Techniques d'enquête, décembre 1985
Vol. 11, n° 2, pp. 135–148
Statistique Canada

Estimation du total pour deux caractères dans les enquêtes à bases de sondage multiples

B.C. SAXENA, P. NARAIN, et A.K. SRIVASTAVA[1]

RÉSUMÉ

Le présent document traite de l'estimation de caractères multiples dans les enquêtes à bases de sondage multiples. Le gain attribuable à l'étude de deux caractères dans une même enquête par rapport à la tenue d'enquêtes distinctes pour chaque caractère est analysé. On fait également des comparaisons de coûts entre une enquête à bases multiples portant sur deux caractères et une enquête à base simple portant sur deux caractères.

MOTS CLÉS: Enquête portant sur plusieurs caractères; estimation par stratification a posteriori; optimisation; comparaison de coûts.

1. INTRODUCTION

La technique des enquêtes à bases multiples a été présentée pour la première fois par Hartley (1962) et examinée en détail par la suite par Lund (1968), Hartley (1974), Vogel (1975), Armstrong (1979), etc. Lund a proposé une solution de rechange à l'estimateur de Hartley, qui comporte la division de l'échantillon entre divers domaines. Hartley (1974) a poursuivi avec une approche plus générale pouvant s'appliquer à divers plans de sondage. Il a constaté que la plupart des cas possibles de bases multiples utilisaient des types d'unités différents dans leurs bases respectives. Bosecker et Ford (1976) ont développé l'estimateur de Hartley pour tirer avantage de la stratification dans le domaine de chevauchement. Serrurier et Phillips (1976) et Armstrong (1978) ont expérimenté les techniques comportant des bases de sondage multiples dans des enquêtes agricoles. L'utilité d'une enquête à bases de sondage multiples a été démontrée dans des cas très divers. Dans des enquêtes par sondage, il arrive parfois que l'intérêt ne réside pas seulement dans l'estimation d'un seul caractère, mais dans l'estimation de plusieurs caractères qui doivent être étudiés simultanément. Pour en arriver à une utilisation efficace des ressources, cela est souvent réalisé par la truchement d'enquêtes intégrées. Par exemple, dans le but d'estimer la production de récoltes de légumes, on peut concevoir une enquête unique qui a pour objet d'estimer la production de plusieurs récoltes de légumes. Également, au lieu d'utiliser une base composée de tous les producteurs de légumes, on peut utiliser une base incomplète mais relativement facilement, accessible notamment une base composée des principaux producteurs de légumes. L'estimation du total pour deux caractères dans les enquêtes à bases multiples a été examinée dans le présent document. L'avantage d'étudier plusieurs caractères dans une seule enquête par rapport au recours à des enquêtes indépendantes portant chacune sur une caractère est également analysé.

2. ESTIMATEUR

Soit A et B deux bases de sondage empiétant l'une sur l'autre de taille N_A et N_B respectivement. Dans des enquêtes à bases multiples, deux échantillons de taille n_A et n_B sont tirés de façon indépendante par échantillonnage aléatoire simple à partir des bases de sondage A et B respectivement. Les bases chevauchantes génèrent des domaines *a, b* et *ab* définis comme suit:

a: domaine composé des unités appartenant à la base A uniquement,
b: domaine composé des unités appartenent à la base B uniquement,
ab: domaine composé des unités appart;enant à la fois à la base A et à la base B.

[1] B.C. Saxena, P. Narain, et A.K. Srivastava, Indian Agricultural Statistics Research Institute, Nouvelle-Delhi, Inde.

Les échantillons de taille n_A et n_B sont divisés en échantillons de taille n_a *et* n_{ab} et de taille n_b *et* n_{ba} de sorte que n_a et n_{ab} correspondent aux nombre d'unités de l'échantillon de taille n_A appartenant aux domaines a et ab respectivement. De la même façon, les échantillons de taille n_b et n_{ba} sont les échantillons provenant de la division de l'échantillon de taille n_B en échantillons composés d'unités appartenant aux domaines b et ab respectivement. Dans une étude à plusieurs caractères, ces domaines seront une fois de plus divisés, générant des sous-domaines comme suit:

Soit $y_{(1)}$ et $y_{(2)}$ deux caractères devant faire l'objet d'une étude. Chacun des domaines habituels a, ab et b sera une fois de plus divisé en $a(1)$, $a(12)$, $a(2)$, $ab(1)$, $ab(12)$, $ab(2)$ et $b(1)$, $b(12)$ et $b(2)$ respectivement. Dans ce cas-ci, $a(1)$, $a(12)$ et $a(2)$ sont les sous-domaines composés des unités du domaine a ayant le caractère $y_{(1)}$, les deux caractères $y_{(1)}$ et $y_{(2)}$ et le caractère $y_{(2)}$ respectivement. La même explication s'applique aux autres sous-domaines $ab(1)$, $ab(12)$, etc. Par conséquent, la division des deux échantillons dans l'étude de deux caractères se fera de la façon suivante:

$$n_A = n_a + n_{ab}$$

où

$$n_a = n_{a(1)} + n_{a(2)} + n_{a(12)} \quad \text{et} \quad n_{ab} = n_{ab(1)} + n_{ab(2)} + n_{ab(12)},$$

et

$$n_B = n_b + n_{ba}$$

où

$$n_b = n_{b(1)} + n_{b(2)} + n_{b(12)} \quad \text{et} \quad n_{ba} = n_{ba(1)} + n_{ba(2)} + n_{ba(12)}.$$

Dans ce cas-ci, les échantillons de taille $n_{a(1)}$, $n_{a(2)}$, etc. sont les échantillons composés des unités de l'échantillon de taille n_a appartenant respectivement aux sous-domaines $a(1)$, $a(2)$, etc. Si on se limite à un caractère, on peut alors définir

$$n_{A(1)} = n_{a(1)} + n_{a(12)} + n_{ab(1)} + n_{ab(12)},$$

$$n_{B(1)} = n_{b(1)} + n_{b(12)} + n_{ba(1)} + n_{ba(12)}.$$

De la même façon, les échantillons de taille $n_{A(2)}$ et $n_{B(2)}$ sont définis pour le deuxième caractère. L'estimation du total pour le premier caractère est donnée par

$$\hat{Y}^{(1)} = \hat{Y}_{a(1)} + \hat{Y}^{(1)}_{a(12)} + p_1\hat{Y}_{ab(1)} + q_1\hat{Y}_{ba(1)} + p_2\hat{Y}^{(1)}_{ab(12)} +$$

$$+ q_2\hat{Y}^{(1)}_{ba(12)} + \hat{Y}_{b(1)} + \hat{Y}^{(1)}_{b(12)} \qquad (1)$$

où $\hat{Y}_{a(1)}$, $\hat{Y}^{(1)}_{a(12)}$, etc . sont les totaux estimatifs pour le caractère $y_{(1)}$ des sous-domaines respectifs. Dans l'analyse qui suit, pour ce qui a trait aux domaines ayant les deux caractères, l'indice supérieur correspond au caractère examiné. Dans le cas des domaines ayant seulement un caractère, l'indice supérieur n'est pas utilisé parce que, de toute évidence, le domaine correspond au caractère.

Aussi, $p_1 + q_1 = 1$ et $p_2 + q_2 = 1$. Définissons $\bar{y}_{a(1)}$, $\bar{y}_{a(2)}$, etc. comme les moyennes des échantillons pour les sous-domaines respectifs ayant les caractères $y_{(1)}$ et $y_{(2)}$ respectivement.

Par conséquent,

$$\hat{Y}^{(1)} = N_{a(1)}\bar{y}_{a(1)} + N_{a(12)}\bar{y}^{(1)}_{a(12)} + N_{ab(1)}\,(p_1\bar{y}_{ab(1)} + q_1\bar{y}_{ba(1)})$$

$$+ N_{ab(12)}(p_2\bar{y}^{(1)}_{ab(12)} + q_2\bar{y}^{(1)}_{ba(12)})$$

$$+ N_{b(12)}\bar{y}^{(1)}_{b(12)} + N_{b(1)}\bar{y}_{b(1)}. \qquad (2)$$

De la même façon, pour le deuxième caractère, on a

$$\hat{Y}^{(2)} = N_{a(2)}\bar{y}_{a(2)} + N_{a(12)}\bar{y}^{(2)}_{a(12)} + N_{ab(2)}(p_3\bar{y}_{ab(2)} + q_3\bar{y}_{ba(2)})$$

$$+ N_{ab(12)}(p_4\bar{y}^{(2)}_{ab(12)} + q_4\bar{y}^{(2)}_{ba(12)}) + N_{b(12)}\bar{y}^{(2)}_{b(12)}$$

$$+ N_{b(2)}\bar{y}_{b(2)} \qquad (3)$$

où

$$p_3 + q_3 = 1 \text{ et } p_4 + q_4 = 1.$$

2.1 Variance de l'estimateur

La variance conditionnelle des estimations $\hat{Y}^{(1)}$ et $\hat{Y}^{(2)}$ fondées sur la stratification a posteriori pour un sous-domaine donné qui ne tient pas compte de la correction d'échantillonnage pour population finie peut s'écrire de la façon suivante

$$V(\hat{Y}^{(1)} \mid n_{a(1)}, n_{a(12)}, \text{ etc.}) = N^2_{a(1)}\frac{\sigma^2_{a(1)}}{n_{a(1)}} + N^2_{a(12)}\frac{\sigma^{(1)2}_{a(12)}}{n_{a(12)}}$$

$$+ N^{(2)}_{ab(1)}\left(p_1^2\frac{\sigma^2_{ab(1)}}{n_{ab(1)}} + q_1^2\frac{\sigma^2_{ab(1)}}{n_{ba(1)}}\right)$$

$$+ N^2_{ab(12)}\left(p_2^2\frac{\sigma^{(1)2}_{ab(12)}}{n_{ab(12)}} + q_2^2\frac{\sigma^{(1)2}_{ab(12)}}{n_{ba(12)}}\right) + N^2_{b(1)}\frac{\sigma^2_{b(1)}}{n_{b(1)}}$$

$$+ N^2_{b(12)}\frac{\sigma^{(1)2}_{b(12)}}{n_{b(12)}} \qquad (4)$$

La variance non conditionnelle de $\hat{Y}^{(1)}$ est donnée approximativement par

$$V(\hat{Y}^{(1)}) = \frac{N_A}{n_A}\left\{N_{a(1)}\sigma^2_{a(1)} + N_{a(12)}\sigma^{(1)2}_{a(12)} + p_1^2N_{ab(1)}\sigma^2_{ab(1)}\right.$$

$$\left. + p_2^2N_{ab(12)}\sigma^{(1)2}_{ab(12)}\right\} + \frac{N_B}{n_B}\left\{N_{b(1)}\sigma^2_{b(1)} + N_{b(12)}\sigma^{(1)2}_{b(12)}\right. \qquad (5)$$

$$\left. + q_1^2N_{ab(1)}\sigma^2_{ab(1)} + q_2^2N_{ab(12)}\sigma^{(1)2}_{ab(12)}\right\}$$

qui est égale à la variance obtenue dans le cas d'un échantillonnage stratifié à répartition proportionnelle.

Pareillement,

$$V(\hat{Y}^{(2)}) = \frac{N_A}{n_A}\left\{(N_{a(2)}\sigma^2_{a(2)} + N_{a(12)}\sigma^{(2)^2}_{a(12)} + p_3^2 N_{ab(2)}\sigma^2_{ab(2)} + p_4^2 N_{ab(12)}\sigma^{(2)^2}_{ab(12)}\right\} + \frac{N_B}{n_B}\left\{N_{b(2)}\sigma^2_{b(2)} + N_{b(12)}\sigma^{(2)^2}_{b(12)} + q_3^2 N_{ab(2)}\sigma^2_{ab(2)} + q_4^2 N_{ab(12)}\sigma^{(2)^2}_{ab(12)}\right\} \quad (6)$$

où $\sigma^2_{a(1)}$, $\sigma^2_{a(2)}$, etc. sont les variances pour les deux caractères dans les sous-domaines respectifs.

Pour optimiser les p_i's ($i = 1, 2, 3, 4$) dans une enquête commune, il faut minimiser une combinaison de variances individuelles sous réserve du maintien du coût fixe total de l'enquête combinée. Prenons la combinaison linéaire la plus simple possible

$$F = V(\hat{Y}^{(1)}) + V(\hat{Y}^{(2)}).$$

Dans le cas de l'enquête commune, une fonction de coût qui convienne pourrait être la suivante:

$$C' = C_1(n_{a(1)} + n_{ab(1)}) + C_2(n_{a(12)} + n_{ab(12)}) + C_3(n_{a(2)} + n_{ab(2)}) + C_4(n_{b(1)} + n_{ba(1)}) + C_5(n_{b(12)} + n_{ba(12)}) + C_6(n_{b(2)} + n_{ba(2)}) \quad (7)$$

où C_1 est le coût par unité dans le sous domaine $a(1)$, $ab(1)$; C_2 dans le sous-domaine $a(12)$, $ab(12)$; C_3 dans le sous-domaine $a(2)$, $ab(2)$ de la base A. De même C_4, C_5 et C_6 sont les coûts par unité dans les sous-domaines de la base B. Dans la fonction de coût ci-haut, la taille des échantillons pris au hasard entre en jeu. Soit le coût prévu

$$C = E(C') = n_A(C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3) + n_B(C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6) \quad (8)$$

où

$$\Phi_1 = \frac{N_{a(1)} + N_{ab(1)}}{N_A}, \quad \Phi_2 = \frac{N_{a(12)} + N_{ab(12)}}{N_A},$$

$$\Phi_3 = \frac{N_{a(2)} + N_{ab(2)}}{N_A}, \quad \Phi_4 = \frac{N_{b(1)} + N_{ba(1)}}{N_B},$$

$$\Phi_5 = \frac{N_{b(12)} + N_{ba(12)}}{N_B}, \quad \Phi_6 = \frac{N_{b(2)} + N_{ba(2)}}{N_B}.$$

Ou

$$C = n_A C_A + n_B C_B \quad (9)$$

où

$$C_A = C_1\Phi_1 + C_2\Phi_2 + C_3\Phi_3 \quad \text{et} \quad C_B = C_4\Phi_4 + C_5\Phi_5 + C_6\Phi_6.$$

Pour obtenir les p_i optimums, tout comme n_A et n_B optimums, ilfaut minimiser la fonction F sous réserve de l'existence de la fonction de coût prévu comme celle calculée dans (9). Les variables de pondération p_i et les tailles des échantillons sont obtenues comme suit à l'aide du multiplicateur de Lagrange:

$$\frac{P_1}{q_1} = \frac{P_2}{q_2} = \frac{P_3}{q_3} = \frac{P_4}{q_4} = \frac{N_B n_A}{n_B N_A} = \frac{P}{q} \text{ (disons)}, \tag{10}$$

et

$$\frac{n_A^2}{N_A} = \gamma \frac{K_5 + K_1 p_1^2 + K_2 p_2^2 + K_3 p_3^2 + K_4 p_4^2}{C_A},$$

$$\frac{n_B^2}{N_B} = \gamma \frac{K_6 + K_1 q_1^2 + K_2 q_2^2 + K_3 q_3^2 + K_4 q_4^2}{C_B}, \tag{11}$$

où γ est déterminé pour satisfaire au coût prévu et où

$$K_1 = N_{ab(1)} \sigma^2_{ab(1)},\ K_2 = N_{ab(12)} \sigma^{(1)2}_{ab(12)},$$

$$K_3 = N_{ab(2)} \sigma^2_{ab(2)},\ K_4 = N_{ab(12)} \sigma^{(2)2}_{ab(12)},$$

$$K_5 = N_{a(1)} \sigma^2_{a(1)} + N_{a(2)} \sigma^2_{a(2)} + N_{a(12)} (\sigma^{(1)2}_{a(12)} + \sigma^{(2)2}_{a(12)}),$$

$$K_6 = N_{b(1)} \sigma^2_{b(1)} + N_{b(2)} \sigma^2_{b(2)} + N_{b(12)} (\sigma^{(1)2}_{b(12)} + \sigma^{(2)2}_{b(12)}). \tag{12}$$

À partir de (10) et (11), on obtient

$$\frac{q^2 N_B C_B}{p^2 N_A C_A} = \frac{K_6 + (K_1 + K_2 + K_3 + K_4) q^2}{K_5 + (K_1 + K_2 + K_3 + K_4) p^2} \tag{13}$$

Il s'agit d'une fonction bi-quadratique en p qui peut être résolue pour p. Les fractions de sondage optimums peuvent être calculées à l'aide de (11). Un des cas qui se présente souvent dans les sondages à bases multiples est celui où les bases englobent la population totale. Prenons le cas où la base A englobe la population totale, alors $N_{b(1)} = N_{b(2)} = N_{b(12)} = 0$.

Dans ce cas, (13) se ramène à

$$p^2 = \frac{\alpha}{\varrho - \alpha} \frac{K_5}{K_1 + K_2 + K_3 + K_4} \tag{14}$$

où

$$\varrho = \frac{C_A}{C_B} \quad \text{and} \quad \alpha = \frac{N_B}{N_A}.$$

Supposons que

$$\sigma^2_{a(1)} = \sigma^{(1)^2}_{a(12)}, \ \sigma^2_{a(2)} = \sigma^{(2)^2}_{a(12)}, \ \sigma^2_{ab(1)} = \sigma^{(1)^2}_{ab(12)}, \ \sigma^2_{ab(2)} = \sigma^{(2)^2}_{ab(12)}. \tag{15}$$

Ces hypothèses semblent plausibles étant donné qu'il est peu probable que la variabilité d'un caractère soit touchée par la présence ou l'absence de l'autre caractère. Alors, p^2 se ramène à

$$p^2 = \frac{\alpha}{\varrho - \alpha}\left\{\frac{\sigma^2_{a(1)}(N_{a(1)} + N_{a(12)}) + \sigma^2_{a(2)}(N_{a(2)} + N_{a(12)})}{\sigma^2_{ab(1)}(N_{ab(1)} + N_{ab(12)}) + \sigma^2_{ab(2)}(N_{ab(2)} + N_{ab(12)})}\right\}$$

ou enfin à

$$p^2 = \frac{(1 - \alpha)\Phi_2'}{(\varrho - \alpha)}\left\{\frac{\Phi_3'(\xi_1 + \xi_2) + (1 - \xi_1)}{\Phi_4'(\xi_3 + \xi_4) + (1 - \xi_3)}\right\} \tag{16}$$

où

$$\Phi_1' = \frac{\sigma^2_{a(1)}}{\sigma^2_{ab(1)}}, \ \Phi_2' = \frac{\sigma^2_{a(2)}}{\sigma^2_{ab(2)}}, \ \Phi_3' = \frac{\sigma^2_{a(1)}}{\sigma^2_{a(2)}}, \ \Phi_4' = \frac{\sigma^2_{ab(1)}}{\sigma^2_{ab(2)}}$$

et

$$\xi_1 = \frac{N_{a(1)}}{N_a}, \ \xi_2 = \frac{N_{a(12)}}{N_a}, \ \xi_3 = \frac{N_{ab(1)}}{N_{ab}}, \ \xi_4 = \frac{N_{ab(12)}}{N_{ab}}.$$

En se servant du fait que $N_{ab} = N_B$, $N_{a(2)} + N_{a(12)} = N_a - N_{a(1)}$ et $N_{ab(2)} + N_{ab(12)} = N_{ab} - N_{ab(1)}$, on peut voir que l'expression ci-dessus se ramène à la forme habituelle du cas à un seul caractère étant donné que $\xi_1 = \xi_3 = 1$ et $\xi_2 = \xi_4 = 0$. Il convient de noter que les variances dans des domaines ne sont en général pas connues étant donné que de telles valeurs sont fondées sur des renseignements antérieurs ou sur des valeurs estimées. L'optimalité de p^2 est affecté dans cette mesure.

3. COMPARISON D'UNE ENQUÊTE PORTANT SUR PLUSIEURS CARACTÈRES ET DES ENQUÊTES INDÉPENDANTES PORTANT SUR UN SEUL CARACTÈRE DANS DES SITUATIONS À BASES MULTIPLES

Les enquêtes portant sur plusieurs caractères sont conçues dans le but d'économiser les ressources disponibles, et on peut s'attendre qu'une enquête commune présente des avantages supérieurs à ceux des enquêtes indépendantes portant sur un seul caractère, sur le plan des coûts et de l'efficacité. Nous allons donc examiner l'importance des avantages (gain) attribuables à l'utilisation d'une enquête commune à bases de sondage multiples.

Soit, dans une étude portant sur un seul caractère, disons $y_{(1)}$, des échantillons aléatoires simples de taille n_A et n_B tirés des bases de sondage A et B respectivement. Nous supposons ici que les seules bases disponibles sont les bases utilisées précédemment, et non pas les bases réduites pour chaque caractère. Définissions N_{A1}, N_{B1}, n^*_{A1}, et n^*_{B1} comme les tailles de la population et les tailles des échantillons portant sur le caractère $y_{(1)}$. Ici, n^*_{A1} and n^*_{b1}

sont les tailles des échantillons aléatoires et $E(n_{A1}^*) = n_A N_{A1}/N_A$ et $E(n_{B1}^*) = n_B N_{B1}/N_B$. Dans ce cas, l'estimateur $\hat{Y}^{(1)*}$ et sa variance sont donnés par:

$$\hat{Y}^{(1)*} = (N_{a(1)} + N_{a(12)})\bar{y}_{(a(1),\, a(12))}$$

$$+ (N_{ab(1)} + N_{ab(12)})(p'\bar{y}_{(ab(1),\, ab(12))} + q'\bar{y}_{(ba(1),\, ba(12))})$$

$$+ (N_{b(1)} + N_{b(12)})\bar{y}_{(b(1),\, b(12))}$$

où p' et q' sont les variables de pondération de sorte que $p' + q' = 1$ et où $\bar{y}_{(a(1),\, a(12))}, \bar{y}_{(ab(1),\, ab(12))}$, etc. sont les moyennes des échantillons tirés des domaines combinés respectifs, c'est-à-dire, par exemple, que $\bar{y}_{(a(1),\, a(12))}$ est la moyenne des unités de l'échantillon tiré des domaines $a(1)$ and $a(12)$.

$$V(\hat{Y}^{(1)*}) = \frac{N_A}{n_A}(N_{a(1)}\sigma^2_{a(1)} + N_{a(12)}\sigma^{(1)2}_{a(12)}$$

$$+ (p'^2\frac{N_A}{n_A} + q'^2\frac{N_B}{n_B})(N_{ab(1)}\sigma^2_{ab(1)} + N_{ab(12)}\sigma^{(1)2}_{ab(12)})$$

$$+ \frac{N_B}{n_B}(N_{b(1)}\sigma^2_{b(1)} + N_{b(12)}\sigma^{(1)2}_{b(12)}). \tag{17}$$

Dans ce cas, la fonction de coût prend la forme

$$C = C_1 n_{A1}^* + C_4 n_{B1}^*$$

et le coût prévu est représenté par C^*

$$C^* = C_1\frac{n_A}{N_A}N_{A1} + C_4\frac{n_B}{N_B}N_{B1} = C'_A n_A + C'_B n_B \tag{18}$$

où $C'_A = C_1 N_{A1}/N_A$ et $C'_B = C_4 N_{B1}/N_B$.

Pour simplifier, nous supposons un champ d'observation de 100% par la base A et l'égalité des variances comme dans (15), et nous supposons que $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$. Suivant ces hypothèses, le coût C^* avec n_A et n_B qui minimisent la variance (17) se calcule comme suit (voir annexe pour le détail du calcul).

$$C^* = \frac{(\xi_1 + \xi_2)[\{C_1(1 + \alpha_1^*)(\Phi'_1 + \alpha_1^* p'^2)\}^{1/2} + \alpha_1^*(C_4 q'^2)^{1/2}]^2}{\frac{1}{1 - \alpha}\left\{\frac{(\Phi'_1 + \alpha_1^* p^2)}{n_A} + \frac{\alpha\alpha_1^* q^2}{n_B}\right\}}$$

où

$$\alpha_1^* = \frac{\alpha}{1-\alpha}\frac{\xi_3 + \xi_4}{\xi_1 + \xi_2}.$$

De la même façon, pour l'enquête distincte portant sur le deuxième caractère, le coût se calcule somme suit:

$$C^{**} = \frac{(1-\xi_1)\left[\left\{C_3(1+\alpha_2^*)(\Phi_2' + \alpha_2^* p''^2)\right\}^{1/2} + \alpha_2^*(C_6 q''^2)^{1/2}\right]^2}{\frac{1}{1-\alpha}\left\{\frac{(\Phi_2' + \alpha_2^* p^2)}{n_A} + \frac{\alpha\alpha_2^* q^2}{n_B}\right\}} \tag{19}$$

où

$$p''^2 = \frac{K\Phi_2'}{1 + \alpha_2^*(1-K)}, \quad \alpha_2^* = \frac{\alpha}{1-\alpha}\frac{1-\xi_3}{1-\xi_1}.$$

Pour l'étude combinée portant sur les deux caractères, le coût total C dans le cas d'un champ d'observation de 100% de la base A est donné par (8).

Par conséquent

$$C = \frac{n_A}{N_A}\left[C_1(N_{a(1)} + N_{ab(1)}) + C_2(N_{a(12)} + N_{ab(12)}) + C_3(N_{a(2)} + N_{ab(2)})\right]$$

$$+ \frac{n_B}{N_B}\left[C_4 N_{ab(1)} + C_5 N_{ab(12)} + C_6 N_{ab(2)}\right].$$

En utilisant les hypothèses formulées au sujet des coûts (c'est-à-dire, en supposant que $C_4/C_1 = C_5/C_2 = C_6/C_3 = K$) on obtient

$$C = C_2 n_A\left[(1-\alpha)\left\{\varrho_1\xi_1 + \xi_2 + \varrho_3(1-\xi_1-\xi_2)\right\} +\right.$$

$$\left.\alpha\left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\} + \frac{K}{r}\left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\}\right] \tag{20}$$

où $r = n_A/n_B$, $\varrho_1 = C_1/C_2$ *and* $\varrho_3 = C_3/C_2$.

Mais dansl'étude combinée portant sur les deux caractères (n_A/n_B) optimal $= p/\alpha q$ où p est donné par (16). Donc, la valeur du gain peut être obtenue à partir du ratio.

$$\frac{C^* + C^{**}}{C} =$$

$$\frac{\frac{(\xi_1+\xi_2)\varrho_1 T_1^2}{(\Phi_1' + \alpha_1^* p)} + \frac{(1-\xi_1)\varrho_3 T_2^2}{(\Phi_3' + \alpha_2^* p)}}{\left\{\varrho_1\xi_1 + \xi_2 + \varrho_3(1-\xi_1-\xi_2)\right\} + \left\{\varrho_1\xi_3 + \xi_4 + \varrho_3(1-\xi_3-\xi_4)\right\}\left\{\frac{r\alpha + K}{r(1-\alpha)}\right\}} \tag{21}$$

où

$$T_1 = \left\{(\Phi_1' + \alpha_1^* p'^2)(1 + \alpha_1^*)\right\}^{1/2} + \alpha_1^* q' \sqrt{K}$$

$$T_2 = \left\{(\Phi_2' + \alpha_2^* p''^2)(1 + \alpha_2^*)\right\}^{1/2} + \alpha_2^* q'' \sqrt{K}.$$

K peut être calculé de la façon suivante: à l'aide de la définition de C_A, C_B, Φ_i et de ($i = 1, \ldots, 6$) et de l'équation (A.1), on obtient

$$\frac{C_A}{C_B} = \frac{1}{K} \frac{\varrho_1 \Phi_1 + \Phi_2 + \varrho_3 \Phi_3}{\varrho_1 \Phi_4 + \Phi_5 + \varrho_3 \Phi_6} = \varrho_3,$$

et donc

$$K = \varrho^{-1}\left\{\alpha + (1 - \alpha)\frac{\varrho_1 \xi_1 + \xi_2 + \varrho_3(1 - \xi_1 - \xi_2)}{\varrho_1 \xi_3 + \xi_4 + \varrho_3(1 - \xi_3 - \xi_4)}\right\}. \tag{22}$$

On peut utiliser l'expression de l'équation (21) pour calculer le gain au niveau des coûts que produit une étude portant sur les deux caractères simultanément par rapport à des enquêtes individuelles indépendantes. Le gain G exprimé en pourcentage est donc donné par

$$G = \left(\frac{C^* + C^{**}}{C} - 1\right) \times 100$$

Dans la comparaison des coûts ci-dessus, les coûts prévus, C, C^* et C^{**} n'incluent pas les frais généraux afférents à la réalisation d'une enquête combinée ou à la réalisation d'enquêtes distinctes. On s'attend toutefois que le total des frais généraux d'enquêtes dinstinctes soit beaucoup plus élevé que le total correspondant se rapportant à la réalisation d'une enquête combinée. Par conséquent, le gain réel attribuable à la réalisation d'enquêtes indépendantes sera supérieur au gain G exprimé en pourcentage, tel qu'il est défini plus haut.

L'expression (21) se simplifie de façon appréciable en fonction des hypothèses $\Phi_1' = \Phi_2' = \Phi$ et $\xi_1 = \xi_2 = \xi_3 = \xi_4 = \xi$ (disons).

À partir de (22) $\varrho = 1/K$ et de (16) puisque $\Phi_1'/\Phi_2' = \Phi_3'/\Phi_4'$, le p^2 se simplifie comme suit:

$$p^2 = \frac{K(1 - \alpha)}{1 - k\alpha}\Phi.$$

En outre $\alpha_1^* = \alpha_2^* = \alpha/(1 - \alpha)$.

Par conséquent, à partir de (A.1)

$$p' = p'' = \left\{\frac{K(1 - \alpha)\Phi}{1 - K\alpha}\right\}^{1/2}.$$

D'où

$$T_1 = T_2 = \left\{\Phi \frac{1 - K\alpha}{1 - \alpha}\right\}^{1/2} + \frac{\alpha\sqrt{K}}{1 - \alpha}.$$

Avec toutes ces substitutions apportées à l'expression (21) $(C^* + C^{**})/C$ se simplifie comme suit:

$$\frac{C^* + C^{**}}{C} = \frac{T_2^2}{\Phi + \alpha_1^* p} \times \frac{2\xi\varrho_1 + (1 - \xi)\varrho_3}{\left\{\frac{r\alpha + K}{K(1 - \alpha)} + 1\right\}\{\varrho_1\xi + \xi + \varrho_3(1 - 2\xi)\}}$$

$$= \frac{T_1^2}{(\Phi + \alpha_1^* p)\left\{\frac{r\alpha + K}{K(1 - \alpha)} + 1\right\}} \times \frac{\varrho_3 + \xi(2\varrho_1 - \varrho_3)}{\varrho_3 + \xi(\varrho_1 + 1 - 2\varrho_3)}$$

$$= \frac{\varrho_3 + \xi(2\varrho_1 - \varrho_3)}{\varrho_3 + \xi(1 + \varrho_1 - 2\varrho_3)}$$

où $r = (n_A/n_B)$ optimal $= p/\alpha q$ de (10).

D'où,

$$G = \frac{\xi(\varrho_1 + \varrho_3 - 1)}{\varrho_3 + \xi(1 + \varrho_1 - 2\varrho_3)} \times 100.$$

L'équalité des ξ_i ne semble pas être une hypothèse réaliste. La valeur de G a donc été calculée à l'aide de (19) pour des combinaisons réalistes et représentatives de paramètres; les résultats obtenus figurent dans le tableau 1.

Ce tableau révèle qu'il y a manifestement un gain attribuable à l'intégration d'enquêtes à bases multiples portant sur deux caractères par rapport à la réalisation d'enquêtes distinctes. Le gain augment à mesure qu'augmentent les valeurs de ϱ_1 et ϱ_3.

4. COMPARISON D'ENQUÊTES À BASE MULTIPLES PORTANT SUR DEUX CARACTÈRES ET D'UNE ENQUÊTE À BASE UNIQUE

La comparaison d'une enquête à deux bases de sondage et des enquêtes à base unique pour l'étude de deux caractères présente de l'intérêt sur le plan pratique. Une étude analogue a été faite par Hartley (1962) pour le cas d'un seul caractère. La réduction relative des coûts calculée suivant une méthode analogue s'exprime

$$R = \left(1 + \frac{\alpha q}{p\varrho}\right)^2 \Big/ \left\{1 + \frac{\alpha q(1 + p)}{p^2\varrho}\right\}$$

où p^2 is given by (16), $\varrho = C_A/C_B$ et $\alpha = N_{ab}/N_A$

La réduction des coûts attribuable à l'utilisation d'une enquête à bases multiples plutôt qu'à l'utilisation d'une enquête à base unique a été calculée pour un ensemble donné de valeurs des paramètres et les résultats figurent dans le tableau 2. D'après ce tableau, la réduction des coûts est considérable.

Tableau 1

Gain en pourcentage au niveau des coûts attribuable à l'utilisation d'une enquête commune à bases multiples portant sur deux caractères plutôt qu'à l'utilisation d'enquêtes distinctes

Lorsque $\varrho = 10$, $\Phi_1' = 0.25$, $\Phi_2' = 0.5$, $\Phi_3' = 1$, $\alpha = 0.5$.

	ϱ_3						
ϱ_1	0.3	0.4	0.5	0.6	0.7	0.8	0.9
			$\xi_1 = 0.2, \xi_2 = 0.2, \xi_3 = 0.4, \xi_4 = 0.2$				
0.3						1.5	3.9
0.4					1.7	4.2	6.4
0.5				1.8	4.4	6.7	8.7
0.6			1.8	4.5	6.9	8.9	10.7
0.7		1.7	4.6	7.0	9.1	10.9	12.6
0.8	1.7	4.6	7.1	9.3	11.2	12.8	14.3
0.9	4.5	7.1	9.4	11.3	13.0	14.5	15.9
			$\xi_1 = 0.2, \xi_2 = 0.4, \xi_3 = 0.2, \xi_4 = 0.4$				
0.3						4.5	9.1
0.4					4.6	9.3	13.6
0.5				4.8	9.6	14.0	17.9
0.6			4.9	9.9	14.3	18.3	22.0
0.7		5.1	10.1	14.7	18.8	22.5	25.9
0.8	5.2	10.4	15.1	19.3	23.1	26.5	29.7
0.9	10.8	15.5	19.8	23.6	27.1	30.3	33.2

Tableau 2

Réduction des coûts pour des variances constantes

Lorsque $\Phi_1' = 0.25$, $\Phi_2' = 0.5$, $\Phi_3' = 1$, et $\xi_1 = 0.2$, $\xi_2 = 0.3$, $\xi_4 = 0.4$.

	α					
ϱ	0.5	0.6	0.7	0.8	0.9	0.95
100	.227	.175	.132	.094	.059	.040
20	.304	.254	.200	.169	.127	.101
10	.367	.321	.279	.238	.193	.164
5	.462	.423	.387	.351	.308	.277
2	.661	.646	.634	.621	.599	.578
1	.876	.895	.918	.943	.971	.985

ANNEXE

La solution optimale pour p' qui minimise la variance (17) sous la contrainte de C^* selon les hypothèses d'un champ d'observation de 100% par la base A et l'égalité des variances est donnée par

$$p'^2 = \frac{1 - \alpha}{\varrho' - \alpha} \left\{ \frac{\sigma^2_{a(1)}(\xi_1 + \xi_2)}{\sigma^2_{ab(1)}(\xi_3 + \xi_4)} \right\}$$

avec

$$\varrho' = \frac{C'_A}{C'_B}.$$

En se servant de ce que $N_{A1} = N_{a(1)} + N_{a(12)} + N_{ab(1)} + N_{ab(12)}$ et de ce que $N_{B1} = N_{ab(1)} + N_{ab(12)}$, on peut écrire ϱ' de la façon suivante

$$\varrho' = \frac{C_1}{C_4} \alpha \frac{N_a(\xi_1 + \xi_2) + N_{ab}(\xi_3 + \xi_4)}{N_{ab}(\xi_3 + \xi_4)}$$

$$= \frac{C_1}{C_4} \alpha \left(\frac{1 - \alpha}{\alpha} \frac{\xi_1 + \xi_2}{\xi_3 + \xi_4} + 1 \right)$$

$$= \frac{\alpha}{K} \left(\frac{1}{\alpha_1^*} + 1 \right)$$

où

$$\alpha_1^* = \frac{\alpha}{1 - \alpha} \frac{\xi_3 + \xi_4}{\xi_1 + \xi_2}.$$

D'où

$$p'^2 = \frac{1 - \alpha}{\frac{\alpha}{K}\left(\frac{1}{\alpha_1^*} + 1 \right) - \alpha} \frac{\xi_1 + \xi_2}{\xi_3 + \xi_4} \Phi'_1$$

$$= \frac{K \Phi'_1}{1 + \alpha_1^*(1 - K)} \qquad \text{(A.1)}$$

Définissons

$$\lambda_1 = (N_{a(1)} + N_{a(12)})\sigma^2_{a(1)} + p^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_2 = q^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_3 = (N_{a(1)} + N_{a(12)})\sigma^2_{a(1)} + p'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)},$$

$$\lambda_4 = q'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)}.$$

Étant donné le p' de l'équation (A.1), les tailles optimales des échantillons seront

$$\frac{n^2_{AO}}{N_A} = \gamma' \frac{(N_{a(1)} + N_{a(12)})\sigma^2_{ab(1)} + p'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(12)}}{C'_A}$$

$$= \gamma' \frac{\lambda_3}{C'_A}$$

$$\frac{n^2_{BO}}{N_B} = \gamma' \frac{q'^2(N_{ab(1)} + N_{ab(12)})\sigma^2_{ab(1)}}{C'_B} = \gamma' \frac{\lambda'_4}{C'_B}$$

avec γ' déterminé en fonction de (18). À partir de cela, on obtient

$$\frac{n_{BO}}{n_{AO}} = \frac{N_B}{N_A}\left(\frac{C_1 N_{A1}\lambda_4}{C_4 N_{B1}\lambda_3}\right)^{1/2}. \tag{A.2}$$

De plus,les variances calculées à l'aide de (5) et (17) pour des tailles d'échantillons optimales peuvent s'écrire

$$V(\hat{Y}^{(1)}) = \frac{N_A}{n_A}\lambda_1 + \frac{N_B}{n_B}\lambda_2 \tag{A.3}$$

$$V(\hat{Y}^{*(1)}) = \frac{N_A}{n_{AO}}\lambda_3 + \frac{N_B}{n_{BO}}\lambda_4.$$

En égalant les variances ci-dessus et en utilisant l'équation (A.2), on obtient l'expression pour n_{AO} et n_{BO} en termes de n_A et n_B de la façon suivante:

$$\frac{n_{AO}}{N_A} = \frac{\lambda_3 + \left(\dfrac{C_4\lambda_3\lambda_4 N_{B1}}{C_1 N_{A1}}\right)^{1/2}}{\dfrac{N_A}{n_A}\lambda_1 + \dfrac{N_B}{n_B}\lambda_2}$$

et

$$\frac{n_{BO}}{N_B} = \frac{\lambda_4 + \left(\dfrac{C_1\lambda_3\lambda_4 N_{A1}}{C_4 N_{B1}}\right)^{1/2}}{\dfrac{N_A}{n_A}\lambda_1 + \dfrac{N_B}{n_B}\lambda_2}.$$

En utilisant ces relations, on peut calculer le coût C^* de la façon suivante:

$$C^* = \frac{(\xi_1 + \xi_2)\left[\left\{C_1(1 + \alpha_1^*)(\Phi_1' + \alpha_1^* p'^2)\right\}^{1/2} + \alpha_1^*\left(C_4 q'^2\right)^{1/2}\right]^2}{\dfrac{1}{1-\alpha}\left\{\dfrac{(\Phi_1' + \alpha_1^* p^2)}{n_A} + \dfrac{\alpha\alpha_1^* q^2}{n_B}\right\}}. \qquad \text{(A.4)}$$

BIBLIOGRAPHIE

ARMSTRONG, B. (1979). Test for multiple frames sampling technique for agricultural survey: Nouveau-Brunswick, 1978. *Techniques d'enquête*, 5, 178-199.

BOSECKER, R.R., et FORD, B.L. (1976). Multiple frame estimation with stratified overlap domain. *American Statistical Association Proceedings of the Social Statistics Section*, 219-224.

HARTLEY, H.O. (1962). Multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 203-206.

HARTLEY, H.O. (1974). Multiple frame methodology and selected application. *Sankhya*, Series C, 36, 99-118.

LUND, R.E. (1968). Estimation in multiple frame surveys. *American Statistical Association Proceedings of the Social Statistics Section*, 282-288.

SERRURIER, D., and PHILLIPS, J. (1976). Double frame Ontario pilot hog surveys. *Survey Methodology*, 2, 138-170.

VOGEL, F.A. (1975). Surveys with overlapping frames, problems in application. *American Statistical Association Proceedings of the Social Statistics Section*, 695-699.