Nº 12-001-XIF au catalogue

Techniques d'enquête

Décembre 2004

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 12-001-XIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1 800 263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2004

Publication autorisée par le ministre responsable de Statistique Canada

Avril 2006

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

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

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

Vol. 30, N° 2, pp. 205-216
Statistique Canada, N° 12-001 au catalogue

Effets de plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage complexe

Inho Park et Hyunshik Lee [1]

Résumé

Nous examinons de nouveau la relation entre les effets de plan pour l'estimateur pondéré du total et l'estimateur pondéré de la moyenne sous échantillonnage complexe. Nous donnons des exemples sous diverses conditions. En outre, au moyen d'exemples, nous corrigeons certaines idées fausses concernant les effets de plan.

Mots clés : Échantillonnage aléatoire simple; échantillonnage ppt; échantillonnage à plusieurs degrés; autopondération; stratification a posteriori; coefficient de corrélation intragrappe.

1. Introduction

L'effet de plan est un paramètre dont l'usage est très répandu en échantillonnage pour élaborer le plan d'échantillonnage et indiquer l'effet de ce dernier sur l'estimation et l'analyse. On le définit comme étant le rapport entre la variance d'un estimateur sous échantillonnage complexe et celle d'un estimateur sous échantillonnage aléatoire simple pour la même taille d'échantillon. Les progiciels applicables aux enquêtes complexes, tels que WesVar et SUDAAN, produisent systématiquement une estimation de l'effet de plan. Au départ, on a défini ce dernier en vue de calculer l'estimateur (par le quotient) pondéré de la moyenne de population (Kish 1995). Cependant, en pratique, il est courant d'appliquer ce concept à d'autres statistiques, comme l'estimateur pondéré du total de population, souvent avec fruit, mais parfois avec confusion et en comprenant mal le problème. Cette dernière situation se présente surtout lorsqu'on applique des résultats simples, mais utiles, obtenus au moyen d'un plan d'échantillonnage assez simple, à des problèmes plus complexes. Dans le présent article, nous examinons la relation entre les effets de plan pour l'estimateur pondéré du total et l'estimateur pondéré de la moyenne sous divers plan d'échantillonnage complexes. À la section 2, nous passons brièvement en revue la définition de l'effet de plan et son utilisation pratique, et nous discutons de certaines idées fausses sur les effets de plan de sondage dans le calcul des estimateurs pondérés du total et de la moyenne. À la section 3, nous analysons la différence entre les effets de plan pour l'estimateur pondéré du total et pour celui de la moyenne sous un plan d'échantillonnage à deux degrés. Puis, à la section 4, nous discutons des effets de plan sous divers plans d'échantillonnage à deux degrés et certains cas plus généraux, et nous essayons de corriger certaines idées fausses au moyen de ces exemples. Enfin, à la section 5, nous résumons notre discussion.

2. Brève revue de la définition et de l'utilisation de l'effet de plan en pratique

Un précurseur de l'effet de plan popularisé par Kish (1965) a été utilisé par Cornfield (1951). Ce dernier a défini l'efficacité d'un plan d'échantillonnage complexe pour l'estimation d'une proportion de population comme étant le rapport de la variance de l'estimateur de la proportion sous échantillonnage aléatoire simple avec remise (easar) à la variance correspondante sous échantillonnage aléatoire simple en grappes pour la même taille d'échantillon. L'inverse du rapport défini par Cornfield (1951) a également été utilisé par d'autres auteurs. Ainsi, Hansen, Hurwitz et Madow (1953, volume I, pages 259 – 270) discutent de l'augmentation de la variance relative d'un estimateur par le quotient due à l'effet de mise en grappes de l'échantillonnage en grappes par rapport à l'échantillonnage aléatoire simple sans remise (eassr). Toutefois, l'expression effet de plan, ou Deff en abrégé, a été inventée et définie officiellement par Kish (1965, section 8.2, page 258) comme étant « le rapport de la variance réelle d'un échantillon à la variance d'un échantillon aléatoire simple contenant le même nombre d'éléments » [traduction] (pour un historique pour détaillé, voir aussi Kish 1995, page 73 et les références mentionnées dans ce texte).

Supposons que nous voulions estimer la moyenne de population $(\bar{Y})$ d'une variable y d'après un échantillon de taille m tiré selon un plan d'échantillonnage complexe représenté par p dans une population de taille M. Le Deff de Kish d'une estimation $(\bar{y}_p)$ est donné par

1. Inho Park et Hyunshik Lee, Westat, Inc. 1650 Research Blvd., Rockville, MD 20850, États-Unis. Courriel : InhoPark@westat.com.

$$\text{Deff} = \frac{V_p(\bar{y}_p)}{(1-f)\,S_y^2/m} \qquad (2.1)$$

où V_p est la variance par rapport à $p, f = m/M$ est la fraction d'échantillonnage globale et $S_y^2 = (M-1)^{-1} \sum_{k=1}^{M}(y_k - \bar{Y})^2$ est la variance d'élément de la population de la variable *y*. Bien qu'au départ, l'effet de plan ait été défini pour un estimateur de la moyenne de population (Kish 1995), on peut le définir pour toute statistique significative calculée d'après un échantillon tiré selon un plan d'échantillonnage complexe.

Le Deff est une quantité de population qui dépend de plan d'échantillonnage et se rapporte à une statistique particulière estimant un paramètre de population d'intérêt donné. Divers estimateurs permettant d'estimer un même paramètre ont des effets de plan différents, même sous des plans d'échantillonnage identiques. Par conséquent, l'effet de plan englobe non seulement l'efficacité de plan d'échantillonnage, mais aussi celle de l'estimateur. Särndal, Swensson et Wretman (1992, page 54) ont bien établi ce point en définissant l'effet de plan comme une fonction de plan de sondage (p) et de l'estimateur $(\hat{\theta})$ du paramètre de population $(\theta = \theta(y))$. Donc, nous pouvons l'écrire sous la forme

$$\text{Deff}_p(\hat{\theta}) = \frac{V_p(\hat{\theta})}{V_{\text{easar}}(\hat{\theta}')}$$

où $\hat{\theta}'$ est la forme habituelle d'un estimateur de θ sous eassr, qui est normalement différent de $\hat{\theta}$. Par exemple, pour estimer la moyenne de population, nous pourrions utiliser la moyenne (quotient) pondérée $\hat{\theta} = \sum_s w_k y_k / \sum_s w_k$ avec les poids d'échantillonnage w_k, tandis que $\hat{\theta}'$ serait la moyenne d'échantillon simple $\sum_s y_k / m$, où la sommation est faite sur l'échantillon *s*. Nous examinerons l'incidence d'estimateurs particuliers $\hat{\theta}$ sur l'effet de plan plus loin.

Plus tard, Kish (1995) a préconisé l'utilisation d'un paramètre défini de façon légèrement différente, qu'il a appelé Deft et dont le dénominateur est la variance sous échantillonnage aléatoire simple avec remise (easar), étant donné que l'échantillonnage sans remise fait partie de plan de sondage et devrait donc être reflété dans la définition. Kish a également argumenté que le paramètre Deft est plus facile à utiliser pour faire des inférences et qu'il vaut mieux définir l'effet de plan sans le facteur de correction pour population finie $(1-f)$, car ce dernier est difficile à calculer dans certaines situations. La nouvelle définition est donnée par

$$\text{Deft}_p(\hat{\theta}) = \sqrt{\frac{V_p(\hat{\theta})}{V_{\text{easar}}(\hat{\theta}')}}$$

ou $\text{Deft}_p^2(\hat{\theta}) = V_p(\hat{\theta}) / V_{\text{easar}}(\hat{\theta}')$. Les logiciels applicables à des données d'enquête, tels que WesVar et SUDAAN, produisent Deft^2 au lieu de Deff. Nous utiliserons cette définition dans le présent article.

Quand le paramètre de population est le total (Y), l'estimateur sans biais est le total pondéré d'échantillon, c'est-à-dire $\hat{Y} = \sum_s w_k y_k$. Si le paramètre d'intérêt est la moyenne de population, on l'estime habituellement par la moyenne pondérée, qui est $\hat{\bar{Y}} = \sum_s w_k y_k / \sum_s w_k$. Il s'agit d'un cas particulier de l'estimateur par le quotient, $\sum_s w_k y_k / \sum_s w_k x_k$, où $x_k \equiv 1$ pour tout $k \in s$.

Une idée fausse courante concernant les effets de plan pour $\hat{Y}$ et $\hat{\bar{Y}}$ est que leurs valeurs sont les mêmes. Or, on a constaté que l'effet de plan pour $\hat{Y}$, $\text{Deft}_p^2(\hat{Y})$, a tendance à être beaucoup plus grand que celui pour $\hat{\bar{Y}}$, $\text{Deft}_p^2(\hat{\bar{Y}})$. Cette différence a aussi été mentionnée, entre autres, par Kish (1987) et par Barron et Finch (1978). Une explication est donnée par Hansen et coll. (1953, volume I, pages 336 – 340) qui montrent que la différence a pour origine la variance relative des tailles de grappe. Plus récemment, Särndal et coll. (1992, pages 315 – 318) ont montré que, contrairement au cas de $\hat{\bar{Y}}$, l'effet de plan pour $\hat{Y}$ dépend de la variation (relative) de la variable *y*. En fait, même l'effet de plan pour $\hat{\bar{Y}}$ peut dépendre de la variation (relative) de la variable *y*, ce dont nous discuterons à la section 4. Cette dépendance est en contradiction avec ce que l'effet de plan est destiné à mesurer comme Kish (1995) l'a décrit explicitement :

> [Traduction] « Deft est utilisé pour exprimer les effets de plan d'échantillonnage au-delà de la variabilité élémentaire (S_y^2/m), en éliminant à la fois les paramètres de nuisance que sont les unités de mesure et la taille d'échantillon. Si l'on élimine S_y, les unités et la taille d'échantillon m, les effets de plan sur les erreurs d'échantillonnage deviennent généralisables (transférables) à d'autres statistiques et à d'autres variables, dans la même enquête, voire même dans d'autres. »

Sa déclaration peut être vaguement correcte pour la moyenne pondérée $\hat{\bar{Y}}$ telle qu'exprimée dans la formule approximative d'échantillon fréquemment utilisée de $\text{Deft}_p^2(p, \hat{\bar{Y}})$ donnée par Kish (1987) :

$$\text{Deft}_p^2(\hat{\bar{Y}}) = \{1 + \rho(\bar{m} - 1)\}\left(1 + \text{cv}_w^2\right) \qquad (2.2)$$

où le plan d'échantillonnage p contient des caractéristiques complexes, telles qu'une pondération inégale et un échantillonnage en grappes, $\rho = \rho_p(y)$ est le coefficient de corrélation intraclasse (souvent appelé mesure de l'homogénéité à l'intérieur des grappes), $\bar{m}$ est la taille

moyenne de l'échantillon de grappes et cv_w^2 est la variance d'échantillon relative des poids. Strictement parlant, cette formule n'est pas indépendante de la variable *y*, parce que ρ dépend de cette variable. En outre, l'effet de plan pourrait ne pas être exempt de l'unité de mesure, à moins que $V_p(\hat{\bar{Y}})$ soit exprimé sous une forme factorielle de S_y^2/m. Voir Park et Lee (2002). La formule (2.2) n'est valide que s'il n'existe aucune corrélation entre les poids d'échantillonnage et la variable étudiée *y*. Par contre, s'il existe une corrélation, il peut être nécessaire de modifier la formule conformément aux études de Spencer (2000) et de Park et Lee (2001). À la section suivante, nous exposons cet aspect en détail pour l'échantillonnage à deux degrés et nous examinons aussi ce point plus en profondeur à la section 4.1.

3. Décomposition de l'effet de plan sous échantillonnage à deux degrés

Considérons un plan d'échantillonnage réalisé en deux degrés. Supposons qu'une population $U = \{k : k = 1, \ldots, M\}$ avec *M* éléments soit regroupée en *N* grappes de taille M_i, telle que $M = \sum_{i=1}^N M_i$. L'échantillon de premier degré $s_a = \{i : i = 1, \ldots, n\}$ de *n* grappes (unités primaire d'échantillonnage ou UPE en abrégé) est tiré avec remise à partir de *N* grappes avec probabilité p_i, où $\sum_{i=1}^N p_i = 1$. Soit $p_a = \Pr(s_a)$ le plan d'échantillonnage de premier degré. L'échantillon de deuxième degré $s_{bi} = \{j : j = 1, \ldots, m_i\}$ de m_i éléments (unités secondaires d'échantillonnage ou USE) est alors tiré indépendamment à partir de chaque UPE *i* sélectionnée à la première étape selon un plan d'échantillonnage arbitraire, disons $p_{bi} = \Pr(s_{bi} | s_a)$, où $i \in s_a$. Représentons l'échantillon total d'éléments et le plan d'échantillonnage global par $s = \cup_{i \in s_a} s_{bi}$ et $p = \Pr(s)$, respectivement. Au j^{e} élément de la i^{e} grappe est associée une caractéristique d'enquête $y_{ij}, j = 1, \ldots, M_i, i = 1, \ldots, N$. Pour un $i \in s_a$, donné, représentons par $w_{j|i}$ les poids d'échantillonnage de deuxième degré tels qu'un estimateur de la forme $\hat{Y}_i = \sum_{j=1}^{m_i} w_{j|i}\, y_{ij}$ soit sans biais pour le total de grappe $Y_i = \sum_{j=1}^{M_i} y_{ij}$, c'est-à-dire $E_b(\hat{Y}_i) = Y_i$, où E_b représente l'espérance par rapport à l'échantillonnage de deuxième degré. Soit $w_i = 1/(np_i)$ les poids d'échantillonnage de premier degré et soit $Y = \sum_{i=1}^N Y_i$ le total de population. Il est facile de montrer que $E_a(Y_i/p_i) = Y$. Supposons que Y_i est connu pour $i \in s_a$, $\sum_{i=1}^n w_i Y_i$ est la moyenne des *n* estimateurs sans biais de *Y* de sorte que $E_a(\sum_{i=1}^n w_i Y_i) = Y$, où E_a représente l'espérance par rapport au plan d'échantillonnage de premier degré. Notons qu'aux deux étapes, l'échantillonnage a lieu avec remise. Par conséquent, il se peut que la même unité d'échantillonnage (une grappe ou un élément) soit sélectionnée plus d'une fois, mais soit traitée différemment. Définissons les poids d'échantillonnage globaux par $w_{ij} = w_i\, w_{j|i}$. Manifestement, $\hat{Y} = \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij}\, y_{ij}$ est sans biais pour *Y*, autrement dit $E_p(\hat{Y}) = E_a\, E_b(\hat{Y}) = E_a(\sum_{i=1}^n w_i Y_i) = Y$, où E_p représente l'espérance par rapport à *p*. La variance de $\hat{Y}$ peut s'écrire sous la forme

$$\begin{aligned} V_p(\hat{Y}) &= V_a E_b(\hat{Y}) + E_a V_b(\hat{Y}) \\ &= \sum_{i=1}^N w_i (Y_i - p_i Y)^2 + \sum_{i=1}^N w_i V_b(\hat{Y}_i) \end{aligned} \qquad (3.1)$$

où V_p, V_a et V_b représentent les variances définies par rapport aux plans d'échantillonnage global, de premier degré et de deuxième degré. Voir Särndal et coll. (1992, pages 151 – 152).

Un estimateur de la moyenne de population $\bar{Y} = Y/M$ utilisé fréquemment est l'estimateur (par le quotient) pondéré donné par $\hat{\bar{Y}} = \hat{Y}/\hat{M}$, où $\hat{M} = \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij}$. En recourant à la linéarisation de Taylor, telle que décrite dans Särndal et coll. (1992, pages 176 – 178), nous pouvons approximer $\hat{\bar{Y}}$ par

$$\hat{\bar{Y}} \cong \bar{Y} + M^{-1}\hat{D} \qquad (3.2)$$

où $\hat{D} = \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij}\, d_{ij}$ est un estimateur sans biais du total de population $D = \sum_{i=1}^N \sum_{j=1}^{M_i} d_{ij}$ de $d_{ij} = y_{ij} - \bar{Y}$, qui représente l'écart de y_{ij} par rapport à la moyenne de population $\bar{Y}$. Notons que $D = 0$. En écrivant $D_i = \sum_{j=1}^{M_i} d_{ij} = Y_i - M_i \bar{Y}$ et $\hat{D}_i = \sum_{j=1}^{m_i} w_{j|i}\, d_{ij}$, nous obtenons la variance approximative de $\hat{\bar{Y}}$ pour l'expression (3.2) sous la forme

$$\mathrm{AV}_p(\hat{\bar{Y}}) = \frac{1}{M^2}\left[\sum_{i=1}^N w_i \left(Y_i - \frac{M_i}{M} Y\right)^2 + \sum_{i=1}^N w_i\, V_b(\hat{D}_i)\right]. \qquad (3.3)$$

Si nous tirons un échantillon aléatoire simple de taille $m = \sum_{i=1}^n m_i$ avec remise à partir de la population *U*, alors une moyenne d'échantillon $\bar{y}_{\mathrm{srs}} = \sum_s y_k / m$ et son développement

$$\hat{Y}_{\mathrm{eas}} = M\, \bar{y}_{\mathrm{eas}} = \frac{1}{f} \sum_s y_k \qquad (3.4)$$

serviraient d'estimateurs de la moyenne $\bar{Y}$ et du total *Y* de population, respectivement, sous easar, où $f = m/M$ est la fraction d'échantillonnage globale. Leur variance sous ce plan d'échantillonnage est donnée par $V_{\mathrm{easar}}(\hat{Y}_{\mathrm{eas}}) = M^2 V_{\mathrm{easar}}(\bar{y}_{\mathrm{eas}})$, où $V_{\mathrm{easar}}(\bar{y}_{\mathrm{eas}}) = m^{-1}\, S_y^2$ et $S_y^2 = (M-1)^{-1} \sum_s (y_k - \bar{Y})^2$. Nous notons que *m* est la taille d'échantillon obtenue, qui, en général, est une quantité aléatoire. À partir de (3.1), de (3.3) et des expressions susmentionnées, en remplaçant *m* par sa valeur prévue m' par rapport au plan d'échantillonnage global *p*, c'est-à-dire $m' = E_p(m)$, nous pouvons représenter les effets de plan pour $\hat{Y}$ et $\hat{\bar{Y}}$ par les expressions

$$\mathrm{Deft}_p^2(\hat{Y}) = \frac{m'}{\mathrm{CV}_y^2}\left\{\sum_{i=1}^{N} w_i\left(\frac{Y_i}{Y} - p_i\right)^2 + \sum_{i=1}^{N} w_i V_b\left(\frac{\hat{Y}_i}{Y}\right)\right\} \quad (3.5)$$

et

$$\mathrm{Deft}_p^2(\hat{\bar{Y}}) \cong \frac{m'}{\mathrm{CV}_y^2}\left\{\sum_{i=1}^{N} w_i\left(\frac{Y_i}{Y} - \frac{M_i}{M}\right)^2 + \sum_{i=1}^{N} w_i V_b\left(\frac{\hat{D}_i}{Y}\right)\right\} \quad (3.6)$$

où $\mathrm{CV}_y^2 = S_y^2/\bar{Y}^2$ représente la variance de population relative de la variable y. Partant de ces expressions, nous pouvons écrire la différence entre les effets de plan pour $\hat{Y}$ et $\hat{\bar{Y}}$ de la façon suivante.

$$\mathrm{Deft}_p^2(\hat{Y}) - \mathrm{Deft}_p^2(\hat{\bar{Y}}) \cong \Delta_a + \Delta_b, \quad (3.7)$$

où

$$\Delta_a = \frac{m'}{\mathrm{CV}_y^2}\left\{\sum_{i=1}^{N} w_i\left[\left(\frac{Y_i}{Y} - p_i\right)^2 - \left(\frac{Y_i}{Y} - \frac{M_i}{M}\right)^2\right]\right\}$$

et

$$\Delta_b = \frac{m'}{\mathrm{CV}_y^2}\left\{\sum_{i=1}^{N} w_i\left[V_b\left(\frac{\hat{Y}_i}{Y}\right) - V_b\left(\frac{\hat{D}_i}{Y}\right)\right]\right\}.$$

Dans l'expression (3.7), deux composantes Δ_a et Δ_b reflètent les différences dues aux sources de variation provenant, respectivement, des premier et deuxième degrés d'échantillonnage. Naturellement, la deuxième composante disparaît si tous les éléments compris dans les grappes sélectionnées sont observés, puisque le plan devient un plan à un seul degré ou qu'un échantillon aléatoire simple est sélectionné au deuxième degré. Il en est ainsi parce que les deux variances $V_b(\hat{Y}_i)$ et $V_b(\hat{D}_i)$ sont équivalentes sous les conditions susmentionnées, c'est-à-dire 1) $V_b(\hat{Y}_i) = V_b(\hat{D}_i) = 0$ si $w_{j|i} = 1$ pour tous i et j, et 2) $V_b(\hat{Y}_i) = V_b(\hat{D}_i) \geq 0$ si $w_{j|i} = M_i/m_i$ pour tous i et j. Autrement dit,

$$\Delta_b = 0 \quad \text{si} \quad w_{j|i} = c_i \text{ pour tous } i \text{ et } j, \quad (3.8)$$

où c_i sont des constantes non négatives et pas forcément égales pour les diverses grappes. Parallèlement, nous pouvons montrer que

$$\Delta_a = \begin{cases} 0 & \text{si} \quad p_i \propto M_i, \\ A_p(y) & \text{si} \quad Y_i \propto M_i, \\ -A_p(y) & \text{si} \quad p_i \propto Y_i, \end{cases} \quad (3.9)$$

pour tout i, où $A_p(y) = (m'/\mathrm{CV}_y^2)\sum_{i=1}^{N} w_i(p_i - M_i/M)^2$. Notons que $A_p(y)$ est une quantité non négative et que les conditions de l'expression (3.9) peuvent être réénoncées, respectivement, sous la forme $p_i = M_i/M, \bar{Y}_i = \bar{Y}$ et $p_i = Y_i/Y$, où $\bar{Y}_i = Y_i/M_i$ pour tout $i = 1, \ldots, N$. Ces résultats révèlent l'effet de l'échantillonnage en grappes sur la précision des deux estimateurs. Par exemple, si $p_i = M_i/M$, l'échantillonnage en grappes n'a pas d'effet sur cette précision. Par contre, si $p_i = Y_i/Y$, l'échantillonnage en grappes rend $\hat{Y}$ plus efficace que $\hat{\bar{Y}}$ en ce qui a trait à la précision, mais il rend $\hat{\bar{Y}}$ plus précis que $\hat{Y}$ si $\bar{Y}_i = \bar{Y}$ pour tout i.

Maintenant, examinons certains exemples des conditions (3.8) et (3.9).

Exemple 3.1 Pour un plan d'échantillonnage en grappes à un ou à deux degrés avec échantillonnage en grappes ppt en utilisant $p_i = M_i/M$ et $w_{j|i} = c_i$ pour tout $i = 1, \ldots, N$, nous obtenons, d'après (3.8) et (3.9), $\Delta_a = \Delta_b = 0$, c'est-à-dire qu'il n'y a aucune différence entre les effets de plan pour $\hat{\bar{Y}}$ et $\hat{Y}$.

Nous pouvons obtenir le même résultat que celui de l'exemple 3.1 au moyen de $\hat{Y} = M\bar{Y}$. Il s'agit de l'estimateur par le quotient, qu'on peut utiliser si l'on connaît M. La situation où les poids d'échantillonnage globaux sont constants pour tous les éléments (c'est-à-dire plan d'échantillonnage autopondéré) est un cas particulier bien connu. Nous y reviendrons à la section 4.

Exemple 3.2 Plan d'échantillonnage aléatoire simple en grappes à un degré ou plan d'échantillonnage à deux degrés avec eas aux deux étapes. Sous ces plans d'échantillonnage, nous avons $w_{j|i} = c_i$ et $p_i = 1/N$ pour tous i et j, et il découle donc de (3.8) et de (3.9) que $\Delta_b = 0$ et

$$\Delta_a = \begin{cases} 0 & \text{si} \quad M_i \equiv M_0 \text{ pour tout } i, \\ \bar{m}'\dfrac{\mathrm{CV}_M^2}{\mathrm{CV}_y^2} & \text{si les} \quad \bar{Y}_i \text{ sont tous égaux}, \\ -\bar{m}'\dfrac{\mathrm{CV}_M^2}{\mathrm{CV}_y^2} & \text{si les} \quad Y_i \text{ sont tous égaux}, \end{cases} \quad (3.10)$$

où $\bar{m}' = m'/n$, $\mathrm{CV}_M^2 = \bar{M}^{-2}\sum_{i=1}^{N}(M_i - \bar{M})^2/N$ représente la variance relative des tailles de grappe M_i et $\bar{M} = M/N$ représente la taille moyenne des grappes. Les conditions de (3.10) satisfont aussi aux conditions de (3.9) et, par conséquent, (3.10) est un cas particulier de (3.9). Notons que la quantité $A_p(y)$ dans l'expression (3.9) se réduit approximativement à $\bar{m}' \cdot \mathrm{CV}_M^2/\mathrm{CV}_y^2$ quand $p_i = 1/N$ pour tout i.

L'exemple 3.2 montre que, si l'inégalité des tailles de grappe n'est pas reflétée dans le plan d'échantillonnage, l'efficacité relative de $\hat{Y}$ par rapport à $\hat{\bar{Y}}$ dépend partiellement de la variabilité relative de ces tailles de grappe. Si les moyennes de grappe sont toutes égales, alors l'échantillonnage en grappes rend $\hat{\bar{Y}}$ plus efficace que $\hat{Y}$, et inversement, si tous les totaux de grappe sont égaux. Par ailleurs, si toutes les grappes sont de même taille, l'échantillonnage

aléatoire simple des grappes ne produit aucune différence entre les effets de plan.

À la section 4, nous utilisons les résultats obtenus à la présente section pour discuter d'autres exemples utilisés dans la littérature sur l'échantillonnage.

4. Exemples de l'effet de plan dans la littérature sur l'échantillonnage

4.1 Échantillonnage d'éléments avec probabilités inégales

Considérons l'échantillonnage d'éléments avec probabilités inégales sans mise en grappes. La discussion de la section 3 s'applique à cet exemple avec $M_i \equiv 1$ pour tout $i = 1, \ldots, N$ et, donc, $m = n$. Par souci de concision, nous utilisons y_i pour représenter la valeur de la variable y et nous supposons que N est grand de sorte que $N/(N-1) \cong 1$. En l'absence de la variation d'échantillonnage de deuxième degré, les effets de plan pour $\hat{Y}$ et $\hat{\overline{Y}}$ donnés par les expressions (3.5) et (3.6) se réduisent à

$$\text{Deft}_p^2(\hat{Y}) \cong \frac{\sum_{i=1}^{N} p_i^{-1}(y_i - p_i Y)^2}{\sum_{i=1}^{N} N(y_i - \overline{Y})^2} \tag{4.1}$$

et

$$\text{Deft}_p^2(\hat{\overline{Y}}) \cong \frac{\sum_{i=1}^{N} p_i^{-1}(y_i - \overline{Y})^2}{\sum_{i=1}^{N} N(y_i - \overline{Y})^2}. \tag{4.2}$$

En outre, considérons un exemple où la variable étudiée y n'est pas corrélée à la probabilité de sélection p_i.

Exemple 4.1 Échantillonnage d'éléments avec probabilités inégales sans corrélation entre y_i et p_i. Si y_i et p_i ne sont pas corrélées, nous pouvons approximer $\sum_{i=1}^{N} p_i^{-1}(y_i - \overline{Y})^2$ par $n\overline{W}\sum_{i=1}^{N}(y_i - \overline{Y})^2$, où $\overline{W} = N^{-1}\sum_{i=1}^{N} w_i$. Notons que $E_p(n^{-1}\sum_{i=1}^{n} w_i) = N/n$, $E_p(n^{-1}\sum_{i=1}^{n} w_i^2) = N\overline{W}/n$ et $E_p(n^{-1}\sum_{i=1}^{n} w_i^2)/E_p^2(n^{-1}\sum_{i=1}^{n} w_i) = n\overline{W}/N$. Donc,

$$\begin{aligned}\text{Deft}_p^2(\hat{\overline{Y}}) &\cong n\overline{W}/N \\ &= E_p\left(n^{-1}\sum_{i=1}^{n} w_i^2\right) \Big/ E_p^2\left(n^{-1}\sum_{i=1}^{n} w_i\right). \end{aligned} \tag{4.3}$$

Il est facile de montrer que $n\overline{W}/N \geq 1$ en utilisant l'inégalité de Cauchy-Schwarz (Apostol 1974, page 14). En outre, des calculs ordinaires montrent, en partant de (4.1) et (4.2), que

$$\begin{aligned}&\text{Deft}_p^2(\hat{Y}) - \text{Deft}_p^2(\hat{\overline{Y}}) \\ &\cong \text{CV}_y^{-2}\left\{\sum_{i=1}^{N} p_i^{-1}(p_i - \overline{p})^2 - 2Y^{-1}\sum_{i=1}^{N} p_i(y_i - \overline{Y})(p_i - \overline{p})\right\} \\ &= \text{CV}_y^{-2}(n\overline{W}/N - 1),\end{aligned}$$

où $\overline{p} = N^{-1}\sum_{i=1}^{N} p_i = 1/N$. Nous obtenons la dernière expression à partir de $\sum_{i=1}^{N} p_i^{-1}(p_i - \overline{p})^2 = n\overline{W}/N - 1$ et $\sum_{i=1}^{N} p_i^{-1}(y_i - \overline{Y})(p_i - \overline{p}) \cong 0$ parce que y_i et p_i ne sont pas corrélées. Par conséquent,

$$\text{Deft}_p^2(\hat{Y}) - \text{Deft}_p^2(\hat{\overline{Y}}) \cong \text{CV}_y^{-2}\left\{\text{Deft}_p^2(\hat{\overline{Y}}) - 1\right\}$$

ou

$$\text{Deft}_p^2(\hat{Y}) \cong (1 + \text{CV}_y^{-2})\,\text{Deft}_p^2(\hat{\overline{Y}}) - \text{CV}_y^{-2}. \tag{4.4}$$

D'après (4.4), il est clair que $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\overline{Y}})$ si $\text{Deft}_p^2(\hat{\overline{Y}}) \geq 1$ et que l'égalité tient si $\text{Deft}_p^2(\hat{\overline{Y}}) = 1$ ou $\overline{W} = N/n$. En outre, $\text{Deft}_p^2(\hat{Y}) < \text{Deft}_p^2(\hat{\overline{Y}})$ si $1/(1 + \text{CV}_y^2) < \text{Deft}_p^2(\hat{\overline{Y}}) < 1$.

L'exemple 4.1 montre que l'effet de plan a tendance à être plus grand pour $\hat{Y}$ que pour $\hat{\overline{Y}}$ si la corrélation entre y_i et p_i est faible et que $\text{Deft}_p^2(\hat{\overline{Y}}) \geq 1$.

La quantification habituelle de l'effet des poids inégaux sur l'efficacité de plan donnée par (2.2) est due à Kish (1965, 11.7). Celui-ci a considéré les cas où les poids inégaux ont une origine « aléatoire », comme des problèmes de base de sondage ou des redressements pour la non-réponse. En supposant que 1) un échantillon aléatoire de taille n tiré avec remise est subdivisé en G classes de pondération telles que le même poids w_g soit attribué à n_g unités d'échantillonnage dans la classe g et que $n = \sum_{g=1}^{G} n_g$, et que 2) les G variances de classe de pondération sont égales à la variance unitaire de y, c'est-à-dire $S_{yg}^2 = S_y^2$ pour tout $g = 1, \ldots, G$, il a proposé une quantité donnée par

$$\text{Deft}_{\text{Kish}}^2(\hat{\overline{Y}}) = n\sum_{g=1}^{G} n_g w_g^2 \Big/ \left(\sum_{g=1}^{G} n_g w_g\right)^2, \tag{4.5}$$

pour mesurer l'augmentation de la variance de $\hat{\overline{Y}}$ comparativement à la variance hypothétique sous easar de taille n. La logique du calcul susmentionné est que la perte de précision de $\hat{\overline{Y}}$ due à une pondération aléatoirement inégale peut être approximée par le rapport de la variance sous échantillonnage stratifié disproportionné à celle sous l'échantillonnage stratifié proportionné.

Plus tard, en posant que, dans (4.5), $n_g = 1$ pour tout g et, donc, $n = G$, Kish (1992) a proposé une formule approximative bien connue donnée par

$$\text{Deft}_{\text{Kish}}^2(\hat{\overline{Y}}) = n\sum_{i=1}^{n} w_i^2 \Big/ \left(\sum_{i=1}^{n} w_i\right)^2 = 1 + \text{cv}_w^2, \tag{4.6}$$

où $\text{cv}_w^2 = n^{-1}\sum_{i=1}^n (w_i - \overline{w})^2 / \overline{w}^2$ est la variance d'échantillon relative et $\overline{w}$ est la moyenne d'échantillon de w_i. Notons que (4.6) est une approximation sur échantillon de (4.3). Dans le cas d'un plan d'échantillonnage inefficace pour l'estimation de *Y*, l'inefficacité diminue avec l'estimation par le quotient. Considérons maintenant le cas opposé d'une corrélation de la variable *y* et de la probabilité de sélection p_i, où l'efficacité de $\hat{Y}$ augmente.

Exemple 4.2 Échantillonnage d'éléments avec probabilités inégales où y_i est corrélée à p_i. Supposons que y_i est reliée linéairement à p_i par $y_i = A + Bp_i + e_i$, où A et B sont les coefficients de régression par les moindres carrés du modèle pour la population (finie) et e_i est le résidu correspondant. En outre, supposons que le modèle de régression soit bien ajusté aux données de population et que la variance de l'erreur soit à peu près homogène, de sorte que $R_{ew} \cong 0$ et $R_{e^2w} \cong 0$, où R_{ew} et R_{e^2w} représentent les corrélations de population des paires (e_i, w_i) et (e_i^2, w_i), respectivement. Par exemple, $R_{ew} = \sum_{i=1}^N (e_i - \overline{E})(w_i - \overline{W})/\{(N-1)S_e S_w\}$, où $\overline{E} = \sum_{i=1}^N e_i / N$, S_e et S_w sont les écarts-types de population de e_i et w_i, respectivement. Alors, les effets de plan donné par (4.1) et (4.2) se réduisent à

$$\text{Deft}_p^2(\hat{Y}) \cong (n\overline{W}/N)(1 - R_{yp}^2) + (n\overline{W}/N - 1)\left(\frac{R_{yp}}{\text{CV}_p} - \frac{1}{\text{CV}_y}\right)^2 \qquad (4.7)$$

et

$$\text{Deft}_p^2(\hat{\overline{Y}}) \cong (n\overline{W}/N)(1 - R_{yp}^2) + (n\overline{W}/N - 1)\left(\frac{R_{yp}}{\text{CV}_p}\right)^2, \qquad (4.8)$$

respectivement, où R_{yp} est la corrélation de population entre y_i et p_i et CV_p est le coefficient de variation de population de p_i (voir Park et Lee (2001), pour la preuve). Il découle de (4.7) et (4.8) que $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\overline{Y}})$ si, et uniquement si

$$2R_{yp} \leq \text{CV}_p / \text{CV}_y, \qquad (4.9)$$

où l'égalité tient si, et uniquement si, $2R_{yp} = \text{CV}_p / \text{CV}_y$. De surcroît, l'inégalité est inversée si l'inégalité (4.9) devient opposée.

La condition (4.9) indique que $\hat{Y}$ a tendance à être moins efficace que $\hat{\overline{Y}}$ en ce qui concerne la précision quand R_{yp} est petit. Donc, nous voyons que R_{yp} est un déterminant important de l'effet de plan d'échantillonnage avec probabilités inégales sur $\hat{Y}$ et $\hat{\overline{Y}}$ et leur efficacité relative.

En s'efforçant d'élaborer une expression approximative de l'effet de plan quand y_i est corrélée à p_i, Spencer (2000) a proposé une formule approximative d'échantillon pour $\hat{Y}$ et l'a comparée à la formule approximative (4.6) de Kish pour le cas particulier où $R_{yp} = 0$. Comme le montre l'exemple 4.2, les deux effets de plan (4.7) et (4.8) ne sont pas égaux à moins que $\overline{W} = N/n$ (voir Park et Lee (2001) pour une discussion plus approfondie et certains exemples numérique). En outre, ce cas particulier donne la même condition que pour l'exemple 4.1 et, par conséquent, les deux formules approximatives de l'effet de plan (4.7) et (4.8) sont équivalentes à (4.4) et (4.3), respectivement.

4.2 Échantillonnage en grappes à un degré

Considérons un échantillonnage en grappes à un degré, où chaque élément compris dans une grappe échantillonnée est inclus dans l'échantillon, c'est-à-dire $m_i \equiv M_i$ pour tout $i \in s_a$. Étant donné l'absence de la variation d'échantillonnage de deuxième degré, la variance de $\hat{Y}$ correspond uniquement au premier terme de l'expression (3.1) et peut être décomposée comme suit

$$\sum_{i=1}^N w_i (Y_i - p_i Y)^2 = \frac{M(N-1)}{n} S_{yB}^2 + \sum_{i=1}^N w_i Q_i \overline{Y}_i^2, \qquad (4.10)$$

où $S_{yB}^2 = (N-1)^{-1}\sum_{i=1}^N M_i (\overline{Y}_i - \overline{Y})^2$ et $Q_i = M_i(M_i - p_i M)$ pour $i = 1, \ldots, N$. Notons que $Q_i = 0$ si $p_i = M_i / M$, c'est-à-dire que p_i est proportionnel à la taille de grappe M_i. Notons aussi que S_{yB}^2 est la moyenne quadratique des écarts entre grappes dans une analyse de variance. En représentant la moyenne quadratique des écarts à l'intérieur des grappes par $S_{yW}^2 = (M-N)^{-1}\sum_{i=1}^N \sum_{j=1}^{M_i} (y_{ij} - \overline{Y}_i)^2$, écrivons $S_{yB}^2 = S_y^2\{1 + \delta(M-N)/(N-1)\}$ avec $\delta = 1 - S_{yW}^2 / S_y^2$. Puisque la taille d'échantillon prévue est $m' = n\overline{M}$, l'effet de plan pour $\hat{Y}$ peut s'écrire, en partant de (4.10), sous la forme

$$\text{Deft}_p^2(\hat{Y}) = \left(\frac{N-1}{N}\right)\left(1 + \frac{M-N}{N-1}\delta\right) + \frac{n\overline{M}}{\text{CV}_y^2}\sum_{i=1}^N \frac{w_i Q_i}{M_i^2}\left(\frac{Y_i}{Y}\right)^2. \qquad (4.11)$$

De la même façon, l'effet de plan pour $\hat{\overline{Y}}$ peut être exprimé par

$$\text{Deft}_p^2(\hat{\overline{Y}}) \cong \left(\frac{N-1}{N}\right)\left(1 + \frac{M-N}{N-1}\delta\right) + \frac{n\overline{M}}{\text{CV}_y^2}\sum_{i=1}^N \frac{w_i Q_i}{M_i^2}\left(\frac{D_i}{Y}\right)^2. \qquad (4.12)$$

Nous observons que l'effet de plan pour $\hat{\bar{Y}}$ diffère de celui pour $\hat{Y}$ dans le deuxième terme contenant $D_i = \sum_{i=1}^{M_i}(y_{ij} - \bar{Y})$ au lieu de Y_i. En outre, notons que la quantité $\delta = \delta_p(y)$ est le coefficient de détermination ajusté (R^2_{adj}) dans le contexte de l'analyse par régression. On peut le considérer comme une mesure d'homogénéité. Pour une discussion plus approfondie de δ, voir Särndal et coll. (1992, pages 130-131) et Lohr (1999, page 140).

Exemple 4.3 Échantillonnage aléatoire simple à un degré de grappes. Dans cet exemple, si $p_i = 1/N$ pour tout $i = 1, \ldots, N$, les deux effets de plan de sondage donnés par (4.11) et (4.12) se réduisent, respectivement, à

$$\text{Deft}_p^2(\hat{Y}) = \left(\frac{N-1}{N}\right)\left(1 + \frac{M-N}{N-1}\delta\right) + \frac{1}{N \cdot \text{CV}_y^2}\sum_{i=1}^{N}(M_i - \bar{M})\left(\frac{M_i}{\bar{M}}\right)\left(\frac{\bar{Y}_i}{\bar{Y}}\right)^2 \quad (4.13)$$

et

$$\text{Deft}_p^2(\hat{\bar{Y}}) \cong \left(\frac{N-1}{N}\right)\left(1 + \frac{M-N}{N-1}\delta\right) + \frac{1}{N \cdot \text{CV}_y^2}\sum_{i=1}^{N}(M_i - \bar{M})\left(\frac{M_i}{\bar{M}}\right)\left(\frac{\bar{D}_i}{\bar{Y}}\right)^2, \quad (4.14)$$

où $\bar{M} = M/N$. Puisque $\text{Deft}_p^2(\hat{Y}) - \text{Deft}_p^2(\hat{\bar{Y}}) \propto \sum_{i=1}^{N} M_i(M_i - \bar{M})(2\bar{Y}_i - \bar{Y})$, l'inégalité entre les effets de plan pour $\hat{Y}$ et $\hat{\bar{Y}}$ dépend de la loi conjointe de $\bar{Y}_i$ et M_i.

Exemple 4.4 Échantillonnage aléatoire simple à un degré de grappes de même taille. Dans ce cas-ci, nous avons $M_i \equiv M_0$ et $p_i = 1/N$ pour tout $i = 1, \ldots, N$ et nous pouvons approximer les deux effets de plan donnés par (4.13) et (4.14) par la même quantité donnée par

$$\left(\frac{N-1}{N}\right)\left[1 + \frac{N(M_0 - 1)}{N-1}\delta\right], \quad (4.15a)$$

puisque $M_i - \bar{M} = 0$ pour tout $i = 1, \ldots, N$.

Pour tenir compte de l'effet de la mise en grappes sur l'estimation de la variance, on utilise souvent la forme la plus simple d'échantillonnage aléatoire simple en grappes à un degré comme dans l'exemple 4.4. Consulter, par exemple, Cochran (1977, section 9.4), Lehtonen et Pahkinen (1995, page 91) et Lohr (1999, section 5.2.2). Bien que ces auteurs aient adopté un plan d'échantillonnage sans remise, par souci de simplicité et de cohérence, nous comparons leurs formules à notre formule avec hypothèse d'échantillonnage avec remise. De surcroît, la comparaison est valide parce que leurs formules intègrent la correction pour population finie au numérateur ainsi qu'au dénominateur, si bien que cet effet s'annule essentiellement. Cochran (1977, section 9.4) obtient la formule

$$\text{Deft}_p^2(\hat{\bar{Y}}) = \frac{NM_0 - 1}{M_0(N-1)}\left[1 + (M_0 - 1)\rho\right] \cong 1 + (M_0 - 1)\rho, \quad (4.15b)$$

où ρ est le coefficient de corrélation intragrappe défini par

$$\rho = \frac{2\sum_{i=1}^{N}\sum_{j>k=1}^{M_0}(y_{ij} - \bar{Y})(y_{ij} - \bar{Y})}{(M_0 - 1)\sum_{i=1}^{N}\sum_{j=1}^{M_0}(y_{ij} - \bar{Y})^2}. \quad (4.15c)$$

En réécrivant $\sum_{i=1}^{N}[\sum_{j=1}^{M_0}(y_{ij} - \bar{Y})]^2 = M_0(N-1)S_{yB}^2$ et $\sum_{i=1}^{N}\sum_{j=1}^{M_0}(y_{ij} - \bar{Y})^2 = (NM_0 - 1)S_y^2 = (N-1)S_{yB}^2 + N(M_0 - 1)S_{yW}^2$, il est facile de montrer que

$$\begin{aligned} &2\sum_{i=1}^{N}\sum_{j>k=1}^{M_0}(y_{ij} - \bar{Y})(y_{ik} - \bar{Y}) \\ &= \sum_{i=1}^{N}\left[\sum_{j=1}^{M_0}(y_{ij} - \bar{Y})\right]^2 - \sum_{i=1}^{N}\sum_{j=1}^{M_0}(y_{ij} - \bar{Y})^2 \\ &= (M_0 - 1)\left[(NM_0 - 1)S_y^2 - NM_0 S_y^2 w\right] \end{aligned}$$

et, donc, partant de (4.15c), que $\rho = 1 - \{NM_0/(NM_0 - 1)\}(S_{yW}^2/S_y^2) \cong \delta$ en supposant que $M_i \equiv M_0$ pour tout $i = 1, \ldots, N$, $NM_0/(NM_0 - 1) \cong 1$. Par conséquent, en supposant de surcroît que $(N-1)/N \cong 1$ et $(NM_0 - 1)M_0^{-1}(N-1)^{-1} \cong 1$, les deux formules de l'effet de plan (4.15a) et (4.15b) sont approximativement équivalentes à $1 + (M_0 - 1)\delta$. D'autres auteurs arrivent à la même formule approximative. Il en est ainsi parce que δ et ρ mesurent essentiellement la même chose, c'est-à-dire l'homogénéité de la grappe. Dans ces conditions, deux estimateurs $\hat{Y}$ et $\hat{\bar{Y}}$ ont le même effet de plan, tel que discuté à l'exemple 3.2. Notons qu'il s'agit d'un cas simple de plan d'échantillonnage autopondéré.

Särndal et coll. (1992, section 8.7) comparent les effets de plan pour deux estimateurs dans les conditions de l'exemple 4.3. Ils établissent aussi une expression simplifiée $1 + (\bar{M} - 1)\delta$ pour (4.13) et (4.14), en supposant qu'on peut ne pas tenir compte des covariances de M_i avec $M_i\bar{Y}_i^2$ et $M_i\bar{D}_i^2$. Leur discussion de la différence entre les estimateurs du total et de la moyenne se résume à Δ_a dans l'exemple 3.2. Ils notent aussi que l'effet de plan peut être beaucoup plus important pour le total que pour la moyenne de population, parce que la perte due à l'échantillonnage de grappes est plus importante quand on estime le total que quand on estime la moyenne.

Une pratique courante lorsque les grappes sont de taille inégale consiste à utiliser une méthode d'échantillonnage plus efficace qui tient compte de la différence de taille,

comme l'échantillonnage en grappes ppt. Nous pouvons appliquer les expressions (4.11) et (4.12) à des probabilités de sélection arbitraires p_i, où les p_i sont fixées de façon à être proportionnelles à une mesure de taille donnée $Z_i \geq 0$. La différence entre les effets de plan pour $\hat{Y}$ et $\bar{\hat{Y}}$ est expliquée par Δ_a dans (3.9), ou autrement

$$\Delta_a = \frac{m'}{\mathrm{CV}_y^2} \sum_{i=1}^{N} \frac{w_i Q_i}{M^2} \left[\left(\frac{\bar{Y}_i}{\bar{Y}} \right)^2 - \left(\frac{\bar{D}_i}{\bar{Y}} \right)^2 \right]. \qquad (4.16)$$

Le terme Q_i de (4.16) représente l'effet de p_i sur l'estimation de la variance lorsqu'on utilise une autre mesure de taille que la taille réelle des grappes M_i. Thomsen, Tesfu et Binder (1986) ont considéré, entre autres facteurs, l'effet d'une mesure de taille périmée sous échantillonnage à deux degrés avec échantillonnage aléatoire simple d'éléments à la deuxième étape. Nous y reviendrons à la section 4.4.

4.3 Plans d'échantillonnage autopondérés

Dans le cas d'un échantillonnage autopondéré, chaque élément échantillonné a le même poids, si bien que les estimateurs du total et de la moyenne ont tous deux une forme simple. Ils sont donnés par $\hat{Y} = y/f$ et $\bar{\hat{Y}} = y/m$, où $f = m/M$ est la fraction d'échantillonnage globale et $y = \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij}$ est le total d'échantillon. Alors, comme dans le cas de l'échantillonnage aléatoire simple, ainsi que le montre (3.4), les deux estimateurs ont le même effet de plan.

Un plan d'échantillonnage autopondéré peut être appliqué de diverses façons par synchronisation des méthodes d'échantillonnage de premier et de deuxième degré (par exemple, Kish 1965, section 7.2). Par exemple, si l'on utilise un échantillonnage avec probabilités égales pour le premier degré, alors l'échantillonnage de deuxième degré devrait se faire selon une méthode à probabilités égales avec une fraction d'échantillonnage uniforme pour toutes les UPE. À titre de cas particulier, où on sélectionne un échantillon aléatoire simple d'UPE de taille égale (c'est-à-dire $M_i = M_0$ pour tout i), Hansen et coll. (1953, volume II, pages 162 – 163) montrent que

$$\mathrm{CV}_p^2(\hat{\bar{Y}}) \cong \frac{1}{m} \mathrm{CV}_y^2 \left[1 + \rho(\bar{m} - 1) \right], \qquad (4.17)$$

où $\mathrm{CV}_p^2(\hat{\bar{Y}}) = V_p(\hat{\bar{Y}}) / \bar{Y}^2$ est la variance relative de $\hat{\bar{Y}}$ sous le plan d'échantillonnage p et ρ est le coefficient de corrélation intragrappe tel que défini dans (4.15c). Puisque la variance relative de $\bar{\hat{Y}}$ sous easar est $m^{-1} \mathrm{CV}_y^2$, la formule approximative bien connue de l'effet de plan pour $\hat{\bar{Y}}$ sous un plan d'échantillonnage autopondéré s'ensuit immédiatement sous la forme

$$\mathrm{Deft}_p^2(\hat{\bar{Y}}) = 1 + \rho(\bar{m} - 1). \qquad (4.18)$$

Pour les plans d'échantillonnage en grappes à un degré, nous avons montré des formes semblables données par (4.15a) et (4.15b) (voir aussi Yamane 1967, section 8.7). Hansen et coll. (1953, volume II. page 204) montrent en outre que $\mathrm{CV}_p^2(\hat{Y}) = \mathrm{CV}_p^2(\hat{\bar{Y}})$ pour un plan d'échantillonnage fondé sur l'échantillonnage aléatoire simple aux deux étapes, ce qui implique que $\hat{Y}$ et $\bar{\hat{Y}}$ ont le même effet de plan.

4.4 Échantillonnage à deux degrés avec probabilités inégales

Considérons l'exemple qui suit.

Exemple 4.5 Plan d'échantillonnage à deux degrés où n UPE sont sélectionnées avec remise avec probabilité p_i et un échantillon aléatoire simple de même taille de $m_0 \geq 2$ éléments est sélectionné avec remise à partir de chaque UPE sélectionnée. À l'aide de calculs et de simplifications ordinaires, nous pouvons montrer que

$$\mathrm{Deft}_p^2(\hat{Y}) \cong 1 + (m_0 - 1)\tau + W_y^*, \qquad (4.19)$$

où

$$\tau = \frac{(N-1)S_{yB}^2 + \sum_{i=1}^{N} (m_0 - 1)^{-1} S_{yi}^2}{(N-1)S_{yB}^2 + \sum_{i=1}^{N} (M_i - 1) S_{yi}^2}, \qquad (4.20)$$

$S_{yi}^2 = (M_i - 1)^{-1} \sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2$, $W_y^* = W_y / V_{\mathrm{srswr}}(\hat{Y}_{\mathrm{srs}}) = (m_0 / \mathrm{CV}_y^2) \sum_{i=1}^{N_i} (Q_i / p_i M^2)(\bar{Y}_i / \bar{Y})^2 (1 + \mathrm{CV}_{yi}^2 / m_0)$, et $\mathrm{CV}_{yi}^2 = S_{yi}^2 / \bar{Y}_i^2$ représente la variance relative intragrappe de la variable y. De la même façon,

$$\mathrm{Deft}_p^2(\hat{\bar{Y}}) \cong 1 + (m_0 - 1)\tau + W_d^*, \qquad (4.21)$$

où $W_d^* = W_d / V_{\mathrm{easar}}(\hat{Y}_{\mathrm{eas}}) = (m_0 / \mathrm{CV}_y^2) \sum_{i=1}^{N} (Q_i / p_i M^2) (\bar{D}_i / \bar{Y})^2 (1 + \mathrm{CV}_{di}^2 / m_0)$, et $\bar{D}_i$ et CV_{di}^2 sont définis à l'aide de la variable transformée $d\,(d_{ij} = y_{ij} - \bar{Y})$ de façon analogue à $\bar{Y}_i$ et CV_{yi}^2, respectivement. (Les calculs détaillés de l'établissement des expressions (4.19) et (4.21) peuvent être obtenus auprès des auteurs.) Dans le cas où $m_i = m_0$ pour tout i, la différence entre les effets de plan donnés par (4.19) et (4.21) se réduit à (3.7) ou (4.16). L'échantillonnage de deuxième degré ne contribue pas à la différence.

Si nous revenons à Thomsen et coll. (1986), qui ont étudié l'effet de l'utilisation d'une mesure périmée de taille sur la variance, nous voyons que la discussion qui précède concernant $\hat{Y}$ est équivalente à la leur. La seule différence tient au fait qu'ils émettent l'hypothèse d'un plan d'échantillonnage sans remise à la deuxième étape. Notons, toutefois, que la définition de τ dans Thomsen et coll. (1986) diffère légèrement de (4.20) et de δ à la section 4.2. Cependant, ces définitions sont étroitement liées. Pour le

voir, écrivons τ sous forme d'une fonction de certaines quantités b_i associées aux UPE comme suit :

$$\tau(b_i) = \frac{(N-1)S_{yB}^2 - \sum_{i=1}^{N} b_i S_{yi}^2}{(N-1)S_{yB}^2 + \sum_{i=1}^{N} (M_i - 1)S_{yi}^2}.$$

Alors, nous obtenons le τ de Thomsen et coll. (1986) avec $b_i = 1$, le τ de l'exemple 4.5 avec $-1/(m_0 - 1)$, et δ de la section 4.2 avec $(M_i - 1)/\{\sum_{i=1}^{N}(M_i - 1)/(N-1)\}$. En égalant la formule de Kish (4.18) pour $\hat{\bar{Y}}$ à (4.19) pour $\hat{Y}$, ils ont manifestement oublié de tenir compte du fait que les effets de plan pour $\hat{Y}$ et $\hat{\bar{Y}}$ peuvent être fort différents.

Pour les cas plus généraux, Kish (1987) a établi la formule bien connue suivante pour $\hat{\bar{Y}}$:

$$\begin{aligned}\mathrm{Deft}_{\mathrm{Kish}}^2(\hat{\bar{Y}}) &= \frac{n\sum_{g=1}^{G} n_g w_g^2}{\left(\sum_{g=1}^{G} n_g w_g\right)^2}\left[1 + \rho(\bar{m} - 1)\right] \\ &= (1 + \mathrm{cv}_w^2)\left[1 + \rho(\bar{m} - 1)\right].\end{aligned}$$

Il l'a obtenue en appliquant (4.5) (ou (4.6)) et (4.18) de façon récursive afin d'intégrer les effets de la mise en grappes ainsi que des poids inégaux. Gabler, Haeder et Lahiri (1999) ont justifié la formule susmentionnée pour $\hat{\bar{Y}}$ en utilisant un modèle de superpopulation défini pour la classification croisée de N grappes et G classes de pondération. Cependant, on ne peut exposer la différence entre les effets de plan pour $\hat{\bar{Y}}$ et $\hat{Y}$ selon une approche fondée sur un modèle de ce type, puisque y_k est traitée comme une variable aléatoire tandis que w_k est fixe. Sous cette approche, $\mathrm{Deft}_p^2(\hat{Y})$ ne diffère de $\mathrm{Deft}_p^2(\hat{\bar{Y}})$ que par un facteur $(\hat{M}/M)^2$, alors que la différence réelle peut être nettement plus prononcée, comme nous l'avons montré dans le présent article (par exemple, expressions (3.7) et (4.23)).

4.5 Cas plus généraux

La pondération des données d'enquête nécessite non seulement des poids d'échantillonnage, mais aussi l'application de diverses méthodes de redressement de la pondération, comme la stratification a posteriori, l'ajustement proportionnel itératif (raking) et la correction pour la non-réponse. Nous considérons ces cas généraux ici.

Nous pouvons réécrire l'approximation de premier ordre de Taylor de l'estimateur pondéré de la moyenne $\hat{\bar{Y}} = \hat{Y}/\hat{M}$ donné par (3.2) sous la forme $(\hat{Y} - Y)/Y \cong (\hat{\bar{Y}} - \bar{Y})/\bar{Y} + (\hat{M} - M)/M$. En prenant la variance des deux membres de l'équation, nous obtenons

$$\begin{aligned}\mathrm{CV}_p^2(\hat{Y}) \cong\ & \mathrm{CV}_p^2(\hat{\bar{Y}}) + \mathrm{CV}_p^2(\hat{M}) \\ & + 2R_p(\hat{\bar{Y}}, \hat{M})\,\mathrm{CV}_p(\hat{\bar{Y}})\mathrm{CV}_p(\hat{M}), \qquad (4.22)\end{aligned}$$

où $\mathrm{CV}_p^2(\hat{Y}), \mathrm{CV}_p^2(\hat{\bar{Y}}), \mathrm{CV}_p^2(\hat{M})$ sont les variances relatives de $\hat{Y}$, $\hat{\bar{Y}}$ et $\hat{M}$ respectivement et $R_p(\hat{\bar{Y}}, \hat{M})$ est le coefficient de corrélation de $\hat{\bar{Y}}$ et $\hat{M}$ par rapport au plan d'échantillonnage complexe p et tout redressement de la pondération. Puisque les variances relatives des simples total et moyenne d'échantillon $\hat{Y}_{\mathrm{eas}}$ et $\bar{y}_{\mathrm{eas}}$ sont $\mathrm{CV}_{\mathrm{easar}}^2(\hat{Y}_{\mathrm{eas}}) = \mathrm{CV}_{\mathrm{easar}}^2(\bar{y}_{\mathrm{eas}}) = m^{-1}\mathrm{CV}_y^2$ sous easar de taille m, il découle de (4.22) que

$$\begin{aligned}\mathrm{Deft}_p^2(\hat{Y}) \cong\ & \mathrm{Deft}_p^2(\hat{\bar{Y}}) \\ & + 2R_p(\hat{\bar{Y}}, \hat{M})\nabla_p(y)\mathrm{Deft}_p(\hat{\bar{Y}}) + \nabla_p^2(y), \qquad (4.23)\end{aligned}$$

où $\nabla_p(y) = \mathrm{CV}_p(\hat{M})/\mathrm{CV}_{\mathrm{easar}}(\bar{y}_{\mathrm{eas}})$ est non négatif. À titre d'illustration, considérons une variable binaire y, où $\mathrm{CV}_y^2 \cong (1 - \bar{Y})/\bar{Y}$ et, donc, $\nabla_p(y)$ peut être arbitrairement grand quand $\bar{Y}$ s'approche de 1 ou petit quand $\bar{Y}$ s'approche de zéro en supposant que $\mathrm{CV}_p(\hat{M}) \neq 0$. Si $\nabla_p(y)$ est quasi nul, les deux effets de plan sont presque égaux. Sinon, l'un est plus grand que l'autre dépendant des valeurs de $\nabla_p(y)$ et de $R_p(\hat{\bar{Y}}, \hat{M})$. Si les poids d'échantillonnage sont calés sur la taille connue de population M, $\hat{Y}$ et $\hat{\bar{Y}}$ ont le même effet de plan, puisque $\hat{M} = M$ et $\mathrm{CV}_p(\hat{M}) = 0$. Dans ce cas, le calage n'influe pas sur $\hat{\bar{Y}}$, mais $\hat{Y} = M\hat{\bar{Y}}$, qui est un estimateur par le quotient. Notons que nous pouvons utiliser les méthodes de stratification a posteriori ou d'ajustement proportionnel itératif (raking) si nous disposons d'information sur la taille de population au niveau de la sous-population et que nous obtenons également des effets de plan équivalent. Néanmoins, en général, nous avons $\mathrm{Deft}_p^2(\hat{Y}) \geq \mathrm{Deft}_p^2(\hat{\bar{Y}})$ si

$$R_p(\hat{\bar{Y}}, \hat{M}) \geq -\frac{1}{2}\frac{\nabla_p(y)}{\mathrm{Deft}_p(\hat{\bar{Y}})} \quad \text{ou}$$

$$R_p(\hat{\bar{Y}}, \hat{M}) \geq -\frac{1}{2}\frac{\mathrm{CV}_p(\hat{M})}{\mathrm{CV}_p(\hat{\bar{Y}})}, \qquad (4.24)$$

et inversement.

Il est instructif d'examiner certains cas particuliers. Par exemple, si $R_p(\hat{\bar{Y}}, \hat{M}) \geq 0$, alors $\mathrm{Deft}_p^2(\hat{Y}) > \mathrm{Deft}_p^2(\hat{\bar{Y}})$, mais une corrélation négative (par exemple $R_p(\hat{\bar{Y}}, \hat{M}) < 0$) ne donne pas nécessairement lieu à $\mathrm{Deft}_p^2(\hat{Y}) \leq \mathrm{Deft}_p^2(\hat{\bar{Y}})$. Pour un cas particulier de $R_p(\hat{\bar{Y}}, \hat{M}) = 0$, la différence est donnée par

$$\mathrm{Deft}_p^2(\hat{Y}) - \mathrm{Deft}_p^2(\hat{\bar{Y}}) \cong \frac{\mathrm{CV}_p^2(\hat{M})}{\mathrm{CV}_{\mathrm{easar}}^2(\bar{y}_{\mathrm{eas}})}. \qquad (4.25)$$

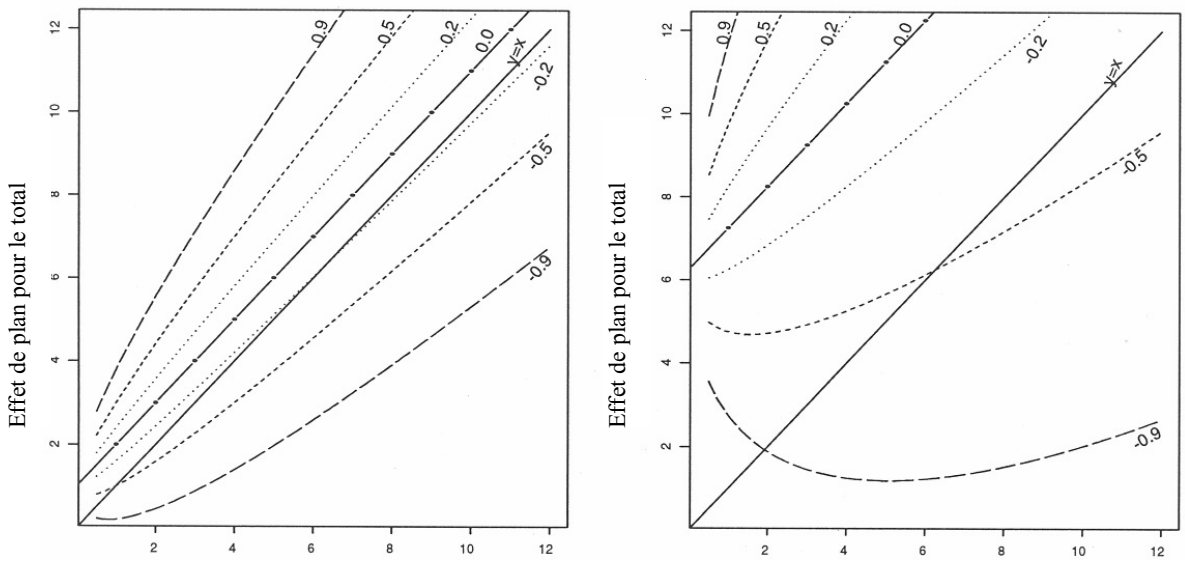
La figure 1 illustre la relation entre les deux effets de plan. L'expression (4.23) est représentée graphiquement pour certaines valeurs fixes de $R_p(\hat{\bar{Y}}, \hat{M})$ et de $\nabla_p(y)$. La droite en trait plein passant par l'origine, qui représente des

effets de plan égaux, est la droite de référence. Comme le montre le graphique, la comparaison n'est pas nette. Si $R_p(\hat{\bar{Y}}, \hat{M}) < 0$, $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\bar{Y}})$ pour un petit $\text{Deft}_p^2(\hat{\bar{Y}})$, mais la relation est inversée à mesure que $\text{Deft}_p^2(\hat{\bar{Y}})$ s'accroît.

Hansen et coll. (1953, volume I, pages 338 – 339) indiquent que $R_p(\hat{\bar{Y}}, \hat{M})$ est souvent proche de 0. Dans ces conditions, l'expression (4.25) se simplifie en $\text{Deft}_p^2(\hat{Y}) \cong \text{Deft}_p^2(\hat{\bar{Y}})\left[1 + \text{CV}_p^2(\hat{M}) / \text{CV}_p^2(\hat{\bar{Y}})\right]$, d'où nous obtenons $\text{Deft}_p^2(\hat{Y}) \geq \text{Deft}_p^2(\hat{\bar{Y}})$. Ce cas particulier a été étudié par Jang (2001). Cependant, cela ne semble pas nécessaire, comme le montre l'exemple suivant.

Exemple 4.6 Pour illustrer la relation entre les effets de plan pour $\hat{\bar{Y}}$ et $\hat{Y}$, nous avons utilisé un ensemble de données sur les adultes provenant de la troisième National Health and Nutrition Examination Survey américaine (NHANES III) qui est fourni comme fichier de démonstration dans la version 4.0 de WesVar. La NHANES III est une enquête avec examen médical nationale à grande échelle fondée sur un plan d'échantillonnage stratifié à plusieurs degrés, pour lequel la méthode des répliques répétées équilibrées (BRR) modifiée de Fay est employée pour l'estimation de la variance. (Voir Judkins 1990, pour plus de précisions sur la méthode de Fay.) Nous n'avons utilisé que 19 793 enregistrements pour lesquels les réponses étaient complètes pour les caractéristiques énumérées au tableau 1. Notons que le poids qui figure dans le fichier de démonstration diffère du poids final de la NHANES III obtenu par stratification a posteriori. Pour plus de renseignements sur le fichier de démonstration, voir Westat (2001).

Le tableau 1 donne les effets de plan pour $\hat{\bar{Y}}$ et $\hat{Y}$, et les termes constituant (4.23) pour les caractéristiques choisies. Notons que $\nabla_p(y)$ diminue de façon monotone en CV_y sachant que $m = 19\,793$ et $\text{cv}_p(\hat{M}) = 3{,}2$ %. Bien que $\nabla_p(y)$ ait tendance à être le facteur déterminant de la différence entre les effets de plan, $R_p(\hat{\bar{Y}}, \hat{M})$ peut jouer un rôle important quand il est négatif. Par exemple, pour deux races/groupes ethniques, à savoir les Afro-américains et les Hispaniques, les valeurs négatives –0,67 et –0,24 de $R_p(\hat{\bar{Y}}, \hat{M})$ sont les causes de $\text{Deft}_p^2(\hat{Y}) < \text{Deft}_p^2(\hat{\bar{Y}})$. Certains effets de plan pour $\hat{Y}$ sont énormes. Il n'en est pas ainsi dans le cas des poids finaux stratifiés a posteriori de la NHANES III, avec lesquels on obtient le même effet de plan pour $\hat{Y}$ et $\hat{\bar{Y}}$. Ce résultat souligne l'importance du calage du redressement des poids pour l'estimation des totaux.



(a) $\nabla_p(y) = 1{,}0$ (b) $\nabla_p(y) = 2{,}5$

Figure 1. Tracés de $\text{Deft}_p^2(\hat{Y})$ en fonction de $\text{Deft}_p^2(\hat{\bar{Y}})$ pour a) $\nabla_p(y) = 1{,}0$, b) $\nabla_p(y) = 2{,}5$. La droite en trait plein correspond à $\text{Deft}_p^2(\hat{\bar{Y}}) = \text{Deft}_p^2(\hat{Y})$. Les autres courbes correspondent à $R_p(\hat{\bar{Y}}, \hat{M}) = -0{,}9, -0{,}5, -0{,}2, 0, 0{,}2, 0{,}5$ et 0,9, respectivement.

Tableau 1
Comparaison des effets de plan pour le total pondéré et la moyenne pondérée à l'aide d'un sous-ensemble du fichier de données sur les adultes provenant de la troisième National Health and Nutrition Examination Survey (NHANES III) américaine

		Moyenne			Total						
Caractéristique		Estimation	Deft^2	cv	Estimation	Deft^2	cv	cv_y	$r_p(\hat{\bar{Y}},\hat{M})$	$\nabla_p(y)$	$-\frac{cv_p(\hat{M})}{2cv_p(\hat{\bar{Y}})}$
A fumé 100+ cigarettes au cours de la vie?	Oui	0,53	4,13	0,014	98 397 795	31,31	0,038	0,944	0,20	4,83	−0,58
Fait du diabète?	Oui	0,05	1,75	0,040	9 783 307	1,92	0,042	4,246	−0,34	1,07	−0,31
	Non	0,95	1,75	0,002	176 341 218	393,47	0,033	0,236	0,34	19,35	−5,53
Fait de l'hypertension	Oui	0,23	3,42	0,024	42 939 866	7,96	0,037	1,826	−0,18	2,50	−0,37
	Non	0,77	3,42	0,007	143 184 660	78,44	0,034	0,548	0,18	8,32	−1,22
Race/Groupe ethnique	Afro-américan*	0,12	7,64	0,054	21 567 028	4,21	0,040	2,762	−0,67	1,65	−0,11
	Hispanique*	0,05	6,70	0,079	9 550 326	6,48	0,078	4,300	−0,24	1,06	−0,08
Sexe	Masculin	0,48	1,40	0,009	88 725 967	19,18	0,033	1,048	−0,11	4,35	−1,55
	Féminin	0,52	1,40	0,008	97 398 559	25,39	0,034	0,954	0,11	4,77	−1,70
Nombre de cigarettes fumées par jour	_	5,25	6,42	0,037	977 225 826	10,51	0,047	2,044	−0,09	2,23	−0,17
Taille de la population	_	_	_	_	186 124 526	_	0,032	_	_	_	_

Nota : * indique les cas où l'effet de plan est plus faible pour $\hat{Y}$ que pour $\hat{\bar{Y}}$.

5. Conclusion

Nous avons étudié les effets de plan des deux estimateurs les plus répandus de la moyenne et du total de population dans le cas des enquêtes par sondage sous divers plans d'échantillonnage avec remise. À notre avis, l'utilisation d'un échantillonnage avec remise ne constitue pas forcément une limite grave, car elle permet de voir les choses plus clairement, sans embrouiller les calculs à cause des complications probablement inutiles des plans d'échantillonnage sans remise. En outre, l'effet de la correction pour population finie s'annule en grande partie dans notre formule de l'effet de plan, si bien que les résultats sont fort comparables aux effets de plan classiques pour l'échantillonnage sans remise. Par conséquent, nos résultats devraient être utiles en pratique. Nous résumons les plus importants ci-après.

La formule approximative bien connue de l'effet de plan proposée par Kish pour les estimateurs pondérés de la moyenne (type quotient) ne se généralise facilement ni en forme et ni en concepts aux problèmes plus généraux, particulièrement les estimateurs pondérés du total, contrairement à ce que pensent de nombreuses personnes. En fait, $\hat{\bar{Y}}$ et $\hat{Y}$ ont souvent des effets de plan fort différents, à moins que le plan d'échantillonnage soit autopondéré ou que les poids d'échantillonnage soient calés sur la taille connue de population. En outre, l'effet de plan n'est généralement pas indépendant de la distribution de la variable étudiée, même pour l'estimateur de la moyenne, sans parler de l'estimateur du total. De surcroît, la corrélation de la variable étudiée et des poids utilisés pour l'estimation peut être un déterminant important de l'effet de plan. Par conséquent, outre ce qu'il est destiné à évaluer au départ, l'effet de plan mesure non seulement l'effet d'un plan d'échantillonnage complexe sur une statistique particulière, mais aussi les effets de la distribution de la variable étudiée et de ses relations avec le plan d'échantillonnage sur la statistique. Puisque les progiciels applicables à des données d'enquête complexes calculent systématiquement l'effet de plan, il semble approprié d'avertir leurs utilisateurs de ces faits assez obscurs au sujet de l'effet de plan.

Remerciements

Les auteurs remercient Louis Rizzo à Westat, un éditeur associé et deux arbitres pour leurs commentaires utiles et leurs suggestions sur une version antérieur de cet article.

Bibliographie

Apostol, T.M. (1974). *Mathematical Analysis.* 2ième Éd. Reading, MA: Addison-Wesley.

Barron, E.W., et Finch, R.H. (1978). Design Effects in a complex multistage sample: The Survey of Low Income Aged and Disabled (SLIAD), *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 400-405.

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

Cornfield, J. (1951). Modern methods in the sampling of human populations. *American Journal of Public Health*, 41, 654-661.

Gabler, S., Haeder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. Vol. I, New York: John Wiley & Sons, Inc.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vol. II, New York: John Wiley & Sons, Inc.

Judkins, D.R. (1990). Fay's method for variance estimation, *Journal of Official Statistics*, 6, 223-239.

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

Kish, L. (1987). Weighting in Deft^2. *The Survey Statistician.* Juin 1987.

Kish, L. (1992). Weighting for unequal p_i. *Journal of Official Statistics*, 8, 183-200.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

Jang, D. (2001). On procedures to summarize variances for survey estimates. *Proceedings of the Survey Research Methods of the American Statistical Association.* CD-ROM.

Lehtonen, R., et Pahkinen, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Brooks/Cole.

Park, I., et Lee, H. (2001). The design effect: do we know all about it? *Proceedings of the Section on Survey Research Methods,* American Statistical Association. CD-ROM.

Park, I., et Lee, H. (2002). A revisit of design effects under unequal probability sampling. *The Survey Statistician*, 46, 23-26.

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

Spencer, B.D. (2000). Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de selection. *Techniques d'enquête*, 26, 137-138.

Thomsen, I., Tesfu, D. et Binder, D.A. (1986). Estimation of Design Effects and Intraclass Correlations When Using Outdated Measures of Size. *International Statistical Review*, 54, 343-349.

Westat (2001). *WesVar 4.0 User's Guide.* Rockville, MD: Westat, Inc.

Yamane, T. (1967). *Elementary Sampling Theory*. New Jersey: Prentice-Hall.