Techniques d'enquête, juin 1985
Vol. 11, no. 1, pp. 17–35
Statistique Canada

Inférence conditionnelle dans les enquêtes par sondage

J.N.K. RAO[1]

RÉSUMÉ

L'auteur jette un regard critique sur les méthodes traditionnelles d'inférence dans les enquêtes par sondage. Il souligne la nécessité de faire reposer l'inférence sur des sous-ensembles identifiables de la population. Utilisant des échantillons aléatoires de tailles diverses, il apporte un certain nombre d'exemples concrets d'inférences dépendant de la configuration réelle de l'échantillon ainsi que les problèmes qui peuvent y être associés. Parmi les exemples choisis, mentionnons l'estimation d'une moyenne de population suivant un échantillonnage aléatoire simple, l'estimation d'une moyenne de population en présence d'observations aberrantes, l'estimation du total et de la moyenne d'un domaine, l'estimation de la moyenne d'une population dans le contexte d'une stratification double et l'estimation d'une moyenne de population suivant des plans de sondage généraux. Enfin, l'auteur analyse le biais conditionnel et la variance conditionnelle des estimateurs de la moyenne d'une population (de la moyenne ou du total d'un domaine) de même que les intervalles de confiance correspondants.

MOTS CLÉS : Inférence conditionnelle, biais conditionnel, variance conditionnelle, moyenne de population, taille d'échantillons aléatoires.

1. INTRODUCTION

D'après la théorie classique de l'inférence dans les enquêtes par sondage, le plan de sondage définit l'espace échantillon S (ensemble des échantillons possibles s) et les probabilités de sélection correspondantes $p(s)$. Le choix d'un estimateur repose sur le critère de convergence ou d'absence de biais et sur la comparaison des erreurs quadratiques moyennes (EQM), suivant un échantillonnage répété avec des probabilités $p(s)$, l'espace échantillon S étant utilisé comme ensemble de référence. Ainsi, l'estimateur $\hat{Y}$ d'une moyenne de population $\bar{Y}$ est sans biais si $E(\hat{Y}) = \sum_{s\in S} p(s)\hat{Y}_s = \bar{Y}$, où $\hat{Y}_s$ est la valeur de $\hat{Y}$ pour l'échantillon s. L'EQM de l'estimateur $\hat{Y}$ est donnée par $\mathrm{EQM}(\hat{Y}) = \sum_{s\in S} p(s)(\hat{Y}_s - \bar{Y})^2$, et $\hat{Y}$ est convergent si son EQM tend vers zéro au fur et à mesure que la taille de l'échantillon augmente. Un estimateur convergent ou sans biais de $\mathrm{EQM}(\hat{Y})$, désigné par $\mathrm{eqm}(\hat{Y})$, renferme une mesure d'incertitude pour la valeur de $\hat{Y}$. Si $\hat{Y}$ est sans biais ou convergent, on obtient pour les valeurs observées $\hat{Y}_s$ et $\mathrm{eqm}(\hat{Y}_s)$ un intervalle de confiance de grand échantillon défini par l'équation suivante (à un seuil de confiance $1 - \alpha$):

$$I_s = \hat{Y}_s \pm z_{\alpha/2}\sqrt{\mathrm{eqm}(\hat{Y}_s)}, \tag{1}$$

où $z_{\alpha/2}$ est la borne supérieure $\alpha/2$-de l'intervalle de confiance pour une variable $N(0, 1)$. L'équation (1) signifie que, dans un échantillonnage répété ayant comme ensemble de référence S, environ $100\ (1 - \alpha)\%$ des intervalles I_s, renfermeront la vraie valeur $\bar{Y}$.

Il est juste de comparer des erreurs quadratiques moyennes inconditionnelles $\mathrm{EQM}(\hat{Y})$, au moment de l'élaboration de l'enquête, mais il se peut que l'espace échantillon S ne soit plus l'ensemble de référence qui convienne à l'inférence une fois l'échantillon s prélevé, si

[1] J.N.K. Rao, département de mathématiques et de statistique, Université Carleton, Ottawa (Ontario), Canada K1S 5B6.

celui-ci contient des sous-ensembles identifiables. Nous illustrerons plus loin la notion de sous-ensemble identifiable par des exemples d'échantillons aléatoires de tailles diverses. Le choix de l'ensemble de référence approprié n'est toutefois pas unique. En fait, on pourrait concrètement considérer l'échantillon prélevé s comme unique, mais alors aucune inférence ne serait possible dans le cadre d'un échantillonnage répété puisque l'ensemble de référence approprié serait un singleton (Holt et Smith 1979).

L'inférence conditionnelle est un sujet qui a été très discuté en statistique classique depuis Fisher (1925). En testant, par exemple, la notion d'indépendance dans un tableau de contingence 2 x 2, Fisher prétendait que l'inférence devait dépendre des totaux marginaux observés des lignes et les colonnes même si les marges n'étaient pas déterminées par le plan de sondage. Yates (1984) a réexaminé ce problème. Le choix d'un ensemble de référence approprié ne s'impose pas toujours à l'évidence, mais il paraît raisonnable d'appliquer les règles suivantes : 1) Il faut choisir une méthode conditionnelle *avant* l'observation des données, surtout dans le domaine public. 2) Il faut répartir conditionnellement l'ensemble S de telle manière que les sous-ensembles obtenus ne renferment pas ou à peu près pas de renseignements sur les paramètres d'intérêt; en d'autres mots, les statistiques servant d'indices aux sous-ensembles devraient être des fonctions auxiliaires des observations (Cox et Hinkley 1974, p. 38). 3) Si la taille des échantillons est aléatoire (par exemple, les tailles d'échantillons de domaines) et que la distribution de la population est entièrement (ou du moins partiellement) connue, les inférences dépendront des tailles d'échantillons observées. Dans ce contexte, Durbin (1969, p. 643) affirme que si la taille de l'échantillon est déterminée de manière aléatoire et qu'il se trouve qu'on obtienne un grand échantillon, on sait parfaitement bien que les mesures recherchées sont plus précises que s'il s'était agi d'un échantillon de petite taille. Il précise que, de toute évidence, on devrait se servir des renseignements connus sur la taille de l'échantillon pour interpréter les résultats. Selon Durbin, du point de vue de l'analyse des données effectivement observées, il semble tout à fait incorrect de faire une moyenne des variations hypothétiques de la taille de l'échantillon lorsqu'en fait, on connaît exactement la taille de l'échantillon.

La présente analyse se limitera à l'inférence conditionnelle en présence de tailles d'échantillon aléatoires, selon la règle 3 énoncée ci-dessus. Malgré cette restriction, nous montrerons qu'il n'est pas toujours facile, en pratique, de faire des inférences conditionnelles. Nous débutons notre analyse avec des exemples simples, puis nous poursuivons avec des problèmes plus complexes. Pour les enquêtes par sondage, ce sont Holt et Smith (1979) qui avancent les arguments les plus convaincants en faveur de l'inférence conditionnelle, bien que leur analyse soit limitée à la stratification a posteriori d'un échantillon aléatoire simple (EAS) (voir section 3.1).

Lahiri (1969) a souligné la difficulté de convaincre les utilisateurs de données statistiques, personnes intelligentes mais peu versées en la matière, de l'importance réelle des estimations provenant des enquêtes par sondage. Il souligne, notamment, l'aberration qui consiste à se servir implicitement de l'erreur type (d'échantillonnage) comme mesure de précision de l'estimation observée (de l'échantillon), en s'inspirant d'un certain nombre d'exemples tirés de la théorie actuelle.

2. ÉCHANTILLONNAGE ALÉATOIRE SIMPLE AVEC REMISE

L'échantillonnage aléatoire simple (EAS) avec remise est rarement utilisé en pratique, mais il offre un moyen facile de s'initier à l'inférence conditionnelle.

Supposons qu'un échantillon aléatoire simple s, de taille n est prélevé avec remise dans une population de taille N de telle sorte que l'ensemble S contient N^n échantillons s. Soit ν le nombre d'unités distinctes dans s. Alors, ν est une variable aléatoire pouvant prendre

les valeurs 1, ..., n. Définissons t_i comme le nombre de fois que la $i$ème unité de la population est incluse dans s. Alors, deux estimateurs courants de la moyenne de population $\bar{Y}$ sont donnés par:

$$\bar{y}_n = \frac{1}{n} \sum_{i \in s} t_i y_i, \tag{2.1}$$

la moyenne de l'échantillon fondée sur les n prélèvements, et

$$\bar{y}_\nu = \frac{1}{\nu} \sum_{i \in s} y_i, \tag{2.2}$$

la moyenne fondée sur les unités distinctes de s. Les deux moyennes $\bar{y}_n$ et $\bar{y}_\nu$ sont inconditionnellement non biaisées selon l'ensemble de référence S, et la variance inconditionnelle de $\bar{y}_\nu$ est toujours moindre que celle de $\bar{y}_n$. Ainsi, pour des raisons d'efficacité, on devrait préférer $\bar{y}_\nu$ à $\bar{y}_n$. L'estimateur d'Horvitz-Thompson

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i} = \frac{\nu}{E(\nu)} \bar{y}_\nu \tag{2.3}$$

est aussi inconditionnellement non biaisé; dans l'équation ci-dessus, π_i désigne la probabilité que l'unité i soit incluse au moins une fois dans l'échantillon:

$$\pi_i = \frac{E(\nu)}{N} = 1 - \left(1 - \frac{1}{N}\right)^n.$$

La comparaison des variances de $\bar{y}_\nu$ et de $\bar{y}_{HT}$ indique que $\bar{y}_\nu$ n'est pas toujours supérieur à $\bar{y}_{HT}$.

À la lumière des affirmations de Durbin (1969), il est clair que l'on devrait faire reposer l'inférence sur la valeur observée de ν, en d'autres mots, l'ensemble de référence approprié est l'ensemble S_ν constitué de $\binom{N}{\nu}$ échantillons de taille réelle ν, et non l'ensemble S. Heureusement, l'ensemble se prête facilement à l'inférence conditionnelle puisque $P(s_\nu | \nu) = \binom{N}{\nu}^{-1}$; en d'autres termes, l'échantillon d'unités distinctes observé s_ν, est, conditionnellement, un échantillon aléatoire simple de taille ν prélevé sans remise. Il s'ensuit que $\bar{y}_\nu$ est conditionnellement non biaisé, c'est-à-dire que $E_2(\bar{y}_\nu) = \bar{Y}$ où E_2 représente l'espérance conditionnelle, tandis que $E_2(\bar{y}_{HT}) = [\nu/E(\nu)]\bar{Y} \neq \bar{Y}$ de sorte que $\bar{y}_{HT}$ est conditionnellement biaisé. Il faudrait donc préférer $\bar{y}_\nu$ à $\bar{y}_{HT}$, en dépit de la comparaison peu concluante de leurs variances inconditionnelles. Soulignons que $\bar{y}_{HT}$ constituerait une sous-estimation marquée si la valeur ν observée était de beaucoup inférieure à $E(\nu)$.

La variance conditionnelle $V_2(\bar{y}_\nu)$, est une mesure d'incertitude convenable; elle est estimée sans distorsion par la formule suivante:

$$v(\bar{y}_\nu) = \left(\frac{1}{\nu} - \frac{1}{N}\right) s^2_{\nu y}, \tag{2.4}$$

où $(\nu - 1)s^2_{\nu y} = \sum_{i \in s} (y_i - \bar{y}_\nu)^2$ et V_2 désigne la variance conditionnelle. L'intervalle de confiance approprié pour $\bar{Y}$ est obtenu par la formule suivante:

$$I_\nu = \bar{y}_\nu \pm z_{\alpha/2} \sqrt{v(\bar{y}_\nu)}. \tag{2.5}$$

Conditionnellement, le seuil de confiance de I_ν est environ $1 - \alpha$ si la valeur de ν est élevée. Un autre estimateur de la variance, soit

$$v^*(\bar{y}_\nu) = \left[E\left(\frac{1}{\nu}\right) - \frac{1}{N}\right]s^2_{\nu y} \tag{2.6}$$

est conditionnellement biaisé sauf lorsqu'il s'étend à l'ensemble de l'espace échantillon S. Les équations (2.4) et (2.6) permettent de déduire que $v(\bar{y}_\nu) < v^*(\bar{y}_\nu)$ si $1/\nu < E(1/\nu)$, et l'inverse si $1/\nu > E(1/\nu)$. Par conséquent, l'intervalle de confiance fondé sur (2.6) serait trop étroit si $E(1/\nu) < 1/\nu$ et correspondrait alors à un seuil de confiance inférieur à $1 - \alpha$, et serait trop étendu si $E(1/\nu) > 1/\nu$, ce qui donnerait un seuil de confiance supérieur à $1 - \alpha$. Il convient de souligner qu'un intervalle conditionnellement juste est également juste en l'absence de toute condition.

3. ÉCHANTILLONNAGE ALÉATOIRE SIMPLE SANS REMISE

Supposons qu'un échantillon aléatoire simple de taille déterminée n est prélevé sans remise. En l'absence de sous-ensembles identifiables, l'ensemble de référence approprié est l'ensemble S constitué de $\binom{N}{n}$ échantillons s, de taille n; la moyenne de l'échantillon $\bar{y}_n$ est non biaisée et sa variance est estimée sans distorsion par l'équation

$$v(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)s^2_{ny} \tag{3.1}$$

où $(n - 1)s^2_{ny} = \sum_{i\in s}(y_i - \bar{y}_n)^2$. L'intervalle de confiance correspondant est donné par I_s: $\bar{y}_n \pm z_{\alpha/2}\sqrt{v(\bar{y}_n)}$ avec un seuil de confiance d'environ $1 - \alpha$ si n est élevé.

Supposons maintenant qu'il existe des sous-ensembles identifiables en ce sens que nous observons un échantillon de configuration $\underline{n} = (n_1, \ldots, n_k)$ réparti dans k strates formées à posteriori et ayant des poids donnés $W_i = N_i/N$. Idéalement, on aurait dû procéder par échantillonnage stratifié, mais il n'existait pas de base de stratification. L'ensemble de référence approprié est désormais l'ensemble $S_{\underline{n}}$, composé de $\prod\binom{N_i}{n_i}$ échantillons de configuration réalisée $\underline{n}$, la distribution de celle-ci étant entièrement connue.

3.1 Tous les $n_i \geq 1$

Si *toutes* les valeurs de $n_i \geq 1$ observées sont supérieures ou égales à 1, l'estimateur habituel de la moyenne de l'échantillon stratifié à posteriori

$$\bar{y}_{pst} = \sum W_i\bar{y}_i \tag{3.2}$$

est conditionnellement non biaisé étant donné $\underline{n}$ puisque $P(s|\underline{n}) = \prod\binom{N_i}{n_i}^{-1}$, en d'autres termes, l'échantillon observé s est, conditionnellement, un échantillon aléatoire stratifié $(s_1,\ldots,s_k)$ avec des strates de taille n_i. Dans l'équation ci-dessus, $\bar{y}_i$ désigne la moyenne de l'échantillon dans la $i$ème strate. La variance conditionnelle $V_2(\bar{y}_{pst})$ offre une mesure d'incertitude convenable; elle est estimée sans biais par l'équation

$$v(\bar{y}_{pst}) = \sum W_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s^2_{iy}, \tag{3.3}$$

pourvu que *toute les valeurs* $n_i \geq 2$, où $(n_i - 1)s_{iy}^2 = \sum_{j \in s_i} (y_{ij} - \bar{y}_i)^2$ (Holt et Smith 1979). L'intervalle de confiance correspondant, I_{pst}: $\bar{y}_{pst} \pm z_{\alpha/2}\sqrt{v(\bar{y}_{pst})}$, est conditionnellement juste. Un autre estimateur de la variance

$$v^*(\bar{y}_{pst}) = \sum W_i^2 \left[E\left(\frac{1}{n_i}\right) - \frac{1}{N_i}\right] s_{iy}^2 \tag{3.4}$$

$$\doteq \left(\frac{1}{n} - \frac{1}{N}\right) \sum W_i s_{iy}^2$$

est conditionnellement biaisé sauf lorsqu'on prend la moyenne sur l'ensemble de l'espace échantillon S (en supposant que $P(n_i \leq 1)$ soit négligeable). L'efficacité conditionnelle de l'intervalle de confiance fondé sur (3.4) dépend évidemment de la mesure où les valeurs $1/n_i$ observées divergent de leur espérance respective $E(1/n_i)$. Il convient de souligner que l'intervalle I_{pst} est également juste en l'absence de toute condition, pourvu que $P(n_i \leq 1)$ soit négligeable pour tous les i.

Si $n_i = 1$ pour certaines valeurs de i, on ne peut obtenir d'estimateur de variance conditionnellement non biaisé. Dans ce cas, toutefois, il suffirait d'utiliser la méthode des strates combinées ou la solution modèle proposée à l'origine par Hartley *et coll.* (1969) pour estimer la variance dans un échantillonnage aléatoire stratifié donnant une unité par strate. Des études empiriques pourraient probablement éclaircir la question de l'applicabilité de ces méthodes.

L'argument invoqué habituellement en faveur de $\bar{y}_{pst}$ de préférence à $\bar{y}$ est que la variance inconditionnelle de $\bar{y}_{pst}$ est approximativement égale à la variance obtenue dans la répartition proportionnelle et,de ce fait, est moindre que la variance inconditionnelle de $\bar{y}$. Il faut également se rappeler que la répartition proportionnelle est susceptible de produire des gains d'efficacité plutôt modestes. Il est toutefois plus important de noter que la moyenne de l'échantillon ($\bar{y}$) est conditionnellement biaisée:

$$E_2(\bar{y}) = \sum w_i \bar{Y}_i \neq \sum W_i \bar{Y}_i = \bar{Y}, \; w_i = \frac{n_i}{n}, \tag{3.5}$$

et que, par conséquent, les inférences correspondantes pourraient être conditionnellement inexactes.

Exemple 1. Supposons $k = 2$ (par exemple, des strates d'hommes et de femmes avec des poids projetés du recensement, W_1 et $W_2 = 1 - W_1$, ou des petits et des grands hôpitaux (Royall 1970)). Royall a utilisé un modèle de superpopulation

$$E_m(y_i) = \beta x_i, \; i = 1, \ldots, N, \; \beta > 0, \; x_i > 0 \tag{3.6}$$

pour démontrer que $\bar{y}$ est conditionnellement biaisé en fonction du modèle. Dans cette équation, E_m désigne l'espérance du modèle, c'est-à-dire,

$$E_m(\bar{y}) = \beta \bar{x} \neq E_m(\bar{Y}) = \beta \bar{X} \tag{3.7}$$

à moins que la moyenne de l'échantillon $\bar{x}$ ne corresponde à la moyenne de la population $\bar{X}$. Dans le modèle de Royall $x_i =$ représente le nombre de lits dans le $i$ème hôpital et $y_i =$ le nombre de lits occupés dans le $i$ème hôpital; de plus, $x_1, \ldots, x_N$ sont connus. Royall soutient que $\bar{y}$ produit une sérieuse sous-estimation si l'échantillon observé contient tous (ou presque tous) les petits hôpitaux puisque $B_m(\bar{y}) = E_m(\bar{y}) - E_m(\bar{Y}) = \beta(\bar{x} - \bar{X})$ et $\bar{x} \ll \bar{X}$. Nous pouvons démontrer la même chose dans notre cadre d'analyse conditionnel sans

supposer l'existence d'un modèle. Il est possible d'exprimer le rapport du biais conditionnel de $\bar{y}$ à la moyenne de la population des grands hôpitaux, $\bar{Y}_2$, comme suit:

$$\frac{B_2(\bar{y})}{\bar{Y}_2} = (W_1 - w_1)\delta = (w_2 - W_2)\delta, \tag{3.8}$$

où $B_2(\bar{y}) = E_2(\bar{y}) - \bar{Y}$ désigne le biais conditionnel de $\bar{y}$, $\delta = (\bar{Y}_2 - \bar{Y}_1)/\bar{Y}_2$ et $0 < \delta < 1$ puisque la moyenne de la population $\bar{Y}_1$, des petits hôpitaux est moindre que $\bar{Y}_2$. Si $w_1 = 1$ (c'est-à-dire que tous les petits hôpitaux sont inclus dans l'échantillon), alors $E_2(\bar{y}) = \bar{Y}_1 << \bar{Y}$ et $\bar{y}$ constitue donc une sérieuse sous-estimation. De même, si $w_1 >> W_1$ (c'est-à-dire que la plupart des petits hôpitaux sont inclus dans l'échantillon), l'équation (3.8) nous amène à déduire que $\bar{y}$ donnerait lieu à une sérieuse sous-estimation.

Dans l'exemple ci-dessus, il faudrait utiliser l'estimateur de la moyenne de l'échantillon stratifié a posteriori $\bar{y}_{pst} = W_1\bar{y}_1 + W_2\bar{y}_2$, qui est conditionnellement non biaisé sauf si $n_1 = 0$ ou $n_2 = 0$. En fait, il serait préférable d'utiliser un estimateur par quotient

$$\bar{y}_{pst,r} = \frac{\bar{y}_{pst}}{\bar{x}_{pst}}\bar{X}, \tag{3.9}$$

où $\bar{x}_{pst} = W_1\bar{x}_1 + W_2\bar{x}_2$ et $\bar{x}_i$ est la moyenne de l'échantillon de x dans la $i$ème strate. L'estimateur (3.9) est, à peu de choses près, conditionnellement non biaisé et plus efficace que $\bar{y}_{pst}$ si n est grand.

Remarque 1. Dans son exemple, Royall devrait en fait utiliser un plan de sondage plus efficace que l'échantillonnage aléatoire simple puisque toutes les valeurs x de la population sont connues; ce pourrait être, par exemple, un échantillonnage aléatorie stratifié suivant x ou, peut-être, une répartition optimale fondée sur les valeurs x.

Remarque 2. Royall justifie l'utilisation de l'estimateur par quotient habituel $\bar{y}_r = (\bar{y}/\bar{x})\bar{X}$ dans son modèle défini par l'équation (3.6); l'utilisation de cet estimateur ne peut toutefois être justifiée dans le cadre d'analyse conditionnel (échantillonnage répété) puisque $\bar{y}_r$ est conditionnellement biaisé:

$$B_2(\bar{y}_r) \doteq \bar{X}\left[\frac{w_2\bar{Y}_1 + w_2\bar{Y}_2}{w_1\bar{X}_1 + w_2\bar{X}_2} - R\right], \; R = \frac{\bar{Y}}{\bar{X}} \tag{3.10}$$

$$\neq 0$$

sauf si $\bar{y}_1/\bar{x}_1 = \bar{y}_2/\bar{x}_2 = R$. Dans le cas extrême où $w_1 = 1$, $B_2(\bar{y}_r) = \bar{X}(R_1 - R)$ où $R_1 = \bar{Y}_1/\bar{X}_1$. Ainsi, $B_2(\bar{y}_r) \lesseqgtr 0$ dans la mesure où $R_1 \lesseqgtr R$.

Remarque 3. Si le poids W_1 n'est pas connu mais que $\bar{X}$ l'est, nous ne pouvons utiliser $\bar{y}_{pst}$ ni $\bar{y}_{pst,r}$. Royall propose d'utiliser $\bar{y}_r$ en faisant reposer l'inférence sur la moyenne observée $\bar{x}$. Toutefois, le choix de $\bar{x}$ est quelque peu arbitraire, et il se pourrait que le biais conditionnel de $\bar{y}_r$ soit assez important à moins que le modèle décrit par l'équation (3.6) ne soit vrai, du moins approximativement.

Si nous disposions a priori de renseignements valables sur W_1, par exemple $W_1^* \leq W_1 \leq W_1^{**}$ où W_1^* et W_1^{**} sont connus, on pourrait alors utiliser le pseudo estimateur de la moyenne de l'échantillon stratifié a posteriori: $\bar{Y}$:

$$\bar{y}_{pst}^* = \tilde{W}_1\bar{y}_1 + \tilde{W}_2\bar{y}_2, \tag{3.11}$$

où $\tilde{W}_1 = w_1$ si $W_1^* \leq w_1 \leq W_1^{**}$, $= W_1^*$ si $w_1 < W_1^*$, $= W_1^{**}$ si $w_1 > W_1^{**}$ et $\tilde{W}_2 = 1 - \tilde{W}_1$. Même biaisés, l'estimateur $\bar{y}_{pst}^*$ et son équivalent sous forme de quotient devraient être plus efficaces conditionnellement, que $\bar{y}$ et $\bar{y}_r$ si (n_1, n_2) sont donnés. En l'absence de toute condition, l'EQM de $\bar{y}_{pst}^*$ devrait être moindre que l'EQM de $\bar{y}$, pourvu que $W_1^* \leq W_1 \leq W_1^{**}$. On pourrait ausi recourir à une méthode bayesienne pour estimer W_1 en définissant une distribution préalable pour W_1.

Exemple 2 (observations aberrantes). L'estimation d'une moyenne de population $\bar{Y}$ en présence d'observations aberrantes pose un problème comparable à celui de l'exemple précédent concernant les hôpitaux. Supposons qu'il est démontré que la population renferme une faible proportion W_2, d'observations aberrantes (valeurs excessives) mais que W_2 est inconnu, c'est-à-dire $W_1 >> W_2$ et $\bar{Y}_2 >> \bar{Y}_1$. Alors, si l'échantillon observé ne contient aucune observation aberrante (c'est-à-dire, $w_2 = 0$), nous dirons que $\bar{y}$ ne reflète aucunement la vraie valeur de $\bar{Y}$ (Chinnappa, 1976) malgré que $\bar{y}$ soit (inconditionnellement) non biaisé. Le sens de cet énoncé découle du fait que $E_2(\bar{y}) = \bar{Y}_1 << \bar{Y}$, où E_2 désigne, comme précédemment, l'espérence conditionnelle.

Par ailleurs, nous considérions $\bar{y}$ comme une surestimation importante si l'échantillon contenait des observations aberrantes. L'équation (3.8) nous permet d'avancer cela en soulignant que $w_2 >> W_2$ (puisque W_2 est très petit). Par exemple, si $N_2 = 1$, nous avons $w_2 = 1/n >> W_2 = 1/N$. Dans ce cas, il nous faut modifier l'estimation $\bar{y}$ en réduisant le poids associé aux observations aberrantes comprises dans l'échantillon. Nous pourrions, par exemple, faire passer ce poids $\bar{y}$ de $1/n$ à $1/N$ et rajuster le poids associé aux autres observations de façon que la somme des n poids soit égale à 1:

$$\bar{y}^* = \frac{N - n_2}{N} \bar{y}_1 + \frac{n_2}{N} \bar{y}_2. \tag{3.12}$$

Le biais relatif conditionnel de $\bar{y}^*$ est calculé à l'aide de l'équation suivante:

$$\frac{B_2(\bar{y}^*)}{\bar{Y}_2} = \left(w_2 \frac{n}{N} - W_2\right)\delta, \tag{3.13}$$

alors que $B_2(\bar{y})/\bar{Y}_2 = (w_2 - W_2)\delta$. Si $w_2 \frac{n}{N} - W_2 < 0$, alors

$$\left|w_2 \frac{n}{N} - W_2\right| = W_2 - w_2 \frac{n}{N} < w_2 - W_2 \text{ si } 2W_2 < w_2\left(1 + \frac{n}{N}\right).$$

L'inéquation $2W_2 < w_2(1 + n/N)$ devrait être satisfaite puisque $w_2 >> W_2$. Si $w_2 n/N - W_2 > 0$, alors

$$\left|w_2 \frac{n}{N} - W_2\right| = w_2 \frac{n}{N} - W_2 < w_2 - W_2.$$

Par conséquent, le biais conditionnel de l'estimateur $\bar{y}^*$ devrait être moindre, en valeur absolue, que celui de l'estimateur $\bar{y}$.

L'estimateur $\bar{y}^*$ découle essentiellement de l'estimateur de la moyenne de l'échantillon stratifié a posteriori $\bar{y}_{pst}$ lorsqu'on pose $N_2 = n_2$. On peut obtenir une solution plus satisfaisante en recueillant, au préalable, des renseignements convenables sur $W_1(= 1 - W_2)$ à partir des données du recensement par exemple, et en utilisant par la suite l'estimateur $\bar{y}_{pst}^*$ ou l'estimateur fondé sur un estimateur de Bayes de W_1.

Hidiroglou et Srinath (1981) ont calculé le biais conditionnel de même que les EQM conditionnelle et inconditionnelle de $\bar{y}$, $\bar{y}^*$ et d'autres variantes de $\bar{y}$, mais ils n'ont pas comparé les biais conditionnels de $\bar{y}$ et de $\bar{y}^*$ comme ci-dessus.

3.2 $n_i = 0$ pour certaines strates i

Si la taille n de l'échantillon est petite ou qu'il y a un trop grand nombre de strates formées a posteriori, n_i pourrait être égal à zéro pour certaines strates i. L'estimateur de la moyenne de l'échantillon stratifié a posteriori (équation 3.2) devient, dans ce cas,:

$$\bar{y}_{pst} = \sum' W_i \bar{y}_i, \tag{3.14}$$

où $\sum'$ représente la sommation sur l'ensemble des strates ayant une valeur n_i non nulle. L'estimateur (3.14) est conditionnellement biaisé:

$$E_2(\bar{y}_{pst}) = \sum' W_i \bar{Y}_i \neq \sum W_i \bar{Y}_i. \tag{3.15}$$

Il demeure conditionnellement biaisé même suivant l'hypothèse idéale que $\bar{Y}_i = \bar{Y}$ pour toutes les i, qui montre, par ailleurs, que $\bar{y}_{pst}$ pourrait produire une sous-estimation importante. Cet estimateur est aussi biaisé inconditionnellement. Pour surmonter ces difficultés, il existe une méthode courante qui consiste à combiner des strates semblables pour faire en sorte que $n_i > 0$ pour toutes les i dans l'ensemble réduit de strates. Fuller (1966) propose une solution plus efficace pour le cas particulier où $k = 2$ strates formées a posteriori mais où son cadre d'analyse est inconditionnel en ce sens qu'il tient compte de la probabilité P_1^*, que n_1 soit égal à 0 étant donné que $n_1 = 0$ ou $n_2 = 0$. L'estimateur proposé par Fuller prend la forme suivante:

$$\begin{aligned} \bar{y}_F &= \frac{W_1}{P_1^*}\bar{y}_1 \text{ si } n_2 = 0 \\ &= \frac{W_2}{P_2^*}\bar{y}_2 \text{ si } n_1 = 0, \end{aligned} \tag{3.16}$$

où $P_2^* = 1 - P_1^*$. L'estimateur $\bar{y}_F$ est conditionnellement non biaisé si $n_1 = 0$ ou $n_2 = 0$, mais il est conditionnellement biaisé si nous avons (n_1, n_2), même lorsque $\bar{Y}_1 = \bar{Y}_2 = \bar{Y}$.

L'équation suivante définit un estimateur inconditionnellement non biaisé:

$$\bar{y}_D = \sum \frac{a_i}{E(a_i)} W_i \bar{y}_i, \tag{3.17}$$

(Doss *et coll.* 1979), où $a_i = 1$ si au moins une unité de la strate i est incluse dans l'échantillon, $= 0$ dans le cas contraire, et $\bar{y}_i$ est défini comme $\bar{Y}_i$ si $n_i = 0$ (à noter que $a_i \bar{y}_i = 0$ si $n_i = 0$ même si l'on ne connaît pas $\bar{Y}_i$). Toutefois, l'estimateur $\bar{y}_D$, est conditionnellement biaisé puisque

$$E_2(\bar{y}_D) = \sum' \frac{W_i \bar{Y}_i}{E(a_i)} \neq \sum W_i \bar{Y}_i = \bar{Y}.$$

Il demeure conditionnellement biaisé même si $\bar{Y}_i = \bar{Y}$ pour tous les i.

Doss *et coll.* ont critiqué l'efficience de $\bar{y}_D$ parce qu'il n'est pas invariant par translation (c'est-à-dire que $\bar{y}_D$ n'augmente pas d'une valeur c lorsque chaque y_i devient $y_i + c$, c étant une constante arbitraire), et que, par conséquent, on peut accroître indûment la variance

de $\bar{y}_D$, en haussant y_i d'une valeur c suffisamment élevée. Par ailleurs, l'estimateur par quotient

$$\bar{y}_{rD} = \frac{\sum \frac{a_i}{E(a_i)} W_i \bar{y}_i}{\sum \frac{a_i}{E(a_i)} W_i}, \tag{3.18}$$

proposé par Doss *et coll.*, est invariant par translation. Il est conditionnellement biaisé, mais le biais est approximativement égal à zéro si $\bar{Y}_i = \bar{Y}$ pour tous les i. L'équation suivante définit un autre estimateur par quotient qui est conditionnellement semblable à $\bar{y}_{rD}$:

$$\bar{y}_{r(pst)} = \frac{\sum' W_i \bar{y}_i}{\sum' W_i}, \tag{3.19}$$

Contrairement à $\bar{y}_{rD}$, celui-ci est inconditionnellement non convergent. Ainsi, on peut préférer $\bar{y}_{rD}$ à $\bar{y}_{r(pst)}$ ou à $\bar{y}_D$.

Si l'on disposait simultanément de renseignements sur *toutes* les strates, il serait possible d'ajuster un modèle en fonction des moyennes de strates observées $\bar{y}_i$ et d'estimer la moyenne de la population des strates ayant $n_i = 0$. Par exemple, s'il existe une relation linéaire entre les moyennes de la population $\bar{X}_i$ d'une variable concomitante et les valeurs $\bar{Y}_i$ correspondantes, on peut calculer la valeur prévue de $\bar{Y}_i$ à l'aide de la formule $\hat{\alpha} + \hat{\beta}\bar{X}_i = \bar{y}_i^*$, où $\hat{\alpha}$ et $\hat{\beta}$ sont les estimateurs par la méthode des moindres carrés obtenus en minimisant $\sum'(\bar{y}_i - \alpha - \beta\bar{X}_i)^2$. L'estimateur de $\bar{Y}$ est donc défini par

$$\bar{y}_{pst}^* = \sum' W_i \bar{y}_i + \sum'' W_i \bar{y}_i^*, \tag{3.20}$$

où $\sum''$ désigne la sommation sur l'ensemble des strates où $n_i = 0$. Si le modèle ajusté s'avère satisfaisant, cet estimateur devrait présenter de bonnes propriétés conditionnelles. Nous pouvons donc conclure à partir de cette analyse qu'il n'y a pas de solution facile si $n_i = 0$ pour quelques-unes des strates.

4. STRATIFICATION DOUBLE

Les ouvrages de statistique renferment des solutions ingénieuses pour accroître l'efficacité des estimateurs. Bryant *et coll.* (1960) ont proposé un plan de sondage à stratification double, suivant lequel la taille n_{ij} de l'échantillon de certaines strates (cellules) est nulle. Par cette méthode, nous sommes censés pouvoir estimer la moyenne de la population même si la taille globale n de l'échantillon est inférieure au nombre total de strates. A l'aide d'une méthode de répartition proportionnelle des tailles marginales $(n_{i.}, n_{.j})$, Bryant *et coll.* ont obtenu une répartition aléatoire n_{ij} telle que $E(n_{ij}) = (n_{i.}n_{.j})/n = nW_{i.}W_{.j}$, où $W_{i.}$ et $W_{.j}$ représentent respectivement les totaux marginaux des lignes et des colonnes pour les poids des cellules W_{ij}. Ces mêmes auteurs ont proposé l'estimateur

$$\bar{y}_U = \frac{1}{n} \sum \sum n_{ij} G_{ij} \bar{y}_{ij}, \tag{4.1}$$

où $G_{ij} = n^2 W_{ij}/(n_{i.}n_{.j})$ et $\bar{y}_{ij}$ peut être assimilé à $\bar{Y}_{ij}$ si $n_{ij} = 0$. L'estimateur $\bar{y}_U$ est inconditionnellement non biaisé. Toutefois, la distribution de n_{ij} est entièrement connue (puisque tous les W_{ij} sont connus) et, par conséquent, l'ensemble de référence approprié est l'ensemble des échantillons à configuration $\{n_{ij}\}$; en d'autres termes, le plan de sondage proposé devrait être considéré, aux fins d'inférence, comme un échantillonnage aléatoire simple

stratifié. L'estimateur $\bar{y}_U$ est conditionnellement biaisé:

$$E_2(\bar{y}_U) = \sum\sum(\frac{n_{ij}G_{ij}}{n})\bar{Y}_{ij} \neq \sum\sum W_{ij}\bar{Y}_{ij} = \bar{Y};$$

soulignons que $E_2(\bar{y}_{ij}) = \bar{Y}_{ij}$ si $n_{ij} > 0$. Ce dernier estimateur présente les mêmes faiblesses que l'estimateur $\bar{y}_D$ défini dans la section précédente; la difficulté peut être contournée par l'emploi de l'estimateur par quotient

$$\bar{y}_r = \frac{\bar{y}_U}{\bar{a}_U} = \frac{\sum\sum n_{ij}G_{ij}\bar{y}_{ij}}{\sum\sum n_{ij}G_{ij}} \tag{4.2}$$

où $\bar{a}_U = \sum\sum n_{ij}G_{ij}/n$. $\bar{y}_r$ L'estimateur (4.2) est aussi entaché d'un biais conditionnel mais celui-ci est approximativement égal à zéro si $\bar{Y}_{ij} = \bar{Y}$ pour tous les (i, j). Cette dernière condition peut toutefois être invraisemblable dans le cas qui nous occupe puisque les strates sont définies différemment dans le plan de sondage.

Comme à la section 3.1, il semble nécessaire d'utiliser un modèle qui met en relation les strates échantillonnées et les strates non échantillonnées. Vu l'absence de renseignements concomitants, nous pouvons raisonnablement supposer le modèle suivant:

$$y_{ijk} = \mu + \beta_j + \tau_i + \varepsilon_{ijk} \tag{4.3}$$

où y_{ijk} est la k-ième observation de la (i, j)-ième cellule, β_j et τ_i sont les effets permanents et ε_{ijk} sont les erreurs indépendantes ayant zéro pour moyenne et σ^2 pour variance commune. Malheureusement, les données de l'échantillon ne permettent pas d'estimer la combinaison linéaire $\mu + \beta_j + \tau_i$ pour les strates non échantillonnées et il est, par conséquent, impossible d'estimer les valeurs $\bar{Y}_{ij}$ correspondantes. On peut remédier à cette situation en supposant que β_j et τ_i sont des variables aléatoires, ce qui nous permet d'obtenir un élément prédictif $\hat{\mu} + \hat{\beta}_j + \hat{\tau}_i$. Dans le cas qui nous occupe, toutefois, le modèle des effets aléatoires peut être moins réaliste que celui qui est défini par l'équation (4.3).

Le défi que ces problèmes posaient a poussé Bankier (1985) a élaboré une méthode itérative appliquée à des échantillons stratifiés indépendants suivant deux critères de stratification distincts. Son estimateur n'est à peu près pas biaisé en fonction du modèle des effets permanents (4.3) tandis que l'estimateur habituel d'Horvitz-Thompson et son équivalent sous forme de quotient le sont.

La méthode de Bankier peut être appliquée au problème de la stratification double. D'après cette méthode, l'estimateur par quotient de $\bar{Y}$ est défini:

$$\bar{y}(p) = \sum\sum\frac{G_{ij}(p)}{n}y_{ij} \tag{4.4}$$

où y_{ij} est le total de l'échantillon dans la (i, j)-ème cellule ($y_{ij} = 0$ si $n_{ij} = 0$) et $G_{ij}(p)$ représente les valeurs obtenues au cours de la p-ième itération, de telle sorte que:

$$\sum_j\frac{G_{ij}(p)}{n}n_{ij} \doteq W_{i.} = \sum_j W_{ij} \tag{4.5}$$

et

$$\sum_i \frac{G_{ij}(p)}{n} n_{ij} \doteq W_{.j} = \sum_i W_{ij}.$$

Les valeurs de $G_{ij}(p)$ sont déterminées comme suit: soit $G_{ij}(0) = G_{ij} > 0 \ \forall(i, j)$, et

$$G_{ij}(p) = G_{ij}(p-1) \frac{W_{i.}}{\sum_j \frac{G_{ij}(p-1)}{n} n_{ij}} \quad \text{si } p \text{ est impair}$$

$$= G_{ij}(p-1) \frac{W_{.j}}{\sum_i \frac{G_{ij}(p-1)}{n} n_{ij}} \quad \text{si } p \text{ est pair.} \tag{4.6}$$

En vertu du modèle des effets permanents (4.3), nous avons

$$E_m[\bar{y}(p)] \doteq \mu + \sum_i W_{i.}\tau_i + \sum_j W_{.j}\beta_j = E_m(\sum\sum W_{ij}\bar{Y}_{ij})$$

$$= E_m(\bar{Y}),$$

c'est-à-dire que $\bar{y}(p)$ n'est à peu près pas biaisé en fonction du modèle. Puisque $E(G_{ij}(0)n_{ij}/n) = W_{ij}$ pour la sélection $G_{ij}(0) = G_{ij}$, les valeurs initiales devraient être bonnes. La méthode itérative pourrait toutefois poser des problèmes de convergence à cause des nombreuses cellules vides ($n_{ij} = 0$) prévues dans le modèle de Bryant *et coll.* Nous analyserons éventuellement ces problèmes de convergence et les propriétés conditionnelles de l'estimateur par quotient (4.4) dans un document ultérieur.

Si les moyennes de population $\bar{X}_{ij}$ d'une variable concomitante x sont connues pour *toutes* les strates, il est alors possible, comme à la section 3.1, d'ajuster un modèle en fonction des moyennes de strates observées $\bar{y}_{ij}$. Par exemple, le modèle $\bar{y}_{ij} = \beta\bar{x}_{ij} + b_j + t_i + \bar{\varepsilon}_{ij}$ où b_j et t_i sont les effets aléatoires et $\bar{\varepsilon}_{ij}$ est la moyenne de l'échantillon des erreurs ε_{ijk} dans la (i, j)-ième cellule, pourrait être un modèle acceptable. À cela pourrait s'ajouter un élément prédictif de $\bar{Y}_{ij}$ pour les strates non échantillonnées $\hat{\beta}\bar{x}_{ij} + \hat{b}_j + \hat{t}_i$, grâce auquel on pourrait déterminer un estimateur de $\bar{Y}$. Cette méthode s'apparente aux méthodes d'estimation pour les petites régions, sauf qu'ici le paramètre d'intérêt est la moyenne globale $\bar{Y}$ plutôt que les molyennes des cellules $\bar{Y}_{ij}$. Nous prévoyons analyser les propriétés conditionnelles d'autres estimateurs de $\bar{Y}$ dans un document ultérieur.

5. NON-RÉPONSE

5.1 Modèle simplifié

Supposons qu'on obtient m réponses dans un échantillon aléatoire simple de taille n. Définissons W_1 comme la proportion correspondant à la strate de réponse et $\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2$ comme la moyenne de la population, où $\bar{Y}_1$ et $\bar{Y}_2$ sont les moyennes se rapportant respectivement à la strate de réponse et à la strate de non-réponse; de plus, $W_2 = 1 - W_1$. Dans le présent modèle, il n'est pas dit que l'inférence doit reposer sur la valeur observée de m puisque la distribution de m dépend de la valeur inconnue W_1, laquelle entre dans le calcul du paramètre d'intérêt. En outre, la moyenne ($\bar{y}_m$) de l'échantillon des

répondants est inconditionnellement biaisée puisque $E(\bar{y}_m) = \bar{Y}_2 \neq \bar{Y}$. Il est donc nécessaire de construire un modèle de réponse, même en l'absence de toute condition, à moins qu'on ne prélève un sous-échantillon de non-répondants. Dans un modèle simplifié, on suppose que la probabilité d'une réponse est la même pour toutes les unités rejointes, disons p^*; en d'autres termes, les données manquantes sont réparties aléatoirement. Suivant ce modèle, la distribution de m dépend uniquement de p^* et nous devrions, par conséquent, faire reposer l'inférence sur m si p^* est supposée connue (ou, du moins, partiellement connue ou non liée à $\bar{Y}$). Oh et Scheuren (1983) ont montré que, moyennant une valuer m donnée, l'échantillon s_m de répondants pouvait être comparé à un échantillon aléatoire simple de taille m prélevé dans la population globale. Ainsi, $\bar{y}_m$ est conditionnellement non biaisé et sa variance conditionnelle est estimée sans distorsion par l'équation suivante:

$$v_2(\bar{y}_m) = (m^{-1} - N^{-1})s^2_{my}, \tag{5.1}$$

où $(m-1)s^2_{my} = \sum_{i \in s_m}(y_i - \bar{y}_m)^2$. L'intervalle de confiance correspondant $\bar{y}_m \pm z_{\alpha/2}\sqrt{v_2(\bar{y}_m)}$ est conditionnellement juste, du moins approximativement, si m est grand.

Par ailleurs, l'estimateur d'Horvitz-Thompson (p^* étant connue):

$$\bar{y}_{HT} = \frac{m}{E(m)}\bar{y}_m = \sum_{i \in s_m} \frac{y_i}{np^*} \tag{5.2}$$

est conditionnellement biaisé, comme dans le cas illustré à la section 2, mais il est non biaisé lorsqu'il s'applique à l'ensemble de la distribution de m. En ce qui concerne les plans de sondage généraux, l'estimateur par quotient

$$\hat{\bar{Y}}_{HT,r} = \frac{\sum_{s_m} \frac{y_i}{\pi_i p_i^*}}{\sum_{s_m} \frac{1}{\pi_i p_i^*}} \tag{5.3}$$

est souvent utilisé pour des raisons d'efficience; dans l'équation ci-dessus, π_i est la probabilité d'inclusion et p_i^* est la probabilité de réponse (supposée connue) de la i-ième unité à condition que celle-ci ait été contactée. Dans le cas d'un échantillonnage aléatoire simple où $p_i^* = p^*$, il est intéressant de constater que $\hat{\bar{Y}}_{HT,r}$ se réduit à $\bar{y}_m$. On peut donc en déduire que l'estimateur par quotient pourrait s'avérer efficace dans un cadre d'analyse conditionnel appliqué à des plans de sondage généraux.

5.2 Modèle plus conforme à la réalité

Dans un modèle plus conforme à la réalité, on suppose que les données manquantes sont réparties aléatoirement dans des strates formées a posteriori et pondérées par des poids W_i connus. Soit n_i et m_i qui désignent respectivement la taille de l'échantillon et la taille de l'échantillon des répondants dans la i-ième strate formée a posteriori. Alors, la distribution à plusieurs variables de (n_i, m_i) dépend uniquement de la valeur de W_i et des probabilités de réponse dans les strates formées a posteriori. L'inférence devrait donc reposer sur les valeurs observées de (n_i, m_i), pourvu que les probabilités de réponse dans les strates formées a posteriori soient connues ou qu'elles n'aient aucun rapport avec les paramètres étudiés (par exemple, les moyennes des strates formées a posteriori). Conditionnellement, l'échantillon observé est comparable à un échantillon aléatoire simple stratifié avec des strates de taille

déterminé m_i (Oh et Scheuren 1983). En conséquence, l'estimateur

$$\bar{y}_{pst,m} = \sum W_i \bar{y}_{mi} \tag{5.4}$$

est conditionnellement non biaisé et sa variance conditionnelle est estimée sans distorsion par l'équation suivante;

$$v_2(\bar{y}_{pst,m}) = \sum W_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{miy}^2 \tag{5.5}$$

où $\bar{y}_{mi}$ et s_{miy}^2 représentent respectivement la moyenne et la variance de l'échantillon des répondants dans la $i$ème strate formée a posteriori.

Si on ne connaît pas les valeurs de W_i, il est de pratique courante de remplacer W_i dans l'équation (5.4) par son estimation $w_i = n_i/n$. On peut alors s'interroger sur la valeur de l'inférence conditionnelle puisque la distribution de (n_i, m_i) dépend des poids W_i, que l'on ne connaît pas, et que ces poids entrent dans le calcul du paramètre $\bar{Y} = \sum W_i \bar{Y}_i$. Si l'on disposait de données partielles sur W_i (par exemple, les bornes de W_i), l'inférence conditionnelle pourrait se faire comme dans le cas décrit à la remarque 3 de l'exemple 1, cependant on obtiendrait encore un estimateur conditionnellement biaisé (mais vraisemblablement plus efficace que l'estimateur (5.4) où W_i est remplacé par w_i).

6. ESTIMATION DE DOMAINES (EAS)

6.1 Moyenne d'un domaine

Dans un échantillonnage aléatoire simple (EAS), l'estimateur habituel de la moyenne d'une sous-population (domaine), $\bar{Y}_i$, est exprimé par la moyenne de l'échantillon

$$\bar{y}_i = \sum_{j \in s_i} \frac{y_j}{n_i}, \; n_i > 0 \tag{6.1}$$

où s_i est l'échantillon prélevé dans le domaine et n_i est la taille correspondante.

Si la taille du domaine, N_i, est connue, on devrait faire reposer l'inférence sur la valeur observée n_i. L'estimateur $\bar{y}_i$ est conditionnellement non biaisé si $n_i > 0$ puisque, conditionnellement, s_i est un échantillon aléatoire simple de taille déterminée n_i prélevé dans le domaine. La variance conditionnelle de cet estimateur est estimée sans distorsion par:

$$v(\bar{y}_i) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) s_{iy}^2, \; n_i > 0 \tag{6.2}$$

et l'intervalle de confiance correspondant $\bar{y}_i \pm z_{\alpha/2} \sqrt{v(\bar{y}_i)}$ est conditionnellement juste.

L'estimateur $\bar{y}_i$ est toutefois irrégulier dans le cas de petits domaines (petites régions) où n_i est petit. De plus, $\bar{y}_i$ n'est pas défini si $n_i = 0$. Pour remédier à cela, les statisticiens proposent d'utiliser un estimateur modifié

$$\bar{y}_i' = \frac{a_i}{E(a_i)} \bar{y}_i, \; n_i \geq 0 \tag{6.3}$$

où $a_i = 1$ si $n_i \geq 1$; $= 0$ si $n_i = 0$ et où $\bar{y}_i$ est assimilé à $\bar{Y}_i$ si $n_i = 0$. L'estimateur défini en (6.3) est toutefois conditionnellement biaisé:

$$E_2(\bar{y}_i') = \frac{a_i}{E(a_i)} \bar{Y}_i.$$

Il constitue une sous-estimation si $n_i = 0$, et une surestimation si $n_i \geq 0$, même s'il est inconditionnellement non biaisé. Le degré de surestimation dépend de la valeur de $E(a_i) = P(n_i \geq 1)$. Si, par exemple, $P(n_i \geq 1) = 0.75$, alors $E_2(\bar{y}_i') = (\frac{4}{3})\bar{Y}_i$ si $n_i \geq 1$.

Sarndal (1984) a propose l'estimateur suivant dans le cas de l'estimation pour les petites régions:

$$\bar{y}_{iS} = \bar{y} + \frac{w_i}{W_i}(\bar{y}_i - \bar{y}),\ n_i \geq 0, \tag{6.4}$$

où $\bar{y} = \sum w_j \bar{y}_j$ est la moyenne globale de l'échantillon et $w_i = n_i/n$. L'estimateur est à peu près inconditionnellement non biaisé, mais il est conditionnellement biaisé à moins que $w_i = W_i$:

$$B_2(\bar{y}_{iS}) = (\frac{w_i}{W_i} - 1)(\bar{Y}_i - \bar{Y}'), \tag{6.5}$$

où $\bar{Y}' = \sum w_i \bar{Y}_i$. Si $n_i = 0$, l'estimateur $\bar{y}_{iS}$ se réduit à l'estimateur synthétique $\bar{y}$. Le degré de sous-estimation (ou de surestimation) de $\bar{y}_{iS}$ dépend à la fois de $w_i/W_i - 1$ et de $\bar{Y}_i - \bar{Y}'$ et est, par conséquent, plus difficile à analyser que le biais de $\bar{y}_i'$. Cependant, le biais conditionnel* de $\bar{y}_{iS}$ serait supérieur, en valeur absolue, à celui de $\bar{y}$ si $w_i > 2W_i$ (l'EQM conditionnelle de $\bar{y}_{iS}$ serait de ce fait plus grande). De plus, la variance conditionnelle de l'estimateur conditionnellement non biaisé $\bar{y}_i$ est moindre que celle de $\bar{y}_{iS}$ si $w_i > W_i$ (si l'on ne tient pas compte de la variance de $\bar{y}$ par rapport à celle de $\bar{y}_i$) et, de ce fait, l'EQM conditionnelle de $\bar{y}_i$ est moindre que celle de $\bar{y}_{iS}$.

Hidiroglou et Sarndal (1985) proposé une variante de $\bar{y}_{iS}$:

$$\bar{y}_{iS}^{**} = \begin{cases} \bar{y}_i \text{ si } w_i \geq W_i \\ \bar{y}_{iS}^{*} = \bar{y} + \left(\frac{w_i}{W_i}\right)^2 (\bar{y}_i - \bar{y}) \text{ si } w_i < W_i. \end{cases} \tag{6.6}$$

L'estimateur $\bar{y}_{iS}^{**}$est conditionnellement non biaisé si $w_i \geq W_i$, tandis que son biais conditionnel, en valeur absolue, est moindre que celui de $\bar{y}$ si $w_i < W_i$. L'estimateur défini en (6.6) est justifié par le fait que la variance conditionnelle de $\bar{y}_{iS}^{*}$ (ou $\bar{y}_{iS}$) est supérieure à celle de $\bar{y}_i$ (si l'on ne tient pas compte de la variance de $\bar{y}$ par rapport à celle de $\bar{y}_i$), si $w_i > W_i$, tandis qu'elle est inférieure à celle de $\bar{y}_{iS}$ si $w_i < W_i$. Sous cette condition, toutefois, le biais conditionnel de $\bar{y}_{iS}^{*}$ est supérieur, en valeur absolue, à celui de $\bar{y}_{iS}$. Ainsi, le choix entre ces deux estimateurs lorsque $w_i < W_i$ ne s'impose pas à l'évidence, et il ne semble y avoir de solution simple.

*L'estimateur de Sarndal devrait toutefois être plus efficace dans le cas d'une stratification simple. On obtient cet estimateur en groupant des estimateurs comme celui défini en (6.4) qui s'appliquent à deux groupes ou plus.

Drew et al. (1982) ont proposé un autre estimateur qui dépend de la taille de l'échantillon ainsi que d'un paramètre K_0. Dans le cas de l'échantillonnage aléatoire simple, si l'on choisit $K_0 = 1$, l'estimateur en question devient:

$$\bar{y}_{iD} = \begin{cases} \bar{y}_i \text{ si } w_i \geq W_i \\ \\ \bar{y}_{iS} \text{ si } w_i < W_i. \end{cases} \tag{6.7}$$

Nous avons souligné plus haut que le choix entre $\bar{y}_{iS}$ et $\bar{y}_{iS}^*$ ne s'imposait pas à l'évidence lorsque $w_i < W_i$. Il en va de même du choix entre $\bar{y}_{iD}$ et $\bar{y}_{iS}^{**}$.

Si N_i est inconnue, l'inférence conditionnelle est encore possible pourvu que N_i n'ait aucun rapport avec le paramètre étudié $\bar{Y}_i$. Elle est également faisable lorsqu'on dispose de renseignements partiels sur N_i (si l'on connaît, par exemple, les bornes de N_i).

S'il existe une variable concomitant x ayant une moyenne de domaine $\bar{X}_i$ connue, l'estimateur par quotient

$$\bar{y}_{ir} = \frac{\bar{y}_i}{\bar{x}_i} \bar{X}_i \tag{6.8}$$

et l'estimateur par régression (Battese et Fuller 1981)

$$\bar{y}_{ir}^* = \bar{y}_i + \frac{\bar{y}}{\bar{x}} (\bar{X}_i - \bar{x}_i) \tag{6.9}$$

sont tous deux conditionnellement non biaisés (approximativement), mais $\bar{y}_{ir}^*$ sera vraisemblablement plus efficient s'il est possible d'appliquer, du moins approximativement, une équation de régression de même pente (dont la courbe passe par l'origine) aux petites régions. Si les pentes diffèrent, il conviendrait mieux d'utiliser un estimateur empirique de Bayes, qui est plus complexe (Dempster *et coll.*, 1981).

6.2 Total d'un domaine

Si N_i est connue, on peut facilement obtenir une estimation du total d'un domaine $Y_i = N_i \bar{Y}_i$ en multipliant un estimateur choisi de $\bar{Y}_i$ par N_i. Si, par ailleurs, N_i est inconnue, on se sert de l'estimateur non biaisé habituel

$$\hat{Y}_i = \hat{N}_i \bar{y}_i = \frac{N}{n} \sum_{j \in s_i} Y_j, \; n_i \geq 1 \tag{6.9}$$

où $\hat{N}_i = N w_i$ est l'estimateur sans biais de N_i et $P(n_i = 0)$ est supposé négligeable.

Supposons, maintenant, que nous disposions de renseignements préalables, par exemple $N_i^* \leq N_i \leq N_i^{**}$. L'inférence conditionnelle est alors possible. Le biais conditionnel de $\hat{Y}_i$ est

$$B_2(\hat{Y}_i) = (\hat{N}_i - N_i)\bar{Y}_i. \tag{6.10}$$

L'équation ci-dessus nous permet d'affirmer (en supposant que $\bar{Y}_i > 0$) que $B_2(\hat{Y}_i) > 0$, c'est-à-dire, qu'il y a surestimation si $\hat{N}_i > N_i$ et que $B_2(\hat{Y}_i)$ s'accroît lorsque la taille n_i de l'échantillon du domaine augmente. De même, $B_2(\hat{Y}_i) < 0$, c'est-à-dire qu'il y a sous-estimation si $\hat{N}_i < N_i$ et $|B_2(\hat{Y}_i)|$ augmente lorsque n_i diminue; le biais conditionnel est nul si $\hat{N}_i = N_i$.

À l'aide des renseignements préalables, il est possible de modifier $\hat{Y}_i$:

$$\hat{Y}_i^* = \begin{cases} N_i^* \bar{y}_i & \text{si } \hat{N}_i < N_i^* \\ \hat{N}_i \bar{y}_i & \text{si } N_i^* \leq \hat{N}_i \leq N_i^{**} \\ N_i^{**} \bar{y}_i & \text{si } \hat{N}_i > N_i^{**}. \end{cases} \tag{6.11}$$

Le biais conditionnel de $\hat{Y}_i^*$est inférieur, en valeur absolue, à celui de $\hat{Y}_i$ si $\hat{N}_i < N_i^*$ ou $\hat{N}_i > N_i^{**}$, tandis que $\hat{Y}_i^* = \hat{Y}_i$ dans l'intervalle $N_i^* \leq \hat{N}_i \leq N_i^{**}$. En conséquence, $\hat{Y}_i^*$ est conditionnellement plus efficace que l'estimateur non biaisé $\hat{Y}_i$. De plus, l'EQM inconditionnelle de $\hat{Y}_i^*$est moindre que celle de $\hat{Y}_i$, même si $\hat{Y}_i^*$ est inconditionnellement biaisé. Malheureusement, il n'existe pas de moyen facile d'améliorer l'efficience de $\hat{Y}_i^*$ dans l'intervalle $N_i^* \leq \hat{N}_i \leq N_i^{**}$. Quoi qu'il en soit, $\hat{Y}_i^*$ devrait être préféré à $\hat{Y}_i$. pour que l'estimation du total d'un domaine soit efficiente, il faut disposer de bons renseignements supplémentaires sur la taille du domaine.

7. PLANS DE SONDAGE GÉNÉRAUX

L'adjustement par stratification a posteriori est couramment utilisé dans les grandes enquêtes complexes afin, surtout, d'accroître l'efficience des estimateurs. Mentionnons, à titre d'exemple, le facteur âge-sexe dans l'enquête sur la population active du Canada (EPA). Il existe également une théorie générale de l'inférence inconditionnelle. L'estimateur du total Y est défini par:

$$\hat{Y}_{pst} = \sum M_i \frac{\hat{Y}_i}{\hat{M}_i} \tag{7.1}$$

où $\hat{Y}_i$ et $\hat{M}_i$ sont respectivement les estimateurs non biaisés habituels du total Y_i et de la taille M_i de la i-ième strate formée a posteriori. Dans l'EPA, on se sert des données projetées du recensement pour les M_i. L'estimateur $\hat{Y}_{pst}$ se réduit à $\sum N_i \bar{y}_i$ dans le cas d'un échantillonnage aléatoire simple (voir (3.2)), et nous avons vu précédemment que $\sum N_i \bar{y}_i$ est conditionnellement non biaisé dans un échantillon aléatoire simple (en supposant que $n_i \geq 1$ pour tous les i). Toutefois, dans les plans de sondage complexes, il semble difficile d'analyser les propriétés conditionnelles de (7.1); même le choix d'un ensemble de référence n'est pas aussi évident que l'on croit. Pour illustrer ce problème, considérons un échantillonnage aléatoire simple stratifié où $L = 2$ strates et $k = 2$ strates formées a posteriori. Si nous faisons reposer l'inférence conditionnelle sur les tailles d'échantillon (n_{h1}, n_{h2}) observées dans chaque strate h formée a posteriori, la théorie s'applique normalement pourvu que l'on connaisse les tailles N_{hi} des strates formées a posteriori. En pratique, toutefois, nous risquons d'éprouver des difficultés si nous avons des tailles n_{hi} d'échantillon nulles; il se peut aussi que nous ne connaissions pas les tailles N_{hi} des strates ou que les projections soient inexactes même si nous avons $N_{.i} = \sum_h N_{hi} = M_i$. il serait alors préférable de faire reposer l'inférence conditionnelle sur la taille totale observée $(n_{.1}, n_{.2})$ des échantillons, où $n_{.i} = \sum_h n_{hi}$.

Dans ce cas particulier d'échantillonnage aléatoire simple stratifié ($L = 2$, $k = 2$), l'estimateur $\hat{Y}_{pst}$ s'exprime comme suit:

$$\hat{Y}_{pst} = N_{.1} \frac{N_{1.} \dfrac{y_{11}}{n_{1.}} + N_{2.} \dfrac{y_{21}}{n_{2.}}}{N_{1.} \dfrac{n_{11}}{n_{1.}} + N_{2.} \dfrac{n_{21}}{n_{2.}}} + N_{.2} \frac{N_{1.} \dfrac{y_{12}}{n_{1.}} + N_{2.} \dfrac{y_{22}}{n_{2.}}}{N_{1.} \dfrac{n_{12}}{n_{1.}} + N_{2.} \dfrac{n_{22}}{n_{2.}}} \tag{7.2}$$

où $N_{h.} = N_{h1} + N_{h2}$ et $n_{h.} = n_{h1} + n_{h2}$ désignent respectivement la taille de la population et la taille de l'échantillon de la strate et où y_{hi} est le total de l'échantillon dans la (h, i)-ième cellule. Étant donné $(n_{.1}, n_{.2})$, l'espérance conditionnelle de (7.2) ne peut être déterminée puisqu'il faut faire la somme

$$E_2(\hat{Y}_{pst}) = \sum_t p(s_t|n_{.1}, n_{.2})\hat{Y}_{pst}(t) \tag{7.3}$$

où s_t est un échantillon probable tel que la taille observée $\tilde{n}_{hi}$ des échantillons satisfait l'équation $\tilde{n}_{1i} + \tilde{n}_{2i} = n_{.i}$ $(i = 1, 2)$, et $\hat{Y}_{pst}(t)$ est la valeur de l'équation (7.2) pour l'échantillon s_t, et $p(s_t|n_{.1}, n_{.2})$ est la probabilité conditionnelle que s_t soit observé étant donné $(n_{.1}, n_{.2})$:

$$p(s_t|n_{.1}, n_{.2}) = \left[\sum_{n_{11}=0}^{n_1} \binom{N_{11}}{n_{11}} \binom{N_{12}}{n_{1.}-n_{11}} \binom{N_{21}}{n_{.1}-n_{11}} \binom{N_{22}}{n_{2.}-n_{.1}+n_{11}} \right]^{-1}. \tag{7.4}$$

Cependant, les équations (7.3) et (7.4) nous indiquent clairement que $E_2(\hat{Y}_{pst}) \neq Y$ puisque, contrairement à $p(s_t|n_{.1}, n_{.2})$, $\hat{Y}_{pst}$ ne dépend pas des totaux N_{hi} des cellules.

Pour ce qui a trait à l'estimation de la variance, la formule utilisée habituellement dans les plans de sondage généraux est exprimée:

$$v^*(\hat{Y}_{pst}) = v(z_t^*) \tag{7.5}$$

où $v(y_t) = v(\hat{Y})$ est l'estimateur habituel de la variance du total estimé $\hat{Y}$; on obtient $v(z_t^*)$ à partir de $v(\hat{Y})$ en remplaçant y_t par

$$z_t^* = y_t - \sum_i \frac{\hat{Y}_i}{M_i} a_t(i) \tag{7.6}$$

où $a_t(i) = 1$ si le t-ième élément appartient à la i-ième strate formée a posteriori; sinon $a_t(i) = 0$ (Williams 1962). Dans le cas d'un échantillonnage aléatoire simple, l'équation (7.5) devient

$$v^*(\hat{Y}_{pst}) \doteq N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \sum n_i s_{iy}^2 \tag{7.7}$$

(en supposant que $(n_i - 1)/(n-1) \doteq n_i/n$); l'équation ci-dessus n'est pas égale à l'équation (3.3) lorsqu'elle est multipliée par N^2. Donc, l'équation (7.5) ne convient pas très bien au cadre d'analyse conditionnel, même lorsqu'il s'agit d'un échantillonnage aléatoire simple. par ailleurs, un nouvel estimateur de variance

$$v(\hat{Y}_{pst}) = v(z_t), \tag{7.8}$$

où

$$z_t = \sum_i \frac{M_i}{\hat{M}_i} \left(y_t(i) - \frac{\hat{Y}_i}{\hat{M}_i} a_t(i) \right) \tag{7.9}$$

et $y_t(i) = y_t$ si le t-ième élément appartient à la i-ième strate formée a posteriori,

sinon $y_t(i) = 0$, pourrait mieux convenir que $v^*(\hat{Y}_{pst})$ puisque, dans le cas d'un échantillonnage aléatoire simple il se réduit à (3.3) lorsque multiplié par N^2 et qu'il n'est pas nécessaire d'introduire une correction d'échantillonnage pour population finie

$$v(\hat{Y}_{pst}) = \sum_i \frac{N_i^2}{n_i} s_{iy}^2. \tag{7.10}$$

En outre, d'après une théorie des estimateurs par quotient appliqués à des modèles, $v(\hat{Y}_{pst})$ serait conditionnellement plus efficient que $v^*(\hat{Y}_{pst})$. Quoi qu'il en soit, il n'y a pas de mal à utiliser l'estimateur (7.8) puisque, en l'absence de toute condition, il est asymptotiquement équivalent à l'estimateur habituel de la variance (équation 7.5).

8. ANALYSE

La présente étude montre clairement que l'inférence conditionnelle dans les plans de sondage complexes soulève d'énormes problèmes. Il ne faut pas pour autant utiliser aveuglément les méthodes classiques. Lorsque l'inférence conditionnelle est possible, comme dans le cas de l'échantillonnage aléatoire simple, nous devrions certes utiliser des méthodes adaptées à un cadre d'analyse conditionnel, comme celles qui sont décrites dans les sections 2 à 6, tandis que dans les cas plus complexes, nous devrions au moins modifier légèrement les méthodes traditionnelles, comme dans le cas de l'équation (7.6), afin qu'elles soient compatibles avec les résultats conditionnellement justes observés dans des cas particuliers. De toute évidence, ce domaine doit faire l'objet d'autres recherches.

REMERCIEMENTS

Ce document est basé sur les exposés que j'ai faits à un atelier de travail sur l'inférence conditionnelle dans les enquêtes par sondage. Je tiens à remercier Mme Nanjamma Chinnappa pour avoir organisé l'atelier à Statistique Canada. Je remercie également mes collègues de Statistique Canada et le professeur D. Holt pour les commentaires utiles et constructifs qu'ils m'ont donnés au cours de la rédaction de ce document.

BIBLIOGRAPHIE

BANKIER, M. (1985). Conditionally unbiased estimators based on any number of independent stratified samples. Mémoire, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

BATTESE, G.E., et FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 500-505.

BRYANT, E.C., HARTLEY, H.O., et JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association,* 55, 105-124.

CHINNAPPA, B.N. (1976). A preliminary note on methods of dealing with unusually large units in sampling from skew populations. Document technique non publié, Division des méthodes d'enquêtes-institutions et agriculture, Statistics Canada.

COX, D.R., et HINKLEY, D.V. (1974). *Theoretical Statistics.* Londres, Chapman et Hall.

DEMPSTER, A.P., RUBIN, D.B., et TSUTAKAWA, R.K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association,* 76, 341-353.

DOSS, D.C., HARTLEY, H.O., et SOMAYAJULU, G.R. (1979). An exact small sample theory for post-stratification. *Journal of Statistical Planning and Inference,* 3, 235-248.

DREW, J.H., SINGH, M.P., et CHOUDHRY, H. (1982). Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada. *Techniques d'enquêtes,* 8, 17-47.

DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. Dans *New Developments in Survey Sampling* (Eds. N.L. Johnson et H. Smith), New York: Wiley - Interscience.

FISHER, R.A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd (5e éd., 1934).

FULLER, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association,* 61, 1172-1183.

HARTLEY, H.O., RAO, J.N.K., et KIEFER, G. (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association,* 64, 841-851.

HIDIROGLOU, M.H., et SÄRNDAL, C.E. (1985). Étude empirique de quelques estimateurs de régression pour petits domaines. *Techniques d'enquête,* 11, 65-77.

HIDIROGLOU, M.H., et SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association,* 76, 690-695.

HOLT, D., et SMITH, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society,* Série A, 142, 33-46.

LAHIRI, D.B. (1969). On the unique sample, the surveyed one. Document technique non publié, Indian Statistical Institute.

OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys,* vol. 2, Academic Press, 142-184.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika,* 57, 377-387.

SARNDAL, C.E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association,* 79, 624-631.

WILLIAMS, W.H. (1962). The variance of an estimator with post-stratified weighting. *Journal of the American Statistical Association,* 57, 622-627.

YATES, F. (1984). Tests of significance for 2 $\times$ 2 contingency tables. *Journal of the Royal Statistical Society,* Série A, 147, 426-463.