VARIANCE ET ESTIMATION DE LA VARIANCE POUR DES STATISTIQUES COMPLEXES DANS LE CAS DE DEUX ÉCHANTILLONS

Jean-Claude DEVILLE, Camélia GOGA

CREST-ENSAI, Campus de Ker Lann

Introduction

Nous voulons estimer des statistiques complexes qui dépendent de plusieurs variables mesurées sur des échantillons différents. C'est par exemple l'évolution d'une statistique complexe quand les échantillons sont observés à différents instants. Il s'agit alors d'enquêtes répétées dans le temps et étudiées par Cochran (1977) dans le chapitre "Sampling on two occasions", Kish (1956) et Tam (1984) et plus récemment Särndal *et al* (1992), Caron et Ravalet (2000), Hidiroglu (2002). Quand l'information auxiliaire est utilisée pour améliorer les résultats, de nouveaux estimateurs composites ont été proposés par Bell (2001), Fuller et Rao (2001), Singh, Kennedy et Wu (2001). Ils donnent également leur application dans certaines enquêtes.

On propose ici une méthode en deux étapes : dans un premier temps on linéarise la statistique complexe par la fonction d'influence (Deville 1999) ; le paramètre d'intérêt est approximé par une somme pondérée avec des poids qui dépendent des échantillons impliqués. La deuxième étape est consacrée à la recherche des poids optimaux, c'est à dire tels que la variance de l'estimateur soit minimale. On donne ainsi une analyse générale du problème à "deux échantillons" en introduisant une nouvelle classe d'estimateurs composites. Notre étude peut s'appliquer à l'estimation d'un ratio en présence de non-réponse partielle où pour l'estimation du coefficient de régression quand l'information auxiliaire est disponible sur un ensemble d'individus et notre variable d'intérêt est connue seulement sur un sous-ensemble de celui-ci. L'évolution d'autres fonctions non-linéaires de totaux plus compliquées, comme l'indice de Gini par exemple, peuvent être estimées par cette approche.

Cet article est structuré de la façon suivante : la section présente un court rappel sur les fonctions d'influence (Hubert 1981) et sur la technique de linéarisation par la fonction d'influence introduite par Deville (1999) dans le cas d'un seul échantillon. Nous proposons ensuite une extension de cette méthode à deux échantillons. La statistique complexe est approximée par une combinaison linéaire de totaux avec des poids qui dépendent des deux échantillons. La section 3 est consacrée à la recherche des poids optimaux au sens de la variance. Plus précisément, on définit dans la section 3.1 un plan de sondage multidimensionnel; le cas bidimensionnel est décrit plus en détail. Dans la section 3.2 on donne les probabilités d'inclusion de premier et deuxième degré bidimensionnelles. Ensuite, dans la section 3.3 nous trouvons le meilleur estimateur de type Horwitz-Thompson pour un paramètre

d'intérêt Z quand l'information provient de deux échantillons différents s_1 et s_2. Un estimateur de la variance est également donné. Afin de contourner les difficultés provenant du fait de considérer un grand nombre de paramètres, on propose dans la section 3.4 une réduction de leur nombre. La section 3.5 donne des résultats pour l'estimation d'un ratio pour un plan de sondage particulier, le sondage de Bernoulli bidimensionnel et conditionnel à la taille.

1. Linéarisation de statistiques complexes par la fonction d'influence

Soit $T = \{t, t+1, ..., t+T-1\}$ un ensemble fini de dates et pour chaque t on considère une population

$$U = \{1, ..., k, ..., N\}$$

que l'on suppose, pour l'instant, la même pendant les T tirages, (on néglige dans un premier temps les naissances et les morts). A chaque individu $k \in U$, nous associons le vecteur x_k de $\mathbf{R}^q$ où $q = p \times T$, p est le nombre de variables et T est le nombre d'instants de mesure. On note $y_{jk} = y_j(x_k) \in R^T$ la valeur de la $j$ème variable pour l'individu k aux différents instants. Enfin on note y_{jkt} la valeur de y_{jk} à l'instant t.

On considère la mesure M sur $\mathbf{R}^q$ qui attribue la masse 1 pour chaque élément de la population. Supposons que le paramètre d'intérêt q (un indice par exemple) est une fonction linéaire ou non, de totaux :

$$q = j(t_{y_1}, ..., t_{y_p}), \quad t_{y_j} = \sum_U y_{jk} \text{ pour } j = 1, ..., p$$

Le paramètre d'intérêt q s'écrit comme une fonctionnelle de M :

$$q = j\left(\int y_1 dM; ...; \int y_p dM\right).$$

Dans la suite on considère $T = 2$, $q = (q_t, q_{t+1})$. Soit $q_t = Q(M), q_{t+1} = S(M)$ les valeurs de la variable d'intérêt q aux instants t et $t+1$. Elles sont estimées grâce aux échantillons $s_t \subset U$ à l'instant t et $s_{t+1} \subset U$ en $t+1$. Soit $\hat{M}_t$ l'estimateur de la mesure M à l'instant t qui attribue un poids $w_{k,t}$ pour tous les éléments $k \in s_t$ et zéro au reste de la population. Un estimateur naturel pour la fonction $Q(M)$ est l'estimateur par substitution $Q(\hat{M}_t)$. De manière similaire, à l'instant $t+1$, on définit un estimateur $\hat{M}_{t+1}$ de M et $S(\hat{M}_{t+1})$ l'estimateur par substitution de $S(M)$. La fonction d'influence au point x d'une fonctionnalité $\Phi(M)$, si elle existe, est définie par

$$I\Phi(M, x) = \lim_{h \to 0} \frac{1}{h}\left(\Phi(M + hd_x) - \Phi(M)\right)$$

Pour une statistique Q homogène de degré a, Deville (1999) montre, sous certaines hypothèses générales, que l'estimateur par substitution vérifie la relation :

$$N^{-a}\left(Q(\hat{M}_t) - Q(M_t)\right) = \sum_{k \in U} z_{k,t} w_{k,t} e_k^t - \sum_{k \in U} z_{kt} + O_p\left(\frac{1}{\text{card}(s_t)}\right)$$

où $e_k^t = \mathbf{1}(k \in s_t)$, et la variable linéarisée $z_{kt} = IQ(M; x_k)$ est la valeur de la fonction d'influence IQ appliquée à Q et au système de points x_k. De manière similaire, on approxime $S(M_{t+1})$.

De plus, la variance de l'estimateur par substitution peut être approximée par la variance de l'estimateur de Horwitz-Thompson pour le total de z_{kt} :

$$\mathrm{Var}\left(N^{-a}\left(Q(\hat{M}_t)-Q(M)\right)\right)\cong \mathrm{Var}\left(\sum_{s_t} z_{k,t}w_{k,t}\right)$$

Il est nécessaire d'estimer les z_{kt} pour obtenir un estimateur de la variance. Deville (1999) a montré qu'on obtenait un estimateur consistant en substituant $z_{k,t}$ par $IQ(\hat{M}_t)$ dans l'expression de la variance.

Le résultat précédent peut être utilisé pour estimer la variance de n'importe quelle combinaison des fonctionnelles $Q(M)$ et $S(M)$, notée $\Phi(Q(M),S(M))$.

On a d'une manière générale :

$$I\Phi(Q(M),S(M))=\Phi'_Q IQ(M)+\Phi'_S IS(M)$$

où Φ'_Q et Φ'_S sont les dérivées partielles de Φ par rapport à Q et à S. Notons $\hat{\Phi}=\Phi(Q(\hat{M}_t),\ S(\hat{M}_{t+1}))$ l'estimateur de Φ par substitution. En généralisant le calcul de Deville (1999), on a alors le résultat :

$$N^{-a}(\hat{\Phi}-\Phi)=\frac{1}{N}\sum_U(\Phi'_Q IQ+\Phi'_S IS)w_k(s_t,s_{t+1})-\frac{1}{N}\sum_U(\Phi'_Q IQ+\Phi'_S IS)+o_p(1) \qquad (1)$$

et une approximation de la variance

$$\mathrm{Var}\left(N^{-a}(\hat{\Phi}-\Phi)\right)\cong \mathrm{Var}\left(\frac{1}{N}\sum_U(\Phi'_Q IQ+\Phi'_S IS)\,w_k(s_t,s_{t+1})\right) \qquad (2)$$

où $w_k(s_t,s_{t+1})$ pour tous $k\in U$ sont les poids à déterminer dépendant à la fois de s_t et s_{t+1}. Nous allons pour cela minimiser le critère suivant

$$\mathrm{Var}\left(\frac{1}{N}\sum_U(\Phi'_Q IQ+\Phi'_S IS)\,w_k(s_t,s_{t+1})\right) \qquad (3)$$

Exemple : estimation d'un ratio

On veut estimer un ratio $R=\frac{Y}{X}$ quand les variables X et Y sont observées sur deux échantillons différents s_1 et s_2. La linéarisée de R a l'expression

$$r_k=-\frac{Y}{X^2}x_k+\frac{1}{X}y_k$$

quand la fonction d'influence de X, $IX=x_k$ est connue sur s_1 et celle de Y, $IY=y_k$ est connue sur s_2. Alors,

$$\hat{R}-R=\sum_U\left\{-\frac{Y}{X^2}x_k+\frac{1}{X}y_k\right\}w_k(s_1,s_2)$$
$$-\sum_U\left\{-\frac{Y}{X^2}x_k+\frac{1}{X}y_k\right\}+O_p\left(\frac{1}{\mathrm{card}\,(s_1\cup s_2)}\right) \qquad (4)$$

Résoudre (3), revient ici à chercher les poids $w_k(s_1,s_2)$ qui minimisent la variance de (4).

2. Théorie de type Horwitz-Thompson pour deux échantillons.

2.1 Le plan de sondage multidimensionnel

On sélectionne un échantillon s_t dans U_t, $t \in T$. Un plan de sondage multidimensionnel est une loi de probabilité $p(s = s_1,...,s_T))$ sur l'espace $[P(U)]^T$ (Cotton & Hesse 1992). Il résulte qu'on peut déduire les lois pour $2^T - 1$ intersections et réunions de s_t dans U_t, $t \in T$. Les plans de sondage $p_t(s_t)$ sur s_t sont définis comme des lois marginales. Pour $T = 2$, on dispose des lois de probabilité de $s_1, s_2, s_{12} = s_1 \cap s_2$ et $s_1 \cup s_2$ aussi bien que les lois des deux "oreilles" $s_{1*} = s_1 - s_{12}$ et $s_{2*} = s_2 - s_{12}$; on note ces lois avec $p_\oplus$ pour $\oplus \in \{1,2,1*,2*,1 \cup 2,1* \cup 2*\}$. Des plans de sondage usuels comme le panel ou le plan en deux phases peuvent être obtenus comme cas particuliers. Par exemple, pour $p(s_1 = s_2) = 1$, on a le panel et le deux-phases pour $p(s_2 | s_1) = 0$ sauf si $s_2 \subset s_1$. De façon concrète, l'échantillonnage bidimensionnel peut se réaliser de différentes manières : tirages disjoints de l'intersection et des deux oreilles, tirage de s_1 suivi de partage en deux et tirage disjoints de la deuxième oreille, tirages coordonnés de divers façons etc.

2.2 Les probabilités d'inclusion bidimensionnelles

Notre objectif est d'étendre à deux échantillons l'estimation de type H-T. Dans le cas d'un seul échantillon , on cherche un estimateur sans biais dont les poids ne dépendent que de l'appartenance ou non d'un individu k à l'échantillon. Si les seuls aléas viennent de l'échantillonnage, on trouve l'estimateur $\sum_U \frac{y_k}{p_k} e_k$ où $\varepsilon_k = 1$ $(k \in s)$ et $\pi_k = Pr(k \in s)$. Après avoir défini les probabilités d'inclusion d'ordre deux, $\pi_{kl} = Pr(k,l \in s)$ la variance de l'estimateur a pour expression $\sum_U \sum_U \Delta_{kl} \frac{y_k}{p_k} \frac{y_l}{p_l}$, $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

La situation est plus compliquée quand on dispose de deux échantillons ; à chaque unité k dans U vont correspondre sept variables aléatoires dont trois indépendantes $\varepsilon_k^\oplus$ avec $\oplus \in \{1,2,12,1*,2*,1 \cup 2,1* \cup 2*\}$. On peut choisir les variables indépendantes parmi 29 choix possibles. Dans notre étude, on a choisi de prendre comme variables indépendantes $\varepsilon_k^{1*}, \varepsilon_k^{12}, \varepsilon_k^{2*}$. Les autres possibilités peuvent être obtenues à l'aide des transformations linéaires.

Chaque unité $k \in U$ a une probabilité d'inclusion de premier degré notée $\pi_k^\oplus$ par rapport au plan de sondage $p_\oplus$; elle satisfait $\pi_k^\oplus = \mathrm{E}(\varepsilon_k^\oplus)$ avec $\oplus \in \{1*,12,2*\}$. Dans le cas multidimensionnel, Cotton & Hesse (1992) donnent une définition pour les probabilités d'inclusion de premier degré. Nous appliquons cette définition dans le cas bidimensionnel et on montre que les quantités ainsi obtenues sont les mêmes $\pi_k^\oplus$ données auparavant. Ensuite, les probabilités d'inclusion de second degré par rapport au plan bidimensionnel sont données.

Considérons le plan bidimensionnel $s = (s_1, s_2)$. Pour chaque unité k dans U nous pouvons définir la trace de l'échantillon s sur k, notée $\mathrm{tr}_k(s)$, comme vecteur d'éléments les traces de s_1 et s_2 sur l'individu k, c'est à dire

$$\mathbf{s_k} = \mathrm{tr}_k(s) = (s_1 \cap \{k\}, s_2 \cap \{k\}) \in \{\emptyset, \{k\}\} \times \{\emptyset, \{k\}\}.$$

Cotton et Hesse (1992) définissent les probabilités d'inclusion de premier degré bidimensionnelles notées $f_k(\mathbf{s_k})$ comme

$$f_k(\mathbf{s_k}) = \sum_{s \in \mathrm{tr}_k^{-1}(\mathbf{s_k})} p(s) \text{ pour } \mathbf{s_k} \in \{\emptyset, \{\mathrm{k}\}\}^2, k \in U$$

Par exemple, $f_k(\{\{k\},\{k\}\}) = \sum_{s \in \mathrm{tr}_k^{-1}(\{\{k\},\{k\}\})} p(s) = \sum_{k \in s_{12}} p(s_1, s_2) = \pi_k^{12}$ et on obtient les probabilités d'inclusion de premier degré classiques par rapport à la loi de l'intersection. On peut obtenir de manière équivalente les $\pi_k^{1*}, \pi_k^{2*}, \pi_k^{**}$ pour tous $k \in U$.

Nous allons définir dans la suite les probabilités d'inclusion de deuxième ordre dérivées par rapport au plan bidimensionnel $s = (s_1, s_2)$. Soit $\mathbf{s_{k,l}}$ la trace de s sur le couple (k,l) où

$$\mathbf{s_{k,l}} = \mathrm{tr}_{k,l}(s) = (s_1 \cap \{k,l\}, s_2 \cap \{k,l\}) \in \{\emptyset, \{k\}, \{l\}, \{k,l\}\}^2$$

et notons avec $f_{k,l}(\mathbf{s_{k,l}})$ la probabilité que les individus k,l appartiennent à l'échantillon bidimensionnel s. Alors, $f_{k,l}(\mathbf{s_{k,l}})$ a l'expression suivante

$$f_{k,l}(\mathbf{s_{k,l}}) = \sum_{s \in \mathrm{tr}_{k,l}^{-1}(\mathbf{s_{k,l}})} p(s) \quad \text{pour} \quad \mathbf{s_{k,l}} \in \{\emptyset, \{k\}, \{l\}, \{k,l\}\}^2$$

Dans le cas particulier où $\mathbf{s_{k,l}} = \{\{k,l\},\{\emptyset,k\}\}$ la probabilité $(k,l) \in (s_1, s_1)$ et $(\mathrm{k,l}) \in (s_2, \emptyset)$ est

$$f_{k,l}(\{\{k,l\},\{\emptyset,k\}\}) = \sum_{s \in \mathrm{tr}_{k,l}^{-1}(\{\{k,l\},\{\emptyset,k\}\})} p(s) = \sum_{k,l \in s_1, k \in s_2, l \notin s_2} p(s_1, s_2) = p_{kl}^{12,1*}.$$

Pour les autres valeurs de $\mathbf{s_{k,l}}$ nous obtenons $\pi_{k,l}^{\oplus,\otimes} = E(\varepsilon_k^{\oplus} \varepsilon_l^{\otimes})$ avec $\oplus, \otimes \in \{1*, 12, 2*\}$.

Notons par analogie avec Särndal *et al.* (1992) dans le cas unidimensionnel, $\Delta_{kl}^{\oplus,\otimes} = p_{k,l}^{\oplus,\otimes} - p_k^{\oplus} p_l^{\otimes}$ pour tous $k,l \in U$. On peut remarquer que les $\pi_{k,l}^{\oplus,\otimes}$ satisfont :

- $p_{k,l}^{\oplus,\oplus} = p_{k,l}^{\oplus}$ les probabilités classiques de second degré par rapport au plan unidimensionnel $p_{\oplus}$;
- $\pi_{k,l}^{\oplus,\otimes} \neq \pi_{l,k}^{\oplus,\otimes}$ généralement, mais $\pi_{k,l}^{\oplus,\otimes} = \pi_{l,k}^{\otimes,\oplus}$.
- pour $\oplus \neq \otimes \in \{1*, 12, 2*\}$ on a $\Delta_{kk}^{\oplus,\otimes} = -\pi_k^{\oplus} \pi_k^{\otimes}$.

La matrice $\Delta_{kl}^{\oplus,\otimes}$ pour tous $k,l \in U$ et $\oplus, \otimes \in \{1*, 12, 2*\}$ aura dans la suite la même signification que dans le cas unidimensionnel, c'est une matrice de variance-covariance de neufs blocs $\left(\Delta_{kl}^{\oplus,\otimes}\right)_{k,l=\overline{1,N}}$ dont seulement six sont différents.

2.3 Estimation linéaire sans biais optimale

On considère les paramètres d'intérêt de la forme $Z = fX + yY$ où X est le total de la variable X mesurée sur un échantillon s_1 et Y le total de la variable Y mesurée sur un autre échantillon s_2.

On estime Z à l'aide d'un estimateur linéaire sans biais $\hat{Z}$, avec

$$\hat{Z} = \sum_U (x_k + y_k) w_k(s_1, s_2)$$

où les poids $w_k(s_1, s_2)$ dépendent à la fois de s_1 et s_2,

$$w_k(s_1, s_2) = w_k^{1*} e_k^{1*} + w_k^{2*} e_k^{2*} + w_k^{12} e_k^{12}.$$

Dans s_{12} on décompose w_k^{12} en deux poids qui dépendent de la variable d'intérêt. Il résulte alors,

$$\hat{Z} = \sum_{k \in s_1^*} w_k^{1*} x_k + \sum_{k \in s_2^*} w_k^{2*} y_k + \sum_{k \in s_{12}^*} (w_k^{12x} x_k + w_k^{12y} y_k) \tag{5}$$

On propose une reparamétrisation des w_k qui va permettre de réduire le nombre de paramètres à déterminer, tout en vérifiant la condition des sans biais pour Z. On prend $w_k(s_1, s_2)$ comme suit

$$\begin{cases} w_k^{1*} = \dfrac{a_k}{\mathbf{p}_k^{1*}} \;;\; w_k^{12x} = \dfrac{f - a_k}{\mathbf{p}_k^{12}} \\ w_k^{2*} = \dfrac{b_k}{\mathbf{p}_k^{2*}} \;;\; w_k^{12x} = \dfrac{y - b_k}{\mathbf{p}_k^{12}} \end{cases} \tag{6}$$

et cela conduit à l'expression suivante $\hat{Z}$:

$$\hat{Z} = q' \operatorname{diag} \begin{pmatrix} x \\ \dots \\ y \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \hat{t}_{H-T} \tag{7}$$

où $q = ((a_k)_{k=1}^N, (b_k)_{k=1}^N)' \in \mathbf{R}^{2N}$ est le vecteur de paramètres inconnus, a, b d'éléments

$a_k = \frac{e_k^{1*}}{p_k^{1*}} - \frac{e_k^{12*}}{p_k^{12}}$, $ß_k = \frac{e_k^{2*}}{p_k^{2*}} - \frac{e_k^{12*}}{p_k^{12}}$, $\hat{t}_{H-T} = \sum_U \frac{f x_k + y y_k}{\mathbf{p}_k^{12}} \mathbf{e}_k^{12}$.

Il existe un grand choix d'estimateurs sans biais parmi vérifiant (7). On va choisir celui qui a la variance minimale. La variance de $\hat{Z}$ a l'expression

$$V(\hat{Z}) = q' \Gamma q + 2q' g + c$$

où Γ, γ matrices de dimensions $2N * 2N$, $2N * 1$ et $c = \mathrm{Var}(\hat{t}_{H-T})$.

On voit que la variance est une forme quadratique en q avec $c \in \mathbf{R}$, $\gamma \in \mathbf{R}^{2N}$ et Γ une matrice symétrique positive qui ne dépend pas de q.

Si la matrice Γ est définie positive alors la variance de $\hat{Z}$ est *minimale* pour $q = -\Gamma^{-1} g$ avec la valeur optimale $V_{opt}(\hat{Z}) = V(\hat{t}_{H-T}) - g' \Gamma^{-1} g \leq V(\hat{t}_{H-T})$. Montanari (1987) obtient la même expression pour le paramètre d'une régression multiple quand l'estimateur de la différence généralisée est utilisé.

Le résultat obtenu a un intérêt théorique mais pas en pratique, si les dimensions de Γ et γ sont grandes, le calcul de q est presque impossible. De plus, ces matrices dépendent de toutes les valeurs inconnues y_k, x_k pour tous $k \in U$.

On peut obtenir également une formule de type $\mathrm{H}-\mathrm{T}$ pour la variance de $\hat{Z}$ qui permettra d'en déduire un estimateur. En écrivant $\hat{Z}$ de la façon suivante :

$$\hat{Z} = (vect)' \begin{pmatrix} e^{1*} \\ e^{2*} \\ e^{12} \end{pmatrix}$$

avec $(vect)' = \left(\left(\frac{a_k x_k}{p_k^{1*}} \right)_{k=1}^{N}, \left(\frac{b_k y_k}{p_k^{2*}} \right)_{k=1}^{N}, \left(\frac{x_k(f-a_k)+y_k(y-b_k)}{p_k^{12}} \right)_{k=1}^{N} \right)$.

Alors la variable a l'expression

$$V(\hat{Z}) = (vect)' \begin{pmatrix} \Delta_{kl}^{1*,1*} & \Delta_{kl}^{1*,2*} & \Delta_{kl}^{1*,12} \\ \Delta_{kl}^{2*,1*} & \Delta_{kl}^{2*,2*} & \Delta_{kl}^{2*,12} \\ \Delta_{kl}^{12,1*} & \Delta_{kl}^{12,2*} & \Delta_{kl}^{12,12} \end{pmatrix} (vect) \qquad (8)$$

où chaque $\Delta_{kl}^{\oplus,\otimes}$ est en effet une matrice de taille $N \times N$ pour $\oplus,\otimes \in \{1*,12,2*\}$.

L'expression (8) est analogue en deux dimensions de la variance d'un estimateur Horvitz-Thompson. On a neuf blocs de variance-covariance $(\Delta_{kl}^{\oplus,\otimes})_{kl \in U}$ par rapport à un seul pour le cas unidimensionnel classique. On va utiliser l'expression (8) dans le but d'obtenir une estimation de la variance de $\hat{Z}$ similaire au cas unidimensionnel, c'est à dire en fonction des quantités de type $\breve{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$. La situation est plus compliquée à cause des termes de covariance. Les produits croisés comme $\sum_{k \in U} \sum_{l \in U} \Delta_{kl}^{1*,2*} \frac{x_k}{\pi_k^{1*}} \frac{y_l}{\pi_l^{2*}}$, peuvent être estimés seulement dans l'espace de l'intersection s_{12}. On obtient $\sum_{k \in s_{12}} \sum_{l \in s_{12}} \frac{\Delta_{kl}^{\oplus,\otimes}}{p_k^{12}} \frac{x_k}{p_k^{\oplus}} \frac{y_l}{p_l^{\otimes}}$ pour $\oplus,\otimes \in \{1*,12,2*\}$. Pour estimer les termes restants, on procède de la manière suivante : les sommes doubles en X sont estimées sur s_1 et celles de Y sur s_2. On obtient $\sum_{k \in s_1} \sum_{l \in s_1} \frac{\Delta_{kl}^{\oplus,\otimes}}{\pi_k^{1}} \frac{x_k}{\pi_k^{\oplus}} \frac{y_l}{\pi_l^{\otimes}}$ pour $\oplus,\otimes \in \{1*,12\}$ et $\sum_{k \in s_2} \sum_{l \in s_2} \frac{\Delta_{kl}^{\oplus,\otimes}}{\pi_k^{2}} \frac{x_k}{\pi_k^{\oplus}} \frac{y_l}{\pi_l^{\otimes}}$ pour $\oplus,\otimes \in \{12,2*\}$.

2.4 Réduction du nombre de paramètres

Comme on l'a précisé dans la section précédente, le fait de considérer beaucoup de paramètres réduit considérablement les possibilités de dériver leurs expressions et implicitement, la variance de $\hat{Z}$. Une réduction du nombre de paramètres est proposée dans Deville & Goga (2002). Tout d'abord est considéré le cas quand les N valeurs de a_k, b_k sont stratifiés dans H classes, respectivement J, suivi du cas extrême quand $a_k = a$ pour tous $k \in U$ et $b_k = b$ pour tous $k \in U$ où a, b sont des constantes. Pour cette dernière situation, (7) devient

$$\begin{aligned} \hat{Z} &= a(\hat{X}^{1*} - \hat{X}^{12}) + b(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{t}_{H-T} \\ &= (a,b) \begin{pmatrix} \boldsymbol{x}'\boldsymbol{\alpha} \\ \boldsymbol{y}'\boldsymbol{\beta} \end{pmatrix} + \hat{t}_{H-T} \end{aligned} \qquad (9)$$

avec $\boldsymbol{x}'\boldsymbol{\alpha} = \hat{X}^{1*} - \hat{X}^{12}$, $\boldsymbol{y}'\boldsymbol{\beta} = \hat{Y}^{2*} - \hat{Y}^{12}$ et $\hat{t}_{H-T}$ garde la même expression $\sum_U \frac{fx_k + yy_k}{p_k^{12}} \boldsymbol{e}_k^{12}$.

En notant $\boldsymbol{?} = (a,b)'$, la valeur optimale $\boldsymbol{?}_{opt} = -\Gamma^{-1} g$ où

$$\Gamma = V\begin{pmatrix} \boldsymbol{x}'\boldsymbol{a} \\ \boldsymbol{y}'\boldsymbol{ß} \end{pmatrix} = \begin{pmatrix} V(\hat{X}^{1*} - \hat{X}^{12}) & \mathrm{Cov}(\boldsymbol{x}'\boldsymbol{a}, \boldsymbol{y}'\boldsymbol{ß}) \\ \mathrm{Cov}(\boldsymbol{x}'\boldsymbol{a}, \boldsymbol{y}'\boldsymbol{ß}) & V(\hat{Y}^{2*} - \hat{Y}^{12}) \end{pmatrix} \tag{10}$$

$$g = \mathrm{Cov}\left(\begin{pmatrix} \boldsymbol{x}'\boldsymbol{a} \\ \boldsymbol{y}'\boldsymbol{ß} \end{pmatrix}, \hat{t}_{H-T}\right) = \begin{pmatrix} \mathrm{Cov}(\boldsymbol{x}'\boldsymbol{a}, \hat{t}_{H-T}) \\ \mathrm{Cov}(\boldsymbol{y}'\boldsymbol{ß}, \hat{t}_{H-T}) \end{pmatrix}$$

Il en résulte que dans ce cas $\boldsymbol{?}_{opt}$ dépend des termes de variance et covariance qui peuvent être estimés dans certaines situations.

2.5 Le cas d'un sondage de Bernoulli bidimensionnel conditionnel à la taille de S_1, S_2, S_{12}

On tire un échantillon bidimensionnel de Bernoulli $s = (s_1, s_2)$ considéré comme un cas particulier d'un sondage de Poisson bidimensionnel introduit par Cotton & Hesse (1992). Comme chaque échantillon s_1, s_2 et s_{12} est de taille variable, on va considérer le plan de sondage de Bernoulli bidimensionnel conditionnel aux taille de s_1, s_2 et s_{12} décrit dans Deville & Goga (2002). On donne dans la suite la valeur optimale du paramètre $q = (a,b)'$ quand on veut estimer une combinaison linéaire $Z = fX + yY$.

Résultat : Pour un échantillon bidimensionnel de Bernoulli n_1, n_2, n_{12} conditionnel aux taille de s_1, s_2 et s_{12} on a

$$\hat{Z} = a\hat{X}^{1*} + (f-a)\hat{X}^{12} + b\hat{Y}^{2*} + (y-b)\hat{Y}^{12} \tag{11}$$

$$\boldsymbol{?}_{opt} = \frac{-h_1 h_2}{1 - r^2 h_1 h_2}\begin{pmatrix} f r^2 + r\, y\,(1 - h_2^{-1})S - f h_2^{-1} \\ y r^2 + r\, f\,(1 - h_1^{-1})S^{-1} - y h_1^{-1} \end{pmatrix} \tag{12}$$

où $h_1 = \frac{n_1^*}{n_1}$, $h_2 = \frac{n_2^*}{n_2}$ sont les taux de renouvellement, $S = \frac{S_y}{S_x}$ et r est le coefficient de corrélation. La variance optimale peut se calculer d'après la formule $\mathrm{Var}_{opt} = V_{H-T} - g'\Gamma g$; il résulte une expression assez compliquée pour la variance, raison pour laquelle on a évité de la donner.

Application à l'estimation d'un ratio

On applique les résultats trouvés ci-dessus pour estimer un ration $R = \frac{Y}{X}$ quand X est mesuré sur s_1 et Y sur s_2. Cette situation peut apparaître par exemple en présence de nonréponse. Par exemple, un échantillon s est sélectionné dans U est la non réponse se produit différemment pour les variables X et Y. Plus précisément, nous avons $s_1 \subset s$ comme ensemble des répondants pour X et $s_2 \subset s$ pour Y avec une intersection assez grande en général. On suppose comme modèle de nonréponse le modèle de Bernoulli bidimensionnel conditionnellement à la taille largement décrit dans Deville & Goga (2002).

Appliquons la technique de linéarisation présentée en section 3.5. La variable linéarisée de R est $z_k = -\frac{Y}{X^2} x_k + \frac{1}{X} y_k$, expression qui ne dépend pas de l'échantillon. On va avoir :

$$\hat{R} - R = \left\{ \sum_U \left(-\frac{Y}{X^2} x_k + \frac{1}{X} y_k \right) w_k(s_1, s_2) - Z \right\} + O_p\left(\frac{1}{\text{card}(s_1 \cup s_2)} \right)$$

Dans ce cas $f = -\frac{Y}{X^2}$ et $y = \frac{1}{X}$ alors

$$\hat{R}_{opt} - R \cong a_{opt}(\hat{X}^{1*} - \hat{X}^{12}) + b_{opt}(\hat{Y}^{2*} - \hat{Y}^{12}) + \hat{t}_{H-T} - Z$$

avec

- $\hat{t}_{H-T} = f\hat{X}^{12} + y\hat{Y}^{12}$;
- a_{opt}, b_{opt} les paramètres optimaux ;
- $Z = \sum_U (fx_k + yy_k)$ est le total de la variable z_k.

L'estimateur par substitution de R a l'expression :

$$\hat{R}_{opt} = \frac{b_{opt}\hat{Y}^{2*} + (y - b_{opt})\hat{Y}^{12}}{a_{opt}\hat{X}^{1*} + (f - a_{opt})\hat{X}^{12}}$$

Comparaison avec des estimateurs naturels de $\hat{R}$

1. On peut estimer R en ne considérant que l'intersection s_{12}. En utilisant la technique de linéarisation, on déduit que

$$\hat{R}^{12} - R \cong \sum_U z_k \frac{\varepsilon_k^{12}}{\pi_k^{12}} - Z = \phi\hat{X}^{12} + \psi\hat{Y}^{12} - Z \quad (13)$$

$$= \hat{t}_{H-T} - Z \quad (14)$$

avec $\hat{R}^{12} = \frac{\hat{Y}^{12}}{\hat{X}^{12}}$

Il résulte qu'on peut exprimer la variance de $\hat{R}^{12}$ par la variance de $\hat{t}_{H-T}$ notée avec V_{H-T}. Alors, on a toujours $AVar_{opt}(\hat{R}_{opt}) \leq AVar(\hat{R}^{12})$ indépendamment du plan de sondage utilisé.

2. Une autre possibilité est d'estimer le total de X sur s_1 et le total de Y sur s_2.

Il résulte que R est estimé par $\hat{\hat{R}}^{1,2} = \frac{\hat{Y}^2}{\hat{X}^1}$ et l'approximation suivante fonctionne

$$\hat{\hat{R}}^{1,2} - R \cong \phi\hat{X}^1 + \psi\hat{Y}^2 - Z$$

Pour un sondage de Bernoulli bidimensionnel à la taille, on a

$$\phi\hat{X}^{1} + \psi\hat{Y}^{2} = \phi h_1(\hat{X}^{1*} - \hat{X}^{12}) + \psi h_2(\hat{Y}^{1*} - \hat{Y}^{12}) + \hat{t}_{H-T}$$

avec $h_1 = \frac{n_1^*}{n_1}$ et $h_2 = \frac{n_2^*}{n_2}$ les taux de renouvellement. Il résulte alors que $\phi\hat{X}^{1} + \psi\hat{Y}^{2}$ fait partie de la classe des estimateurs considérés (9) avec $a = \phi h_1$ et $b = \psi h_2$ et par conséquent la variance approximative de $\hat{\hat{R}}^{1,2}$ est supérieure à la variance approximative de $\hat{R}_{opt}$:

$$AVar_{opt}(\hat{R}_{opt}) \leq AVar(\hat{\hat{R}}^{1,2})$$

où *AVar* représente la variance approximative.

2.6 Cas général $\mathbf{Z} = f\mathbf{X} + y\mathbf{Y} + d\mathbf{V} + c\mathbf{T}$

On est confronté à une situation plus générale mais tout à fait réaliste. (cf exemple du coefficient de régression) quand on désire estimer une quantité de type $Z = fX + yY + dV + cT$ où chaque lettre majuscule représente le total d'une variable d'intérêt ; X et Y sont comme précédemment, la variable V du total V est mesuré sur $s_{12} = s_1 \cap s_2$ et la variable T du total T est mesuré sur $s_1 \cup s_2$. On cherche à estimer Z par

$$\hat{Z} = \sum_U (x_k + y_k + v_k + t_k) w_k(s_1, s_2)$$

où $w_k(s_1, s_2) = w_k^{1*} e_k^{1*} + w_k^{2*} e_k^{2*} + w_k^{12} e_k^{12}$. Cette fois-ci, comme les variables X et T sont mesurées sur s_{1*} on décompose w_k^{1*} en deux poids qui dépendent de la variable d'intérêt et qui sont notés $w_k^{1*,x}$ et $w_k^{1*,t}$. On procède de la même façon avec w_k^{2*} et w_k^{12} ; on obtient les ensembles de poids $w_k^{2*,y}$ et $w_k^{2*,t}$ pour w_k^{2*} et $w_k^{12,x}$, $w_k^{12,y}$, $w_k^{12,v}$, $w_k^{12,t}$ pour w_k^{12} respectivement.

Considérons la paramétrisation suivante pour $w_k(s_1, s_2)$:

$$\begin{cases} w_k^{1*,x} = \dfrac{a_k}{p_k^{1*}} \;;\; w_k^{12,x} = \dfrac{f - a_k}{p_k^{12}} \\ w_k^{2*,y} = \dfrac{b_k}{p_k^{2*}} \;;\; w_k^{12,y} = \dfrac{y - b_k}{p_k^{12}} \\ w_k^{12,v} = \dfrac{d - b_k}{p_k^{12}} \\ w_k^{1*,t} = \dfrac{c_k}{p_k^{1*}} \;;\; w_k^{2*,t} = \dfrac{d_k}{p_k^{2*}} \;;\; w_k^{12t} = \dfrac{c - c_k - d_k}{p_k^{12}} \end{cases} \tag{15}$$

Elle vérifie la condition de sans biais pour $\hat{Z}$. On peut écrire $\hat{Z}$ sous la forme matricielle suivante

$$\hat{Z} = q' \begin{pmatrix} \text{diag}x & 0 \\ 0 & \text{diag}y \\ \text{diag}t & 0 \\ 0 & \text{diag}t \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \hat{t}_{H-T} \tag{16}$$

où $q'' = \left((a_k)_{k=1}^N, (b_k)_{k=1}^N, (c_k)_{k=1}^N, (d_k)_{k=1}^N\right)' \in \mathbf{R}^{\mathbf{4N}}$ et α, β sont des vecteurs de dimension N d'éléments $\alpha_k = \left(\frac{\varepsilon_k^{1*}}{\pi_k^{1*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}}\right)$, $\beta_k = \left(\frac{\varepsilon_k^{2*}}{\pi_k^{2*}} - \frac{\varepsilon_k^{12}}{\pi_k^{12}}\right)$ pour tous $k \in U$

et $\hat{t}_{H-T} = \sum_U \frac{\phi x_k + \psi y_k + \delta v_k + \chi t_k}{\pi_k^{12}} \varepsilon_k^{12}$.

Si on note $H = \begin{pmatrix} \text{diag}x & 0 \\ 0 & \text{diag}y \\ \text{diag}t & 0 \\ 0 & \text{diag}t \end{pmatrix}$ alors $\theta_{opt} = -\Gamma^{-1}\gamma$ avec

$$\begin{cases} \Gamma = H \times \text{Var}\left(\begin{pmatrix} a \\ b \end{pmatrix}\right) \times H' \\ g = H \times \text{Cov}\left(\begin{pmatrix} a \\ b \end{pmatrix} \times \hat{t}_{H-T}\right) \end{cases} \tag{17}$$

Cas particuliers

1. Si V et T ne sont pas disponibles, on estime $Z = \phi X + \psi Y$ par $\hat{Z}$ obtenu de (16) après avoir éliminé les $2N$ dernières lignes de la matrice H et les termes qui contiennent t_k et v_k dans $\hat{t}_{H-T}$. L'expression (7) de $\hat{Z}$ obtenue dans la section 5 est ainsi retrouvée.

2. Si V et T sont disponibles, alors en posant $\delta = \chi = 0$ on estime $Z = \phi X + \psi Y$ en bénéficiant de la connaissance de V et T qui peut être considérée comme une information auxiliaire.

$$\hat{Z} = \theta' \times H + \hat{t}_{H-T}$$

avec le même paramètre θ de dimension $4N$ et la matrice H mais cette fois-ci $\hat{t}_{H-T} = \phi \hat{X}^{12} + \psi \hat{Y}^{12}$.

3. Si on ne dispose que des variables X et T alors $Z = \phi X + \chi T$

$$\text{et } \hat{Z} = q' \begin{pmatrix} \text{diag}x & 0 \\ \text{diag}t & 0 \\ 0 & \text{diag}t \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \hat{t}_{H-T} \tag{18}$$

où $q'' = \left((a_k)_{k=1}^N, (c_k)_{k=1}^N, (d_k)_{k=1}^N\right)' \in \mathbf{R}^{\mathbf{3N}}$ et $\hat{t}_{H-T} = \phi \hat{X}^{12} + \chi \hat{T}^{12}$

4. Si on dispose de X , de Y et de V ou de X , de Y et de T :

- Dans le premier cas, on estime $Z = \phi X + \psi Y$ par $\hat{Z}$ obtenu de (16) après avoir éliminé les $2N$ dernières lignes de la matrice H , et $\hat{t}_{H-T} = \phi \hat{X}^{12} + \psi \hat{Y}^{12} + \delta \hat{V}^{12}$.
- Dans le deuxième cas , la matrice H ne change pas et $\hat{t}_{H-T} = \phi \hat{X}^{12} + \psi \hat{Y}^{12} + \chi \hat{T}^{12}$.

Exemple de coefficient de régression

Pour le coefficient de régression $B = \dfrac{\sum_U x_k t_k}{\sum_U t_k^2}$ quand la variable X est connue sur un échantillon s_1 et la variable T sur $s_1 \cup s_2$ la variable linéarisée est $t^{-1} t_k (x_k - t_k B)$ où $\tau = \sum_U t_k^2$

$$\hat{B} - B \cong \tau^{-1} \sum_U \left(t_k x_k - t_k^2 B\right) w_k (s_1, s_2) - \tau^{-1} \sum_U \left(t_k x_k - t_k^2 B\right)$$

ce qui correspond au cas 3. ci-dessus. On peut déduire a_{opt} , c_{opt} , d_{opt} optimaux et

$$\hat{B} - B \cong \tau^{-1} a_{opt} \left(\widehat{X * T^{1*}} - \widehat{X * T^{12}}\right) + \tau^{-1} \left[c_{opt} \left(\hat{\tau}^{1*} - \hat{\tau}^{12}\right) + d_{opt} \left(\hat{\tau}^{2*} - \hat{\tau}^{12}\right)\right] \quad (19)$$

$$- \tau^{-1} \sum_U \left(t_k x_k - t_k^2 B\right) \quad (20)$$

où $X * T = \sum_U x_k t_k$, $\tau = \sum_U t_k^2$ avec leur estimateurs de Horwitz-Thompson sur s_{1*} , s_{12} , s_{2*} .

Références

[1] Bell, P. (2001). Comparaison d'autres estimateurs pour l'Enquête sur la population active. *Technique d'enquête*, 27, 1, pp. 57-68

[2] Caron N., and Ravalet, P. (2000). Estimation dans les enquêtes répétées : Application à l'Enquête Emploi en continu. *Document de travail INSEE de la Direction des Statistiques Démographiques et Sociales N° 0005.*

[3] Cassel, C.M, Särndal, C.E., and Wretman, JH. (1992). *Model assisted survey sampling.* Springer-Verlag

[4] Cochran, W. G., (1977). *Sampling Technique*. Wiley, New-York

[5] Cotton, F. , Hesse, C. (1992). Tirages coordonnes d'échantillons. *Document de travail INSEE de la Direction des Statistiques Economiques E9206.*

[6] Deville, J.C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. *Survey Methodology*, 25, 193-203.

[7] Deville, J.C., Goga, C. (2002). The Horwitz-Thompson theory fot two samples, preprint

[8] Fuller, W., Rao, J.N.K (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquêtes*, 27, 1, pp 49-56.

[9] Gourieroux, C. and Roy, G. (1978). Enquête en deux vagues : renouvellement de l'échantillon. *Annales de l'INSEE*, 29, 115-144.

[10] Hidiroglu, M. A. (2001) Double sampling. *Survey methodology*, to appear.

[11] Huber, P. J. (1981). *Robust Statistics.* John Wiley and Sons

[12] Kish, L. (1965). *Survey Sampling*. New-York, Wiley.

[13] Montanari, G. E. (1987). Post-sampling efficient prediction in large scale surveys. International Statistical Review, 55, pp.191-202.

[14] Singh, A. C., Kennedy, B. Wu, S. (2001). Estimation composite par régression pour l'Enquête sur la populationactive du Canada avec plan de sondage à renouvellement de panel. Techniques d'enquêtes. 27,1,pp.35-48.

[15] Tam, S. M. (1984). On Covariances From Overlapping Samples. The American Statistican, 38, 4, pp.288-289.