ESSAYS ON MODEL ASSISTED SURVEY PLANNING

BY

ANDERS HOLMBERG
Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, Uppsala, Tuesday, June 3, 2003 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in Swedish.

Abstract

The quality of sample survey results is to a large degree dependent on decisions made by survey statisticians at the planning stage. The first paper studies two issues related to the planning stage: (i) the sensitivity of model assumptions concerning the relation between the size measure and a study variable in without replacement probability proportional-to-size sampling (wrs sampling), and (ii) properties of practicable sample selection schemes for fixed size wrs sampling. These two issues are also addressed in the second paper, which furthermore discusses the consequences of the presence of more than one study variable and to what extent the auxiliary information used in the design and that used in the estimators interact.

The evident problem in both the first and the second paper is how to choose an overall efficient sampling design when there are several important study variables with various relationships to the available auxiliary variables. The third paper suggests a diagnostic tool to support the choice of design, and on the basis of three criteria of overall efficiency optimal designs are derived.

The optimal designs presented in the third paper may not be fully satisfactory in meeting specified precision requirements for separate estimators. To achieve a design that is tailor-made to meet such requirements, optimisation must be done under restrictions. Though the underlying optimisation problem is only outlined in paper III, a solution involving non-linear programming methods is given in the fourth paper. By way of an example based on an application to a Swedish business population, the fourth paper compares a design obtained through non-linear programming algorithms with designs discussed in paper III as well as designs based on the same principal ideas as those discussed in the first two papers. The paper suggests a flexible solution regarding how to use auxiliary information exhaustively and to provide diagnostic support for the final design choice.

Anders Holmberg, Department of Information Science, Box 513, Uppsala University, SE-75120 Uppsala, Sweden

© Anders Holmberg 2003

ISSN 0282-7492
ISBN 91-554-5623-5
urn:nbn:se:uu:diva-3417 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-3417)
PAPERS INCLUDED IN THE THESIS

This thesis is based on the following papers, which are referred to in the text by their Roman numerals:


Reprints were made with the permission of the publishers.
Contents

Acknowledgements ........................................................................................................... 5

1 Introduction ..................................................................................................................... 7
  1.1 Concepts of survey planning in this thesis................................................................. 7
    1.1.1 On the use of auxiliary, or supplementary, information .................................. 9
  1.2 Using models to assist survey planning .................................................................... 12
  1.3 General definitions .................................................................................................... 14
    1.3.1 Sampling design and inclusion probabilities .................................................. 14
    1.3.2 Descriptive planning stage models ..................................................................... 15
  1.4 Auxiliary variables and the strategy choice ............................................................. 16
    1.4.1 The estimator choice .......................................................................................... 17
    1.4.2 The GREG estimator for population totals ...................................................... 17
    1.4.3 Use of auxiliary variables in the sampling design, the single parameter case ... 18
    1.4.4 Probability proportional-to-size sampling (πps sampling) ................................ 19
    1.4.5 The Anticipated Variance ................................................................................. 22
    1.4.5 Planning values and guesstimates ....................................................................... 24
  1.5 The multivariate sampling problem .......................................................................... 25

2 Summary of the papers ................................................................................................. 28
  2.1 Reflections on unequal probability sampling strategies ........................................... 28
  2.2 On the choice of strategy in unequal probability sampling ..................................... 30
  2.3 A multiparameter perspective on the choice of sampling design in surveys ............. 32
  2.4 On the choice of optimal sampling design in business surveys with several important study variables ................................................................. 34
  2.5 Minor comments and concluding remarks .................................................................. 35

References ......................................................................................................................... 38
Acknowledgements

It is a special pleasure to write this part of the thesis, namely the expression of my indebtedness to people and institutions whose help has made this work possible. This thesis marks the end of my graduate studies, and over the years many have directly or indirectly contributed to the outcome. However, as always the author accepts the full responsibility for any inadequacies.

The most important person to whom I wish to express my true and deepest gratitude is Professor Bengt Swensson. From the beginning, he has had a major influence on this work and on my development within the field of statistics. With great knowledge, insight, scrutiny and patience he has read and commented on numerous manuscripts. This thesis and the papers on which it is based would not have developed to their present standard, had it not been for Bengt. I also wish to thank Bengt’s wife Lena for her patience and understanding of my need to sometimes consult Bengt outside office hours.

Bengt has also co-authored one paper. Special credit must be given to him as well as to Patrik Flisberg and Mikael Rönnqvist co-authoring another paper. Their efforts have had a direct impact on paper IV and it has been stimulating and rewarding working with them.

This work is the fruit of efforts I have made both at the Department of Statistics, Örebro University, and Statistics Sweden’s R&D Department. Consequently, there are many people who have actively and morally supported my efforts. I am especially grateful to my present and former managers at Statistics Sweden, Eva Elvers, Agneta Lissmats, Evert Blom and Lars Lyberg, for their encouragement and for allowing me to spend parts of my working hours on this thesis. In a similar way, I am also indebted to Olle Häggbom and Erik Wallgren at Örebro University. Olle, as head of department, helped by finding means for adequate funding and Erik, as director of studies, helped by planning and administering reasonable teaching schedules. I also gratefully acknowledge the financial support provided by Adolf Lindgrens Stiftelse.
In addition, I want to thank personnel at Uppsala University, my formal supervisor Anders Christoffersson (who took important initial steps regarding my admission as a graduate student) and Gunilla Klaar as well as Anders Ågren (for appreciated help on practical as well as administrative matters throughout the years).

There are many colleagues and friends I wish to thank for showing understanding and interest in my work. In particular, I thank Martin Axelson. Over the years we have shared statistical problems, offices, train compartments and leisure, and without his sense of humour and readiness to discuss both serious and tiny matters, this work would have been less joyful. As dear colleagues and friends (also teachers) at Örebro University, Olle Carlsson, Hans Hedén, Åke Holmén (who, alas, did not live to read this), Niklas Karlsson and Erik Wallgren have given me invaluable experience in various fields of statistics as well as explaining how it is taught and practised.

One of the privileges at Statistics Sweden has been the close access to inspiring discussions with skilful survey statisticians. I am obliged to Claes Andersson (for sharing his vast experience and all his tips and solutions helpful to me when I have got stuck in daily work), to Lennart Nordberg (for encouraging detailed discussions and for providing me with realistic data sets), to Sixten Lundström and Håkan L. Lindström (for their excellent tutoring which inspired my career choice) and to Carl-Erik Särndal (for always showing an interest in my work, giving me constructive criticism, discussing innovative ideas and sharing his knowledge). Collectively, I also acknowledge the support from my colleagues at the R&D Department and those I have closely worked with from other departments of Statistics Sweden as well as at Örebro University. Without your assistance and fine co-operation in different projects, this work would have been harder.

Finally, I wish to give tribute to my family and closest friends, especially my parents who did not live to see this, my two sons John and Filip and (at the final stages) Britta. To have your trust and support has been priceless. Thank you!

Vadstena, March 2003
Anders Holmberg
1 Introduction

Whenever there is a need to obtain information on characteristics of a large finite population of people, households, business enterprises or other elementary units, it is natural to consider a sample survey. In most cases, a carefully planned and well-executed sample survey will provide more precise and timelier information at a lower cost than a complete enumeration would. Obviously, taking a sample means that less time to collect and process data is required, but the huge and unwieldy organization required for a complete enumeration often also implies less control of data collection, resulting in many errors. The present thesis treats some of the statistical questions that arise in the planning process of a sample survey, in particular questions concerning the combined choice of a sampling design, a sample selection scheme and estimator(s).

1.1 Concepts of survey planning in this thesis

To plan a medium-sized or large sample survey is a complex undertaking involving a large number of considerations, decisions and choices on the part of a team of – among others – subject matter experts, survey measurement specialists, computer specialists and survey statisticians. In the discussion of sample survey planning in the present thesis the focus is on some of the sta-
istical tasks and the assumption is that much planning work already has been done, which means that far from all aspects of survey planning will be covered. For example, neither tasks such as the formulation of a subject matter problem and its translation to a statistical problem, nor the choice of measurement device and editing procedures (although they are all fundamental in the planning process) will be discussed here. The point of departure for the discussion in the thesis is as follows: a sample survey is to be taken in order to estimate certain well-defined characteristics (parameters) of a well-defined finite population for which an up-to-date good sampling frame exists, i.e. a frame which besides information that makes it possible to get observational contact with selected elements, contains useful data on a set of auxiliary variables. The focus will be on the case where the frame is a direct listing of the population elements, making direct sampling of elements possible; examples are business registers kept by many countries, and registers of the total population found in some European countries.

A typical survey has many study variables and aims to estimate many parameters. Common parameters are totals (e.g. the total monthly value of imported commodities) and functions of totals such as means (e.g. the average income of adult wage-earners), proportions (e.g. the proportion of cultivated area devoted to wheat) and ratios (e.g. the ratio of rental income of commercial premises to that of rented area).

Hence, one important goal for the survey statistician in the planning phase of the survey is to find a strategy – a combination of sampling design and estimators for the most important parameters – which gives the best possible (most precise) parameter estimates for a given cost, or conversely to obtain estimates with a desired precision at lowest possible cost. Survey sampling
theory offers instruments to achieve such a goal. However, for the purpose of coming to a decision on the best feasible strategy, survey theory should be applied as early as possible in the planning process, and it must be balanced by practical and budgetary restrictions.

1.1.1 On the use of auxiliary, or supplementary, information

In a broad sense, auxiliary information – supplementary information is an often-used essentially synonymous expression – is any information of assistance about the population available prior to a survey. For instance, in questionnaire design the language and the ordering of questions could be tailored to suit some characteristics of the respondents, and when data collection methods are specified, decisions about the measurement instrument and decisions on how to implement technical equipment could be greatly improved.

In the present thesis, ‘supplementary information’ will be used in a broad sense. In contrast, ‘auxiliary information’ will be reserved for the following specific kind of information, consisting of two related components: (i) data on a set of variables (which will be called auxiliary variables) available at the planning phase as well as at the estimation phase for every population element; (ii) appropriate a priori knowledge of the structure of relationships between important study variables and (subsets of) the auxiliary variables.

In the search for estimators of high precision, survey sampling theory strongly emphasizes the use of auxiliary information. Often, the sampling frame is equipped with at least a few auxiliary variables that may be useful, while subject matter experts and the survey statistician may have knowledge of the relationship between study variables and auxiliary variables from
similar surveys of similar populations. If no such useful knowledge is available, it will often be worthwhile to make a pilot survey.

The use of auxiliary information can be traced back to the early days when the foundations of sampling theory were laid. In Neyman’s classical paper from 1934, one of the major contributions is the explicit use of auxiliary information. Neyman suggested a partition of the population into strata followed by the selection of a probability sample in each stratum. Cochran (1942) is an early reference which introduces auxiliary information in the estimation procedure. Cochran discussed the regression method of estimation in detail.

Since Neyman, much of the development of sampling theory has evolved around the use of auxiliary information and auxiliary variables. In an introduction to Neyman’s paper, Tore Dalenius states:

“As demonstrated by the developments in the last half-century, supplementary information may be exploited for all aspects of the sample design, the definition of sampling units, the selection design and the estimation method.” (Dalenius 1992)

Because auxiliary variables can be used in a wide variety of ways at the design stage as well as at the estimation stage, efforts to use this information to find efficient estimators lead to many techniques for sampling and estimation. Results on optimum allocation in stratified sampling, on probability proportional-to-size sampling, and on regression or calibration estimators all rely on properties of auxiliary variables. Recent references that illustrate this wide variety are Laaksonen (2002), who attempts to systematize auxiliary variables into different types, and Estevao and Särndal (2002), who show
that there are ten different cases where auxiliary information may be used for calibration in two-phase sampling.

In all of the papers in the present thesis, auxiliary information plays a key role in the planning process. The first paper studies two issues related to the planning stage: (i) the sensitivity of model assumptions concerning the relation between the size measure and a study variable in without replacement probability proportional-to-size sampling (πps sampling), and (ii) properties of practicable sample selection schemes for fixed size πps sampling. These two issues are also addressed in the second paper, which furthermore discusses the consequences of the presence of more than one study variable and to what extent the auxiliary information used in the design and that used in the estimators interact.

The evident problem in both the first and the second paper is how to choose an overall efficient sampling design when there are several important study variables with various relationships to the available auxiliary variables. The third paper suggests a diagnostic tool to support the choice of design, and on the basis of three criteria of overall efficiency optimal designs are derived.

The optimal designs presented in the third paper may not be fully satisfactory in meeting specified precision requirements for separate estimators. To achieve a design that is tailor-made to meet such requirements, optimisation must be done under restrictions. Though the underlying optimisation problem is only outlined in paper III, a solution involving non-linear programming methods is given in the fourth paper. By way of an example based on an application to a Swedish business population, the fourth paper also compares a design obtained through non-linear programming algorithms with
designs discussed in paper III as well as designs based on the same basic ideas as those discussed in the first two papers.

The results presented in the thesis are applicable in multipurpose surveys with several important study variables, where auxiliary information is available and a probability proportional-to-size sampling design is considered, e.g. business and agricultural surveys repeated over time. The discussion, notation and methods presuppose that without replacement direct element sampling is used. However, with proper adjustments, the results should also be of potential use in multipurpose, multistage area sampling, where the first-stage units are chosen proportional to size.

In the following, there is first given what would appear to be a necessary and more detailed theoretical background. In conclusion, there are then offered a summary of the major contributions of the four papers and some comments and remarks.

1.2 Using models to assist survey planning

“Models are appropriately used to guide and evaluate the design of probability samples, but with large samples the inferences should not depend on the model.”

Hansen, Madow and Tepping (1983)

One thing that distinguishes survey sampling theory from the mainstream of statistical theory is that inference concerns parameters of an actual finite population and not unknown parameters of an otherwise (assumed to be) known frequency function of some hypothetical population. The latter is commonly the case in general statistical theory and means that the interest
lies in inference about model parameters. However, since the usual target of our inference in survey sampling is unknown parameters of the finite population, models are usually used differently.

Two main types of inferences are considered in the survey literature, design-based and model-based inference. In design-based inference, the finite population is considered fixed and the variable values are fixed. In model-based inference, the values of the variables in the finite population are assumed realizations from a superpopulation model.

In this thesis, the concern is with design-based inference. The population is fixed, and the study variable values are fixed (but unknown). The stochastic structure is induced by the sampling design, and the sample itself is the random element. However, the superpopulation concept will be used as well, but differently from its use in model-based inference. Superpopulation models will be used merely (i) as a tool for giving compact descriptions of useful a priori knowledge about relations between the study variables and available auxiliary variables, e.g. for describing basic features of finite population scatterplots, and (ii) as guides in the search for efficient strategies. Hence, inference will not depend on the model, but the model will assist it in the sense that the better the superpopulation model depicts the finite population, the better we can hope to succeed in our task of estimating the finite population parameters.
1.3 General definitions

Let $U = \{1, \ldots, k, \ldots, N\}$ be our finite population consisting of $N$ elements. Consider $Q$ study variables $y_1, \ldots, y_q, \ldots, y_Q$, and let $y_{qk}$ denote the unknown value of $y_q$ for population element $k$. In addition, we suppose that there are $P$ auxiliary variables at hand. They are denoted $u_1, \ldots, u_p, \ldots, u_P$, and their values $u_p \ (p = 1, \ldots, P)$ are known for every element $k$ in the population. Furthermore, we will concentrate on the estimation of population totals, i.e. $t_q = \sum_{k \in U} y_{qk} = \sum_{U} y_{qk}$.

1.3.1 Sampling design and inclusion probabilities

To select a set sample $s \subseteq U$ of size $n_s$, a without replacement sampling design, $p(\cdot)$, will be used. The sampling design is the function that gives each possible sample its probability of being selected, and it is essential for determining the properties of estimators to be used in the survey.

Fundamental characteristics of a given sampling design are the inclusion probabilities of first and second order. The first-order inclusion probability of element $k$ is the probability that a sample containing the $k$:th population element is realized. We denote this by $\pi_k \ (k = 1, \ldots, N)$. The second-order inclusion probability (the probability that a sample containing elements $k$ and $l$ is realized) will be denoted by $\pi_{kl} \ (k, l = 1, \ldots, N)$. They are mainly of
importance for variance calculations. Throughout the thesis it is assumed that $\pi_k > 0$ and $\pi_{kl} > 0$ for every $k$ and $l$.

Since $p(\cdot)$ often is a complicated function it is of little direct practical use. However, knowledge of the $\pi_k$ and the $\pi_{kl}$ alone is normally sufficient for our needs, e.g. for determining expected values and variances of estimators.

The planning approach to come up with a good strategy for a sample survey adopted in this thesis results in a decision on a class of preferred sampling designs characterized by a preferred set of first-order inclusion probabilities. To implement a design conforming to this set we need a sample selection scheme. Often, but not always, we can find such a scheme. Sometimes, however, there is no (practical) scheme that implements a design with a set of first-order inclusion probabilities which exactly match the preferred ones. The latter is the case for (medium to large) fixed size probability proportional-to-size sampling designs.

Without replacement, probability proportional-to-size sampling designs ($\mathcal{P}ps$ designs) are present in all four papers of this thesis. Brief summaries of the basic ideas behind these designs are given in the papers and can be found in most standard textbooks.

1.3.2 Descriptive planning stage models

There must be a useful and well-guessed relationship between the auxiliary variables and the study variables if the auxiliary variables are to be of assistance in the choice of strategy. In the following, we assume that the structure of such relationships can be caught by linear models $\xi_q$ such that
$$y_{qk} = x'_{qk} \beta_q + e_{qk}$$ with \( E_{\xi_q} \left( e_{qk} \right) = 0, \quad V_{\xi_q} \left( e_{qk} \right) = \sigma^2_{qk} \) and \( E_{\xi_q} \left( e_{qk} e_{qj} \right) = 0 \ (k \neq j), \) i.e. for \( (k = 1, \ldots, N), \)

$$\begin{cases} E_{\xi_q} \left( y_{qk} \right) = x'_{qk} \beta_q \\ V_{\xi_q} \left( y_{qk} \right) = \sigma^2_{qk} \end{cases}$$ (1)

where \( x'_{qk} = \left( x_{i_k}, \ldots, x_{j_k}, \ldots, x_{j_k} \right) \) is a suitable set of \( J_q \) (positive) auxiliary variables formed from \( u_1, u_2, \ldots, u_p, \) and where \( \beta_q = \left( \beta_{i_k}, \ldots, \beta_{i_k}, \ldots, \beta_{i_k} \right)' \) and \( \sigma^2_{q1}, \ldots, \sigma^2_{qN} \) are model parameters.

### 1.4 Auxiliary variables and the strategy choice

As mentioned previously, the available auxiliary information will influence the survey statistician’s final choice of strategy. The general problem of finding the, in some sense, best strategy has been studied by different authors. In such studies, it seems necessary to accept certain restrictions of the general problem. Sometimes restrictions are made regarding specific classes of estimators, and sometimes the emphasis is placed solely on finding an efficient sampling design with certain properties. Horvitz and Thompson (1952) and Godambe (1955) are classical references that approach the problem, restricting themselves to a class of linear estimators, whilst Neyman (1934) and Dalenius (1950) are examples of references where the sampling design has been focused upon.
1.4.1 The estimator choice

In this thesis, one restriction is that we consider the family of estimators known as GREG (generalized regression) estimators (for details see Särndal, Swensson and Wretman (1992), sections 6.4-6.7). It is a wide family of estimators, however, which includes several well-known estimators used in practice. With use of a model like the one given in equation (1) as a starting point for a GREG estimator, different model expressions give us estimators such as the post-stratified estimator and the simple ratio estimator.

A GREG estimator is approximately unbiased even with poorly chosen models, but with a strong relationship between $y$ and $x$, and a fair knowledge of that relationship, the GREG estimator will, as far as efficiency is concerned, outperform the well-known Horvitz-Thompson estimator which does not use auxiliary information.

1.4.2 The GREG estimator for population totals

The GREG estimator for a population total is defined as

$$\hat{t}_{x_{q'}} = \hat{t}_{y_{q'}} + \left( t_{x_q} - \hat{t}_{x_{q'}} \right) \hat{B}_q. \quad (2)$$

Here, $\hat{t}_{y_{q'}} = \sum_{k \in S} x_{qk} / \pi_k = \sum_{k} y_{qk} / \pi_k$ is the Horvitz-Thompson, or $\pi$ estimator, $t_{x_q} = (t_{x_{q_1}}, \ldots, t_{x_{q_j}}, \ldots, t_{x_{q_k}})$ is a $J_q$-dimensional vector of $x_{q_i}$ totals, $\hat{t}_{x_{q'}}$ is a vector of the corresponding $\pi$ estimators and

$$\hat{B}_q = \left( \sum_s \frac{x_{qsk}x_{qk}'}{c_{qk}\pi_k} \right)^{-1} \sum_s \frac{x_{qsk}y_{qk}}{c_{qk}\pi_k} \quad (3)$$
is an estimated vector of regression coefficients, where $c_{qk}$ is a suitable constant.

The Taylor expansion (design) variance of $\hat{t}_{y_d}$ is given by

$$V_T(\hat{t}_{y_d}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \left( \frac{\pi_k \pi_l}{\pi_k \pi_l} \right) E_{qk} E_{ql}$$

where $E_{qk} = y_{qk} - x_{qk}' B_{q} \ (k = 1, \ldots, N)$ are finite population fit residuals, with $B_{q} = \left( \sum_{k=1}^{K} x_{qk} x_{qk}' / c_{qk} \right)^{-1} \sum_{k=1}^{K} y_{qk} x_{qk}' / c_{qk}$ a finite population regression coefficient. The issue of variance estimation will not be discussed in detail in this thesis. However, for the strategies we consider, we require that at least approximately unbiased estimators for (4) are available.

### 1.4.3 Use of auxiliary variables in the sampling design, the single parameter case

By far the most important application of auxiliary variables in the sampling design is that of stratification. As mentioned above, it goes back to Neyman and it can be justified for several reasons. General reasons given in Särndal et al. (section 3.7) are:

1. Precision requirements of estimates for certain subpopulations, domains, can be assured by using domains as strata.
2. Practical aspects such as nonresponse rates, method of measurement and the quality of auxiliary information may differ between subpopulations, and can be efficiently handled by stratification.
3. The survey organization may be divided into geographical districts, which makes it natural for administrative reasons to let each district be a stratum.

In addition, when the stratification is made on an auxiliary variable which is well-correlated with the important study variables, stratified sampling ensures a good basis for obtaining efficient estimates. Although it is elementary on its own, stratified sampling may be combined with other methods of sampling yielding rather complex sampling designs, e.g. one may consider stratified sampling with probability proportional-to-size sampling (πps sampling) within strata.

1.4.4 Probability proportional-to-size sampling (πps sampling)

As the name proportional-to-size indicates, the elements for a πps sample are selected with probabilities proportional to a size measure. The size measure, \( z \), is one of the auxiliary variables \( u \), or a function (transformation) of one or several auxiliary variables, i.e. \( z = h(u) \). Its value has to be known for every population element, and ideally, if the simple \( \pi \) estimator is to be used, it should be proportional to the study variable.

We illustrate the ideas behind πps sampling by first looking at the \( \pi \) estimator for a single study variable \( y_q \), \( \hat{y}_{\pi} = \sum_s y_{qk} / \pi_k \). We note that, if the inclusion probabilities \( \pi_k \) could be made proportional to \( y_{qk} \), then the ratio \( y_{qk} / \pi_k \) is constant, say \( c \), and for every (fixed size design) sample we would get \( \hat{y}_{\pi} = \sum_s y_{qk} / \pi_k = nc \), which has zero variance. However, proportional inclusion probabilities cannot be derived using the
study variable since all $y_{qk}$ are unknown. Instead, we choose inclusion probabilities proportional to $z$, whose values in turn should be nearly proportional to the values of $y_q$. The first-order inclusion probabilities are then given by

$$\pi_k = \frac{n z_k}{\sum_U z_k} \quad (k = 1, \ldots, N)$$

where $z$ is strictly positive, $\sum_U \pi_k = n$ and $\pi_k \leq 1$.

Occasionally, some $z_k$ values are very large and equation (5) violates $\pi_k \leq 1$. One common method to deal with this is to create a ‘take-all’ stratum, $A$, of those elements and choose them with certainty (i.e. $\pi_k = 1$ for $k \in A$). Then, for the remaining elements ($k \in U - A$), the $\pi_k$ are chosen to be proportional to $z$. The method presented in the fourth paper of this thesis offers another solution, where the numerical optimization method automatically fulfills the restriction $\pi_k \leq 1$.

Henceforth, without-replacement probability proportional-to-size sampling designs based on the size variable $z$ will be denoted by $\pi ps(z)$, and we will assume that $\pi_k \leq 1$.

There are a large number of sample selection schemes to implement a $\pi ps(z)$ design. For example, Brewer and Hanif (1983) list 50 schemes. However, if we exclude random size designs, it has turned out to be hard to devise a scheme for arbitrary sample size $n$ that has a number of desirable properties, e.g. (a) the actual selection of the sample is relatively simple, (b) all first-order inclusion probabilities are strictly proportional to the size vari-
able, (c) the design admits (at least approximately) unbiased estimation of estimator variances. If we also want to be able to base the sample selection on the technique of permanent random numbers (PRN), which is desirable in large survey organizations taking many surveys, some of which are repeated over time, it will be even harder.

Relatively new \( \pi \) designs, attractive to practitioners, are the order sampling designs like sequential Poisson sampling (see Ohlsson (1990, 1998)) and Pareto \( \pi \) (see Rosén (1997a,b) and Saavedra (1995)), and the PoMix design proposed by Kröger, Särndal and Teikari (1999, 2000). In papers I–II, the properties of strategies based on these designs are compared with model-based stratified simple random sampling (mb-STSI) as proposed by Wright (1983).

When \( \pi \) sampling is relevant, we have no reason to restrict ourselves to the \( \pi \) estimator. Since we have auxiliary information at hand we can use the GREG estimator. Let us study \( \pi \) sampling in the light of a special case of model (1) and suppose that the following first and second moments of the superpopulation make sense (where \( \sigma^2 \) is a (possibly unknown) constant, \( \gamma \) is a constant reflecting the degree of heteroscedasticity and \( u_k > 0 \)):

\[
\begin{align*}
E_{\pi_k} (y_{q_k}) &= \beta u_k \\
V_{\pi_k} (y_{q_k}) &= \sigma^2 u_k^\gamma
\end{align*}
\] (6)

With this description of a population structure, the observations in a finite population scatterplot of \( y_{q_k} \) and \( u_k \) would gather around the line \( \beta u_k \), with a larger spread around the line for larger values of \( u_k \) if \( \gamma > 0 \). When \( \gamma = 1 \) and \( c_{q_k} = \sigma_{q_k}^2 = \sigma^2 u_{q_k} \), the GREG estimator following from this
model (sometimes called the common ratio model) is the well-known ratio estimator

\[ \hat{t}_{y_{qy}} = \sum_{u} u \sum_{y_{ik}} y_{ik} / \pi_{ik}. \]

If the model structure given by (6) reasonably well portrays the finite population scatter, it indicates a nearly proportional relationship between \( y_{ik} \) and \( u_k \) in the finite population. The \( u_k \) \((k = 1, \ldots, N)\) are known and a \( \pi ps(u) \) design yields

\[ \hat{t}_{y_{qy}} = \sum_{s} y_{qk} / \pi_k. \]

(7)

Hence, the ratio estimator in this case coincides with the \( \pi \) estimator.

Heteroscedastic patterns are common in actual populations. Brewer (1963, 2002) discusses this in detail and suggests that \( \gamma \) often is in the interval \([1, 2]\).

Remark: Textbooks often refer to a \( \pi ps(u) \) design combined with the \( \pi \) estimator as a very efficient strategy, particularly when the variance structure is strongly heteroscedastic, say \( V_{y_{qk}}(y_{qk}) = \sigma^2 u_k^2 \) i.e. \( \gamma = 2 \). However, from the studies in this thesis, we note that it usually is possible to find other strategies that work better.

1.4.4 The Anticipated Variance

When we plan a survey, we need some measure that can help us to discriminate and choose between alternative strategies. Obviously, since only auxiliary information is available, this measure must be a function of auxiliary information. One such measure is the anticipated variance, defined by Isaki
If $\hat{t}_{y_{x'}}$ is an estimator of $t_{y_{x}}$ and if (1) is interpreted as moments of a superpopulation from which a finite population has been selected, the anticipated variance is the variance of $\hat{t}_{y_{x'}} - t_{y_{x}}$ over both the model $\xi_q$ and the sampling design, i.e.

$$E_{\xi_q}E_p\left[ (\hat{t}_{y_{x'}} - t_{y_{x}})^2 \right] - \left[ E_{\xi_q}E_p\left( \hat{t}_{y_{x'}} - t_{y_{x}} \right) \right]^2.$$  

If the model $\xi$ is well specified, then an approximation to the anticipated variance can be written as (see Särndal et al. pp 450–451),

$$ANV_q(\hat{t}_{y_{x'}}) = \sum_{\ell} \left( \pi_k^{-1} - 1 \right) \sigma^2_{qk}.$$  \hspace{1cm} (8)

By using equation (8), we find a justification for choosing a $\pi$p$s$ design with a certain size measure. For a sampling design such that $E_p(n_{x'}) = n$, Result 12.2.1 in Särndal et al. shows that equation (8) is minimized if the inclusion probabilities, $\pi_k$, are chosen proportional to $\sigma_{qk}$, i.e. a $\pi$p$s(\sigma_q)$ design. The minimum value $ANV_q(\hat{t}_{y_{x'}})$ is then

$$ANV_{q\min}(\hat{t}_{y_{x'}}) = \sum_{\ell} \left( \pi_k^{-1(\text{opt})} - 1 \right) \sigma^2_{qk} = \frac{1}{n} \left( \sum_{\ell} \sigma_{qk} \right)^2 - \sum_{\ell} \sigma^2_{qk}. \hspace{1cm} (9)$$

If we apply this result to the common ratio model in equation (6), the optimal inclusion probabilities that minimize $ANV_{q\min}(\hat{t}_{y_{x'}})$ are,
\[ \pi_k = \pi_{q^{(opt)}} = \frac{m_i^{\gamma/2}}{\sum_i u_i^{\gamma/2}} \quad (k = 1, \ldots, N) \]

hence, \( \pi_k \propto \pi_k = u_i^{\gamma/2} \).

For fixed size designs, other authors have presented results also indicating that an optimal design is attained when \( \pi_k \propto \sigma_{q_k} \) – Hajek (1959) (linear design-unbiased estimators), Brewer (1963) (ratio estimation) and Cassel, Särndal and Wretman (1976, 1977) (generalized difference estimators). When the \( \pi \) estimator coincides with the ratio estimator (see equation (7)), Godambe also gives a similar result. Rosén (2000) used the result by Cassel, Särndal and Wretman to study optimal strategies by combining GREG estimation and Pareto \( \pi \) sampling. Wright’s proposal of \( mb-STSI \) is inspired by the wish to find a simplified design with inclusion probabilities that come close to the optimal.

1.4.5 Planning values and guesstimates

We need a notation to distinguish quantities which are used in the planning process from values which hypothetically are considered to be true values. Let us call these planning values ‘guesstimates’, since they reflect quantifications of the surveyor’s belief or guess about the structure in the finite population. In this summary and in papers II–IV, guesstimates that are used for calculations will be marked by the tilde symbol. For example, a guesstimate of the variance structure will be denoted by \( \tilde{\sigma}_{q_k}^2 \) and a guesstimate of the \( \gamma \) parameter discussed earlier will be denoted by \( \tilde{\gamma} \). In the first paper a somewhat different notation is used.
1.5 The multivariate sampling problem

“It is often better to have reasonably good solutions of the proper problems than optimum solutions of the wrong problems.”

Herbert Robbins

Most sample surveys are multivariate. When a multivariate survey is planned, the design choice is – since it affects all estimates – relatively more important than the choice of estimators. Although a good estimator sometimes can compensate for a poorly chosen design and give an acceptable precision, it is also likely that precision is lost compared to what is achievable with a better design.

The theoretical results described in section 1.4 and in the survey literature are often not sufficient guides for finding a good sampling design for the multivariate situation. In most presentations, the concept of ‘optimum’ is tied to a single study variable. Examples are to be found in the theory above (choice of unequal inclusion probabilities), in the theory of optimum stratification and optimum allocation in stratified sampling and in the theory of multi-stage sampling, where the number of units in each stage can be determined optimally.

A simple reason why few papers deal with the multivariate survey problem might be that there is no evident criterion of optimality. For example, a design that is optimal or close to optimal in the single parameter case, may not be the best choice in an overall multiparameter sense. Obviously, when we choose our design we must find a compromise based on some kind of multivariate consideration.
It seems as if there are three often used ways to approach the multivariate problem in practice: (i) Ignore it and reduce it to a single variable problem through a snap judgment based on ‘experience’. (ii) For each of the important parameters, study the effect of various single-variable dependent actions, and then select a design that seems to be the best compromise. (iii) Find some kind of overall multivariate criterion, perhaps mechanically by generalizing single variable concepts, and choose the action that optimises this criterion.

The first choice may not be as bad as it may sound. In many situations, a few important study variables may have similar properties, and the total loss in precision following from using one of them might be acceptable. However, at the planning stage, time should be spent carefully studying the overall effects of choosing such a simplistic approach. Otherwise, it might turn out to be a bad practice.

The second approach is better since it means that a design choice is made on the basis of considerations of the multivariate situation. An example that fits into this approach is described by Kott and Bailey (2000) called maximal Brewer selection. It is not an optimisation-based method, but it is simple to implement and it guarantees the desired precision of all estimators considered in the planning. Various methods of averaging key measures may also be placed in this category of approaches. One is to select a stratified sample with an allocation based on the average of various single-variable optimum allocations, another, in the case of \( \pi ps(\sigma_q) \) sampling, is to choose the size measure as the average or median of \( Q \) different single variable \( \sigma_q \)'s.

For multivariate stratified sampling a number of authors have considered the third (optimisation-based) approach category, e.g. Dalenius (1957), Chat-

In this thesis, the multivariate problem is also treated using an optimisation-based approach. The results on univariate optimal designs from section 1.4 are extended to the multivariate case. The advantage of using this kind of approach is that, at least theoretically, optimal solutions are obtained. Moreover, as we choose the optimisation criterion, we also attain certain control of the desired design properties, as opposed to relying on intuitive ad hoc solutions. However, as for all other approaches the success will ultimately depend on the chosen criterion and the underlying assumptions made by the survey designer.

A possible drawback with an optimisation-based approach is cumbersome computations. However, in the present case the methods proposed for the multivariate situation are only slightly more complicated than single-variable methods. In addition, the methods suggested in papers III–IV are to a large degree based on computations that should be executed anyway in any serious planning effort.
2 Summary of the papers

In this section, the aims and the main conclusions of the four papers are summarized in chronological order.

2.1 Reflections on unequal probability sampling strategies*

In section 1.4.4, it was noted that an optimal design in the sense of minimizing $ANV_q(\hat{t}_{y_g})$ (an approximation to the anticipated variance of a GREG estimator) is such that the first-order inclusion probabilities are chosen proportional to $\sigma_{q_k}$, i.e. a $\pi ps(\sigma_{q_k})$ design. If a $\pi ps(\sigma_{q_k})$ design is a preferred design, the designs (sampling schemes) studied in paper I can be applied in an attempt to implement it. However, none of the fixed size designs will yield an exact $\pi ps(\sigma_{q_k})$ design, and therefore it is interesting to compare their properties. Another aim is to study how robust the sampling designs are against misjudgements of $\sigma_{q_k}$ made at the planning stage.

* Paper I is a reprint of Holmberg and Swensson (2001), which (when originally published) mistakenly was entitled ‘On Pareto $\pi ps$ sampling: Reflections on unequal probability sampling’. 
The paper compares the properties of \( \pi \)s sampling designs such as Poisson \( \pi \)s and Pareto \( \pi \)s, with Poisson mixture (\( \text{PoMix} \)) sampling and model-based simple random stratified sampling (\( \text{mb-STS} \)). Combined with a GREG estimator, these sampling designs are serious alternatives for survey designers who wish to use auxiliary information in the sampling design as well as at the estimation stage.

We compare the design properties by defining a penalty measure as the ratio between the \( \text{ANV}_q(\hat{\gamma}_{y'}) \) of a specific design and the minimum anticipated variance \( \text{ANV}_{q_{\min}}(\hat{\gamma}_{y'}) \), i.e.,

\[
\text{AVP}(\tilde{\gamma}, \omega) = \frac{\text{ANV}_q(\hat{\gamma}_{y'})_{\omega(q)}}{\text{ANV}_{q_{\min}}(\hat{\gamma}_{y'})}.
\] (9)

By using (9), we can study the increase in approximate anticipated variance obtained when a non-optimal \( \pi \)s(\( z \)) design is used. \( \tilde{\gamma} \) is a guessimate of the degree of heteroscedasticity in the supposed variance structures, and \( \omega \) is used to differentiate between the studied PoMix designs. \( \omega \) is a (transformation) parameter called the Bernoulli width. (If \( \omega = 0 \), PoMix is equal to the original design induced by the sampling scheme, e.g., Pareto or Poisson sampling.) For \( \text{mb-STS} \) a corresponding penalty measure without \( \omega \) is used.

In tables 1-5 of the paper, we see the effects of different assumed variance structures in terms of the penalty measure. We note the following: (i) Compared to the \( \pi \)s designs, \( \text{mb-STS} \) seems to be less sensitive to a large value of \( |\tilde{\gamma} - \gamma| \) (the absolute error of the heteroscedastic pattern guesstimate). (ii) The variance increase from using \( \text{PoMix} \) sampling is larger when \( \tilde{\gamma} < \gamma \).
than when $\gamma > \gamma$. In fact, it is only when $\gamma > \gamma$ that a value other than $\omega = 0$ is of assistance, i.e. PoMix is only of potential interest if the survey designer in the planning process happens to overestimate $\gamma$. (iii) However, when $\gamma > \gamma$ there exists an optimal $\omega$ which makes the variance penalty of a PoMix design very small. Moreover, when $\gamma > \gamma$, PoMix sampling with $\omega = 0.3 f$ (where $f = n / N$, i.e. the sampling fraction) is less sensitive to deviations of $\gamma$ from $\gamma$ than direct $\pi ps(z^$ sampling. Finally, (iv) since different study variables may have different degrees of heteroscedasticity, planning for good variance properties of one estimator may be bought at a high price for estimators of other study variable totals.

2.2 On the choice of strategy in unequal probability sampling

The second paper is a follow-up study of the first paper. Again, Pareto $\pi ps$, PoMix and mb-STS are compared. However, in this paper different strategy choices in a bivariate situation are looked at. There are two key parameters and strong auxiliary information at hand for the design as well as for estimation purposes.

Besides a discussion of Pareto $\pi ps$, PoMix and mb-STS, there is an examination of some other questions. When we have strong auxiliary information, will there be any significant variance decrease from using a $\pi ps(\sigma_v)$ design and a GREG estimator instead of (i) using auxiliary information for estimation purposes only, (ii) using a classical (non-optimal) $\pi ps$ design
combined with the $\pi$ estimator, as discussed in connection with equation (7) in section 1.4.3? Moreover, is it possible to more or less fully compensate a non-optimal design choice by a suitable estimator choice? As a final point, which of the studied strategies is the best in an overall perspective?

As in paper I, the comparisons indicate that we cannot rule out the use of $mb$-$STSI$. By means of results from a Monte Carlo study, we note that among the studied GREG estimators it does not seem to matter much which estimator we choose. Using auxiliary information in the estimator only is usually better than using auxiliary information in the design only. However, using auxiliary information in both is the best strategy, yielding the smallest variances. $PoMix$ works better than Pareto $\pi s$ when the heteroscedastic pattern is stronger. Nevertheless, for most strategies both these sampling designs yield somewhat higher variances than $mb$-$STSI$.

In paper II as well as in the first paper, we see that we obtain small variances by using the auxiliary information and choosing a $\pi s(\sigma_q)$ design. However, if we specify the variance structure badly and, as a result, choose a non-optimal design, the variance will increase and a good estimator choice is not enough to fully compensate. Another concern appears when we have several study variables with different variances structures. Then, a $\pi s(\sigma_q)$ design based on a single study variable, $y_q$, might work well for that variable, but for other variables the variances may become larger than acceptable. In the third and fourth papers there is a discussion of designs adapted to the multivariate situation.
2.3 A multiparameter perspective on the choice of sampling design in surveys

When we choose a strategy in the multivariate situation, we can always choose different estimators for different study variable totals, \( t_{y_q} \) \((q = 1, \ldots, Q)\). However, we can only choose one design and that choice will affect all the parameter estimates of the survey. One goal might be to try to find a design such that the estimator variances deviate as little as possible from the estimator variances resulting from single-variable optimal designs. A good design in a multivariate situation should not yield very good precision for some estimators at the expense of unacceptably high variances for others. In other words, we should search for a 'best' compromise.

Three approaches based on different criteria are presented in order to achieve a compromise design for a multivariate situation. In principle, all methods presented are generalizations of the single variable solution outlined in section 1.4.4. Hence, they are based on approximations of the anticipated variance of the GREG estimator and they are optimal under the given criterion.

For each study variable, auxiliary information is used to specify a linear model with a specific variance structure, as in equation (1). Thereafter, we derive the anticipated variances, \( ANV_q(\hat{t}_{y_q}) \) \((q = 1, \ldots, Q)\), and present three approaches based on functions of \( ANV_q(\hat{t}_{y_q}) \), to attain a compromise design.
The approach that seems most promising is a design that minimizes a loss function, the anticipated overall relative efficiency loss (ANOREL), defined as

$$\text{ANOREL} = \sum_{q=1}^{Q} \frac{\sum_{q=1}^{Q} H_q \left( \text{ANV}_q \left( \hat{\mu}_q \right) - \text{ANV}_{q_{\text{min}}} \left( \hat{\mu}_q \right) \right)}{\text{ANV}_{q_{\text{min}}} \left( \hat{\mu}_q \right)}$$

(10)

where $H_q$ ($q = 1, \ldots, Q$) are weights (summing up to unity) that reflect the relative importance of the parameters to be estimated.

The measure in equation (10) is similar to the measure discussed by Dalenius (1957, chapter 9) to find a compromise allocation for stratified sampling in the multiparameter case. Furthermore, we show that equation (10) is minimized if the design is such that the first-order inclusion probabilities are given by

$$\pi_k = \frac{n}{\sum_{q=1}^{Q} H_q} \left( \prod_{l} \left( \frac{\sigma_{q_k}^2}{\sigma_{q_{\text{opr}} k}^2} \right) \right).$$

(11)

This expression is not as cumbersome as it may appear. It involves the single-variable optimal inclusion probabilities $\pi_{q_{\text{opr}} k}$ and the variance structure $\sigma_{q_{k}}^2$. In a carefully planned multiparameter survey, guesses, i.e. $\tilde{\pi}_{q_{(\cdot) k}}$ and $\tilde{\sigma}_{q_{k}}^2$, of both these quantities should be computed in any case.

We study the suggested compromise approaches for a practical case. Subsequently, auxiliary information is used to make computations and illustrate suitable planning phase analysis. A diagnostic table is used for comparisons between the designs at our disposal and for a final choice of design.
2.4 On the choice of optimal sampling design in business surveys with several important study variables

A general aim of a multiparameter survey is to find a design that meets the desired levels of precision for all important target parameters simultaneously. If we choose our design according to one of the criteria presented in the third paper, we will achieve a minimization of a summary measure without taking into account that precision requirements may be different for different estimates. In paper IV, the approaches of paper III are generalized (and slightly extended) by applying non-linear programming algorithms.

Non-linear programming methods have previously been used in the case of similar problems related to optimal allocation in stratified sampling. Cochran (1977, section 5A4) provides a list of early references, Chromy (1987) and Bethel (1989) are references of later date.

For Poisson $\pi$ sampling and the $\pi$ estimator, Sigman and Monsour (1995) sketched a procedure using non-linear programming similar to that of paper IV. Saavedra (1999) applied these ideas using the algorithm proposed by Chromy to determine probabilities to be used for Pareto $\pi$ sampling in a price and volume petroleum product survey.

The general problem studied in paper IV can be described as follows. A sampling design is to be determined to minimize a function $f$ for all $Q$ estimator variances and fulfill specified restrictions, $v_q$, made on function $g$ of each estimator variance (or variance approximation), i.e. minimize

$$f\left(g\left(V\left(i_{y_1}\right)\right), g\left(V\left(i_{y_2}\right)\right), \ldots, g\left(V\left(i_{y_0}\right)\right)\right)$$

with restrictions
\[ g\left(V\left(\hat{t}_{y_q}\right)\right) \leq v_q \quad (q = 1, \ldots, Q). \] 

For the general problem above, the paper suggests solutions for the cases when \( f \) is either the weighted arithmetic or geometric mean and \( g \) is one of the functions given by the three approaches in paper III, i.e. (i) a variance, (ii) a relative variance, or (iii) the (recommended) relative efficiency loss. Moreover, since the methods are planning stage means, the unknown estimator variances in (12) are replaced by \( ANV_q\left(\hat{t}_{y_q}\right) \quad (q = 1, \ldots, Q) \).

The theoretical background of the applied non-linear programming model is given as well as an application using a Swedish business population. The studies indicate that the methods work both theoretically and practically, hence giving us the best compromise design in a multivariate situation, at least for not too large populations. The two foremost points of the paper are to suggest a flexible solution regarding how to use auxiliary information exhaustively in the design planning, and to provide diagnostic support for the final design choice.

### 2.5 Minor comments and concluding remarks

Finally, some minor but interesting observations made during the work on this thesis are offered.

The practical applicability of the studied methods has been an underlying guide in the work on these four papers. The results and methods of the last two papers are general in the sense that they allow a flexible use of the auxiliary information, estimators and sampling schemes. On the issue of finding a
‘best’ strategy for multiparameter surveys, the design should be in focus, primarily because the overall loss in precision due to a poorly chosen design seems to be relatively larger than the loss obtained by using a less efficient member (-s) of the GREG estimator family. We can implement the presented compromise designs by taking advantage of the recent developments in πps sampling, which provide us with practically feasible sampling schemes, for instance the Pareto πps sampling scheme. The first and the second paper give us experience of advantages and drawbacks of such schemes.

As far as the estimators are concerned, the results have been presented on the basis of the class of GREG estimators. Different GREG estimators are explicitly compared in paper II, but in the other papers the specific expression of the GREG estimator is of secondary importance, and in the third and fourth papers it has been tacitly understood that the models are well specified and that the survey designer can make a good estimator choice. In our setting, i.e. with well-chosen models and reasonably large, direct element samples, the designs presented will provide efficient estimates combined with a GREG estimator. It is not likely that there is much to gain in variance reduction by using estimators outside the GREG family, for instance calibration estimators (Deville and Särndal (1992), Lundström and Särndal (1999) and Estevao and Särndal (2000)) or the (less practical) optimal regression estimator (see Rao (1992, 1994, 1997) and Montanari (1987, 1998)).

Although the anticipated variance and the \( \text{ANV}_q(\hat{t}_{xy}) \) are conceptually different from the design variance of an estimator (e.g. \( V_{\hat{y}}(\hat{t}_{xy}) \) in equation (4) in section 1.4.2), it is of interest to note the similarities between table 2 and table 3 given in paper III (as well as similarities between the tables in
paper IV). They suggest that planning stage computations such as those illustrated could be a helpful tool to anticipate how different design choices affect estimator variances.
References


