N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

TECHNIQUE D'ESTIMATION À DOUBLE BASE POUR LES ENQUÊTES À SUIVI DE NON-RÉPONSE

Avinash C. Singh, Vincent G. Iannacchione et Jill A. Dever[1]

RÉSUMÉ

Dans des enquêtes où les taux de réponse sont bas, on peut faire une enquête de suivi auprès des non-répondants pour augmenter le nombre de répondants à l'enquête principale. Pour une estimation efficiente reposant sur l'enquête principale et l'enquête complémentaire et dans un mode analogue à celui de la technique d'estimation pour de petites régions, nous proposons d'élaborer un estimateur composite qui ménagerait un équilibre entre la variance de l'estimateur instable fondé sur les échantillons principal et complémentaire, d'une part, et le biais de l'estimateur stable fondé sur le seul échantillon principal, d'autre part. Il reste que, comme il s'agit d'un problème de grandes et non de petites régions, le cadre d'estimation à double base peut servir à sa formulation. De plus, on peut dégager une pondération composée d'un calage où s'intègrent des contrôles de valeurs extrêmes de pondération, tout en préservant les totaux de contrôle connus de population et les totaux de contrôle à valeur nulle pour les estimations de différences entre les deux échantillons dans le cas d'un jeu clé de variables étudiées. Nous illustrons la méthode que nous proposons par une enquête auprès de vétérans de la guerre du Golfe ayant comporté une enquête de suivi de non-réponse.

MOTS CLÉS : Biais de non-réponse, erreur quadratique moyenne, estimation à double base, estimation pour de petites régions, pondération par calage, valeurs de pondération extrêmes.

1. INTRODUCTION

Une importante application de l'échantillonnage double pour la stratification implique le recours à une enquête complémentaire de suivi pour modérer l'effet de non-réponse dans l'enquête principale (Hansen et Hurvitz, 1946; Cochran,1977, p. 370). Dans une telle situation, on forme une strate distincte avec les non-répondants d'après les résultats de l'enquête initiale (de la première phase) auprès de l'échantillon. Cette application se justifie par le modèle de réponse de la population (pour des références récentes, voir, par exemple, Fay, 1991, et Shao et Steel, 1999) où, dans des conditions déterminées d'enquête, on pose qu'un indicateur de réponse aléatoire peut être attribué à chaque unité de la population avant que l'échantillonnage n'ait lieu. Toutefois, l'appartenance d'unités aux strates de non-réponse et de réponse est inconnue tant que la collecte initiale de données durant la première phase n'est pas achevée. Lors de la seconde phase, on tire un sous-échantillon de non-répondants à l'enquête initiale qui est ensuite soumis à une enquête de suivi, avec d'ordinaire un effort plus intense que celui qui s'exerce dans le cadre de l'enquête initiale.

En théorie, on peut faire appel à des échantillons d'enquête de suivi pour atténuer les limitations de corrections de non-réponse par modèle dans des estimations d'enquête où on utilise seulement l'échantillon principal, les corrections de biais par modélisation de non-réponse pouvant être inadéquates dans le cas d'enquêtes où le taux de non-réponse est élevé. La réalité est cependant qu'une réponse incomplète à l'enquête complémentaire se trouve à diluer l'apport effectif d'une réduction de biais, d'où la nécessité d'une certaine modélisation de non-réponse. Les considérations de coût limitent habituellement la taille de l'échantillon de suivi, ce qui accroît la variabilité des estimations à cause de l'effet marqué de pondération inégale (EPI) dans l'échantillon composite. Même là, un suivi de non-réponse peut nous livrer des indications importantes sur les non-répondants et le biais de non-réponse, plus particulièrement dans les enquêtes où les taux de réponse sont bas. Une enquête de suivi présente un autre avantage de taille dont il sera principalement question dans le présent document. Nous démontrons que, par rapport à l'estimateur habituel fondé sur un double échantillonnage pour la stratification, un estimateur plus efficace peut être

[1] Avinash C. Singh, Vincent G. Iannacchione et Jill A. Dever, RTI International, Statistics Research Division, 3040, chemin Cornwallis, Research Triangle Park, NC 27709, États-Unis.

conçu par un autre mode de combinaison des données des échantillons principal et complémentaire. C'est dans cette optique que nous nous inspirons de l'estimation composite dans le cadre d'enquêtes à double base de sondage pour extraire plus d'information des données.

Nous justifions la méthode que nous proposons à la **section 2** par analogie avec l'estimation pour de petites régions, mais nous la formulons en nous reportant à la méthode d'étalonnage à double base (EDB) de Singh et Wu (1996, 2003) après modifications pour tenir compte de la dépendance entre échantillons. Ayant d'abord donné un aperçu de la méthode EDB, nous examinons comment on peut se servir de la méthode de calage MEG pour produire une pondération à restriction de plage de valeurs (non négatives, par exemple) et à contrôle intégré de valeurs extrêmes de pondération d'échantillonnage. Nous décrivons en outre comment la méthode a été adaptée de manière à inclure des contrôles supplémentaires fournis par les deux estimateurs. À la **section 3**, nous exposons pas à pas l'application de la méthode EDB proposée. À la **section 4**, nous traitons du problème de l'estimation de variance. Nous montrons que, après linéarisation de Taylor, l'estimateur de variance standard dans un double échantillonnage pour stratification peut être utilisé. Nous proposons également une estimation de variance plus simple correspondant au cas d'un échantillonnage à une phase et à degrés multiples là où l'échantillon complémentaire est emboîté dans les unités primaires d'échantillonnage (UPE). À la **section 5** enfin, nous présentons les résultats empiriques d'une application à une enquête sur les vétérans de la guerre du Golfe.

2. JUSTIFICATION ET FORMULATION DE LA MÉTHODE EDB

Pour une plus grande efficience des estimateurs d'enquête, il nous faut y intégrer le plus de renseignements utiles possible au stade de l'estimation. Nous considérons à cette fin les deux estimateurs suivants d'un total de population T_y correspondant à une variable étudiée y. Soit s_A désignant les répondants à la première phase (enquête principale) et p_1, le plan d'échantillonnage probabiliste correspondant. Nous définissons un estimateur $\hat{T}_{y(A)}$ convenablement corrigé de la non-réponse à la première phase par le modèle ξ_1, où l'indice A signifie que seuls les répondants de la première phase entrent dans l'estimation. Il est clair que, pour que cet estimateur soit utile compte tenu de la faible réponse prévue à la première phase, il nous faut obtenir de bons prédicteurs de non-réponse (peut-être au moyen de sources administratives), de sorte que la mise en correspondance répondants-non-répondants puisse se faire dans le modèle ξ_1.

Soit s_B désignant les répondants aux deux phases (enquêtes principale et complémentaire) pour les plans d'échantillonnage probabiliste correspondants p_1 et p_2. Nous définissons un second estimateur $\hat{T}_{y(B)}$ convenablement corrigé de la non-réponse à la seconde phase par le modèle ξ_2, où l'indice B signifie que tous les répondants entrent dans l'estimation. Cet estimateur ressemble à l'estimateur habituel en cas de double échantillonnage pour la stratification. L'estimateur $\hat{T}_{y(A)}$ est approximativement sans biais pour le plan d'échantillonnage p_1 et le modèle de non-réponse correspondant ξ_1. Il en va de même de l'estimateur $\hat{T}_{y(B)}$ pour le plan $p = p_1p_2$ et le modèle ξ_2. L'estimateur $\hat{T}_{y(A)}$ est stable, mais sera probablement entaché d'un biais à cause des limites de ξ_1; en revanche, $\hat{T}_{y(B)}$ sera instable, mais probablement presque sans biais. On peut aborder le problème de la combinaison de ces deux estimateurs en prenant pour justification l'estimation pour de petites régions ER où l'estimateur composite représente un bon compromis entre le biais de $\hat{T}_{y(A)}$ et la variance de $\hat{T}_{y(B)}$. Comme dans l'estimation ER, un estimateur composite peut être formé par une combinaison linéaire convexe telle que l'EQM en définition mixte pour le modèle $\xi = (\xi_1, \xi_2)$ et le plan d'échantillonnage $p = (p_1, p_2)$ s'en trouve minimisée. À noter qu'il est impossible de dégager une estimation stable du biais sans modèle, aussi supposons-nous que les modèles ξ_1 et ξ_2 sont valides pour le calcul de l'EQM. Il reste que, comme le modèle ξ_1 est jugé « fragile », l'estimateur composite résultant devrait être biaisé, mais d'une valeur plus proche – avec plus de stabilité – de celle de $\hat{T}_{y(B)}$.

D'ordinaire, les deux bases sont en chevauchement partiel dans une telle situation et les deux échantillons sont indépendants. Dans le cas qui nous préoccupe cependant, ces bases sont identiques et les deux échantillons sont

dépendants, partageant les répondants s_A de la première phase. Singh et Wu (1996, 2003) ont envisagé un calage à double base (EDB) pour l'élaboration d'estimations composites comme estimateurs avec facteur d'extension, et ce, en utilisant une estimation par régression linéaire pour une poststratification simultanée des échantillons tirés des deux bases. On fait en sorte de respecter les totaux de contrôle correspondant aux variables auxiliaires habituelles (x), de même qu'un certain nombre de nouveaux totaux de contrôle à valeur nulle correspondant à de nouvelles variables auxiliaires (z). Dans ce cas, les variables z dénotent certaines variables clés étudiées pour la méthode d'estimation composite. Ce sont des variables sur lesquelles des données sont recueillies dans les deux échantillons. La variable y représente une variable à l'étude arbitraire qui peut être une des variables z directement prises en compte dans l'estimation composite.

Nous passerons brièvement en revue les éléments de la méthode EDB avant d'examiner les modifications à prévoir en fonction d'échantillons dépendants. Pour la simplicité de l'illustration, nous examinerons comment le problème de la combinaison de deux estimateurs non biaisés et indépendants, $\hat{T}_{y(A)}$ et $\hat{T}_{y(B)}$, d'un total commun de population T_y peut être reformulé en un problème de calage selon une nouvelle contrainte définie par rapport à la variable étudiée y et un total de contrôle à valeur nulle. Posons deux échantillons aléatoires simples (EAS) de tailles respectives n_A et n_B (la taille globale est désignée par n) prélevés sur une population de taille N. Une combinaison linéaire optimale pour la minimisation de la variance nous est donnée par

$$\hat{T}_{y,opt} = \alpha_{opt}\,\hat{T}_{y(A)} + (1+\alpha_{opt})\hat{T}_{y(B)}, \tag{2.1a}$$

où

$$\alpha_{opt} = V(\hat{T}_{y(B)})\left[V(\hat{T}_{y(A)}) + V(\hat{T}_{y(B)})\right]^{-1}. \tag{2.1.b}$$

Pour des EAS, nous avons $\hat{T}_{y(A)} = \sum_{s_A} y_{kA}\, d_{kA}$, où $d_{ka} = N / n_A$ et où l'estimation de variance habituelle est

$$\hat{V}(\hat{T}_{y(A)}) = \left(1-\frac{n_A}{N}\right)\left(\frac{N}{n_A-1}\right)\sum_{s_A} y_{kA} d_{kA}(y_{kA}-\bar{y}_A). \tag{2.1.c}$$

L'estimation de variance de $\hat{V}_{y(B)}$ reçoit la même définition. Ainsi,

$$1-\hat{\alpha}_{opt} = \left(\sum\nolimits_{s_A} y_{kA} d_{kA}\,(y_{kA}-\bar{y}_A)(1-n_A/N)(N/(n_A-1))\right)\left(\hat{V}(\hat{T}_{y(A)}) + \hat{V}(\hat{T}_{y(B)})\right)^{-1}. \tag{2.1d}$$

Nous obtenons après reformulation de $\hat{T}_{y,opt}$ comme

$$\hat{T}_{y,opt} = \hat{T}_{y(A)} + (1-\hat{\alpha}_{opt})\left(\hat{T}_{y(B)} - \hat{T}_{y(A)}\right), \tag{2.2a}$$

$$\begin{aligned}\hat{T}_{y,opt} &= \sum_{s_A} y_{kA} d_{kA}\left[1 + c_{kA}(y_{kA}-\bar{y}_A)\hat{V}^{-1}\left(0-(\hat{T}_{y(A)}-\hat{T}_{y(B)})\right)\right] \\ &\approx \sum_{s_A} y_{kA} d_{kA}\left[1+(n_A/n)^{-1}(\hat{\lambda}_{yA}/n)(y_{kA}-\bar{y}_A)\right] = \sum_{s_A} y_{kA} d_{kA} a_{kA} = \sum_{s_A} y_{kA} w_{kA}\end{aligned}, \tag{2.2b}$$

où $\hat{V} = \hat{V}(\hat{T}_{y(A)}) + \hat{V}(\hat{T}_{y(B)})$, $c_{kA} = \left(1-\frac{n_A}{N}\right)\frac{N}{n_A-1} \approx \frac{N}{n_A}$ (sans la correction de population finie ou cpf),

$\hat{\lambda}_{yA} = N\left[\hat{V}(\hat{T}_{y(A)}) + \hat{V}(\hat{T}_{y(B)})\right]^{-1}\left(0-(\hat{T}_{y(A)}-\hat{T}_{y(B)})\right)$ et $a_{kA} = 1+(n_A/n)^{-1}(\hat{\lambda}_{yA}/n)(y_{kA}-\bar{y}_A)$.

Cette formule ressemble à celle d'un estimateur par calage par régression linéaire. On peut voir le facteur $\hat{\lambda}_{yA}/n$ comme $O_p(n^{-1/2})$ dans les conditions habituelles. Il convient de noter que, contrairement à ce qui se passe dans une estimation habituelle par régression pour des enquêtes à base unique avec variables auxiliaires *x*, le paramètre λ du facteur de correction est ici à une échelle déterminée par l'inverse de la taille effective relative de l'échantillon n_A / n. Il est bon de noter que, comme on pouvait s'y attendre, l'échantillon de taille supérieure a généralement des corrections inférieures (les facteurs sont plus proches de l'unité). En d'autres termes, il y a des corrections différenciées de la pondération pour les deux échantillons; les corrections sont moindres pour l'échantillon de plus grande taille relative, de sorte que les deux estimations deviennent identiques. De même, nous pouvons écrire $\hat{T}_{y,opt}$ en termes de w_{kB}. Nous obtenons

$$\begin{aligned}\hat{T}_{y,opt} &= \hat{T}_{y(B)} + \hat{\alpha}_{opt}\left(\hat{T}_{y(A)} - \hat{T}_{y(B)}\right) \approx \sum_{s_B} y_{kB} d_{kB}\left[1 + (n_B/n)^{-1}\left(\hat{\lambda}_{yB}/n\right)\left(y_{kB} - \bar{y}_B\right)\right] \\ &= \sum_{s_B} y_{kB} d_{kB} a_{kB} = \sum_{s_B} y_{kB} w_{kB};\ \hat{\lambda}_{yB} = -\hat{\lambda}_{yA}\,. \end{aligned} \tag{2.3}$$

Les équations (2.2b) et (2.3) impliquent que les valeurs initiales de pondération d_{kA} pour s_A et d_{kB} pour s_B sont respectivement calées à w_{kA} et w_{kB} et que, par conséquent, les estimations de l'un et l'autre des échantillons sont identiques et égales à $\hat{T}_{y,opt}$. En d'autres termes, la différence entre les deux estimations devient nulle après calage, zéro étant le nouveau total de contrôle. Au cœur même de la technique d'étalonnage proposée, il y a cette idée de totaux de contrôle à valeur nulle pour les nouvelles variables auxiliaires définies par les variables clés à l'étude (ce que nous désignerons désormais par *z*) avec des différences réduites par contrainte à zéro entre les deux estimations pour la base de sondage en chevauchement. Dans le cas présent, les deux estimateurs sont dépendants à cause des échantillons communs, et il faut donc apporter une modification appropriée au paramètre α_{opt} en combinaison en tenant compte de la covariance.

Dans le cas de plans d'échantillonnage complexes, il est généralement difficile d'exprimer la combinaison linéaire optimale en une formule de calage. Voilà pourquoi Singh et Wu (1996, 2003) ont pris une extension de l'estimateur REGG (régression généralisée) pour proposer un estimateur composite sous-optimal sous forme d'estimateur par calage. Les formes des facteurs de correction sont données par

$$a_{kA} = 1 + \eta_A^{-1}(\mathbf{x}_A'\lambda_A + \mathbf{z}_A'\lambda_z)\,,\ a_{kB} = 1 + \eta_B^{-1}(\mathbf{x}_B'\lambda_B - \mathbf{z}_B'\lambda_z)\,, \tag{2.4}$$

où **x** désigne les covariables auxiliaires habituelles à totaux de contrôle connus T_x et où **z** est le jeu de variables clés à l'étude. On notera avec intérêt que les facteurs de correction en question peuvent s'obtenir par minimisation de la fonction de la distance suivante en fonction des contraintes de calage et en termes de variables x et z :

$$\Delta(w,d) = \eta_A \sum\nolimits_{s_A} d_{kA}(a_{kA} - 1)^2 + \eta_B \sum\nolimits_{s_B} d_{kB}(a_{kB} - 1)^2\,. \tag{2.5}$$

À noter que ***z*** figure avec des signes différents dans les deux facteurs de correction, puisque les totaux de contrôle sont nuls dans ce cas. Les paramètres λ_z sont toutefois communs aux deux facteurs. Les paramètres (ou facteurs d'échelle) η_A et η_B sont préspécifiés et indiquent les tailles relatives d'échantillon selon les plans propres aux deux échantillons. Ils sont analogues aux tailles relatives d'échantillon du cas EAS. Dans la pratique, on peut déterminer les paramètres η par une recherche en grille permettant de minimiser la variance de $\hat{T}_{z,comp}$ (ou la trace de la matrice des covariances pour ***z*** multidimensionnel). Ainsi, le choix de η_A et η_B sera automatiquement le reflet d'une éventuelle dépendance des deux échantillons. À la prochaine section, nous parlerons de l'utilisation du modèle exponentiel généralisé (MEG) de Folsom et Singh (2000) au lieu de l'estimation REGG, ce modèle comportant un contrôle intégré de valeurs extrêmes de pondération pour des domaines appropriés de ces valeurs et donnant des facteurs de correction sur des plages préspécifiées. Enfin, le MEG offre un mode unifié de correction tant de non-réponse que de poststratification.

3. ESTIMATEUR EDB PROPOSÉ

3.1 Définition de l'EDB

Pour les deux échantillons de répondants en chevauchement, s_A (première phase) et s_B (première et seconde phases), définis à la section 2, soit $s_{A\|B}$ désignant l'échantillon en concaténation. Notre but est d'élaborer l'estimateur composite suivant d'un total de population :

$$\hat{T}_{y(A\|B)} = \zeta_A \sum_{s_A} w_{kA}\, y_{kA} + (1-\zeta_A) \sum_{s_B} w_{kB}\, y_{kB}\ , \tag{3.1}$$

où y_{kA} = résultat d'intérêt pour le k^e répondant de s_A, y_{kB} = résultat d'intérêt pour le k^e répondant de s_B, w_{kA} = poids de calage du k^e répondant de s_A et w_{kB} = poids de calage du k^e répondant de s_B.

$$\zeta_A = \left(\sum_{s_B} w_{kB}^2 - \sum_{s_A} w_{kA} w_{kB} \right) \left(\sum_{s_A} w_{kA}^2 + \sum_{s_B} w_{kB}^2 - 2 \sum_{s_A} w_{kA} w_{kB} \right)^{-1} \tag{3.2}$$

Il faut en outre que soient respectés les contrôles habituels sur $\boldsymbol{x}$ et les nouveaux contrôles à valeur nulle sur $\boldsymbol{z}$. Le paramètre ζ_A traduit la dépendance de s_A et s_B. On le choisit pour minimiser la variance lorsque celle-ci est approchée par une constante en multiplication avec la formule EPI $(1+CV^2(w))$ par un modèle simple de moyenne commune de super-population (Kish, 1965). Ainsi, ζ_A est une approximation de a_{opt} livrée par l'équation (2.1b).

Dans le cas des variables z, le paramètre ζ_A n'a aucune incidence, les estimations de s_A et s_B étant les mêmes. Toutefois, le choix du facteur d'échelle η_A influe sur la combinaison qui est implicite dans le calage de pondération (2.5). Le choix se fait *a priori* ou par voie empirique par une recherche en grille sur la plage $0 < \eta_A < 1$ qui permet de minimiser la variance. En revanche, pour une variable arbitraire y, les deux estimations sont différentes, et il faut le paramètre ζ_A pour leur combinaison. Le choix (3.2a) pour ζ_A est simple et heuristique. Autre possibilité : pour diverses variables y, on peut choisir un ζ_A approprié qui est commun à tous les y et tel que la variance généralisée est minimisée. Dans le calcul de variance, les facteurs η_A et ζ_A sont considérés comme préspécifiés, puisqu'on suppose que des données antérieures ont servi à leur estimation.

L'estimateur composite peut aussi s'exprimer sous la forme suivante :

$$\hat{T}_{y(A\|B)} = \sum_{s_B} w_k^* y_{kB}; w_k^* = \zeta_A w_{kA} + \zeta_B w_{kB},\ \text{si}\ k \in s_A,\ \text{et}\ \zeta_B w_{kB},\ \text{si}\ k \in (s_B - s_A), \tag{3.3}$$

où w_k^* désigne un ensemble unique de valeurs finales de pondération par calage pour tout l'échantillon s_B.

Dans les formules qui précèdent, les poids de calage sont ainsi définies :

$$w_{kA} = d_{kA} a_{kA,nr} a_{kA,ps}, \quad w_{kB} = d_{kB} a_{kB,nr} a_{kB,ps}, \tag{3.4}$$

où d_{kA} et d_{kB} sont les valeurs de pondération déjà définies pour le plan d'échantillonnage et les facteurs a, les corrections de non-réponse (nr) et de poststratification (ps). À noter qu'une correction par quotient du type Hajek (Hajek, 1971, dans ses observations sur l'étude de Basu) est utile avant toute correction de pondération (pour nr ou ps), car elle atténue l'effet des valeurs extrêmes de pondération comme dans la fable de l'éléphant de Basu. En fait, elle produit le centrage souhaité (Singh et Sarndal, 2003) des coefficients de régression dans l'estimation REGG habituelle pour la poststratification, si bien que l'estimateur optimal par régression EAS (voir la section 2) peut s'obtenir comme cas particulier de l'estimation REGG.

3.2 Recours au calage MEG pour l'EDB

Comme nous l'avons mentionné à la section 2, le MEG offre un mode unifié de correction tant de non-réponse que de poststratification. Nous nous servons de ce modèle pour déterminer les facteurs de correction de pondération de la forme suivante (nous illustrons ici par les facteurs de poststratification) pour les paramètres bornés préspécifiés $\ell_{kA} < c_{kA} < u_{kA}, \quad \ell_{kB} < c_{kB} < u_{kB}$

$$a_{kA} = \frac{\ell_{kA}(u_{kA} - c_{kA}) + u_{kA}(c_{kA} - \ell_{kA})\exp_A}{(u_{kA} - c_{kA}) + (c_{kA} - \ell_{kA})\exp_A}, \; a_{kB} = \frac{\ell_{kB}(u_{kB} - c_{kB}) + u_{kB}(c_{kB} - \ell_{kB})\exp_B}{(u_{kB} - c_{kB}) + (c_{kB} - \ell_{kB})\exp_B}, \tag{3.5}$$

où $\exp_A = \exp[\eta_A^{-1} A_{kA}(\mathbf{x}'_{kA}\boldsymbol{\lambda}_{xA} + \mathbf{z}'_{kA}\boldsymbol{\lambda}_z)]$, $A_{kA} = \dfrac{m_{kA}(u_{kA} - \ell_{kA})}{(u_{kA} - c_{kA})(c_{kA} - \ell_{kA})}$, $m_{kA} = b_{kA} / d_{kA}$ et les expressions correspondantes pour l'échantillon B reçoivent une définition analogue.

Le paramètre b_k (pour un traitement plus commode, nous avons supprimé les indices A et B) désigne les limites de définition des valeurs extrêmes de pondération. Ainsi, les valeurs extrêmes après calage satisfont l'inégalité $b_k \ell_k < w_k < b_k u_k$ ou les facteurs de correction correspondants, l'inégalité $m_k \ell_k < a_k < m_k u_k$, où $a_k = w_k / d_k$. En d'autres termes, les poids extrêmes corrigés ne sont pas nécessairement tronqués aux limites b_k, mais restent au voisinage de ces limites selon les données et les totaux de contrôle de la procédure de calage. On notera que, pour les valeurs non extrêmes, $m_k = 1$. Il est clair que, dans des applications réelles, il est loin d'être pratique de varier les limites selon les unités k. À en juger par notre expérience, il suffit de disposer de trois ensembles de limites (l_1, c, u_1), (l_2, c, u_2) et (l_3, c, u_3) pour les hautes valeurs extrêmes, les valeurs non extrêmes et les basses valeurs extrêmes. Le centre c est fixé respectivement à 1,0 pour la correction de poststratification et à plus de 1,0 (propension de réponse inverse globale, par exemple) pour la correction de non-réponse. Les facteurs d'échelle η_A et η_B $(=1-\eta_A)$ figurant dans les facteurs de correction peuvent s'interpréter comme des paramètres qui reflètent la différence de tailles effectives d'échantillon entre les deux échantillons, ainsi que l'effet de dépendance de ces derniers. On obtient des estimations des paramètres λ des facteurs de correction (3.5) en résolvant simultanément les équations suivantes :

$$\sum_{s_A} \mathbf{x}_{kA}\mathbf{d}_{kA}\mathbf{a}_{kA} = \mathbf{T}_\mathbf{x}, \sum_{s_B} \mathbf{x}_{kB}\mathbf{d}_{kB}\mathbf{a}_{kB} = \mathbf{T}_\mathbf{x}, \sum_{s_A} \mathbf{z}_{kA}\mathbf{d}_{kA}\mathbf{a}_{kA} - \sum_{s_B} \mathbf{z}_{kB}\mathbf{d}_{kB}\mathbf{a}_{kB} = \mathbf{0}, \tag{3.6}$$

où $T_\mathbf{x}$ est le vecteur des totaux habituels de poststratification pour $\mathbf{x}$. La macro MEG de RTI applique la méthode de Newton-Raphson à cette fin. Les facteurs de correction (3.5) peuvent directement s'obtenir par minimisation d'une fonction de la distance (indiquée ci-après) à l'aide des multiplicateurs de Lagrange et selon les contraintes de calage (3.6).

$$\begin{aligned} &\Delta(w, d) = \eta_A \Delta_A(w_A, d_A) + \eta_B \Delta_B(w_B, d_B), \\ &\Delta_A(w_A, d_A) = \sum_{s_A} \frac{d_{kA}}{A_{kA}} \left[(a_{kA} - l_{kA}) \log \frac{a_{kA} - l_{kA}}{c_{kA} - l_{kA}} + (u_{kA} - a_{kA}) \log \frac{u_{kA} - a_{kA}}{u_{kA} - c_{kA}} \right] \end{aligned} \tag{3.7}$$

et Δ_B reçoit une définition semblable.

3.3 Étapes de l'EDB

Voici une description sommaire des étapes de l'application de la méthode EDB proposée :

1. On définit deux échantillons d'unités sélectionnées, à savoir l'échantillon A (s_A* = ensemble des unités sélectionnées pour l'enquête principale) et l'échantillon B (s_B* = répondants à l'enquête principale, plus ensemble des unités sélectionnées pour l'enquête complémentaire). On fixe aussi les valeurs des paramètres η et les limites des facteurs de correction de non-réponse et de poststratification.

2. On apporte une correction par quotient du type Hajek à la pondération d'échantillonnage pour chaque échantillon de sorte que chaque ensemble de valeurs de pondération donne par addition les totaux de population spécifiés.
3. On apporte une correction de non-réponse par le modèle MEG aux poids dégagés à l'étape 2 pour les ensembles de répondants s_A dans s_A* et s_B dans s_B*. Les facteurs de correction sont respectivement définis par $a_{kA,nr}$ et $a_{kB,nr}$. À noter que la sommation des valeurs corrigées de pondération pour les deux ensembles de répondants correspond au chiffre total de population spécifié à l'étape 2 par correction de Hajek.
4. Il y a ensuite correction de poststratification par le modèle MEG pour l'échantillon concaténé $s_{A\|B}$ avec les contrôles démographiques habituels et les nouveaux contrôles de calage à valeur nulle (3.7); les facteurs de correction sont définis par $a_{kA,ps}$ et $a_{kB,ps}$ pour l'un et l'autre des échantillons.
5. On choisit η_A de manière à minimiser la fonction objective définie par la variance ou la trace de la matrice des covariances (voir la section suivante). On calcule les poids de calage pour l'échantillon A (w_{kA}) et l'échantillon B (w_{kB}). On se sert de la formule en 3.2 (ou minimisation de variance généralisée) pour calculer le paramètre ζ_A et, par là, w_k* selon la définition qui en a été donnée (3.3).

4. ESTIMATION DE VARIANCE

4.1 Plans d'échantillonnage simples avec facteurs de correction aléatoires

Pour bien prendre en compte les facteurs de correction aléatoires, nous pouvons procéder par linéarisation de Taylor (voir, par exemple, Singh et Folsom, 2000, Binder, 1996, et Binder et coll., 2000) et ainsi obtenir un estimateur de variance qui soit approximativement sans biais. Nous supposons que les paramètres η_A, η_B, ζ_A et ζ_B sont donnés *a priori* et peuvent donc être considérés comme non aléatoires. L'hypothèse est acceptable si ces paramètres sont estimés à partir de données chronologiques. Pour la simplicité de l'illustration, nous considérerons uniquement les facteurs de correction de poststratification comme aléatoires. Pour $\hat{T}_{z(A\|B)}$, nous présentons l'estimateur linéarisé qui contient les écarts de Taylor comme résidus. L'opération se fait de deux façons, l'une pour l'échantillon A et l'autre pour l'échantillon B. Nous avons

$$\hat{T}_{z(A\|B)} = \sum_{s_A} z_{kA} w_{kA} = \sum\nolimits_{s_A} z_{kA} d_{kA} a_{kA,nr} a_{kA,ps} \approx \sum_{s_{A\|B}} \Delta_k + const \equiv \hat{T}^{linA}_{z(A\|B)} + const\,, \tag{4.1}$$

où $\Delta_k = \delta_{kA} w_{kA} z_{kA} - A_k'(H^{-1})' \sum\nolimits_{s_A} w_{kA} B_k z_{kA}$, $A_k' = (\delta_{kA} w_{kA} x_{kA}', \quad \delta_{kB} w_{kB} x_{kB}', \quad \delta_{kA} w_{kA} z_{kA}' - \delta_{kB} w_{kB} z_{kB}')$

$$B_k' = (\delta_{kA} a_{kA,ps}^{-1} \phi_{kA,ps} x_{kA}', \quad \delta_{kB} a_{kB,ps}^{-1} \phi_{kB,ps} x_{kB}', \delta_{kA} a_{kA,ps}^{-1} \phi_{kA,ps} z_{kA}' - \delta_{kB} a_{kB,ps}^{-1} \phi_{kB,ps} z_{kB}')\,. \tag{4.2a}$$

Ajoutons que δ_{kA} est un si l'unité est dans s_A et zéro dans les autres cas; δ_{kB} reçoit une définition semblable. La variable $\phi_{kA,ps}$ se définit comme

$$\phi_{kA,ps} = \eta_A^{-1} m_{kA} \frac{(u_{kA,ps} - a_{kA,ps})(a_{kA,ps} - l_{kA,ps})}{(u_{kA,ps} - c_{kA,ps})(c_{kA,ps} - l_{kA,ps})} \tag{4.2b}$$

et la matrice H est $\sum_{s_{A\|B}} x_{k,ps}^{*} x_{k,ps}'^{*} d_k a_{k,nr} \phi_{k,ps}$, où $x_{kA,ps}' = (\delta_{kA} x_{kA}', \delta_{kB} x_{kB}', \delta_{kA} z_{kA}' - \delta_{kB} z_{kB}')$.

D'une autre manière équivalente, nous pouvons formuler l'estimateur linéarisé en termes d'échantillon B comme

$$\hat{T}_{z(A\|B)} = \sum_{s_B} z_{kB} w_{kB} = \sum\nolimits_{s_B} z_{kB} d_{kB} a_{kB,nr} a_{kB,ps} \approx \sum_{s_{A\|B}} \Delta_k + const \equiv \hat{T}^{linB}_{z(A\|B)} + const\,. \tag{4.3}$$

Pour une variable *y* arbitraire, l'expression linéarisée de $\hat{T}_{y(A\|B)}$ pour l'estimation de variance comporterait ζ_A, car les deux échantillons peuvent ne pas donner des estimations calées identiques. On peut aisément l'obtenir sous la forme suivante

$$\hat{T}_{y(A\|B)} \approx \zeta_A \hat{T}^{linA}_{y(A\|B)} + (1-\zeta_A)\hat{T}^{linB}_{y(A\|B)}, \tag{4.4}$$

ce qu'on pourrait réexprimer comme la somme de deux termes, l'un concernant l'échantillon A et l'autre, l'échantillon B. Si *y* devait être une des variables *z*, nous obtiendrions la même estimation de variance qu'en (4.3) ou (4.5), parce que

$$\begin{aligned}&\zeta_A^2 Var\left(\hat{T}^{linA}_{z(A\|B)}\right) + (1-\zeta_A)^2 Var\left(\hat{T}^{linB}_{z(A\|B)}\right) + 2\zeta_A(1-\zeta_A)Cov\left(\hat{T}^{linA}_{z(A\|B)}, \hat{T}^{linB}_{z(A\|B)}\right)\\ &= \left[\zeta_A^2 + (1-\zeta_A)^2 + 2\zeta_A(1-\zeta_A)\right]Var\left(\hat{T}_{z(A\|B)}\right) = Var\left(\hat{T}_{z(A\|B)}\right).\end{aligned} \tag{4.5}$$

Une fois qu'on dispose de la version linéarisée de l'estimateur, on peut dégager une estimation de la variance de la formule de Rao (1973) ou de Lohr (1999) en remplaçant la variable *y* par les résidus appropriés. On pourrait concevoir des formules semblables pour tenir compte de la variation due aux corrections tant de non-réponse que de poststratification. Il convient de noter que la méthode itérative du quotient habituelle est un cas particulier de la méthode MEG et que la correction simple de poststratification par quotient en est aussi un cas particulier. Ainsi, les résidus mentionnés précédemment représentent une généralisation des résidus servant à la linéarisation des estimations par quotient. À noter aussi que des méthodes utilisant des répliques comme la méthode jackknife constituent une solution de rechange à la méthode de linéarisation de Taylor (Kott et Stukel, 1997; Fuller, 1998; Kim et Sitter, 2003).

4.2 Plans d'échantillonnage complexes

Pour les plans d'échantillonnage à une phase et à plusieurs degrés, on dispose de formules simples d'estimation de variance si les UPE sont considérées comme étant tirées avec remise et qu'il existe des estimations non biaisées du total de population de chaque UPE. Sous cette hypothèse, des formules du type EAS simple ou stratifié peuvent servir à l'estimation de variance. La chose est possible parce que, pour les plans d'échantillonnage habituels à plusieurs degrés, l'échantillonnage au deuxième degré et aux degrés supérieurs respecte les hypothèses d'invariance (les probabilités de sélection aux degrés supérieurs ne dépendent pas des résultats du premier degré) et d'indépendance (la sélection des USE est indépendante d'une UPE à l'autre); voir, par exemple, Sarndal, Swensson et Wretman (1992). Ces conditions sont suffisantes, mais elles peuvent être assouplies. Les principales exigences sont (1) qu'en conditionnant par l'échantillon du premier degré, les sélections soient indépendantes entre UPE aux degrés supérieurs et (2) que l'estimation de chaque UPE soit conditionnellement sans biais pour le total de l'UPE. Pour les plans d'échantillonnage à deux phases, l'hypothèse de l'indépendance de la sélection d'USE dans les UPE n'est pas respectée, mais comme dans notre cas du double échantillonnage pour la stratification, si la stratification respecte les limites des UPE et que l'échantillonnage de la seconde phase est conçu pour être emboîté dans les UPE (en d'autres termes, le plan de traitement considère les UPE comme des sous-strates), la méthode simplifiée d'estimation de variance à une phase serait applicable dans l'hypothèse habituelle d'une sélection des UPE avec remise. À noter que, dans le cas des enquêtes comportant un suivi de non-réponse, cela implique que les unités de l'échantillon complémentaire font l'objet d'une sélection indépendante entre UPE. Si l'échantillonnage de la première phase ne comporte pas d'UPE, il peut être raisonnable de construire des pseudo-UPE par souci de simplifier l'estimation de variance. Pour tenir compte des facteurs aléatoires de correction de calage, on peut appliquer la formule simplifiée reposant sur l'hypothèse de la sélection d'UPE avec remise à l'estimateur linéarisé présenté à la section 4.1.

5. APPLICATION

La 10e enquête anniversaire sur la santé des vétérans de la guerre du Golfe (GWHS) est une enquête probabiliste menée à l'échelle nationale auprès des hommes et des femmes qui ont servi dans toutes les branches des forces armées américaines dans la guerre du golfe Persique en 1991. Cette étude vise principalement (1) à fournir des estimations nationales de dénombrement des combattants de la guerre du Golfe qui ont d'importants ennuis de santé et (2) à modéliser les grandes variables en corrélation avec les problèmes de santé déclarés. Comme autres objectifs, il y a notamment des comparaisons entre les militaires en service actif et les réservistes et l'élaboration de modèles explicatifs distincts de la fréquence des ennuis de santé pour les hommes et les femmes ayant combattu dans le golfe Persique. L'objectif du plan d'échantillonnage de cette étude était la sélection dans la population cible d'un échantillon probabiliste d'ex-combattants d'une taille suffisante pour la réalisation des objectifs d'analyse. La *population cible* du GWHS est formée de plus de 685 000 hommes et femmes qui ont combattu en 1991 dans la guerre du Golfe dans toutes les branches des forces armées américaines. Nous avons tiré un échantillon systématique stratifié de 10 301 combattants de la base de sondage maintenue par le Defense Manpower Data Center. Nous avons délimité quatre strates principales en subdivisant selon le sexe les militaires en service actif et les réservistes. Dans chacune de ces strates, nous avons suréchantillonné les combattants qui s'étaient inscrits au Gulf War Comprehensive Clinical Evaluation Program (CCEP) du Département de la Défense et qui avaient reçu un diagnostic médical fondé sur la 9e Révision de la Classification internationale des maladies, le but étant d'obtenir un nombre suffisant d'intéressés déclarant d'importants ennuis de santé. La base de sondage a en outre fait l'objet d'un tri selon la race et l'ethnicité pour l'obtention d'un échantillon représentatif.

L'enquête initiale a eu lieu par la poste en 2001. Nous sommes parvenus à un taux global de réponse de 54,4 % (selon la définition RR3 de l'AAPOR) après trois envois postaux du questionnaire, l'envoi d'une carte postale de rappel et un appel téléphonique de relance. Les taux de réponse à l'enquête postale ont été plus élevés chez les femmes, les réservistes et les gens qui avaient été évalués dans le cadre du programme CCEP. Le taux général de réponse a été inférieur de 20 % à ce qui était prévu. Soucieuse de réduire le biais susceptible d'être introduit par la non-réponse, l'équipe de projet a décidé d'effectuer un suivi téléphonique auprès d'un sous-échantillon de non-répondants. Nous avons arrêté une taille de sous-échantillon de suivi de 1 000 non-répondants, soit environ le cinquième de tous les répondants par la poste, en fonction des fonds disponibles pour l'étude. La répartition de l'échantillon de suivi a été inversement proportionnelle aux taux de réponse postale des diverses strates. Préalablement à la sélection, chaque non-répondant a été caractérisé comme probablement « facile » ou « difficile » à joindre selon que l'intervieweur avait pu ou non prendre contact avec un membre du ménage de l'ex-combattant à l'occasion des tentatives d'obtention d'une réponse à l'enquête postale. Les non-répondants caractérisés comme « faciles à joindre » ont été suréchantillonnés de manière à accroître la taille de l'échantillon effectif du suivi. Pour alléger le fardeau de réponse, nous avons limité le suivi téléphonique à 69 des 151 questions du questionnaire postal. Nous avons obtenu un taux global de réponse de 55,5 % (définition RR3 de l'AAPOR) pour les 1 000 non-répondants sélectionnés en vue de l'enquête téléphonique complémentaire. Les tendances de la réponse ont été les mêmes que pour l'enquête postale, mais le taux de réponse s'est le plus accru chez les hommes en service actif qui n'avaient pas été évalués dans le cadre du programme CCEP. Au total, 5 709 unités de l'échantillon admissibles ont répondu soit à l'enquête par la poste, soit à l'enquête complémentaire au téléphone. Pour les deux enquêtes, le taux global de réponse après pondération (ce qu'on appelle le *taux effectif de réponse*) pour les unités de l'échantillon admissibles s'est établi à 70,5 % à un intervalle de confiance à 95 % de +/- 3,3 %.

Nous avons appliqué la méthode EDB décrite dans les sections qui précèdent pour calculer la pondération d'analyse GWHS avec $s_A^* = 10\,301$ pour les ex-combattants initialement sélectionnés aux fins de l'enquête et avec $s_B^* = 5\,182$ répondants par la poste, plus 1 000 non-répondants choisis pour le suivi téléphonique. À noter que s_B^* est un sous-ensemble propre de s_A^*. La pondération d'échantillonnage est telle que $\Sigma_{sA^*} \; d_{kA} = \Sigma_{sB^*} \; d_{kB} = 685\,074$ unités de la base de sondage. Les deux échantillons de répondants en chevauchement sont s_A avec 5 182 pour l'enquête postale et s_B avec 5 182 pour cette même enquête et 527 pour l'enquête complémentaire. À noter aussi que s_A est un sous-ensemble propre de s_B. Pour estimer les variances des résultats d'enquête en toute cohérence par rapport au plan d'échantillonnage, nous avons créé 294 répliques (aussi appelées groupes aléatoires), ce qui nous a permis de combiner les données des enquêtes principale et complémentaire. Dans chacune des huit strates de la première phase, nous avons attribué au hasard 35 unités d'échantillon à chaque répétition en exigeant que chacune compte un nombre à peu près égal de répondants à l'enquête postale et au moins un répondant à l'enquête de suivi. Le grand avantage avec des groupes aléatoires est qu'on peut se servir de progiciels d'enquête standard (SUDAAN[MD], par exemple) pour analyser les données. En fait, les estimations de variance obtenues pour

les résultats des répondants à l'enquête postale (le suivi téléphonique étant exclu) par groupes aléatoires étaient seulement un peu « conservatrices » à comparer aux estimations établies pour des plans d'échantillonnage stratifié (à degré unique). Dans le cas de l'estimation de variance à deux phases, nous avons calculé 294 ensembles de valeurs de pondération pour les répliques en vue d'une utilisation avec la méthode jackknife d'estimation de variance à une suppression (Lohr, 1999, p. 298). Nous avons élaboré chaque ensemble de pondération itérative en supprimant en série une répétition de l'échantillon et en corrigeant ensuite la pondération EDB pour tenir compte de la répétition supprimée.

Les totaux de contrôle de poststratification $T_{\boldsymbol{x}}$ correspondent aux 17 chiffres suivants : strates du premier degré (8) : sexe x composante x évaluation CCEP; branche des forces armées (4) : armée, force navale et garde côtière, Marine Corps et force aérienne; race et ethnicité (3) : Blancs, Noirs et autres; hiérarchie militaire (2) : officiers et militaires du rang. Ensuite, nous avons appliqué le modèle MEG à chaque échantillon séparément pour ainsi calculer des facteurs de correction de poststratification $a_{kA,ps}$ et $a_{kB,ps}$, que nous avons appliqués à la pondération d'échantillonnage corrigée pour la non-réponse de sorte que leurs sommes correspondent aux 17 totaux de contrôle. Il pourrait être bon de faire porter séparément ces corrections de poststratification sur chaque échantillon si on veut réduire le biais de couverture avant de procéder à l'estimation finale EDB pour un gain d'efficience. Ces valeurs de pondération servent de pondération d'entrée à la procédure EDB. Les 10 résultats d'enquête suivants forment le vecteur $\boldsymbol{z}$ de contrôles à valeur nulle : maladie multisymptomatique CDC (2 niveaux); stress posttraumatique (2 niveaux); indicateur de fatigue chronique (2 niveaux); échelle continue de déficit physique SF 36; échelle continue de dépression symptomatique de Hopkins; échelle de fatigue à 13 éléments de Chalder (2 niveaux); inconfort sexuel du partenaire (3 niveaux); tabagisme actuel (3 niveaux); alcoolisme actuel (5 niveaux); état matrimonial actuel (2 niveaux).

Pour obtenir chacun des ensembles de pondération des 294 répétitions EDB, nous avons calculé par le MEG des facteurs de correction EDB $a_{kA,DFC}$ et $a_{kB,DFC}$, que nous avons ensuite appliqués aux poids d'échantillonnage corrigés pour la non-réponse et la poststratification pour que les différences entre les 10 grands résultats par w_{KA} et w_{kB} soient nulles, tout en respectant les 17 totaux de contrôle. À l'aide d'une recherche en grille, nous avons établi qu'une constante d'échelle $\eta_A = 0{,}80$ minimisait la somme des variances des grandes variables de résultat. Par la constante d'échelle, nous avons calculé les valeurs de pondération EDB w_{KA} et w_{kB}, puis un facteur composé $\zeta_A = 0{,}82$ pour la détermination de w^*_k, qui est l'ensemble des valeurs finales de pondération EDB pour tout l'échantillon de répondants s_B.

5.1 Effets de pondération inégale

La combinaison d'un taux de sous-échantillonnage 1 sur 5 pour le suivi téléphonique et d'un taux de réponse de 55 % à l'enquête complémentaire a donné pour les répondants à l'enquête de suivi téléphonique une pondération d'analyse qui était approximativement décuple de celle des répondants à l'enquête postale. Ainsi, la réduction du biais par l'enquête de suivi a subi l'incidence négative de l'accroissement de la variance d'échantillonnage consécutif à l'augmentation de la variabilité de la pondération d'échantillonnage de l'échantillon composite. Au tableau 1, nous montrons que, avant l'EDB, la taille effective d'échantillon décroît significativement, passant de 1 672 à 535 lorsque les répondants à l'enquête de suivi sont inclus dans l'analyse. En d'autres termes, les variances des estimations relatives à l'échantillon combiné sont supérieures à celles de l'échantillon de la seule enquête postale. Après l'EDB, les tailles effectives d'échantillon pour l'échantillon global excèdent celles de l'échantillon de l'enquête postale pour tout grand domaine de déclaration.

Tableau 1. Comparaison des effets de pondération inégale (EPI) et des tailles effectives d'échantillon (Eff. n) avant et après l'étalonnage à double base (EDB)

Domaine	Répondants		Avant l'EDB				Après l'EDB			
			Enquête postale		Enquêtes postale et complémentaire		Enquête postale		Enquêtes postale et complémentaire	
	Enquête postale	Enquêtes postale et complémentaire	EPI	Eff. n	EPI	Eff. n	EPI	Eff. n	EPI	Eff. n
Ensemble	5 182	5 709	3,10	1 672	10,67	535	3,11	1 666	3,09	1 850
Militaires en service actif	3 214	3 566	2,62	1 227	9,01	396	2,63	1 223	2,62	1 362
Réservistes	1 968	2 143	2,32	848	7,50	286	2,33	845	2,33	919
Hommes	3 382	3 735	2,31	1 464	7,98	468	2,32	1 458	2,31	1 620
Femmes	1 800	1 974	1,76	1 021	6,51	303	1,77	1 014	1,79	1 100
Hommes en service actif	2 100	2 339	1,91	1 101	6,58	355	1,91	1 097	1,91	1 223
Femmes en service actif	1 114	1 227	1,72	649	6,60	186	1,73	645	1,75	700
Hommes de la réserve	1 282	1 396	1,93	666	6,26	223	1,93	664	1,94	721
Femmes de la réserve	686	747	1,79	383	5,81	129	1,80	381	1,82	411

5.2 Effets de l'étalonnage à double base (EDB) sur les estimations d'enquête

Au **tableau 2**, nous présentons les estimations d'enquête et les erreurs d'échantillonnage correspondantes avant et après l'étalonnage EDB. Nous livrons deux ensembles de résultats d'enquête. Le premier comprend les 10 grands résultats qui forment le vecteur z de contrôles à valeur nulle. Après l'EDB, la différence est nulle entre ces estimations pour les variables visées par w_{KA} et w_{kB}. Le second ensemble de résultats « autres » illustre les effets de la procédure EDB sur les résultats (variables y) qui n'entrent pas explicitement dans la procédure de calage. Pour ces résultats, l'estimateur EDB est l'estimateur composite w_{KA} et w_{kB} par le facteur $\zeta_A = 0{,}82$.

Tableau 2. Effets de la procédure EDB sur les résultats d'enquête

Résultats d'enquête	Avant l'EDB					Après l'EDB					
	Échantillon A		Échantillon B			Échantillon A		Échantillon B		EDB	
						w_{kA}		w_{kB}		w^*_k	
	Moyenne	ET	Moyenne	ET	Diff.	Moyenne	ET	Moyenne	ET	Moyenne	ET
Grands résultats (variables z) :											
% Maladie multisymptomatique	68	1,16	63,6	1,91	4,5 §	67,3	1,05	67,3	1,05	67,3	1,05
% Stress posttraumatique	7,6	0,63	8,5	1,17	-0,9	7,8	0,59	7,8	0,59	7,8	0,59
Indicateur de fatigue chronique	10,3	0,64	11,7	1,49	-1,3	10,5	0,65	10,5	0,65	10,5	0,65
Échelle de fatigue de Chalder	60,5	1,26	57,9	2,05	2,6	60,1	1,16	60,1	1,16	60,1	1,16
% Inconfort sexuel	10,1	0,77	9,5	1,44	0,5	10,1	0,75	10,1	0,75	10,1	0,75
% Tabagisme actuel	24,3	1,16	28,6	1,93	-4,3 §	25,0	1,05	25,0	1,05	25,0	1,05
% Alcoolisme actuel (mult. X/semaine)	31,3	1,20	32,1	2,05	-0,8	31,4	1,10	31,4	1,10	31,4	1,10
% Union de droit ou de fait	76,4	0,96	73,8	1,94	2,6 §	76,1	0,92	76,1	0,92	76,1	0,92
Sous-échelle de la dépression	1,7	0,02	1,7	0,03	0,0	1,7	0,01	1,7	0,01	1,7	0,01
Échelle de déficit physique	7,0	0,04	7,0	0,07	0,1	7,0	0,04	7,0	0,04	7,0	0,04
Autres résultats :											
% Santé générale = excellente	8,8	0,63	11,7	1,53	-3,0	8,8	0,60	11,3	1,50	9,2	0,61
% Santé générale = très bonne	29,4	1,14	27,1	1,74	2,3	29,5	1,11	26,8	1,77	29,0	1,05
% Santé générale = bonne	38,5	1,21	39,2	2,31	-0,7	38,3	1,20	40,7	2,18	38,7	1,18
% Santé générale = passable	20,0	1,13	18,4	1,69	1,5	20,0	1,09	17,9	1,50	19,7	1,05
% Santé générale = médiocre	3,3	0,42	3,5	1,07	-0,2	3,4	0,44	3,3	0,96	3,4	0,42
% Heures de travail réduites	16,5	0,92	18,6	1,70	-2,1	16,7	0,96	17,2	1,19	16,8	0,93
% Limitation des activités	22,9	1,14	24,5	2,09	-1,6 §	23,1	1,11	23,0	1,37	23,1	1,10
% Perte d'intérêt pour les AVQ	57,7	1,27	61,5	1,93	-3,8 §	57,7	1,19	61,3	1,73	58,4	1,18

§ Différence significative au niveau 0,05.

REMERCIEMENTS

Les travaux du premier des auteurs ont été financés en partie par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada à l'Université Carleton (Ottawa) dans le cadre d'une chaire de recherche à titre de professeur adjoint. Les auteurs remercient les D[rs] Lori Ebert et Elizabeth Federman, du RTI, de les avoir aidés à appliquer ce cadre méthodologique. Les données exploitées dans cet article viennent d'une enquête réalisée par RTI International pour le Medical Research and Material Command des forces armées américaines.

RÉFÉRENCES

Binder, D.A. (1996), "Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach", *Survey Methodology*, 22, pp. 17-22.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M., et W. Jocelyn (2000), "Variance Estimation for Two-Phase Stratified Sampling", *The Canadian Journal of Statistics*, Vol. 28, No. 3.

Cochran, W.G. (1977), *Sampling Techniques, Second Edition*, John Wiley & Sons, New York.

Deville, J.-C., et C.-E Sarndal (1992), "Calibration Estimation in Survey Sampling", *Journal of the American Statistical Association*, 87, pp. 376-382.

Fay, R.E. (1991), "A Design-Based Perspective on Missing Data Variance", In Proceedings of the Annual Research Conference, US Bureau of the Census, pp.429-440.

Folsom, R.E. Jr., et A.C. Singh (2000), "A Generalized Exponential Model for Sampling Weight Calibration for a Unified Approach to Nonresponse, Post-stratification, and Extreme Weight Adjustments", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 598-603.

Fuller, W.A. (1998), "Replication Variance Estimation for Two-Phase Samples", *Statistica Sinica*, 8, pp. 1153-1164.

Hajek, J. (1971), Comments on the paper "An Essay on the Logical Foundations of Survey Sampling, Part One by D.Basu", in *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott, Holt, Rinehart, and Winston of Canada, Ltd, Toronto.

Hansen, M. H., et W.N. Hurvitz (1946), "The Problem of Nonresponse in Surveys", *Journal of the American Statistical Association*, 41, pp. 517-529.

Kim, J.K., et R.R. Sitter (2003), "Efficient Replication Variance Estimation for Two-Phase Sampling", *Statistica Sinica*, 13, pp. 641-653.

Kish, L. (1965), *Survey Sampling*, John Wiley & Sons, New York.

Kott, P.S., et D.M. Stukel (1997), "Can the Jackknife Be Used With a Two-Phase Sample", *Survey Methodology*, 23:2, pp. 81-89.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, Duxbury Press, New York.

Potter, F. (1990), "A Study of Procedures to Identify and Trim Extreme Sampling Weights", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 225-230.

Rao, J.N.K. (1973), "On Double Sampling for Stratification and Analytical Surveys", *Biometrika*, 60:125-133, pp. 669.

Sarndal, C.E., Swensson, et Wretman. (1992), *Model-assisted survey sampling*, Springer-Verlag.

Shao, J., et P. Steel (1999), "Variance Estimation for Survey Data with Composite Imputation and Non-Negligible Sampling Fractions", *Journal of the American Statistical Association*, 94, pp. 254-265.

Singh, A.C., et C.-E. Sarndal (2003), "Optimal Regression, Generalized Regression, and Modified Regression", manuscript under preparation.

Singh, A.C., et R.E. Folsom (2000), "Bias Corrected Estimating Function Approach for Calibration Adjusted Variance Estimation", *Journal of the American Statistical Association, Section on Survey Research Methods*, pp. 610-615.

Singh, A.C., et S. Wu (2003), "An Extension of Generalized Regression Estimator to Dual-Frame Surveys", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, in print.

Singh, A.C., et S. Wu (1996), "Estimation for Multiframe Complex Surveys by Modified Regression", *Proceedings of the Statistical Society of Canada, Survey Methods Section*, pp. 69-77.