Open Journal of Statistics, 2014, 4, 178-187
Published Online April 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.43017

Sampling Designs with Linear and Quadratic Probability Functions

Lennart Bondesson[1], Anton Grafström[2], Imbi Traat[3]

[1]Department of Mathematics and Mathematical Statistics, Umeå University, Umeå, Sweden
[2]Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden
[3]Institute of Mathematical Statistics, University of Tartu, Tartu, Estonia
Email: lennart.bondesson@math.umu.se, anton.grafstrom@slu.se, imbi.traat@ut.ee

Received 20 January 2014; revised 20 February 2014; accepted 27 February 2014

Abstract

Fixed size without replacement sampling designs with probability functions that are linear or quadratic functions of the sampling indicators are defined and studied. Generality, simplicity, remarkable properties, and also somewhat restricted flexibility characterize these designs. It is shown that the families of linear and quadratic designs are closed with respect to sample complements and with respect to conditioning on sampling outcomes for specific units. Relations between inclusion probabilities and parameters of the probability functions are derived and sampling procedures are given.

Keywords

Complementary Midzuno Design; Conditional Sample; Inclusion Probability; Midzuno Design; Mixture of Designs; Parameters of Design; Sample Complement; Sinha Design

1. Introduction

In the first part of the paper, we consider the most general fixed size without replacement sampling design with a probability function that is linear in the sampling inclusion indicators—the linear sampling design. The linear design, being an unequal probability design, is remarkable due to simple explicit relations between inclusion probabilities and parameters of the probability function. This enables sampling with desired inclusion probabilities, design-unbiased estimation and variance estimation. As special cases, the design covers the classical Midzuno [1] and the complementary Midzuno (see [2]) designs. It is shown that the linear design can be seen as a mixture of the two types of Midzuno designs. It is also shown that the family of linear designs is closed with respect to conditioning on sampling outcomes. This property, as well as the mixture representation, offers easy

How to cite this paper: Bondesson, L., Grafström, A. and Traat, I. (2014) Sampling Designs with Linear and Quadratic Probability Functions. *Open Journal of Statistics*, **4**, 178-187. http://dx.doi.org/10.4236/ojs.2014.43017

methods for sampling from the linear design. The family is also closed with respect to sample complements, *i.e.* the complement of a sample from a linear design is a sample from another linear design.

In the second part of the paper, the fixed size without replacement design with quadratic probability function—the quadratic sampling design—is defined and studied. It is the natural extension of the linear design. A classical design by Sinha [3] is a special case of the quadratic design. His design aimed at sampling with prescribed second-order inclusion probabilities. Together with Sinha's design, two other special quadratic designs are studied more closely. These three designs are easy to use for sampling. There are explicit expressions for the second-order inclusion probabilities of Sinha's design. In the general case, the formulas are more complicated. A lemma is proved which relates the second-order inclusion probabilities and the parameters of the quadratic design. Like the family of linear designs, the family of quadratic designs is closed with respect to sample complements and with respect to conditioning on sampling outcomes. The last property makes list-sequential sampling from these designs efficient.

The linear and the quadratic designs are simple but somewhat restricted when the aim is sampling with prescribed first- or second-order inclusion probabilities. They cannot be used for all such probabilities. In the final section of the paper, possible extensions are mentioned.

2. Linear Sampling Designs

We treat without replacement (WOR) sampling designs of size n from a population of size N. Let $\boldsymbol{I}=(I_1,I_2,\cdots,I_N)$ be the binary random sampling inclusion vector. A sampling design is given by its probability function $p(\boldsymbol{x})=\Pr(\boldsymbol{I}=\boldsymbol{x})$. Let $S_n=\left\{\boldsymbol{x}\in\{0,1\}^N;\ |\boldsymbol{x}|=\sum_{k=1}^N x_k=n\right\}$, *i.e.* S_n is the set of all possible "samples" $\boldsymbol{x}$ of size n.

Definition 1. *A sampling design of size* $n\ (1\le n\le N-1)$ *is called linear if there are real constants* $c_1,c_2,\cdots,c_N$ *such that* $p(\boldsymbol{x})\propto\sum_{k=1}^N c_k x_k$, $\boldsymbol{x}\in S_n$.

Of course, equal c_ks give SRSWOR. Below it is always assumed that the c_ks are normalized to have sum 1. Then $p(\boldsymbol{x})=C\sum_{k=1}^N c_k x_k$, where $C=1\Big/\binom{N-1}{n-1}$. In fact,

$$1=\sum_{\boldsymbol{x}\in S_n}p(\boldsymbol{x})=C\sum_{k=1}^N c_k\sum_{\boldsymbol{x}\in S_n}x_k=C\sum_{k=1}^N c_k\binom{N-1}{n-1}=C\binom{N-1}{n-1}.$$

The inclusion probabilities $\pi_i=\Pr(I_i=1)$ of the linear design are given by

$$\pi_i=c_i+(1-c_i)\frac{n-1}{N-1}. \tag{1}$$

Indeed, since $\sum_{k=1}^N c_k=1$,

$$\pi_i=\sum_{\boldsymbol{x}\in S_n}x_i p(\boldsymbol{x})=C\sum_{k=1}^N c_k\sum_{\boldsymbol{x}\in S_n}x_i x_k=C\left(c_i\binom{N-1}{n-1}+(1-c_i)\binom{N-2}{n-2}\right).$$

Since $0\le\pi_i\le 1$, (1) shows that we must have $-\frac{n-1}{N-n}\le c_i\le 1$. The c_i s can be expressed in terms of the π_i s as

$$c_i=\frac{N-1}{N-n}\left(\pi_i-\frac{n-1}{N-1}\right). \tag{2}$$

By similar algebra, the second-order inclusion probabilities $\pi_{ij}=\Pr(I_iI_j=1)$ are given by

$$\pi_{ij}=\frac{n-1}{N-1}(c_i+c_j)+(1-c_i-c_j)\frac{(n-1)(n-2)}{(N-1)(N-2)},\quad i\ne j,$$

and hence

$$\pi_{ij} = \frac{n-1}{N-2}\left(\pi_i + \pi_j - \frac{n}{N-1}\right), \quad i \neq j, \tag{3}$$

i.e. the π_{ij} s are linear in the first-order inclusion probabilities.

Without restriction we may assume that the c_ks are given in increasing order. Obviously, in order to get $p(\boldsymbol{x}) \geq 0$, it is then necessary and sufficient that $\sum_{k=1}^{n} c_k \geq 0$. We obtain the following theorem.

Theorem 1. *Let* $0 \leq \pi_1 \leq \pi_2 \leq \cdots \leq \pi_N \leq 1$ *be given numbers with sum n. Then there is a linear sampling design with these numbers as inclusion probabilities if and only if* $\frac{1}{n}\sum_{k=1}^{n} \pi_k \geq \frac{n-1}{N-1}$.

Proof. By the relation (2), we see that $\sum_{k=1}^{n} c_k \geq 0 \Leftrightarrow \frac{1}{n}\sum_{k=1}^{n} \pi_k \geq \frac{n-1}{N-1}$. □

To sample from a linear sampling design, we may use the well-known acceptance-rejection (AR) technique. A constant *A* such that $p(\boldsymbol{x}) = C\sum_{k=1}^{N} c_k x_k \leq A p_{SRS}(\boldsymbol{x}) = A \Big/ \binom{N}{n}$ for all $\boldsymbol{x} \in S_n$ is first found. Assuming that the c_ks are in increasing order, it suffices to put $A = \frac{N}{n}\sum_{k=N-n+1}^{N} c_k$. Then we generate a tentative sample $\boldsymbol{x}$ according to SRSWOR of size *n* and a random number $U \sim U(0,1)$. The sample is accepted as a sample from the linear design if

$$U \leq \frac{C\sum_{k=1}^{N} c_k x_k}{A p_{SRS}(\boldsymbol{x})} = \frac{N}{n}\frac{\sum_{k=1}^{N} c_k x_k}{A} = \frac{\sum_{k=1}^{N} c_k x_k}{\sum_{k=N-n+1}^{N} c_k}.$$

The full procedure is repeated until a sample is accepted. The acceptance rate equals $\sum_{k=1}^{N} c_k \pi_k \Big/ \sum_{k=N-n+1}^{N} c_k$. To be seen later on, also other sampling techniques exist.

There are two special linear sampling designs: the Midzuno design and the complementary Midzuno design.

The *Midzuno design* has the probability function $p_M(\boldsymbol{x}) \propto \sum_{k=1}^{N} a_k x_k$, $\boldsymbol{x} \in S_n$, with $\sum_{k=1}^{N} a_k = 1$ and $a_k \geq 0$ for all *k*. Apparently it is a linear design. We may sample from the design by selecting one unit according to the probabilities a_k and then $n-1$ further units according to SRSWOR from the remaining $N-1$ ones. The design was introduced in [1] (considering the a_k s as proportional to auxiliary variables). Tillé ([4], p. 117) generalizes it and gives many further references concerning the design. Brewer and Hanif ([5], p. 25) remark that the inequalities on the inclusion probabilities are restrictive.

The *complementary Midzuno design* (see [2]) has the probability function

$$p_{cM}(\boldsymbol{x}) = \binom{N-1}{n}^{-1} \sum_{k=1}^{N} b_k (1 - x_k), \; \boldsymbol{x} \in S_n,$$

with $\sum_{k=1}^{N} b_k = 1$ and $b_k \geq 0$ for all k. This design is considered in [6] with the b_k s proportional to auxiliary variables. We may sample from the design by removing one unit according to the probabilities b_k and sample n units by SRSWOR among the remaining $N-1$ ones. For the complementary Midzuno design, we then get $\pi_i = \frac{n}{N-1}(1-b_i)$. The formula for the second-order inclusion probabilities coincides with (3). Since

$$\sum_{k=1}^{N} b_k (1 - x_k) = \sum_{k=1}^{N} \left(n^{-1} - b_k\right) x_k, \; \boldsymbol{x} \in S_n,$$

we also see that the complementary Midzuno design is a linear design with coefficients $\tilde{c}_k = \frac{n}{N-n}\left(\frac{1}{n} - b_k\right)$ with sum 1. If $b_k \leq 1/n$ for all *k*, it is even a Midzuno design and hence the two designs overlap each other.

Remark 1. *Let* $0 \leq \pi_1 \leq \pi_2 \leq \cdots \leq \pi_N \leq 1$ *have sum n, i.e.* $\sum_{k=1}^{N} \pi_k = n$. *Then, by* (2), *these numbers are inclusion probabilities for a Midzuno design iff* (*a*) $\pi_1 \geq \frac{n-1}{N-1}$. *Obviously* (*a*) *implies the less restrictive inequa-*

lity in Theorem 1. *The* π_i *s are inclusion probabilities of a complementary Midzuno design iff* (*b*) $\pi_N \le \frac{n}{N-1}$. *Of course, also* (*b*) *implies the inequality in Theorem* 1. *In fact, assuming the contrary that* $\frac{1}{n}\sum_{k=1}^{n}\pi_k < \frac{n-1}{N-1}$, *we get by* (*b*) *that* $\sum_{k=1}^{N}\pi_k < \frac{n(n-1)}{N-1} + \frac{(N-n)n}{N-1} = n$, *which is a contradiction as the sample size is n.*

The family of linear designs can be considered as closed with respect to sample complements. More precisely, the complement of any sample of size n from a linear design, is a sample of size $N-n$ from another linear design. This follows from the relation

$$\sum_{k=1}^{N} c_k x_k = \sum_{k=1}^{N}\left(\frac{1}{N-n} - c_k\right)(1-x_k), \quad \boldsymbol{x} \in S_n.$$

Another interesting property of the family of linear designs is that it is closed with respect to conditioning on the sampling outcomes for specific units. For instance, if we know that unit 1 is selected (*i.e.* $x_1 = 1$), then the probability function for the remaining size $n-1$ sampling among $N-1$ units is still linear. In fact, given that $\pi_1 > 0$ and that $I_1 = x_1 = 1$, we have by (1) that $c_1 > -\frac{n-1}{N-n}$. The conditional probability function of $\boldsymbol{I}_2 = (I_2, I_3, \cdots, I_N)$ is then given by

$$\Pr(\boldsymbol{I}_2 = \boldsymbol{x}_2) \propto \frac{1}{\pi_1}\left(\sum_{k=2}^{N} c_k x_k + c_1\right) \propto \sum_{k=2}^{N}\left(c_k + c_1\frac{1}{n-1}\right)x_k, \ |\boldsymbol{x}_2| = n-1.$$

The new coefficients $c_k + \frac{c_1}{n-1}, k \ge 2,$ have sum $1 - c_1 + \frac{N-1}{n-1}c_1 = 1 + \frac{N-n}{n-1}c_1 > 0$ by the inequality for c_1 above. Normalizing them, we get coefficients $c_k^1 = \frac{(n-1)c_k + c_1}{(N-n)c_1 + n - 1}$ with sum 1. If it is known that unit 1 is not selected ($x_1 = 0$ and $\pi_1 < 1$, $c_1 < 1$), the new coefficients are simply given by $c_k^0 = c_k \big/ \sum_{i=2}^{N} c_i$, $k \ge 2$.

It follows that samples from a linear design can easily be generated *list-sequentially*. The first unit in the population is sampled with probability $\pi_1 = c_1 + (1-c_1)\frac{n-1}{N-1}$. If it is selected, then the second unit is sampled with probability $c_2^1 + (1-c_2^1)\frac{n-2}{N-2}$. Else the second unit is sampled with probability $c_2^0 + (1-c_2^0)\frac{n-1}{N-2}$ etc.

3. A Mixture Result

A linear design can also be called a *mixed* Midzuno design because of the following theorem.

Theorem 2. *A linear sampling design* $p(\boldsymbol{x}) \propto \sum_{k=1}^{N} c_k x_k$, $\boldsymbol{x} \in S_n$, *where* $\sum_{k=1}^{N} c_k = 1$, *is a mixture of a Midzuno design and a complementary Midzuno design. It is a Midzuno design if* $c_k \ge 0$ *for all k and it is a complementary one if* $c_k \le 1/(N-n)$ *for all k.*

The components of the mixture are not unique. The very last statement in the theorem follows from the re-writing $\sum_{k=1}^{N} c_k x_k = \sum_{k=1}^{N}\left(\frac{1}{N-n} - c_k\right)(1-x_k)$, $\boldsymbol{x} \in S_n$. The full proof of the theorem is somewhat technical and is given in an appendix. The proof yields the following procedure for finding suitable components of the mixture. It is assumed that the design is not a pure Midzuno design and that the units in the population are ordered such that $c_1 \le c_2 \le \cdots \le c_m < 0 \le c_{m+1} \le \cdots \le c_N$, where $1 \le m < n$ (since $\sum_{k=1}^{n} c_k \ge 0$).

Procedure:

Let $\nu = m$. While $c_{\nu+1} < -\frac{1}{n-\nu}\sum_{k=1}^{\nu} c_k$, let $\nu = \nu + 1$.

(As $\sum_{k=1}^{n} c_k \geq 0$, the final ν is less than n.)

Then put $\beta = -\frac{N-n}{n-\nu}\sum_{k=1}^{\nu} c_k$, $\alpha = 1-\beta$, and $\gamma = \beta\frac{n}{N-n}$.

Finally, let the parameters for the components be defined by

$$a_k = \begin{cases} 0, & k = 1,2,\cdots,\nu \\ \left(c_k - \frac{\gamma}{n}\right)\Big/\alpha, & k = \nu+1,\cdots,N \end{cases} \quad \text{and} \quad b_k = \begin{cases} \frac{1}{n} - \frac{c_k}{\gamma}, & k = 1,2,\cdots,\nu \\ 0, & k = \nu+1,\cdots,N. \end{cases}$$

The first component, the Midzuno one, is used with probability α and the second component, the complementary Midzuno one, is used with probability β.

Example 1. *Let* $N = 8$, $n = 6$, *and* $\boldsymbol{c} = \left[-\frac{1}{4}, -\frac{1}{10}, -\frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{1}{4}, \frac{7}{20}, \frac{1}{2}\right]$. *Obviously* $m = 3$. *It is not possible to use* $\nu = 3$ *since* $c_4 = \frac{1}{10} < -\frac{1}{n-3}\sum_{k=1}^{3} c_k = \frac{8}{60}$. *The procedure gives us* $\nu = 4$, $\beta = \frac{3}{10}$, $\alpha = \frac{7}{10}$ *and* $\gamma = \frac{9}{10}$ *and* $\boldsymbol{a} = \left[0,0,0,0,\frac{1}{14},\frac{1}{7},\frac{2}{7},\frac{1}{2}\right]$ *and* $\boldsymbol{b} = \left[\frac{4}{9},\frac{5}{18},\frac{2}{9},\frac{1}{18},0,0,0,0\right]$. *However, since* $c_k \leq \frac{1}{N-n} = \frac{1}{2}$, *the design is also a pure complementary Midzuno design with* $b_k = \frac{1}{n} - \frac{N-n}{n}c_k$ *and hence* $\boldsymbol{b} = \left[\frac{1}{4},\frac{1}{5},\frac{11}{60},\frac{2}{15},\frac{1}{10},\frac{1}{12},\frac{1}{20},0\right]$.

Theorem 2 implies another simple way of generating samples from the linear design. First it is decided by a random choice whether a Midzuno or a complementary Midzuno design should be used. Then one of these designs is applied. In Tillé's ([4], pp. 99-104) terminology, we can see this technique as a special *splitting technique* which after two steps ends with SRSWOR.

4. Quadratic Sampling Designs

There is a natural extension of the linear designs. In this section we look at fixed size designs with *quadratic* probability functions. We could use ordinary quadratic forms. However, it is more appropriate to sum over sets of size 2 than over pairs with ordered elements. Below $\sum_{\{i,j\}}$ means a sum over the $\binom{N}{2}$ sets $\{i,j\}$, $j \neq i$.

Definition 2. *A sampling design of size* $n\,(2 \leq n \leq N-2)$ *is called quadratic if there are real constants* d_{ij} *with* $d_{ji} = d_{ij}$ *and* $\sum_{\{i,j\}} d_{ij} = 1$ *such that* $p(\boldsymbol{x}) \propto \sum_{\{i,j\}} d_{ij}x_i x_j$, $\boldsymbol{x} \in S_n$.

In particular, if $n = 2$ then d_{ij} is the probability to select the sample $\{i,j\}$. The normalization constant can be shown to be equal to the reciprocal of $\binom{N-2}{n-2}$. It can also be shown that the parameters (d_{ij}) are uniquely determined by the design. To get $p(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in S_n$, it is necessary and sufficient that $\sum_{\{i,j\}\subseteq A} d_{ij} \geq 0$ for all subsets A of $\{1,2,\cdots,N\}$ of size n. This means that the possible parameters (d_{ij}) form a convex set in $\mathbb{R}^M$, where $M = \binom{N}{2}$. However, this set is rather complicated if n is not very small.

For $2 \leq n \leq N-2$, a linear design can also be seen as a quadratic design, because $(n-1)x_i = \sum_{j;\,j\neq i} x_i x_j$, $\boldsymbol{x} \in S_n$. There are three natural quadratic designs with their basic parameters symmetric and *nonnegative*:

(a) $p_a(\boldsymbol{x}) \propto \sum_{\{i,j\}} a_{ij} x_i x_j,\ \boldsymbol{x} \in S_n.$

(b) $p_b(\boldsymbol{x}) \propto \sum_{\{i,j\}} b_{ij} (1-x_i)(1-x_j),\ \boldsymbol{x} \in S_n.$

(c) $p_c(\boldsymbol{x}) \propto \sum_{\{i,j\}} c_{ij} \dfrac{x_i(1-x_j)+x_j(1-x_i)}{2},\ \boldsymbol{x} \in S_n.$

Assuming that $\sum_{\{i,j\}} a_{ij} = 1$, $\sum_{\{i,j\}} b_{ij} = 1$, and $\sum_{\{i,j\}} c_{ij} = 1$, we readily find that for the three cases the normalization constants are the reciprocals of $\binom{N-2}{n-2}$, $\binom{N-2}{n}$, and $\binom{N-2}{n-1}$, respectively. It is not clear that the designs (b) and (c) are quadratic according to Definition 2 but it will be obvious later on. It will also follow that the complement of a sample from a quadratic design of size *n* can be seen as a sample from a quadratic design of size $N-n$.

The design (a) corresponds to selecting one pair of units according to the probabilities a_{ij} and then $n-2$ further ones by SRSWOR. The design (b) was considered by Sinha in [3]. Remove one pair of units according to the probabilities b_{ij} with sum 1 and then choose *n* units by SRSWOR among the remaining ones. The design (c) corresponds to selecting one pair according to the probabilities c_{ij} and then keeping one of the units for the sample and removing the other one. Then $n-1$ units are selected by SRSWOR among the $N-2$ non-selected units.

Remark 2. *An extension of the design* (*c*) *would be to let* $p_c(\boldsymbol{x}) \propto \sum_{i,j} c_{ij} x_i (1-x_j)$, $\boldsymbol{x} \in S_n$, *with* $c_{ij} \geq 0$ *and* $c_{ii} = 0$ *but not necessarily* $c_{ij} = c_{ji}$. *In this case the units i and j have different roles. In the symmetric case the new parameters* c_{ij} *equal the previous ones multiplied by* 1/2. *This extension is not considered further here.*

Using the facts that for $\boldsymbol{x} \in S_n$, $(n-1)x_i = \sum_{j;\, j\neq i} x_i x_j$, and $\binom{n}{2} = \sum_{\{i,j\}} x_i x_j$ and

$$\sum_{\{i,j\}} b_{ij}(x_i + x_j) = \sum_{\{i,j\}} \frac{B_i + B_j}{n-1} x_i x_j, \text{ where } B_i = \sum_{j;\, j\neq i} b_{ij},$$

it is not difficult to see that also the designs (b) and (c) are indeed quadratic designs according to Definition 2 with parameters d_{ij} that are not necessarily nonnegative. More explicitly we have, with $B = \sum_{i=1}^{N} B_i = 2\sum_{\{i,j\}} b_{ij} = 2$,

$$\sum_{\{i,j\}} b_{ij}(1-x_i)(1-x_j) = \sum_{\{i,j\}} \left(b_{ij} - \frac{B_i + B_j}{n-1} + \frac{B}{n(n-1)} \right) x_i x_j,$$

$$\sum_{\{i,j\}} c_{ij} \frac{x_i(1-x_j)+x_j(1-x_i)}{2} = \sum_{\{i,j\}} \left(\frac{C_i + C_j}{2(n-1)} - c_{ij} \right) x_i x_j, \text{ where } C_i = \sum_{j;\, j\neq i} c_{ij}.$$

One can then realize that $p_b(\boldsymbol{x}) \propto \sum_{\{i,j\}} d_{ij}^b x_i x_j$, where

$$d_{ij}^b = \frac{n(n-1)}{(N-n)(N-n-1)} \left(b_{ij} - \frac{B_i + B_j}{n-1} + \frac{B}{n(n-1)} \right) \text{ and } \sum_{\{i,j\}} d_{ij}^b = 1. \tag{4}$$

Further

$$p_c(\boldsymbol{x}) \propto \sum_{\{i,j\}} d_{ij}^c x_i x_j, \text{ where } d_{ij}^c = \frac{n-1}{N-n} \left(\frac{C_i + C_j}{2(n-1)} - c_{ij} \right) \text{ and } \sum_{\{i,j\}} d_{ij}^c = 1.$$

Given parameters d_{ij} with sum 1 we may solve the equations $d_{ij}^b = d_{ij}$ for possible nonnegative solutions b_{ij} and the equations $d_{ij}^c = d_{ij}$ for possible nonnegative solutions c_{ij}. In fact, we have

$$b_{ij} = \frac{(N-n)(N-n-1)}{n(n-1)} \left(d_{ij} - \frac{D_i + D_j}{N-n-1} + \frac{D}{(N-n)(N-n-1)} \right), \tag{5}$$

where $D_i = \sum_{j;\, j\neq i} d_{ij}$ and $D = \sum_{i=1}^{N} D_i = 2$. This result is a consequence of formula (4) but with the sample of size n replaced by its complement of size $N-n$. Somewhat more complicated calculations show that, at least if $2n \neq N$,

$$c_{ij} = \frac{N-n}{(N-2n)(n-1)}\left(D_i + D_j\right) - \frac{D}{(N-2n)(n-1)} - \frac{N-n}{n-1} d_{ij}. \tag{6}$$

This gives us the following theorem.

Theorem 3. *A quadratic sampling design* $p(\boldsymbol{x}) \propto \sum_{\{i,j\}} d_{ij} x_i x_j$, $\boldsymbol{x} \in S_n$, *with* $\sum_{\{i,j\}} d_{ij} = 1$ *is a design of type* (*a*) *iff* $d_{ij} \geq 0$. *It is design of type* (*b*) (*Sinha design*) *iff*

$$d_{ij} - \frac{D_i + D_j}{N-n-1} + \frac{D}{(N-n)(N-n-1)} \geq 0.$$

It is a design of type (*c*) *iff*

$$\frac{1}{N-2n}\left(D_i + D_j\right) - \frac{D}{(N-2n)(N-n)} - d_{ij} \geq 0.$$

It can happen that a design is of all the three types simultaneously. For example, SRSWOR is such a design.

Example 2. *Let* $N = 5, n = 3$ *and let the design be defined by*

$$d_{12} = \frac{1}{6}, d_{13} = \frac{1}{12}, d_{14} = -\frac{1}{12}, d_{15} = \frac{1}{6}, d_{23} = d_{24} = \frac{1}{4}, d_{25} = d_{34} = d_{35} = 0, d_{45} = \frac{1}{6}.$$

By (5) *and* (6), *the corresponding b- and c-parameters are then given by*

$$b_{12} = \frac{1}{18}, b_{13} = \frac{5}{36}, b_{14} = \frac{1}{12}, b_{15} = \frac{1}{6}, b_{23} = b_{24} = \frac{1}{12}, b_{25} = 0, b_{34} = b_{35} = \frac{1}{9}, b_{45} = \frac{1}{6}$$

$$c_{12} = -\frac{1}{6}, c_{13} = \frac{1}{4}, c_{14} = \frac{5}{12}, c_{15} = \frac{1}{6}, c_{23} = c_{24} = -\frac{1}{4}, c_{25} = 0, c_{34} = c_{35} = \frac{1}{3}, c_{45} = \frac{1}{6}.$$

Thus in this case the design is not of type (*a*) *or* (*c*) *but of type* (*b*).

In the linear case we saw that a general linear design can be represented as a mixture of a Midzuno and a complementary Midzuno design. It would be desirable that a general quadratic design could be represented as a mixture of the three designs (a), (b), and (c) (or its extension). However, such a result is not likely to hold.

For the design (b) the second-order inclusion probabilities π_{ij} (which determine the first-order ones since $(n-1)\pi_i = \sum_{j;\, j\neq i} \pi_{ij}$) are easy to find. We have

$$\pi_{ij} = \frac{n(n-1)}{(N-2)(N-3)} \sum_{\{k,l\};\{k,l\}\cap\{i,j\}=\varnothing} b_{kl}.$$

It is also easy to find the *b*-parameters that correspond to (π_{ij}). In fact,

$$b_{ij} = \frac{(N-2)(N-3)}{n(n-1)} \pi_{ij} - \frac{N-2}{n}\left(\pi_i + \pi_j\right) + 1.$$

This result is due to Sinha [3] but more generally also a special case of Lemma 1 below. Sinha used his result to find a method to sample with prescribed second-order inclusion probabilities. However, for the method to work with nonnegative b_{ij}, strong restrictions on the π_{ij} s are needed.

Lemma 1. *Let* $q_{ij}, i \neq j$, *be given numbers with* $q_{ij} = q_{ji}$. *Then the equations*

$$\sum_{\{k,l\};\{k,l\}\cap\{i,j\}=\varnothing} c_{kl} = q_{ij}, \quad i \neq j, \tag{7}$$

in the unknowns $c_{ij} = c_{ji}, i \neq j$, *have the solution*

$$c_{ij} = q_{ij} - \frac{Q_i + Q_j}{N-3} + \frac{Q}{(N-2)(N-3)}, \text{ where } Q_i = \sum_{j;j\neq i} q_{ij}, Q = \sum_{i=1}^{N} Q_i. \quad (8)$$

We only have to put $q_{ij} = \pi_{ij}\frac{(N-2)(N-3)}{n(n-1)}$, to get the earlier solution $b_{ij}\ (= c_{ij})$ from the lemma. The proof of Lemma 1 is given in the appendix.

For a design given by $p(\boldsymbol{x}) \propto \sum_{\{i,j\}} d_{ij}x_i x_j$, the formulas for the second-order inclusion probabilities π_{ij} are somewhat complicated. However, we can first find b_{ij} from formula (5) and then use the π_{ij} formula above. The inverse procedure $(\pi \to b \to d)$ can also be used to find *d*-parameters corresponding to given second-order inclusion probabilities. The "mixed case" (c) can be handled in the same way. A computer program makes these procedures simple to use.

To sample from the general quadratic design, we could use an AR-technique by dominating $p(\boldsymbol{x})$ by a multiple of $p_{SRS}(\boldsymbol{x})$. However, it is probably simpler to use list-sequential sampling. In fact, given $I_1 = x_1$ (0 or 1), the probability function for the remaining sample of size n or $n-1$ is again quadratic, cf. Section 2. But we must recalculate parameters (coefficients) and calculate first-order inclusion probabilities. The first-order inclusion probability π_1 is given by (assuming that the *d*-sum is 1)

$$\pi_1 = \sum_{l;l\geq 2} d_{1l} + \frac{n-2}{N-2}\sum_{\{k,l\},l>k\geq 2} d_{kl}.$$

Given that $I_1 = x_1$ with $x_1 = 0$, the new *d*-coefficients (with sum 1) are just

$$d_{ij}^0 = \frac{d_{ij}}{\sum_{\{k,l\},l>k\geq 2} d_{kl}}, \quad i,j \geq 2, i \neq j.$$

If $x_1 = 1$ and $n > 2$, the new *d*-coefficients (with sum 1) are instead given by

$$d_{ij}^1 \propto d_{ij} + \frac{d_{1i} + d_{1j}}{n-2}, \quad i,j \geq 2, i \neq j.$$

For $n = 2$ the conditional design is a linear design with parameters proportional to d_{1i}.

5. Additional Comments

As recently becoming more common in sampling articles, we have used sampling indicators and focused on the probability function of these. The names *linear design* and *quadratic design* become very natural for this reason.

There are several advantages of the linear design: there are simple relations between the parameters and the inclusion probabilities of first and second order. It is also easy to sample. Thus the design is easy to use and for the Horvitz-Thompson estimate of a population total, we can also readily find a variance estimate. The drawback of the design is its lack of sufficient flexibility: although much more general than SRSWOR, it is not able to cover all possible first-order inclusion probabilities.

The quadratic design is more complicated than the linear one. By using a quadratic design, it is possible to sample with prescribed second-order inclusion probabilities. However, it cannot be used for all such possible inclusion probabilities and therefore this design is also not flexible enough.

To get more flexibility, the linear and quadratic functions need to be complemented by binomial factors. To sample with prescribed first-order inclusion probabilities, we may generally use a probability function of the form

$$p(\boldsymbol{x}) \propto \prod_{i=1}^{N} p_i^{x_i}(1-p_i)^{1-x_i} \times \sum_{k=1}^{N} c_k x_k, \quad \boldsymbol{x} \in S_n,$$

where the parameters are suitably chosen. At least three well-known designs are of this form (with nonnegative c_k): the conditional Poisson design, the Sampford design, and the Pareto design, cf. [7]. For these three cases, the linear function $\sum_{k=1}^{N} c_k x_k$ is proportional to $\sum_{k=1}^{N} b_k(1-x_k)$ with nonnegative parameters b_k.

To sample with prescribed second-order inclusion probabilities, we may use a design with

$$p(\boldsymbol{x}) \propto \prod_{i=1}^{N} p_i^{x_i} (1-p_i)^{1-x_i} \times \sum_{\{k,l\}} b_{kl} (1-x_k)(1-x_l), \quad \boldsymbol{x} \in S_n.$$

Letting $p_i = \pi_i$ (the desired first-order inclusion probabilities) and using Lemma 1, it is possible to calculate the suitable parameters b_{kl} without much effort. Details will be presented in a planned forthcoming paper. This design is much more flexible than the quadratic design but it has not full flexibility. A fully flexible design uses a probability function of the form $p(\boldsymbol{x}) \propto \exp\left(\sum_{\{i,j\}} a_{ij} x_i x_j\right)$, $\boldsymbol{x} \in S_n$, but it is not easy to find the parameters [8].

Acknowledgements

This research was supported by the Estonian Science Foundation grant 8789.

References

[1] Midzuno, H. (1952) On the Sampling System with Probability Proportional to Sum of Sizes. *Annals of the Institute of Statistical Mathematics*, **3**, 99-107. http://dx.doi.org/10.1007/BF02949779

[2] Bondesson, L. and Traat, I. (2013) On Sampling Designs with Ordered Conditional Inclusion Probabilities. *Scandinavian Journal of Statistics*, **40**, 724-733. http://dx.doi.org/10.1111/sjos.12024

[3] Sinha, B.K. (1973) On Sampling Schemes to Realize Preassigned Sets of Inclusion Probabilities of First Two Orders. *Calcutta Statistical Association Bulletin*, **22**, 89-110.

[4] Tillé, Y. (2006) Sampling Algorithms. Springer, New York.

[5] Brewer, K.R.W. and Hanif, M. (1983) Sampling with Unequal Probabilities. *Lecture Notes in Statistics*, *No.* 15, Springer-Verlag, New York. http://dx.doi.org/10.1007/978-1-4684-9407-5

[6] Wywiał, J. (2000) On Precision of Horvitz-Thompson Strategies. *Statistics in Transition*, **4**, 779-798.

[7] Bondesson, L., Traat, I. and Lundqvist, A. (2006) Pareto Sampling versus Conditional Poisson and Sampford Sampling. *Scandinavian Journal of Statistics*, **33**, 699-720. http://dx.doi.org/10.1111/j.1467-9469.2006.00497.x

[8] Bondesson, L. (2012) On Sampling with Prescribed Second-Order Inclusion Probabilities. *Scandinavian Journal of Statistics*, **39**, 813-829. http://dx.doi.org/10.1111/j.1467-9469.2012.00808.x

Appendix

Here proofs are given of Theorem 2 and Lemma 1.

Proof of Theorem 2. Excluding the case that $c_k \geq 0$ for all k, and then ordering the c_ks as $c_1 \leq c_2 \leq \cdots \leq c_m < 0 \leq c_{m+1} \leq \cdots \leq c_N$, where $1 \leq m < n$, we shall show that

$$c_k = \alpha a_k + \gamma\left(\frac{1}{n} - b_k\right), \quad k = 1, 2, \cdots, N, \tag{9}$$

where $a_k \geq 0$ and $\sum_{k=1}^N a_k = 1$ and $b_k \geq 0$ and $\sum_{k=1}^N b_k = 1$ and $\alpha \geq 0$, $\gamma > 0$. Re-writing then $\gamma\left(\frac{1}{n} - b_k\right)$ as $\beta \frac{n}{N-n}\left(\frac{1}{n} - b_k\right) = \beta \tilde{c}_k$ (with $\sum_{k=1}^N \tilde{c}_k = 1$), we have the desired mixture representation

$$\sum_{k=1}^N c_k x_k = \alpha \sum_{k=1}^N a_k x_k + \beta \sum_{k=1}^N \tilde{c}_k x_k, \quad \alpha + \beta = 1.$$

We shall show that (9) can be achieved by putting for some ν such that $m \leq \nu < n$, $a_k = 0$ and $b_k = \frac{1}{n} - \frac{c_k}{\gamma}$ for $k = 1, \cdots, \nu$ and $b_k = 0$ and $a_k = \left(c_k - \frac{\gamma}{n}\right)\Big/\alpha$ for $k = \nu+1, \cdots, N$. To get $\sum_{k=1}^N b_k = 1$ and $\gamma > 0$, we must have $\sum_{k=1}^\nu c_k < 0$ and $\gamma = -\sum_{k=1}^\nu c_k \Big/ \left(1 - \frac{\nu}{n}\right)$. To get $b_k \geq 0$ and $a_k \geq 0$, we must have

$$c_\nu \leq \gamma/n = -\frac{\sum_{k=1}^\nu c_k}{n-\nu} \quad \text{and} \quad c_{\nu+1} \geq \gamma/n = -\frac{\sum_{k=1}^\nu c_k}{n-\nu}. \tag{10}$$

(Then there is no problem to choose $\alpha \geq 0$ such that $\sum_{k=1}^N a_k = 1$.) We start by trying $\nu = m$. Since $c_m < 0$ and $\sum_{k=1}^m c_k < 0$, the first inequality in (10) is certainly satisfied with strict inequality. If also the second one is satisfied, we can stop and use $\nu = m$. Otherwise, the second inequality is not satisfied and then $c_{m+1} < -\frac{\sum_{k=1}^m c_k}{n-m}$. Equivalently $c_{m+1} < -\frac{\sum_{k=1}^{m+1} c_k}{n-m-1}$ (and $\sum_{k=1}^{m+1} c_k < 0$). If now $c_{m+2} \geq -\frac{\sum_{k=1}^{m+1} c_k}{n-m-1}$, we can stop and put $\nu = m+1$. Otherwise $c_{m+2} < -\frac{\sum_{k=1}^{m+1} c_k}{n-m-1}$. Equivalently $c_{m+2} < -\frac{\sum_{k=1}^{m+2} c_k}{n-m-2}$. If now $c_{m+3} \geq -\frac{\sum_{k=1}^{m+2} c_k}{n-m-2}$, we can stop and put $\nu = m+2$. The procedure may continue until $c_{n-1} < -\frac{\sum_{k=1}^{n-1} c_k}{n-(n-1)} = -\sum_{k=1}^{n-1} c_k$. But $c_n \geq -\sum_{k=1}^{n-1} c_k$ (as $p(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in S_n$) and we can then put $\nu = n-1$ and stop. □

Proof of Lemma 1. Let $C_i = \sum_{j;\, j \neq i} c_{ij}$ and $C = \sum_{i=1}^N C_i$. Then (7) can be re-expressed as $\frac{1}{2}C - C_i - C_j + c_{ij} = q_{ij}$. It follows that

$$Q_i = \frac{1}{2}(N-1)C - (N-1)C_i - (C - C_i) + C_i = (N-3)\left(\frac{1}{2}C - C_i\right),$$

and hence $C_i = \frac{1}{2}C - \frac{Q_i}{N-3}$. Thus $c_{ij} = q_{ij} - \frac{Q_i + Q_j}{N-3} + \frac{1}{2}C$. Summing then over i and $j, i \neq j$, we get $C = Q - (N-1)\frac{Q+Q}{N-3} + \frac{1}{2}N(N-1)C$ and hence $\frac{1}{2}C = Q\big/\big((N-2)(N-3)\big)$. This shows that the solution must be of the form (8). It is not difficult to check that (8) also is a solution. □