

Techniques d'enquête

Formules exactes et d'approximation pour les probabilités d'inclusion de deuxième ordre dans l'échantillonnage systématique randomisé avec probabilités inégales et sans remise

par Kees van Berkel et Erwin Vondenhoff

Date de diffusion : le 29 juin 2026



Statistique
Canada

Statistics
Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- | | |
|---|----------------|
| • Service de renseignements statistiques | 1-800-263-1136 |
| • Service national d'appareils de télécommunications pour les malentendants | 1-800-363-7629 |
| • Télécopieur | 1-514-283-9350 |

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « [Normes de service à la clientèle](#) ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Publication autorisée par la ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par la ministre de l'Industrie, 2026

L'utilisation de la présente publication est assujettie aux modalités de l'[entente de licence ouverte](#) de Statistique Canada.

Une [version HTML](#) est aussi disponible.

This publication is also available in English.

Formules exactes et d'approximation pour les probabilités d'inclusion de deuxième ordre dans l'échantillonnage systématique randomisé avec probabilités inégales et sans remise

Kees van Berkel et Erwin Vondenhoff¹

Résumé

L'échantillonnage avec probabilité proportionnelle à la taille est largement utilisé par les organismes nationaux de statistique. Dans leur cas, les unités de la population sont sélectionnées avec des probabilités proportionnelles à une variable auxiliaire. Les formules de variance dans ce genre de plan requièrent à la fois des probabilités d'inclusion de premier et de deuxième ordre. Le calcul des probabilités d'inclusion de deuxième ordre est particulièrement difficile pour les grandes populations et a fait l'objet de nombreuses études. Le présent article expose de nouvelles formules exactes et d'approximation pour les probabilités d'inclusion de deuxième ordre dans l'échantillonnage systématique randomisé avec probabilités inégales et sans remise.

Mots-clés : Échantillonnage avec probabilité proportionnelle à la taille (PPT); estimateur de Horvitz-Thompson; estimation de la variance.

1. Introduction

Les organismes nationaux de statistique utilisent souvent des plans d'échantillonnage dans lesquels les unités sont sélectionnées avec des probabilités proportionnelles à une variable auxiliaire. Un exemple courant est la sélection de municipalités avec des probabilités proportionnelles à la taille de leur population. Ce genre de plan avec probabilité proportionnelle à la taille (PPT) présente un intérêt particulier, car il améliore l'efficacité lorsque la variable auxiliaire est fortement liée à la variable d'intérêt de l'enquête.

Lors de l'estimation de la variance, les probabilités d'inclusion jouent un rôle central. L'expression exacte de la variance de l'estimateur de Horvitz-Thompson requiert non seulement les probabilités d'inclusion de premier ordre, mais aussi celles de deuxième ordre, désignées par π_{gh} et représentant la probabilité que deux unités g et h soient toutes deux incluses dans l'échantillon. Bien que dans un certain nombre d'approximations d'estimateurs de la variance, on n'utilise que les probabilités d'inclusion de premier ordre (voir Matei et Tillé, 2005), il demeure important de mettre au point des formules exactes et d'approximation pour π_{gh} . Les formules exactes permettent l'estimation de la variance sans approximation et servent de référence pour évaluer les méthodes d'approximation.

Le présent article porte sur l'échantillonnage systématique randomisé avec PPT. Pour ce plan, Hartley et Rao (1962) ont établi des expressions analytiques exactes pour π_{gh} dans le cas particulier d'une taille d'échantillon égale à deux et d'une taille de population égale à trois ou à quatre. Ils ont mentionné qu'à mesure que la taille de la population augmente, l'évaluation exacte de π_{gh} devient de plus en plus laborieuse et les

1. Kees van Berkel et Erwin Vondenhoff, Statistics Netherlands, Division of Data Collection, C.P. 4481, 6401 CZ Heerlen, Pays-Bas. Courriel : cam.vanberkel@cbs.nl.

expressions résultantes se complexifient considérablement. Connor (1966) a mis au point une procédure générale pour le calcul de π_{gh} pour des tailles d'échantillon et de population arbitraires, étendant ainsi les résultats de Hartley et Rao. Cette procédure a été améliorée par Gray (1971) et ensuite mise en œuvre dans un algorithme informatique par Hidioglou et Gray (1975). Cependant, les calculs qui y sont associés demeurent exigeants, et la complexité croît de façon exponentielle avec la taille de la population.

Pour pallier cela, des chercheurs ont étudié des approximations de π_{gh} . Hartley et Rao (1962) ont proposé une approximation asymptotique, et Kott (2005) a montré qu'elle donnait de bons résultats avec l'estimateur de variance de Horvitz-Thompson. Hájek (1964) a présenté une approche fondée sur l'échantillonnage de Poisson réjectif, tandis que Brewer et Donadio (2003) et Knottnerus (2003, 2011) ont examiné des approximations de forme particulière, conformément aux travaux antérieurs de Brewer (1963) et de Durbin (1967) pour des échantillons de taille deux. Thompson et Wu (2008) ont proposé une approche par simulation pour l'estimation de π_{gh} .

Dans le présent article, de nouvelles formules exactes et d'approximation pour π_{gh} sont dérivées et comparées aux approches existantes dans la littérature. L'article est structuré comme suit. La section 2 présente le plan d'échantillonnage systématique randomisé avec PPT et la section 3 fournit une notation et des définitions supplémentaires. La section 4 contient de nouvelles formules exactes pour π_{gh} et la section 5 contient de nouvelles formules d'approximation. À la section 6, nous passons en revue les résultats connexes de la littérature et, à la section 7, nous comparons la précision de nos approximations à celle des approximations existantes. La section 8 illustre l'application des expressions de π_{gh} à l'estimation de la variance de l'estimateur de Horvitz-Thompson. Enfin, la section 9 conclut l'article.

2. Échantillonnage systématique randomisé avec probabilités inégales et sans remise

Considérons une population U composée de N unités. On associe à chaque $g \in U$ un nombre $X_g > 0$ qui représente sa taille. L'objectif du plan de sondage est de tirer un échantillon de n unités distinctes de sorte que la probabilité π_g de sélectionner l'unité $g \in U$ dans l'échantillon soit proportionnelle à sa taille X_g . Voici une procédure d'échantillonnage correspondante décrite par Hartley et Rao (1962). On dispose les unités dans un ordre aléatoire, désigné par $1, 2, \dots, N$. On calcule les sous-totaux $T_k = \sum_{i=1}^k X_i$, $k = 1, 2, \dots, N$. Ainsi, $T_N = \sum_{i=1}^N X_i = X$ est la taille totale de U . L'intervalle $(0, T_N]$ est divisé en N sous-intervalles $(0, T_1], (T_1, T_2], \dots, (T_{N-1}, T_N]$ correspondant aux unités $1, 2, \dots, N$, respectivement. L'intervalle d'échantillonnage T est défini par $T = X/n$. Un départ aléatoire b est sélectionné selon la loi uniforme sur $(0, T]$. L'unité k appartient à l'échantillon si et seulement s'il existe un entier non négatif i tel que $T_{k-1} < b + iT \leq T_k$, où $T_0 = 0$. Une condition nécessaire pour cette procédure d'échantillonnage pour sélectionner n unités est que $X_g < T$ pour toutes les unités $g \in U$. Par conséquent, seules les tailles satisfaisant à cette condition sont considérées. Les probabilités d'inclusion de premier ordre satisfont $\pi_g = X_g / T$ pour toutes les unités $g \in U$.

3. Notation et définitions supplémentaires

Supposons que $g, h \in U$, où $g < h$, sont ordonnées selon la procédure d'échantillonnage de la section 2 et que les intervalles correspondants sont $(T_{g-1}, T_g]$ et $(T_{h-1}, T_h]$. La distance entre g et h est définie comme la longueur de l'écart entre les intervalles, soit $T_{h-1} - T_g$.

Pour tout sous-ensemble W de U , le nombre de ses unités est désigné par $|W|$. La taille de W , désignée par $\gamma(W)$, est définie par la somme des tailles de ses unités, $\gamma(W) = \sum_{i \in W} X_i$ et $\gamma(\emptyset) = 0$.

Pour $g \in U$, on suppose que $U_g = U \setminus \{g\}$, c'est-à-dire la population U de laquelle l'unité g a été retirée. Pour $g, h \in U$, on suppose que $U_{gh} = U \setminus \{g, h\}$, c'est-à-dire la population U de laquelle les unités g et h ont été retirées.

Pour $g, h \in U$ et $l \in \{2, \dots, N\}$, on définit $V(g, h, l) = \{W \subset U_{gh} \mid |W| = l - 2\}$, l'ensemble de tous les sous-ensembles de U_{gh} contenant exactement $l - 2$ unités. Dans les cas limites où $l = 2$ ou $l = N$, on a $V(g, h, 2) = \{\emptyset\}$ et $V(g, h, N) = U_{gh}$.

4. Formules exactes pour les probabilités d'inclusion de deuxième ordre

Pour $2 \leq n \leq N$, la probabilité π_{gh} que les deux unités g et h soient sélectionnées dans l'échantillon est appelée « probabilité d'inclusion de deuxième ordre de g et de h ». Selon le plan de sélection, la sélection de g et de h dans l'échantillon dépend de la position de g par rapport à h . Par conséquent, sans perte de généralité, on peut supposer que l'unité g est placée en première position. Soit Ω_g , l'ensemble de tous les ordonnancements possibles de U où g est en première position. Pour $\omega \in \Omega_g$, soit T_{h-1}^ω , la borne gauche de l'intervalle correspondant à h et soit $L_{gh}(d)$, la longueur de $(0, X_g] \cap \bigcup_{l=1}^{n-1} (d - lT, d - lT + X_h]$ pour $d \in [X_g, X - X_h]$. La probabilité conditionnelle que les deux unités g et h soient sélectionnées dans l'échantillon étant donné l'ordonnement $\omega \in \Omega_g$ est égale à $L_{gh}(T_{h-1}^\omega) / T$. Étant donné que Ω_g possède $(N-1)!$ éléments, tous équiprobables, la loi de probabilité totale donne :

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

Le calcul de cette formule peut être simplifié en séparant Ω_g en sous-ensembles non vides et disjoints assurant que T_{h-1}^ω est constant à l'intérieur de chaque sous-ensemble.

Théorème 4.1

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

Démonstration. L'unité g étant placée en première position, l'unité h peut occuper les positions $2, \dots, N$. Supposons que h occupe la position l pour un certain $l \in \{2, \dots, N\}$. Alors, $l-2$ unités de U_{gh} sont placées entre g et h , et les $N-l$ unités restantes, après h . On remarque que T_{h-1}^ω ne dépend ni de l'ordonnement des $l-2$ unités entre g et h , ni de l'ordonnement des $N-l$ unités après h . Le nombre total de permutations de ces unités est $(l-2)!(N-l)!$. De plus, $T_{h-1}^\omega = X_g + \gamma(W)$, où W désigne l'ensemble des $l-2$ unités entre g et h . Donc, $T(N-1)!\pi_{gh} = \sum_{l=2}^N \sum_{W \in V(g,h,l)} (l-2)!(N-l)! L_{gh}(X_g + \gamma(W))$. ■

Étant donné les $\binom{N-2}{l-2}$ éléments dans $V(g,h,l)$, la somme du Théorème 4.1 comprend 2^{N-2} termes. L'addition de ces termes entraîne une complexité en temps exponentielle à mesure que la taille de population N augmente. De la Propriété 4.1 ci-après, il s'ensuit que l'évaluation de la fonction L_{gh} augmente de façon linéaire en fonction de la taille d'échantillon n . Ainsi, la complexité de calcul de la somme du Théorème 4.1 est d'ordre $n \cdot 2^N$.

La fonction L_{gh} peut être exprimée comme une somme de parties positives de fonctions linéaires et est symétrique par rapport au point médian de son domaine.

Propriété 4.1

Pour $g, h \in U$ et $d \in [X_g, X - X_h]$, on a :

- a) $L_{gh}(d) = \sum_{k=1}^{n-1} \{\mathbb{P}(d - kT + X_h) - \mathbb{P}(d - kT) - \mathbb{P}(d - kT - X_g + X_h) + \mathbb{P}(d - kT - X_g)\}$, où $\mathbb{P}(x) = \max(x, 0)$, la partie positive de $x \in \mathbb{R}$.
- b) $L_{gh}(d) = L_{gh}(X + X_g - X_h - d)$.

La démonstration de la Propriété 4.1 est présentée à l'annexe A.

En utilisant la symétrie de L_{gh} et les coefficients binomiaux, la somme dans la formule de π_{gh} du Théorème 4.1 peut être réduite.

Théorème 4.2

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

où $\lfloor x \rfloor$ est la partie entière de x et $\delta_{i,j}$ est le delta de Kronecker qui vaut 1 si $i = j$ et 0 sinon.

Démonstration. Pour chaque $W \in V(g,h,l)$, son complément W^C dans U_{gh} est le sous-ensemble de U_{gh} contenant toutes les unités n'appartenant pas à W . Ainsi, $W^C \in V(g,h,N-l+2)$ et $X_g + X_h + \gamma(W) + \gamma(W^C) = X$. À partir de cette démonstration et de b) de la Propriété 4.1, il s'ensuit que $L_{gh}(X_g + \gamma(W)) = L_{gh}(X_g + \gamma(W^C))$ et donc :

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

À partir de cela et de la propriété $\binom{n}{k} = \binom{n}{n-k}$ des coefficients binomiaux, il s'ensuit que le l^e terme de la somme sur l dans le Théorème 4.1 est égal au $(N-l+2)^e$ terme pour $l = 2, \dots, N$. Ainsi, la somme sur l est égale au double de la somme sur la première moitié des termes lorsque N est impair et au double de la somme sur les $N/2$ premiers termes plus le terme central lorsque N est pair. ■

L'application du Théorème 4.2 et de a) de la Propriété 4.1 rend le calcul de π_{gh} simple. Toutefois, pour de grandes valeurs de N , le calcul devient laborieux. Bien que le nombre de termes de la somme du Théorème 4.2 ait été réduit de moitié par rapport à la somme du Théorème 4.1, la somme comporte toujours 2^{N-3} termes. Ainsi, le temps de calcul nécessaire croît encore de façon exponentielle en fonction de la taille de la population N et demeure d'ordre $n \cdot 2^N$.

L'expression de π_{gh} du Théorème 4.2 peut être simplifiée davantage, notamment lorsque certaines tailles d'unité dans U sont identiques. À cette fin, la notation vectorielle suivante est utilisée pour $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ et $\mathbf{k}, \mathbf{n} \in \mathbb{N}_0^m$:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^m a_i b_i, |\mathbf{a}| = \sum_{i=1}^m |a_i|, \binom{\mathbf{n}}{\mathbf{k}} = \prod_{i=1}^m \binom{n_i}{k_i}, 0 \leq \mathbf{k} \leq \mathbf{n} \text{ signifie } 0 \leq k_i \leq n_i \text{ pour toute } i.$$

Théorème 4.3

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

où les lettres en gras désignent des vecteurs de dimension égale au nombre de tailles distinctes d'unité dans U_{gh} . Le vecteur $\mathbf{X}_{gh}$ contient ces tailles, le vecteur $\mathbf{n}_{gh}$ comptabilise le nombre d'occurrences de chaque taille dans U_{gh} et le vecteur $\mathbf{k}$ contient des entiers non négatifs.

Démonstration. Soient $g, h \in U$ et soit m , le nombre de tailles distinctes d'unité dans U_{gh} . Après réindexation, au besoin, ces tailles sont $X_1, X_2, \dots, X_m$. Pour chaque sous-ensemble W de U_{gh} , sa taille $\gamma(W)$ est égale à $\mathbf{k} \cdot \mathbf{X}_{gh}$, où $\mathbf{k}$ est le vecteur à m dimensions dont la i^e composante est égale au nombre d'unités dans W ayant la taille X_i . Pour $l \geq 2$, le nombre d'ensembles $W \in V(g, h, l)$ contenant exactement k_i unités de taille X_i pour $i = 1, \dots, m$ est égal à $\binom{\mathbf{n}_{gh}}{\mathbf{k}}$, sous réserve que $0 \leq \mathbf{k} \leq \mathbf{n}_{gh}$ et que $|\mathbf{k}| = |W| = l - 2$. ■

S'il y a m tailles distinctes d'unité dans U_{gh} , la complexité de calcul de l'évaluation de π_{gh} selon le Théorème 4.3 est polynomiale d'ordre $n \cdot N^m$ pour une valeur de m fixe. Pour les petites valeurs de m , cela représente une réduction importante par rapport à la complexité exponentielle $n \cdot 2^N$ du Théorème 4.2. Par conséquent, le calcul exact de π_{gh} n'est réalisable avec le Théorème 4.2 que pour des petites valeurs de N et avec le Théorème 4.3 pour des valeurs de N un peu plus grandes, à condition que la valeur de m soit petite. Pour les grandes valeurs de N , cependant, aucune des deux formules exactes n'est calculable en temps raisonnable. Des formules d'approximation sont donc présentées à la section suivante.

5. Formules d'approximation pour les probabilités d'inclusion de deuxième ordre

Si toutes les unités de U_{gh} sont de taille égale pour certaines $g, h \in U$, alors on peut calculer π_{gh} au moyen du Théorème 4.3. Par conséquent, dans la présente section, on suppose que les unités de U_{gh} ne sont pas toutes de même taille.

Théorème 5.1

Pour $N \geq 8$, l'approximation $\tilde{\pi}_{gh}$ de π_{gh} est donnée par :

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

où

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

avec

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

ϕ_{ghl} et Φ_{ghl} désignant, respectivement, la fonction de densité et la fonction de distribution cumulative de moyenne μ_{ghl} et de variance σ_{ghl}^2 .

La démonstration du Théorème 5.1 est présentée à l'annexe B.

La complexité des approximations du Théorème 5.1 est d'ordre $n \cdot N$, soit linéaire. Il s'agit d'une réduction substantielle par rapport à la complexité d'ordre $n \cdot 2^N$ des formules exactes des Théorèmes 4.1 et 4.2, et à la complexité d'ordre $n \cdot N^m$ de la formule exacte du Théorème 4.3, où m est le nombre de tailles distinctes d'unité dans U_{gh} .

Pour une taille d'échantillon n fixe, les probabilités d'inclusion exactes de deuxième ordre π_{gh} ont la propriété :

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

Les approximations $\tilde{\pi}_{gh}$ du Théorème 5.1 n'ont pas nécessairement cette propriété. Par conséquent, elles sont mises à l'échelle de manière que les approximations mises à l'échelle ont bien la propriété :

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

où les multiplicateurs α_{gh} satisfont $\alpha_{gh} = \alpha_{hg}$ pour toutes les $g \neq h \in U$. Il en résulte N équations pour $N(N-1)/2$ multiplicateurs inconnus. Pour $N > 3$, le nombre d'inconnues dépasse le nombre d'équations, de sorte que les multiplicateurs ne sont pas déterminés de façon unique. Pour remédier à cela, une condition supplémentaire est imposée. Les multiplicateurs doivent être aussi proches que possible de un. On formalise cette condition en minimisant :

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

sous les contraintes définies dans les équations (5.2).

Théorème 5.2

Pour $N \geq 8$, l'approximation π_{gh}^{BV} de π_{gh} est donnée par :

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

où $\tilde{\pi}_{gh}$ provient du Théorème 5.1 et $\alpha_{gh} = 1 - \frac{1}{2}(\lambda_g + \lambda_h) \tilde{\pi}_{gh}$. Dans ce cas-ci, λ est la solution de $\mathbf{A} \cdot \lambda = \mathbf{b}$, où $\mathbf{A}$ est la matrice $N \times N$ ayant des entrées $A_{gh} = \tilde{\pi}_{gh}^2$ pour $g \neq h \in U$ et $A_{gg} = \sum_{h \in U_g} \tilde{\pi}_{gh}^2$ pour $g \in U$, et $\mathbf{b}$ est le vecteur à N dimensions ayant des éléments $b_g = 2 \left(\sum_{h \in U_g} \tilde{\pi}_{gh} - (n-1) \pi_g \right)$ pour $g \in U$.

La démonstration du Théorème 5.2 est présentée à l'annexe C. En général, trouver une solution explicite à $\mathbf{A} \cdot \lambda = \mathbf{b}$ est difficile, mais cela n'est pas nécessaire pour nos besoins. Des méthodes numériques précises peuvent être utilisées pour résoudre les équations, et généralement la complexité est d'ordre N^3 . Les équations ne doivent être résolues qu'une seule fois pour calculer tous les facteurs d'échelle.

6. Expressions des probabilités d'inclusion de deuxième ordre dans la littérature

La littérature sur les probabilités d'inclusion de deuxième ordre dans l'échantillonnage systématique randomisé avec probabilités inégales et sans remise est abondante. En effet, comme cela a été mentionné en introduction, Hartley et Rao (1962), Connor (1966) et Gray (1971) ont élaboré des procédures pour le calcul des valeurs exactes de π_{gh} . La formule (2.4) de Connor fait intervenir une somme comportant le même nombre de termes que la somme figurant dans notre Théorème 4.1. Cependant, contrairement à notre expression explicite du terme de la somme que l'on obtient à l'aide de a) de la Propriété 4.1 de la fonction L_{gh} , Connor ne fournit pas de représentation explicite des termes individuels. La modification de la formule de Connor par Gray donne également une somme comportant autant de termes que dans notre Théorème 4.2,

mais sa formulation est procédurale plutôt qu'explicite. L'algorithme correspondant a ensuite été implanté dans un programme informatique par Hidirolou et Gray (1975).

Au fil des années, de nombreuses formules d'approximation ont été publiées. L'un des résultats particulièrement intéressants est l'approximation asymptotique de π_{gh} dérivée par Hartley et Rao (1962), formule (5.16) :

$$\pi_{gh}^{\text{HR}} = n(n-1) \left\{ \begin{aligned} & X_g X_h X^{-2} + (X_g^2 X_h + X_g X_h^2) X^{-3} - X_g X_h A X^{-4} + \\ & 2(X_g^3 X_h + X_g X_h^3 + X_g^2 X_h^2) X^{-4} - 3(X_g^2 X_h + X_g X_h^2) A X^{-5} + \\ & 3X_g X_h A^2 X^{-6} - 2X_g X_h B X^{-5} \end{aligned} \right\}, \quad (6.1)$$

où

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

Cette approximation est valable pour les valeurs de n fixes et les grandes valeurs de N , ou les petites valeurs de n/N , et est exacte jusqu'à l'ordre de N^{-4} . Brewer et Donadio (2003) ont proposé des approximations sous la forme :

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

Pour $c_g = c$, il découle de $\sum_{g \in U} \sum_{h \in U_g} \pi_{gh} = n(n-1)$ que :

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

S'inspirant du ratio de sommes de π_{gh} et de $\pi_g \pi_h$ pour $h \in U_g$, ils ont également proposé :

$$c_g = (n-1) / (n - \pi_g). \quad (6.4)$$

En développant l'approximation (6.1) de Hartley et Rao, ils ont obtenu :

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

Knottnerus (2003 et 2011) a retenu une expression différente pour c_g :

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

pour aboutir, pour $X_g, X_h < X/2$, à la formule d'approximation :

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

Cette approximation coïncide avec celles proposées par Brewer (1963) et Durbin (1967) pour les échantillons avec PPT de taille deux. Une autre formule d'approximation intéressante, proposée par Hájek (1964), est issue du plan de sondage de Poisson réjectif connexe :

$$\pi_{gh}^H = \pi_g \pi_h \{1 - (1 - \pi_g)(1 - \pi_h) / d\}, \quad (6.8)$$

lorsque $d = \sum_{i \in U} \pi_i (1 - \pi_i)$ a une grande valeur, ce qui suppose que n et $N - n$ ont une grande valeur. Thompson et Wu (2008) ont mis au point une approche par simulation pour le calcul des probabilités d'inclusion de deuxième ordre. Pour $n = 10$ et $N = 20$, les approximations de Hartley et Rao sont proches des approximations issues des simulations. Elles concordent jusqu'à la deuxième décimale pour la majorité des π_{gh} .

7. Précision des formules d'approximation pour les probabilités d'inclusion de deuxième ordre

Pour évaluer la précision des formules d'approximation pour les probabilités d'inclusion de deuxième ordre, on utilise l'erreur absolue moyenne (EAM) :

$$\text{EAM} = \frac{2}{N(N-1)} \sum_{g < h \in U} |\pi_{gh} - \hat{\pi}_{gh}|, \quad (7.1)$$

où $\hat{\pi}_{gh}$ est une approximation de la valeur exacte de π_{gh} . Étant donné que $\pi_{gh} = \pi_{hg}$, cette mesure de la qualité ne tient compte que des probabilités π_{gh} quand $g < h$. Dans les exemples suivants, les probabilités d'inclusion exactes de deuxième ordre ont été calculées à l'aide du Théorème 4.2. Diverses méthodes d'approximation, notamment celles attribuables à Van Berkel et Vondenhoff (BV), à Kottnerus (K), à Hartley et Rao (HR), à Hájek (H) et à Brewer et Donadio (BD), sont évaluées pour différentes distributions de population et différentes tailles d'échantillon. Pour les approximations de BV, le Théorème 5.2 a été appliqué et, pour les approximations de BD, les formules (6.2) et (6.3) ont été utilisées.

Exemple 7.1

On considère l'exemple de Brewer obtenu par extension, où la population est composée de 8 unités avec $X_k = k$ pour $k = 1, \dots, 8$. Cette population est semblable à celle correspondant à la figure B.1 de l'annexe B. Un échantillon systématique avec PPT de taille $n = 2$ est tiré. La taille de population X est égale à 36 et l'intervalle d'échantillonnage T vaut $X / n = 18$. Les probabilités d'inclusion exactes de deuxième ordre sont présentées dans la matrice (7.2) et sont arrondies à 5 décimales. Par souci d'exhaustivité, les probabilités d'inclusion de premier ordre $\pi_g = \pi_{gg}$ figurent sur la diagonale.

$$(\pi_{gh})_{g \leq h \in U} = \begin{pmatrix} 0,05556 & 0,00238 & 0,00423 & 0,00608 & 0,00794 & 0,00979 & 0,01164 & 0,01349 \\ & 0,11111 & 0,00794 & 0,01349 & 0,01534 & 0,02090 & 0,02275 & 0,02831 \\ & & 0,6676 & 0,02275 & 0,02460 & 0,03016 & 0,03571 & 0,04127 \\ & & & 0,22222 & 0,03016 & 0,03942 & 0,05053 & 0,05979 \\ & & & & 0,27778 & 0,05423 & 0,06534 & 0,08016 \\ & & & & & 0,33333 & 0,08016 & 0,09868 \\ & & & & & & 0,38889 & 0,12275 \\ & & & & & & & 0,44444 \end{pmatrix}. \quad (7.2)$$

Le Théorème 5.2 fournit les approximations des probabilités d'inclusion de deuxième ordre :

$$(\pi_{gh}^{BV})_{g < h \in U} = \begin{pmatrix} 0,00264 & 0,00441 & 0,00615 & 0,00789 & 0,00967 & 0,01148 & 0,01333 \\ & 0,00893 & 0,01339 & 0,01576 & 0,02025 & 0,02274 & 0,02741 \\ & & 0,02164 & 0,02471 & 0,03011 & 0,03549 & 0,04139 \\ & & & 0,03134 & 0,03963 & 0,05044 & 0,05964 \\ & & & & 0,05345 & 0,06482 & 0,07981 \\ & & & & & 0,08064 & 0,09959 \\ & & & & & & 0,12329 \end{pmatrix}. \quad (7.3)$$

Les approximations concordent avec les probabilités exactes jusqu'à au moins la deuxième décimale : 2 concordent jusqu'à la deuxième décimale, 20 jusqu'à la troisième et 6 jusqu'à la quatrième. L'EAM est de 0,000389. Le tableau 7.1 présente l'EAM $\times 10^6$ des différentes méthodes d'approximation pour toutes les tailles d'échantillon satisfaisant à la condition de non-autosélection selon laquelle l'intervalle d'échantillonnage T doit être supérieur à la taille X_k pour toutes les valeurs de k .

Tableau 7.1

EAM $\times 10^6$ des approximations des probabilités d'inclusion de deuxième ordre

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 et Donadio (BD); Van Berkel et Vondenhoff (BV); Hájek (H); Hartley et Rao (HR); Knottnerus (K); erreur absolue moyenne (EAM).

Exemple 7.2

On considère la population correspondant à la partie de droite de la figure B.1 de l'annexe B. Les tailles des 13 unités sont 16, 20, 6, 8, 19, 8, 17, 13, 12, 5, 11, 18 et 15, et sont générées au moyen de la loi uniforme discrète sur l'intervalle $[5, 20]$. Le tableau 7.2 présente l'EAM $\times 10^6$ des différentes méthodes d'approximation pour toutes les tailles d'échantillon satisfaisant à la condition de non-autosélection.

Tableau 7.2

EAM $\times 10^6$ des approximations des probabilités d'inclusion de deuxième ordre

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 et Donadio (BD); Van Berkel et Vondenhoff (BV); Hájek (H); Hartley et Rao (HR); Knottnerus (K); erreur absolue moyenne (EAM).

Exemple 7.3

On considère la population correspondant à la partie de gauche de la figure B.2 de l'annexe B. Les tailles des unités sont 10, 11, 6, 9, 8, 10, 8, 7, 13, 13, 11, 10 et 14, et sont obtenues en arrondissant 13 tirages de la loi normale de moyenne 10 et de variance 4. Le tableau 7.3 présente l'EAM $\times 10^6$ des différentes méthodes d'approximation pour toutes les tailles d'échantillon satisfaisant à la condition de non-autosélection.

Tableau 7.3
EAM $\times 10^6$ des approximations des probabilités d'inclusion de deuxième ordre

<i>n</i>	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	3 092	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 et Donadio (BD); Van Berkel et Vondenhoff (BV); Hájek (H); Hartley et Rao (HR); Knottnerus (K); erreur absolue moyenne (EAM).

Exemple 7.4

On considère la population correspondant à la partie de droite de la figure B.2 de l'annexe B. Les tailles des unités sont 20, 48, 17, 111, 82, 8, 72, 4, 15, 9, 59, 30 et 35, et sont obtenues en arrondissant 13 tirages de la loi du khi carré à 3 degrés de liberté. Le tableau 7.4 présente l'EAM $\times 10^6$ des différentes méthodes d'approximation pour toutes les tailles d'échantillon satisfaisant à la condition de non-autosélection.

Tableau 7.4
EAM $\times 10^6$ des approximations des probabilités d'inclusion de deuxième ordre

<i>n</i>	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 et Donadio (BD); Van Berkel et Vondenhoff (BV); Hájek (H); Hartley et Rao (HR); Knottnerus (K); erreur absolue moyenne (EAM).

Exemple 7.5

On considère la population de Connor composée de 10 unités de taille 0,3, 0,1, 0,2, 0,15, 0,05, 0,4, 0,08, 0,18, 0,34 et 0,2. Le tableau 7.5 présente l'EAM $\times 10^6$ des différentes méthodes d'approximation pour toutes les tailles d'échantillon satisfaisant à la condition de non-autosélection.

Tableau 7.5**EAM $\times 10^6$ des approximations des probabilités d'inclusion de deuxième ordre**

<i>n</i>	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 et Donadio (BD); Van Berkel et Vondenhoff (BV); Hájek (H); Hartley et Rao (HR); Knottnerus (K); erreur absolue moyenne (EAM).

Dans les cinq exemples, la méthode de BV surpasse systématiquement les autres méthodes en ce qui concerne l'EAM, en particulier pour les échantillons de grande taille où les écarts liés à l'ampleur de l'erreur deviennent plus prononcés. La méthode de BD affiche les erreurs les plus élevées dans tous les tableaux. Parmi les trois méthodes restantes, on observe que les méthodes de K et de HR donnent des résultats comparables. Dans les exemples 7.1 et 7.4, les méthodes de K et de HR surpassent la méthode de H ou lui sont équivalentes, tandis que dans les exemples 7.2, 7.3 et 7.5, elles lui sont supérieures pour les petites tailles d'échantillon, mais lui sont inférieures pour les grandes. Pour toutes les méthodes, l'EAM augmente en fonction de la taille d'échantillon; il y a de légères déviations locales par rapport à cette tendance et une déviation notable pour la méthode de H dans les exemples 7.2, 7.3 et 7.5 pour les grandes tailles d'échantillon.

8. Estimation de la variance

Dans leur article fondamental, Horvitz et Thompson (1952) ont présenté un estimateur sans biais du total d'une variable cible Y dans la population, désormais connu sous le nom d'estimateur de Horvitz-Thompson :

$$\hat{Y}_{HT} = \sum_{g \in S} y_g / \pi_g, \quad (8.1)$$

où S est un échantillon tiré selon un plan de sondage probabiliste assurant que toutes les probabilités d'inclusion de premier ordre π_g sont positives et connues, et y_g est la valeur observée de Y pour l'unité g . La variance de cet estimateur est donnée par :

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

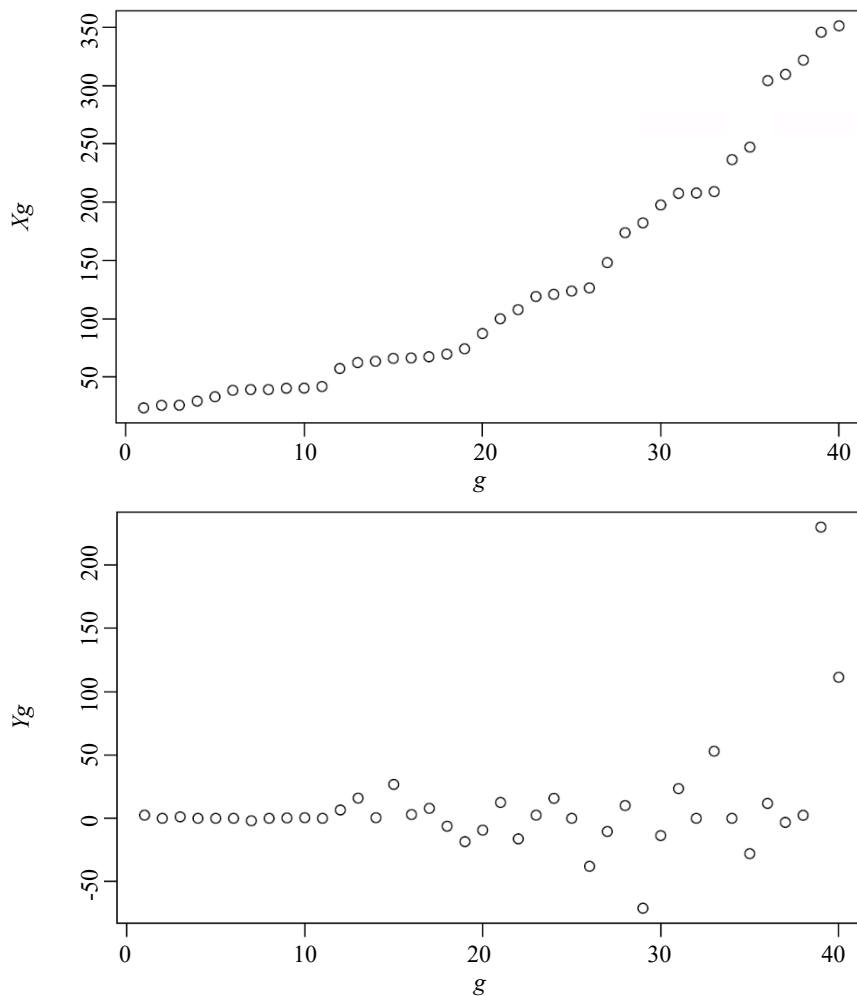
où N est la taille de la population et Y_g est la valeur réelle de la variable cible Y pour l'unité g . Pour les échantillons de taille fixe, cette variance peut être réécrite comme suit :

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

Le calcul de la variance de l'estimateur de Horvitz-Thompson nécessite la connaissance de toutes les probabilités d'inclusion de premier et de deuxième ordre, ainsi que les valeurs de Y pour toutes les unités de la population.

À titre d'application, on considère l'Indice des prix à la production aux Pays-Bas. À la section 4 de Knottnerus (2011), 70 indices des prix des produits de base sur environ 2 500 ont été retenus. Afin de réduire le temps de calcul des formules de variance, certaines simplifications ont été apportées. Les produits de base ont d'abord été triés selon la taille de la variable auxiliaire X , mise à l'échelle par un facteur de 1 000. Ensuite, les 23 produits de base ayant les plus petites valeurs de X et les 7 ayant les plus grandes valeurs de X ont été retirés, ce qui donne une population de $N = 40$. La figure 8.1 illustre les valeurs de X et de Y pour ces produits de base.

Figure 8.1 Valeurs de la variable auxiliaire X et de la variable cible Y pour les unités de la population



Afin de simplifier davantage le calcul des probabilités d'inclusion exactes de deuxième ordre, le partitionnement à k -moyennes a été appliqué à X , ce qui a formé sept classes. Pour chaque classe c , la valeur moyenne de X , désignée par $\bar{X}_c$, a été calculée et arrondie à l'entier le plus proche; voir le tableau 8.1.

Tableau 8.1

Classe (c), taille (X), moyenne de X par classe ($\bar{X}_c$) et variable cible (Y) pour 40 produits de base (unités)

unité	classe c	X	$\bar{X}_c$	Y	unité	classe 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

Pour la suite de la présente section, les valeurs originales de X sont remplacées par les moyennes de leur classe respective. Après cet ajustement, le coefficient de corrélation de Pearson entre X et Y est 0,42. Pour les tailles d'échantillon allant de 8 à 13, les variances de l'estimateur de Horvitz-Thompson ont été calculées de deux façons. Premièrement, nous avons substitué les probabilités d'inclusion exactes de deuxième ordre du Théorème 4.3 dans la formule (8.2) et, deuxièmement, nous avons substitué les approximations des probabilités d'inclusion de deuxième ordre du Théorème 5.2 dans la formule (8.2). Dans les deux cas, les probabilités d'inclusion exactes de premier ordre ont été utilisées.

Le tableau 8.2 présente les résultats et montre que les approximations des variances concordent étroitement avec les variances exactes. Pour les tailles d'échantillon 8 et 9, les approximations surestiment légèrement les variances exactes, tandis que pour les tailles d'échantillon plus grandes, elles les sous-estiment légèrement. Cependant, les écarts relatifs ne dépassent jamais 1,5 %, ce qui confirme la grande fiabilité des approximations.

Tableau 8.2

Variances exactes (V) et approximations des variances $\hat{V}$ de l'estimateur de Horvitz-Thompson selon la taille d'échantillon

n	V	$\hat{V}$	$\hat{V} - V$	$(\hat{V} - V) / V$
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

Dans les applications pratiques, la variable cible Y n'est observée que pour les unités de l'échantillon, de sorte que la variance doit être estimée. Pour des échantillons de taille $n > 1$, un estimateur sans biais de la variance est donné par :

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

Sen (1953) et Yates et Grundy (1953) ont proposé un estimateur sans biais de la variance pour les échantillons de taille fixe :

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

Ces deux estimateurs nécessitent les probabilités d'inclusion de premier et de deuxième ordre ainsi que les valeurs de Y pour toutes les unités échantillonnées. Matei et Tillé (2005) ont passé en revue 20 estimateurs de variance applicables aux plans avec probabilités inégales et ont constaté que les estimateurs (8.4) et (8.5) donnent des résultats similaires, sauf quand les populations présentent une forte corrélation entre X et Y , auquel cas (8.4) est légèrement inférieur. Leurs simulations ont également montré que les estimateurs n'utilisant que les probabilités d'inclusion de premier ordre donnent des résultats comparables, quelle que soit la corrélation entre X et Y . Par conséquent, ils ont recommandé l'estimateur plus simple proposé par Deville (1993) :

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

Par ailleurs, le cas $c_g = n / (n-1)$ correspond à l'échantillonnage avec remise. Cependant, comme l'ont mentionné Matei et Tillé, cet estimateur est fortement biaisé et surestime considérablement la variance lorsqu'il est appliqué à un plan sans remise. Pour cette raison, il n'est pas examiné davantage dans la présente section.

En plus des estimateurs de variance (8.4) et (8.5), nous définissons les estimateurs $\hat{V}_{HT}^{BV}$ et $\hat{V}_{SYG}^{BV}$ en remplaçant les probabilités d'inclusion exactes de deuxième ordre π_{gh} par leurs approximations π_{gh}^{BV} du Théorème 5.2. Avec l'estimateur (8.6), on obtient au total cinq estimateurs de variance.

La précision des cinq estimateurs a été évaluée par simulation. Étant donné que des unités dont la taille dépasse l'intervalle d'échantillonnage apparaissent pour des tailles d'échantillon supérieures à 13, l'analyse a été limitée aux six plus grandes tailles d'échantillon réalisables : 8, 9, ..., 13. Pour chaque taille d'échantillon, 1 000 échantillons indépendants ont été tirés selon le plan d'échantillonnage systématique randomisé avec des probabilités proportionnelles aux moyennes de classe des valeurs de X du tableau 8.1. Les moyennes des estimations de Horvitz-Thompson pour le total de la population 322,3 de Y étaient de 319,2, de 323,7, de 320,5, de 324,4, de 321,5 et de 321,4 pour les tailles d'échantillon 8 à 13, respectivement.

La précision de chaque estimateur de variance $\hat{V}$ est évaluée à l'aide de deux indicateurs de la qualité : le biais relatif (BR_{sim}) et l'erreur-type relative (ETR_{sim}), définis comme suit. Fixons la taille d'échantillon n . Soit V_{sim} , le vecteur dont la i^{e} composante est l'estimation de la variance obtenue par $\hat{V}$ pour la i^{e} simulation, $i = 1, \dots, 1\,000$. On désigne la moyenne et la variance de ces estimations par $\bar{V}_{\text{sim}}$ et $\text{var}(V_{\text{sim}})$, respectivement, et V désigne la variance réelle indiquée au tableau 8.2. Alors,

$$BR_{\text{sim}}(\hat{V}) = \frac{\bar{V}_{\text{sim}} - V}{V} \text{ et } ETR_{\text{sim}}(\hat{V}) = \frac{\sqrt{\text{var}(V_{\text{sim}})}}{V}. \quad (8.7)$$

Le tableau 8.3 présente les biais relatifs et les erreurs-types relatives des cinq estimateurs. Les biais relatifs sont très faibles, ce qui confirme que les estimateurs sont effectivement sans biais. En particulier, les résultats des estimateurs utilisant les approximations des probabilités d'inclusion de deuxième ordre (les colonnes 2 et 3, ainsi que les colonnes 4 et 5) sont très similaires. L'estimateur de Deville présente des biais relatifs de la même ampleur que ceux des autres estimateurs. Les erreurs-types relatives permettent avant tout de déterminer la précision des estimateurs. Les valeurs des colonnes 7 et 8 sont pratiquement identiques, de même que celles des colonnes 9 et 10, ce qui indique un rendement comparable entre les approches exacte et approximative. Les erreurs-types de l'estimateur de Deville sont comparables à celles des autres estimateurs.

Tableau 8.3

Biais relatif et erreur-type relative des estimateurs de variance dans les simulations selon la taille d'échantillon

n	Biais relatif					Erreur-type relative				
	$\hat{V}_{\text{HT}}$	$\hat{V}_{\text{HT}}^{\text{BV}}$	$\hat{V}_{\text{SYG}}$	$\hat{V}_{\text{SYG}}^{\text{BV}}$	$\hat{V}_D$	$\hat{V}_{\text{HT}}$	$\hat{V}_{\text{HT}}^{\text{BV}}$	$\hat{V}_{\text{SYG}}$	$\hat{V}_{\text{SYG}}^{\text{BV}}$	$\hat{V}_D$
	%					%				
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 et Vondenhoff (BV); Horvitz-Thompson (HT); Sen et Yates et Grundy (SYG).

Aucun estimateur ne permet d'atteindre de façon systématique le plus petit biais ou la plus petite erreur-type pour toutes les tailles d'échantillon. Il n'y a pas de différence notable entre l'utilisation de nos approximations ou l'utilisation de nos valeurs exactes des probabilités d'inclusion de deuxième ordre dans les estimations de la variance. De plus, pour toute taille d'échantillon fixe, aucun estimateur ne surpasse les autres, ce qui indique que tous les estimateurs de variance montrent des résultats similaires.

9. Conclusion

Dans le présent article, de nouvelles formules exactes et d'approximation sont dérivées pour les probabilités d'inclusion de deuxième ordre dans l'échantillonnage systématique randomisé avec probabilités inégales et sans remise. Nos formules exactes sont plus explicites et moins laborieuses que celles proposées dans la littérature. Lorsque toutes les tailles d'unité de la population sont distinctes, la complexité de calcul de nos expressions exactes des Théorèmes 4.1 et 4.2 est identique à celle des formules existantes : elle croît de façon exponentielle en fonction de la taille de la population et de façon linéaire en fonction de la taille d'échantillon. Cependant, lorsque la population contient des unités de tailles identiques, la complexité de calcul de notre formule exacte du Théorème 4.3 croît de façon polynomiale en fonction de la taille de la population, tout en augmentant de façon linéaire en fonction de la taille de l'échantillon.

La formule d'approximation du Théorème 5.1 repose sur l'observation clé selon laquelle la distance entre deux unités de la population, pour un nombre fixe d'unités entre elles, suit une loi normale asymptotique. Cela découle du théorème de la limite centrale pour l'échantillonnage aléatoire simple sans remise à partir de populations finies (Erdős et Rényi, 1959). Comme le montre l'annexe B, cette approximation donne de bons résultats, même pour les populations de petite taille (par exemple pour $N = 13$) et pour diverses configurations de tailles d'unité. La complexité de calcul de cette approximation croît de façon linéaire tant en fonction de la taille de la population qu'en fonction de la taille de l'échantillon.

La formule d'approximation finale du Théorème 5.2 est une version ajustée de celle du Théorème 5.1. La complexité de calcul de l'ensemble des facteurs d'échelle est d'ordre N^3 . Comparativement aux méthodes d'approximation existantes, la nôtre fournit des estimations plus précises des probabilités d'inclusion de deuxième ordre, en particulier pour les échantillons de grande taille, quelle que soit la configuration sous-jacente de la taille des unités.

Compte tenu de la grande précision de nos approximations des probabilités d'inclusion de deuxième ordre, celles-ci conviennent bien à l'estimation de la variance de l'estimateur de Horvitz-Thompson. En pratique, l'utilisation des probabilités exactes ou approximatives donne des estimations de la variance d'une précision comparable. Néanmoins, l'estimateur de la variance de Deville, qui ne requiert que les probabilités d'inclusion de premier ordre, atteint le même niveau de précision tout en restant calculable en temps raisonnable pour des populations de grande taille. Par conséquent, l'estimateur de Deville demeure le choix privilégié pour l'estimation de la variance à grande échelle.

Remerciements

Les auteurs remercient les examinateurs et le rédacteur associé pour leur lecture attentive du manuscrit et pour leurs précieux commentaires qui ont contribué à l'améliorer. Les auteurs remercient également Sander Scholtus de Statistics Netherlands pour son appui en matière de mathématiques et pour ses nombreuses observations et suggestions pertinentes.

Annexe

A. Démonstration de la Propriété 4.1

Pour $g, h \in U$, la fonction L_{gh} est définie par :

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

où $l_{gh}(x)$ désigne la longueur de $(0, X_g] \cap (x, x + X_h]$, pour n'importe quelle valeur de $x \in \mathbb{R}$,

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

Les figures A.1 et A.2 illustrent la forme du graphique de l_{gh} pour $X_g > X_h$ et $X_g < X_h$, respectivement. Dans le cas où $X_g = X_h$, le sommet du graphique n'est pas un segment horizontal, mais l'unique point $(0, X_g)$.

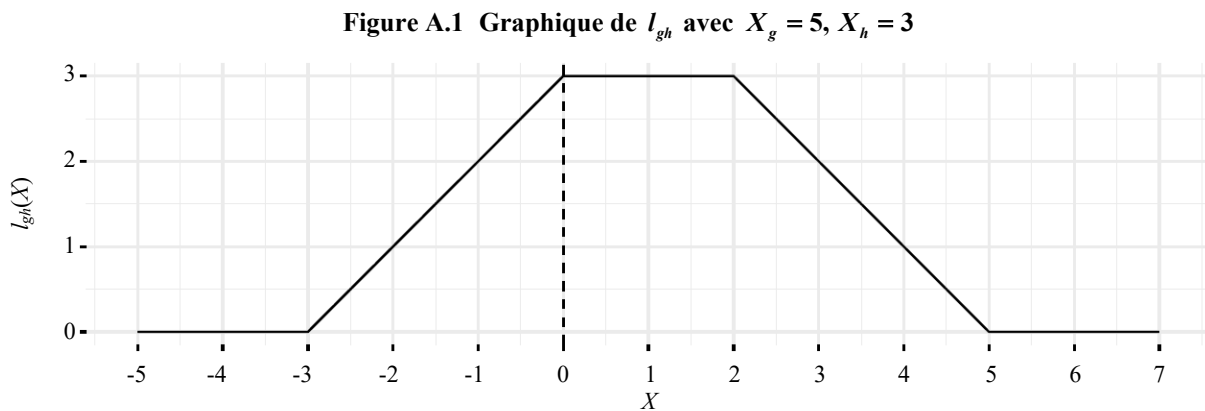
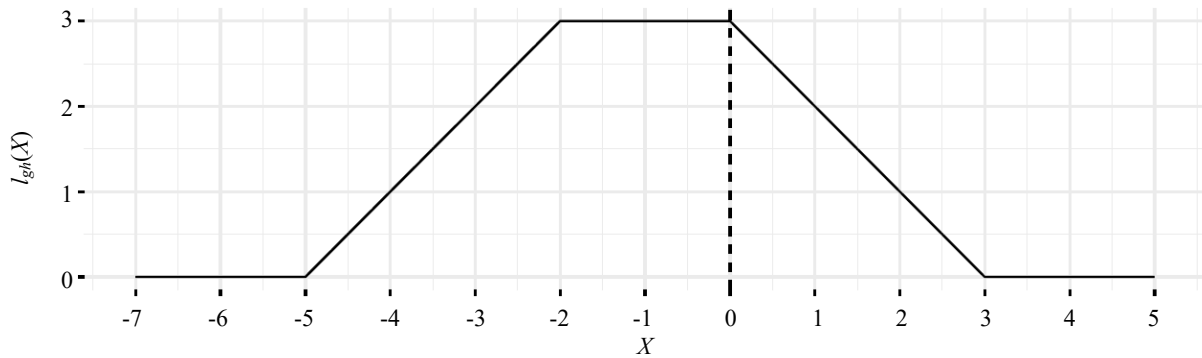


Figure A.2 Graphique de l_{gh} avec $X_g = 3, X_h = 5$ 

Le graphique de L_{gh} est construit en additionnant $n-1$ graphiques décalés de l_{gh} , et chaque décalage est de kT unités vers la droite, $k=1, \dots, n-1$. Pour la forme du graphique de L_{gh} , il y a une distinction entre $T \geq X_g + X_h$ et $T < X_g + X_h$. Dans le premier cas, les parties positives des graphiques décalés de l_{gh} ne se chevauchent pas, ce qui permet de placer les graphiques décalés côte à côte. Dans le second cas, les parties positives de deux graphiques décalés consécutifs de l_{gh} se chevauchent, de sorte que le minimum de L_{gh} est égal à $X_g + X_h - T$, ce qui représente également le minimum aux bornes du domaine de L_{gh} . Les figures A.3 et A.4 illustrent la forme du graphique de L_{gh} pour $T \geq X_g + X_h$ et $T < X_g + X_h$, respectivement. Les paramètres de la figure A.3 sont $X_g = 3, X_h = 2, n = 3, T = 6$. Dans la figure A.4, ils sont $X_g = 5, X_h = 4, n = 3, T = 7$.

La fonction l_{gh} est une fonction linéaire continue par segments. En utilisant de l'algèbre élémentaire, l_{gh} peut être exprimée comme une combinaison linéaire de parties positives de fonctions linéaires :

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

où $\mathbb{P}$ désigne la partie positive, $\mathbb{P}(x) = \max(x, 0)$, $x \in \mathbb{R}$. En combinant les formules (A.1) et (A.3), on obtient l'expression explicite de $L_{gh}(d)$, pour $d \in [X_g, X - X_h]$,

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

Il découle de la formule (A.4) que L_{gh} est symétrique par rapport au point médian $(X + X_g - X_h) / 2$ de son domaine quand l'on utilise $X = nT$, quand l'on change la variable de sommation k en $l = n - k$ et quand $\mathbb{P}(x) = \mathbb{P}(-x) + x$. ■

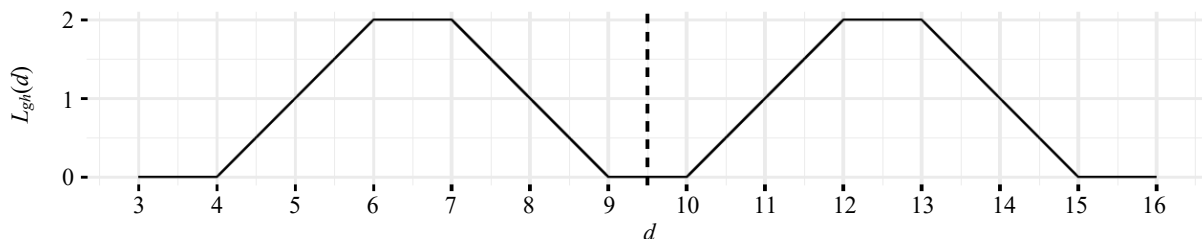
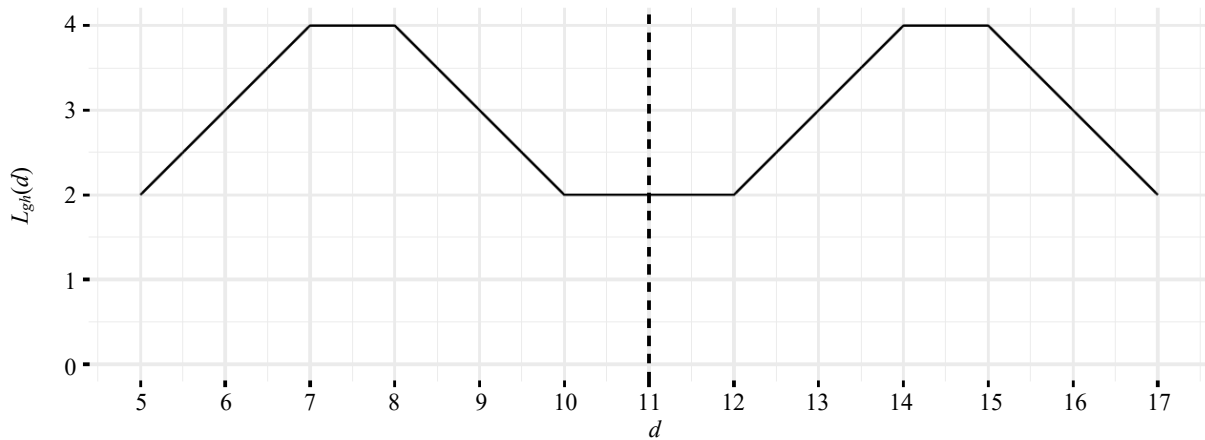
Figure A.3 Graphique de L_{gh} avec $n = 3$ et $T > X_g + X_h$ 

Figure A.4 Graphique de L_{gh} avec $n = 3$ et $T < X_g + X_h$ 

B. Démonstration du Théorème 5.1

Le point de départ est la formule exacte de π_{gh} du Théorème 4.2. Soit $N \geq 8$. La somme de cette formule comporte alors au moins quatre termes, dont les trois premiers peuvent être calculés facilement et de façon exacte. Des approximations sont développées pour les termes d'ordre supérieur. On considère donc le terme de la somme pour $l \geq 5$,

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

On définit $D_{ghl} = \{X_g + \gamma(W) \mid W \in V(g,h,l)\}$. Pour chaque $d \in D_{ghl}$, soit $n_{ghl}(d)$, le nombre d'ensembles $W \in V(g,h,l)$ pour lesquels $X_g + \gamma(W) = d$ et soit $f_{ghl}(d) = n_{ghl}(d) / \binom{N-2}{l-2}$. Alors :

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

Étant donné que $V(g,h,l)$ compte $\binom{N-2}{l-2}$ éléments, la fonction f_{ghl} représente la loi de probabilité des distances possibles entre les unités g et h , conditionnellement au fait que h occupe la l^{e} position. Ainsi, t_{ghl} est l'espérance de L_{gh} sous cette condition.

Chaque ensemble $W \in V(g,h,l)$ peut être interprété comme un échantillon aléatoire simple sans remise de taille $l-2$ tiré de U_{gh} . En conséquence, $\gamma(W)$ est le total sur l'échantillon des tailles d'unité dans W . En utilisant le théorème de la limite centrale pour l'échantillonnage aléatoire simple sans remise à partir de populations finies (Erdős et Rényi, 1959), $\gamma(W)$ suit asymptotiquement une loi normale de moyenne et de variance :

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

où μ_{gh} et σ_{gh}^2 désignent, respectivement, la moyenne et la variance des valeurs de X des unités de U_{gh} ,

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

Le théorème de la limite centrale s'applique à condition que la fraction d'échantillonnage $(l-2)/(N-2)$ ne soit pas trop proche de zéro ou de un, et qu'aucune taille d'unité ne domine la variance totale de la population. Ainsi, pour des valeurs élevées de l et de N , on peut donner une approximation de la fonction de probabilité, f_{ghl} , en utilisant la fonction de densité φ_{ghl} de la loi normale de moyenne μ_{ghl} et de variance σ_{ghl}^2 données par :

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

Soit :

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

Chaque $d \in D_{ghl}$ appartenant à l'intervalle $[X_g, X - X_h]$, on peut obtenir une approximation de la somme dans (B.2) en utilisant l'intégrale :

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

On procède ensuite à l'évaluation de l'intégrale de (B.7). Pour alléger la notation, l'indice ghl est temporairement omis, c'est-à-dire que $\varphi(x) = \varphi_{ghl}(x)$, $\mu = \mu_{ghl}$ et $\sigma^2 = \sigma_{ghl}^2$. La fonction de distribution cumulative Φ , associée à φ , peut être exprimée en fonction de la fonction d'erreur de Gauss (erf) :

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

Selon la définition de φ et de Φ , les dérivées sont :

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

On en déduit les primitives suivantes :

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

La suite de la démonstration suppose que $X_g \geq X_h$. Pour $X_g < X_h$, les calculs sont analogues et mènent au même résultat. À l'annexe A, on a montré que la forme du graphique de la fonction L_{gh} dépend de la position de $X_g + X_h$ par rapport à T . On considère d'abord $T \geq X_g + X_h$. Puisque $X_g \geq X_h$, il découle de la formule (A.2) que pour $x, \lambda \in \mathbb{R}$:

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

Puisque $T \geq X_g + X_h$, il s'ensuit que $X_g \leq kT \leq kT + X_g \leq X - X_h$, pour toute valeur de $k \in \{1, \dots, n-1\}$, et donc :

$$I(k) = \int_{X_g}^{X-X_h} \varphi(x) \cdot l_{gh}(x-kT) dx = \int_{kT-X_h}^{kT} \varphi(x) \cdot (x-kT+X_h) dx + X_h \cdot \int_{kT}^{kT+X_g-X_h} \varphi(x) dx + \int_{kT+X_g-X_h}^{kT+X_g} \varphi(x) \cdot (-x+kT+X_g) dx. \quad (\text{B.12})$$

En appliquant la formule (B.10) et en effectuant quelques calculs, on obtient :

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

$$F(x) = (\mu - x) \Phi(x) - \sigma^2 \varphi(x). \quad (\text{B.14})$$

Puisque $L_{gh}(x) = \sum_{k=1}^{n-1} l_{gh}(x-kT)$, l'intégrale recherchée peut être évaluée :

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

On considère ensuite le cas $T < X_g + X_h$. Puisque $T - X_h < X_g$, la borne inférieure de la deuxième intégrale de la formule (B.12) doit être remplacée par X_g pour $k=1$. En outre, puisque $(n-1)T + X_g > X - X_h$, la borne supérieure de la dernière intégrale de la formule (B.12) doit être remplacée par $X - X_h$ pour $k=n-1$. Cela implique que de la formule (B.15), il faut soustraire :

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

En combinant (B.13)-(B.16), on obtient l'expression recherchée pour $\tilde{t}_{ghl} = \int_{X_g}^{X-X_h} \varphi_{ghl}(x) \cdot L_{gh}(x) dx$. ■

Remarque

Pour illustrer que cette approximation donne également de bons résultats pour les petites valeurs de N , nous présentons quelques exemples. Dans la partie gauche de la figure B.1, l'histogramme représente la fonction de probabilité empirique f_{ghl} , tandis que la courbe rouge montre son approximation par la loi normale φ_{ghl} . Dans ce cas-ci, la population U est composée de 13 unités de taille $X_k = k + 4$, pour $k = 1, 2, \dots, 13$, et le triplet (g, h, l) est fixé à $(1, 2, 6)$. Dans la partie droite de la figure B.1, les tailles des unités ont été remplacées par les valeurs 16, 20, 6, 8, 19, 8, 17, 13, 12, 5, 11, 18 et 15, générées par la loi uniforme discrète sur l'intervalle $[5, 20]$.

Lorsque les tailles X_k sont générées par une variable aléatoire suivant une loi normale ou même asymétrique, on peut toujours obtenir une approximation pour f_{ghl} en utilisant φ_{ghl} comme l'illustre la figure B.2. Sa partie gauche correspond à U composée de 13 unités de tailles 10, 11, 6, 9, 8, 10, 8, 7, 13, 13, 11, 10 et 14, obtenues en arrondissant 13 tirages de la loi normale de moyenne 10 et de variance 4. Dans la partie droite, les tailles ont été remplacées par les valeurs 20, 48, 17, 111, 82, 8, 72, 4, 15, 9, 59, 30 et 35, obtenues en arrondissant 13 tirages de la loi du khi carré à trois degrés de liberté.

Figure B.1 Fonction de probabilité f_{ghl} et approximation φ_{ghl} , où $X_k = k + 4$ (à gauche) et les X_k proviennent d'une loi uniforme (à droite)

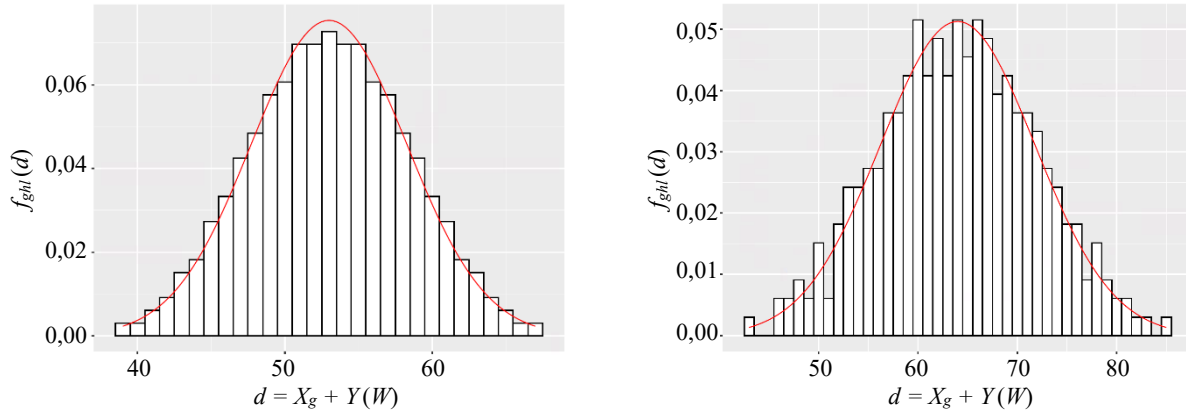
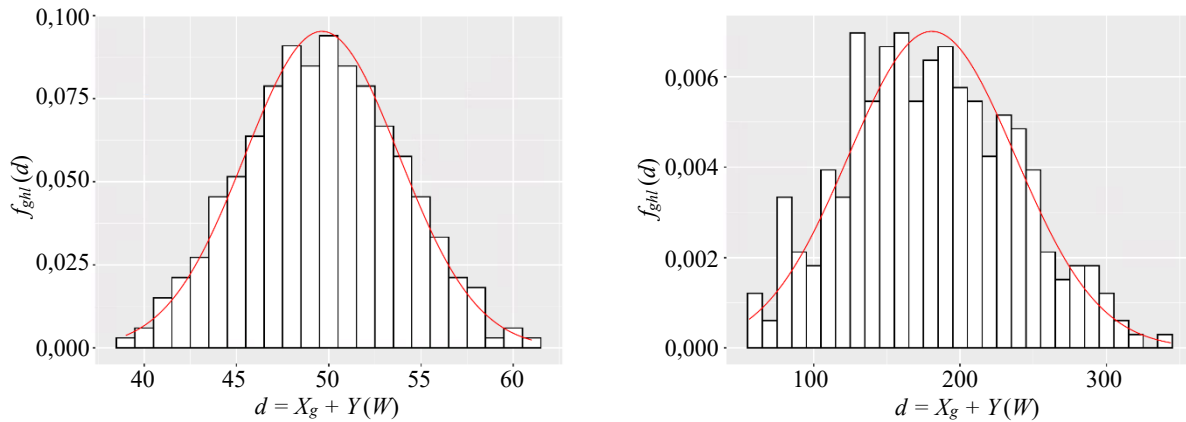


Figure B.2 Approximation de la fonction de probabilité f_{ghl} par φ_{ghl} où les X_k proviennent d'une loi normale (à gauche) et d'une loi du khi carré (à droite)



C. Démonstration du Théorème 5.2

La minimisation de (5.3) sous les contraintes sur les limites dans (5.2) peut être résolue par la méthode des multiplicateurs de Lagrange. Le lagrangien correspondant est égal à :

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

En réglant à zéro les dérivées partielles du lagrangien par rapport à α_{gh} , les α_{gh} peuvent être exprimés en fonction des multiplicateurs de Lagrange λ_g ,

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

En substituant ces valeurs de α_{gh} dans les contraintes sur les limites dans (5.2), on obtient N équations pour les N multiplicateurs de Lagrange inconnus. En notation matricielle, ces équations peuvent s'écrire $\mathbf{A} \cdot \boldsymbol{\lambda} = \mathbf{b}$, où $\mathbf{A}$ est la matrice $N \times N$ ayant des éléments $A_{gh} = \tilde{\pi}_{gh}^2$ pour $g \neq h \in U$ et $A_{gg} = \sum_{h \in U_g} \tilde{\pi}_{gh}^2$ pour $g \in U$, $\boldsymbol{\lambda}$ est le vecteur des multiplicateurs de Lagrange et $\mathbf{b}$ est le vecteur à N dimensions ayant $b_g = 2 \left(\sum_{h \in U_g} \tilde{\pi}_{gh} - (n-1)\pi_g \right)$ pour $g \in U$. ■

Bibliographie

- 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. et Donadio, M.E. (2003). [La variance sous grande entropie de l'estimateur de Horvitz-Thompson](https://www150.statcan.gc.ca/n1/pub/12-001-x/2003002/article/6778-fra.pdf). *Techniques d'enquête*, 29(2), 213-220. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/pub/12-001-x/2003002/article/6778-fra.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. Accessible à l'adresse : https://www.jstor.org/stable/2282827#metadata_info_tab_contents.
- Deville, J.-C. (1993). *Estimation de la Variance pour les Enquêtes en Deux Phases*. Note manuscrite interne. France : INSEE. [En français].
- Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164. Accessible à l'adresse : https://www.jstor.org/stable/2985777#metadata_info_tab_contents.
- Erdős, P. et Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Magyar Tudosanyos Akademia Budapest Matematikai Kutato Intezet Koezlemenyei*, 4, 49-57. Accessible à l'adresse : 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. Accessible à l'adresse : <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. Accessible à l'adresse : https://www.jstor.org/stable/2238287#metadata_info_tab_contents.

- Hartley, H.O. et Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374. Accessible à l'adresse : <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. et Gray, G.B. (1975). [Un algorithme informatique pour les probabilités de sélection conjointes \(échantillonnage systématique avec probabilité proportionnelle à la taille\)](#). *Techniques d'enquête*, 1(1), 1-9. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1975001/article/00002-fra.pdf>.
- Horvitz, D.G. et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685. Accessible à l'adresse : <https://www.jstor.org/stable/pdf/2280784.pdf>.
- Knottnerus, P. (2003). Sample survey theory: Some Pythagorean perspectives. New York: *Springer-Verlag*. Demande du PDF en texte intégral à l'adresse suivante : https://www.researchgate.net/publication/4741910_Sample_Survey_Theory_Some_PythagoreanPerspectives_Paul_Knottnerus.
- Knottnerus, P. (2011). [À propos de l'efficacité de l'échantillonnage à probabilité proportionnelle à la taille aléatoire](#). *Techniques d'enquête*, 37(1), 103-110. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/pub/12-001-x/2011001/article/11450-fra.pdf>.
- Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439. Accessible à l'adresse : <https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/a-note-on-the-hartley-rao-variance-estimator.pdf>.
- Matei, A. et 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. Accessible à l'adresse : <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. et Wu, C. (2008). [Échantillonnage PPT systématique randomisé basé sur la simulation en cas de substitution d'unités](#). *Techniques d'enquête*, 34(1), 3-11. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2008001/article/10613-fra.pdf>.

Yates, F. et 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. Accessible à l'adresse : <https://www.jstor.org/stable/pdf/2983772.pdf>.