Techniques d'enquête, décembre 1996
Vol. 22, n° 2, pp. 169-175
Statistique Canada

Variance asymptotique pour un plan séquentiel sans remise à probabilités inégales

YVES G. BERGER[1]

RÉSUMÉ

Nous proposons une approximation des probabilités d'inclusion d'ordre deux pour le plan de Chao (1982) en vue d'obtenir un estimateur de variance approché pour l'estimateur de Horvitz et Thompson. Ensuite, nous comparerons cette variance à d'autres approximations données pour le plan systématique randomisé (Hartley et Rao 1962), le plan réjectif (Hájek 1964) et le plan de Rao-Sampford (Rao 1965 et Sampford 1967). Nous conclurons que ces approximations sont équivalentes si les probabilités d'inclusion d'ordre un sont petites et si la taille de l'échantillon est grande.

MOTS CLÉS: Sondage avec remise; plan systématique randomisé; plan réjectif; plan de Rao-Sampford; probabilités d'inclusion; Horvitz-Tompson; Yates-Grundy.

1. INTRODUCTION

Soit U_N une population finie contenant N unités. Considérons U_k un sous-ensemble de U_N constitué des k premières unités de U_N. Nous notons $\pi_{(k;i)}$ les probabilités d'inclusion d'ordre un pour une population U_k. Nous supposons qu'elles sont proportionnelles à une variable auxiliaire. Ces probabilités ont deux arguments: la taille k de la population et le numéro d'ordre i de l'unité dans la population. Nous supposons que $\pi_{(k;i)} < 1$ pour tous les i et tous les $k > n$. Cette hypothèse a plus de chance d'être violée dans le cas où k est petit, c'est-à-dire proche de n. Nous pouvons remédier à ce problème en supposant que les valeurs de la variable auxiliaire sont peu dispersées pour les unités se trouvant au début de la population.

Nous notons $\pi_{(k;i,j)}$ la probabilité d'inclusion d'ordre deux des unités i et j pour une population U_k. Ces probabilités dépendent du plan de sondage utilisé.

Nous utiliserons l'estimateur de Horvitz-Thompson (1951) pour estimer le total $\sum_{i=1}^{N} Y_i$ d'une variable Y. Cet estimateur est donné par

$$t_{HT} = \sum_{i \in S_N} \frac{Y_i}{\pi_{(N;i)}}; \qquad (1)$$

où S_N est un échantillon de U_N. Nous supposons que la taille de S_N est constante et égale à n.

Étant donné que la taille de l'échantillon est fixée, un estimateur de la variance de (1) est donné par l'estimateur de Yates-Grundy (1953),

$$\hat{V} = \sum_{j \in S_N} \sum_{i \in S_N; i<j} \frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \left[\frac{Y_i}{\pi_{(N;i)}} - \frac{Y_j}{\pi_{(N;j)}} \right]^2, \qquad (2)$$

où

$$\Delta_{(N;i,j)} = \pi_{(N;i,j)} - \pi_{(N;i)} \pi_{(N;i)}. \qquad (3)$$

Considérons la séquence de taille d'échantillon $\{n_1, n_2, ..., n_\nu, ...\}$ et la séquence de taille de population $\{N_1, N_2, ..., N_\nu, ...\}$, où n_ν et N_ν augmentent lorsque $\nu \to \infty$. Pour simplifier le problème nous supprimons l'indice ν.

L'approche asymptotique adoptée ici est celle d'Hájek (1964):

$$d = \sum_{j=1}^{N} \pi_{(N;j)}[1 - \pi_{(N;j)}] \to \infty,$$

ce qui signifie que $n \to \infty$ et $(N - n) \to \infty$, étant donné que $d \leq \sum_{j=1}^{N} [1 - \pi_{(N;j)}] = N - n$ et $d \leq \sum_{j=1}^{N} \pi_{(N;j)} = n$.

Dans la section 2, nous présentons le plan de sondage de Chao (1982) ainsi que trois résultats se rapportant aux probabilités d'inclusion d'ordre un et deux. Dans la section 3, nous donnons une approximation des $\pi_{(N;i,j)}$. Dans la section 4, nous proposons une approximation de la variance de Yates-Grundy. La section 5 est consacrée à la comparaison de cette approximation de variance avec d'autres approximations proposées pour le plan systématique randomisé, le plan réjectif et le plan de Rao-Sampford. Deux exemples numériques sont présentés dans la section 6.

2. PLAN DE SONDAGE DE CHAO

Il s'agit d'un plan de sondage sans remise à probabilités inégales, à taille fixée. Cette méthode est la généralisation de la méthode de McLeod et Bellhouse (1983) pour un plan simple.

Soit S_k un échantillon de taille n de U_k avec un ensemble $\{\pi_{(k;i)}: i \in U_k\}$ de probabilités d'inclusion d'ordre un. Le plan de Chao permet d'avoir un échantillon S_{k+1} de taille n de U_{k+1} avec un ensemble $\{\pi_{(k+1;i)}: i \in U_{k+1}\}$ de probabilités d'inclusion d'ordre un. La méthode consiste à sélectionner la $(k+1)$-ième unité avec une probabilité $\pi_{(k+1;k+1)}$. Si cette unité n'est pas sélectionnée, alors nous prenons $S_{k+1} = S_k$; sinon nous prenons $S_{k+1} = S_k \cup \{k+1\} \backslash \{j\}$, où j est une unité sélectionnée au

[1] Yves Berger, Université Libre de Bruxelles, Laboratoire de Méthodologie du Traitement des Données, C.P. 124, Avenue Jeanne, 44, Bruxelles, Belgique, E-Mail : yvberger@ulb.ac.be

hasard dans S_k. La procédure démarre à partir d'un échantillon initial, $S_n = U_n$, constitué des n premières unités de la population.

Le plan de Chao a l'avantage d'être séquentiel. En effet, il permet de sélectionner un échantillon par un simple parcours séquentiel de la population. Le plan systématique est un autre plan séquentiel souvent utilisé. Mais ce dernier a le désavantage d'induire des probabilités d'inclusion d'ordre deux nulles. On peut éviter ce problème en randomisant le plan systématique. Dans ce cas, la population est triée de manière aléatoire avant que l'échantillon ne soit sélectionné. Cette opération élimine partiellement le problème des probabilités d'inclusion d'ordre deux nulles. Par contre, comme nous le verrons à la fin de cette section, le plan de Chao a l'avantage de ne pas avoir de probabilités d'inclusion d'ordre deux nulles. Une randomisation n'est donc pas nécessaire pour ce dernier.

Le plan réjectif et le plan de Rao-Sampford ont le désavantage de ne pas être séquentiels. En effet, les unités sont sélectionnées au hasard avec remise dans la population. Si une unité est sélectionnée deux fois, on est obligé de sélectionner un nouvel échantillon. Ces deux plans, bien que plus simple à comprendre, sont plus difficiles à mettre en oeuvre que le plan de Chao.

Le théorème suivant, qui est une application directe du théorème donné par Chao (1982), donne une relation entre la probabilité d'inclusion d'ordre un, $\pi_{(k;i)}$, de la i-ième unité de U_k et la probabilité d'inclusion d'ordre un, $\pi_{(k+1;i)}$, de la i-ième unité de U_{k+1}.

Théorème 1

$$\pi_{(k+1;i)} = \begin{cases} [1 - \pi_{(k+1;k+1)} R_{(k;i)}]\, \pi_{(k;i)}, & \text{pour} \quad i < k+1; \\ \pi_{(k+1;k+1)} & \text{, pour} \quad i = k+1; \end{cases} \tag{4}$$

où

$$R_{(k;i)} = \begin{cases} \dfrac{1 - \pi_{(n+1;i)}}{\pi_{(n+1;n+1)}}, & \text{pour} \quad k = n, \\ \dfrac{1}{n} & \text{, pour} \quad k \geq n+1. \end{cases} \tag{5}$$

Les probabilités d'inclusion d'ordre deux peuvent être calculées de manière itérative en utilisant le théorème suivant:

Théorème 2 (Chao, 1982)

$$\pi_{(k;i,j)} = \begin{cases} \{1 - \pi_{(k;k)}[R_{(k-1;i)} + R_{(k-1;j)}]\}\pi_{(k-1;i,j)}, & \text{pour} \quad i < j < k; \\ \pi_{(k;k)}[1 - R_{(k-1;i)}]\,\pi_{(k-1;i)} & \text{, pour} \quad i < j = k. \end{cases}$$

Bethlehem et Schuerhoff (1984) donnent une condition nécessaire et suffisante pour que les probabilités d'inclusion d'ordre deux soient strictement positives, pour une population U_k:

$\#\{i : i \leq \ell \text{ et } \pi_{(\ell;i)} = 1\} \neq n - 1$, pour tout ℓ tel que $n < \ell \leq k$.

Comme $\pi_{(\ell;i)} < 1$ pour tous i et ℓ tels que $i \leq \ell \leq k$, cette condition est toujours satisfaite. Donc, dans le cadre de cet article, nous n'aurons jamais de probabilités d'inclusion d'ordre deux nulles.

De plus, la quantité $\Delta_{(N;i,j)}$ est toujours négative si on utilise le plan de Chao (Chao 1982, p. 656). Dès lors, la variance de Yates-Grundy a l'avantage d'être toujours positive.

3. APPROXIMATION DES PROBABILITÉS D'INCLUSION D'ORDRE 2

Le théorème suivant nous donne une expression asymptotique pour les probabilités d'inclusion d'ordre deux, pour le plan de Chao.

Théorème 3

$$\pi_{(N;i,j)} \approx \begin{cases} \pi_{(N;i)}\,\pi_{(N;j)}\,\dfrac{n-1}{n - p_{(j)}} & \text{, si } j > n+1; \\ \pi_{(N;i)}\,\pi_{(N;j)}\,\dfrac{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1}{\pi_{(n+1;i)}\,\pi_{(n+1;j)}}, & \text{si } j \leq n+1; \end{cases} \tag{6}$$

où $p_{(j)} = \pi_{(j;j)}$ et $i < j$.

La preuve de ce théorème se trouve en appendice I.

Nous constatons que cette approximation a une structure différente suivant que $j > n+1$ ou que $j \leq n+1$. Pour éviter ce problème, nous allons imposer une condition plausible sur la variable auxiliaire, de manière à ce qu'il y ait équivalence entre ces deux structures. Considérons l'hypothèse énoncée dans l'introduction, selon laquelle les valeurs de la variable auxiliaire sont peu dispersées pour les $n+1$ premières unités de la population. Plus précisément, nous supposons que la variable auxiliaire est constante pour les $n+1$ premières unités, c'est-à-dire:

$$\pi_{(n+1;i)} = \frac{n}{n+1} \quad \text{pour } i \leq n+1.$$

Dans ce cas,

$$\frac{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1}{\pi_{(n+1;i)}\,\pi_{(n+1;j)}} = \frac{n-1}{n - \pi_{(n+1;j)}}.$$

En utilisant (6), nous avons l'approximation suivante pour les probabilités d'inclusion d'ordre deux

$$\pi_{(N;i,j)} \approx \pi_{(N;i)}\,\pi_{(N;j)}\,\frac{n-1}{n - p_{(j)}} \quad \text{si } i < j; \tag{7}$$

où

$$p_{(j)} = \begin{cases} \pi_{(j;j)} & \text{, si} \quad j > n+1, \\ \pi_{(n+1;j)}, & \text{si} \quad j \leq n+1. \end{cases} \tag{8}$$

4. ESTIMATEUR DE VARIANCE

La relation (7) conduit à l'approximation suivante des $\Delta_{(N;i,j)}$:

$$\tilde{\Delta}_{(N;i,j)} = \pi_{(N;i)}\pi_{(N;j)}\frac{p_{(j)} - 1}{n - p_{(j)}}, \quad \text{si} \quad i < j. \tag{9}$$

(2), (7) et (9) permettent d'établir une expression asymptotique pour l'estimateur de Yates-Grundy.

$$\tilde{V}_C = \frac{1}{[n-1]}\sum_{j\in S_N}[1 - p_{(j)}]\sum_{i\in S_N; i<j}\left[\frac{Y_i}{\pi_{(N;i)}} - \frac{Y_j}{\pi_{(N;j)}}\right]^2. \tag{10}$$

Mais cette expression a tendance à sous-estimer la variance. En effet, pour établir la relation (6), on utilise l'approximation (19) de l'appendice I. Cette approximation implique toujours que:

$$\pi_{(N;i,j)} < \pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n - p_{(j)}}. \tag{11}$$

On peut s'en rendre compte facilement en constatant que (20) est obtenu à partir de (18) en utilisant l'approximation (19). L'inégalité (11) est donc vraie pour $j > n + 1$. Pour $j \le n + 1$, il suffit de constater que (21) est également obtenu à partir de (19). L'inégalité (11) implique que:

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} > \frac{1 - p_{(j)}}{n-1}, \tag{12}$$

étant donnée que $\Delta_{(N;i,j)} < 0$. Par (2), (10) et (12), nous avons effectivement

$$\hat{V} > \tilde{V}_C.$$

Pour remédier à ce problème de sous-estimation de la variance, nous proposons de faire un redressement sur (9). Il est bien connu que:

$$\sum_{i=1;i\neq j}^{N}\pi_{(N;i,j)} = (n-1)\pi_{(N;j)}. \tag{13}$$

L'approximation (7) ne respecte pas la contrainte (13). Le redressement consiste à supposer les $p_{(j)}$ inconnus et à les choisir de manière à ce que (13) soit satisfaite pour l'approximation des probabilité d'ordre deux. C'est-à-dire:

$$\sum_{i=1}^{j-1}\pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n-p_{(j)}} + \sum_{i=j+1}^{N}\pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n-p_{(i)}} = (n-1)\pi_{(N;j)}.$$

Cette contrainte peut s'écrire:

$$\sum_{i=1}^{j-1}\pi_{(N;i)} + \sum_{i=j+1}^{N}\pi_{(N;i)}\frac{n - p_{(j)}}{n - p_{(i)}} = n - p_{(j)}. \tag{14}$$

Étant donné $\sum_{j=1}^{N}\pi_{(N;j)} = n$, la contrainte (14) est pratiquement vérifiée si

$$p_{(i)} = \pi_{(N;i)} \tag{15}$$

$$\sum_{i=j+1}^{N}\pi_{(N;i)}\frac{n - \pi_{(N;j)}}{n - \pi_{(N;i)}} \approx \sum_{i=j+1}^{N}\pi_{(N;i)}. \tag{16}$$

La relation (16) est plausible étant donné que la différence entre la partie gauche et la partie droite de (16) a comme borne inférieure

$$\frac{1}{n}\sum_{i=j+1}^{N}\pi_{(N;i)}[\pi_{(N;i)} - \pi_{(N;j)}],$$

et comme borne supérieure

$$\frac{1}{n-1}\sum_{i=j+1}^{N}\pi_{(N;i)}[\pi_{(N;i)} - \pi_{(N;j)}].$$

Ces deux bornes sont proches de zéro lorsque les $\pi_{(N;i)}$ sont peu dispersées. Ceci signifie que la solution (15) est appropriée lorsque les $\pi_{(N;i)}$ sont petits. Ces deux bornes sont également d'autant plus proches de zéro que j est grand. Donc la solution (15) vérifie d'autant plus (13) que j est grand. Ceci implique que notre approximation (9) est très bonne pour les couples d'unités (i, j) $(i < j)$ tels que l'unité j est située à la fin de la population. En fait, nous voulons que l'approximation (9) soit la meilleure pour les couples d'unités (i, j) ayant une grande probabilité de se trouver dans l'échantillon (c'est-à-dire, pour les couples (i, j) $(i < j)$ dont $\pi_{(N;j)}$ est la plus grande). Il est donc préférable de placer les unités possédant des grandes probabilités d'inclusion d'ordre un à la fin de la population.

Si on choisi de prendre $p_{(i)} = \pi_{(N;i)}$, nous avons des $p_{(i)}$ plus petit que (8). Ceci conduit à une approximation de la variance plus grande. Cette solution est d'autant plus acceptable qu'elle correspond au résultat du plan simple sans remise. En effet, si on remplace dans (7) $\pi_{(N;i)}$, $\pi_{(N;j)}$ et $p_{(j)}$ par n/N, on obtient

$$\pi_{(N;i,j)} \approx \frac{n(n-1)}{N(N-1)}, \quad \text{si} \quad i > n + 1.$$

Cette expression correspond, bien évidemment, au résultat du plan simple sans remise.

En conclusion, nous approximons $\Delta_{(N;i,j)}$ par (9) avec $p_{(i)} = \pi_{(N;i)}$. Nous supposons que la population est triée de manière à ce que les unités ayant des petites $\pi_{(N;i)}$ se situent au début de la population et que les unités ayant des grandes $\pi_{(N;i)}$ se situent à la fin de la population. Nous supposons également que les $\pi_{(N;i)}$ ne sont pas trop dispersées pour les $n + 1$ premières unités de la population.

5. COMPARAISON AVEC D'AUTRES PLANS

Au lieu de comparer les probabilités d'inclusion d'ordre deux, nous allons comparer les quantités $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ qui sont utiles pour calculer la variance de Yates-Grundy. Nous allons regarder ce que donnent ces quantités pour le plan de Chao, le plan systématique randomisé (Hartley et Rao 1962), le plan réjectif (Hájek 1964) et le plan de Rao-Sampford (Rao 1965 et Sampford 1967).

Théorème 4

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \begin{cases} \dfrac{1-\pi_{(N;j)}}{n-1}, & \text{pour le plan de Chao;} \\ \dfrac{1-\pi_{(N;i)}-\pi_{(N;j)}}{n-1}, & \text{pour le plan systématique randomisé;} \\ \dfrac{n[1-\pi_{(N;i)}][1-\pi_{(N;j)}]}{d(n-1)}, & \text{pour le plan réjectif et pour le plan de Rao-Sampford.} \end{cases}$$

La preuve de ce théorème se trouve en appendice II.

Il est important de noter que l'approximation proposée pour le plan systématique randomisé provient de l'approximation de Deville (p. 21) et non de la célèbre approximation de Hartley-Rao (1962). Nous n'avons pas pu utiliser la formule de Hartley-Rao étant donné que celle-ci repose sur l'hypothèse asymptotique, n fixé et $N \to \infty$, différente de celle adoptée dans cet article.

Nous constatons que si les $\pi_{(N;i)}$ sont petites, $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ est équivalent pour le plan de Chao et pour le plan systématique. Mais, nous constatons que $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ est toujours plus petit dans le cas systématique que dans le cas de Chao. Cela est certainement dû au fait que l'approximation pour le plan systématique sous-estime $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$. On peut s'en rendre compte en remplaçant $\pi_{(N;i)}$ et $\pi_{(N;j)}$ par n/N. On obtient alors

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \frac{N-2n}{N(n-1)},$$

pour le plan systématique randomisé. Or, un tel plan est équivalent au plan simple. Il faudrait dès lors que,

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} = \frac{N-n}{N(n-1)}.$$

Nous proposons de redresser l'approximation de $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ pour le plan systématique en la multipliant par

$$\frac{N-n}{N-2n} = \frac{1-f}{1-2f},$$

où $f = n/N$ est le taux de sondage.

L'approximation de $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ pour le plan de Chao est également du même ordre que celle du plan réjectif. En effet, si les $\pi_{(N;i)}$ sont petites, nous avons l'approximation

$$\frac{n[1-\pi_{(N;i)}]}{d} \approx \frac{n[1-\pi_{(N;i)}]}{[1-\pi_{(N;i)}]\sum_{j=1}^{N}\pi_{(N;j)}} \approx 1.$$

Donc l'estimateur de Yates-Grundy est approximativement le même, que l'on utilise le plan de Chao, le plan systématique randomisé, le plan réjectif ou le plan de Rao-Sampford, pour n grand et des $\pi_{(N;i)}$ petites.

6. EXEMPLES NUMÉRIQUES

Les deux exemples suivants correspondent à deux cas extrêmes. Dans le premier exemple, les $\pi_{(N;i)}$ sont peu dispersées; dans le second, elles le sont beaucoup plus. Prenons un petit échantillon de taille 20. La population aura une taille de 50 pour que les $\pi_{(N;i)}$ ne soient pas trop petites. Nous nous sommes volontairement mis dans de mauvaises conditions, pour montrer que même avec un échantillon de taille 20 et une petite population, les résultats asymptotiques constituent déjà une bonne approximation.

Exemple 1

Prenons les probabilités d'inclusion d'ordre un représentées par la figure 1.

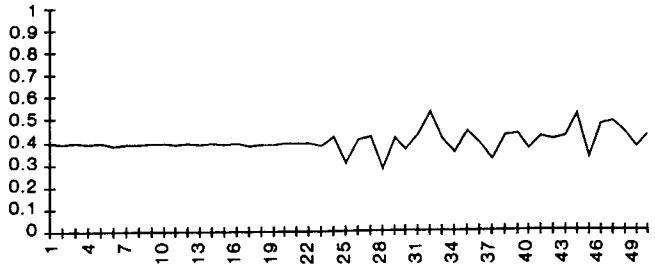

Figure 1. Probabilités d'inclusion d'ordre un, dans le cas de l'exemple 1

La figure 2 représente en ordonnée les vraies valeurs de $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$ pour le plan de Chao et en abscisse les approximations. Nous avons également représenté la droite où les approximations sont égales aux vraies valeurs. Les approximations seront d'autant meilleures que les points sont proches de la droite.

Nous avons une erreur moyenne de -0.000569 avec un écart type de 0.0015996. Ceci est très faible par rapport à l'ordre de grandeur des approximations. Le centre de gravité du nuage de point se trouve en (0.0313; 0.0318). Il peut paraître surprenant qu'il y ait moins de points à gauche du centre de gravité qu'à droite. Cela est simplement dû au fait que la plupart des points à gauche du centre de gravité sont superposés.

Nous constatons que les couples (i,j) avec $i<j$ tel que $\pi_{(N;j)}$ est grand correspondent à des points se trouvant à gauche. Ce sont les couples ayant la meilleure approximation. De plus,

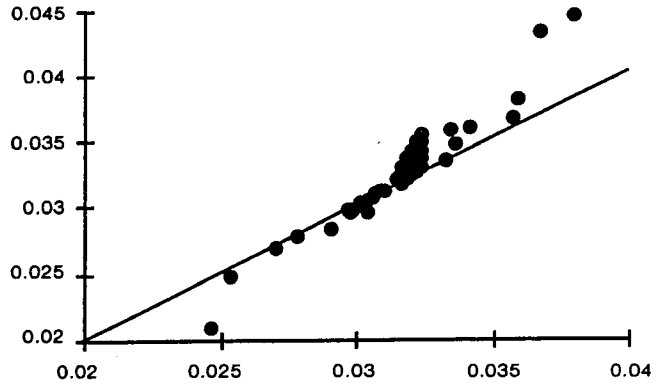

Figure 2. Approximations et vraies valeurs de $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$, dans le cas de l'exemple 1

ces couples ont une grande probabilité de se trouver dans l'échantillon étant donné que $\pi_{(N;j)}$ est grand. Donc notre variance approchée (10) est tout à fait acceptable.

Exemple 2

Les probabilités d'inclusion d'ordre un sont données dans la figure 3. Ici nous constatons que ces probabilités sont plus dispersées que dans l'exemple 1. La figure 4 donne les vraies valeurs ainsi que les approximations de $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$.

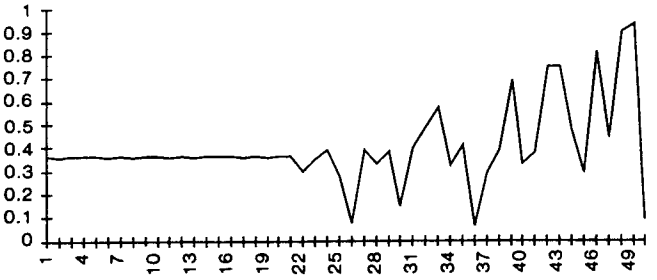

Figure 3. Probabilités d'inclusion d'ordre un, dans le cas de l'exemple 2

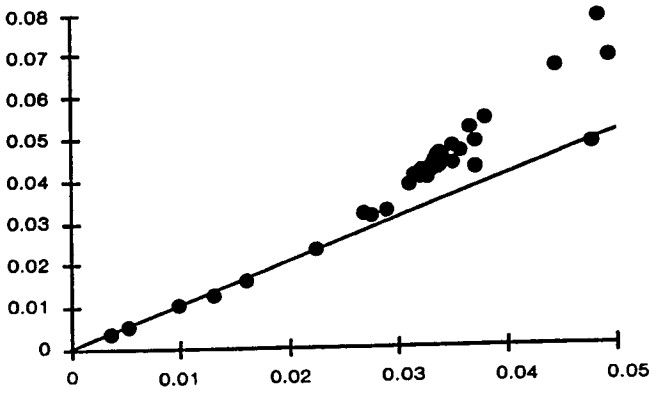

Figure 4. Approximations et vraies valeurs de $-\Delta_{(N;i,j)}/\pi_{(N;i,j)}$, dans le cas de l'exemple 2

Nous avons une erreur moyenne de -0.006999 avec un écart type de 0.006438. Le centre de gravité du nuage de points se trouve en (0.02957; 0.036606).

Nous aboutissons à la même conclusion que celle de l'exemple 1. Le second exemple conduit à de plus mauvaises approximations. Cela est simplement dû aux grandes probabilités d'inclusion d'ordre un.

7. CONCLUSION

Le plan de Chao possède plusieurs avantages: (i) il est séquentiel, (ii) les probabilités d'inclusion d'ordre deux sont positives et (iii) la variance de Yates-Grundy est toujours positive. Par contre, les probabilités d'inclusion d'ordre deux sont difficiles à calculer. C'est pourquoi, nous proposons de les approximer. Nous avons constaté que cette approximation est d'autant meilleure que le début de la population est constitué d'unités ayant de petites $\pi_{(N;i)}$ et que la fin de la population est constitué d'unités ayant de grandes $\pi_{(N;i)}$. Nous avons comparé notre approximation avec d'autres approximations données pour le plan systématique randomisé, le plan réjectif et le plan de Rao-Sampford. Nous avons conclu que ces approximations sont équivalentes si les probabilités d'inclusion d'ordre un sont petites et si la taille de l'échantillon est grande. Les deux exemples numériques qui terminent cet article, confirment les bons résultats de notre approximation.

ANNEXE I

Preuve du théorème 3

Avant d'entamer la preuve de ce théorème, on peut démontrer les deux lemmes suivants.

Lemme 1

$$\pi_{(k;i)} = p^*_{(i)} \prod_{\ell=a^*_i}^{k} \left[1 - \pi_{(\ell;\ell)} \frac{1}{n}\right];$$

où

$$p^*_{(i)} = \begin{cases} \pi_{(i;i)} & \text{si} \quad i > n+1; \\ \pi_{(n+1;i)} & \text{si} \quad i \leq n+1; \end{cases}$$

$$a^*_i = \begin{cases} i+1 & \text{si} \quad i > n+1; \\ n+2 & \text{si} \quad i \leq n+1. \end{cases} \qquad (17)$$

Lemme 2

$$\pi_{(k;ij)} = q^*_{(j)} \prod_{\ell=a^*_j}^{k} \left[1 - \pi_{(\ell;\ell)} \frac{2}{n}\right];$$

où $i < j$,

$$q^*_{(j)} = \begin{cases} \pi_{(j-1;i)}\pi_{(j;j)}\left(1 - \dfrac{1}{n}\right) & \text{si} \quad j > n+1; \\ \pi_{(n+1;i)} + \pi_{(n+1;j)} - 1 & \text{si} \quad j \le n+1; \end{cases}$$

et a_i^* est défini par (17).

Maintenant, avec ces deux lemmes, nous pouvons démontrer le théorème 3.

Preuve du théorème 3

Cas 1: Si $j > n+1$, par le lemme 2, nous avons

$$\pi_{(N;i,j)} = \pi_{(j-1;i)}\pi_{(j;j)}\left(1 - \frac{1}{n}\right)\prod_{\ell=j+1}^{N}\left[1 - \pi_{(\ell;\ell)}\frac{2}{n}\right].$$

Par le lemme 1, cette dernière expression devient

$$\pi(N;i,j) = p^*_{(i)}\pi_{(j;j)}\left(1 - \frac{1}{n}\right)\prod_{\ell=a_i^*}^{j-1}\left[1 - \pi_{(\ell;\ell)}\frac{1}{n}\right]\prod_{q=j+1}^{N}\left[1 - \pi_{(q;q)}\frac{2}{n}\right].$$

En multipliant cette dernière expression par

$$\left[\frac{1 - \pi_{(j;j)}\frac{1}{n}}{1 - \pi_{(j;j)}\frac{1}{n}}\right]\prod_{\ell=j+1}^{N}\left[\frac{1 - \pi_{(\ell;\ell)}\frac{1}{n}}{1 - \pi_{(\ell;\ell)}\frac{1}{n}}\right]\left[\frac{1 - \pi_{(\ell;\ell)}\frac{2}{n}}{1 - \pi_{(\ell;\ell)}\frac{2}{n}}\right] = 1,$$

et en regroupant certains termes, nous obtenons

$$\pi_{(N;i,j)} = \pi_{(j;j)}p^*_{(i)}\left[\frac{n-1}{n - \pi_{(j;j)}}\right]\prod_{\ell=a_i^*}^{N}\left[1 - \pi_{(\ell;\ell)}\frac{1}{n}\right]\prod_{q=j+1}^{N}\left[\frac{1 - \pi_{(q;q)}\frac{2}{n}}{1 - \pi_{(q;q)}\frac{1}{n}}\right].$$

Par le lemme 1, cette dernière expression devient

$$\pi_{(N;i,j)} = \left[\frac{n-1}{n - \pi_{(j;j)}}\right]\pi_{(N;i)}\pi_{(j;j)}\prod_{\ell=j+1}^{N}\left[\frac{1 - \pi_{(\ell;\ell)}\frac{2}{n}}{1 - \pi_{(\ell;\ell)}\frac{1}{n}}\right]. \qquad (18)$$

Si n est suffisamment grand

$$\frac{1 - \pi_{(\ell;\ell)}\frac{2}{n}}{1 - \pi_{(\ell;\ell)}\frac{1}{n}} \approx \left[1 - \pi_{(\ell;\ell)}\frac{2}{n}\right]\left[1 + \pi_{(\ell;\ell)}\frac{1}{n}\right];$$

$$\approx 1 + \frac{\pi_{(\ell;\ell)}}{n} - \frac{2\pi_{(\ell;\ell)}}{n} - \frac{2\pi^2_{(\ell;\ell)}}{n^2};$$

$$\approx 1 - \frac{\pi_{(\ell;\ell)}}{n}. \qquad (19)$$

Dès lors (18) devient,

$$\pi_{(N;i,j)} \approx \left[\frac{n-1}{n - \pi_{(j;j)}}\right]\pi_{(N;i)}\pi_{(j;j)}\prod_{\ell=j+1}^{N}\left[1 - \pi_{(\ell;\ell)}\frac{1}{n}\right]. \qquad (20)$$

Finalement, par le lemme 1, cette dernière expression s'écrit:

$$\pi_{(N;i,j)} \approx \pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n - \pi_{(j;j)}}.$$

Cas 2: Si $j \le n+1$, le lemme 2 donne

$$\pi_{(N;i,j)} = [\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1]\prod_{\ell=n+2}^{N}\left[1 - \pi_{(\ell;\ell)}\frac{2}{n}\right];$$

c'est-à-dire

$$\pi_{(N;i,j)} = \prod_{\ell=n+2}^{N}\left[1 - \pi_{(\ell;\ell)}\frac{1}{n}\right]\prod_{q=n+2}^{N}\left[\frac{1 - \pi_{(q;q)}\frac{2}{n}}{1 - \pi_{(q;q)}\frac{1}{n}}\right][\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1].$$

En utilisant l'approximation (19), nous obtenons

$$\pi_{(N;i,j)} \approx \left\{\prod_{\ell=n+2}^{N}\left[1 - \pi_{(\ell;\ell)}\frac{1}{n}\right]\right\}^2 \pi_{(n+1;i)}\pi_{(n+1;j)}\frac{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1}{\pi_{(n+1;i)}\pi_{(n+1;j)}}.$$

Par le lemme 1, nous obtenons finalement

$$\pi_{(N;i,j)} \approx \pi_{(N;i)}\pi_{(N;j)}\frac{\pi_{(n+1;i)} + \pi_{(n+1;j)} - 1}{\pi_{(n+1;i)}\pi_{(n+1;j)}}. \qquad (21)$$

Cqfd.

ANNEXE II

Preuve du Théorème 4

- Pour le plan de Chao, it suffit d'utiliser (6), (9) et (15).
- Pour le plan systèmatique randomisé, il suffit d'utiliser l'approximation des $\pi_{(N;i,j)}$ donnée par Deville (p. 21)

$$\pi_{(N;i,j)} \approx \pi_{(N;i)}\pi_{(N;j)}\frac{n-1}{n-\pi_{(N;i)}-\pi_{(N;j)}}. \qquad (22)$$

Cette expression est obtenue à partir de l'hypothèse

$$\text{Max}_{1\le i\le N}\left\{\frac{\pi_{(N;i)}}{n}\right\} \to 0.$$

Cette dernière hypothèse est vérifiée car $n \to \infty$.

- Pour le plan réjectif, par le résultat d'Hájek (1964, p. 1508), nous avons

$$\frac{-\Delta_{(N;i,j)}}{\pi_{(N;i,j)}} \approx \frac{[1-\pi_{(N;i)}][1-\pi_{(N;j)}]}{d-[1-\pi_{(N;i)}][1-\pi_{(N;j)}]}, \qquad (23)$$

pour $d \to \infty$. Notons que (23) reste valable pour le plan de Rao-Sampford (*cf.* Hájek 1981, Théorème 8.2, p. 82). En utilisant l'approximation (Hájek 1964, p. 1521),

$$\{d-[1-\pi_{(N;i)}][1-\pi_{(N;j)}]\}^{-1} \approx \frac{n}{d(n-1)},$$

nous obtenons le résultat du théorème.

Cqfd.

REMERCIEMENTS

L'auteur tient à remercier les arbitres qui ont émis de nombreux commentaires constructifs, permettant d'améliorer considérablement cet article.

BIBLIOGRAPHIE

BETHLEHEM, J.G., et SCHUERHOFF, H. (1984). Second-order inclusion probabilities in sequential sampling without replacement with unequal probabilities. *Biometrika*, 71, 642-644.

CHAO, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.

DEVILLE, J.-C. (Sans date). Cours de sondage, Chapitre III: les outils de bases. Polycopié de l'ENSAE, Paris.

HÁJEK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

HÁJEK, J. (1981). *Sampling from a Finite Population*. New York et Bassel: Marcel Dekker, Inc.

HARTLEY, H.O., et RAO, J.N.K. (1962). Sampling with unequal probabilities without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

HORVITZ, D.G., et THOMPSON, D.J. (1951). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47,663-685.

McLEOD, A.I., et BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sample. *Applied Statistics*, 32, 2.

RAO, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 494-513.

YATES, F., et GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, série B, 1, 253-261.