Techniques d'enquête, juin 1986
Vol. 12, n° 1, pp. 31–39
Statistique Canada

Résultats optimaux en situation de non-réponse

V.P. GODAMBE et M.E. THOMPSON[1]

RÉSUMÉ

L'application de fonctions d'estimation optimales aux enquêtes par sondage (Godambe et Thompson 1986) produit des résultats optimaux en situation de non-réponse.

MOTS CLÉS: Fonction d'estimation optimale; non-réponse.

1. INTRODUCTION ET GÉNÉRALITÉS

Une enquête par sondage est habituellement caractérisée par une population **P** de N individus i; $\mathbf{P} = \{i: i = 1, \ldots, N\}$. À chaque individu i correspond une valeur réelle y_i. Le vecteur $\mathbf{y} = (y_1, \ldots, y_i, \ldots, y_N)$ est appelé vecteur de population. Tout sous-ensemble s de **P** est appelé échantillon. Posons $S = \{s\}$. On appelle plan de sondage une distribution de probabilité p sur S. On prélève un échantillon s au moyen d'un plan de sondage p et on détermine les valeurs y_i: $i \epsilon s$ par une enquête. En l'occurrence, les données sont représentées par χ_s où

$$\chi_s = \{s, (i, y_i): i \in s\}. \tag{1.1}$$

À l'aide des données définies par (1.1), on cherche à estimer un paramètre de population d'enquête θ_N, qui est une fonction réelle particulière du veteur de population $\mathbf{y}$; $\theta_N = \theta_N(\mathbf{y})$.

À cet égard, nous supposons un modèle de superpopulation suivant lequel $y_1, \ldots, y_N$ sont indépendantes et, pour certaines valeurs *connues* x_i de covariables $i = 1, \ldots, N$,

$$\epsilon(y_i - \theta x_i) = 0, \; i = 1, \ldots, N, \tag{1.2}$$

ϵ étant l'espérance mathématique relative au modèle. Dans le modèle (1.2), θ est le paramètre de régression *inconnu* habituel, l'espérance mathématique étant supposée maintenir les valeurs de x_i fixes. L'ordonnée à l'origine du modèle de régression n'est pas précisée dans (1.2) car il est souvent possible de l'éliminer par une méthode de stratification appropriée (Godambe, 1982). Il est à noter que le modèle (1.2) *ne* définit *pas* de fonction de variance.

Selon Godambe et Thompson (1986), pour des nombres *particuliers* α_i, $i = 1, \ldots, N$, nous définissons le paramètre de population d'enquête θ_N comme la solution de l'équation

$$\tilde{g} = \sum_{i=1}^{N} (y_i - \theta x_i)\alpha_i = 0. \tag{1.3}$$

[1] V.P. Godambe et M.E. Thompson, Département de la statistique et des sciences actuarielles, Université de Waterloo, Waterloo (Ontario), Canada, N2L 3G1

C'est-à-dire,

$$\theta_N = \sum_{i=1}^{N} y_i \alpha_i / \sum_{i=1}^{N} x_i \alpha_i. \tag{1.4}$$

Le paramètre θ_N se rattache au modèle (1.2) par l'équation

$$\epsilon \tilde{g} = 0. \tag{1.5}$$

Toute fonction réelle h des données χ_s dans (1.1) et du paramètre θ est désignée *une fonction d'estimation sans biais* des paramètres θ_N et θ si

$$E(h - \tilde{g}) = 0 \text{ pour toutes les valeurs de } \mathbf{y} \text{ et } \theta, \tag{1.6}$$

E étant l'espérance mathématique s'appliquant au plan de sondage p utilisé pour former l'échantillon s. Étant donné les équations (1.5) et (1.6), nous disons que la solution de l'équation

$$h(\chi_s, \theta) = 0,$$

pour l'ensemble de données χ_s est une estimation des paramètres θ et θ_N définis respectivement par (1.2) et (1.4). Pour la fonction $\tilde{g}$ dans (1.4), suivant le plan de sondage p, posons $H(p)$ comme la classe de toutes les fonctions d'estimation sans biais h. C'est-à-dire

$$H(p) = \{h: E(h - \tilde{g}) = 0 \text{ pour toutes les valeurs de } \mathbf{y} \text{ et } \theta\}. \tag{1.7}$$

Alors, nous disons qu'une *fonction d'estimation* $h^* \in H(p)$ est optimale si

$$\epsilon E(h^*)^2 \leq \epsilon E(h)^2 \text{ pour tout } h \in H(p) \tag{1.8}$$

(Godambe et Thompson 1986). De plus, lorsque l'inéquation (1.8) est satisfaite,

$$h^* = 0 \tag{1.9}$$

est considérée comme *l'équation d'estimation optimale* pour l'estimation du paramètre θ_N défini par (1.3) et (1.4).

Pour le plan de sondage p utilisé pour le prélèvement de l'échantillon s, posons π_i, $i = 1, \ldots, N$ comme les probabilités d'inclusion. En d'autres termes,

$$\pi_i = \sum_{s \ni i} p(s), \; i = 1, \ldots, N, \tag{1.10}$$

où $s \ni i$ désigne tous les échantillons s qui renferment l'individu i. Nous posons par hypothèse

$$\pi_i > 0, \; i = 1, \ldots, N. \tag{1.11}$$

Théorème 1.1. (Godambe et Thompson 1986). Pour tout plan de sondage p qui satisfait à l'inéquation (1.11), aux termes du modèle (1.2), dans la classe $H(p)$ de toutes les fonctions d'estimation sans biais (équation (1.7)), la fonction optimale h^*, c'est-à-dire h^* satisfaisant à l'inéquation (1.8), est définie par

$$h^* = \sum_{i \in s} (y_i - \theta x_i)\, \alpha_i / \pi_i, \tag{1.12}$$

π_i étant la probabilité d'inclusion définie en (1.10). L'équation d'estimation optimale s'exprime donc:

$$\sum_{i \in s} (y_i - \theta x_i)\, \alpha_i / \pi_i = 0. \tag{1.13}$$

L'estimateur $\hat{\theta}_s$ du paramètre de population d'enquête θ_N défini en (1.4) et du paramètre de superpopulation θ défini en (1.2) est défini comme

$$\hat{\theta}_s = \frac{\sum_{i \in s} y_i \alpha_i / \pi_i}{\sum_{i \in s} x_i \alpha_i / \pi_i}. \tag{1.14}$$

Brewer (1963) et Hajek (1971) ont déjà proposé cet estimateur selon des critères de 'vraisemblance'.

Pour faire le lien entre le théorème 1.1 et les résultats optimaux obtenus antérieurement (voir, par exemple, Godambe, 1982) nous posons $\alpha_i \equiv 1$ dans l'équation (1.3) et, partant, dans l'équation (1.2). De plus, nous considérons un modèle de superpopulation que l'on tire de (1.2) en posant $\theta = \theta_0$, celle-ci étant une valeur définie. Alors, pour tout plan de sondage avec des probabilités d'inclusion π_i satisfaisant à l'inéquation (1.11), dans la classe de tous les estimateurs sans biais de θ_N du plan de sondage (dans (1.4)), $\alpha_i = 1$, $i = 1, \ldots, N$), l'espérance mathématique de la variance du plan de sondage dans ce modèle de superpopulation est minimisée pour l'estimation

$$e = \frac{1}{X}\left\{ \sum_{i \in s} \frac{y_i - \theta_0 x_i}{\pi_i} + \theta_0 \sum_{i=1}^{N} x_i \right\} \tag{1.15}$$

où $X = \Sigma_1^N x_i$. Le caractère optimal de l'estimation e à $\theta = \theta_0$ vaut pour toutes les valeurs de θ si et seulement si le plan de sondage est tel que

$$\text{Probability}\left\{ s: \left(\sum_{i \in s} \frac{x_i}{\pi_i} - \sum_{i=1}^{N} x_i \right) = 0 \right\} = 1. \tag{1.16}$$

Or, lorsque le plan de sondage satisfait à la condition (1.16), $\hat{\theta}_s$ dans (1.14) est égal à e dans (1.15). Ainsi, le théorème 1.1 explique tous les résultats optimaux obtenus antérieurement, mais il ne s'arrête pas là. Dans beaucoup de cas, notamment lorsqu'il s'agit de plans de sondage où $\pi_i \propto x_i$, la condition (1.16) suppose un plan de sondage où *la taille de l'échantillon est fixe*. Par ailleurs, selon le théroème 1.1, on peut obtenir des résultats

optimaux, peu importe que la taille de l'échantillon soit fixe ou variable. Autrement dit, on peut obtenir des résultats optimaux dans des plans de sondage où *la taille de l'échantillon est aléatoire*; ce genre de plans de sondage se présente souvent dans les cas de non-réponse analysés plus loin.

2. NON-RÉPONSE ET RÉSULTATS OPTIMAUX

Supposons qu'un échantillon s est prélevé dans la population d'enquête P au moyen d'un plan de sondage p. Supposons aussi qu'à cause de la non-réponse, nous connaissons uniquement les valeurs de la variable y_i relatives aux sous-ensembles $s' \subset s$; $s - s'$ étant les non-répondants. Ainsi, au lieu d'être représentés par χ_s (équation 1.1), les données correspondantes sont désormais représentées par l'expression

$$\chi_{s,s'} = (s,s'\{(i,y_i):i \in s'\}). \tag{2.1}$$

Nous pouvons alors considérer deux problèmes d'estimation:

(I) En l'absence de non-réponse, c'est-à-dire si nous conaissions toutes les données χ_s définies en (1.1), nous estimerions le paramètre de population d'enquête θ_N défini en (1.4) en résolvant l'équation d'estimation optimale définie en (1.12), c'est-à-dire $h^* = 0$. Lorsque les données hypothétiques sont remplacées par $\chi_{s,s'}$ défini en (2.1), on peut chercher à estimer h^* au moyen d'une fonction $h'(\chi_{s,s'})$. Cela est en accord avec une proposition de Rubin (1976). Conformément à l'équation (1.7), nous définissons la classe des fonctions d'estimation sans biais h' (pour h^*, étant donné l'échantillon s) comme

$$H'(p,.,s) = \{h': E(h' - h^*|s) = 0, \text{ pour toutes les valeurs de } \mathbf{y} \text{ et } \theta\}; \tag{2.2}$$

le point (.) dans H' indique que la classe H' ne peut être définie qu'une fois que le *mécanisme de réponse* a été précisé. De nouveau, nous définisson h'^* comme la *fonction d'estimation optimale* dans H' (équation 2.2) si $h'^* \epsilon H'$ et si, aux termes du modèle (1.2), $\epsilon E(h'^*)^2 \leq \epsilon E(h'^*)^2$ pour tout $h' \in H'$.

(II) Par ailleurs, nous pourrions chercher à estimer *directement* (c'est-à-dire, sans estimer h^* comme en (I)) le paramètre de population d'enquête θ_N à l'aide des données $\chi_{s,s'}$. Conformément à l'équation (1.7), nous définissons la classe des fonctions d'estimation sans biais $h''(\chi_{s,s'})$ comme

$$H''(p,.) = \{h'': E(h'' - \bar{g}) = 0, \text{ pour toutes les valeurs de } \mathbf{y} \text{ et } \theta\}; \tag{2.3}$$

comme dans le cas précédent, le point (.) dans H'' indique que la classe H'' ne peut être définie qu'une fois que le *mécanisme de réponse* a lui-même été défini. À nouveau, nous définissons h'' comme la *fonction d'estimation optimale* dans H'' si $h''^* \in H''$ et si, selon le modèle (1.2), $\epsilon E(h''^*)^2 \leq \epsilon E(h'')^2$ pour toutes les fonctions d'estimation $h'' \in H''$.

Dans les expressions $H'(p,.,s)$ et $H''(p,.)$, le mécanisme de réponse (.) n'est pas défini. Nous nous attachons maintenant à le définir.

MÉCANISME DE RÉPONSE: Si l'individu i de la population d'enquête **P** est inclus dans l'échantillon s

$$\begin{aligned} &\text{la probabilité que } i \text{ fournisse une réponse est } q_i \\ &\text{et la probabilité qu'il ne fournisse pas de réponse est } 1-q_i, \end{aligned} \tag{2.4}$$

$i = 1, \ldots, N$; nous posons que $q_i > 0$, $i = 1, \ldots, N$.

À l'aide du mécanisme de réponse $\mathbf{q} = (q_1, \ldots, q_N)$ défini en (2.4), nous pouvons définir entièrement la classe $H'(p,.,s)$ dans (2.2) comme $H'(p, \mathbf{q}, s)$ et la classe $H''(p,.)$ dans (2.3) comme $H''(p, \mathbf{q})$.

Le théorème 2.1 ci-dessous est l'application du cas (I) tandis que les théroèmes 2.2, 2.3 et 2.4 sont l'application du cas (II).

Théroème 2.1. Pour tout plan de sondage p satisfaisant à l'inéquation (1.11) et pour tout échantillon s dans la classe $H'(p, \mathbf{q}, s)$ des fonctions d'estimation (équation 2.2), selon le modèle de superpopulation (1.2), $\epsilon E\{(h')^2 | s\}$ est minimisée pour $h' = h'^*$ où

$$h'^* = \sum_{i \epsilon s'} (y_i - \theta x_i)\alpha_i/\pi_i q_i; \tag{2.5}$$

en d'autres termes, h'^* est la fonction d'estimation optimale dans $H'(p, \mathbf{q}, s)$.

Preuve. Comme il a été souligné dans la section 1, le caractère optimal de h^* dans (1.12) tient pour les plans de sondage où la *taille de l'échantillon est aléatoire* et pour n'importe quelle valeur de α_i, $i = 1, \ldots, N$ dans (1.3). On peut donc faire la preuve du théroème 2.1 en remplaçant dans la théorème 1.1 la population **P** par s *et* α_i par α_i/π_i, $i \epsilon s$ et en précisant que les probabilités d'inclusion sont désormais désignées q_i, $i \in s$.

Théorème 2.2. Soit $\bar{H}''$ la sous-classe de H'' dans (2.3) de telle sorte que toute fonction d'estimation $h''(\chi_{s,s'})$ dans $\bar{H}''$ ne dépend que de s dans (s,s'). Alors, pour tout plan de sondage p satisfaisant à l'inéquation (1.11), dans la classe $\bar{H}''(p, \mathbf{q})$, selon le modèle de superpopulation (1.2), $\epsilon E\{(h'')^2\}$ est minimisée pour $h'' = h''^*$ où

$$h''^* = \sum_{i \epsilon s'} (y_i - \theta x_i)\alpha_i/\pi_i; \tag{2.6}$$

en d'autres termes, h''^* est la fonction d'estimation optimale dans $\bar{H}''(p, \mathbf{q})$.

Preuve. Elle découle directement du théroème 1.1, dans lequel il suffit de remplacer s par s' et les probabilités d'inclusion π_i par $\pi_i q_i$, $i = 1, \ldots, N$.

Théorème 2.3. La fonction d'estimation h''^* définie en (2.6) est optimale dans *toute* la classe $H''(p, \mathbf{q})$ définie en (2.3). En d'autres termes, l'application du théorème 2.2 ne se limite pas à la sous-classe $\bar{H}''$ de H''.

Preuve. Pour n'importe quelle probabilité de réponse $\mathbf{q}$ en (2.4) et n'importe quel plan de sondage p, le paramètre statistique $(\{i, y_i\}: i \in s')$ est *suffisant* pour le vecteur de population $\mathbf{y}$. En termes plus précis, si nous nous reportons aux équations (1.1) et (2.1), la probabilité conditionnelle $\text{Prob}(\chi_{s,s'} \mid \chi_{s'}, \mathbf{y})$ est indépendante de $\mathbf{y}$. Ainsi, pour toute fonction d'estimation $h'' \in H''(p, \mathbf{q})$ dans (2.3), nous avons la fonction d'estimation $E(h'' \mid \chi_{s'}) = \bar{h}'' \in \bar{H}''$ et $\epsilon E(\bar{h}'')^2 < \epsilon E(h'')^2$, ce qui démontre le théorème 2.3.

Lorsque $s \equiv s'$, c'est-à-dire lorsqu'il n'y a aucun non-répondant, doit-on encore estimer h^* par $h'^* = h''^*$? Il est évident que non et Godambe (1986) le montre par des *conditions* appropriées. On tend à se poser la même question dans les cas où il y a très peu de non-répondants et là encore, on peut trouver une réponse a cette question en posant des conditions appropriées. En définitive, le caractère optimal formel de la relation $h'^* = h''$ donne à croire que cette relation est utile et qu'elle est de nature à produire une bonne estimation lorsque le taux de non-réponse est élevé et que les valeurs relatives de q_i sont connues. Toutefois, elle est loin d'être aussi efficace lorsque le taux de non-réponse est peu élevé; dans un tel cas, cette relation gagnerait à être améliorée car on pourrait alors en tirer naturellement des conditions. Il est encore plus important de poser les hypothèses appropriées lorsque les probabilités de réponse sont inconnues. C'est le cas que nous illustrons ci-dessous.

Supposons que la population d'enquête $\mathbf{P}$ est divisé en k strates $\mathbf{P}_j$ de taille N_j, $j = 1,\ldots, k$. Supposons de plus que les probabilités de réponse dans chaque strate sont constantes, c'est-à-dire

$$q_i = q^{(j)} \text{ for all } i \in \mathbf{P}_j;\ j = 1, \ldots, k. \tag{2.7}$$

Contrairement à (2.4), où l'on supposait que les probabilités de réponse étaient connues, l'expression (2.7) suppose des probabilités de réponse $q^{(j)}$, $j = 1, \ldots, k$ *inconnues*. Désignons par p_0 le plan de sondage stratifié qui consiste à prélever dans le strate $\mathbf{P}_j$ un échantillon aléatoire simple (sans remise) de taille n_j, $j = 1, \ldots, k$. Comme en (2.3), nous définissons maintenant la classe des fonctions d'estimation sans biais $h_1(\chi_{s,s'})$

$$H_1(p_0) = \{h_i: E(h_1 - \bar{g}) = 0 \text{ pour toutes les valeurs de } \mathbf{y}, \theta \text{ et } q^{(j)},\ j = 1, \ldots, k\}, \tag{2.8}$$

où $q^{(j)}$ est défini comme en (2.7). Posons $s'_j = s' \cap \mathbf{P}_j$ et $|s'_j| = n'_j$, c'est-à-dire la taille de l'échantillon de répondants de la strate $\mathbf{P}_j$, $j = 1, \ldots, k$.

Théorème 2.4. Pour le plan de sondage p_0 dans la classe $H_1(p_0)$ des fonctions d'estimation définie en (2.8), selon le modèle de superpopulation (1.2), $\epsilon E(h_1^2)$ est minimisée pour $h_1 = h_1^*$ où

$$h_1^* = \sum_{j=1}^{k} \sum_{i \in s'_j} (y_i - \theta x_i)\, \alpha_i / \left(\frac{n'_j}{N_j}\right); \tag{2.9}$$

et d'autres termes, h_1^* est la fonction d'estimation optimale dans $H_1(p_0)$.

Preuve. La distribution d'échantillonnage des données $\chi_{s,s'}$ définies en (2.1) dépend non seulement du vecteur de population inconnu $\mathbf{y}$, mais aussi du paramètre inconnu $q^{(j)}$, $j = 1, \ldots, k$. Or, pour chaque vecteur $\mathbf{y}$ donné, le paramètre statistique n'_j, $j = 1, \ldots, k$ est *entièrement suffisant* pour le paramètre $q^{(j)}$, $j = 1, .., k$. Donc, pour des valeurs données de $\mathbf{y}$ et de θ dans (2.8),

$$[E(h_1 - \bar{g}) = 0, \text{ pour tout } q^{(j)},\ j = 1, \ldots, k]$$

$$\Rightarrow E\{(h_1 - \bar{g}) | n'_j,\ j = 1, \ldots, k\} = 0, \tag{2.10}$$

ne tenant pas compte des séries de valeur nulle. De plus, *étant donné* le nombre de répondants n_j' de la strate $\mathbf{P}_j$, la probabilité de $i \epsilon s_j'$ est $(n_j/N_j)(n_j'/n_j) = (n_j'/N_j)$. Par conséquent, pour toute fonction d'estimation $h_1 \epsilon H_1$ dans (2.8), nous avons par la théroème 2.3 l'inéquation.

$$\epsilon E\{(h_1^*)^2 | n_j', j = 1, \ldots, k\} \leq \epsilon E\{(h_1)^2 | n_j', j = 1, \ldots, k\}, \tag{2.11}$$

h_1^* étant définie en (2.9). On démontre le théorème 2.4 en calculant les espérance mathématiques de part et d'autre de l'inéquation (2.11) pour les variations de n_j', $j = 1, \ldots, k$.

On peut interpréter intuitivement la fonction d'estimation optimale h_1^* définie en (2.9) de la façon suivante: si les probabilités de réponse $q^{(j)}$, $j = 1, \ldots, k$ en (2.7) était *connues*, alors, selon le théorème 2.3, la fonction d'estimation optimale pour le plan de sondage p_o serait donnée par

$$h'' = \sum_{j=1}^{k} \sum_{i \in s_j'} (y_i - \theta x_i)\alpha_i / (\frac{n_j}{N_j} q^{(j)}).$$

Or, lorsque $q^{(j)}$ est inconnue (ce qui est effectivement le cas dans le théorème 2.4), nous l'*estimons* par la formule (n_j'/n_j), $j = 1, \ldots, k$. Si nous substituons cette expression à $q^{(j)}$ dans h'', nous obtenons la fonction d'estimation h_1^* définie en (2.9).

Les estimations découlant de la résolution des équations $h'^* = 0$, $h''^* = 0$ et $h_1^* = 0$ définies respectivement en (2.5), (2.6) et (2.9) ont déjà été proposées par plusieurs auteurs selon des critères de vraisemblance. À ce sujet, l'ouvrage de Cassel et coll. (1983) mérite d'être cité. L'hypothèse relative aux probabilités de réponse (2.4) semble avoir évoluée peu à peu dans les ouvrages de statistique. Hartley (1946) est une source intéressante à cet égard.

3. PROBABILITÉS D'INCLUSION OPTIMALE

À ce stade-ci, il convient de souligner que le caractère optimal de la fonction d'estimation h''^* définie en (2.6) a été établi suivant le modèle de superpopulation (1.2), lequel *ne* définit *pas* la fonction de variance. Il faudrait pourtant définir cette fonction dans le modèle précité afin de connaître les probabilités d'inclusion optimale. Nous posons par hypothèse

$$\epsilon(y_i - \theta x_i)^2 = \sigma^2 f(x_i), \; i = 1, \ldots, N, \tag{3.1}$$

où f est une fonction *connue* de x, et σ^2 peut être inconnue. Si maintenant nous appliquons l'équation (3.1) à la fonction d'estimation h''^* définie en (2.6), nous obtenons

$$\epsilon E(h''^*)^2 = \sum_{i=1}^{N} \frac{\epsilon(y_i - \theta x_i)^2 \alpha_i^2}{\pi_i q_i} = \sigma^2 \sum_{i=1}^{N} \frac{f(x_i)\alpha_i^2}{\pi_i q_i}. \tag{3.2}$$

Dans l'équation ci-dessus, les probabilités de réponse q_i définies en (2.4) sont des valeurs *données* (*fixes*). Par ailleurs, on peut établir les probabilités d'inclusion optimale d'un plan de sondage en minimisant $\epsilon E(h''^*)^2$ dans (3.2) moyennant la restriction (A) ou (B).

$$(A): \sum_{i=1}^{N} \pi_i = \text{constante},$$

$$(B): \sum_{i=1}^{N} \pi_i q_i = \text{constante}. \tag{3.3}$$

Selon (A), nous maintenons constante la taille moyenne de l'échantillon s car $E|s| = \Sigma_i^N \pi_i$. Selon (B), nous maintenons constante la taille moyenne de l'échantillon effectif s' car $E|s'| = \Sigma_i^N \pi_i q_i$. Or, comme les valeurs de q_i sont *fixes*, $\epsilon E(h''^*)^2$ peut être minimisée selon les conditions respectives suivantes:

$$(A): \pi_i \propto \{\frac{f(x_i)}{q_i}\}^{1/2}\alpha_i,$$

$$(B): \pi_i \propto \frac{(f(x_i))^{1/2}}{q_i}\alpha_i. \tag{3.4}$$

En désignant par n' la taille de l'échantillon effectif s', c'est-à-dire $|s'| = n'$, nous tirons de (B) en (3.4) l'équation

$$\pi_i = \frac{(f(x_i))^{1/2}\alpha_i}{\{\Sigma_1^N (f(x_i))^{1/2}\alpha_i\}} \quad \frac{E(n')}{q_i}, \; i = 1, \ldots, N. \tag{3.5}$$

En outre, pour un plan de sondage à taille d'échantillon fixe tel que

$$\text{Probabilité}\{s: |s| \neq n\} = 0,$$

nous déduisons de (3.5).

$$\sum_{i=1}^{n} \pi_i = n = \sum_{i=1}^{N} \frac{(f(x_i))^{1/2}\alpha_i}{\{\Sigma_1^N (f(x_i))^{1/2}\alpha_i\}} \quad \frac{1}{q_i} E(n'). \tag{3.6}$$

Lorsque dans l'équation ci-dessus toutes les probabilités de réponses q_i, $i = 1, \ldots, N$ sont égales, par exemple $q_i = q$, $i = 1, \ldots, N$,

$$n = E(n')/q; \tag{3.7}$$

si, par exemple, $q = 1/2$, la taille de l'échantillon (initial) s devra être le double de l'espérance mathématique de la taille de l'échantillon effectif (s')!

Supposons maintenant que la population d'enquête $\mathbf{P}$ est divisée en strates $\mathbf{P}_j$, $j = 1. \ldots, k$ de sorte que les probabilités de réponse dans chaque strate soient constantes, c'est-à-dire que l'équation (2.7) soit satisfaite. Pour un plan de sondage stratifié qui consiste à prélever un échantillon de taille n_j dans la strate $\mathbf{P}_j$, $j = 1, \ldots k$ nous tirons de (3.5)

$$n_j = \frac{E(n')}{q^{(j)}} \quad \frac{\sum_{i \in \mathbf{P}_j} (f(x_i))^{1/2}\alpha_i}{\sum_{i \in \mathbf{P}} (f(x_i))^{1/2}\alpha_i}, \; j = 1, \ldots, k. \tag{3.8}$$

Si les valeurs de $(f(\chi_i))^{1/2}\alpha_i$ sont constantes pour $i = 1, \ldots N$, il est clair d'après l'équation ci-dessus qu'il faudra prélever un échantillon relativement plus grand dans la strate et que la probabilité de réponse devra être moins élevée si l'on veut obtenir une répartition optimale. De fait, nous aurions dans un tel cas

$$E(n_j') = E(n')/k,$$

où n_j' est la taille de l'échantillon effectif s_j' prélevé dans la strate $\mathbf{P}_j$, $j = 1, \ldots, k$.

BIBLIOGRAPHIE

BREWER, K.R.W. (1963). Ratio estimation in finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

CASSEL, C.M., SARNDA, C.E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problems. Dans *Incomplete Data in Sample Surveys 3*, (colligé par W.G. Madow et Ingram Olkin), Academic Press: New York, 143-160.

GODAMBE, V.P. (1982). Estimation in survey sampling: Robustness and optimality. *Journal of the American Statistical Association*, 77, 393-403.

GODAMBE, V.P. (1986). Quasi-score function, quasi-observed Fisher information and conditioning in survey sampling (non publié).

GODAMBE, V.P. and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Institute Review* (à paraître).

HAJEK, J. (1971). Contribution to discussion of paper by D. Basu. Dans *Foundations of Statistical Inference* (colligé par V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.

HARTLEY, H.O. (1946). Discussion of paper by F. Yates. *Journal of the Royal Statistical Society* Série A, 109, 37.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-589.