UMEÅ UNIVERSITET

Contributions to the Theory of Unequal Probability Sampling

Anders Lundquist

Doctoral Dissertation
Department of Mathematics and Mathematical Statistics
Umeå University
SE-90187 Umeå
Sweden

ISBN:978-91-7264-760-2
Printed by: Print & Media
Umeå 2009

Contents

List of papers

The thesis is based on the following papers:

I. Bondesson, L., Traat, I. & Lundqvist, A. (2006). Pareto sampling versus Sampford and Conditional Poisson sampling. *Scand. J. Statist.* **33**, 699-720.

II. Lundqvist, A. & Bondesson, L. (2005). On sampling with desired inclusion probabilities of first and second order. Research Report in Mathematical Statistics No. 3, 2005, Department of Mathematics and Mathematical Statistics, Umeå University.

III. Lundqvist, A. (2007). On the distance between some πps sampling designs. *Acta Appl. Math.* **97**, 79-97.

IV. Lundqvist, A. (2009). Balanced unequal probability sampling with maximum entropy. Manuscript.

V. Lundqvist, A. (2009). A note on choosing sampling probabilities for conditional Poisson sampling. Manuscript.

Paper I is reprinted with the kind permission from the Board of the foundation of the Scandinavian Journal of Statistics. Paper III is reprinted with the kind permission from Springer Netherlands.

Abstract

This thesis consists of five papers related to the theory of unequal probability sampling from a finite population. Generally, it is assumed that we wish to make model-assisted inference, i.e. the inclusion probability for each unit in the population is prescribed before the sample is selected. The sample is then selected using some random mechanism, the sampling design.

Mostly, the thesis is focused on three particular unequal probability sampling designs, the conditional Poisson (CP-) design, the Sampford design, and the Pareto design. They have different advantages and drawbacks: The CP design is a maximum entropy design but it is difficult to determine sampling parameters which yield prescribed inclusion probabilities, the Sampford design yields prescribed inclusion probabilities but may be hard to sample from, and the Pareto design makes sample selection very easy but it is very difficult to determine sampling parameters which yield prescribed inclusion probabilities. These three designs are compared probabilistically, and found to be close to each other under certain conditions. In particular the Sampford and Pareto designs are probabilistically close to each other. Some effort is devoted to analytically adjusting the CP and Pareto designs so that they yield inclusion probabilities close to the prescribed ones. The result of the adjustments are in general very good. Some iterative procedures are suggested to improve the results even further.

Further, balanced unequal probability sampling is considered. In this kind of sampling, samples are given a positive probability of selection only if they satisfy some balancing conditions. The balancing conditions are given by information from auxiliary variables. Most of the attention is devoted to a slightly less general but practically important case. Also in this case the inclusion probabilities are prescribed in advance, making the choice of sampling parameters important. A complication which arises in the context of choosing sampling parameters is that certain probability distributions need to be calculated, and exact calculation turns out to be practically impossible, except for very small cases. It is proposed that Markov Chain Monte Carlo (MCMC) methods are used for obtaining approximations to the relevant probability distributions, and also for sample selection. In general, MCMC methods for sample selection does not occur very frequently in the sampling literature today, making it a fairly novel idea.

Keywords: balanced sampling, conditional Poisson sampling, inclusion probabilities, maximum entropy, Markov chain Monte Carlo, Pareto sampling, Sampford sampling, unequal probability sampling.

2000 Mathematics Subject Classification: 62D05, 62E15, 65C05.

Preface

Although there is only one name on the cover of a PhD thesis, this one included, a large number of people are included in the process of completing it. I would like to mention some of those people. Since they are selected using the truly random sampling procedure known as "Anders' memory" there is the risk of forgetting someone, but here it goes.

Firstly, I would like to thank my supervisor Lennart Bondesson, the man with a seemingly never-ending and steadily flowing stream of ideas which also, perhaps even more impressing, converge almost surely to a solution. It has been a privilege to share your vast knowledge about mathematical statistics, academic writing, and old famous statisticians.

Continuing on the subject of people with a large statistical knowledge, another person who springs to mind is my co-supervisor, Göran Arnoldsson from the Department of Statistics. I have seen that you can learn a lot by just keeping quiet and eating your lunch, about statistics as well as boats.

Thanks to Peter Anton and Lennart Nilsson at the Department of Mathematics and Mathematical Statistics for having the courage to employ an engineer who did not really know what he wanted to become when he grew up.

Special thanks to my sampling colleague Anton Grafström for reading, commenting and thus improving the introduction of the thesis. Others who have supported me in different ways are Berith Melander, Ingrid Svensson, and Ingrid Westerberg-Eriksson.

Thank you to all the great colleagues at the Department of Mathematics and Mathematical Statistics and Department of Statistics for making both places a pleasure to work at in general and to have coffee breaks at in particular.

Thanks to all of my friends, with a special mention for Erik Lampa and Jon Hallander for helping me to achieve the only perfect score during my student years. As the TV-commercial in those days put it: "13 rätt är respekt!"

And finally, saving the best for last: My family. Thanks to my parents and my sister, Bengt, Eva and Kristina, for always loving and having faith in me, even though you understood everything, nothing, and something about what I was doing. Thank you to my wife and best friend, Helena, and our lively and lovely daughters Emma and Sara, for love, support, and for generally making my non-work life such a great pleasure!

Umeå, April 2009
Anders Lundquist

1 Introduction

Suppose that we want to investigate one or more characteristics of a finite population. A population consists of a number of units, which may be for instance citizens of Sweden, manufacturing companies in Europe, the trees of a forest stand, and so on. Examples of interesting characteristics are, e.g., the proportion of people in favour of changing the currency from SEK to EUR, average expenditure on raw materials and total timber volume. When investigating such characteristics, we define variables of which we measure the numeric value. The variables we are interested in are called interesting variables or study variables.

In some situations it is possible to obtain information on the interesting variables from all units in the population, in which case we have performed a total count or census. In practice, it is usually not possible to perform a census due to time and/or economic reasons. In that case we perform a sample survey, i.e. we select a sample of units from a sampling frame, the list of units in the population which are available for selection. For the units selected in the sample, we measure the value of the interesting variables, and generalize the result to the whole population, that is, from the sample we estimate the characteristics of the population as a whole. If a sample survey is performed, there is some uncertainty in the result for the whole population, we have a sampling error.

If we perform random sampling, the size of the sampling error can be estimated from the sample. Random sampling also reduces the risk of systematic errors in the estimation procedure. When random sampling is performed, the random mechanism used for selecting the sample is called the sampling design.

Of course, there are many other errors, called nonsampling errors, which can occur even if a total count is carried out. Examples of such errors are nonresponse, measurement errors, frame errors, etc. When performing a survey all these possible errors should be taken into account.

This thesis deals exclusively with problems connected to random sampling and the sampling error. It contains five papers, mostly devoted to comparing and improving three different sampling designs, namely the conditional Poisson, the Sampford, and the Pareto designs. Some theoretical background is given in the following sections. In section 2, we give some general background on inference and introduce some notation. In section 3, we present some more theoretical considerations and results which are specific for this thesis. In section 4, the included papers are summarized. Finally, in section 5 we give some conclusions and mention some open problems.

2 Random sampling and inference

We will first introduce some basic notations and definitions, and then move on to inference aspects. That is, how to draw conclusions about a population based on the information in a sample.

2.1 Definitions and notation

A unique number, or some other kind of label, is assigned to each unit in the population, thus making identification possible. Let $U = \{1, 2, ..., N\}$ denote the population of size N. From the population we wish to select a random sample containing n units. The sampling can be performed with replacement (WR) or without replacement (WOR). If the sampling is WR, the same unit in the population may be selected several times in the sample. When sampling WOR an already selected unit cannot be selected again, thus guaranteeing that all selected units are distinct. In this case, the sample s is a subset of the population.

Now, we want to randomly select WOR a sample s of size n from a population U which has size N. Such a random selection is described by some probability scheme, called the sampling design. Mathematically, the sampling design is defined as a probability distribution on the set of all possible samples, and $p(s)$ is the probability of obtaining a specific sample s. A sampling design can be implemented in different ways.

If the sampling is performed WOR, which this thesis is focused on, we can describe the sample by

$$\mathbf{I} = (I_1, I_2, ..., I_N)\,,$$

where I_k is a so-called inclusion indicator for unit k. It is a Bernoulli random variable such that

$$I_k = \begin{cases} 1 & \text{if unit k is selected in the sample.} \\ 0 & \text{otherwise.} \end{cases}$$

These I_k satisfy $\sum_1^N I_k = n$. There are two important events to consider, namely the event that unit k is included in the sample, and the event the units j and k are both included in the sample. We use the following notation:

$$\pi_k = E(I_k) = Pr(I_k = 1), \qquad \pi_{jk} = E(I_j I_k) = Pr(I_j = 1, I_k = 1).$$

These are the first and second-order inclusion probabilities, respectively. Hence

$$Cov(I_j, I_k) = \pi_{jk} - \pi_j \pi_k, \qquad Var(I_k) = \pi_k(1 - \pi_k).$$

We can also see that the π_k:s satisfy $\sum_1^N \pi_k = n$, by noting that $\sum_1^N I_k = n$, and taking the expectation on both sides.

The first and second-order inclusion probabilities are the most important characteristics of a sampling design. A WOR sampling design where the inclusion probabilities are not all equal is called a πps sampling design, or simply a πps design (see, e.g., Särndal *et al.* 1992, pp. 90-97). In many cases sampling designs are chosen with the primary goal of obtaining specified first order inclusion probabilities in particular, but second-order inclusion probabilities may be considered as well. This is due to the fact that the estimators we want to use are functions of the inclusion indicators. The moments of the estimators depend on the inclusion probabilities up to the order of the moment of interest. Most often we only need the first and second moment of an estimator, since then we know its mean and variance. It follows that the first and second-order inclusion probabilities are the most important ones to know.

2.2 Inference in survey sampling

Assume for simplicity that we have only one study variable (i.e. we are only interested in one characteristic of the population), and denote it by y. Each unit in the population has a numerical value of y. These values are denoted by $\{y_1, ..., y_N\}$. The y_k:s can be seen as fixed but unknown values or as realizations of some random variables. It is often the case that we also have access to information on one or several auxiliary variables, usually denoted by x, but denoted by z in the introduction as well as in some of the papers of the current thesis. Auxiliary variables are variables which are not our primary interest, but it is reasonable to assume that they are connected to our study variable in some way. If there is such a connection, we can use it for improving the estimators. For instance, if the study variable is the mean amount spent by household on consumer electronics, the income of the household is a reasonable auxiliary variable, since there should be some connection between income and spending. When selecting auxiliary variables, there are two basic requirements: They should be related to y, and they should be easy to obtain information on. A common way of selecting and obtaining information on auxiliary variables is by using registers. Registers are useful because it is better if we know the values of the auxiliary variables for all units in the population, not just for the sampled ones.

The topic of inference from surveys started attracting attention in the 1930's, and is still an active research topic. Usually the aim of the inference is to estimate the population total or the population mean. The total is usually denoted by Y. The mean is commonly denoted by $\bar{Y}$, or sometimes by μ. The definitions are

$$Y = \sum_{k=1}^{N} y_k, \qquad \bar{Y} = \frac{Y}{N}.$$

From here on, all sums $\sum$ stated without restrictions are summations over the entire population, i.e. from 1 to N.

Over the years, there have been two major inference approaches which have been somewhat unified in more recent years.

First there is the *design-based* approach. Here, all randomness originates from the sampling design, while the values of the y-variable are considered to be fixed but unknown. The uncertainty associated with our observed estimates is only due to the fact that we do not study the entire population. Even though the values of the study variable are fixed we will observe different estimates for different samples. The most widely used estimator for design-based inference is the Horvitz-Thompson (HT-) estimator (or π-estimator) for the population total:

$$\hat{Y}_{HT} = \sum \frac{y_k}{\pi_k} I_k.$$

Sometimes, $1/\pi_k$ is denoted by d_k and called the design weight. This terminology originates from the fact that each sampled unit can be considered to represent d_k units in the population, the observations are "inflated" to match the magnitude of the population. Since $E(I_k) = \pi_k$, the HT-estimator is unbiased with respect to the sampling design. The variance of the HT-estimator is, given in Sen-Yates-Grundy form,

$$Var(\hat{Y}_{HT}) = -\frac{1}{2} \sum \sum (\pi_{jk} - \pi_j \pi_k) \left(\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k} \right)^2 .$$

The HT-estimator works best if the π_k:s are approximately proportional to the y_k:s. One advantage of the design-based approach to survey sampling inference is that we can derive estimators with desirable properties while making almost no assumptions at all about the population.

For those willing to make more assumptions about the population, there is the *model-based* approach. Here we consider the actual population values as realizations of the *random variables* $y_1, ..., y_N$. We then need a suitable model, often called a superpopulation model. The specification of the model is in principle up to the researcher and only limited to what assumptions he or she considers appropriate to make. In general, a model-based estimator of a population total can be written as

$$\hat{Y} = \sum_{k \in s} y_k + \sum_{k \notin s} \hat{y}_k,$$

where the values of the non-sampled units are predicted using the model. The estimation then relies on finding good predictions of the y-values for the unsampled units, given the model. The actual sampling design is not that important. If the model is correctly specified, the model-based approach may yield better results than the design-based approach. If the model is incorrectly specified, all conclusions are more or less invalid, depending on the degree of misspecification. In recent years there has been a lot interest focused on small-area estimation, where we do not have so much data and thus we need to develop appropriate models to be able to make any inference. Attention has also been directed towards using models for dealing with nonsampling errors such as nonresponse (see, e.g., Särndal & Lundström 2005).

In practice, "the middle way" appears to be most common in survey sampling inference. The design- and model-based approaches have been combined in the model-assisted inference approach (Särndal *et al.* 1992). Roughly speaking, we use the design-based estimators and try to improve them by introducing simple models. For example, if there is an auxiliary variable z, which we think is correlated to the study variable y, we may choose a sampling design which yields desired inclusion probabilities $\pi_k^d,\ k = 1, .., N$, where

$$\pi_k^d = n\frac{z_k}{\sum z_i}.$$

We then use the HT-estimator

$$\hat{Y}_{HT} = \sum \frac{y_k}{\pi_k^d} I_k.$$

The idea here is that $\pi_k = \pi_k^d \propto z_k \propto y_k$. The last proportionality does not usually hold exactly, but the variance of the HT-estimator usually decreases even if it only holds approximately. Does it hold exactly, the variance of the HT-estimator becomes zero. Also, it must be possible to select a sampling design which yields

$$E(I_k) = \pi_k^d,$$

since otherwise the HT-estimator will be biased.

Of course, other ideas about how to choose the π_k^d:s have been proposed. One example is connected to generalized regression estimation (GREG, see, e.g., Särndal *et al.* 1989, 1992 sections 6.4-6.7, and Holmberg 2003). Here we have a model for y,

$$y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k,$$

where we further assume

$$E(\varepsilon_k) = 0,\ Var(\varepsilon_k) = \sigma_k^2, \text{ and } Cov(\varepsilon_j, \varepsilon_k) = 0.$$

The π_k^d:s are then chosen such that $\pi_k^d \propto \sigma_k$, and regression estimation is applied.

In summary, model-assisted inference is probably the most widely-used approach today. However, it relies on the possibility of obtaining fairly large samples, so that large-sample theory and approximations may be utilized. In situations where only small samples are available, such as small-area estimation, the use of appropriate models is essential.

2.3 Further considerations regarding inference

Assume that we want to apply model-assisted inference, using the HT-estimator. The desired π_k^d:s are derived from using some kind of model. The first problem we encounter is to find a sampling design which yields inclusion probabilities π_k^d.

However, if the model is not correctly specified there will be some problems. One way of "protecting" ourselves against misspecification of the model is to use a sampling design which has a high entropy, where the entropy is given by

$$E = -\sum_s p(s) \cdot \log p(s).$$

Having high entropy corresponds to spreading the probability mass as much as possible over the allowed samples, while making sure that we obtain the correct inclusion probabilities. The allowed samples are usually all possible samples having fixed size, n. Sometimes, for instance in real time sampling (Meister 2004, Bondesson & Thorburn 2008) we may allow the sample size to be random, but that is not considered here.

In recent years, the methods of calibration and balancing have been suggested for improving either the HT-estimate (calibration) or the sample selection procedure (balancing). They both rely on the possibility of utilizing information provided by m auxiliary variables, $\mathbf{z} = \{z^{(1)}, ..., z^{(m)}\}$. We assume that the population total for each $z^{(j)}, j = 1, ..., m$ is known. The totals are denoted $Z^{(j)}, j = 1, ..., m$.

Calibration (see, e.g., Deville & Särndal 1992) is a method which adjusts the design weights $d_k = 1/\pi_k$ when a sample already has been selected. In calibration the design weights are changed from d_k to w_k. The w_k:s are chosen to be as close as possible to the d_k:s, in some metric, under the restrictions

$$\sum_{k \in s} w_k z_k^{(j)} = Z^{(j)}, \ j = 1, ..., m,$$

for all auxiliary variables $Z^{(j)}$. The idea is that if the study variable and the auxiliary variables are correlated, weights which yield perfect estimates of the known totals for the auxiliary variables will also bring the estimate of the population total for the study variable closer to the true value.

The idea behind balancing (Tillé 2006) is similar to the one behind calibration. Instead of changing the design weights after a sample has been selected, we change the probability of selection, $p(s)$, for the samples that are available for selection. It is possible that some selection probabilities are set to zero. The balancing conditions are based on the auxiliary variables, $z^{(1)}, ..., z^{(m)}$, and they are

$$\sum_{k \in s} d_k z_k^{(j)} = Z^{(j)}, \ j = 1, ..., m,$$

for all samples s such that $p(s) > 0$, where $Z^{(j)}$ is a known total. Only samples satisfying the balancing conditions are given a positive probability of selection.

3 Some πps designs and related results

This thesis is devoted to studying various properties of πps sampling designs. To illustrate, in figure 1 we have a small population consisting of five units, i.e. $N = 5$. The units have different sizes (circle areas), and we wish to select $n = 2$ units with probability proportional to size. For instance in forestry, trees are sampled in this way, where the diameter at breast height is used to measure the basal area which is the size measure used for the sampling. The desired inclusion probabilities $\pi_k^d,\ k = 1, ..., 5$, are given within each circle.

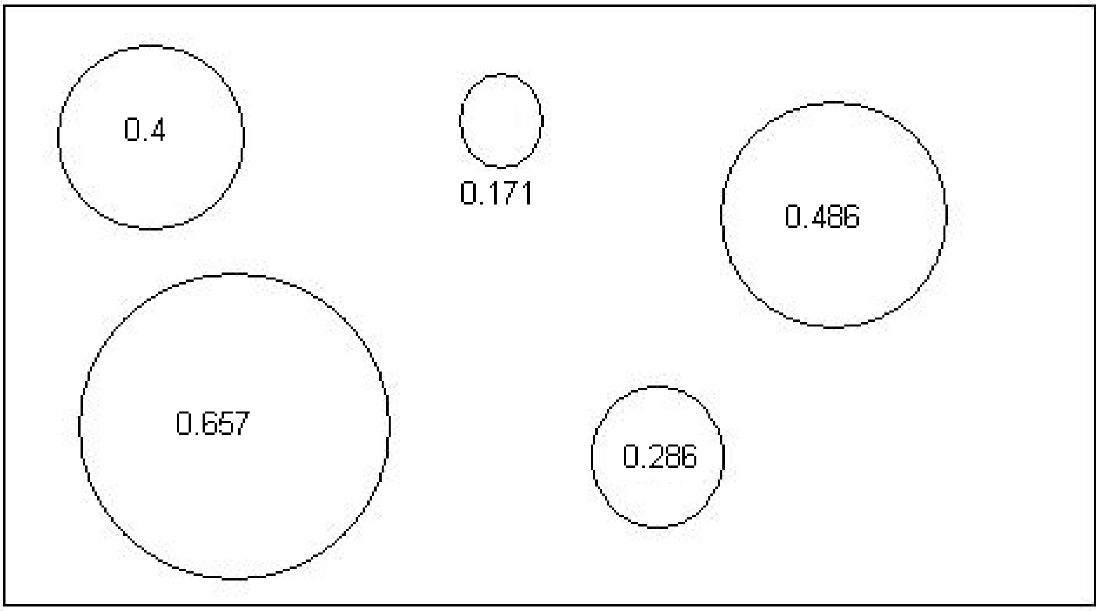

Figure 1: A small population of trees with N=5 and n=2. Inclusion probabilities proportional to disc areas.

Many different πps sampling designs have been suggested over the years. The books by Brewer & Hanif (1983) and Tillé (2006) give quite comprehensive reviews of available methods. In this thesis, we will concentrate mainly on three different πps designs, namely the conditional Poisson (CP) design (Hájek 1964, 1981), the Sampford design (Sampford 1967) and the recent Pareto design (Rosén 1997a,b) .

3.1 The CP design

We perform CP sampling by sampling unit k in the population independently of all the other units with sampling probability p_k, where usually $\sum p_k = n$. In the end, we only accept samples of size n. This may take some time, but less time-consuming procedures, for instance list-sequential procedures, have been suggested by, e.g., Chen & Liu (1997) and Traat *et al.* (2004). In Grafström (2009b), another very efficient implementation is discussed. It is possible to show (Hájek 1981, pp. 28-31) that it is a maximum entropy design, i.e. that CP sampling yields maximum entropy among all designs having fixed sample size n and factual inclusion probabilities π_k. One drawback is that $\pi_k \neq p_k$, and thus we need to determine somehow which sampling probabilities to use in order to achieve our desired inclusion probabilities.

3.2 The Sampford design

Sampford sampling (Sampford 1967, Hajek 1981, pp. 85-87) is performed as follows. We have sampling probabilities p_k, such that $\sum p_k = n$. To start with, one unit is selected with replacement and selection probabilities equal to p_k/n, $k = 1, \ldots, N$. Then $n - 1$ further units are selected with replacement according to probabilities $p'_k \propto p_k/(1-p_k)$, such that $\sum p'_k = 1$. We accept the sample if all n units are distinct, otherwise we repeat until acceptance. This procedure is potentially time consuming. A rejection-free method of selecting a Sampford sample has been introduced by Grafström (2009a). This is not a maximum entropy design, although quite close to being one. Further, it has the distinct advantage that $\pi_k = p_k$, which makes it easy to determine what sampling probabilities to use, we just set $p_k = \pi_k^d$.

3.3 The Pareto design

Pareto sampling (Rosén 1997a,b) is an order sampling design with sampling parameters p_k, $k = 1, ..., N$. To be comparable with the CP and Sampford designs, the sampling parameters should be specified so that $\sum p_k = n$. In general, an order sampling design bases the sample selection on order statistics. To each unit in the population is assigned a value of a ranking variable,

$$Q_k = \frac{F^{-1}(U_k)}{F^{-1}(p_k)}, \ k = 1, ..., N,$$

where F is a probability distribution function on $(0, \infty)$ and the U_k:s are i.i.d. $U(0, 1)$. The n units with smallest values of Q_k constitute the sample. Ohlsson (1998) used the uniform distribution function, $F(x) = x, \ 0 < x < 1$, but this is not the best choice. For Pareto sampling, F is chosen as the Pareto distribution function, $F(x) = x/(1 + x), x \geq 0$. The ranking variables Q_k are then

$$Q_k = \frac{U_k/(1 - U_k)}{p_k/(1 - p_k)}.$$

In some texts, including Rosén's original ones, the parameters are denoted by λ_k. This design is very simple to sample from, since there are no rejections. We always obtain a sample directly, and $\pi_k \approx p_k$. The approximation is quite good for large sample sizes but not so good for smaller sample sizes. Further, it is not a maximum entropy design, but rather close to being so. We need some method to determine which sampling parameters to use in order to achieve exactly our desired π_k^d:s.

3.4 Choosing $\mathbf{p_k}$:s for the CP and Pareto designs

The topic of choosing sampling probabilities p_k for the CP and Pareto designs is mainly treated in papers I, II, III, and V. We assume that we have desired inclusion probabilities, π_k^d, $k = 1, ..., N$. A first naive choice of p_k:s is to let

$$p_k = \pi_k^d,$$

which unfortunately yields, for both designs, that

$$\pi_k \neq \pi_k^d$$

and we introduce some bias in the HT-estimator. There are now two options: we can use an analytical approach or a computer-intensive iterative approach, or a combination. The main contributions in this thesis are in the analytical field. We have the π_k^d:s to work with, and it is reasonable to assume that there exist functions, f and g for CP and Pareto respectively, such that if we choose

$$p_k = f(\pi_k^d, d) \text{ and } p_k = g(\pi_k^d, d),$$

where

$$d = \sum \pi_k^d (1 - \pi_k^d),$$

we will obtain almost correct inclusion probabilities, i.e. $\pi_k \approx \pi_k^d$, where the approximation is very close. We need some idea about how to choose f and g. Further, it is in fact easier to work with the "sampling odds"

$$r_k = p_k/(1 - p_k)$$

than with the p_k:s themselves. In the first three papers we consider an asymptotic case, namely the case where d is large. In paper I it is shown that in this asymptotic case, the CP, Pareto and Sampford designs are equal from a probabilistic point of view. We use this similarity in papers I, II, and III. In paper III, we also derive measures of the probabilistic distances between these designs, based on asymptotic considerations, and make some illustrations for a few cases. It is shown that the Pareto design with adjusted p_k:s is very close to the Sampford design. In paper V, we consider the CP design for the asymptotic case where d is close to zero. This is a rather extreme case corresponding to, for a population ordered by decreasing desired inclusion probabilities,

$$\pi_k^d \approx \begin{cases} 1 & k = 1, ..., n \\ 0 & k = n+1, ..., N. \end{cases}$$

3.4.1 Analytic results - the CP design

For the asymptotic case where d is large, we exploit the similarity with the Sampford design and make some approximations. We finally obtain

$$r_k = \alpha \cdot \frac{\pi_k^d}{1 - \pi_k^d} \exp\left(\frac{\frac{1}{2} - \pi_k^d}{d}\right),$$

where α is chosen so that $\sum p_k = n$ is satisfied. In papers I and II, the exponential factor is presented as $\exp((1 - \pi_k^d)/d)$, which just means another α than here. Since we make some approximations, we do not obtain $\pi_k = \pi_k^d$ exactly, but come very close, and much closer than with the naive choice of sampling probabilities.

When considering the other asymptotic case, d close to 0, we obtain,

$$r_k = \begin{cases} \beta \frac{\pi_k^d}{1-\pi_k^d}\sqrt{\frac{d}{2}} & \text{if } \pi_k^d \text{ is close to 1} \\ \\ \beta \frac{\pi_k^d}{1-\pi_k^d}\sqrt{\frac{2}{d}} & \text{if } \pi_k^d \text{ is close to 0.} \end{cases}$$

The value of β is chosen such that $\sum p_k = n$.

In paper V we unify the two asymptotic cases by suggesting that, in general for the CP design, the sampling probabilities may be chosen according to (we omit the proportionality constant here)

$$r_k \propto \frac{\pi_k^d}{1-\pi_k^d} \cdot h(\pi_k^d, d),$$

where $h(\pi_k^d, d)$ is chosen as follows. Let $x_k = \left(\frac{1}{2} - \pi_k^d\right)/d$, to shorten expressions. Then we let

$$h(\pi_k^d, d) = h(x_k) = \exp\left(a \cdot \operatorname{arcsinh}(x_k/a)\right),$$

where $\operatorname{arcsinh}(x) = x + \sqrt{1+x^2}$, and a must be chosen somehow. Based on theoretical considerations and simulation results, we recommend in paper V that a is chosen as $a = 1/2 + d^3/2$. This choice of a implies that, as $d \to \infty$, $h(x) \to e^x$. This is consistent with the previous results regarding the case where d is large.

3.4.2 Analytic results - the Pareto design

When considering the asymptotic case of large d, we once again exploit the similarity with the Sampford design and make some approximations. We obtain for the p_k:s,

$$r_k = \gamma \cdot \frac{\pi_k^d}{1-\pi_k^d} \cdot \exp\left(\frac{\pi_k^d(1-\pi_k^d)(\frac{1}{2}-\pi_k^d)}{d^2}\right),$$

where γ is chosen so that $\sum p_k = n$ is satisfied. Since we make approximations, we do not obtain $\pi_k = \pi_k^d$ exactly, but we come extremely close, even closer than in the corresponding CP case. Of course, we also come much closer than with the naive choice of sampling probabilities. It may be noted that this is very easily implemented in practice. The Pareto ranking variables for the naive choice of sampling probabilities are

$$Q_k = \frac{U_k/(1-U_k)}{\pi_k^d/(1-\pi_k^d)},$$

and the new ranking variables are then given by

$$\widetilde{Q}_k = Q_k \cdot \exp\left(\frac{\pi_k^d(1-\pi_k^d)(\pi_k^d - \frac{1}{2})}{d^2}\right).$$

We just need to add one simple factor in the calculations.

For the asymptotic case where d is close to zero, we have no analytical results for the Pareto design. It is suggested in paper V that the earlier function h may be used *ad hoc* for the Pareto design, replacing the x_k:s with

$$y_k = x_k \cdot \left(\frac{\pi_k^d (1 - \pi_k^d)}{d} \right),$$

with x_k as before. Some initial numerical experiments suggest that the π_k:s are closer to the π_k^d:s compared to the naive choice of sampling probabilities. The improvement is most obvious when d is very small, about 0.4 or smaller. For d-values larger than 1 the refined adjustment is of no importance.

3.4.3 A little about iterative procedures

As mentioned previously, another way of finding appropriate p_k:s is to use some iterative procedure. Such procedures rely heavily on the possibility of rapid calculation of the π_k:s for the design in question. For this purpose, recursion formulas have been derived by, e.g., Aires (1999), Chen *et al.* (1994) and Tillé (2006). We also have presented some recursion formulas for the CP and Pareto designs, in papers I, II, IV and V. For instance, the inclusion probabilities for the CP design of size n can be calculated using the inclusion probabilities for the CP design of size $n - 1$ by using

$$\pi_k^{(n)} = n \frac{r_k (1 - \pi_k^{(n-1)})}{\sum_i r_i (1 - \pi_i^{(n-1)})}.$$

We will present a brief overview of two iterative procedures.

The first procedure. This procedure has been suggested and used by Aires (1999) and Tillé (2006, pp. 81-84). We select some starting values p_k^0 for the sampling probabilities, usually $p_k^0 = \pi_k^d$, and use the procedure

$$p_k^{(t+1)} = p_k^{(t)} + c \left(\pi_k^d - \pi_k^{(t)} \right), \; t = 0, 1, 2, ...$$

where the $\pi_k^{(t)}$:s are the factual inclusion probabilities in iteration step t. A common choice is $c = 1$. Provided that the procedure converges, the desired p_k:s are obtained. It may be noticed that, for CP sampling,

$$\frac{\partial \pi_k}{\partial p_j} = \frac{Cov(I_k, I_j)}{p_j (1 - p_j)}.$$

This implies that, if p_k can be chosen close to π_k for all $k = 1, ..., N$, the Jacobian matrix of the transformation from $\mathbf{p} = \{p_1, ..., p_N\}$ to $\boldsymbol{\pi} = \{\pi_1, ..., \pi_N\}$ has diagonal elements approximately equal to one. Also, if all the covariances $Cov(I_i, I_j), i \neq j$ are small, the Jacobian matrix is close to the identity transformation, which motivates the use of this procedure with $c = 1$.

The second procedure. Another possible choice is iterative proportional fitting. Here we let

$$\frac{p_k^{(t+1)}}{1-p_k^{(t+1)}} = \frac{\pi_k^d/(1-\pi_k^d)}{\pi_k^{(t)}/(1-\pi_k^{(t)})} \cdot \frac{p_k^{(t)}}{1-p_k^{(t)}}, \; t = 0, 1, 2, \dots ,$$

which illustrates how the name was chosen. For this procedure Chen *et al.* (1994) prove convergence for fixed size CP sampling. However, the procedure may be applied for Pareto sampling as well.

3.5 Sampling with desired second-order inclusion probabilities

It may be of interest to have a sampling design which yields desired second-order inclusion probabilities, π_{jk}^d, rather than first-order ones (Sinha 1973). The second-order inclusion probabilities appear in the variance of the HT-estimator. If we have a superpopulation model, the expected variance of the HT-estimator can be minimized by choosing the π_{jk}^d:s properly. It may be noticed that for a design of fixed size n the second-order inclusion probabilities π_{jk}^d, $j \neq k$, determine the first-order ones since $(n-1)\pi_k^d = \sum_{\{j,\ j\neq k\}} \pi_{jk}^d$.

In paper II, we consider some possible designs for sampling with desired second-order inclusion probabilities. The first proposal is a design belonging to the exponential family with sampling parameters p_{jk}. It is not obvious how to choose the sampling parameters in order to obtain desired π_{jk}^d:s. It is suggested that iterative procedures such as iterative proportional fitting (cf. section 3.4.3) are used. Another suggestion is to use an ordinary CP design with sampling parameters $p_k = \pi_k^d$, and modify its probability function $p(\mathbf{x}) = Pr(\mathbf{I} = \mathbf{x})$ by multiplication by a quadratic form, $\mathbf{x^T A x}$, where $\mathbf{A}$ is a symmetric $N \times N$ matrix with zeros on the diagonal and the other entries given by a system of $N(N-1)/2$ equations involving inclusion probabilities of order up to four for the CP design.

Both the suggested methods in paper II require a lot of calculations. There are no rigorous existence proofs regarding the sampling parameters or the matrix $\mathbf{A}$. In the examples considered in paper II, both sampling parameters and matrix could be calculated and verified to yield correct first and second-order inclusion probabilities - thus they existed. Thus it seems as though these quantities exist if we do not choose the inclusion probabilities in too an extreme way.

3.6 On measuring probabilistic distances

The issue of measuring the distance between two probability distributions is an old one. Many different types of measures have been suggested over the years. An overview is given by Gibbs & Su (2002). Here we will give a short introduction of the measures used in this thesis.

Two of the most widely used distance measures are the Kullback-Leibler (KL) divergence (Kullback & Leibler 1951) and the Hellinger distance (see, e.g., Pollard 2002, p. 61). Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be discrete probability distributions. For simplicity, we will just write f_1 and f_2. The expressions for the respective distance measures are

$$D_{KL}(f_1, f_2) = \sum_{\mathbf{x}} f_1 \ \log\left(\frac{f_1}{f_2}\right) = E_{f_1} \log\left(\frac{f_1}{f_2}\right),$$

$$D_H(f_1, f_2) = \sqrt{2 \cdot \sum_{\mathbf{x}} \left(\sqrt{f_1} - \sqrt{f_2}\right)^2}\ .$$

When we take asymptotic considerations into account, we use a Taylor expansion of both the above probability metrics. The goal is to get an expression which is easier to deal with analytically. Using a Taylor expansion of order two, we obtain the same result for both the KL divergence and the squared Hellinger distance. We denote the common result by D^2_χ. The formula resembles the expression of a Chi-Square Goodness-of-fit test statistic, which motivates the notation. The χ^2 distance is given by

$$D_{\chi^2}(f_1, f_2) = \sqrt{\frac{1}{2} \sum_{\mathbf{x}} \frac{(f_2 - f_1)^2}{f_1}}.$$

It may be noted that all the distance measures above are in fact special cases of the more general Cressie-Read distance (Cressie & Read 1984). The constants in the Hellinger and χ^2 distances may vary between authors. In some cases they are also defined without taking the square root. We can also see that the Hellinger distance is symmetric and satisfies the triangle inequality, but the KL divergence and χ^2 distance are both non-symmetric. Among these three probability metrics, the Hellinger distance is the only true metric.

In paper III, the χ^2 distance measure is used for theoretical considerations and the Hellinger distance is utilized when performing principle coordinate analysis (PCO, see, e.g., Cox & Cox 1994) in a few examples. Performing PCO means that we try to draw a map of how, in this case, the sampling designs are positioned with respect to each other. This map is based on pairwise distances between all the designs. The map is usually, as in paper III, in two dimensions. This means that the pairwise distances may not be reproduced exactly. The results of both theory and PCO suggest that especially the Pareto design with adjusted ranking variables is probabilistically very close to the Sampford design. The CP design is a bit away from the other two, even if adjusted sampling parameters are used.

3.7 On balanced sampling with maximum entropy

We begin by briefly restating the idea of balanced sampling. The balancing is based on auxiliary variables $z^{(1)}, ..., z^{(m)}$ with known population totals $Z^{(1)}, .., Z^{(m)}$. The balancing conditions are

$$\sum_{k \in s} d_k z_k^{(j)} = Z^{(j)}, \ j = 1, ..., m,$$

for all samples s such that $p(s) > 0$. Only samples satisfying the balancing conditions are given a positive probability of being selected.

In paper IV, we consider balancing conditions with some restrictions. We consider a cross-stratified population, i.e. a population stratified in $t \geq 2$ different ways. We have fixed sample sizes for each stratum but not for each cross-stratum. This means that the balancing conditions are of integer type, since the auxiliary variables are stratum indicators

$$z_k^{(j)} = \begin{cases} 1 & \text{if unit } k \text{ belongs to stratum } j \\ 0 & \text{otherwise,} \end{cases}$$

and the known totals are the stratum sizes. To illustrate, consider an example.

Example 1 (Stratification by age and gender). Suppose that we have a population of size N, stratified with respect to age and gender. We wish to sample n_M males, n_F females, n_Y people below 50 years of age, and n_O people of age 50 or above. The total sample size should be n, where $n_M + n_F = n_Y + n_O = n$. The resulting cross-stratification table is seen in table 1 below.

Table 1: Cross-stratification by age and gender

	Age < 50	**Age ≥ 50**	*Sample size*
Male	*Males* < 50	*Males* ≥ 50	n_M
Female	*Females* < 50	*Females* ≥ 50	n_F
Sample size	n_Y	n_O	n

If we denote the set of all possible samples satisfying the balancing conditions by Ω, it can be shown that in order to obtain maximum entropy, we should perform conditional Poisson (CP-) sampling within the set Ω. We give a proof in paper IV, and note that the maximum entropy property of CP sampling has been proven before in another way by, e.g., Hájek (1981, pp. 28-31).

As a further requirement on the balanced design, we still want to obtain desired inclusion probabilities, π_k^d. In order to achieve that, we somehow need to determine what sampling probabilities to use when we carry out the CP sampling. The situation becomes complicated since the sample size for at least one cross-stratum is random, i.e. the sample size follows some (discrete) probability distribution. This means that we cannot use the results from papers I, II and V, since then we require that all sample sizes are fixed. We propose the use of iterative methods instead. If we want to apply iterative methods, there are two problems to solve.

1. We need to calculate the probability distribution for the sample sizes.
2. For all possible sample sizes in each cross-stratum, we must calculate the actual inclusion probabilities given the sample size in the cross-stratum and the sampling probabilities.

The first problem turns out to be computationally challenging. The required probability functions may be written down explicitly, but in practice the number of combinations for which calculations need to be done grows rapidly. For very small examples, however, we can calculate the probability function exactly by using the expression for it. For more realistic situations, we propose the use of Markov Chain Monte Carlo (MCMC) techniques, or Gaussian approximations, as a way of obtaining approximations to the distribution in question. The MCMC procedures are very general and may be used in almost any possible setting. Further, we advocate the possibility of using MCMC methods for sample selection as well. Currently MCMC methods are not widely used in sampling theory. Now, once the first problem stated above has been solved, we can solve the second one by combining the solution to the first problem with our previously known recursion formulas for CP sampling. We can thus use some iterative procedure to determine the sampling probabilities.

There is then the question if there always exists a sampling design that yields desired inclusion probabilities π_k^d satisfying $\sum_{k\in S}\pi_k^d = n_S$, where S denotes a stratum and n_S the stratum sample size, for each stratum. We give, in paper IV, such a proof for the case of no more than three balancing conditions. For the case of four balancing conditions, we present a counterexample to show that in general there is no guarantee that an appropriate sampling design exists.

4 Summary of the papers

Note that in the summaries of the papers, the notation has been adapted to be coherent with the rest of the introduction. The notation in the actual papers may differ somewhat from notation used here in the summaries.

4.1 Paper I. Pareto sampling versus Sampford and conditional Poisson sampling

In this paper we first compare the Pareto and Sampford sampling designs. We use their respective probability functions and Laplace approximations, and show that from a probabilistic viewpoint these two designs are very close to each other. In fact, they can be shown to be asymptotically identical. The rejective method of Sampford sampling may be time consuming. We show that a Sampford sample may be generated by passing a Pareto sample through an acceptance−rejection filter. It is still a rejective method, but the number of rejections are reduced substantially. Most often, there are no rejections at all. This technique can be modified to give us an efficient method to generate conditional Poisson (CP) samples as well. We also show how to calculate inclusion probabilities of any order for the Pareto design given the inclusion probabilities for the CP design. Finally, we derive a new explicit approximation of the second-order inclusion probabilities. The new approximation is valid for several designs, and we show how to apply it to get single sum type variance estimates of the Horvitz-Thompson estimator.

4.2 Paper II. On sampling with desired inclusion probabilities of first and second order

We present a new simple approximation for obtaining sampling probabilities p_k for conditional Poisson sampling to yield given inclusion probabilities π_k^d. This approximation is based on the fact that the Sampford design yields the desired inclusion probabilities. We present a few alternative routines to calculate exact p_k-values, and carry out some numerical comparisons. Further we derive two methods for achieving desired second-order inclusion probabilities π_{jk}^d. This might be interesting if we want to control (and minimize) the variance of the Horvitz-Thompson estimator. The first method is to use an exponential family probability function. We determine the parameters of this probability function by using an iterative proportional fitting algorithm. The second method we use is to modify the conditional Poisson probability function with a quadratic factor. We perform a small numerical study on these two methods as well.

4.3 Paper III. On the distance between some πps sampling designs

Here, we derive an approximation for obtaining sampling parameters p_k for Pareto sampling in order to obtain given inclusion probabilities π_k^d. We continue by investigating the distances between some probability distributions arising from different πps sampling designs. The designs in question are Poisson, Conditional Poisson (CP), Sampford, Pareto, Adjusted CP (cf. Paper II), and Adjusted Pareto sampling, as derived in this paper. We begin by using the Kullback-Leibler divergence and the Hellinger distance. Then we use a Taylor expansion of order two on both distance measures. The common result is a simpler distance measure to work with theoretically. This measure of χ^2-type is evaluated first theoretically and then numerically in examples with small populations. We further illustrate the numerical examples by a multidimensional scaling technique called principal coordinate analysis (PCO). From both the theoretical analysis and the illustrated examples we see that Adjusted CP, Sampford, and adjusted Pareto are rather close to each other. In particular, Sampford and adjusted Pareto are very close. Pareto is located slightly further away from these, while CP and especially Poisson are quite far from all the others.

4.4 Paper IV. Balanced unequal probability sampling with maximum entropy

We investigate how to perform balanced unequal probability sampling with maximum entropy. We focus on some particular balancing conditions, namely the conditions of having specified marginal sums of the sample sizes in a cross-stratification table, but unspecified sample sizes in each cross-stratum of the table. When only the marginal sums are fixed, the sample sizes for one or more cross-stratums in the table are random. In principle, it is possible to express the probability distribution for those sample sizes explicitly. However, the computations quickly become difficult, except for very small cases. It is crucial to determine the probability distribution somehow, otherwise we are not able to calculate the inclusion probabilities for the design. We propose the use of Markov Chain Monte Carlo (MCMC) methods for obtaining good approximations to the probability distributions in question. It is proposed that the MCMC methods are used for sample selection as well. As another alternative, we consider large-sample Gaussian approximations. As usual when conditional Poisson sampling is used, the inclusion probabilities do not equal the sampling ones. Regardless of which method one uses for distributional calculation, iterative procedures may be used for obtaining sampling probabilities yielding inclusion probabilities very close or equal to the specified ones. We suggest a few different such iterative methods.

4.5 Paper V. A note on choosing sampling probabilities for conditional Poisson sampling

In this paper, the starting point is conditional Poisson sampling, and the fact that the sampling probabilities and the factual inclusion probabilities are not identical, even if $\sum p_k = n$. This is a problem. We present a new method for choosing the sampling probabilities, which may be used under more general conditions than previously suggested methods (cf. paper II). The new method uses only the desired, predetermined, inclusion probabilities π_k^d and the number $d = \sum \pi_k^d(1 - \pi_k^d)$. We then compare the performance of this new method to other reasonable choices of sampling probabilities. Finally we note that this new method could also be used as an *ad hoc* method for determining the parameters for Pareto sampling.

5 Conclusion and some open problems

This thesis is very much about the CP, Sampford and Pareto sampling designs. They have been thoroughly compared both theoretically and with examples (Papers I, II and III), and we have derived adjustments for the CP and Pareto designs in order to obtain desired inclusion probabilities (Papers I,II and V for CP, paper III for Pareto). In particular, the adjustment for Pareto sampling is quite simple to use. The adjusted Pareto design has been found to be probabilistically very close to the Sampford design. We have considered the CP design in connection with balanced maximum entropy sampling (Paper IV). It might seem strange, but there is still much more to be done. A few things will be mentioned below.

The area of balanced πps sampling has been introduced in the literature fairly recently, and there are plenty of things to do. For instance, the results about maximum entropy balanced sampling given in paper IV should be possible to generalize and extend quite a bit, both regarding existence of and how to determine the sampling probabilities. The gaussian approximations could perhaps be replaced with some approximate discrete probability distribution, since we are trying to approximate a discrete distribution. The problem of finding a convenient and practical way of performing maximum entropy balanced sampling needs more attention. Possibly some generalization of Pareto sampling could be applied.

Regarding sampling with desired second-order inclusion probabilities a lot remains to be done. There are theoretical problems, such as determining if a set of given π_{jk}^d:s really are second-order inclusion probabilities corresponding to some sampling design. There is also the problem of finding a practical way of performing the actual sampling. It is possible to use MCMC methods for sampling.

The analytical expressions for the sampling probabilities/parameters p_k in CP and Pareto sampling could possibly be improved, which will of course also be of benefit when applying iterative procedures, since we will have access to excellent starting values for the iterations. For the case where d is small an explicit expression for the p_k:s, if such an expression even exists, has not yet been derived.

Finally, an issue which is not really a research problem. Considering the methods of adjusting CP and Pareto sampling, they are not well known to the general sampling public. In particular the modification of Pareto sampling for large d is very simple to use in practice, and for most populations found in practice, d is large enough to make sure that the modification yields inclusion probabilities closer to the desired ones. Pareto sampling is one of the methods suggested by Statistics Sweden (2008) for performing πps sampling.

References

Aires, N. (1999). Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. *Methodol. Comput. Appl. Probab.* **4**, 457-469.

Bondesson, L. & Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scand. J. Statist.* **35**, 466-483.

Brewer, K.R.W. & Hanif, M. (1983). *Sampling with unequal probabilities.* Lecture Notes in Statistics, No. 15. Springer-Verlag, New York.

Chen, S-X, Dempster, A.P. & Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81**, 457-469.

Chen, S-X. & Liu, J.S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875-892.

Cox, T.F. & Cox, M.A.A. (1994). *Multidimensional scaling.* Chapman & Hall, London.

Cressie, N.A.C. & Read, T.R.C. (1984). Multinomial Goodness-of-fit tests. *J. R. Statist. Soc. B.* **46**, 440-464.

Deville, J-C. & Särndal, C-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87**, 376-382.

Gibbs, A.L. & Su, F.E. (2002). On choosing and bounding probability metrics. *Internat. Statist. Rev.* **70**, 419-435.

Grafström, A. (2009a). Non-rejective implementations of the Sampford sampling design. *J. Statist. Plann. Inference* **139**, 2111-2114.

Grafström, A. (2009b). Repeated Poisson sampling. *Statist. Probab. Lett.* **79**, 760-764.

Hájek, J. (1964). Theory of rejective sampling. *Ann. Math. Statist.* **35**, 1491-1523.

Hájek, J. (1981). *Sampling from a finite population.* Marcel Dekker, New York.

Holmberg, A. (2003). *Essays on model assisted survey planning.* PhD Thesis. Department of Information Sciences, Uppsala University.

Kullback, S. & Leibler, R.A. (1951). On information and sufficency. *Ann. Math. Statist.* **22**, 79-86.

Meister, K. (2004). *On methods for real time sampling and distributions in sampling.* PhD Thesis. Department of Mathematical Statistics, Umeå University.

Ohlsson, E. (1998) Sequential Poisson sampling. *J. Off. Statist.* **14**, 149-162.

Pollard, David E. (2002). *A user's guide to measure theoretic probability.* Cambridge University Press, Cambridge, UK.

Rosén, B. (1997a). Asymptotic theory for order sampling. *J. Statist. Plann. Inference* **62**, 135-158.

Rosén, B. (1997b). On sampling with probability proportional to size. *J. Statist. Plann. Inference* **62**, 159-191.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499-513.

Sinha, B.K. (1973). On sampling schemes to realize pre-assigned sets of inclusion probabilities of the first two orders. *Calcutta Statist. Assoc. Bull.* **22**, 89-110.

Statistics Sweden (2008) *Urval - från teori till praktik (Sampling, from theory to practice).* Handbook 2008:1. Statistics Sweden, Stockholm.

Särndal, C-E. & Lundström, S. (2005). *Estimation in surveys with nonresponse.* Wiley, Chichester.

Särndal, C-E., Swensson, B. & Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika.* **76**, 527-537.

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

Tillé, Y. (2006). *Sampling algorithms.* Springer science + Business media, Inc., New York.

Traat, I., Bondesson, L. & Meister, K. (2004). Sampling design and sample selection through distribution theory. *J. Statist. Plann. Inference* **123**, 395-413.