UMEÅ UNIVERSITET

On Unequal Probability Sampling Designs

Anton Grafström

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

ISBN: 978-91-7264-999-6
Printed by Print & Media
Umeå 2010

To Katrin

Contents

List of papers

The thesis is based on the following six papers:

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

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

III. Grafström, A. (2010). On a generalization of Poisson sampling. *J. Statist. Plann. Inference* **140**, 982-991.

IV. Grafström, A. (2010). Entropy of unequal probability sampling designs. *Statist. Methodol.* **7**, 84-97.

V. Bondesson, L. & Grafström, A. (2010). An extension of Sampford's method for unequal probability sampling. To appear in *Scand. J. Statist.*

VI. Grafström, A. (2009). Efficient sampling when the inclusion probabilities do not sum to an integer. Manuscript.

Papers I, II, III and IV are reprinted with the kind permission from Elsevier. Paper V is reprinted with the kind permission from the Board of the Foundation of the Scandinavian Journal of Statistics and Blackwell Publishing.

Abstract

The main objective in sampling is to select a sample from a population in order to estimate some unknown population parameter, usually a total or a mean of some interesting variable. When the units in the population do not have the same probability of being included in a sample, it is called unequal probability sampling. The inclusion probabilities are usually chosen to be proportional to some auxiliary variable that is known for all units in the population. When unequal probability sampling is applicable, it generally gives much better estimates than sampling with equal probabilities. This thesis consists of six papers that treat unequal probability sampling from a finite population of units.

A random sample is selected according to some specified random mechanism called the sampling design. For unequal probability sampling there exist many different sampling designs. The choice of sampling design is important since it determines the properties of the estimator that is used. The main focus of this thesis is on evaluating and comparing different designs. Often it is preferable to select samples of a fixed size and hence the focus is on such designs.

It is also important that a design has a simple and efficient implementation in order to be used in practice by statisticians. Some effort has been made to improve the implementation of some designs. In Paper II, two new implementations are presented for the Sampford design.

In general a sampling design should also have a high level of randomization. A measure of the level of randomization is entropy. In Paper IV, eight designs are compared with respect to their entropy. A design called adjusted conditional Poisson has maximum entropy, but it is shown that several other designs are very close in terms of entropy.

A specific situation called real time sampling is treated in Paper III, where a new design called correlated Poisson sampling is evaluated. In real time sampling the units pass the sampler one by one. Since each unit only passes once, the sampler must directly decide for each unit whether or not it should be sampled. The correlated Poisson design is shown to have much better properties than traditional methods such as Poisson sampling and systematic sampling.

Key words: conditional Poisson sampling, correlated Poisson sampling, entropy, extended Sampford sampling, Horvitz-Thompson estimator, inclusion probabilities, list-sequential sampling, non-rejective implementation, Pareto sampling, Poisson sampling, probability functions, ratio estimator, real-time sampling, repeated Poisson sampling, Sampford sampling, sampling designs, splitting method, unequal probability sampling.

Preface

If someone had told me ten years ago that I would write a PhD thesis in Mathematical Statistics I would not have believed it. I had no plans to stay in school for such a long time. However, a lot has changed since then and I found that studying can be both fun and rewarding. Today when I look back, I am very glad for this wonderful opportunity to learn more about mathematics, statistics and sampling. Yet this work would never have been possible without the help and inspiration I got from a number of people.

First I would like to thank Lennart Bondesson, my supervisor, for all valuable help and guidance. Your great knowledge about statistics and your enthusiasm for helping and solving problems are admirable. What makes you an outstanding supervisor is that you also have had the courage to stand back sometimes and let me find my own way. Still you have managed to guide me in the right direction whenever needed.

Another person I would like to thank is Sara de Luna, my co-supervisor, for rewarding discussions and great general advise. Thanks also to Daniel Thorburn for reading and commenting on paper III.

I have found the Department of Mathematics and Mathematical Statistics to be a stimulating environment to work at. For that I thank all my great colleagues. Some of you deserve a special mention and one name that comes to my mind is Peter Anton who has inspired me as a teacher and who encouraged me to apply for a PhD-student position. Thanks also to Anders Lundqvist, my sampling colleague, for valuable discussions. Others that have helped me in various ways are Ingrid Westerberg-Eriksson and Berith Melander.

Special thanks to Lina Schelin and Niklas Lundström for being such great friends. It has been a pleasure to share my ups and downs with both of you during these years. Thanks also for reading and improving the introduction of the thesis.

To my mother and father, thanks for raising me to believe that everything is possible. To all my family and friends, thanks for always encouraging me in my studies. Finally, to my wonderful wife Katrin, thanks for all your love and support. You are the source of my inspiration and without you I would never had become the person I am today.

Umeå, April 2010
Anton Grafström

1 Introduction

In sampling we are interested in some characteristics of a finite population of units. A forester may be interested in the total volume of timber in a forest stand, in which case each tree is a unit in the population of trees. For an upcoming election we may be interested in the proportion of people in favour of some political party among the eligible voters. A company may be interested to find out how satisfied their customers are with the service or product that is provided. Sampling is used to gain such information without measuring all units in the population.

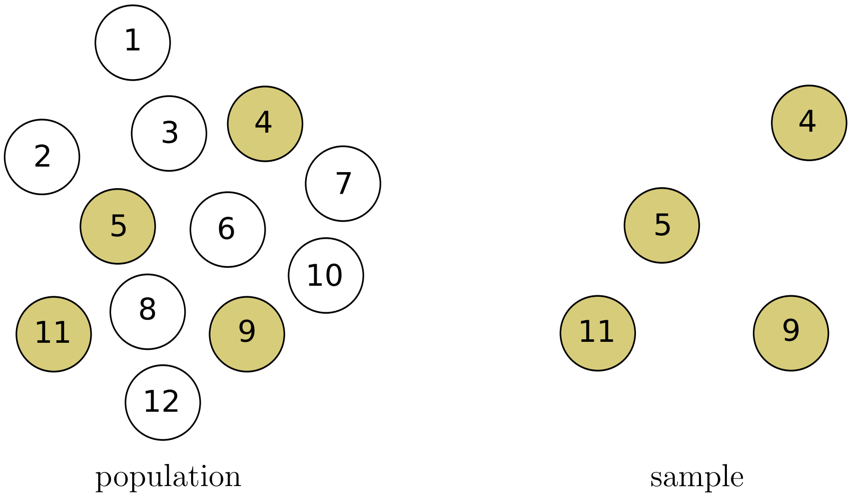

Figure 1: Illustration of population and sample.

By using sampling theory, it is possible to get a sufficiently good estimate of the parameter of interest at a reasonably low cost. Of course, the low cost is the main reason why sampling is so widely used. We are daily presented with the results from different statistical surveys. Most of these surveys, all the serious ones, are based on the theory of sampling. This advantage of sampling is also a problem since the number of surveys has increased to a level that has become a burden for the respondents. As a result there is a problem with non-response in many statistical surveys. Non-response occurs when some of the units in the selected sample cannot be measured or refuse to be measured. The problem of non-response is not treated in this thesis. For different methods to handle non-response, see eg. Särndal & Lundström (2005).

Before a sample can be selected, we usually need to list the units in the population. This list is called the *sampling frame*. It is important that the frame is correct and matches the population of interest. Otherwise there will be errors in the estimates due to the frame imperfections. It is assumed throughout this thesis that the frame is perfect. It is also important that the selected units are correctly measured, otherwise there will be measurement errors.

The only type of error that we focus on in this thesis is the sampling error. The sampling error comes from the fact that only a sample is observed and not the entire population. Of course, when performing a statistical survey it is important to consider all possible sources of error.

A simple way to take a sample of size n is to let all the possible samples have the same probability of being selected. This is called simple random sampling and then all units have the same probability of being chosen. Each unit can be represented by a numbered ball, as in figure 1. Then we put all the balls in an urn and draw n balls without replacement to select a sample.

When the units do not have the same probability of being selected we call it *unequal probability sampling*, which is a part of the title and the main topic of this thesis. When unequal probability sampling is applicable, it usually produces much better estimates than sampling with equal probabilities. When the inclusion probabilities, π_i, are prescribed for all units, unequal probability sampling is also called πps-sampling, where ps stands for proportional to size.

A common belief among non-samplers is that good samples should be miniature versions of the population, i.e. if the population consists of 50% males, then the sample should also do so. In general this is not true. If it was true, there would be no use for unequal probability sampling. Since the goal most often is to estimate a population parameter, a sampling procedure is good if it allows for efficient unbiased estimation of the parameter of interest.

A *sampling design* describes the probability mechanism used to select a sample. For unequal probability sampling there exist many different sampling designs that can be used. Unfortunately there exists no universally best design. In general it depends on the population and the sampling situation which design is the best one. However, in practice we never have complete information about the population since then there would be no need for sampling. Hence other more general criteria, such as the level of randomness, must be used when choosing a sampling design. Many different designs are presented and evaluated in this thesis.

During the last 15 years several new designs for πps-sampling have been presented. The splitting method introduced by Deville & Tillé (1998) is the most general one of them all. It can reproduce all other designs, though not always in a simple way. The fine idea behind the splitting method has led to several new designs.

This thesis contains six papers, much of the focus is on comparing different designs and also on improving the implementation of some designs. In section 2, some background and notation are given. The πps sampling situation is described in section 3 and some important designs and results are presented in section 4. A case called real-time sampling, which is treated in Paper III, is introduced in section 5. Section 6 gives a short introduction to the sampling situation in Papers V and VI. Sampling designs can have different degrees of randomization and a measure of randomness, called entropy, is introduced in section 7. The six papers are summarized in section 8. In section 9, conclusions and open problems are presented.

2 Definitions and notation

The finite population of N units is denoted by $\mathcal{U} = \{1, 2, ..., N\}$. We are interested in selecting a sample from $\mathcal{U}$ in order to estimate some parameter, often a total or a mean of some variable. In this thesis sampling without replacement (WOR) is treated, i.e. each unit can only be selected once. Thus a sample s is a subset of the population $\mathcal{U}$. It is also possible to sample units with replacement (WR) but such methods generally give less efficient estimation and are not treated here.

A random sample is selected according to a sampling design. Formally, a sampling design is a discrete probability distribution on a support Q of possible samples $s \subset \mathcal{U}$. The probability of getting the sample s is denoted by $p(s)$ and we have $p(s) > 0$ for all $s \in Q$. Since it is a probability distribution we also have $\sum_{s \in Q} p(s) = 1$. The following example illustrates two different sampling designs.

Example 1. If the population has four units $\mathcal{U} = \{1, 2, 3, 4\}$, there are six possible samples of size $n = 2$:

$$s_1 = \{1,2\},\ s_2 = \{1,3\},\ s_3 = \{1,4\},\ s_4 = \{2,3\},\ s_5 = \{2,4\},\ s_6 = \{3,4\}.$$

In figure 2, two different designs for selecting one of the 6 samples are illustrated. Design 1 corresponds to simple random sampling where each possible sample has the probability 1/6 of being selected. For this design we can select one of the samples by spinning wheel 1. Design 2 has different probabilities for the samples. Samples 1-3 are each selected with probability 1/9 and the samples 4-6 are each selected with probability 2/9. A sample from design 2 can be selected by spinning wheel number 2. In practice there are often too many possible samples to directly select a sample. Instead a sample is often selected by randomly choosing the units in a suitable way. □

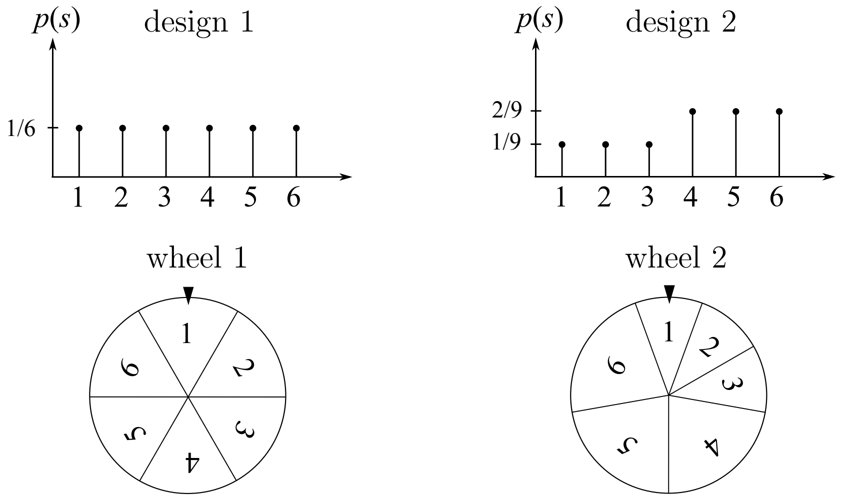

Figure 2: Illustration of two different sampling designs for selecting one of the 6 possible samples in example 1.

An important event in sampling is the inclusion of unit i in the sample. That event is usually indicated by the inclusion indicator I_i, defined as

$$I_i = \begin{cases} 1 & \text{if unit } i \text{ is included in the sample} \\ 0 & \text{otherwise.} \end{cases}$$

Thus I_i is a Bernoulli random variable. A random sample can be described by the vector of inclusion indicators $\mathbf{I} = (I_1, I_2, ..., I_N)$ and a sample which is the outcome of $\mathbf{I}$ is usually denoted by $\mathbf{x}$. Hence there are two different notations for a sample, we use s to denote a subset of $\mathcal{U}$ and each s corresponds uniquely to a binary vector $\mathbf{x} \in \{0,1\}^N$.

The inclusion probabilities of the units are important characteristics of a sampling design. The inclusion probability for unit i is defined as

$$\pi_i = \Pr(I_i = 1) = E(I_i) = \sum_{\mathbf{x} \in Q} x_i p(\mathbf{x}).$$

These π_i are called first-order inclusion probabilities. Generally the inclusion probability for unit i can be calculated by summing the probabilities of the samples that contain unit i.

Example 2. For design 1 in example 1, we see that each unit is included in three samples and every sample has probability $1/6$. Hence the inclusion probabilities are 1/2 for all four units. For design 2, the inclusion probability is 3/9 for unit 1 and 5/9 for the units 2, 3 and 4. □

The second-order inclusion probabilities of a sampling design are defined as

$$\pi_{ij} = \Pr(I_i = 1,\ I_j = 1) = E(I_i I_j) = \sum_{\mathbf{x} \in Q} x_i x_j p(\mathbf{x}).$$

Thus π_{ij} is the probability that both unit i and unit j are included in the sample. The inclusion probabilities of first and second-order are needed for estimation and variance estimation.

3 Basics about πps sampling

Usually the goal is to estimate the total of some variable y, which has value y_i for unit i. Thus we want to estimate $Y = \sum_{i=1}^{N} y_i$. All the y_is are unknown before a sample has been selected. In order to use unequal probability sampling we need some auxiliary information. It is often the case that we know the value of another variable $z_i > 0$ for each unit $i \in \mathcal{U}$ and we suspect that y is approximately proportional to z. The following example illustrates one possible situation.

Example 3. If the objective is to estimate the total amount of pollution from a number of factories, then we may know or strongly suspect that larger factories generate more pollution than smaller factories. If we have access to some auxiliary information z about the size of the factories, that information can be used. Such information may be the number of employees, the size of the buildings or the number of units produced last year and so on. In this situation we want to sample large factories with higher probabilities than small factories since large factories will contribute more to the total amount of pollution. By doing so we can get a much better estimate than if the factories are selected with equal probabilities. □

The information available to us before a sample is selected is the labels, $i = 1, 2, ..., N$, of the units and the value of z_i for each unit i. Then we want to select each unit with probability $\pi_i = cz_i$, where c is a positive constant. Usually it is preferable to select samples of fixed size n, since that often leads to more efficient estimators and it becomes easier to control the cost of collecting the sample. When the sample size n is fixed it is required that $\sum_{i=1}^{N} \pi_i = n$.

Now assume that the π_is are known and that $\sum_{i=1}^{N} \pi_i = n$. If we can select a sample so that the inclusion probabilities are π_i, $i = 1, 2, ..., N$, then it is possible to use the Horvitz-Thompson (HT) estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} I_i \tag{1}$$

of the unknown total Y. It is easily shown that this estimator is unbiased, i.e. $E(\hat{Y}_{HT}) = Y$. For a fixed sample size, the variance of the HT-estimator can be written as

$$\operatorname{var}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{i,j \in \mathcal{U}} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \tag{2}$$

If y is approximately proportional to z, then the variance of the HT-estimator will be low. This can be seen since if there is perfect proportionality, all the ratios y_i/π_i are equal and $\operatorname{var}(\hat{Y}_{HT}) = 0$.

It is important to notice that (2) in practice never can be calculated since it requires full knowledge of all the y_is. Hence we must be able to estimate the variance of the HT-estimator from a single sample, otherwise we have no clue about the precision of the estimate. For this purpose it is possible to use the Sen-Yates-Grundy estimator and it can be written as

$$\widehat{\operatorname{var}}_{SYG}(\hat{Y}_{HT}) = -\frac{1}{2} \sum_{i,j \in \mathcal{U}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 I_i I_j . \tag{3}$$

If $\pi_{ij} > 0$ for all $i, j \in \mathcal{U}$, this is an unbiased estimator of $\operatorname{var}(\hat{Y}_{HT})$.

4 Some πps designs and results

In this section, some of the important designs for πps-sampling are presented together with related results.

4.1 The Poisson design

A simple way to select a sample with unequal inclusion probabilities is by a method known as Poisson sampling. For Poisson sampling, each unit i is selected independently of the others with probability π_i. As a result the sample size is random when using Poisson sampling. A random sample size is usually not desirable, since it often leads to less efficient estimators. However, an advantage of Poisson sampling is its very simple implementation. It is also easy to estimate the variance of the HT-estimator due to the fact that the inclusion indicators are independent. The Poisson design has the following probability function

$$p(\mathbf{x}) = \prod_{i=1}^{N} \pi_i^{x_i}(1-\pi_i)^{1-x_i}, \quad \mathbf{x} \in \{0,1\}^N. \tag{4}$$

The Poisson design is important since the implementation of some of the other designs is based on Poisson sampling.

4.2 The conditional Poisson design

If a fixed sample size n is desired, it is possible to generate Poisson samples and accept the sample only if the sample size is n. The resulting design is called conditional Poisson (CP) sampling and it was studied by Hájek (1964) and it is also treated in his posthumous book (Hájek, 1981). Since not all samples are accepted, this procedure affects the inclusion probabilities. Let p_i, $i = 1, 2, ..., N$, be the parameters for Poisson sampling, i.e. each unit i is included independently of the others with probability p_i. Also let I_i, $i = 1, 2, ..., N$, be the inclusion indicators for Poisson sampling, i.e. the I_is are independent and $I_i \sim Be(p_i)$. If only samples of size n are accepted, we get the inclusion probabilities

$$\pi_i^{(n)} = \Pr\left(I_i = 1 \left| S_N = n\right.\right), \tag{5}$$

where $S_N = \sum_{j=1}^N I_j$. The inclusion probabilities $\pi_i^{(n)}$ can be calculated recursively by the following formula

$$\pi_i^{(n)} = n \frac{p_i/(1-p_i)(1-\pi_i^{(n-1)})}{\sum_{j=1}^N p_j/(1-p_j)(1-\pi_j^{(n-1)})}, \tag{6}$$

where $\pi_i^{(0)} = 0$, $i = 1, 2, ..., N$. Formula (6) is essentially due to Chen *et al.* (1994) and can be found in e.g. Tillé (2006, p. 81). We give a proof of this formula.

Proof of (6). First we notice that

$$\begin{aligned} \pi_i^{(n)} &= \Pr\left(I_i = 1 \left| S_N = n\right.\right) = \frac{\Pr\left(I_i = 1, S_N = n\right)}{\Pr\left(S_N = n\right)} \\ &= \frac{\Pr(I_i = 1, S_N^{(-i)} = n-1)}{\Pr\left(S_N = n\right)} = p_i \frac{\Pr(S_N^{(-i)} = n-1)}{\Pr\left(S_N = n\right)}, \end{aligned}$$

where $S_N^{(-i)} = \sum_{j \neq i} I_j$. The last equality follows from the fact that I_i and $S_N^{(-i)}$

are independent. With the same notation and technique we also have

$$\begin{aligned}1-\pi_i^{(n-1)} &= \Pr\left(I_i=0\,|S_N=n-1\right)=\frac{\Pr\left(I_i=0,S_N=n-1\right)}{\Pr\left(S_N=n-1\right)}\\ &= \frac{\Pr(I_i=0,S_N^{(-i)}=n-1)}{\Pr\left(S_N=n-1\right)}=(1-p_i)\frac{\Pr(S_N^{(-i)}=n-1)}{\Pr\left(S_N=n-1\right)}.\end{aligned}$$

Then we get

$$\frac{\pi_i^{(n)}}{1-\pi_i^{(n-1)}} = \frac{p_i}{1-p_i}\cdot\frac{\Pr\left(S_N=n-1\right)}{\Pr\left(S_N=n\right)},$$

from which (6) follows since $\sum_{i=1}^{N}\pi_i^{(n)}=n$. □

If the parameters p_i, $i=1,2,...,N$, with $\sum_{i=1}^{N}p_i=n$ are used we only have $\pi_i^{(n)}\approx p_i$. We need to adjust the p_is in order to get the inclusion probabilities π_i, $i=1,2,...,N$. The parameters can be adjusted by a simple iterative procedure due to Aires (2000),

$$p_i(t+1)=p_i(t)+(\pi_i-\pi_i^{(n)}(t)),\;\; t=0,1,2,... \tag{7}$$

where $\pi_i^{(n)}(t)$ corresponds to the inclusion probabilities of CP-sampling when the parameters $p_i(t)$ are used. These $\pi_i^{(n)}(t)$ must be calculated in each step, preferably by using (6). However, if $p_i(0)=\pi_i$, we usually only need to do a few number of iterations. If adjusted parameters are used, we call it adjusted CP-sampling, which yields correct inclusion probabilities. The probability function for CP-sampling is

$$p(\mathbf{x})=C\prod_{i=1}^{N}p_i^{x_i}(1-p_i)^{1-x_i},\;\; \mathbf{x}\in\{0,1\}^N,\;\; |\mathbf{x}|=\sum_{i=1}^{N}x_i=n, \tag{8}$$

where C is a normalizing constant.

The presented implementation of CP-sampling can be slow since in some situations many Poisson samples must be generated before we get a sample of size n. There also exist other implementations of the CP-design. One of them is a list-sequential implementation, cf. Traat *et al.* (2004) or Tillé (2006). In a list-sequential implementation the sampling outcome is first decided for unit 1, then for unit 2 and so forth. The list-sequential implementation of CP-sampling is usually very efficient for moderate size populations. For very large populations there exists no efficient implementation of CP-sampling.

In Paper I, a new design called repeated Poisson (RP) sampling is presented. The RP-design is extremely close to the CP-design and has an efficient implementation even for very large populations. The implementation of the RP-design uses Poisson sampling to repeatedly add or remove units until the sample size becomes n, which usually happens after only a few iterations.

4.3 The Pareto design

Pareto sampling was introduced by Rosén (1997a, b). Let p_i, $i = 1, 2, ..., N$, with $\sum_{i=1}^{N} p_i = n$ be the parameters. To select a sample we generate $U_1, U_2, ..., U_N$, where the U_is are independent $U(0, 1)$ random variables. Then the Pareto ranking variables

$$Q_i = \frac{U_i/(1 - U_i)}{p_i/(1 - p_i)}, \quad i = 1, 2, ..., N,$$

are calculated. The sample consists of the n units with the smallest Q_i values. If p_i, $i = 1, 2, ..., N$, are used as parameters, this procedure only yields inclusion probabilities $\pi_i \approx p_i$. As for CP-sampling, the parameters must be adjusted in order to give exactly the prescribed inclusion probabilities. It is rather complicated to do exact adjustment for this design. Different methods for adjusting the parameters for both CP-sampling and Pareto sampling have been derived and studied by Lundqvist (2009). A simple approximation of the adjusted parameters is given by Bondesson *et al.* (2006). The approximation corresponds to using the new ranking variables

$$\tilde{Q}_i = Q_i \cdot \exp\left(\frac{p_i(1 - p_i)(p_i - \frac{1}{2})}{d^2}\right),$$

where $d = \sum_{i=1}^{N} p_i(1 - p_i)$. Then the inclusion probabilities will be even closer to the p_is. Most often Pareto sampling is used without adjustment since the π_is will be very close to the p_is for fairly large populations. In this case the resulting bias of the HT-estimator (1) will usually be negligible. The advantage of the Pareto design is that the implementation is simple and very efficient. Samples can be rapidly generated even for large populations. For Pareto sampling the probability function can be written as

$$p(\mathbf{x}) = \prod_{i=1}^{N} p_i^{x_i}(1 - p_i)^{1-x_i} \times \sum_{k=1}^{N} c_k x_k, \quad |\mathbf{x}| = n, \tag{9}$$

where the constants c_k are defined by integrals and approximately $c_k \propto 1 - p_k$, see e.g. Bondesson *et al.* (2006) for details.

4.4 The Sampford design

The Sampford design was introduced by Sampford (1967) and is one of the first πps designs for fixed sample size. This design is exact, i.e. it yields exactly the prescribed inclusion probabilities π_i, $i = 1, 2, ..., N$, with $\sum_{i=1}^{N} \pi_i = n$. No adjustment is needed for the parameters. The probability function for the design is

$$p(\mathbf{x}) = C \prod_{i=1}^{N} \pi_i^{x_i} (1 - \pi_i)^{1-x_i} \times \sum_{k=1}^{N} (1 - \pi_k) x_k, \quad \mathbf{x} \in \{0, 1\}^N, \quad |\mathbf{x}| = n, \tag{10}$$

where C is a normalizing constant.

The first implementation of this design was given by Sampford (1967) and it can be described in the following way. First one unit is drawn with replacement according to the probabilities π_i/n, $i = 1, 2, ..., N$. Then $n - 1$ further units are drawn with replacement according to the probabilities $p_i' \propto \pi_i/(1 - \pi_i)$, with $\sum_{i=1}^{N} p_i' = 1$. If all the n units are distinct, then the sample is accepted. Of course, the algorithm may be restarted as soon as a doublet is drawn. In general this is a very slow procedure since a large proportion of the samples will be rejected.

Another implementation is to first select one unit according to the probabilities π_i/n, $i = 1, 2, ..., N$. Then a Poisson sample is selected among all units with the probabilities π_i, $i = 1, 2, ..., N$. If the Poisson sample has size $n - 1$ and all n units are distinct, then the sample is accepted. Otherwise the procedure is repeated from the beginning. Traat *et al.* (2004) made an improvement of this implementation. The first unit is selected in the same way. Then a conditional Poisson sample of size $n - 1$ is selected by using a list-sequential method. If all n units are distinct the sample is accepted.

Bondesson *et al.* (2006) presented another rejective implementation of the Sampford design by noticing the fact that the Sampford design is very close to the Pareto design. A Pareto sample can often be accepted as a Sampford sample by using an acceptance-rejection technique. This implementation is rather technical and some approximations must be used in practice.

The main results of Paper II are two new implementations of the Sampford design that are non-rejective. The idea behind the most efficient of these two new implementations is to adjust the drawing probabilities for the first selected unit and then to generate a Poisson sample under the conditions that the sample size is n and that the first selected unit is included. The procedure can be described as follows. The first unit should be selected according to the drawing probabilities

$$q_i = \frac{\pi_i (1 - \pi_i^{(n-1)})}{\sum_{j=1}^{N} \pi_j (1 - \pi_j^{(n-1)})}, \quad i = 1, 2, ..., N,$$

where $\pi_i^{(n-1)}$ corresponds to the inclusion probabilities for conditional Poisson sampling with parameters $p_i = \pi_i$ and sample size $n-1$. These $\pi_i^{(n-1)}$ can rapidly be calculated by using formula (6). Assume that unit k was selected in the first draw. Then a Poisson sample should be selected under the conditions that the sample size is n and that unit k is selected. This corresponds to selecting a conditional Poisson sample of size n using the parameters $p_i(k)$, where

$$p_i(k) = \begin{cases} \pi_i, & i \neq k \\ 1, & i = k. \end{cases}$$

If a list-sequential method is used to select the conditional Poisson sample, this is a non-rejective implementation of the Sampford design. Thus the efficiency of this implementation is not dependent of the parameters π_i, which is the main advantage of a non-rejective method.

4.5 The splitting method

The general splitting method was introduced by Deville & Tillé (1998) and is also treated in Tillé (2006, Ch. 6). The idea is to start with the vector $\boldsymbol{\pi} = \boldsymbol{\pi}(0) = (\pi_1, \pi_2, ..., \pi_N)$ of inclusion probabilities and then split this vector into two or more new vectors. Then one of the new vectors is chosen randomly in such a way that the expected value of the new vector $\boldsymbol{\pi}(1)$ equals the previous vector $\boldsymbol{\pi}(0)$. When a coordinate of $\boldsymbol{\pi}(t)$ becomes 0 or 1, it cannot be further changed. The splitting is continued until all coordinates of the vector are 0 or 1. In each step we have $E(\boldsymbol{\pi}(t+1)|\boldsymbol{\pi}(t)) = \boldsymbol{\pi}(t)$, thus this method always respects the inclusion probabilities.

Every πps design can be implemented by the splitting method. In general it can be difficult to determine how the splits should be performed. Different special cases have been introduced. One is the pivotal method, proposed by Deville & Tillé (1998). For the pivotal method the inclusion probabilities are updated for two units at a time, in such a way that the sampling outcome is determined for at least one of the units. The pivotal method is presented and compared to other designs in Paper IV, where it is found to have good properties. The pivotal method has also appeared in other fields, see e.g. Dubhashi *et al.* (2007).

5 Real-time sampling and correlated Poisson sampling

In real-time sampling the units of the population pass the sampler one by one and the sampler must instantly decide for each unit whether or not it should be sampled. When the sampler makes a decision for unit i, there is no information available for the units $i+1, ..., N$. Even the population size N may be unknown. Thus unit i may be the last unit that arrive. Different methods for real-time sampling with equal and unequal inclusion probabilities were studied by Meister (2004). Here it is assumed that the value of some auxiliary size-variable becomes known for the sampler at sight of the units. Thus the desired inclusion probability for unit i becomes known at least when unit i arrive. Real-time sampling is a much more complicated situation since less information is available in advance. There are two obvious ways of taking a πps sample in this situation. One is to use Poisson sampling and accept a large variation in sample size and the other is to use systematic sampling. Systematic sampling does not include much randomness and the order of the units is sometimes randomized to overcome this problem. In real-time sampling that is not possible.

In Paper III a new and general method for real-time sampling, called correlated Poisson sampling, is investigated. Correlated Poisson sampling was introduced by Bondesson & Thorburn (2008). It is a list-sequential method where the inclusion probabilities are successively updated. At step 1 of the procedure, unit 1 is included with probability $\pi_1^{(0)} = \pi_1$. Then, at step i when the value of I_{i-1} has been recorded, unit i is included with the updated probability

$$\pi_i^{(i-1)} = \pi_i - \sum_{j=1}^{i-1} \left(I_j - \pi_j^{(j-1)} \right) w_j^{(i)}. \tag{11}$$

The $w_j^{(i)}$s are weights that can be chosen in many different ways, cf. Paper III or Bondesson & Thorburn (2008) for details.

In Paper III, it was found that if units that are close in the ordering have similar values of the variable of interest, a lot of efficiency can be gained by using correlated Poisson sampling instead of Poisson sampling. Another advantage of correlated Poisson sampling is that the variation of the sample size can be reduced. Sometimes it is even possible to have a fixed sample size. In Paper IV, the probability function for correlated Poisson sampling is derived. It is shown that, since it is a list sequential procedure, the probability function can be written in terms of the updated inclusion probabilities

$$p(\mathbf{x}) = \prod_{i=1}^{N} \left(\pi_i^{(i-1)}\right)^{x_i} \left(1-\pi_i^{(i-1)}\right)^{1-x_i}, \quad \mathbf{x} \in \{0,1\}^N. \tag{12}$$

The updated inclusion probabilities are always known and given by (11) for a generated sample $\mathbf{x}$.

6 πps-sampling when the inclusion probabilities do not sum to an integer

Most methods for πps sampling are used to select samples of a fixed size. Then it is required that the inclusion probabilities sum to an integer. However it is not always a good idea to rescale preliminary inclusion probabilities to sum to an integer. Assume that $\sum_{i=1}^{N} \pi_i = n + a$, where $n \geq 0$ is integer and $a \in (0,1)$.

One approach to get correct inclusion probabilities is to use a technique called random rounding. With probability a the inclusion probabilities are rounded upwards to $\pi_i^U = \frac{n+1}{n+a}\pi_i$, where $\sum_{i=1}^{N} \pi_i^U = n+1$. Otherwise, with probability $1-a$ the inclusion probabilities are rounded downwards to $\pi_i^L = \frac{n}{n+a}\pi_i$, where $\sum_{i=1}^{N} \pi_i^L = n$. After this random rounding any πps-design for fixed sample size may be used. Unfortunately this technique does not always work since some of the π_i^Us may be larger than 1.

In Papers V and VI we give different solutions to this problem that always work. In Paper V it is shown that some designs (Sampford, conditional Poisson and Pareto) can be extended to this situation. In Paper VI, we go further and show that every design for fixed size πps sampling can be used to select a sample when $\sum_{i=1}^{N} \pi_i = n + a$. The simple trick that is used here is to add a phantom unit $N+1$, so that $\sum_{i=1}^{N+1} \pi_i = n+1$. Now any design for fixed size πps sampling can be used to select a sample of size $n+1$ from this extended population. If the phantom unit is selected it is dismissed so that the sample size becomes n, otherwise the sample size is $n+1$. Thus we have shown that it is always possible to get correct inclusion probabilities with any fixed size πps-design even if the inclusion probabilities do not sum to an integer.

7 Comparing different designs

For a sampling design to be generally applicable it is necessary that it includes much randomness. Otherwise the estimator can have a very large variance and a very skew distribution. One measure of randomness is Shannon's entropy, which for a design with support Q is defined as

$$H = -\sum_{\mathbf{x} \in Q} p(\mathbf{x}) \log(p(\mathbf{x})) = -E_p \left[\log(p(\mathbf{x}))\right]. \tag{13}$$

The adjusted CP-design has maximum entropy among all fixed size πps-designs, as was shown by Hájek (1981). In Paper IV, eight different designs are compared with respect to their entropy. In order to calculate the entropy, the probability function must be known. A general method to derive the probability function from a sampling algorithm is also presented in Paper IV.

Several designs have nearly maximum entropy. The top four designs are adjusted CP, adjusted Pareto, a design called Brewer's method, and Sampford. Also the pivotal method has high entropy if the units, for which the inclusion probabilities are updated, are chosen randomly in each step. Systematic sampling has the lowest entropy if it is used without first randomizing the order of the units. If the order of the units is first randomized, systematic sampling has higher entropy but is not close to having maximum entropy.

In order to compare different designs it is also possible to look at some measure for the distance between designs. One such measure is the Hellinger distance, d_H, which is defined as

$$d_H^2(p_1, p_2) = \frac{1}{2} \sum_{\mathbf{x} \in Q} \left(\sqrt{p_1(\mathbf{x})} - \sqrt{p_2(\mathbf{x})} \right)^2,$$

where $Q = Q_1 \cup Q_2$ and Q_1, Q_2 are the supports of $p_1(\cdot)$ and $p_2(\cdot)$.

Lundqvist (2007) compared some designs for πps sampling by deriving expressions for asymptotic distances between the designs. It was found that adjusted CP, adjusted Pareto and the Sampford design are close. Here it suffices with a small example to illustrate the distances between the designs.

Example 4. A population of size $N = 10$, known as the Sampford-Hájek population, is used. Let $n = 5$ and

$$\boldsymbol{\pi} = (0.2, 0.25, 0.35, 0.4, 0.5, 0.5, 0.55, 0.65, 0.7, 0.9).$$

For this population, the Hellinger distances have been calculated between eight different designs and the result is presented in Table 1. The eight designs are adjusted CP (ACP), adjusted Pareto (APareto), Brewer, Sampford, Pivotal, splitting into simple random sampling (SSRS), systematic πps and two variants of correlated Poisson sampling, cf. Paper IV for a description of the different designs.

Table 1: Hellinger distances between the different designs for the population in example 4. The probability function of the Pivotal design has been estimated with 10^7 samples.

	ACP	APareto	Brewer	Sampf	Corr(p)	Pivotal	SSRS	Corr(m)
ACP	0							
APareto	0.0026	0						
Brewer	0.0030	0.0044	0					
Sampf	0.0032	0.0006	0.0048	0				
Corr(p)	0.0137	0.0128	0.0143	0.0127	0			
Pivotal	0.0210	0.0230	0.0211	0.0235	0.0282	0		
SSRS	0.2396	0.2419	0.2396	0.2424	0.2448	0.2240	0	
Corr(m)	0.5518	0.5505	0.5517	0.5501	0.5491	0.5577	0.6750	0
Syst	0.8534	0.8533	0.8533	0.8532	0.8517	0.8528	0.8667	0.8191

The result shows that the designs that yield high entropy also have probability functions that are very close. The two closest designs are Sampford and adjusted Pareto. ☐

8 Summary of the papers

In this section short summaries of the six papers are presented.

8.1 Paper I: Repeated Poisson sampling

In this paper a new design for fixed size unequal probability sampling is presented. The new design, Repeated Poisson (RP) sampling, is based on Poisson sampling. Units are successively added to or removed from the sample until the desired sample size is achieved. The probability function of the RP-design is derived, not in a closed form but as the limit distribution of a Markov chain.

It is shown, by examples and simulation, that the RP-design is very close to the conditional Poisson (CP) design. The advantage of the RP-design over the CP-design is that it is more efficient in selecting samples from large populations.

Also, a variant of the RP-design can be used to efficiently select samples of fixed size within strata in the case of several stratifications.

8.2 Paper II: Non-rejective implementations of the Sampford sampling design

Sampford sampling, introduced by Sampford (1967), is a method for fixed size unequal probability sampling. The Sampford design has nearly maximum entropy and also the advantage of giving exactly the prescribed inclusion probabilities. A major drawback of the Sampford design has been that the implementations has been rejective and slow. In some situations the rejective implementations may even fail to produce a sample due to the low acceptance rate.

In this paper different non-rejective implementations are presented. The main advantage of these implementations is that the efficiency is not dependent on the inclusion probabilities. Thus a sample can always be generated. One of the non-rejective implementations is rather efficient and that method is a modification of a rejective list-sequential method introduced by Traat *et al.* (2004). The other method is a list-sequential method where updated conditional inclusion probabilities are calculated in each step. That method requires somewhat more calculations. These new implementations make the Sampford design more practical and usable.

8.3 Paper III: On a generalization of Poisson sampling

In this paper a new design for real time sampling is studied. The new design, called correlated Poisson sampling, was introduced by Bondesson & Thorburn (2008). In real time sampling the units pass the sampler successively one by one. Each unit passes the sampler only once and at that time it must be decided whether or not it should be included in the sample. There is no information available for the units that have not yet passed. Even the population size may be unknown in advance. Of course, the population size becomes known when all units have passed the sampler.

Two traditional πps-sampling methods that can be used in this situation are Poisson sampling and systematic sampling. Poisson sampling has the disadvantage of giving a random sample size with a large variation. The drawback of systematic sampling is that the design has low entropy, i.e. a low level of randomization.

The new method, correlated Poisson sampling, is very general and certain weights can be chosen in many different ways. Some different strategies for choosing the weights are compared. It is shown that in many cases it is possible to get more efficient estimators than we get by using Poisson sampling. It is also possible to reduce the variation of the sample size compared to Poisson sampling. In some situations it is possible to have a fixed sample size. Also, this design generally gives a much higher level of randomization than systematic sampling.

8.4 Paper IV: Entropy of unequal probability sampling designs

Eight designs for unequal probability sampling are compared with respect to their entropy. The entropy is a measure of randomness and a high entropy is usually preferred. Both old and more recent designs are compared.

In order to calculate exactly the entropy of a design, the probability function must be known. For some designs the probability function had not been presented previously. An approach to derive the probability function from a sampling algorithm is presented and also used to derive the probability function for some of the designs. One of them is the correlated Poisson design presented in Paper III. Also several general estimators of the entropy are presented and compared by simulation. It is shown by two different examples that several designs are close to having maximum entropy. Some designs yield low entropy and one should be careful when choosing these designs.

8.5 Paper V: An extension of Sampford's method for unequal probability sampling

The Sampford design is extended to the case where the inclusion probabilities do not sum to an integer. A modified version of Sampford's algorithm is presented. The sampling outcome is left open for one randomly chosen unit that gets a new inclusion probability. A generalized vector of inclusion indicators is introduced, where one coordinate is allowed to be in the interval $(0, 1)$, i.e. the outcome is undecided for exactly one unit. The probability function is derived for this generalized vector. It is proved that the prescribed inclusion probabilities are achieved with the new algorithm. Moreover, the conditional Poisson and Pareto designs are extended. Three different applications are presented. Variance estimation for some different sampling situations is also treated in Appendices.

8.6 Paper VI: Efficient sampling when the inclusion probabilities do not sum to an integer

It is shown that every unequal probability sampling design for fixed sample size can be extended to the case when the inclusion probabilities do not sum to an integer. The cost for not re-scaling the inclusion probabilities is that we have to accept a small variation in sample size. Let N be the population size and let $\pi_1, \pi_2, ..., \pi_N$, with $\sum_{i=1}^{N} \pi_i = n + a$, where $n \geq 0$ is integer and $a \in (0, 1)$, be the prescribed inclusion probabilities. By adding a phantom unit $N+1$ with inclusion probability $1 - a$, the inclusion probabilities will sum to $n + 1$ for the extended population. Then any design for fixed size πps-sampling can be applied. If the phantom unit is selected, it is dismissed and the sample size becomes n, otherwise the sample size is $n + 1$. It is clear that the prescribed inclusion probabilities are achieved with this procedure. By choosing the adjusted conditional Poisson design it is possible to sample with maximum entropy in this situation. Different strategies for estimation under these circumstances are also given.

9 Conclusions and open problems

In general we advocate the use of a high entropy design. Having high entropy is particularly important if some assumptions do not hold. An example of such an assumption is that we assume that the inclusion probabilities are approximately proportional to the variable of interest. Other assumptions that sometimes are made concern the ordering of the units in the population. When the entropy is high, the probability mass is well distributed over a large number of samples. When the entropy is low it is possible that most or all of the probability mass is put on bad samples, where bad means that the HT-estimate is far from the true total. Variance estimation also becomes easier with a high entropy design, since then the variance of the HT-estimator can be well estimated without the use of second-order inclusion probabilities, see e.g. Tillé (2006, pp. 137-142).

However, there are several high entropy πps designs that are very close to each other. Since these designs will produce similar results, it does not matter much which is used. All these designs have different advantages. The ACP-design has maximum entropy, but it has a somewhat complicated implementation. The Sampford design is slightly easier to implement since no adjustment of the parameters is needed. Pareto sampling has a very efficient implementation but the parameters need to be adjusted, which can be rather difficult to do exactly. Brewer's method and the pivotal design are very simple to implement but does not allow for exact calculation of second-order inclusion probabilities. Thus one may choose different designs depending on what property is important for the specific situation.

For real-time sampling, the correlated Poisson design gave very promising results for the simulated populations in Paper III. Here the assumption that units that are close with respect to order also have similar values of the variable of interest is an important factor for the improvement. The method should be further evaluated to see what happens when that assumption does not hold. It should also be tested with some real applications.

An interesting problem for the future is when several auxiliary variables are available and several totals are to be estimated, which is a common situation in practice. How should the inclusion probabilities be chosen? Which design should be used to draw the sample? Which estimators should be used? There are some different approaches that can be used.

One possibility is to use essentially the additional auxiliary information after the sample has been collected. Then it is possible to use a generalized regression estimator (GREG) instead of the HT-estimator, see e.g. Särndal (1996). The main idea for this estimator is to fit a regression model to the observed y-values by using the auxiliary information. The GREG-estimator uses both observed and estimated y-values. If the regression relationship is strong, the GREG-estimator will be nearly unbiased and will have a low variance. More than one auxiliary variable may be used to determine what inclusion probabilities to use. Some proposals of how to choose the inclusion probabilities in the multivariate case are given and discussed by Holmberg (2003).

The additional information can also be more directly incorporated in the sample selection procedure. Deville & Tillé (2004) introduced the cube method that can be used to select balanced samples with given inclusion probabilities. Then the HT-estimator reproduces the known totals for the auxiliary variables, at least approximately. By only allowing balanced samples to be selected, the support and thus the entropy may be heavily reduced. Usually this procedure give much better estimates but reducing the support can have a negative effect also. Thus, the cube method may be a good alternative but its entropy needs to be evaluated. The effect on the estimator also needs to be evaluated for different situations.

Selecting πps-samples with general balancing conditions and maximum entropy is an unsolved problem. Lundqvist (2009) treats this problem for some specific balancing conditions. CP-sampling in this situation corresponds to only accepting samples that fulfil the balancing conditions. This implementation becomes inefficient if the restricted support is small. It is also a very difficult problem to adjust the parameters in order to get correct inclusion probabilities with the restrictions caused by the balancing conditions.

Restricted Pareto sampling, cf. Bondesson (2010), is another possibility. Pareto sampling (with adjustment) has been shown to be very close to adjusted CP-sampling when the restriction is fixed sample size. The restricted Pareto design can handle several restrictions on the sample and has high entropy even in such cases. As for CP-sampling, it is a difficult problem to determine what parameters to use in order to get correct inclusion probabilities.

It would be interesting to compare and evaluate these very different approaches to use additional auxiliary information.

References

Aires, N. (2000). Comparisons between conditional Poisson sampling and Pareto πps sampling designs. *J. Statist. Plann. Inference* **88**, 133-147.

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

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

Bondesson, L. (2010) Conditional and restricted Pareto sampling; Two new methods for unequal probability sampling. *Scand. J. Statist.* **37**, doi:10.1111/j.1467-9469.2010.00700.x.

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

Deville, J-C. & Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89-101.

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling; the cube method. *Biometrika* **91**, 893-912

Dubhashi, D., Jonasson, J. & Ranjan, D. (2007). Positive influence and negative dependence. *Combin. Probab. Comput.* **16**, 29-41.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *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.

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

Lundqvist, A. (2009). Contributions to the theory of unequal probability sampling. Ph.D-thesis, Dept. of mathematics and mathematical statistics, Umeå university.

Meister, K. (2004). On methods for real time sampling and distributions in sampling. Ph.D-thesis, Dept. of mathematics and mathematical statistics, Umeå university.

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.

Särndal, C-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *J. Amer. Statist. Assoc.* **91**, 1289-1300.

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

Tillé, Y. (2006). *Sampling algorithms.* Springer series in statistics, 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.