

Effets de plan de sondage dus aux bases de sondage dans les enquêtes auprès des établissements

Monroe G. Sirken¹

Résumé

Lorsqu'on dispose de bases de sondage indépendantes énumérant tous les établissements et la mesure de leur taille, on utilise habituellement pour les enquêtes auprès des établissements l'estimateur PPT de Hansen–Hurwitz (HH) pour estimer le volume des transactions faites par les établissements avec les populations. Le présent article décrit la version par échantillonnage en réseau (NS pour *network sampling*) de l'estimateur HH proposée pour remplacer éventuellement la version PPT. L'estimateur NS dépend d'une liste d'établissements établie d'après une enquête démographique qui énumère les ménages et leur probabilité de sélection dans une enquête démographique par sondage et le nombre de transactions, si tant est qu'il y en ait, que fait chaque ménage avec chaque établissement. Un modèle statistique est élaboré en vue de comparer l'efficacité des estimateurs HH et NS en cas d'enquête par échantillonnage à un degré ou à deux degrés auprès des établissements en supposant que la base de sondage autonome et celle établie d'après une enquête démographique ne comportent aucune erreur de couverture ni de mesure de taille.

Mots clés : Bases de sondage autonome d'établissements; listes d'établissements établies d'après une enquête démographique; estimateur de Hansen-Hurwitz; estimateur par échantillonnage en réseau.

1. Introduction

Les listes d'établissements qui font des transactions avec les ménages visés par les enquêtes démographiques par sondage servent de bases de sondage aux enquêtes auprès des établissements lorsque l'on peut établir la correspondance entre les transactions déclarées par les ménages lors des enquêtes démographiques et les enregistrements des établissements pertinents. Par exemple, les listes d'établissements qui font des transactions avec les ménages participant à la National Medical Expenditure Panel Survey (MEPS), une enquête nationale auprès d'un échantillon de population, servent de bases de sondage pour les enquêtes sur les fournisseurs de soins médicaux qui permettent de compléter et de vérifier les données sur les dépenses pour soins médicaux correspondant aux transactions déclarées par les membres répondants des ménages de la MEPS (Cohen 1998). Cependant, les listes d'établissements qui font des transactions avec les ménages qui participent à des enquêtes démographiques par sondage servent rarement de bases de sondage aux enquêtes auprès des établissements destinées à recueillir des renseignements sur les transactions que les établissements font avec tous les ménages. Le Current Price Index (CPI) produit par le Bureau of Labor Statistics est un cas rare, qui mérite d'être mentionné, d'enquête auprès d'établissements fédéraux fondée sur une base de sondage produite à partir d'une enquête auprès d'un échantillon de population. La CPI Pricing Survey, une enquête nationale auprès des établissements de détail qui a pour but de recueillir des données sur les prix pour un panier de biens de consommation achetés par l'ensemble des consommateurs, a pour base de sondage les listes

d'établissements de détail qui font des transactions avec les ménages qui participent à la CPI Continuing Point of Purchase Survey (Leaver et Valliant 1995).

Après avoir examiné les plans de restructuration de la famille d'enquêtes nationales indépendantes auprès des fournisseurs de soins de santé (hôpitaux, médecins, cliniques, *etc.*) du National Center for Health Statistics (NCHS), un groupe d'experts du Committee on National Statistics a proposé (Wunderlich 1992) d'utiliser les listes de fournisseurs de soins de santé déclarés par les ménages lors de la National Health Interview Survey (NHIS), qui est une enquête par sondage nationale permanente auprès des ménages (Massey, Moore, Parsons et Tadros 1991), comme bases de sondage des enquêtes nationales auprès des fournisseurs de soins de santé. Selon le Comité, étant donné l'évolution rapide des listes de fournisseurs de soins de santé due à l'évolution rapide du système national de prestation des services de santé, les listes de fournisseurs de soins de santé établies d'après la NHIS seraient plus exactes, et plus faciles et moins coûteuses à produire et à tenir à jour que les listes indépendantes de fournisseurs de soins de santé utilisées à l'époque. Peu après la diffusion du rapport du groupe d'experts, le NCHS a lancé un projet de recherche sur les bases de sondage produites d'après des enquêtes démographiques que nous résumons brièvement plus bas.

Au départ, l'étude s'est concentrée presque exclusivement sur les propriétés statistiques des listes de fournisseurs de soins de santé établies d'après la NHIS. Judkins, Berk, Edwards, Mohr, Stewart et Waksberg (1995) ont étudié la qualité des listes autonomes de fournisseurs de soins de santé utilisées à l'époque ou susceptibles d'être utilisées, et

1. Monroe G. Sirken, Senior Research Scientist, National Center for Health Statistics, U.S.A.

cherché à déterminer les catégories de fournisseurs de soins pour lesquels les listes établies d'après la NHIS offrirait le plus de potentiel. Subséquemment, Judkins, Marker, Waksberg, Botman et Massey (1999) ont comparé grossièrement l'efficacité des enquêtes sur les soins dentaires fondées sur une base de sondage établie d'après la NHIS et celle des enquêtes fondées sur la base de sondage autonome, et ont conclu que les listes de fournisseurs de soins de santé établies d'après la NHIS méritent d'être prises sérieusement en considération lorsqu'on ne dispose d'aucune liste autonome raisonnablement complète comportant des mesures de taille raisonnablement bonnes.

Ces dernières années, la recherche s'est concentrée sur les propriétés statistiques des estimateurs fondés sur des bases de sondage établies d'après des enquêtes démographiques et a pris une orientation plus théorique qu'auparavant. Les difficultés conceptuelles qu'a posé au départ l'élaboration d'estimateurs non biaisés pour une base de sondage établie d'après une enquête démographique, parce qu'un même établissement fait des transactions avec plusieurs ménages, ont été surmontées grâce à l'application de la théorie de l'échantillonnage en réseau (Sirken 1997; Thompson 1992). Sirken, Shimizu et Judkins (1995) ont élaboré la version par échantillonnage en réseau de l'estimateur HH, version que nous appellerons dans le présent article estimateur NS (pour *network sampling*), et Sirken et Shimizu (1999) ont élaboré la version par échantillonnage en réseau de l'estimateur d'Horwitz-Thompson (HT). Le présent article décrit le développement d'un modèle d'erreur statistique permettant de comparer l'efficacité de l'estimateur NS qui dépend de la base de sondage établie d'après une enquête démographique et celle de l'estimateur HH qui dépend de la base de sondage autonome. Le modèle d'erreur suppose que les bases de sondage ne contiennent aucune erreur de couverture ni de mesure de taille des établissements et que les coûts de leur production et tenue à jour sont équivalents. Le modèle est fondé sur l'hypothèse que l'enquête démographique d'après laquelle est établie la base de sondage est réalisée par échantillonnage aléatoire simple (EAS), mais il peut être appliqué à d'autres plans d'échantillonnage non considérés dans le présent article.

La présentation de l'article est la suivante. La notation est décrite à la section 2. Les sections 3.1 et 3.2 donnent, respectivement, une description de l'estimateur HH auto-pondéré PPT et de sa variance pour une enquête par échantillonnage à deux degrés auprès des établissements fondée sur la base de sondage autonome, et de l'estimateur NS et de sa variance pour une enquête par échantillonnage à deux degrés auprès d'établissements fondée sur la base de sondage établie d'après une enquête démographique. L'élaboration du modèle d'erreur est présentée aux sections 4.1 à 4.4. La différence entre les variances HH et NS de deuxième degré pour des tailles d'échantillon prévues équivalentes est établie à la section 4.1. À la section 4.2, la composante de premier degré de la variance de l'estimateur

NS à deux degrés est subdivisée en composante de la variance représentant les effets des ménages qui font et qui ne font pas de transactions, et la section 4.3 montre les effets de plan de sondage de l'estimateur NS en cas d'échantillonnage à un seul degré. Les composantes de deuxième degré de la variance des estimateurs NS et HH sont comparées à la section 4.4. Pour conclure, la section 5 résume les principaux résultats de la comparaison de l'efficacité des estimateurs HH et NS dans le cas d'enquêtes par échantillonnage à un degré et à deux degrés auprès des établissements au moyen du modèle d'erreur et expose brièvement les limites du modèle. La preuve d'un énoncé statistique figurant à la section 4.2 est donnée en annexe.

2. Notation

Représentons par N_j le nombre de ménages faisant des transactions avec l'établissement j ($j=1, 2, \dots, R$), par N_o , le nombre de ménages ne faisant pas de transactions avec un établissement et par N^* , le nombre de ménages distincts faisant des transactions avec R établissements. Alors, $N = N^* + N_o$ représente le nombre total de ménages.

Représentons par M_{ij} le nombre de transactions de l'établissement j ($j=1, 2, \dots, R$) avec le ménage i ($i=1, 2, \dots, N$), où $M_{ij} \geq 0$ si l'établissement j fait des transactions avec le ménage i , et $M_{ij} = 0$ si l'établissement j et le ménage i ne font pas de transactions. Alors, $M_j = \sum_{i=1}^N M_{ij}$ représente le nombre de transactions de l'établissement j avec N ménages, et $M = \sum_{j=1}^R M_j$, le nombre de transactions de M établissements avec N ménages, et $\bar{M} = M/N$, le nombre moyen de transactions par ménage.

Représentons par X_{jk} la valeur de la variable x pour la transaction k ($k=1, \dots, M_j$) de l'établissement j ($j=1, 2, \dots, R$). Alors, $X_j = \sum_{k=1}^{M_j} X_{jk}$ représente la somme de la variable x sur les M_j transactions de l'établissement j , et $X = \sum_{j=1}^R X_j$ la somme de la variable x sur les M transactions de R établissements. Représentons par $\bar{X}_j = X_j/M_j$ la valeur moyenne de la variable x sur les M_j transactions de l'établissement j , et $\bar{X} = X/M$ la valeur moyenne de la variable x sur M transactions.

3. Estimateurs et variances

3.1 Estimateur et variance HH

Considérons une enquête par échantillonnage à deux degrés auto-pondéré auprès des établissements réalisée à l'aide d'une liste autonome d'établissements qui énumère les R établissements et leur mesure de taille, M_j ($j=1, 2, \dots, R$). Les établissements sont les unités primaires d'échantillonnage (UPE) et les transactions sont les unités secondaires d'échantillonnage. On sélectionne un échantillon PPT de r établissements avec remise à partir de la liste autonome, et on sélectionne indépendamment un

échantillon de taille $t_{HH} < \min(M_1, \dots, M_j, \dots, M_R)$ transactions, où t_{HH} est un entier positif, par échantillonnage aléatoire simple sans remise pour chaque établissement j ($j=1, 2, \dots, r$) de l'échantillon.

L'estimateur HH non biaisé autopondéré PPT de X est

$$X'_{HH} = \frac{M}{r} \sum_{j=1}^r \bar{X}'_j \quad (1)$$

où $\bar{X}'_j = \sum_{k=1}^{t_{HH}} X_{ij}/t_{HH}$ est l'estimation non biaisée de $\bar{X}_j = X_j/M_j$ ($j=1, 2, \dots, R$). Comme les établissements sont sélectionnés avec remise, l'estimateur HH compte \bar{X}'_j autant de fois que l'établissement j est sélectionné dans l'échantillon.

La variance de X'_{HH} est (Thompson 1992)

$$\text{Var}(X'_{HH}) = \frac{M^2}{r} \sigma_{HHI}^2 + \frac{M}{rt_{HH}} \sum_{j=1}^R (M_j - t_{HH}) \sigma_j^2 \quad (2)$$

où les premier et deuxième termes du deuxième membre représentent, respectivement, les composantes de premier et de deuxième degré de la variance,

$$\sigma_{HHI}^2 = \frac{1}{M} \sum_{j=1}^R M_j (\bar{X}_j - X/M)^2 \quad (3)$$

est la variance entre établissements, et

$$\sigma_j^2 = \frac{1}{M_j - 1} \sum_{k=1}^{M_j} (X_{jk} - X_j/M_j)^2 \quad (4)$$

est la variance à l'intérieur de l'établissement j .

3.2 Estimateur et variance NS

Considérons une enquête par échantillonnage à deux degrés auprès des établissements fondée sur une base de sondage établie d'après une enquête démographique. La base de sondage est une liste de n ménages échantillonnés H'_i ($i=1, 2, \dots, n$) qui ont participé à une enquête démographique par sondage. Pour chaque ménage énuméré H'_i , la base fournit π_i , c'est-à-dire la probabilité de sélection dans l'enquête-ménage et M_{ij} , c'est-à-dire le nombre de transactions du ménages avec chaque établissement j ($j=1, 2, \dots, R$) (les M_{ij} sont déclarés par les membres répondants des ménages lors de l'enquête démographique par sondage).

Chacun des n ménages figurant sur la liste établie d'après l'enquête démographique représente une grappe d'établissements dont la taille varie de 0 à R établissements avec lesquels le ménage a fait des transactions. Les n grappes d'établissements représentent les unités primaires d'échantillonnage et les M_j ($j=1, 2, \dots, r$) transactions des r établissements échantillonnés représentent les unités d'échantillonnage secondaires. On sélectionne l'échantillon de transactions pour l'établissement j ($j=1, 2, \dots, R$) de la façon suivante: un échantillon aléatoire simple de transactions de taille $t_{NS} M_{ij} < \text{Min}(M_1, M_2, \dots, M_r)$ est sélectionné indépendamment sans remise pour chaque ménage échantillonné H'_i ($i=1, 2, \dots, n$), où t_{NS} est un entier positif. La taille de l'échantillon de transactions de

l'établissement j ($j=1, 2, \dots, R$) est égale à $t_{NS} \sum_{i=1}^n M_{ij}$, et la taille totale de l'échantillon de transactions est égale à τt_{NS} , où $\tau = \sum_{i=1}^n \sum_{j \in A_i} M_{ij}$, c'est-à-dire la somme des transactions sur n ménages échantillonnés, est une variable aléatoire.

L'estimateur NS de X est

$$X'_{NS} = \sum_{i=1}^n \frac{1}{\pi_i} \sum_{j \in A_i} M_{ij} \bar{X}'_j(i)$$

où A_i représente la grappe d'établissements distincts qui font des transactions avec le ménage échantillonné H'_i , et

$$\bar{X}'_j(i) = \sum_{k=1}^{t_{NS} M_{ij}} X_{jk} / (t_{NS} M_{ij})$$

est une estimation non biaisée \bar{X}'_j pour un échantillon de $t_{NS} M_{ij}$ transactions de l'établissement j . Comme les ménages sont sélectionnés avec remise, l'estimateur NS compte la quantité $\sum_{j \in A_i} M_{ij} \bar{X}'_j(i)$ chaque fois que le ménage H'_i ($i=1, 2, \dots, n$) est sélectionné dans l'échantillon et, comme un même établissement fait des transactions avec plusieurs ménages, l'estimateur NS compte la quantité $M_{ij} \bar{X}'_j(i)$ chaque fois qu'un ménage échantillonné i ($i=1, 2, \dots, n$) fait des transactions avec l'établissement j .

Si l'on suppose que l'enquête démographique est réalisée par EAS, $\pi_i = n/N$, et l'estimateur basé sur l'échantillonnage en réseau est

$$X'_{NS} = \frac{N}{n} \sum_{i=1}^n \sum_{j \in A_i} M_{ij} \bar{X}'_j(i). \quad (5)$$

L'estimateur NS est un estimateur non biaisé de X .

$$\begin{aligned} E(X'_{NS}) &= \sum_{i=1}^n E \sum_{j \in A_i} M_{ij} \bar{X}'_j(i) = \sum_{i=1}^n \sum_{j \in A_i} M_{ij} \bar{X}'_j \\ &= \sum_{i=1}^n M_j \bar{X}_j = \sum_{j=1}^R X_j = X. \end{aligned}$$

L'estimateur NS représenté par l'équation (5) est autopondéré, parce que nous avons supposé que les n ménages sont sélectionnés par EAS. L'estimateur sera autopondéré si le plan d'échantillonnage de l'enquête démographique par sondage utilisée pour établir la liste d'établissements est autopondéré. Si $N = N^* = M$, ce qui sous-entend que N^* ménages ne font chacun qu'une seule transaction et que $N_0 = N - N^*$ ménages ne font aucune transaction, et si $n=r$ et $t_{NS} = t_{HH}$, les estimateurs HH et NS sont équivalents.

$$\begin{aligned} X'_{NS} &= \frac{N}{n} \sum_{i=1}^n \sum_{j \in A_i} M_{ij} \bar{X}'_j(i) = \frac{N}{n} \sum_{i=1}^n \sum_{j \in A_i} \bar{X}'_j \\ &= \frac{M}{r} \sum_{j=1}^R \bar{X}'_j = X_{HH}. \end{aligned} \quad (6)$$

Dans les conditions d'EAS avec remise de n ménages et de sélection indépendante de $t_{NS} M_{ij}$ transactions par EAS

sans remise pour chaque établissement j lié au ménage H_i , la variance de l'estimateur NS (5) est donnée par (Sirken et coll. 1995) :

$$\text{Var}(X'_{\text{NS}}) = \frac{N^2}{n} \sigma_{\text{NS1}}^2 + \frac{N}{nt_{\text{NS}}} \sum_{i=1}^N \sum_{j=1}^R M_{ij} \frac{M_j - t_{\text{NS}} M_{ij}}{M_j} \sigma_j^2 \quad (7)$$

où les premier et deuxième termes du deuxième membre représentent, respectivement, les composantes de premier et de deuxième degré de la variance,

$$\sigma_{\text{NS1}}^2 = \frac{1}{N} \sum_{i=1}^N \left(\sum_{j \in A_i} M_{ij} \bar{X}'_j - X/N \right)^2 \quad (8)$$

est la variance de population entre les ménages, et σ_j^2 , la variance de population à l'intérieur de l'établissement j telle que définie en (4). Une estimation non biaisée de la variance NS est donnée par

$$\text{Var}(X'_{\text{NS}}) = \frac{N^2}{n(n-1)} \sum_{i=1}^n \left[\sum_{j \in A_i} M_{ij} \bar{X}'_j(i) - \bar{X}' \right]^2 \quad (9)$$

où $\bar{X}' = X'/N$.

4. Le modèle d'erreur

4.1 Variances HH et NS pour des tailles prévues d'échantillon équivalentes

Obtenu en soustrayant(2) de (7), la différence entre les variances des estimateurs HH et NS de X est

$$\begin{aligned} \text{Var}(X'_{\text{NS}}) - \text{Var}(X'_{\text{HH}}) &= \left[\frac{N^2}{n} \sigma_{\text{NS1}}^2 - \frac{M^2}{r} \sigma_{\text{HH1}}^2 \right] \\ &+ \left[\frac{N}{nt_{\text{NS}}} \sum_{i=1}^N \sum_{j=1}^R M_{ij} \frac{M_j - t_{\text{NS}} M_{ij}}{M_j} \sigma_j^2 \right. \\ &\quad \left. - \frac{M}{rt_{\text{HH}}} \sum_{j=1}^R (M_j - t_{\text{HH}}) \sigma_j^2 \right] \quad (10) \end{aligned}$$

où les premier et deuxième ensembles de termes entre crochets dans le deuxième membre représentent, respectivement, les différences entre les composantes de premier et de deuxième degré de la variance des estimateurs HH et NS de X .

Représentons par $m_{\text{NS}} = \tau t_{\text{NS}}$ la taille de l'échantillon de transactions dans l'enquête auprès des établissements fondée sur la base de sondage établie d'après une enquête démographique, où t_{NS} , un entier positif, est la taille de l'échantillon de transactions sélectionné par transaction des n ménages échantillonnés, et par $\tau = \frac{\sum_{i=1}^n \sum_{j \in A_i} M_{ij}}{M}$, la somme des transactions sur n ménages échantillonnés.

Manifestement, τ est une variable aléatoire et son espérance conditionnelle sur l'ensemble des échantillons de n ménages est $E(\tau|n) = n\bar{M}$ où $\bar{M} = M/N$ est la taille moyenne des transactions des ménages. Il s'ensuit que $E(m_{\text{NS}}|n) = t_{\text{NS}} E(\tau|n) = n\bar{M} t_{\text{NS}}$ est la taille prévue de l'échantillon de transactions de l'estimateur NS conditionnelle sur l'ensemble des échantillons de n ménages.

Représentons par $m_{\text{HH}} = r t_{\text{HH}}$ la taille de l'échantillon de transactions dans le cas de l'enquête auprès des établissements fondée sur la base de sondage autonome, où r est la taille de l'échantillon d'établissements et t_{HH} , la taille de l'échantillon de transactions par établissement sélectionné. Posons que $r = E(\tau|n) = n\bar{M}$ et $t_{\text{HH}} = t_{\text{NS}} = t$, et il s'ensuit que les tailles prévues des échantillons de transactions des estimateurs NS et HH conditionnelles sur l'ensemble des échantillons de n ménages sont équivalentes, à savoir $E(m_{\text{HH}}|n) = t E(\tau|n) = nt\bar{M} = E(m_{\text{NS}}|n)$.

Un tel calage des tailles des échantillons d'établissements et de transactions assure que les enquêtes HH et NS auprès des établissements soient réalisées dans des conditions à peu près équivalentes de contraintes financières si les coûts par établissement et par transaction sur le terrain sont à peu près les mêmes pour les deux enquêtes. Il convient toutefois de mentionner que cette équation de coût ne tient pas compte des différences entre les coûts de création et de tenue à jour des listes d'établissements autonomes et des listes d'établissements produites d'après une enquête démographique.

Si l'on substitue $r = n\bar{M}$, $t_{\text{HH}} = t_{\text{NS}} = t$ et $M = N\bar{M}$ dans la formule (9), la différence entre les variances NS et HH pour des tailles prévues équivalentes d'échantillons d'établissements et de transactions conditionnelles sur l'ensemble des échantillons de n de ménages est

$$\begin{aligned} \text{Var}(X'_{\text{NS}}) - \text{Var}(X'_{\text{HH}}) &= \frac{N^2}{n} \left[\sigma_{\text{NS1}}^2 - \bar{M} \sigma_{\text{HH1}}^2 \right] \\ &- \frac{N}{nt} \sum_{j=1}^R \sigma_j^2 \left[(M_j - t) - \sum_{i=1}^N \frac{M_{ij}(M_j - M_{ij})}{M_j} \right]. \quad (11) \end{aligned}$$

Les premier et deuxième termes du deuxième membre de l'équation (11) représentent, respectivement, la différence entre les composantes de premier et de deuxième degrés des variances des estimateurs NS et HH pour des tailles prévues équivalentes d'échantillons conditionnelles sur l'ensemble des échantillons de n ménages.

4.2 Décomposition de la variance de population NS à un seul degré

Habituellement, certains ménages ne font de transaction avec aucun établissement et la proportion varie selon le type d'établissement. Par exemple, aux États-Unis, l'utilisation des soins médicaux par les familles varie fortement selon la catégorie de fournisseurs de soins de santé (Dicker et Sunshine 1987). Pour une période de 12 mois, 70 % de familles n'ont pas déclaré d'hospitalisation, 7 % n'ont pas déclaré de visite à un médecin en consultation externe et 28 % n'ont pas déclaré de visite chez un dentiste.

Soit

$P = \frac{N^*}{N}$ = fraction de N ménages comptant au moins une transaction,

$P_0 = 1 - P \frac{N_0}{N}$ = fraction de N ménages sans aucune transaction.

Nous démontrons à l'annexe que la variance de population à un seul degré de l'estimateur NS de X , si on l'exprime sous forme de fonction de P , se décompose en deux parties

$$\sigma_{NS1}^2(P) = \frac{1}{N} \sum_{i=1}^N \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N} \right)^2 = P\sigma_{NS1^*}^2 + \sigma^2(P)E_{NS1^*}^2, \quad 0 < P \leq 1 \quad (12)$$

où

$$\sigma_{NS1^*}^2 = \frac{1}{N^*} \sum_{i=1}^{N^*} \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2 \quad (13)$$

représente la variance de population à un seul degré de la variable x sur la population tronquée de N^* ménages comptant au moins une transaction,

$$E_{NS1^*}^2 = \left(\frac{X}{N^*} \right)^2 = \frac{1}{N^*} \sum_{i=1}^{N^*} \left(\sum_{j \in A_i} M_{ij} \bar{X}_j \right)^2 - \sigma_{NS1^*}^2 \quad (14)$$

représente le carré de l'espérance de la variable x sur la population tronquée de N^* ménages et

$$\sigma^2(P) = P(1-P) \quad (15)$$

est la variance de la variable binômiale P . Pour une valeur fixe de M , la fonction $\sigma_{NS1}^2(P|M)$ est maximale quand

$$P = P_{\max} = \frac{1}{2} \left[\left(\frac{\sigma_{NS1^*}^2}{E_{NS1^*}^2} + 1 \right) \right] \leq 1.$$

Si $\sigma_{NS1^*}^2 \geq E_{NS1^*}^2$, $P_{\max} = 1$ et si $\sigma_{NS1^*}^2 < E_{NS1^*}^2$, $1/2 < P_{\max} < 1$.

Quand $P = 1$, $\sigma^2(P = 1) = 0$ et, par conséquent, $\sigma_{NS1}^2(P = 1) = \sigma_{NS1^*}^2$. Si $P = \bar{M} = (M/N) = 1$, ce qui sous-entend que chacun des N ménages compte une seule transaction,

$$\sigma_{NS1}^2(P = \bar{M} = 1) = \sigma_{NS1^*}^2(N^* = M) = \sigma_{HH1}^2 \quad (16)$$

car

$$\begin{aligned} \sigma_{NS1^*}^2(N^* = M) &= \frac{1}{N^*} \sum_{i=1}^{N^*} \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2 \\ &= \frac{1}{M} \sum_{j=1}^R M_j \left(\bar{X}_j - \frac{X}{M} \right)^2 = \sigma_{HH1}^2, \end{aligned} \quad (17)$$

et $\sigma^2(P = 1) = 0$. Autrement dit, si $P = \bar{M} = 1$, ce qui sous-entend que chacun des N ménages compte une seule transaction, la variance de l'estimateur NS1, qui dépendrait alors de l'EAS de transactions avec remise, est équivalente à la variance de l'estimateur HH1, qui dépend d'un échantillon en grappe PPT de taille équivalente sélectionné avec remise.

4.3 Effet de plan de sondage en cas d'échantillonnage à un seul degré

Soit

$$X'_{NS1} = \frac{N}{n} \sum_{i=1}^N \sum_{j \in A_i} M_{ij} \bar{X}_j = \text{l'estimateur NS non biaisé}$$

de X en cas d'échantillonnage à un seul degré,

$$X'_{HH1} = \frac{M}{R_{HH}} \sum_{j=1}^{r_{HH}} \bar{X}_j = \text{l'estimateur HH non biaisé de } X$$

en cas d'échantillonnage à un seul degré.

Définissons l'effet de plan de sondage total de l'échantillonnage à un seul degré pour l'estimateur NS1 comme étant le rapport des variances des estimateurs NS1 et HH1 pour des tailles équivalentes d'échantillons conditionnelles sur l'ensemble des échantillons de n ménages.

$$\lambda(P) = \frac{\text{Var}(X'_{NS1})}{\text{Var}(X'_{HH1})} = \frac{\sigma_{NS1}^2(P)}{\bar{M} \sigma_{HH1}^2} \quad (18)$$

où $\lambda(P) < 1$ indique que l'estimateur NS1 est plus efficace que l'estimateur HH1, et $\lambda(P) > 1$ indique que l'estimateur HH1 est plus efficace que l'estimateur NS1.

Nous avons noté dans (12) et (15) que $\sigma_{NS1}^2(P) = P\sigma_{NS1^*}^2 + P(1-P)(X/N^*)^2$, et dans (16), que $\sigma_{HH1}^2 = \sigma_{NS1^*}^2(N^* = M)$. Si nous faisons ces substitutions dans (18), l'effet total de plan de sondage devient

$$\lambda(P) = \text{deft}_{NS1}^2 + (1-P) Z_{NS1}, \quad 0 < P \leq 1 \quad (19)$$

où

$$Z_{NS1} = \frac{P(X/N^*)^2}{\bar{M} \sigma_{NS1^*}^2(N^* = M)} \quad (20)$$

est l'effet attribuable aux N_0 ménages ne comptant aucune transaction, et

$$\text{deft}_{NS1}^2 = \left[\frac{P \sigma_{NS1^*}^2}{\bar{M} \sigma_{HH1}^2} \right] = \left[\frac{P \sigma_{NS1^*}^2}{\bar{M} \sigma_{NS1^*}^2(N^* = M)} \right] \quad (21)$$

est l'effet attribuable aux N^* ménages comptant des transactions. Autrement dit, deft_{NS1}^2 est l'effet de plan de sondage de l'échantillonnage en réseau d'une population de N^* grappes de ménages comptant au moins une transaction, avec probabilités égales et remise, comparativement à l'échantillonnage en réseau d'une population de M

transactions donnant lieu à une taille prévue d'échantillon équivalente, par EAS avec remise. [Le lecteur consultera Kish (1982) pour la définition de deft^2].

Dans (19), l'effet total de plan de sondage dépend de $\text{deft}_{\text{NS}}^2$, de Z_{NS1} et de P , et les valeurs de ces paramètres, ainsi que les relations entre eux, varient probablement considérablement d'une enquête à l'autre et d'une variable et d'un domaine de population à l'autre lors d'une même enquête. Bien que, théoriquement, l'estimateur NS1 puisse être plus efficace que l'estimateur HH1, en pratique, ce résultat paraît fort improbable, parce que l'échantillonnage en grappes est habituellement moins efficace que l'échantillonnage aléatoire simple. Une condition nécessaire pour que l'estimateur NS1 soit aussi efficace ou plus efficace que l'estimateur HH1 est que $\text{deft}_{\text{NS1}} \leq 1 - (1 - P)Z_{\text{NS1}}$, et cette condition n'est vraisemblablement pas satisfaite, particulièrement si la valeur de P est faible, et si le regroupement des transactions dans les ménages est principalement dû aux ménages faisant des transactions multiples avec un même établissement plutôt qu'aux ménages faisant des transactions avec plusieurs établissements.

4.4 Comparaison des efficacités en cas d'échantillonnage à deux degrés

Dans le cas de l'échantillonnage à deux degrés, la différence entre les composantes de deuxième degré des variances HH et NS pour des tailles d'échantillon prévues équivalentes de $nt\bar{M}$ transactions conditionnelles sur l'ensemble des échantillons de n ménages, c'est-à-dire le deuxième terme du deuxième membre de l'équation (11), se réduit à

$$\begin{aligned} \frac{N}{nt} \sum_{j=1}^R \sigma_j^2 \left\{ (M_j - t) - \sum_{j=1}^N \frac{M_{ij}(M_j - tM_{ij})}{M_j} \right\} \\ = \frac{N}{n} \sum_{j=1}^R \frac{\rho_j}{M_j} \sigma_j^2 \end{aligned} \quad (22)$$

où $\rho_j/M_j = 1/M_j \sum_{i=1}^N M_{ij}(M_{ij} - 1)$ est la différence entre les corrections HH et NS de deuxième degré pour une population finie pour l'établissement j . Si aucun des N ménages ne compte plusieurs transactions avec l'établissement j , les variances HH et NS de deuxième degré pour l'établissement j sont équivalentes et $\rho_j = 0$. Sinon, $\rho_j > 0$ et la variance de deuxième degré pour l'établissement j est plus grande pour l'estimateur HH que pour l'estimateur NS. La valeur de ρ_j est maximale si l'établissement j compte M_j transactions avec un seul ménage.

Les composantes de deuxième degré de la variance des estimateurs HH et NS sont équivalentes, c'est-à-dire $\sum_{j=1}^R \rho_j = 0$, lorsqu'aucun des H ménages ne fait de transactions multiples avec aucun des R établissements. Naturellement, les variances de deuxième degré sont équivalentes si les transactions sont sélectionnées avec remise ou que la variance à l'intérieur d'un établissement est

$\sigma_j^2 = 0 (j = 1, 2, \dots, R)$. Cependant, à part ces contingences, la variance de deuxième degré est toujours plus importante pour l'estimateur HH que pour l'estimateur NS, et la grandeur de la différence dépend de l'importance du regroupement des transactions avec un même établissement à l'intérieur des ménages, et de l'importance des variances à l'intérieur des établissements.

Si aucun des N^* ménages ne fait plusieurs transactions avec le même établissement, la différence entre les variances des estimateurs HH et NS est la même pour les enquêtes sur échantillon d'établissements à un seul degré et à deux degrés. Sinon, la différence entre les variances HH et NS est plus faible pour l'échantillonnage à deux degrés que celui à un seul degré des établissements, car, quand les ménages font des transactions multiples avec le même établissement, la variance de second degré est plus importante pour l'estimateur HH que pour l'estimateur NS.

5. Résumé et conclusion

Le modèle d'erreur présenté ici permet de comparer l'efficacité de deux estimateurs du volume de transactions entre les établissements et les particuliers dans le cas d'enquêtes sur échantillons à un degré et à deux degrés auprès des établissements. L'estimateur de Hansen-Hurwitz (HH) est fondé sur une base de sondage autonome énumérant les établissements et, pour chacun, le volume de ses transactions avec tous les ménages durant une période particulière de l'année civile. L'estimateur par échantillonnage en réseau (NS pour *network sampling*) dépend d'une base de sondage établie d'après une enquête démographique qui fournit la liste des ménages et de leurs probabilités de sélection lors d'une enquête démographique par sondage et, pour chaque ménage, donne le nombre de transactions avec chaque établissement durant la période précisée de l'année civile.

En outre, les estimateurs NS et HH sont basés sur des plans de sondage différents. En cas d'échantillonnage à un seul degré, l'estimateur HH dépend d'un plan de sondage où les établissements sont les unités d'échantillonnage sélectionnées avec PPT et avec remise, et l'estimateur NS dépend d'un plan de sondage où les ménages sont les unités d'échantillonnage que l'on sélectionne avec la même probabilité que dans l'enquête démographique, enquête que l'on suppose, dans le modèle d'erreur, être réalisée par EAS avec remise. Dans le cas de l'échantillonnage à deux degrés, les transactions sont les unités d'échantillonnage de deuxième degré des estimateurs HH et NS. L'estimateur HH est basé sur des échantillons de transactions de taille fixe qui sont sélectionnés par EAS indépendamment, sans remise. L'estimateur NS est basé sur des échantillons de transactions dont la taille est proportionnelle au nombre de transactions que fait chaque ménage avec chaque établissement et qui sont sélectionnés indépendamment par EAS sans remise.

Les estimateurs NS et HH sont aussi efficaces l'un que l'autre si, et uniquement si, chaque ménage de la population complète fait une seule transaction. Sinon, ni l'estimateur NS ni l'estimateur HH n'est nécessairement plus efficace que l'autre. Néanmoins, il semble probable que l'estimateur HH soit plus efficace que l'estimateur NS en cas d'échantillonnage à un seul degré des établissements, voire même considérablement plus efficace, particulièrement lorsque des fractions importantes de ménages ne font aucune transaction et (ou) lorsque le regroupement des transactions à l'intérieur des ménages est dû principalement aux ménages qui font plusieurs transactions avec le même établissement plutôt qu'aux ménages qui font des transactions avec plusieurs établissements. Dans le cas de l'échantillonnage à deux degrés, le résultat n'est pas aussi évident que dans le cas de l'échantillonnage à un degré, car la composante de deuxième degré de la variance de l'estimateur HH excède celle de l'estimateur NS d'une valeur qui dépend de l'importance du regroupement des transactions multiples avec un même établissement à l'intérieur des ménages.

On pourrait soutenir que la principale limite du modèle d'erreur présenté ici est la présomption que les bases de sondage autonome et établie d'après une enquête démographique ne comportent aucune erreur de couverture ni de mesure de taille. Toutefois, les coûts comparatifs de la création et de la tenue à jour d'une liste d'établissements autonome ou établie d'après une enquête démographique varient vraisemblablement fortement d'une enquête à l'autre. Quoique le modèle vise à égaliser les coûts des enquêtes auprès des établissements fondées sur chaque catégorie de bases de sondage, il ne tient pas compte des coûts différents de création et de tenue à jour de chacune des catégories de bases de sondage.

Même si l'on ne dispose pas de données empiriques sur les coûts comparatifs de la création et de la tenue à jour des bases de sondage, il est juste de dire que la base de sondage établie d'après une enquête démographique devrait être considérée sérieusement comme plan de sondage de rechange éventuel lorsque la création et la tenue à jour de bases de sondage indépendantes de bonne qualité est impossible, exorbitante ou demande trop de temps, et (ou) lorsque la création et la tenue à jour de listes d'établissements établies d'après une enquête démographique est relativement peu coûteuse. Par exemple, la base de sondage établie d'après une enquête démographique serait particulièrement intéressante comme remplacement éventuel de la base de sondage indépendante lorsque cette dernière est difficile à créer et à tenir à jour à cause de changements rapides dus à la création, à la fermeture et à la fusion d'établissements, et que le coût de la base de sondage établie d'après une enquête démographique est assez faible parce qu'elle peut être créée et tenue à jour sous forme de sous-produit d'une enquête démographique par sondage en cours (Wunderlich 1992) et (ou) sous forme de sous-produit d'un programme permanent d'appariement des transactions des ménages recensés lors d'une enquête

démographique avec les enregistrements des établissements avec lesquels ils font ces transactions (Cohen 1998).

Une autre limite du modèle tient à l'hypothèse irréaliste selon laquelle l'enquête démographique à partir de laquelle est produite la liste d'établissements est fondée sur un plan d'échantillonnage à un degré en vertu duquel les ménages sont sélectionnés avec probabilité égale et avec remise. En fait, les enquêtes démographiques sont presque toujours fondées sur un plan d'échantillonnage à plusieurs degrés selon lequel les ménages sont sélectionnés sans remise à l'étape finale d'échantillonnage. Habituellement, l'hypothèse que la sélection se fait par EAS a tendance à produire une estimation considérablement sous-estimée de la variance de l'estimateur NS et, par conséquent, à pour effet d'exagérer l'efficacité relative de cet estimateur comparativement à l'estimateur HH. Par contre, l'hypothèse selon laquelle les ménages sont échantillonnés avec remise a les effets opposés, mais ceux-ci sont modérés (Sirken 2001) comparativement à ceux de l'hypothèse d'EAS. Toutefois, le modèle d'erreur peut être appliqué aux autres plans d'échantillonnage, non considérés ici, que l'on utilise pour réaliser les enquêtes démographiques.

Le modèle d'erreur présenté précise les paramètres critiques qui déterminent l'efficacité relative des estimateurs établis pour les enquêtes auprès des établissements basées sur des bases de sondage autonomes, d'une part, et établies d'après une enquête démographique, d'autre part. La valeur de ces paramètres varie considérablement d'une enquête à l'autre et d'une variable ou d'un domaine de population à l'autre dans le cas d'une même enquête. Malheureusement, on ne dispose à l'heure actuelle d'aucune donnée empirique et il serait fort important d'en obtenir pour estimer les paramètres du modèle dans des conditions d'enquêtes très variées. Le présent article, faut-il l'espérer, suscitera un intérêt pour la réalisation d'enquêtes auprès des établissements basées sur des listes d'établissements établies d'après une enquête démographique et mènera à des améliorations de la conception des enquêtes auprès des établissements visant à estimer le volume de transactions entre les établissements et les populations.

Remerciements

L'auteur remercie les deux évaluateurs et, surtout, un rédacteur adjoint, pour leurs commentaires très constructifs. Les opinions exprimées dans le présent article n'engagent que l'auteur et ne représentent pas nécessairement les vues ou positions officielles du National Center for Health Statistics.

Annexe

Si on l'exprime sous forme de fonction de P , c'est-à-dire la fraction de ménages comptant au moins une transaction, la variance de population de premier degré de l'estimateur par échantillonnage en réseau (NS) de X

$$\sigma_{NSI}^2 = \frac{1}{N} \sum_{i=1}^N \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N} \right)^2$$

se décompose en deux parties

$$\sigma_{NSI}^2(P) = P\sigma_{NSI^*}^2 + \sigma^2(P)E_{NSI^*}^2 \quad 0 < P \leq 1$$

où

$$P = \frac{N^*}{N}$$

$$\sigma_{NSI^*}^2 = \frac{1}{N^*} \sum_{i=1}^{N^*} \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2$$

est la variance de la population tronquée de premier degré de l'estimateur NS, à l'exclusion des $N_0 = N - N^*$ ménages ne comptant aucune transaction avec les établissements,

$$\sigma^2(P) = P(1 - P)$$

est la variance de la variable binômiale P , et

$$E_{NSI^*}^2 = (X/N^*)^2$$

est le carré de l'espérance de la variable x répartie sur les N^* ménages.

Preuve

$$\begin{aligned} \sigma_{NSI}^2 &= \frac{1}{N} \sum_{i=1}^N \left(\sum_{j \in A_i} M_{ij} \bar{X}_j - \frac{X}{n} \right)^2 \\ &= \frac{1}{N} \sum_{i=1}^{N^*} \left(\sum_{j=1}^R M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2 + \frac{N_0}{N} \left(\frac{X}{N} \right)^2. \quad (A.1) \end{aligned}$$

Ajoutons X/N^* au premier terme du deuxième membre de (A.1) et soustrayons l'en.

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^{N^*} \left(\sum_{j=1}^R M_{ij} \bar{X}_j - \frac{X}{n} \right)^2 \\ &= \frac{P}{N^*} \sum_{i=1}^{N^*} \sum_{j \in A_i} \left(M_{ij} \bar{X}_j - \frac{X}{N^*} \right)^2 + P \left(\frac{X}{N^*} - \frac{X}{N} \right)^2 \\ &= P\sigma_{NSI^*}^2(P) + P \left(\frac{X}{N^*} - \frac{X}{N} \right)^2. \quad (A.2) \end{aligned}$$

Remplaçons le premier terme du deuxième membre de (A.1) par (A.2).

$$\begin{aligned} \sigma_{NSI}^2(P) &= P\sigma_{NSI^*}^2 + P \left(\frac{X}{N^*} - \frac{X}{N} \right)^2 + (1 - P) \left(\frac{X}{N} \right)^2 \\ &= P\sigma_{NSI^*}^2(P) + \sigma^2(P)E_{NSI^*}^2 \quad (A.3) \end{aligned}$$

où

$$\sigma^2(P) = P(1 - P), \text{ et } E_{NSI^*}^2 = \left(\frac{X}{N^*} \right)^2.$$

Bibliographie

- Cohen, S.B. (1998). Sample design of the 1996 medical expenditure panel survey medical provider component. *Journal of Economic and Social Measurement*, 24, 25-53.
- Dicker, M., et Sunshine, J.H. (1987). Family use of health care, United States, 1980. *National Health Care Utilization and Expenditure Survey*. Rapport No. 10. DHHS Pub. 87-20210.
- Judkins, D., Berk, M., Edwards, S., Mohr, P., Stewart, K. et Waksberg, J. (1995). National Health Care Survey: List versus Network Sampling, Rapport non-publié. National Center for Health Statistics.
- Judkins, D., Marker, D., Waksberg, J., Botman, S. et Massey, J. (1999). National Health Interview Survey: Research for the 1995-2004 redesign. National Center for Health Statistics. *Vital and Health Statistics*. Washington, DC: Government Printing Office, Series 2. 126, 76-89.
- Kish, L. (1982). Design effect. *Encyclopedia of the Statistical Sciences*. John Wiley & Sons, Inc. 2, 347-348.
- Leaver, S., et Valliant, R. (1995). Statistical problems in estimating the U.S. consumer price index. Dans *Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott). New York: John Wiley & Sons, Inc.
- Massey, L.T., Moore, T.F., Parsons, V. et Tadro, W. (1991). Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics, *Vital and Health Statistics*. Washington, DC: Government Printing Office, Series 2, 110.
- Sirken, M., et Shimizu, I. (1999). Enquêtes auprès des établissements fondées sur un échantillon représentant de ménages : l'estimateur de Horvitz-Thompson. *Techniques d'enquête*, 25, 213-218.
- Sirken, M., Shimizu, I. et Judkins, D. (1995). The population based establishments surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1, 470-473.
- Sirken, M.G. (1997). Network sampling. *Encyclopedia of Biostatistics*. John Wiley & Sons, Inc. 4, 2977-2986.
- Sirken, M.G. (2001). The Hansen-Hurwitz estimator revisited: PPS sampling without replacement. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Sous presse.
- Thompson, S. (1992). *Sampling*. New York: John Wiley & Sons, Inc. 117-118.
- Wunderlich, G.S. (Ed.) (1992). *Toward a National Health Care Survey: A Data System for the 21st Century*. Washington, DC: National Academy Press.