Techniques d'enquête, juin 1990
Vol. 16, n° 1, pp. 113–125
Statistique Canada

Estimation au moyen d'un échantillonnage double et d'une stratification duale

DONALD B. WHITE[1]

RÉSUMÉ

Nous cherchons ici à estimer le total d'une population finie qui est stratifiée à deux niveaux: un niveau secondaire, caractérisé par une faible variabilité intrastrate mais inconnu avant la première phase du sondage, et un niveau préliminaire (stratification a priori), qui permet de prévoir avec assez d'efficacité, pour chaque unité, les résultats de la stratification a posteriori. Dans un sondage à deux phases visant à tenir compte de la non-réponse, par exemple, la stratification a posteriori peut servir à diviser une sous-population en deux groupes: répondants et non-répondants. Nous appliquons les estimateurs de Vardeman et Meeden (1984) suivant diverses hypothèses concernant le genre d'information préalable utilisée. Au moyen d'une simulation, nous analysons l'erreur type de ces estimateurs en regard de celle des méthodes courantes. Lorsqu'il n'existe pas d'information préalable et que l'on a recours à l'échantillonnage proportionnel, l'estimateur est non biaisé et sa variance est calculée par approximation. Dans ce cas, la variance est toujours inférieure à celle de l'échantillonnage double ordinaire pour stratification. Par ailleurs, lorsqu'il n'y a pas d'information préalable mais que l'on a recours à l'échantillonnage non proportionnel, il est possible, moyennant une légère modification du plan d'échantillonnage de la seconde phase, de déterminer un estimateur non biaisé de même que la variance correspondante, un estimateur non biaisé de cette variance et un mode de répartition optimale pour les deux phases du sondage. Enfin, nous examinons des applications de ces méthodes.

MOTS CLÉS: Sondage à deux phases; information préalable; estimation de la variance; répartition optimale; non-réponse.

1. INTRODUCTION

Les plans d'échantillonnage stratifié n'utilisent pas tous le même genre d'information préalable. Par exemple, le modèle de stratification habituel suppose que nous disposons de toute l'information voulue sur les membres de chaque strate. La stratification a posteriori est utile lorsque nous disposons d'information globale sur les tailles de strates mais que nous n'avons aucune information sur les unités proprement dites. Par ailleurs, l'échantillonnage double pour stratification suppose qu'il n'existe aucune information préalable sur les strates. En outre, il est nécessaire d'avoir une certaine connaissance des valeurs de la population si l'on veut, par exemple, répartir entre les strates les ressources consacrées à l'échantillonnage (voir, par exemple, Cochran, 1977, p. 96-99 et 331-332).

Les hypothèses rigoureuses qui sont propres à ces plans de sondage et à ces modèles de population ne sont pas toujours vérifiées à cause de la divergence qui peut exister entre la population à l'étude et l'information préalable (parfois périmée). Cherchant à résoudre cet écart, Vardeman et Meeden (1984) ont défini deux estimateurs qui combinent les données relatives aux membres des strates et aux tailles et moyennes de strates avec les données correspondantes pour l'échantillon courant. Les deux estimateurs sont utilisés dans des circonstances fondamentalement différentes. Le premier est utilisé lorsque l'information préalable est uniquement globale, c'est-à-dire lorsqu'elle ne porte que sur les tailles et les moyennes de strates, tandis que le second estimateur est utilisé lorsqu'on dispose en plus de données partielles sur les

[1] Donald B. White, Département de statistique State University of New York, Buffalo 249 Farber Hall Buffalo, New York 14214.

membres de chaque strate. Dans ce dernier cas, la population est stratifiée selon divers facteurs, dont certains sont connus et d'autres non; ces derniers peuvent toutefois être déterminés sans trop de frais dans une première phase de sondage.

Par exemple, nous pouvons évaluer la propagation d'une maladie infectieuse en ayant recours à l'échantillonnage. Si le dépistage de la maladie est une opération coûteuse, il est sage de recourir à la stratification (par catégorie de risque) pour réduire la taille de l'échantillon de la seconde phase. Le sexe, l'âge, le lieu de résidence, l'origine ethnique, les habitudes d'hygiène et le contact avec des porteurs de la maladie sont tous des facteurs qui peuvent déterminer les catégories de risque. Comme certains de ces facteurs ne sont pas connus avant l'échantillonnage, on peut utiliser le modèle de Vardeman et Meeden puisqu'on peut déduire les véritables catégories de risque à l'aide des facteurs connus.

Prenons comme autre exemple le sondage à deux phases pour non-réponse. Si nous appliquons une version élargie de la méthode de Hansen et Hurwtiz (1946), nous avons une poulation qui est subdivisée en deux strates: les répondants et les non-répondants. Les méthodes que nous analysons ici s'appliquent lorsqu'on dispose d'information dans des strates de premier niveau, lesquelles servent ensuite à déterminer si une unité est susceptible ou non de faire partie du groupe des répondants.

L'utilisation d'information préalable dans les plans de sondage à deux phases est un sujet qui a déjà été traité dans des ouvrages de l'échantillonnage. Han (1973), par exemple, s'est servi de l'information préalable sur une variable de régression auxiliaire (qui devait être mesurée dans un échantillon de première phase) pour construire une hypothèse simple (disons H_0) concernant la moyenne de cette variable. Il a ensuite testé H_0. à l'aide des observations de l'échantillon de première phase. Si l'hypothèse H_0 était vérifiée, il se servait de la valeur définie par celle-ci H_0 dans l'estimateur; dans le cas contraire, il se servait de la moyenne de l'échantillon.

L'utilisation du premier estimateur de Vardeman et Meeden (information globale seulement) fait l'objet d'une analyse dans White (1987). Celui-ci s'attache à optimaliser le choix des constantes de pondération pour l'information préalable par rapport à l'information contenue dans l'échantillon courant. En ce qui concerne le présent article, nous envisageons le cas où il existe aussi de l'information préalable sur les membres des strates. Après avoir présenté la notation pertinente dans la section 2, nous analysons un exemple simulé dans la section 3. Par la suite, nous analysons des estimateurs non biaisés au point de vue de la variance, de l'estimation non biaisée de cette variance et de la répartition optimale des ressources affectées à l'échantillonnage dans deux contextes différents. Enfin dans la section 5, nous voyons des applications de ces méthodes.

2. MODÈLE DE POPULATION ET PLAN DE SONDAGE

Nous allons maintenant exposer le modèle de population et le plan de sondage proposé. Soit une population finie P d'unités identifiées 1, 2, ..., N et les valeurs inconnues correspondantes $y_1, y_2 \ldots, y_N$. Désignons le total de population par $\tau = \sum_{i=1}^{N} y_i$. Pour $1 \leq i \leq N$, l'unité i appartient à la strate de second niveau j_i, $1 \leq j_i \leq J$, que nous ne connaissons pas, et appartient en même temps à la strate de premier niveau (strate préliminaire) k_i, $1 \leq k_i \leq K$, que nous connaissons.

Il y a des paramètres de population qui exigent une notation spéciale. Ces paramètres sont la taille, la moyenne et la variance des groupes. Ceux-ci sont identifiés par les indices inférieurs pertinents: l'absence d'indice sert à désigner la population totale, "$k\cdot$." renvoie à la strate préliminaire k, $1 \leq k \leq K$, "$\cdot j$" renvoie à la strate de second niveau j, $1 \leq j \leq J$, et l'indice "kj" sert à désigner l'intersection de la strate préliminaire k et de la strate de second niveau j. Les symboles N, $\bar{Y}$, et S^2 représentent le nombre d'éléments, la moyenne et la variance de population finie respectivement. Enfin, désignons par P, $P_{k\cdot}$, $P_{\cdot j}$ et P_{kj} les sous-ensembles de P qui correspondent aux quatre groupes définis ci-dessus. Par exemple, nous avons

$$S_{k.}^2 = \frac{1}{N_{k.} - 1} \sum_{i \in P_{k.}} (y_i - \bar{Y}_{k.})^2.$$

Nous pouvons aussi écrire l'équation suivante:

$$\tau = \sum_j N_{.j} \bar{Y}_{.j}. \tag{1}$$

Enfin, soit $W_{kj} = N_{kj}/N_{k.}$, c.-à-d. que W_{kj} est la proportion d'unités de la strate préliminaire k qui appartiennent à la strate j.

Voyons maintenant la méthode d'échantillonnage. Dans la première phase, on tire sans remise un échantillon aléatoire simple stratifié s'; cet échantillon est constitué de $n'_{k.}$ unités (fraction de sondage de la première phase: $f'_{k.} = n'_{k.}/N_{k.}$) prélevées dans la strate préliminaire k. Les échantillons prélevés dans des strates préliminaires différentes sont indépendants. Pour ces $n' = \sum_k n'_{k.}$ unités, on observe des strates de second niveau j_i. Suivant la notation définie plus haut, n'_{kj} désigne le nombre d'unités de s' tirées de la strate préliminaire k, qui appartiennent aussi à la strate de second niveau j. Par ailleurs, $n'_{.j} = \sum_k n'_{kj}$ est le nombre total d'unités de s' qui font partie de la strate de second niveau j. Cet ensemble est désigné par $s'_{.j}$. Ces quantités sont connues tandis que celles qui reposent sur les valeurs y, comme $\bar{y}'$ et $s^{2'}$ (avec les quatre genres d'indices), ne le sont toujours pas. Pour les besoins de cet article, la moyenne d'un ensemble vide est définie comme nulle et la variance s^2 d'un groupe dont l'effectif est un ou zéro est aussi définie comme nulle. Notons que pour $1 \leq k \leq K$, les vecteurs aléatoires $(n'_{k1}, \ldots, n'_{kJ})$ sont indépendants et suivent chacun une distribution hypergéométrique multidimensionnelle.

Pour la seconde phase d'échantillonnage, nous décomposons s' en $\cup_{j=1}^{J} s'_{.j}$, c.-à-d. que nous procédons à une stratification a posteriori. Pour chaque j, posons $v_j(\cdot)$ comme une application connue dans et sur les nombres entiers positifs, où $v_j(0) = 0$ et $1 \leq v_j(x) \leq x$ si $x \geq 1$. L'échantillon de la seconde phase s est ausi stratifié mais il est un sous-ensemble de s' et est stratifié en fonction de la stratification a posteriori. L'échantillon prélevé dans $s'_{.j}$ est désigné par $s_{.j}$ et sa taille est $n_{.j} \equiv v_j(n'_{.j})$. À ce stade-ci, nous pouvons observer des valeurs y ainsi que les paramètres établis à l'aide de ces valeurs, $\bar{y}_{.j}$ et $s^2_{.j}$, qui sont la moyenne et la variance de population finie des unités de la strate j dans l'échantillon de la seconde phase.

Lorsqu'on calcule les valeurs estimées de τ selon la formule de Vardeman et Meeden, on peut faire intervenir des estimations préliminaires des tailles relatives de strates pour chaque strate préliminaire et des moyennes de strates. Ici, nous disposons d'estimations préliminaires pour les valeurs W_{kj} et $\bar{Y}_{.j}$ et ces estimations sont désignées par Π_{kj} et $\mu_{.j}$, respectivement. Dans l'estimateur défini plus bas, ces estimations sont affectées de constantes de pondération qui reflètent le rapport entre le degré de fiabilité de l'estimation et celui de l'information correspondante tirée de l'échantillon courant. On désigne par $\tilde{M}_{k.} \in [0,\infty]$ le degré de certitude attribué à l'ensemble $(\Pi_{k1}, \ldots, \Pi_{kJ})$ pour chaque k et par $M_{.j} \in [0,\infty]$ le degré de certitude attribué à $\mu_{.j}$, pour chaque j. Dans l'échantillon courant, l'ensemble $(W_{k1}, \ldots, W_{kJ})$ est estimé par $(n'_{k1}/n'_{k.}, \ldots, n'_{kJ}/n'_{k.})$ et repose sur un échantillon aléatoire simple de taille $n'_{k.}$. Par conséquent, le rapport entre la valeur de $\tilde{M}_{k.}$ et celle de $n'_{k.}$ par exemple reflétera le rapport entre le degré de fiabilité de Π_{kj} et celui de $n'_{kj}/n'_{k.}$. De même, dans l'échantillon courant, $\bar{Y}_{.j}$ est estimé par $\bar{y}_{.j}$ et repose sur un échantillon de taille $n_{.j}$; le rapport entre les valeurs de $M_{.j}$ et de $n_{.j}$ reflète donc, là aussi, le rapport entre le degré de fiabilité de l'estimation préliminaire et celui de l'estimation tirée de l'échantillon courant. Un poids de fiabilité (ou degré de fiabilité) nul signifie que l'on n'utilise pas l'information préalable et un poids égal à l'infini signifie, comme pour l'utilisation des tailles de strates dans le modèle de stratification a posteriori habituel, que l'on ne se sert pas de l'information correspondante de l'échantillon courant.

En nous servant des estimations préliminaires, des estimations de l'échantillon courant et des poids de fiabilité, nous estimons W_{kj} et $\bar{Y}_{.j}$ par $\hat{\Pi}_{kj} = (\tilde{M}_{k.}\Pi_{kj} + n'_{kj})/(\tilde{M}_{k.} + n'_{k.})$ et $\hat{\mu}_{.j} = (M_{.j}\mu_{.j} + n_{.j}\bar{y}_{.j})/(M_{.j} + n_{.j})$ respectivement. Enfin, nous construisons un estimateur $\hat{\tau}$ du total de population τ en remplaçant n'importe quelle quantité non observée de l'équation (1) par la valeur estimée correspondante, calculée ci-dessus. Ainsi, nous avons

$$\hat{\tau} = \sum_{j=1}^{J} \left\{ n_{.j}\bar{y}_{.j} + (n'_{.j} - n_{.j})\hat{\mu}_{.j} + \sum_{k=1}^{K} (N_{k.} - n'_{k.})\hat{\Pi}_{kj}\hat{\mu}_{.j} \right\}. \tag{2}$$

Vardeman et Meeden ne traitent pas le calcul du biais et de la variance de $\hat{\tau}$ pour le cas général. White (1987) examine le cas où $K = 1$ et $M_{.j} = 0$, $1 \leq j \leq J$. Avant d'analyser des cas plus complexes, nous allons examiner les résultats d'une simulation appliquée à une population hypothétique.

3. ÉTUDE DE MONTE CARLO

Pour les besoins de cette simulation, nous nous servons d'une population et d'un plan d'échantillonnage particuliers pour estimer, comme dans l'exemple du début, le taux de propagation d'une maladie infectieuse. Pour une population de 10,000 personnes qui sont prédisposées à cette maladie, on suppose que la prévalence est plus élevée parmi les 5,000 personnes qui demeurent dans la partie ouest de la région étudiée. Par conséquent, nous divisons la population en $K = 2$ strates (régions est et ouest). Ensuite, nous supposons que le sondeur est en mesure, grâce à de l'information supplémentaire obtenue facilement, de classer les personnes en fonction du degré de risque de contracter la maladie. Voir le tableau 1 pour les données détaillées de la population.

Pour l'estimation du nombre total de personnes infectées ($\tau = 2302$), nous supposons qu'il n'existe aucune information préalable sur les proportions de strates $\bar{Y}_{.1}$, $\bar{Y}_{.2}$ et $\bar{Y}_{.3}$ et nous posons par conséquent $M_{.1} = M_{.2} = M_{.3} = 0$. Nous devons considérer quatre autres éléments essentiels pour l'estimation: 1) les estimations préliminaires $\{\Pi_{kj}: k = 1, 2; j = 1, 2, 3\}$ concernant la répartition des personnes entre les strates de premier niveau et de second niveau, 2) les constantes de pondération $\tilde{M}_1$ et $\tilde{M}_2$ attribuées à ces estimations préliminaires, 3) le plan d'échantillonnage et l'échantillon effectif de la première phase et 4) le plan d'échantillonnage et l'échantillon effectif de la seconde phase. Ces quatre éléments sont exposés en détail ci-dessous.

En premier lieu, White (1987) a observé qu'une solution efficace pour $K = 1$ était de choisir une constante de pondération M égale à la taille de l'échantillon sur lequel était fondée l'information antérieure. Compte tenu de cette observation, nous avons prévu, pour chaque simulation, la formation d'un ensemble d'estimations préliminaires $\{\Pi_{kj}\}$ à partir d'un échantillon

Tableau 1

Nombre de cas de maladies/taille du groupe pour les strates de premier et de second niveau

Lieu de résidence \ Groupe de risque j	Faible 1	Moyen 2	Élevé 3	Total
Région est ($k = 1$)	40/4000	80/800	100/200	220/5000
Région ouest ($k = 2$)	2/200	80/800	2000/4000	2082/5000
Total	42/4200	160/1600	2100/4200	2302/10000

préliminaire de taille m (= 500 ou 2500) prélevé dans chaque strate de premier niveau. En d'autres termes Π_{kj} est défini comme la proportion des m unités de la strate de premier niveau k qui appartiennent aussi à la strate de second niveau j.

En deuxième lieu, nous avons défini les constantes de pondération comme suit pour chaque simulation: $\tilde{M}_{1\cdot} = \tilde{M}_{2\cdot} = M$ pour tous $M \in \{0, 100, 200, 300, \ldots, 10{,}000, \infty\}$ Il convient de rappeler que $M = \infty$ correspond au modèle de stratification a posteriori habituel, où on n'utilise pas du tout les données de l'échantillon courant pour estimer la taille des groupes.

En troisième lieu, l'échantillon de la première phase est stratifié suivant les strates préliminaires, les fractions de sondage $f'_{k\cdot}$ étant définies $f'_{1\cdot} = f'_{2\cdot} = f, f \in \{.10, .20, .30, .40, .50\}$. Rappelons que dans cette phase d'échantillonnage, seules les données relatives aux strates de second niveau sont observées. La collecte de ces données est une opération vraisemblablement peu coûteuse.

En revanche, l'échantillonnage d'une unité à la phase 2, où l'on détermine la présence du virus, est une opération qui est supposée assez coûteuse. Les unités échantillonnées forment un sous-échantillon de l'échantillon de la première phase, stratifié suivant les strates de second niveau. Là encore, les fractions de sondage sont définies comme étant égales ($v_j(n'_{\cdot j}) = [c_j n'_{\cdot j}]$ pour une valeur $n'_{\cdot j}$ suffisamment grande, et $c_1 = c_2 = c_3 = c$) et pour permettre la comparaison de simulations, on choisit la valeur de c de manière que la proportion de la population totale incluse dans l'échantillon de la phase 2 soit toujours la même (.10).

Le processus suivant est répété $R = 50{,}000$ fois: formation d'un échantillon préliminaire de taille m à partir duquel sont établies des estimations préliminaires Π_{kj} de W_{kj}. Ensuite, formation d'un échantillon stratifié suivant les strates de premier niveau avec fractions de sondage f. Seules les données relatives aux strates de second niveau sont observées. Ensuite, formation d'un sous-échantillon stratifié suivant les strates de second niveau avec fractions de sondage c et classement des unités de ce sous-échantillon en deux groupes: infectées et non infectées. Finalement, calcul de $\hat{\tau}$ pour chaque valeur de M considérée. L'erreur type de $\hat{\tau}$ est estimée au moyen des R valeurs simulées de $\hat{\tau}$. Rappelons toutefois que dans la réalité, l'erreur type d'une valeur estimée dépendra des valeurs de Π_{kj} utilisées. Dans la présente simulation, ces valeurs diffèrent à chaque itération et il vaut mieux ici considérer l'erreur type estimée comme une moyenne à long terme pour diverses distributions de $\hat{\tau}$ combinées suivant la distribution des valeurs Π_{kj} fondées sur l'échantillon préliminaire.

Les simulations ont été exécutées sur un ordinateur IBM3031. Pour cet exemple, où $y_i \in \{0,1\}$ pour tout i, toutes les quantités aléatoires sont des fonctions de variables hypergéométriques ou de variables hypergéométriques multidimensionnelles indépendantes. Compte tenu de ce que la distribution conditionnelle d'un élément d'une distribution hypergéométrique multidimensionnelle, étant donné n'importe quel sous-ensemble des autres coordonnées, est elle-même hypergéométrique, toutes les quantités aléatoires ont été simulées au moyen du sous-programme de simulation hypergéométrique GGHPR du IMSL 92DP. Pour la première combinaison de valeurs de m et f (500 et .10), on a répété la simulation cinq fois afin de vérifier la cohérence.

On retrouve dans les tableaux 2 et 3 des caractéristiques pertinentes de la variation de l'ET($\hat{\tau}$) simulée; dans le premier cas, il s'agit de la variation en fonction de M pour les cinq simulations répétées et dans le second cas, de la variation en fonction de M pour des simulations faites pour diverses valeurs de f et de m . Le tableau 2 ne donne que les résultats saillants de la simulation, qui confirment la cohérence et indiquent que le nombre d'itérations choisi est suffisammenté levé. Le tableau 3 permet d'établir une comparaison entre d'une part, les meilleures méthodes en usage actuellement (sondage à deux phases ordinaire ou échantillonnage stratifié suivant les strates de premier niveau) et d'autre part, la méthode idéale, c'est-à-dire celle où les strates réelles sont connues. L'erreur type d'un estimateur fondé sur un échantillon stratifié suivant les strates de premier niveau (strates préliminaires) seulement est de 113.27 tandis que celle d'un estimateur fondé sur un échantillon stratifié suivant les strates réelles est de 105.47. Par conséquent, si nous désignons par $\hat{\tau}_2$ l'estimateur fondé sur le sondage à deux phases ordinaire et que nous reconnaissons que ET($\hat{\tau}_2$) varie en fonction de f et de c, les valeurs

Tableau 2

Principaux résultats des simultations répétées $m = 500, f = .10$ et $c = 1.0$

Simulation #	M_0	ET($\hat{\tau}$)			
		$M = 0$	$M = m$	$M = M_0$	$M = \infty$
1	600	113.55	109.67	109.62	112.00
2	700	113.42	109.50	109.45	111.80
3	700	113.92	109.86	109.78	112.00
4	600	113.61	109.71	109.66	112.07
5	600	113.56	109.74	109.70	112.17

Tableau 3

Principales caractéristiques et de ET($\hat{\tau}$) comme fonction de M

					ET($\hat{\tau}$)				Réduction relative de ET($\hat{\tau}$) en pourcentage		
m	f'	c	ET($\hat{\tau}_2$)	M_0	$M = M_0$	$M = 0$	$M = m$	$M = \chi$	$M = 0$	$M = m$	$M = \infty$
500	.10	1.00	126.29	600	109.62	113.55	109.67	112.00	−3.6	46.2	16.3
500	.20	.50	115.19	600	107.95	109.02	107.97	110.72	54.5	67.9	32.7
500	.30	.33	111.80	600	107.87	108.25	107.87	110.38	56.1	62.1	22.4
500	.40	.25	109.22	750	106.51	106.76	106.52	108.29	65.6	72.0	24.8
500	.50	.20	107.98	700	106.17	106.28	106.18	107.55	67.7	71.7	17.1
2500	.10	1.00	126.29	*	≤ 106.20	113.33	106.42	106.20	−0.8	87.8	90.6
2500	.20	.50	115.19	*	≤ 105.76	108.67	106.02	105.76	59.0	92.9	96.3
2500	.30	.33	111.80	*	≤ 106.63	108.18	106.87	106.63	57.2	77.9	81.7
2500	.40	.25	109.22	*	≤ 105.77	106.59	105.94	105.77	70.1	87.5	92.0
2500	.50	.20	107.98	*	≤ 105.81	106.34	105.96	105.81	65.3	80.5	86.5

* -- > 10,000

qui figurent dans la colonne "Réduction relative de ET($\hat{\tau}$)" sont calculées au moyen de la formule 100 [min(ET($\hat{\tau}_2$), 113.27) − ET($\hat{\tau}$)] / [min(ET($\hat{\tau}_2$), 113.27) − 105.47].

Le tableau 3 nous permet de faire plusieurs constatations importantes. Premièrement, lorsque $m = 500$, M_0 se rapproche sensiblement de cette valeur tout en demeurant légèrement supérieur. C'est ce que prévoit White (1987) lorsqu'il analyse le cas où $K = 1$. Pour $m = 2500$, malgré que $M_0 > 10{,}000$ dans chaque cas, on remarque que ET($\hat{\tau}$) à $M = m$ se rapproche sensiblement de l'erreur type minimum à $M = M_0$.

Deuxièmement, lorsque $M = m$, la réduction relative de ET($\hat{\tau}$) varie de 46% à plus de 90%. En outre, lorsque $M = 0$, ce qui correspond à la stratification duale sans information préalable sur aucune caractéristique de la population, la réduction relative de ET($\hat{\tau}$) est toujours supérieure à 50% sauf pour ce qui a trait à la plus petite fraction de sondage de première phase, $f = .10$. Lorsque, pour $f = .10$, il n'y a pas d'information préalable et que l'échantillon de la première phase est petit, il est préférable d'utiliser les strates préliminaires et de ne pas tenir compte des strates réelles. En revanche, si l'on dispose d'un ensemble d'estimations préliminaires pour W_{kj} mais qu'on ne sait trop quels poids attribuer à ces valeurs, on peut recourir à la stratification a posteriori habituelle, où $M = \infty$. Si l'information préalable est juste (en l'occurrence, $m = 2500$), la réduction relative de ET($\hat{\tau}$) est toujours supérieure à 80%. Même lorsque l'information préalable n'est qu'à demi exacte (en l'occurrence, $m = 500$), la réduction relative de l'erreur type se situe entre 16 et 33%.

En résumé, s'il est possible d'appliquer une constante de pondération à des éléments d'information préalable sur la répartition des unités entre les strates, on peut réduire sensiblement l'erreur type de l'estimateur grâce aux méthodes décrites ci-dessus. Même si on ne peut trouver de constante de pondération ou qu'on ne dispose pas de l'information préalable voulue, il est encore possible de réduire l'erreur type de l'estimateur par la stratification duale en choisissant $M = 0$ si l'information préalable sur W_{kj} est maigre ou inexistante, ou $M = \infty$ si l'information préalable est exacte. Cela démontre l'importance du cas où $M = 0$. Celui-ci est analysé en détail dans la section suivante.

4. BIAIS, ERREUR TYPE ET RÉPARTITION OPTIMALE SANS INFORMATION PRÉALABLE

Lorsqu'il n'y a pas d'information préalable, nous posons $M_j = 0$ et $\tilde{M}_k = 0$ pour chaque $1 \leq j \leq J$, et chaque $1 \leq k \leq K$. Dans cette section, nous supposons aussi au départ que l'échantillon, dans les deux phases, est proportionnel à la taille du groupe d'où il est tiré, c'est-à-dire pour chaque k, $n'_{k.} = fN_{k.}$ (*c.-à-d.* $f'_{k.} = f$, pour tous k) et pour chaque j, $n_{.j} = cn'_{.j}$ (*c.-à-d.*, $v_j(x) = cx$, pour tous j). Cela implique automatiquement une approximation (désignée ci-dessous par A1) puisque la taille des échantillons ainsi obtenus n'est pas nécessairement un nombre entier. Toutefois, pour des populations assez grandes et des fractions de sondage f et c suffisamment élevées, cette approximation a peu d'effet sur les calculs qui vont suivre.

Dans le cas qui nous occupe, $\hat{\mu}_{.j}$ se ramène à $\bar{y}_{.j}$ et $\hat{\Pi}_{kj}$ à $n'_{kj}/n'_{k.}$; par conséquent, nous avons $\hat{\tau} = 1/f \sum_{j=1}^{J} n'_{.j}\bar{y}_{.j}$. Le calcul de l'espérance mathématique et de la variance de $\hat{\tau}$ est décrit en annexe. Dans cette annexe, nous posons deux arguments de condition; nous posons tout d'abord une condition en fonction de s' puisque l'échantillon de la seconde phase est une fonction de s' puis, étant donné la nature hypergéométrique multidimensionnelle de l'échantillon de la première phase, une autre condition en fonction de n'_{kj}, qui est le nombre d'unités de s' tirées de la strate préliminaire k, qui appartiennent aussi à la strate de second niveau j.

Nous montrons tout d'abord que l'estimateur $\hat{\tau}$ est non biaisé (abstraction faite de l'approximation A1) et que sa variance peut être calculée approximativement par la formule suivante:

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f} \sum_{k} N_{k.} S^2_{k.} + \frac{1-c}{fc} \sum_{j} N_{.j} S^2_{.j}. \tag{3}$$

Comme nous le voyons plus en détail dans l'annexe, l'équation (3) produit des valeurs proches des valeurs simulées, repose sur des approximations dont l'erreur est faible pour de grandes populations et des échantillons suffisamment grands, et correspond à l'équation exacte dans les trois cas courants. En outre, on peut montrer facilement que la variance calculée à l'aide de cette équation est toujours inférieure à celle calculée pour le sondage à deux phases ordinaire.

Comme dans n'importe quel modèle de stratification, il faut penser à un plan optimal. Il s'agit ici de déterminer une variance minimum pour un coût donné. À cette fin, posons $T_1 = \sum_k N_{k.} S^2_{k.}$ et $T_2 = \sum_j N_{.j} S^2_{.j}$. Pour les besoins de la cause, nous supposons que ces valeurs sont connues. En réalité, nous ne disposons bien sûr que d'estimations. Dans un deuxième temps, définissons D comme le budget total, d_0 comme les frais de mise en marche, d_1 comme le coût unitaire pour l'échantillon de la première phase et d_2 comme le coût unitaire pour l'échantillon de la seconde phase. Si nous définissons D_a comme le montant (en dollars) consacré à l'échantillonnage par unité de population, nous avons

$$D_a = \frac{D - d_0}{N} = f(d_1 + cd_2). \tag{4}$$

Les paramètres f et c devant satisfaire l'équation (4), nous cherchons à minimiser l'équation (3), var($\hat{\tau}$), maintenant définie

$$\text{var}(\hat{\tau}) \approx \frac{1 - f}{f} T_1 + \frac{1 - c}{fc} T_2. \tag{5}$$

Nous en arrivons assez facilement à la solution suivante:

$$c = \left[\frac{d_1 T_2}{d_2(T_1 - T_2}\right]^{1/2} \tag{6}$$

f étant calculé à l'aide de l'équation (4). Si $T_1 \leq T_2$, nous choisissons automatiquement $c = 1$ puisque dans les circonstances, la stratification a priori est plus efficace que la stratification a posteriori.

Dans le cas de l'échantillonnage non proportionnel, l'estimateur donné est biaisé et le calcul du biais et de l'erreur type de cet estimateur est très long. Toutefois, en modifiant légèrement le plan d'échantillonnage de la seconde phase et, du même coup, l'estimateur $\hat{\tau}$, on obtient un estimateur sans biais. Après avoir décrit les modifications requises, nous calculons la variance de l'estimateur ainsi qu'un estimateur non biaisé de cette variance et nous déterminons une méthode optimale pour répartir les ressources affectées à l'échantillonnage entre les strates de premier niveau et de second niveau.

La modification du plan d'échantillonnage consiste à effectuer l'échantillonnage de la seconde phase à l'intérieur des strates de premier niveau, plutôt que grouper les éléments des strates de premier niveau par strate de second niveau. Ainsi, étant donné n'_{kj} unités qui appartiennent à $s' \cap P_{kj}$, nous avons une fonction $v_{kj}(\cdot)$ (comme $v_{\cdot j}(\cdot)$ définie dans la section 2) qui appelle un échantillon aléatoire simple de taille $n_{kj} = v_{kj}(n'_{kj}) = c_{kj} n'_{kj}$ qui doit être prélevé dans $s' \cap P_{kj}$. À partir de cet échantillon, nous calculons les valeurs $\bar{y}_{kj}$ et s^2_{kj}, qui ont été définies dans la section 2. La formule de l'estimateur est maintenant $\hat{\tau} = \sum_k 1/f'_{k\cdot} \sum_j n'_{kj} \bar{y}_{kj}$.

Or, comme les échantillons (et par voie de conséquence les estimateurs) sont indépendants d'une strate préliminaire à l'autre, $\hat{\tau}$ est la somme des estimateurs indépendants des K totaux de strates de premier niveau, où chaque estimateur repose sur un sondage à deux phases ordinaire. Ainsi, les résultats de Rao (1973) s'appliquent à chaque strate de premier niveau et nous constatons en premier lieu que $\hat{\tau}$ est non biaisé parce que ses termes sont des estimateurs non biaisés des totaux de strates de premier niveau correspondants. En deuxième lieu, si nous nous servons des résultats de Rao, nous avons

$$\text{var}(\hat{\tau}) = \sum_k \frac{1}{f_{k\cdot}} \left[(N_{k\cdot} - n'_{k\cdot}) S^2_{kj} + \sum_j N_{kj} S^2_{kj}(1/c_{kj} - 1)\right]. \tag{7}$$

En outre, un estimateur non biaisé de var($\hat{\tau}$) est défini par l'expression:

$$\hat{\text{var}}(\hat{\tau}) = \sum_k N_{k\cdot} \left[(N_{k\cdot} - 1) \sum_j \left(\frac{n'_{kj} - 1}{n'_{k\cdot} - 1} - \frac{n_{kj} - 1}{N_{k\cdot} - 1}\right) \frac{n'_{kj} s^2_{kj}}{n'_{k\cdot} n_{kj}}\right.$$

$$+ \frac{N_{k\cdot} - n'_{k\cdot}}{N_{k\cdot}(n'_{k\cdot} - 1)} \sum_j \frac{n'_{kj}}{n'_{k\cdot}} \left(\bar{y}_{kj} - \sum_{j'} \frac{n'_{kj'}}{n'_{k\cdot}} \bar{y}_{kj'} \right)^2 \Bigg] . \tag{8}$$

Lorsque nous avons examiné le cas de l'échantillonnage proportionnel au début de cette section, nous avons proposé deux estimateurs pour τ, un fondé sur un échantillon de seconde phase combiné et l'autre, sur un échantillon non combiné. Dans les deux cas, il s'agissait d'un estimateur sans biais; de plus, lorsque $f'_{k\cdot} = f$ pour tous k et $c_{kj} = c$ pour tous k et tous j, l'équation (7) se ramène à l'équation (3), c'est-à-dire à la formule de la variance approximative de l'estimateur fondé sur un échantillon de seconde phase combiné.

Enfin, nous servant une fois de plus des résultats de Rao, nous déterminons un plan optimal pour répartir les ressources consacrées à l'échantillonnage. Supposons que nous disposons de D dollars pour les deux phases du sondage; il en coûte $d'_{k\cdot}$ dollars pour échantillonner une unité dans $P_{k\cdot}$ à la phase 1 et $d_{\cdot j}$ dollars pour échantillonner une unité dans $P_{\cdot j}$ à la phase 2. Étant donné ces coûts, nous voulons déterminer les valeurs $f'_{k\cdot}$ et c_{kj} qui minimisent la variance de $\hat{\tau}$. En utilisant l'inégalité de Cauchy pour l'échantillon de la phase 2 dans chaque strate de premier niveau, nous constatons que, quelle que soit la valeur de $f'_{k\cdot}$, la fraction de sondage pour la strate de second niveau j est définie par l'expression

$$c_{kj} = S_{kj} \left(\frac{d'_{k\cdot}}{d_{\cdot j}(S^2_{k\cdot} - \sum_j W_{kj} S^2_{kj})} \right)^{1/2} . \tag{9}$$

Par ailleurs, le coût prévu effectif (pour les deux phases d'échantillonnage) pour chaque unité tirée de la strate de premier niveau k à la phase 1 est

$$d^{(e)}_{k\cdot} = d'_{k\cdot} + \sum_j W_{kj} c_{kj} d_j . \tag{10}$$

De ce point de vue, on peut considérer la première phase du sondage comme un échantillonnage stratifié ordinaire, où le coût (effectif) du prélèvement de l'unité dans $P_{k\cdot}$ est déterminé à l'aide de l'équation (10). Cochran (1977, p.97) définit la formule de répartition voulue pour la première phase:

$$\frac{n'_{k\cdot}}{n'} = \frac{N_{k\cdot} S_{k\cdot} / \sqrt{d^{(e)}_{k\cdot}}}{\sum_{k'} N_{k'\cdot} S_{k'\cdot} / \sqrt{d^{(e)}_{k'\cdot}}} \tag{11}$$

où

$$n' = \sum_k n'_{k\cdot} = D \sum_k \frac{N_{k\cdot} S_{k\cdot} / \sqrt{d^{(e)}_{k\cdot}}}{\sum_{k'} N_{k'\cdot} S_{k'\cdot} / \sqrt{d^{(e)}_{k'\cdot}}} . \tag{12}$$

Grâce aux modifications proposées par Rao, il existe désormais une solution pour les cas où l'un ou plusieurs des c_{kj} seraient supérieurs à un. On peut aussi modifier les résultats de la façon habituelle afin de réduire au maximum le coût d'échantillonnage lorsque la variance est connue.

5. APPLICATIONS

La méthode de stratification duale que nous venons de décrire peut être utilisée à deux niveaux. À un premier niveau, on peut recourir à l'échantillonnage double avec strates préliminaires sans utiliser d'information préalable sur les tailles ou les moyennes de strates. À un second niveau, plus complexe, pourvu que l'on dispose d'information préalable sur le nombre d'unités dans chaque strate de second niveau qui proviennent de chacune des strates de premier niveau et que l'on connaisse le degré de fiabilité de cette information, il sera possible de réduire davantage l'erreur type de l'estimateur grâce à cette information.

Cette méthode de sondage à deux phases et la méthode d'estimation correspondante pourraient être utilisées dans l'enquête nationale que l'on propose pour évaluer la propagation du VIH (syndrome d'immunodéficience acquise). La période d'incubation de la maladie, que l'on estime à 4.5 ans en moyenne (Lui et coll., 1986), rend cette enquête d'autant plus nécessaire; en revanche, elle peut être très difficile à réaliser à cause de facteurs psychosociaux et financiers. C'est pourquoi il est nécessaire de rechercher des méthodes qui permettent de réduire la taille de l'échantillon tout en conservant le même niveau de précision.

Allen (1984) présente des données qui permettent de croire qu'il est possible de répartir la population des États-Unis selon une série de facteurs qui peuvent servir à définir des catégories de risque. Parmi les facteurs connus, qui peuvent servir à définir les strates de premier niveau, notons l'âge, le sexe, la présence de certaines maladies, la nationalité, le statut d'immigrant et le lieu géographique. Parmi les facteurs inconnus, qui peuvent être déterminés au moyen d'une interview, notons l'orientation sexuelle et la toxicomanie. Les données sur la prévalence du VIH dans divers sous-groupes peuvent aussi bien être incluses dans l'estimation globale de la prévalence que servir à déterminer la répartition de l'échantillon. On dispose de données de ce genre pour les donneurs de sang par exemple (Kuritsky et coll., 1986), les militaires (Redfield et Burke, 1987), les toxicomanes consommateurs de drogues injectables de Queens, New York (Robert- Guroff et coll., 1986) et les homosexuels de Greenwich Village (Casareale et coll., 1984/1985). Bien que cette information préalable puisse servir à réduire le coût d'échantillonnage et à accroître le niveau de précision, le problème de la confidentialité des tests ne se pose pas avec moins d'acuité; du reste, il faudra étudier ce problème sous tous ses angles avant de pouvoir tirer des résultats probants de cette étude.

REMERCIEMENTS

L'auteur tient à exprimer sa gratitude à l'arbitre pour les précieux commentaires que celui-ci lui a transmis en ce qui concerne l'échantillonnage non proportionnel.

ANNEXE

Calcul de l'espérance et de la variance sans information préalable (échantillonnage proportionnel)

À l'aide de la notation définie dans la section 2, nous calculons tout d'abord $E(\hat{\tau})$. L'espérance conditionnelle, étant donné s', est $E(\hat{\tau} \mid s') = 1/f \sum_j n'_{\cdot j} \bar{y}'_{\cdot j}$. En exprimant $n'_{\cdot j} \bar{y}'_{\cdot j}$ par $\sum_k n'_{kj} \bar{y}'_{kj}$, nous obtenons $E(\hat{\tau}) = E(E(\hat{\tau} \mid s')) = 1/f \sum_j \sum_k E\big(n'_{kj} E(\bar{y}'_{kj} \mid n'_{kj})\big) = 1/f \sum_j \sum_k E(n'_{kj}) \bar{Y}_{kj} = \tau$ puisque n'_{kj} est hypergéométrique, étant donné la fraction de sondage f et N_{kj} unités contenues dans la strate de premier niveau k et la strate de second niveau j. Par conséquent, $\hat{\tau}$ est non biaisé en l'occurrence (abstraction faite de l'approximation A1).

Le calcul de la variance s'effectue selon le même raisonnement sauf qu'il est beaucoup plus détaillé. Nous ne retiendrons ici que certaines étapes du calcul, plus particulièrement celles où il est question d'approximation. Tout d'abord, en appliquant les deux arguments de condition définis plus haut, nous obtenons

$$\text{var}(E(\hat{\tau} \mid s')) = \frac{1-f}{f} \sum_k N_k . S_{k.}^2 . \tag{13}$$

Nous obtenons ensuite

$$\text{var}(\hat{\tau} \mid s') = \frac{1-c}{f^2 c} \sum_j \frac{n'_{.j}}{n'_{.j}-1} \left[\sum_k (n'_{kj}-1) s'^2_{kj} + \sum_k n'_{kj} (\bar{y}'_{kj} - \bar{y}'_{.j})^2 \right]. \tag{14}$$

Les seconde et troisième approximations consistent à substituer un (1) à $n'_{.j}/(n'_{.j}-1)$ (A2) et n'_{kj} à $(n'_{kj}-1)$ (A3) dans l'équation (14). En calculant l'espérance du premier terme de l'équation (14), nous obtenons

$$E\left[\frac{1-c}{f^2 c} \sum_j \sum_k n'_{kj} s'^2_{kj} \right] \approx \frac{1-c}{fc} \sum_j \sum_k N_{kj} S^2_{kj} . \tag{15}$$

Une approximation additionnelle (A4) est nécessaire dans l'équation ci-dessus; nous ne tenons pas compte de la possibilité que $n'_{kj} \leq 1$ pour tout k,j. Nous devons aussi calculer l'espérance du second terme de l'équation (14). La formule exacte pour calculer cette espérance est

$$\frac{1-c}{fc} \sum_j \left\{ \sum_k N_{kj} (\bar{Y}_{kj} - \bar{Y}_{.j})^2 + a_1 \sum_k S^2_{kj} - a_2 \right\} \tag{16}$$

où $a_1 = 1 - f - E[n'_{kj}(1 - n'_{kj}/N_{kj})/n'_{.j}]$ et $a_2 = E[(\sum_k n'_{kj}(\bar{Y}_{kj} - \bar{Y}_{.j}))^2/n'_{.j}]$. Nous remarquons en premier lieu que $| a_1 | \leq 1$ et qu'on peut ne pas en tenir compte lorsqu'il est combiné à N_{kj} dans l'équation (15) (approximation A5). En outre, si dans a_2, on remplace $n'_{.j}$ par son espérance mathématique, $fN_{.j}$ (approximation A6), étant donné que $E[\sum_k n'_{kj} (\bar{Y}_{kj} - \bar{Y}_{.j})] = 0$, nous avons

$$a_2 \approx \frac{1}{fN_{.j}} \text{var}\left(\sum_k n'_{kj} (\bar{Y}_{kj} - \bar{Y}_{.j}) \right) \approx (1-f) \sum_k \frac{N_{kj}}{N_{.j}} (1 - W_{kj}) (\bar{Y}_{kj} - \bar{Y}_{.j})^2$$

où nous avons finalement substitué $N_{k.}$ à $(N_{k.} - 1)$ (approximation A7) en calculant la variance de la variable hypergéométrique n'_{kj}. Lorsque nous comparons a_2 au terme analogue de l'équation (16) affecté du coefficient N_{kj}, nous constatons que a_2 est en soi à peu près négligeable. Enfin, si nous ne tenons pas compte des différences entre N_{kj} et $(N_{kj} - 1)$ ou entre $N_{.j}$ et $(N_{.j} - 1)$ (approximation A8), nous pouvons combiner les équations (15) et (16) pour obtenir

$$E(\text{var}(\hat{\tau} \mid s')) \approx \frac{1-c}{fc} \sum_j \frac{N_{.j}}{N_{.j}-1} \sum_k \left[(N_{kj}-1) S^2_{kj} + N_{kj} (\bar{Y}_{kj} - \bar{Y}_{.j})^2 \right]$$

$$= \frac{1-c}{fc} \sum_j N_{.j} S^2_{.j} . \tag{17}$$

En combinant les équations (13) et (17), nous obtenons finalement

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f} \sum_k N_k . S_{k.}^2 + \frac{1-c}{fc} \sum_j N_{.j} S_{.j}^2 . \qquad (18)$$

La validité de cette approximation repose sur trois constatations. Premièrement, lorsqu'on applique l'équation (18) dans les cinq exemples pour lesquels il existe des données simulées, on obtient des résultats très comparables. L'erreur type approximative calculée à l'aide de l'équation (12) est 113.25, 108.97, 108.09, 106.77 et 106.32 pour $f' = .10, .20, .30, .40$ et $.50$ respectivement. Ces valeurs se rapprochent sensiblement de celles indiquées dans le tableau 3 sous la rubrique $M = 0$ de la colonne $ET(\hat{\tau})$ avec $m = 500$ ou $m = 2500$. Deuxièmement, après avoir analysé l'erreur engendrée par chaque approximation, on a constaté que cette erreur était négligeable (sauf peut-être en ce qui a trait à l'approximation A6) pour des populations et des échantillons relativement grands. En ce qui concerne l'approximation A6, la loi des grands nombres indique néanmoins que $n'_{.j}$ pourra être remplacée adéquatement par son espérance mathématique si l'échantillon est suffisamment grand. Enfin, comme nous le décrivons ci-dessous, l'équation (18) se ramène à l'équation exacte pour les trois plans d'échantillonnage courants. Premièrement, nous sommes en présence d'un échantillonnage stratifié ordinaire (selon les strates de premier niveau) lorsque $J = K, P_{.j} = P_{k.}$ pour $j = k$, et $c = 1$. L'équation (18) se réduit alors à $\text{var}(\hat{\tau}) \approx (1-f)/f \sum_k N_{k.} S_{k.}^2$, qui est bien l'équation exacte. Par ailleurs, nous sommes en présence d'un sondage à deux phases ordinaire pour stratification lorsque $K = 1$ et là encore, l'équation (18) se ramène à l'équation exacte (voir Cochran, 1977, p.329). De même, nous sommes en présence d'un échantillonnage stratifié ordinaire selon les strates de second niveau lorsque $f = 1$ (K et la stratification a priori ne sont plus vraiment pertinents dans ce cas) et une fois de plus, l'équation (18) se réduit à l'équation exacte.

BIBLIOGRAPHIE

ALLEN, J.R. (1984). Epidemiology of the Acquired Immunodeficiency Syndrome (AIDS) in the United States. *Seminars in Oncology*, 11, 4-11.

CASAREALE, D. *et al.* (1984/5). Prevalence of AIDS-associated retrovirus and antibodies among male homosexuals at risk for AIDS in Greenwich Village. *AIDS Research*, 1, 407-421.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.

HAN, C. (1973). Double sampling with partial information on auxiliary variables. *Journal of the American Statistical Association*, 68, 914-918.

HANSEN, M.H., et HURWITZ, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

KURITSKY, J.N. *et al.* (1986). Results of nationwide screening of blood and plasma for antibodies to HTLV-III. *Transfusion*, 26, 205-207.

LUI, K. *et al.* (1986). A model based approach for estimating the mean incubation period of transfusion-associated acquired immuno-deficiency symdrome. *Proceedings of the National Academy of Sciences*, 83, 3051-3055.

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

REDFIELD, R.R., et BURKE, D.S. (1987). Shadow on the land: the epidemiology of HIV infection. *Viral Immunology*, 1, 69-81.

ROBERT-GUROFF, M. (1986). Prevalence of antibodies to HTLV-I, -II, and -III in intravenous drug abusers from an AIDS endemic region. *Journal of the American Medical Association*, 255, 3133-3137.

VARDEMAN, S., et MEEDEN, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.

WHITE, D. (1987). Mean squared error of estimators using two stage sampling for stratification and prior information. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.