On Methods for Real Time Sampling and Distributions in Sampling

Kadri Meister

Doctoral Dissertation
Department of Mathematical Statistics
Umeå University
SE-901 87 Umeå
Sweden

ISBN 91-7305-795-9
Printed by Print & Media
Umeå 2004

Emale

To my mother

Ei jõua kirjutada puhtandit
me selles elus.
Nagu on, nii jääb
see paranduste mitmekordne räga.

⋆⋆⋆

No time to write the final draft
within this lifetime.
Leave as it is
the thorny tangled thicket of corrections.

⋆⋆⋆

Vi hinner inte med en renskrift
här i livet.
Som det var, så får det vara
med alla rättelsers härva.

Doris Kareva

Contents

List of papers

The present thesis is based on the following papers.

A. Meister, K. and Bondesson, L. (2001). Some real time sampling methods. *Research Report 2001–2*, Department of Mathematical Statistics, Umeå University. Revised version.

B. Meister, K. (2002). Asymptotic considerations concerning real time sampling methods. *Statistics in Transition*, **5**, 1037–1050.

C. Meister, K. (2004). Some different methods to get stationary Bernoulli sequences with negative correlations for sampling applications. Manuscript.

D. Traat, I., Meister, K. and Sõstra, K. (2001). Statistical inference in sampling theory. *Theory of Stochastic Processes*, **7**, 301–316.

E. Traat, I., Bondesson, L. and Meister, K. (2004). Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference*, **123**, 395–413.

F. Meister, K. and Traat, I. (1999). On the design-based distribution of some estimators in survey sampling. *Theory of Stochastic Processes*, **3**, 324–329.

Abstract

This thesis is composed of six papers, all dealing with the issue of sampling from a finite population. We consider two different topics: real time sampling and distributions in sampling. The main focus is on Papers A–C, where a somewhat special sampling situation referred to as real time sampling is studied. Here a finite population passes or is passed by the sampler. There is no list of the population units available and for every unit the sampler should decide whether or not to sample it when he/she meets the unit. We focus on the problem of finding suitable sampling methods for the described situation and some new methods are proposed. In all, we try not to sample units close to each other so often, i.e. we sample with negative dependencies. Here the correlations between the inclusion indicators, called sampling correlations, play an important role. Some evaluation of the new methods are made by using a simulation study and asymptotic calculations. We study new methods mainly in comparison to standard Bernoulli sampling while having the sample mean as an estimator for the population mean. Assuming a stationary population model with decreasing autocorrelations, we have found the form for the nearly optimal sampling correlations by using asymptotic calculations. Here some restrictions on the sampling correlations are used. We gain most in efficiency using methods that give negatively correlated indicator variables, such that the correlation sum is small and the sampling correlations are equal for units up to lag m apart and zero afterwards. Since the proposed methods are based on sequences of dependent Bernoulli variables, an important part of the study is devoted to the problem of how to generate such sequences. The correlation structure of these sequences is also studied.

The remainder of the thesis consists of three diverse papers, Papers D–F, where distributional properties in survey sampling are considered. In Paper D the concern is with unified statistical inference. Here both the model for the population and the sampling design are taken into account when considering the properties of an estimator. In this paper the framework of the sampling design as a multivariate distribution is used to outline two-phase sampling. In Paper E, we give probability functions for different sampling designs such as conditional Poisson, Sampford and Pareto designs. Methods to sample by using the probability function of a sampling design are discussed. Paper F focuses on the design-based distributional characteristics of the π-estimator and its variance estimator. We give formulae for the higher-order moments and cumulants of the π-estimator. Formulae of the design-based variance of the variance estimator, and covariance of the π-estimator and its variance estimator are presented.

Key words: Finite population sampling, inferential issues, real time sampling, sequential sampling methods, negative sampling correlations, model-design-based inference, multivariate Bernoulli and multinomial designs.

2000 Mathematics Subject Classifications: 62D05, 62E15, 60G10.

Preface

PhD-studies are like a roller coaster ride, with eagerness to experience new things as well as learn more about ones own limits. However, one never really knows where it may lead. Just when you think there is nowhere else to go but down, you are pulled back up and everything is turned around. Occasionally, the ride gives you a butterflies-in-the-stomach feeling but not for long. You are borne away by negative G-forces that may throw you off if you do not hold on tight. The steepest drop can be several dozen degrees, so to experience the thrills you have to put up with the screams.

As much as you have enjoyed the intellectual (and emotional) roller coaster ride, you will be relieved when it is all over. In the end, when you look back – with pride – you are going to think it was a great ride! I want to thank the many passengers in the roller coaster cars for taking this memorable ride with me.

On the first seat, my supervisor Professor Lennart Bondesson. It has been great to work with you and take advantage of your broad knowledge and enthusiasm for solving new problems. I am grateful to you for all helpful discussions and suggestions. There is still lots more to learn from you about mathematical statistics ... and even more about mushrooms!

Docent Imbi Traat, my co-supervisor at the University of Tartu. Thank you for all the encouragement through the years, and for sharing your ideas and knowledge, both about statistics and life in general.

Professor Emeritus Gunnar Kulldorff. You have played a central role in the choice of sampling theory as the topic of my thesis. Thank you for your support in 'practical matters' during my first years of studies, and for putting Umeå on my map.

My colleagues at the Department of Mathematical Statistics. You have contributed to this work by providing a comfortable and inspiring work atmosphere. Ingrid Westerberg Eriksson and Lennart Nilsson, a special warm thanks for all your help and care!

All current and past (counted from my first visit to Umeå) PhD-students. You have been a great inspiration for me through all these years! Thank you for sharing all the laughs and screams! Special thanks to Mykola for taking one of the photos on page 9.

Everyone who has commented on and suggested improvements for different parts of this thesis. I am grateful to Paul Haemig for improving my English.

My friends everywhere. For all the enjoyable times we have together! Warmest thanks especially to the ones from Estonia, for always asking when I will be back there again.

My mother, my brother and his family. Thank you for always being supportive and understanding, even if you did not fully understand what I was occupied with.

And last but not least, Magnus, on the seat next to me. Thank you for taking all these roller coaster loops with me! Without you I would never have even thought about going through PhD-studies.

Umeå, December 2004
Kadri Meister

1 Introduction

We often want to have some information about a specified set of units, a finite population. What is the proportion of unemployed in a country as a whole and in various regions of the country? What is the average expenditure for food in households? What is the total volume of a forest stand? One way to get answers to such questions is to collect information from every unit that belongs to the population. Another way is to use some sampling technique. Sampling is choosing in some way just part of a population – called *a sample* – so that with an appropriate study of the sample we may say something about the whole population.

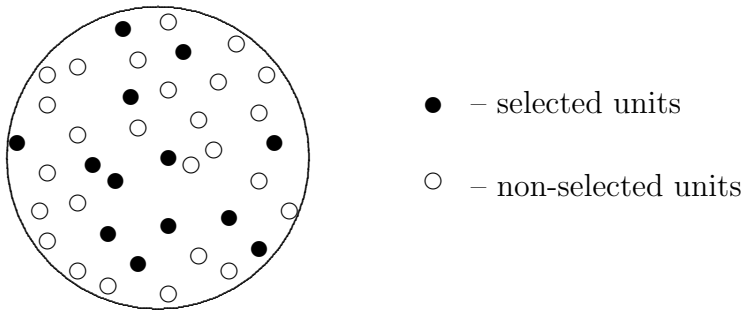

Figure 1: Illustration of a population and a sample

One of the obvious questions here is how to choose the sample. There are many different more or less advanced techniques. As a simple example, we may select every 10th unit from the sampling frame – the list of population units. The sample may be drawn by a probability mechanism, called *sampling design*. The latter then plays a central role by determining how the units are selected from the population and also the essential statistical properties of the random quantities calculated from the sample. The usual inference problem in sampling is to estimate some summary characteristic of the population, such as the total or the mean, after observing the sample. Additionally, we would like to say something about the precision of the estimate, i.e. the size of the sampling error. The latter results from the fact that the sample is part of the population and estimates from the sample may not be identical to the corresponding population quantities. A sampling method that is easy to implement and gives estimates with good precision should be used.

Furthermore, nonsampling errors such as imperfections in the sampling frame, nonresponse, measurement errors, etc., may occur whether the entire population or a sample of the population is studied and all should be taken into account.

Survey sampling theory has a long history. The idea of observing a representative sample instead of the entire population goes back to the late 19th century, and the work of the Norwegian statistician A. N. Kiær was influential at that time. It is generally agreed that a fundamental step for development of the probability sampling approach was made by Neyman (1934). Early developments in theory and methods concentrated on efficient sampling designs and associated estimation techniques for population totals or means. More recently, the methods for analysis of survey data that take into account the complexity of the sampling design – both sampling and nonsampling errors – have developed rapidly.

We will not make any attempt to summarize the history of the subject – this has already been done by many excellent review papers published in recent years. For a brief history of the development of survey sampling, and in particular the probability sampling approach, see Hansen, Dalenius & Tepping (1985) and Rao & Bellhouse (1990). Rao (1999) covers some current topics on survey sampling, including developments in survey design, data collection and processing, issues related to inference from survey data, resampling methods for analysis of survey data, and small area estimation.

This thesis consists of six papers dealing with the issue of sampling from a finite population. Two different topics are treated: real time sampling and distributions in sampling. In the following sections the theoretical background is given. Section 2 starts with basic notation and gives definitions used further in the thesis. Then, different approaches to inference when sampling from a finite population are shortly outlined and some estimators are given. In Section 3, a real time sampling situation is described and some suitable sampling methods are presented. In Section 4, we give a brief overview of different problems that form the base for the work in Papers D–F. In Section 5, the contents of the different papers in the thesis are summarized. In Section 6, some conclusions are given and open problems are discussed.

2 Inferential issues

2.1 Basic notation

The fundamental problem in sampling theory is to make inference (estimation, variance estimation, confidence intervals), for example about the population total by observing a sample selected according to a specified sampling method.

Let $U = \{1, 2, \ldots, N\}$ denote a finite population of N units. Traditionally, there are two distinguished definitions of a sample in the sampling literature

(e.g. Särndal, Swensson & Wretman, 1992, pp. 27–28, 49–50). In the case of sampling without replacement (WOR) where a unit, once sampled can not be sampled again, a sample s is defined as a subset of population U. For sampling with replacement (WR), where a unit is allowed to be sampled several times, we can look at the vector of selected units. Here the unit selected at the kth draw is the kth element of the vector. Hence this vector includes information on both the selecting order of the units and the number of times each unit is selected. A sample s can be defined as the set of distinct units in this vector.

Random selection of a sample s of size n from a finite population U can usually be described by some probability mechanism, called sampling design. A sampling design $p(s)$ is defined as a probability distribution on sets; $p(s)$ is the probability to get the sample (set) s.

The inclusion of a given unit i in a sample is a random event indicated by the random variable I_i, an inclusion indicator. Here $I_i = 0$ if unit i is not included in the sample and $I_i = k_i$, where $k_i > 0$ is some integer, if it is included (k_i shows the number of times unit i is selected). The N indicators can be summarized in vector form as $\boldsymbol{I} = (I_1, I_2, \ldots, I_N)$. WR-sampling designs are not used so often and if not stated otherwise, we have a WOR-design in mind in the remaining discussion. In the case of WOR-design, the inclusion indicator I_i takes only the values 0 or 1. It is a random variable from a Bernoulli distribution with

$$\begin{aligned} E(I_i) &= \pi_i, \quad Var(I_i) = \pi_i(1-\pi_i), \\ Corr(I_i, I_j) &= R_{ij} = (\pi_{ij} - \pi_i\pi_j)/\sqrt{\pi_i\pi_j(1-\pi_i)(1-\pi_j)}, \quad i \neq j, \end{aligned}$$

where $\pi_i = Pr(I_i = 1)$ and $\pi_{ij} = Pr(I_i = 1, I_j = 1)$ are the inclusion probabilities of first- and second-order, respectively, and are fundamental characteristics of a given sampling design.

The sampling design is often chosen to yield certain desired first- and second-order inclusion probabilities. The estimators used in survey sampling are functions of the inclusion indicators, hence the moments of the estimators are functions of the first- and higher-order inclusion probabilities, respectively. Knowledge of the π_i and π_{ij} alone is normally sufficient for one of the primary goals in survey sampling, namely to determine exact or approximate expected values and variances of the estimators, and to determine variance estimators.

The correlations $R_{ij} = Corr(I_i, I_j)$, here referred to as sampling correlations, are sometimes used instead of π_{ij} for describing different sampling designs in the present thesis.

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)$ be the vector of values of the study variable Y for the population units. Sometimes there is some information available about another variable X prior to sampling. This variable, called an auxiliary variable, can carry information about the study variable and hence assist in the estimation. For example, in many surveys of human populations, we can have the value of the study variable from a previous census as an auxiliary variable. The goal is to obtain estimators with increased precision for the study variable.

In the following discussion, we use Y (and X) to represent both the variable and the population total. The exact meaning is given in the context.

There have been some different viewpoints on how to make inference in survey sampling, depending on the source of randomness. The major approaches are briefly described in the following section.

2.2 Different approaches to inference when sampling from a finite population

Foundational aspects of inference from sample survey data have attracted a lot of attention since the 1960's.

Finite population sampling is the area of statistics in which the primary mode of the analysis is based on the sampling design, the distribution of $\boldsymbol{I}$, rather than on statistical models for the variable Y. This is called *the design-based approach*. Here the population units have fixed but unknown values of the study variable Y. The uncertainty in estimates obtained by sampling thus stems from the fact that only part of the population is studied. This means that randomness is only coming from the sampling design, i.e. the sole random quantities are the inclusion indicators. While the population characteristic remains fixed, the estimate of it depends on which sample is selected.

Example 1. An often used estimator for the population total Y is

$$\hat{Y}_{HT} = \sum \frac{I_i Y_i}{\pi_i},$$

the well-known Horvitz-Thompson estimator (called the HT-estimator or the π-estimator) introduced by Horvitz & Thompson (1952). Here and in the following discussion, any sum $\sum$ without summation restrictions means a sum over the whole population. Since the values of Y_i/π_i are given constants, we can easily see that this estimator is design-unbiased, i.e. $E_p(\hat{Y}) = Y$, where the index p denotes that the expectation is with respect to the sampling design. Design-unbiased variance estimators have been derived by Horvitz & Thompson (1952), Yates &

Grundy (1953) and Sen (1953). The goodness of the HT-estimator depends on the values of π_i, working best when π_i is approximately proportional to the value of Y_i.

In general, using design-based inference with careful attention to the sampling design and using a suitable estimation method, we can obtain estimates that have good properties without relying on any assumptions about the population itself.

In some sense an opposite approach is *the model-based approach*, which requires some probability model ξ for the N-dimensional distribution of $\boldsymbol{Y}$. Here the population values $Y_1, Y_2, \ldots, Y_N$ are random variables, generated by a model ξ, often called a superpopulation model. The specification of ξ can vary from something crude and basic to a very detailed description, depending on what assumptions the model maker feels are rational to make. For example, a simple model to adopt is that $Y_1, Y_2, \ldots, Y_N$ are independent with $E_\xi(Y_i) = \mu$ and $Var_\xi(Y_i) = \sigma^2$. The actual finite population values are considered to be realizations of $Y_1, Y_2, \ldots, Y_N$ and the inference still concerns the finite population and its parameters.

Example 2. We can write the population total Y as

$$Y = \sum Y_i = \sum_{i \in s} Y_i + \sum_{i \notin s} Y_i.$$

Here the values of non-selected units are predicted using the model and then a suitable estimator $\hat{Y}$ is derived. Hence the estimators are dependent on the chosen model. Now, however, the properties of the $\hat{Y}$, e.g. unbiasedness, are derived with respect to the model ξ and not with respect to the sampling design p.

Model-based inference may have advantages if the model is appropriate. The challenge with the model-based approach lies in the question of how to specify the model exactly. The major weakness of the model-based approach is that if the model is misspecified, it may lead to invalid conclusions.

Models are widely used also within the design-based inference, both in sampling design and in estimation, but in a *model-assisted* way using the terminology of Särndal, Swensson & Wretman (1992, p. 227). The values of the study variable are still fixed, but assumptions about a possible model that has generated these values are made. For example, we can assume some correlation structure in the model. In the case of available auxiliary information, a relation between the study variable and the auxiliary variable can be assumed. These assumptions are not expected to hold exactly.

Example 3. Let $\hat{Y}$ be some estimator of the population total. We can study how this estimator behaves under different population models, for example by calcu-

lating $E_\xi(Var_p(\hat{Y}))$. Here the estimator is design-based and the formula of the variance estimator is also derived with respect to the sampling design. However, we can get reduced variance in situations where the assumed model is approximately valid. We use this framework when considering suitable methods for the real time sampling situation.

Example 4. Let X be an auxiliary variable. Then we can use the Horvitz-Thompson ratio estimator for estimating the total Y

$$\hat{Y}_{HTR} = X\frac{\hat{Y}_{HT}}{\hat{X}_{HT}} = X\hat{R},$$

where X is a population total, $R = Y/X$ and $\hat{Y}_{HT}$ and $\hat{X}_{HT}$ are the HT-estimators for the Y- and X-totals. It is most appropriate, i.e. has small design-based variance, when the following model is approximately valid: $E_\xi(Y_i) = \beta X_i, Var_\xi(Y_i) = \sigma^2 X_i$, that is, when there is an approximate proportionality between the variables X and Y in the population.

Hence, in the model-assisted approach, the model is used as a tool for motivating the choice of a sampling method or an estimator. Inference remains design-based and the design-based properties of the estimators are not dependent on the chosen model. Here, increased precision of the estimators may be achieved.

Strengths and weaknesses of the above-mentioned approaches have been discussed in several articles. A paper by Royall (1970) can be considered a starting point for the "design-based versus model-based" debate, which continues e.g. in Särndal (1978), Hansen, Madow & Tepping (1983), and Kalton (2002). For an overview of the debate, the reader is referred to Little (2004).

To sum up, different approaches are often used depending on the context. Design-based methods are used when calculating descriptive statistics, such as totals and means, based on large probability samples. To handle nonsampling errors, e.g. nonresponse, models are necessary even in the design-based approach. Many of the developments in survey sampling during recent years have been concerned with the application of model-based methods for small-area estimation and nonsampling errors.

Remark. In this thesis, design-based inference is mainly considered. However, sometimes some superpopulation models are used for studying properties of different designs or estimators. In Papers A–B, our notation of the study variable values differs from the notation in well-known sampling textbooks such as Cochran (1977) and Särndal, Swensson & Wretman (1992). The symbols y_i are commonly used to denote fixed, but unknown, population values. We use the

symbol Y_i to denote both the random variable associated with the ith population unit (if some model is used) and a fixed finite population value. This notation is also used e.g. in Raj (1968).

2.3 More about some estimators

In this section, the concern is with design-based inference. We assume a sampling design that ensures positive first-order inclusion probabilities, π_i, and also positive second-order inclusion probabilities, π_{ij}, for all $i \neq j$. Such designs permit design-unbiased estimators and variance estimators.

The sample size n, is given by $n = \sum I_i$. Hence $E(n) = \sum \pi_i$ and

$$Var(n) \;=\; \sum \pi_i(1-\pi_i) + 2\sum\sum_{i<j}(\pi_{ij} - \pi_i\pi_j).$$

Sampling methods that give random sample size are often avoided in practice since the variable sample size will cause an increase in variance for certain types of estimators. The double sum in $Var(n)$ depends on the correlations between the inclusion indicators and it is clear that one must sample with negative dependencies in order to get low sample size variation.

Further, we consider estimation of the population mean $\bar{Y}$. The Horvitz-Thompson estimator of $\bar{Y}$ is

$$\hat{\bar{Y}}_{HT} \;=\; \frac{1}{N}\sum I_i \breve{Y}_i = \frac{1}{N}\hat{Y}_{HT}, \tag{1}$$

where $\breve{Y}_i = Y_i/\pi_i$, which is the notation introduced in Särndal, Swensson & Wretman (1992, p. 42). There are different forms for the variance of the HT-estimator. For fixed size sampling designs, it can be given in the Sen-Yates-Grundy form (Sen, 1953; Yates & Grundy, 1953) by

$$Var(\hat{\bar{Y}}_{HT}) \;=\; -\frac{1}{2N^2}\sum\sum(\pi_{ij} - \pi_i\pi_j)(\breve{Y}_i - \breve{Y}_j)^2$$

and its unbiased estimator is given by

$$\widehat{Var}(\hat{\bar{Y}}_{HT}) \;=\; -\frac{1}{2N^2}\sum\sum \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}(\breve{Y}_i - \breve{Y}_j)^2 I_i I_j,$$

where the summation is effectively over the sample.

Since we can view $\bar{Y}$ as a ratio of two population totals, Y and N, respectively, another possible estimator of $\bar{Y}$ is

$$\hat{\bar{Y}}_{HTR} \;=\; \frac{\hat{Y}_{HT}}{\hat{N}} = \frac{\sum I_i \breve{Y}_i}{\sum \breve{I}_i}, \tag{2}$$

where $\breve{I}_i = I_i/\pi_i$. For $\pi_i \equiv \pi$ the estimator (2) reduces to the sample mean $\bar{y}$.

Since $\bar{y}$ is a nonlinear function of the inclusion indicators, it has a slight bias. An approximate MSE of the estimate $\bar{y}$ is given by

$$MSE(\bar{y}) \approx -\frac{1}{2(E(n))^2} \sum\sum a_{ij}(Y_i - Y_j)^2,$$

where $E(n) = N\pi$ and the coefficients $a_{ij} = E((I_i - n/N)(I_j - n/N))$ are functions of the second-order inclusion probabilities.

The estimator $\hat{\bar{Y}}_{HTR}$ performs better than (1) in cases where the sample size is variable. Therefore it is used as an estimator for real time sampling methods.

3 Real time sampling situations

In the first half of the thesis, Papers A–C, the concern is with the real time sampling situation and the corresponding sampling methods. The sampling situation under study may have been considered before by others. However, searches in the sampling literature and databases have not revealed any systematic attention and research about the underlying case.

3.1 Background

When taking a sample from a finite population, there is often a sampling frame available. Units to be measured are selected from the frame by some procedure corresponding to a chosen sampling design. There are many different methods to use for this case, depending for example on the amount of accessible auxiliary information.

Consider now a sampling situation where there is *no sampling frame available* and where *units come*, one by one, *in real time to a sampler*. For every unit the sampler should decide immediately whether or not to sample it by using some sequential selection method. Alternatively, *the sampler visits the units* in some order chosen by the sampler. The population size N is usually unknown before the sampling but will eventually become known. This kind of sampling is here referred to as *real time sampling*.

Example 5. The units may be passengers using the public transportation in some city. It might be possible that the units in the population can order themselves. In an extreme case, the units can order themselves by taking into account

the sampling scheme. Some units, like people coming to a customs control, may want to avoid being sampled. In such a case, systematic sampling of every fifth unit may be a bad sampling method.

Figure 2: Passengers as sampling units

Example 6. The units may be every tree in a forest stand, with some of the trees sampled by a forester walking around. The visiting order of the trees in the stand is chosen more or less subjectively by the sampler. Therefore, the selection of units for the sample is influenced by the sampler's subjective choice of order.

Figure 3: A forest stand where some trees are sampled

In the following section we present some methods that partly avoid the negative effects of ordering.

3.2 Suitable sampling methods

As stated before, several well-known sampling methods need a list of the population and are not suitable to use in real time sampling situations. Here, some sequential selection method is needed. Systematic sampling and Poisson sampling are two possible well-known methods (see e.g. Särndal, Swensson & Wretman, 1992, Chapter 3) to apply for sampling a finite population that passes or is passed by the sampler. Both methods are also easy to use for the sampler.

For systematic sampling in its basic form, the first unit in the sample is drawn by simple random sampling from among the first μ units in the population. Then, every μth population unit is chosen. However, for systematic sampling there is a problem with the variance estimator because the condition about positive second-order inclusion probabilities for every pair of units is not fulfilled. This method is included for comparisons in Paper A but excluded in Papers B and C.

For Poisson sampling, independent $U(0,1)$ random variables $U_1, U_2, \ldots, U_N$ are generated to perform the sampling. The selection or non-selection of unit i is decided by the following rule: if $U_i < \pi_i$, where π_i is a predetermined inclusion probability, unit i is selected, otherwise not. Because the π_is can be specified in a variety of ways, Poisson sampling corresponds to a whole class of designs. In the general case, the inclusion probabilities are often chosen to be proportional to some size measure. A special case of Poisson sampling is Probability Proportional to Prediction (3P) sampling, as described by, for example, Husch, Miller & Beers (2003, p. 355). Here no auxiliary information is accessible prior to sampling and the inclusion probabilities are based on predicted values of the study variable. This method is commonly recommended as a sampling method in forestry, at least in the US. When all the units have the same inclusion probability, $\pi_i \equiv \pi$, Poisson sampling is called Bernoulli sampling. Since sampling methods with equal inclusion probabilities are studied in Papers A–C, just the latter is considered for comparisons in the following discussion.

For Bernoulli sampling, population units are selected independently of each other. Neighbouring units may have similar study variable values, and therefore it may be wise not to sample units close to each other too often. Hence sampling with negative dependencies, i.e. with negative sampling correlations, would be more efficient. We introduce some sampling methods that partly take into account *the pre-history of the sampling*.

These methods, as well as Bernoulli sampling, usually give random sample size. However, for sampling methods with negative sampling correlations, the variability of the sample size is smaller than for Bernoulli sampling.

There are several alternative methods with which to perform the sampling. When applying Bernoulli sampling, we sample population units for which the corresponding random numbers are below some level. By permitting these random numbers to be dependent, two simple extensions of this method are obtained: *sampling according to a stationary process* and *sampling according to some function of independent uniform random variables*.

For sampling according to a stationary process, a strictly stationary process $\{Z_i\}$ with given correlations $r_k = Corr(Z_i, Z_{i+k})$ is used as a tool for defining the values of the inclusion indicators. We set

$$I_i \;=\; \begin{cases} 1 \text{ if } Z_i \leq c \\ 0 \text{ otherwise} \end{cases},$$

where c is a given constant determined by the desired inclusion probability π, i.e. $c = F^{-1}(\pi)$, where F is the distribution function of Z_i. The process $\{Z_i\}$ must be easy to simulate, so a stationary standard normal process is mainly considered, both in Papers A and C.

Sampling according to some function of independent uniform random variables is introduced in Paper C. We generate independent $U(0,1)$ random variables $U_{1-m}, U_{2-m}, U_{3-m}, \ldots$, where $m \geq 1$ is some fixed integer. The general rule for defining the value of I_i is

$$I_i \;=\; \begin{cases} 1 \text{ if } U_i \leq h(U_{i-m}, \ldots, U_{i-2}, U_{i-1}) \\ 0 \text{ otherwise} \end{cases},$$

where $h(\cdot)$ is some function that gives negative correlations between the inclusion indicators. The choice of $h(\cdot)$ is a delicate task. We look at three cases – a linear, a product and a minimum function – where an explicit formula can be derived for the sampling correlations. An attempt to generalize this method by using a geometric approach is made. The focus is on the 2-dependent case, i.e. sets of inclusion indicators more than lag 2 from each other are independent. We look at a subset B of the unit cube and set $I_i = 1$ if $(U_{i-2}, U_{i-1}, U_i) \in B$. Here the question is how to choose B to get an efficient sampling method.

Figure 4 shows the form of a suitable B when using a unit square in the 1-dependent case.

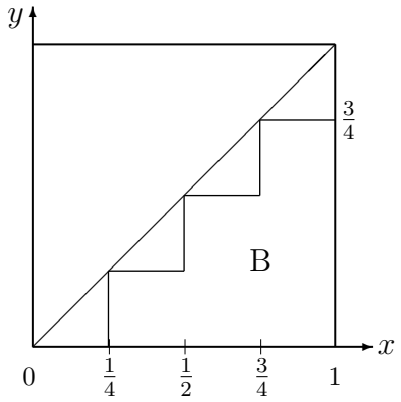

Figure 4: The optimal subset B for $\pi = 3/8$

We can also look at the step length between two sampled units. Instead of a one point distribution as for systematic sampling, or a geometric distribution as for Bernoulli sampling, we can use other step length distributions to get different sampling methods. An attempt to use such a method was made by Fan, Muller & Rezucha (1962) with a truncated geometric distribution but it did not give encouraging results. Methods used in this thesis were introduced by Bondesson (1986), where conditions for suitable step length distributions were given. Results from renewal theory (see e.g. Feller, 1957, Chapter 13) are used and therefore the obtained method is called *renewal sampling*. The latter is considered in papers A and C, where different methods are derived by using different step length distributions.

A wish to get a sampling design with almost fixed sample size for the real time sampling situation is behind the idea for *stratified sampling with a random start* described in Paper C. This method is distinct from the other suggested methods. Here the dependence structure for the inclusion indicators is not the same as for the other methods and this explains somewhat different results.

3.3 Comparison of methods

The evident problem in Papers A–C is how the suggested methods for the real time sampling situation work compared to suitable well-known methods, especially compared to Bernoulli sampling.

Using the sample mean $\bar{y}$ for estimating the population mean, we are interested in the variability of the estimate when using different sampling methods. The mean square error of the sample mean, $\mathrm{MSE}(\bar{y})$, is used as an efficiency measure.

In Paper A, some of the methods are compared in a simulation study for a special population model. Here renewal sampling and sampling according to a stationary process are studied among the new methods. Some good results are observed, but some final conclusions are that one should look at larger populations as well as different population models.

For getting better insight when the new methods work well, some asymptotic calculations are made in Paper B. We assume that the sequence of inclusion indicators, $\{I_i\}$, is a real stationary Bernoulli process in discrete time. Besides, some assumptions are made about the population structure. We look at the asymptotic model-based expectation of MSE($\bar{y}$) that depends on both model correlations and sampling correlations. We compare this quantity in the case of Bernoulli sampling and general sampling with negative sampling correlations. The question here is in which form the sampling correlations should be to gain most in efficiency by using more advanced real time sampling methods.

Some restrictions are used for finding the best correlations. At first, we use the condition $0 \geq R_k \geq -d$, $k = 1, 2, 3, \ldots$, where d is some constant depending on the predetermined π-value. This condition is important for getting stable MSE-estimates. Also, a condition on the sum of sampling correlations is used; see below.

It appears that the sampling method with negative sampling correlations has advantages for a stationary population model with decreasing autocorrelations. The resulting optimal sampling correlations have a simple form. A sampling method with equal negative correlations $R_1 = R_2 = \ldots = R_m$ and $R_k = 0$ for $k > m$, where m is some integer, has approximately an optimal structure. Therefore, to find sampling methods that allow such a correlation structure is of great interest.

3.4 About stationary Bernoulli processes

In the following discussion the focus is mainly on m-dependent processes, i.e. sets of variables more than lag m apart are independent.

Let $\{I_i\}$ be a stationary Bernoulli process in discrete time. Hence, we sample with equal inclusion probabilities and the sampling correlations

$$R_k = Corr(I_i, I_{i+k}) = \frac{\pi_{i,i+k} - \pi^2}{\pi(1-\pi)} \tag{3}$$

are equal for every two units k steps apart.

Since the proposed sampling methods are based on such Bernoulli processes as described above, it is important to pay some attention to the properties of these processes. Let m be some fixed integer. We want to sample units with a lag up to m apart dependently, and those more than lag m apart independently. Here, negative dependencies are desired and hence the sequences $\{I_i\}$ such that $R_k < 0$ for $k = 1, 2, \ldots, m$, and 0 for $k > m$ are of interest. The possible values of the sampling correlations play an important role. We would like to have these as negative as possible for gaining most in efficiency compared to Bernoulli sampling. As seen from (3) there is an obvious lower bound for the correlation R_k, that is $R_k \geq -\pi/(1-\pi)$. However, more conditions are needed for getting a sequence $\{I_i\}$ with desired properties.

A condition on the sum of sampling correlations is used due to the following reasoning: Under which conditions is a sequence $\{R_k\}$ with $R_0 = 1, R_k \leq 0, k = 1, 2, 3, \ldots$, a correlation sequence for a stationary Bernoulli process? In the corresponding stationary normal process case with correlations $r_0 = 1, r_k \leq 0, k = 1, 2, 3, \ldots$, the problem has a simple solution. In this case, it is shown in Bondesson (2003) that a necessary and sufficient condition is

$$\sum_{k=1}^{\infty} r_k \geq -\frac{1}{2}.$$

In the Bernoulli case, the problem is more complicated. A necessary condition is that the correlation sum is not lower than $-1/2$. This can be shown by looking at the variance of the sum of the Bernoulli variables and using the fact that it has to be nonnegative. The correlation sum $\sum_{k=1}^{m} R_k$ should preferably be close to its lower bound. In Bondesson (2003), it is conjectured that for $m = 1$ the lower bound for the correlation sum is $-1/3$ at least for a 1-dependent sequence and this still unproved conjecture is used in Paper B for finding the form of the optimal correlations.

For $m \geq 2$, however, the correlation sum can actually be below $-1/3$. There is no general result about the lower limit in this case. To get better insight, correlation sums are studied for different sampling methods in Paper C. Numerical examples about possible negative correlations and minimum values of the correlation sum with given values of π are presented for different sampling methods.

There are many different ways to derive sequences $\{I_i\}$ with negative correlations. Many good results have been achieved but at the moment we can not give a simple general solution of how to define a sequence $\{I_i\}$ with predetermined π, and with equal negative correlations $R_1 = R_2 = \ldots = R_m$, and $R_k = 0$ for $k > m$. More research is certainly needed.

4 On distributional characteristics in sampling

The second half of the thesis includes three diverse papers about distribution theory in survey sampling. As mentioned before, sampling design is a basic notion in sampling theory when using design-based and model-assisted approaches. In Section 2.1, a sampling design is defined as a probability distribution on sets. One can also consider the vector of inclusion indicators, $\boldsymbol{I}$, here called a design vector, and define the multivariate distribution of $\boldsymbol{I}$, $p(\boldsymbol{k}) = Pr(\boldsymbol{I} = \boldsymbol{k})$, as a sampling design, see e.g. Traat (2000). Here, $\boldsymbol{k} = (k_1, k_2, \ldots, k_N)$ is the outcome of $\boldsymbol{I}$ and is an indicator of the realized sample, being always of dimension N.

This definition of a sampling design has some advantages. It covers both WR- and WOR-sampling designs and also uses the knowledge and tools worked out for multivariate distributions (see e.g. Johnson, Kotz & Balakrishnan, 1997). The multivariate distribution of $\boldsymbol{I}$ is a multivariate Bernoulli distribution for WOR-sampling designs and some other multivariate discrete distribution for WR-sampling designs. Drawing a sample from a population U according to some sampling design means generating an outcome from a multivariate design distribution $p(\boldsymbol{k})$.

The vector form of a sample can be naturally incorporated into the inference process. The design vector can be used in the definition of the survey data and different statistics are functions of $\boldsymbol{I}$ and study variable values. This definition of sampling design is used in Paper D for considering statistical inference in sampling. Here both the population model and the sampling design is included in the inference, and characteristics of some estimators are considered.

A vector form of survey data makes it possible to use matrix tools in the sampling theory. The way of expressing the sampling design as a multivariate distribution, as well as expressing the samples as vectors, is used by Ollila (2004), who considers different sample resampling methods for variance estimation.

Using the framework of the design vector, it is possible to give the probability function of a sampling design. Probability functions of several important sampling designs, e.g. the conditional Poisson, Sampford, and Pareto, are given in Paper E. Some different ways to use probability functions for drawing a sample are described.

In Paper F, the main focus is on the π-estimator for the population total. We consider a design-based distribution of this estimator, and give design-based higher-order moments and cumulants which have not received so much attention in the sampling literature. Since the π-estimator is a linear function of the inclusion

indicators, its higher-order moments and cumulants depend on the respective quantities of the inclusion indicators. Some attention is also paid to the variance estimator of the π-estimator. It is well known that it is design-unbiased, but in this paper its design-based variance is considered.

5 Summary of the papers

The present thesis contains six papers that consider different topics in survey sampling. In Papers A–C, we focus on the real time sampling situation and corresponding sampling methods. Some sampling methods suitable for this case are proposed and conditions giving efficient methods are considered. In Papers D–F, distribution theory in survey sampling is considered from different viewpoints. In Papers D and E, we use the design vector in the definition of the sampling design. In Paper D, unified inference is considered where both the population model and the sampling design are taken into account. In Paper E, the focus is on probability functions of different sampling designs. Formulae of the probability functions of several sampling designs are given. In Paper F, we consider the design-based distributional characteristics of the π-estimator of the population total and its variance estimator.

Paper A. Some real time sampling methods

Let us consider a sampling situation where a finite population passes or is passed by the sampler. There is no list of the population and for every population unit the sampler has to decide whether or not to sample a unit when he/she meets it. The population size N is most often unknown before the sampling but eventually it becomes known. Among the well-known methods, systematic sampling and Poisson (Bernoulli) sampling can be adopted for taking a sample in this sampling situation. The aim of this paper is to find more suitable methods. We consider sampling methods with equal inclusion probabilities. Two general classes of sampling methods are introduced: renewal sampling and sampling according to a stationary process, which in some way take into account the pre-history of the sampling. To make some evaluation of these methods, a simulation study is performed. Different methods with equal first-order inclusion probabilities are compared numerically. In comparison with systematic sampling and Bernoulli sampling, some promising results are derived for the new methods in the case of a specific population model.

Paper B. Asymptotic considerations concerning real time sampling methods

A general real time sampling method with negative sampling correlations is considered in this paper. This method is compared with Bernoulli sampling, having independent inclusion indicators and hence zero sampling correlations. The sampling correlations enter into the formulae of variances of different estimators and hence negative sampling correlations can contribute to reduced variances.

The aim of this paper is to find conditions under which the sampling method with negative sampling correlations should be used in favor of the well-known Bernoulli sampling. In other words: in which form should the correlations be for gaining most in efficiency using some more advanced sampling method than Bernoulli sampling? Some asymptotic calculations are made for finding the solution to this problem. We assume some stationary model for the study variable. The asymptotic model-based expectation of the mean square error (MSE) of the sample mean is studied. It depends on both population model correlations and sampling correlations. Bernoulli sampling and general sampling with negative correlations are compared with respect to expected MSE, with some restrictions on the values of possible correlations.

It appears that the latter sampling method has advantages for a stationary population model with decreasing autocorrelations. The optimal sampling correlations have a simple form and approximately optimal sampling designs should have equal negative correlations for units up to lag m apart and zero correlations afterwards. The achieved gain in efficiency depends also on how strongly correlated the population values are.

Paper C. Some different methods to get stationary Bernoulli sequences with negative correlations for sampling applications

In the previous paper we gave an approximately optimal form for sampling correlations to use to achieve more efficient sampling than standard Bernoulli sampling. The main question in this paper is how to get the nearly optimal sampling method with given first-order inclusion probabilities and negative sampling correlations. A stationary sequence of Bernoulli variables can be used as a tool for defining different methods. Hence, problems such as how to get these Bernoulli sequences and the correlation structure of these sequences are the main focus of this paper. Special attention is paid to the lowest possible correlation sum for given inclusion probabilities and for different methods.

The emphasis is on finding a way to get a Bernoulli sequence with negative correlations and with as low correlation sum as possible. Special cases of sampling according to a stationary process and renewal sampling are studied. A generalization of Bernoulli sampling is obtained by using some function of independent uniform random variables for defining the Bernoulli variables. Further, a stratification method with random start is described. Numerical examples showing the lowest possible correlation sum for different methods are presented.

Paper D. Statistical inference in sampling theory

The framework with sampling design as a discrete multivariate distribution is used in this paper. This covers both with and without replacement sampling designs. We look at a unified approach where both the sampling design and the population model are taken into account in inference. A general linear estimator and its variance formula are given. It includes, for example, well-known design-based estimators as special cases. A two-phase sampling design is studied on a more general level where with and without replacement designs are allowed in both phases.

Paper E. Sampling design and sample selection through distribution theory

We use a multivariate approach with a unifying treatment of with and without replacement designs. The probability functions of several important sampling designs, such as hypergeometric, conditional Poisson, Sampford, and general order sampling designs are presented. A list-sequential method for generating a sample from any given design using probability function is developed.

Paper F. The design-based distribution of some estimators in survey sampling

In this paper, we consider design-based distributional characteristics of the π-estimator for the population total. Usually only the design-based expected value and variance are given in the sampling literature. General formulae for the design-based kth-order moments and cumulants of the π-estimator are presented in this paper. The π-estimator is a linear function of the inclusion indicators, hence its kth-order moments and cumulants depend on the kth-order moments and cumulants of the inclusion indicators, respectively.

For estimating the variance of the π-estimator, an unbiased variance estimator is used. We give formulae for the design-based variance of this estimator and the covariance of the π-estimator and its variance estimator.

6 Conclusions and open problems

The recent work of the author has concentrated on the real time sampling situation. Hence, only questions arising in Papers A–C are treated in this section. The topics considered in the other papers have been developed further by others. For example, probability functions of some sampling designs are treated and applied in Bondesson, Traat & Lundqvist (2004).

Consider real time sampling with equal inclusion probabilities. The initial goal of the work was to find suitable sampling methods (e.g. renewal sampling) and compare these to well-known methods that are possible to adopt in this case. Several new methods are proposed. All have in common that we in some way take into account how the previous units are sampled and try not to sample units close to each other so often, i.e. we sample with negative dependencies. It makes sense intuitively that in the real time sampling population, units close to each other may have similar study variable values. Hence, the proposed methods should improve the estimation.

We have chosen to study sampling methods with negative correlations mainly in comparison to standard Bernoulli sampling, while using the sample mean as an estimator for the population mean. Also, we assume a stationary population model with decreasing autocorrelations. In this case, we have found the form for the nearly optimal sampling correlations by using asymptotic calculations. Here some restrictions on the sampling correlations are used. We gain most in efficiency using methods that give negatively correlated indicator variables and such that *the correlation sum is small and that the correlations R_k are equal for units up to lag m apart and zero afterwards.*

Instead of giving further attention to the estimation problem, the focus has changed to study sequences of negatively correlated Bernoulli variables. It is of main interest to study how to generate such sequences with desired properties.

As shown in Paper C, there are many different ways to get negative correlations. However, despite many attempts, a nice, simple, optimal solution for choosing which method to use for a given π-value to obtain the desired sampling correlations has not been found. Thus, practical suggestions for the sampler are not given. More research is certainly needed.

There is a difficult unsolved problem concerning the minimum value of the correlation sum. We have presented some numerical calculations of the correlation sum for different methods that give some possibility to compare these methods with each other.

We do not assume that the population model is exactly true. Hence it is not always clear what the best value of m should be. It depends both on conditions of the sampling correlations and also on the population model. For larger π-values, i.e. π close to $1/2$, only units close to each other are usually sampled dependently. In the case of small π-values, we can have more correlations negative. It would be of interest to study more how the estimates and variance estimates behave, depending on how many sampling correlations are negative. Here different population structures should give different results.

Would the form of the optimal sampling correlations hold also for other estimators? In the design-based framework, different estimators are functions of the inclusion indicators, hence sampling correlations enter into the variance formulae. If it is possible to have the variance estimator in Sen-Yates-Grundy form, then under some restrictions sampling with negative correlations should result in reduced variance.

Not much has been said about sampling with unequal inclusion probabilities. Sampling according to a stationary process can easily be extended to the case of unequal probabilities. For the other methods it is not so obvious how to act for sampling with unequal probabilities. For instance, we can first apply sampling with equal inclusion probabilities and then decide, for a selected unit, if it will be sampled or not by using some predictions or auxiliary variables.

If the sampler determines the order of the units in the population (or the population units can order themselves), then Poisson sampling gives total protection against subjective ordering bias. The methods proposed in this thesis protect partially against such ordering. It would be of interest to find a measure of this protection.

The inclusion probability π is assumed to be determined before taking the sample. However, the way of choosing the value of π has not been much discussed. Some guess about the population size N should be made.

The preliminary plans of this work also included a plan to use different methods on real data, for instance data from forestry. However, since the focus of the work changed over time, this plan was set aside. Still, the author's belief is that the ideas behind the presented methods will eventually find good practical applications.

References

Bondesson, L. (1986). Sampling of a linearly ordered population by selection of units at successive random distances. *Report No. 25*, Section of Forest Biometry, Swedish University of Agricultural Sciences, Umeå.

Bondesson, L. (2003). On a Minimum Correlation Problem. *Statistics & Probability Letters*, **62**, 361–370.

Bondesson, L., Traat, I. and Lundqvist, A. (2004). Pareto Sampling versus Sampford and Conditional Poisson Sampling. *Research Report 2004–6*, Department of Mathematical Statistics, Umeå University.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications.* Vol. I, 2nd ed. New York: Wiley.

Hansen, M. H., Dalenius, T. and Tepping, B. J. (1985). The development of sample surveys of finite populations. In: A. C. Atkinson and S. E. Fienberg (eds.), *A Celebration of Statistics: The ISI Centenary Volume*, 327–354. New York: Springer–Verlag.

Hansen, M. H., Madow, W. G. and Tepping, B. G. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys. *Journal of the American Statistical Association*, **78**, 776–793.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Husch, B., Miller, C. I. and Beers, T. W. (2003). *Forest Mensuration.* 4th ed. New York: Wiley.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* New York: John Wiley.

Kalton, G. (2002). Models in the Practice of Survey Sampling (Revisited). *Journal of Official Statistics*, **18**, 129–154.

Little, R. J. (2004). To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*, **99**, 546–556.

Neyman, J. (1934). On the different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of The Royal Statistical Society*, **97**, 558–625.

Ollila, P. (2004). A Theoretical Overview for Variance Estimation in Sampling Theory with Some New Techniques for Complex Estimators. *Research Report 240, September 2004*, Statistics Finland.

Raj, D. (1968). *Sampling theory.* New York: McGraw–Hill.

Rao, J. N. K. (1999). Some current trends in sample survey theory and methods. *Sankhyā, Ser. B*, **61**, 1–57.

Rao, J. N. K. and Bellhouse, D. R. (1990). History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis. *Survey Methodology* , **16**, 3–29.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377–387.

Särndal, C.–E. (1978). Design-based and Model-based Inference in Survey Sampling. *Scandinavian Journal of Statistics*, **5**, 27–52.

Särndal, C.–E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer–Verlag.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119–127.

Traat, I. (2000). Sampling design as a multivariate distribution. In: T. Kollo, E.-M. Tiit and M. Srivastava, (eds), *New trends in Probability and Statistics 5, Multivariate Statistics*, 195–208. Vilnius, Utrecht: TEV/VSP.

Yates, F. and Grundy, P. M. (1953). Selection without replacement within strata with probability proportional to size. *Journal of the Royal Statistical Society, Ser. B*, **15**, 235–261.