Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Combinaison d’échantillons probabilistes indépendants

par Anton Grafström, Magnus Ekström, Bengt Gunnar Jonsson,
Per-Anders Esseen et Göran Ståhl

Date de diffusion : le 27 juin 2019

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 STATCAN.infostats-infostats.STATCAN@canada.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

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

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.

Techniques d'enquête, juin 2019
Vol. 45, nº 2, pp. 371-387
Statistique Canada, nº 12-001-X au catalogue

Combinaison d'échantillons probabilistes indépendants

Anton Grafström, Magnus Ekström, Bengt Gunnar Jonsson, Per-Anders Esseen et Göran Ståhl[1]

Résumé

Dans divers domaines, il est de plus en plus important de fusionner les sources d'information disponibles pour améliorer les estimations des caractéristiques de la population. En présence de plusieurs échantillons probabilistes indépendants d'une population finie, nous examinons plusieurs solutions d'estimateur combiné du total de la population, basé soit sur une combinaison linéaire d'estimateurs distincts, soit sur une méthode par échantillon combiné. L'estimateur en combinaison linéaire fondé sur des variances estimées est susceptible d'être biaisé, car les estimateurs distincts du total de la population peuvent être fortement corrélés à leurs estimateurs de la variance respectifs. Nous illustrons la possibilité d'utiliser un échantillon combiné pour estimer les variances des estimateurs distincts, ce qui donne des estimateurs de la variance groupés généraux. Ces estimateurs de la variance groupés utilisent tous les renseignements disponibles et peuvent réduire considérablement le biais d'une combinaison linéaire d'estimateurs distincts.

Mots-clés : Estimateur de Horvitz-Thompson; probabilités d'inclusion; estimateur en combinaison linéaire; estimation de la variance.

1 Introduction

Le fait d'utiliser toute l'information disponible pour produire de meilleures estimations est une idée très séduisante, mais on sait rarement comment procéder exactement pour obtenir les meilleurs résultats. Une littérature abondante traite de ce qu'on appelle maintenant la méta-analyse, qui se fonde sur l'idée de la combinaison des résultats de plusieurs études. Cochran et Carroll (1953) et Cochran (1954) ont publié deux des premiers articles portant sur la combinaison d'estimations provenant d'expériences différentes. Koricheva, Gurevitch et Mengersen (2013) et Schmidt et Hunter (2014) ont rédigé deux ouvrages présentant un traitement plus actuel et plus complet de la méta-analyse. Dans le présent article, nous n'étudierons pas la combinaison de résultats d'expériences classiques, mais plutôt de résultats provenant d'échantillons probabilistes multiples. Nous présentons tous les éléments de plan requis, comme les probabilités d'inclusion du premier et du second ordre, pour une combinaison générale de multiples échantillons indépendants provenant de différents plans d'échantillonnage. Nous présentons également de nouveaux estimateurs de la variance d'estimateurs distincts basés sur le plan des échantillons combinés. Les estimateurs de la variance proposés peuvent être considérés comme des estimateurs de la variance groupés généraux utilisant toute l'information disponible. En particulier, on peut utiliser ces estimateurs de la variance groupés dans une combinaison linéaire d'estimateurs distincts pour réduire l'erreur quadratique moyenne (EQM) par rapport à l'utilisation d'estimateurs de la variance distincts et donc indépendants.

1. Anton Grafström, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Suède. Courriel : anton.grafstrom@slu.se; Magnus Ekström, Department of Statistics, USBE, Umeå University, SE-90187 Umeå, Suède, et Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Suède. Courriel : magnus.ekstrom@umu.se; Bengt Gunnar Jonsson, Department of Natural Sciences, Mid Sweden University, SE-85170 Sundsvall, Suède. Courriel : bengt-gunnar.jonsson@miun.se; Per-Anders Esseen, Department of Ecology and Environmental Science, Umeå University, SE-90187 Umeå, Suède. Courriel : per-anders.esseen@umu.se; Göran Ståhl, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183 Umeå, Suède. Courriel : goran.stahl@slu.se.

Nous posons une restriction, à savoir que nous traitons uniquement une combinaison d'échantillons probabilistes indépendants provenant d'une même population au même moment ou, si ce n'est pas le cas, en supposant que la variation de la variable cible n'est pas significative. De plus, nous supposons que chaque plan d'échantillonnage est suffisamment connu pour que les probabilités d'inclusion du premier et du second ordre soient connues pour toutes les unités. En général, nous devons également être en mesure d'identifier de façon unique chaque unité afin de détecter si une même unité est sélectionnée dans plus d'un échantillon, ou plusieurs fois dans le même échantillon. Quelques-unes de ces hypothèses pourraient être très restrictives, car elles ne se vérifient pas nécessairement dans certaines circonstances pratiques.

Soit $U = \{1, 2, \ldots, N\}$ l'ensemble d'étiquettes de N unités dans la population. Notre objectif consiste à estimer le total d'une variable cible y, qui prend comme valeur y_i pour l'unité $i \in U$. Nous cherchons ainsi à estimer $Y = \sum_{i=1}^{N} y_i$. Nous supposons que nous avons accès à k échantillons probabilistes indépendants $S^{(\ell)}$, $\ell = 1, \ldots, k$, de U, où les échantillons peuvent provenir de différents plans d'échantillonnage. Sous ces hypothèses, nous examinons différentes possibilités d'estimation du total de la population en utilisant tous les renseignements disponibles. Dans certains cas, il faut savoir quelles sont les unités incluses dans plusieurs échantillons différents. De nos jours, on le sait plus facilement dans les enquêtes relatives à la surveillance environnementale et aux ressources naturelles, en raison de l'utilisation répandue de systèmes de positionnement par satellite précis (Næsset et Gjevestad, 2008). Dans les études environnementales, on peut souvent considérer les unités comme des emplacements ayant des coordonnées données, mais la situation est différente dans les sondages menés auprès, par exemple, de personnes pouvant être anonymes ou non identifiables. De plus, des programmes de surveillance du paysage et des forêts sont réalisés dans plusieurs pays (Tomppo, Gschwantner, Lawrence et McRoberts, 2009; Ståhl, Allard, Esseen, Glimskår, Ringvall, Svensson, Sundquist, Christensen, Gallegos Torell, Högström, Lagerqvist, Marklund, Nilsson et Inghe, 2011; Fridman, Holm, Nilsson, Nilsson, Ringvall et Ståhl, 2014) et ils doivent parfois être complétés par des programmes spéciaux d'échantillonnage pour que des cibles d'exactitude données soient atteintes pour certaines régions ou années (Christensen et Ringvall, 2013).

Dans la section 2, nous rappelons d'abord la théorie de la combinaison linéaire optimale d'estimateurs indépendants distincts. Ensuite, à la section 3, nous présentons la théorie de combinaison d'échantillons indépendants. Comme une unité peut être incluse dans plus d'un échantillon ou plusieurs fois dans le même échantillon, nous devons décider s'il faut dénombrer une ou plusieurs inclusions. Si l'on compte une seule inclusion, le plan est un plan sans remise, alors que si l'on choisit plusieurs inclusions, on obtient une forme de plan avec remise. Dans la section 4, nous présentons deux exemples comparant différentes possibilités d'estimation. Enfin, à la section 5, nous concluons l'article par une discussion.

2 Combiner des estimations distinctes

Supposons que nous avons $k \geq 2$ estimateurs, $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$ d'un total de population Y, résultant de k échantillons indépendants de la même population. Nos possibilités dépendent fortement des renseignements

disponibles. Si nous avons des estimations et les estimations de la variance correspondante, une combinaison linéaire fondée sur des poids calculés à partir de variances estimées pourrait alors être une solution intéressante. Nous pourrions aussi pondérer les estimateurs en fonction de la taille de l'échantillon, si elle est connue, mais on sait que cela est loin d'être optimal dans certaines situations. Rappelons la théorie de combinaison linéaire optimale d'estimateurs indépendants sans biais. La combinaison linéaire de $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$ ayant la plus petite variance est

$$\hat{Y}_L = \alpha_1 \hat{Y}_1 + \alpha_2 \hat{Y}_2 + \ldots + \alpha_k \hat{Y}_k,$$

où

$$\alpha_i = \frac{1/V_i\left(\hat{Y}_i\right)}{\sum_{j=1}^{k} 1/V_j\left(\hat{Y}_j\right)}$$

sont des poids positifs dont la somme est égale à 1. La variance de $\hat{Y}_L$ est

$$V\left(\hat{Y}_L\right) = \frac{1}{\sum_{j=1}^{k} 1/V_j\left(\hat{Y}_j\right)}.$$

Il est courant d'utiliser les estimations de la variance à la place des variances inconnues dans le calcul des poids α, comme dans Cochran et Carroll (1953) et Cochran (1954). Si les estimateurs de la variance sont convergents, cette méthode fournira asymptotiquement la pondération optimale. De plus, si on suppose que les estimateurs de la variance sont indépendants des estimateurs $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$, l'estimateur qui en résulte

$$\hat{Y}_L^* = \hat{\alpha}_1 \hat{Y}_1 + \hat{\alpha}_2 \hat{Y}_2 + \ldots + \hat{\alpha}_k \hat{Y}_k,$$

est sans biais et sa variance dépend uniquement de la variance de $\hat{Y}_L$ et des EQM des $\hat{\alpha}_i$, voir Rubin et Weisberg (1974). Toutefois, comme nous le montrerons bientôt, l'hypothèse d'indépendance est susceptible de ne pas être respectée dans de nombreuses applications d'échantillonnage. Dans le cas de corrélations positives entre les estimateurs et leurs estimateurs de variance, nous donnons un poids plus important aux petites estimations, car elles ont tendance à avoir des variances estimées plus petites. Ainsi, l'estimateur combiné (utilisant des poids basés sur des variances estimées) aura un biais négatif, susceptible d'augmenter à mesure qu'augmente le nombre d'enquêtes indépendantes que nous combinons (voir l'exemple 1). L'opposé est également vrai en cas de corrélation négative, mais cette situation est probablement plus rare dans les applications d'échantillonnage.

Exemple 1 : *Voici un exemple simpliste illustrant la possibilité d'augmentation du biais à mesure qu'augmente le nombre d'enquêtes indépendantes que nous combinons. Supposons que l'estimateur sans biais $\hat{Y}$ pour un échantillon a pour valeur 1 ou 2 avec des probabilités égales et soit un estimateur de la variance dont la valeur est c fois l'estimateur* (*corrélation parfaite*) *et qui est sans biais* $(c = 1/6)$. *De toute évidence, la valeur espérée de $\hat{Y}$ est 1,5. Considérons ensuite la combinaison linéaire de deux estimateurs indépendants $\left(\hat{Y}_1, \hat{Y}_2\right)$ du même type que $\hat{Y}$ au moyen des variances estimées. La paire $\left(\hat{Y}_1, \hat{Y}_2\right)$ a les quatre résultats possibles suivants* (*1,1*), (*1,2*), (*2,1*), (*2,2*), *chacun ayant une probabilité de 1/4. Les*

résultats correspondants pour la combinaison linéaire $\hat{Y}_L^*$ *avec les variances estimées sont 1, 4/3, 4/3, 2 avec une espérance de* $17/12 \approx 1{,}4167$. *Le biais est négatif. Si un troisième estimateur indépendant du même type est ajouté, nous avons les huit résultats suivants* (*1,1,1,1*), (*1,1,2*), (*1,2,1*), (*1,2,2*), (*2,1,1*), (*2,1,2*), (*2,2,1*), (*2,2,2*), *chacun ayant une probabilité égale de 1/8. Les résultats correspondants pour* $\hat{Y}_L^*$ *sont 1, 6/5, 6/5, 3/2, 6/5, 3/2, 3/2, 2 avec une espérance de* $111/80 = 1{,}3875$. *Le biais est encore plus négatif et il continue de croître à mesure que d'autres estimateurs indépendants du même type sont ajoutés à la combinaison.*

2.1 Pourquoi a-t-on fréquemment une corrélation positive entre l'estimateur et l'estimateur de la variance dans les applications d'échantillonnage ?

La question de la corrélation positive entre l'estimateur d'un total et l'estimateur de sa variance a été constatée auparavant par Gregoire et Schabenberger (1999) quand l'échantillonnage rend asymétriques des populations biologiques, mais nous montrons qu'une forte corrélation peut apparaître dans des applications d'échantillonnage plus générales. Supposons que la variable cible n'est pas négative et que $y_i > 0$ pour exactement N' unités. La proportion de y_i non nuls (positifs) est désigné par $p = N'/N$. Cette situation est très courante dans l'échantillonnage et nous obtenons une variable cible de ce type si nous estimons un total de domaine ($y_i = 0$ à l'extérieur du domaine) ou si un sous-ensemble de la population seulement a la propriété d'intérêt.

L'estimateur de Horvitz-Thompson (HT) sans biais fondé sur le plan est donné par

$$\hat{Y} = \sum_{i \in S} \frac{y_i}{\pi_i},$$

où S désigne l'ensemble aléatoire d'unités échantillonnées et $\pi_i = \Pr(i \in S)$. Selon des plans à taille fixe, la variance de $\hat{Y}$ est

$$V\left(\hat{Y}\right) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2,$$

où $\pi_{ij} = \Pr(i \in S, j \in S)$ est la probabilité d'inclusion du second ordre. L'estimateur de la variance correspondant est

$$\hat{V}\left(\hat{Y}\right) = -\frac{1}{2}\sum_{i \in S}\sum_{j \in S}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Si tous les π_{ij} sont strictement positifs, l'estimateur de la variance est alors un estimateur sans biais de $V\left(\hat{Y}\right)$.

Le nombre de y_i non nuls dans S (et par conséquent dans $\hat{Y}$) est noté ici par n' et est habituellement un nombre aléatoire. On peut montrer que le nombre d'éléments non nuls dans $\hat{V}\left(\hat{Y}\right)$ est approximativement proportionnel à n' si p est petit, ce qui indique la possibilité d'une forte corrélation entre $\hat{Y}$ et $\hat{V}\left(\hat{Y}\right)$ en général si p est petit. Pour montrer que le nombre de termes non nuls dans $\hat{V}\left(\hat{Y}\right)$ est approximativement proportionnel à n', nous observons trois cas, le troisième étant le plus général.

Cas 1 : Supposons que toutes les valeurs y_i/π_i non nulles sont différentes, c'est-à-dire $y_i/\pi_i \neq y_j/\pi_j$ pour $i \neq j$, et $\pi_{ij} \neq \pi_i\pi_j$ pour tous les i, j. La double somme dans $\hat{V}\left(\hat{Y}\right)$ contient alors $2n'(n-n')$ termes non nuls ayant la forme

$$\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\left(\frac{y_k}{\pi_k}\right)^2,$$

où k est égal à i ou j et $i \neq j$. Il y a $n'(n'-1)$ termes non nuls ayant la forme

$$\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2,$$

où $i \neq j$. Le nombre total de termes non nuls est $n'(2n-n'-1)$. Si n est plutôt grand et p est petit, alors $n' << n$ et nous avons à peu près $n'(2n-n'-1) \approx 2n'n$. Ainsi, le nombre de termes non nuls est approximativement proportionnel à n'.

Cas 2 : Supposons que toutes les valeurs y_i/π_i non nulles sont égales, par exemple y_i est une variable d'indicateur et $\pi_i = n/N$, et $\pi_{ij} \neq \pi_i\pi_j$ pour tous les i, j. Alors la double somme dans $\hat{V}\left(\hat{Y}\right)$ contient $2n'(n-n')$ termes non nuls ayant la forme

$$\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\left(\frac{y_k}{\pi_k}\right)^2,$$

où k est égal à i ou j et $i \neq j$. Si n est plutôt grand et p est petit, alors $n' << n$ nous avons à peu près $2n'(n-n') \approx 2n'n$. Ainsi, le nombre de termes non nuls est encore approximativement proportionnel à n'.

Cas 3 : Si certaines valeurs y_i/π_i non nulles sont égales et que les autres sont différentes, alors le nombre de termes non nuls sera entre $2n'(n-n')$ (cas 2) et $n'(2n-n'-1)$ (cas 1). Ainsi, le nombre de termes non nuls dans $\hat{V}\left(\hat{Y}\right)$ est toujours approximativement proportionnel à n' si p est petit.

Si $\pi_{ij} \leq \pi_i\pi_j$ pour tous les $i \neq j$, alors tous les termes non nuls sont positifs. Cette condition se vérifie, par exemple, pour les plans d'échantillonnage aléatoire simple (EAS) et avec probabilités inégales à grande entropie, comme sous un échantillonnage de Poisson conditionnel, de Sampford et de Pareto. Pour un traitement plus détaillé de l'entropie des plans d'échantillonnage, voir par exemple Grafström (2010). Il est peu probable que la taille moyenne des termes positifs dans $\hat{V}\left(\hat{Y}\right)$, ou $\hat{Y}$ dépende beaucoup de n'. Par conséquent, si $\hat{Y}$ contient n' termes positifs, et $\hat{V}\left(\hat{Y}\right)$ contient un nombre de termes positifs proportionnel à n', leurs tailles sont principalement déterminées par n'. Une variance relative élevée dans n' peut causer une forte corrélation entre $\hat{Y}$ et $\hat{V}\left(\hat{Y}\right)$, voir l'exemple 2.

Des plans courants peuvent produire une variance relative élevée pour n'. Si nous utilisons un échantillonnage aléatoire simple sans remise, nous obtenons $n' \sim \text{Hyp}(N, N', n)$ et $V(n')/E(n') = (1-p)(N-n)/(N-1) \approx (1-p)(1-n/N)$, ce qui signifie qu'il nous faut un grand p ou une grande fraction d'échantillon n/N afin d'obtenir une petite variance relative pour n'. Dans de nombreuses

applications, nous aurons une valeur p plutôt petite et une petite fraction d'échantillon n/N et, par conséquent, pour de nombreux plans (qui n'utilisent pas de renseignements antérieurs pouvant expliquer dans une certaine mesure si $y_i \neq 0$ ou non), il y aura une variance relative élevée pour n'. Afin d'illustrer la grandeur de la corrélation résultant entre l'estimateur et son estimateur de variance, voici un exemple d'échantillonnage aléatoire simple sans remise.

Exemple 2 : *Dans cet exemple, nous simulerons d'abord une population de taille* $N = 1\,000$ *où* $N' = 100$, *c'est-à-dire* $p = 0{,}1$. *Les 100 valeurs* y *non nulles sont simulées à partir de* $N(\mu, \sigma^2)$ *sachant que* $\mu = 10$ *et* $\sigma = 2$. *Nous sélectionnons des échantillons de taille* $n = 200$ *avec échantillonnage aléatoire simple, de sorte que* $\pi_i = n/N$ *et* $\pi_{ij} = n(n-1)/(N(N-1))$ *pour* $i \neq j$. *La corrélation observée entre* $\hat{Y}$ *et* $\hat{V}(\hat{Y})$ *était de 0,974 pour* 10^6 *échantillons, voir la figure 2.1 pour les 1 000 premières observations de* $(\hat{Y}, \hat{V}(\hat{Y}))$. *Si nous augmentons* p *à 0,3, la corrélation est encore supérieure à 0,9. Les résultats ne changent pas si le rapport* σ/μ *demeure inchangé; nous obtenons par exemple les mêmes corrélations si* $\mu = 100$ *et* $\sigma = 20$.

Supposons maintenant que nous avons accès à plus d'un échantillon pour l'estimation de Y. Comme on l'a indiqué précédemment, en cas de corrélations positives élevées entre les estimateurs et les estimateurs de variance correspondants, il y a un risque de biais important quand on utilise une combinaison linéaire avec des variances estimées. L'intérêt de l'utilisation de données combinées peut être plus grand pour les petits domaines ou les propriétés rares, cas dans lesquels le problème de forte corrélation est le plus probable. Nous aborderons ensuite d'autres solutions d'utilisation de renseignements combinés provenant de plusieurs échantillons.

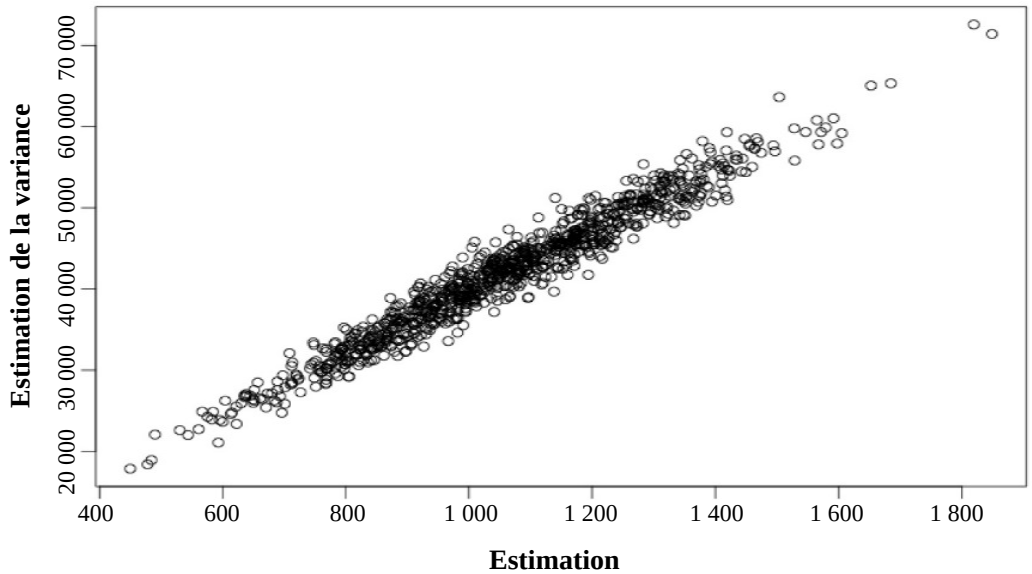

Figure 2.1 Relation entre l'estimateur de Horvitz-Thompson et son estimateur de variance pour une variable à 90 % nulle.

3 Combiner des échantillons

Nous calculons ici les éléments du plan (par exemple probabilités d'inclusion du premier et du second ordre) pour l'échantillon combiné. Il existe toutefois différentes façons de combiner des échantillons. Nous devons par exemple choisir entre un comptage multiple ou un comptage unique pour le plan combiné. Quand on combine des échantillons indépendants provenant de la même population, nous devons connaître les probabilités d'inclusion de toutes les unités des échantillons, pour tous les plans. Les probabilités d'inclusion du second ordre sont nécessaires pour l'estimation de la variance. Dans certains cas, il nous faut aussi des identificateurs uniques (étiquettes) pour les unités de façon afin qu'elles puissent être appariées, par exemple, quand on utilise un comptage unique ou quand au moins un plan distinct a des probabilités inégales. Bankier (1986) a examiné la méthode du comptage unique pour le cas particulier de la combinaison de deux échantillons aléatoires simples stratifiés sélectionnés indépendamment et tirés de la même base de sondage. Roberts et Binder (2009) et O'Muircheartaigh et Pedlow (2002) ont discuté des différentes possibilités pour combiner des échantillons indépendants tirés d'une même base de sondage, mais pas avec des plans d'échantillonnage généraux.

Un problème quelque peu semblable est l'estimation fondée sur des échantillons provenant de bases multiples chevauchantes, voir par exemple les articles de synthèse de Lohr (2009, 2011) et les articles qu'il cite. Bien que le fait d'avoir la même base de sondage puisse être considéré comme un cas particulier de bases multiples, nous n'avons pas trouvé de calculs des éléments du plan (en particulier des probabilités d'inclusion du second ordre et l'ordre deux du nombre espéré d'inclusions) pour la combinaison des plans d'échantillonnage généraux. Pour le cas des plans de sondage probabiliste généraux, nous présentons ci-dessous en détail deux façons principales de combiner les échantillons probabilistes et de calculer les caractéristiques de plan correspondantes nécessaires pour l'estimation sans biais et l'estimation de la variance sans biais.

3.1 Combinaison à comptage unique

Ici, nous combinons d'abord deux échantillons indépendants $S^{(1)}$ et $S^{(2)}$ tirés de la même population, et nous étudions l'union des deux échantillons dans notre échantillon combiné. Ainsi, l'inclusion d'une unité n'est comptée qu'une seule fois, même si l'unité est incluse dans plus d'un échantillon. Les probabilités d'inclusion du premier ordre sont :

$$\pi_i^{(1,2)} = \pi_i^{(1)} + \pi_i^{(2)} - \pi_i^{(1)}\pi_i^{(2)}, \tag{3.1}$$

où $\pi_i^{(1,2)} = \Pr\left(i \in S^{(1)} \cup S^{(2)}\right)$ et $\pi_i^{(\ell)} = \Pr\left(i \in S^{(\ell)}\right)$ pour $\ell = 1, 2$. Soit $I_i^{(1)}$, $I_i^{(2)}$ et $I_i^{(1,2)}$ l'indicateur d'inclusion de l'unité i dans $S^{(1)}$, $S^{(2)}$ et $S^{(1)} \cup S^{(2)}$ respectivement. Le plan qui en résulte n'est plus un plan à taille fixe (bien que les plans distincts soient de taille fixe). La taille espérée de l'union $S^{(1)} \cup S^{(2)}$ est donnée par $E\left(n^{(1,2)}\right) = \sum_{i=1}^{N} \pi_i^{(1,2)}$, où $n^{(1,2)} = \sum_{i=1}^{N} I_i^{(1,2)}$ désigne la taille aléatoire de l'union. Si nous

voulons savoir dans quelle mesure les échantillons se chevaucheront en moyenne, la taille espérée du chevauchement est donnée par la somme $\sum_{i=1}^{N} \pi_i^{(1)} \pi_i^{(2)}$.

Les probabilités d'inclusion du second ordre $\pi_{ij}^{(1,\,2)}$ pour l'union $S^{(1)} \cup S^{(2)}$ peuvent s'exprimer en termes des probabilités d'inclusion du premier et du second ordre des deux plans respectifs. Soit $B = \left(i \in S^{(1)} \cup S^{(2)},\ j \in S^{(1)} \cup S^{(2)}\right)$, alors $\pi_{ij}^{(1,\,2)} = \Pr(B)$. En conditionnant sur les résultats pour i et j dans $S^{(1)}$, on obtient les quatre cas suivants

m	A_m	$\Pr(A_m)$	$\Pr(B \mid A_m)$
1	$i \in S^{(1)},\ j \in S^{(1)}$	$\pi_{ij}^{(1)}$	1
2	$i \in S^{(1)},\ j \notin S^{(1)}$	$\pi_i^{(1)} - \pi_{ij}^{(1)}$	$\pi_j^{(2)}$
3	$i \notin S^{(1)},\ j \in S^{(1)}$	$\pi_j^{(1)} - \pi_{ij}^{(1)}$	$\pi_i^{(2)}$
4	$i \notin S^{(1)},\ j \notin S^{(1)}$	$1 - \pi_i^{(1)} - \pi_j^{(1)} + \pi_{ij}^{(1)}$	$\pi_{ij}^{(2)}$

où $\pi_{ij}^{(\ell)} = \Pr\left(i \in S^{(\ell)},\ j \in S^{(\ell)}\right)$ pour $\ell = 1, 2$. Les événements A_m, $m = 1, 2, 3, 4$, sont disjoints et $\sum_{m=1}^{4} \Pr(A_m) = 1$. Par conséquent, on a par la formule des probabilités totales $\pi_{ij}^{(1,\,2)} = \Pr(B) = \sum_{m=1}^{4} \Pr(B \mid A_m) \Pr(A_m)$. Cela donne

$$\pi_{ij}^{(1,\,2)} = \pi_{ij}^{(1)} + \pi_j^{(2)} \left(\pi_i^{(1)} - \pi_{ij}^{(1)}\right) + \pi_i^{(2)} \left(\pi_j^{(1)} - \pi_{ij}^{(1)}\right) + \pi_{ij}^{(2)} \left(1 - \pi_i^{(1)} - \pi_j^{(1)} + \pi_{ij}^{(1)}\right). \qquad (3.2)$$

On peut généraliser les équations (3.1) et (3.2) pour obtenir de façon récursive les probabilités d'inclusion du premier et du second ordre de l'union d'un nombre arbitraire k d'échantillons indépendants. Après avoir calculé les probabilités de l'union des deux premiers échantillons, nous pouvons combiner le résultat avec les probabilités du troisième plan en utilisant les mêmes formules, et ainsi de suite. À titre d'exemple, soit $\pi_i^{(1,\,\ldots,\,\ell)}$ la probabilité d'inclusion du premier ordre de l'unité i dans l'union des premiers échantillons ℓ. On a alors

$$\pi_i^{(1,\,\ldots,\,\ell+1)} = \pi_i^{(1,\,\ldots,\,\ell)} + \pi_i^{(\ell+1)} - \pi_i^{(1,\,\ldots,\,\ell)} \pi_i^{(\ell+1)},$$

comme probabilité d'inclusion du premier ordre de l'unité i dans l'union des $\ell + 1$ premiers échantillons. De même, pour les probabilités d'inclusion du second ordre, nous obtenons la formule récursive

$$\begin{aligned} \pi_{ij}^{(1,\,\ldots,\,\ell+1)} = {} & \pi_{ij}^{(1,\,\ldots,\,\ell)} + \pi_j^{(\ell+1)} \left(\pi_i^{(1,\,\ldots,\,\ell)} - \pi_{ij}^{(1,\,\ldots,\,\ell)}\right) + \pi_i^{(\ell+1)} \left(\pi_j^{(1,\,\ldots,\,\ell)} - \pi_{ij}^{(1,\,\ldots,\,\ell)}\right) \\ & + \pi_{ij}^{(\ell+1)} \left(1 - \pi_i^{(1,\,\ldots,\,\ell)} - \pi_j^{(1,\,\ldots,\,\ell)} + \pi_{ij}^{(1,\,\ldots,\,\ell)}\right). \end{aligned}$$

Désormais, pour la combinaison de k échantillons indépendants, nous utilisons la notation simplifiée $\pi_i = \pi_i^{(1,\,\ldots,\,k)}$, $\pi_{ij} = \pi_{ij}^{(1,\,\ldots,\,k)}$ et $I_i = I_i^{(1,\,\ldots,\,k)}$. Étant donné que les échantillons individuels sont suceptibles de se chevaucher, le plan qui en résulte n'est pas de taille fixe. L'estimateur combiné sans biais à comptage unique, sous forme d'estimateur de Horvitz-Thompson, est donné par

$$\hat{Y}_{\mathrm{CS}} = \sum_{i \in S} \frac{y_i}{\pi_i}.$$

La variance est

$$V\left(\hat{Y}_{\mathrm{CS}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j},$$

et l'estimateur de la variance sans biais est

$$\hat{V}\left(\hat{Y}_{\mathrm{CS}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\frac{I_i I_j}{\pi_{ij}}.$$

Pour la combinaison d'échantillons indépendants avec des probabilités d'inclusion positives du premier ordre, nous avons toujours $\pi_{ij} > 0$ pour toutes les paires (i, j), ce qui est nécessaire pour que l'estimateur de la variance ci-dessus soit sans biais. En ce qui concerne l'EQM, il peut être avantageux de ne pas utiliser l'estimateur à un comptage unique, mais plutôt un estimateur qui tienne compte de la taille de l'échantillon aléatoire. Toutefois, dans ce cas, nous nous limitons à n'utiliser que des estimateurs sans biais.

3.2 Combinaison à comptage multiple

Nous examinons d'abord la manière de combiner deux échantillons indépendants $S^{(1)}$ et $S^{(2)}$ sélectionnés dans la même population, où nous permettons que chaque unité puisse être incluse plusieurs fois. Le nombre d'inclusions de l'unité i dans l'échantillon combiné est noté par $S_i^{(1,\,2)}$, et il est égal à la somme du nombre d'inclusions de l'unité i dans les deux échantillons combinés, c'est-à-dire $S_i^{(1,\,2)} = S_i^{(1)} + S_i^{(2)}$, où $S_i^{(\ell)}$ est le nombre d'inclusions de l'unité i dans l'échantillon ℓ. Le nombre espéré d'inclusions de l'unité i dans la combinaison est donné par

$$E\left(S_i^{(1,\,2)}\right) = E_i^{(1,\,2)} = E_i^{(1)} + E_i^{(2)}, \tag{3.3}$$

où $E_i^{(\ell)} = E\left(S_i^{(\ell)}\right)$ est le nombre espéré d'inclusions de l'unité i dans l'échantillon $S^{(\ell)}$, $\ell = 1, 2$. La taille de l'échantillon (pouvant être aléatoire) est la somme $\sum_{i=1}^{N} S_i^{(1,\,2)}$ de toutes les inclusions individuelles et la taille espérée de l'échantillon est la somme $\sum_{i=1}^{N} E_i^{(1,\,2)}$ de tous les nombres individuels espérés d'inclusions. On peut montrer que

$$E\left(S_i^{(1,\,2)}S_j^{(1,\,2)}\right) = E_{ij}^{(1,\,2)} = E_{ij}^{(1)} + E_i^{(1)}E_j^{(2)} + E_i^{(2)}E_j^{(1)} + E_{ij}^{(2)}, \tag{3.4}$$

où $E_{ij}^{(\ell)} = E\left(S_i^{(\ell)}S_j^{(\ell)}\right)$, $\ell = 1, 2$ sont le nombre espéré d'inclusions du second ordre dans l'échantillon ℓ. De toute évidence, $E_{ij}^{(\ell)} = \pi_{ij}^{(\ell)}$ si le plan pour l'échantillon ℓ est sans remise. Notons que, parce que $S_i^{(\cdot)}$ peut prendre d'autres valeurs que 0 ou 1, $E_{ii}^{(\cdot)}$ est en général non égal à $E_i^{(\cdot)}$, mais $\pi_{ii}^{(\cdot)} = \pi_i^{(\cdot)}$. On peut utiliser les équations (3.3) et (3.4) de façon récursive pour obtenir $E_i^{(\cdot)}$ et $E_{ij}^{(\cdot)}$ pour la combinaison d'un nombre arbitraire k d'échantillons indépendants. Nous obtenons alors les formules récursives

$$E_i^{(1,\ldots,\ell+1)} = E_i^{(1,\ldots,\ell)} + E_i^{(\ell+1)}$$

et

$$E_{ij}^{(1,\ldots,\ell+1)} = E_{ij}^{(1,\ldots,\ell)} + E_i^{(1,\ldots,\ell)}E_j^{(\ell+1)} + E_j^{(1,\ldots,\ell)}E_i^{(\ell+1)} + E_{ij}^{(\ell+1)}.$$

Les résultats précédents et (3.4) découlent du fait que $S_i^{(1,\ldots,\ell+1)} = S_i^{(1,\ldots,\ell)} + S_i^{(\ell+1)}$ et que $S_i^{(1,\ldots,\ell)}$ et $S_i^{(\ell+1)}$ sont indépendants. Nous avons par exemple

$$\begin{aligned} E_{ij}^{(1,\ldots,\ell+1)} = E\left(S_i^{(1,\ldots,\ell+1)}S_j^{(1,\ldots,\ell+1)}\right) &= E\left(\left(S_i^{(1,\ldots,\ell)} + S_i^{(\ell+1)}\right)\left(S_j^{(1,\ldots,\ell)} + S_j^{(\ell+1)}\right)\right) = \\ &E\left(S_i^{(1,\ldots,\ell)}S_j^{(1,\ldots,\ell)} + S_i^{(1,\ldots,\ell)}S_j^{(\ell+1)} + S_j^{(1,\ldots,\ell)}S_i^{(\ell+1)} + S_i^{(\ell+1)}S_j^{(\ell+1)}\right) = \\ &E_{ij}^{(1,\ldots,\ell)} + E_i^{(1,\ldots,\ell)}E_j^{(\ell+1)} + E_j^{(1,\ldots,\ell)}E_i^{(\ell+1)} + E_{ij}^{(\ell+1)}. \end{aligned}$$

Pour la combinaison de k échantillons indépendants, nous utilisons maintenant la notation simplifiée $E_i = E_i^{(1,\ldots,k)}$, $E_{ij} = E_{ij}^{(1,\ldots,k)}$, et $S_i = S_i^{(1,\ldots,k)}$. On peut estimer le total Y sans biais avec l'estimateur à comptage multiple, l'estimateur de Hansen-Hurwitz (Hansen et Hurwitz, 1943) en étant un cas particulier. Il est donné par

$$\hat{Y}_{\mathrm{CM}} = \sum_{i=1}^{N} \frac{y_i}{E_i} S_i.$$

Nous obtenons l'estimateur de Hansen-Hurwitz si $E_i = np_i$, où n est le nombre d'unités tirées et p_i, avec $\sum_{i=1}^{N} p_i = 1$, sont les probabilités d'un tirage indépendant unique. La variance de $\hat{Y}_{\mathrm{CM}}$ peut être exprimée comme étant

$$V\left(\hat{Y}_{\mathrm{CM}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(E_{ij} - E_iE_j\right)\frac{y_i}{E_i}\frac{y_j}{E_j}.$$

Un des estimateurs de la variance est

$$\hat{V}\left(\hat{Y}_{\mathrm{CM}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(E_{ij} - E_iE_j\right)\frac{y_i}{E_i}\frac{y_j}{E_j}\frac{S_iS_j}{E_{ij}}.$$

Il s'ensuit directement que l'estimateur de la variance ci-dessus est sans biais, car quand on combine des échantillons indépendants avec des probabilités d'inclusion du premier ordre positives, nous avons toujours $E_{ij} > 0$ pour toutes les paires (i, j).

3.3 Comparer des estimateurs combinés et distincts

Deux exemples montrent que l'estimateur combiné n'est pas nécessairement aussi bon que le meilleur estimateur distinct.

Exemple 3 : *Supposons que le premier échantillon, $S^{(1)}$, est de taille fixe avec $\pi_i^{(1)} \propto y_i$, et que le deuxième est un échantillon aléatoire simple avec $\pi_i^{(2)} = n/N$. Alors, l'estimateur de Horvitz-Thompson $\hat{Y}_1 = \sum_{i\in S^{(1)}} y_i/\pi_i^{(1)}$, a une variance nulle, mais l'estimateur à comptage unique combiné avec $\pi_i = \pi_i^{(1)} + \pi_i^{(2)} - \pi_i^{(1)}\pi_i^{(2)}$ a une variance positive. Par conséquent, l'estimateur combiné est moins performant que le meilleur estimateur distinct.*

Exemple 4 : *Supposons que le plan du premier échantillon est stratifié de telle sorte qu'il n'y a pas de variation à l'intérieur des strates. Alors, l'estimateur distinct* $\hat{Y}_1 = \sum_{i \in S^{(1)}} y_i \big/ \pi_i^{(1)}$ *a une variance nulle. Si le premier échantillon est combiné à un deuxième échantillon non stratifié, alors le plan d'échantillonnage qui en résulte n'a pas de tailles d'échantillon fixes dans les strates. Par conséquent, l'estimateur combiné a une variance positive.*

Ces exemples montrent qu'il faut faire preuve de prudence avant de combiner des plans d'échantillonnage très différents, comme un plan à probabilités inégales avec un plan à probabilités égales, ou un plan de sondage stratifié avec un plan de sondage non stratifié. Il faut se montrer particulièrement prudent si nous voulons estimer le total directement à partir de l'échantillon combiné. Notons toutefois qu'en cas de combinaison d'échantillons de plans relativement semblables, il est probable que l'estimateur combiné soit meilleur que le meilleur des estimateurs distincts.

Nous examinerons ensuite comment utiliser la méthode combinée pour l'estimation des variances distinctes, puis l'estimateur en combinaison linéaire. En fait, comme nous le verrons ultérieurement, l'utilisation de la méthode combinée pour l'estimation de la variance de variances distinctes peut agir comme stabilisateur des poids de la combinaison linéaire quand les poids sont basés sur des variances estimées. On a une sorte d'effet de regroupement pour les estimateurs de variance quand ils sont estimés avec le même ensemble de renseignements.

3.4 Utiliser un échantillon combiné pour l'estimation des variances d'estimateurs distincts

Au lieu d'estimer directement le total Y à partir du plan combiné, on peut utiliser le plan combiné pour estimer les variances des estimateurs distincts, puis continuer avec une combinaison linéaire des estimateurs distincts. Supposons que nous avons accès à k échantillons indépendants et que nous voulons estimer la variance d'un estimateur distinct, dont la variance est une double somme des unités de population. Il y a deux possibilités principales pour l'estimateur de variance : multiplier par

$$\frac{I_i I_j}{\pi_{ij}} \quad \text{ou} \quad \frac{S_i S_j}{E_{ij}}$$

dans la formule de variance pour obtenir un estimateur sans biais de la variance basé sur la combinaison de la totalité des k échantillons $S^{(\ell)}$, $\ell = 1, \ldots, k$. Par exemple, en supposant que la variance de $\hat{Y}_1$ est

$$V\left(\hat{Y}_1\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}\right)\frac{y_i}{\pi_i^{(1)}}\frac{y_j}{\pi_j^{(1)}},$$

nous pouvons utiliser la combinaison de $S^{(\ell)}$, $\ell = 1, \ldots, k$, pour estimer $V\left(\hat{Y}_1\right)$ par l'estimateur à comptage unique

$$\hat{V}_{\text{CS}}\left(\hat{Y}_1\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}\right)\frac{y_i}{\pi_i^{(1)}}\frac{y_j}{\pi_j^{(1)}}\frac{I_i I_j}{\pi_{ij}}$$

ou l'estimateur à comptage multiple

$$\hat{V}_{\mathrm{CM}}\left(\hat{Y}_1\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}\right)\frac{y_i}{\pi_i^{(1)}}\frac{y_j}{\pi_j^{(1)}}\frac{S_i S_j}{E_{ij}}.$$

Notons que $\pi_{ij} = \pi_{ij}^{(1,\ldots,k)}$, $I_i = I_i^{(1,\ldots,k)}$, $E_{ij} = E_{ij}^{(1,\ldots,k)}$ et $S_i = S_i^{(1,\ldots,k)}$, de sorte que les estimateurs de la variance ci-dessus utilisent tous les renseignements disponibles sur la variable cible. Par conséquent, ces estimateurs de la variance peuvent être considérés comme des estimateurs de la variance groupés généraux. Il s'ensuit directement que les deux estimateurs sont sans biais, car tous les plans ont des probabilités d'inclusion du premier ordre positives, ce qui implique que tous les π_{ij} et tous les E_{ij} sont strictement positifs. De façon intéressante, les estimateurs de la variance ci-dessus sont sans biais bien que le plan distinct 1 ait certaines probabilités d'inclusion du second ordre nulles, ce qui empêche l'estimation de la variance sans biais basée sur le seul échantillon $S^{(1)}$.

Bien que la production d'un estimateur de la variance sans biais pour tout plan soit une propriété séduisante, on ne peut pas recommander les estimateurs de la variance ci-dessus pour les plans ayant un degré élevé de probabilités d'inclusion du second ordre nulles (comme en cas d'échantillonnage systématique). En effet, les estimateurs peuvent être très instables pour ces plans et produire une proportion élevée d'estimations de la variance négatives.

Comme nous le verrons, si nous souhaitons utiliser un estimateur en combinaison linéaire, il est important que toutes les variances soient estimées de la même manière. Il est alors probable que les rapports, par exemple

$$\frac{\hat{V}_{\mathrm{CS}}\left(\hat{Y}_1\right)}{\hat{V}_{\mathrm{CS}}\left(\hat{Y}_2\right)} \quad \text{et} \quad \frac{\hat{V}_{\mathrm{CM}}\left(\hat{Y}_1\right)}{\hat{V}_{\mathrm{CM}}\left(\hat{Y}_2\right)}$$

soient stables (qu'ils aient une petite variance). Les rapports sont plus stables parce que les estimateurs dans le numérateur et le dénominateur se fondent sur les mêmes informations et sont estimés avec les mêmes poids pour toutes les paires $(i,\ j)$ dans tous les estimateurs. Avec les variances estimées, nous obtenons

$$\hat{\alpha}_i = \left[\sum_{j=1}^{k}\frac{\hat{V}\left(\hat{Y}_i\right)}{\hat{V}\left(\hat{Y}_j\right)}\right]^{-1},$$

donc si les rapports des estimateurs de la variance ont une petite variance, alors $\hat{\alpha}_i$ a une petite variance. La pondération dans la combinaison linéaire $\hat{Y}_L^*$ se stabilise alors. Comme le montre l'exemple suivant, le rapport des estimateurs de la variance peut même avoir une variance nulle. Par conséquent, il peut parfois donner une pondération optimale y compris quand les variances sont inconnues.

Exemple 5 : *Supposons que nous voulons combiner des estimations résultant de deux échantillons aléatoires simples de taille différente. Nous pouvons bien entendu le faire de façon optimale sans estimer les variances, mais à titre d'exemple, nous utiliserons la méthode ci-dessus pour estimer les variances distinctes au moyen de l'échantillon combiné. Dans ce cas, l'utilisation des estimateurs* $\hat{V}_{\mathrm{CS}}\left(\hat{Y}_1\right)$ *et* $\hat{V}_{CS}\left(\hat{Y}_2\right)$

donne la pondération optimale, de même que $\hat{V}_{\text{CM}}\left(\hat{Y}_1\right)$ *et* $\hat{V}_{\text{CM}}\left(\hat{Y}_2\right)$. *Ce résultat découle du fait que si les deux plans sont un échantillonnage aléatoire simple, nous avons* :

$$\frac{\hat{V}_{\text{CS}}\left(\hat{Y}_1\right)}{\hat{V}_{\text{CS}}\left(\hat{Y}_2\right)} = \frac{\hat{V}_{\text{CM}}\left(\hat{Y}_1\right)}{\hat{V}_{\text{CM}}\left(\hat{Y}_2\right)} = \frac{V\left(\hat{Y}_1\right)}{V\left(\hat{Y}_2\right)},$$

qui se vérifie simplement. Si on a deux échantillons aléatoires simples, la situation correspond à l'utilisation d'une estimation groupée pour S^2 (*la variance de la population de* y) *dans les expressions des estimations de la variance, et cette estimation groupée est ensuite annulée dans le calcul des poids.*

On en conclut que cette procédure est aussi susceptible de fournir une pondération plus stable pour les plans qui s'écartent de l'échantillonnage aléatoire simple, dans la mesure où les plans concernés ont une grande entropie (un caractère aléatoire élevé). Le problème du biais de l'estimateur en combinaison linéaire avec des variances estimées sera réduit par rapport à l'utilisation d'estimateurs de la variance distincts et donc indépendants.

Nous pensons que cela peut être une solution de substitution très intéressante, car l'estimateur du total basé sur le plan combiné ne donne pas nécessairement une plus petite variance que le meilleur des estimateurs distincts. Au moyen de cette stratégie, nous pouvons améliorer les estimateurs de la variance distincts, particulièrement pour un plus petit échantillon (si les données sont disponibles pour un plus grand échantillon). Ainsi, la combinaison linéaire résultante avec des variances estimées conjointement peut être une stratégie très avantageuse.

En cas de comptage unique, nous pourrions utiliser un estimateur de la variance de type ratio comme

$$\hat{V}_R\left(\hat{Y}_1\right) = \frac{N^2}{\gamma_{1,\ldots,k}} \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_j^{(1)}\right)\frac{y_i}{\pi_i^{(1)}}\frac{y_j}{\pi_j^{(1)}}\frac{I_i I_j}{\pi_{ij}},$$

où $\gamma_{1,\ldots,k} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{I_i I_j}{\pi_{ij}}$. En cas de comptage multiple, nous pouvons remplacer $I_i I_j / \pi_{ij}$ par $S_i S_j / E_{ij}$. Cet estimateur par le ratio utilise la taille connue de la population des paires $(i, j) \in \{1, 2, \ldots, N\}^2$, qui est N^2, et divise par la somme des poids d'échantillon des paires. Il faut noter que $E\left(\gamma_{1,\ldots,k}\right) = N^2$. Cette correction est utile, car le nombre de paires dans l'estimateur peut être aléatoire (puisque l'union des échantillons peut avoir une taille aléatoire). Cela permet de rééchelonner les poids de l'échantillon (des paires) pour les additionner à N^2. Ceci introduira un certain biais (comme toujours avec les estimateurs par le ratio), mais l'objectif est de réduire la variance de l'estimateur de la variance. Toutefois, cette méthode n'est utile que si nous nous intéressons à la variance distincte, car le terme de correction sera identique pour tous les estimateurs de la variance distincts. Par conséquent, cela ne change pas la pondération d'un estimateur en combinaison linéaire avec des variances estimées.

4 Exemples de simulations

Nous présentons ici deux exemples de la méthode de simulation de Monte-Carlo. Dans le premier exemple, nous combinons deux échantillons tirés d'une loi de Poisson en utilisant des probabilités

d'inclusion approximativement proportionnelles à la variable cible. Dans le deuxième exemple, nous combinons un échantillon aléatoire simple non stratifié à un échantillon aléatoire simple stratifié.

4.1 Combinaison de deux échantillons tirés d'une loi de Poisson

Nous produisons une population de taille $N = 200$, avec une variable auxiliaire $X_i \sim N(\mu = 20, \sigma^2 = 16)$. La variable cible est produite comme étant $(Y_i \mid X_i = x_i) = x_i + \epsilon_i$, où $\epsilon_i \sim N\left(0, (x_i/20)^2\right)$. Les deux ensembles de probabilités d'inclusion sont produits $\pi_i^{(1)} \propto \pi_i^{(2)} \propto x_i$, où $\sum_{i=1}^{N} \pi_i^{(1)} = n_1$ et $\sum_{i=1}^{N} \pi_i^{(2)} = n_2$. Soit les tailles d'échantillon espérées $n_1 = 15$ et $n_2 = 25$. Par souci de simplicité, les deux plans seront des plans de Poisson (où les unités sont sélectionnées indépendamment). Cela nous permet de calculer exactement les variances pour les estimateurs distincts (et par conséquent la combinaison linéaire optimale) et pour les échantillons combinés à comptage unique et multiple. Pour ce qui est des stratégies à combinaison linéaire utilisant des variances estimées, nous avons réalisé une simulation de Monte-Carlo avec 1 000 000 de sélections d'échantillons répétées. Les variances réelles pour les deux estimateurs de HT distincts, les estimateurs à comptage simple et comptage multiple pour les échantillons combinés, et la combinaison linéaire optimale des estimateurs distincts sont présentés dans le tableau 4.1. Les résultats de la simulation pour les différentes combinaisons linéaires avec les variances des estimations sont également présentés dans le tableau 4.1.

Tableau 4.1
Résultats pour la combinaison des échantillons tirés d'une loi de Poisson. Variances réelles pour les deux estimateurs de HT distincts, estimateurs à comptage simple (CS) et comptage multiple (CM) pour les échantillons combinés, et combinaison linéaire optimale des estimateurs distincts. Résultats de la simulation, pour ce qui est du biais estimé et de l'EQM, pour trois estimateurs de combinaison linéaire avec variances estimées

Estimateur	Biais (biais relatif)	EQM
$\hat{Y}_1$	0	1 053 083
$\hat{Y}_2$	0	596 069
$\hat{Y}_{\text{CS}}$	0	361 088
$\hat{Y}_{\text{CM}}$	0	380 929
$\hat{Y}_L$ Optimal	0	380 626
$\hat{Y}_L^*$ Distinct	-92,8 (-2,24 %)	412 248
$\hat{Y}_L^*$ CS groupé	1,6 (+0,04 %)	381 106
$\hat{Y}_L^*$ CM groupé	1,6 (+0,04 %)	381 106

L'utilisation d'estimateurs de variance combinés (groupés) a réduit à la fois le biais et la variance pour une combinaison linéaire par rapport à l'utilisation d'estimateurs de variance distincts. Dans cet exemple, la combinaison linéaire avec l'estimation de la variance groupée donne des résultats à la performance très

proche de celle des résultats de la combinaison linéaire optimale. Le biais négatif avec les estimateurs de variance distincts est principalement attribuable à une corrélation positive entre l'estimateur du total et l'estimateur de la variance correspondant dans le plan de Poisson. Pour ce scénario, on a obtenu le meilleur résultat par la combinaison des échantillons avec un comptage unique.

4.2 Combiner un échantillonnage aléatoire simple non stratifié à un échantillonnage aléatoire simple stratifié

Ici, nous avons produit une population de taille $N = 1\,000$, avec deux strates de taille $N_1 = 600$ et $N_2 = 400$. La variable cible y_i, $i = 1, \ldots, N$, a été produite comme suit. Dans la strate 1, il y a 500 y_i égaux à zéro et 100 autres y_i ont été tirés de $N(\mu_1 = 10, \sigma^2 = 4)$. Dans la strate 2, il y a 300 y_i égaux à zéro et les 100 autres y_i ont été tirés de $N(\mu_2 = 15, \sigma^2 = 4)$ Le premier échantillon est un échantillonnage aléatoire simple non stratifié de taille $n = 50$ et le deuxième échantillon est un échantillonnage aléatoire simple stratifié où les tailles d'échantillons de la strate sont $n_1 = 30$ et $n_2 = 20$ Les variances des estimateurs de HT distincts et des échantillons combinés avec comptage simple et multiple ont été calculés exactement. On a réalisé une simulation de Monte-Carlo avec 10 000 répétitions pour évaluer la performance d'un estimateur en combinaison linéaire avec des variances estimées. Les résultats sont présentés au tableau 4.2. On réduit le biais par l'utilisation d'une combinaison linéaire avec des estimateurs de variance groupés par rapport à l'utilisation d'estimateurs de variance distincts. De plus, pour ce scénario, on a obtenu le meilleur résultat par la combinaison des échantillons avec un comptage unique.

Tableau 4.2
Résultats pour la combinaison d'un échantillonnage aléatoire simple et d'un échantillonnage aléatoire simple stratifié. Variances réelles pour les deux estimateurs de HT distincts, estimateurs à comptage simple (CS) et comptage multiple (CM) pour l'échantillon combiné, et combinaison linéaire optimale des estimateurs distincts. Résultats de la simulation, pour ce qui est du biais estimé et de l'EQM, pour trois estimateurs de combinaison linéaire avec variances estimées

Estimateur	Biais (biais relatif)	EQM
$\hat{Y}_1$	0	516 835
$\hat{Y}_2$	0	498 321
$\hat{Y}_{\text{CS}}$	0	248 888
$\hat{Y}_{\text{CM}}$	0	253 789
$\hat{Y}_L$ Optimal	0	253 704
$\hat{Y}_L^*$ Distinct	-77 (-3 %)	287 680
$\hat{Y}_L^*$ CS groupé	9 (+0,4 %)	257 229
$\hat{Y}_L^*$ CM groupé	9 (+0,4 %)	257 217

5 Discussion

Par les exemples de simulation de la section précédente, nous cherchons seulement à montrer les différentes méthodes et nous n'affirmons pas la généralité du résultat. Cependant, nous pensons fort

probable que l'utilisation d'estimateurs de variance groupés soit préférable à l'utilisation d'estimateurs de variance distincts dans un estimateur en combinaison linéaire, particulièrement dans les cas où les estimateurs du total distincts sont fortement corrélés à leurs estimateurs de variance.

Nous avons présenté en détail une méthode de combinaison d'échantillons probabilistes indépendants et calculé les caractéristiques de plan correspondantes nécessaires à la réalisation d'une estimation sans biais et d'une estimation de la variance. Nous avons aussi illustré le danger présenté par l'utilisation d'échantillons combinés en cas de plans d'échantillonnage très différents. De plus, nous avons montré qu'il y a souvent un risque de forte corrélation positive entre l'estimateur de HT et l'estimateur de la variance correspondant. Une telle dépendance peut causer un biais si les variances estimées sont utilisées dans une combinaison linéaire. Nous avons par conséquent proposé une autre méthode, à savoir l'utilisation d'un échantillon combiné pour estimer des variances distinctes. Cette autre solution peut conduire à des poids plus stables dans une combinaison linéaire d'estimateurs distincts et elle peut réduire le biais et la variance.

Il faut évidemment tenir compte de certaines limites en ce qui concerne la possibilité d'application de cette méthodologie, car nous supposons des plans d'échantillonnage entièrement connus et l'utilisation d'une même base de sondage avec des unités identifiables. Il faudrait réaliser d'autres études pour comprendre la sensibilité des écarts par rapport à certaines de ces hypothèses, comme le fait d'avoir des unités non identifiables ou d'utiliser des probabilités d'inclusion approximatives du second ordre.

Cette méthodologie est particulièrement importante quand un premier effort d'échantillonnage s'est révélé insuffisant. Ces situations sont courantes dans la surveillance environnementale, par exemple (Christensen et Ringvall, 2013). On peut alors concevoir un échantillon complémentaire de manière à permettre une combinaison plus efficace.

Remerciements

Nous remercions le relecteur anonyme et le rédacteur adjoint dont les précieux commentaires ont permis l'amélioration de notre article. Ces travaux ont bénéficié d'un financement du *Swedish Science Council* (subvention 340-2013-5076).

Bibliographie

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81(396), 1074-1079.

Christensen, P., et Ringvall, A.H. (2013). Using statistical power analysis as a tool when designing a monitoring program: Experience from a large-scale Swedish landscape monitoring program. *Environmental Monitoring and Assessment*, 185(9), 7279-7293.

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10(1), 101-129.

Cochran, W.G., et Carroll, S.P. (1953). A sampling investigation of the efficiency of weighting inversely as the estimated variance. *Biometrics*, 9(4), 447-459.

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H. et Ståhl, G. (2014). Adapting National Forest Inventories to changing requirement-the case of the Swedish National Forest Inventory at the turn of the 20th century. *Silva Fennica*, 48(3), article id 1095.

Grafström, A. (2010). Entropy of unequal probability sampling designs. *Statistical Methodology*, 7(2), 84-97.

Gregoire, T.G., et Schabenberger, O. (1999). Sampling-skewed biological populations: Behavior of confidence intervals for the population total. *Ecology*, 80, 1056-1065.

Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4), 333-362.

Koricheva, J., Gurevitch, J. et Mengersen, K. (Éds.) (2013). Handbook of meta-analysis in ecology and evolution. Princeton University Press.

Lohr, S.L. (2009). Multiple-frame surveys. *Handbook of Statistics*, 29, 71-88.

Lohr, S.L. (2011). Autres plans de sondage : échantillonnage avec bases de sondage multiples chevauchantes. *Techniques d'enquête*, 37, 2, 213-232. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11608-fra.pdf.

Næsset, E., et Gjevestad, J.G. (2008). Performance of GPS precise point positioning under conifer forest canopies. *Photogrammetric Engineering & Remote Sensing*, 74(5), 661-668.

O'Muircheartaigh, C., et Pedlow, S. (2002). Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLSY97. Dans *Proceedings of the 2002 Joint Statistical Meetings*, American Statistical Association, 2557-2562.

Roberts, G., et Binder, D. (2009). Analyses based on combining similar information from multiple surveys. Dans *Joint Statistical Meetings 2009 Survey Research Methods Section*, 2138-2147.

Rubin, D.B., et Weisberg, S. (1974). The variance of a linear combination of independent estimators using estimated weights. *ETS Research Bulletin Series*, 1974(2), i-5.

Schmidt, F.L., et Hunter, J.E. (2014). Methods of meta-analysis: Correcting error and bias in research findings. Sage publications.

Ståhl, G., Allard, A., Esseen, P.-A., Glimskår, A., Ringvall, A., Svensson, J., Sundquist, S., Christensen, P., Gallegos Torell, Å., Högström, M., Lagerqvist, K., Marklund, L., Nilsson, B. et Inghe, O. (2011). National Inventory of Landscapes in Sweden (NILS)-scope, design, and experiences from establishing a multiscale biodiversity monitoring system. *Environmental Monitoring and Assessment*, 173(1-4), 579-595.

Tomppo, E., Gschwantner, T., Lawrence, M. et McRoberts, R.E. (Éds.) (2009). National Forest Inventories: Pathways for Common Reporting. *Springer Science & Business Media*.