Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Exact and approximation formulas for second-order inclusion probabilities in randomized systematic sampling with unequal probabilities and without replacement

by Kees van Berkel and Erwin Vondenhoff

Release date: June 29, 2026

Statistics Canada
Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Survey Methodology, June 2026
Vol. 52, No. 1, pp. 235-258
Statistics Canada, Catalogue No. 12-001-X

Exact and approximation formulas for second-order inclusion probabilities in randomized systematic sampling with unequal probabilities and without replacement

Kees van Berkel and Erwin Vondenhoff[1]

Abstract

Probability-proportional-to-size sampling is widely used by national statistical offices. Here population units are selected with probabilities proportional to an auxiliary variable. Variance formulas in such designs require both first- and second-order inclusion probabilities. The computation of second-order inclusion probabilities is particularly challenging for large populations, and has been the subject of extensive research. This article presents some new exact and approximation formulas for second-order inclusion probabilities in randomized systematic sampling with unequal probabilities and without replacement.

Key Words: Horvitz-Thompson estimator; Probability-proportional-to-size (PPS) sampling; Variance estimation.

1. Introduction

National statistical offices often employ sampling designs in which units are selected with probabilities proportional to an auxiliary variable. A common example is the selection of municipalities with probabilities proportional to their population sizes. Such probability-proportional-to-size (PPS) designs are attractive because they improve efficiency when the auxiliary variable is strongly related to the survey variable of interest.

In variance estimation, inclusion probabilities play a central role. The exact variance expression of the Horvitz-Thompson estimator requires not only first-order inclusion probabilities but also second-order inclusion probabilities, denoted by π_{gh} representing the probability that two units g and h are jointly included in the sample. While a number of approximate variance estimators make use only of first-order inclusion probabilities, see Matei and Tillé (2005), it remains important to develop exact and approximation formulas for π_{gh}. Exact formulas allow variance estimation without approximation and provide benchmarks against which approximation methods can be evaluated.

This article focuses on randomized systematic PPS sampling. For this design, Hartley and Rao (1962) derived exact analytical expressions for π_{gh} in the special case of sample size two and population size three or four. They noted that as the population size increases, the exact evaluation of π_{gh} becomes increasingly laborious and the resulting expressions grow very complex. Connor (1966) developed a general procedure for calculating π_{gh} for arbitrary sample and population sizes, extending the results of Hartley and Rao. This procedure was elaborated by Gray (1971) and later implemented in a computer algorithm by Hidiroglou and Gray (1975). However, the associated computations remain demanding, with complexity that grows exponentially in population size.

1. Kees van Berkel and Erwin Vondenhoff, Statistics Netherlands, Division of Data Collection, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. E-mail: cam.vanberkel@cbs.nl.

To address this, researchers have investigated approximations to π_{gh}. Hartley and Rao (1962) proposed an asymptotic approximation, which Kott (2005) showed to perform well with the Horvitz-Thompson variance estimator. Hájek (1964) introduced an approach based on rejective Poisson sampling, while Brewer and Donadio (2003) and Knottnerus (2003, 2011) explored special-form approximations, in line with earlier work by Brewer (1963) and Durbin (1967) for samples of size two. Thompson and Wu (2008) proposed a simulation-based approach for obtaining estimates for π_{gh}.

In this article, new exact and approximation formulas for π_{gh} are derived and compared with existing approaches in the literature. The article is organized as follows. Section 2 introduces the randomized systematic PPS sampling design, and Section 3 establishes additional notation and definitions. Section 4 contains new exact formulas for π_{gh}, and Section 5 contains new approximation formulas. Section 6 reviews related results from the literature, and Section 7 compares the accuracy of our approximations with existing ones. Section 8 illustrates the application of expressions for π_{gh} to variance estimation of the Horvitz-Thompson estimator. Finally, Section 9 concludes the article.

2. Randomized systematic sampling with unequal probabilities without replacement

Consider a population U consisting of N units. Assign to each $g \in U$ a number $X_g > 0$ that is referred to as its size. The purpose of the sampling design is to draw a sample of n distinct units in such a way that the probability π_g of selecting unit $g \in U$ in the sample is proportional to its size X_g. A corresponding sampling procedure described by Hartley and Rao (1962) is as follows. Arrange the units in a random order, denoted by $1,2,\ldots,N$. Compute the subtotals $T_k = \sum_{i=1}^{k} X_i$, $k = 1,2,\ldots,N$. Then $T_N = \sum_{i=1}^{N} X_i = X$ is the total size of U. The interval $(0, T_N]$ is divided into N subintervals $(0, T_1], (T_1, T_2], \ldots, (T_{N-1}, T_N]$ corresponding to units $1,2,\ldots,N$ respectively. The sampling interval T is defined by $T = X / n$. A random start b is selected from the uniform distribution on $(0, T]$. Unit k belongs to the sample if and only if there exists a nonnegative integer i with $T_{k-1} < b + iT \le T_k$, where $T_0 = 0$. A necessary condition for this sampling procedure to select n units is that $X_g < T$ for all units $g \in U$. Therefore, only sizes that satisfy this condition are considered. The first order inclusion probabilities satisfy $\pi_g = X_g / T$ for all $g \in U$.

3. Additional notation and definitions

Let $g, h \in U$ with $g < h$, ordered according to the sampling procedure of Section 2, with associated intervals $(T_{g-1}, T_g]$ and $(T_{h-1}, T_h]$. The distance between g and h is defined as the length of the gap between their intervals, that is $T_{h-1} - T_g$.

For any subset W of U, the number of its units is denoted by $|W|$. The size of W, denoted by $\gamma(W)$, is defined by the sum of the sizes of its units, $\gamma(W) = \sum_{i \in W} X_i$ and $\gamma(\emptyset) = 0$.

For $g \in U$, let $U_g = U \setminus \{g\}$, that is population U from which unit g has been removed. For $g,h \in U$, let $U_{gh} = U \setminus \{g, h\}$, that is population U from which units g and h have been removed.

For $g,h \in U$ and $l \in \{2,\ldots,N\}$, define $V(g,h,l) = \{W \subset U_{gh} \mid |W| = l-2\}$, the set of all subsets of U_{gh} containing exactly $l-2$ units. In the marginal cases where $l=2$ or $l=N$, we have $V(g,h,2) = \{\emptyset\}$ and $V(g,h,N) = U_{gh}$.

4. Exact formulas for second-order inclusion probabilities

For $2 \le n \le N$, the probability π_{gh} that both units g and h are selected in the sample is called the second-order inclusion probability of g and h. The selection scheme implies that it depends on the position of g relative to h whether they are both selected in the sample. Therefore, without loss of generality, we may assume that unit g is positioned at the first position. Let Ω_g be the set of all possible orderings of U with g at the first position. For $\omega \in \Omega_g$ let T^{ω}_{h-1} be the left boundary of the interval corresponding to h, and let $L_{gh}(d)$ denote the length of $(0, X_g] \cap \bigcup_{l=1}^{n-1}(d - lT,\ d - lT + X_h]$ for $d \in [X_g, X - X_h]$. The conditional probability that both units g and h are selected in the sample given ordering $\omega \in \Omega_g$ equals $L_{gh}(T^{\omega}_{h-1})/T$. Since Ω_g has $(N-1)!$ elements, each equally likely to occur, the law of total probability yields

$$\pi_{gh} = \frac{1}{T(N-1)!} \sum_{\omega \in \Omega_g} L_{gh}(T^{\omega}_{h-1}).$$

The calculation of this formula can be simplified by partitioning Ω_g into non-empty disjoint subsets ensuring that T^{ω}_{h-1} is constant within each subset.

Theorem 4.1

$$\pi_{gh} = \frac{1}{T(N-1)} \sum_{l=2}^{N} \frac{1}{\binom{N-2}{l-2}} \sum_{W \in V(g,h,l)} L_{gh}\left(X_g + \gamma(W)\right), \text{ for } g,h \in U.$$

Proof. As unit g is positioned in the first position, unit h can occupy positions $2,\ldots,N$. Suppose h occupies position l for some $l \in \{2,\ldots,N\}$. Then $l-2$ units from U_{gh} are placed between g and h, and the remaining $N-l$ units after h. Note that T^{ω}_{h-1} depends neither on the ordering of the $l-2$ units between g and h, nor on the ordering of the $N-l$ units after h. The total number of permutations of these units is $(l-2)!(N-l)!$. Moreover, $T^{\omega}_{h-1} = X_g + \gamma(W)$, where W denotes the set of the $l-2$ units between g and h. So $T(N-1)!\,\pi_{gh} = \sum_{l=2}^{N} \sum_{W \in V(g,h,l)} (l-2)!\,(N-l)!\,L_{gh}(X_g + \gamma(W))$. ■

With $\binom{N-2}{l-2}$ elements in $V(g,h,l)$, the sum in Theorem 4.1 comprises 2^{N-2} terms. Summing these terms straightforwardly, results in exponential time complexity as the population size N increases. From Property 4.1 below, it follows that evaluating the function L_{gh} scales linearly with the sample size n. So the computational complexity of the sum in Theorem 4.1 is of order $n \cdot 2^N$.

The function L_{gh} can be expressed as a sum of positive parts of linear functions and is symmetric with respect to the midpoint of its domain.

Property 4.1

For $g,h \in U$ and $d \in [X_g, X - X_h]$ we have

(a) $L_{gh}(d) = \sum_{k=1}^{n-1} \{\mathbb{P}(d - kT + X_h) - \mathbb{P}(d - kT) - \mathbb{P}(d - kT - X_g + X_h) + \mathbb{P}(d - kT - X_g)\}$, where $\mathbb{P}(x) = \max(x, 0)$, the positive part of $x \in \mathbb{R}$.

(b) $L_{gh}(d) = L_{gh}(X + X_g - X_h - d)$.

The proof of Property 4.1 is included in Appendix A.

By symmetry of L_{gh} and binomial coefficients, the summation in the formula for π_{gh} in Theorem 4.1 can be reduced.

Theorem 4.2

$$\pi_{gh} = \frac{1}{T(N-1)} \sum_{l=2}^{\lfloor \frac{N}{2} \rfloor + 1} \frac{\left(2 - \delta_{l, \frac{N}{2}+1}\right)}{\binom{N-2}{l-2}} \sum_{W \in V(g,h,l)} L_{gh}(X_g + \gamma(W)), \text{ for } g,h \in U,$$

where $\lfloor x \rfloor$ is the integer part of x, and $\delta_{i,j}$ is the Kronecker delta which equals 1 if $i = j$ and 0 otherwise.

Proof. For each $W \in V(g,h,l)$ its complement W^C in U_{gh} is the subset of U_{gh} containing all units that do not belong to W. So $W^C \in V(g,h,N-l+2)$, and $X_g + X_h + \gamma(W) + \gamma(W^C) = X$. From this and Property 4.1.(b), it follows that $L_{gh}(X_g + \gamma(W)) = L_{gh}(X_g + \gamma(W^C))$ and therefore,

$$\begin{aligned} \sum_{W \in V(g,h,N-l+2)} L_{gh}(X_g + \gamma(W)) &= \sum_{Z \in V(g,h,l)} L_{gh}(X_g + \gamma(Z^C)) \\ &= \sum_{Z \in V(g,h,l)} L_{gh}(X_g + \gamma(Z)). \end{aligned}$$

From this and property $\binom{n}{k} = \binom{n}{n-k}$ of binomial coefficients, it follows that the l-th term in the sum over l in Theorem 4.1 equals the $(N-l+2)$-th term for $l = 2, \ldots, N$. Thus, the sum over l equals twice the sum over the first half of the terms when N is odd, and twice the sum over the first $N/2$ terms plus the middle term when N is even. ∎

Applying Theorem 4.2 and Property 4.1.(a), the computation of the π_{gh} is straightforward. However, for large N, the computation becomes laborious. Although the number of terms in the sum of Theorem 4.2 has been halved compared to the sum in Theorem 4.1, it contains 2^{N-3} terms. So the computation time required still grows exponentially with the population size N and remains of order $n \cdot 2^N$.

The expression for π_{gh} in Theorem 4.2 can be further simplified, in particular if some unit sizes in U are equal. For that purpose the following vector notation is used for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ and $\mathbf{k}, \mathbf{n} \in \mathbb{N}_0^m$,

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^m a_i b_i, \ |\mathbf{a}| = \sum_{i=1}^m |a_i|, \ \binom{\mathbf{n}}{\mathbf{k}} = \prod_{i=1}^m \binom{n_i}{k_i}, \ 0 \le \mathbf{k} \le \mathbf{n} \text{ means } 0 \le k_i \le n_i \text{ for all } i.$$

Theorem 4.3

$$\pi_{gh} = \frac{1}{T(N-1)} \sum_{l=2}^{\left\lfloor \frac{N}{2} \right\rfloor + 1} \frac{\left(2 - \delta_{l, \frac{N}{2}+1}\right)}{\binom{N-2}{l-2}} \sum_{\substack{\mathbf{0} \le \mathbf{k} \le \mathbf{n}_{gh} \\ |\mathbf{k}| = l-2}} \binom{\mathbf{n}_{gh}}{\mathbf{k}} L_{gh}(X_g + \mathbf{k} \cdot \mathbf{X}_{gh}), \text{ for } g,h \in U,$$

where bold letters are vectors of dimension equal to the number of distinct unit sizes in U_{gh}. Vector $\mathbf{X}_{gh}$ stores those sizes, vector $\mathbf{n}_{gh}$ counts how many times each size occurs in U_{gh}, and vector $\mathbf{k}$ contains nonnegative integers.

Proof. Let $g,h \in U$, and let m be the number of distinct unit sizes in U_{gh}. By re-indexing, if necessary, those sizes are $X_1, X_2, \ldots, X_m$. For each subset W of U_{gh}, its size $\gamma(W)$ equals $\mathbf{k} \cdot \mathbf{X}_{gh}$ where $\mathbf{k}$ is the m-vector whose i-th component equals the number of units in W with size X_i. For $l \ge 2$, the number of sets $W \in V(g,h,l)$ containing exactly k_i units of size X_i for $i = 1, \ldots, m$, equals $\binom{\mathbf{n}_{gh}}{\mathbf{k}}$, restricted to $\mathbf{0} \le \mathbf{k} \le \mathbf{n}_{gh}$ and $|\mathbf{k}| = |W| = l-2$. ■

If there are m distinct unit sizes in U_{gh}, the computational complexity of evaluating π_{gh} via Theorem 4.3 is polynomial of order $n \cdot N^m$ in N, for fixed m. For small m, this represents a substantial reduction compared to the exponential complexity $n \cdot 2^N$ in Theorem 4.2. Consequently, exact evaluation of π_{gh} is feasible with Theorem 4.2 only for small N, and with Theorem 4.3 for somewhat larger N, provided that m is small. For large N, however, neither exact formula is computationally tractable. Therefore, approximation formulas are presented in the next section.

5. Approximation formulas for second-order inclusion probabilities

If all units in U_{gh} are of equal size for certain $g,h \in U$, then π_{gh} can be computed by Theorem 4.3. Therefore, in this section it is assumed that the units in U_{gh} are not all of equal size.

Theorem 5.1

For $N \ge 8$, approximation $\tilde{\pi}_{gh}$ for π_{gh} is given by,

$$\begin{aligned} T(N-1) \cdot \tilde{\pi}_{gh} = {} & 2\, L_{gh}(X_g) + \frac{2}{N-2} \sum_{i \in U_{gh}} L_{gh}(X_g + X_i) \\ & + \frac{4}{(N-2)(N-3)} \sum_{i<j \in U_{gh}} L_{gh}(X_g + X_i + X_j) \\ & + \sum_{l=5}^{\left\lfloor \frac{N}{2} \right\rfloor + 1} \left(2 - \delta_{l, \frac{N}{2}+1}\right) \tilde{t}_{ghl}, \text{ for } g,h \in U, \end{aligned}$$

where

$$\begin{aligned} \tilde{t}_{ghl} = {} & \sum_{k=1}^{n-1} \left\{ \mathrm{F}_{ghl}(kT) - \mathrm{F}_{ghl}(\max(X_g, kT - X_h)) \right. \\ & + \left. \mathrm{F}_{ghl}(kT + X_g - X_h) - \mathrm{F}_{ghl}(\min(X - X_h, kT + X_g)) \right\} \\ & + \mathbb{P}(X_g + X_h - T) \cdot (\Phi_{ghl}(X - X_h) - \Phi_{ghl}(X_g)), \end{aligned}$$

with

$$\begin{aligned} \mathrm{F}_{ghl}(x) &= (\mu_{ghl} - x)\,\Phi_{ghl}(x) - \sigma^2_{ghl}\varphi_{ghl}(x), \quad x \in \mathbb{R}, \\ \mu_{ghl} &= X_g + (l-2)\cdot\mu_{gh}, \qquad \sigma^2_{ghl} = \frac{(l-2)\cdot(N-l)}{N-3}\cdot\sigma^2_{gh}, \\ \mu_{gh} &= \frac{1}{N-2}(X - X_g - X_h), \qquad \sigma^2_{gh} = \frac{1}{N-2}\sum_{i\in U_{gh}}(X_i - \mu_{gh})^2, \end{aligned}$$

φ_{ghl} and Φ_{ghl} the probability density function and the cumulative distribution function, respectively, of the normal distribution with mean μ_{ghl} and variance σ^2_{ghl}.

The proof of Theorem 5.1 is included in Appendix B.

The complexity of the approximations in Theorems 5.1 is of order $n \cdot N,$ which is linearly in both n and $N.$ That is a substantial reduction compared to the complexity of $n \cdot 2^N$ for the exact formulas in Theorems 4.1 and 4.2, and of $n \cdot N^m$ for the exact formula in Theorem 4.3, where m is the number of distinct unit sizes in U_{gh}.

For a fixed sample size $n,$ the exact second-order inclusion probabilities π_{gh} have the property

$$\sum_{h\in U_g}\pi_{gh} = (n-1)\,\pi_g, \text{ for } g \in U. \tag{5.1}$$

The approximations $\tilde{\pi}_{gh}$ of Theorem 5.1 do not necessarily have this property. Therefore, they are scaled such that the scaled approximations do have the property

$$\sum_{h\in U_g}\alpha_{gh}\cdot\tilde{\pi}_{gh} = (n-1)\,\pi_g, \text{ for } g \in U. \tag{5.2}$$

where the multipliers α_{gh} satisfy $\alpha_{gh} = \alpha_{hg}$ for all $g \neq h \in U.$ This results in N equations for $N(N-1)/2$ unknown multipliers. For $N > 3,$ the number of unknowns exceeds the number of equations, so the multipliers are not uniquely determined. To address this, an additional condition is imposed. The multipliers should be as close to one as possible. This is formalized by minimizing

$$\sum_{g\in U}\sum_{g<h\in U}(\alpha_{gh} - 1)^2, \tag{5.3}$$

subject to the constraints given in equations (5.2).

Theorem 5.2

For $N \geq 8,$ approximation π^{BV}_{gh} for π_{gh} is given by

$$\pi^{\mathrm{BV}}_{gh} = \alpha_{gh}\cdot\tilde{\pi}_{gh}, \text{ for } g,h \in U,$$

with $\tilde{\pi}_{gh}$ from Theorem 5.1, and $\alpha_{gh} = 1 - \frac{1}{2}(\lambda_g + \lambda_h)\,\tilde{\pi}_{gh}.$ Here $\boldsymbol{\lambda}$ is the solution of $\mathbf{A}\cdot\boldsymbol{\lambda} = \mathbf{b},$ with $\mathbf{A}$ the $N \times N$- matrix with entries $A_{gh} = \tilde{\pi}^2_{gh}$ for $g \neq h \in U,$ and $A_{gg} = \sum_{h\in U_g}\tilde{\pi}^2_{gh}$ for $g \in U,$ and $\mathbf{b}$ the N- vector with elements $b_g = 2\left(\sum_{h\in U_g}\tilde{\pi}_{gh} - (n-1)\,\pi_g\right)$ for $g \in U.$

The proof of Theorem 5.2 is included in Appendix C. In general, finding a closed-form solution to $\mathbf{A}\cdot\boldsymbol{\lambda} = \mathbf{b}$ is challenging, but this is not necessary for our purposes. Accurate numerical methods are

available for solving the equations, in general with complexity of order N^3. The equations only need to be solved once to calculate all scale factors.

6. Expressions for second-order inclusion probabilities in the literature

There is extensive literature on second-order inclusion probabilities in randomized systematic sampling with unequal probabilities without replacement. As mentioned in the introduction, Hartley and Rao (1962), Connor (1966) and Gray (1971) derived procedures for computing the exact values of π_{gh}. Connor's formula (2.4) involves a sum with the same number of terms as the sum appearing in our Theorem 4.1. However, unlike our explicit expression for the summand using Property 4.1(a) of function L_{gh}, Connor does not provide an explicit closed-form representation of the individual terms. Gray's modification of Connor's formula likewise yields a sum with as many terms as in our Theorem 4.2, but his formulation is procedural rather than explicit. The corresponding algorithm was subsequently implemented in a computer program by Hidiroglou and Gray (1975).

Over the years, many approximation formulas have been published. One particularly intriguing result is the asymptotic approximation for π_{gh} derived by Hartley and Rao (1962), formula (5.16),

$$\pi_{gh}^{\mathrm{HR}} = n(n-1)\left\{\begin{array}{l} X_gX_hX^{-2} + (X_g^2X_h + X_gX_h^2)\,X^{-3} - X_gX_h\mathrm{A}X^{-4} + \\ 2(X_g^3X_h + X_gX_h^3 + X_g^2X_h^2)X^{-4} - 3(X_g^2X_h + X_gX_h^2)\,\mathrm{A}X^{-5} + \\ 3X_gX_h\mathrm{A}^2X^{-6} - 2X_gX_h\mathrm{B}X^{-5} \end{array}\right\}, \tag{6.1}$$

where

$$\mathrm{A} = \sum_{i\in U} X_i^2 \text{ and } \mathrm{B} = \sum_{i\in U} X_i^3.$$

This approximation holds for fixed n and large N, or small n/N, and is correct to the order of N^{-4}. Brewer and Donadio (2003) proposed approximations of the form

$$\pi_{gh}^{\mathrm{BD}} = \pi_g\pi_h(c_g + c_h)/2. \tag{6.2}$$

For $c_g = c$, it follows from $\sum_{g\in U}\sum_{h\in U_g}\pi_{gh} = n(n-1)$ that

$$c = (n-1)\Big/\left(n - \frac{1}{n}\sum_{i\in U}\pi_i^2\right). \tag{6.3}$$

Inspired by the ratio of sums of π_{gh} and $\pi_g\pi_h$ for $h\in U_g$ they also proposed

$$c_g = (n-1)\big/(n-\pi_g). \tag{6.4}$$

Elaborating on Hartley and Rao's approximation (6.1), they also came up with

$$c_g = (n-1)\Big/\left(n - 2\pi_g + \frac{1}{n}\sum_{i\in U}\pi_i^2\right). \tag{6.5}$$

Knottnerus (2003 and 2011) chose a different expression for c_g:

$$c_g = (n-1) \,/\, (n\gamma(1-2X_g / X)) \text{ with } \gamma = \frac{1}{2}+\frac{1}{2}\sum_{i\in U}\frac{X_i}{X-2X_i}, \tag{6.6}$$

to arrive for $X_g, X_h < X/2$ at the approximation formula

$$\pi_{gh}^{\mathrm{K}} = n(n-1)\, X_g X_h (X - X_g - X_h) / (\gamma X (X-2X_g)(X-2X_h)). \tag{6.7}$$

This approximation coincides with those proposed by Brewer (1963) and Durbin (1967) for PPS samples of size 2. Another interesting approximation formula comes from the related rejective Poisson sampling design, proposed by Hájek (1964):

$$\pi_{gh}^{\mathrm{H}} = \pi_g \pi_h \left\{1-(1-\pi_g)(1-\pi_h)/d\right\}, \tag{6.8}$$

if $d = \sum_{i\in U}\pi_i(1-\pi_i)$ is large, which implies that n and $N-n$ are large. Thompson and Wu (2008) developed a simulation-based approach for computing the second-order inclusion probabilities. For $n=10$ and $N=20,$ the Hartley-Rao approximations are close to the simulated approximations. They match to the second decimal point for the majority of the π_{gh}.

7. Accuracy of approximation formulas for second-order inclusion probabilities

To investigate the accuracy of approximation formulas for second-order inclusion probabilities, the mean absolute error (MAE) is used,

$$\mathrm{MAE} = \frac{2}{N(N-1)}\sum_{g<h\in U}\left|\pi_{gh}-\hat{\pi}_{gh}\right|, \tag{7.1}$$

where $\hat{\pi}_{gh}$ is an approximation of the exact value π_{gh}. Since $\pi_{gh}=\pi_{hg},$ this quality measure includes only the probabilities π_{gh} where $g<h.$ In the following examples, the exact second-order inclusion probabilities were calculated using Theorem 4.2. Various approximation methods, including those by Van Berkel and Vondenhoff (BV), Knottnerus (K), Hartley and Rao (HR), Hájek (H), and Brewer and Donadio (BD), are evaluated across different population distributions and sample sizes. For the BV-approximations Theorem 5.2 was applied, and for the BD-approximations formulas (6.2) and (6.3).

Example 7.1

Consider Brewer's extended example, where the population consists of eight units with $X_k = k$ for $k=1,\ldots,8.$ This population is similar to the population corresponding to Figure B.1 of Appendix B. A systematic PPS sample of size $n=2$ is drawn. The population size X equals 36 and the sampling interval T equals $X/n=$ 18. The exact second-order inclusion probabilities are shown in matrix (7.2), rounded to five decimal places. For completeness, the first-order inclusion probabilities $\pi_g = \pi_{gg}$ are on the diagonal.

$$(\pi_{gh})_{g\le h\in U} = \begin{pmatrix} 0.05556 & 0.00238 & 0.00423 & 0.00608 & 0.00794 & 0.00979 & 0.01164 & 0.01349 \\ & 0.11111 & 0.00794 & 0.01349 & 0.01534 & 0.02090 & 0.02275 & 0.02831 \\ & & 0.16676 & 0.02275 & 0.02460 & 0.03016 & 0.03571 & 0.04127 \\ & & & 0.22222 & 0.03016 & 0.03942 & 0.05053 & 0.05979 \\ & & & & 0.27778 & 0.05423 & 0.06534 & 0.08016 \\ & & & & & 0.33333 & 0.08016 & 0.09868 \\ & & & & & & 0.38889 & 0.12275 \\ & & & & & & & 0.44444 \end{pmatrix} \quad (7.2)$$

Theorem 5.2 provides approximations for the second-order inclusion probabilities,

$$(\pi_{gh}^{BV})_{g<h\in U} = \begin{pmatrix} 0.00264 & 0.00441 & 0.00615 & 0.00789 & 0.00967 & 0.01148 & 0.01333 \\ & 0.00893 & 0.01339 & 0.01576 & 0.02025 & 0.02274 & 0.02741 \\ & & 0.02164 & 0.02471 & 0.03011 & 0.03549 & 0.04139 \\ & & & 0.03134 & 0.03963 & 0.05044 & 0.05964 \\ & & & & 0.05345 & 0.06482 & 0.07981 \\ & & & & & 0.08064 & 0.09959 \\ & & & & & & 0.12329 \end{pmatrix} \quad (7.3)$$

The approximations match the exact probabilities at least to the second decimal place: 2 match to the second decimal, 20 to the third, and 6 to the fourth. The MAE is 0.000389. Table 7.1 contains the MAE × 10^6 of the different approximation methods, for all sample sizes satisfying the non-self-selecting condition that the sampling interval T exceeds size X_k for all k.

Table 7.1
MAE × 10^6 of approximations of second-order inclusion probabilities

n	BV	K	HR	H	BD
2	389	1,111	1,195	3,882	3,163
3	469	5,189	5,521	7,360	10,246
4	906	10,550	11,361	10,729	22,071

Note: Brewer and Donadio (BD); Van Berkel and Vondenhoff (BV); Hájek (H); Hartley and Rao (HR); Knottnerus (K); mean absolute error (MAE).

Example 7.2

Consider the population corresponding to the right panel of Figure B.1 of Appendix B. The sizes of its thirteen units are 16, 20, 6, 8, 19, 8, 17, 13, 12, 5, 11, 18, 15, generated by the discrete Uniform distribution over the interval [5,20]. Table 7.2 contains the MAE × 10^6 of the different approximation methods, for all sample sizes satisfying the non-self-selecting condition.

Table 7.2
MAE × 10^6 of approximations of second-order inclusion probabilities

n	BV	K	HR	H	BD
2	17	26	31	900	437
3	126	132	167	1,232	1,316
4	87	545	598	1,581	2,863
5	111	1,599	1,601	2,074	3,734
6	245	2,947	2,974	4,143	8,038
7	382	3,748	3,850	4,715	10,806
8	232	6,473	6,623	4,479	13,594

Note: Brewer and Donadio (BD); Van Berkel and Vondenhoff (BV); Hájek (H); Hartley and Rao (HR); Knottnerus (K); mean absolute error (MAE).

Example 7.3

Consider the population corresponding to the left panel of Figure B.2 of Appendix B. The sizes of its units are 10, 11, 6, 9, 8, 10, 8, 7, 13, 13, 11, 10, 14, generated by rounding 13 draws from the Normal distribution with mean 10 and variance 4. Table 7.3 contains the MAE × 10^6 of the different approximation methods, for all sample sizes satisfying the non-self-selecting condition.

Table 7.3
MAE × 10^6 of approximations of second-order inclusion probabilities

n	BV	K	HR	H	BD
2	35	29	28	857	290
3	113	129	115	1,150	857
4	112	479	445	1,380	1,669
5	80	997	1,024	1,537	3,229
6	318	2,704	2,675	3092	4,974
7	97	3,278	3,398	3,304	7,779
8	196	3,024	3,177	2,156	8,970
9	292	3,562	3,792	1,643	11,466

Note: Brewer and Donadio (BD); Van Berkel and Vondenhoff (BV); Hájek (H); Hartley and Rao (HR); Knottnerus (K); mean absolute error (MAE).

Example 7.4

Consider the population corresponding to the right panel of Figure B.2 of Appendix B. The sizes of its units are 20, 48, 17, 111, 82, 8, 72, 4, 15, 9, 59, 30, 35, generated by rounding 13 draws from the Chi-squared distribution with three degrees of freedom. Table 7.4 contains the MAE × 10^6 of the different approximation methods, for all sample sizes satisfying the non-self-selecting condition.

Table 7.4
MAE × 10^6 of approximations of second-order inclusion probabilities

n	BV	K	HR	H	BD
2	105	198	253	1,141	1,210
3	341	1,059	1,005	2,045	3,576
4	623	2,527	2,701	3,938	8,468

Note: Brewer and Donadio (BD); Van Berkel and Vondenhoff (BV); Hájek (H); Hartley and Rao (HR); Knottnerus (K); mean absolute error (MAE).

Example 7.5

Consider Connor's population consisting of ten units with sizes 0.3, 0.1, 0.2, 0.15, 0.05, 0.4, 0.08, 0.18, 0.34, 0.2. Table 7.5 contains the MAE × 10^6 of the different approximation methods, for all sample sizes satisfying the non-self-selecting condition.

Table 7.5
MAE × 10^6 of approximations of second-order inclusion probabilities

n	BV	K	HR	H	BD
2	176	235	273	2,010	1,682
3	280	1,427	1,281	3,025	4,962
4	806	3,953	4,117	5,859	12,111
5	966	10,095	11,155	5,775	20,968

Note: Brewer and Donadio (BD); Van Berkel and Vondenhoff (BV); Hájek (H); Hartley and Rao (HR); Knottnerus (K); mean absolute error (MAE).

Across the five examples, method BV consistently outperforms the other methods in terms of MAE, especially for larger samples where the differences in error magnitude become more pronounced. Method BD shows the highest errors across all tables. Among the remaining three methods, it is noticeable that methods K and HR perform about the same. In Examples 7.1 and 7.4, methods K and HR outperform or match method H, whereas in Examples 7.2, 7.3, and 7.5 they perform better than method H for small sample sizes but worse for larger ones. For all methods, the MAE increases with the sample size, with minor local deviations from this trend, and a noticeable deviation for method H in Examples 7.2, 7.3 and 7.5 for large sample sizes.

8. Variance estimation

In their seminal article, Horvitz and Thompson (1952) introduced an unbiased estimator for the population total of a target variable Y, now known as the Horvitz-Thompson estimator,

$$\hat{Y}_{\mathrm{HT}} = \sum_{g\in S} y_g \big/ \pi_g , \tag{8.1}$$

where S is a sample drawn under a probability-based sampling design ensuring that all first-order inclusion probabilities π_g are positive and known, and y_g is the observed value of Y for unit g. The variance of this estimator is given by

$$\mathrm{V}(\hat{Y}_{\mathrm{HT}}) = \sum_{g=1}^{N} \frac{1-\pi_g}{\pi_g} Y_g^2 + 2\cdot \sum_{g=1}^{N-1}\sum_{h=g+1}^{N} \frac{\pi_{gh}-\pi_g\pi_h}{\pi_g\pi_h} Y_g Y_h, \tag{8.2}$$

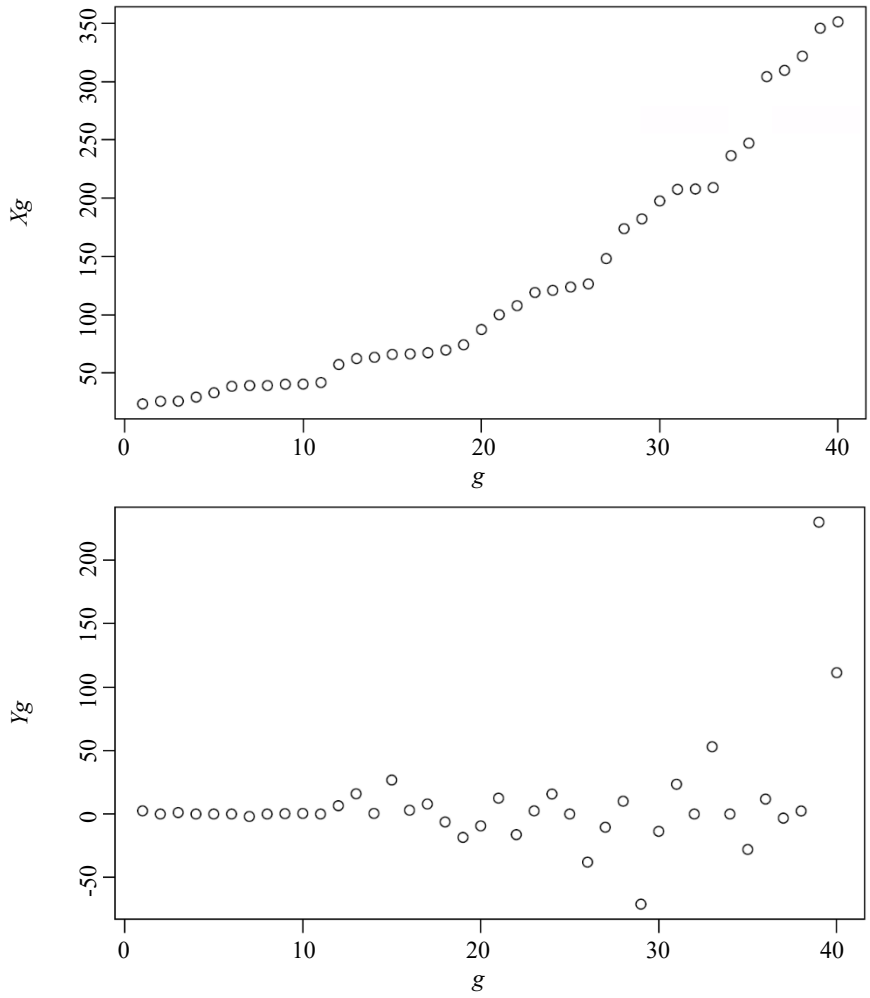
where N is the population size and Y_g is the true value of target variable Y for unit g. For fixed-size samples, this variance can be rewritten as

$$\mathrm{V}(\hat{Y}_{\mathrm{HT}}) = -\sum_{g=1}^{N-1}\sum_{h=g+1}^{N} (\pi_{gh}-\pi_g\pi_h)\left(\frac{Y_g}{\pi_g}-\frac{Y_h}{\pi_h}\right)^2 . \tag{8.3}$$

Calculating the variance of the Horvitz-Thompson estimator requires knowledge of all first- and second-order inclusion probabilities, as well as the values of Y for all population units.

As an application, the Producer Price Index in the Netherlands is considered. In Section 4 of Knottnerus (2011), 70 out of about 2,500 commodity price indexes were taken. To reduce computation time for the variance formulas, we made some simplifications. First, the commodities were sorted by size of the auxiliary variable X, scaled by a factor of thousand. Then, the 23 commodities with the smallest X-values and the 7 with the largest X-values were removed, resulting in a population of $N = 40$. Figure 8.1 illustrates the values of X and Y for these commodities.

Figure 8.1 Values of auxiliary variable X and target variable Y for units in the population



To further simplify the calculation of the exact second-order inclusion probabilities, k-means clustering was applied to X, forming seven clusters. For each cluster c, the mean X-value $\overline{X}_c$ was computed and rounded to the nearest integer, see Table 8.1.

Table 8.1
Cluster (c), size (X), cluster mean of X ($\bar{X}_c$), and target variable (Y) of 40 commodities (units)

unit	cluster c	X	$\bar{X}_c$	Y	unit	cluster c	X	$\bar{X}_c$	Y
1	1	23.6	34	2.5	21	3	100.0	121	12.5
2	1	25.8	34	0.0	22	3	107.8	121	-16.2
3	1	25.9	34	1.2	23	3	119.1	121	2.5
4	1	29.4	34	0.0	24	3	120.9	121	15.8
5	1	33.1	34	0.0	25	3	123.7	121	0.0
6	1	38.6	34	0.0	26	3	126.4	121	-37.9
7	1	39.3	34	-1.9	27	3	148.1	121	-10.4
8	1	39.3	34	0.0	28	4	173.8	196	10.1
9	1	40.4	34	0.3	29	4	182.2	196	-71.0
10	1	40.5	34	0.5	30	4	197.5	196	-13.6
11	1	41.9	34	0.0	31	4	207.5	196	23.5
12	2	57.3	68	6.6	32	4	207.8	196	0.0
13	2	62.4	68	16.0	33	4	209.0	196	53.0
14	2	63.5	68	0.5	34	5	236.4	242	0.0
15	2	66.0	68	26.8	35	5	247.1	242	-27.9
16	2	66.4	68	3.0	36	6	304.1	312	11.8
17	2	67.4	68	7.9	37	6	309.5	312	-3.2
18	2	69.7	68	-6.2	38	6	321.7	312	2.4
19	2	74.2	68	-18.4	39	7	345.6	348	230.0
20	2	87.3	68	-9.3	40	7	351.1	348	111.4

For the remainder of this section, the original X-values are replaced by their corresponding cluster means. With this adjustment, Pearson's correlation coefficient between X and Y is 0.42. For sample sizes ranging from 8 to 13, the variances of the Horvitz-Thompson estimator were calculated in two ways. First, by substituting the exact second-order inclusion probabilities of Theorem 4.3 into formula (8.2), and second, by substituting the approximated second-order inclusion probabilities of Theorem 5.2 into formula (8.2). In both cases, the exact first-order inclusion probabilities were used.

Table 8.2 presents the results and demonstrates that the approximated variances closely match the exact variances. For sample sizes 8 and 9, the approximations slightly overestimate the exact variances, while for larger sample sizes they slightly underestimate them. However, the relative differences never exceed 1.5%, confirming that the approximations are highly reliable.

Table 8.2
Exact variances (V) and approximated variances $\hat{\mathrm{V}}$ of the Horvitz-Thompson estimator by sample size

n	V	$\hat{\mathrm{V}}$	$\hat{\mathrm{V}}-V$	$(\hat{\mathrm{V}}-V)/V$
8	99,727	100,122	395	0.004
9	80,394	80,615	221	0.003
10	64,419	64,331	-88	-0.001
11	51,328	50,965	-363	-0.007
12	40,102	39,609	-493	-0.012
13	31,005	30,554	-451	-0.015

In practical applications, the target variable Y is observed only for sample units, so the variance must be estimated. For samples with size $n > 1$, an unbiased variance estimator is given by

$$\hat{\mathrm{V}}_{\mathrm{HT}}(\hat{Y}_{\mathrm{HT}}) = \sum_{g=1}^{n} \frac{1-\pi_g}{\pi_g^2} y_g^2 + 2 \cdot \sum_{g=1}^{n-1} \sum_{h=g+1}^{n} \frac{\pi_{gh} - \pi_g \pi_h}{\pi_{gh} \pi_g \pi_h} y_g y_h. \tag{8.4}$$

Sen (1953) and Yates and Grundy (1953) proposed an unbiased variance estimator for fixed-size samples,

$$\hat{\mathrm{V}}_{\mathrm{SYG}}(\hat{Y}_{\mathrm{HT}}) = -\sum_{g=1}^{n-1} \sum_{h=g+1}^{n} \frac{\pi_{gh} - \pi_g \pi_h}{\pi_{gh}} \left(\frac{y_g}{\pi_g} - \frac{y_h}{\pi_h} \right)^2. \tag{8.5}$$

Both of these estimators require first- and second-order inclusion probabilities and the values of Y for all sampled units. Matei and Tillé (2005) reviewed twenty variance estimators applicable to unequal probability designs and found that estimators (8.4) and (8.5) perform similarly, except in populations with a high correlation between X and Y, where (8.4) is slightly inferior. Their simulations also indicated that estimators using only first-order inclusion probabilities perform comparably regardless of correlation between X and Y. Consequently, they recommended the simpler estimator proposed by Deville (1993),

$$\hat{\mathrm{V}}_{D}(\hat{Y}_{\mathrm{HT}}) = \sum_{g=1}^{n} \frac{c_g}{\pi_g^2} \left\{ y_g - \pi_g \cdot \left(\sum_{h=1}^{n} c_h y_h \Big/ \pi_h \right) \Big/ \sum_{h=1}^{n} c_h \right\}^2, \text{ with } c_g = (1-\pi_g) \frac{n}{n-1}. \tag{8.6}$$

In addition, the case $c_g = n/(n-1)$ corresponds to sampling with replacement. However, as noted by Matei and Tillé, this estimator is highly biased and substantially overestimates the variance when applied to a design without replacement. For this reason, it is not considered further in this section.

In addition to variance estimators (8.4) and (8.5), we define the estimators $\hat{V}_{\mathrm{HT}}^{\mathrm{BV}}$ and $\hat{V}_{\mathrm{SYG}}^{\mathrm{BV}}$ by replacing the exact second-order inclusion probabilities π_{gh} with their approximations π_{gh}^{BV} from Theorem 5.2. Together with estimator (8.6), this results in a total of five variance estimators.

The accuracy of the five estimators was evaluated through simulation. Since units with sizes exceeding the sampling interval arise for sample sizes greater than 13, the analysis was restricted to the six largest feasible sample sizes: 8, 9, …, 13. For each sample size, 1,000 independent samples were drawn using the randomized systematic sampling design with probabilities proportional to the cluster mean X-values from Table 8.1. The averages of the Horvitz-Thompson estimates for the population total 322.3 of Y were 319.2, 323.7, 320.5, 324.4, 321.5, and 321.4 for sample sizes 8 through 13, respectively.

The accuracy of each variance estimator $\hat{\mathrm{V}}$ is assessed using two quality indicators: relative bias ($\mathrm{RB}_{\mathrm{sim}}$) and relative standard error ($\mathrm{RSE}_{\mathrm{sim}}$), which are defined as follows. Fix the sample size n. Let V_{sim} be the vector whose i-th component is the variance estimate from $\hat{\mathrm{V}}$ for the i-th simulation, $i = 1, \ldots, 1{,}000$. Denote the mean and variance of these estimates by $\overline{V}_{\mathrm{sim}}$ and $\mathrm{var}(V_{\mathrm{sim}})$, respectively, and let V be the true variance as reported in Table 8.2. Then,

$$\mathrm{RB}_{\mathrm{sim}}(\hat{\mathrm{V}}) = \frac{\overline{V}_{\mathrm{sim}} - V}{V} \text{ and } \mathrm{RSE}_{\mathrm{sim}}(\hat{\mathrm{V}}) = \frac{\sqrt{\mathrm{var}(V_{\mathrm{sim}})}}{V}. \tag{8.7}$$

Table 8.3 presents the relative biases and relative standard errors for the five estimators. The relative biases are very small, confirming that the estimators are effectively unbiased. In particular, the results for the estimators using the approximated second-order inclusion probabilities (columns 2 and 3, as well as columns 4 and 5) are very similar. Deville's estimator shows relative biases of the same order of magnitude as the other estimators. The relative standard errors primarily determine estimator accuracy. The figures in columns 7 and 8 are nearly identical, as are those in columns 9 and 10, indicating consistent performance between the exact and approximated approaches. The standard errors for Deville's estimator are comparable to those of the other estimators.

Table 8.3
Relative bias and relative standard error of the variance estimators in the simulations by sample size

	Relative bias					Relative standard error				
n	$\hat{V}_{HT}$	$\hat{V}^{BV}_{HT}$	$\hat{V}_{SYG}$	$\hat{V}^{BV}_{SYG}$	$\hat{V}_{D}$	$\hat{V}_{HT}$	$\hat{V}^{BV}_{HT}$	$\hat{V}_{SYG}$	$\hat{V}^{BV}_{SYG}$	$\hat{V}_{D}$
	%					%				
8	-0.9	-0.9	2.3	3.2	0.3	45.9	46.1	50.1	50.8	46.7
9	-0.7	-0.6	1.4	2.0	0.4	52.3	52.4	51.8	52.0	51.8
10	2.8	2.7	3.8	3.7	3.3	53.1	53.0	55.2	55.2	53.3
11	-1.0	-1.8	0.2	-0.6	-0.8	27.1	26.5	25.6	25.2	26.4
12	-1.7	-2.8	-0.9	-2.3	-0.4	47.0	46.3	48.3	47.5	47.1
13	2.7	1.4	4.7	3.2	3.1	55.2	55.2	58.8	58.4	53.5

Note: Van Berkel and Vondenhoff (BV), Horvitz-Thompson (HT); Sen et Yates and Grundy (SYG).

No single estimator consistently achieves the lowest bias or smallest standard error across all sample sizes. It does not make much difference whether our approximations or exact values of second-order inclusion probabilities are used in variance estimates. Moreover, for any fixed sample size, no estimator outperforms the others, suggesting that all variance estimators perform at a similar level.

9. Conclusion

In this article, new exact and approximation formulas are derived for the second-order inclusion probabilities in randomized systematic sampling with unequal probabilities and without replacement. Our exact formulas are more explicit and less cumbersome than those available in the literature. When all population unit sizes are distinct, the computational complexity of our exact expressions in Theorems 4.1 and 4.2 matches that of the existing formulas: growing exponentially with the population size and linearly with the sample size. However, when the population contains units with identical sizes, the computational complexity of our exact formula in Theorem 4.3 increases polynomially with the population size, while still scaling linearly with the sample size.

The approximation formula in Theorems 5.1 is based on the key observation that the distance between two population units, given a fixed number of units between them, is asymptotically normally distributed.

This follows from the Central Limit Theorem for simple random sampling without replacement from finite populations (Erdős and Rényi, 1959). As shown in Appendix B, this approximation performs well even for small populations (e.g. for $N = 13$) and for a variety of unit size patterns. The computational complexity of the approximation grows linearly in both the population size and the sample size.

The ultimate approximation formula in Theorem 5.2 is a scaled version of that in Theorem 5.1. The computational complexity of computing all scaling factors is of order N^3. Compared with existing approximations, our method provides more accurate estimates for the second-order inclusion probabilities, particularly for larger sample sizes, regardless of the underlying unit size configuration.

Given the high accuracy of our approximate second-order inclusion probabilities, they are well suited for estimating the variance of the Horvitz-Thompson estimator. In practice, using either the exact or approximate probabilities yields variance estimates of comparable accuracy. Nevertheless, Deville's variance estimator, requiring only first-order inclusion probabilities, achieves the same level of accuracy while remaining computationally feasible for large populations. Therefore, Deville's estimator remains the practical choice for large-scale variance estimation.

Acknowledgements

The authors thank the referees and the Associate Editor for their careful reading of the manuscript and for providing valuable comments that helped improve it. The authors also thank Sander Scholtus of Statistics Netherlands for mathematical support and for many insightful discussions and suggestions.

Appendix

A. Proof of Property 4.1

For $g, h \in U$, the function L_{gh} is defined by

$$L_{gh}(d) = \sum_{k=1}^{n-1} l_{gh}(d - kT), \text{ for } d \in [X_g, X - X_h], \tag{A.1}$$

where $l_{gh}(x)$ denotes the length of $(0, X_g] \cap (x, x + X_h]$, for any $x \in \mathbb{R}$,

$$l_{gh}(x) = \begin{cases} 0, & x \le -X_h \\ x + X_h, & -X_h < x \le \min(X_g - X_h, 0) \\ \min(X_g, X_h), & \min(X_g - X_h, 0) < x \le \max(X_g - X_h, 0) \\ X_g - x, & \max(X_g - X_h, 0) < x \le X_g \\ 0, & x > X_g. \end{cases} \tag{A.2}$$

Figures A.1 and A.2 show the shape of the graph of l_{gh} for $X_g > X_h$ and $X_g < X_h$ respectively. In the case where $X_g = X_h$, the top of the graph is not a horizontal line but the single point $(0, X_g)$.

Figure A.1 Graph of l_{gh} with $X_g = 5, X_h = 3$

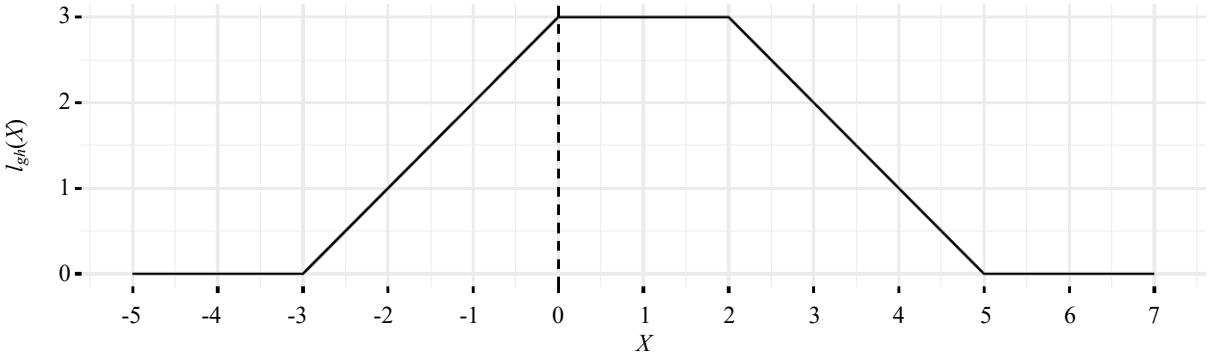
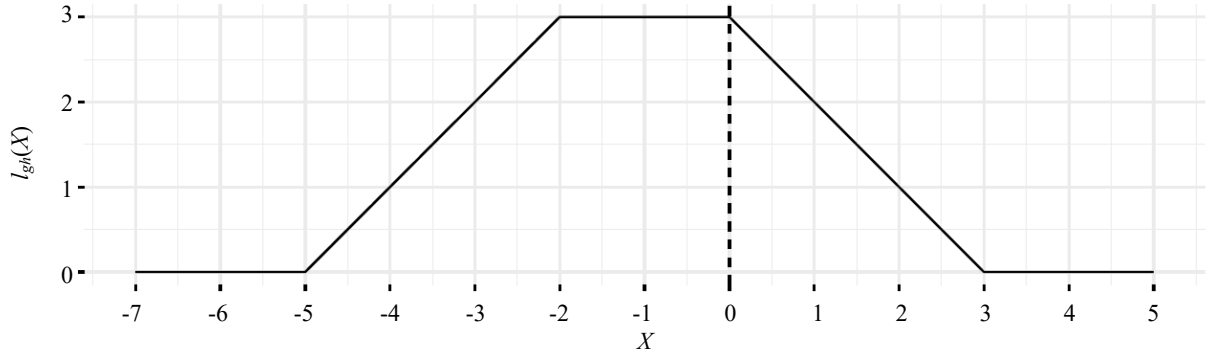


Figure A.2 Graph of l_{gh} with $X_g = 3, X_h = 5$



The graph of L_{gh} is constructed by summing $n-1$ shifted graphs of l_{gh}, where each shift being kT units to the right, $k = 1,\ldots,n-1$. For the shape of the graph of L_{gh}, a distinction is made between $T \geq X_g + X_h$ and $T < X_g + X_h$. In the first case, there is no overlap between positive parts of the shifted graphs of l_{gh}, allowing these shifted graphs to be placed side by side. In the second case, positive parts of two consecutive shifted graphs of l_{gh} overlap, resulting in the minimum of L_{gh} being equal to $X_g + X_h - T$, which also represents the minimum at the boundaries of the domain of L_{gh}. Figures A.3 and A.4 illustrate the shape of the graph of L_{gh} for $T \geq X_g + X_h$ and $T < X_g + X_h$ respectively. The parameters in Figure A.3 are $X_g = 3,\ X_h = 2, n = 3, T = 6$. In Figure A.4, they are $X_g = 5,\ X_h = 4, n = 3, T = 7$.

The function l_{gh} is continuous and piecewise linear. By elementary algebra, l_{gh} can be expressed as a linear combination of positive parts of linear functions,

$$l_{gh}(x) = \mathbb{P}(x + X_h) - \mathbb{P}(x) - \mathbb{P}(x - X_g + X_h) + \mathbb{P}(x - X_g),\ x \in \mathbb{R}, \tag{A.3}$$

where $\mathbb{P}$ denotes the positive part, $\mathbb{P}(x) = \max(x, 0),\ x \in \mathbb{R}$. Combining formulas (A.1) and (A.3) yields the explicit expression for $L_{gh}(d)$, for $d \in [X_g, X - X_h]$,

$$L_{gh}(d) = \sum_{k=1}^{n-1} \{\mathbb{P}(d - kT + X_h) - \mathbb{P}(d - kT) - \mathbb{P}(d - kT - X_g + X_h) + \mathbb{P}(d - kT - X_g)\}. \tag{A.4}$$

It follows from formula (A.4) that L_{gh} is symmetric with respect to the midpoint $(X + X_g - X_h)/2$ of its domain, using $X = nT$, changing summation variable k into $l = n - k$, and $\mathbb{P}(x) = \mathbb{P}(-x) + x$. ■

Figure A.3 Graph of L_{gh} with $n = 3$ and $T > X_g + X_h$

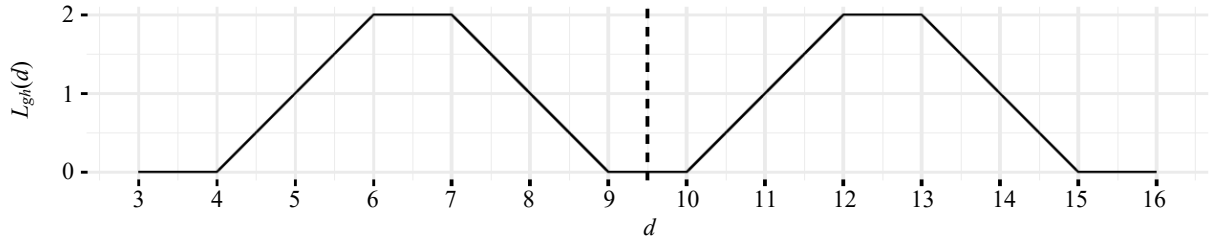
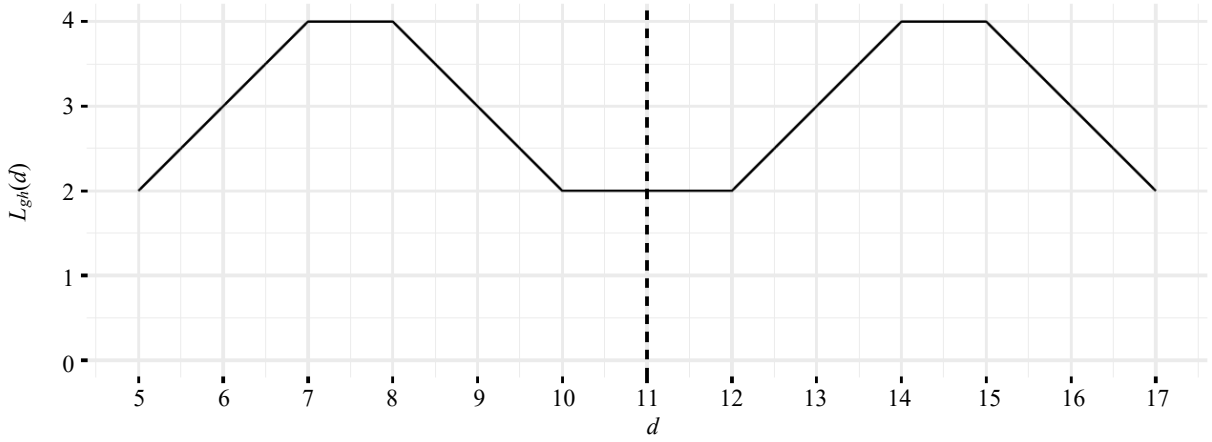


Figure A.4 Graph of L_{gh} with $n = 3$ and $T < X_g + X_h$



B. Proof of Theorem 5.1

The starting point is the exact formula for π_{gh} in Theorem 4.2. Let $N \geq 8$. Then the sum of that formula has at least four terms, the first three of which can be easily calculated exactly. For the higher terms approximations are developed. So consider the summand for $l \geq 5$,

$$t_{ghl} = \frac{1}{\binom{N-2}{l-2}} \sum_{W \in V(g,h,l)} L_{gh}(X_g + \gamma(W)). \tag{B.1}$$

Define $D_{ghl} = \{X_g + \gamma(W) \mid W \in V(g,h,l)\}$. For each $d \in D_{ghl}$, let $n_{ghl}(d)$ denote the number of sets $W \in V(g,h,l)$ for which $X_g + \gamma(W) = d$ and let $f_{ghl}(d) = n_{ghl}(d) \Big/ \binom{N-2}{l-2}$. Then

$$t_{ghl} = \sum_{d \in D_{ghl}} f_{ghl}(d) \cdot L_{gh}(d). \tag{B.2}$$

As $V(g,h,l)$ has $\binom{N-2}{l-2}$ elements, the function f_{ghl} represents the probability distribution of the possible distances between units g and h, conditional on h occupying the l-th position. Thus t_{ghl} is the expected value of L_{gh} under this condition.

Each set $W \in V(g,h,l)$ can be interpreted as a simple random sample without replacement of size $l-2$ from U_{gh}. Accordingly, $\gamma(W)$ is the sample total of the unit sizes in W. By the Central Limit Theorem for

simple random sampling without replacement from finite populations (Erdős and Rényi, 1959), $\gamma(W)$ is asymptotically normally distributed, with mean and variance

$$E(\gamma(W)) = (l-2)\cdot\mu_{gh} \text{ and } \mathrm{Var}(\gamma(W)) = \frac{(l-2)\cdot(N-l)}{N-3}\cdot\sigma_{gh}^2, \tag{B.3}$$

where μ_{gh} and σ_{gh}^2 denote, respectively, the mean and variance of the X- values of units in U_{gh},

$$\mu_{gh} = \frac{1}{N-2}(X - X_g - X_h) \text{ and } \sigma_{gh}^2 = \frac{1}{N-2}\sum_{i\in U_{gh}}(X_i - \mu_{gh})^2. \tag{B.4}$$

The Central Limit Theorem applies provided that the sampling fraction $(l-2)/(N-2)$ is not too close to 0 or 1, and that no unit size dominates the total population variance. Hence, for large l and N, the probability distribution f_{ghl} can be approximated by the probability density function φ_{ghl} of the normal distribution with mean μ_{ghl} and variance σ_{ghl}^2 given by

$$\mu_{ghl} = X_g + (l-2)\cdot\mu_{gh} \text{ and } \sigma_{ghl}^2 = \frac{(l-2)\cdot(N-l)}{N-3}\cdot\sigma_{gh}^2. \tag{B.5}$$

That is,

$$\varphi_{ghl}(x) = \frac{1}{\sigma_{ghl}\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\left(\frac{x-\mu_{ghl}}{\sigma_{ghl}}\right)^2}, \quad x\in\mathbb{R}. \tag{B.6}$$

As each $d\in D_{ghl}$ is in the interval $[X_g, X - X_h]$, the sum in (B.2) can be approximated by the integral

$$\tilde{t}_{ghl} = \int_{X_g}^{X-X_h}\varphi_{ghl}(x)\cdot L_{gh}(x)\,dx. \tag{B.7}$$

Next, the integral in (B.7) is evaluated. For readability, the subscript ghl is temporarily omitted, writing $\varphi(x)=\varphi_{ghl}(x)$, $\mu=\mu_{ghl}$, and $\sigma^2=\sigma_{ghl}^2$. The cumulative distribution function Φ, associated with φ, can be expressed in terms of the Gauss error function erf,

$$\Phi(x) = \frac{1}{2}\left(1+\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right), \qquad \mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt, \quad x,z\in\mathbb{R}. \tag{B.8}$$

By the definitions of φ and Φ their derivatives are

$$\Phi'(x) = \varphi(x), \qquad \varphi'(x) = \frac{\mu-x}{\sigma^2}\varphi(x), \quad x\in\mathbb{R}. \tag{B.9}$$

From these, the following antiderivatives are obtained,

$$\int\varphi(x)\,dx = \Phi(x)+C, \quad \int x\cdot\varphi(x)\,dx = \mu\cdot\Phi(x)-\sigma^2\varphi(x)+C, \quad C\in\mathbb{R}. \tag{B.10}$$

The remainder of the proof assumes that $X_g \geq X_h$. For $X_g < X_h$, the calculations are analogous and lead to the same result. In Appendix A it was shown that the shape of the graph of the function L_{gh} depends on the position of $X_g + X_h$ relative to T. First consider $T \geq X_g + X_h$. As $X_g \geq X_h$, it follows from formula (A.2) that for $x,\lambda\in\mathbb{R}$,

$$l_{gh}(x-\lambda) = \begin{cases} 0, & x \le \lambda - X_h \\ x-\lambda+X_h, & \lambda - X_h < x \le \lambda \\ X_h, & \lambda < x \le \lambda + X_g - X_h \\ -x+\lambda+X_g, & \lambda + X_g - X_h < x \le \lambda + X_g \\ 0, & x > \lambda + X_g. \end{cases} \tag{B.11}$$

Since $T \ge X_g + X_h,$ it follows that $X_g \le kT \le kT + X_g \le X - X_h,$ for all $k \in \{1,\ldots,n-1\},$ and so

$$\begin{aligned} I(k) = \int_{X_g}^{X-X_h} \varphi(x) \cdot l_{gh}(x-kT)\,dx &= \int_{kT-X_h}^{kT} \varphi(x) \cdot (x - kT + X_h)\,dx \\ &+ X_h \cdot \int_{kT}^{kT+X_g-X_h} \varphi(x)\,dx + \int_{kT+X_g-X_h}^{kT+X_g} \varphi(x) \cdot (-x + kT + X_g)\,dx. \end{aligned} \tag{B.12}$$

Applying formula (B.10) and performing some arithmetic yields

$$I(k) = \mathrm{F}(kT) - \mathrm{F}(kT - X_h) + \mathrm{F}(kT + X_g - X_h) - \mathrm{F}(kT + X_g), \text{ with} \tag{B.13}$$

$$\mathrm{F}(x) = (\mu - x)\,\Phi(x) - \sigma^2\,\varphi(x). \tag{B.14}$$

As $L_{gh}(x) = \sum_{k=1}^{n-1} l_{gh}(x-kT),$ the desired integral can be evaluated,

$$\int_{X_g}^{X-X_h} \varphi(x) \cdot L_{gh}(x)\,dx = \sum_{k=1}^{n-1} \int_{X_g}^{X-X_h} \varphi(x) \cdot l_{gh}(x-kT)\,dx = \sum_{k=1}^{n-1} I(x). \tag{B.15}$$

Next consider $T < X_g + X_h.$ Since $T - X_h < X_g,$ the lower bound in the second integral of formula (B.12) must be replaced by X_g for $k = 1.$ Furthermore, since $(n-1)T + X_g > X - X_h,$ the upper bound in the last integral of formula (B.12) must be replaced by $X - X_h$ for $k = n-1.$ This implies that from formula (B.15) the following must be subtracted,

$$\begin{aligned} &\int_{T-X_h}^{X_g} \varphi(x) \cdot (x - T + X_h)\,dx + \int_{X-X_h}^{(n-1)T+X_g} \varphi(x) \cdot (-x + (n-1)\,T + X_g)\,dx \\ &\quad = \mathrm{F}(X_g) - \mathrm{F}(T - X_h) + \mathrm{F}(X - X_h) - \mathrm{F}((n-1)\,T + X_g) \\ &\qquad - (X_g + X_h - T) \cdot (\Phi(X - X_h) - \Phi(X_g)). \end{aligned} \tag{B.16}$$

Combining (B.13)-(B.16) yields the desired expression for $\tilde{t}_{ghl} = \int_{X_g}^{X-X_h} \varphi_{ghl}(x) \cdot L_{gh}(x)\,dx.$ ■

Remark

To illustrate that this approximation also works well for small values of $N,$ we provide a few examples. In the left panel of Figure B.1, the histogram represents the empirical probability distribution $f_{ghl},$ while the red curve shows its normal approximation $\varphi_{ghl}.$ Here, the population U consists of 13 units with sizes $X_k = k + 4,$ for $k = 1,2,\ldots,13,$ and the triplet (g,h,l) is fixed at $(1,2,6).$ In the right panel of Figure B.1, the unit sizes have been replaced by the values 16, 20, 6, 8, 19, 8, 17, 13, 12, 5, 11, 18, 15, generated from the discrete Uniform distribution on the interval [5, 20].

If the sizes X_k are generated by a normally or even skewed distributed random variable, f_{ghl} can still be reasonably approximated by φ_{ghl} which is illustrated in Figure B.2. Its left panel relates to U consisting of 13 units with sizes 10, 11, 6, 9, 8, 10, 8, 7, 13, 13, 11, 10, 14 generated by rounding 13 draws from the Normal distribution with mean 10 and variance 4. In the right panel the sizes have been replaced by values 20, 48, 17, 111, 82, 8, 72, 4, 15, 9, 59, 30, 35, generated by rounding 13 draws from the Chi-squared distribution with three degrees of freedom.

Figure B.1 Probability distribution f_{ghl} and approximation φ_{ghl}, where $X_k = k + 4$ (left) and X_k from a Uniform distribution (right)

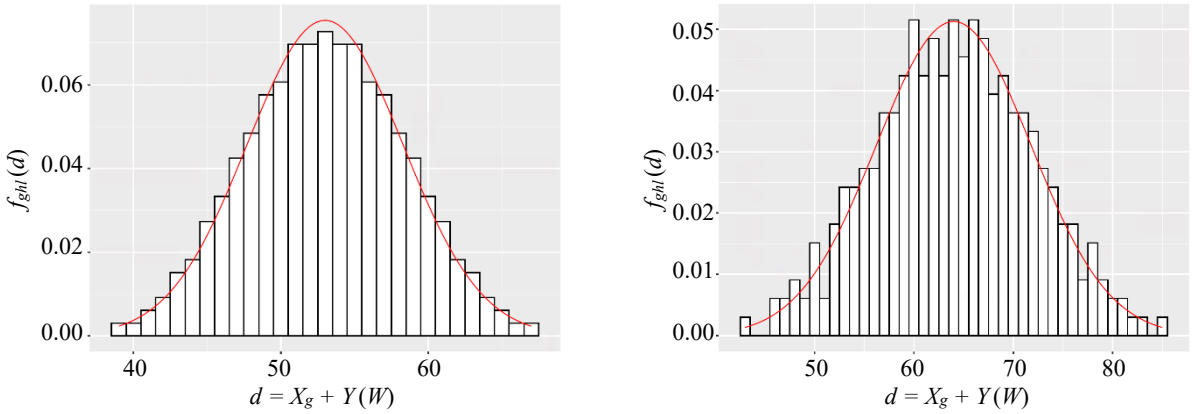
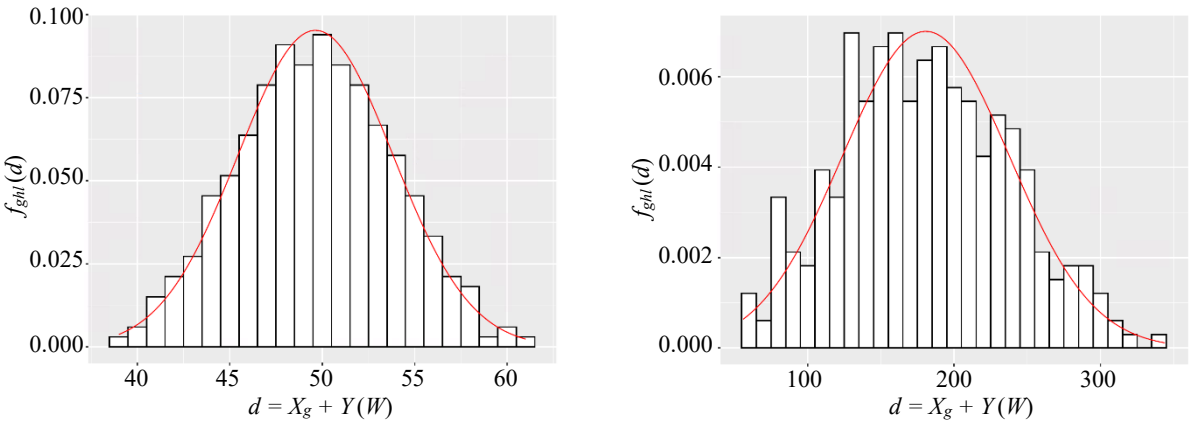


Figure B.2 Probability distribution f_{ghl} approximated by φ_{ghl} where X_k from a Normal distribution (left) and from a Chi-squared distribution (right)



C. Proof of Theorem 5.2

Minimizing (5.3) with boundary conditions (5.2) can be solved using the Lagrange multiplier method. The corresponding Lagrangian equals

$$\sum_{g\in U}\sum_{g<h\in U}(\alpha_{gh}-1)^2+\sum_{g\in U}\lambda_g\left(\sum_{h\in U_g}\alpha_{gh}\cdot\tilde{\pi}_{gh}-(n-1)\pi_g\right).$$

By setting the partial derivatives of the Lagrangian with respect to α_{gh} to zero, the α_{gh} can be expressed in terms of the Lagrange multipliers λ_g,

$$2(\alpha_{gh}-1)+(\lambda_g+\lambda_h)\,\tilde{\pi}_{gh}\;=\;0\;\text{ so }\;\alpha_{gh}\;=\;1-\frac{1}{2}(\lambda_g+\lambda_h)\,\tilde{\pi}_{gh}.$$

By substituting these values of α_{gh} into boundary conditions (5.2), N equations for the N unknown Lagrange multipliers arise. In matrix notation these equations can be written as $\mathbf{A}\cdot\boldsymbol{\lambda}=\mathbf{b}$, with $\mathbf{A}$ the $N\times N$- matrix with entries $A_{gh}=\tilde{\pi}_{gh}^2$ for $g\neq h\in U$, and $A_{gg}=\sum_{h\in U_g}\tilde{\pi}_{gh}^2$ for $g\in U$, and $\boldsymbol{\lambda}$ the vector of Lagrange multipliers, and $\mathbf{b}$ the N- vector with $b_g=2\left(\sum_{h\in U_g}\tilde{\pi}_{gh}-(n-1)\pi_g\right)$ for $g\in U$. ■

References

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13. https://doi.org/10.1111/j.1467-842X.1963.tb00132.x.

Brewer, K.R.W. and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29(2), 189-196. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2003002/article/6778-eng.pdf.

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390. Available at: https://www.jstor.org/stable/2282827#metadata_info_tab_contents.

Deville, J.-C. (1993). *Estimation de la Variance pour les Enquêtes en Deux Phases*. Internal handwritten note. France: INSEE. [In French].

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164. Available at: https://www.jstor.org/stable/2985777#metadata_info_tab_contents.

Erdős, P. and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Magyar Tudoanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei*, 4, 49-57. Available at: https://real.mtak.hu/200900/1/cut_MATKUTINT_4_1_1959_pp49_-_61.pdf.

Gray, G.B. (1971). Joint probabilities of selection of units in systematic samples. *Proceedings of the American Statistical Association*, 271-276. Available at: http://www.asasrms.org/Proceedings/y1972/Joint%20Probabilities%20Of%20Selection%20Of%20Units%20In%20Systematic%20Samples.pdf.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523. Available at: https://www.jstor.org/stable/2238287#metadata_info_tab_contents.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374. Available at: https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-33/issue-2/Sampling-with-Unequal-Probabilities-and-without-replacement/10.1214/aoms/1177704564.full.

Hidiroglou, M.A. and Gray, G.B. (1975). A computer algorithm for joint probabilities of selection (Systematic PPS sampling). *Survey Methodology*, 1(1), 99-108. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1975001/article/00002-eng.pdf.

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. Available at: https://www.jstor.org/stable/pdf/2280784.pdf.

Knottnerus, P. (2003). Sample survey theory: Some Pythagorean perspectives. New York: *Springer-Verlag*. Request full-text pdf at https://www.researchgate.net/publication/4741910_Sample_Survey_Theory_Some_PythagoreanPerspectives_Paul_Knottnerus.

Knottnerus, P. (2011). On the efficiency of randomized probability proportional to size sampling. *Survey Methodology*, 37(1), 95-102. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2011001/article/11450-eng.pdf.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439. Available at: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/a-note-on-the-hartley-rao-variance-estimator.pdf.

Matei, A. and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4), 543-570. Available at: http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/evaluation-of-variance-approximations-and-estimators-in-maximum-entropy-sampling-with-unequal-probability-and-fixed-sample-size.pdf.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of Indian Society for Agricultural Statistics*, 5, 119-127.

Thompson, M.E. and Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Survey Methodology*, 34(1), 3-10. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10613-eng.pdf.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Series B, 15, 235-261. Available at: https://www.jstor.org/stable/pdf/2983772.pdf.