

REVUE DE STATISTIQUE APPLIQUÉE

F. BERTRAND

R. PALM

Échantillonnage avec probabilités inégales et sans remise : comparaison de l'estimateur de Horvitz et Thompson à des alternatives plus simples

Revue de statistique appliquée, tome 38, n° 3 (1990), p. 5-21

http://www.numdam.org/item?id=RSA_1990__38_3_5_0

© Société française de statistique, 1990, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

ÉCHANTILLONNAGE AVEC PROBABILITÉS INÉGALES ET SANS REMISE : COMPARAISON DE L'ESTIMATEUR DE HORVITZ ET THOMPSON A DES ALTERNATIVES PLUS SIMPLES

F. BERTRAND (1) et R. PALM (2)

(1) *Unité d'Epidémiologie - Université Catholique de Louvain,
30-34 Clos Chapelle-aux-Champs B-1200 Bruxelles.*

(2) *Unité de Statistique et d'Informatique - Faculté des Sciences Agronomiques,
2 Passage des Déportés B-5800 Gembloux.*

RÉSUMÉ

Pour un échantillonnage réalisé avec probabilités inégales et sans remise, les probabilités d'inclusion des unités d'échantillonnage permettant l'utilisation de l'estimateur du total de HORVITZ et THOMPSON sont estimées par simulation. Le biais auquel on peut s'attendre si on analyse un tel échantillon comme s'il avait été prélevé avec probabilités égales et sans remise, ou avec probabilités inégales et avec remise, est également étudié par simulation. On discute enfin l'intérêt de l'échantillonnage avec probabilités inégales et sans remise par rapport à des alternatives plus simples.

Mots-clés : Echantillonnage, Estimateur de Horvitz-Thompson, Probabilité d'inclusion, Simulation.

SUMMARY

When sampling with unequal probabilities and without replacement, the inclusion probabilities of units in the sample which allow the use of HORVITZ and THOMPSON's estimator of a total, are obtained by simulation. The biases encountered when analysing such a sample as if drawn with equal probabilities and without replacement, or with unequal probabilities but with replacement, are also studied. Finally, the interest of using sampling with unequal probabilities rather than simpler sampling methods is discussed.

Key-words : Sampling, Horvitz-Thompson estimator, Probability of selection, simulation.

1. Introduction

L'échantillonnage avec probabilités inégales et sans remise a fait l'objet de très nombreuses études, dont le but est le plus souvent de proposer aux utilisateurs des méthodes de tirages des unités telles que les probabilités d'inclusion

des unités dans l'échantillon soient connues *a priori* ou faciles à calculer. VOS (1974 a; 1974 b) en détaille une vingtaine et HANIF et BREWER (1980) en recensent une cinquantaine. Dans ces cas, l'estimation du total de la population ne pose guère de problèmes. Des informations à ce sujet sont données dans de nombreux ouvrages consacrés aux sondages et notamment dans COCHRAN (1977), DESABIE (1966), DROESBEKE *et al.* (1987) et GROSBRAS (1987). Par contre, lorsque les probabilités d'inclusion des unités dans l'échantillon ne sont pas connues *a priori*, l'estimation du total soulève généralement des difficultés.

Une telle situation peut se présenter, par exemple, à la suite de la mise en oeuvre incorrecte d'un échantillonnage à deux degrés, lorsque le plan d'échantillonnage prévoit le tirage des unités du premier degré avec remise et probabilités de sélection proportionnelles à la taille des unités, mais, qu'en pratique, le tirage se fait sans remise. Ainsi, dans le cadre d'une étude relative à la prévention des maladies cardio-vasculaires, une enquête a été réalisée en Belgique, dans la province de Luxembourg (VAN GELDEREN, 1983). Dans ce but, les 229 communes de cette province ont été regroupées en trois strates, en fonction de leur degré d'urbanisation. La première strate a été constituée de 11 villes, la seconde strate de 86 communes urbanisées et la troisième strate de 132 communes rurales. Un échantillon aléatoire et simple de personnes a ensuite été prélevé dans chacune des villes de la première strate avec un taux de sondage constant. Par contre, dans les deux autres strates, on a sélectionné des communes par un tirage sans remise avec des probabilités inégales et, dans ces communes, on a procédé à un tirage aléatoire d'un nombre constant d'individus. Pour la sélection des communes, on a utilisé la méthode classique des totaux cumulés décrite, par exemple, par DEROO et DUSSAIX (1980) pour la sélection d'unités avec probabilités proportionnelles à la taille, mais l'échantillonnage a été réalisé sans remise. A chaque commune, on a d'abord attribué un nombre de nombres aléatoires égal au nombre d'habitants de la commune, les nombres aléatoires attribués à chaque commune étant évidemment différents. A l'aide d'un générateur de nombres aléatoires, on a prélevé des nombres aléatoires a_1, a_2, \dots , et on a inclus dans l'échantillon les communes auxquelles ces nombres a_1, a_2, \dots , ont été associés. Toutefois, si deux ou plusieurs nombres aléatoires prélevés ont été associés à la même commune, celle-ci n'a été incluse dans l'échantillon qu'une seule fois, puisqu'on effectue un échantillonnage sans remise.

Si, dans cet échantillonnage, les unités du premier degré avaient été sélectionnées avec remise, les probabilités de sélection des communes auraient été proportionnelles au nombre d'habitants des communes et chaque habitant aurait finalement eu la même probabilité de faire partie de l'échantillon. L'analyse d'un tel échantillon n'aurait soulevé aucune difficulté, l'échantillon étant autopondéré (DESABIE, 1966). Du fait de la non-remise, les probabilités de sélection des communes ne sont plus proportionnelles aux nombres d'habitants et l'analyse de l'échantillon est nettement plus complexe.

L'objectif de cette étude est triple. Il s'agit, d'abord, de trouver une solution correcte en vue de l'analyse de l'échantillon présenté ci-dessus. Ensuite, on se propose d'évaluer l'erreur commise lorsqu'on analyse un tel échantillon comme s'il avait été tiré avec remise ou avec probabilités égales. Enfin, on souhaite déterminer l'efficacité de l'échantillonnage réalisé par rapport à des alternatives plus simples.

Dans un but de simplification, l'étude se limitera cependant à l'échantillonnage du premier degré.

Nous présenterons d'abord les différents estimateurs comparés (paragraphe 2) et les simulations réalisées (paragraphe 3). Ensuite, nous discuterons les résultats obtenus, en ce qui concerne l'estimation du total (paragraphe 4), l'estimation de la variance du total (paragraphe 5) et, enfin, l'intérêt de l'échantillonnage avec probabilités inégales et sans remise (paragraphe 6). Les conclusions de l'étude seront présentées au paragraphe 7.

2. Estimateurs comparés

2.1. Estimateurs de HORVITZ et THOMPSON

Un estimateur non biaisé du total, Y , d'une population, dans le cas du tirage d'un nombre n fixé d'unités avec probabilités inégales et sans remise, est proposé par HORVITZ et THOMPSON (1952) :

$$\widehat{Y}_1 = \sum_{i=1}^n \frac{y_i}{\pi_i}, \quad (1)$$

y_i étant la valeur que prend la variable étudiée en chaque unité i de la population considérée et π_i étant la probabilité que l'unité i fasse partie de l'échantillon de n unités tirées parmi N , sans remise et avec probabilités inégales.

La variance de cet estimateur est donnée par la formule suivante (cf, par exemple, COCHRAN, 1977) :

$$V(\widehat{Y}_1) = \sum_{i=1}^N \frac{(1 - \pi_i)}{\pi_i} y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j, \quad (2)$$

π_{ij} étant la probabilité que les unités i et j fassent simultanément partie de l'échantillon sélectionné. Cette relation peut encore s'écrire (COCHRAN, 1977) :

$$V(\widehat{Y}_1) = \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \quad (3)$$

Des estimateurs de cette variance sont proposés par HORVITZ et THOMPSON (1952) :

$$v_{HT}(\widehat{Y}_1) = \sum_{i=1}^n \frac{(1 - \pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j, \quad (4)$$

et par YATES et GRUNDY (1953) :

$$v_{YG}(\widehat{Y}_1) = \sum_{i=1}^n \sum_{j>i}^n \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \quad (5)$$

Ces deux estimateurs sont non biaisés, mais très instables, et peuvent conduire à des valeurs négatives. GOURIEROUX (1981) signale que les estimations qui en découlent sont en général différentes, mais qu'aucune n'est systématiquement préférable à l'autre.

L'utilisation des formules données ci-dessus nécessite la détermination des probabilités d'inclusion π_i et π_{ij} . En pratique, le calcul exact de ces probabilités est très complexe pour le type d'échantillonnage décrit dans l'introduction; elles seront par conséquent estimées à partir de la sélection, par simulation, d'un grand nombre d'échantillons.

D'autres relations permettant d'estimer le total de la population et la variance de ce total sont données par MURTHY (COCHRAN, 1977), mais leur utilisation entraîne un volume considérable de calculs; pour cette raison, elles n'ont pas été retenues.

2.2. Solutions approchées

Pour évaluer l'erreur commise en traitant un échantillon tiré avec probabilités inégales et sans remise par des méthodes d'analyse plus simples, les données ont été traitées comme si elles provenaient, d'une part, d'un échantillonnage avec probabilités égales et sans remise et, d'autre part, d'un échantillonnage avec probabilités inégales et avec remise.

a) Tirage avec probabilités égales et sans remise.

Les relations suivantes permettent d'obtenir l'estimation du total, la variance de cet estimateur et l'estimation de la variance de l'estimateur (COCHRAN, 1977) :

$$\widehat{Y}_2 = N \left(\frac{1}{n} \sum_{i=1}^n y_i \right), \quad (6)$$

$$V(\widehat{Y}_2) = N^2 \frac{\sigma_y^2}{n} \frac{N-n}{N-1} \quad (7)$$

et

$$v(\widehat{Y}_2) = N^2 \frac{\widehat{\sigma}_y^2}{n} \left(1 - \frac{n}{N} \right). \quad (8)$$

Dans ces relations, σ_y^2 est la variance de la caractéristique observée :

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^N \left(y_i - \frac{\sum_{i=1}^N y_i}{N} \right)^2,$$

et $\hat{\sigma}_y^2$ est une estimation de cette variance :

$$\hat{\sigma}_y^2 = \frac{1}{n-1} \sum_{i=1}^n \left(y_i - \frac{\sum_{i=1}^n y_i}{n} \right)^2.$$

b) Tirage avec probabilités inégales et avec remise.

On considère que la probabilité de sélection d'une unité à chaque tirage, z_i , est proportionnelle à une caractéristique connue, x_i , qui est, par exemple, la taille de l'unité :

$$z_i = \frac{x_i}{\sum_{i=1}^N x_i}.$$

On a alors (COCHRAN, 1977) :

$$\hat{Y}_3 = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i}, \quad (9)$$

$$V(\hat{Y}_3) = \frac{1}{n} \sum_{i=1}^N z_i \left(\frac{y_i}{z_i} - \sum_{i=1}^N y_i \right)^2 \quad (10)$$

et

$$v(\hat{Y}_3) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n z_i \left(\frac{y_i}{z_i} - \sum_{i=1}^n \frac{y_i}{z_i} \right)^2. \quad (11)$$

3. Simulations réalisées

Trois paramètres ont essentiellement été pris en considération lors du choix des simulations : la distribution des tailles des unités de la population étudiée, la

corrélation liant la variable étudiée et les tailles des unités des populations choisies et, enfin, la fraction sondée.

Six distributions de tailles ont été retenues. Les deux premières correspondent respectivement à la distribution des 132 communes rurales (distribution 1) et des 86 communes urbanisées (distribution 2) constituant les bases de sondages de l'enquête présentée dans l'introduction (VAN GELDEREN, 1983). La troisième distribution provient d'une enquête similaire réalisée dans la région d'Augsburg, en Allemagne (KEIL *et al.*, 1985) et comprend 64 villages (distribution 3). Les trois autres distributions de tailles sont artificielles et comportent 150 unités. Elles ont été obtenues par un générateur de nombres aléatoires de distribution normale (distribution 4), log-normale (distribution 5) et triangulaire (distribution 6). Ces six distributions, représentées dans la figure 1, ont été standardisées afin de rendre, pour chacune d'elles, la somme des tailles égale à 100.000.

En ce qui concerne les observations proprement dites, y_i , elles ont été générées de manière à couvrir un large éventail de corrélations positives avec la taille x_i des unités. La génération repose essentiellement sur le principe suivant. On génère d'abord des valeurs aléatoires initiales y'_i , distribuées selon une loi normale réduite et indépendantes de x_i et on effectue l'opération suivante :

$$y_i = ax_i + y'_i.$$

Des valeurs croissantes de a permettent d'augmenter la corrélation entre y_i et x_i puisque la corrélation des variables x et y est égale à :

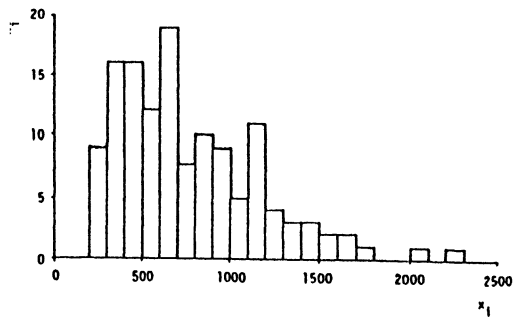
$$\text{corr}(x, y) = \frac{1}{\sqrt{1 + \frac{1}{a^2 \sigma_x^2}}},$$

σ_x^2 étant la variance de la variable x . D'autres modes de génération, basés notamment sur des valeurs de départ y_i réellement observées, ont également été utilisés, mais de manière plus limitée (BERTRAND, 1986).

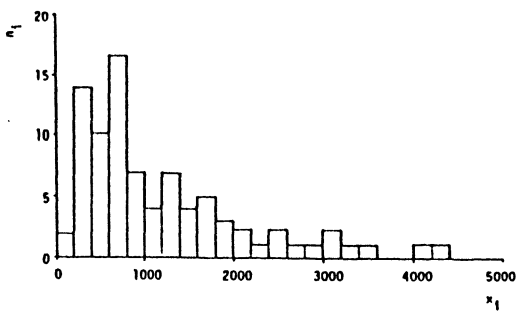
Les distributions de ces variables ont ensuite été standardisées de manière à rendre la moyenne nulle et l'écart-type unitaire, afin d'obtenir des résultats de simulations directement comparables.

Le tableau 1 donne les nombres de variables y générées pour chacune des six distributions de tailles, en fonction de la corrélation avec les tailles x . On constate que le nombre de variables n'est pas constant, ni pour les différentes distributions de tailles, ni pour les différentes classes de corrélation.

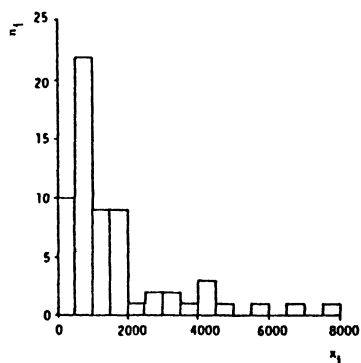
Enfin, lors de l'échantillonnage dans les 88 populations à deux dimensions ainsi générées, les quatre taux de sondage suivants ont été considérés : 0,10, 0,25, 0,50 et 0,75.



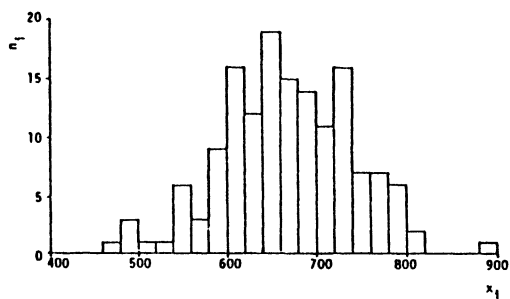
Distribution 1



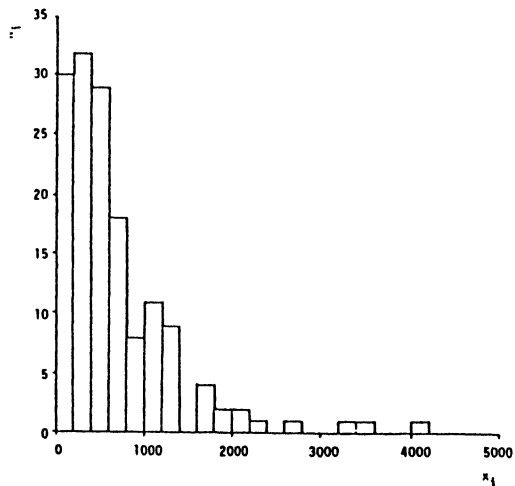
Distribution 2



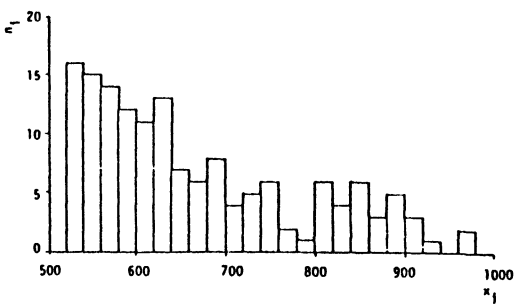
Distribution 3



Distribution 4



Distribution 5



Distribution 6

FIGURE 1

Distribution de fréquences des tailles des unités des six distributions considérées (après standardisation).

TABLEAU 1
Répartition du nombre de distributions à deux dimensions
qui ont été générées pour chaque distribution de tailles,
en fonction du coefficient de corrélation entre la caractéristique
étudiée y et la taille des unités x

Distributions des tailles	Coefficients de corrélation					Totaux
	0,0-0,2	0,2-0,4	0,4-0,6	0,6-0,8	0,8-1,0	
Distribution 1	1	2	2	1	2	8
Distribution 2	2	3	4	3	5	17
Distribution 3	16	3	3	3	4	29
Distribution 4	3	2	2	1	4	12
Distribution 5	2	3	0	1	4	10
Distribution 6	2	2	1	1	6	12
Totaux	26	15	12	10	25	88

Pour chacun des 352 cas, on a d'abord estimé les probabilités d'inclusion π_i et π_{ij} définies au paragraphe 2, par la sélection d'un très grand nombre d'échantillons selon la technique présentée au paragraphe 1. Les estimations des π_i sont données par le rapport entre le nombre d'échantillons comportant l'unité i et le nombre total d'échantillons prélevés. De même, les estimations de π_{ij} sont données par le rapport entre le nombre d'échantillons comportant à la fois l'unité i et l'unité j et le nombre total d'échantillons prélevés. Des informations complémentaires à ce sujet sont données par BERTRAND (1986).

Disposant des estimations des π_i et des π_{ij} , on a alors prélevé, pour chacun des 352 cas présentés ci-dessus, 100 échantillons indépendants et on a déterminé les différents estimateurs décrits au paragraphe 2. Pour chacun de ceux-ci, on dispose ainsi de 100 valeurs dont on peut calculer, notamment, la moyenne et la variance, qui représentent des estimations de la moyenne et de la variance des distributions d'échantillonnage de ces estimateurs.

4. Estimation du total

Afin de mettre en évidence le biais éventuel des trois estimateurs du total, des tests de conformité des moyennes ont été réalisés. Les caractéristiques y_k ayant été centrées, leur total est égal à zéro et l'hypothèse nulle s'écrit :

$$H_0 : m_{\widehat{Y}_k} = 0 \quad (k = 1, 2 \text{ ou } 3)$$

et on rejette l'hypothèse nulle, au niveau 0,05, si :

$$|t_{\text{obs}}| = \frac{|\widehat{Y}_k|}{\widehat{\sigma}_{\widehat{Y}_k} / \sqrt{100}} \geq t_{0,975(99)} = 1,984,$$

avec

$$\widehat{Y}_k = \frac{1}{100} \sum_{i=1}^{100} \widehat{Y}_{ki}$$

et

$$\widehat{\sigma}_{\widehat{Y}_k}^2 = \frac{1}{99} \sum_{i=1}^{100} (\widehat{Y}_{ki} - \widehat{Y}_k)^2.$$

Le tableau 2 reprend les moyennes des valeurs t_{obs} relatives à ces tests de conformité, en fonction de la fraction sondée et du coefficient de corrélation entre y et x . Des résultats plus détaillés sont donnés par BERTRAND (1986).

Pour l'estimateur du total donné par HORVITZ et THOMPSON, \widehat{Y}_1 , les résultats ont montré qu'aucune moyenne des totaux estimés n'est significativement différente de zéro, ce qui confirme le caractère non biaisé de cet estimateur.

Par contre, pour l'estimation du total par la formule (6), relative aux tirages avec probabilités égales et sans remise, on a obtenu des moyennes de totaux significativement différentes de zéro dans 300 cas sur les 352. On a constaté que cette méthode de traitement des données surestime le total, le biais étant d'autant plus important que la corrélation entre y et x est grande, du moins lorsqu'elle est positive. En effet, étant donné que le tirage avec probabilités inégales, tel qu'il a été pratiqué, donne plus de chance aux grosses unités de faire partie de l'échantillon et que, pour ces unités, la variable étudiée présente, en moyenne, des valeurs d'autant plus élevées que la corrélation est forte, il s'en suit que l'estimation du total obtenue en accordant à toutes les valeurs un poids égal est d'autant plus biaisée que la corrélation est grande.

Les estimations des totaux basées sur les formules relatives au tirage avec remise et probabilités proportionnelles à la taille (formule (9)), conduisent à des résultats biaisés dans 181 cas sur 352 mais, contrairement à la situation précédente, les totaux sont sous-estimés.

La comparaison des valeurs π_i et nz_i intervenant dans les formules relatives à \widehat{Y}_1 et \widehat{Y}_3 permet d'expliquer cette sous-estimation. A titre d'exemple, la figure 2 donne les valeurs des π_i en fonction des z_i pour la distribution des tailles des 132 communes rurales et pour les quatre fractions sondées. Afin de comparer les poids accordés aux différentes unités dans l'estimation du total, on a également tracé, sur le graphique, les quatre droites d'« iso-poids », d'équations $\pi_i = nz_i$.

Pour une fraction sondée faible, on constate que le nuage reste concentré autour de cette droite, l'écart entre les deux pondérations étant faible. Plus la

TABLEAU 2
Valeurs de t_{obs} relatives aux tests de conformité des moyennes
des estimateurs \hat{Y}_1 , \hat{Y}_2 et \hat{Y}_3 : valeurs moyennes
en fonction de la fraction sondée et du coefficient de corrélation

Fractions sondées	Estimateurs	Coefficients de corrélation					Moyennes
		0,0-0,2	0,2-0,4	0,4-0,6	0,6-0,8	0,8-1,0	
0,10	\hat{Y}_1	-0,63	0,20	-0,40	-0,05	-0,14	-0,25
	\hat{Y}_2	0,31	7,37	9,79	14,76	15,58	8,79
	\hat{Y}_3	-0,75	-0,13	-0,83	-0,75	-1,05	-0,74
0,25	\hat{Y}_1	-0,31	-0,03	-0,17	-0,01	0,07	-0,10
	\hat{Y}_2	1,15	10,59	15,71	23,55	26,46	14,48
	\hat{Y}_3	-0,50	-1,23	-1,80	-2,84	-3,89	-2,03
0,50	\hat{Y}_1	-0,28	0,32	-0,36	0,14	0,22	-0,00
	\hat{Y}_2	0,94	12,48	19,44	31,58	44,50	21,29
	\hat{Y}_3	-1,01	-3,51	-5,84	-9,02	-14,04	-6,71
0,75	\hat{Y}_1	0,36	-0,14	0,27	-0,03	-0,07	0,09
	\hat{Y}_2	0,78	8,54	17,44	27,07	59,62	24,08
	\hat{Y}_3	-1,20	-8,69	-13,09	-22,05	-38,33	-17,01
Moyennes	\hat{Y}_1	-0,22	0,09	-0,17	0,01	0,02	-0,07
	\hat{Y}_2	0,80	9,74	15,59	24,24	36,54	17,16
	\hat{Y}_3	-0,86	-3,39	-5,39	-8,66	-14,33	-6,62

fraction sondée augmente, plus le nuage se courbe et s'écarte de la droite d'«isopoids». Il s'en écarte d'autant plus que les unités sont de taille importante. On remarque également que la droite coupe nettement le nuage en deux, séparant ainsi les unités pour lesquelles π_i est supérieur à nz_i des unités pour lesquelles π_i est inférieur à nz_i . Dans le cas où π_i est supérieur à nz_i , les unités sont de petite taille et leur poids dans l'estimation du total par la relation (9) est supérieur à celui qu'elles ont dans la relation (1) et la situation inverse se présente pour les unités de grande taille, lorsque π_i est inférieur à nz_i . De plus, l'augmentation de la fraction sondée déplace le point d'intersection de la droite et du nuage de points vers les unités de taille plus petite, ce qui augmente le nombre d'unités qui ont une pondération plus faible lors de l'utilisation de la relation (9). L'examen des graphiques relatifs aux autres populations a conduit aux mêmes observations, et, de plus, on a pu constater que les divergences entre les π_i et les nz_i s'accroissent avec l'augmentation de la dissymétrie de la population des tailles.

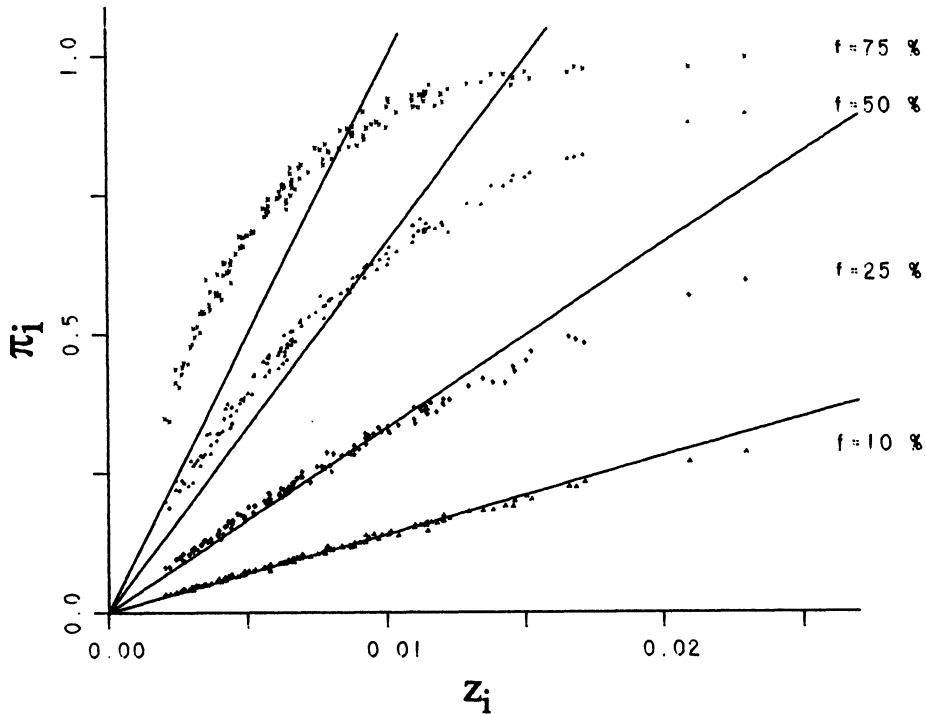


FIGURE 2

Diagrammes de dispersion des probabilités d'inclusion (π_i) et des tailles relatives (z_i) des unités de la distribution 1 et droite «iso-poids» ($\pi_i = nz_i$) dans l'estimation d'un total, pour les quatre fractions sondées (f) considérées.

De la comparaison des π_i et des nz_i , on peut conclure que l'estimateur \hat{Y}_3 accorde un poids trop faible aux unités de grande taille et un poids trop élevé aux unités de petite taille. En présence de corrélation positive entre la taille et le caractère observé, on doit donc s'attendre à obtenir des valeurs sous-estimées par l'utilisation de l'estimateur \hat{Y}_3 .

5. Variance de l'estimateur de HORVITZ-THOMPSON

Pour chacun des 352 groupes de 100 échantillons prélevés dans des conditions identiques, les quatre variances suivantes ont été calculées :

- (a) $V(\hat{Y}_1)$: variance théorique de l'estimateur de HORVITZ et THOMPSON (relation (2));
- (b) $\hat{V}(Y_1)$: estimation de la variance de l'estimateur de HORVITZ et THOMPSON, calculée à partir des 100 totaux estimés :

$$\widehat{V}(\widehat{Y}_1) = \frac{1}{99} \sum_{i=1}^{100} \left[\widehat{Y}_{1i} - \frac{1}{100} \left(\sum_{i=1}^{100} \widehat{Y}_{1i} \right) \right]^2;$$

- (c) $\bar{v}_{HT}(\widehat{Y}_1)$: moyenne arithmétique des 100 variances estimées par la relation (4);
- (d) $\bar{v}_{YG}(\widehat{Y}_1)$: moyenne arithmétique des 100 variances estimées par la relation (5).

On a ensuite calculé la moyenne arithmétique de ces valeurs par fraction sondée et par classe de corrélation. Les résultats sont donnés dans le tableau 3. On constate que toutes ces mesures de la variance concordent bien dans l'ensemble et qu'aucune des deux formules d'estimation de la variance de l'estimateur de HORVITZ et THOMPSON (relations (4) et (5)) ne semble supérieure à l'autre. D'autre part, parmi les 352 distributions d'échantillonnage de ces deux estimations de variance, on en a relevé respectivement 52 et 46 présentant au moins une estimation de variance négative sur les 100 échantillons simulés. Pour ces distributions-là, le nombre de valeurs négatives est toujours inférieur ou égal à 5, sauf dans un seul cas, relatif à la formule (4), où le nombre était compris entre 5 et 10. Les distributions présentant des valeurs négatives ont toujours été observées pour les fractions sondées faibles. On a constaté également que seulement 15 distributions sur les 352 présentent simultanément une ou plusieurs estimations de la variance négatives pour chacune des deux formules, ces concordances ne s'observant d'ailleurs que pour la fraction sondée la plus faible.

Afin de comparer la précision des deux estimateurs de la variance, on a également déterminé, pour les 352 cas, l'écart-type des 100 variances estimées. Les valeurs moyennes de ces écarts-types sont données dans le tableau 4, par fractions sondées et par classes de coefficient de corrélation. L'examen de ce tableau ne fait pas apparaître de différence importante entre les deux estimateurs.

On peut donc conclure que les deux estimateurs de la variance ne se distinguent pas nettement l'un de l'autre, tant pour le biais que pour la précision. Ceci confirme l'affirmation de GOURIEROUX (1981).

6. Intérêt de l'échantillonnage avec probabilités inégales et sans remise

L'échantillonnage avec probabilités inégales et sans remise, tel qu'il a été défini au paragraphe 1, exige, pour l'analyse, un volume de calcul plus important que l'échantillonnage aléatoire et simple ou que l'échantillonnage avec probabilités proportionnelles à la taille et avec remise, puisqu'il est indispensable d'estimer, par simulation, les probabilités d'inclusion π_i et π_{ij} . On peut se demander si ce surcroît de calcul est compensé par une précision plus grande de l'estimation du total.

Pour répondre à cette question, on a calculé les variances théoriques $V(\widehat{Y}_2)$

TABLEAU 3

Variances de l'estimateur de HORVITZ et THOMPSON :
moyennes obtenues pour les différentes variances en fonction de la fraction sondée
et du coefficient de corrélation

Fractions sondées	Types de variance	Coefficients de corrélation				
		0,0-0,2	0,2-0,4	0,4-0,6	0,6-0,8	0,8-1,0
0,10	$V(\hat{Y}_1)$	1.422	1.639	1.506	1.350	1.299
	$\hat{V}(\hat{Y}_1)$	1.356	1.586	1.612	1.294	1.364
	$\bar{v}_{HT}(\hat{Y}_1)$	1.391	1.632	1.583	1.235	1.335
	$\bar{v}_{YG}(\hat{Y}_1)$	1.071	1.525	1.509	1.183	1.559
0,25	$V(\hat{Y}_1)$	485	542	507	448	415
	$\hat{V}(\hat{Y}_1)$	446	477	496	408	420
	$\bar{v}_{HT}(\hat{Y}_1)$	469	530	620	440	419
	$\bar{v}_{YG}(\hat{Y}_1)$	473	533	525	443	413
0,50	$V(\hat{Y}_1)$	159	179	171	143	126
	$\hat{V}(\hat{Y}_1)$	146	164	157	129	117
	$\bar{v}_{HT}(\hat{Y}_1)$	148	177	173	138	122
	$\bar{v}_{YG}(\hat{Y}_1)$	149	178	173	137	125
0,75	$V(\hat{Y}_1)$	45	53	53	40	35
	$\hat{V}(\hat{Y}_1)$	48	57	55	42	37
	$\bar{v}_{HT}(\hat{Y}_1)$	48	55	54	41	36
	$\bar{v}_{YG}(\hat{Y}_1)$	49	56	55	42	36

et $V(\hat{Y}_3)$ par les relations (7) et (10). Toutefois, dans cette dernière relation, on a remplacé l'effectif n d'unités différentes correspondant à la fraction sondée de 0,10, 0,25, 0,50 ou 0,75 par l'effectif n' ($n' > n$) représentant l'effectif moyen de l'échantillon avec remise qu'il faut prélever pour obtenir n unités différentes. Les valeurs de n' ont été déterminées à partir des simulations, en gardant en mémoire le nombre de fois que chaque unité sélectionnée sans remise aurait été sélectionnée dans un tirage avec remise. Le tableau 5 donne, pour chaque distribution de tailles et par fractions sondées, le rapport moyen n'/n .

Les valeurs moyennes des trois variances théoriques $V(\hat{Y}_1)$, $V(\hat{Y}_2)$ et $V(\hat{Y}_3)$ sont données dans le tableau 6, en fonction de la fraction sondée et du coefficient de corrélation.

La variance $V(\hat{Y}_2)$ relative au tirage aléatoire et simple est évidemment indépendante de la valeur du coefficient de corrélation, et les faibles variations

TABLEAU 4
 Ecarts-types des variances estimées $v_{HT}(\hat{Y}_1)$ et $v_{YG}(\hat{Y}_1)$:
 moyennes obtenues en fonction de la fraction sondée
 et du coefficient de corrélation

Fractions sondées	Types de variance	Coefficients de corrélation					Moyennes
		0,0-0,2	0,2-0,4	0,4-0,6	0,6-0,8	0,8-1,0	
0,10	$v_{HT}(\hat{Y}_1)$	3.675	2.396	2.534	2.162	1.710	2.571
	$v_{YG}(\hat{Y}_1)$	2.216	2.258	2.397	1.626	3.532	2.554
0,25	$v_{HT}(\hat{Y}_1)$	773	479	541	444	271	511
	$v_{YG}(\hat{Y}_1)$	906	565	629	522	318	599
0,50	$v_{HT}(\hat{Y}_1)$	143	102	104	85	53	99
	$v_{YG}(\hat{Y}_1)$	158	120	113	91	58	109
0,75	$v_{HT}(\hat{Y}_1)$	24	19	19	15	12	18
	$v_{YG}(\hat{Y}_1)$	26	21	22	17	12	20
Moyennes	$v_{HT}(\hat{Y}_1)$	1.154	749	800	677	511	800
	$v_{YG}(\hat{Y}_1)$	826	741	790	564	980	821

TABLEAU 5
 Rapport moyen n'/n du nombre total d'unités sélectionnées
 avec remise au nombre d'unités distinctes souhaitées lors du tirage
 avec probabilité proportionnelle à la taille

Distributions des tailles	Fractions sondées			
	0,10	0,25	0,50	0,75
Distribution 1	1,07	1,19	1,50	2,24
Distribution 2	1,08	1,25	1,70	2,86
Distribution 3	1,13	1,31	1,93	3,23
Distribution 4	1,05	1,16	1,40	1,87
Distribution 5	1,10	1,32	1,93	3,27
Distribution 6	1,05	1,15	1,38	1,88

qui s'observent pour les valeurs moyennes proviennent du déséquilibre qui existe au niveau du plan de simulation (paragraphe 3). On constate aussi que la variance

du tirage aléatoire et simple, $V(\hat{Y}_2)$, est en moyenne inférieure à la variance des tirages avec probabilités inégales et que le tirage avec probabilités inégales et sans remise est en moyenne plus précis que le tirage avec probabilités inégales et avec remise.

En fait, un examen plus détaillé des 352 cas simulés montre que $V(\hat{Y}_1)$ n'est inférieur à $V(\hat{Y}_2)$ que dans 53 cas, correspondant en général à des fractions sondées élevées et à de fortes corrélations. Au vu de ces résultats, l'échantillonnage avec probabilités inégales ne semble donc pas préférable au tirage avec probabilités égales sauf dans ces situations extrêmes où la fraction sondée est très élevée (supérieure à 0,75) et la corrélation très forte (supérieure à 0,8). De même, $V(\hat{Y}_1)$ n'est supérieur à $V(\hat{Y}_3)$ que dans 18 cas sur 352, ce qui confirme bien l'intérêt de n'utiliser que les unités distinctes pour l'estimation du total. On trouve donc, dans le cas de l'échantillonnage avec probabilités inégales, un principe connu et démontré depuis longtemps pour l'échantillonnage avec probabilités égales (DES RAJ et KHAMIS, 1958). Toutefois, le gain de précision reste faible, du moins pour des valeurs raisonnables de la fraction sondée.

TABLEAU 6

Valeurs moyennes des variances théoriques dans le cas de l'échantillonnage avec probabilités inégales et sans remise ($V(\hat{Y}_1)$), de l'échantillonnage avec probabilités constantes et sans remise ($V(\hat{Y}_2)$) et de l'échantillonnage avec probabilités proportionnelles à la taille et avec remise ($V(\hat{Y}_3)$) : moyennes obtenues en fonction de la fraction sondée et du coefficient de corrélation

Fractions sondées	Variances	Coefficients de corrélation				
		0,0-0,2	0,2-0,4	0,4-0,6	0,6-0,8	0,8-1,0
0,10	$V(\hat{Y}_1)$	1.422	1.639	1.506	1.350	1.299
	$V(\hat{Y}_2)$	792	1.045	907	902	1.084
	$V(\hat{Y}_3)$	1.488	1.729	1.569	1.421	1.377
0,25	$V(\hat{Y}_1)$	485	542	507	448	415
	$V(\hat{Y}_2)$	351	311	308	363	321
	$V(\hat{Y}_3)$	543	609	569	511	486
0,50	$V(\hat{Y}_1)$	159	179	171	143	126
	$V(\hat{Y}_2)$	119	105	104	123	109
	$V(\hat{Y}_3)$	196	230	221	192	189
0,75	$V(\hat{Y}_1)$	45	53	53	40	35
	$V(\hat{Y}_2)$	39	35	34	40	36
	$V(\hat{Y}_3)$	84	99	95	82	83

7. Conclusions

L'objectif de la présente étude était triple. Il s'agissait d'abord de proposer une analyse correcte d'un échantillon prélevé sans remise et avec probabilités inégales et inconnues *a priori*. Ensuite, on souhaitait étudier l'erreur commise lors de l'emploi de solutions approchées. Enfin, on désirait comparer la précision de l'échantillonnage sans remise et avec probabilités inégales avec la précision de méthodes d'échantillonnage plus simples.

Pour le premier point, on a montré que les probabilités d'inclusion π_i et π_{ij} peuvent être estimées par simulation d'un grand nombre d'échantillons. Ensuite, l'estimateur de HORVITZ et THOMPSON (relation(1)) peut être utilisé pour l'estimation du total. Quant à la variance de cet estimateur, elle peut être obtenue par l'estimateur de HORVITZ et THOMPSON (relation (4)) ou par l'estimateur de YATES et GRUNDY (relation (5)), ces deux estimateurs étant de qualité comparable.

En ce qui concerne les solutions approchées, on a montré, d'une part, que le traitement de l'échantillon comme s'il avait été prélevé de manière aléatoire et simple (relation (6)) conduit à une surestimation et, d'autre part, que le traitement de l'échantillon comme s'il provenait d'un tirage avec remise et probabilités proportionnelles à la taille (relation (9)) donne lieu à une sous-estimation. Les biais sont d'autant plus importants que la corrélation entre la caractéristique y et la taille x des unités est élevée et que la fraction sondée est grande. Ces résultats ne sont toutefois valables que pour des corrélations positives.

Enfin, l'étude a également montré que, pour les simulations réalisées, l'échantillonnage avec probabilités inégales n'était, en général, pas supérieur à l'échantillonnage avec probabilités égales et que l'échantillonnage sans remise et avec probabilités inégales était supérieur à l'échantillonnage avec remise et probabilités inégales, la différence entre ces deux méthodes n'étant sensible que pour des fractions sondées supérieures à 0,10.

8. Remerciements

Les auteurs remercient vivement le professeur P. DAGNELIE pour tous les conseils et remarques dont il leur a fait part.

Bibliographie

- BERTRAND F. (1986). *Echantillonnage avec probabilités inégales et sans remise*. Gembloux, Faculté des Sciences Agronomiques, 129 p.
- COCHRAN W.G. (1977). *Sampling techniques*. 3ème éd. New-York, Wiley, 428 p.
- DEROO M. et DUSSAIX A.M. (1980). *Pratique et analyse des enquêtes par sondage*. Paris, Presses Universitaires de France, 302 p.
- DESABIE J. (1966). *Théorie et pratique des sondages*. Paris. Dunod, 481 p.

- DES RAJ et KHAMIS S.H. (1958). Some remarks on sampling with replacement. *Ann. Math. Statist.*, 29, 550-557.
- DROESBEKE J.J., FICHET B. et TASSI P. (eds) (1987). *Les sondages*. Paris, Economica, 310 p.
- GOURIEROUX C. (1981). *Théorie des sondages*. Paris, Economica, 272 p.
- GROSBRAS J.-M. (1987). *Méthodes statistiques des sondages*. Paris, Economica, 331 p.
- HANIF M. et BREWER K.R.W. (1980). Sampling with unequal probabilities without replacement : a review. *Intern. Statist. Rev.*, 48, 317-335.
- HORVITZ D.G. et THOMPSON D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Ass.*, 47, 663-685.
- KEIL U., CAIRNS V., DORING A., HARTEL U., JORCIK J., PERZ S. et STIEBER J. (1985). *Manual of operations. Survey*. München, Gesellschaft für Strahlen - und Umweltforschung, Bericht 20/85, 210 p.
- VAN GELDEREN C. (1983). *Enquête sur les comportements et attitudes liés à la morbidité cardio-vasculaire dans la province de Luxembourg. Constitution d'un échantillon probabiliste stratifié à deux degrés*. Louvain-La-Neuve. Association interuniversitaire pour la prévention des maladies cardio-vasculaires. Document de recherche n° 1, 22 p.
- VOS J.W.E. (1974a). Steekproeftrekking met ongelijke kansen (Deel I). *Statist. Neerl.*, 28, 11-49.
- VOS J.W.E. (1974b). Steekproeftrekking met ongelijke kansen (Deel II). *Statist. Neerl.*, 28, 69-108.
- YATES F. et GRUNDY P.M. (1953). Selection without replacement from within strata with probabilities proportional to size. *J. Roy. Statist. Soc., Series B*, 15, 253-261.