

(ii) The order distribution is $F_k(t)=H(t\,\theta_k)$, $0 \le k < \infty$, k=1,2,...,N.

Sequential Poisson Sampling, as defined by Ohlsson, is an order sampling with $z_k=r_k$ and $\theta_k=p_k$ and

$$q_k = \frac{z_k}{\theta_k} = \frac{r_k}{p_k} = \eta_k ,$$

where r_k is the permanent random number of the unit k and p_k and η_k are defined as in (8.3.9) and (8.3.10).

Pareto sampling is an order sampling scheme with $z_k = r_k/(1-r_k)$ and $\theta_k=\lambda_k/(1-\lambda_k)$, where $\lambda_k=np_k$ and

$$q_k = \frac{r_k(1-\lambda_k)}{\lambda_k(1-r_k)} .$$

We note that λ_k is the desired inclusion probability for a Poisson πps sampling of an expected size n without replacement. Rosen has shown that Pareto πps sampling is optimal in the class $OS(n;F)$, in that it gives the minimum variance for the HT estimator.

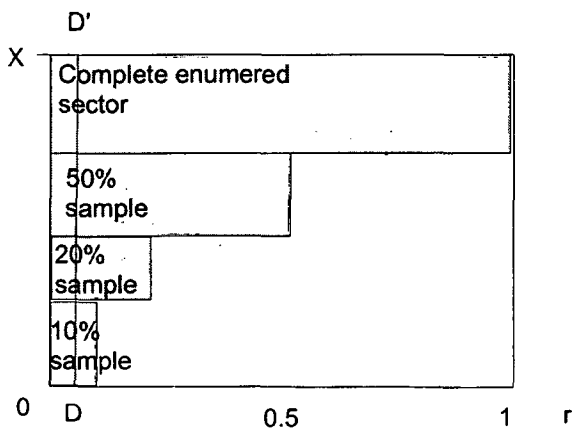
8.4. The Constant Shift method in rotation

Rotation is a procedure which makes the panel survey more fair, releasing some units from the preceding survey and replacing these with new ones. A rotation round is a sequence during which all units are surveyed once.

A compromise between negative and positive coordination is to rotate units in successive samples so that the response burden is evenly distributed among them all. A rotation system based on rotation groups which did not use PRNs was described in section 1.1. Using the constant shift method it is possible to rotate units in succeeding samples using procedures based on use of the PRN technique.

Set the constant shift value, D, at zero for the first round of a repeated survey and increase it by a fixed amount before each subsequent round. This should cause the small units to be rotated more rapidly than the larger ones. The overall proportion of the population rotated depends on the size distribution of the units and is an increasing function of the value D. In the case of Bernoulli sampling stratified according to the measure of size, rotation operates as follows.

Figure 8.4.1. Rotation using Bernoulli sampling stratified by size



The inclusion probabilities π_k are functions of the size measures x_h in the stratum h. The points in Figure 8.4.1 specified by the random number r_k and the size measure x_k correspond to units in the population. Units are included in the sample if $\pi_k > r_k$. Rotation takes place when we shift the sampling area to the right by the constant shift interval D. This means that we release units lying to the left of the vertical line DD'. As can be seen, the proportion of the rotation is the greater the smaller is the sampling fraction. Let D=0.05. The rotation proportion will then be 50%in the lowest sample stratum, 25% in the next sample stratum and 10%in the next. That in the complete enumerated sector will be 0%. The small units are rotated more rapidly than the large ones, as should be the case. Let us now move on to Poisson πps sampling.

Rotation can be carried out in two ways using Poisson πps sampling. In Figure 8.4.2 below it takes place by rotating the sampling area around the origin. This gives all the units the same rotation ratio. In Figure 8.4.3 the constant shift method is used, which gives faster rotation for small units. Although it bypasses some small units in every rotation round, it must be regarded as the better means of evening out the response burden. We will return later to the problem of bypassed units.

Figure 8.4.2. Rotation around the origo

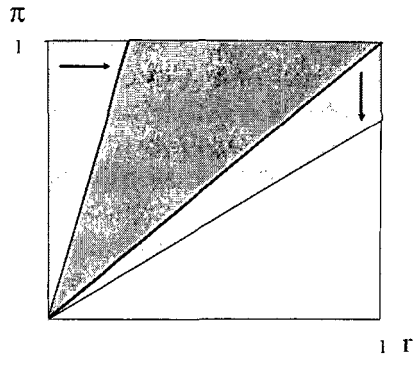
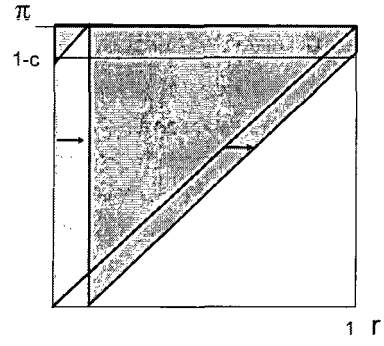


Figure 8.4.3. Constant shift method



To formulate the constant shift method, let the required selection probability be π_{kj} , which is the probability of unit k, $k=1,...,N$, being included in survey j, $j=1,...,J$. Unit k will then be selected in the survey j if $\pi_{kj} \geq r_k$.

By control we mean here the capacity to influence the probability of one unit being included in successive surveys by changing the constant shift interval D. Consider the joint probability of a unit being included in both samples j = 1,2. The probabilities of inclusion of a unit k are π_{k1} and π_{k2}. Assume that $\pi_{k2} \geq \pi_{k1}$ and make the realistic assumption that $D \leq 1 - D$. It follows that the joint inclusion probability in successive surveys will take the modified form proposed by Sunter (1977).

$$\pi_{k12} = \begin{cases} 2\pi_{k2} - 1, & if \quad 1\text{-}D < \pi_{k2} \leq 1 \\ \pi_{k2}\text{-}D, & if \quad D \leq \pi_{k2} \leq 1\text{-}D \\ 0, & if \quad 0 \leq \pi_{k2} \leq D \end{cases} . \qquad (8.4.1)$$

As we can see, there is an area where the inclusion probability with respect to each sample is zero. This means that we do not achieve the panel effect, repetition of units, for the smallest units. PoMix sampling as introduced by Särndal, Kröger and Teikari (1999) eliminates this problem.

It is easy to handle the formation and closure of businesses in a rotation based on Poisson sampling. The frame has to be updated before each draw, new units being assigned a PRN of r_k and the no longer existing units together with their PRN's being removed from the frame.

A brief summary of Chapter 8

There are three methods available for coordinating business samples: one based on rotation groups and other two on permanent random numbers (PRN). A brief history of these methods is presented in this chapter. Some important ways to stratify skewed data are presented, and also some methods for mastering the problem of changes between strata occurring in the time

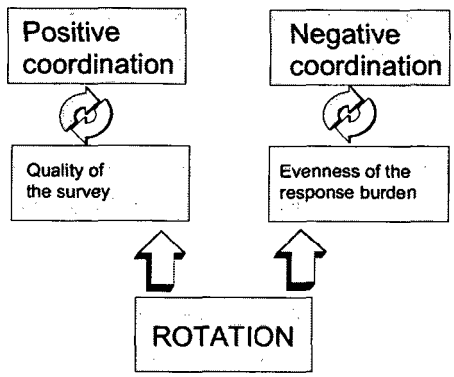
elapsing from one sampling to the next. These observations are presented in the context of rotation groups, but they are important in other methods too. An application of Simple random sampling in randomly formed rotation groups is described in section 9.1. The rest of this thesis will concentrate on the use of PRNs. The idea of using random numbers in sequential sampling comes from Fan et al. (1962) and was elaborated on by Cassel and later by Atmer and Sjöberg. The latter also introduced the JALES method, subsequently termed sequential Simple random sampling.

The method of Poisson sampling introduced by Hajek in 1964 forms the basis of PoMix sampling, which was introduced by Kröger, Särndal and Teikari (1999) and is one of the main themes pursued in the remaining part of this thesis.

9.
APPLICATIONS OF SAMPLE COORDINATION

A general theory of sampling design was presented briefly in Chapter 2 and sampling schemes for use in sample coordination were presented in more detail in Chapter 4. This chapter presents some existing coordination systems. All of them work for both positive and negative coordination, but the emphasis on positive coordination is greater in some systems and that on negative coordination greater in others. Rotation is a compromise between these two extremes of emphasis, as described in the figure below.

Figure 9.1 Coordination of samples



The smallest overlap between samples, and thus the greatest evenness in the distribution of the response burden, is achieved by negative coordination, Whereas positive coordination improves the quality of the survey. A compromise is therefore required, in the form of rotation.

9.1 Applications primarily based on the Positive Coordination

There are two systems which rely mainly on positive coordination of business samples: the Canadian system, based on the use of rotation groups, and the system used in New Zealand, based on the use of permanent random numbers in a Bernoulli sampling scheme. The system employed in Australia was based on equal probability collocated sampling until 1983, since then it has been based on the synchronized sampling technique.

The Canadian co-ordination system is based on the same theory of "Simple Random Sample of randomly formed rotation groups" that was

presented in the last chapter. The response burden is distributed by rotating the units in succeeding surveys using a rotation groups method. The system is described by Hidiroglou et al. (1991) and Hidiroglou and Srinath (1993). The units in each stratum are grouped into a certain number of rotation groups, depending on the sampling fraction in the stratum and the time-in / time-out constraints. The procedure is described below.

According to the Hidiroglou method described in the previous chapter, the population is stratified into a certainty part Y^{TA} and the rest of the population Y^{R}, the latter being further stratified into parts, called take-some strata, of sizes N_h. Let $f_h = n_h/N_h$ be the desired sampling fraction in the take-some stratum h. Denote by t_{in} the desired number of occasions on which the unit should stay in the sample, and by t_{out} , the minimum required number of occasions on which it should stay out of the sample once it has been rotated out. If there is no time-out constraint, then the number of rotation groups is simply determined by multiplying the inverse of the sampling fraction f_h^{-1} by the number of occasions on which the units is to be in the sample. This cannot, however, ensure that the unit stays out of the sample for the desired period after it has been rotated out. The following method, which ensure that units stay out of the sample on at least t_{out} occasions, also determines the number of rotation groups G_h in the take-some stratum population and the number of rotation groups g_h in the sample G_h.

Compute

$$X_h = \text{int}\left|\frac{1-f}{f} t_{in} + 0.5\right|,$$

where *int* /* / denotes the integer part of the argument. Two conditions arise:

1. If $X_h > t_{out}$, which means that the integer part of the argument is greater than the minimum number of occasions on which the unit should stay out of the sample, then the number of rotation groups in the sample is

   $g_h = t_{in}$

   and the number of rotation groups is

   $G_h = t_{in} + X_h$.

2. If $X_h < t_{out}$, which means that the integer part of the argument is smaller than the minimum number of occasions on which the unit should stay out of the sample, then the number of rotation groups in the sample is

$$g_h = \text{int}\left|\frac{1-f_h}{f_h} t_{out} + 0.5\right|$$

and the number of rotation groups is

$$G_h = g_h + t_{out}.$$

Once the number of rotation groups has been determined, we must allocate the units to these groups. The rotation groups are initially numbered 1,2, ... ,G, and this is the order in which they are selected in the sample. We shall call this the rotation ordering. A random permutation of this ordering we will call the assign ordering . The method of allocation depends on whether the number of units N_h is greater or less than the number of rotation groups G_h.

1. If $N_h \geq G_h$, then at least one unit can be allocated to each rotation group. The 1^{st} unit is assigned to the 1^{st} rotation group, the 2^{nd} unit to the 2^{nd} rotation group ... P^{th} unit to the P^{th} rotation group, the $(G+1)^{th}$ unit to the first rotation group and so on. Let $N_h = mG_h + q$, where $m > 0$ and $q > 0$ are integers. Then the first q rotation groups will include $(m+1)$ units each and the last $(G - q)$ rotation groups will includes m units.

2. If $N_h<G_h$ the rotation groups which are non-empty must be determined. These groups must be as equispaced as possible, to ensure that the expected sample size $n_h = f_h N_h$ will be achieved. A random number ρ is selected between 1 and G_h / N_h. Then the rotation groups

   $$\rho,\ \rho + G_h / N_h ,\ \rho + 2G_h / N_h ,\ldots,\ \rho + (N_h - 1)G_h/N_h$$

   will be selected to be non-empty. A random permutation of units N_h will be performed, After which the 1^{st} unit will be assigned to the 1^{st} non-empty rotation group, the 2^{nd} unit to the 2^{nd} non-empty rotation group, and so on. Finally all the N_h units have been assigned to N_h non-empty rotation groups.

After the units have been allocated to rotation groups, the rotation using these groups is simple. The groups numbered 1 to g_h in rotation ordering are selected for the sample on the first occasion, and on the second occasion group 1 is dropped and group number $g_h + 1$ is added. The non-empty rotation groups are rotated so that at the same as an empty group is dropped a new empty group is included. On these occasions no rotation takes place.

All enterprise births falling into the take-all stratum are certain to be

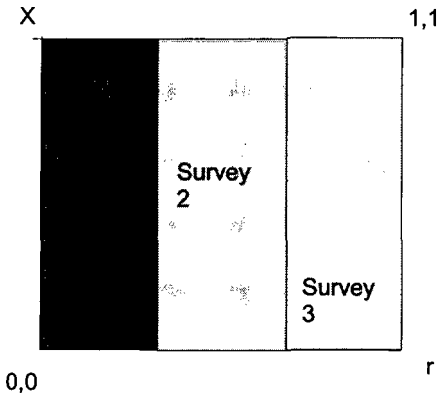
surveyed, while those falling into the take-some stratum are stratified and given an assign ordering number. Assume that the last assignment rotation number was q, here $0 \leq q \leq G_h$. The first new enterprise will be given the assign rotation number $(q+1)$ in the case $N_h \geq G_h$, and subsequent ones will be assigned to rotation groups starting from the next to it. If $N_h < G_h$, there will only be N_h rotation groups. Empty panels are never assigned new enterprises.

The changes in the classification variables are reflected in the estimation process by the use of a domain estimate. Within one time survey, if a change is found after the selection, the latest classification will be assigned to the data and the weight originally assigned will be retained for estimation purposes. For tabulation purposes, however, the unit will belong to the new domain. Over a period of time the changes in classification may become sufficiently important to require examination of the stratification and subsequent sampling rates. To maximize the overlap between the current sample and the new sample, an adaptation of the Kish and Scott (1971) method presented in Chapter 8 can be used.

The shortcoming of this method is that it is much more complicated than the constant shift method.

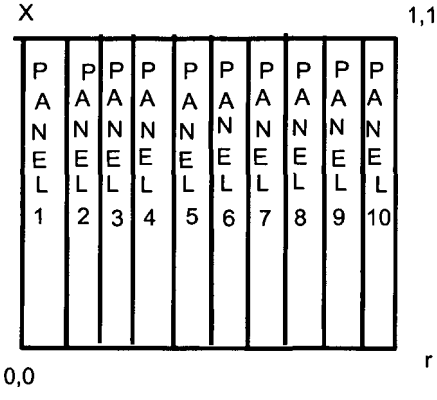
The **New Zealand Department of Statistics** has adopted a panel-based approach to sample designs for business surveys. This employs Bernoulli sampling. Each unit is assigned a Permanent Random Number (PRN) in the interval (0,1). To control the overlap and rotation a Poisson chart is used which has the PRNs of the units on the horizontal dimension and the size of the unit on the vertical dimension. The parts of the survey are marked out in the chart, so as to ensure that no unit is selected for more than one sample in a given survey period.

Figure 9.2 Parts of the survey defined in the Poisson chart (New Zealand)



When no unit is to be included in the sample for more than an expected number of survey periods, the Poisson chart is divided into rotation groups (panels) according the random numbers of the units.

Figure 9.3. Rotation groups in the Poisson chart (New Zealand)



Initially the sample may include panels 1 and 2. After one year, if it is wished to rotate out half of the original sample, panels 2 and 3 are taken.

The method looks the same as the JALES-method as used in the SAMU system (described later), but it uses Bernoulli sampling where SAMU is based on the sequential SRS sampling procedure.

The Australian coordination system has been based since 1982 on the Synchronized Sampling System. The selection mechanism is adapted to the JALES method. Each unit is given a random number $r \in$ unif(0,1), and the first n units are selected for the first sample using the sequential SRS scheme. For the next selection we describe this sample with an interval open to the

right $[a_1,e_1)$. This is used as a trial sample interval, and if it contains exactly n units it is approved. Otherwise, if it has been affected by the formation or closure of units, the starting or end point has to be moved. If $[a_1,e_1)$ contains more than n units, the starting point is moved to the right, and if it contains less than n units, the end point is moved to the right. The same procedure is employed if the sample size has changed.

When we want a rotation to take place we partition the first sample into rotation groups of sizes $n_1,\ldots, n_R$ with starting points $p_0=a_1, \ldots, p_R= e_1$. At the next selection, $[p_1,e_1)$ is used as a trial interval and process goes on as earlier.

When a unit changes stratum, synchronized sampling treats it as a case of death in the old stratum and a case of birth in the new one.

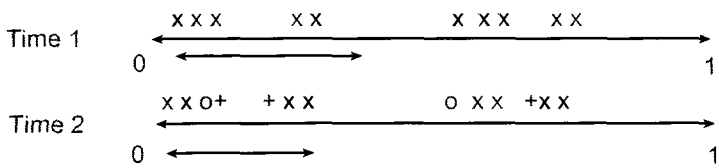
9.2 Applications primarily based on the Negative Coordination

The JALES technique for sampling coordination was described in the previous chapter. The SAMU system used by Statistics Sweden, as described by Thulin I (1976) and Ohlsson (1995), is based on this technique. In the SAMU system each unit is assigned a random number drawn independently from a uniform distribution (0,1). As the random numbers apply permanently to their units, they are called Permanent Random Numbers (PRN). The frame units are arranged in ascending order of their PRNs. New units are assigned PRNs when they enter the frame and are placed in order accordingly.

The sample design used in the SAMU system, known as Sequential SRS allows both negative and positive coordination when used with PRNs. Although negative coordination has been considered more important in this system, it is useful to begin by describing positive coordination.

In two successive samples we start both draws from the beginning of the sampling line using the same PRN's. Denote the persistent units by x, closed units by o and newly formed units by (+).

Figure 9.2.1. Positive coordination in SAMU -system



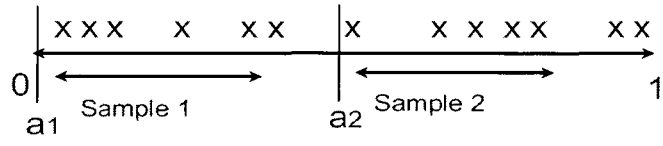
Since the PRNs are not equispaced, the intervals between the random numbers of the units on the line (0,1) will not be equal. It is this property that creates the random sample size in Bernoulli and Poisson sampling. In sequential SRS the sample size is always equal to the expected sample size, because we draw exactly n successive units.

We draw a sample of five units at time 1 and again at time 2 and hope to have the overlap between them as great as possible. We can see that due to the formation of new units and closure of old ones there is natural rotation

between samples. If, say, one of the first five units has closed between time 1 and time 2 and two units have entered the scheme, the sample that we draw at time 2 will contain three constant units and two new ones.

When we co-ordinate samples for cross-sectional surveys, we hope that the overlap between them will be as small as possible. To reduce the overlap between two samples, we choose two constants a_1 and a_2 in the interval (0,1). To make it sure that the two samples do not overlap, the interval between the constants should be 0.5. Then choose n units to the right (or left) beginning from the points a_1 and a_2.

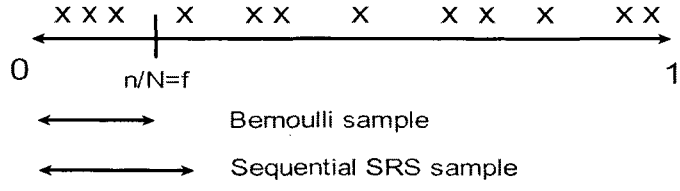
Figure 9.2.2. negative coordination in SAMU



When we use sequential SRS, we need not group units in panels as in the New Zealand case. Samples can be rotated by drawing the first n units beginning from the origin and placing a_1 for the next sample at the $(n+1)^{st}$ unit. This gives the maximum rotation. Different rates of rotation can be obtained by locating a_2 between units 1 and (n+1).

The difference between Sequential SRS and Bernoulli sampling is described below.

Figure 9.2.3. Difference between Bernoulli sampling and Sequential Simple SRS



Here the size of the population is $N = 12$ and the expected sample size is $n = 4$, so that the inclusion probability of a unit in both Sequential SRS and Bernoulli sampling is n/N = 1/3. Sequential SRS includes exactly 4 units, but Bernoulli sampling includes all the units which have been assigned a PRN < 1/3. In this case the actualized sample size using Bernoulli sampling is $n = 3$.

A system called EDS is used in Statistics Netherlands to coordinate business samples. The sampling frame is a file extracted from the Central Business Register and the units are enterprises. Coordination takes place by accumulating a measure of the response burden in the sampling frame. Comparable to the response load index introduced by Sunter (1977) to measure response burden, as described in chapter 7, the EDS system measures the response burden by the estimated time required to complete the questionnaire (van Huis, Koeijers, de Ree 1994), arranged in six classes, as below.

Table 9.2.4. Classes of response burden assessed in terms of questionnaire completion time

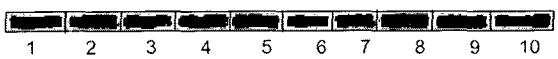
Class	Completion time (min)	Response burden
1	1 – 30	1
3	31 – 60	2
2	61 – 120	4
4	121 – 180	6
5	181 – 240	8
6	241 –	10

Response burden values per enterprise are accumulated in the sampling frame. If an enterprise is selected, the value of this questionnaire is added to the enterprise's total.

Since each enterprise is assigned a PRN in the interval (0,1), the enterprises are recorded in the framework together their identification number, PRN, a size class, code for the branch of economic activity concerned and RB total. Before sampling, the enterprises in each stratum are sorted into ascending order of their RB total and in the case of equal RB totals according to their PRN.

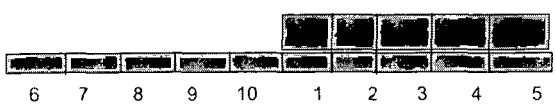
Before selection of the very first sample the enterprises will have been sorted entirely according to their PRNs. In Figure 9.2.4. we suppose there are ten enterprises.

Figure 9.2.4. Enterprises placed in random order



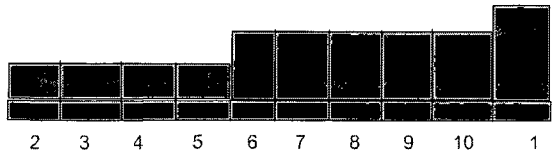
Let the sampling fraction n/N be 5/10. The first 5 units are then selected and gain an RB value according the completion time of the questionnaire (say 2). Before the next selection the units in the frame are sorted as follows.

Figure 9.2.5 Enterprises after first selection



Let the sampling fraction of the next selection be 6/10, and let these be given the RB index 4. The least burdened enterprises are drawn. The sequence of units in the frame after the second selection is presented in Figure 9.2.6. The units are first arranged in ascending order of their response burden load, and if the loads are identical, according to their random numbers.

Figure 9.2.6 Enterprises after the second selection



It can be seen that after the same rounds the enterprises are purely in ascending order of their response load, so that it remains obscure whether the system is based on PRN usage at all.

For the rotation of samples, the EDS creates a file containing the identification numbers of the enterprises that were included in the sample selected previously for the same survey. This is done by means of a dummy variable which indicates whether the enterprise was included in that sample. The inclusion probability for each enterprise is included.

The rotation fraction determines the proportion of the total number of enterprises that can be relieved of participation in the survey. The rotation fraction must be selected in the interval [0,1]. If the rotation fraction is 0, the resulting sample will have maximum overlap with the previous sample, whereas the value 1 will result in minimum overlap.

The Statistical Agency of France (INSEE) maintain a system of coordinated selection of stratified samples called **OCEAN**. This system, described by Cotton and Hesse (1992), is based on assigning units random numbers which are recalculated after each selection. The unit can be either an enterprise or an establishment. Units to be surveyed are selected by Simple random sampling Stratified by size (SSRS).

To describe the system, we can consider the selection of two samples s^1 and s^2. Assume that no changes in the population occur between these occasions. Denote by n^{1h} and n^{2h} the numbers of units we want to select from stratum h for samples 1 and 2 respectively. Each unit k is assigned a random number, $r^{1h}(k)$, selected in the interval (0,1) and arranged in the stratum h according to their size measure. We then determine for each stratum h a random starting point, c^{1h}. When we reach the endpoint of the line (0,1), we choose (n_{1h}-r) units to the right from the starting point c_{1h} and the remaining r units to the right of the origin.

To obtain a negative coordination, the selection sequence for the next sample must be moved to the right, as in the SAMU system. OCEAN takes into account the possibility of changes in strata by using an algorithm which moves the selected units to the end of line (0,1) and selects sample s^2 from the beginning of the line. The procedure is as follows. The random numbers of the units in sample s^1 are changed using the transformation that moves the units selected in the first sample to the end of the stratum h and when the second sample is drawn from the beginning of this stratum there is no overlapping. Maximum rotation takes place when the random numbers of all units included in s^1 are transformed, and different rotation rates can be achieved by transforming only some of the random numbers. A pure panel approach would make this equal to the SAMU system.

The coordination system used in Statistics Finland is named OTKO, an acronym for the Finnish words meaning sample coordination. OTKO is not yet very widely used, because it is a rather new system, in addition to which extensive use is made of administrative data and only the largest enterprises are surveyed. OTKO is based on the use of PRNs and permits various sampling designs, including PoMix sampling, which is described in the rest of this thesis. Each unit is given the response index 100 when it comes into the frame, and every questionnaire reduces this index. When it falls below a fixed level the unit is removed from the frame for a time, and when it is reintroduced as a new unit it receives the response index 100 again.

A brief summary of Chapter 9

This chapter describes briefly some existing coordination systems which are based on the theory and history described in chapter 8. The rest of this thesis will be concerned with describing PoMix sampling.

10.
POISSON MIXTURE (POMIX) SAMPLING

Poisson sampling, as we have seen, has some important strengths as regards sampling coordination, but also certain weaknesses. Its advantages include the ease with which it allows us to handle the formation of new units and elimination of old ones, to estimate totals and ratios, and to handle co-ordination in general, because it is based on PRN. The weaknesses of Poisson sampling include the fact that sample sizes are random and that some small units are missed in each round of the rotation (more on this later). Moreover, the remaining small units are included only once, so that some longitudinal information on small enterprises is missed. The contribution of the smallest enterprises to the total estimates is not significant, but sometimes longitudinal information is required on small and medium-sized enterprises (SMEs). Furthermore, it is unfair that some enterprises should be required to give statistical information and others not.

Some solutions to the first weakness, i.e. random sample size, were suggested in Chapter 2. However, as Sunter (1986) writes,

> "There is usually no particular reason to insist on a fixed n. In most cases, the almost inevitable nonresponse would make such a requirement absurd."

Let me leave this problem aside for the time being and address the second problem instead. The solution proposed for this is the Poisson Mixture (PoMix) sampling scheme (Kröger, Särndal and Teikari 1999). We will return to the problem of random sample size in Chapter 12, where a form of fixed sample size PoMix sampling, known as order PoMix sampling, is introduced.

10.1 Two special cases of Poisson Mixture sampling

Poisson Mixture (PoMix) sampling can be described as a general case of rotating successive samples. It is a family of sampling schemes in which Poisson equiprobability sampling and Poisson πps sampling represent the two extreme cases. All the other members of the family are mixtures of these extremes. Following Särndal et al. (1992), I will use the term Bernoulli sampling to refer to Poisson equiprobability sampling (see Chapter 5). The rectangular area which includes the units belonging to Bernoulli sampling is called the Bernoulli part. PoMix sampling is based on Permanent Random Numbers (PRN) attached to each unit, and it uses sample rotation by the constant shift method.

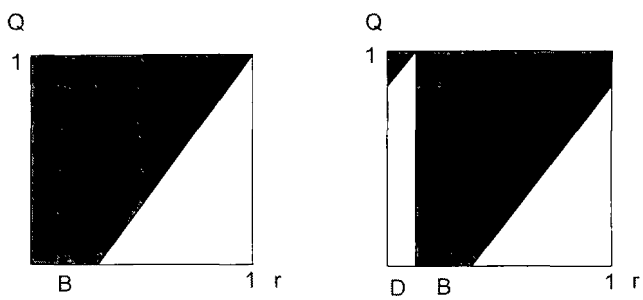
The expected sample size in PoMix sampling consists of n^{BE} units drawn from the Bernoulli part and n^{PO} units drawn from the rest of the sampling area. It is thus

$$E(n) = n^{BE} + n^{PO}. \qquad (10.1.1)$$

A permanent random number (PRN), r_k, is assigned to each population unit. This number is generated from the uniform distribution *Unif(0,1)*. A size measure Q is calculated for every unit in such a way that the sampling area is rectangular, the PRN being the horizontal dimension and the size measure Q the vertical dimension. For simplicity, let Q be the size measure normed so that there are no negative values and no values exceeding 1.

Figure 10.1.1 (A.) below presents the general case of PoMix sampling in a two-dimensional area, where each unit k is represented by a point (r_k, Q_k), where r_k (on the horizontal axis) is the random number and Q_k (on the vertical axis) is the size measure. The rotation for the next sampling case is shown in Figure 10.1.1 (B.). The starting-point is moved to the right by the constant shift D. Thus the starting-point for the third occasion will be 2D, and so on

Figure 10.1.1 Sample rotation using PoMix sampling, $D < B$.

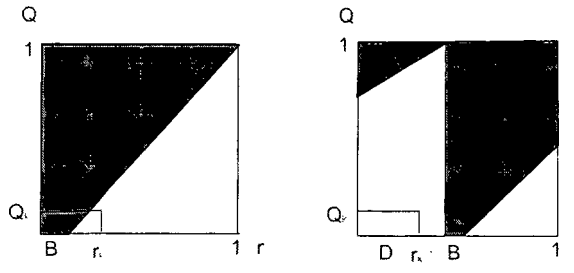


A. Sampling at time 1

B. Sampling at time 2

The width of the Bernoulli part, B, and the width of the constant shift, D, are fixed. The units included in the first draw lie in the shaded area of Figure 10.1.1 (A), and those included in the second draw in the shaded area of Figure 10.1.1 (B). As we can see, there are no units in the area above $D+B$, which has zero probability of being included in the sample drawn at time 1 or at time 2. This is because the width of the constant shift, D, is smaller than the width of the Bernoulli part, B. In the opposite case, presented in Figures 10.1.2 (A) and (B), unit k has zero probability of being included in the sample drawn at time 1 or at time 2.

Figure 10.1.2 Sample rotation using PoMix sampling, $D > B$.



A. Sampling at time 1 B. Sampling at time 2

The size of the area where the probability of being included in either of these successive samples depends on the difference between the width of the Bernoulli part (B) and the width of the constant shift (D) is highest in the case of Poisson πps sampling. When $B \geq D$, this area vanishes. On the other hand, in the extreme case (which is Poisson πps rotation with $B = 0$), the base of this triangle has the length of the constant shift D.

10.2 A take-all stratum and introduction of the size measure Q

We assumed in section 7.1 that the size measure Q must be normed so that there are no negative values and no values greater than one. To achieve this, one or more take-all units must first be identified. If the population is highly skewed, as is often the case in business populations, some units are selected with a probability of one because they contribute most to the total estimates. Some procedures were described in Chapter 4 for determining this take-all stratum. The following procedure is used here. Let n be the expected sample size, fixed in advance. We assume that the largest units are assigned to a take-all stratum, denoted U^{TA}, to be selected with a probability of one. Let n^{TA} be the size of U^{TA}. The rest of the population is

$$U^R = U - U^{TA},$$

of size

$$N^R = N - n^{TA}.$$

A random sample is to be selected from U^R, of expected size

$$n^R = n - n^{TA}.$$

Using (8.3.4) we define a size measure for unit k in U^R as follows:

$$A_k = \frac{n^R x_k}{\sum_{U^R} x_k}. \qquad (10.2.1)$$

We have $0 \le A_k < 1$ for all $k \in U^R$. This follows from the principle used to construct the take-all stratum U^{TA}, which we will now describe.
First we compute

$$A_k^* = \frac{nx_k}{\sum_U x_k}, \text{for } k \in U .$$

If $A_k^* \ge 1$ for some units k, say n^0 units, then these units are assigned to a preliminary take-all stratum U^0. The procedure is then repeated to see if additional units should be assigned to this stratum. Compute

$$A_k^{**} = \frac{(n-n^0)x_k}{\sum_{U-U^0} x_k} .$$

Those units k for which $A_k^{**} \ge 1$ are also included in the take all stratum. The procedure is repeated until no further units are assigned to the take all stratum. As a result we have a take-all stratum U^{TA} of size n^{TA} while the rest of the population is denoted by U^R.
We can now assume that

$$A_k = \frac{(n-n^{TA})x_k}{\sum_{U^R} x_k} < 1$$

for all $k \in U^R$, because if $A_k \ge 1$ had been true for any of these units they would have been assigned to the take-all stratum.
We can now rewrite 10.1.1 using U^R as the sampling frame:

$$n^R = n^{RBE} + n^{RPO} = BN^R + n^{RPO} \qquad (10.2.2)$$

where B is the width of Bernoulli part, N^R is the size of the rest of the population and n^R is the expected size of the sample taken from the rest of the population. Correspondingly n^{RBE} is the expected size of the Bernoulli sample from the rest of the population and n^{RPO} of that of the Poisson πps sample.
We fix the constant B so that $0 \le B \le f^R$, where

$$f^R = \frac{n^R}{N^R} \quad \text{and} \quad B = f^{RBE} = \frac{n^{RBE}}{N^R}.$$

Using (10.2.1) and (10.2.2), we define the size measure Q_k for each unit $k \in U^R$ as follows:

$$Q_k = \frac{n^R - N^R B}{1-B}\frac{x_k}{\sum_{U^R} x_k} = \frac{N^R(f^R - B)}{1-B}\frac{x_k}{\sum_{U^R} x_k} = \frac{(f^R - B)}{(1-B)}\frac{x_k}{\bar{x}^R}, \qquad (10.2.3)$$

where

$$\bar{x}^R = \frac{\sum_{U^R} x_k}{N^R}.$$

From the fact that that $A_k = \frac{n^R x_k}{\sum_{U^R} x^k} < 1$ for all $k \in U^R$ we can prove that $Q_k < 1$ for all $k \in U^R$. Rewrite (10.2.3) as follows:

$$Q_k = \frac{1 - B/f^R}{1-B} A_k.$$

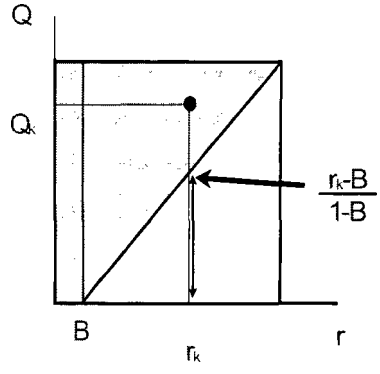
Since $A_k < 1$, we have

$$Q_k = \frac{1 - B/f^R}{1-B} A_k < 1 \text{ for all } k \in U^R.$$

10.3. Algorithm for PoMix sampling

Let us concentrate for a moment on time one only. Using PoMix sampling, unit k is included in the sample if $k \in U^R$ and it falls in the shaded area in Figure 10.3.1 below.

Figure 10.3.1 PoMix sampling



Thus unit k is included in the sample if

$$0 < r_k \leq B$$

or if

$$B < r_k \leq 1 \text{ and } Q_k \geq \frac{r_k - B}{1 - B}.$$

It follows that the probability of the inclusion of unit k in the sample consists of two conditional probabilities: the conditional probability that the unit will be drawn when $r_k \leq B$ (Bernoulli sampling) and the conditional probability that the unit will be drawn when $r_k > B$ (Poisson πps sampling). The inclusion probability π_k thus is given by

$$\pi_k = P(k \in s) = P(k \in s | r_k \leq B)P(r_k \leq B) + P(k \in s | r_k > B)P(r_k > B)$$

It is easy to see that $P(k \in s | r_k \leq B) = 1$, because we select all units for which $r_k \leq B = n^{RBE} / N^R$,
where n^{RBE} is the expected sample size in the Bernoulli part.

It is also clear that $P(r_k \leq B) = B$, which means that B is the inclusion probability of unit k with respect to the Bernoulli part.
It is also easy to see that

$$P(r_k > B) = 1 - B,$$

but the inclusion probability with respect to the first part of Poisson πps sampling is somewhat more complicated:

$$P(k \in s | r_k > B) = P(Q_k > (r_k - B)/(1 - B) = P(r_k \leq B + Q_k(1 - B)).$$

We showed in section 10.2 that $0 \leq Q_k < 1$ for all k included in U^R.
We can now see that, because
$[Q_k > (r_k - B)/(1 - B) \Leftrightarrow r_k < B + Q_k(1 - B)]$, the inclusion probability can be presented as

$$\pi_k = B + P(B < r_k < \min(1, B + Q_k(1 - B)). \qquad (10.3.1)$$

From (10.3.1) we get the first order inclusion probability for unit k in the rest of the population:

$$\pi_k = \begin{cases} B + Q_k(1 - B), & \text{if } k \in U^R \\ 1 & \text{if } k \in U^{TA} \end{cases}. \qquad (10.3.2)$$

It follows that unit k is included in the sample if

$$0 < r_k \leq Q_k(1 - B).$$

The second order inclusion probabilities are easy to compute, because r_k and r_l are independent random numbers. We have

$$\pi_{kl} = \pi_k \pi = [B + Q_k(1-B)][B + Q_l(1-B)], \text{ for } k \neq l \in U_R \text{ .(10.3.3)}$$

We use the name PoMix sampling because (10.3.1) can be written as

$$\pi_k = \frac{B}{f^R} f^R + (1 - \frac{B}{f^R}) A_k \text{ . From this equation we obtain}$$

$$\pi_k = \begin{cases} A_k, \text{ if } B = 0 \\ f^R, \text{ if } B = f^R \end{cases}, \text{ for all k} \in U^R .$$

Thus, if we set $B=0$, we obtain Poisson πps sampling, and if we set $B=f^R$ we obtain Bernoulli sampling.

10.4. Estimation based on a PoMix sample

The unbiased Horwitz-Thompson estimator for a PoMix sample is easy to obtain by replacing (6.7) with (10.3.2)

$$\hat{Y} = \sum_s \frac{y_k}{\pi_k} = \sum_{k \in s} \frac{y_k}{[B + Q_k(1-B)]} .$$

However, as both Bernoulli and Poisson sampling are list sequential samplings with random sample size *n,* the unbiased HT estimator gives a large variance. It is therefore better to use an alternative which gives approximately unbiased estimators but much smaller variances. The first such alternative is called the weighted sample mean (Särndal et al. 1993). An approximately unbiased estimator for the population mean is

$$\tilde{Y}^s = \frac{\sum_s y_k / \pi_k}{\sum_s 1/\pi_k},$$

where the numerator is the HT estimator for the population total, $\hat{Y}$, and the denominator is the HT estimator for the population size, $\hat{N}$. Multiplying this by the known population size *N* we obtain the first alternative estimator for *Y*

$$\hat{Y}^{alt1} = N\tilde{Y}^S = N \frac{\sum_s y_k / \pi_k}{\sum_s 1/\pi_k} = N \frac{\hat{Y}}{\hat{N}}$$

To estimate the approximate variance for PoMix sampling we must keep in

mind that in the case of independence the population covariance (π_{kl} - $\pi_k\pi_l$) is equal to zero. This means that we have the following general formula for variance estimator

$$\hat{V} = \sum_s \frac{1}{\pi_k}(\frac{1}{\pi_k} - 1)(g_k e_k)^2, \qquad (10.4.1)$$

where g_k is the g-weight, which incorporates the auxiliary information into the estimator, and $e_k = y_k - \hat{y}_k$. In this case the g-weight is $N/\hat{N}$ and

$$e_k = y_k - \frac{\hat{Y}}{\hat{N}} = y_k - \tilde{Y}^s.$$

We thus have

$$\hat{V}(\hat{Y}^{alt1}) = \frac{N^2}{\hat{N}^2}\sum_s \frac{1}{\pi_k}(\frac{1}{\pi_k} - 1)(y_k - \tilde{Y}^s)^2.$$

If we have auxiliary data which are correlated with the variable of interest, it is possible to include these in the estimator by means of the g-weight. We can now try to improve the estimation by using the regression estimator

$$\hat{Y}^{alt2} = \mathrm{X}\hat{b} = \sum_s \frac{1}{\pi_k} g_k y_k,$$

where $\mathrm{X} = \sum_U x_k$, $\hat{b} = \dfrac{\sum_s \frac{y_k}{\pi_k}}{\sum_s \frac{x_k}{\pi_k}}$ and $g_k = \dfrac{\mathrm{X}}{\sum_s \frac{x_k}{\pi_k}}$.

For the variance estimator we have $e_k = y_k - \hat{b}x_k$. Thus the estimator for the approximated variance takes the form

$$\hat{V}(\hat{Y}^{alt2}) = \frac{X^2}{\hat{X}^2}\sum_s \frac{1}{\pi_k}(\frac{1}{\pi_k} - 1)(y_k - \hat{b}x_k)^2,$$

where

$$\hat{X}^2 = (\sum_s \frac{x_k}{\pi_k})^2.$$

10.5. Joint probability in two successive samples

As we have seen, the rotation of Poisson samples causes some units to be missed in each rotation round. Where PoMix sampling is used, it is possible to control the number of units that have zero probability of being included in two successive samples s^1 and s^2.

Let π_{k12} be the probability that of unit k being included in both s^1 and s^2 and let the Bernoulli interval be B and the constant shift D ($B,D \le \frac{1}{2}$). Depending on the relation between D and B, we have two cases.

CASE 1:

$$D \le B \le \frac{1}{2}.$$

In this case we can show that the probability of inclusion in two successive samples takes the form

$$\pi_{k12} = P(k \in s^1 \wedge k \in s^2) = \begin{cases} 1 - D - (1-B)(1-Q_k), if\ 0 \le Q_k < 1 - \frac{D}{1-B} \\ 1 - 2(1-B)(1-Q_k), if\ 1 - \frac{D}{1-B} \le 1 \end{cases}$$

Corresponding, for

CASE 2: $$B \le D \le \frac{1}{2},$$

we can show that the probability is

$$\pi_{k12} = P(k \in s^1 \wedge k \in s^2) = \begin{cases} 0, if\ 0 \le Q_k < \frac{D-B}{1-B} \\ 1 - D - (1-B)(1-Q_k), if\ \frac{D-B}{1-B} \le Q_k < 1 - \frac{D}{1-B} \\ 1 - 2(1-B)(1-Q_k), if\ 1 - \frac{D}{1-B} \le Q_k \le 1 \end{cases} \quad (10.5.1)$$

As can be seen from (10.5.1), we have $\pi_{k12} = 0$, if $B \le D, \text{and}, 0 \le Q_k < \frac{D-B}{1-B}$. We can also see that by changing B, we can control the number of units which have a zero probability of being included in two successive samples. This means that by changing B we can add or cut units included in two or more successive cases. It is also possible to study changes with respect to small enterprises.

Equating $B=0$ in (10.5.1) we reach the same result as Sunter (1977) on page 56. We must recall that $Q_k = A_k$ if $B = 0$.

$$\pi_{k12} = P(k \in s^1 \wedge k \in s^2) = \begin{cases} 0, & if\, 0 \le Q_k < D \\ 1 - D - (1 - Q_k), & if D \le Q_k < 1 - D \\ 1 - 2(1 - Q_k), & if\, 1 - D \le Q_k \le 1 \end{cases}$$

This means that bysetting $B=0$ in PoMix rotation we achieve Poisson πps rotation.

10.6 Order PoMix sampling

If a random sample size is considered undesirable, this can be avoided by using the method of Ohlsson (1990, 1996, 1998) that was described in chapter 8.3 above. In this method the inclusion probabilities are normed so that we can draw a Poisson πps sample by choosing exactly the n units with the smallest normed random numbers. As was shown in chapter 8.3, the sequential Poisson sample was achieved as follows. The condition for inclusion in the Poisson πps sample was

$$r_k \le A_k = n^R p_k, \qquad \text{where} \qquad p_k = \frac{x_k}{\sum_{U^R} x_k}. \quad (10.6.1)$$

The normed random numbers η_k are obtained by dividing both sides by x_k

$$\eta_k = \frac{r_k}{p_k} \le n^R.$$

In the case of order PoMix sampling we must first introduce a modified size measure

$$A_k^{\,\mathrm{mod}} = B + (1 - \frac{B}{f^R}) A_k = B + (1 - \frac{B}{f^R}) n p_k$$

which is a linear transformation of A_k.

Because $A_k^{\,mod}$ is a modification of A_k (by linear transformation), we cannot use exactly the same procedure as Ohlsson for order sampling. We define

$$\zeta_k = \frac{r_k}{A_k^{\,\mathrm{mod}}} = \frac{r_k}{B + (1 - \frac{B}{f^R}) A_k}. \quad (10.6.2)$$

There are two interesting special cases (10.6.2). If B is zero, we have just the order Poisson sampling of Ohlsson (r_k/A_k). If $B = f^R = n^R / N^R$ we obtain the result r_k/f_R. As PRNs divided by a constant give the same order of units as the original PRNs, we are in effect performing Sequential Simple random sampling with constant sample size.

A brief summary of chapter 10

Poisson Mixture (PoMix) sampling was introduced to improve Poisson rotation, which ignores some small units and does not give any longitudinal information for others. This was done by adding part of the equiprobable Bernoulli sampling procedure to that of Poisson πps sampling. The take-all stratum and the size measure Q were introduced and a sampling algorithm was prepared. The joint probability of inclusion in two successive samples in PoMix sampling was found to be exactly the same as that obtained earlier by Sunter with a Bernoulli part of width zero.

To ensure a fixed sample size, order PoMix sampling was introduced on the basis of the method described by Ohlsson. The resulting sampling led to two interesting observations. When the value of the Bernoulli part was set at zero, we obtained Ohlsson's sequential Poisson sampling. Otherwise,when the size of the Bernoulli part was set at exactly n/N, we did not get Bernoulli sampling but Sequential Simple random sampling.

Unexpectedly, we found that the addition of part of the equiprobable Bernoulli sampling procedure to Poisson πps sampling reduced the variance of the estimators that used auxiliary information. The next chapter describes simulation studies carried out on order PoMix sampling.

11.
A MONTE CARLO STUDY OF POMIX SAMPLING

A Monte Carlo experiment was conducted in order to see the effects of the Bernoulli width on PoMix sampling. For this purpose we used a real population of 1,000 Finnish enterprises. For enterprise k, $k = 1, ..., 1,000$, y_k is the number of employees in the enterprise and x_k is the total wages and salaries paid by the enterprise. In Chapter 12 we use an artificial data set to test these effects with different skewnesses of the population distributions. The 1,000 units were selected randomly from an originally larger population of enterprises. Each unit k is assigned a random number r_k.

Since the assignment of PRNs to population units is a random procedure, a proper Monte Carlo study also requires repetitions of PRN assignments in the same population. Therefore, for each 100 assignments of PRNs, 100 samples were drawn using PoMix sampling with a fixed value of the Bernoulli width B. The Monte Carlo experiment thus consisted of 100 × 100 = 10,000 combinations. For each combination, we computed four point estimators, the corresponding four variance estimators, and the corresponding four confidence intervals. To establish the effect of the Bernoulli width B, we carried out one of these experiments (with 10,000 combinations) for each of a range of values of B situated in the interval $0 \leq B \leq f_R$, , where $f_R = n_R/N_R$ denotes the expected sampling rate in U_R. The simulation results are obviously not free of Monte Carlo error, but it is fair to argue that the 10,000 combinations give sufficient reliability.

The Monte Carlo population had the following characteristics: The total y to be estimated was $Y = \sum_U y_k = 169{,}168$, and the expected sample size was fixed at $E(n) = 100$. The procedure described in Chapter 10.2 was used to determine the take-all part n^{TA}, and the formula (10.2.1) to calculate inclusion probabilities for the units in the take-some part. This resulted in a take-all stratum of 29 units and a take-some part consisting of 971 units, with a population total $Y^R = 46.138$ employees. This gives a take-some population comprising 97.1% of the units but accounting for only 27.3 % of the total population (= 169 168 employees). The coefficient of variation (see 6.1) is 1.78 for the variable y and 1.94 for the variable x; the correlation coefficient between y and x being 0.965.

Three estimators of the population total were compared. The experiment involved repeated draws of samples as well as repeated assignments of PRN's to the N population units.

In the following expressions for the three estimators, $a_k = 1/\pi_k$ denotes the sampling weight of unit k, where π_k is given by (10.3.2).

We can expect the simulation to show that $\hat{Y}^{GREG}$ (6.11) and $\hat{Y}^{SRAT}$ (6.10), which use auxiliary information at both the design stage and the estimation stage, will improve the HT estimator $\hat{Y}^{HT}$ (6.9), which uses auxiliary information only at the sampling stage, but the extent of the improvement is unpredictable and interesting to observe.

In a survey, auxiliary information can be exploited at the sampling stage, at the estimation stage, or at both stages. With PoMix sampling, auxiliary information is used at the design stage; the extent to which this occurs being dependent on the value of B. Despite this use of auxiliary information, it is expected in our simulation that $\hat{Y}^{GREG}$ (6.11) and $\hat{Y}^{SRAT}$ (6.10), which make further use of the auxiliary variable at the estimation stage, will perform better than the HT estimator, $\hat{Y}^{HT}$, which only benefits from auxiliary information at the sampling stage.

The simulation for a range of different values of B was carried out in such a way that the maximum value of $B = n^R / N^R$ was $71/971 = 0.0073$, which gives the Bernoulli sample. For each value of B = 0, 0.01, ..., 0.07 10 000 PRN/Sample pairs were produced. The results were used to calculate five Monte Carlo summary outcomes for these four point estimators. The simulation results are shown in Tables 11.1, which displays the Monte Carlo coverage rates (MCRTE) for four estimates with nominal 90% (MCRTE90) and 95% (MCRTE95) confidence intervals, and Table 11.2, which displays two simulation quantities for each of the four estimators, namely:

(1) $MCV(\hat{Y})$ = Monte Carlo variance of the point estimator $\hat{Y}$, i.e. the variance of the 10 000 point estimates

(2) $MVE(\hat{V})$ = Monte Carlo expectation of the variance estimator $\hat{V}$, i.e. the arithmetic mean of the 10 000 variance estimates;

Table 11.1.

Monte Carlo coverage rates for four estimates with nominal 90% (MCRTE90) and 95% (MCRTE95) confidence intervals

Bernoulli width D	**MCRTE95**			**MCRTE90**		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	94.50	92.75	92.48	89.70	87.35	86.74
0.010	95.20	93.47	93.52	90.43	87.93	88.02
0.020	95.06	93.88	93.88	90.36	88.49	88.55
0.025	95.06	94.56	94.70	90.64	89.73	89.72
0.030	94.63	94.09	94.19	89.85	88.70	88.86
0.040	93.84	94.47	94.64	88.77	89.41	89.60
0.050	93.97	93.76	93.82	88.67	88.08	88.53
0.060	93.54	92.12	92.69	89.10	85.99	87.27
0.070	92.93	90.67	92.03	88.40	84.27	86.11
0.073	91.03	88.03	90.46	86.53	81.26	83.86

We do not know the bias ratio presented in (6.1) and we cannot check the coverage probability presented in (6.2) and (6.3). The results of the Monte Carlo simulation in Table 11.1 indicate that all three estimators for MCRTE90 and MCRTE95 are close to their theoretical values, which are 90% and 95%, respectively. Only for $\hat{Y}^{GREG}$ and $\hat{Y}^{SRAT}$, and when B comes close to the upper limit, is it possible to see any marked tendency for the MCRTE to drop below the nominal value.

Preliminary tests indicated that the variance is not smallest in the case of Poisson πps sampling with $B=0$. This surprising observation prompted the addition of a Bernoulli width of 0.025 to the tables. Table 11.2 shows the Monte Carlo variances for four estimates. The Monte Carlo expectations for the four point estimators are not shown because they are all very close to the target parameter value $Y = 169\ 168$.

Table 11.2.

Results of Monte Carlo simulation for different Bernoulli widths B.

Bernoulli Width B	$MCV(\hat{Y})*10^{-6}*$			$MCE(\hat{V})*10^{-6}$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	24.56	3.43	3.46	24.92	3.43	3.46
0.010	22.74	1.84	1.85	23.53	1.86	1.87
0.020	24.75	1.77	1.78	25.37	1.77	1.78
0.025	25.51	1.78	1.79	26.86	1.81	1.82
0.030	28.03	1.80	1.81	28.58	1.87	1.88
0.040	35.17	2.03	2.06	33.54	2.11	2.15
0.050	42.25	2.64	2.67	41.42	2.51	2.59
0.060	56.08	3.65	3.67	55.70	3.24	3.44
0.070	90.73	5.47	5.59	91.28	4.72	5.37
0.073	119.13	7.09	7.43	116.27	5.34	6.49

Table 11.2 shows very little difference between, $\hat{Y}^{GREG}$ and $\hat{Y}^{SRAT}$. By contrast, $\hat{Y}^{HT}$ has considerably greater variance. This confirms that the HT estimator is a poor choice compared with the alternative that uses a closely correlated auxiliary variable. This is true most particularly for Bernoulli sampling, but it is also the case for B values near the lower end of the interval $\left[0, f^R\right]$, which shows that the sampling design alone does not remove all the power from the auxiliary variable, even though we are close to Poisson πps sampling (B=0).

The variance estimator performs well in the sense that $MCE(\hat{V})$ is generally very close to $MCV(\hat{Y})$, which measures the variance of $\hat{Y}$. This holds for all estimators and all values of B, with a few notable exceptions, namely, in the case of $\hat{Y}^{GREG}$ when B is close to the upper extreme (Bernoulli sampling). Then the variance estimators for these two estimators only clearly overestimate the true variance.

Interestingly, the minimum variance for $\hat{Y}^{HT}$, $\hat{Y}^{REG}$ and $\hat{Y}^{SRAT}$ is not obtained for Poisson πps with $B = 0$, as one might expect, but rather for a value of B apparently somewhere between 0.02 and 0.03. The improvements achieved in the case $B = 0.02$ relative to $B = 0$ are substantial for $\hat{Y}^{CRAT}$, $\hat{Y}^{GREG}$ and $\hat{Y}^{SRAT}$. As we can see in Table 11.3, the variance ratio

$$\frac{(\mathrm{MCV}(\hat{Y})\,|\,\mathrm{B}=0.02)}{(\mathrm{MCV}(\hat{Y})\,|\,\mathrm{B}=0)}$$

is close to 50% for $\hat{Y}^{GREG}$ and $\hat{Y}^{SRAT}$. More precisely, this ratio is 0.52 for $\hat{Y}^{GREG}$ and 0.51 for $\hat{Y}^{SRAT}$. Once the simulation results had been examined, we carried out one additional simulation for the case B = 0.025. This confirmed that, with our data set, the minimum variance occurs at around this value of B. The results in Table 11.3 confirm that the minimum variance for two estimators $\hat{Y}^{GREG}$ and $\hat{Y}^{SRAT}$ is obtained at a point which lies in the neighbourhood of B=0.025. One possible explanation for this surprising result is that when B is close to zero, the units with the lowest x values, when selected, will have unduly large weights, which implies high variability. This can be avoided by choosing a B that is a long way from zero.

Table 11.3.

Results of simulations with different Bernoulli widths, *B*. Improvement of Pomix sampling for different Bernoulli widths *B*. $MCV(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC expectation for the variance estimator $\hat{V}$. Each row in Table 11.2 is divided by the first row

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.93	0.54	0.53	0.94	0.54	0.54
0.020	1.01	0.52	0.51	1.02	0.52	0.51
0.025	1.04	0.52	0.52	1.08	0.53	0.53
0.030	1.14	0.52	0.52	1.15	0.55	0.54
0.040	1.43	0.59	0.60	1.35	0.62	0.62
0.050	1.72	0.77	0.77	1.66	0.73	0.75
0.060	2.28	1.06	1.06	2.24	0.94	0.99
0.070	3.69	1.59	1.62	3.66	1.38	1.55
0.073	4.85	2.07	2.15	4.67	1.56	1.88

The Monte Carlo results showed, somewhat surprisingly, that PoMix sampling combined with a regression estimator is more efficient for certain Bernoulli widths B within the interval $0 \leq B \leq f^R = n^R/N^R$ than for $B = 0$, corresponding to Poisson πps sampling. One possible explanation is that this result is caused by our particular data set. To check this we examined the Taylor linearized variance of $\hat{Y}$ (see Särndal et al. 1992, Ch. 6):

$$V^{TAY} = \sum_{U^R} (a_k - 1) E_k^{\ 2}$$

where $a_k = 1/\pi_k$ and E_k is the population analogue of the sample-based residual e_k used in the variance estimator (10.4.1). This residual for the estimator $\hat{Y}^{GREG}$, for example, is

$$E_k = y_k - b^{GREG} x_k \qquad (11.1)$$

with

$$b^{GREG} = \frac{\sum_{U^R} (a_k - 1) y_k x_k}{\sum_{U^R} (a_k - 1) x_k^{\ 2}}.$$

The residual (11.1) comes from the regression model

$$\begin{cases} E(y_k) = \beta x_k \\ V(y_k) = \sigma^2 x_k \end{cases}.$$

It is reasonable to model the squared residual pattern as

$$E_k^{\,2} = \sigma^2 x_k^{\,p}(1+\delta_k), \qquad (11.2)$$

whereupon, using an approximation

$$E_k^{\,2} \approx \sigma^2 x_k^{\,p}$$

and (10.3.2.), we have

$$V^{TAY} = \sigma^2 \sum_{U^R} (a_k - 1) x_k^{\,p} = \sigma^2 H(B,p) \sum_{U^R} x_k^{\,p},$$

where

$$H(B,p) = \bar{x}^R \sum_{U^R} \frac{x_k^{\,p}}{B\bar{x}^R + (f^R - B)x_k},$$

where

$$\bar{x}^R = \sum_{U^R} x_k / N^R.$$

Consider the fixed value of p in the interval $0 \le p \le 2$. We wish to find out whether $H(B,p)$ has a smaller value for some B within the interval $0 \le B \le f_R$ than at $B=0$, which is $H(0,p)$.

We find

$$H'(B,p) = \bar{x}^R \sum_{U^R} \frac{x_k^{\,p}(x_k - \bar{x}^R)}{(B\bar{x}^R + (f^R - B)x_k)^2},$$

and the value of which at $B = 0$ is

$$H'(0,p) = (\bar{x}_{U^R} / f_R^2) \sum_{U^R} x_k^{\,p-2} (x_k - \bar{x}_{U^R}). \qquad (11.3)$$

The sign of (11.2) is the same as that of $\sum_{U^R} x_k^{p-2} (x_k - \bar{x}_{U^R})$. But this quantity equals the covariance between x_k^{p-2} and $x_k - \bar{x}_{U^R}$ in U_R (note that $x_k - \bar{x}_{U^R}$ has zero mean), apart from the factor $1/(N_R - 1)$. When p satisfies $0 \leq p < 2$ this covariance is negative: when x_k increases, $x_k - \bar{x}_{U^R}$ increases steadily, but x_k^{p-2} decreases steadily (and always remains positive). The sign of $H'(B,p)$ is therefore negative; and consequently, it is not at $B = 0$ that H(B,p) attains its minimum value but at some B in the interval $\lfloor 0, f^R \rfloor$. For $p = 2$, $H(0,p)$ now has its minimum at $B = 0$.

These considerations raise the question of whether the population used for the simulation corresponds to a value of $0<p<2$, but distinctly less than 2, so that we can expect significant gains from PoMix sampling. To obtain an answer, we estimated p by fitting the logarithmic version of the model (11.2) to the data available for $U^R = U - U^{TA}$. That is, we fitted

$$w_k = a + pz_k,$$

where

$$w_k = log(E_k^2),$$

and

$$E_k = y_k - b^{SRAT} x_k$$

with

$$b^{SRAT} = \frac{\sum_{U^R} y_k}{\sum_{U^R} x_k};$$

and

$$z_k = log(x_k).$$

We obtained the value p=1.45 by treating p as a linear regression slope estimated as

$$p = \frac{\sum_{U^R} (w_k - \bar{w}^R)(z_k - \bar{z}^R)}{\sum_{U^R} (z_k - \bar{z}^R)^2}.$$

Since the value of p is considerably less than 2, our Monte Carlo population is indeed one in which one can expect significant gains from the use of PoMix sampling with a value of B within the interval $\lfloor 0, f^R \rfloor$.

A brief summary of chapter 11

A Monte Carlo simulation of PoMix sampling was carried out. A total of 10,000 sampling combinations were drawn from the frame of 1000 enterprises, which were derived at random from the Business Register. The findings confirmed that the addition of some part of the equiprobable Bernoulli sampling procedure to Poisson πps sampling reduces the variance of the estimators when auxiliary information is used. An examination of the Taylor linearized variance showed that the parameter p, measuring the heterogeneous error term, must be smaller than 2 in order to achieve the desired results. Linearizing the error term using a logarithmic version, we obtained the value of p 1.45 for our data.

12.
MC SIMULATION STUDIES OF POMIX SAMPLING IN TWO WEIBULL-DISTRIBUTED ARTIFICIAL POPULATIONS

12.1 Preparing the artificial data sets

We have seen that it is advantageous to add some part of the equiprobability sampling procedure to that of PoMix sampling. That is, by taking $B > 0$, the estimators based on auxiliary information show less variance than is the case when $B = 0$. If the heteroscedasticity is very large, $p \geq 2$, we lose this improvement. We are now interested in the distribution of the auxiliary variable x. To see the impact of PoMix sampling on data with a very high skewness and kurtosis we prepared two artificial data sets (x_k, y_k), $k = 1,...,N$, one with the values x_k following an exponential distribution and the other with a more skewed distribution.

The Weibull distribution with parameters $\alpha > 0$ and $c > 0$ is defined by the distribution function

$$F(x) = P(X \leq x) = 1 - e^{-cx^{\alpha}} \qquad \text{for } x > 0.$$

If we set $\alpha = 1$, we get the exponential distribution with parameter c,

$$F(x) = 1 - e^{-cx}, \tag{12.1.1}$$

with the density function

$$f(x) = ce^{-cx}. \tag{12.1.2}$$

Setting $\alpha = 1/2$ we get

$$F(x) = 1 - e^{-c\sqrt{x}}, \tag{12.1.3}$$

with the density function

$$f(x) = \frac{c}{2}\frac{1}{\sqrt{x}}e^{-c\sqrt{x}}. \tag{12.1.4}$$

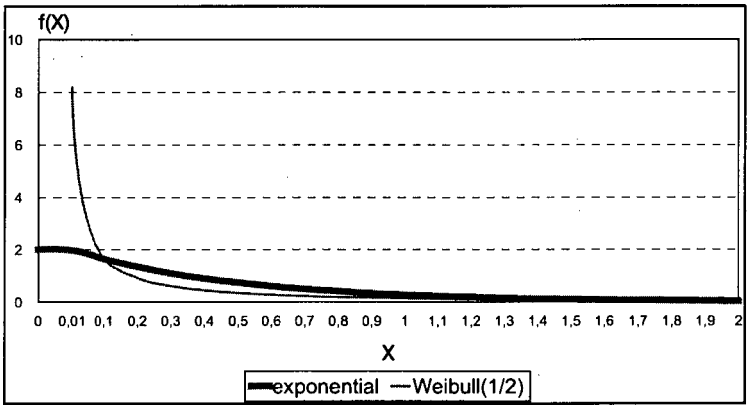
We decided to fix c so as to equalize the expected values of (12.1.2) and (12.1.4). That is, we determined c so that

$$\int_0^\infty xce^{-cx}dx = \int_0^\infty x\frac{c}{2}\frac{1}{\sqrt{x}}e^{-c\sqrt{x}}dx, \qquad (12.1.5)$$

which gives c=2 (see annex 1). In the following we use the terms "*Weibull*(1/2)" and "exponential" to refer to the Weibull distribution with α=1/2 and c=2 and the exponential distribution with α=1 and c=2. The density functions of these two distributions are presented in Figure 12.1.1, which shows that *Weibull*(1/2) is more skewed and has a higher frequency of very small units, than the *exponential* distribution.

Figure 12.1.1

The exponential distribution (12.1.2) with c=2 and the *Weibull*(1/2) distribution (12.1.4)



We wish to create 1000 artificial unit values x_k for each of the two distributions. To obtain values x_k which obey the exponential distribution, we define for k=1,...,1000

$$u_k = 1 - e^{-2x_k} \qquad (12.1.6)$$

Solving this for x_k, we obtain

$$x_k = \frac{1}{2}\ln\left(\frac{1}{1-u_k}\right) \text{ for k=1,...,1000} \qquad (12.1.7)$$

Correspondingly, we obtain for the *Weibull*(1/2) distribution

$$u_k = 1 - e^{-2\sqrt{x_k}}$$

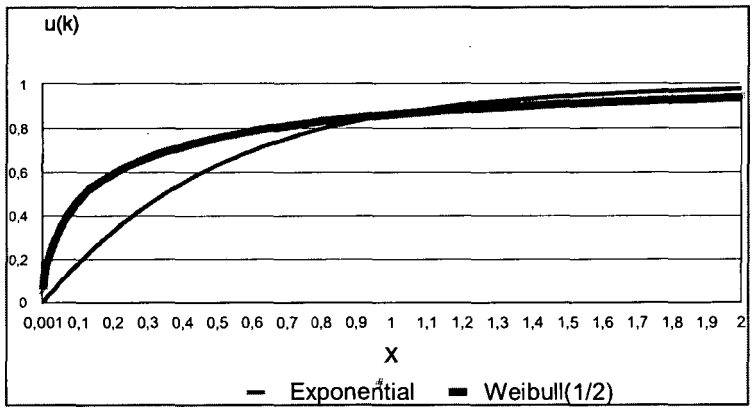
and

$$x_k = \left(\frac{1}{2}\ln\frac{1}{1-u_k}\right)^2 \text{ for } k=1,\ldots,1000. \qquad (12.1.8)$$

The values of u_k for given values of x are shown in Figure 12.1.2. We can see that the *Weibull*(1/2) distribution gives more small values for x and fewer large ones than the exponential distribution.

Figure 12.1.2.

Graph of the exponential (thinner line) and *Weibull*(1/2) distribution functions (thicker line).



Our next step is to fix the principle for choosing the values of u_k in (12.1.7) and (12.1.8). One possibility is to choose the M as random numbers obeying the *Unif*(0,1) distribution. Another is to choose equidistant values.

For convenience, we have chosen the equidistant alternative, with values for u_k determined by the formula

$$u_k = \frac{k-0.5}{1000} \qquad \text{for } k = 1,\ldots,1000. \qquad (12.1.9)$$

Using (12.1.9) as values for the u_k, we derive 1000 values for x_k by (12.1.7) when α=1 and by (12.1.8) when α=1/2. The first four moments of x_k in these two sets are presented in table 12.1. 1.

Table 12.1.1

Characteristics of the two distributions of values of x, exponential and *Weibull*(1/2)

	Exponential	**Weibull(1/2)**
Expected value	0.5	0.5
Variance	0.25	1.25
Skewness	0.25	9.25
Kurtosis	-2.4375	134.063

As seen in Table 12.1.1, the two distributions have the same mean, μ_x = 0,5, While the *Weibull*(1/2) distribution has greater variance, skewness and kurtosis than the exponential distribution.

The next step is to construct values for the variable Y that are closely correlated with X. We decided to fix the conditional expected value of y_k so that $E(y_k|x_k) = 2x_k$ and so that the value of the correlation coefficient is 0.9, that is

$$corr(x,y) = \frac{\text{cov}(x,y)}{\sqrt{V(x)V(y)}} = 0.9$$

The procedure for obtaining the values y_k is explained in detail in Annex 2.

For the exponential distribution, we generate y_k; k=1 , ... , 1000 such that

$$y_k|x_k \sim Gamma(0{,}23457, 8.52624x_k);$$

The derivations of the parameter values 0.23457 and 8.52624 are shown in Annex 2.

We then have

$$E(y_k|x_k) = 0.23457 * 8.52624x_k = 2x_k$$
$$V(y_k|x_k) = (0.23457)^2 * 8.52624x_k = 0.46914x_k$$
$$Corr(x,y) = 0.9$$

For *Weibull*(1/2) we generate y_k , k = 1,...,1000, so that

$$y_k|x_k \sim Gamma(0{,}39095, 5.115814x_k) \ .$$

The derivation of the parameter values is shown in Annex 2.
We then have

$$E(y_k|x_k) = 0.39095 * 5.11581 x_k = 2x_k$$

$$V(y_k|x_k) = (0.39095)^2 * 511581 x_k = 0.78190 x_k$$

$$Corr(x, y) = 0.9$$

We now have two population data sets (x_k, y_k), k=1,...,1000 and can use these as frames for PoMix sampling.

12.2 Simulation results for random size PoMix sampling

Using the procedure described in section 12.1, we have constructed two data sets, one with *exponentially distributed values x and y* and the other with *Weibull(1/2) distributed values for x.* As in chapter 10, we conducted a MC experiment on these two data sets involving four estimators of the population total of y. The experiment involved repeated draws of samples and repeated assignments of the set of PRNs to the population units. For each of 100 assignments of the PRN's, 100 samples were drawn using PoMix sampling with a fixed value of the Bernoulli width B. For each of the 10.000 combinations, three point estimators, the corresponding three variance estimators and the corresponding three confidence intervals were computed. To see the effect of the Bernoulli width B, experiments were carried out for a range of values $0 \leq B \leq f$, where $f = n/N$. For the exponential case the size of each sample was n=100 out of N=1000 so that the sampling fraction was 100/1000=0.1. For the more skew *Weibull (½)* case some units was put into the take-all stratum so that the size of each sample was n=87 out of N=987 so that the sampling fraction is 87/987=0.088

As seen in Table 12.2.1, the variance is only 50% of that in Poisson πps sampling when the Bernoulli width B lies in the interval 0.03 - 0.04 and α = 1/2. The percentages are calculated by dividing each row ($B \neq 0$) in Table 1 of Annex 3 by the first row (B=0).

$$\frac{MCV(\hat{Y}|B)}{MCV(\hat{Y}|B=0)} \quad \text{and} \quad \frac{MCE(\hat{V}|B)}{MCE(\hat{V}|B=0)} . \qquad (12.2.1)$$

Table 12.2.1.

Results of the simulation experiment using a *Weibull*(1/2) distribution.

Improvement gained with Pomix sampling for different Bernoulli widths *B*.

$MCV(\hat{Y})$ = MC variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ =

MC expectation of the variance estimator $\hat{V}$.

Each row of Table 1 in Annex 3 is divided by the first row.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.72	0.59	0.59	0.73	0.59	0.60
0.020	0.70	0.51	0.52	0.70	0.52	0.53
0.030	0.72	0.50	0.50	0.71	0.47	0.48
0.040	0.76	0.48	0.49	0.73	0.47	0.47
0.050	0.83	0.49	0.50	0.80	0.49	0.49
0.060	0.95	0.51	0.52	0.91	0.52	0.53
0.070	1.12	0.55	0.57	1.14	0.56	0.57
0.080	1.44	0.61	0.63	1.48	0.66	0.67
0.088	2.01	0.70	0.74	2.01	0.77	0.79

For the case of exponential distribution, Table 12.2.2 shows that the variance is about 80% of that in Poisson πps sampling when the Bernoulli width B lies in the interval 0.03 - 0.04.

Table 12.2.2. Results of the simulation experiment using an exponential distribution . Improvement gained with Pomix sampling for different Bernoulli widths *B*.

$MCV(\hat{Y})$ = MC variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC expectation of the variance estimator $\hat{V}$. Each row in Table 2 of Appendix 4 is divided by the first row.

Bernoulli width B	$MCV(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.99	0.88	0.88	0.95	0.90	0.90
0.020	1.00	0.83	0.83	0.88	0.83	0.84
0.030	1.04	0.81	0.81	1.02	0.76	0.76
0.040	1.08	0.80	0.80	0.94	0.80	0.80
0.050	1.14	0.80	0.81	1.01	0.81	0.81
0.060	1.21	0.81	0.82	1.14	0.79	0.80
0.070	1.31	0.83	0.84	1.25	0.84	0.84
0.080	1.45	0.87	0.87	1.34	0.88	0.89
0.090	1.63	0.92	0.93	1.60	0.92	0.92
0.100	1.93	1.00	1.00	1.90	1.04	1.03

The relative improvement achieved when the value of the parameter α in the Weibull distribution is changed from 1 to 1/2 is shown in Table 12.2.3 below. This is measured by

$$\frac{MCV(\hat{Y}|B,\alpha=1/2)/MCV(\hat{Y}|B=0,\alpha=1/2)}{MCV(\hat{Y}|B,\alpha=1)/MCV(\hat{Y}|B=0,\alpha=1)} \quad (12.2.2).$$

By definition, this formula gives the value 1 in the first row, corresponding to Poisson πps. The results obtained with different values of *B* show an improvement if they are smaller than one. The smaller the value, the greater this improvement is.

With these two values of α we can see that the improvement achieved by PoMix sampling over Poisson πps is better for the distribution with the greater skewness, i.e. *Weibull*(1/2). Otherwise, we can see from annex 3 how the heterogeneity of the population affects the improvement achieved by PoMix sampling.

We note that the relative improvement is greater for α=1/2 than for α=1. The efficiency increases about 40% when we move from the exponential distribution to the *Weibull*(1/2) distribution. It seems that PoMix sampling is suitable for a business population that is very skew and has a high kurtosis value. There are no significant differences in the relative efficiency of the estimators $\hat{Y}^{GREG}$ and $\hat{Y}^{SRAT}$, while the relative improvement with the

estimator $\hat{Y}^{HT}$ is smaller although still positive, except in the case of Bernoulli sampling (B=0.088).

Table 12.2.3.

Results of simulations for different Bernoulli widths *B*. Relative improvement moving from the exponential to the *Weibull*(1/2) distribution. $MCV(\hat{Y})$ = MC variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC expectation of the variance estimator $\hat{V}$

Bernoulli width B	$MCV(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.73	0.67	0.67	0.77	0.66	0.67
0.020	0.70	0.62	0.63	0.80	0.62	0.63
0.030	0.69	0.61	0.62	0.69	0.63	0.64
0.040	0.71	0.61	0.61	0.78	0.58	0.59
0.050	0.73	0.61	0.62	0.79	0.60	0.61
0.060	0.78	0.63	0.64	0.79	0.65	0.67
0.070	0.86	0.66	0.68	0.91	0.67	0.68
0.080	0.99	0.70	0.72	1.10	0.75	0.76
0.088	1.23	0.77	0.79	1.26	0.83	0.86

So, as far we have found three good properties of PoMix sampling:

1. We have seen that when Poisson πps samples are rotated, not all the units have a positive probability greater than zero of being selected in any of the successive samples. PoMix sampling can help to reduce the number of these bypassed units. When we set the value of *B* equal to or greater than the value of the constant shift, *D*, we get the positive inclusion probability (greater than zero) for each unit. This is necessary if we want to update the frame with a sample, and it is necessary to update every unit at some point in time. Otherwise it is fair that all units should be evenly subjected to questionnaires.
2. By rotating PoMix samples we can follow even the smallest units in two or more successive time intervals. This is important when surveying SME's (small and medium-sized enterprises). With Poisson πps sampling we do not have this panel effect for the smallest units.
3. PoMix sampling gives smaller variance than Poisson πps sampling with some values of *B* .

We have also found two factors that affect the improvement achieved with PoMix sampling. First, we found in chapter 10 that an increasing value of *p* weakens the improvement brought about by PoMix sampling. Second, we

found that the greater the skewness and kurtosis in the data, the more effective PoMix sampling is compared with Poisson πps sampling. This means that PoMix sampling is suitable for business surveys.

12.3 Simulation results for Order PoMix sampling

It is interesting to see how order PoMix sampling is improved with B≠0 compared with the case of $B = 0$. Two order PoMix sampling schemes were introduced in Chapter 10: one which gives the sequential Poisson sampling of Ohlsson (1990, 1996, 1998) when $B = 0$, using (10.6.2), and one which gives the Pareto sampling of Rosen (1996a, 1996b). To see this improvement, we conducted MC experiments on the two data sets described in the previous chapter.

There was not found essential difference in variances between sequential Poisson sampling and Pareto sampling. We therefore concentrate below on sequential Poisson sampling only.

Table 12.3.1 below shows how the efficiency improves when we add elements of equiprobability sampling to the Sequential Poisson sampling scheme. When we set $p=1$ and $\alpha=1$ we can achieve an improvement of about 25 percent in the sampling scheme if we choose B between 0.02 and 0.05. The formula used in the calculations is presented in 12.2.2.

Table 12.3.1

Results of simulations using different Bernoulli widths B when p=1 and $\alpha=1$.

Improvements in Order PoMix sampling achieved with different Bernoulli width B.

$MVC(\hat{Y})$ = MC variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC expectation of the variance estimator $\hat{V}$. Each row of Table 2 in Annex 4 is divided by the first row.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.94	0.89	0.89	0.91	0.85	0.85
0.020	1.01	0.84	0.84	0.94	0.77	0.77
0.030	1.14	0.82	0.82	1.11	0.76	0.76
0.040	1.33	0.81	0.81	1.19	0.73	0.73
0.050	1.58	0.81	0.81	1.54	0.76	0.76
0.060	1.89	0.82	0.82	2.05	0.81	0.81
0.070	2.32	0.84	0.84	2.31	0.82	0.82
0.080	2.91	0.87	0.87	2.78	0.86	0.86
0.090	3.75	0.92	0.92	3.82	0.87	0.87
0.100	5.12	1.00	1.00	5.01	0.98	0.98

As demonstrated by Table 12.3.2 below, the efficiency is increased when we move to a *Weibull*(1/2) distribution. As in the case of random size PoMix sampling, we achieve an improvement of over 50% when adding equiprobable sampling with B between 0.02 and 0.03. As in the case of random size PoMix sampling, it seems that fixed size order PoMix sampling also improves the efficiency when we move to a distribution with higher skewness and kurtosis; always assuming that $p < 2$.

Table 12.3.2.

Results of simulations with different Bernoulli widths B when $p=1$ and $\alpha=1/2$.

Improvements in Order PoMix sampling achieved with different Bernoulli widths B.

$MVC(\hat{Y})$ = MCvariance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC expectation of the variance estimator $\hat{V}$. Each row of Table 1 in annex 4 is divided by the first row.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.66	0.64	0.64	0.68	0.67	0.67
0.020	0.61	0.55	0.56	0.61	0.55	0.55
0.030	0.63	0.53	0.53	0.64	0.53	0.53
0.040	0.69	0.53	0.53	0.68	0.50	0.50
0.050	0.78	0.53	0.53	0.78	0.55	0.55
0.060	0.92	0.55	0.55	0.92	0.53	0.53
0.070	1.16	0.59	0.60	1.20	0.60	0.61
0.080	1.59	0.66	0.67	1.54	0.67	0.68
0.088	2.43	0.77	0.78	2.31	0.75	0.74

As seen in Table 12.3.3, the same effect could be observed for order PoMix sampling as was found earlier for random sample size PoMix sampling. Efficiency was lost with increasing heteroscedasticity:

$$\frac{MCV(\hat{Y}|B,\alpha=1,p=2)}{MCV(\hat{Y}|B,\alpha=1,p=1)}$$

The deterioration increases at greater cell values.

Table 12.3.3.

Results of simulations with different Bernoulli widths B when $\alpha=1$. Relative improvement between $p=2$ and $p=1$. Sequential Poisson sampling follows if $B=0$. $MVC(\hat{Y})$ = MCvariance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC expectation of the variance estimator $\hat{V}$.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.01	1.49	1.19	1.19	1.42	1.16	1.16
0.02	2.01	1.41	1.40	1.89	1.46	1.46
0.03	2.49	1.63	1.63	2.35	1.64	1.63
0.04	2.92	1.89	1.88	3.10	1.95	1.95
0.05	3.36	2.21	2.20	3.49	2.16	2.15
0.06	3.94	2.66	2.64	3.38	2.47	2.46
0.07	4.64	3.24	3.21	4.07	3.16	3.13
0.08	5.85	4.20	4.18	5.72	4.00	3.93
0.09	7.72	5.45	5.59	7.14	6.04	5.84

When we move from random size PoMix sampling to fixed size order PoMix sampling the efficiency is greatly improved if we use the Horwid-Thompson estimator, as shown in Table 12.3.4, although the improvement is close to zero if we use auxiliary information in the ratio or regression estimators. The first row shows that this holds good even with sequential Poisson sampling. The use of auxiliary information improves the estimates so that no further improvement is achieved by fixing the sample size.

Table 12.3.4.

Results of simulations with different Bernoulli widths B when $\alpha=1$. Change in efficiency between Order (fixed sample size) PoMix sampling and non-fixed size PoMix sampling. Sequential Poisson sampling follows if $B=0$ in Order PoMix sampling. $MVC(\hat{Y})$ = MCvariance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC expectation of the variance estimator $\hat{V}$.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	0.20	0.99	0.99	0.20	1.03	1.03
0.010	0.19	1.00	1.00	0.19	0.97	0.97
0.020	0.20	1.00	1.00	0.21	0.95	0.95
0.030	0.22	0.99	0.99	0.21	1.03	1.03
0.040	0.25	1.00	1.00	0.25	0.95	0.95
0.050	0.28	0.99	0.99	0.30	0.97	0.97
0.060	0.32	0.99	0.99	0.35	1.06	1.05
0.070	0.36	1.00	1.00	0.36	1.00	1.00
0.080	0.41	0.99	0.99	0.41	1.00	1.00
0.090	0.47	0.98	0.99	0.47	0.97	0.97
0.100	0.54	0.99	0.99	0.51	0.96	0.97

A brief summary of chapter 12

To see the effect of skewness on PoMix sampling, two artificial data sets were introduced: a highly skewed *Weibull* (1/2) distributed data set and an exponentially distributed data set. A greater improvement in estimation was obtained for the more skewed distribution, provided that the value of heterogeneity p was smaller than two. We thus know that a major improvement is achieved with PoMix sampling if the population is highly skewed and if the heterogeneity of the error term is not very great.

The improvement in estimation was also tested using order PoMix sampling and comparing this with fixed size sequential Poisson sampling and Pareto sampling. Since it was found that there are no significant differences between the results of sequential Poisson and Pareto sampling, the latter was dropped out. PoMix was found to produce the same improvement in variances as random size PoMix sampling. Using the *Weibull*(1/2) distribution, the variance was reduced by about 50% when a Bernoulli part was included. It was also found that sequential Poisson sampling improved the variances compared with Poisson πps sampling only when the ordinary HT estimator was used without any auxiliary information.

SUMMARY

When we decide to carry out a statistical survey we must determine the kind of survey and the kind of information needed in order to obtain answers to the problems we are interested in. Whether we use a census or a survey sampling is determined by questions of cost. The information needs tell us what estimates we need, and it is these that determine what type of survey must be conducted, which in turn defines what kind of coordination system is possible and reasonable to use in order to lessen the response burden. If a census is used, no coordination is needed, as the inclusion probabilities are one. Once the type of survey has been determined, the next step is to decide what are the units of which the population is composed and to look for a frame for the population we are interested in.

The quality of the frame is a critical part of a business survey, as it must include and define the units and classifications that are needed. The units must be suitable for statistical purposes and coincide with real units in the population. For this purpose we commonly need information on the geographical level or activity level concerned.

Due to the continuous formation of new businesses and closure of some existing ones, the frame is never the same between two points in time. Apart from causing overcoverage and undercoverage in the frame, changes in the business population also cause some problems of sample coordination. We may send a questionnaire to a business which has changed its ID number and was in the same inquiry on the last occasion but with a different number. It is essential to pay attention to business demography in order to reduce these difficulties.

Some units change strata between two time points. One mechanism which takes note of the formation and closure of businesses is the constant shift method performed using permanent random numbers. Keyfitz provides a method to control for changes in strata. A frame is often linked to coordinating system, which contains the update data necessary for coordination purposes, such as permanent random numbers.

A probability sampling design gives us a device for making generalizations from a sample to the whole population. The use of permanent random numbers (PRN) in list-sequential sampling schemes such as Bernoulli sampling, Sequential SRS or Poisson sampling makes the coordination of sampling easy. Bernoulli sampling is a special case of Poisson sampling, i.e. equiprobable Poisson sampling. Due to the highly skewed distribution of business populations, stratification is needed for all equiprobable sampling schemes.

Users of survey data have an insatiable demand for detail. This means that the burden imposed on respondents will cause sample fatigue if it is not controlled by the statistical agency. Sunter has presented a model which

includes three important parts: the response obligation, which reflects the agency's assessment of what is a respondent's reasonable share of the total burden, and the response load and inclusion probability, which together form the expected response burden, which must not exceed the response load. The inclusion probability makes the response burden a random variable, and this means that the distribution of the response burden is not even. Coordination enables this distribution to be made more even, and thus the main topic of this thesis is the theory of coordination of business samples.

There are three methods of coordinating business samples. One is based on the use of rotation groups and other two on the use of permanent random numbers (PRN). The first of these, Simple random sampling in randomly formed rotation groups, is the one used in Statistics Canada. Another one which is based on the idea of the use of random numbers in sequential sampling originates from Fan et al. (1962) and was improved later by Cassel and then by Atmer and Sjöberg, who introduced the JALES method, later called sequential Simple random sampling.

Hajek introduced Poisson sampling in 1964, and this served as a basis for Poisson Mixture (PoMix) sampling, which was introduced by Kröger, Särndal and Teikari in 1999 to improve on Poisson rotation, which ignores some small units and does not give longitudinal information on others. The improvement was achieved by adding part of the technique of equiprobable Bernoulli sampling to Poisson πps sampling. The take-all stratum and the size measure Q were introduced and a sampling algorithm prepared. The joint probability in two successive samples in PoMix sampling was found to be exactly the same as had been obtained earlier by Sunter with a zero width for the Bernoulli part.

To obtain a fixed sample size, order PoMix sampling was introduced on the basis of the method described by Ohlsson. The resulting sampling gave two interesting results. When the value for the Bernoulli part was set at zero, we obtained exactly Ohlsson's sequential Poisson sampling. Otherwise, when the size of the Bernoulli part was set at exactly n/N, we did not get Bernoulli sampling but sequential Simple random sampling.

Unexpectedly, we found when we carried out a Monte Carlo test that the addition of part of the equiprobable Bernoulli sampling routine to Poisson πps sampling reduced the variance of the estimators that made use of auxiliary information. A total of 10,000 sampling combinations were drawn from a frame of 1000 enterprises extracted at random from the Business Register. An examination of the Taylor linearized variance showed that parameter p, measuring the heterogeneous error term, must be smaller than 2 in order to achieve the desired results. Linearizing the error term by means of a logarithmic version we arrived at a p-value of 1.45 in our data.

To investigate the effect of skewness on PoMix sampling, two artificial data sets were introduced, one highly skewed, called *Weibull* (1/2)-distributed, and the other exponentially distributed. The result was that the improvement in the estimates was greater for the more skewed distribution, provided that the heterogeneity value p was smaller that two. Thus we know that the improvement achievable by using PoMix sampling is great if the

population is highly skewed and the heterogeneity of the error term is not very great.

The improvement in the estimates was also tested using order PoMix sampling as opposed to fixed size Sequential Poisson sampling or Pareto sampling. As no significant differences were found between the results of Sequential Poisson and Pareto sampling, the latter was dropped. It was found that PoMix gave the same improvement in variances as in the case of random size PoMix sampling. Using the *Weibull* (1/2) distribution with some part of the Bernoulli routine the variance was reduced by about 50 percent. It was also found that Sequential Poisson sampling improved the variances compared with Poisson πps sampling only when an ordinary HT estimator without auxiliary information was used.

REFERENCES

Atmer, J., Thulin, G., Bäcklund, S. (1975). Coordination of Samples with the JALES Technique. *Statistisk Tidskrift*, 13, pp. 443-350.

Australian Bureau of Statistics (1985), ABS Computing network Systems manual, VIII, unpublished report, Belconnen: *Australian Bureau of Statistics.*

Badsley, P., Chambers, R.L. (1984). Multipurpose Estimation from Unbalanced Samples. *Applied Statistics*, 33, pp. 290-299.

Bailar, B.A. (1989). Information needs, Surveys, and Measurement errors. *Panel Surveys.* Edited by Kasprzyk, Duncan, Kalton and Singh. U.S. 1989.

Baltagi, B.H. (1996). *Econometric Analysis of Panel Data.* New York 1996.

Bankier, M.D. (1988). Power Allocations: Determining Sample Sizes for Sub-national Areas, *The American Statistician,* 42, pp. 174-177.

Brewer, *k*.R.W., Early L.J. and Joyce S.F. (1972). Selecting several samples from a single population, *Austral. J. Statist,* 14, pp. 231-239.

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

Brewer, *k*.R.W. (1994). Survey Sampling Inference: Some Past Perspectives and Present Prospects (1994). *Pakistan Journal of Statistics,* 1994, 10(1)A, pp. 213-233.

Cassel, P.G. (1967). Urvalsdragning på datamaskin med sekventiella metoder. *Statistisk Tidskrift,* 6, pp. 486-492.

Colledge, M.J. (1995). Frames and Business Register: An overview. *Business Survey Methods.* Ed. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York 1995.

Cotton, F., Hesse, C. (1992). Coordinated Selection of Stratified Samples. *Proceedings of Statistics Canada Symposium 92. Design and Analysis of Longitudinal Surveys.* November 1992.

Council regulation, (EEC) No 3037/90 of 9 October 1990 on the statistical classification of economic activities in the European Community, *Official Journal of the European Communities,* L 293, 33, 24 October 1990.

Council Regulation, (EEC) No696/93 of 15 march 1993 on the statistical units for the observation and analysis of the production system in the Community, *Official Journal of the European Communities* L76/1, 30 March 1993.

Cox, B.G., Chinnappa, B.N. (1995). Unique Features of Business Surveys. *Business Survey methods.* Ed. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York 1995.

Dalenius, T. (1952). The Problem of Optimum Stratification in a special Type of Design. *Skandinavisk Aktuerietidskrift,* 35, pp. 61-70.

Deming,W. E. (1960). *Sampling Design in Business research.* New York: Wiley 1960.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in Survey sampling. *Journal of the American statistical association,* 87, pp. 376-382.

Duncan, G., Kalton, G. (1987). Issues of Design and Analysis of Surveys across Time. *International Statistical Review*, 55(1), pp. 97-117.

Fan, C.T., Muller, M.E., Rezucha, I (1962). Development of Sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of American Statistical Association,* 57, pp. 387-402

Glasser, G.J. (1962). On the Complete Coverage of Large Units in a Statistical Study. *Review of the International Statistical Institute,* 30, pp. 28-32.

Hajek, J. (1981). *Sampling from a Finite Population.* Edited by Vaclav Dupac, New York 1981.70

Hajek, J. (1960). Limiting distributions in Simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.* 5, pp. 361 - 374.

Hajek, J. (1964). Asymptotic Theory of Rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.* 35, pp. 1431-1523.

Hansen, M. H., Hurwitz, W.N., Madow, W.G. (1953) *Sample Survey Methods and Theory (Vol.2)* New York: John Wiley.

Hansen, M.H. and Hurwitz, W.N (1943). On the Theory of Sampling from a Finite Population. *Annals of Mathematical Statistics,* 14, pp. 333-362.

Hidiroglou, M.A., Särndal, C-E., Binder, D.A. (1995). Weighting and Estimation in Business Surveys. *Business Survey Methods.* Ed. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York 1995.

Hidiroglou, M.A (1986). The Construction of a Self-Representing Stratum of large Units in Survey Design, *The American Statistician,* 40, pp. 27-31.

Hidiroglou, M.A., Choudhry, G.H., and Lavallee, P. (1991). A Sampling and Estimation Methodology for Sub-Annual Business Surveys. *Survey Methodology*, 17, pp. 195 - 210.

Hidiroglou, M.A., Srinath, *k*.P. (1993). Problems Associated with designing Sub-annual Business Surveys. *Journal of Business & Economic Statistics,* 11, pp. 397 - 405

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

Kröger, H., Särndal, C-E., and Teikari, I. (1999). Poisson Mixture Sampling: A Family of designs for Coordinated selection Using permanent Random Numbers. *Survey Methodology,* 25, pp. 3-11

Lavallee, P., Hidiroglou, M.A. (1988). On the Stratification of Skewed Population. *Survey Methodology,* 14, pp. 33-43.

Kish, L., Scott, A. (1971). Retaining Units after Changing Strata and Probabilities. *Journal of the American Statistical Association,* 66, pp. 461-470.

Laaksonen, S., Teikari, I. (1999). Employment and productivity in Finnish Manufacturing. *Micro and Macrodata of Firms,* Ed. Biffignandi Silvia, New York 1999.

Lessler, J. T. (1982). Frame Errors., *A taxonomy of Error Sources and Error Measures for Surveys.* Ed. J. T. Lessler, R. E. Folsom and W. D. Kalsbeek. Final report. Research Triangle Park, NC: Research Triangle Institute.

Malinen, P. (1994). Pk-yritysten hallintomenettely. *Turun kauppakorkeakoulu. ritystoiminnan tutkimuskeskus.* Sarja C. Keskustelua 2/94. Turku 1994.

Mustaniemi, T. (1997). Enterprise Demography as a Method of Studying Real Enterprise Births. *The Evolution of Firms and Industries.* Research Reports 223, pp. 207-216. Statistics Finland.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, *Journal of Royal Statistical Society*, 97, pp. 558-606.

Ogus, J.L. and Clark, D.F. (1971). *The annual survey of manufactures:* A report on methodology, Technical papers No. 24, U.S. Bureau of the Census, Washington, D.C.

Ohlsson, E. (1990). Sequential Poisson Sampling from a Business Register and its Application to Swedish Consumer Price Index. *R&D Report, Statistics Sweden.* Research - Methods - Development 1990:6.

Ohlsson, E. (1995). Sequential Poisson Sampling. *Institute of Actuarial Mathematics and Mathematical Statistics,* Stockholm University, Report No. 182, June 1995.

Ohlsson, E. (1995). Coordination of Samples Using Permanent Random Numbers. *Business survey Methods.* Ed. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York 1995.

Struits, P., Willeboordse, A. (1995). Changes in Populations of Statistical Units. *Business Survey Methods.* Ed. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott. New York 1995.

Sunter, A.B. (1977). Response Burden, sample Rotation, and Classification renewal in Economic Surveys. *International Statistical Review,* 45, pp. 209-222.

Sunter, A.B., (1977). List Sequential Sampling with Equal or Unequal Probabilities without Replacement. *Appl. Statist,* 26, pp 261-268.

Sunter, A.B. (1986), Solutions to the Problem of Unequal Probability Sampling Without Replacement. *International Statistical Review*, 54, pp. 33-50.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1992). *Model assisted survey sampling.* Springer, New York, 1992.

Templeton, R. (1990). Poisson Meets the new Zealand Business Directory. *The New Zealand Statistician,* 25. pp. 2-9.

Thulin, G. (1976). Samordnade urval inom företagsstatistiken (SAMU), Praktisk tillämpning av JALES-metoden. *Statistisk tidskrift,* 14, pp. 85-101.

Toimialaluokitus 1995. *Tilastokeskuksen käsikirjoja* 4, Helsinki 1993. (In Finnish)

Tuominen, T. (1999). *Vastausrasite ja sen mittaaminen Tilastokeskuksen yrityssurveyssä 1998 (Measuring the Response Burden of Business Surveys in Statistics Finland in 1998).* Unpublished Master's thesis, Department of Statistics. University of Jyväskylä. (In Finnish)

U.N. Statistical office (1990). *International Standard Industrial Classification of all Economic Activities - Revision 3* (ISIC Rev.3) Series M, No. 4 Rev 3 , New York: United Nations.

Van, H M., Koeijers, E., and De Ree, J. (1994). EDS, Sampling System for The Central Business Register at Statistics Netherlands. Unpublished manuscript, Voorburg: Netherlands Official Statistics

Van Huis, M., Koeijers, E., and De Ree, J. (1994). Response Burden and Coordinated Sampling for Economic Surveys. *Netherlands Official Statistics*, pp. 17 – 26.

ANNEX I

Determining the parameter c in (12.1.1) and (12.1.3)

To see the impact of the effectivity of Pomix sampling in the data with different skewness it was prepared artificial data. For that we defined and equalized two exponential functions

$$\int_0^\infty xce^{-cx}dx = \int_0^\infty x\frac{c}{2}\frac{1}{\sqrt{x}}e^{-c\sqrt{x}}dx \qquad (1)$$

for witch c was solved. In the left hand side of (1) it was set $cx=z$ which gives after differentiating both sides $dx = \frac{dz}{c}$. In the right hand side it was set $c\sqrt{x} = z$ which gives after differentiating both sides $dx = \frac{2zdz}{c^2}$. After putting these values in (1) we get

$$\int_0^\infty xe^{-z}\frac{dz}{c} = \int_0^\infty \frac{c}{2}\frac{z}{c}e^{-z}\frac{2xdz}{c^2} \qquad (2)$$

which can be written in the form

$$\frac{1}{c}\int_0^\infty z^{2-1}e^{-z}dz = \frac{1}{c^2}\int_0^\infty z^{3-1}e^{-z}dz \qquad (3)$$

So we have got two gamma functions. The left hand side can be written

$$\frac{1}{c} * \Gamma(2) = \frac{1}{c} * 1 \qquad (4)$$

and the right hand side

$$\frac{1}{c^2} * \Gamma(3) = \frac{1}{c^2} * 2! \qquad (5)$$

From (3) and (4) we get

$$\frac{1}{c} = \frac{2}{c^2} \Rightarrow c = 2 \qquad (6)$$

Which is the desired value of c for which the left and the right hand side of (1) has the same distribution. After putting (6) to the left hand side of (1) we get

$$E(x) = \int_0^\infty x 2e^{-2x} dx$$

and considering that $cx=z$ and $dx = \frac{dz}{c}$ we get

$$E(x) = \int_0^\infty z e^{-z} \frac{dz}{c} = \frac{1}{2}.$$

Thus for the Weibull distribution with $\alpha=1$ we get

$$E(x) = ab = \frac{1}{2} * 1 = \frac{1}{2} = \mu_x \qquad (7)$$

and

$$V(x) = a^2 b = \left(\frac{1}{2}\right)^2 1 = \frac{1}{4} = {\sigma_x}^2 \qquad (8)$$

ANNEX 2

Derivation of y_k-values for given x_k-values

Next we define a variable $X \sim Gamma(a,b)$. This means that the density function of X is

$$f(x) = \frac{x^{b-1}}{a^b \Gamma(b)} e^{-\frac{x}{a}} \quad \text{for } x > 0$$

We have: $\begin{cases} E(X) = ab \\ V(X) = a^2 b \end{cases}$

Because we need a study variable y_k, $k= 1,\ldots,N$ and with that highly correlated (ρ=0.9) auxiliary variable X_k $k=1,\ldots,N$, we must create y_k, $k=1,\ldots,N$ so that, given X_k

$$y_k \Big| x_k \sim Gamma\left(\frac{\sigma^2}{\beta}, \frac{\beta^2 x_k}{\sigma^2}\right) \qquad (1)$$

Then

$$\begin{aligned} E(y_k \Big| x_k) &= ab = \frac{\sigma^2}{\beta}\frac{\beta^2 x_k}{\sigma 2} = \beta x_k \\ V(y_k \Big| x_k) &= a^2 b = \left(\frac{\sigma^2}{\beta}\right)^2 \frac{\beta^2 x_k}{\sigma^2} = \sigma^2 x_k \end{aligned} \qquad (2)$$

We want $corr(x,y) = \dfrac{Cov(x,y)}{\sqrt{V(x)V(y)}}$ to have a fixed value, say 0.9.

Using (2) we get

$$\begin{aligned} E(Y) &= EE(Y|X) = E(\beta x) = \beta E(X) = \beta \mu_x \\ E(XY) &= E(XE(Y|X)) = E(X\beta X) = \beta E(X^2) = \beta(\sigma^2{}_x + \mu^2{}_x) \\ V(Y) &= EV(Y|X) + VE(Y|X) = \sigma^2 \mu_x + \beta^2 \sigma^2{}_x \end{aligned}$$

and further

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \beta(\sigma^2{}_x + \mu^2{}_x) - \mu_x \beta \mu_x = \beta\sigma^2{}_x$$

and

$$Corr(X;Y) = \frac{Cov(X;Y)}{\sqrt{V(X)V(Y)}} = \frac{\beta\sigma^2{}_x}{\sqrt{\sigma^2{}_x(\sigma^2\mu_x + \beta^2\sigma^2{}_x}} = \frac{\beta}{\sqrt{\beta^2 + \sigma^2 \frac{\mu_x}{\sigma^2{}_x}}} \quad (3)$$

We define the Weibull distribution with $\alpha = 1$. From (7) and (8) in Annex 1we got

$$\begin{cases} \mu_x = \frac{1}{2} \\ \sigma^2{}_x = \frac{1}{4} \end{cases} \quad (4)$$

Then we fix $\beta = 2$ and find σ^2 so that $corr(x,y) = 0.9$. Using (3) and (4) we get

$$0.9 = corr(x,y) = \frac{\beta}{\sqrt{\beta^2 + \sigma^2 \frac{1/2}{1/4}}} = \frac{\beta}{\sqrt{\beta^2 + 2\sigma^2}} \quad (5)$$

We can solve σ^2 from (5)

$$(0.9)^2(\beta^2 + 2\sigma^2) = 0.81 * 2\sigma^2 = \beta^2(1 - 0.81) = 0.19\beta^2$$

$$\sigma^2 = \frac{0.19}{0.81 * 2}\beta^2 = 0.46914$$

For (1) we can now calculate two parameter

$$\begin{cases} \frac{\sigma^2}{\beta} = \frac{0.46914}{2} = 0.23457 \\ \frac{\beta^2}{\sigma^2} = \frac{4}{0.46914} = 8.52624 \end{cases} \quad (6)$$

For Weibull distribution with $\alpha=1$ we have the exponential distribution

$$f(X) = 2e^{-2X}$$

We must then generate Y_k, $k=1,\ldots,N$, so that

$$y^k \left| x^k \right. \sim Gamma(0.23457, 8.52624x_k).$$

which gives

$$E(y_k|x_k) = 0.23457 * 8.52624 x_k = 2x_k$$
$$V(y_k|x_k) = (0.23457)^2 * 8.52624 x_k = 0.46914 x_k$$
$$Corr(x,y) = 0.9$$

Putting (6) in Annex 1to the right hand side of (1) in Annex 1gives

$$f(x) = \frac{1}{\sqrt{x}} e^{-2\sqrt{x}}$$

$$E(x^m) = \int_0^\infty x^m \frac{1}{\sqrt{x}} e^{-2\sqrt{x}} dx \qquad (7)$$

If we set $z = 2\sqrt{x}$ it follows $x = z^2/4$ and dx = z/2dz. Putting these values in (7) we get

$$E(x^m) = \int_0^\infty \frac{1}{4^m} z^{2m} e^{-z} dz = \frac{\Gamma(2m+1)}{4^m}$$

If we put *m*=2 we get

$$\begin{cases} \mu = 1.5 \\ \sigma^2 = 1.25 \end{cases}$$

To obtain *corr(x,y) =0.9* we get as in (5) before

$$0.9 = \frac{\beta}{\sqrt{\beta^2 + \sigma^2 \frac{1.5}{1.25}}} \qquad (8)$$

When we fix β=2 we get

$$\begin{cases} \sigma^2 = 0.78189 \\ \sigma^2/2 = 0.39095 \\ \beta^2/\sigma^2 = 5.11581 \end{cases}$$

We create y_k *k=1,...,N*, so that

$$y_k|x_k \sim Gamma(0.39095, 5.11581 x_k)$$

which gives

$$\begin{cases} E(y_k|x_k) = 0.39095 * 5.11581 x_k \approx 2x_k \\ V(y_k|x_k) = (0.39095)^2 * 511581 x_k = 0.78190 x_k \\ Corr(x, y) = 0.9 \end{cases}$$

ANNEX 3

Results of simulation study using the frame of artificial units

Results of simulation study using the frame of artificial units obeying Weibull distibutions with parameter values α=1 and α=1/2. Size of frame is 1000 units from which 100*100 Pomix samples of expected size of 100 are drawn. The measure of heteroscedasticity p=1. The effectiveness in tables 3 and 4 are measured by ratios

$$\frac{MVC(\hat{Y})|B=w}{MVC(\hat{Y})|B=0} \quad \text{and} \quad \frac{MCE(\hat{V})|B=w}{MCE(\hat{V})|B=0},$$

where w is the width of B in corresponding Tables 1 and 2. The change of this effectiveness when α changes from 1 to 1/2 is measured by dividing the values in Table 3 by the values in Table 4.

Table 1.

Results of simulation study with different Bernoulli widths, B, when p=1 and α=1/2 .

Values of $MVC(\hat{Y})$ =MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ =

MC-Expectation of the variance estimator $\hat{V}$. Values are multiplied by 10^6

Bernoulli width B	$MVC(\hat{Y})*10^{-10}$			$MCE(\hat{V})*10^{-10}$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	15.73	11.29	11.16	16.00	11.70	11.53
0.010	11.34	6.63	6.64	11.74	6.93	6.95
0.020	10.98	5.80	5.81	11.23	6.08	6.08
0.030	11.31	5.60	5.63	11.30	5.53	5.56
0.040	12.00	5.46	5.49	11.67	5.44	5.45
0.050	13.11	5.51	5.55	12.81	5.68	5.71
0.060	14.87	5.78	5.85	14.50	6.03	6.10
0.070	17.69	6.21	6.31	18.28	6.52	6.59
0.080	22.67	6.91	7.03	23.69	7.74	7.75
0.088	31.62	7.96	8.21	32.16	8.99	9.07

Table 2.

Results of simulation study for different Bernoulli widths B when p=1 and α=1 . Values of $MVC(\hat{Y})$ =MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-Expectation of the variance estimator r $\hat{V}$. Values are multiplied by 10^6

Bernoulli width B	$MVC(\hat{Y})*10^{-10}$			$MCE(\hat{V})10^{-10}$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	10.31	2.11	2.10	11.15	2.13	2.12
0.010	10.19	1.85	1.85	10.59	1.91	1.91
0.020	10.32	1.75	1.75	9.81	1.77	1.77
0.030	10.69	1.71	1.71	11.39	1.61	1.61
0.040	11.14	1.68	1.69	10.43	1.69	1.70
0.050	11.70	1.69	1.70	11.29	1.71	1.73
0.060	12.50	1.72	1.72	12.77	1.68	1.69
0.070	13.53	1.75	1.76	13.94	1.78	1.78
0.080	14.94	1.83	1.84	14.97	1.88	1.88
0.090	16.80	1.94	1.95	17.85	1.96	1.95
0.100	19.88	2.10	2.10	21.24	2.21	2.19

Table 3.

Results of simulation study for different Bernoulli widths B when p=1 and α=1/2 . Effectiveness of Pomix sampling with different Bernoulli width B. $MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-Expectation of the variance estimator $\hat{V}$. Each row of table 1 is divided by the first row.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.72	0.59	0.59	0.73	0.59	0.60
0.020	0.70	0.51	0.52	0.70	0.52	0.53
0.030	0.72	0.50	0.50	0.71	0.47	0.48
0.040	0.76	0.48	0.49	0.73	0.47	0.47
0.050	0.83	0.49	0.50	0.80	0.49	0.49
0.060	0.95	0.51	0.52	0.91	0.52	0.53
0.070	1.12	0.55	0.57	1.14	0.56	0.57
0.080	1.44	0.61	0.63	1.48	0.66	0.67
0.088	2.01	0.70	0.74	2.01	0.77	0.79

Table 4.

Results of simulation study for different Bernoulli widths B when p=1 and α=1 .

Effective of Pomix sampling with different Bernoulli width B. $MVC(\hat{Y})$ =

MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-Expectation of

the variance estimator $\hat{V}$. Each row of table 2 is divided by the first row.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.99	0.88	0.88	0.95	0.90	0.90
0.020	1.00	0.83	0.83	0.88	0.83	0.84
0.030	1.04	0.81	0.81	1.02	0.76	0.76
0.040	1.08	0.80	0.80	0.94	0.80	0.80
0.050	1.14	0.80	0.81	1.01	0.81	0.81
0.060	1.21	0.81	0.82	1.14	0.79	0.80
0.070	1.31	0.83	0.84	1.25	0.84	0.84
0.080	1.45	0.87	0.87	1.34	0.88	0.89
0.090	1.63	0.92	0.93	1.60	0.92	0.92
0.100	1.93	1.00	1.00	1.90	1.04	1.03

Table 5.

Results of simulation study for different Bernoulli widths B when p=1. Ratio of efectiviness

betveen cases α=1/2 and α=1. $MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and

$MCE(\hat{V})$ = MC-Expectation of the variance estimator $\hat{V}$. Table 3 is divided by table 4

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.73	0.67	0.67	0.77	0.66	0.67
0.020	0.70	0.62	0.63	0.80	0.62	0.63
0.030	0.69	0.61	0.62	0.69	0.63	0.64
0.040	0.71	0.61	0.61	0.78	0.58	0.59
0.050	0.73	0.61	0.62	0.79	0.60	0.61
0.060	0.78	0.63	0.64	0.79	0.65	0.67
0.070	0.86	0.66	0.68	0.91	0.67	0.68
0.080	0.99	0.70	0.72	1.10	0.75	0.76
0.088	1.23	0.77	0.79	1.26	0.83	0.86

ANNEX 4

The effectiviness of Order PoMix sampling with different values of p using parameter values α=1/5 or α=1

Results of simulation study using the frame of artificial units obeying Weibull distibutions with p-values (measure of heteroscedasticity) 1, 1.5 and 2 when parameter value α=1. The effectiveness in tables are measured by ratios

$$\frac{MVC(\hat{Y})|B=w}{MVC(\hat{Y})|B=0} \quad \text{and} \quad \frac{MCE(\hat{V})|B=w}{MCE(\hat{V})|B=0}, \text{ where w is the}$$

width if B.

The change of this effectiveness in table 3 is got by dividing the figures in Table 1 by Table 2 . Correspondinly are calculated the changes in Tables 6 and 7 using Annexes 1 and 4.

Table 1.

Results of simulation study for different Bernoulli widths B when p=1 and α=0.5 .

Effectiveness of Order Pomix sampling with different Bernoulli width B. Sequential Poisson sampling follows if B=0. $MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-Expectation of the variance estimator $\hat{V}$ Values are multiplied by 10^3.

Bernoulli width B	$MVC(\hat{Y})*10^{-3}$			$MCE(\hat{V})10^{-3}$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	10534.26	10394.51	10413.17	10617.86	10617.86	10617.86
0.010	6919.03	6662.45	6684.55	7238.63	7100.80	7115.41
0.020	6418.27	5758.74	5782.64	6427.61	5809.25	5829.02
0.030	6658.29	5520.50	5541.25	6742.39	5669.22	5677.17
0.040	7280.42	5494.41	5522.86	7229.89	5329.96	5346.37
0.050	8184.33	5501.70	5530.79	8287.33	5793.76	5811.50
0.060	9688.50	5712.92	5761.36	9717.46	5632.22	5666.09
0.070	12182.09	6128.90	6220.48	12762.72	6422.85	6476.78
0.080	16760.61	6829.13	6976.39	16386.67	7108.69	7175.13
0.088	25557.85	7989.33	8165.24	24486.61	7944.45	7904.44

Table 2.

Results of simulation study for different Bernoulli widths *B* when p=1 and α=1 . Effective of Order Pomix sampling with different Bernoulli width *B*. Sequential Poisson sampling follows if B=0. $MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-Expectation of the variance estimator $\hat{V}$.

Bernoulli width B	$MVC(\hat{Y})*10^{-3}$			$MCE(\hat{V})*10^{-3}$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	2100.51	2078.91	2079.50	2182.85	2182.85	2182.85
0.010	1979.32	1847.81	1848.52	1981.08	1847.41	1847.84
0.020	2112.85	1741.90	1743.40	2062.40	1677.07	1678.31
0.030	2385.76	1694.61	1696.47	2424.22	1660.50	1661.30
0.040	2786.55	1681.25	1684.29	2604.13	1601.09	1603.16
0.050	3310.28	1680.36	1684.15	3351.10	1667.09	1669.27
0.060	3972.87	1697.44	1701.95	4471.66	1775.83	1776.24
0.070	4880.47	1745.04	1751.00	5040.89	1785.25	1786.18
0.080	6113.68	1809.35	1815.34	6078.00	1876.73	1876.00
0.090	7883.64	1910.11	1918.70	8335.85	1899.30	1894.04
0.100	10745.80	2071.53	2081.60	10936.51	2131.68	2131.59

Table 3.

Results of simulation study for different Bernoulli widths *B* when p=1 and α=0.5 . Effective of Order Pomix sampling with different Bernoulli width *B* between Sequential Poisson sampling and Pareto sampling which follows if B=0. $MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-Expectation of the variance estimator $\hat{V}$.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	1.02	1.02	1.02	1.02	1.04	1.04
0.020	0.99	0.98	0.98	0.92	0.93	0.93
0.030	0.99	0.99	0.99	0.96	0.98	0.98
0.040	1.00	1.00	1.00	1.05	0.97	0.96
0.050	1.00	1.00	1.00	0.93	0.94	0.94
0.060	0.99	0.99	0.98	1.02	0.97	0.97
0.070	1.00	1.00	1.00	1.07	0.97	0.97
0.080	1.00	1.00	1.00	1.01	0.98	0.98
0.088	1.00	1.00	0.99	0.90	0.92	0.93

Table 4.

Results of simulation study for different Bernoulli widths B when p=1 and α=0.5 . Effective of Order Pomix sampling with different Bernoulli width B. Sequential poisson sampling follows if B=0. $MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-Expectation of the variance estimator $\hat{V}$. Each row of table 2 is divided by the first row.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.66	0.64	0.64	0.68	0.67	0.67
0.020	0.61	0.55	0.56	0.61	0.55	0.55
0.030	0.63	0.53	0.53	0.64	0.53	0.53
0.040	0.69	0.53	0.53	0.68	0.50	0.50
0.050	0.78	0.53	0.53	0.78	0.55	0.55
0.060	0.92	0.55	0.55	0.92	0.53	0.53
0.070	1.16	0.59	0.60	1.20	0.60	0.61
0.080	1.59	0.66	0.67	1.54	0.67	0.68
0.088	2.43	0.77	0.78	2.31	0.75	0.74

Table 5.

Results of simulation study for different Bernoulli widths B when p=1 and α=1 . Effective of Order Pomix sampling with different Bernoulli width B. Sequential Poisson sampling follows if B=0. $MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-Expectation of the variance estimator $\hat{V}$. Each row of table 2 is divided by the first row.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	1.00	1.00	1.00	1.00	1.00	1.00
0.010	0.94	0.89	0.89	0.91	0.85	0.85
0.020	1.01	0.84	0.84	0.94	0.77	0.77
0.030	1.14	0.82	0.82	1.11	0.76	0.76
0.040	1.33	0.81	0.81	1.19	0.73	0.73
0.050	1.58	0.81	0.81	1.54	0.76	0.76
0.060	1.89	0.82	0.82	2.05	0.81	0.81
0.070	2.32	0.84	0.84	2.31	0.82	0.82
0.080	2.91	0.87	0.87	2.78	0.86	0.86
0.090	3.75	0.92	0.92	3.82	0.87	0.87
0.100	5.12	1.00	1.00	5.01	0.98	0.98

Table 6.

Results of simulation study for different Bernoulli widths *B* when α=1 . Change of effectiviness between *p*=2 and *p*=1. Sequential poisson sampling follows if B=0. $MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-expectation of the variance estimator $\hat{V}$

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	1.49	1.19	1.19	1.42	1.16	1.16
0.02	2.01	1.41	1.40	1.89	1.46	1.46
0.03	2.49	1.63	1.63	2.35	1.64	1.63
0.04	2.92	1.89	1.88	3.10	1.95	1.95
0.05	3.36	2.21	2.20	3.49	2.16	2.15
0.06	3.94	2.66	2.64	3.38	2.47	2.46
0.07	4.64	3.24	3.21	4.07	3.16	3.13
0.08	5.85	4.20	4.18	5.72	4.00	3.93
0.088	7.72	5.45	5.59	7.14	6.04	5.84

Table 7.

Results of simulation study for different Bernoulli widths *B* when α=1 . Change of effectiviness between Order (fixed sample size) Pomix sampling and non fixed size Pomix sampling. Sequential poisson sampling follows if *B*=0 in Order Pomix sampling.

$MVC(\hat{Y})$ = MC-variance of the point estimator $\hat{Y}$ and $MCE(\hat{V})$ = MC-expectation of the variance estimator $\hat{V}$. Table 1 in annex4 is divided by 2 in annex 2.

Bernoulli width B	$MVC(\hat{Y})$			$MCE(\hat{V})$		
	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$	$\hat{Y}^{HT}$	$\hat{Y}^{GREG}$	$\hat{Y}^{SRAT}$
0.000	0.20	0.99	0.99	0.20	1.03	1.03
0.010	0.19	1.00	1.00	0.19	0.97	0.97
0.020	0.20	1.00	1.00	0.21	0.95	0.95
0.030	0.22	0.99	0.99	0.21	1.03	1.03
0.040	0.25	1.00	1.00	0.25	0.95	0.95
0.050	0.28	0.99	0.99	0.30	0.97	0.97
0.060	0.32	0.99	0.99	0.35	1.06	1.05
0.070	0.36	1.00	1.00	0.36	1.00	1.00
0.080	0.41	0.99	0.99	0.41	1.00	1.00
0.090	0.47	0.98	0.99	0.47	0.97	0.97
0.100	0.54	0.99	0.99	0.51	0.96	0.97

ANNEX 5

Notations in formulas

U	*Population*
U_p	*Model group*
U^{TA}	*Take-all part of population*
n^{TA}	*Size of take-all population*
N^R	*Size of Take-some population*
U^R	*Take-some part of population*
N	*Population size*
k, l	*Index for population unit (In section 8.2 k means also the order number of which is under experiment)*
s	*Sample*
n_s	*Realized sample size*
n^{BE}, n^{PO}	*Units drawn from Bernoulli and Poisson sampling areas*
n^R	*Sample size of the rest of population*
r_k	*Random number of unit k*
PRN	*Permanent random number*
D	*Constant shift interval*
B	*Width of Bernoulli part in PoMix sampling*
Q_k	*Size measure for PoMix sampling (In section 8.3 also ranking variables for order sampling)*
A_k	*Size measure for PoMix sampling in* U^R
ϕ_k, η_k	*Modified random numbers for unit k*

$$\sum_U y_k = \sum_{k=1}^{N} y_k$$

$$\sum_s y_k = \sum_{k=1}^{n} y_k$$

$$\sum_{s_p} y_k = \sum_{k=1}^{s_p} y_k$$

s_p	*A sample of a model group*
x_k	*Value of auxiliary variable of unit k*
$\mathbf{x}_k$	*Auxiliary variable vector of unit k*
X	*Population total of auxiliary variable*
$\mathbf{X}$	*Population total of auxiliary variable vector*

$p(s)$	*Sample design*
I_k	*Inclusion indicator*
π_k	*First order inclusion probability of unit k*
π_{kl}	*Second order inclusion probability of units k and l*
a_k	*Sample weight* $1/\pi_k$
g_k	*g-weight*
$w_k = a_k g_k$	
$E(\hat{\theta})$	*Expected value of estimator* $\hat{\theta}$
$V(\hat{\theta})$	*Variance of estimator* $\hat{\theta}$
$B(\hat{\theta})$	*Bias of estimator* $\hat{\theta}$
S^2	*Population variance*
$MSE(\hat{\theta})$	*Mean square error of estimator* $\hat{\theta}$
$BR(\hat{\theta})$	*Bias ratio of estimator* $\hat{\theta}$
E_0, E_{00}	*Permutations*
$\hat{Y}$	*Horwitz-Thompson estimator for population total*
$\hat{Y}^{\pi ps}$	*Total estimator for Poisson* πps *sampling*
$\hat{Y}^{\pi psMo}$	*Total estimator for Modified Poisson* πps *sampling*
$\hat{Y}^{SEpps}$	*Total estimator for Sequential Poisson sampling*
$\hat{Y}_r$	*Ratio estimator of population total*
$\hat{Y}^{HT}$	*Horwitz-Thompson estimator of population total (Take-all part separated)*
$\hat{Y}^{SRAT}$	*Separate ratio estimator of population total (Take-all part separated)*
$\hat{Y}^{GREG}$	*Generalized regression estimator of population total (Take-all part separated)*
RB	*response burden*
β_j	*Response load of questionnaire j*
u	*Number of units already accepted in the sample*
P_0	*Probability of empty sample*
$F(.)$	*Order distribution*
$OS(n,F)$	*Order sampling of size n of order distribution F*

Notes used in section 8.1

X^{TOT}	*Sample total*
r_h	*Ratio of sample totals in stratum h*
R_h	*Ratio of population totals in stratum h*
w_h	*weight for stratum h*
$\overline{X}$	*population mean*
$S^2{}_h$	*Approximated variance of* R_h
$\hat{\overline{Y}}$	*Unbiased estimator for population mean*
t	*Number of Large units or take-all units*
c	*Coefficient of variation = desired level of precision*
p_d	*Probability of drawing decreasing unit in sample*
P_d	*Original inclusion probability of* p_d
p_I	*Probability of drawing decreasing unit in sample*
P_I	*Original inclusion probability of* p_I

TUTKIMUKSIA-SARJA
RESEARCH REPORTS SERIES

Tilastokeskus on julkaissut Tutkimuksia v. 1966 alkaen, v. 1990 lähtien ovat ilmestyneet seuraavat:

164. **Henry Takala,** Kunnat ja kuntainliitot kansantalouden tilinpidossa. Tammikuu 1990. 60 s.

165. **Jarmo Hyrkkö,** Palkansaajien ansiotasoindeksi 1985=100. Tammikuu 1990. 66 s.

166. **Pekka Rytkönen,** Siivouspalvelu, ympäristöhuolto ja pesulapalvelu 1980-luvulla. Tammikuu 1990. 70 s.

167. **Jukka Muukkonen,** Luonnonvaratilinpito kestävän kehityksen kuvaajana. 1990. 119 s.

168. **Juha-Pekka Ollila,** Tieliikenteen tavarankuljetus 1980-luvulla. Helmikuu 1990. 45 s.

169. **Tuovi Allén – Seppo Laaksonen – Päivi Keinänen – Seija Ilmakunnas,** Palkkaa työstä ja sukupuolesta. Huhtikuu 1990. 90 s.

170. **Ari Tyrkkö,** Asuinolotiedot väestölaskennassa ja kotitaloustiedustelussa. Huhtikuu 1990. 63 s.

171. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Nuorten koulutus ja kotitausta. Toukokuu 1990. 115 s.

171b. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Education and the family background of the young in Finland. 1990. 115 pp.

172. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Eriarvoisuus kuoleman edessä. Sosioekonomiset kuolleisuuserot Suomessa 1971–85. Kesäkuu 1990. 145 s.

173. **Jukka Muukkonen,** Sustainable development and natural resource accounting. August 1990. 96 pp.

174. **Iiris Niemi – Hannu Pääkkönen,** Time use changes in Finland in the 1980s. August 1990. 118 pp.

175. **Väinö Kannisto,** Mortality of the elderly in late 19th and early 20th century Finland. August 1990. 50 pp.

176. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Socio-economic mortality differences in Finland 1971–85. December 1990. 108 pp.

177. **Jaana Lähteenmaa – Lasse Siurala,** Nuoret ja muutos. Tammikuu 1991. 211 s.

178. **Tuomo Martikainen – Risto Yrjönen,** Vaalit, puolueet ja yhteiskunnan muutos. Maaliskuu 1991. 120 s.

179. **Seppo Laaksonen,** Comparative Adjustments for Missingness in Short-term Panels. April 1991. 74 pp.

180. **Ágnes Babarczy – István Harcsa – Hannu Pääkkönen,** Time use trends in Finland and in Hungary. April 1991. 72 pp.

181. **Timo Matala,** Asumisen tuki 1988. Kesäkuu 1991. 64 s.

182. **Iiris Niemi – Parsla Eglite – Algimantas Mitrikas – V.D. Patrushev – Hannu Pääkkönen,** Time Use in Finland, Latvia, Lithuania and Russia. July 1991. 80 pp.

183. **Iiris Niemi – Hannu Pääkkönen,** Vuotuinen ajankäyttö. Joulukuu 1992. 83 s.

183b. **Iiris Niemi – Hannu Pääkkönen – Veli Rajaniemi – Seppo Laaksonen – Jarmo Lauri,** Vuotuinen ajankäyttö. Ajankäyttötutkimuksen 1987–88 taulukot. Elokuu 1991. 116 s.

184. **Ari Leppälahti – Mikael Åkerblom,** Industrial Innovation in Finland. August 1991. 82 pp.

185. **Maarit Säynevirta,** Indeksiteoria ja ansiotasoindeksi. Lokakuu 1991. 95 s.

186. **Ari Tyrkkö,** Ahtaasti asuvat. Syyskuu 1991. 134 s.

187. **Tuomo Martikainen – Risto Yrjönen,** Voting, parties and social change in Finland. October 1991. 108 pp.

188. **Timo Kolu,** Työelämän laatu 1977–1990. Työn ja hyvinvoinnin koettuja muutoksia. Tammikuu 1992. 194 s.

189. **Anna-Maija Lehto,** Työelämän laatu ja tasa-arvo. Tammikuu 1992. 196 s.

190. **Tuovi Allén – Päivi Keinänen – Seppo Laaksonen – Seija Ilmakunnas,** Wage from Work and Gender. A Study on Wage Differentials in Finland in 1985. 88 pp.

191. **Kirsti Ahlqvist,** Kodinomistajaksi velalla. Maaliskuu 1992. 98 s.

192. **Matti Simpanen – Irja Blomqvist,** Aikuiskoulutukseen osallistuminen. Aikuiskoulutustutkimus 1990. Toukokuu 1992. 135 s.

193. **Leena M. Kirjavainen – Bistra Anachkova – Seppo Laaksonen – Iiris Niemi – Hannu Pääkkönen – Zahari Staikov,** Housework Time in Bulgaria and Finland. June 1992. 131 pp.

194. **Pekka Haapala – Seppo Kouvonen,** Kuntasektorin työvoimakustannukset. Kesäkuu 1992. 70 s.

195. **Pirkko Aulin-Ahmavaara,** The Productivity of a Nation. November 1992. 72 pp.

196. **Tuula Melkas,** Valtion ja markkinoiden tuolla puolen. Joulukuu 1992. 150 s.

197. **Fjalar Finnäs,** Formation of unions and families in Finnish cohorts born 1938–67. April 1993. 58 pp.

198. **Antti Siikanen – Ari Tyrkkö,** Koti – Talous – Asuntomarkkinat. Kesäkuu 1993. 167 s.

199. **Timo Matala,** Asumisen tuki ja aravavuokralaiset. Kesäkuu 1993. 84 s.

200. **Arja Kinnunen,** Kuluttajahintaindeksi 1990=100. Menetelmät ja käytäntö. Elokuu 1993. 89 s.

201. **Matti Simpanen,** Aikuiskoulutus ja työelämä. Aikuiskoulutustutkimus 1990. Syyskuu 1993. 150 s.

202. **Martti Puohiniemi,** Suomalaisten arvot ja tulevaisuus. Lokakuu 1993. 100 s.

203. **Juha Kivinen – Ari Mäkinen,** Suomen elintarvike- ja metallituoteteollisuuden rakenteen, kannattavuuden ja suhdannevaihteluiden yhteys; ekonometrinen analyysi vuosilta 1974 – 1990. Marraskuu 1993. 92 s.

204. **Juha Nurmela,** Kotitalouksien energian kokonaiskulutus 1990. Marraskuu 1993. 108 s.

205a. **Georg Luther,** Suomen tilastotoimen historia vuoteen 1970. Joulukuu 1993. 382 s.

205b. **Georg Luther,** Statistikens historia i Finland till 1970. December 1993. 380 s.

206. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Naiset huipulla. Huhtikuu 1994. 64 s.

207. **Wangqiu Song,** Hedoninen regressioanalyysi kuluttajahintaindeksissä. Huhtikuu 1994. 100 s.

208. **Anne Koponen,** Työolot ja ammattillinen aikuiskoulutus 1990. Toukokuu 1994. 118 s.

209. **Fjalar Finnäs,** Language Shifts and Migration. May 1994. 37 pp.

210. **Erkki Pahkinen – Veijo Ritola,** Suhdannekäänne ja taloudelliset aikasarjat. Kesäkuu 1994. 200 s.

211. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Women at the Top. July 1994. 66 pp.

212. **Olavi Lehtoranta,** Teollisuuden tuottavuuskehityksen mittaminen toimialatasolla. Tammikuu 1995. 73 s.

213. **Kristiina Manderbacka,** Terveydentilan mittarit. Syyskuu 1995. 121 s.

214. **Andres Vikat,** Perheellistyminen Virossa ja Suomessa. Joulukuu 1995. 52 s.

215. **Mika Maliranta,** Suomen tehdasteollisuuden tuottavuus. Helmikuu 1996. 189 s.

216. **Juha Nurmela,** Kotitaloudet ja energia vuonna 2015. Huhtikuu 1996. 285 s.

217. **Rauno Sairinen,** Suomalaiset ja ympäristöpolitiikka. Elokuu 1996. 179 s.

218. **Johanna Moisander,** Attitudes and Ecologically Responsible Consumption. August 1996. 159 pp.

219. **Seppo Laaksonen** (ed.), International Perspectives on Nonresponse. Proceedings of the Sixth International Workshop on Household Survey Nonresponse. December 1996. 240 pp.

220. **Jukka Hoffrén,** Metsien ekologisen laadun mittaaminen. Elokuu 1996. 79 s.

221. **Jarmo Rusanen – Arvo Naukkarinen** – Alfred Colpaert – Toivo Muilu, Differences in the Spatial Structure of the Population Between Finland and Sweden in 1995 – a GIS viewpoint. March 1997. 46 pp.

222. **Anna-Maija Lehto,** Työolot tutkimuskohteena. Marraskuu 1996. 289 s.

223. **Seppo Laaksonen** (ed.), The Evolution of Firms and Industries. June 1997. 505 pp.

224. **Jukka Hoffrén,** Finnish Forest Resource Accounting and Ecological Sustainability. June 1997. 132 pp.

225. **Eero Tanskanen**, Suomalaiset ja ympäristö kansainvälisestä näkökulmasta. Elokuu 1997. 153 s.

226. **Jukka Hoffrén**, Talous hyvinvoinnin ja ympäristöhaittojen tuottajana –Suomen ekotehokkuuden mittaaminen. Toukokuu 1999. 154 s.

227. **Sirpa Kolehmainen,** Naisten ja miesten työt. Työmarkkinoiden segregoituminen Suomessa 1970–1990. Lokakuu 1999. 321 s.

228. **Seppo Paananen**, Suomalaisuuden armoilla. Ulkomaalaisten työnhakijoiden luokittelu. Lokakuu 1999. 152 s.

229. **Jukka Hoffrén,** Measuring the Eco-efficiency of the Finnish Economy. October 1999. 80 pp.

230. **Anna-Maija Lehto – Noora Järnefelt** (toim.), Jaksaen ja joustaen. Artikkeleita työolotutkimuksesta. Joulukuu 2000. 264 s.

231. **Kari Djerf**, Properties of some estimators under unit nonresponse. January 2001. 76 pp.

232. **Ismo Teikari,** Poisson mixture sampling in controlling the distribution of response burden in longitudinal and cross section business surveys. March 2001. 120 pp.

The Research Report series describe the Finnish society in the light of up-to-date research results. Scientific studies that are carried out at Statistics Finland or that are based on the data sets of Statistics Finland are published in the series. The series also includes collections of scientific articles such as conference proceedings.

In business surveys the response burden is an important aspect since one business may fall in surveys many times in a given time interval. This raises a question. How to even out this burden as fairly as possible? Poisson sampling, with Bernoulli sampling and strict Poisson sampling as special cases, has been found in the early seventies to have good properties as regards sample co-ordination.

A new approach for sample co-ordination, the Poisson Mixture (PoMix) sampling is introduced in this study. This is a sampling scheme which partly uses Bernoulli sampling scheme and partly strict Poisson sampling scheme. The study also proves this sampling scheme to be more efficient than the traditional Poisson sampling.

Tilastokeskus, myyntipalvelu
PL 4V
00022 TILASTOKESKUS
puh. (09) 1734 2011
faksi (09) 1734 2500
myynti.tilastokeskus@tilastokeskus.fi
www.tilastokeskus.fi

Statistikcentralen, försäljningstjänsten
PB 4V
00022 STATISTIKCENTRALEN
tfn (09) 1734 2011
fax (09) 1734 2500
myynti.tilastokeskus@stat.fi
www.stat.fi

Statistics Finland, Sales Services
P.O.Box 4V
FIN-00022 STATISTICS FINLAND
Tel. +358 9 1734 2011
Fax +358 9 1734 2500
myynti.tilastokeskus@stat.fi
www.stat.fi

ISSN 0355–2071
=Tutkimuksia
ISBN 951–727–873
Tuotenumero 9456