Techniques d'enquête, décembre 2001
Vol. 27, N° 2, pp. 157-169
Statistique Canada, N° 12-001 au catalogue

L'échantillonnage double

M.A. Hidiroglou [1]

Résumé

La théorie de l'échantillonnage double est d'habitude présentée en supposant que l'un des échantillons est imbriqué dans l'autre. Ce genre de sondage est appelé sondage à deux phases. L'échantillon de première phase fournit de l'information auxiliaire ($\boldsymbol{x}$) relativement peu chère à obtenir, alors que l'échantillon de deuxième phase contient les variables d'intérêt. On se sert des données de la première phase de plusieurs façons : (a) pour stratifier l'échantillon de deuxième phase; (b) pour améliorer l'estimation en se servant de l'estimateur par différence, quotient ou régression; (c) pour sous échantillonner un ensemble d'unités non répondantes. Cependant, il n'est pas nécessaire que l'un des échantillons soit imbriqué dans l'autre ou soit sélectionné dans la même base de sondage. Le cas de l'échantillonnage double *non-imbriqué* est légèrement abordé dans les livres classiques du sondage (Des Raj 1968, Cochran 1977). Cette méthode est présentement utilisée dans plusieurs organismes nationaux d'enquêtes.

Cet article consolide l'échantillonnage double en le présentant sous forme unifiée. Plusieurs exemples de sondage utilisées à Statistique Canada illustrent cette unification.

Mots clés : L'échantillonnage double; données auxiliaires; régression; optimale.

1. Introduction

La théorie de l'échantillonnage double est habituellement présentée en supposant que l'un des échantillons est imbriqué dans l'autre. Ce genre de sondage est appelé sondage à deux phases. L'échantillon de première phase fournit de l'information auxiliaire ($\boldsymbol{x}$) relativement peu chère à obtenir, alors que l'échantillon de deuxième phase contient les variables d'intérêts. On se sert des données de la première phase de plusieurs façons : (a) pour stratifier l'échantillon de deuxième phase ; (b) pour améliorer l'estimation en se servant de l'estimateur par différence, quotient ou régression ; (c) pour sous échantillonner un ensemble d'unités non répondantes. L'échantillonnage à deux phases est une technique puissante et rentable qui a une longue histoire. Neyman (1938) a été le premier à la proposer. Rao (1973) a étudié l'échantillonnage double dans le contexte de la stratification et des études analytiques. Cochran (1977) présentait les résultats fondamentaux de l'échantillonnage à deux phases, y compris les estimateurs de régression les plus simples pour les plans d'échantillonnage de ce genre. Des travaux plus récents en la matière comprennent ceux de Breidt et Fuller (1993), qui ont mis au point des méthodes d'estimation efficaces sur le plan des calculs pour l'échantillonnage à trois phases, en présence de données auxiliaires. Chaudhuri et Roy (1994) se sont, pour leur part, penchés sur les propriétés optimales des estimateurs de régression plus simples mais bien connues de l'échantillonnage à deux phases. Hidiroglou et Särndal (1998) ont suggéré des estimateurs basés sur le calage et la régression pour l'échantillonnage à deux phases afin de tenir compte de la disponibilité de données auxiliaires aux deux niveaux du plan de sondage.

Les deux cas de l'échantillonnage double, imbriqué et non-imbriqué sont traités séparément dans la littérature. Cependant, il n'est pas nécessaire que l'un des échantillons soit imbriqué dans l'autre, ou même sélectionné dans la même base de sondage. Ce cas est appelé l'échantillonnage double *non-imbriqué*. Il est légèrement abordé dans les livres classiques du sondage (Des Raj 1968, Cochran 1977). Cette méthode est présentement utilisée dans plusieurs organismes de sondage. L'Enquête canadienne sur l'emploi, la rémunération et les heures (EERH) menée à Statistique Canada en est un exemple (Rancourt et Hidiroglou 1998). Dans cette enquête, deux échantillons indépendants sont sélectionnés à partir de deux bases de sondage différentes, mais représentent le même univers. Les données auxiliaires ($\boldsymbol{x}$), qui comprennent le nombre d'employés et sont obtenues d'un échantillon sélectionné d'une base administrative provenant de l'Agence des douanes et du Revenu du Canada. Ces mêmes variables ainsi que les variables d'intérêts (y), le nombre d'heures travaillées par les employés et le sommaire de la rémunération, sont recueillies à partir d'un échantillon sélectionné du Registre des entreprises de Statistique Canada. Deville (1999) décrit une situation semblable à l'INSEE pour une enquête auprès des ménages.

Cependant, ils découlent d'un seul estimateur et ne diffèrent que par leurs variances. Cet article se structure comme suit. La partie 2 expose la notation. La partie 3 présente l'unification de ces deux méthodes de l'échantillonnage double. À la partie 4, on présente la variance estimée, l'estimateur de calage imbriqué et non-imbriqué. On donne plusieurs exemples à la partie 5. La partie 6 récapitule brièvement ce qui a été appris.

1. M.A. Hidiroglou, Division des méthodes d'enquêtes auprès des entreprises, Édifice R.-H.-Coats, 11-A, Statistique Canada, Ottawa (Ontario), K1A 0T6. Courriel : hidirog@statcan.ca.

2. Notation

2.1 Cas imbriqué

La population est représentée par $U = \{1, ..., k, ..., N\}$. On prélève un premier échantillon probabiliste s_1 ($s_1 \subseteq U$) sélectionné de la population U selon un plan d'échantillonnage pour lequel la probabilité de sélection est $\pi_{1k} = P(k \in s_1)$ pour la $k^{\text{ième}}$ unité choisie dans s_1. Étant donné s_1 un deuxième échantillon s_2 ($s_2 \subseteq s_1 \subseteq U$) est tiré de s_1 selon un plan d'échantillonnage pour lequel la probabilité conditionnelle de sélection est $\pi_{2k|s_1} = P(k \in s_2 \mid s_1)$ pour la $k^{\text{ième}}$ unité choisie dans s_2. Notons qu'il s'agit de probabilités conditionnelles puisqu'elles supposent qu'on connaît s_1. La figure 1 représente un exemple d'échantillons imbriqués.

On suppose que $\pi_{1k} > 0$ pour toutes les valeurs $k \in U$ et que $\pi_{2k|s_1} > 0$ pour toutes les valeurs $k \in s_1$. Le poids de l'unité échantillonnée k sera dénoté par $w_{1k} = 1/\pi_{1k}$ pour l'échantillon de la première phase et $w_{2k} = 1/\pi_{2k|s_1}$ pour celui de la deuxième phase. Le poids d'échantillonnage global d'une unité de deuxième phase sélectionnée, $k \in s_2$, sera donc $w_k^* = w_{1k}\, w_{2k}$.

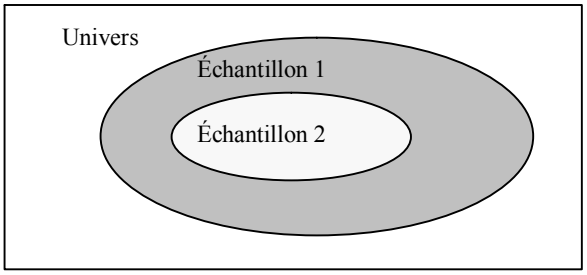

Figure 1. Échantillons imbriqués.

Appelons $\boldsymbol{x}$ le vecteur auxiliaire disponible de l'échantillon de la première phase, et $\boldsymbol{x}_k$ la valeur pour l'unité k. Comme le font Hidiroglou et Särndal (1998), divisons $\boldsymbol{x}_k$ en deux parties $\boldsymbol{x}_{1k}$ et $\boldsymbol{x}_{2k}$. Les valeurs du vecteur $\boldsymbol{x}_{1k}$ sont supposées connues pour l'ensemble de la population U, alors que les valeurs du vecteur $\boldsymbol{x}_{2k}$ ne sont disponibles que pour l'échantillon de la première phase s_1.

2.2 Cas non-imbriqué

Il est possible que les deux échantillons soient sélectionnés indépendamment de la même base de sondage, ou même de bases de sondage différentes (mais équivalentes). Les figures 2 et 3 représentent des exemples de cas non-imbriqués.

Le cas non-imbriqué représenté par la figure 3 n'est pas considéré dans cet article. Ce cas peut être compliqué pour des plans de sondages arbitraires, puisqu'il faut calculer des probabilités de sélection conjointes entre les deux échantillons s_1 et s_2. Ce calcul se simplifie si l'on considère que deux échantillons s_1 et s_2 ont été sélectionnés en se servant d'un plan de sondage tel que le plan aléatoire simple (avec ou sans remise). On peut alors se servir des résultats de Tam (1984) afin d'obtenir les probabilités conjointes requises pour le calcul de l'estimation de la variance pour un estimateur de total $Y = \sum_U y_k$.

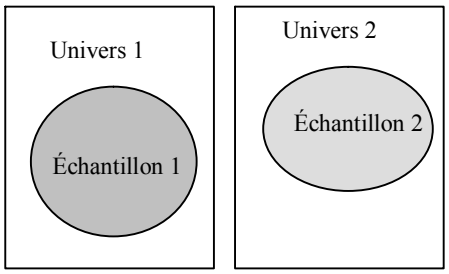

Figure 2. Deux échantillons indépendants bases de sondages différentes.

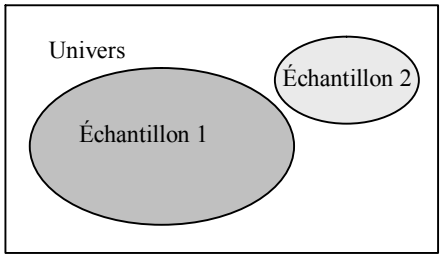

Figure 3. Deux échantillons sélectionnés indépendamment dans la même base de sondage.

Pour le cas que nous allons étudier, nous supposons que les échantillons s_1 et s_2 ont été sélectionnés indépendamment à partir de deux bases de sondage différentes (Figure 2), $U_1 = \{1, ..., k, ..., N_1\}$ et $U_2 = \{1, ..., k, ..., N_2\}$. Les échantillons s_1 ($s_1 \subseteq U_1$) et s_2 ($s_2 \subseteq U_2$) sont sélectionnés tel que leur probabilité de sélection est respectivement $\pi_{1k}^{(1)} = P(k \in s_1) > 0$ et $\pi_{2k}^{(2)} = P(k \in s_2) > 0$. Le poids de l'unité k est $w_{1k}^{(1)} = 1/\pi_{1k}^{(1)}$ pour le premier échantillon s_1 et $w_{2k}^{(2)} = 1/\pi_{2k}^{(2)}$ pour le deuxième échantillon s_2. Les indices supérieurs (1) et (2) sont utilisés pour différencier les probabilités de sélection du cas imbriqué. Il se peut que les unités de sondage diffèrent entre les deux bases, tout en ayant la même couverture. Des exemples de telles procédures d'échantillonnage ont été mentionnés dans l'introduction, et plus de détails sont donnés dans le deuxième exemple de la section 5.3.

Soit $\boldsymbol{x}_k = (\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}')'$, un vecteur de donnés auxiliaires. On suppose que $\boldsymbol{x}_{1k}^{(1)}$ est connue pour toutes les unités appartenant à la base de sondage U_1, tandis que $\boldsymbol{x}_k^{(1)}$ est seulement observé pour l'échantillon s_1. Nous observons $y_k^{(2)}$, $\boldsymbol{x}_k^{(2)}$ pour l'échantillon s_2. Le degré de divergence entre les valeurs des données varie selon la complexité de l'unité d'échantillonnage, et combien ces unités diffèrent en concept entre les deux bases. Pour les unités plus « simples », les données rapportées pour des unités « semblables » dans s_1 et s_2 devraient être à peu près égales. Une différence de similarité entre les même unités

dans s_1 et s_2 serait fort probablement causé par une différence dans le questionnaire ou par le fait que différents répondants remplissent le questionnaire. Notons aussi que $\boldsymbol{X}_1 = \sum_{U_1} \boldsymbol{x}_{1k}^{(1)} = \sum_{U_2} \boldsymbol{x}_{1k}^{(2)}$ puisque U_1 et U_2 ont la même couverture.

3. Estimateur optimal pour les échantillons imbriqués et non-imbriqués

Pour les deux cas, imbriqué et non-imbriqué, l'objectif est d'estimer le total de population $Y = \sum_U y_k$ où y_k représente la valeur de l'unité $k \in U$. Un estimateur non biaisé de Y est $\hat{Y}_{\mathrm{HT}} = \sum_{s_2} w_k^* y_k$, où $w_k^* = w_{1k} w_{2k}$ pour le cas imbriqué et $w_k^* = w_{2k}^{(2)}$ pour le cas non-imbriqué.

Les données auxiliaires nous permettent de modifier les poids d'échantillonnage et à l'aide de facteurs de calage calculés d'après l'information complémentaire à différents niveaux (univers, échantillon de première phase). On modifie le poids d'échantillonnage d'une unité en le multipliant par un facteur de calage, et on appelle le produit « poids de calage ». Le tableau 1 résume la représentation des données disponibles pour les cas imbriqués et non-imbriqués, correspondant aux figures 1 et 2.

Tableau 1
Données disponibles pour la population et les échantillons

Niveau	Cas imbriqué	Cas non-imbriqué
Population	$\boldsymbol{x}_{1k}$: connu pour $k \in U$	$\boldsymbol{x}_{1k}^{(1)}$: connu pour $k \in U_1$
Premier échantillon	$\boldsymbol{x}_k$: observé pour $k \in s_1$	$\boldsymbol{x}_k^{(1)}$: observé pour $k \in s_1$
Deuxième échantillon	y_k, $\boldsymbol{x}_k$: observé pour $k \in s_2$	$y_k^{(2)}$, $\boldsymbol{x}_k^{(2)}$: observé pour $k \in s_2$

Un estimateur de régression qui peut servir à estimer le total Y de la population pour les échantillons imbriqués et non-imbriqués est :

$$\hat{\hat{Y}}_{\mathrm{REG}} = \hat{\hat{Y}}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \boldsymbol{B}_1 + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})' \boldsymbol{B}. \quad (3.1)$$

Les différentes valeurs des totaux des données auxiliaires $\boldsymbol{x}$ et y dans l'équation (3.1) sont présentées dans le tableau 2.

On suppose aussi que les variances, $V(\hat{\hat{Y}}_{\mathrm{HT}})$, ainsi que les covariances $\mathrm{Cov}(\hat{\boldsymbol{X}}, \hat{\hat{\boldsymbol{X}}}')$, $\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}')$, $\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{X}}}')$, $\mathrm{Cov}(\hat{\hat{Y}}_{\mathrm{HT}}, \hat{\boldsymbol{X}}')$ et $\mathrm{Cov}(\hat{\hat{Y}}_{\mathrm{HT}}, \hat{\hat{\boldsymbol{X}}}')$ sont connues ou estimables.

Afin de ne pas alourdir la notation on élimine les indices supérieurs pour le reste de cette section. L'estimation des paramètres, $\boldsymbol{B}$ et $\boldsymbol{B}_1$ ainsi que celles de leurs variances associées, refléteront le fait que l'on sonde différemment pour les deux cas imbriqué et non-imbriqué. On minimise la variance de $\hat{\hat{Y}}_{\mathrm{REG}}$. afin d'obtenir les estimateurs de $\boldsymbol{B}$ et $\boldsymbol{B}_1$. Cette variance est :

$$\begin{aligned} V(\hat{\hat{Y}}_{\mathrm{REG}}) = {} & V(\hat{\hat{Y}}_{\mathrm{HT}}) + \boldsymbol{B}_1' V(\hat{\boldsymbol{X}}_1) \boldsymbol{B}_1 + \boldsymbol{B}' V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}) \boldsymbol{B} \\ & - 2\,\mathrm{Cov}(\hat{\hat{Y}}_{\mathrm{HT}}, \hat{\boldsymbol{X}}_1') \boldsymbol{B}_1 + 2\,\mathrm{Cov}(\hat{\hat{Y}}_{\mathrm{HT}}, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})') \boldsymbol{B} \\ & - 2\,\boldsymbol{B}_1' \mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})') \boldsymbol{B}. \end{aligned} \quad (3.2)$$

Les paramètres sont obtenus en prenant la dérivée de (3.2) par rapport à $\boldsymbol{B}$ et $\boldsymbol{B}_1$. Nous obtenons les deux équations suivantes à résoudre :

$$\begin{aligned} & V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}) \boldsymbol{B} + \mathrm{Cov}((\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}), \hat{\hat{Y}}_{\mathrm{HT}}) \\ & \qquad - \mathrm{Cov}((\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}), \hat{\boldsymbol{X}}_1') \boldsymbol{B}_1 = \boldsymbol{0} \end{aligned} \quad (3.3)$$

et

$$\begin{aligned} & - \mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})') \boldsymbol{B} - \mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) \\ & \qquad + V(\hat{\boldsymbol{X}}_1) \boldsymbol{B}_1 = \boldsymbol{0}. \end{aligned} \quad (3.4)$$

Tableau 2
Sommes des données auxiliaires $\boldsymbol{x}$ et y pour les cas imbriqués et non-imbriqués

Niveau	Cas imbriqué	Cas non-imbriqué
Population	$\boldsymbol{X}_1 = \sum_U \boldsymbol{x}_{1k}$	$\boldsymbol{X}_1 = \sum_{U_1} \boldsymbol{x}_{1k}^{(1)}$
Premier échantillon	$\hat{\boldsymbol{X}}_1 = \sum_{s_1} w_{1k} \boldsymbol{x}_{1k}$; $\hat{\boldsymbol{X}} = \sum_{s_1} w_{1k} \boldsymbol{x}_k$	$\hat{\boldsymbol{X}}_1 = \sum_{s_1} w_{1k} \boldsymbol{x}_{1k}^{(1)}$; $\hat{\boldsymbol{X}} = \sum_{s_1} w_{1k} \boldsymbol{x}_k^{(1)}$
Deuxième échantillon	$\hat{\hat{\boldsymbol{X}}}_1 = \sum_{s_2} w_k^* \boldsymbol{x}_{1k}$; $\hat{\hat{\boldsymbol{X}}} = \sum_{s_2} w_k^* \boldsymbol{x}_k$	$\hat{\hat{\boldsymbol{X}}}_1 = \sum_{s_2} w_{2k} \boldsymbol{x}_{1k}^{(2)}$; $\hat{\hat{\boldsymbol{X}}} = \sum_{s_2} w_{2k} \boldsymbol{x}_k^{(2)}$
	$\hat{\hat{Y}}_{\mathrm{HT}} = \sum_{s_2} w_k^* y_k$	$\hat{\hat{Y}}_{\mathrm{HT}} = \sum_{s_2} w_{2k}^* y_k^{(2)}$

En résolvant le système d'équations (3.3) et (3.4), nous obtenons les paramètres requis $\boldsymbol{B}$ et $\boldsymbol{B}_1$. C'est-à-dire :

$$\boldsymbol{B} = \boldsymbol{T}^{-1}\boldsymbol{H} \tag{3.5}$$

où

$$\boldsymbol{T} = V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}) - (\mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'))' V^{-1}(\hat{\boldsymbol{X}}_1)(\mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')),$$

$$\boldsymbol{H} = (\mathrm{Cov}((\hat{\hat{\boldsymbol{X}}} - \hat{\boldsymbol{X}}), \hat{\hat{Y}}_{\mathrm{HT}})) + (\mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'))' V^{-1}(\hat{\boldsymbol{X}}_1)\,\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{Y}}}_{\mathrm{HT}})$$

et

$$\boldsymbol{B}_1 = \boldsymbol{T}_1^{-1}\boldsymbol{H}_1 \tag{3.6}$$

où

$$\boldsymbol{T}_1 = V(\hat{\boldsymbol{X}}_1),$$

et

$$\boldsymbol{H}_1 = \mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) + \mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')'\boldsymbol{B}.$$

Résultat 1 : Un estimateur de régression optimal pour les échantillons imbriqués et non-imbriqués est :

$$\hat{\hat{Y}}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)'\,\hat{\boldsymbol{B}}_{1,\mathrm{OPT}} + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'\,\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}} \tag{3.7}$$

où

$$\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}} = \hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{H}} \tag{3.8}$$

et

$$\hat{\hat{\boldsymbol{B}}}_{1,\mathrm{OPT}} = \hat{\boldsymbol{T}}_1^{-1}\hat{\boldsymbol{H}}_1. \tag{3.9}$$

$\hat{\boldsymbol{T}}_1$, $\hat{\boldsymbol{H}}_1$, $\hat{\boldsymbol{T}}$ et $\hat{\boldsymbol{H}}$ sont les valeurs estimées de $\boldsymbol{T}_1$, $\boldsymbol{H}_1$, $\boldsymbol{T}$ et $\boldsymbol{H}$ obtenues à l'aide d'un cadre de travail menant à l'inférence basée sur le plan de sondage. Ces valeurs dépendent de la méthode de sélection des échantillons. La variance de $\hat{\hat{Y}}_{\mathrm{OPT}}$ et de son estimateur associé dépendent de l'imbrication ou non du plan de sondage. Puisque l'estimateur $\hat{\hat{Y}}_{\mathrm{OPT}}$ est optimal, il s'ensuit que l'estimateur par régression $\hat{\hat{Y}}_{\mathrm{OPT}}$ est aussi optimal. Cette forme a été discutée par Montanari (1987) pour le cas d'un plan de sondage à une phase.

3.1 Application à l'échantillonnage double imbriqué

La théorie pour ce cas est développée en se servant d'une approche conditionnelle. Si deux paramètres quelconques $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$ sont estimés par $\hat{\hat{\boldsymbol{\theta}}}_1$ et $\hat{\hat{\boldsymbol{\theta}}}_2$ à partir de l'échantillon s_2, il en suit que si on conditionne sur l'échantillon réalisé s_1, on obtient le résultat bien connu suivant :

(i) L'espérance mathématique de $\hat{\hat{\boldsymbol{\theta}}}$ est $E(\hat{\hat{\boldsymbol{\theta}}}) = E_1 E_2(\hat{\hat{\boldsymbol{\theta}}} \mid s_1)$, où E_2 dénote l'espérance de $\hat{\hat{\boldsymbol{\theta}}}$ étant donné s_1.

(ii) La variance de $\hat{\hat{\boldsymbol{\theta}}}$ est :

$$V(\hat{\hat{\boldsymbol{\theta}}}) = E_1 V_2(\hat{\hat{\boldsymbol{\theta}}} \mid s_1) + V_1 E_2(\hat{\hat{\boldsymbol{\theta}}} \mid s_1). \tag{3.10}$$

(iii) La covariance entre $\hat{\hat{\boldsymbol{\theta}}}_1$ et $\hat{\hat{\boldsymbol{\theta}}}_2$ est :

$$\mathrm{Cov}(\hat{\hat{\boldsymbol{\theta}}}_1, \hat{\hat{\boldsymbol{\theta}}}_2') = E_1\,\mathrm{Cov}_2((\hat{\hat{\boldsymbol{\theta}}}_1, \hat{\hat{\boldsymbol{\theta}}}_2') \mid s_1) + \mathrm{Cov}_1(E_2(\hat{\hat{\boldsymbol{\theta}}}_1 \mid s_1), E_2(\hat{\hat{\boldsymbol{\theta}}}_2' \mid s_1)).$$

Les différentes composantes de $\hat{\boldsymbol{T}}$, $\hat{\boldsymbol{H}}$, $\hat{\boldsymbol{T}}_1$ et de $\hat{\boldsymbol{H}}_1$ seront estimées en supposant un plan de sondage ayant une taille d'échantillon non-fixe. Le cas de la taille l'échantillon fixe suit facilement, étant un cas spécial de ce dernier. En se servant des expressions (i)–(iii), on peut réexprimer les termes qui définissent le paramètre $\boldsymbol{B}$ comme :

$$\mathrm{Cov}(\hat{\boldsymbol{X}}, \hat{\hat{\boldsymbol{X}}}') = \mathrm{Cov}(\hat{\boldsymbol{X}}, \hat{\boldsymbol{X}}') = V(\hat{\boldsymbol{X}});$$

$$\mathrm{Cov}(\hat{\hat{Y}}_{\mathrm{HT}}, \hat{\boldsymbol{X}}') = \mathrm{Cov}(\hat{Y}_{\mathrm{HT}}, \hat{\boldsymbol{X}}');$$

$$V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}) = E_1\left[\sum\sum_{s_1} c_{2k\ell|s_1}\,\boldsymbol{x}_k\,\boldsymbol{x}_\ell'\right];$$

$$\mathrm{Cov}\left[\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'\right] = \boldsymbol{0};$$

et

$$\mathrm{Cov}(\hat{\hat{\boldsymbol{X}}}, \hat{\hat{\boldsymbol{Y}}}_{\mathrm{HT}}) = \mathrm{Cov}(\hat{\mathbf{X}}, \hat{Y}_{\mathrm{HT}}) + E_1\left[\sum\sum_{s_1} c_{2k\ell|s_1}\,\boldsymbol{x}_k\,\boldsymbol{x}_\ell'\right]; \tag{3.11}$$

où $c_{2k\ell|s_1} = (\pi_{2k\ell|s_1} - \pi_{2k|s_1}\,\pi_{2\ell|s_1})/\pi_k^*\,\pi_\ell^*$ et $\hat{Y}_{\mathrm{HT}} = \sum_{s_1} y_k/\pi_{1k}$. Les probabilités d'inclusion dans ces expressions sont $\pi_{2k\ell|s_1} = \Pr(k, \ell \in s_2 \mid s_1)$ et $\pi_k^* = \pi_{1k}\,\pi_{2k|s_1}$. On peut exprimé $\boldsymbol{B}$ plus simplement comme :

$$\boldsymbol{B} = \left[E_1\left(\sum\sum_{s_1} c_{2k\ell|s_1}\,\boldsymbol{x}_k\,\boldsymbol{x}_\ell'\right)\right]^{-1} E_1\left[\sum\sum_{s_1} c_{2k\ell|s_1}\,\boldsymbol{x}_k\,y_\ell\right] \tag{3.12}$$

et son estimateur optimal est :

$$\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}} = \left[\sum\sum_{s_2} \hat{c}_{2k\ell|s_1}\boldsymbol{x}_k\boldsymbol{x}_\ell'\right]^{-1}\left[\sum\sum_{s_2} \hat{c}_{2k\ell|s_1}\boldsymbol{x}_k y_\ell\right] \tag{3.13}$$

où $\hat{c}_{2k\ell|s_1} = c_{2k\ell|s_1} / \pi_{2k\ell|s_1}$.

Rappelons que l'estimateur optimal par régression $\hat{\mathbf{B}}_{1,\mathrm{OPT}}$, est donné par l'expression (3.9), où

$$\hat{\boldsymbol{T}}_1 = \hat{\boldsymbol{V}}(\hat{\boldsymbol{X}}_1)$$

et

$$\hat{\mathbf{H}}_1 = \hat{\mathrm{Cov}}(\hat{\mathbf{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) + \hat{\mathrm{Cov}}(\hat{\mathbf{X}}_1, \hat{\mathbf{X}}')\,\hat{\hat{\mathbf{B}}}_{\mathrm{OPT}} - \hat{\mathrm{Cov}}(\hat{\mathbf{X}}_1, \hat{\hat{\mathbf{X}}}')\,\hat{\hat{\mathbf{B}}}_{\mathrm{OPT}}.$$

Nous définissons les composantes des termes $\hat{\mathbf{T}}_1$ et de $\hat{\mathbf{H}}_1$ comme suit. En premier lieu, nous estimons $V(\hat{\boldsymbol{X}}_1) = \sum\sum_{s_1} c_{1k\ell}\, \boldsymbol{x}_{1k}\, \boldsymbol{x}'_{1\ell}$ par

$$\hat{V}(\hat{\boldsymbol{X}}_1) = \sum\sum_{s_1} \hat{c}_{1k\ell}\, \boldsymbol{x}_{1k}\, \boldsymbol{x}'_{1\ell} \tag{3.14}$$

où $c_{1k\ell} = (\pi_{1k\ell} - \pi_{1k}\,\pi_{1\ell}) / (\pi_{1k}\,\pi_{\ell})$ et $\hat{c}_{1k\ell} = c_{1k\ell} / \pi_{1k\ell}$.

Aussi puisque

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) &= E_1\,\mathrm{Cov}_2\,[(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) \mid s_1] \\ &\quad + \mathrm{Cov}_1[E_2\,(\hat{\boldsymbol{X}}_1 \mid s_1), E_2\,(\hat{\hat{Y}}_{\mathrm{HT}} \mid s_1)] \\ &= \mathrm{Cov}_1\,(\hat{\boldsymbol{X}}_1, \hat{Y}_{\mathrm{HT}}) \\ &= \sum\sum_{U_1} c_{1k\ell}\, \mathbf{x}_{1k}\, y_{1\ell} \qquad (3.15)\end{aligned}$$

on estime $\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}})$ par

$$\hat{\mathrm{Cov}}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) = \sum\sum_{s_2} c^*_{1k\ell}\, \boldsymbol{x}_{1k}\, y_{\ell} \tag{3.16}$$

où

$$c^*_{1k\ell} = c_{1k\ell} / \pi^*_{k\ell}, \pi^*_{k\ell} = \pi_{1k\ell}\,\pi_{2k\ell|s_1},$$

$$\pi_{1k\ell} = \Pr\,(k, \ell \in s_1),$$

$$\pi_{2k\ell|s_1} = \Pr\,(k, \ell \in s_1)$$

et

$$\pi^*_k = \pi_{1k}\pi_{2k|s_1}.$$

De même,

$$\hat{\mathrm{Cov}}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{X}}}') = \sum\sum_{s_2} c^*_{1k\ell}\, \boldsymbol{x}_{1k}\, y'_{\ell} \tag{3.17}$$

et

$$\hat{\mathrm{Cov}}(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}') = \sum\sum_{s_1} \hat{c}_{1k\ell}\, \boldsymbol{x}_{1k}\, x'_{\ell}. \tag{3.18}$$

Dans le cas de l'échantillonnage double imbriqué, l'estimateur optimal de $\boldsymbol{B}_1$ est donc :

$$\hat{\boldsymbol{B}}_{1,\mathrm{OPT}} = (\hat{V}\,(\hat{\boldsymbol{X}}_1))^{-1} \begin{bmatrix} \hat{\mathrm{Cov}}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{\mathrm{HT}}) \\ \quad + (\hat{\mathrm{Cov}}(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}') \\ \quad - \hat{\mathrm{Cov}}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{X}}}'))\,\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}} \end{bmatrix} \tag{3.19}$$

où les composantes de $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ ont été définies par les expressions (3.14) – (3.18).

La forme optimale des estimateurs $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ et $\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}$ a ses avantages et désavantages. Un des plus grand avantage de la forme optimale, tel que rapporté par Cassady et Valliant (1993), Rao (1994), ainsi que Montanari (2000), est qu'elle a de bonnes inférences conditionnelles (en conditionnant sur la variable auxiliaire $\boldsymbol{x}$). Telle que le remarque Montanari (2000), l'optimalité asymptotique de $\hat{\hat{Y}}_{\mathrm{OPT}}$ est strictement une propriété basée sur le plan de sondage, et réalisé conditionnellement sur la population finie. Le plus grand désavantage de la forme optimale est qu'elle requiert le calcul de probabilités conjointes.

Nous pouvons, cependant, tirer profit de la forme optimale, et l'exprimer de façon plus simple pour de nombreux plans de sondage. Pour des plans de sondage où la sélection de l'échantillon est avec probabilité inégale et sans remise, on peut contourner le calcul de probabilité conjointe par approximation de la variance exacte. Plusieurs auteurs, dont Hartley et Rao (1962), Deville (1999), Berger (1998), Rósen (2000), ainsi que Brewer (2000). Pour un échantillon s (tiré sans remise et à une phase), Tillé (2001) propose l'approximation suivante de la variance estimée de $\hat{Y}_{\mathrm{HT}} = \sum_s y_k / \pi_k$, où π_k est la probabilité de sélection de l'unité k :

$$\begin{aligned}\hat{V}(\hat{Y}_{\mathrm{HT}}) &= \sum_s \frac{c_k}{\pi_k^2}(y_k - y_k^*)^2 \\ &= \sum_s c_k \left(\frac{y_k}{\pi_k} - \tilde{y}\right)^2. \qquad (3.20)\end{aligned}$$

Ici, c_k est la variable qui sert à l'approximation, $y_k^* = \pi_k \sum_s c_\ell\, y_\ell / \pi_\ell / \sum_s c_\ell$, $\tilde{y} = y_k^* / \pi_k$ et π_k est la probabilité de sélection de l'unité k. Tillé (2001) fournit plusieurs exemples de c_k pour divers plan de sondage.

Notons que cette formule est exacte dans le cas d'un plan de sondage stratifié, où l'échantillon est aléatoire simple et sans remise dans chaque strate U_h $(h = 1, ..., L)$ de la population U. Pour ce cas, si k dénote une unité appartenant à l'échantillon s_h sélectionné dans le strate U_h, alors $c_k = n_h / (n_h - 1)\,(1 - n_h/N_h)$ si $k \in U_h$ et 0 autrement, $\pi_k = n_h / N_h$ si $k \in U_h$ et 0 autrement. Nous obtenons ainsi la variance estimée exacte, $\hat{V}(\hat{Y}_{\mathrm{HT}}) = \sum_{h=1}^{L} N_h^2 (1 - n_h / N_h)$ $\sum_{s_h} (y_k - \bar{y}_h)^2 / n_h (n_h - 1)$. La formule est aussi exacte dans le cas d'un plan de sondage stratifié où l'échantillon a été sélectionné avec remise. Ici $c_k = 1$ pour toutes les unités appartenant à la strate U_h et zéro autrement. En ce servant de cette approximation, les doubles sommes qui apparaissent dans $\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}$ ainsi que $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ peuvent être exprimées comme de simples sommes. Hidiroglou et Särndal (1998) ont contourné le problème de doubles sommes dans l'estimation de $\boldsymbol{B}$ et $\boldsymbol{B}_1$ en proposant l'estimateur GREG, $\hat{\hat{Y}}_{\mathrm{GREG}}$, pour un plan de sondage à deux phases imbriquées :

$$\hat{\hat{Y}}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \hat{\boldsymbol{B}}_{1,\text{GREG}} + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})' \hat{\hat{\boldsymbol{B}}}_{\text{GREG}}$$

où

$$\hat{\hat{\boldsymbol{B}}}_{\text{GREG}} = \left(\sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \boldsymbol{x}_k \boldsymbol{x}_k'}{\sigma_{2k}^2} \right)^{-1} \sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \boldsymbol{x}_k y_k}{\sigma_{2k}^2}, \quad (3.21)$$

$$\hat{\boldsymbol{B}}_{1,\text{GREG}} = \left(\sum_{s_1} \frac{w_{1k} \boldsymbol{x}_{1k} \boldsymbol{x}_{1k}'}{\sigma_{1k}^2} \right)^{-1} \left\{ \begin{array}{l} \sum_{s_2} \frac{w_k^* \boldsymbol{x}_{1k} y_k}{\sigma_{1k}^2} + \sum_{s_1} \frac{w_{1k} \boldsymbol{x}_{1k} \boldsymbol{x}_k'}{\sigma_{1k}^2} \hat{\hat{\boldsymbol{B}}}_{\text{GREG}} \\ \qquad - \sum_{s_2} \frac{w_k^* \boldsymbol{x}_{1k} \boldsymbol{x}_k'}{\sigma_{1k}^2} \hat{\hat{\boldsymbol{B}}}_{\text{GREG}} \end{array} \right\} \quad (3.22)$$

où $\{\sigma_{1k}^2 : k \in s_1\}$ et $\{\sigma_{2k}^2 : k \in s_2\}$ sont des facteurs préétablis et positifs.

Les estimateurs $\hat{\hat{\boldsymbol{B}}}_{\text{GREG}}$ et $\hat{\boldsymbol{B}}_{1,\text{GREG}}$ peuvent être justifiés soit en supposant des modèles de régression différents pour chaque phase, ou en se servant de deux calages successifs. Pour l'approche calage, on obtient premièrement des poids de calage $\tilde{w}_{1k}$ pour la première phase qui respectent l'équation de calage $\sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_{1k} = \sum_U \boldsymbol{x}_{1k}$. Ces poids de calages peuvent exprimer comme le produit des poids de sondage w_{1k} et un facteur de sondage de calage g_{1k} où :

$$g_{1k} = 1 + \left(\sum_U \boldsymbol{x}_{1k} - \sum_{s_1} w_{1k} \boldsymbol{x}_{1k} \right)' \left(\sum_{s_1} w_{1k} \frac{\boldsymbol{x}_{1k} \boldsymbol{x}_{1k}'}{\sigma_{1k}^2} \right)^{-1} \frac{\boldsymbol{x}_{1k}}{\sigma_{1k}^2} \quad (3.23)$$

pour $k \in s_1$.

Les poids de calage de première phase $\tilde{w}_{1k}$ servent ensuite de poids initiaux pour calculer les poids de calage globaux $\tilde{w}_k^*$. Ces poids de calage globaux respectent l'équation de calage basée sur la deuxième phase $\sum_{s_2} \tilde{w}_{1k}^* \boldsymbol{x}_k = \sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k$. L'estimateur du total, $\hat{\hat{Y}}_{\text{GREG}}$, peut être aussi exprimé en fonction du poids de calage global $\tilde{w}_k^*$ et des valeurs y associées. C'est-à-dire $\hat{\hat{Y}}_{\text{GREG}} = \sum_{s_2} \tilde{w}_k^* y_k$. Les poids globaux calés sont $\tilde{w}_k^* = w_k^* g_k^*$, où $g_k^* = g_{1k} g_{2k}$, et g_{1k} est issu de (3.23), tandis que g_{2k} correspond à

$$g_{2k} = 1 + \left(\sum_{s_1} \tilde{w}_{1k} \boldsymbol{x}_k - \sum_{s_2} \tilde{w}_{1k} w_{2k} \boldsymbol{x}_k \right)' \left(\sum_{s_1} \frac{\tilde{w}_{1k} w_{2k} \boldsymbol{x}_k \boldsymbol{x}_k'}{\sigma_{2k}^2} \right)^{-1} \frac{\boldsymbol{x}_k}{\sigma_{2k}^2} \quad (3.24)$$

pour $k \in s_2$.

Remarque : Les estimateurs $\hat{\boldsymbol{B}}_{1,\text{GREG}}$ (3.21) et $\hat{\hat{\boldsymbol{B}}}_{\text{GREG}}$ (3.22), correspondent au *cas additif* d'Hidiroglou et Särndal (1998), et ont la même forme que les estimateurs de régression optimaux $\hat{\boldsymbol{B}}_{1,\text{OPT}}$ (3.8) et $\hat{\hat{\boldsymbol{B}}}_{\text{OPT}}$ (3.9). En effet, les composantes de l'estimateur de $\boldsymbol{B}$ sont estimées en remplaçant $\boldsymbol{T}$ par $(\sum_{s_2} w_{1k} w_{2k} \boldsymbol{x}_k \boldsymbol{x}_k' / \sigma_{2k}^2)$ et $\boldsymbol{H}$ par $\sum_{s_2} w_{1k} w_{2k} \boldsymbol{x}_k y_k / \sigma_{2k}^2$. Les seconds termes de $\boldsymbol{H}$ et de $\boldsymbol{T}$ sont exactement égaux à zéro. De même, afin d'estimer $\boldsymbol{B}_1$, on remplace $\boldsymbol{T}_1$ par $\sum_{s_1} w_{1k} \boldsymbol{x}_{1k} \boldsymbol{x}_{1k}' / \sigma_{1k}^2$ et $\boldsymbol{H}_1$ par

$$\sum_{s_2} \frac{w_k^* \boldsymbol{x}_{1k} y_{2k}}{\sigma_{1k}^2} + \left(\sum_{s_1} \frac{w_{1k} \boldsymbol{x}_{1k} \boldsymbol{x}_k'}{\sigma_{1k}^2} - \sum_{s_2} \frac{w_k^* \boldsymbol{x}_{1k} \boldsymbol{x}_k'}{\sigma_{1k}^2} \right)' \hat{\hat{\boldsymbol{B}}}_{\text{GREG}}.$$

La variance estimée de $\hat{\hat{Y}}_{\text{GREG}} = \hat{\hat{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \hat{\boldsymbol{B}}_{1,\text{GREG}} + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})' \hat{\hat{\boldsymbol{B}}}_{\text{GREG}}$ est présentée dans Hidiroglou et Särndal (1998).

Remarque : L'efficacité de la forme GREG, tel qu'énoncé dans Särndal, Swensson, et Wretman (1992), requiert que le modèle postulé soit juste. Aussi, si la taille de l'échantillon est assez grande, les estimateurs optimaux sont plus efficaces (Rao 1994) que le GREG. Cependant, si la taille de l'échantillon est relativement petite, un désavantage de la forme optimale OPT est qu'elle est généralement moins stable, et plus complexe à calculer que le GREG. Aussi, tel que rapporté par Särndal (1996), et illustré sous forme de simulation par Montanari (2000), la forme optimale n'est pas tellement plus efficace que GREG, si la taille de l'échantillon est relativement petite. Il se peut même que la variance estimée soit plus grande que celle associée au GREG.

3.2 Application à l'échantillonnage double non-imbriqué

Deville (1999) a considéré le cas non-imbriqué (figure 2), en supposant que $\boldsymbol{x}_{2k}$ est connu pour s_1 et s_2. L'estimateur optimal de régression est :

$$\hat{\hat{Y}}_{\text{OPT}} = \hat{\hat{Y}}_{\text{HT}} + (\hat{\boldsymbol{X}}_2 - \hat{\hat{\boldsymbol{X}}}_2)' \hat{\hat{\boldsymbol{B}}}_{2,\text{OPT}} \quad (3.25)$$

où $\hat{\hat{Y}}_{\text{HT}} = \sum_{s_2} w_{2k} y_k^{(2)}$, $\hat{\boldsymbol{X}}_2 = \sum_{s_1} w_{1k} \boldsymbol{x}_{2k}^{(1)}$, $\hat{\hat{\boldsymbol{X}}}_2 = \sum_{s_2} w_{2k} \boldsymbol{x}_{2k}^{(2)}$. Ici $\hat{\hat{\boldsymbol{B}}}_{2,\text{OPT}} = (\hat{V}(\hat{\boldsymbol{X}}_2) + \hat{V}(\hat{\hat{\boldsymbol{X}}}_2))^{-1} \hat{\text{Cov}}(\hat{\hat{Y}}_{\text{HT}}, \hat{\hat{\boldsymbol{X}}}_2')$ est l'estimateur optimal de $\boldsymbol{B}_2 = (\sum_{U_2} \boldsymbol{x}_{2k} \boldsymbol{x}_{2k}')^{-1} \sum_{U_2} \boldsymbol{x}_{2k} y_k$ si les deux bases de sondage U_1 et U_2 sont indépendantes. La forme des termes de variance et de covariance qui définissent $\hat{\hat{\boldsymbol{B}}}_{2,\text{OPT}}$ dépend des plan de sondage de s_1 et s_2.

On peut améliorer la précision de l'estimateur de $\boldsymbol{X}_2$ en déterminant $\boldsymbol{A}_2$, tel que $\tilde{\boldsymbol{X}}_2 = \boldsymbol{A}_2 \hat{\boldsymbol{X}}_2 + (\boldsymbol{I} - \boldsymbol{A}_2) \hat{\hat{\boldsymbol{X}}}_2$ soit de variance minimum. Ainsi $\boldsymbol{A}_2 = (V(\hat{\boldsymbol{X}}_2) + V(\hat{\hat{\boldsymbol{X}}}_2))^{-1} V(\hat{\boldsymbol{X}}_2)$. En supposant que $V(\hat{\boldsymbol{X}}_2)$ est approximativement un multiple de $V(\hat{\hat{\boldsymbol{X}}}_2)$, c'est-à-dire $V(\hat{\boldsymbol{X}}_2) \doteq \alpha_2 V(\hat{\hat{\boldsymbol{X}}}_2)$, on obtient que $\boldsymbol{A}_2 \doteq \boldsymbol{I}/(1+\alpha_2)$ où $\boldsymbol{I}$ est la matrice d'identité ayant la même dimension que la matrice de covariance $V(\hat{\boldsymbol{X}}_2)$. La valeur

optimale de α_2 est obtenue en minimisant la variance de $\tilde{\boldsymbol{X}}_2$. Un choix sous optimal pour α_2, mais adéquat (Deville 1999), est $\alpha_2 = n_1/(n_1+n_2)$, où n_1 et n_2 sont les tailles respectives des échantillons s_1 et s_2. Il est à noter que Korn et Graubart (1999) ont aussi fait un tel choix pour combiner deux totaux étant estimés de deux sources différentes. En substituant $\tilde{\mathbf{X}}_2$ à la place de $\hat{\mathbf{X}}_2$ dans l'expression (3.25), on obtient

$$\tilde{\boldsymbol{X}}_2 - \hat{\boldsymbol{X}}_2 = (\hat{\boldsymbol{X}}_2 - \hat{\tilde{\boldsymbol{X}}}_2)/(1+\alpha_2). \tag{3.26}$$

L'estimateur de Y du total de la population est donc :

$$\tilde{\tilde{Y}}_{\mathrm{OPT}} = \hat{Y}_{\mathrm{HT}} + (\tilde{\boldsymbol{X}}_2 - \hat{\boldsymbol{X}}_2)'\,\tilde{\tilde{\boldsymbol{B}}}_{2,\mathrm{OPT}} \tag{3.27}$$

où

$$\tilde{\tilde{\boldsymbol{B}}}_{2,\mathrm{OPT}} = -[\hat{V}(\tilde{\boldsymbol{X}}_2 - \hat{\tilde{\boldsymbol{X}}}_2)]^{-1}\,\hat{\mathrm{Cov}}\,(\hat{\tilde{Y}}_{\mathrm{HT}}, (\tilde{\boldsymbol{X}}_2 - \hat{\boldsymbol{X}}_2)'). \tag{3.28}$$

Si l'on substitute (3.26) dans (3.28), on peut réexprimer $\tilde{\tilde{\boldsymbol{B}}}_{2,\mathrm{OPT}}$ comme :

$$\tilde{\tilde{\boldsymbol{B}}}_{2,\mathrm{OPT}} = [\hat{V}(\hat{\boldsymbol{X}}_2)]^{-1}\,\hat{\mathrm{Cov}}\,(\hat{\tilde{Y}}_{\mathrm{HT}}, \hat{\boldsymbol{X}}_2). \tag{3.29}$$

Remarque : L'estimateur $\hat{Y}_{\mathrm{OPT}}$ (3.25) est exactement égal à $\tilde{\tilde{Y}}_{\mathrm{OPT}}$ (3.27). Ceci implique qu'il n'y a eu aucun gain vis-à-vis de l'estimation de Y, en substituant un meilleur estimateur de $\boldsymbol{X}_2$. Cependant, l'estimateur $\tilde{\tilde{\boldsymbol{B}}}_{2,\mathrm{OPT}}$ associé à $\tilde{\tilde{Y}}_{\mathrm{OPT}}$ ressemble beaucoup plus à un estimateur de régression auquel on s'attend, que $\hat{\boldsymbol{B}}_{2,\mathrm{OPT}}$ associé à $\hat{Y}_{\mathrm{OPT}}$.

L'estimateur GREG correspondant pour le cas où on sert de $\tilde{\boldsymbol{X}}_2$ au lieu de $\tilde{\boldsymbol{X}}_2$ est :

$$\tilde{\tilde{Y}}_{\mathrm{GREG}} = \hat{Y}_{\mathrm{HT}} + (\tilde{\boldsymbol{X}}_2 - \hat{\boldsymbol{X}}_2)'\,\tilde{\tilde{\boldsymbol{B}}}_{2,\mathrm{GREG}} \tag{3.30}$$

où

$$\tilde{\tilde{\boldsymbol{B}}}_{2,\mathrm{GREG}} = \left(\sum\nolimits_{s_2} w_{2k}\,\frac{\boldsymbol{x}_k^{(2)}\,\boldsymbol{x}_k'^{(2)}}{\sigma_{2k}^2}\right)^{-1} \sum\nolimits_{s_2} w_{2k}\,\frac{\boldsymbol{x}_k^{(2)}\,y_k^{(2)}}{\sigma_{2k}^2}.$$

En plus, si l'on connaît aussi $\boldsymbol{x}_{1k}^{(1)}$ pour $k \in U_1$ où $\boldsymbol{X}_1 = \sum_{U_1} \boldsymbol{x}_{1k}^{(1)}$, on peut considérer l'estimateur de régression

$$\tilde{\tilde{Y}}_{\mathrm{OPT}} = \hat{\tilde{Y}}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)'\,\tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} + (\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})'\,\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}}. \tag{3.31}$$

C'est-à-dire que $\tilde{\boldsymbol{X}}$ est obtenu en minimisant la combinaison linéaire $\boldsymbol{A}\hat{\boldsymbol{X}} + (\boldsymbol{I} - \boldsymbol{A})\,\hat{\tilde{\boldsymbol{X}}}$ et $V(\hat{\boldsymbol{X}}) \doteq \alpha\, V(\hat{\tilde{\boldsymbol{X}}})$. La différence entre $\tilde{\boldsymbol{X}}$ et $\tilde{\tilde{\boldsymbol{X}}}$ peut s'exprimer comme

$$\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}} = (\hat{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})/(1+\alpha). \tag{3.32}$$

Étant donné que s_1 et s_2 sont des échantillons indépendants, on peut aisément démontrer que :

$$\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}} = [\hat{V}(\hat{\tilde{\boldsymbol{X}}})]^{-1}\,\hat{\mathrm{Cov}}(\hat{\tilde{\boldsymbol{X}}}, \hat{\tilde{Y}}_{\mathrm{HT}}) \tag{3.33}$$

et que

$$\tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} = [\hat{V}(\hat{\tilde{\boldsymbol{X}}}_1)]^{-1}\,[\hat{\mathrm{Cov}}(\hat{\tilde{\boldsymbol{X}}}_1, \hat{\tilde{Y}}_{\mathrm{HT}})]. \tag{3.34}$$

Les composantes de $\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}}$ sont estimées par :

$$\hat{V}(\hat{\tilde{\boldsymbol{X}}}) = \sum\sum\nolimits_{s_2} \hat{c}_{2k\ell}\,\boldsymbol{x}_k^{(2)}\boldsymbol{x}_\ell'^{(2)} \tag{3.35}$$

et

$$\hat{\mathrm{Cov}}(\hat{\tilde{\boldsymbol{X}}},\ \hat{\tilde{Y}}_{\mathrm{HT}}) = \sum\sum\nolimits_{s_2} \hat{c}_{2k\ell}\,\boldsymbol{x}_k^{(2)}\,y_\ell^{(2)} \tag{3.36}$$

tandis que celles de $\tilde{\tilde{\boldsymbol{B}}}_{1,\mathrm{OPT}}$ sont estimées par :

$$\hat{V}(\hat{\tilde{\boldsymbol{X}}}_1) = \sum\sum\nolimits_{s_2} \hat{c}_{2k\ell}\,\boldsymbol{x}_{1k}^{(2)}\,\boldsymbol{x}_{1\ell}'^{(2)} \tag{3.37}$$

et

$$\hat{\mathrm{Cov}}(\hat{\tilde{\boldsymbol{X}}}_1, \hat{\tilde{Y}}_{\mathrm{HT}}) = \sum\sum\nolimits_{s_2} \hat{c}_{2k\ell}\,\boldsymbol{x}_{1k}^{(2)}\,y_\ell^{(2)} \tag{3.38}$$

où

$$\hat{c}_{2k\ell} = \frac{\pi_{2k\ell} - \pi_{2k}\,\pi_{2\ell}}{(\pi_{2k\ell})\,(\pi_{2k}\,\pi_{2\ell})}.$$

Nous pouvons donc nous servir de l'approximation (3.20) afin d'estimer les termes (3.35) – (3.38). La forme correspondante du GREG qui évite le calcul des probabilités conjointes de sélection est donnée par :

$$\tilde{\tilde{Y}}_{\mathrm{GREG}} = \hat{\tilde{Y}}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)'\,\tilde{\boldsymbol{B}}_{1,\mathrm{GREG}} + (\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})'\,\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{GREG}} \tag{3.39}$$

où $\boldsymbol{X}_1 = \sum_{U_1} \boldsymbol{x}_{1k}^{(1)}$, $\hat{\boldsymbol{X}}_1 = \sum_{s_1} w_{1k}\,\boldsymbol{x}_{1k}^{(1)}$, $\hat{\boldsymbol{X}} = \sum_{s_1} w_{1k}\,\boldsymbol{x}_k^{(1)}$ et $\hat{\tilde{\boldsymbol{X}}} = \sum_{s_2} w_{2k}\,\boldsymbol{x}_k^{(2)}$.

Les estimateurs de régression du type GREG dans l'équation (3.39) sont estimés par

$$\tilde{\boldsymbol{B}}_{1,\mathrm{GREG}} = \left(\sum\nolimits_{s_2} w_{2k}\,\frac{\boldsymbol{x}_{1k}^{(2)}\,\boldsymbol{x}_{1k}'^{(2)}}{\sigma_{1k}^2}\right)^{-1} \sum\nolimits_{s_2} w_{2k}\,\frac{\boldsymbol{x}_{1k}^{(2)}\,y_k^{(2)}}{\sigma_{1k}^2} \tag{3.40}$$

et

$$\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{GREG}} = \left(\sum\nolimits_{s_2} w_{2k}\,\frac{\boldsymbol{x}_k^{(2)}\,\boldsymbol{x}_k'^{(2)}}{\sigma_{2k}^2}\right)^{-1} \sum\nolimits_{s_2} w_{2k}\,\frac{\boldsymbol{x}_k^{(2)}\,y_k^{(2)}}{\sigma_{2k}^2}. \tag{3.41}$$

4. Estimation de la variance de l'estimateur optimale de regression

4.1 L'échantillonnage double imbriqué

Rappelons que l'estimateur de régression optimal de Y est

$$\hat{\hat{Y}}_{\text{OPT}} = \hat{\hat{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \, \hat{\boldsymbol{B}}_{1,\text{OPT}} + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})' \, \hat{\hat{\boldsymbol{B}}}_{\text{OPT}}. \quad (4.1)$$

Afin d'obtenir la variance estimée de (4.1), nous réexprimons les composantes définissant $\hat{\hat{\boldsymbol{B}}}_{\text{OPT}}$ ainsi que $\hat{\boldsymbol{B}}_{1,\text{OPT}}$ en regroupant les termes associés à la variable y, afin d'exprimer la double somme comme une simple somme. Cette manipulation a été décrite par Montanari (1998) pour un plan de sondage à une phase, ayant des probabilités de sélection arbitraires. En suivant Montanari (1998) et en adaptant l'algèbre du cas d'une phase à celui deux phases on obtient :

$$\begin{aligned}\hat{\hat{\boldsymbol{B}}}_{\text{OPT}} &= \left[\sum\sum\nolimits_{s_2} \hat{c}_{2k\ell|s_1} \boldsymbol{x}_k \boldsymbol{x}_\ell'\right]^{-1} \left[\sum\sum\nolimits_{s_2} \hat{c}_{2k\ell|s_1} \boldsymbol{x}_k y_\ell\right] \\ &= \left[\sum\sum\nolimits_{s_2} \hat{c}_{2k\ell|s_1} \boldsymbol{x}_k \boldsymbol{x}_\ell'\right]^{-1} \left[\sum\nolimits_{s_2} \frac{a_{2k}}{\pi_k^*} y_k\right] \end{aligned} \quad (4.2)$$

où

$$a_{2k} = \frac{1-\pi_{2k|s_1}}{\pi_k^*} \boldsymbol{x}_k + \sum_{\substack{\ell \neq k \\ \ell \in s_2}} \frac{(\pi_{2k\ell|s_1} - \pi_{2k|s_1}\pi_{2\ell|s_1})}{\pi_{2k\ell|s_1}\pi_\ell^*} \boldsymbol{x}_\ell.$$

On approxime $\hat{\boldsymbol{B}}_{1,\text{OPT}}$ donné en (3.15) par $[\hat{V}(\hat{\boldsymbol{X}}_1)]^{-1}\,[\hat{\text{Cov}}(\hat{\boldsymbol{X}}_1,\, \hat{Y}_{\text{HT}})]$, c'est-à-dire,

$$\begin{aligned}\hat{\boldsymbol{B}}_{1,\text{OPT}} &\doteq [\hat{V}(\hat{\boldsymbol{X}}_1)]^{-1}\left[\hat{\text{Cov}}(\hat{\boldsymbol{X}}_1, \hat{Y}_{\text{HT}})\right] \\ &= \left[\sum\sum\nolimits_{s_1} \hat{c}_{1k\ell} \boldsymbol{x}_{1k} \boldsymbol{x}_{1\ell}'\right]^{-1} \left[\sum\nolimits_{s_1} \frac{a_{1k}}{\pi_{1k}} y_k\right] \end{aligned} \quad (4.3)$$

où

$$a_{1k} = \frac{1-\pi_{1k}}{\pi_{1k}} \boldsymbol{x}_{1k} + \sum_{\substack{\ell \neq k \\ \ell \in s_1}} \frac{(\pi_{1k\ell} - \pi_{1k}\,\pi_{1\ell})}{\pi_{1\ell}\pi_{1k\ell}} \boldsymbol{x}_{1\ell}.$$

En substituant (4.2) et (4.3) dans (4.1), et en soustrayant le total de la population Y, on obtient :

$$\begin{aligned}\hat{\hat{Y}}_{\text{OPT}} - Y &\doteq \left(\sum\nolimits_{s_1} g_{1k} \frac{y_k}{\pi_{1k}} - \sum\nolimits_U y_k\right) \\ &\quad + \left(\sum\nolimits_{s_2} g_{2k} \frac{y_k}{\pi_k^*} - \sum\nolimits_{s_1} \frac{y_k}{\pi_{1k}}\right) \end{aligned} \quad (4.4)$$

où

$$g_{1k} = 1 + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \, (\hat{V}(\hat{\boldsymbol{X}}_1))^{-1} \, a_{1k} \text{ pour } k \in s_1 \quad (4.5)$$

et

$$g_{2k} = 1 + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})' \, (\hat{V}(\hat{\hat{\boldsymbol{X}}}))^{-1} \, a_{2k} \text{ pour } k \in s_2. \quad (4.6)$$

Résultat 2 : La variance estimée de $\hat{\hat{Y}}_{\text{OPT}}$ en nous servant de l'équation (4.1) est :

$$\begin{aligned}\hat{V}(\hat{\hat{Y}}_{\text{OPT}}) &= \sum\sum\nolimits_{s_2} c_{1k\ell}^* \, g_{1k} \, g_{1\ell} \, e_{1k} \, e_{1\ell} \\ &\quad + \sum\sum\nolimits_{s_2} c_{2k\ell}^* \, g_{2k} \, g_{2\ell} \, e_{2k} \, e_{2\ell} \end{aligned} \quad (4.7)$$

où

$$c_{1k\ell}^* = \frac{(\pi_{1k\ell} - \pi_{1k}\,\pi_{1\ell})}{\pi_{k\ell}^*\,\pi_{1k}\,\pi_{1\ell}};$$

$$c_{2k\ell}^* = \frac{(\pi_{2k\ell|s_1} - \pi_{2k|s_1}\,\pi_{2\ell|s_1})}{\pi_{2k\ell|s_1}\,\pi_k^*\,\pi_\ell^*};$$

$$e_{1k} = y_k - \boldsymbol{x}_{1k}' \, \hat{\boldsymbol{B}}_{1,\text{OPT}};$$

et

$$e_{2k} = y_k - \boldsymbol{x}_k' \, \hat{\hat{\boldsymbol{B}}}_{\text{OPT}}.$$

4.2 L'échantillonnage double non-imbriqué

Nous obtenons la variance estimée de $\tilde{\hat{Y}}_{\text{OPT}}$ en nous servant de l'approximation suivante.

$$\begin{aligned}\tilde{\hat{Y}}_{\text{OPT}} &= \hat{\hat{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)' \, \tilde{\boldsymbol{B}}_{1,\text{OPT}} + (\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})' \, \tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}} \\ &= \ddot{Y}_{\text{OPT}} + O_p(n_1^{-1/2}) \end{aligned} \quad (4.8)$$

où

$$\ddot{Y}_{\text{OPT}} = \hat{\hat{Y}}_{\text{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)' \, \boldsymbol{B}_{1,\text{OPT}} + (\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})' \, \boldsymbol{B}_{\text{OPT}}. \quad (4.9)$$

La décomposition de $\ddot{Y}_{\text{OPT}}$ en termes plus élémentaires donne :

$$\begin{aligned}\ddot{Y}_{\text{OPT}} &= \hat{\hat{Y}}_{\text{HT}} + \left(\boldsymbol{X}_1 - \frac{\hat{\boldsymbol{X}}_1 + \alpha\,\hat{\hat{\boldsymbol{X}}}_1}{1+\alpha}\right)' \boldsymbol{B}_{1,\text{OPT}} \\ &\quad + \frac{(\tilde{\boldsymbol{X}} - \hat{\tilde{\boldsymbol{X}}})'}{1+\alpha} \boldsymbol{B}_{\text{OPT}} \\ &= \left(\hat{\hat{Y}}_{\text{HT}} - \frac{1}{1+\alpha}(\alpha\,\hat{\hat{\boldsymbol{X}}}_1' \, \boldsymbol{B}_{1,\text{OPT}} + \hat{\tilde{\boldsymbol{X}}}' \, \boldsymbol{B}_{1,\text{OPT}})\right) \\ &\quad + \left(\boldsymbol{X}_1' \, \boldsymbol{B}_{1,\text{OPT}} - \frac{1}{1+\alpha}(\hat{\boldsymbol{X}}_1' \boldsymbol{B}_{1,\text{OPT}} - \hat{\tilde{\boldsymbol{X}}}' \boldsymbol{B}_{\text{OPT}})\right). \end{aligned} \quad (4.10)$$

La variance de $\ddot{Y}_{\text{OPT}}$ est :

$$V(\ddot{Y}_{\text{OPT}}) = V\left(\hat{\bar{Y}}_{\text{HT}} - \frac{1}{1+\alpha}(\alpha\,\hat{\boldsymbol{X}}_1'\,\boldsymbol{B}_{1,\text{OPT}} + \hat{\boldsymbol{X}}'\,\boldsymbol{B}_{\text{OPT}})\right)$$

$$+\frac{1}{(1+\alpha)^2}\begin{bmatrix}\alpha\,\boldsymbol{B}_{1,\text{OPT}}'\,V(\hat{\boldsymbol{X}}_1)\,\boldsymbol{B}_{1,\text{OPT}}\\ +\boldsymbol{B}_{\text{OPT}}'\,V(\hat{\boldsymbol{X}})\,\boldsymbol{B}_{\text{OPT}}\\ +2\alpha\,(\boldsymbol{B}_{\text{OPT}}'\,V(\hat{\boldsymbol{X}})\,\tilde{\boldsymbol{B}}_{1,\text{OPT}}'\\ +\text{Cov}(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}'))\,\boldsymbol{B}_{\text{OPT}}\end{bmatrix}. \qquad (4.11)$$

Résultat 3 : La variance estimées de $\tilde{\bar{Y}}_{\text{OPT}}$, $\hat{V}(\tilde{\bar{Y}}_{\text{OPT}})$, défini par l'équation (4.8) est approximativement égale à :

$$\hat{V}\left(\hat{\bar{Y}}_{\text{HT}} - \frac{1}{1+\alpha}\,(\alpha\,\hat{\boldsymbol{X}}_1'\,\tilde{\boldsymbol{B}}_{1,\text{OPT}} + \hat{\boldsymbol{X}}'\,\tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}})\right)$$

$$+\frac{1}{(1+\alpha)^2}\begin{bmatrix}\alpha\,\tilde{\boldsymbol{B}}_{1,\text{OPT}}'\,\hat{V}(\hat{\boldsymbol{X}}_1)\,\tilde{\boldsymbol{B}}_{1,\text{OPT}}\\ +\,\tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}}'\,\hat{V}(\hat{\boldsymbol{X}})\,\tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}}\\ +\,2\alpha\,(\tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}}'\,\hat{V}(\hat{\boldsymbol{X}})\,\tilde{\boldsymbol{B}}_{1,\text{OPT}}\\ +\,\hat{\text{Cov}}(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}')\,\tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}})\end{bmatrix}. \qquad (4.12)$$

Le calcul du premier terme de (4.12) est basé sur les résidus $y_k - (\alpha\,\boldsymbol{x}_{1k}'\,\tilde{\boldsymbol{B}}_{1,\text{OPT}} + \boldsymbol{x}_k'\,\tilde{\tilde{\boldsymbol{B}}}_{\text{OPT}})/(1+\alpha)$. Les autres termes de (4.12) sont assez simples à calculer, se basant principalement sur les variances estimées de $\hat{\bar{X}}_1$ et de $\hat{\bar{X}}$, ainsi que sur leurs covariances. On peut se servir de l'approximation de la variance, telle qu'énoncé par Tillé (2001) pour ce cas, et aussi l'appliquer aux covariances.

5. Quelques exemples spécifiques

Trois exemples traditionnels pour l'échantillonnage double sont présentés pour les deux cas (imbriqué et non-imbriqué). En plus, nous décrivons brièvement deux grandes enquêtes entreprises menées à Statistique Canada qui se servent de ces méthodes.

5.1 Échantillons imbriqués

Exemple 1 : Supposons qu'un échantillon s_1 aléatoire simple de taille n_1 est sélectionné de la population U de taille N. L'échantillon est ensuite stratifié en L strates s_{1h} de taille respective n_{1h}. Des échantillons aléatoires s_{2h} de taille n_{2h} sont ensuite sélectionnés sans remise dans chaque strate s_{1h}. L'estimateur du total est $\hat{Y}_{\text{EXP}} = N\sum_{h=1}^{L} p_{1h}\,\bar{y}_{2h} = N\,\bar{y}_{2,st}$, où $p_{1h} = n_{1h}/n_1$. En se servant de (4.7), on peut démontrer que la variance estimée de $\hat{\bar{Y}}_{\text{EXP}}$, $\hat{V}(\hat{\bar{Y}}_{\text{EXP}})$, se décompose en la somme de, $\hat{V}_1(\hat{\bar{Y}}_{\text{EXP}})$ et $\hat{V}_2(\hat{\bar{Y}}_{\text{EXP}})$ qui correspondent aux phases 1 et 2 du plan de sondage. Ainsi :

$$\hat{V}(\hat{\bar{Y}}_{\text{EXP}}) = \hat{V}_1(\hat{\bar{Y}}_{\text{EXP}}) + \hat{V}_2(\hat{\bar{Y}}_{\text{EXP}})$$

où

$$\hat{V}_1(\hat{\bar{Y}}_{\text{EXP}}) = N^2\,\frac{(1-f_1)}{n_1}\sum_{h=1}^{L} p_{1h}\begin{bmatrix}(1-a_h)\;\hat{S}_{2yh}^2\\ +\dfrac{n_1}{n_1-1}(\bar{y}_{2h} - \bar{y}_{2,st})^2\end{bmatrix};$$

$$\hat{V}_2(\hat{\bar{Y}}_{\text{EXP}}) = N^2\sum_{h=1}^{L}\frac{(1-f_{2h})}{n_{2h}}\,p_{1h}^2\,\hat{S}_{2yh}^2;$$

et

$$a_h = \frac{(n_1 - n_{1h})}{n_{2h}(n_1-1)};\; f_1 = \frac{n_1}{N};\; f_{2h} = \frac{n_{2h}}{n_{1h}};$$

$$\hat{S}_{2yh}^2 = \frac{1}{n_{2h}-1}\sum_{s_{2h}}(y_k - \bar{y}_{2h})^2;$$

$$\bar{y}_{2h} = \frac{1}{n_{2h}}\sum_{s_{2h}} y_k$$

et

$$\bar{y}_{2,\,st} = \sum_{h=1}^{L} p_{1h}\,\bar{y}_{2h}.$$

Exemple 2 : Supposons que pour le plan de sondage de l'exemple 1, nous avons aussi des données auxiliaires, $\boldsymbol{x}_k$, disponibles à la première phase s_1. Si nous supposons que les pentes ($\boldsymbol{\beta}_h$) diffèrent entre les strates, nous postulons le modèle $y_k = \boldsymbol{x}_k'\,\boldsymbol{\beta}_h + \varepsilon_k$, où $E(\varepsilon_k) = 0$, $E(\varepsilon_k^2) = \sigma_k^2$, $k \in s_{1h}$, $h = 1, \ldots, L$, et $E(\varepsilon_k\varepsilon_\ell) = 0$ pour $k \neq \ell$, où k, $\ell \in s_{1h}$ $h = 1, \ldots, L$. Ce modèle nous amène à l'estimateur par régression séparé, c'est-à-dire,

$$\hat{\boldsymbol{Y}}_{\text{SEP, REG}} = \sum_{h=1}^{L}\frac{N}{n_1}\,\frac{n_{1h}}{n_{2h}}\sum_{s_2} g_{2k}\,y_k$$

où

$$g_{2k} = 1 + \left(\sum_{s_{1h}}\boldsymbol{x}_k' - \sum_{s_{2h}}\frac{n_{1h}}{n_{2h}}\,\boldsymbol{x}_k'\right)$$

$$\left(\sum_{s_{2h}}\frac{n_{1h}}{n_{2h}}\,\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\sigma_k^2}\right)^{-1}\frac{\boldsymbol{x}_k}{\sigma_k^2}$$

si $k \in s_{2h}$. Pour chaque strate h, les pentes $\boldsymbol{\beta}_h$ sont estimées par

$$\hat{\tilde{\boldsymbol{B}}}_{2h} = \left(\sum_{s_{2h}}\frac{n_{1h}}{n_{2h}}\,\frac{\boldsymbol{x}_k\boldsymbol{x}_k'}{\sigma_k^2}\right)^{-1}\left(\sum_{s_{2h}}\frac{n_{1h}}{n_{2h}}\,\frac{\boldsymbol{x}_k y_k}{\sigma_k^2}\right).$$

La variance de $\hat{Y}_{\text{SEP, REG}}$ est estimée comme étant la somme des composantes de chaque phase, c'est-à-dire $\hat{V}_1(\hat{Y}_{\text{EXP}})$ et $\hat{V}_2(\hat{Y}_{\text{SEP, REG}})$. La variance $\hat{V}_2(\hat{Y}_{\text{SEP, REG}})$ est obtenue en substituant la variable y_k par $e_k = g_k(y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}_h)$ dans l'expression $\hat{V}_2(\hat{\bar{Y}}_{\text{EXP}})$. Ceci implique que la variance estimée de $\hat{Y}_{\text{SEP, REG}}$ est :

$$\hat{V}(\hat{Y}_{\text{SEP, REG}}) = \frac{N^2(1-f_1)}{n_1}\sum_{h=1}^{L} p_{1h}\left[\begin{array}{l}(1-a_h)\,\hat{S}^2_{2yh} \\ \\ +\dfrac{n_1}{n_1-1}(\bar{y}_{2h}-\bar{y}_{2,st})^2\end{array}\right]$$

$$+\sum_{h=1}^{L}\frac{N^2(1-f_{2h})}{n_{2h}}\,p_{1h}^2\,\hat{S}^2_{2eh}$$

où

$$\hat{S}^2_{2eh} = \sum_{s_{2h}}\frac{(e_k-\bar{e}_h)^2}{n_{2h}-1}$$

et

$$\hat{S}^2_{2yh} = \frac{1}{n_{2h}-1}\sum_{s_{2h}}(y_k-\bar{y}_{2h})^2.$$

5.2 Échantillons non-imbriqués

Ces deux exemples sont extraits de Des Raj (1968, pages 142–149). Nous nous en servirons afin d'illustrer les résultats des sections 3 et 4. On considère deux plans de sondage différents.

Le premier plan de sondage, on suppose que : (i) le premier échantillon s_1 de taille n_1 est sélectionné avec un plan de sondage aléatoire simple et sans remise de la population U; et (ii) le deuxième échantillon s_2 de taille n_2 est sélectionné soit en se servant des mesures de tailles x_i observées à partir du premier échantillon s_1 (cas imbriqué), ou en le sélectionnant indépendamment (cas non-imbriqué) du premier échantillon s_1 de manière proportionnel à la taille x_i (connu pour toutes les unités de la population).

L'estimateur qui en découle est

$$\hat{\hat{Y}}_{\text{EPTAR}} = \frac{N}{n_1}\frac{\sum_{s_1} x_i}{n_2}\sum_{s_2}\frac{y_i}{x_i}.$$

Pour le deuxième plan de sondage on suppose que les deux échantillons s_1 et s_2 ont été sélectionnés avec un plan de sondage aléatoire simple et sans remise. Encore une fois, on considère les cas imbriqués et non-imbriqués. On suppose que l'on observe l'observation auxiliaire x_i pour toute unité sélectionnée dans le premier échantillon s_1. L'estimateur est $\hat{\bar{Y}}_{\text{RAT}} = (N/n_1\sum_{s_1} x_i)(\sum_{s_2} y_i/\sum_{s_2} x_i) = \hat{\bar{X}}\hat{\bar{R}}$. Le tableau 3 résume ces deux plans de sondage, et présente les deux estimateurs ainsi que leurs variances estimées pour les cas imbriqués et non-imbriqués.

Les termes non définis dans le tableau 3 sont $p_{1i} = x_i/\sum_{s_1} x_i$; $p_i = x_i/\sum_U x_i$; $V(\hat{Y}_p) = 1/n_1\sum_U p_i(y_i/p_i - Y)^2$; $S^2_{y-Rx} = (N-1)^{-1}\sum_U (y_i - Rx_i)^2$; $f_2 = n_2/N$ $f_1 = n_1/N$, et $R = Y/X$.

On remarque à partir du tableau 3 est que les variances ne sont pas très différentes entre les cas imbriqués et non-imbriqués. Pour $\hat{\bar{Y}}_{\text{EPTAR}}$, la variance sera plus petite pour le cas imbriqué, si le coefficient de variation (CV) de la variable y est plus petit que celui de la variable x. Pour $\hat{\bar{Y}}_{\text{RAT}}$, la variance sera plus petite pour le cas imbriqué si $\rho\,\text{CV}(\bar{y}) < \text{CV}(\bar{x})$ où ρ est la corrélation entre y et x.

Tableau 3
Deux plans de sondage avec échantillonnage imbriqué et non-imbriqué

	Premier plan de sondage	**Deuxième plan de sondage**
Sondage	$N \to n_1$ (EASSR)	$N \to n_1$ (EASSR)
Plan de sondage	$n_1 \to n_2$ (EPTAR)	$n_1 \to n_2$ (EASSR)
Estimateur	$\hat{\hat{Y}}_{\text{EPTAR}} = \dfrac{N}{n_1}\sum_{s_2}\dfrac{y_i}{n_2\,p_{1i}}$	$\hat{\hat{Y}}_{\text{RAT}} = \sum_{s_1}\dfrac{\sum_{s_2} y_i}{\sum_{s_2} x_i} = \hat{\bar{X}}\,\hat{\bar{R}}$
Variance Imbriqué	$\dfrac{N^2(1-f_1)}{n_1}S_y^2 + \dfrac{V(\hat{Y}_p)}{n_2}$	$\dfrac{N^2(1-f_1)}{n_1}(2RS_{xy} - R^2S_x^2) + N^2\dfrac{(1-f_2)}{n_2}S^2_{y-Rx}$
Non-imbriqué	$\dfrac{N^2(1-f_1)}{n_1}R^2\,S_x^2 + \dfrac{V(\hat{Y}_p)}{n_2}\left[1 + \dfrac{1}{n_1}(1-f_1)\dfrac{S_x^2}{\bar{X}^2}\right]$	$\dfrac{N^2(1-f_1)}{n_1}R^2S_x^2 + N^2\dfrac{(1-f_2)}{n_2}S^2_{y-Rx}$

5.3 Deux enquêtes à Statistique Canada

Plusieurs enquêtes à Statistique Canada se servent de l'échantillonnage double. On illustre les idées de cet article par deux enquêtes menées auprès des entreprises. Ces enquêtes sont l'Enquête sur les marchandises du détail (EMD), et l'Enquête sur l'Emploi, la rémunération, et les heures (EERH). L'Enquête sur les marchandises de détail (EMD), se sert de l'échantillonnage double imbriqué, tandis que l'Enquête sur l'emploi, la rémunération et les heures (EERH) se sert de l'échantillonnage double non-imbriqué.

L'enquête trimestrielle sur les marchandises de détail : L'objectif de l'Enquête trimestrielle sur les marchandises de détail (ETMD) est d'obtenir de l'information détaillée sur les ventes des marchandises de détail tous les trois mois. L'EMD est un sous-échantillon de l'Enquête mensuelle du commerce de détail (EMCD), une enquête mensuelle. L'EMCD mesure principalement les ventes par groupe de commerce (regroupement de codes de classification type industrielle (CTI) à trois ou quatre chiffres de 1980), par province et pour certaines régions métropolitaines de recensement (RMR). La population cible se compose des compagnies statistiques ayant des emplace-ments statistiques identifiés sur le Registre des entreprises qui sont actifs dans le commerce de détail. Environ 16 000 compagnies sont interviewées chaque mois. La population est stratifiée par province, territoire, certaines RMR et par groupe de commerce.

L'EMCD est stratifiée en H strates, basées sur la taille (2 – 3 groupes), la géographie (10 provinces, 2 territoires) et le type d'industrie (16 groupes principaux). Cet échantillon est re-stratifié indépendamment pour l'ETDC. La stratification de l'ETDC diffère celle de l'EMDC au niveau géographie, taille et industrie. Un sous échantillon est sélectionné en se servant de la « nouvelle » stratification de l'échantillon de l'EMCD. L'estimation de l'ETDC est basée sur un estimateur à double quotient qui se sert des données auxiliaires (les ventes) de l'EMCD. L'unité d'échantillonnage de deuxième phase (ETMD) demeure la compagnie statistique. L'échantillon de première phase est re-stratifié par groupe de commerce, par province et taille selon l'information la plus récente provenant de l'EMCD. Pour les besoins de la stratification, chaque compagnie se voit attribuer une province et un groupe de commerce dominant c'est-à-dire ceux pour lesquels elle génère le plus de ventes. L'estimateur par rapport à deux phases est utilisé par l'EMCD. Binder, Babyak, Brodeur, Hidiroglou, et Jocelyn (2000) ont dérivé l'estimateur de variance qui tenait compte du plan de sondage ainsi que de la méthode d'estimation. Ils ont réduit les estimateurs de variance du total à de simples sommes des termes (résidus) appropriés pour l'estimateur par ratio.

Les résultats de Binder et coll. (2000) peuvent être adaptés pour incorporer l'estimateur de régression optimal à chaque phase. On suppose que l'on connaît de l'information auxiliaire $(\boldsymbol{x}_{1k})$ au niveau de la population U, soit pour chaque unité $k \in U$ ou pour le total $\boldsymbol{X}_{1k} = \sum_U \boldsymbol{x}_{1k}$. Le plan de sondage de l'ETMC peut être formellement décrit comme suit. La population est stratifiée en H strates U_h; $h = 1, ..., H$, et des échantillons aléatoires simples s_{1h} et sans remise, de taille n_{1h}, sont sélectionnés dans chaque strate U_h. La variable $\boldsymbol{x}_k$ est observée pour chaque unité appartenant à s_1. L'échantillon de première phase qui en résulte, $s_1 = \boldsymbol{U}_{h=1}^{H} s_{1h}$, est ensuite stratifié en strates s_{1g}, $g = 1, ..., G$. La stratification de s_1 est indépendante de la stratification de l'univers U. Un échantillon aléatoire simple s_{2g} de taille n_{2g} est ensuite sélectionné de chaque strate s_{1g}, $g = 1, ..., G$. On observe $(y_k, \boldsymbol{x}_k')$, où $\boldsymbol{x}_k = (\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}')'$ pour chaque unité appartenant à l'échantillon $s_2 = U_{g=1}^{G} s_{2g}$. On suppose que les modèles $y_k = \boldsymbol{x}_{1k}' \boldsymbol{\beta}_1 + \varepsilon_{1k}$ et $y_k = \boldsymbol{x}_k' \boldsymbol{\beta} + \varepsilon_{2k}$ tiennent pour s_1 et s_2 respectivement. Pour chacun de ces modèles $\varepsilon_{1k} \sim (0, \sigma_1^2 z_{1k})$ et $\varepsilon_{2k} \sim (0, \sigma_2^2 z_{2k})$ où z_{1k} et z_{1k} sont des facteurs positifs connus. Si $z_{1k} \neq 1$ ou $z_{1k} \neq 1$ pour tous les unités $k \in U$, on peut standardiser les données en les divisant soit par $\sqrt{z_{1k}}$ ou $\sqrt{z_{2k}}$. L'estimateur de régression optimal pour le total Y qui en résulte est :

$$\tilde{\hat{Y}}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + (\boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1)' \, \tilde{\boldsymbol{B}}_{1,\mathrm{OPT}} + (\hat{\boldsymbol{X}} - \tilde{\boldsymbol{X}})' \, \tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}}$$

où les composantes de $\tilde{\hat{Y}}_{\mathrm{OPT}}$ ont été définies dans la section 3.1. La forme simplifiée (sans doubles sommes) de la variance de $\tilde{\hat{Y}}_{\mathrm{OPT}}$ est :

$$\hat{V}(\tilde{\hat{Y}}_{\mathrm{OPT}}) = \sum_{h=1}^{H} N_h^2 \, (1 - f_{1h}) \, \frac{\hat{S}_{1h}^2}{n_{1h}} + \sum_{g=1}^{G} n_{1g}^2 \, (1 - f_{2g}) \, \frac{\hat{S}_{2g}^2}{n_{2g}} + \sum_{h=1}^{H} \sum_{g=1}^{G} \frac{N_h^2 \, (1 - f_{1h}) \, n_{2g}^2 \, (1 - f_{2g})}{n_{1h}^2 \, (n_{1h} - 1)} \, \frac{\hat{S}_{2hg}^2}{n_{2h}}$$

où les variances sont définies par

$$\hat{S}_{1h}^2 = \frac{1}{n_{1h} - 1} \left\{ \sum_{g=1}^{G} \sum_{k=1}^{n_{2gh}} \frac{n_{1g}}{n_{2g}} \tilde{e}_{1k}^2 - \frac{1}{n_{1h}} \left(\sum_{g=1}^{G} \sum_{k=1}^{n_{2gh}} \frac{n_{1g}}{n_{2g}} \tilde{e}_{1k} \right)^2 \right\};$$

$$\hat{S}_{2hg}^2 = \frac{1}{n_{2hg} - 1} \sum_{k=1}^{n_{2hg}} (\tilde{e}_{1k} - \bar{\tilde{e}}_{1(hg)})^2$$

et

$$\hat{S}_{2g} = \frac{1}{n_{2g} - 1} \sum_{k=1}^{n_{2g}} (\tilde{e}_{2k} - \bar{\tilde{e}}_{2h})^2.$$

Les moyennes qui leur correspondent sont

$$\bar{\tilde{e}}_{1(hg)} = \frac{1}{n_{2hg}} \sum_{k=1}^{n_{2hg}} \tilde{e}_{1k},\ \bar{\tilde{e}}_{1(hg)} = \frac{1}{n_{2hg}1} \sum_{k=1}^{n_{2hg}} \tilde{e}_{1k}$$

et

$$\bar{\tilde{e}}_{2h} = \frac{1}{n_{2g}} \sum_{k=1}^{n_{2g}} \tilde{e}_{2k}.$$

Ici, n_{2hg} est le nombre d'unités sélectionnées dans l'échantillon s_2 qui appartiennent à l'intersection des strates U_h et s_{1g}. Les résidus qui rentrent en jeu dans chacune de ces expressions sont $\tilde{e}_{1k} = g_{1k}\,(y_k - \boldsymbol{x}'_{1k}\,\tilde{\boldsymbol{B}}_{1,\mathrm{OPT}})$ et $\tilde{e}_{2k} = g_{2k}\,(y_k - \boldsymbol{x}'_k\,\tilde{\tilde{\boldsymbol{B}}}_{\mathrm{OPT}})$. Les facteurs d'ajustement g_{1k} et g_{2k} sont tels que définis dans la section 4.1.

L'Enquête sur l'emploi, la rémunération et les heures : L'objectif de cette enquête est d'obtenir des estimations sur le nombre d'employés rémunérés, la rémunération hebdomadaire moyenne et autres variables connexes, selon diverses combinaisons de branche d'activité et province. Cette enquête a été remaniée afin de permettre l'utilisation de dossiers administratifs sur toutes les entreprises comprises dans l'univers de l'enquête; ainsi, l'enquête produit des estimations basées à la fois sur les dossiers administratifs (échantillon ADMIN) et les données recueillies sur le terrain par l'Enquête sur la rémunération auprès des entreprises (ERE). Plus de détails sur l'ERE sont donnés dans Rancourt et Hidiroglou (1998).

L'échantillon ADMIN s_1 de quelques 200 000 unités est sélectionné de l'univers U_1 des (comptes de retenue sur la paye) pour obtenir les données administratives. Le plan de sondage pour cet échantillon est Bernoulli stratifié (par région), et le taux d'échantillonnage dans les strates varie de 10 % à 100 %. La taille de l'échantillon représente environ 20 % du nombre total de comptes de retenue sur la paye. Seulement deux variables $(\mathbf{x}_{1k}^{(1)})$ sont disponibles sur la source administrative : le nombre d'employés rémunérés et la rémunération mensuelle brute.

L'échantillon de l'ERE, s_2, est constitué d'environ 10 000 établissements sélectionnés à partir du Registre des entreprises U_2. L'ERE recueille les mêmes deux variables que la source administrative soient le nombre d'employés rémunérés et la rémunération mensuelle brute, $(\mathbf{x}_{1k}^{(2)})$, plusieurs autres variables $(\mathbf{x}_{2k}^{(2)})$ d'intérêt définies par type d'employés (employés payés à l'heure, salariés, propriétaires actifs, autres employés), ainsi que les variables d'intérêt telles le nombre d'heures payées et les gains hebdomadaires, $(y_k^{(2)})$. On trouve plus d'information sur l'ERE dans Rancourt et Hidiroglou (1998).

L'ERE est stratifiée par type d'industrie, par région géographique et par taille (variant de deux à trois groupes définis selon le nombre d'employés). Ces strates ont été conçues en tenant compte du fait que nous allions bâtir des modèles de régression entre $y_k^{(2)}$ et $\boldsymbol{x}_k^{(2)}$ afin de prédire $\hat{\boldsymbol{y}}_k$ pour les données administratives. L'estimation s'effectue en deux étapes. Premièrement, on cale les poids de sondage $w_k^{(1)}$ des données administratives à des comptes connus, N_i, pour des régions U_{1i}, $i = 1, \ldots, I$, de la population des unités administratives. Pour une unité k appartenant à U_{1i}, le poids modifié du premier échantillon est $\tilde{w}_k^{(1)} = w_k^{(1)}\,g_{1i}$, où $g_{1i} = N_i / \sum_{s_{1i}} w_k^{(1)}$ et $s_{1i} = s_1 \cap U_{1i}$. Deuxièmement, on régresse $y_k^{(2)}$ sur $\boldsymbol{x}_k^{(2)}$ pour des groupes de calage $s_{2,j}$, $j = 1, \ldots, J$ de l'échantillon d'établissements. Ces groupes, qui reposent sur une classification industrielle, régionale et parfois par taille, sont formés a priori afin d'obtenir les meilleures relations de régression possibles. Ceci nous donne pour chaque groupe $s_{2,j}$ un vecteur de régression $\hat{\boldsymbol{B}}_j$,

$$\hat{\boldsymbol{B}}_j = \left(\sum_{s_{2,j}} w_k^{(2)} \boldsymbol{x}_k^{(2)} \boldsymbol{x}_k'^{(2)} / \hat{\sigma}_k^2 \right)^{-1} \sum_{s_{2,j}} w_k^{(2)} \boldsymbol{x}_k^{(2)} y_k^{(2)} / \hat{\sigma}_k^2 ;$$
$$j = 1, \ldots, J$$

où $w_k^{(2)}$ est le poids de sondage pour chaque établissement sélectionné, et $\hat{\sigma}_k^2$ sont des facteurs positifs connus qui contrôlent l'impact des données aberrantes ou le type d'estimateur. Donc, si $\hat{\sigma}_k^2$ est proportionnel à une des composantes de $\boldsymbol{x}_k^{(2)}$, on obtient l'estimateur par quotient. L'estimateur d'une variable y est donc $\hat{Y} = \sum_{j=1}^{J} \sum_{s_{1,j}} \tilde{w}_k^{(1)}\,\boldsymbol{x}_k'^{(1)}\,\hat{\boldsymbol{B}}_j$, où $s_{1,j}$ correspond aux groupes de calage utilisé pour $s_{2,j}$. L'EERH se sert donc d'un plan de sondage à échantillonnage double non-imbriqué. Plus de détails sur le remaniment de l'EERH sont disponibles dans Hidiroglou (1995), et Hidiroglou, Latouche, Armstrong et Gossen (1995).

6. Conclusion

Une seule formulation décrit les cas imbriqués et non-imbriqués. Ces deux cas sont habituellement traités séparément dans la littérature. Cet article a unifié ces deux méthodes de sondage en se basant sur la régression optimale. Pour le cas imbriqué, on a observé que l'estimateur de régression obtenu par Hidiroglou et Särndal (1998) ressemble beaucoup à la forme correspondante de l'estimateur de régression optimale. Pour le cas non-imbriqué, on a adapté la méthode de Deville (1999) lorsqu'il y a des données auxiliaires au niveau de la population. Finalement, on a illustré cette théorie par des exemples pratiques.

Bibliographie

Berger, Y. (1998). Rate of Convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74, 149-168.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M.A. et Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 4, 751-764.

Breidt, J., et Fuller, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā*, 55, 297-309.

Brewer, K. (2000). Deriving and estimating an approximate variance for the Horvitz-Thompson estimator using only first order inclusion probabilities. Dans *Proceedings of the Second International Conferences on Establishment Surveys*. Buffalo, New York, 1417-1422.

Cassady, R.J., et Valliant, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.

Chaudhuri, A., et Roy, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika*, 41, 355-362.

Cochran, W.G. (1977). *Sampling Techniques*, 3ième Éd. New York : John Wiley & Sons, Inc.

Des Raj (1968). *Sampling Theory*. TMH Édition.

Deville, J.-C. (1999). Estimation de la Variance pour des Statistiques et des estimateurs complexes : Technique de résidus et de linéarisation. *Techniques d'enquête*, 25, 219-230.

Deville, J.-C. (1999). Calage simultané de plusieurs enquêtes. *Recueil du Symposium 1999 : Combiner des données de sources différentes*, 225-230.

Hartley, H.O., et Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Hidiroglou, M.A. (1995). Sampling and estimation for stage one of the canadian survey of employment, payrolls and hours survey. *Recueil de la Section des méthodes d'enquête*, Société Statistique du Canada, 123-128.

Hidiroglou, M.A., Latouche, M., Armstrong, B. et Gossen, M. (1995). Improving survey information using administrative records: The case of the canadian employment survey. *Proceedings of the 1995 Annual Research Conference*. U.S. Bureau of the Census, 171-197.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. Wiley series in *Probability and Statistics*.

Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.

Montanari, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 71-79.

Montanari, G.E. (2000). Conditioning on auxiliary variables means in finite population inference. Australian New Zealand *Journal of Statistics*, 42, 407-421.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

Rancourt, E., et Hidiroglou, M.A. (1998). L'utilisation de dossiers administratifs dans l'Enquête canadienne sur l'emploi, la rémunération, et les heures. *Recueil de la Section des Méthodes d'enquêtes*, Société Statistique du Canada, 39-48.

Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika*, 60, 125-133.

Rao, J.N.K. (1994). Estimation of totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-166.

Rósen, B. (2000). A User's guide to Pareto πps sampling. Dans *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, New York, 289-294.

Särndal, C.-E. (1996). Efficient estimators with simple variances in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

Särndal, C.-E., Swensson, B. et Wretman, Y. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288-289.

Tillé, Y. (2001). *Théorie des sondages : Échantillonnage et estimation en population finies*. Dumond.