Techniques d'enquête, juin 1987
Vol. 13, n° 1, pp. 93-100
Statistique Canada

Méthode d'échantillonnage avec probabilités de sélection proportionnelles à la taille

A. DEY et A.K. SRIVASTAVA[1]

RÉSUMÉ

On propose ici un nouveau plan d'échantillonnage sans remise avec probabilités inégales de sélection de n unités ($n > 2$) dans une population finie. Ce plan assure que les probabilités de sélection sont proportionnelles à la taille. Il offre l'avantage de simplifier le processus de sélection et d'estimation et de produire un estimateur de la variance non négatif. On montre que la variance de l'estimateur de Horvitz-Thompson obtenu à l'aide de ce nouveau plan est plus petite que celle des estimateurs produits habituellement selon un plan d'échantillonnage avec probabilités proportionnelles à la taille et avec remise. Le plan proposé donne également de bons résultats par rapport au plan d'échantillonnage sans remise élaboré par Sampford (1967).

MOTS CLÉS: Échantillonnage avec probabilités inégales; estimateur de Horvitz-Thompson.

1. INTRODUCTION

Dans un plan d'échantillonnage sans remise avec probabilités inégales de sélection de n unités à partir d'une population finie comprenant N unités, si π_i désigne la probabilité d'inclusion de la i-ième unité dans l'échantillon, $i = 1, 2, \ldots, N$, l'estimateur de Horvitz-Thompson (1952) (l'estimateur H-T) de Y, qui est la valeur totale de la variable d'intérêt, y, dans la population étudiée, s'exprime

$$\hat{Y} = \sum_{i \in s} (y_i / \pi_i), \tag{1.1}$$

où y_i est la valeur de y chez la i-ième unité et où la sommation porte sur les unités incluses dans l'échantillon. La variance de $\hat{Y}$ est

$$Var\ (\hat{Y}) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij})\ (y_i/\pi_i - y_j/\pi_j)^2 \tag{1.2}$$

où π_{ij} désigne la probabilité d'inclusion du couple d'unités i et j dans l'échantillon ($i \neq j$; $i, j = 1, 2, \ldots, N$).

On peut s'attendre que la variance de $\hat{Y}$ sera beaucoup moins grande si on utilise un plan d'échantillonnage qui assure que les π_i sont proportionnels à une mesure donnée de la taille, disons, x_i, pour $i = 1, 2, \ldots, N$, où on suppose que les x_i sont presque proportionnels aux y_i. Les plans d'échantillonnage dans lesquels les π_i sont proportionnels aux x_i sont appelés des plans d'échantillonnage avec probabilités de sélection proportionnelles à la taille (PSPT). Pour une présentation détaillée des techniques d'échantillonnage avec probabilités inégales, y compris les plans d'échantillonage avec PSPT, voir la monographie de Brewer et Hanif (1983).

Les plans d'échantillonnage avec probabilités inégales et sans remise, en général, et les plans d'échantillonnage avec PSPT, en particulier, devraient posséder entre autres caractéristi-

[1] A. Dey et A.K. Srivastava, Indian Agricultural Statistics Research Institute, Library Avenue, Nouvelle-Delhi 110012, Inde.

ques souhaitables l'avantage de simplifier le processus de sélection et d'estimation, de permettre de calculer un estimateur de la variance non négatif et d'être plus efficaces que les techniques d'échantillonnage avec probabilités proportionnelles à la taille (PPT) et avec remise. Malheureusement, pour des échantillons de taille plus grande que deux, il n'y a pas encore beaucoup de techniques qui satisfont pleinement à toutes ces exigences.

Dans le présent article, on propose un plan d'échantillonnage pour des échantillons de taille arbitraire n où $n > 2$. La technique est plutôt simple tant pour la sélection de l'échantillon que pour l'estimation, étant donné qu'on peut disposer d'expressions compactes pour les π_i. Elle donne également la possibilité de calculer un estimateur positif de la variance de l'estimateur H-T de Y. La performance réalisée par l'estimateur H-T au moyen du plan proposé est comparée à celle de l'estimateur résultant de l'utilisation d'une technique d'échantillonnage avec PPT et avec remise. À partir de cette comparaison, on trouve une condition suffisante simple qui, si elle est satisfaite, assure une performance du nouveau plan supérieure à celle de l'autre technique. Les résultats d'une étude empirique faite avec quelques populations naturelles indiquent que le plan proposé se compare avantageusement à celui élaboré par Sampford (1967).

2. MÉTHODE D'ÉCHANTILLONNAGE

Soient une population de N unités, y la variable d'intérêt et x une variable auxiliaire prise comme mesure de la taille. On suppose que les valeurs de x sont connues pour toutes les unités de la population. On veut maintenant tirer un échantillon de taille n $(n > 2)$. Pour commencer, on suppose que n est *pair*.

On divise la population en m groupes m $(> n/2)$ de telle sorte que le i-ième groupe contienne N_i unités $(N_i > 2)$ et $i = 1, 2, \ldots, m)$ et que, pour chaque groupe,

$$X_i/X > (n - 2)/[n(m - 1)], \tag{2.1}$$

où

$$X_i = \sum_{u=1}^{N_i} x_{i_u},$$

x_{i_u} est la valeur de x chez la u-ième unité du i-ième groupe et $X = X_1 + X_2 + \ldots + X_m$.

La condition (2.1) est satisfaite si les X_i $(i = 1, 2, \ldots, m)$ sont presque égaux. Il a été montré que, dans des populations réelles, telles que celles considérées par Rao et Bayless (1969) et d'autres auteurs, cette condition est satisfaite pour plusieurs valeurs de m si les groupes sont formés de manière que leurs tailles X_i respectives sont presque égales entre elles. Rao et Lanke (1984) ont proposé une méthode de groupement des unités dans laquelle N unités sont groupées en R groupes de telle sorte que la valeur totale de chaque groupe, X_i, est presque égale d'un groupe à un autre et que la taille des groupes est soit $[N/R]$ ou $[N/R] + 1$, où $[x]$ est le plus grand nombre entier contenu dans x. On peut également appliquer la méthode de Rao-Lanke pour former les groupes.

Une fois les m groupes formés, la méthode d'échantillonnage proposée comprend les étapes suivantes:

Étape 1. Sélection de $n/2$ groupes parmi les m groupes, au moyen de la méthode d'échantillonnage de Midzuno (1951) avec probabilités $\{P_i'\}$, c'est-à-dire sélection d'un groupe avec probabilité

$$P_i' = [n(m-1)P_i - (n-2)]/(2m-n), \text{ où } P_i = X_i/X,$$

et des $(n/2) - 1$ autres groupes avec probabilités égales et sans remise.

Étape 2. Sélection de deux unités, dans chacun des groupes choisis, suivant l'une quelconque des méthodes avec PSPT, par exemple à l'aide de la méthode de Durbin (1967), c'est-à-dire sélection dans le i-ième groupe choisi ($i = 1, 2, \ldots, n/2$) d'une unité avec probabilité

$$p_{i_u|\,i} = x_{i_u}/X_i$$

et de la deuxième unité avec probabilité révisée

$$p_{i_u|\,i_v} = x_{i_v}[1/(X_i - 2x_{i_v}) + 1/(X_i - 2x_{i_u})]/D_i$$

où
$$D_i = \left[1 + \sum_{u=1}^{N_i} x_{i_u}/(X_i - 2x_{i_u})\right].$$

Avec cette méthode d'échantillonnage, la probabilité de sélection de la i_u-ième unité est évidemment

$$\pi_{i_u} = n\,p_{i_u} \tag{2.2}$$

où

$$p_{i_u} = x_{i_u}/X.$$

De plus, les probabilités de sélection d'une paire d'unités sont

$$\pi_{i_u i_v} = \frac{n\,p_{i_u}\,p_{i_v}\,(P_i - p_{i_u} - p_{i_v})}{D_i(P_i - 2\,p_{i_u})\,(P_i - 2\,p_{i_v})} \tag{2.3}$$

et

$$\pi_{i_u j_v} = \frac{n\,(n-2)\,p_{i_u}p_{j_v}}{(m-1)(m-2)\,P_iP_j}\;[(m-1)(P_i + P_j) - 1], \tag{2.4}$$

$$i \neq j;\ i, j = 1, 2, \ldots, m.$$

Ainsi, nous voyons que le plan proposé est bel et bien un plan d'échantillonnage avec PSPT.

Tel qu'il a été mentionné précédemment, on peut utiliser à l'étape 2 de la méthode proposée n'importe quel plan avec PSPT pour sélectionner deux unités. Comme la méthode

de Durbin (1967), qui est équivalente à celle de Rao (1963) et de Brewer (1963), donne en général de bons résultats, c'est elle qu'on a décidé d'appliquer à l'étape 2.

3. ESTIMATEUR DE LA VARIANCE

Deux estimateurs sans biais bien connus de la variance de $\hat{Y}$, $Var(\hat{Y})$, ont été élaborés par Horvitz et Thompson (1952) et par Yates et Grundy (1953). Ces deux estimateurs présentent l'inconvénient de prendre parfois des valeurs négatives. Dans la présente section, on examine un estimateur positif de la variance qui met à profit le fait que le plan d'échantillonnage proposé est à deux degrés.

À l'aide d'un résultat Des Raj (1966), un estimateur non biaisé de $Var(\hat{Y})$ prend la forme suivante:

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n/2} \pi_i^{-1} \sum_{u<v} \sum \left[\frac{\pi_{i_u|i}\, \pi_{i_v|i}}{\pi_{i_u i_v|i}} - 1 \right] \left[\frac{y_{i_u}}{\pi_{i_u|i}} - \frac{y_{i_v}}{\pi_{i_v|i}} \right]^2$$

$$+ \sum_{i<j}^{n/2} \sum^{n/2} \left(\frac{\pi_i \pi_j}{\pi_{ij}} - 1 \right) \left[\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j} \right]^2, \tag{3.1}$$

où

$$\pi_i = n P_i / 2 ,$$

$$\pi_{ij} = \frac{n\,(n-2)}{4\,(m-2)} \{ (P_i + P_j) - 1/(m-1) \} ,$$

$$\pi_{i_u|i} = 2 p_{i_u} / P_i ,$$

$$\pi_{i_u i_v|i} = \frac{2 p_{i_u} p_{i_v} (P_i - p_{i_u} - p_{i_v})}{D_i P_i (P_i - 2p_{i_u})(P_i - 2p_{i_v})} ,$$

et

$$\hat{Y}_i = \sum_{u=1}^{2} y_{i_u} / \pi_{i_u|i} , \tag{3.2}$$

y_{i_u} étant la valeur de y chez la u-ième unité dans le i-ième groupe.

Les deux termes du membre de droite de l'équation (3.1) correspondent à l'estimateur de la variance de Yates-Grundy calculés respectivement pour la méthode de Durbin (étape 2) et pour la méthode de Midzuno (étape 1). Comme l'estimateur de la variance de Yates-Grundy est toujours positif pour chacune de ces deux méthodes d'échantillonnage, il s'ensuit que l'estimateur de la variance de l'équation (3.1) est également positif. Cependant, l'estimateur en (3.1) n'est ni l'estimateur de la variance de Horvitz-Thompson ni l'estimateur de la variance de Yates-Grundy.

4. COMPARAISON DU PLAN PROPOSÉ ET DE LA STRATÉGIE AVEC PPT AVEC REMISE

Dans la présente section, nous comparons l'efficacité des deux stratégies suivantes:

Stratégie 1. Plan d'échantillonnage proposé utilisé avec l'estimateur de Horvitz-Thompson.

Stratégie 2. Plan d'échantillonnage avec PPT et avec remise utilisé avec l'estimateur habituel.

La stratégie 1 est plus efficace (variance plus petite) que la stratégie 2 si et seulement si

$$\sum_{i=1}^{m} \sum_{u \neq v}^{N_i} \pi_{i_u i_v} (y_{i_u}/p_{i_u} - Y)(y_{i_v}/p_{i_v} - Y) + \sum_{i \neq j}^{m} \sum_{u}^{N_i} \sum_{v}^{N_j} \pi_{i_u j_v} (y_{i_u}/p_{i_u} - Y)(y_{j_v}/p_{j_v} - Y) < 0. \tag{4.1}$$

Après un certain nombre de manipulations algébriques longues mais élémentaires, l'inégalité (4.1) se ramène à ceci

$$- \sum_{i=1}^{n} (n/D_i) \sum_{u=1}^{N_i} (y_{i_u} - Y_i p_{i_u}/P_i)^2/(P_i - 2p_{i_u}) - n(n-2) \left[\sum_{i=1}^{m} (Y_i/P_i - Y) \right]^2 / [(m-2)(m-1)] - n(m-2)^{-1} \sum_{i=1}^{m} [\{(2n - m - 2)P_i - (n-2)(m-1)^{-1}\}(Y_i/P_i - Y)^2] < 0, \tag{4.2}$$

où

$$Y_i = \sum_{u} y_{i_u}$$

De toute évidence, (4.2) se vérifie si

(i) $(2n - m - 2) > 0$, et si

(ii) $P_i > (n-2)/[(m-1)(2n-m-2)]$. (4.3)

De plus, comme pour la première étape du plan d'échantillonnage nous utilisons la méthode de Midzuno avec des probabilités révisées $\{P_i'\}$, chacun des P_i doit satisfaire la condition (2.1), c'est-à-dire que chaque P_i doit satisfaire

$$P_i > (n-2)/[n(m-1)].$$

Par conséquent, (4.2) se vérifie si

$$m \leq (n-2). \tag{4.4}$$

Tableau 1

Description des populations

Numéro de population	Source	N	y	x
1.	Des Raj (1965)	20	Nombre de ménages	Nombre de ménages estimé à l'oeil
2.	Rao (1963)	14	Nombre d'acres de maïs en 1960	Nombre d'acres de maïs en 1958
3.	Cochran (1963, p. 204)	10	Poids des pêches	Poids des pêches estimé à l'oeil
4.	Hanurav (1967)	20	Population en 1967	Population en 1957
5.	Hanurav (1967)	19	Population en 1967	Population en 1957
6.	Hanurav (1967)	16	Population en 1967	Population en 1957
7.	Hanurav (1967)	17	Population en 1967	Population en 1957
8.	Cochran (1963, p. 325)	10	Nombre de personnes par îlot de logements	Nombre de pièces par îlot de logements
9.	Cochran (1963, p. 156, villes 1-16)	16	Population en 1930	Population en 1920
10.	Cochran (1963, p. 156, villes) 33-49)	17	Population en 1930	Population en 1920
11.	Sampford (1962, p. 61)	35	Nombre d'acres d'avoine en 1957	Nombre d'acres d'avoine en 1947
12.	Sukhatme et Sukhatme (1970, p. 256, cercles 1-20)	20	Nombre d'acres de blé	Nombre de villages
13.	Sukhatme et Sukhatme (1970, p. 256, cercles 21-40)	20	Nombre d'acres de blé	Nombre de villages
14.	Yates (1960, p. 163)	20	Volume de bois d'oeuvre	Volume de bois d'oeuvre estimé à l'oeil

Tableau 2

Efficacité relative en pourcentage des stratégies 1 et 3 par rapport à la stratégie 2, avec les populations décrites au tableau 1 ($n = 4$)

Population Numéro	Stratégie 1				Stratégie 3
	$m = 3$	4	5	6	
1.	130.1	118.7	120.8	124.5	127.8
2.	132.6	130.2	–	–	127.1
3.	149.1	–	–	–	147.9
4.	120.7	120.6	122.7	129.7	117.8
5.	129.1	138.7	158.7	–	125.1
6.	158.0	173.1	–	–	139.5
7.	151.9	144.8	169.2	–	131.9
8.	168.5	–	–	–	145.5
9.	118.3	116.3	–	–	109.5
10.	126.6	–	–	–	112.2
11.	113.8	116.2	135.6	129.9	113.8
12.	117.4	128.0	119.0	–	119.3
13.	122.2	120.6	–	–	119.7
14.	124.8	123.1	115.4	113.2	116.3

Il semble donc que, pour que la stratégie 1 soit supérieure à la stratégie 2, il faille choisir m de manière que

$$n/2 < m \leq (n - 2). \tag{4.5}$$

Il est toutefois clair que (4.4) n'est qu'une condition suffisante mais non nécessaire. Dans le cas où $n > 6$, la condition (4.5) donne plusieurs choix pour la valeur de m, tandis que si $n = 6$, (4.5) implique que $m = 3$. Pour $n = 4$, aucune valuer de m ne peut satisfaire (4.5). On s'est donc employé à étudier l'efficacité de la stratégie 1, lorsque $n = 4$, avec un certain nombre de populations naturelles pour diverses valeurs de m qui ne satisfont pas (4.5). Une description des populations est donnée au tableau 1, tandis que le tableau 2 donne l'efficacité relative de la stratégie 1 par rapport à celle de la stratégie 2 avec les populations décrites dans le tableau 1. Le tableau 2 compare également la performance de l'estimateur H-T utilisé de concert avec le plan d'échantillonnage de Sampford (1967), ce qu'on désigne ici par la stratégie 3, et la performance de la stratégie 2.

Il ressort du tableau 2 que la performance de la stratégie proposée (stratégie 1) se compare avantageusement à celle de la stratégie 3 dans le cas de la plupart des populations étudiées. Bien entendu, les deux stratégies sont toutes deux supérieures à la stratégie 2.

Pour obtenir l'efficacité relative de la stratégie 1, les unités ont été groupées en ne veillant uniquement qu'à ce que la condition (2.1) soit satisfaite. On a également tenté d'utiliser la méthode de Rao et Lanke (1984) pour former les groupes. Cette méthode n'a toutefois pas conduit à une grande efficacité dans tous les cas. D'autres recherches sont nécessaires pour décider du «meilleur» choix des groupes. Pour certaines populations, il a été impossible de former des groupes satisfaisant à la condition (2.1) avec des valeurs élevées de m; c'est pour cette raison que l'efficacité relative dans ces cas-là n'apparaît pas dans le tableau 2.

En conclusion, il semble indiquer de faire un bref commentaire au sujet des cas où la taille de l'échantillon voulue, n, est *impaire*. On peut tirer un échantillon avec PSPT pour n impair en sélectionnant $(n + 1)$ unités à l'aide de la méthode proposée, puis en rejetant au hasard une unité. Les expressions pour les π_i et π_{ij} sont également assez simples dans ce cas-là. De

toute évidence, quand une des $(n + 1)$ unités de l'échantillon est rejetée au hasard, l'échantillon définitif comprend deux unités de chacun de $(n - 1)/2$ groupes et seulement une unité d'un des groupes. Un estimateur non biaisé et positif de $Var(\hat{Y})$, semblable à celui de l'expression (3.1), peut ensuite être calculé à partir des $(n - 1)/2$ groupes qui contiennent deux unités de l'échantillon.

REMERCIEMENTS

Les auteurs remercient l'arbitre pour ses conseils pertinents concernant la première version.

BIBLIOGRAPHIE

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

BREWER, K.R.W. et HANIF, M. (1983). *Sampling with Unequal Probabilities*. Lecture Notes in Statistics, No. 15, New York: Springer-Verlag.

COCHRAN, W.G. (1963). *Sampling Techniques*, (2e éd.). New York: John Wiley.

DES RAJ (1965). Variance estimation in randomized systematic sampling with probability proportional to size, *Journal of the American Statistical Association*, 60, 278-284.

DES RAJ (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Journal of the Royal Statistical Society*, Sér. C, 16, 152-164.

HANURAV, T. (1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *Journal of the Royal Statistical Society*, Sér. B, 29, 379-391.

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.

MIDZUNO, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 2, 99-108.

RAO, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58, 202-215.

RAO, J.N.K. et BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units by stratum. *Journal of the American Statistical Association*, 64, 540-559.

RAO, J.N.K. et LANKE, J. (1984). Simplified unbiased variance estimation for multistage designs. *Biometrika*, 71, 387-395.

SAMPFORD, M.R. (1962). *An Introduction to Sampling Theory*. Édimbourg: Oliver and Boyd.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

SUKHATME, P.V. et SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*, (2e éd.). Ames, Iowa: Iowa State University Press.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3e éd.). Londres: Griffin.

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