Techniques d'enquête, décembre 1990
Vol. 16, n° 2, pp. 291–295
Statistique Canada

Estimation d'un rapport sans connaître le plan d'échantillonnage

SHIBDAS BANDYOPADHYAY[1]

RÉSUMÉ

L'auteur montre que, pour une catégorie d'estimateurs linéaires sans biais appliqués à une catégorie de plans d'échantillonnage, il est possible d'estimer un rapport de population au moyen d'un rapport d'estimateurs non biaisés sans connaître les poids d'échantillonnage. Cette catégorie de plans d'échantillonnage comprend des plans aussi courants que l'échantillonnage avec probabilités inégales avec ou sans remise, l'échantillonnage stratifié avec répartition proportionnelle et probabilités inégales de sélection sans remise dans chaque strate, *etc.*

MOTS CLÉS: Rapport de totaux non pondérés; échantillonnage symétrique.

1. INTRODUCTION

Soit m le nombre d'adultes qui savent lire et écrire parmi t adultes dans un échantillon de n familles tiré d'une population donnée. Nous supposons que le taux d'alphabétisation des adultes dans la population R est estimé au moyen de la formule $r = m/t$. De même, pour un tableau à double entrée qui donne la répartition (en pourcentage) d'individus selon l'âge et le sexe, nous supposons qu'une fréquence de case est estimée au moyen du rapport entre le nombre d'individus dans la case et le nombre total d'individus pour l'échantillon de n familles (ce rapport étant multiplié par 100 pour obtenir un pourcentage).

Peu importe la méthode de sélection des familles, un rapport formé de deux totaux non pondérés et qui vise à estimer un ratio ou une répartition en pourcentage répond aux besoins de nombreux utilisateurs non spécialistes. De fait, certains rapports d'enquêtes renferment des tableaux dont les éléments sont estimés de cette façon, comme si le plan d'échantillonnage était un plan autopondéré.

En revanche, si les n familles devaient être choisies suivant un plan d'échantillonnage (à un degré) avec PPTSR, il faudrait normalement déterminer des totaux pondérés pour obtenir des estimateurs non biaisés des numérateurs et des dénominateurs correspondants avant de calculer un rapport ou une répartition en pourcentage.

Dans cette étude, nous montrons que, pour des plans d'échantillonnage comme le plan à un degré avec PPTSR et sans aucune autre hypothèse,

i) un rapport formé de deux totaux non pondérés peut servir à estimer le rapport de population correspondant; le premier doit être vu comme le rapport d'un *estimateur sans biais* du numérateur à un *estimateur sans biais* du dénominateur;

ii) il existe une catégorie de plans d'échantillonnage, qui ne sont pas autopondérés, pour lesquels la proposition (i) se vérifie. Cette catégorie comprend les plans d'échantillonnage à un degré avec probabilités inégales avec ou sans remise et les plans d'échantillonnage stratifié avec répartition proportionnelle et probabilités inégales de sélection sans remise dans chaque strate.

[1] Shibdas Bandyopadhyay, Applied Statistics, Surveys and Computing Division, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, Inde.

2. PLANS D'ÉCHANTILLONNAGE SYMÉTRIQUES

Considérons une population finie composée de N unités $U_1, U_2, \ldots, U_N$. Désignons par Y_i et X_i, les valeurs de deux variables d'intérêt, Y et X qui se rapportent à l'unité U_i, $i = 1, 2, \ldots, N$.

Le problème consiste à estimer un taux ou ratio $R = T(Y)/T(X)$ où $T(Y) = Y_1 + Y_2 + \ldots + Y_N$, et $T(X)$ est défini de la même façon avec la variable X.

La procédure normale est de calculer des estimations sans biais de $T(Y)$ et $T(X)$ puis d'utiliser le rapport de ces estimations comme estimateur de R. Dans cet article, nous nous proposons d'appliquer la même méthode de telle manière que le rapport des estimations fasse abstraction des probabilités de sélection des unités d'échantillonnage.

Définissons un plan d'échantillonnage.

Soit S l'ensemble de tous les échantillons possibles tel que $p(s) > 0$, où $p(s)$ est la probabilité de tirer l'échantillon s, et $\sum_{s \in S} p(s) = 1$.

Pour s dans S et $i = 1, 2, \ldots, N$,

$n(i,s)$ = le nombre de fois que U_i est incluse dans s, et $\alpha_i = \sum_{s \in S} n(i,s)$, le nombre de fois que U_i est incluse dans tous les échantillons possibles.

S, $p(s)$, et α_i dépendent du plan d'échantillonnage.

Définition 2.1. Un plan d'échantillonnage est dit symétrique si $\alpha_i = \alpha$, pour tous $i = 1, 2, \ldots, N$.

L'estimateur ci-dessous, qui est fondé sur l'échantillon s, et qui appartient à la catégorie d'estimateurs linéaires sans biais de Godambe (1955) pour $T(Y)$, a été analysé par Bandyopadhyay et coll. (1977).

$$T(Y,s) = \sum_{i=1}^{N} Y_i n(i,s) \alpha_i^{-1} p^{-1}(s). \tag{2.1}$$

De toute évidence, $T(Y,s)$ est un estimateur non biaisé de $T(Y)$. Un estimateur du ratio $R = T(Y)/T(X)$, fondé sur un échantillon s et défini comme le rapport d'un estimateur sans biais de $T(Y)$ à un estimateur sans biais de $T(X)$, est

$$R(s) = T(Y,s)/T(X,s) = \sum_{i=1}^{N} Y_i n(i,s) \alpha_i^{-1} \Big/ \sum_{i=1}^{N} X_i n(i,s) \alpha_i^{-1}. \tag{2.2}$$

Dans le cas de plans d'échantillonnage symétriques, $\alpha_i = \alpha$ pour tous i et (2.2) devient

$$R(s) = \sum_{i=1}^{N} Y_i n(i,s) \Big/ \sum_{i=1}^{n} X_i n(i,s) = \frac{\text{total non pondéré des valeurs } Y \text{ dans l'échantillon}}{\text{total non pondéré des valeurs } X \text{ dans l'échantillon}}. \tag{2.3}$$

Les observations précédentes sont résumées dans le théorème suivant.

Théorème principal. Étant donné un plan d'échantillonnage symétrique, le rapport de deux totaux non pondérés est un estimateur du rapport de population correspondant dans la mesure où il est défini comme le rapport d'un estimateur sans biais à un autre estimateur sans biais, sauf que le rapport estimé fait abstraction des probabilités de sélection des unités de la population.

Il convient de préciser que dans le cas de plans symétriques, il n'est pas nécessaire que les probabilités de sélection des unités soient égales. Les plans d'échantillonnage symétriques ne sont donc pas nécessairement autopondérés. Les plans autopondérés exigent que $\alpha_i p(s)$ ait la même valeur pour tous i et s, alors que cette condition n'est pas caractéristique des plans symétriques.

Dans le cas de plans non symétriques, l'équation (2.2) est facile à résoudre puisque les α_i sont faciles à calculer dans la majorité des cas et qu'il n'est pas nécessaire de calculer les probabilités de sélection.

En ce qui concerne l'échantillonnage de n unités sans remise, on compte $\binom{N-1}{n-1}$ échantillons (non ordonnés) qui renferment une unité U_i, donnée, de sorte que $\alpha_i = \binom{N-1}{n-1}$ pour tous i et par conséquent, le plan d'échantillonnage avec PPTSR est symétrique. Il convient de souligner qu'il n'y a pas toujours $\binom{N}{n}$ échantillons possibles avec les plans PPTSR. Comme l'indique Connor (1966), il arrivera que dans des plans d'échantillonnage systématique avec PPT (randomisé ou non), un ensemble de n unités ait une probabilité de sélection nulle. Les résultats de notre analyse s'appliquent lorsque le plan PPTSR est tel que, pour aucun ensemble de n unités, il n'existe de probabilité de sélection composée nulle.

Pour ce qui a trait à l'échantillonnage de n unités avec remise, on compte N^n échantillons (ordonnés), de sorte que $\alpha_i = nN^{n-1}$ pour tous i et par conséquent, le plan PPTAR est symétrique.

Dans le cas d'un échantillonnage par strate dans k strates, avec PPTSR, la valeur α pour chaque unité de la strate j est

$$\alpha_j = \frac{n_j}{N_j} \prod_{i=1}^{K} \binom{N_i}{n_i}$$

qui devient une constante lorsque la répartition est proportionnelle et qu'il n'existe de probabilité de sélection composée nulle pour aucun ensemble d'unités dans aucune strate, N_j et n_j étant respectivement l'effectif de la population et de l'échantillon pour la strate $j = 1, 2, \ldots, k$. On peut aussi avoir des plans à plusieurs degrés symétriques en appliquant le même mode de répartition.

Dans le cas de l'échantillonnage avec PPTAR, soulignons que l'estimateur non biaisé de $T(Y)$ défini en (2.1) n'est pas acceptable. On peut améliorer cet estimateur en remplaçant $n(i,s)$ et α_i par $n^*(i,s)$ et α_i^*, respectivement, où $n^*(i,s)$ est égal à 1 si $n(i,s)$ est au moins égal à 1, et égal à 0 si $n^*(i,s)$ est nul, et où α_1^* est défini par rapport à $n^*(i,s)$. En l'occurrence, $\alpha_i^* = N^n - (N-1)^n$, soit le nombre d'échantillons (ordonnés) contenant une unité U_i donnée. En ce qui a trait à la comparaison d'estimateurs, il n'est pas possible d'obtenir une expression en forme analytique fermée pour l'efficacité relative, même lorsqu'il s'agit de plans PPTAR.

Pour pouvoir comparer les biais relatifs et l'efficacité d'estimateurs, nous proposons une étude empirique réalisée au moyen d'un plan PPTSR. Une autre formule tout aussi intéressante serait d'analyser la variance et le biais de gros échantillons au moyen du développement de Taylor.

Il est clair que nous ne pouvons estimer la variance de $R(s)$ sans connaître les poids d'échantillonnage ou sans poser d'autres hypothèses. Cependant, si s_1 et s_2 sont deux demi-échantillons prélevés au moyen du même plan symétrique (comme deux échantillons PPTSR indépendants de même taille), R est estimé au moyen de la formule $[R(s_1) + R(s_2)]/2$, et l'estimateur non biaisé de la variance de l'estimateur est $[R(s_1) - R(s_2)]^2/4$.

Si $T(X)$ est connu, un estimateur par quotient pour $T(Y)$ est $T(X)T(Y,s)/T(X,s)$, cet estimateur peut être amélioré, comme dans Bandyopadhyay (1980), selon que la fraction de sondage est supérieure ou non à 0.5.

Lorsque les unités de la population sont réparties en k grappes non chevauchantes et que la probabilité de sélection de la grappe j est p_j, nous sommes alors en présence d'un plan symétrique ($\alpha_i = 1$ pour toutes les unités de chaque grappe). Précisons que la taille de l'échantillon correspond à la taille de la grappe choisie et que, par conséquent, les plan symétriques ne produisent pas nécessairement des échantillons de taille fixe.

3. ÉTUDE EMPIRIQUE DU BIAIS ET DE L'ERREUR QUADRATIQUE MOYENNE

Yates et Grundy (1953) ont considéré les trois populations hypothétiques suivantes, formées chacune de quatre unités.

	Population A				Population B				Population C			
X	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
Y	0.5	1.2	2.1	3.2	0.8	1.4	1.8	2.0	0.2	0.6	0.9	0.8

Le plan d'échantillonnage consiste à prélever avec PPTSR un échantillon de taille $n = 2$ en utilisant la valeur X comme mesure de taille des unités. On veut comparer le biais et l'erreur quadratique moyenne de $R(s)$ avec ceux de $R_{HT}^{(s)}$ où $R_{HT}^{(s)}$ est le rapport de l'estimateur d'Horvitz-Thompson (1952) de $T(Y)$ à l'estimateur d'Horvitz-Thompson de $T(X)$. Les résultats de la comparaison sont présentés dans le tableau ci-dessous.

Populations:	A	B	C
Biais relatif de $R(s)$	0.02456	−0.02785	−0.00496
Biais relatif de $R_{HT}(s)$	−0.00379	0.00552	0.00232
EQM de $R(s)$	0.2946	0.2946	0.0824
EQM de $R_{HT}(s)$	0.3159	0.3642	0.0690
Efficacité de $R(s)$par rapport à $R_{HT}(s)$	1.0723	1.2362	0.8374

Bien que le biais de $R(s)$ par rapport à R soit supérieur à celui de $R_{HT}^{(s)}$ pour les trois populations, les différences sont minimes. $R(s)$ est plus efficace que $R_{HT}(s)$ pour les populations A et B mais moins efficace pour la population C.

Comme les trois populations ci-dessus sont des cas extrêmes par rapport aux situations que l'on retrouve dans la pratique, on peut penser que l'estimateur $R(s)$ sera utile lorsqu'on ne connaîtra pas les données du plan d'échantillonnage au moment de l'estimation.

REMERCIEMENTS

L'auteur remercie sincèrement les arbitres qui, par leurs commentaires judicieux et constructifs, ont contribué à la version définitive de cet article.

BIBLIOGRAPHIE

BANDYOPADHYAY, S., CHATTOPADHYAY, A.K., et KUNDU, S.C. (1977). On estimation of population total. *Sankhyā*, Sér. C, 39, 28-42.

BANDYOPADHYAY, S. (1980). Improved ratio and product estimators. *Sankhyā*, Sér. C, 42, 45-49.

CONNOR, W.S. (1966). An exact formula for the probability that two specificed sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-396.

GODAMBE, V.P. (1955). A unified theory of sampling from finite population. *Journal of the Royal Statistical Society*, Sér. B, 17, 269-278.

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

YATES, F., et GRUNDY, P.M. (1953). Selections without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Sér. B, 15, 253-261.