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

Article

Estimateurs de variance par linéarisation pour les paramètres de modèles à partir de données d'enquêtes complexes

par Abdellatif Demnati et J.N.K. Rao

Décembre 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2010
Vol. 36, N° 2, pp. 211-220
Statistique Canada, N° 12-001-X au catalogue

Estimateurs de variance par linéarisation pour les paramètres de modèles à partir de données d'enquêtes complexes

Abdellatif Demnati et J.N.K. Rao [1]

Résumé

Les méthodes de linéarisation de Taylor sont souvent utilisées pour obtenir des estimateurs de la variance d'estimateurs par calage de totaux et de paramètres de population finie (ou de recensement) non linéaires, tels que des ratios ou des coefficients de régression et de corrélation, qui peuvent être exprimés sous forme de fonctions lisses de totaux. La linéarisation de Taylor s'applique généralement à tout plan d'échantillonnage, mais elle peut produire de multiples estimateurs de la variance qui sont asymptotiquement sans biais par rapport au plan en cas d'échantillonnage répété. Le choix parmi les estimateurs de variance doit donc s'appuyer sur d'autres critères, tels que i) l'absence approximative de biais dans la variance par rapport au modèle de l'estimateur obtenu sous un modèle hypothétique et ii) la validité sous échantillonnage répété conditionnel. Demnati et Rao (2004) ont proposé une méthode unifiée de calcul des estimateurs de variance par linéarisation de Taylor produisant directement un estimateur de variance unique qui satisfait aux critères susmentionnés pour des plans de sondage généraux. Dans l'analyse des données d'enquête, on suppose généralement que les populations finies sont générées au moyen de modèles de superpopulation et l'on s'intéresse aux inférences analytiques concernant les paramètres de ces modèles. Si les fractions d'échantillonnage sont faibles, la variance d'échantillonnage reflète presque toute la variation due aux processus aléatoires liés au plan de sondage et au modèle. Par contre, si les fractions d'échantillonnage ne sont pas négligeables, il faut tenir compte de la variance du modèle pour construire des inférences valides concernant les paramètres du modèle sous le processus combiné de génération de la population finie à partir du modèle hypothétique de superpopulation et de sélection de l'échantillon conformément au plan de l'échantillonnage spécifié. Dans le présent article, nous obtenons un estimateur de la variance totale selon l'approche de Demnati-Rao en supposant que les caractéristiques d'intérêt sont des variables aléatoires générées au moyen d'un modèle de superpopulation. Nous illustrons la méthode à l'aide d'estimateurs par le ratio et d'estimateurs définis comme des solutions d'équations d'estimation pondérées par calage. Nous présentons aussi les résultats de simulations en vue de déterminer la performance de l'estimateur de variance proposé pour les paramètres du modèle.

Mots clés : Calage ; estimateurs par le ratio ; variance totale ; régression logistique ; équations d'estimation pondérées.

1. Introduction

Dans les sondages, on s'intéresse souvent à l'estimation d'un total de population finie $Y = \sum_{k=1}^{N} y_k \equiv Y(y)$, où N est la taille de la population finie. Sous un plan d'échantillonnage général avec probabilités d'inclusion positives π_k, un estimateur sans biais sous le plan du total Y est habituellement donné par $\hat{Y} = \sum_{i \in s} y_i/\pi_i \equiv \sum_{k=1}^{N} d_k(s) y_k$, où s est un échantillon, $d_k(s) = a_k(s)/\pi_k$ sont les poids de sondage avec $a_k(s) = 1$ si $k \in s$ et $a_k(s) = 0$ autrement. Nous utilisons la notation opérationnelle et écrivons $\hat{Y}(z) = \sum_{k=1}^{N} d_k(s) z_k$ de sorte que $\hat{Y} = \hat{Y}(y)$. Donc, toutes les sommes étant considérées sur l'ensemble de la population, nous écrivons $\sum_{k=1}^{N} y_k = \sum y_k$ et $\hat{Y}(z) = \sum d_k(s) z_k$ pour simplifier la notation. De nouveau, en utilisant la notation opérationnelle, nous notons un estimateur sans biais de la variance de $\hat{Y}(z)$ comme une fonction quadratique, $\vartheta(z)$, en termes des z_k.

Des estimateurs plus complexes d'un total Y basés sur des données auxiliaires connues de population, tels que les estimateurs par le ratio et par la régression, et des estimateurs de paramètres plus complexes obtenus comme solutions d'équations d'estimation pondérées d'échantillon, tels que les estimateurs des coefficients de régression logistique en population finie, sont aussi utilisés souvent en pratique. Les estimateurs qui peuvent être exprimés sous forme d'une fonctionnalité générale $T(\hat{M})$, où $\hat{M}$ désigne une mesure qui attribue le poids $d_k(s)$ à y_k, ont également été étudiés ; par exemple, $T(\hat{M}) = \int x d\hat{M}(x) = \sum d_k(s) y_k$ si le paramètre de population est le total $T(M) = \int x dM(x) = Y$, où la mesure M attribue une masse unitaire à chaque y_k (Deville 1999). L'estimation en grand échantillon de la variance d'estimateurs aussi complexes, disons $\hat{\theta}$, a été discutée abondamment dans la littérature spécialisée. En particulier, les méthodes d'estimation de la variance de $\hat{\theta}$ par linéarisation de Taylor sont généralement applicables à tout plan d'échantillonnage qui permet d'utiliser un estimateur de variance sans biais $\vartheta(z)$ de $\hat{Y}(z)$. Binder (1983) a étudié les estimateurs $\hat{\theta}$ qui sont des solutions d'équations d'estimation pondérées et a appliqué la linéarisation de Taylor pour obtenir un estimateur de variance qui peut être exprimé comme $\vartheta(\tilde{z})$, où la variable linéarisée $\tilde{z}_k$ dépend de paramètres inconnus, et $\tilde{z}_k$ est remplacée par un

1. Abdellatif Demnati, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6. Courriel : Abdellatif.Demnati@statcan.gc.ca ; J.N.K. Rao, École de mathématique et de statistique, Université Carleton, Ottawa (Ontario) Canada, K1S 5B6. Courriel : JRao@math.carleton.ca.

estimateur z_k qui peut être fondé sur la méthode de substitution. Deville (1999) a dérivé un estimateur de variance par linéarisation de Taylor de la fonctionnelle $T(\hat{M})$ de la forme $\vartheta(\tilde{z})$, où $\tilde{z}_k = I_T(M\,;\, y_k)$ désigne la fonction d'influence de T à la valeur y_k, puis a remplacé $\tilde{z}_k$ par l'estimateur d'échantillon $z_{k1} = I_T(\hat{M}\,;\, y_k)$. Par exemple, quand $\hat{\theta}$ est l'estimateur par le ratio $(\hat{Y}/\hat{X})X = \hat{R}X$ du total Y, où $\hat{X} = \hat{Y}(x)$ et $X = Y(x)$ est le total connu d'une variable auxiliaire x, nous obtenons $\tilde{z}_k = y_k - Rx_k$ et $z_{k1} = y_k - \hat{R}x_k$. Cependant, $z_k = (X/\hat{X})(y_k - \hat{R}x_k)$ est aussi un candidat pour estimer $\tilde{z}_k$ et l'estimateur de variance résultant $\vartheta(z)$ est souvent préféré à $\vartheta(z_1)$; voir Demnati et Rao (2004). Donc, sous l'approche de Deville, le choix d'un estimateur de $\tilde{z}_k$ est dans une certaine mesure arbitraire.

Demnati et Rao (2004) ont étudié des estimateurs généraux qui peuvent être exprimés sous forme de fonctions lisses des poids $\boldsymbol{d}(s) = \{d_1(s), ..., d_N(s)\}^T$, disons $\hat{\theta} = f(\boldsymbol{d}(s))$, et ont obtenu un estimateur de variance par linéarisation de Taylor directement sous la forme $\vartheta(z)$ avec des variables linéarisées connues $z_k = \partial f(b)/\partial b_k\,|_{\boldsymbol{b}=\boldsymbol{d}(s)}$ sans estimer la variable $\tilde{z}_k$ au préalable, puis la remplacer par un estimateur. Par exemple, dans le cas de l'estimateur par le ratio, leur méthode mène automatiquement à z_k susmentionné. Cette méthode peut être appliquée à divers estimateurs, y compris ceux des paramètres de régression logistique en population finie fondés sur les poids de calage (Demnati et Rao 2004). Les travaux antérieurs sur l'estimation directe de la variance comprennent ceux de Binder (1996).

Dans l'analyse des données d'enquête, on suppose fréquemment que les valeurs de population y_k, $k = 1, ..., N$, sont issues d'un modèle de superpopulation, et l'utilisateur cherche souvent à faire des inférences au sujet des paramètres du modèle. Soit θ_N un paramètre de population finie, c'est-à-dire un estimateur d'un paramètre du modèle θ quand les valeurs de population y_k sont toutes connues, et soit $\hat{\theta}$ un estimateur sans biais sous le plan de θ_N, le paramètre de population finie. Supposons que $\hat{\theta}$ est sans biais sous le plan et sous le modèle pour θ, c'est-à-dire que $E_m E_p(\hat{\theta}) = \theta$, où E_m et E_p désignent les espérances sous le plan et sous le modèle, respectivement. Alors, la variance totale de $\hat{\theta}$ est $V(\hat{\theta}) = E_m E_p(\hat{\theta} - \theta)^2$, qui peut être décomposée comme il suit

$$V(\hat{\theta}) = E_m V_p(\hat{\theta}) + V_m(\theta_N), \qquad (1.1)$$

où $V_p(\hat{\theta}) = E_p(\hat{\theta} - \theta_N)^2$ est la variance de $\hat{\theta}$ sous le plan et $V_m(\theta_N)$ est la variance de θ_N sous le modèle. Il découle de (1.1) que la variance totale peut être estimée en utilisant un estimateur fondé sur le plan de $V_p(\hat{\theta})$ si le dernier terme $V_m(\theta_N)$ est négligeable comparativement à $E_m V_p(\hat{\theta})$. Dans ce cas, la distinction entre θ_N et θ peut être ignorée (Skinner, Holt et Smith 1989, page 14). Par ailleurs, il est nécessaire d'estimer la variance totale $V(\hat{\theta})$ quand la variance sous le modèle $V_m(\theta_N)$ n'est pas négligeable comparativement à $E_m V_p(\hat{\theta})$. Il faut pour cela prendre en considération conjointement les processus aléatoires sous le plan et sous le modèle. Molina, Smith et Sugden (2001) soutiennent que le processus combiné de génération de la population finie et de sélection de l'échantillon devrait servir de fondement aux inférences analytiques concernant les paramètres du modèle. Rubin-Bleuer et Şchiopu-Kratina (2005) ont donné un cadre mathématique pour l'inférence conjointe sous le modèle et sous le plan. Cependant, une méthode dont l'application est générale est nécessaire pour l'estimation de la variance totale. Le principal objectif du présent article est de proposer une telle méthode, en étendant l'approche de Demnati-Rao aux paramètres de population finie.

À la section 2, nous considérons le cas d'un paramètre scalaire θ et présentons des estimateurs de variance par linéarisation, en élargissant l'approche de Demnati et Rao (2004). Nous illustrons la méthode pour le cas particulier d'un estimateur par le ratio d'une moyenne de superpopulation θ. À la section 3, nous étendons les résultats de la section 2 aux estimateurs d'un paramètre vectoriel $\boldsymbol{\theta}$ dont les valeurs sont les solutions d'équations d'estimation pondérées et nous illustrons la méthode pour le cas particulier des paramètres d'un modèle de régression logistique. Nous présentons aussi les résultats de simulations.

2. Paramètre du modèle scalaire

2.1 Estimateurs ponctuels

Considérons une population finie U de N éléments, et soit $d_k(s) = a_k(s)/\pi_k$ le poids de sondage attaché à l'élément de population k, où $a_k(s) = 1$ si l'élément k est compris dans l'échantillon s et $a_k(s) = 0$ autrement, et π_k est la probabilité d'inclusion associée à l'élément k. Nous considérons les estimateurs $\hat{\theta}$ d'un paramètre scalaire θ qui peuvent être exprimés comme des fonctions de variables aléatoires sous le plan et le modèle supposé. En particulier, $\hat{\theta} = f(\boldsymbol{A}_d)$, où $\boldsymbol{A}_d$ est une matrice $(p+1) \times N$ avec les colonnes $\boldsymbol{d}_k = (d_k h_{1k}, d_k h_{2k}, ..., d_k h_{(p+1)k})^T \equiv (d_{1k}, ..., d_{(p+1)k})^T$ où $d_k = d_k(s)$ est aléatoire sous le plan, $h_{1k} = 1$ et h_{ik} $(i = 2, ..., p+1)$ sont aléatoires sous le modèle.

Par exemple, considérons le modèle du ratio avec covariables x_k fixes :

$$E_m(y_k) = \beta x_k,\ V_m(y_k) = \sigma^2 x_k,\ \mathrm{Cov}_m(y_k,\ y_t) = 0,\quad k \neq t,\ k, t = 1, ..., N, \qquad (2.1)$$

où E_m, V_m et Cov_m désignent l'espérance sous le modèle, la variance sous le modèle et la covariance sous le modèle,

respectivement, et $\sigma^2 > 0$. Supposons que nous voulons estimer la moyenne de superpopulation $\theta = E_m(\bar{Y}) = N^{-1}\sum E_m(y_k) = \beta\bar{X}$, où $\bar{Y}$ est la moyenne de population finie de y. Dans ce cas, un estimateur par le ratio de θ est donné par

$$\hat{\theta} = \bar{X}(\hat{Y}/\hat{X}) \equiv \bar{X}\hat{R}, \qquad (2.2)$$

où $\hat{Y} = \sum d_k(s)y_k$ et $\hat{X} = \sum d_k(s)x_k$ sont les estimateurs sans biais sous le plan des totaux Y et X, et $\bar{X}$ est la moyenne de population connue de x. Nous pouvons écrire l'estimateur par le ratio (2.2) sous la forme $\hat{\theta} = \bar{X}(\sum d_{2k})/\sum d_{1k}x_k$, où $d_{1k} = d_k(s)$ et $d_{2k} = d_k(s)y_k$. Il s'agit d'un cas particulier de $f(\boldsymbol{A}_d)$ avec $p = 1$ et $h_{2k} = y_k$.

Soit E_p l'espérance sous le plan et $E = E_m E_p$, l'espérance totale. Alors, nous avons $E(d_{1k}) = E_m(1) = 1 \equiv \mu_{1k}$ et $E(d_{ik}) = E_m(g_{ik}) \equiv \mu_{ik}$, $i = 2, ..., p+1$, en notant que $E_p(d_k(s)) = 1$. Nous supposons que $f(\boldsymbol{A}_\mu) = \theta$, où $\boldsymbol{A}_\mu$ est une matrice $(p+1)\times N$ avec les colonnes $\boldsymbol{\mu}_k = (\mu_{1k}, \mu_{2k}, ..., \mu_{(p+1)k})^T$. Donc, $\hat{\theta}$ est asymptotiquement sans biais sous le plan p et sous le modèle m pour θ. Dans le cas particulier de l'estimateur par le ratio, nous avons $f(\boldsymbol{A}_\mu) = \beta\bar{X} = \theta$, en notant que $\mu_{1k} = 1$ et $\mu_{2k} = \beta x_k$.

2.2 Estimateur de variance par linéarisation

Nous commençons par dériver un estimateur de la variance totale d'un estimateur linéaire $\hat{U} = \sum \boldsymbol{u}_k^T \boldsymbol{d}_k$, où $\boldsymbol{u}_k$ est un vecteur de constantes. La variance totale de $\hat{U}$ peut être décomposée comme il suit

$$V(\hat{U}) = E_m V_p(\hat{U}) + V_m E_p(\hat{U}) \equiv I + II, \qquad (2.3)$$

où V_p et V_m désigne la variance sous le plan et la variance sous le modèle, respectivement. Un estimateur sans biais sous le plan de la composante I de la variance totale (2.3) s'obtient en estimant la variance sous le plan $V_p(\hat{U})$ pour des $\boldsymbol{h}_k = (h_{1k}, ..., h_{(p+1)k})^T$ fixes. Maintenant, en notant que $\hat{U} = \sum b_k d_k(s)$ est l'estimateur classique de Narain-Horvitz-Thompson (NHT) du total $U = \sum b_k$ quand les $b_k = \boldsymbol{u}_k^T \boldsymbol{h}_k$ sont fixes conditionnellement, nous pouvons utiliser soit l'estimateur de variance de Sen-Yates-Grandy (SYG) pour des plans avec tailles d'échantillon fixes ou l'estimateur de variance d'Horvitz-Thompson (HT) pour des plans arbitraires. L'estimateur SYG est donné par

$$\text{est}(I) = \vartheta_{\text{SYG}}(\hat{U}) = \sum\sum_{k<t} d_{kt}(s)\frac{(\pi_k\pi_t - \pi_{kt})}{\pi_k\pi_t}(b_k - b_t)^2, \qquad (2.4)$$

où $d_{kt}(s) = \{a_k(s)a_t(s)\}/\pi_{kt}$ et π_{kt} est la probabilité d'inclusion des unités k et t ($k \neq t$). L'estimateur de variance HT est donné par

$$\text{est}(I) = \vartheta_{\text{HT}}(\hat{U}) = \sum\sum d_{kt}(s)\frac{(\pi_{kt} - \pi_k\pi_t)}{\pi_k\pi_t}b_k b_t, \qquad (2.5)$$

où $d_{kk}(s) = d_k(s)$. Pour le cas particulier de l'échantillonnage aléatoire stratifié, (2.4) et (2.5) sont identiques.

Si nous nous tournons vers la composante II de la variance totale (2.3), nous avons $V_m E_p(\hat{U}) = V_m(\sum \boldsymbol{u}_k^T \boldsymbol{h}_k) = \sum\sum \boldsymbol{u}_k^T \text{Cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\boldsymbol{u}_t$ et un estimateur sans biais sous p et m est alors donné par

$$\text{est}(II) = \sum\sum d_{kt}(s)\boldsymbol{u}_k^T \text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\boldsymbol{u}_t, \qquad (2.6)$$

après avoir remplacé $\text{Cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)$ par un estimateur $\text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)$. L'estimateur de la variance totale (2.3) est maintenant donné par $\text{est}(I) + \text{est}(II)$. Nous le désignons, en notation opérationnelle, par $\vartheta(\boldsymbol{u})$.

Penchons-nous maintenant sur l'estimation de la variance totale de $\hat{\theta}$. À l'instar de Demnati et Rao (2004), nous pouvons écrire un développement en série de Taylor de $\hat{\theta} - \theta$ sous la forme

$$\hat{\theta} - \theta \approx \sum \tilde{\boldsymbol{z}}_k^T(\boldsymbol{d}_k - \boldsymbol{\mu}_k) \qquad (2.7)$$

où $\tilde{\boldsymbol{z}}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k|_{\boldsymbol{A}_b = \boldsymbol{A}_\mu}$ et $\boldsymbol{A}_b$ est une matrice $(p+1)\times N$ dont la k^{e} colonne $\boldsymbol{b}_k$ est un vecteur de nombres réels arbitraires. L'approximation (2.7) est valide pour tout $\hat{\theta}$ qui peut être exprimé sous forme d'une fonction lisse des totaux estimés. En suivant l'exemple de Demnati et Rao (2004), nous pouvons maintenant écrire un estimateur par linéarisation de la variance totale sous la forme

$$\vartheta_{\text{DR}}(\hat{\theta}) = \vartheta(\boldsymbol{z}), \qquad (2.8)$$

que nous obtenons à partir de $\vartheta(\boldsymbol{u})$ en remplaçant $\boldsymbol{u}_k$ par la « variable linéarisée » $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$. Une justification théorique rigoureuse de (2.8) ressemble à celle de Deville (1999).

2.3 Cas particulier de l'estimateur par le ratio

Pour l'estimateur par le ratio $\hat{\theta} = \bar{X}\hat{R}$ du paramètre du modèle $\theta = \beta\bar{X}$, $\boldsymbol{z}_k$ se réduit à

$$\boldsymbol{z}_k = (\bar{X}/\hat{X})(-\hat{R}x_k, 1)^T = (z_{1k}, z_{2k})^T. \qquad (2.9)$$

En outre, dans (2.4) ou (2.5), b_k est remplacé par

$$\boldsymbol{z}_k^T \boldsymbol{h}_k = z_{1k} + z_{2k}y_k = (\bar{X}/\hat{X})(y_k - \hat{R}x_k) \equiv (\bar{X}/\hat{X})e_k,$$

en utilisant (2.9). De plus, en remplaçant $\boldsymbol{u}_k$ par $\boldsymbol{z}_k$ dans (2.6), nous obtenons

$$\boldsymbol{z}_k^T \operatorname{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\, \boldsymbol{z}_t = z_{2k} z_{2t} \operatorname{cov}_m(y_k, y_t).$$

Sous le modèle de ratio (2.1) avec la variance sous le modèle $V_m(y_k) = \sigma_k^2,\ k = 1, ..., N$ non spécifié, nous pouvons estimer $\sigma_k^2 = E_m(y_k - \beta x_k)^2$ par $(y_k - \hat{R} x_k)^2$ et poser que $\operatorname{cov}_m(y_k,\ y_t) = 0$, pour $k \neq t$.

Nous étudions maintenant le cas particulier de l'échantillonnage aléatoire simple sans remise. Dans ce cas, aussi bien (2.4) que (2.5) se réduit à

$$\operatorname{est}(I) = \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{1}{n}\left(1 - \frac{n}{N}\right) s_e^2, \qquad (2.10)$$

où $s_e^2 = \sum a_k(s) e_k^2 / (n-1)$, et (2.6) se réduit à

$$\operatorname{est}(II) = \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{(n-1)}{nN} s_e^2. \qquad (2.11)$$

D'où, en utilisant (2.10) et (2.11), l'estimateur de variance (2.8) se réduit à

$$\begin{aligned} \vartheta_{\mathrm{DR}}(\hat{\theta}) &= \operatorname{est}(I) + \operatorname{est}(II) \\ &= \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{1}{n}\frac{N-1}{N} s_e^2. \end{aligned} \qquad (2.12)$$

Il est intéressant de noter que le « poids g » $\bar{X}/\bar{x}$ apparaît automatiquement dans $\vartheta_{\mathrm{DR}}(\hat{\theta})$, donné par (2.12), et que la correction pour population finie $1 - n/N$ est absente dans $\vartheta_{\mathrm{DR}}(\hat{\theta})$ contrairement à $\operatorname{est}(I)$ donné par (2.10).

Suivant l'approche habituelle de l'estimation de la variance totale (voir, par exemple, Korn et Graubard 1998), $V(\hat{\theta})$ s'écrit d'abord

$$\begin{aligned} V(\hat{\theta}) &= E_m V_p(\hat{\theta}) + V_m E_p(\hat{\theta}) \\ &\approx E_m V_p(\hat{\theta}) + V_m(\bar{Y}) \\ &= E_m V_p(\hat{\theta}) + N^{-2} \sum E_m(y_k - \beta\, x_k)^2, \end{aligned} \qquad (2.13)$$

sous le modèle du ratio avec $\sigma_k^2,\ k = 1, ..., N$ non spécifié. Le premier terme $E_m V_p(\hat{\theta})$ de (2.13) est alors estimé au moyen d'un estimateur convergent sous le plan de $V_p(\hat{\theta})$, habituellement (2.10) sans le facteur $g\ (\bar{X}/\bar{x})^2$. Le deuxième terme est estimé par $N^{-2} \sum d_k(s)(y_k - \hat{R} x_k)^2 = (nN)^{-1}(n-1) s_e^2$. La somme des deux termes estimés est alors égale à (2.12) sans le facteur g. Nous désignons cet estimateur de variance habituel par $\vartheta_{\mathrm{cus}}(\hat{\theta})$. Par ailleurs, si (2.10) avec le facteur g est utilisé pour estimer $V_p(\hat{\theta})$, la somme de ce terme estimé et de l'estimateur précédent du deuxième terme mène à un estimateur de variance « hybride »

$$\vartheta_{\mathrm{mix}}(\hat{\theta}) = \operatorname{est}(I) + (nN)^{-1}(n-1) s_e^2,$$

où le terme g est absent dans le dernier terme. Les résultats qui précèdent montrent clairement que le choix de l'estimateur de la variance totale sous l'approche habituelle n'est pas unique, contrairement à la situation sous l'approche proposée.

Si le paramètre d'intérêt est $\beta = \theta / \bar{X}$ au lieu de θ, alors $\hat{\beta} = \hat{\theta} / \bar{X} = \hat{R}$ et $\vartheta_{\mathrm{DR}}(\hat{\beta})$ sous l'échantillonnage aléatoire simple est donné par

$$\vartheta_{\mathrm{DR}}(\hat{\beta}) = \bar{X}^{-2} \vartheta_{\mathrm{DR}}(\hat{\theta}) = \bar{x}^{-2} \frac{1}{n}\frac{N-1}{N} s_e^2. \qquad (2.14)$$

L'approche habituelle aboutit au même estimateur de variance (2.14).

2.4 Étude en simulation

Nous avons effectué une petite étude en simulation pour examiner les propriétés des divers estimateurs de variance, conditionnellement et inconditionnellement à $\hat{X}$. Nous avons d'abord généré $R = 2\,000$ populations finies $\{y_1, ..., y_N\}$, chacune de taille $N = 393$, au moyen du modèle de ratio

$$y_k = 2\, x_k + x_k^{1/2} \varepsilon_k, \qquad (2.15)$$

avec les valeurs indépendantes ε_k tirées de $N(0, 1)$, où les x_k fixes sont les « nombres de lits » pour la population d'hôpitaux étudiée dans Valliant, Dorfman et Royall (2000, page 424-427). Un échantillon aléatoire simple de taille spécifiée n est tiré de chaque population générée. Notre paramètre d'intérêt est $\theta = \beta \bar{X}$, où $\beta = 2$.

L'EQM totale simulée de l'estimateur par le ratio $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$ est calculée comme $M(\hat{\theta}) = R^{-1} \sum_{r=1}^{2\,000} (\hat{\theta}_r - \theta)^2$, où $\hat{\theta}_r$ est la valeur de $\hat{\theta}$ pour le r^{e} échantillon simulé et $(\bar{y},\ \bar{x})$ sont les moyennes d'échantillon. Nous avons estimé la variance totale $\vartheta_{\mathrm{DR}}(\hat{\theta})$, et ses composantes $\vartheta_s = \operatorname{est}(I)$ et $\vartheta_m = \operatorname{est}(II)$ à partir de chaque échantillon simulé r, ainsi que leurs moyennes $\bar{\vartheta}_{\mathrm{DR}}$, $\bar{\vartheta}_s$ et $\bar{\vartheta}_m$ sur r. La figure 1 représente graphiquement la moyenne des estimations de la variance, $\bar{\vartheta}_{\mathrm{DR}}$ et $\bar{\vartheta}_s$, ainsi que l'EQM totale simulée pour $n = 20, 40, ..., 380, 393$. Dans le cas où $n = N$, $\bar{\vartheta}_s = 0$. L'examen de la figure 1 montre que ϑ_{DR} est approximativement sans biais, tandis que ϑ_s donne lieu à une sous-estimation grave à mesure que la taille d'échantillon, n, augmente.

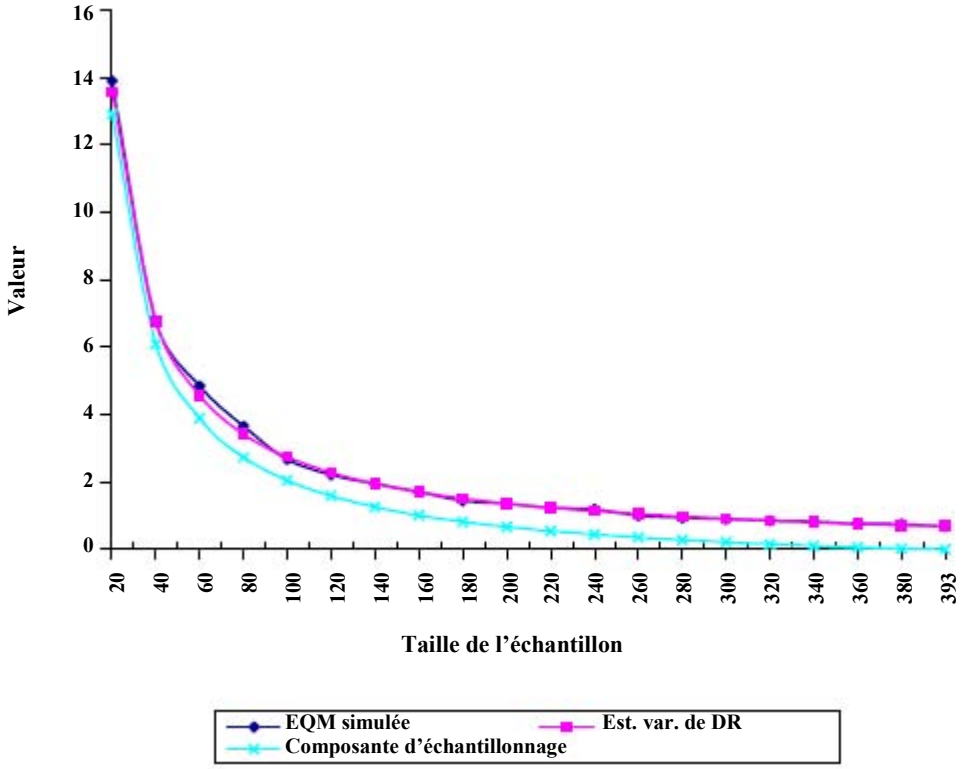

Figure 1 **Moyennes des estimations de variance pour certaines tailles d'échantillons comparativement à l'EQM estimée de l'estimateur par le ratio. ϑ_{DR} = Est. var. de DR, ϑ_s = Composante d'échantillonnage : modèle de ratio**

Nous avons également examiné les propriétés conditionnelles des estimateurs de variance sous échantillonnage aléatoire simple sachant $\bar{x}$, au moyen d'une autre étude en simulation pour l'inférence au sujet de θ, en utilisant le modèle (2.15). L'étude est semblable à celle de Royall et Cumberland (1981) pour l'inférence au sujet de la moyenne de population finie $\theta_N = \bar{Y}$ à partir d'une population fixe $\{y_1, ..., y_N\}$. Nous avons généré $R = 20\,000$ populations finies $\{y_1, ..., y_N\}$, chacune de taille $N = 393$, à partir de (2.15) en utilisant le nombre de lits comme x_k et, pour chaque population, nous avons ensuite sélectionné un échantillon aléatoire simple de taille $n = 100$. Nous avons classé les 20 000 échantillons par ordre croissant de valeur de $\bar{x}$, puis nous les avons regroupés en 20 groupes, chacun de taille 1 000, de façon que le premier groupe, G_1, contienne les 1 000 échantillons ayant les plus petites valeurs de $\bar{x}$, que le groupe suivant, G_2, contienne les 1 000 plus petites valeurs de $\bar{x}$ suivantes, et ainsi de suite pour obtenir $G_1, ..., G_{20}$. Pour chacun des 20 groupes ainsi formés, nous avons calculé la moyenne des estimations du ratio $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$ et l'estimation moyenne $\bar{y}$, ainsi que le biais relatif conditionnel (BRC) résultant en estimant $\theta = 2\bar{X}$; voir la figure 2. Il est clair, si l'on examine cette figure, que $\bar{y}$ est conditionnellement biaisé, contrairement à $\hat{\theta}$: BRC négatif (–14 %) pour G_1 augmentant pour passer à un BRC positif (+14 %) pour G_{20}. Notons que $\bar{y}$ et $\hat{\theta}$ sont tous deux inconditionnellement sans biais pour θ. Le biais conditionnel de $\hat{\theta}$ et de $\bar{y}$ lorsque nous estimons le paramètre du modèle θ est semblable au biais conditionnel dans l'estimation du paramètre de population finie $\theta_N = \bar{Y}$, comme l'ont observé Royall et Cumberland (1981).

Nous avons également calculé l'EQM conditionnelle de $\hat{\theta}$ et le BRC associé des estimateurs de variance ϑ_{DR}, ϑ_{cus} et ϑ_{mix} basés sur les valeurs moyennes de ϑ_{DR}, ϑ_{cus} et ϑ_{mix} dans chaque groupe ; voir la figure 3. Il est évident, si l'on examine cette figure, que le BRC de ϑ_{cus} varie de –28 % à 20 % lorsque l'on passe d'un groupe à l'autre, alors que ϑ_{DR} ne manifeste pas ce genre de tendance et que son BRC est inférieur à 5 % en valeur absolue, sauf pour G_6 et G_{20}. En outre, le BRC de ϑ_{mix} est en grande partie négatif et inférieur à celui de ϑ_{DR} pour la première moitié des groupes et supérieur pour la deuxième moitié, mais ϑ_{mix} ne présente aucune tendance discernable, contrairement à ϑ_{cus}.

La figure 4 donne les taux de couverture conditionnelle (TCC) des intervalles de confiance de la théorie normale fondés sur ϑ_{DR}, ϑ_{cus}, ϑ_{mix} et ϑ_s (en ignorant la composante ϑ_m) au niveau nominal de 95 %. Comme prévu,

l'utilisation de ϑ_s donne lieu à un défaut de couverture important, parce que la fraction d'échantillonnage, 100/393, est grande. Par ailleurs, le TCC associé à ϑ_{DR} est plus proche du niveau nominal dans les divers groupes, tandis que ϑ_{cus} présente une tendance à travers les groupes, le TCC variant de 91 % à 97 %. De surcroît, le TCC associé à ϑ_{mix} est légèrement inférieur à celui associé à ϑ_{DR} pour la première moitié des groupes, mais ϑ_{mix} et ϑ_{DR} donnent des résultats comparables.

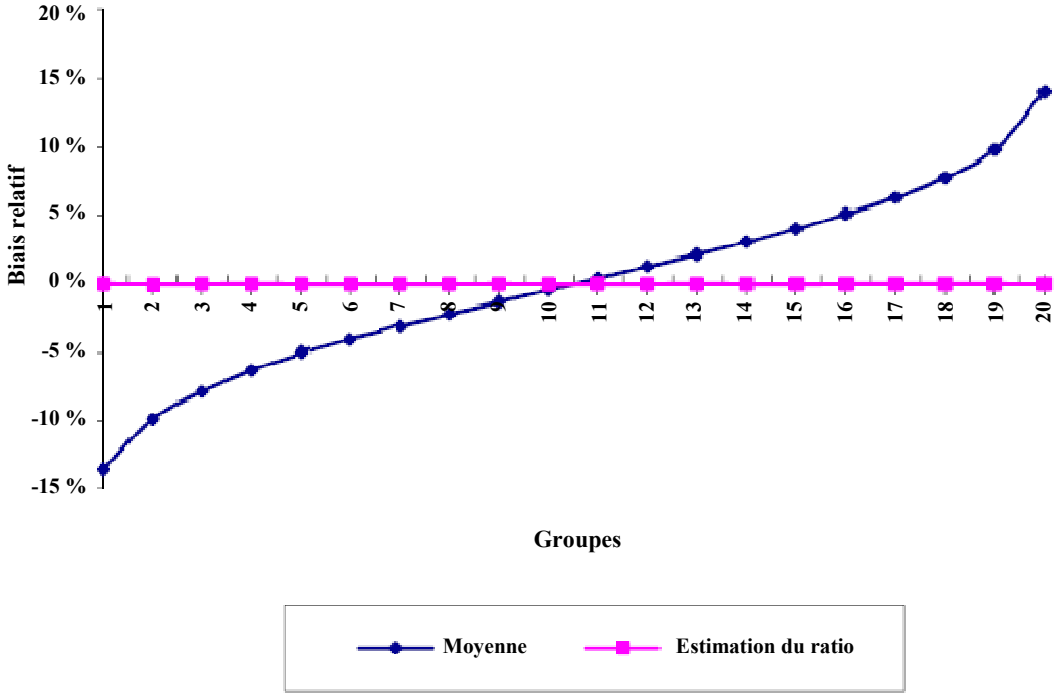

Figure 2 Biais relatif conditionnel de l'estimateur par facteur d'extension et de l'estimateur par le ratio : modèle de ratio

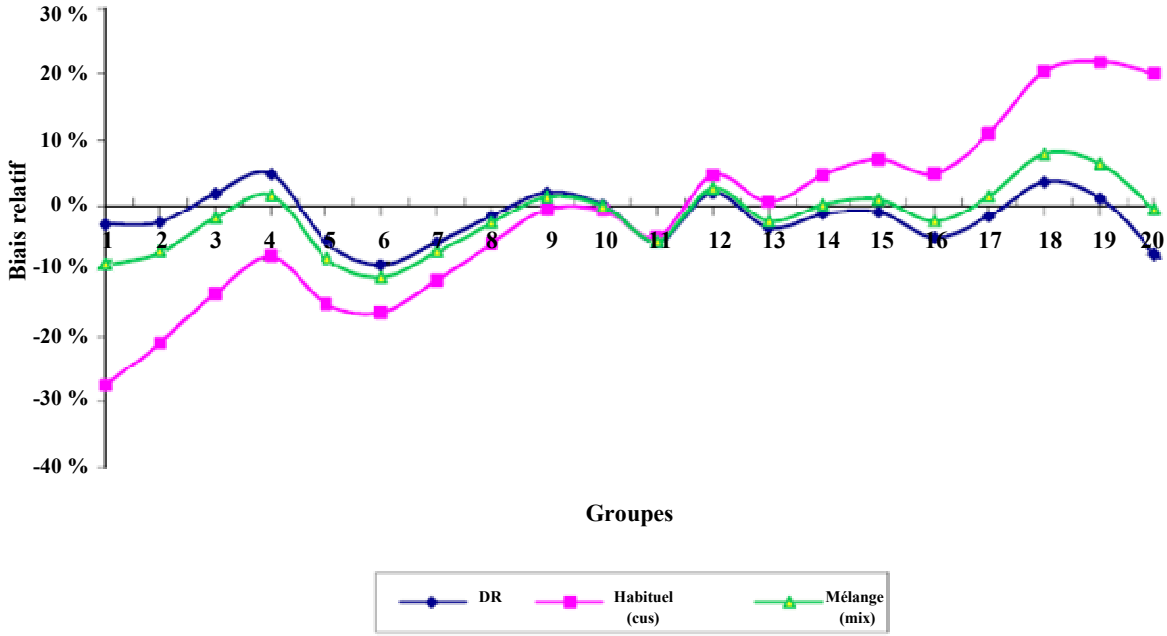

Figure 3 Biais relatif conditionnel des estimateurs de variance ϑ_{DR}, ϑ_{cus} et ϑ_{mix} : modèle de ratio

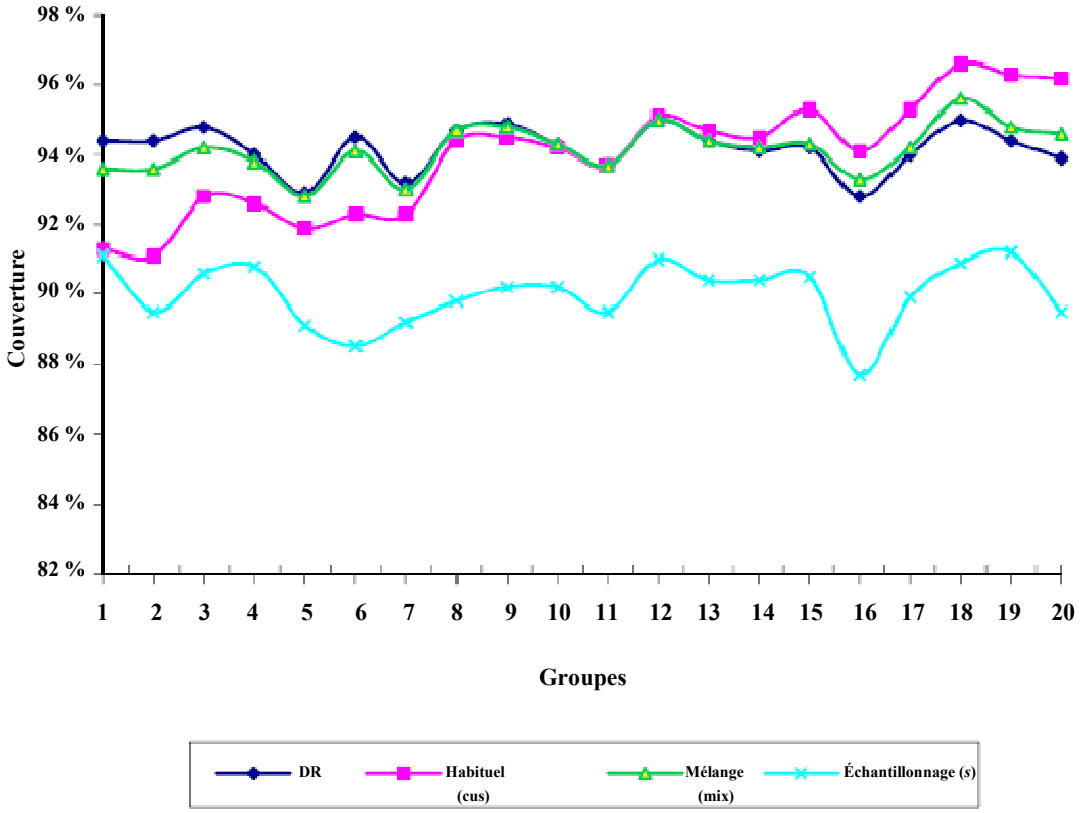

Figure 4 Taux de couverture conditionnels des intervalles de confiance de la théorie normale basés sur ϑ_{DR}, ϑ_{cus}, ϑ_{mix} et ϑ_s pour le niveau nominal de 95 % : modèle de ratio

3. Équations d'estimation pondérées par calage

3.1 Estimateurs des paramètres du modèle

Supposons que le modèle de superpopulation appliqué aux réponses y_k est spécifié par un modèle linéaire généralisé (McCullagh et Nelder 1989) de moyenne $E_m(y_k) = \mu_k(\boldsymbol{\theta}) = h(\boldsymbol{x}_k^T\boldsymbol{\theta})$, où $\boldsymbol{x}_k$ est un vecteur $p \times 1$ de variables explicatives, $\boldsymbol{\theta}$ est le vecteur p des paramètres du modèle et $h(.)$ est une fonction « lien ». Par exemple, $h(a) = a$ donne un modèle de régression linéaire et $h(a) = e^a / (1 + e^a)$, donne un modèle de régression logistique pour les réponses binaires y_k.

Nous définissons des équations d'estimation en population finie (EEPF), basées sur les fonctions d'estimation $\boldsymbol{l}_k(\boldsymbol{\theta})$, comme étant $\boldsymbol{l}(\boldsymbol{\theta}) = \sum \boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0}$ avec $E_m \boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0}$, et la solution de ces EEPF donne le vecteur de paramètres de population finie $\boldsymbol{\theta}_N$. Par exemple, $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{x}_k(y_k - \boldsymbol{\mu}_k(\boldsymbol{\theta}))$ pour les modèles de régression linéaire et logistique. Nous utilisons les poids de régression généralisée (GREG) $w_k(s) = d_k(s) g_k(\boldsymbol{d}(s))$, où les poids g sont donnés par

$$g_k(\boldsymbol{d}(s)) = 1 + (\boldsymbol{T} - \hat{\boldsymbol{T}})^T \left[\sum d_k(s) c_k \boldsymbol{t}_k \boldsymbol{t}_k^T \right]^{-1} c_k \boldsymbol{t}_k,$$

pour des c_k spécifiés, où $\hat{\boldsymbol{T}} = \sum d_k(s) \boldsymbol{t}_k$ est l'estimateur HT du total connu $\boldsymbol{T}$ d'un vecteur $q \times 1$ de variables de calage $\boldsymbol{t}_k$ et $\boldsymbol{d}(s)$ est le vecteur $N \times 1$ des poids $d_k(s)$. Les poids GREG, $w_k(s)$, ont la propriété de calage $\sum w_k(s) \boldsymbol{t}_k = \boldsymbol{T}$ et produisent des estimateurs efficaces $\tilde{Y} = \sum w_k(s) y_k$ des totaux $Y = \sum y_k$, quand la relation entre y_k et t_k est linéaire (Särndal, Swensson et Wretman 1989, chapitre 6).

Nous utilisons les poids de calage, $w_k(s)$, pour estimer les EEPF. Les équations d'estimation pondérée par calage sont données par

$$\tilde{\boldsymbol{l}}(\boldsymbol{\theta}) = \sum w_k(s) \boldsymbol{l}_k(\boldsymbol{\theta}) = \sum d_k(s) g_k(\boldsymbol{d}(s)) \boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0}. \quad (3.1)$$

La solution de (3.1), obtenue par la méthode itérative de type Newton-Raphson, donne l'estimateur pondéré par calage $\tilde{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$, et $\tilde{\boldsymbol{\theta}}$ est approximativement sans biais sous le plan et le modèle pour $\boldsymbol{\theta}$, c'est-à-dire $E(\tilde{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$. Il découle de (3.1) que $\tilde{\boldsymbol{\theta}}$ est de la forme $\boldsymbol{f}(\boldsymbol{A}_d)$ avec $\boldsymbol{d}_k = (d_k(s), d_k(s) \boldsymbol{l}_k^T(\boldsymbol{\theta}))^T$, où $\boldsymbol{f}(\boldsymbol{A}_d)$ est un vecteur $p \times 1$ et $\boldsymbol{A}_d$ est une matrice $(p+1) \times N$ dont la k^e colonne est $\boldsymbol{d}_k$. Ici, nous avons $h_{1k} = 1$ et $(h_{2k}, ..., h_{(p+1)k}) = \boldsymbol{l}_k(\boldsymbol{\theta})$.

3.2 Estimateurs de variance linéarisés

Nous commençons par étendre le résultat de l'estimation de la variance pour le cas scalaire $\hat{U} = \sum \boldsymbol{b}_k^T \boldsymbol{d}_k$ (section 2.2)

au cas vectoriel $\hat{\boldsymbol{U}} = \sum \boldsymbol{U}_k \boldsymbol{d}_k = \sum \boldsymbol{b}_k^T d_k(s)$, où $\boldsymbol{b}_k = \boldsymbol{U}_k \boldsymbol{h}_k$ est un p-vecteur et $\boldsymbol{U}_k$ est une matrice $p \times (p+1)$ dont les lignes sont $\boldsymbol{u}_{jk}^T$, $j = 1, ..., p$. Dans ce cas, l'estimateur de variance SYG (2.4) devient

$$\mathbf{est}(I) = \vartheta_{\mathrm{SYG}}(\hat{U}) = \sum\sum_{k<t} d_{kt}(s) \frac{(\pi_k \pi_t - \pi_{kt})}{\pi_k \pi_t} (\boldsymbol{b}_k - \boldsymbol{b}_t)(\boldsymbol{b}_k - \boldsymbol{b}_t)^T. \quad (3.2)$$

De même, l'estimateur de variance HT (2.5) devient

$$\mathbf{est}(I) = \vartheta_{\mathrm{HT}}(\hat{U}) = \sum\sum d_{kt}(s) \frac{(\pi_{kt} - \pi_k \pi_t)}{\pi_k \pi_t} \boldsymbol{b}_k \boldsymbol{b}_t^T. \quad (3.3)$$

Si nous passons à la composante II de la variance totale de $\hat{\boldsymbol{U}}$, (2.6) devient

$$\mathbf{est}(II) = \sum\sum d_{kt}(s) \boldsymbol{U}_k \operatorname{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t) \boldsymbol{U}_t^T. \quad (3.4)$$

La variance totale de $\hat{\boldsymbol{U}}$ est estimée par la somme de (3.2) et (3.4) pour les plans à taille d'échantillon fixe ou par la somme de (3.3) et (3.4) pour les plans arbitraires.

Un estimateur de variance par linéarisation de la variance totale de $\tilde{\boldsymbol{\theta}}$ s'obtient à partir de l'estimateur de variance totale estimé de $\hat{\boldsymbol{U}}$ en remplaçant $\boldsymbol{U}_k$ par la variable linéarisée $\boldsymbol{Z}_k = \partial \boldsymbol{f}(\boldsymbol{A}_b) / \partial \boldsymbol{b}_k |_{\boldsymbol{A}_b = \boldsymbol{A}_d}$. Si nous suivons la méthode de dérivation implicite de Demnati et Rao (2004), $\boldsymbol{Z}_k$ se réduit à

$$\boldsymbol{Z}_k = [\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1} g_k(\boldsymbol{d}(s)) (-\hat{\boldsymbol{B}}_l^T \boldsymbol{t}_k, \boldsymbol{I}_p),$$

avec

$$\hat{\boldsymbol{B}}_l = \Big[\sum d_k(s) c_k \boldsymbol{t}_k \boldsymbol{t}_k^T\Big]^{-1} \sum d_k(s) c_k \boldsymbol{t}_k \boldsymbol{l}_k^T(\tilde{\boldsymbol{\theta}}),$$

$$\tilde{\boldsymbol{J}}(\boldsymbol{\theta}) = -\sum d_k(s) g_k(\boldsymbol{d}(s)) (\partial \boldsymbol{l}_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T),$$

et $\boldsymbol{I}_p$ est la matrice identité $p \times p$.

Après certaines simplifications, la première composante $\mathbf{est}(I)$ est donnée par (3.2) ou (3.3) avec $\boldsymbol{b}_k$ transformé en

$$\boldsymbol{Z}_k \boldsymbol{h}_k = [\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1} e_k(\tilde{\boldsymbol{\theta}}) g_k(\boldsymbol{d}(s)), \quad (3.5)$$

où

$$\boldsymbol{e}_k(\tilde{\boldsymbol{\theta}}) = \boldsymbol{l}_k(\tilde{\boldsymbol{\theta}}) - \hat{\boldsymbol{B}}_l^T \boldsymbol{t}_k.$$

De même, la deuxième composante $\mathbf{est}(II)$ se réduit à

$$\mathbf{est}(II) = [\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1} \sum d_k(s) g_k^2(\boldsymbol{d}(s)) \boldsymbol{l}_k(\tilde{\boldsymbol{\theta}}) \boldsymbol{l}_k^T(\tilde{\boldsymbol{\theta}}) [\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1}. \quad (3.6)$$

si $\operatorname{Cov}_m[\boldsymbol{l}_k(\tilde{\boldsymbol{\theta}}) \boldsymbol{l}_t^T(\tilde{\boldsymbol{\theta}})] = \boldsymbol{0}$ pour $k \neq t$.

L'estimateur de la variance totale de $\tilde{\boldsymbol{\theta}}$ est maintenant estimé par

$$\vartheta_{\mathrm{DR}}(\tilde{\boldsymbol{\theta}}) = \mathbf{est}(I) + \mathbf{est}(II). \quad (3.7)$$

Cet estimateur de variance de $\tilde{\boldsymbol{\theta}}$ tient automatiquement compte des poids g comme à la section 2.

Un estimateur de variance habituel de $\tilde{\boldsymbol{\theta}}$, $\vartheta_{\mathrm{cus}}(\tilde{\boldsymbol{\theta}})$, est obtenu à partir de (3.7) en ignorant les poids g dans (3.5) et (3.6). De même, un estimateur de variance hybride, $\vartheta_{\mathrm{mix}}(\tilde{\boldsymbol{\theta}})$, est obtenu à partir de (3.7) en gardant les poids g dans $\mathbf{est}(I)$ et en les ignorant dans $\mathbf{est}(II)$.

3.3 Étude en simulation

Nous avons effectué une étude en simulation pour comparer les propriétés relatives des trois estimateurs de variance ϑ_{DR}, ϑ_{cus} et ϑ_{mix}, dans le cas particulier d'un modèle de régression logistique :

$$E_m(y_k) = \mu_k(\boldsymbol{\theta}) = \exp(\boldsymbol{x}_k^T \boldsymbol{\theta}) / \{1 + \exp(\boldsymbol{x}_k^T \boldsymbol{\theta})\} \quad (3.8)$$

$$V_m(y_k) = \mu_k(\boldsymbol{\theta})(1 - \mu_k(\boldsymbol{\theta})), \operatorname{Cov}_m(y_k, y_t) = 0, k \neq t.$$

Dans ces conditions, nous avons $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$, et

$$\tilde{\boldsymbol{J}}(\boldsymbol{\theta}) = \sum d_k(s) g_k(\boldsymbol{d}(s)) \boldsymbol{x}_k \boldsymbol{x}_k^T \mu_k(\boldsymbol{\theta})(1 - \mu_k(\boldsymbol{\theta})).$$

Pour l'étude en simulation, nous posons que $\boldsymbol{x}_k = (1, x_k)^T$, où les x_k désignent le nombre de lits pour la population d'hôpitaux de taille $N = 393$ étudiée à la section 2.2. Nous avons exécuté une poststratification en divisant la population en deux classes, avec $N_1 = 171$ hôpitaux k ayant $x_k < 350$ dans la classe 1 et $N_2 = 122$ hôpitaux k avec $x_k \geq 350$ dans la classe 2. Ici, $g_k(\boldsymbol{d}(s)) = N_h / \hat{N}_h$, $h = 1, 2$, si k appartient à la classe h, où $\hat{N}_h = \sum d_k(s) t_{hk}$ est l'estimateur pondéré par les poids de sondage de N_h, et $\boldsymbol{t}_k = (t_{1k}, t_{2k})^T$ est le vecteur de variables indicatrices de classe t_{hk}.

Nous avons généré $R = 40\,000$ populations finies $\{y_1, ..., y_N\}$, chacune de taille $N = 393$, en émettant l'hypothèse du modèle de régression logistique (3.8) avec $\boldsymbol{\theta} = (\theta_0, \theta_1)^T = (-1,\ 0,005)^T$. Le paramètre d'intérêt est $\theta_1 = 0{,}005$. Dans chacune des populations produites, nous avons sélectionné un échantillon aléatoire simple de taille $n = 150$, puis obtenu l'estimateur $\tilde{\theta}_1$ estimé et pondéré par calage et les estimateurs de variance connexes $\mathrm{est}(I) = \vartheta_s(\tilde{\theta}_1), \vartheta_{\mathrm{DR}}(\tilde{\theta}_1), \vartheta_{\mathrm{cus}}(\tilde{\theta}_1)$ et $\vartheta_{\mathrm{mix}}(\tilde{\theta}_1)$ à partir de chaque échantillon r. Nous avons obtenu les moyennes des estimations et des estimations de variance comme étant $\mathrm{moy}(\hat{\theta}_1) \approx 0{,}00514$, $\mathrm{moy}(\vartheta_{\mathrm{DR}}) \approx 0{,}0989$, $\mathrm{moy}(\vartheta_{\mathrm{cus}}) \approx 0{,}0987$, $\mathrm{moy}(\vartheta_{\mathrm{mix}}) \approx 0{,}0988$ et $\mathrm{moy}(\vartheta_s) \approx 0{,}0613$. En outre, l'EQM totale estimée de $\hat{\theta}_1$ est égale à 0,0998. Donc, conditionnellement, l'estimateur $\tilde{\theta}_1$ est approximativement sans biais pour θ_1, et le biais des trois estimateurs de variance ϑ_{DR}, ϑ_{cus} et ϑ_{mix} est négligeable. Par ailleurs, ignorer la deuxième composante et utiliser seulement la

première, $\text{est}(I) = \vartheta_s(\tilde{\theta}_1)$, donne lieu à une sous-estimation très importante, comme prévu.

Nous avons également examiné les propriétés conditionnelles des trois estimateurs de variance de la même façon qu'à la section 2.2. Nous avons classé les 40 000 échantillons par ordre croissant de taille, n_1, dans la classe 1, puis les avons groupés en 20 groupes, chacun de taille 2 000, de manière que le premier groupe, G_1, contienne les 2 000 échantillons ayant les plus petites valeurs n_1, que le deuxième, G_2, contienne les 2 000 échantillons ayant les plus petites valeurs n_1 suivantes, et ainsi de suite, pour obtenir 20 groupes, $G_1, ..., G_{20}$.

Nous avons calculé l'EQM conditionnelle de $\tilde{\theta}_1$ et le biais relatif conditionnel (BRC) connexe des estimateurs de variance ϑ_{DR}, ϑ_{cus} et ϑ_{mix} basés sur les valeurs moyennes de ϑ_{DR}, ϑ_{cus} et ϑ_{mix} dans chaque groupe ; voir la figure 5. L'examen de cette figure montre que le BRC de ϑ_{cus} varie de 20 % à -20 % lorsque l'on passe d'un groupe à l'autre, tandis que ϑ_{DR} ne présente pas ce genre de tendance et son BRC est inférieur à 5 % en valeur absolue, sauf pour deux groupes. En outre, le BRC de ϑ_{mix} présente une tendance, mais moins prononcée que celle du BRC de ϑ_{cus}. La figure 6 donne les taux de couverture conditionnels (TCC) des intervalles de la théorie normale basés sur ϑ_{DR}, ϑ_{cus} et ϑ_{mix} pour le niveau nominal de 95 %. La figure 6 montre que ϑ_{cus} présente une tendance selon le groupe, le TCC variant de 97 % à 92 %, tandis que le TCC associé à ϑ_{DR} est proche du niveau nominal dans les divers groupes. En outre, le TCC associé à ϑ_{mix} est légèrement supérieur à celui associé à ϑ_{DR} pour la première moitié des groupes et légèrement inférieur, pour les autres.

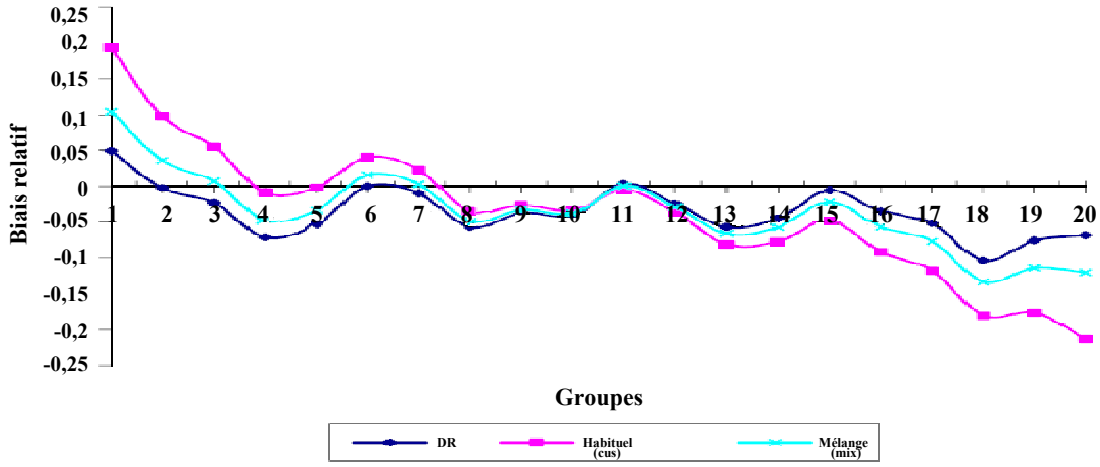

Figure 5 Biais relatif conditionnel des estimateurs de variance : régression logistique

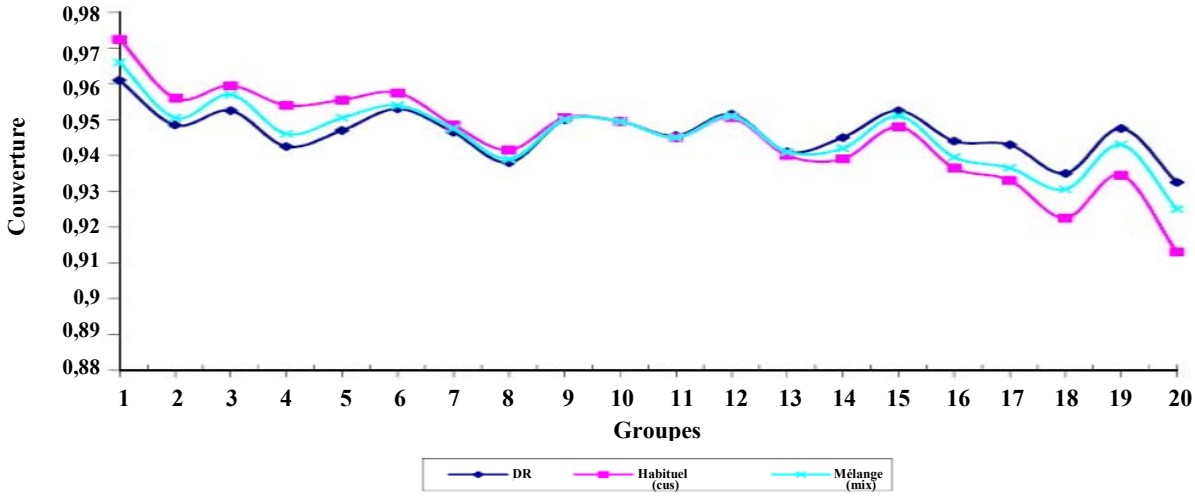

Figure 6 Taux de couverture conditionnel des intervalles de confiance de la théorie normale pour le niveau nominal de 95 % : régression logistique

Conclusion

Nous avons étudié l'estimation de la variance totale des estimateurs des paramètres du modèle sous l'hypothèse d'un modèle de superpopulation. Notre approche mène directement à un estimateur de variance par linéarisation qui, comme nous le montrons, donne de bons résultats dans un cadre conditionnel quand les poids de calage sont utilisés pour l'estimation. Nous sommes en train d'étudier des extensions de notre méthode à l'estimation de la variance totale sous imputation pour la non-réponse partielle et sous intégration de deux enquêtes indépendantes.

Remerciements

Nous remercions deux examinateurs de leurs suggestions et commentaires constructifs. Les travaux de J.N.K. Rao ont été financés en partie par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

Bibliographie

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

Binder, D. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases : une approche de type « recette ». *Techniques d'enquête*, 22, 17-22.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête (avec discussion). *Techniques d'enquête*, 30, 17-37.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*, New York : John Wiley & Sons, Inc.

McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models*, 2ième Éd. Chapman & Hall, Londres.

Molina, E.A., Smith, T.M.F. et Sugden, R.A. (2001). Modeling overdispersion for complex survey data. *Revue Internationale de Statistique*, 69, 373-384.

Royall, R.M., et Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

Rubin-Bleuer, S., et Şchiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, 33, 2789-2810.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

Skinner, C.J., Holt, D. et Smith, T.M.F. (1989). *Analysis of Complex Surveys*, New York : John Willey & Sons, Inc.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite population sampling and inference*: *A prediction approach*, New York : John Wiley & Sons, Inc.