Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Nouveau regard sur les intervalles de confiance en échantillonnage

par V.P.Godambe et M.E.Thompson

Décembre 1999

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 1999
Vol. 25, N° 2, pp. 185-198
Statistique Canada, N° 12-001 au catalogue

Nouveau regard sur les intervalles de confiance en échantillonnage

V.P. Godambe et M.E. Thompson [1]

Résumé

En échantillonnage, comme dans d'autres domaines conventionnels de la statistique, les intervalles de confiance pour un paramètre sont souvent obtenus en inversant la distribution d'une quantité approximative servant de pivot, {(estimation – paramètre)/(variance estimatée)$^{1/2}$}. Par ailleurs, la théorie des fonctions d'estimation suggère une méthode plus directe de construction d'une quantité servant de pivot, et donc d'intervalles de confiance. Ces autres intervalles de confiance fonctionnent beaucoup mieux que les intervalles de confiance conventionnels dans des études de simulation.

Mots clés : Intervalles de confiance; fonctions d'estimation; optimalité; stratification; échantillonnage.

1. Introduction historique

Le thème des intervalles de confiance a d'abord été discuté dans une communication bien connue de Neyman (1934) présentée à la Royal Statistical Society. La communication portait sur l'échantillonnage. Pourtant, l'exposé de Neyman n'a pas décrit la construction comme telle des intervalles de confiance pour un schéma d'échantillonnage. À l'époque la distinction entre les paramètres d'une population observée, d'une part, et ceux d'une population hypothétique, d'autre part, n'était peut-être pas bien comprise du tout (Deming 1950, Godambe 1976, Godambe 1997, Smith 1997). Nous pouvons affirmer aujourd'hui que l'exposé de Neyman sur les intervalles de confiance concernait principalement les paramètres d'une population hypothétique. Dans une publication subséquente (1937), Neyman a explicitement montré comment on pouvait obtenir des intervalles de confiance à partir d'une quantité servant de pivot, une fonction d'observations et du paramètre d'intérêt ayant une distribution fixe (connue). La disponibilité de tels « pivots » (ou de pivots approximatifs) pour certaines populations hypothétiques caractérisées par quelques paramètres scalaires se laisse facilement démontrer. Par contre, afin de caractériser une population observée de taille N, il faut un paramètre de N dimensions (Basu 1958, Hájek 1959). Dans cette situation, généralement, aucune fonction non triviale des observations et du paramètre d'intérêt ne peut servir exactement de pivot pour la distribution induite par un plan de sondage probabiliste.

La section VI de la communication de Neyman de 1934 était une annexe. Entre autres choses, l'annexe comportait une note I, portant sur des intervalles de confiance, suivie d'une note II intitulée « The Markoff Method and Markoff Theorem on Least Squares ». Le théorème dont il est question ici, selon la terminologie moderne, est le théorème de Gauss-Markoff de l'estimation de la variance minimale non biaisée. Or il est vrai que la « variance » d'un estimateur non biaisé, si elle est connue, permet de construire un intervalle de confiance approximatif en supposant qu'une distribution normale approximative pour l'estimateur : (estimateur – paramètre)/(variance)$^{1/2}$ représente un pivot approximatif. Cela n'est guère utile toutefois, car la « variance » mentionnée ci-dessus n'est jamais connue dans le contexte d'un échantillonnage. D'après les publications qui traitent de ce sujet (par exemple, Chaudhuri et Vos 1988), une pratique courante consiste à remplacer la variance inconnue par une « estimation ». La question de base, qui n'est généralement pas abordée dans la littérature, est celle-ci : laquelle des nombreuses estimations de la variance servirait de pivot (ou de pivot approximatif) menant à une série d'intervalles de confiance plausibles ? Ce problème, pour une population hypothétique comportant un modèle paramétrique sous-jacent, est résolu à l'aide de l'information de Fisher observée (Efron et Hinkley 1978). Une généralisation de l'information de Fisher observée, pour un modèle semi-paramétrique, fournie par la théorie des fonctions d'estimation optimales (Godambe 1985, Godambe et Thompson 1986) fournit une réponse à cette question dans le contexte de l'échantillonnage.

Jusqu'à présent le thème des « intervalles de confiance en échantillonnage » a été envisagé dans le cadre des « fonctions d'estimation » dans quelques exposés seulement. Woodruff (1952) a présenté une première démonstration d'intervalles de confiance pour des « mesures de position » d'une population observée, à l'aide de fonctions d'estimation utilisées de façon informelle. Cet argument a été commenté en détail par Godambe (1991). Un deuxième exposé a été préparé par Binder et Patak (1994). Le premier auteur, dans une publication antérieure (Binder 1983), avait utilisé de façon informelle des fonctions d'estimation pour des enquêtes complexes. Dans ce cas, toutefois, les intervalles de confiance présentés étaient de type plutôt conventionnel. L'exposé de Binder et Patak (1994) et le présent article se fondent sur la théorie des fonctions d'estimation. La différence de base entre les deux exposés est que ce dernier se rattache essentiellement au critère d'optimalité des

1. V.P. Godambe et M.E. Thompson, Statistics and Acturial Science, University of Waterloo, Waterloo (Ontario) Canada N2L 3G1.

fonctions d'estimation. Ce critère d'optimalité relie les populations observées présentement à un modèle de superpopulation semi-paramétrique, bien que de façon très souple. Cette relation, comme le montre le présent exposé, oriente le choix de la fonction d'estimation et des intervalles de confiance implicites à utiliser pour un problème donné. À part ce renvoi à un modèle de superpopulation, les intervalles de confiance présentés dans le présent exposé se fondent sur le plan. De même, les deux articles, celui de Binder et Patak (1994) et le présent exposé (section 5), abordent le cas important des « paramètres dérangeants ». Toutefois, les problèmes abordés dans les deux articles sont différents et les résultats ne chevauchent aucunement.

2. Intervalles de confiance robustes pour le modèle

Nous entamons la discussion en décrivant l'estimation d'un paramètre unique θ d'un modèle de superpopulation. Nous supposons que, pour le modèle, les observations y_i pour i de l'échantillon s sont indépendantes, avec

$$\varepsilon g_i(y_i, \theta) = 0, \; i \in s;$$

$$\varepsilon g_i^2(y_i, \theta) = \sigma_i^2, \; i \in s, \qquad (2.1)$$

où les g_i sont des fonctions d'estimation élémentaires.

Nous pouvons obtenir des intervalles de confiance approximatifs « robustes pour le modèle » pour θ en inversant

$$\left| \left\{ \sum_{i \in s} g_i(y_i, \theta) \right\} \Big/ \left\{ \sum_{i \in s} g_i^2(y_i, \theta) \right\}^{1/2} \right| = z \qquad (2.2)$$

où z est un percentile de la distribution $N(0, 1)$

Une version très générale de (2.2) a été présentée à titre de pivot approximatif, dans le cadre de processus aléatoires, par un des auteurs (Godambe 1985). Des versions antérieures de (2.2), de Fisher (1925), d'Efron et Hinkley (1978) et de Royall (1986), avaient toutes utilisé le numérateur (estimation $\hat{\theta}$ – paramètre θ); dans les trois cas, les dénominateurs étaient la racine carrée de différents estimateurs de la variance de $(\hat{\theta} - \theta)$. On trouvera une étude précoce des propriétés de (2.2) dans Mach (1988). Une étude subséquente a été menée par Vinod (1998).

Le bien-fondé de l'utilisation de (2.2) comporte différents aspects :

a) $\sum_{i \in s} g_i^2(y_i, \theta)$ est un estimateur, non biaisé pour le modèle, de $\mathrm{Var}(\sum_{i \in s} g_i(y_i, \theta))$, peu importe la forme de $\{\sigma_i^2\}$;

b) étant analogue à l'information observée, $\sum_{i \in s} g_i^2(y_i, \theta)$ peut être considéré comme non biaisé dans certaines conditions, compte tenu d'aspects importants de la structure de l'échantillon; plus précisément, $\sum_{i \in s} g_i^2(y_i, \theta)$ est la « variance » du numérateur « à la condition » que l'on ait le même partitionnement qui sous-tend l'optimalité du numérateur (Godambe 1985; Godambe et Thompson 1986);

c) si le modèle est mal précisé, $\sum_{i \in s} g_i^2(y_i, \theta)$ englobera dans une certaine mesure le biais de la fonction d'estimation de même que sa variabilité;

d) si les y_i, $i \in s$ sont des $N(\theta, \sigma^2)$ indépendants et distribués de façon identique, alors

$$\tau = \frac{\sum_{i \in s} (y_i - \theta)}{\sqrt{\sum_{i \in s} (y_i - \theta)^2}} \qquad (2.3)$$

se rapproche davantage de $N(0, 1)$ que la statistique t, puisque $\mathrm{Var}(\tau) = 1$, et l'aplatissement de τ est $3 - 6/(n + 2)$; les intervalles de confiance pour θ fondés sur l'inversion de $|\tau| = z$ sont

$$\bar{y} \pm \sqrt{\frac{n - 1}{n - z^2}} \frac{z s_y}{\sqrt{n}} \qquad (2.4)$$

où s_y est l'écart-type de l'échantillon.

Supposons maintenant que $\boldsymbol{\theta}$ est un paramètre de valeur-vecteur et que $\psi(\boldsymbol{\theta})$ est un paramètre d'intérêt scalaire. Supposons que les y_i sont indépendants dans le cadre du modèle de superpopulation, que $g_i(y_i, \boldsymbol{\theta})$ a le même caractère dimensionnel que $\boldsymbol{\theta}$, et que le système d'équation d'estimation non biaisé

$$\sum_{i \in s} g_i(y_i, \boldsymbol{\theta}) = 0 \qquad (2.5)$$

résulte de la minimisation d'une fonction objective scalaire

$$\sum_{i \in s} G_i(y_i, \boldsymbol{\theta}). \qquad (2.6)$$

Pour l'estimation de $\psi(\boldsymbol{\theta})$, on pourrait procéder par « profilage », c'est-à-dire en trouvant un $\tilde{\boldsymbol{\theta}}(\psi)$ qui permettrait de minimiser (2.6) pour une valeur fixe $\psi(\boldsymbol{\theta}) = \psi$. Alors, la fonction d'estimation pour ψ permettrait de trouver $\hat{\psi}$ pour minimiser

$$\sum_{i \in s} G_i(y_i, \tilde{\boldsymbol{\theta}}(\psi)).$$

La forme vectorielle du système permettant de trouver $\boldsymbol{\theta}(\psi)$ serait

$$\sum_{i \in s} g_i(y_i, \boldsymbol{\theta}) - \lambda \frac{\partial \psi}{\partial \boldsymbol{\theta}} = 0, \qquad (2.7)$$

où λ est un multiplicateur de Lagrange (scalaire), avec la restriction voulant que $\psi(\boldsymbol{\theta}) = \psi$. Il existe une correspondance univoque entre ψ et λ, λ étant 0 lorsque ψ est $\hat{\psi}$. L'estimation $\hat{\psi}$ permet de résoudre

$$\sum_{i \in s} g_i(y_i, \tilde{\boldsymbol{\theta}}(\psi)) = 0 \qquad (2.8)$$

ou une combinaison linéaire de ses composants.

Si $\boldsymbol{a}$ est un vecteur de la dimension de $\boldsymbol{\theta}$, et si

$$\boldsymbol{a}^{\tau} \sum_{i \in s} g_i(y_i, \boldsymbol{\theta}(\psi)) = 0 \qquad (2.9)$$

est une équation d'estimation (possiblement sous-optimale) pour $\hat{\psi}$, il semble raisonnable d'obtenir des intervalles de confiance approximatifs pour ψ en inversant

$$\frac{\left| \boldsymbol{a}^{\tau} \sum_{i \in s} g_i(y_i, \boldsymbol{\theta}(\psi)) \right|}{\sqrt{\boldsymbol{a}^{\tau} \left[\sum_{i \in s} g_i(y_i, \boldsymbol{\theta}(\psi)) g_i^{\tau}(y_i, \boldsymbol{\theta}(\psi)) \right] \boldsymbol{a}}} = z. \qquad (2.10)$$

Remarques :

i) Les équations d'estimation (2.8) et (2.9) ne sont qu'approximativement sans biais, et leurs termes ne sont qu'approximativement indépendants. Le recours à (2.10) sera donc plus facile à justifier théoriquement pour de grands échantillons.

ii) Même si le système (2.5) ne provient pas de la minimisation d'une fonction objective (2.6), le processus d'estimation de ψ à l'aide de (2.7), (2.9) et (2.10) peut tout de même se poursuivre, et il demeure significatif.

iii) Lorsque $\boldsymbol{\theta}(\boldsymbol{y})$ et $\psi(\boldsymbol{y})$ sont des paramètres de population finie analogues à $\boldsymbol{\theta}$ et à ψ, le même processus peut se produire, (2.10) étant remplacé par

$$\frac{\left| \boldsymbol{a}^{\tau} \sum_{i \in s} g_i(y_i, \boldsymbol{\theta}(\psi(\boldsymbol{y}))) \right|}{\sqrt{v(\chi_s, \boldsymbol{\theta}(\psi(\boldsymbol{y})))}} = z, \qquad (2.11)$$

où $\boldsymbol{y} = (y_1, ..., y_N)$, s est un échantillon, $\chi_s = \{(i, y_i) : i \in s\}$, et le dénominateur est une estimation appropriée de l'écart-type du numérateur. (Voir la notation et l'élaboration aux sections 3 et 7.)

En réalité, l'objectif du présent exposé est d'explorer l'adaptation de (2.2) et de (2.10) et de leur bien-fondé au contexte de l'échantillonnage. Ainsi, supposons que nous avons une population finie de taille N, que y_i, $i = 1, ..., N$ sont des $N(\theta, \sigma^2)$ indépendants et distribués de façon identique, et que nous voulons estimer, ou « prédire » de façon équivalente, la moyenne de la population finie $\bar{Y} = \sum_{i=1}^{n} y_i/N$ à partir des observations d'un échantillon. La même théorie de distribution qu'en (d) ci-dessus permet d'établir que

$$\tau = \frac{n^{1/2} b(\bar{y} - \bar{Y})}{\sqrt{(n-1)s_y^2 + b^2(\bar{y} - \bar{Y})^2}}, \qquad (2.12)$$

où $b = (1/n - 1/N)^{1/2}$, est approximativement $N(0, 1)$. Le fait d'inverser $|\tau| = z$ donne lieu à des intervalles « de prédiction » pour $\bar{Y}$ sous la forme

$$\bar{y} \pm \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} s_y \, z \sqrt{\frac{n-1}{n - z^2}}. \qquad (2.13)$$

Pour le modèle hypothétique, ceux-ci auront des propriétés de prédiction améliorées relativement aux intervalles de confiance fondés sur l'échantillonnage aléatoire simple habituel

$$\bar{y} \pm \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} s_y \, z. \qquad (2.14)$$

Lorsque le plan de sondage est un échantillonnage aléatoire simple, un estimateur sans biais pour l'échantillonnage de la variance de l'échantillonnage de $n^{1/2} b(\bar{y} - \bar{Y})$ est

$$\frac{N}{N-1} \sum_{i \in s} (y_i - \bar{Y})^2. \qquad (2.15)$$

Lorsque N est grand, le pivot τ de (2.12) est approximativement

$$\frac{n^{1/2} b(\bar{y} - \bar{Y})}{\sqrt{\frac{N}{N-1} \sum_{i \in s} (y_i - \bar{Y})^2}},$$

ce qui représente approximativement

$$\frac{\sum_{i \in s} (y_i - \bar{Y})}{\sqrt{n\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i \in s} (y_i - \bar{Y})^2}}. \qquad (2.16)$$

En résumé, donc, le pivot de (2.12) adopte une forme semblable à celle de (2.2), θ étant remplacé par le paramètre de la population finie $\bar{Y}$. Le pivot de (2.12) comporte à la fois une interprétation de prédiction et une justification (échantillonnage aléatoire simple) fondée sur le plan.

3. Fonctions d'estimation optimales pour les populations observées

Assez souvent, l'estimation pour des populations observées se fonde sur le plan aussi bien que sur le modèle. Godambe et Thompson (1986) ont proposé le cadre suivant pour l'estimation optimale de quantités de la population finie qui correspondent à des paramètres de la superpopulation.

Soit θ, un paramètre de la superpopulation; soit $\varphi_1, ..., \varphi_N$ des fonctions d'estimation élémentaires indépendantes telles que $\varepsilon \varphi_i(y_i, \theta) = 0$ pour $i = 1, ..., N$. Soit $\boldsymbol{y} = (y_1, ..., y_N)$, le vecteur de population des réponses; soit aussi $\theta(\boldsymbol{y})$, le paramètre de la population finie qui représente la solution en θ de

$$\sum_{i=1}^{N} \varphi_i(y_i, \theta) = 0. \qquad (3.1)$$

Nous considérons θ et $\theta(\boldsymbol{y})$ comme des paramètres associés, l'un de la superpopulation, l'autre de la population finie. Nous les considérons comme réels par souci de simplicité.

Supposons que $p = \{p(s), s \in S\}$ est un plan de sondage, ou une fonction de probabilité en S, la collection d'échantillons ou de sous-ensembles s de $\{1, ..., N\}$. Le plan de sondage p entraîne une distribution pour le résultat $\chi_s = \{(i, y_i) : i \in s\}$. Soit E_p, l'espérance pour le plan de sondage. Une fonction d'estimation d'échantillon est une fonction $g_s(\chi_s, \theta)$, et une stratégie de fonction d'estimation (g, p) est considérée comme sans biais si

$$E_p\, g_s(\chi_s, \theta) \equiv \sum_{i=1}^{N} \varphi_i(y_i, \theta) \qquad (3.2)$$

pour tous les $\boldsymbol{y}$, θ. L'estimation ponctuelle pour $\theta(\boldsymbol{y})$ se poursuit si l'on trouve la solution $\hat{\theta}_s$ de

$$g_s(\chi_s, \theta) = 0. \qquad (3.3)$$

Parmi les stratégies non biaisées (g, p) (p, fixe), la stragégie (g^*, p) peut être considérée comme optimale si

$$\varepsilon E_p \left(g_s^*(\chi_s, \theta) - \sum_{i=1}^{N} \varphi_i(y_i, \theta) \right)^2 \qquad (3.4)$$

est minimale. Godambe et Thompson (1986) ont pu montrer que la fonction d'estimation optimale est donnée par

$$g_s^*(\chi_s, \theta) = \sum_{i \in s} g_i^*(y_i, \theta) \qquad (3.5)$$

où $g_i^*(y_i, \theta) = \varphi_i(y_i, \theta)/\pi_i$ et π_i est la probabilité (sous p) d'inclusion de i dans l'échantillon. À noter que l'optimalité de g_s^* en ce sens est indépendante de la structure de la variance $\{\varepsilon \varphi_i^2(y_i, \theta)\}$. L'absence de biais (3.2) est une contrainte très sévère, mais elle est également importante dans le contexte de l'échantillonnage.

Supposons maintenant que nous avons une fonction d'estimation d'échantillonnage $g_s(\chi_s, \theta) = \sum_{i \in s} g_i(y_i, \theta)$ qui satisfait (3.2), et qui peut être ou ne pas être optimale pour l'estimation ponctuelle de $\theta(\boldsymbol{y})$ suivant le critère (3.4). Puisque $g_s(\chi_s, \theta)$ est également une fonction d'estimation pour θ, l'inversion de (2.2) :

$$\left| \frac{\sum_{i \in s} g_i(y_i, \theta)}{\sqrt{\sum_{i \in s} g_i^2(y_i, \theta)}} \right| = z$$

devrait fournir des intervalles de confiance pour θ avec une bonne couverture, pour le modèle de superpopulation.

Lorsque $\theta(\boldsymbol{y})$ est l'objet de l'inférence, il est de nouveau tentant d'adopter une stratégie de prédiction et d'envisager l'inversion

$$\left| \frac{\sum_{i \in s} g_i(y_i, \theta(\boldsymbol{y}))}{\sqrt{v_m(\chi_s, \theta(\boldsymbol{y}))}} \right| = z \qquad (3.6)$$

où

$$\varepsilon v_m(\chi_s, \theta(\boldsymbol{y})) = \varepsilon \left(\sum_{i \in s} g_i(y_i, \theta(\boldsymbol{y})) \right)^2. \qquad (3.7)$$

Une autre possibilité est d'adopter une stratégie d'essai inverse fondée sur le plan et d'envisager l'inversion

$$\left| \frac{\sum_{i \in s} g_i(y_i, \theta(\boldsymbol{y}))}{\sqrt{v_p(\chi_s, \theta(\boldsymbol{y}))}} \right| = z \qquad (3.8)$$

où

$$E_p v_p(\chi_s, \theta) = E_p \left(\sum_{i \in s} g_i(y_i, \theta) - \sum_{i \in s}^{N} \varphi_i(y_i, \theta) \right)^2. \qquad (3.9)$$

Pour un plan bien choisi qui correspond à des éléments appropriés du modèle, ces deux stratégies devraient donner des résultats qui se rapprochent les uns des autres; les intervalles pour $\theta(\boldsymbol{y})$ seront différents des intervalles fournis par (3.6), d'un montant qui tient compte du caractère fini de la population.

Des considérations semblables s'appliquent à un paramètre multidimensionnel $\theta(\boldsymbol{y})$. Toutefois, on trouvera à la prochaine section une description de la stratégie générale comportant un paramètre unidimensionnel, pour la situation un peu artificielle de strates comportant une moyenne commune.

4. Échantillonnage aléatoire simple stratifié

Nous utilisons la notation habituelle. La population étiquetée de N individus (unités) est notée $\mathcal{P} = \{i : i = 1, ..., N\}$. La population $\mathcal{P}$ est divisée en k strates non chevauchantes $\mathcal{P}_j$ de taille N_j, $j = 1, ..., k$. Une variable aléatoire définie pour la population $\mathcal{P}$ est y, supposée scalaire par souci de simplicité. Pour l'individu i, $y = y_i$, $i = 1, ..., N$. Le vecteur population $\boldsymbol{y} = (y_1, ..., y_N)$. Pour obtenir une estimation de la moyenne de la population $\bar{Y} = \sum_1^N y_i / N$ on tire un échantillon s de taille n à partir de $\mathcal{P}$, avec un plan d'échantillonnage aléatoire simple stratifié sans remise. Les échantillons de différentes strates $\mathcal{P}_j$ sont notées s_j, $|s_j| = n_j$, $j = 1, ..., k$; $\sum n_j = n$. De plus, $\bar{Y}_j$ et $\bar{y}_j$ représentent les moyennes de y dans la strate $\mathcal{P}_j$ et l'échantillon s_j respectivement, $j = 1, ..., k$. Ainsi

$$\bar{Y} = \sum_{j=1}^{k} N_j \bar{Y}_j / N. \qquad (4.1)$$

De façon analogue, nous définissons

$$\bar{y} = \sum_{j=1}^{k} N_j\, \bar{y}_j / N. \qquad (4.2)$$

Si nous supposons que les composants du vecteur population $\boldsymbol{y} = (y_1, ..., y_N)$ ont été tirés indépendamment à partir d'une superpopulation comportant une moyenne commune $\varepsilon(y_i) = \theta$, $i = 1, ..., N$ mais des variances possiblement différentes $\varepsilon(y_i - \theta)^2$, $i = 1, ..., N$, et si nous considérons $\bar{Y}$ comme la solution de $\sum_{i=1}^{N} \varphi_i(y_i, \theta) = 0$, où $\varphi_i(y_i, \theta) = y_i - \theta$, alors la fonction d'estimation optimale pour l'estimation de la moyenne de la population $\bar{Y}$ en (4.1) est donnée par

$$g = \sum_{j=1}^{k} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in s_j} (y_i - \bar{Y}) \qquad (4.3)$$

(Godambe et Thompson 1986; Godambe 1995). Ce g se présente sous la forme $g_s^*(\chi_s, \theta(\boldsymbol{y}))$ où g_s^* est donné en (3.5). Ainsi l'estimation optimale de $\bar{Y}$ est donnée par $\bar{y}$, la solution en $\bar{Y}$ de l'équation $g = 0$. Les modèles de superpopulation définis uniquement par les quelques premiers moments, comme le modèle mentionné ci-dessus défini par $\varepsilon(y_i) = \theta$, $i = 1, ..., N$, sont appelés semi-paramétriques, contrairement aux modèles entièrement paramétriques précisés par les fonctions de densité.

La « fonction d'estimation optimale » pour un modèle semi-paramétrique partage de nombreuses propriétés statistiquement importantes avec la « fonction de caractérisation » pour un modèle paramétrique. Dans le modèle semi-paramétrique, donc, la fonction d'estimation optimale s'appelle une fonction de quasi-caractérisation. (Godambe 1985; Godambe et Heyde 1987; Godambe et Thompson 1989). Pour un modèle paramétrique, on peut construire des intervalles de confiance à l'aide de l'information de Fisher (définie comme la variance de la fonction de caractérisation), ou de son estimation naturelle l'information de Fisher observée. De même, dans le cas d'un modèle semi-paramétrique, on peut obtenir les intervalles de confiance à partir de la fonction de quasi-caractérisation, c'est-à-dire la fonction d'estimation optimale, et de sa variance estimatée. Dans le contexte d'une enquête, bien que le critère d'optimalité soit lié au modèle de superpopulation $\varepsilon(y_i) = \theta$, les propriétés des intervalles de confiance donnés ci-dessous se fondent surtout, sinon entièrement, sur le plan.

La variance induite par le plan de la fonction d'estimation optimale g en (4.3) est donnée par

$$V(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{N_j - 1} \sum_{i \in \mathcal{P}_j} (y_i - \bar{Y}_j)^2. \quad (4.4)$$

De plus, puisque notre paramètre d'intérêt est $\bar{Y}$, les y_i inobservés et les moyennes de strates $\bar{Y}_j$ en (4.4) sont des paramètres dérangeants. Le modèle de superpopulation qui sous-tend la fonction d'estimation g en (4.3), c'est-à-dire $\varepsilon(y_i) = \theta$, $i = 1, ..., N$, suggère que pour de grandes tailles de strates N_j nous pouvons laisser de côté les différences $\bar{Y}_j - \bar{Y}$, et remplacer $\bar{Y}_j$ par $\bar{Y}$, $j = 1, ..., k$ en (4.4). (On trouvera à la section 7 une discussion des modèles pour lesquels les différences $\bar{Y}_j - \bar{Y}$ ne peuvent pas être laissées de côté.) Grâce à ce remplacement, une estimation de la variance $V(g)$ en (4.4) peut être donnée par

$$\hat{V}(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \frac{N_j}{(N_j - 1)} \frac{1}{n_j} \sum_{i \in s_j} (y_i - \bar{Y})^2. \quad (4.5)$$

$$= \left\{ \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{(n_j - 1)} \sum_{i \in s_j} (y_i - \bar{y}_j)^2 \right\} + R = \hat{V}_0 + R \quad (4.6)$$

où $\bar{y}_j$ est la moyenne de l'échantillon s_j, $j = 1, ..., k$ comme en (4.2). Dans le second membre de l'équation (4.6) le premier terme est $O(1/n_j)$ tandis que le deuxième terme R est $O(1/n_j^2)$. Pour de grands échantillons, donc, si on laisse de côté le terme R, $\hat{V}$ en (4.5) se réduit à l'estimation conventionnelle $\hat{V}_0$. Cela entraîne les intervalles de confiance conventionnels pour $\bar{Y}$ d'après l'inversion de la distribution de $\{g/(\hat{V}_0)^{1/2}\}$. Toutefois, lorsque la taille de l'échantillon n_j, $j = 1, ..., k$ n'est pas très grande, la théorie de la fonction d'estimation suggère des intervalles de confiance pour $\bar{Y}$ fondés sur la distribution $N(0, 1)$ asymptotique de $\{g/(\hat{V})^{1/2}\}$.

Pour un plan d'échantillonnage aléatoire simple stratifié, nous écrivons la fonction d'estimation g en (4.3) sous la forme

$$g = \sum_{i \in s} g_i, \qquad (4.7)$$

où comme auparavant s représente l'échantillon (d'individus tirés de toutes les strates). Alors, pour de grands n_j et N_j, si on laisse de côté la correction de strates finies, et si $(n_j - 1)$ est remplacé par n_j, $j = 1, ..., k$ en (4.5), nous avons

$$\hat{V}(g) \simeq \hat{V}_\alpha(g) = \sum_{i \in s} g_i^2 = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{i \in s_j} (y_i - \bar{Y})^2. \quad (4.8)$$

On constate facilement que, pour l'échantillonnage aléatoire simple de chaque strate (encore une fois pour des n_j et N_j, $j = 1, ..., k$ raisonnablement grands), la distribution d'échantillonnage et la distribution de superpopulation pour la quantité $\{g/(\hat{V}_a)^{1/2}\}$ tendraient à être identiques, c'est-à-dire à peu près $N(0, 1)$. Nous avons déjà identifié la fonction d'estimation optimale g à la fonction de quasi-caractérisation. De plus, tout comme pour un modèle paramétrique l'inversion de la distribution de la {fonction de caractérisation/(information de Fisher observée)$^{1/2}$} fournit asymptotiquement les intervalles de confiance les plus courts, pour le modèle semi-paramétrique $\{g/(\hat{V}_a)^{1/2}\}$ fournit asymptotiquement les intervalles de confiance les plus courts (Wilks 1938; Godambe et Heyde 1987).

L'analyse ci-dessus se laisse facilement prolonger de façon à inclure une covariable aléatoire. Supposons que, pour la population $\mathcal{P} = \{i : i = 1, ..., N\}$ en plus de la variable aléatoire y à l'étude, nous définissons une covariable

aléatoire x, elle aussi supposée scalaire comme y par souci de simplicité. Pour l'individu i, $x = x_i$ est connu, $i = 1, ..., N$. Le modèle de superpopulation $\varepsilon(y_i - \theta) = 0$ qui sous-tend la discussion ci-dessus s'étend alors à $\varepsilon(y_i - \theta x_i) = 0$, $i = 1,..., N$. Dans le sens de (4.1) et (4.2) nous définissons

$$\bar{X} = \sum_{j=1}^{k} N_j \bar{X}_j / N \qquad (4.9)$$

et

$$\bar{x} = \sum_{j=1}^{k} N_j \bar{x}_j / N \qquad (4.10)$$

où $\bar{X}_j$ et $\bar{x}_j$ sont les moyennes de x dans la strate $\mathcal{P}_j$ et l'échantillon s_j respectivement. Pour l'estimation de $\bar{Y}/\bar{X}$, la solution de $\sum_{i=1}^{N}(y_i - \theta x_i) = 0$, la fonction d'estimation optimale g en (4.3) est alors remplacée par

$$g = \sum_{j=1}^{k} \frac{N_j}{N} \cdot \frac{1}{n_j} \sum_{i \in s_j} \left(y_i - \frac{\bar{Y}}{\bar{X}} x_i \right). \qquad (4.11)$$

Comme auparavant, la solution de l'équation $g = 0$ fournit l'estimation optimale (ou approximativement optimale) pour $\bar{Y}$. Encore une fois, le modèle de superpopulation $\varepsilon(y_i - \theta x_i) = 0$, $i = 1, ..., N$, suggère que l'on prenne

$$\frac{\bar{Y}_j}{\bar{X}_j} = \frac{\bar{Y}}{\bar{X}}$$

lorsque la taille des strates N_j est grande, et que l'on laisse de côté les différences

$$\bar{Y}_j - \frac{\bar{Y}}{\bar{X}} \bar{X}_j, \; j = 1, ..., k.$$

Cela donne l'estimateur suivant de la variance g en (4.11) :

$$\hat{V}(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \frac{N_j}{(N_j - 1)} \frac{1}{n_j} \sum_{i \in s_j} \left(y_i - \frac{\bar{Y}}{\bar{X}} x_i \right)^2. \qquad (4.12)$$

(On trouvera à la section 7 une discussion des modèles pour lesquels on ne peut laisser de côté les différences $\bar{Y}_j - (\bar{Y}/\bar{X})\bar{X}_j$.) À noter que (4.12) se laisse réduire en (4.6) si $x_i = \text{constant}$ $i = 1, ..., N$. Encore une fois, d'après la théorie des fonctions d'estimation, on peut obtenir les intervalles de confiance pour $\bar{Y}$ en inversant la distribution d'échantillonnage du pivot (approximatif) $g/\{\hat{V}(g)\}^{1/2}$; asymptotiquement la distribution est $N(0, 1)$.

5. Échantillonnage en grappes stratifié

Dans la présente section, nous supposons que la population entière d'individus (unités) est divisée comme auparavant en strates non chevauchantes. Or maintenant, de plus, chaque strate est divisée en grappes d'individus non chevauchantes. Au premier degré d'échantillonnage, on tire de chaque strate un petit nombre de grappes par échantillonnage aléatoire simple. Ensuite, on tire de chaque grappe retenue un échantillon d'individus (ultimes), possiblement à l'aide d'un « plan de sondage » à plusieurs degrés. Ce plan de sondage est « particulier » à la « grappe » et ne dépend pas des autres grappes qui ont été retenues lors de la sélection du premier degré.

Pour intégrer la situation ci-dessus à notre cadre de travail, nous utilisons l'extension suivante de la notation précédente. Comme auparavant i représente un « individu » Une « grappe » est notée c. Les éléments de strates $\mathcal{P}_j$, $j = 1, ..., k$ sont maintenant des grappes c; la strate $\mathcal{P}_j$ est constituée de N_j grappes, $j = 1, ..., k$. Un échantillon d'individus de la grappe c est noté s^c, et l'ensemble de grappes tiré de la strate $\mathcal{P}_j$ est noté s_j, avec $|s_j| = n_j$ et $|\mathcal{P}_j| = N_j$, $j = 1, ..., k$. Autrement, nous utilisons la même notation qu'auparavant. Encore une fois, le modèle de superpopulation est $\varepsilon(y_i - \theta x_i) = 0$ pour tous les individus i de la population.

Supposons maintenant que le plan de sondage pour la grappe c est tel que (une fois la grappe tirée) la probabilité d'inclusion d'un individu $i \in c$ dans l'échantillon est π_i'. Si donc la grappe $c \in \mathcal{P}_j$, la probabilité d'inclusion inconditionnelle de i est $\pi_i'(n_j / N_j)$. Si les moyennes de la population de y et x sont notées $\bar{Y}$ et $\bar{X}$ respectivement, la fonction d'estimation optimale pour $\bar{Y}$ ou $(\bar{Y}/\bar{X})$, relativement au modèle de superpopulation mentionné ci-dessus, est donnée en remplaçant g en (4.11) par

$$g = \sum_{j=1}^{k} \frac{N_j}{n_j} \sum_{c \in s_j} \sum_{i \in s^c} \left\{ \frac{y_i \left(\frac{\bar{Y}}{\bar{X}} \right) x_i}{\pi_i'} \right\}. \qquad (5.1)$$

De plus, si $\bar{Y}_c$ et $\bar{X}_c$ représentent les moyennes des grappes de y et de x respectivement, la fonction d'estimation optimale (Godambe 1995) pour l'estimation de $\bar{Y}_c$ ou de $(\bar{Y}_c / \bar{X}_c)$ est tirée de (5.1) sous la forme

$$g_c = \sum_{i \in s^c} \left\{ \frac{\left(y_i - \frac{\bar{Y}_c}{\bar{X}_c} x_i \right)}{\pi_i'} \right\}. \qquad (5.2)$$

Nous supposons maintenant que, pour chaque grappe c, le plan de sondage est calé; autrement dit pour chaque échantillon s^c comportant une probabilité de tirage non nulle,

$$\sum_{i \in s^c} \frac{x_i}{\pi_i'} = X_c,$$

où X_c est le total de la grappe pour x. Pour ce genre de plan de sondage calé, si $\hat{Y}_c$ représente l'estimation correspondante de Y_c, le total de la grappe pour y, nous avons d'après (5.1)

$$g = \frac{1}{N}\sum_{j=1}^{k} \frac{N_j}{n_j} \sum_{c \in s_j} \left\{ \hat{Y}_c - \left(\frac{\overline{Y}}{\overline{X}} \right) X_c \right\}. \quad (5.3)$$

Il est possible de montrer, à l'aide d'une algèbre assez simple, que la variance de g en (5.3) satisfait

$$V(g) = E\left\{ \frac{1}{N^2}\sum_{j=1}^{k} \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left(\hat{Y}_c - \frac{\overline{Y}}{\overline{X}} X_c \right)^2 \right\} + O\left(\frac{1}{N} \right).$$

(Voir l'annexe.) Si donc la taille de toutes les strates N_j est assez grande nous avons

$$V(g) \simeq E\left\{ \frac{1}{N^2}\sum_{j=1}^{k} \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left(\hat{Y}_c - \frac{\overline{Y}}{\overline{X}} X_c \right)^2 \right\}. \quad (5.4)$$

Une estimation naturelle de la variance $V(g)$ en (5.4) est donnée par

$$\hat{V}(g) = \frac{1}{N^2}\sum_{j=1}^{k} \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left(\hat{Y}_c - \frac{\overline{Y}}{\overline{X}} X_c \right)^2; \quad (5.5)$$

on peut obtenir les intervalles de confiance pour $(\overline{Y}/\overline{X})$ ou $\overline{Y}$ comme auparavant en inversant la distribution d'échantillonnage du pivot approximatif $g/\sqrt{\{\hat{V}(g)\}}$; asymptotiquement la distribution est $N(0, 1)$.

Les intervalles de confiance dont il est question ci-dessus n'exigent aucune connaissance du plan de sondage pour l'une ou l'autre grappe, à la condition qu'au niveau de la grappe les estimations $\hat{Y}_c$ de Y_c soient disponibles pour $c \in s_j$, $j = 1, ..., k$. Ces intervalles de confiance, bien que valables, ne sauraient être aussi efficaces que ceux qui se fondent sur les données entières, le cas échéant.

Il importe de distinguer les estimations $\hat{V}(g)$ en (4.5), (4.12) et (5.5) (des variances $V(g)$ de la fonction d'estimation g) des estimations conventionnelles des variances des estimateurs. Généralement, celles-là comportent essentiellement le paramètre d'intérêt Y ou $\overline{Y}$. Ces dernières doivent par définition être libres du paramètre. Nous pourrions lancer l'hypothèse que la distribution de $g/\sqrt{\{\hat{V}(g)\}}$ tendrait généralement vers sa limite plus rapidement que la distribution correspondante de

$$(\hat{\overline{Y}} - \overline{Y}) / \{\text{estimation de la variance de } \hat{\overline{Y}}\}^{1/2}. \quad (5.6)$$

Car contrairement à {l'estimation de la variance de $\hat{\overline{Y}}$} en (5.6), $\hat{V}(g)$ serait la somme de variables aléatoires distribuées de façon indépendante et serait plus stable.

L'estimation $\hat{V}(g)$ en (5.5) dépend des variables aléatoires de l'échantillon uniquement par les estimations des totaux ou moyennes de grappes; cette propriété est partagée aussi par l'estimation traditionnelle de la variance de $\hat{\overline{Y}}$ en (5.6). À l'égard de cette dernière, des références initiales peuvent être retracées aux échantillons superposés de Mahalanobis dans les années 1930, tandis que l'on trouve des exemples plus récents dans Särndal, Swensson et Wretman (1992), et dans Yung et Rao (1996).

6. Estimation de la variance par la méthode *bootstrap*

Nous présentons dans la présente section les versions *bootstrap* des estimations de la variance $\hat{V}(g)$ données en (4.5), (4.12) et (5.5). Nous illustrons la méthode dans le cas de (4.12) en détail; on pourrait obtenir les estimations (4.5) et (5.5) à titre de cas spéciaux.

Notre méthode *bootstrap* est différente de la méthode habituelle en ce sens que nous obtenons la variance *bootstrap* de la fonction d'estimation g en (4.11), en gardant la valeur du paramètre $(\overline{Y}/\overline{X})$ fixe. Comme auparavant nos données sont constituées de (y_i, x_i): $i \in s_j$, $j = 1$. Le rééchantillonnage stratifié se fait comme suit : un nombre n_j de tirages est réalisé avec remise à partir de (y_i, x_i): $i \in s_j$, $j = 1, ..., k$. Si q représente un tirage générique, une valeur g_b *boostrap* générique de la fonction d'estimation g est donnée par

$$g_b = \sum_{j=1}^{k} \frac{N_j}{N} \frac{1}{n_j} \sum_{q=1}^{n_j} \left(y_q - \frac{\overline{Y}}{\overline{X}} x_q \right). \quad (6.1)$$

Si l'on note E_B et V_B l'espérance *bootstrap* et la variance respectivement, nous avons

$$E_B(g_b) = g.$$

Et

$$V_B(g_b) = E_B(g_b^2) - \{E_B(g_b)\}^2 = E_B(g_b^2) - g^2. \quad (6.2)$$

En (6.2),

$$E_B(g_b^2) = A + B + C \quad (6.3)$$

où

$$A = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{q=1}^{n_j} E_B \left(y_q - \frac{\overline{Y}}{\overline{X}} x_q \right)^2$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{i \in s_j} \left(y_i - \frac{\overline{Y}}{\overline{X}} x_i \right)^2.$$

$$B = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{\substack{q \neq q' \\ q, q' = 1 \\ \text{strate } j}}^{n_j} E_B \left(y_q - \frac{\overline{Y}}{\overline{X}} x_q \right) \left(y_q - \frac{\overline{Y}}{\overline{X}} x_{q'} \right), q', q$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} n_j (n_j - 1) \left(\overline{y}_j - \frac{\overline{Y}}{\overline{X}} \overline{x}_j \right)^2.$$

$$C = \sum_{\substack{j \neq j' \\ j, j' = 1}}^{k} \frac{N_j N_j'}{N^2} \frac{1}{n_j n_j'}$$

$$\times \sum_{\substack{q=1 \\ \text{strate } j}}^{n_j} \sum_{\substack{q'=1 \\ \text{strate } j'}}^{n_{j'}} E_B \left(y_q - \frac{\overline{Y}}{\overline{X}} x_q \right) \left(y_{q'} - \frac{\overline{Y}}{\overline{X}} x_{q'} \right),$$

$$= g^2 - \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\overline{y}_j - \frac{\overline{Y}}{\overline{X}} \overline{x}_j \right)^2.$$

Nous avons, à partir des égalités ci-dessus,

$$(B + C) = g^2 - \sum_{j=1}^{k} \frac{N_j^2}{n_j^2}\frac{1}{n_j}\left(\bar{y}_j - \frac{\bar{Y}}{\bar{X}}\bar{x}_j\right)^2.$$

Vu l'hypothèse que les variables aléatoires y_i sont tirées d'une superpopulation qui satisfait $\varepsilon(y_i - \theta x_i) = 0$, $i = 1, ..., N$, dans l'expression ci-dessus pour $(B+C)$, $\bar{y}_j - (\bar{Y}/\bar{X})\,\bar{x}_j \simeq 0(1/\sqrt{n_j})$, $j = 1, ..., k$. Par conséquent en (6.3), pour de grands n_j, $j = 1, ..., k$,

$$E_B(g_b^2) \simeq A + g^2.$$

Autrement dit en (6.2)

$$V_B(g_b) \simeq A = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\frac{1}{n_j^2}\sum_{i\in s_j}\left(y_i - \frac{\bar{Y}}{\bar{X}}x_i\right)^2 \simeq \hat{V}(g) \quad (6.4)$$

en (4.12).

On obtient l'estimation de la variance $\hat{V}_a(g)$ en (4.8) à titre de cas spécial de (6.4) lorsque $x_i = 1$, $i = 1, ..., N$. De même, on obtient l'estimation de la variance $\hat{V}(g)$ en (5.5) en remplaçant en (6.4) l'individu « i » par une grappe « c » et, de façon correspondante, en remplaçant y_i et x_i par $\hat{Y}_c$ et X_c respectivement.

7. Strates à moyennes différentes

L'optimalité des fonctions d'estimation g en (4.3), (4.11) et (5.3) dépend essentiellement de la condition de superpopulation $\varepsilon(y_i - \theta x_i) = 0$, $i = 1, ..., N$. Puisque c'est le paramètre de la population finie qui d'intérêt, l'optimalité de $\varepsilon(y_i - \theta x_i)^2$, $i = 1, ..., N$ n'est pas influencée par la variance de la superpopulation (Godambe 1995).

Dans le cas où $x_i \equiv 1$ il est intéressant de noter que l'optimalité de la fonction d'estimation

$$g = \sum_{j=1}^{k} \frac{N_j}{N}\frac{1}{n_j}\sum_{i\in s}(y_i - \bar{Y})$$

en (4.3) reste valable même lorsque le modèle de superpopulation $\varepsilon(y_i - \theta) = 0$, $i \in \mathcal{P}$ est remplacé par le modèle étendu $\varepsilon(y_i - \theta_j) = 0$, $i \in \mathcal{P}_j$, $j = 1, ..., k$. C'est-à-dire que maintenant εy_i peut varier d'une strate à l'autre (Godambe 1995). L'optimalité restera valable parce que $\sum_{j=1}^{k}(N_j/N)(\theta_j - \varepsilon\bar{Y}) = 0$. Toutefois, on ne peut plus approcher de la variance de g en remplaçant la moyenne de la strate $\bar{Y}_j$ par la moyenne de la population $\bar{Y}$ en (4.4). L'approximation antérieure et l'estimation subséquente $\hat{V}(g)$ en (4.6) se fondaient sur l'hypothèse que (pour de grandes strates) la différence $\bar{Y}_j - \theta$ ou $\bar{Y}_j - \bar{Y}$, $j = 1, ..., k$ pouvait être laissée de côté. Vu le remplacement de θ dans la strate $\mathcal{P}_j$ par θ_j, les termes $\bar{Y}_j - \bar{Y}$ ne peuvent plus être laissés de côté, $j = 1, ..., k$. Notons ici que le fait de stratifier la population de façon à donner à chaque strate une homogénéité « interne » tend à élargir les différences $\bar{Y}_j - \bar{Y}$, $j = 1, ..., k$. Le titre de cette section reflète justement cette situation.

Afin d'obtenir une estimation $\hat{V}(g)$, dans le cadre du modèle étendu $\varepsilon(y_i - \theta_j) = 0$, nous notons que

$$g = \sum_{j=1}^{k} \frac{N_j}{N}\frac{1}{n_j}\sum_{i\in s}(y_i - \bar{Y}_j)$$

et nous cherchons à estimer les paramètres dérangeants ou $\bar{Y}_j$, $j = 1, ..., k$, en maintenant $\bar{Y}$ fixe. À noter que ce problème d'estimation est tout à fait différent, conceptuellement aussi bien que mathématiquement, de celui de l'estimation de la variance de $\bar{y}$ en (4.4). La procédure est celle qui a été établie à la section 2, où $\psi(\boldsymbol{y})$ est $\bar{Y}$ et $\theta(\boldsymbol{y})$ est constitué de $\bar{Y}_j$, $j = 1, ..., k$.

Le problème de l'estimation de $\bar{Y}_j$, $j = 1, ..., k$, sous réserve que la moyenne de la population $\bar{Y}$ reste fixe, se résout à l'aide de la technique habituelle du multiplicateur de Lagrange. Nous adaptons comme hypothèse de travail que les variances de modèle, c'est-à-dire $\varepsilon(y_i - \theta_j)^2 = \sigma_i^2$, sont constantes (σ^2) pour tous les $i \in \mathcal{P}$. Pour des variations de $\bar{Y}_j$, $j = 1, ..., k$, nous trouvons un point critique de la fonction

$$\varphi = \sum_{j=1}^{k}\sum_{i\in s_j}(y_i - \bar{Y}_j)^2 - \lambda\left\{\left(\sum_{j=1}^{k}\frac{N_j\bar{Y}_j}{N}\right) - \bar{Y}\right\}, \quad (7.1)$$

où λ est le multiplicateur de Lagrange. Cette technique d'estimation est intuitivement attrayante indépendamment du modèle de superpopulation mentionné ci-dessus. Il est facile de vérifier que (7.1) est minimisé pour l'estimation $\hat{\bar{Y}}_j$ de $\bar{Y}_j$ où

$$\hat{\bar{Y}}_j = \bar{y}_j - \frac{N_j/(n_j N)}{\sum_{j=1}^{k}[N_j^2/(n_j N^2)]}(\bar{y} - \bar{Y}),\ j = 1,...,k, \quad (7.2)$$

n_j étant comme auparavant la taille des échantillons de la strate j, $j = 1, ..., k$. À noter que, lorsque la taille des strates N_j et la taille des échantillons n_j sont « proportionnelles », c'est-à-dire que $(n_j/N_j) = (n/N)$, $j = 1, ..., k$, les équations (7.2) se réduisent à

$$\hat{\bar{Y}}_j = \bar{y}_j - (\bar{y} - \bar{Y}),\ j = 1, ..., k. \quad (7.3)$$

Cette simple relation peut également servir lorsque la taille des strates et la taille des échantillons ne sont pas exactement proportionnelles, mais seulement de façon approximative. À noter, relativement à (7.3), que la fonction d'estimation $(\bar{Y}_j - \bar{y}_j) - (\bar{Y} - \bar{y})$ est sans biais pour ce qui est du plan.

La discussion ci-dessus suggère également l'estimation des moyennes de strates $\bar{Y}_j$ lorsqu'il existe une covariable aléatoire x non constante. Le modèle de superpopulation qui sous-tend la fonction d'estimation g en (4.11), comme nous l'avons noté, était $\varepsilon(y_i - \theta x_i) = 0$, pour tous les individus $i \in \mathcal{P}$, avec un paramètre commun θ. Supposons que ce modèle

doive être remplacé par un modèle plus souple et réaliste et que le paramètre θ puisse varier d'une strate à l'autre. C'est-à-dire que maintenant $\varepsilon(y_i - \theta_j x_i) = 0$, $i \in \mathcal{P}_j$, $i \in \mathcal{P}_j$, $\mathcal{P}_j$ désignant comme auparavant la strate j, $j = 1, ..., k$. Comme en (4), les moyennes de strates $\bar{Y}_j$ sont incorporées dans la variance $V(g)$ de la fonction d'estimation g en (4.11) :

$$V(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right) \frac{1}{N_j - 1} \sum_{i \in \mathcal{P}_j} \left\{(y_i - \bar{Y}_j) - \frac{\bar{Y}}{\bar{X}}(x_i - \bar{X}_j)\right\}^2, \quad (7.4)$$

$\bar{X}_j$, $j = 1, ..., k$ désignant comme auparavant les moyennes de strates des x. Nous avons obtenu l'estimation $\hat{V}(g)$ en (4.12) en laissant de côté les termes $\bar{Y}_j = \bar{Y}/\bar{X}\,\bar{X}_j$ en supposant une grande taille pour les strates N_j et le modèle de superpopulation $\varepsilon(y_i - \theta x_i) = 0$ pour tous les individus $i \in \mathcal{P}$, comportant un paramètre « commun » θ. En présence du nouveau modèle plus souple $\varepsilon(y_i - \theta_j x_i) = 0$, $i \in \mathcal{P}_j$, $j = 1, ..., k$, les termes $\bar{Y}_j = \bar{Y}/\bar{X}\,\bar{X}_j$ ne peuvent plus être laissés de côté. L'estimation appropriée, c'est-à-dire $\hat{V}(g)$, de la variance $V(g)$ en (7.4) est donnée par

$$\hat{V}(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right) \frac{N_j}{N_j - 1} \frac{1}{n_j} \sum_{i \in s_j} \left\{(y_i - \bar{Y}_j) - \frac{\bar{Y}}{\bar{X}}(x_i - \bar{X}_j)\right\}^2. \quad (7.5)$$

Toutefois, les intervalles de confiance habituels pour $(\bar{Y}/\bar{X})$ obtenus en inversant la distribution du pivot approximatif $\{g/\sqrt{\{\hat{V}(g)\}}\}$ comportent désormais des paramètres dérangeants $\bar{Y}_j$, $j = 1, ..., k$, à supposer que les moyennes de strates des covariables aléatoires $\bar{X}_j$, $j = 1, ..., k$, soient connues.

Comme auparavant, nous devons estimer les paramètres dérangeants, c'est-à-dire les moyennes de strates $\bar{Y}_j$, $j = 1, ..., k$, pour une moyenne de la population $\bar{Y}$ fixe. À noter que le modèle de superpopulation sous-jacent précise $\varepsilon(y_i - \theta_j x_i) = 0$, $i \in \mathcal{P}_j$, $j = 1, ..., k$. En notant de plus les variances de superpopulation $\varepsilon(y_i - \theta_j x_i)^2 = \sigma_i^2$, $i \in \mathcal{P}_j$, et en supposant comme auparavant que la taille des strates $|\mathcal{P}_j| = N_j$, $j = 1, ..., k$ est grande, nous remplaçons la fonction φ en (7.1) par

$$\psi = \sum_{j=1}^{k} \sum_{i \in s_j} (\sigma_i^2)^{-1}\left(y_i - \frac{\bar{Y}_j}{\bar{X}_j} x_i\right)^2 - \lambda\left\{\left(\sum_{j=1}^{k} \frac{\bar{Y}_j}{\bar{X}_j} N_j \bar{X}_j\right) - N\bar{Y}\right\}, \quad (7.6)$$

λ étant comme auparavant le multiplicateur de Lagrange. En formulant ψ en (7.6) nous adoptons comme hypothèse de travail que pour le modèle de superpopulation comportant $\varepsilon(y_i - \theta_j x_i) = 0$, les fonctions de variance $\varepsilon(y_i - \theta_j x_i)^2 = \sigma_i^2 = \sigma^2 x_i$, $i \in \mathcal{P}$. Autrement dit, σ_i^2 est proportionnel à la valeur de la covariable aléatoire x_i, $i \in \mathcal{P}$. Comme il a été mentionné au début de la présente section, l'hypothèse de travail que nous venons de décrire est adoptée par souci de simplicité surtout et ne comporte aucune conséquence statistique importante (Godambe 1995). Il est facile de vérifier que les valeurs (estimations) $\hat{\bar{Y}}_j$ de $\bar{Y}_j$ optimisées en (7.6) sont données par

$$\frac{\bar{y}_j}{\bar{x}_j} - \frac{\hat{\bar{Y}}_j}{\bar{X}_j} = \left[\left\{\left(\sum_{j=1}^{k} N_j \bar{X}_j \frac{\bar{y}_j}{\bar{x}_j}\right) - N\bar{Y}\right\} \Big/ \left\{\sum_{j=1}^{k} \frac{(N_j \bar{X}_j)^2}{2n_j \bar{x}_j}\right\}\right] \frac{N_j \bar{X}_j}{2n_j \bar{x}_j}, \quad (7.7)$$

$j = 1, ..., k$.

Pour ce qui est de la fonction d'estimation g en (4.3), la variance $V(g)$ est donnée par (4.4). De plus, si $\hat{V}_1(g)$ est l'estimation de $V(g)$ fondée sur les estimations $\hat{\bar{Y}}_j$ de $\bar{Y}_j$, $j = 1, ..., k$ donné par (7.3), dès lors, par analogie avec (4.5) (mais en tenant compte du fait que, puisque $\bar{Y}_j$ est estimé, la somme des carrés comporte moins de n_j degrés de liberté) nous avons

$$\hat{V}_1(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right) \cdot \frac{1}{n_j - 1} \cdot \sum_{i \in s_j} \{y_i - (\bar{y}_j - \bar{y} + \bar{Y})\}^2. \quad (7.8)$$

Les intervalles de confiance pour $\bar{Y}$ sont obtenus grâce à l'inversion de la distribution du pivot approximatif $[g/\{\hat{V}_1(g)\}^{1/2}]$; asymptotiquement

$$g/\{\hat{V}_1(g)\}^{1/2} \sim N(0, 1). \quad (7.9)$$

De même, dans le cas d'une covariable aléatoire, pour une fonction d'estimation g en (4.11), si $\hat{V}_2(g)$ désigne l'estimation de la variance $V(g)$ en (7.4) d'après les estimations $\hat{\bar{Y}}_j$ données par (7.7), dès lors

$$\hat{V}_2(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right) \frac{1}{(n_j - 1)} \sum_{i \in s_j} \left\{\left(y_i - \frac{\bar{y}_j}{\bar{x}_j}\bar{X}_j + A_j\right) - \frac{\bar{Y}}{\bar{X}}(x_i - \bar{X}_j)\right\}^2$$

où

$$A_j = w_j' N(\bar{y}_R - \bar{Y})/N_j,$$

$$w_j' = [N_j^2 \bar{X}_j^2/n_j\bar{x}_j] \Big/ \sum_{j=1}^{k} [N_j^2 \bar{X}_j^2/n_j\bar{x}_j],$$

$$\bar{y}_R = \sum_{j=1}^{k} (N_j/N)\bar{X}_j \bar{y}_j/\bar{x}_j.$$

Toutefois, une forme moins compliquée qui demeure une fonction de $\bar{Y}_j$ uniquement par l'entremise de $\bar{Y}/\bar{X}$ est

$$\hat{V}_2(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{1}{(n_j - 1)} \sum_{i \in s_j} \left\{\left(y_i - \frac{\bar{y}_j}{\bar{x}_j}\bar{X}_j\right) - \frac{\bar{Y}}{\bar{X}}(x_i - \bar{X}_j)\right\}^2. \quad (7.10)$$

Encore une fois, les intervalles de confiance pour $\bar{Y}$ se fondent sur l'inversion de la distribution de $[g/\{V_2(g)\}^{1/2}]$; asymptotiquement

$$g / \{\hat{V}_2(g)\}^{1/2} \sim N(0, 1). \quad (7.11)$$

8. Propriétés empiriques

À la section précédente, nous avons décrit la construction d'intervalles de confiance lorsque le modèle de superpopulation $\varepsilon(y_i - \theta) = 0$ ou $\varepsilon(y_i - \theta x_i) = 0$, avec une valeur « commune » de θ pour tous les individus $i \in \mathcal{P}$, est remplacé par le modèle $\varepsilon(y_i - \theta_j) = 0$ ou $\varepsilon(y_i - \theta_j x_i) = 0$, $i \in \mathcal{P}_j$, $j = 1, ..., k$. Autrement dit, θ peut maintenant varier d'une strate à l'autre. Généralement, dans la pratique, nous ne pouvons savoir avec certitude si, pour la population observée en question, le paramètre θ a une valeur « commune » pour tous les individus $i \in \mathcal{P}$. Des considérations théoriques aussi bien que numériques indiquent clairement que le rendement des intervalles de confiance calculés en fonction de l'hypothèse d'une valeur commune de θ (par exemple, ceux qui se fondent sur les pivots $[g/\{\hat{V}(g)\}^{1/2}]$ de la section 4) est très sensible même à de « petits écarts » de θ, d'une strate à l'autre. Bien entendu, il arrive typiquement que, lorsque l'on stratifie une population, une évaluation préalable des valeurs moyennes θ pour différents individus entraîne la construction de strates $\mathcal{P}_j$ comportant des valeurs moyennes différentes θ_j, $j = 1, ..., k$.

Pour les raisons décrites ci-dessus, nous proposons le recours général à des intervalles de confiance fondés sur le pivot (7.9), lorsqu'il n'existe pas de covariable aléatoire, et à des intervalles de confiance fondés sur le pivot (7.11) dans le cas d'une covariable aléatoire. Dans les exemples qui suivent, les intervalles de confiance ci-dessus sont comparés à des intervalles de confiance conventionnels : ceux-ci sont obtenus, en l'absence d'une covariable aléatoire, du pivot $N(0, 1)$ approximatif

$$\frac{\left\{\sum_{j=1}^{k} \frac{N_j}{N}\frac{1}{n_j}\sum_{i \in s_j}(y_i - \bar{Y})\right\}}{\left\{\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{1}{(n_j - 1)}\sum_{i \in s_j}(y_i - \bar{y}_j)^2\right\}^{1/2}}; \quad (8.1)$$

en présence d'une covariable aléatoire, le pivot $N(0, 1)$ approximatif est

$$\frac{\sum_{j=1}^{k} \frac{N_j}{N} \cdot \frac{1}{n_j}\sum_{i \in s_j}\left(y_i - \frac{\bar{Y}}{\bar{X}}x_i\right)}{\left[\sum_{j=1}^{k} \frac{N_j^2}{N^2}\cdot\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{1}{n_j - 1}\sum_{i \in s_j}\left\{(y_i - \bar{y}_j) - \frac{\bar{y}}{\bar{x}}(x_i - \bar{x}_j)\right\}^2\right]^{1/2}}. \quad (8.2)$$

(Cochran 1977). En général, nous parlons de (7.9) et (7.11) comme des nouveaux pivots et de (8.1) et (8.2) comme des pivots conventionnels.

Des expériences de simulation poussées ont été menées afin de comparer les intervalles de confiance fondés sur les pivots nouveaux et conventionnels. Toutefois, les résultats décrits ci-dessous se rapportent surtout à de petits échantillons. Nous avons ici 16 populations observées dont chacune est divisée en quatre strates, des échantillons de taille 2, 3, 4, 2 étant tirés des strates respectives. De tels échantillons de taille (totale) aussi faible que 11 font le mieux ressortir, comme dans les tableaux 1 et 2 qui suivent, le rendement supérieur des nouveaux pivots relativement aux pivots conventionnels. À un degré moindre que pour les échantillons de petite taille que nous venons de mentionner, la supériorité des nouveaux pivots relativement aux pivots conventionnels reste valable pour des échantillons de taille moyenne (25), comme dans les tableaux 3 et 4. Nos études de simulation non diffusées comprenaient des populations divisées en 16 strates chacune, avec une taille d'échantillon totale de 50 environ. Même pour des échantillons aussi grands, les nouveaux pivots semblent donner un meilleur rendement que les pivots conventionnels. Tôt ou tard, bien sûr, pour des échantillons de taille très grande, la distinction entre les deux pivots, les nouveaux et les conventionnels, tend à disparaître pour ce qui est du rendement.

Les seize populations observées (1) – (16) des tableaux 1 et 2 ci-dessous, exception faite pour les populations (7), (8), comportent chacune une taille de 1 000; les populations (7), (8) comportent chacune une taille de 2 000. Comme il a été mentionné, chacune des seize populations est divisée en quatre strates. Les tableaux 1 et 2 comportent chacun six colonnes (i), (ii), (vi). La colonne (i) indique le numéro de la population (·). La colonne (ii) indique, selon les quatre strates de la population (·), les distributions de la superpopulation dont sont tirées les strates. La distribution peut être de type chi carré (*C*), normal (*N*) ou uniforme (*U*). Lorsqu'il n'existe pas de covariable aléatoire comme dans le tableau 1, la colonne (ii) indique simplement la distribution de la variable aléatoire *y*; dans le tableau 2, par contre, elle indique la distribution de la variable aléatoire *y* aussi bien que de la covariable aléatoire *x*, la moyenne de *y* (dépendant de *x*) étant θx. La colonne (iii) décrit la taille de l'échantillon pour les différentes strates. La colonne (iv) donne la probabilité de couverture nominale. Les colonnes (v) et (vi) fournissent la probabilité de couverture réelle et la longueur moyenne des intervalles de confiance, pour 4 000 simulations. Ainsi, une ligne horizontale type du tableau 1, pour (6) par exemple, se lirait comme suit. Les

quatre strates de la population (6) sont tirées, respectivement, de distributions de la superpopulation normale, chi carré, normale, chi carré; la taille des échantillons des différentes strates sont (2, 3, 4, 2) respectivement. L'interprétation des colonnes (iv), (v), (vi) est simple. Contrairement aux populations (1) – (16) ci-dessus, les populations (17) et (18) des tableaux 3 et 4 sont divisées en huit strates chacune, la population (17) ne comportant pas de covariable aléatoire et (18) en comportant une.

Tableau 1

(i) Population	(ii) Distribution de la superpopulation y	(iii) Taille des échantillons	(iv) Probabilité de couverture nominale	(v) Probabilité de couverture réelle		(vi) Longueur moyenne	
				pivot (7.9)	pivot (8.1)	pivot (7.9)	pivot (8.1)
(1)	{*N*,*C*,*U*,*C*}	(2,3,4,2)	0,95	0,967	0,86	19,83	11,33
(2)	{*N*,*C*,*U*,*C*}	(2,3,4,2)	0,90	0,90	0,80	13,11	9,51
(3)	{*N*,*N*,*N*,*N*}	(2,3,4,2)	0,90	0,90	0,807	4,85	3,52
(4)	{*U*,*U*,*U*,*U*}	(2,3,4,2)	0,90	0,90	0,817	4,90	3,55
(5)	{*N*,*C*,*N*,*C*}	(2,3,4,2)	0,95	0,946	0,82	34,34	19,62
(6)	{*N*,*C*,*N*,*C*}	(2,3,4,2)	0,90	0,866	0,76	22,71	16,46
(7)	{*N*,*U*,*C*,*N*}	(2,3,4,2)	0,95	0,97	0,869	20,76	11,80
(8)	{*N*,*U*,*C*,*N*}	(2,3,4,2)	0,90	0,908	0,81	13,69	9,96

Pour les populations numérotées de (1) à (4) ci-dessous, la valeur moyenne θ reste fixe d'une strate à l'autre, $\theta = 100$, l'écart-type varie entre 2,0 et $\sqrt{200.00}$. Pour les autres populations, (5) à (8), la valeur moyenne θ varie d'une strate à l'autre, entre $\theta = 100$ et $\theta = 400$

Tableau 2

(i) Population	(ii) Distribution de la superpopulation		(iii) Taille des échantillons	(iv) Probabilité de couverture nominale	(v) Probabilité de couverture réelle		(vi) Longueur moyenne	
	x	y			pivot (7.11)	pivot (8.2)	pivot (7.11)	pivot (8.2)
(9)	{*U*,*U*,*U*,*U*}	{*C*,*C*,*C*,*C*}	(2,3,4,2)	0,90	0,879	0,82	91,12	54,12
(10)	{*U*,*U*,*U*,*U*}	{*C*,*C*,*C*,*C*}	(2,3,4,2)	0,95	0,926	0,876	113,49	64,49
(11)	{*U*,*U*,*U*,*U*}	{*N*,*N*,*N*,*N*}	(2,3,4,2)	0,90	0,88	0,84	12,21	6,85
(12)	{*C*,*C*,*C*,*C*}	{*N*,*N*,*N*,*N*}	(2,3,4,2)	0,90	0,83	0,83	7,07	6,84
(13)	{*C*,*C*,*C*,*C*}	{*C*,*C*,*C*,*C*}	(2,3,4,2)	0,95	0,926	0,87	113,53	100,10
(14)	{*C*,*C*,*C*,*C*}	{*C*,*C*,*C*,*C*}	(2,3,4,2)	0,90	0,869	0,84	35,89	33,01
(15)	{*C*,*C*,*C*,*C*}	{*C*,*C*,*C*,*C*}	(2,3,4,2)	0,95	0,92	0,89	42,88	39,34
(16)	{*C*,*C*,*C*,*C*}	{*U*,*C*,*C*,*U*}	(2,3,4,2)	0,95	0,959	0,909	31,17	26,68

Pour les populations numérotées (9) à (12) ci-dessous, le coefficient de régression θ reste fixe pour toutes les strates, $\theta = 3$. Pour les autres populations (13) - (15), le coefficient de régression θ varie d'une strate à l'autre entre $\theta = 2$ et $\theta = 4$

Tableau 3

(i) Population	(ii) Distribution de la superpopulation	(iii) Taille des échantillons	(iv) Couverture nominale	(v) Probabilité de couverture réelle		(vi) Longueur moyenne	
				pivot (7.9)	pivot (8.1)	pivot (7.9)	pivot (8.1)
(17)	{*N*,*U*,*C*,*U*,*C*,*N*,*U*,*U*}	(2,3,4,2,3,4,3,4)	0,95	0,93	0,889	12,76	10,94

Pour la population (17) ci-dessous, la valeur moyenne θ varie d'une strate à l'autre entre $\theta = 100$ et $\theta = 800$

Tableau 4

(i) Population	(ii) Distribution de la superpopulation	(iii) Taille des échantillons	(iv) Couverture nominale	(v) Probabilité de couverture réelle		(vi) Longueur moyenne	
				pivot (7.11)	pivot (8.2)	pivot (7.11)	pivot (8.2)
(18)	x : {*C*,*C*,*C*,*C*,*C*,*C*,*C*,*C*} y : {*N*,*U*,*N*,*C*,*C*,*C*,*C*,*N*}	(2,3,4,2,3,4,3,4)	0,95	0,937	0,90	22,8	20,89

Pour la population (18) ci-dessous, le coefficient de régression θ varie d'une strate à l'autre entre $\theta = 3$ et $\theta = 6$

9. Conclusions

Les conclusions ci-dessous se fondent sur les considérations théoriques des sections précédentes et les résultats de simulation décrits à la section 8, de même que sur de nombreux autres résultats de simulation, comme il a été mentionné antérieurement, mais ne figurant pas dans le présent exposé.

La situation qui se présente lorsqu'il n'existe pas de covariable aléatoire semble assez claire d'après les tableaux 1 et 3 de la section 8. Pour de petits échantillons, les intervalles de confiance conventionnels, c'est-à-dire ceux qui se fondent sur le pivot (8.1), peuvent être très trompeurs : la probabilité de couverture « déclarée » peut être très différente de la probabilité de couverture « réelle ». Par ailleurs, cet écart entre les probabilités de couverture déclarée et réelle, pour ce qui est des intervalles de confiance conventionnels, semble s'accroître à mesure que la variation des moyennes de strates augmente. Il est intéressant de constater, comme nous l'avons fait remarquer à la section 7, que cette variation accrue des moyennes de strates peut souvent résulter d'une stratification de la population en strates comportant une homogénéité (interne) pour assurer une estimation ponctuelle efficace. Les intervalles de confiance fondés sur le nouveau pivot (7.9), comme nous pouvons le constater dans les tableaux 1 et 3 de la section 8, ont un bien meilleur rendement que ceux qui se fondent sur le pivot conventionnel (8.1). D'après nos simulations fondées sur trois distributions, c'est-à-dire normale, chi carré et uniforme, il semble que la comparaison des rendements du nouveau pivot (7.9) et du pivot conventionnel (8.1) dépend des distributions surtout pour ce qui est de la variation de leurs valeurs moyennes d'une strate à l'autre. Plus précisément, la comparaison n'est pas très touchée par la variance ni la forme des distributions. On pouvait s'y attendre compte tenu de notre modèle semi-paramétrique sous-jacent, $\varepsilon(y_i - \theta_j) = 0,\ i \in \mathcal{P}_j,\ j = 1, ..., k$. Cela vient prolonger la conclusion déjà tirée au début de la section 8. Nous tenons à souligner ici que l'optimalité de la fonction d'estimation g en (4.3) reste valable même lorsque θ varie d'une strate à l'autre.

Pour de grands échantillons, d'après les résultats de nos simulations mentionnés antérieurement (mais non décrits ici), la différence entre les deux ensembles d'intervalles de confiance, l'un fondé sur le pivot (7.9) et l'autre sur le pivot (8.1), tend à diminuer. Cela aussi est conforme à la théorie.

Les tableaux 2 et 4 de la section 8 fournissent des résultats au sujet d'intervalles de confiance pour des populations pouvant comporter une covariable aléatoire. La comparaison du rendement du nouveau pivot (7.11) et du pivot conventionnel (8.2) est alors assez subtile. Nous envisageons deux situations : la première pour un coefficient de régression θ identique pour toutes les strates, la deuxième pour θ qui varie (mais pas beaucoup) d'une strate à l'autre. La fonction d'estimation g en (4.11) n'est optimale que dans la première situation. Dans cette situation (c'est-à-dire pour θ identique pour toutes les strates), qui offre un intérêt surtout théorique, le pivot présenté à la fin de la section 4, comme l'indiquent nos études de simulation (non présentées ici), fonctionne très bien. La deuxième situation, elle, est plus réaliste. Concrètement, donc, il est très important d'étudier le rendement de la fonction d'estimation g, c'est-à-dire le rendement des intervalles de confiance fondés sur le nouveau pivot (7.11), lorsque θ varie d'une strate à l'autre. Dans une telle situation, il est clair, d'après les tableaux 2 et 4, que les intervalles de confiance fondés sur le nouveau pivot (7.11) fournissent des probabilités de couverture « réelles » plus proches des probabilités de couverture « déclarées » que les intervalles de confiance fondés sur le pivot conventionnel (8.2). De plus, dans la première situation, c'est-à-dire pour θ identique pour toutes les strates, comme l'indique le tableau 2, le rendement du nouveau pivot (7.11) est au moins égal à celui du pivot conventionnel (8.2). Le phénomène semble encore plus remarquable à mesure que la variation des valeurs de la covariable aléatoire prend de l'ampleur. En réalité, puisque les valeurs de la covariable aléatoire au sein de chaque strate tendent à l'uniformité, la différence de rendement entre les deux pivots, le nouveau (7.11) et le conventionnel (8.2), tend à diminuer. La différence diminue également à mesure que la taille de l'échantillon augmente.

Un examinateur a suggéré que, puisque dans le cas d'une strate unique notre nouveau pivot (7.11) ressemble étroitement au pivot de Fieller, dont il est question dans Cochran (1977), nous abordions les observatins de Cochran : pour certaines distributions paramétriques (par exemple, la normalité à deux variables de (x, y), une des moyennes se rapprochant de zéro), les intervalles de confiance pour un ratio de moyennes fondé sur le pivot de Fieller risque d'avoir des propriétés indésirables (pour ce qui est de la probabilité de couverture et de la longueur des intervalles) comparativement aux intervalles de confiance fondés sur le pivot conventionnel (8.2). Dans le contexte d'un sondage, de telles circonstances seraient exceptionnelles, comme l'indiquent les résultats de nos simulations. De plus, notre nouveau pivot (7.11) est « valide » pour un modèle semi-paramétrique; une grande classe de modèles paramétriques sous-tend cette situation.

Remerciements

Nous tenons à remercier A.C. Singh de ses précieuses remarques au sujet d'une version antérieure du présent exposé. Nous remercions également Jiahua Chen et Lianxiang Wang de leurs conseils et de leur assistance pour les calculs.

Annexe

La variance de la fonction d'estimation g en (5.3) c'est-à-dire $V(g)$, est $E(g^2)$ puisque $E(g) = 0$.
De plus

$$E(g^2) = E\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{\sum_{c\in s_j}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)\Big/ n_j\right\}^2 + \sum_{\substack{j\neq j'\\ j,j'=1}} \frac{N_j N_{j'}}{N^2} E\left[\left\{\sum_{c\in s_j}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)\Big/ n_j\right\}\left\{\sum_{c\in s_{j'}}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)\Big/ n_j\right\}\right], = A + B. \quad \text{(I)}$$

Donc

$$A = \sum_{j=1}^{k} \frac{N_j^2}{N^2} E\left\{\sum_{c\in s_j}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)^2 \Big/ n_j^2\right\} + \sum_{j=1}^{k} \frac{N_j^2}{N^2} E\left\{\sum_{\substack{c\neq c'\\ c,c'\in s_j}} E_{c,c'}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)\left(\hat{Y}_{c'} - \frac{Y}{X}X_{c'}\right)\Big/ n_j^2\right\}, \quad \text{(II)}$$

où $E_{c,c'}$ représente l'espérance qui maintient les grappes c, c' fixes. Or comme nous l'avons signalé au début de la section 5, les plans de sondage pour différentes grappes sont indépendants. De plus, le deuxième terme du second membre de « A » est égal à

$$\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{E\left[\frac{1}{n_j}\sum_{c\in s_j}\left(Y_c - \frac{Y}{X}X_c\right)^2\right] - E\left[\frac{1}{n_j^2}\sum_{c\in s_j}\left(Y_c - \frac{Y}{X}X_c\right)^2\right]\right\}$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{1}{N_j - 1}\sum_{c\in\mathcal{P}_j}\left[(Y_c - \bar{Y}_j) - \frac{Y}{X}(X_c - \bar{X}_j)\right]^2 + \left(\bar{Y}_j - \frac{Y}{X}\bar{X}_j\right)^2 - \frac{1}{n_j N_j}\sum_{c\in\mathcal{P}_j}\left(Y_c - \frac{Y}{X}X_c\right)^2\right\}$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{\left[\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{1}{N_j - 1}\sum_{c\in\mathcal{P}_j}\left(Y_c - \frac{Y}{X}X_c\right)^2\right] + \frac{N_j(n_j - 1)}{n_j(N_j - 1)}\left(\bar{Y}_j - \frac{Y}{X}\bar{X}_j\right)^2 - \frac{1}{n_j N_j}\sum_{c\in\mathcal{P}_j}\left(Y_c - \frac{Y}{X}X_c\right)^2\right\}$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{\frac{N_j(n_j - 1)}{n_j(N_j - 1)}\left(\bar{Y}_j - \frac{Y}{X}\bar{X}_j\right)^2 - \frac{n_j - 1}{N_j(N_j - 1)}\frac{1}{n_j}\cdot\sum_{c\in\mathcal{P}_j}\left(Y_c - \frac{Y}{X}X_c\right)^2\right\} \quad \text{(III)}$$

où $\bar{Y}_j = \sum_{c\in\mathcal{P}_j} Y_c/N_j$ et $\bar{X}_j = \sum_{c\in\mathcal{P}_j} X_c/N_j$, $j = 1, ..., k$. Le terme B en $E(g^2)$ se simplifie en

$$B = \left[\sum_{j=1}^{k} \frac{N_j}{N}\left(\bar{Y}_j - \frac{Y}{X}\bar{X}_j\right)\right]^2 - \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\bar{Y}_j - \frac{Y}{X}\bar{X}_j\right)^2 = -\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\bar{Y}_j - \frac{Y}{X}\bar{X}_j\right)^2. \quad \text{(IV)}$$

À noter que, à cause du modèle de superpopulation $\varepsilon(y_i - \theta x_i) = 0$, $i \in \mathcal{P}$ en (III) et (IV) le terme

$$\left(\bar{Y}_j - \frac{Y}{X}\bar{X}_j\right)^2 \text{ est } O\left(\frac{1}{N_j}\right),\ j = 1, ..., k.$$

On peut donc montrer que le deuxième terme de A et le terme B sont tous deux d'ordre $O(1/N)$, tandis que le premier terme de A est d'ordre $O(N^2/n^3)$. Par conséquent, à partir de (I) – (IV) pour des strates de grande taille N_j, $j = 1, ..., k$ nous avons des approximations de la variance $V(g)$ et de son estimation $\hat{V}(g)$ comme en (5.4) et (5.5). Ces approximations sont valides, peu importe le modèle de superpopulation qui vient d'être mentionné, à la condition que la taille des échantillons n_j ainsi que la taille des strates N_j, $j = 1, ..., k$ soient suffisamment grandes.

Bibliographie

Basu, D. (1958). On sampling with and without replacement. *Sankhyā*, 20, 287-294.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D.A., et Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 39, 1035-1043.

Chaudhuri, A., et Vos, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. Amsterdam : North Holland.

Cochran, W.G. (1977). *Sampling Techniques*, (3ième Éd.). New York : John Wiley & Sons, Inc.

Deming, W.E. (1950). *Some Theory of Sampling*. New York : John Wiley & Sons, Inc.

Efron, B., et Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, (avec discussion). *Biometrika*, 65, 457-487.

Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700-725.

Godambe, V.P. (1976). A historical perspective of recent developments in the theory of sampling from actual populations. *Journal of the Indian Society of Agricultural Statistics*, 28, 1-12.

Godambe, V.P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika*, 72, 419-428.

Godambe, V.P. (1991). Orthogonality of estimating functions and nuisance parameters. *Biometrika,* 78, 143-151.

Godambe, V.P. (1995). Estimation of parameters in survey sampling: Optimality. *La revue canadienne de statistique*, 23, 227-243.

Godambe, V.P. (1997). Estimation of Parameters in Survey Sampling. *Recueil 1996 de la Section des méthodes d'enquête, Société Statistique du Canada*.

Godambe, V.P., et Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *Revue Internationale de Statistique*, 55, 231-244.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Godambe, V.P., et Thompson, M.E. (1989). An Extension of quasi-likelihood estimation, (avec discussion). *Journal of Statistical Planning and Inference*, 22, 137-172.

Hájek, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pro Péstování Matematiky*, 84, 387-423.

Mach, L. (1988). The Use of Estimating Functions for Confidence Interval Construction: The Case of the Population Mean. Document de travail No. BSMD-88-028 E, Direction de la méthodologie, Statistique Canada.

Neyman, J. (1934). On two different aspects of representative method: The method of stratified sampling and the method of purposive selection, (avec discussion). *Journal of the Royal Statistical Society*, 97, 558-652.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of Royal Society*, Séries A, 236, 333-380.

Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *Revue Internationale de Statistique*, 54, 221-226.

Särndal, C.-E., Swensson, B., et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Smith, T.M.F. (1997). Social surveys and social science, (avec discussion). *La revue canadienne de statistique*, 25, 23-44.

Vinod, H.D. (1988). Foundations of statistical inference based on numerical roots of robust pivot functions. *Journal of Econometrics*, 81, 387-396.

Wilks, S.S. (1938). Shortest average confidence intervals from large samples. *Annals of Mathematical Statistics*, 9, 166-175.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Yung, W., et Rao, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrès multiples. *Techniques d'enquête*, 22, 23-31.