

SAMPLING DESIGNS ON FINITE POPULATIONS WITH SPREADING CONTROL PARAMETERS

Yves Tillé¹, Lionel Qualité^{1,2} and Matthieu Wilhelm¹

¹*University of Neuchâtel* and ²*Swiss Federal Statistical Office*

Abstract: We present new sampling methods in finite population that allow one to control the joint inclusion probabilities of units and especially the spreading of sampled units in the population. They are based on the use of renewal chains and multivariate discrete distributions to generate the difference of population ranks between successive selected units. With a Bernoulli sampling design, these differences follow a geometric distribution, and with a simple random sampling design they follow a negative hypergeometric distribution. We propose to use other distributions and introduce a large class of sampling designs with and without fixed sample size. The choice of the rank-difference distribution allows us to control units joint inclusion probabilities with a relatively simple method and closed form formula. Joint inclusion probabilities of neighboring units can be chosen to be larger, or smaller, compared to those of Bernoulli or simple random sampling, thus allowing more or less spread of the sample in the population. This can be useful when neighboring units have similar characteristics or, on the contrary, are very different. A set of simulations illustrates the qualities of this method.

Key words and phrases: Sampling design, survey sampling.

1. Introduction

In this paper, we propose sampling methods for fixed and random sample sizes. We more particularly focus on the spacings that are the difference of population ranks between two successive selected units. We propose a large set of new methods that allows one to control the spacings and thus the joint inclusion probabilities of population units in the sample. These methods are useful in that they allow one to make less (or more) likely the selection of neighboring units. Indeed, when the variable of interest takes similar values on neighboring units, spreading the sample improves estimation because the selection of similar units is avoided.

A sampling design is a probability distribution on all the finite subsets of a population. It can be implemented by means of sampling algorithms. Several different sampling algorithms can implement the same sampling designs. Examples

are given in Tillé (2006) where a large number of algorithms is given for designs like Simple Random Sampling (SRS) with and without replacement or maximum entropy sampling designs. Algorithms such that the decision of selecting or not a unit into the sample is taken for each population unit successively according to the order of the population sampling frame are called “sequential” or “one-pass” algorithms. These algorithms are particularly useful when the population list is dynamic, like on a production chain or in real time sampling applications.

Systematic sampling is one of the most common sampling designs. It has been studied among others by Madow and Madow (1944), Cochran (1946), Madow (1949), Bellhouse and Rao (1975), Iachan (1982), Iachan (1983), Murthy and Rao (1988), Bellhouse (1988), Bellhouse and Sutradhar (1988), and Pea, Qualité and Tillé (2007). One advantage of systematic sampling is that it spreads the sample very well over the population, thus allowing one to get precise estimators for totals and averages in the case of “auto-correlated” interest variables. Indeed, it can be shown to be an optimal design in this case under some conditions (Bondesson (1986)). However, it presents the important drawback that lots of unit couples have null joint inclusion probabilities. This makes impossible an unbiased estimation of the variance.

This drawback has led to a quest for other sampling designs that would retain good estimation properties. Deville (1998) proposed the Deville-systematic method, also called ordered pivotal method by Chauvet (2012) (see also Tillé (2006, pp.128-130)). Tillé (1996) proposed a moving stratification algorithm that avoids the selection of neighboring units. Bondesson and Thorburn (2008) and Grafström (2010) also proposed a method that allows one to control joint-inclusion probabilities. Recently, Loonis and Mary (2015) proposed using determinantal point processes that are known for their repulsiveness property (see for example Daley and Vere-Jones (2002, p.138)). This last method necessitates work with a huge matrix.

We advocate the use of point processes with simple specifications, motivated by usual sampling designs: the systematic design has deterministic spacings between selected units, the Bernoulli sampling design (see for example Tillé (2006, pp.43–44)) has geometrically distributed spacings, and circular spacings of the simple random sampling design follow a negative hypergeometric distribution (see Vitter (1984, 1985, 1987)). In this paper, we will use other distributions to tune the joint selection probability of neighboring units. For each of these methods, we are able to compute positive joint inclusion probabilities and unbiased variance estimators. Special attention to edge effects must be given to ensure

correct first-order inclusion probabilities. Part of these sampling designs, with independent and identically distributed spacings, were introduced by Bondesson (1986).

The paper is organized as follows: Section 2 is devoted to the main definitions of survey sampling theory. Sections 3 and 4 present renewal chain sampling designs for random size samples. In Sections 5 and 6, we discuss fixed size sampling obtained through the generation of circular spacings with multivariate discrete distributions. Simulation results are given in Section 7. The paper ends with our conclusions in Section 8.

2. Sampling from a Finite Population

Consider the finite population of N units, $U = \{1, \dots, N\}$. A sample without replacement of U is a subset $s \subset U$. A sampling design $P(\cdot)$ is a probability distribution on samples,

$$P(s) \geq 0, s \subset U, \text{ such that } \sum_{s \subset U} P(s) = 1.$$

Let S denote the random sample, so that $\Pr(S = s) = P(s)$. The sample size $n = \#S$ can be random or not. The inclusion probability of unit k is its probability of being selected into a sample

$$\pi_k = \Pr(k \in S) = \sum_{s \ni k} P(s).$$

The joint inclusion probability of units k and ℓ is their probability of being selected together into a sample

$$\pi_{k\ell} = \Pr(k \text{ and } \ell \in S) = \sum_{s \ni k, \ell} P(s).$$

Let Y be a variable of interest and let y_k be the value of Y associated to unit k of the population. The Horvitz and Thompson (1952) estimator is defined by

$$\hat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

It is an unbiased estimator of the population total

$$t_Y = \sum_{k \in U} y_k,$$

provided that $\pi_k > 0, k \in U$. Let

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell & \text{if } k \neq \ell, \\ \pi_k(1 - \pi_k) & \text{if } k = \ell. \end{cases}$$

The variance of the HT-estimator is

$$\text{var}(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.$$

If the sampling design has a fixed size, the variance can also be written as (see Sen (1953); Yates and Grundy (1953)):

$$\text{var}(\widehat{Y}) = -\frac{1}{2} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell}.$$

Estimators can be derived from these two expressions. For the general case, the Horvitz and Thompson (1952) variance estimator is given by

$$\widehat{\text{var}}_{HT}(\widehat{Y}) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}}, \quad (2.1)$$

where $\pi_{kk} = \pi_k$. When the sample size is fixed, the Sen (1953) and Yates and Grundy (1953) variance estimator is given by

$$\widehat{\text{var}}_{SYG}(\widehat{Y}) = -\frac{1}{2} \sum_{k \in S} \sum_{\substack{\ell \in S \\ \ell \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \quad (2.2)$$

These estimators are unbiased provided that $\pi_{k\ell} > 0$, $k \neq \ell \in U$. Estimator (2.2) is non-negative when $\Delta_{k\ell} \leq 0$, $k \neq \ell$ (Sen-Yates-Grundy conditions).

3. Renewal Chain Sampling Designs

The idea of selecting samples through the use of renewal processes is not new. It can be traced back at least to Bondesson (1986) (see also Meister (2004)). We give a different presentation in this section in that we focus on the parametrization of the distribution of spacings between selected units whereas Bondesson (1986) and Meister (2004) focus on the parametrization of the so-called *renewal sequence*, i.e. the conditional inclusion probabilities given the past. Their aim was to provide solutions for real time sampling, and the proposed methods are intrinsically sequential, allowing one to spread the sample by introducing a negative correlation between the sample inclusion indicators. Bondesson and Thorburn (2008) generalize this idea using a splitting method (see Deville and Tillé (1998)) that allows use of unequal probability sampling designs for real time sampling.

3.1. Definition

In this section, we present a family of sampling algorithms that are parametrized by a discrete probability distribution. By a careful choice of the generating dis-

tribution, we obtain sampling designs with desirable properties. Consider a sequence J_1, \dots, J_N of independently and identically distributed (i.i.d.) random variables in $\mathbb{N}^* = \{1, 2, 3, \dots\}$. The partial sums $S_j = \sum_{i=1}^j J_i$, $j \geq 1$, form a discrete process that is called a simple renewal chain (see for example Feller (1971), Barbu and Limnios (2008, p.18)) by analogy with renewal processes (see Cox (1962), Daley and Vere-Jones (2002), Mitov and Omey (2014)). Using these J_i 's as jumps (or spacings) between successive units selected into the sample, we obtain the family of sampling designs of Definition 1.

Definition 1. *A sampling design is said to be a (simple) renewal chain sampling design if its random sample can be written*

$$\tilde{S} = \{1, \dots, N\} \cap \left\{ \sum_{i=1}^j J_i, 1 \leq j \leq N \right\},$$

where J_1, \dots, J_N are i.i.d. random variables in \mathbb{N}^* .

The first-order inclusion probability of a renewal chain design can be obtained from the common distribution $f(\cdot)$ of the J_i 's:

$$\pi_k = \Pr(k \in \tilde{S}) = \sum_{j=1}^k f^{j*}(k), \tag{3.1}$$

where $f^{j*}(\cdot)$ is the distribution of the sum of j i.i.d. variables with distribution $f(\cdot)$. Indeed, unit k is selected if $J_1 = k$, or $J_1 + J_2 = k$, or \dots , or $J_1 + \dots + J_k = k$. These events are non-overlapping thanks to the J_i 's being positive. We obtain that:

$$\pi_k = \sum_{j=1}^k \Pr\left(\sum_{i=1}^j J_i = k\right),$$

which is exactly Equation (3.1). It is a well-known property of renewal process theory given, for example in Barbu and Limnios (2008, p.21), Cox (1962, p.53), or in Mitov and Omey (2014, pp.44-47).

Even with i.i.d. spacings, a simple renewal chain sampling design usually has unequal first order inclusion probabilities, as we can see in Example 1.

Example 1. Let J_i , $i \in \mathbb{N}^*$ be a sequence of i.i.d. variables such that $\Pr(J_i = 1) = 1/2$ and $\Pr(J_i = 2) = 1/2$. Then,

$$\begin{aligned} \pi_1 &= \Pr(J_1 = 1) = \frac{1}{2}, \\ \pi_2 &= \Pr(J_1 = 2) + \Pr(J_1 + J_2 = 2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}, \end{aligned}$$

$$\begin{aligned} \pi_3 &= \Pr(J_1 + J_2 = 3) + \Pr(J_1 + J_2 + J_3 = 3) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}, \\ \pi_4 &= \Pr(J_1 + J_2 = 4) + \Pr(J_1 + J_2 + J_3 = 4) + \Pr(J_1 + J_2 + J_3 + J_4 = 4) \\ &= \frac{11}{16}, \\ &\vdots \end{aligned}$$

3.2. Equilibrium renewal chains

A delayed renewal chain is a discrete process $(S_j)_{j \in \mathbb{N}}$ with $S_j = \tilde{J}_0 + \sum_{i=1}^j J_i$, where the J_i 's, $i \geq 1$ are i.i.d. random variables taking values in \mathbb{N}^* and \tilde{J}_0 is an independent random variable taking values in \mathbb{N} (see e.g. Barbu and Linnios (2008, p.31)). Of particular interest is the delayed renewal chain obtained when the distribution of \tilde{J}_0 is obtained from the distribution of J_1 using

$$\Pr(\tilde{J}_0 = k) = \frac{\Pr(J_1 \geq k + 1)}{\mathbb{E}(J_1)}, \quad k \in \mathbb{N}, \quad (3.2)$$

provided that $\mathbb{E}(J_1)$ exists. The distribution of \tilde{J}_0 is called the stationary or equilibrium distribution of the renewal chain and the resulting delayed renewal chain is called an equilibrium renewal chain. As written by Barbu and Linnios (2008, Proposition 2.2), this choice of the initial distribution \tilde{J}_0 of the delayed renewal chain is the only one where all $k \in \mathbb{N}$ have the same probability of being in the sample path. Proposition 1 is a general result of renewal process theory (see for example Mitov and Omey (2014, p.46)) that we applied to the discrete case. We propose a direct proof of Proposition 1 in Appendix.

Proposition 1. *If $f(\cdot)$ is a probability distribution on \mathbb{N}^* with cumulative distribution function $F(\cdot)$, expectation μ , and if $f_0(\cdot)$ is defined by $f_0(k) = f(\{k + 1, \dots\})/\mu$, $k \in \mathbb{N}$, then $f_0(\cdot)$ is a probability distribution and*

$$f_0(k) + \sum_{t=1}^k f_0(k-t) \sum_{j=1}^t f^{j*}(t) = \frac{1}{\mu}, \quad \text{for all } k \geq 1. \quad (3.3)$$

Corollary 1. *Let S_j , $j \in \mathbb{N}$ be a delayed renewal chain with $\mathbb{E}(J_1) = \mu$ and \tilde{J}_0 have the distribution of (3.2). For all $k \in \mathbb{N}$, if π_k is the probability that k is in the sample path, then*

$$\pi_k = f_0(k) + \sum_{t=1}^k f_0(k-t) \sum_{j=1}^t f^{j*}(t), \quad (3.4)$$

and is equal to $1/\mu$.

Proof. The event $\{\exists i \in \mathbb{N} \text{ such that } S_i = k\}$ can be decomposed as

$$\{\exists i \in \mathbb{N} \text{ such that } S_i = k\} = \bigcup_{t=0}^k \left\{ \tilde{J}_0 = k - t \right\} \cap \bigcup_{j=1}^t \left\{ \sum_{i=1}^j J_i = t \right\},$$

where all the events in the union are non-overlapping. It follows that

$$\pi_k = f_0(k) + \sum_{t=1}^k f_0(k - t) \sum_{j=1}^t f^{j*}(t) = \frac{1}{\mu},$$

by Proposition 1.

Definition 2. Let X be a random variable with values in \mathbb{N} and finite expectation. A random variable X_F is called a forward transform of X if its distribution is given by

$$\Pr(X_F = k) = \frac{\Pr(X \geq k)}{\mathbb{E}(X + 1)}, \quad k \in \mathbb{N}.$$

Remark 1. Moments of X_F can be derived from those of X using the property, proven in the Appendix, that if X is a random variable on \mathbb{N} with finite moment of order $m + 1$, $\mathbb{E}(X^{m+1})$, $m \geq 0$, then its forward transform X_F has a finite moment of order m and

$$\mathbb{E}(X_F^m) = \frac{\mathbb{E}[F_m(X)]}{\mathbb{E}(X + 1)},$$

where $F_m(x)$ is the Faulhaber polynomial integer function of degree $m + 1$: $F_m(x) = \sum_{k=0}^x k^m$.

The equilibrium distribution \tilde{J}_0 is the forward transform of the distribution of $J_1 - 1$, according to Definition 2. Spacing distributions considered in Section 4 are defined as shifted variables $J_1 = 1 + X$ where X follows a classical probability distribution on \mathbb{N} . The reader can find in Table 4 a collection of distributions that are used in Sections 4 and 6, as well as their forward transforms.

3.3. Equilibrium renewal chain sampling designs

By taking the intersection of the sample path of an equilibrium renewal chain with the population $U = \{1, \dots, N\}$, one obtains a random sampling design. Corollary 1 ensures that all units of the population have the same inclusion probability. The distribution of the first selected unit index X_1 satisfies

$$\Pr(X_1 = k) = \Pr(\tilde{J}_0 = k) + \Pr(\tilde{J}_0 = 0)\Pr(J_1 = k) = \frac{\Pr(J_1 \geq k)}{\mathbb{E}(J_1)}, \quad k \in U. \quad (3.5)$$

By definition, the following sampled units are obtained by adding independent variables distributed like J_1 .

Definition 3. *An equilibrium renewal chain sampling design is the distribution of a random sample S with*

$$S = \{1, \dots, N\} \cap \left\{ \sum_{i=0}^j J_i, 0 \leq j \leq N - 1 \right\},$$

where J_1, \dots, J_{N-1} are i.i.d random variables in \mathbb{N}^* with finite expectation, and J_0 is an independent variable with distribution given by (3.5), $\Pr(J_0 = k) = \Pr(J_1 \geq k)/E(J_1)$, $k \in U$.

For $J_1 = 1 + X$, the random variable J_0 of (3.5) has the same distribution as $1 + X_F$ where X_F is a forward transform of X .

The equilibrium renewal chain design that corresponds to the renewal distribution of Example 1 is given in Example 2. Its first order inclusion probabilities are equal.

Example 2. Consider the sequence $J_i, i \in \mathbb{N}^*$ of Example 1, and define J_0 to be independent of the J_i 's, with $P(J_0 = 1) = 2/3$ and $P(J_0 = 2) = 1/3$ according to (3.5). The new inclusion probabilities $\tilde{\pi}_i$ of this equilibrium renewal sampling design are related to those of Example 1 by:

$$\begin{aligned} \tilde{\pi}_1 &= \Pr(J_0 = 1) &&= \frac{2}{3}, \\ \tilde{\pi}_2 &= \Pr(J_0 = 1)\pi_1 + \Pr(J_0 = 2) &&= \frac{1}{3} + \frac{1}{3} = \frac{2}{3}, \\ \tilde{\pi}_3 &= \Pr(J_0 = 1)\pi_2 + \Pr(J_0 = 2)\pi_1 &&= \frac{1}{2} + \frac{1}{6} = \frac{2}{3}, \\ &\vdots && \end{aligned}$$

3.4. Joint inclusion probabilities

Joint inclusion probabilities of a renewal chain sampling design can be derived from the Probability Mass Function (PMF) $f(\cdot)$ of J_1 . Indeed, the selection of unit ℓ given that unit $k, 0 < k < \ell$, is selected can be decomposed according to the number of selected units between k and ℓ , and this number does not depend on J_0 . We can write that:

$$\Pr(\ell \in S | k \in S) = \sum_{j=1}^{\ell-k} f^{j*}(\ell - k).$$

The joint inclusion probability of units $k \neq \ell$ is thus given by Equation (3.6).

$$\pi_{k\ell} = \pi_k \sum_{j=1}^{\ell-k} f^{j*}(\ell - k), \quad k < \ell. \tag{3.6}$$

3.5. Bernoulli sampling

The Bernoulli sampling design with inclusion probabilities π is obtained by selecting or not units into the sample through independent Bernoulli trials with parameter π (see for example Tillé (2006, p.43)). Its probability distribution is given by:

$$P(s) = \pi^n(1 - \pi)^{N-n}, \quad s \subset U,$$

where $n = \#s$ is the size of sample s . The joint inclusion probabilities are equal to $\pi_{k\ell} = \pi^2, k \neq \ell$. The usual algorithm used to select a sample according to the Bernoulli sampling design simply consists of generating N independent Bernoulli variables and selecting units according to the observed values.

Bernoulli sampling can also be implemented using Definition 1. Indeed, it is clear that spacings of a Bernoulli sampling design are i.i.d. distributed variables with shifted geometric distributions $J_i = 1 + X_i$ where $\Pr(X_i = k) = (1 - \pi)^k\pi, k \geq 0$. Bernoulli sampling is thus a simple renewal chain sampling design satisfying Definition 1. On the other hand it is easy to prove that, if X_i follows a geometric distribution, then X_i has the same distribution as its forward transform (it is the only distributions on \mathbb{N} that enjoy this property). The random variable J_0 of Definition 3.5 has the same distribution as J_1 in this particular case. Consequently, Bernoulli sampling is also an equilibrium renewal chain sampling design according to Definition 3.

Using (3.6), we find the second order inclusion probabilities $\pi_{k\ell} = \pi^2, k \neq \ell$. Indeed, the sum of j i.i.d. geometric random variables with parameter π follows a negative binomial distribution with parameters j and π . The negative binomial distribution with parameters $j \geq 1$ and π in $(0, 1)$ is defined by its PMF:

$$f_{\mathcal{NB}}(x) = \binom{j+x-1}{x} (1-\pi)^x \pi^j, \quad x \in \mathbb{N}, \tag{3.7}$$

where $\binom{a}{b} = a!/[b!(a-b)!]$ if $b \leq a$ are non-negative integers, and $\binom{a}{b} = 0$ if a, b or $a-b$ is negative. Considering that

$$\sum_{i=1}^j J_i = j + \sum_{i=1}^j X_i, \quad j \geq 1,$$

we have that

$$f^{j*}(x) = \binom{x-1}{x-j} (1-\pi)^{x-j} \pi^j, \quad x \geq j.$$

From Equation (3.6), we get that the joint inclusion probabilities are equal to:

$$\begin{aligned}
\pi_{k\ell} &= \pi \sum_{j=1}^{\ell-k} f^{j*}(\ell-k), \quad k < \ell \\
&= \pi \sum_{j=1}^{\ell-k} \binom{\ell-k-1}{\ell-k-j} (1-\pi)^{\ell-k-j} \pi^j \\
&= \pi^2 (\pi + 1 - \pi)^{\ell-k-1} = \pi^2.
\end{aligned}$$

3.6. Systematic sampling

Systematic sampling with rate $1/r$, $r \in \mathbb{N}^*$, from a population $U = \{1, \dots, N\}$ is obtained by generating a random start u with a uniform discrete distribution between 1 and r , and selecting units k of U such that $k \equiv u \pmod{r}$ into the sample (see Madow and Madow (1944)). The first-order inclusion probabilities of this sampling design are given by $\pi_k = 1/r$, $k \in U$, and its joint inclusion probabilities by

$$\pi_{k\ell} = 1/r \text{ if } k \equiv \ell \pmod{r} \text{ and } 0 \text{ otherwise.}$$

If $N = mr$, with $m, r \in \mathbb{N}^*$, the sample size is deterministic and equal to m .

Systematic sampling is an equilibrium renewal chain sampling design, agreeing with Definition 3 where the J_i 's, $i \geq 1$ are deterministic and equal to r . Indeed, the forward transform X_F of $X = r - 1$, $r \in \mathbb{N}^*$ is such that:

$$\Pr(X_F = k) = \frac{\Pr(X \geq k)}{\mathbb{E}(X + 1)} = \frac{\mathbf{1}_{\{r-1 \geq k\}}}{r}, \quad k \in \mathbb{N},$$

and X_F follows a uniform distribution on $\{0, \dots, r - 1\}$. Hence J_0 follows a uniform distribution on $\{1, \dots, r\}$.

The joint inclusion probabilities are obtained from (3.6). Indeed, the sum of j spacings J_i , $i \geq 1$ is deterministic, equal to jr , and

$$f^{j*}(x) = \mathbf{1}_{\{jr=x\}}, \quad x \geq 1.$$

We then have that, for $k < \ell$,

$$\pi_{k\ell} = \pi \sum_{j=1}^{\ell-k} f^{j*}(\ell-k) = \frac{1}{r} \sum_{j=1}^{\ell-k} \mathbf{1}_{\{jr=\ell-k\}} = \frac{1}{r} \mathbf{1}_{\{k \equiv \ell \pmod{r}\}}.$$

We confirm with this expression that most of the joint inclusion probabilities are null, making it impossible to estimate the variance of Horvitz-Thompson estimators without bias.

4. Spreading Renewal Chain Sampling Designs

We have seen in Sections 3.5 and 3.6 two examples of renewal chain sampling

designs with very different spreading properties. In Bernoulli sampling, the selection of units are independent, even if they are adjacent in the population list. In systematic sampling, the selection of adjacent units is impossible, provided that the sampling rate is smaller than 1. This translates to the variance of the spacings distribution: it is null for systematic sampling, that has perfect spreading properties, and it is quite large, equal to $(1 - \pi)/\pi^2$, for Bernoulli sampling.

Using Definition 3, we can build sampling designs with any given spacing distribution on \mathbb{N}^* . The expectation of this distribution is forced by the sampling rate, which is usually itself decided as a function of cost or precision constraints. In Section 4.1, we give an application with shifted negative binomial spacings, allowing for a limited control on the variance and spreading properties of the design. As a limiting case, we find the shifted Poisson spacings of Section 4.2. To have a variance that is arbitrarily small, in Section 4.3 we use shifted binomial distributions that have a variance always smaller than their expectation.

These are only examples, and any distribution or family of distributions on \mathbb{N}^* that offers sufficient control on its shape can be used. Table 4 in Appendix contains a list of useful discrete probability distributions with their probability mass functions, their supports, means and variances.

4.1. Negative binomial spacings

The definition of the negative binomial distribution in (3.7) can be extended to parameters $r > 0$ and p in $(0, 1)$ by considering the PMF:

$$f_{\mathcal{NB}(r,p)}(x) = \frac{\Gamma(r+x)}{x!\Gamma(r)} p^r (1-p)^x, \quad x \in \mathbb{N},$$

where $\Gamma(r) = \int_0^{+\infty} t^{r-1} e^{-t} dt$, $r > 0$ and $\Gamma(k) = (k-1)!$, $k \in \mathbb{N}^*$. The expectation of this distribution is $r(1-p)/p$ and its variance is $r(1-p)/p^2$.

We consider equilibrium renewal sampling designs with positive spacings J_i , $i \geq 1$ such that $J_i - 1$ follows a negative binomial distribution with parameters r and p , denoted $\mathcal{NB}(r, p)$. For a given sampling rate $\pi \in (0, 1)$, we find that $E(J_i) = 1/\pi$ implies that

$$p = \frac{r\pi}{r\pi + 1 - \pi}.$$

It follows that

$$\text{var}(J_i) = \frac{1 - \pi}{\pi} + \frac{1}{r} \left(\frac{1 - \pi}{\pi} \right)^2. \tag{4.1}$$

When $r = 1$, we find, as a special case, the Bernoulli sampling design. From (4.1), we deduce that the variance of spacings is smaller than that of Bernoulli sam-

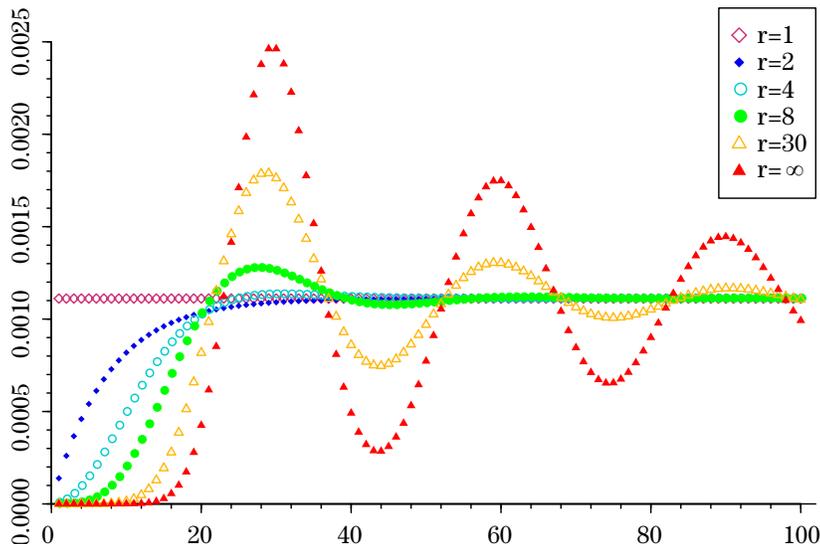


Figure 1. $\pi_{k, k+i}$ in function of i for negative binomial spacings with $\pi = 1/30$, $r = 1, 2, 4, 8, 30$ and $r = +\infty$ (Poisson spacings). When $r = 1$, we obtain the Bernoulli sampling design and a flat line on the plot. Oscillations are stronger for larger values of r .

pling when $r > 1$ and in that case there is a repulsion between selected units: the sample is spread more evenly on the population than if drawings were independent. On the contrary, if $r < 1$, there is an attraction between units and selecting neighbor units together is more likely.

The sum of j independent random variables with negative binomial distribution and parameters r, p , has a negative binomial distribution with parameters jr and p .

Proposition 2. *The second order inclusion probabilities of an equilibrium renewal chain sampling design with shifted negative binomial spacings, sampling rate π and first parameter r are*

$$\pi_{k\ell} = \pi \sum_{j=1}^{\ell-k} \frac{\Gamma(jr + \ell - k - j)}{(\ell - k - j)! \Gamma(jr)} p^{jr} (1 - p)^{\ell - k - j}, \quad k < \ell,$$

where $p = r\pi / (r\pi + 1 - \pi)$.

Proof. Since $J_i - 1$ has a $\mathcal{NB}(r, p)$ distribution, $(\sum_{i=1}^j J_i) - j$ has a $\mathcal{NB}(jr, p)$ distribution. Using (3.6) one gets the result.

These joint inclusion probabilities remain positive for any value of r . They are plotted in Figure 1 for $\pi = 1/30$ and different values of r .

In order to have an equilibrium renewal chain sampling design and equal first order inclusion probabilities, the first sample unit index has to be generated from a shifted forward negative binomial. We get that $J_0 - 1 \sim \text{For}\mathcal{NB}(r, p)$, where the definition of $\text{For}\mathcal{NB}(r, p)$ can be found in Table 4.

4.2. Poisson spacings

The limit of negative binomial distributions when r tends to infinity and p tends to 0, while keeping a constant expectation $\lambda = r(1 - p)/p$, is a Poisson distribution $\mathcal{P}(\lambda)$ with parameter λ , which is also its expectation and its variance.

We consider the equilibrium renewal chain sampling design with shifted Poisson spacings: $J_i - 1 \sim \mathcal{P}(\lambda)$, where $\lambda = (1 - \pi)/\pi$, $i \in \mathbb{N}^*$. The first inter arrival J_0 is selected using a shifted forward Poisson distribution: $J_0 - 1 \sim \text{For}\mathcal{P}(\lambda)$, where the definition of distribution $\text{For}\mathcal{P}(\lambda)$ can be found in Table 4.

Proposition 3. *The second order inclusion probabilities of an equilibrium renewal chain sampling design with shifted Poisson spacings and sampling rate π are, with $\lambda = (1 - \pi)/\pi$,*

$$\pi_{k\ell} = \pi \sum_{j=1}^{\ell-k} \frac{e^{-j\lambda} (j\lambda)^{\ell-k-j}}{(\ell - k - j)!}, \quad k < \ell.$$

4.3. Binomial spacings

The variance of spacings in Sections 4.1 and 4.2 are bounded from below, by $(1 - \pi)/\pi$. However, to get a sample spread close to that of systematic sampling, we need to be able to have a variance that is arbitrarily close to 0. For this, we consider the equilibrium renewal chain sampling design that is obtained with shifted binomial spacings: $J_i - 1 \sim \text{Bin}(r, p)$, $i \in \mathbb{N}^*$, $r \in \mathbb{N}$, $p \in [0, 1]$. The first inter-arrival J_0 is selected using a shifted forward binomial distribution: $J_0 - 1 \sim \text{For}\text{Bin}(r, p)$, where the definition of distribution $\text{For}\text{Bin}(r, p)$ can be found in Table 4.

We find that with a sampling rate equal to π , r must necessarily be greater or equal to $(1 - \pi)/\pi$, and that

$$p = \frac{1}{r} \left(\frac{1 - \pi}{\pi} \right).$$

The variance of spacings is then

$$\text{var}(J_i) = \frac{1 - \pi}{\pi} - \frac{1}{r} \left(\frac{1 - \pi}{\pi} \right)^2, \quad i \in \mathbb{N}^*. \tag{4.2}$$

Considering the constraints on r and p , this variance is minimal when r is the

Table 1. Renewal chain sampling designs and variance of their spacings.

Distribution of $J_1 - 1$	$\text{var}(J_1)$
Negative binomial, $r > 0$	$\frac{1 - \pi}{\pi} + \frac{1}{r} \left(\frac{1 - \pi}{\pi} \right)^2$
Poisson	$\frac{1 - \pi}{\pi}$
Binomial, $r \geq (1 - \pi)/\pi$	$\frac{1 - \pi}{\pi} - \frac{1}{r} \left(\frac{1 - \pi}{\pi} \right)^2$
Systematic or binomial with $r = (1 - \pi)/\pi$	0

smallest integer that is greater or equal to $(1 - \pi)/\pi$. With this r , the variance of spacings is always smaller than 1, which is really small for an integer valued random variable with a usually very large expectation $1/\pi$. When $1/\pi$ is an integer number, the variance of spacings is null when $r = (1 - \pi)/\pi$ and $p = 1$. The sampling design obtained then is just the systematic sampling design.

If r tends to infinity and $p = (1 - \pi)/r\pi$, the binomial distribution with parameters r and p converges in distribution toward the Poisson distribution with parameter $(1 - \pi)/\pi$. Hence, the sampling design of Section 4.2 is also the limiting case of Binomial spacings renewal chain sampling design when r tends to infinity.

Proposition 4. *The second order inclusion probabilities of an equilibrium renewal chain sampling design with shifted binomial spacings, sampling rate π and first parameter r are equal to*

$$\pi_{k\ell} = \pi \sum_{j=1}^{\ell-k} \binom{j r}{\ell - k - j} p^{\ell-k-j} (1-p)^{j(r+1)-\ell-k}, \quad k < \ell,$$

where $p = (1 - \pi)/r\pi$.

4.4. Summary

The different renewal chain sampling designs we considered are listed in Table 1 with the variance of their spacings. If $1/\pi$ is not an integer, the variance of spacings cannot be null and is at least $(\lceil 1/\pi \rceil - 1/\pi)(1/\pi - \lfloor 1/\pi \rfloor)$. This lower bound is not reached with shifted binomial spacings but the binomial renewal chain sampling design enjoys the desirable property of having positive joint inclusion probabilities. Other spacing distributions can be used but we retained only common families of distribution that have useful properties such as stability under convolution.

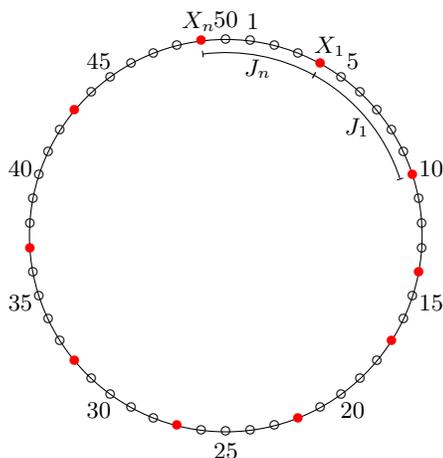


Figure 2. Illustration of a sample of 10 units, in a population of 50 units. The selected units are in red.

5. Fixed Size Sampling Designs with Exchangeable Circular Spacings

Except in very special situations, renewal chain sampling designs do not have fixed sample size. This is due to the independence of spacings. However, in many applications fixed size is required. In this section, we propose to define sampling designs using exchangeable instead of independent spacings. We obtain fixed size designs with equal inclusion probabilities, and we are able to control the sample spread by the choice of the random spacings distribution.

5.1. Circular spacings

A sampling design of fixed size n in a population $U = \{1, \dots, N\}$ is entirely specified by the joint distribution of one of the unit indexes, e.g. X_1 , and the “circular” spacings $J_i = X_{i+1} - X_i$, $i = 1, \dots, n - 1$, and $J_n = N + X_1 - X_n$, where X_1 is the smallest sample unit index and X_n the largest. If we represent the population U around a table, as in Figure 2, the J_i ’s are the difference of units position. Note that considering the population as circular is not new in survey sampling and goes back at least to Fuller (1970).

We intend to work with Equation (5.1) that defines without loss of generality the random sample S of a fixed size sampling design,

$$S = \{S_j \pmod N, j = 1, \dots, n\}, \tag{5.1}$$

where J_1, \dots, J_n are positive integer random variables that sum to N , J_0 is a random variable in U , and

$$S_j = \sum_{i=0}^j J_i.$$

5.2. First order inclusion probabilities

The first order inclusion probabilities of a sampling design that results from Equation (5.1) depend on the joint distribution of the J_i 's, $i = 0, \dots, n$. Intuitively, one sees that, for a given joint distribution of J_1, \dots, J_n with $S_n = N$, choosing J_0 to be independent of the other J_i 's and uniform on $\{1, \dots, N\}$ allows to obtain equal first order inclusion probabilities. To prove this assertion, one can compute in the general case the inclusion probability of a unit k . Consider an independent J_0 , and let

$$f_0(t) = \Pr(J_0 = t), \quad f_j(k) = \Pr(S_j \pmod N = k), \quad t, k \in U, \quad 1 \leq j \leq n.$$

By conditioning on the event $\{J_0 = t\}$ and using the law of total probability on the disjoint events $\{k > t\}$, $\{k = t\}$ and $\{k < t\}$, we get that:

$$\pi_k = \sum_{t=1}^N f_0(t) \left[\mathbf{1}_{t=k} + \mathbf{1}_{t < k} \sum_{j=1}^{k-t} f_j(k-t) + \mathbf{1}_{t > k} \sum_{j=1}^{N+k-t} f_j(N+k-t) \right],$$

with the convention that $f_j(k) = 0$ if $j > n$. Hence we can write that vectors $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)$ and $\mathbf{f}_0 = (f_0(1), \dots, f_0(N))$ are solutions of the linear equation

$$\boldsymbol{\pi} = \mathbf{A}\mathbf{f}_0, \tag{5.2}$$

where \mathbf{A} is the square matrix of size N with general term

$$a_{kt} = \mathbf{1}_{t=k} + \mathbf{1}_{t < k} \sum_{j=1}^{k-t} f_j(k-t) + \mathbf{1}_{t > k} \sum_{j=1}^{N+k-t} f_j(N+k-t), \quad 1 \leq k, t \leq N. \tag{5.3}$$

It is not our purpose to solve the system and find designs with any given inclusion probabilities, especially since solutions depend on the $f_j(k)$'s, but one arrives rapidly to the result that all the lines of \mathbf{A} sum to n (see Proposition 7 in Appendix). Hence, if $f_0(t) = 1/N$ for all t , then the inclusion probabilities are all equal to n/N .

5.3. Joint inclusion probabilities

Let $\ell > k$ in $\{1, \dots, N\}$, and consider $\Pr(\ell \in S | k \in S)$ so that, by definition, $\pi_{k\ell} = \pi_k \Pr(\ell \in S | k \in S)$. Knowing that $k \in S$, the event $\ell \in S$ can be decomposed according to the number of units selected between k and ℓ :

$$\pi_{k\ell} = \pi_k \sum_{j=1}^{\ell-k} \Pr(\ell \in S \text{ and } \#S \cap \{k+1, \dots, \ell\} = j | k \in S). \tag{5.4}$$

The term $\Pr(\ell \in S \text{ and } \#S \cap \{k+1, \dots, \ell\} = j | k \in S)$ is usually difficult to compute: one must decompose according to which S_i is equal to k . However, if the joint distribution of J_1, \dots, J_n has some additional properties, as in Proposition 5, one can obtain a simple expression.

Proposition 5. *Consider a positive integer random vector (J_1, \dots, J_n) that sums to N and such that the distributions of any sum of k successive J_i 's are equal, this condition also holding for the “circular” sums of J_{n-i} up to J_n and J_1 up to J_{k-i} if $i < k$. Then, the second order inclusion probabilities are*

$$\pi_{k\ell} = \pi_k \sum_{j=1}^{\ell-k} f_j(\ell - k), \quad k < \ell. \tag{5.5}$$

Proof. Indeed, we then have that:

$$\Pr(\ell \in S \text{ and } \#S \cap \{k+1, \dots, \ell\} = j | k \in S) = f_j(\ell - k), \quad j = 1, \dots, \ell - k,$$

and the result follows immediately.

In the situation of Proposition 5, we also get that the conditional inclusion probability $\Pr(\ell \in S | k \in S)$ is a function of $\ell - k \pmod N$:

$$\Pr(\ell \in S | k \in S) = \sum_{j=1}^{\ell-k} f_j(\ell - k), \text{ if } k < \ell,$$

$$\Pr(\ell \in S | k \in S) = \sum_{j=1}^{N+\ell-k} f_j(N + \ell - k), \text{ if } \ell < k.$$

It also follows in that case that if the first order inclusion probabilities are all equal, for example when J_0 has a uniform distribution, then the joint inclusion probabilities $\pi_{k\ell}$ depend only on $\ell - k \pmod N$.

Actually, all distributions considered for spacings J_1, \dots, J_n in this paper enjoy a stronger property, they are exchangeable distributions (Aldous (1985), Kallenberg (2005)).

Definition 4. *A family J_1, \dots, J_n of random variables is said to be exchangeable if, for all $1 \leq k \leq n$ and permutation σ of $\{1, \dots, n\}$, the joint distribution of $(J_{\sigma(1)}, \dots, J_{\sigma(k)})$ is equal to the joint distribution of (J_1, \dots, J_k) . If the J_i 's are discrete distributions, this is equivalent to say that $\Pr(J_1 = a_1, \dots, J_n = a_n)$ is a symmetric function of (a_1, \dots, a_n) .*

Exchangeable integer distributions J_1, \dots, J_n that sum to N clearly satisfy the conditions of Proposition 5. They are the natural equivalents to i.i.d. spacing distributions used in Section 3 when the spacings are constrained, by the fixed sample size, to sum to N .

Definition 5. *Fixed size sampling designs with exchangeable circular spacings and uniform inclusion probabilities are the sampling designs with random samples $S = \{S_j \pmod{N}, j = 1, \dots, n\}$ where $S_j = \sum_{i=0}^j J_i$, J_0 is a uniform random distribution on $\{1, \dots, N\}$ independent from $\mathbf{J} = (J_1, \dots, J_n)$, and the J_i 's, $i = 1, \dots, n$, are exchangeable positive integer distributions that sum to N .*

The PMF of a fixed size sampling design with exchangeable circular spacings and uniform inclusion probabilities J_1, \dots, J_n is given simply by

$$P(s) = \frac{n}{N} \Pr(J_1 = x_2 - x_1, \dots, J_{n-1} = x_n - x_{n-1}, J_n = N + x_1 - x_n), \quad (5.6)$$

where x_1, \dots, x_n are the ordered indexes of units sampled in s . Indeed, $P(s)$ can be decomposed according to the value of J_0 into:

$$\begin{aligned} & \Pr(S = \{x_1, \dots, x_n\}) \\ &= \Pr(J_0 = x_1) \Pr(J_1 = x_2 - x_1, \dots, J_{n-1} = x_n - x_{n-1}, J_n = N + x_1 - x_n) \\ & \quad + \Pr(J_0 = x_2) \Pr(J_1 = x_3 - x_2, \dots, J_{n-1} = N + x_1 - x_n, J_n = x_2 - x_1) \\ & \quad \vdots \\ & \quad + \Pr(J_0 = x_n) \Pr(J_1 = N + x_1 - x_n, \dots, J_n = x_n - x_{n-1}), \end{aligned}$$

and the $\Pr(J_0 = x_i)$ are all equal to $1/N$ while the probabilities involving J_1, \dots, J_n are all equal due to the exchangeability of the circular spacings.

5.4. Simple Random Sampling

The Simple Random Sampling (SRS) without replacement design of fixed sample size n is defined by:

$$P(s) = \binom{N}{n}^{-1} \text{ if } \#s = n, \text{ and } P(s) = 0 \text{ otherwise.}$$

A SRS sample can be selected using the following algorithm (Fan, Muller and Rezucha (1962), see also Tillé (2006, p.46)): define a counter $j = 0$, then, for $k = 1$ to N , select unit k with probability $(N-j)/(N-k-1)$ and update $j = j+1$ if k is selected. It is also possible to obtain this design by generating successive jumps according to negative hypergeometric distributions with parameters that depend on the previously selected units (see Vitter (1984, 1985, 1987)).

Proposition 6 asserts that SRS is a sampling design with exchangeable cir-

cular spacings, where the spacings follow a shifted multivariate negative hypergeometric distribution. The (singular) multivariate negative hypergeometric distribution (see for example Johnson, Kotz and Balakrishnan (1997, pp.171-199)) of size $n \geq 1$, with parameters $m \in \mathbb{N}$ and $\mathbf{r} = (r_1, \dots, r_n)$, $r_i > 0$, $i = 1, \dots, n$ is a probability distribution on integer vectors (x_1, \dots, x_n) that sum to m . It is denoted here by $\mathcal{MNH}(m, \mathbf{r})$, and has a PMF given by:

$$f_{\mathcal{MNH}(m, \mathbf{r})}(x_1, \dots, x_n) = \frac{m! \Gamma(R)}{\Gamma(m + R)} \prod_{i=1}^n \frac{\Gamma(r_i + x_i)}{\Gamma(r_i) x_i!}, \tag{5.7}$$

where $R = \sum_{i=1}^n r_i$.

Proposition 6. *SRS is the sampling design of (5.1), where J_0 has a uniform distribution on U , is independent of the J_i 's, $i \geq 1$, and the integer random vector $\mathbf{J} = (J_1, \dots, J_n)$ follows a shifted multivariate negative hypergeometric distribution: $\mathbf{J} - \mathbf{1}_n \sim \mathcal{MNH}(N - n, \mathbf{1}_n)$, where $\mathbf{1}_n$ is the n -vector of ones.*

Proof. With parameter $N - n$ and $\mathbf{1}_n$, the PMF given in (5.7) reduces to

$$f_{\mathcal{MNH}(N-n, \mathbf{1}_n)}(x_1, \dots, x_n) = \binom{N-1}{n-1}^{-1},$$

where x_1, \dots, x_n are non-negative integers that sum to $N - n$. Hence, \mathbf{J} has a uniform distribution on the vectors of positive integer numbers that sum to N ,

$$\Pr[\mathbf{J} = (j_1, \dots, j_n)] = \binom{N-1}{n-1}^{-1},$$

for all positive integers (j_1, \dots, j_n) that sum to N . Moreover, this PMF is symmetric in its arguments and the J_i 's are exchangeable. Applying (5.6), we get that

$$\Pr(S = \{x_1, \dots, x_n\}) = \frac{n}{N} \binom{N-1}{n-1}^{-1} = \binom{N}{n}^{-1},$$

for all $x_1 < \dots < x_n$.

The marginal distributions of the circular spacings are shifted negative hypergeometric distributions. Indeed, the marginal distributions of a $\mathcal{MNH}(m, \mathbf{r})$ -distributed vector are negative hypergeometric distributions (see for example Janardan and Patil (1972)) with respective parameters m , r_i , $R = \sum_j r_j$, and PMF:

$$f_{\mathcal{NH}(m, r_i, R)}(x) = \frac{m! \Gamma(R)}{\Gamma(m + R)} \frac{\Gamma(r_i + x)}{\Gamma(r_i) x!} \frac{\Gamma(R - r_i + m - x)}{\Gamma(R - r_i) (m - x)!}, \quad x \leq m, \quad x \in \mathbb{N}.$$

Their expectation and variance are, respectively, mr_i/R and $m(r_i/R)(1-r_i/R)(R+m)/(R+1)$. It follows that $J_k - 1$ has a negative hypergeometric distribution

with parameters $N - n$, 1 , and n , $k = 1, \dots, n$. In particular we have that

$$\begin{aligned} \mathbb{E}(J_k) &= \frac{N - n}{n} + 1 = \frac{N}{n}, \\ \text{var}(J_k) &= \frac{N - n}{n} \left(1 - \frac{1}{n}\right) \frac{N}{n + 1}. \end{aligned}$$

The second order inclusion probabilities can be derived from (5.5). Indeed, the sum of components of a multivariate negative hypergeometric distribution follows a negative hypergeometric distribution (see Janardan and Patil (1972)). Its parameters are derived from the parameters m and \mathbf{r} by summing the r_i 's that correspond to the components that are in the sum. Hence we have that

$$\sum_{i=1}^j J_i = j + K_j,$$

where K_j follows a negative hypergeometric distribution with parameters $N - n$, j and n . We can deduce that

$$\begin{aligned} f_j(\ell - k) &= \frac{(N - n)! \Gamma(n)}{\Gamma(N)} \frac{\Gamma(j + \ell - k - j)}{\Gamma(j)(\ell - k - j)!} \frac{\Gamma(n - j + N - n - \ell + j + k)}{\Gamma(n - j)(N - n - \ell + j + k)!} \\ &= \frac{(N - n)!(n - 1)!(\ell - k - 1)!(N - \ell + k - 1)!}{(N - 1)!(j - 1)!(\ell - k - j)!(n - j - 1)!(N - n - \ell + k + j)!} \\ &= \frac{n - 1}{N - 1} \binom{N - 2}{\ell - k - 1}^{-1} \binom{n - 2}{j - 1} \binom{N - n}{\ell - k - j}, \end{aligned}$$

and that

$$\pi_{k\ell} = \frac{n(n - 1)}{N(N - 1)} \sum_{j=1}^{\ell - k} \binom{N - 2}{\ell - k - 1}^{-1} \binom{n - 2}{j - 1} \binom{N - n}{\ell - k - j}, \quad k < \ell,$$

via (5.5). However, if we rename $u = j - 1$, $v = \ell - k - 1$, $t = n - 2$ and $s = N - 2$, this last sum is

$$\binom{s}{v}^{-1} \sum_{u=0}^v \binom{t}{u} \binom{s - t}{v - u},$$

and Vandermonde's identity ensures that it is equal to 1. Hence we find the well known result:

$$\pi_{k\ell} = \frac{n(n - 1)}{N(N - 1)}, \quad k \neq \ell.$$

5.5. Systematic sampling

If $N = rn$ with $r \in \mathbb{N}$, the systematic sampling design presented in Section 3.6 is a fixed size sampling design with exchangeable circular spacings. It is trivially obtained by taking J_0 uniform on U and $J_i = r$, $i = 1, \dots, n$.

The joint inclusion probabilities can also easily be derived from (5.5) using that $f_j(\ell - k) = \mathbf{1}_{\{\ell=k+jr\}}$.

6. Spreading Fixed Size Sampling Designs with Exchangeable Circular Spacings

Similar to what we did in Section 4, we introduce in Sections 6.1, 6.2 and 6.3 new sampling designs with spreading properties by choosing different circular spacings distributions.

Following the structure of Section 4, we work, in Section 6.1, on sampling designs with multivariate negative hypergeometric spacings and a spreading control parameter $r > 0$. When $0 < r < 1$, there is an attraction between the selected units: if a unit is selected, then its neighbors are more likely to be selected. If $r = 1$, the design is SRS and if $r > 1$, the sampling is better spread than SRS. As a limit case when r is large, we obtain the multinomial circular spacings design of Section 6.2. The spacings variance of these sampling designs is bounded from below. Smaller variances and better spreading properties are obtained with multivariate hypergeometric circular spacings in Section 6.3, furthering the parallel with binomial spacings of Section 4.3.

6.1. Multivariate negative hypergeometric circular spacings

The multivariate negative hypergeometric distribution $\mathcal{MNH}(m, \mathbf{r})$ has exchangeable marginals exactly when $r_1 = \dots = r_n$, $\mathbf{r} = r\mathbf{1}_n$ for some positive real number r . If $\mathbf{J} - \mathbf{1}_n \sim \mathcal{MNH}(N - n, r\mathbf{1}_n)$, the sampling design of Definition 5 has circular spacings with a variance given by:

$$\text{var}(J_k) = \frac{N - n}{n} \left(1 - \frac{1}{n}\right) \frac{rn + N - n}{rn + 1}, \quad k = 1, \dots, n.$$

These variances are decreasing functions of r , with SRS corresponding to $r = 1$.

According to (5.6), the sampling design PMF is given by:

$$P(s) = \frac{n}{N} \frac{\Gamma(nr)}{[\Gamma(r)]^n} \frac{(N - n)!}{\Gamma[N + n(r - 1)]} \frac{\Gamma(r + N + x_1 - x_n - 1)}{(N + x_1 - x_n - 1)!} \prod_{i=1}^{n-1} \frac{\Gamma(r + x_{i+1} - x_i - 1)}{(x_{i+1} - x_i - 1)!},$$

where x_1, \dots, x_n are the ordered indexes of units sampled in s . The second order inclusion probabilities are obtained from (5.5):

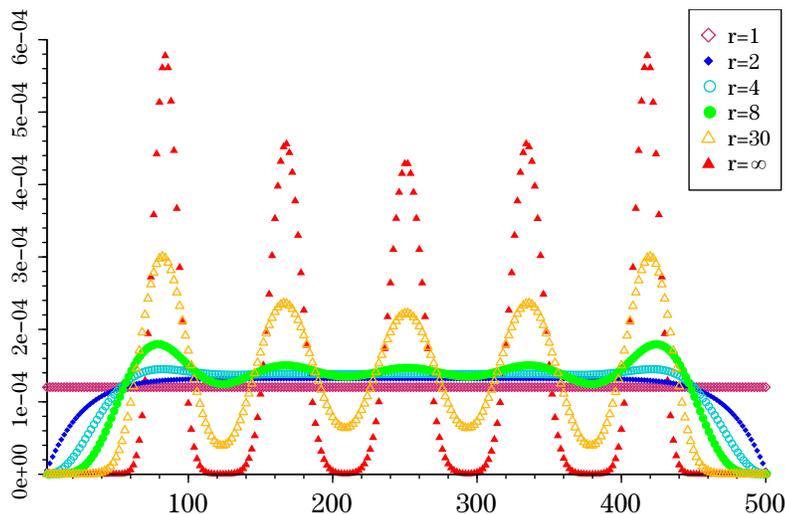


Figure 3. $\pi_{1,1+i}$ in function of i for a fixed size sampling design with shifted multivariate negative hypergeometric circular spacings, $n = 6$, $N = 500$ and $r = 1, 2, 4, 8, 30$ and $r = +\infty$ (multinomial circular spacings). When $r = 1$, we obtain the SRS design and constant joint inclusion probabilities. The larger r is, the more contrasted are the joint inclusion probabilities.

$$\pi_{k\ell} = \frac{n}{N} \sum_{j=1}^{\ell-k} \binom{N-n}{\ell-k-j} \frac{B[\ell-k+j(r-1), N+n(r-1)-\ell+k-j(r-1)]}{B(jr, nr-jr)},$$

$$k < \ell,$$

where $B(\cdot, \cdot)$ denotes the beta function, defined by:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1} dt,$$

if a and b are positive real numbers.

These joint inclusion probabilities are plotted in Figure 3 for different values of r , including their limit when $r \rightarrow \infty$. On this plot, we see strong oscillations of the joint inclusion probabilities when r is large.

6.2. Multinomial circular spacings

Let $\mathcal{MN}_{\text{om}}(m, \mathbf{p})$ be the multinomial distribution with parameters $m \in \mathbb{N}$ and $\mathbf{p} = (p_1, \dots, p_n) \in [0, 1]^n$, $\sum_i p_i = 1$. It is the probability distribution on integer vectors (x_1, \dots, x_n) such that $\sum_i x_i = m$ with PMF:

$$f_{\mathcal{MN}_{\text{om}}(m, \mathbf{p})}(x_1, \dots, x_n) = \binom{m}{x_1 \dots x_n} \prod_{i=1}^n p_i^{x_i}, \tag{6.1}$$

where

$$\binom{m}{x_1 \cdots x_n} = \frac{m!}{x_1! \cdots x_n!}.$$

Its marginal distributions are binomial with respective parameters m and p_i . They are exchangeable exactly when $\mathbf{p} = n^{-1}\mathbf{1}_n$

When r tends to infinity, the multivariate negative hypergeometric distribution with parameters m and $r\mathbf{1}_n$ tends to a multinomial distribution with parameters m and $\mathbf{1}/n$ (see, for instance, Terrell (1999, p.182)). Thus the fixed size sampling design with shifted multinomial exchangeable circular spacings is the limit case of multivariate negative hypergeometric spacings of Section 6.1 when r tends to infinity.

Spacings J_k follow a shifted binomial distribution with parameters $N - n$ and $1/n$. The corresponding sampling design PMF is given by

$$P(s) = \frac{n}{N} \frac{1}{n^{N-n}} \frac{(N - n)!}{(N + x_1 - x_n - 1)! \prod_{j=2}^n (x_j - x_{j-1} - 1)!},$$

where x_1, \dots, x_n are the ordered indexes of units sampled in s . The sum of any j components of a $\mathcal{MN}(om(m, \mathbf{p}))$ multinomial vector follows a binomial distribution with parameters m and p where p is the sum of corresponding p_i 's. Hence we have here that

$$f_i(\ell - k) = \binom{N - n}{\ell - k - j} \left(\frac{j}{n}\right)^{\ell - k - j} \left(1 - \frac{j}{n}\right)^{N - n - \ell + k + j},$$

$$\pi_{k\ell} = \frac{n}{N} \sum_{j=1}^{\ell - k} \binom{N - n}{\ell - k - j} \left(\frac{j}{n}\right)^{\ell - k - j} \left(1 - \frac{j}{n}\right)^{N - n - \ell + k + j}, \quad k < \ell.$$

6.3. Multivariate hypergeometric circular spacings

Variances of circular spacings in Sections 6.1 and 6.2 are bounded from below. In order to have smaller variances, one can use shifted multivariate hypergeometric spacings.

Let $\mathcal{MH}(m, \mathbf{r})$ be the (singular) multivariate hypergeometric distribution with parameters $m \in \mathbb{N}$ and $\mathbf{r} = (r_1, \dots, r_n)$, \mathbf{r} is an integer vector that sums to R and $m \leq R$. It is a probability distribution on integer vectors (x_1, \dots, x_n) that sum to m , $x_i \leq r_i$ and has a PMF given by:

$$f_{\mathcal{MH}(m, \mathbf{r})}(x_1, \dots, x_n) = \binom{R}{m}^{-1} \prod_{i=1}^n \binom{r_i}{x_i}. \tag{6.2}$$

The marginal distributions of $\mathcal{MH}(m, \mathbf{r})$ are hypergeometric variables with re-

spective parameters m , r_i and R , and their variance is $m(r_i/R)(1 - r_i/R)(R - m)/(R - 1)$.

The multivariate hypergeometric distribution has exchangeable marginals exactly when $\mathbf{r} = r\mathbf{1}_n$ for some integer r larger than m/n . Here, we consider the fixed size sampling design with circular spacings \mathbf{J} such that $\mathbf{J} - \mathbf{1}_n \sim \mathcal{MH}(N - n, r\mathbf{1}_n)$ with $r \geq N/n - 1$. We get that the spacing variances are given by

$$\text{var}(J_k) = \frac{N - n}{n} \left(1 - \frac{1}{n}\right) \frac{rn - N + n}{rn - 1}.$$

The parameter r can be used to tune the variance. If $r = (N - n)/n$ is integer, we have $\text{var}(J_k) = 0$ and we obtain the systematic sampling design.

The sampling design PMF is obtained via (5.6):

$$P(s) = \frac{n}{N} \binom{rn}{N - n}^{-1} \binom{r}{x_{i+1} - x_i - 1} \prod_{i=1}^{n-1} \binom{r}{N + x_1 - x_n - 1},$$

where x_1, \dots, x_n are the ordered indexes of units sampled in s . The sum of components of a $\mathcal{MH}(m, \mathbf{r})$ vector follows a hypergeometric distribution. In the present case we find that

$$f_j(x) = \binom{rn}{N - n}^{-1} \binom{jr}{x - j} \binom{rn - jr}{N - n - x + j}, \quad j \leq x \leq N - n,$$

and that the joint inclusion probabilities are

$$\pi_{k\ell} = \frac{n}{N} \binom{rn}{N - n}^{-1} \sum_{j=1}^{\ell-k} \binom{jr}{\ell - k - j} \binom{rn - jr}{N - n - \ell + k + j}, \quad k < \ell.$$

Note that some joint inclusion probabilities may be null, even when $r > (N - n)/n$ and the design is not the systematic sampling design. For example, if $N = 10$, $n = 2$, $r = 5$ and $\ell = k + 1$, then

$$\pi_{k\ell} = \frac{1}{5} \binom{10}{8}^{-1} \binom{5}{0} \binom{5}{8} = 0.$$

6.4. Summary

The different fixed size sampling designs with exchangeable circular spacings that we considered are listed in Table 2 with the variance of their spacings. Other exchangeable circular spacings may be used, and are easily obtained as distributions of vectors of i.i.d. random variables conditioned on the sum of the vectors components. The families of distribution of Section 6 encompass the SRS and fixed-size systematic designs. They allow one to use designs with low spacings variance, but it is not always possible to avoid having null joint inclusion prob-

Table 2. Fixed size sampling designs and variance of their spacings.

Distribution of $\mathbf{J} - \mathbf{1}_n$	$\text{var}(J_i), i = 1, \dots, n$
Multivariate negative hypergeometric, $r > 0$	$\frac{N-n}{n} \left(1 - \frac{1}{n}\right) \frac{rn + N - n}{rn + 1}$
Multinomial	$\frac{N-n}{n} \left(1 - \frac{1}{n}\right)$
Multivariate hypergeometric, $r \geq N/n - 1$	$\frac{N-n}{n} \left(1 - \frac{1}{n}\right) \frac{rn - N + n}{rn - 1}$
Systematic or hypergeometric with $r = N/n - 1$	0

abilities with shifted multivariate hypergeometric spacings. Finally, the designs are not strictly sequential as the population list may need to be run over twice in order to finish selecting a sample.

7. Simulations

A single artificial population of $N = 200$ units was generated with an interest variable Y that had a trend and are autocorrelated, $y_k = k + z_k$, where $z_k = 0.6z_{k-1} + \epsilon_k$ and $\epsilon_k \sim N(0, \sigma_\epsilon = 0.3)$. With this kind of autocorrelation, having well spread samples ought to be an efficient strategy. The “spacing” $N + x_1 - x_n$ between the last sampled unit and the first one is treated like any other spacing, so that ideally one would also want to have some similarity between units at the beginning of the population list and those at the end. This feature can easily be obtained in a setting of continuous population sampling (see Wilhelm, Qualité and Tillé (2017)), but is not common in finite population applications.

For each situation, a set of 100,000 samples was generated. All samples were of fixed size $n = 50$ and were selected using the following sampling designs: Multivariate negative hypergeometric (MNH) with $r = 0.5, r = 1$ (SRS), $r = 5, r = 10, r = 50$; Multinomial (MULT); Multivariate hypergeometric (MH) with $r = 50, r = 10, r = 6, r = 4$. We used different values for the tunings parameter r in all kinds of sampling design in order to show the effect of this tuning parameter.

For each sample an estimate \widehat{Y} of the mean and of the variance $\widehat{\text{var}}_{SYG}(\widehat{Y})$ (with the Sen-Yates-Grundy formula) were produced. Compiling our simulation results, we computed the following values, presented in Table 3: the Bias Ratio

$$\text{BR} = 100 \frac{\text{E}_{sim}(\widehat{Y} - \bar{Y})}{\left[\text{var}_{sim}(\widehat{Y})\right]^{1/2}},$$

Table 3. Results of the 100,000 simulations. The designs are ordered in decreasing order of the variance of the spacings.

	BR	SE	REVAR	CV	Coverage
MNH $r = 0.5$	-0.25	0.46	0.45	0.48	93.97
SRS	-0.12	0.35	0.35	0.23	94.52
MNH $r = 5$	0.08	0.23	0.23	0.21	94.39
MNH $r = 10$	-0.22	0.21	0.21	0.26	94.08
MNH $r = 50$	0.13	0.19	0.19	0.33	93.90
MULT	0.36	0.19	0.19	0.35	93.64
MH $r = 50$	-0.17	0.18	0.18	0.37	93.58
MH $r = 10$	-0.35	0.16	0.16	0.52	92.05
MH $r = 6$	-0.74	0.14	0.14	0.72	83.97
MH $r = 4$	-0.52	0.11	0.15	1.60	40.55

where $E_{sim}(\cdot)$ and $\text{var}_{sim}(\cdot)$ denote the empirical means and variances of the simulation results; the standard error

$$\text{SE} = \left[\text{var}_{sim}(\widehat{Y}) \right]^{1/2};$$

the square root of the variance estimator average

$$\text{REVAR} = \left\{ E_{sim} \left[\widehat{\text{var}}_{SYG}(\widehat{Y}) \right] \right\}^{1/2};$$

the coefficient of variation of the variance estimator

$$\text{CV} = \frac{\left\{ \text{var}_{sim} \left[\widehat{\text{var}}_{SYG}(\widehat{Y}) \right] \right\}^{1/2}}{\text{var}_{sim}(\widehat{Y})};$$

and the coverage rate of the 95% confidence interval.

The simulation results in Table 3 confirm, with column BR, that the estimator of the mean is unbiased. The accuracy of the mean estimator improves as the circular spacings variance decreases, from Design MNH $r = 0.5$ to Design NH $r = 4$.

The conclusions are different for the variance estimator. For all situations in our simulations, the joint inclusion probabilities are positive. The variance estimator is unbiased, and this is confirmed by the fact that columns SE and REVAR are mostly equal. However, when the variance of the spacings are close to 0, the variance estimator is unstable. Indeed, with these parameters, some joint inclusion probabilities are very small (less than 1/1,000) compared to others (on average 0.0625). In column CV that the accuracy of the variance estimator improves at first as the circular spacings variance decreases, from Design MNH $r = 0.5$ to Design MNH $r = 5$, and then the coefficient of variation goes up again

from Design MNH $r = 0.5$ to Design MH $r = 4$. The coverage rate deviates strongly from its nominal value of 95% in the last couple of designs. Thus the design that performs best for the point estimation of the mean does not allow one to properly estimate the precision, and even gives seriously misleading confidence intervals. The same kind of problem arises when a stratified sampling design is used with too many strata.

An arbitration needs to be made between the accuracy of the point estimator and that of its variance estimator. In our simulations, a reasonable solution consists in choosing the sampling design with shifted multinomial distribution (MULT). This method is simple to implement, more so than the MNH or MH. It allows for accurate point estimation while presenting a correct coverage rate of its confidence intervals.

8. Conclusions

In Sections 3 and 5, we proposed general methods to generate uniform inclusion probabilities sampling designs with i.i.d. or exchangeable spacings. We used them in Sections 4 and 6 to obtain sample selection methods with controlled spreading properties and gave, in Section 7, an example where such methods are useful. If the response variable is similar among units that are close in the population list, the choice of the spreading parameter allows one to make a trade-off between precision of the point estimator and precision of variance estimator.

Some of the designs that we consider have concentrated spacings, but, unlike systematic sampling, they retain positive joint inclusion probabilities, and thus allow for an unbiased estimation of variance. These joint inclusion probabilities have computable closed-form expressions and depend only on the “distance” between units in the population list, thus at most $N - 1$ joint inclusion probabilities need to be computed. However, the ranks of sampled units in the population must be known in order to compute a variance estimator.

We do not have a clear solution to extending these results in all generality to unequal first order inclusion probabilities sampling designs. One partial solution is to work on the distribution of J_0 . The choice of a different distribution for J_0 allows one to have a limited control on the inclusion probabilities via Equations (3.4) and (5.2), while leaving the spacings untouched. Another solution is the thinning approach. It consists in selecting a large enough first phase sample with a spreading design and uniform inclusion probabilities and selecting a second phase sub-sample with appropriate inclusion probabilities. However, this

does not preserve the spreading properties.

Supplementary Materials

Some of the proofs, as well as a Table containing the definitions of some distributions, are given in the Supplementary Materials.

Acknowledgment

This work was supported in part by the Swiss Federal Statistical Office. The views expressed in this paper are solely those of the authors. M. W. was partially supported by a Doc.Mobility fellowship of the Swiss National Science Foundation (grant no. P1NEP2_162031). The authors would like to thank three referees and an associate editor for their constructive comments that helped us improve this paper, and Prof. Lennart Bondesson who has kindly sent us copies of his work on a similar topic.

Appendix A: Proof of Proposition 1 and Remark 1

Lemma 1. *If $f(\cdot)$ is a probability distribution on $\{1, 2, \dots\}$ with cumulative distribution function $F(\cdot)$, and $k, j \geq 1$, then*

$$\sum_{t=1}^k f^{(j+1)*}(t) = \sum_{t=1}^k f^{j*}(t)F(k-t).$$

Proof. Indeed, if $\mathbf{1}_A$ is the indicator function of set A ,

$$\begin{aligned} \sum_{t=1}^k f^{(j+1)*}(t) &= \sum_{t=1}^k \sum_{u=1}^t f^{j*}(u)f(t-u) \\ &= \sum_t \sum_u f^{j*}(u)f(t-u)\mathbf{1}_{\{1 \leq u \leq t\}} \mathbf{1}_{\{1 \leq t \leq k\}} \\ &= \sum_u \sum_t f^{j*}(u)f(t-u)\mathbf{1}_{\{1 \leq u \leq k\}} \mathbf{1}_{\{1 \leq u \leq t\}} \mathbf{1}_{\{1 \leq t \leq k\}} \\ &= \sum_u f^{j*}(u)\mathbf{1}_{\{1 \leq u \leq k\}} \left[\sum_t f(t-u) \mathbf{1}_{\{1 \leq u \leq t\}} \mathbf{1}_{\{1 \leq t \leq k\}} \right] \\ &= \sum_{u=1}^k f^{j*}(u) \left[F(k-u) - \underbrace{F(0)}_{=0} \right] \\ &= \sum_{t=1}^k f^{j*}(t)F(k-t). \end{aligned}$$

Proof of Proposition 1. $f_0(\cdot)$ is a well-defined non-negative function on \mathbb{N} . It is sufficient to prove that $\sum_{k \geq 0} f(\{k + 1, \dots\}) = \mu$, but

$$\begin{aligned} \sum_{k \geq 0} f(\{k + 1, \dots\}) &= \sum_{k \geq 0} \sum_{j \geq k+1} f(j) \\ &= \sum_{j \geq 0} \sum_{k \geq 0} f(j) \mathbf{1}_{k+1 \leq j} \\ &= \sum_{j \geq 0} j \cdot f(j) \\ &= \mu. \end{aligned}$$

As $f_0(k - t) = [1 - F(k - t)] / \mu$, to prove (3.3), it is sufficient to note that

$$\begin{aligned} &\sum_{t=1}^k [1 - F(k - t)] \sum_{j=1}^t f^{j*}(t) \\ &= \sum_t \sum_j [1 - F(k - t)] f^{j*}(t) \mathbf{1}_{1 \leq t \leq k} \mathbf{1}_{1 \leq j \leq t} \\ &= \sum_j \sum_t [1 - F(k - t)] f^{j*}(t) \mathbf{1}_{1 \leq t \leq k} \mathbf{1}_{1 \leq j \leq t} \\ &= \sum_j \left[\sum_t f^{j*}(t) \mathbf{1}_{1 \leq t \leq k} \mathbf{1}_{1 \leq j \leq t} - \sum_t F(k - t) f^{j*}(t) \mathbf{1}_{1 \leq t \leq k} \mathbf{1}_{1 \leq j \leq t} \right] \\ &= \sum_{j=1}^k \left[\sum_{t=j}^k f^{j*}(t) - \sum_{t=j}^k F(k - t) f^{j*}(t) \right] \\ &= \sum_{j=1}^k \left[\sum_{t=1}^k f^{j*}(t) - \sum_{t=1}^k F(k - t) f^{j*}(t) \right] \quad (\text{indeed, } f^{j*}(t) = 0 \text{ if } t < j) \\ &= \sum_{t=1}^k f^{1*}(t) - \sum_{t=1}^k F(k - t) f^{k*}(t) \text{ via Lemma 1} \\ &= F(k) - \sum_{t=1}^k f^{(k+1)*}(t) = F(k), \end{aligned}$$

since $f^{(k+1)*}(t) = 0$ if $t \leq k$, and the result follows immediately.

Proof of Remark 1. Consider X a random variable on \mathbb{N} with finite moment of order $m + 1$, $E(X^{m+1})$, $m \geq 0$, and its forward transform X_F according to Definition 2. Then we can write:

$$\sum_{k \geq 0} k^m \Pr(X_F = k) = \sum_{k \geq 0} k^m \frac{\Pr(X \geq k)}{E(X + 1)} = \sum_{k \geq 0} \sum_{i \geq k} \frac{k^m \Pr(X = i)}{E(X + 1)}$$

$$\begin{aligned}
&= \frac{1}{\mathbb{E}(X+1)} \sum_{i \geq 0} \sum_{k \geq 0} \mathbf{1}_{k \leq i} k^m \Pr(X=i) \\
&= \frac{1}{\mathbb{E}(X+1)} \sum_{i \geq 0} \left(\sum_{k=0}^i k^m \right) \Pr(X=i) \\
&= \frac{\mathbb{E}[F_m(X)]}{\mathbb{E}(X+1)},
\end{aligned}$$

where $F_m(x) = \sum_{k=0}^x k^m$.

Appendix B

Proposition 7. *The lines of matrix \mathbf{A} with general term a_{kt} given at (5.3) all sum to n .*

Proof. We have

$$a_{kt} = \mathbf{1}_{t=k} + \mathbf{1}_{t < k} \sum_{j=1}^{k-t} f_j(k-t) + \mathbf{1}_{t > k} \sum_{j=1}^{N+k-t} f_j(N+k-t),$$

with $f_j(t) = 0$ if $j < t$, $t \leq 1$, $t > N$ or $j > n$. We also have that $f_n(N) = 1$ and $f_j(N) = 0$ if $j < n$. The conclusion follows from

$$\begin{aligned}
&\sum_{t=1}^N \mathbf{1}_{t < k} \sum_{j=1}^{k-t} f_j(k-t) \\
&= \sum_{t=1}^N \sum_{j=1}^N f_j(k-t) \mathbf{1}_{j \leq k-t} \mathbf{1}_{j \leq n} \\
&= \sum_{j=1}^n \sum_{t=1}^N f_j(k-t) = \sum_{j=1}^n \Pr(S_j \leq k-1), \text{ and} \\
&\sum_{t=1}^N \mathbf{1}_{t > k} \sum_{j=1}^{N+k-t} f_j(N+k-t) \\
&= \sum_{t=1}^N \sum_{j=1}^N f_j(N+k-t) \mathbf{1}_{j \leq N+k-t} \mathbf{1}_{j \leq n} \mathbf{1}_{t > k} \\
&= \sum_{j=1}^n \sum_{t=1}^N f_j(N+k-t) \mathbf{1}_{t > k} = \sum_{j=1}^n [\Pr(S_j \geq k) - f_j(N)].
\end{aligned}$$

Table 4. Discrete distributions of probability.

Name	Notation	PMF	Support	Parameters	Mean	Variance
Bernoulli	$\mathcal{B}ern(p)$	$p^x(1-p)^{1-x}$	$\{0, 1\}$	$p \in [0, 1]$	p	$p(1-p)$
Forward Bernoulli	$\mathcal{F}or\mathcal{B}ern(p)$	$\frac{p^x}{p+1}$	$\{0, 1\}$	$p \in [0, 1], n \in \mathbb{N}$	(see below the table)	(see below the table)
Binomial	$\mathcal{B}in(n, p)$	$\binom{n}{x} p^x (1-p)^{n-x}$	$\{0, \dots, n\}$	$p \in [0, 1], n \in \mathbb{N}$	np	$np(1-p)$
Forward Binomial	$\mathcal{F}or\mathcal{B}in(n, p)$	$\frac{I_p(x, n-x+1)}{np+1}$	$\{0, \dots, n\}$	$p \in [0, 1], n \in \mathbb{N}$	(see below the table)	(see below the table)
Geometric	$\mathcal{G}(1-p)$	$p(1-p)^x$	\mathbb{N}	$p \in [0, 1]$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Negative Binomial	$\mathcal{N}\mathcal{B}(r, p)$	$\frac{\Gamma(r+x)}{x! \Gamma(r)} p^r (1-p)^x$	\mathbb{N}	$p \in [0, 1], r \in \mathbb{N}^*$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$
Forward Negative Binomial	$\mathcal{F}or\mathcal{N}\mathcal{B}(r, p)$	$\frac{p I_{(1-p)}(x, r)}{r(1-p) + p}$	\mathbb{N}	$p \in [0, 1], r \in \mathbb{N}^*$	(see below the table)	(see below the table)
Poisson	$\mathcal{P}(\lambda)$	$\frac{e^{-\lambda} \lambda^x}{x!}$	\mathbb{N}	$\lambda \in \mathbb{R}_+$	λ	λ
Forward Poisson	$\mathcal{F}or\mathcal{P}(\lambda)$	$\frac{1}{\lambda+1} \left[\mathbf{1}_{x=0} + \frac{\gamma(x, \lambda)}{(x-1)!} \mathbf{1}_{x \geq 1} \right]$	\mathbb{N}	$\lambda \in \mathbb{R}_+$	(see below the table)	(see below the table)
Hypergeometric	$\mathcal{H}(m, r, R)$	$\frac{\binom{r}{x} \binom{R-r}{m-x}}{\binom{R}{m}}$	$\{0, \dots, m\} \cap \{r+m-R, \dots, r\}$	$m, r, R \in \mathbb{N}^*, m, r \leq R$	$\frac{mr}{R}$	$\frac{mr(R-r)}{R^2} \frac{R-m}{R-1}$
Negative Hypergeometric	$\mathcal{N}\mathcal{H}(m, r, R)$	$\frac{\Gamma(r+x) \Gamma(R-r+m-x)}{\Gamma(r)x! \Gamma(R-r)(m-x)!}$	$\{0, \dots, m\}$	$m, r, R \in \mathbb{N}^*, 1 \leq R-r$	$\frac{mr}{R}$	$\frac{mr(R-r)}{R^2} \frac{R+m}{R+1}$
Uniform	$\mathcal{U}(0, a)$	$\frac{1}{a+1}$	$\{0, \dots, a\}$	$a \in \mathbb{N}$	$\frac{a}{2}$	$\frac{(a+1)^2 - 1}{12}$

Expectations and variances of forward distributions are easily computed in function of the first three moments of the original distribution (see Remark 1).

Appendix C: Discrete probability distributions

Let \mathbb{R}_+ denote the set of positive real numbers,

$$\Gamma(r, x) = \int_x^{+\infty} t^{r-1} e^{-t} dt, \quad \gamma(r, x) = \int_0^x t^{r-1} e^{-t} dt,$$

where $r > 0, x > 0$ and

$$B_x(a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt, \quad I_x(a, b) = \frac{B_x(a, b)}{B(a, b)},$$

with $a > 0, b > 0, 0 < x < 1$.

References

- Aldous, D. (1985). Exchangeability and related topics. In Hennequin, P., editor, *cole d't de Probabilits de Saint-Flour XIII 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer Berlin Heidelberg.
- Barbu, V. and Limmios, N. (2008). *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis*. Springer, New York.
- Bellhouse, D. R. (1988). Systematic sampling. In Krishnaiah, P. R. and Rao, C. R., editors, *Handbook of Statistics Volume 6: Sampling*, pages 125–145, Amsterdam. Elsevier/North-Holland.
- Bellhouse, D. R. and Rao, J. N. K. (1975). Systematic sampling in the presence of a trend. *Biometrika* **62**(3), 694–697.
- Bellhouse, D. R. and Sutradhar, B. C. (1988). Variance estimation for systematic sampling when autocorrelation is present. *The Statistician* **37**(3), 327–332.
- Bondesson, L. (1986). Sampling of linearly ordered population by selection of units at successive random distances. Technical Report 25, Swedish University of agricultural sciences, Section of forest biometry, Umeå.
- Bondesson, L. and Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics* **35**(3), 466–483.
- Chauvet, G. (2012). On a characterization of ordered pivotal sampling. *Bernoulli* **18**(4), 1320–1340.
- Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of population. *Annals of Mathematical Statistics* **17**(2), 164–177.
- Cox, D. R. (1962). *Renewal Theory*. Methuen, London.
- Daley, D. and Vere-Jones, D. (2002). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, New York.
- Deville, J.-C. (1998). Une nouvelle (encore une!) méthode de tirage à probabilités inégales. Technical Report 9804, Méthodologie Statistique, INSEE, Paris.
- Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**(1), 89–101.
- Fan, C. T., Muller, M. E. and Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *Journal of the American Statistical Association* **57**(298), 387–402.

- Feller, W. (1971). *An introduction to Probability Theory and its Applications*. Wiley, New-York.
- Fuller, W. A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **32**(2), 209–226.
- Grafström, A. (2010). On a generalization of poisson sampling. *Journal of Statistical Planning and Inference* **140**(4), 982–991.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**(260), 663–685.
- Iachan, R. (1982). Systematic sampling: A critical review. *International Statistical Review* **50**(3), 293–303.
- Iachan, R. (1983). Asymptotic theory of systematic sampling. *Annals of Statistics* **11**(3), 959–969.
- Janardan, K. G. and Patil, G. P. (1972). A unified approach for a class of multivariate hypergeometric models. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)* **34**(4), 363–376.
- Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.
- Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer, New York.
- Loonis, V. and Mary, X. (2015). Determinantal Sampling Designs. *ArXiv e-prints*.
- Madow, L. H. and Madow, W. G. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics* **15**(1), 1–24.
- Madow, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics* **20**(3), 333–354.
- Meister, K. (2004). On Methods for Real Time Sampling and Distributions in Sampling. PhD thesis, Department of Mathematical Statistics, Umeå University.
- Mitov, K. V. and Omev, E. (2014). *Renewal Processes*. Springer, New York.
- Murthy, M. N. and Rao, T. J. (1988). Systematic sampling with illustrative examples. In Krishnaiah, P. R. and Rao, C. R., editors, *Handbook of Statistics Volume 6: Sampling*, pages 147–185, Amsterdam. Elsevier/North-Holland.
- Pea, J., Qualité, L., and Tillé, Y. (2007). Systematic sampling is a minimal support design. *Computational Statistics & Data Analysis* **51**(12), 5591–5602.
- Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.
- Terrell, G. R. (1999). *Mathematical Statistics: A Unified Introduction*. Springer, New York.
- Tillé, Y. (1996). A moving stratification algorithm. *Survey Methodology* **22**, 85–94.
- Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York.
- Vitter, J. S. (1984). Faster methods for random sampling. *Communications of the ACM* **27**(7), 703–718.
- Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software* **11**(1), 37–57.
- Vitter, J. S. (1987). An efficient algorithm for sequential random sampling. *ACM Transactions on Mathematical Software* **13**(1), 58–67.
- Wilhelm, M., Qualité, L., and Tillé, Y. (2017). Quasi-systematic sampling from a continuous population. *Computational Statistics & Data Analysis* **105**, 11–23.
- Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with

probability proportional to size. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **15**, 235–261.

Institute of Statistics, University of Neuchâtel, Bellevaux 51, 2000 Neuchâtel, Switzerland

E-mail: yves.tille@unine.ch

Institute of Statistics, University of Neuchâtel, Bellevaux 51, 2000 Neuchâtel, Switzerland

E-mail: lionel.qualite@unine.ch

Institute of Statistics, University of Neuchâtel, Bellevaux 51, 2000 Neuchâtel, Switzerland

E-mail: matthieu.wilhelm@unine.ch

(Received February 2016; accepted January 2017)