N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

UNE MÉTHODE FONDÉE SUR LE PLAN DE SONDAGE POUR L'ANALYSE D'EXPÉRIENCES INTÉGRÉES DANS LES ENQUÊTES PAR SONDAGE COMPLEXES

Jan A. van den Brakel[1]

RÉSUMÉ

Les expériences intégrées dans des enquêtes par sondage régulières sont particulièrement appropriées pour évaluer les effets de diverses méthodes d'enquête sur les estimations des paramètres d'une population finie. Ne pas tenir compte du plan d'échantillonnage dans l'analyse de ce genre d'expérience produit habituellement des estimations des paramètres et de la variance biaisées par rapport au plan de sondage qui rendent les résultats de l'analyse non comparables aux estimations calculées d'après l'enquête ordinaire. Pour tester les hypothèses au sujet des différences entre les estimations d'échantillon observées sous diverses méthodes d'enquête, nous proposons une approche fondée sur le plan de sondage pour les expériences analytiques intégrées dans les enquêtes par sondage complexes.

MOTS CLÉS : Dispositifs en randomisation totale; estimateur par la régression généralisée; modèles de l'erreur de mesure; dispositifs en blocs randomisés; statistique de Wald.

1. INTRODUCTION

Nombre de travaux de recherche dans le domaine de la méthodologie d'enquête visent à améliorer la qualité et l'efficacité des processus d'enquête par sondage. Une part importante de cette recherche consiste à considérer et à évaluer des méthodes d'enquête de rechange. Les expériences à grande échelle sur le terrain intégrées dans les enquêtes par sondage régulières sont particulièrement appropriées pour quantifier l'effet de diverses mises en œuvre d'une enquête sur le comportement de réponse ou sur les estimations. Les bureaux nationaux de la statistique s'efforcent généralement de ne pas modifier les enquêtes par sondage pendant des périodes aussi longues que possible afin de produire des séries chronologiques ininterrompues d'estimations des paramètres de population. Toutefois, il est inévitable que l'on doive rajuster les processus d'enquête de temps en temps. Le cas échéant, on peut recourir à des expériences intégrées pour déceler et quantifier les ruptures de tendance éventuelles dans les séries chronologiques imputables aux ajustements des processus d'enquête. Des exemples et des applications d'expériences intégrées sont décrits dans Van den Brakel et Renssen (1998) et dans Van den Brakel (2001). Pour comparer les effets de K approches d'enquêtes différentes sur les estimations principales des paramètres de population finie d'une enquête régulière, un échantillon tiré à partir d'une population finie est subdivisé aléatoirement en K sous-échantillons conformément à un plan d'expérience. Les unités d'échantillonnage de chaque sous-échantillon sont assignées à l'une des K approches d'enquête, ou traitements. En général, on assigne à l'enquête régulière un grand sous-échantillon qui est utilisé pour produire les publications officielles et comme groupe témoin dans l'expérience. L'objectif de ce genre d'expériences intégrées est d'estimer les paramètres de population finie sous diverses mises en œuvre de l'enquête, ou traitements, et de tester des hypothèses quant aux écarts entre ces estimations. Dans le présent article, nous proposons une méthode d'analyse fondée sur le plan de sondage pour les expériences intégrées réalisées selon des dispositifs en randomisation totale (DRT) ou en blocs randomisés (DBR). Nous calculons des estimateurs sans biais par rapport au plan de sondage pour les paramètres de population observés sous chacun des K traitements, ainsi que des estimateurs sans biais par rapport au plan de sondage de la matrice des covariances des contrastes entre ces estimations au moyen de l'estimateur d'Horvitz-Thompson ou de l'estimateur par la régression généralisée. Ces estimateurs tiennent compte du plan d'échantillonnage, du plan d'expérience et de la méthode de pondération de l'enquête régulière, et donnent une

[1] Jan A. van den Brakel, Département des méthodes statistiques, Statistique Pays-Bas, C.P. 4481, 6401 CZ, Heerlen, Pays-Bas, courriel : JBRL@CBS.NL.

statistique de Wald fondée sur le plan de sondage pour tester les hypothèses au sujet des différences entre les estimations des paramètres de population finie de l'enquête par sondage.

2. INTÉGRATION D'EXPÉRIENCES DANS LES ENQUÊTES

L'intégration d'expériences dans les enquêtes par sondage est parfois appelée superposition d'expériences à des enquêtes par sondage. Pour commencer, on tire un échantillon à partir d'une population finie cible conformément à un plan d'échantillonnage généralement complexe. Puis, on subdivise aléatoirement l'échantillon conformément à un plan d'expérience en K sous-échantillons. L'approche la plus simple consiste à utiliser un dispositif en randomisation totale (DRT). Cependant, l'application de la randomisation sans contrainte n'est généralement pas le dispositif le plus efficace disponible. Fienberg et Tanur (1987, 1988) soutiennent que l'application d'un dispositif en blocs randomisés (DBR) contenant des structures d'échantillonnage, comme les strates, les unités primaires d'échantillonnage (UPE), les grappes, les intervieweurs et ainsi de suite, pourrait améliorer considérablement la précision de l'expérience. De surcroît, la randomisation sans contraintes n'est pas toujours faisable d'un point de vue pratique. Par exemple, dans les enquêtes par IPAO, où les intervieweurs recueillent les données dans des secteurs géographiques autour de leur lieu de résidence, la randomisation contrainte des unités d'échantillonnage à l'intérieur des intervieweurs ou des régions géographiques qui sont des unions de régions d'intervieweurs adjacentes pourrait être nécessaire pour éviter une augmentation inacceptable de la distance que doivent parcourir les intervieweurs. Cette approche mène naturellement au DBR avec les intervieweurs ou les régions représentant les variables de bloc.

L'un des principaux avantages des expériences intégrées est la sélection aléatoire des unités d'échantillonnage à partir d'une population cible finie qui augmente la possibilité de généraliser les résultats observés sur l'échantillon de l'expérience à de plus grandes populations (Fienberg et Tanur, 1987, 1988). En outre, les expériences intégrées constituent une transition sûre d'un ancien à un nouveau plan de sondage. L'exécution en parallèle de l'ancienne et de la nouvelle approche grâce à une expérience intégrée donne la possibilité de quantifier les ruptures de tendances éventuelles et permet encore de retourner à l'ancienne approche aux fins des publications régulières de données s'il s'avère que la nouvelle approche échoue. Néanmoins, il faut être conscient que deux objectifs plus ou moins concurrents sont combinés dans une expérience intégrée. Conformément à l'objectif de l'enquête régulière, c'est-à-dire l'estimation de paramètres de population aussi précise que possible, le sous-échantillon assigné à l'enquête régulière devrait être maximisé. Conformément à l'objectif de l'expérience, c'est-à-dire l'estimation aussi précise que possible des contrastes entre les estimations sur les sous-échantillons, ces derniers devraient être de préférence de même taille, puisque les plans de sondage équilibrés maximisent la puissance des tests au sujet des effets des traitements (voir, p. ex., Montgomery, 1997).

3. ANALYSE DES EXPÉRIENCES INTÉGRÉES

L'objectif principal des expériences intégrées dans les enquêtes par sondage est d'estimer les paramètres de population observés sous divers traitements et de tester les hypothèses au sujet des différences entre ces estimations sur les sous-échantillons. La réalisation de cet objectif nécessite une méthode d'analyse qui tient compte du plan d'échantillonnage, du plan d'expérience et de la méthode d'estimation appliquée dans le contexte de l'enquête par sondage régulière. L'application d'une telle méthode d'analyse est également une conséquence naturelle de la sélection aléatoire des unités d'échantillonnage à partir d'une population cible dans le but de généraliser les résultats observés à des populations plus grandes que l'échantillon inclus dans l'expérience.

3.1 Modèles de l'erreur de mesure

Bien que la méthode d'analyse proposée dans le présent article soit principalement fondée sur le plan de sondage, elle comporte l'utilisation de modèles de l'erreur de mesure. La théorie de l'échantillonnage fondée sur un plan de sondage est basée en grande partie sur la notion classique selon laquelle les observations obtenues d'après les unités d'échantillonnage sont des valeurs fixes réelles observées sans erreur (p. ex. Cochran, 1977). Cependant, cette approche n'est pas défendable si l'on réalise des expériences pour tester des différences systématiques entre estimations de paramètres de population finie dues à des mises en application différentes de l'enquête ou à des erreurs non dues à l'échantillonnage. Par conséquent, il faut introduire un modèle de l'erreur de mesure pour relier

les différences systématiques entre les estimations des paramètres de population finie aux diverses mises en œuvre de l'enquête, ou traitements. On suppose que les observations obtenues durant l'expérience sont une réalisation du modèle de l'erreur de mesure suivant :

$$y_{ikl}^{\alpha} = u_i + \beta_k + \gamma_l^{\alpha} + \varepsilon_{ik}^{\alpha}.$$

Ici y_{ikl}^{α} est la réponse de l'unité d'échantillonnage *i* assignée au traitement *k* et à l'intervieweur *l* lors de la α^{e} occasion que l'unité d'échantillonnage *i* est observée ou interviewée, u_i est la valeur réelle intrinsèque de l'unité d'échantillonnage *i*, β_k est un effet additif du traitement *k*, γ_l^{α} est un effet de l'intervieweur *l* et $\varepsilon_{ik}^{\alpha}$ est une composante d'erreur à la α^{e} occasion que la valeur intrinsèque de l'unité d'échantillonnage *i* est mesurée sous le traitement *k*. Le modèle permet des effets d'intervieweur mixtes, c'est-à-dire $\gamma_l^{\alpha} = \psi_l + \xi_l^{\alpha}$, où ψ_l et ξ_l^{α} dénotent les effets fixes et aléatoires de l'intervieweur *l*. L'exposant α exprime quelles variables sont aléatoires par rapport au modèle de l'erreur de mesure.

Puisque *K* variables de réponse sont définies sous *K* traitements différents pour chaque unité d'échantillonnage, le modèle d'erreur de mesure peut être exprimé en notation matricielle sous la forme

$$\mathbf{y}_{il}^{\alpha} = \mathbf{j}u_i + \boldsymbol{\beta} + \mathbf{j}\gamma_l^{\alpha} + \boldsymbol{\varepsilon}_i^{\alpha}, \tag{3.1}$$

où $\mathbf{y}_{il}^{\alpha} = (y_{i1l}^{\alpha}, \ldots, y_{iKl}^{\alpha})^t$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^t$, $\boldsymbol{\varepsilon}_i^{\alpha} = (\varepsilon_{i1}^{\alpha}, \ldots \varepsilon_{iK}^{\alpha})^t$ et $\mathbf{j}$ est un vecteur de dimension *K* dont chaque élément est égal à l'unité. Soit E_{α} et Cov_{α} l'espérance et la covariance par rapport au modèle de l'erreur de mesure. On suppose que $E_{\alpha}(\xi_l^{\alpha}) = 0$ et que les effets aléatoires d'intervieweur entre les intervieweurs sont indépendants. En outre, on suppose que $E_{\alpha}(\boldsymbol{\varepsilon}_i^{\alpha}) = \mathbf{0}$ et que les erreurs de mesure entre unités d'échantillonnage différentes sont indépendantes. Donc

$$E_{\alpha}(\mathbf{y}_{il}^{\alpha}) = \mathbf{j}u_i + \boldsymbol{\beta} + \mathbf{j}\psi_l,$$

et

$$\mathrm{Cov}_{\alpha}(\mathbf{y}_{il}^{\alpha}, \mathbf{y}_{i'l'}^{\alpha}) = \begin{cases} \mathrm{Var}_{\alpha}(\boldsymbol{\varepsilon}_i^{\alpha}) + \mathbf{j}\mathbf{j}^t \mathrm{Var}_{\alpha}(\xi_l^{\alpha}) & : \ i = i' \text{ et } l = l' \\ \mathbf{j}\mathbf{j}^t \mathrm{Var}_{\alpha}(\xi_l^{\alpha}) & : \ i \neq i' \text{ et } l = l'. \\ \mathbf{O} & : \ i \neq i' \text{ et } l \neq l' \end{cases}$$

Manifestement, le modèle de l'erreur de mesure permet qu'il existe une corrélation entre les réponses de diverses unités d'échantillonnage assignées au même intervieweur.

3.2 Tests d'hypothèses

Après avoir défini un modèle de l'erreur de mesure pour les observations obtenues dans l'expérience, nous pouvons relier les différences systématiques entre les paramètres de population aux différentes mises en œuvre de l'enquête. Supposons que l'on dispose de *L* intervieweurs pour réaliser le travail sur le terrain. La population *U* de taille *N* peut être subdivisée conceptuellement en *L* groupes U_l de taille N_l tels que toutes les unités d'échantillonnage à l'intérieur d'un groupe soit potentiellement interviewées par le même intervieweur. Soit $\overline{\mathbf{Y}}^{\alpha} = (\overline{Y}_1^{\alpha}, \ldots, \overline{Y}_K^{\alpha})^t$ le vecteur de dimension *K* des moyennes de population de $\mathbf{y}_i^{\alpha}$, c'est-à-dire

$$\overline{\mathbf{Y}}^{\alpha} = \mathbf{j}\frac{1}{N}\sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}\sum_{l=1}^{L}\frac{N_l}{N}\psi_l + \mathbf{j}\sum_{l=1}^{L}\frac{N_l}{N}\xi_l^{\alpha} + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\varepsilon}_i^{\alpha}.$$

Puisque nous nous intéressons aux différences systématiques entre les moyennes de population observées sous les *K* traitements différents, les écarts aléatoires entre les composantes de $\overline{\mathbf{Y}}^{\alpha}$ ne devraient pas donner lieu à des différences significatives dans l'analyse de l'expérience. Nous accomplissons ceci en formulant des hypothèses au sujet de

$$E_{\alpha}\overline{\mathbf{Y}}^{\alpha} = \mathbf{j}\frac{1}{N}\sum_{i=1}^{N} u_i + \boldsymbol{\beta} + \mathbf{j}\sum_{l=1}^{L}\frac{N_l}{N}\psi_l \equiv \overline{\mathbf{Y}}\ ,$$

ce qui aboutit conséquemment à la formulation des hypothèses suivantes :

$$\begin{aligned} H_0 &: \mathbf{C}\overline{\mathbf{Y}} = \mathbf{0} \\ H_1 &: \mathbf{C}\overline{\mathbf{Y}} \neq \mathbf{0} \end{aligned}\ . \tag{3.2}$$

Ici $\mathbf{C} = (\mathbf{j}\,|\,-\mathbf{I})$ dénote une matrice de contrastes $(K-1)\times K$. Puisque $\mathbf{C}\overline{\mathbf{Y}} = \mathbf{C}\boldsymbol{\beta}$, il s'ensuit que l'hypothèse (3.2) concerne les effets du traitement tels que représentés par $\boldsymbol{\beta}$ dans le modèle de l'erreur de mesure. Par conséquent, les différences entre les moyennes de population correspondent exactement aux effets de traitement. Maintenant, l'hypothèse au sujet des effets de traitement peut être testée en estimant $\overline{\mathbf{Y}}$ au lieu de $\boldsymbol{\beta}$, où nous tenons compte du plan d'échantillonnage, du plan d'expérience et de la méthode de pondération de l'enquête par sondage régulière. Si $\hat{\overline{\mathbf{Y}}}^{\alpha}$ dénote un estimateur sans biais par rapport au plan de sondage de ce genre pour $\overline{\mathbf{Y}}$ et que $\mathbf{V}$ dénote la matrice des covariances de $\hat{\overline{\mathbf{Y}}}^{\alpha}$, alors l'hypothèse (3.2) peut être testée au moyen de la statistique de Wald $W = \hat{\overline{\mathbf{Y}}}^{\alpha t}\mathbf{C}^t(\mathbf{C}\mathbf{V}\mathbf{C}^t)^{-1}\mathbf{C}\hat{\overline{\mathbf{Y}}}^{\alpha}$.

3.3 Estimation des moyennes de sous-échantillon

Pour tester l'hypothèse (3.2), nous disposons d'un échantillon *s* tiré à partir d'une population finie *U* de taille *N*. Soit π_i et $\pi_{ii'}$ les probabilités d'inclusion de premier et de deuxième ordre des i^{e} et i,i'^{e} unités d'échantillonnage, respectivement, par rapport à ce plan d'échantillonnage. Dans le cas d'un DRT, *s* est subdivisé aléatoirement en *K* sous-échantillons s_k de taille n_k. Si $n_+ = \sum_{k=1}^{K} n_k$ représente le nombre d'unités d'échantillonnage de *s*, alors la probabilité conditionnelle que l'unité d'échantillonnage *i* soit assignée au traitement *k*, sachant la réalisation de *s*, est égale à n_k / n_+. Dans le cas d'un DBR, les unités d'échantillonnage de *s* sont subdivisées de façon déterministe en *B* blocs s_b. Dans chaque bloc, les unités d'échantillonnage sont réparties aléatoirement sur les *K* traitements. Soit n_{bk} le nombre d'unités d'échantillonnage assignées au traitement *k* dans le bloc *b*. Alors $n_{b+} = \sum_{k=1}^{K} n_{bk}$ représente la taille de s_b, $n_{+k} = \sum_{b=1}^{B} n_{bk}$ représente la taille du sous-échantillon s_k assigné au traitement *k* et $n_{++} = \sum_{b=1}^{B}\sum_{k=1}^{K} n_{bk}$ représente la taille de l'échantillon *s*. La probabilité conditionnelle que l'unité d'échantillonnage *i* soit assignée au traitement *k*, sachant la réalisation de *s* et que $i \in s_b$, est égale à n_{bk} / n_{b+}. Chaque sous-échantillon s_k peut être considéré comme un échantillon à deux phases, où la première phase correspond au plan d'échantillonnage utilisé pour tirer l'échantillon *s* et la deuxième phase, au plan d'expérience utilisé pour subdiviser *s* en *K* sous-échantillons s_k. Par conséquent, les probabilités d'inclusion de premier ordre par rapport aux sous-échantillons sont données par $\pi_i^* = (n_k / n_+)\pi_i$ dans le cas d'un DRT et par $\pi_i^* = (n_{bk} / n_{b+})\pi_i$ dans le cas d'un DBR. Maintenant, l'estimateur d'Horvitz-Thompson pour $\overline{Y}_k^{\alpha}$ fondé sur les observations obtenues pour le sous-échantillon s_k est donné par

$$\hat{\overline{Y}}_k^{\alpha} = \frac{1}{N}\sum_{i\in s_k}\frac{y_{ik}^{\alpha}}{\pi_i^*}\ .$$

Afin de tenir compte de la méthode de pondération de l'enquête régulière, l'analyse est fondée sur l'estimateur par la régression généralisée. L'utilisation d'information auxiliaire au moyen de cet estimateur offre l'avantage de réduire éventuellement la variance par rapport au plan de sondage des estimations de sous-échantillon et de corriger, du

moins partiellement, le biais dû à la non-réponse sélective. Dans le présent contexte, l'estimateur par la régression généralisée représente une analogie fondée sur le plan de sondage de l'analyse de covariance dans la méthodologie type des plans d'expérience. Soit $\mathbf{x}_i = (x_{i1}, \ldots, x_{iH})^t$ un vecteur d'ordre *H* dont chaque élément x_{ih} est une variable auxiliaire de l'unité d'échantillonnage *i*. Conformément à l'approche assistée par modèle de Särndal et coll. (1992), nous supposons que chaque valeur intrinsèque u_i du modèle de l'erreur de mesure pour chaque unité de la population est une réalisation indépendante du modèle de régression linéaire :

$$u_i = \mathbf{b}^t \mathbf{x}_i + e_i , \tag{3.3}$$

où $\mathbf{b}$ représente un vecteur d'ordre *H* de coefficients de régression et e_i représente les résidus du modèle de régression. Soit ω_i^2 la variance de e_i. Il est requis que tous les ω_i^2 soient connus jusqu'à un facteur d'échelle commun; autrement dit $\omega_i^2 = v_i \omega^2$, avec v_i connu. Les moyennes de population finie de ces variables sont représentées par le vecteur $\overline{\mathbf{X}}$. Soit

$$\mathbf{b}_k^{\alpha} = \left(\sum_{i \in U} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\omega_i^2} \right)^{-1} \sum_{i \in U} \frac{\mathbf{x}_i y_{ik}^{\alpha}}{\omega_i^2}$$

les coefficients de régression de population de $\mathbf{b}$ dans (3.3) observés sous le traitement *k*,

$$\hat{\mathbf{b}}_k^{\alpha} = \left(\sum_{i \in s_k} \frac{\mathbf{x}_i \mathbf{x}_i^t}{\pi_i^* \omega_i^2} \right)^{-1} \sum_{i \in s_k} \frac{\mathbf{x}_i y_{ik}^{\alpha}}{\pi_i^* \omega_i^2} \tag{3.4}$$

l'estimateur d'Horvitz-Thompson de $\mathbf{b}_k^{\alpha}$ fondé sur les observations obtenues dans s_k et

$$\hat{\overline{\mathbf{X}}}_k = \frac{1}{N} \sum_{i \in s_k} \frac{\mathbf{x}_i}{\pi_i^*} \tag{3.5}$$

l'estimateur d'Horvitz-Thompson de $\overline{\mathbf{X}}$ fondé sur les unités d'échantillonnage comprises dans s_k. Maintenant, l'estimateur par la régression généralisée de $\overline{Y}_k^{\alpha}$ fondé sur les observations obtenues dans le sous-échantillon s_k est donné par

$$\hat{\overline{Y}}_{k;greg}^{\alpha} = \hat{\overline{Y}}_k^{\alpha} + \hat{\mathbf{b}}_k^{\alpha t} (\overline{\mathbf{X}} - \hat{\overline{\mathbf{X}}}_k) . \tag{3.6}$$

L'estimateur par la régression généralisée peut être approximé par une linéarisation de Taylor de premier ordre dans le voisinage de $(\overline{Y}_k, \mathbf{b}_k, \overline{\mathbf{X}})$, où $\mathbf{b}_k = E_{\alpha}(\mathbf{b}_k^{\alpha})$ et est donné par

$$\hat{\overline{Y}}_{k;greg}^{\alpha} \approx \hat{\overline{Y}}_k^{\alpha} + \mathbf{b}_k^t (\overline{\mathbf{X}} - \hat{\overline{\mathbf{X}}}_k) . \tag{3.7}$$

Notons que $\hat{\overline{Y}}_{k;greg}^{\alpha}$ est un estimateur approximativement sans biais par rapport au plan de sondage de $\overline{Y}_k^{\alpha}$. Puisque $E_{\alpha} \overline{Y}_k^{\alpha} = \overline{Y}_k$, $\hat{\overline{Y}}_{k;greg}^{\alpha}$ est aussi un estimateur approximativement sans biais de $\overline{Y}_k$. Le vecteur $\hat{\overline{\mathbf{Y}}}_{\mathbf{GREG}}^{\alpha} = (\hat{\overline{Y}}_{1;greg}^{\alpha}, \ldots, \hat{\overline{Y}}_{K;greg}^{\alpha})^t$ est un estimateur approximativement sans biais de $\mathbf{Y}$.

3.4 Estimation de la variance des contrastes entre les moyennes de sous-échantillon

Soit $\mathbf{V}$ la matrice des covariances de $\hat{\overline{\mathbf{Y}}}_{\mathbf{GREG}}^{\alpha}$. Puisque les sous-échantillons sont tirés à partir d'une population finie sans remise, les estimations à partir de sous-échantillons sont dépendantes. Par conséquent, il existe une covariance

de plan de sondage non nulle entre les éléments de $\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}}$. Un estimateur de $\mathbf{V}$ nécessite des vecteurs $\mathbf{y}_i^{\alpha}$ contenant les observations des *K* traitements provenant de chaque unité d'échantillonnage. Cependant, dans les plans d'expérience étudiés, chaque unité d'échantillonnage est assignée à un seul des *K* traitements et, donc, on n'observe effectivement qu'une seule des composantes de $\mathbf{y}_i^{\alpha}$, pour $i \in s$. Par conséquent, on ne peut obtenir un estimateur sans biais par rapport au plan de sondage de $\mathbf{V}$. Le problème de l'observation manquante peut être contré en établissant un estimateur approximativement sans biais par rapport au plan de sondage de la matrice des covariances des *K-1* contrastes $\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}}$, dénoté $\mathbf{CVC}^t$. Soit E_s et E_e les espérances par rapport au plan d'échantillonnage et par rapport au plan d'expérience, respectivement. De façon équivalente, Cov_s et Cov_e représentent les covariances par rapport au plan d'échantillonnage et par rapport au plan d'expérience. Considérons la décomposition de la variance suivante :

$$\mathbf{CVC}^t = Cov_{\alpha}E_sE_e(\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}}) + E_{\alpha}Cov_sE_e(\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}}) + E_{\alpha}E_sCov_e(\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}}) \tag{3.8}$$

À condition qu'il existe un vecteur constant **a** d'ordre *H*, tel que $\mathbf{a}^t\mathbf{x}_i = 1$ pour tout $i \in U$, autrement dit qu'on utilise au moins la taille de la population finie comme information auxiliaire dans le scénario de pondération, Van den Brakel (2001) prouve, sous les hypothèses du modèle de l'erreur de mesure (3.1), que

$$\begin{aligned} Cov_{\alpha}E_sE_e(\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}}) &= \frac{1}{N^2}\sum_{i\in U}\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{C}^t \\ E_{\alpha}Cov_sE_e(\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}}) &= \frac{1}{N^2}\sum_{i\in U}\left(\frac{1}{\pi_i}-1\right)\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{C}^t \\ E_{\alpha}E_sCov_e(\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}}) &= E_{\alpha}E_s(\mathbf{CDC}^t) - \frac{1}{N^2}\sum_{i\in U}\frac{\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{C}^t}{\pi_i} \end{aligned} \tag{3.9}$$

où $\boldsymbol{\Sigma}_i = Var_{\alpha}(\boldsymbol{\varepsilon}_i^{\alpha})$ représente la matrice des covariances des erreurs de mesure et $\mathbf{D}$ représente une matrice diagonale de dimensions $K \times K$ avec les éléments

$$d_k = \frac{1}{n_k}\frac{1}{(n_+-1)}\sum_{i\in s}\left(\frac{n_+(y_{ik}^{\alpha}-\mathbf{b}_k^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_+}\sum_{i'\in s}\frac{n_+(y_{i'k}^{\alpha}-\mathbf{b}_k^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2 \equiv \frac{S_k^2}{n_k} \tag{3.10}$$

dans le cas d'un DRT et

$$d_k = \sum_{b=1}^{B}\frac{1}{n_{bk}}\frac{1}{(n_{b+}-1)}\sum_{i\in s_b}\left(\frac{n_{b+}(y_{ik}^{\alpha}-\mathbf{b}_k^t\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{b+}}\sum_{i'\in s_b}\frac{n_{b+}(y_{i'k}^{\alpha}-\mathbf{b}_k^t\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2 \equiv \sum_{b=1}^{B}\frac{S_{bk}^2}{n_{bk}} \tag{3.11}$$

dans le cas d'un DBR. L'insertion des trois composantes données par (3.9) dans (3.8) donne $\mathbf{CVC}^t = E_{\alpha}E_s\mathbf{CDC}^t$. Conditionnellement à α et *s*, nous pouvons tirer directement un estimateur approximativement sans biais par rapport au plan de sondage pour $\mathbf{D}$. Par conséquent, $\mathbf{CVC}^t$ peut, de façon commode, être énoncé implicitement comme étant l'espérance sur le modèle de l'erreur de mesure et le plan d'échantillonnage. Les expressions pour $E_{\alpha}E_sd_k$ sont établies dans Van den Brakel (2001) sous divers plans d'échantillonnage pour les DRT et les DBR. Étant donné la réalisation de α et *s*, l'affectation des unités d'échantillonnage au sous-échantillon s_k peut être considérée comme un échantillonnage aléatoire simple sans remise à partir de *s* dans le cas d'un DRT et comme un échantillonnage aléatoire simple sans remise à partir de s_b dans le cas d'un DBR. Par conséquent, un estimateur sans biais par rapport au plan de sondage pour d_k est donné par

$$\hat{d}_k = \frac{1}{n_k}\frac{1}{(n_k-1)}\sum_{i\in s_k}\left(\frac{n_+(y_{ik}^{\alpha}-\hat{\mathbf{b}}_k^{\alpha t}\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_k}\sum_{i'\in s_k}\frac{n_+(y_{i'k}^{\alpha}-\hat{\mathbf{b}}_k^{\alpha t}\mathbf{x}_{i'})}{N\pi_{i'}}\right)^2 \equiv \frac{\hat{S}_k^2}{n_k} \tag{3.12}$$

dans le cas d'un DRT et par

$$\hat{d}_k = \sum_{b=1}^{B} \frac{1}{n_{bk}} \frac{1}{(n_{bk}-1)} \sum_{i \in s_{bk}} \left(\frac{n_{b+}(y_{ik}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{bk}} \sum_{i' \in s_{bk}} \frac{n_{b+}(y_{i'k}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\mathbf{x}_{i'})}{N\pi_{i'}} \right)^2 \equiv \sum_{b=1}^{B} \frac{\hat{S}_{bk}^2}{n_{bk}} \tag{3.13}$$

dans le cas d'un DBR où s_{bk} représente les n_{bk} unités d'échantillonnage du bloc *b* assignées au traitement *k*. Donc, un estimateur approximativement sans biais de $\mathbf{CVC}^t$ est donné par $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$, où les éléments diagonaux de $\hat{\mathbf{D}}$ sont définis par (3.12) ou (3.13).

Les résultats pour l'estimateur d'Horvitz-Thompson découle, à titre de cas spécial, des résultats obtenus pour l'estimateur par la régression généralisée avec le modèle commun de la moyenne utilisé comme scénario de pondération (Särndal et coll., 1992, section 7.4), c'est-à-dire $(x_i) = 1$ et $\omega_i^2 = \omega^2$. Le modèle commun de la moyenne utilise uniquement le total de population comme information auxiliaire et satisfait donc à la condition selon laquelle il existe un vecteur *H* constant, tel que $\mathbf{a}^t\mathbf{x}_i = 1$ pour tout $i \in U$. Sous ce scénario de pondération, il s'ensuit que

$$\hat{\bar{Y}}_{k;greg}^{\alpha} = \left(\sum_{i \in s_k} \frac{1}{\pi_i^*} \right)^{-1} \left(\sum_{i \in s_k} \frac{y_{ik}^{\alpha}}{\pi_i^*} \right) \equiv \tilde{Y}_k^{\alpha}, \tag{3.14}$$

et $\hat{\mathbf{b}}_k^{\alpha} = \tilde{Y}_k^{\alpha}$. Un estimateur approximativement sans biais par rapport au plan de sondage de la matrice des covariances des contrastes entre les moyennes de sous-échantillon est donné par (3.12) ou (3.13) pour un DRT ou un DBR, respectivement, où $\hat{\mathbf{b}}_k^{\alpha t}\mathbf{x}_i = \tilde{Y}_k^{\alpha}$.

La méthode d'estimation de la variance proposée à la présente section est établie sous des plans d'échantillonnage complexes généraux qui permettent aussi qu'il existe une dépendance entre les estimations d'après les sous-échantillons. Par conséquent, il est remarquable qu'aucune probabilité d'inclusion de deuxième ordre ne soit nécessaire dans les estimateurs de la variance. Un examen approfondi montre qu'il s'agit d'une conséquence 1) de l'hypothèse d'effets de traitement constants, 2) de l'application d'un scénario de pondération qui satisfait à la condition $\mathbf{a}^t\mathbf{x}_i = 1$ pour tout $i \in U$, 3) de l'hypothèse que les erreurs de mesure $\boldsymbol{\varepsilon}_i^{\alpha}$ sont indépendantes, 4) du mécanisme particulier de randomisation du DRT ou du DBR et 5) du calcul des variances des contrastes entre moyennes de sous-échantillon. Par conséquent, les estimateurs de la variance ont la même structure que si les K sous-échantillons avaient été tirés indépendamment l'un de l'autre par sélection des unités d'échantillonnage avec probabilités d'inclusion inégales avec remise. Voir Van den Brakel (2001) pour une discussion plus détaillée de ce résultat. Une autre méthode d'estimation de la variance est proposée par Van den Brakel et Binder (2000).

3.5 Test de Wald

Pour tester l'hypothèse (3.2), les estimateurs fondés sur le plan de sondage des moyennes de sous-échantillon et la matrice des covariances des contrastes entre les moyennes de sous-échantillon donnent lieu à la statistique de Wald fondée sur le plan de sondage suivante :

$$W = \hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}^{\alpha t} \mathbf{C}^t (\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t)^{-1} \mathbf{C} \hat{\bar{\mathbf{Y}}}_{\mathbf{GREG}}^{\alpha}.$$

Étant donné la structure diagonale de $\hat{\mathbf{D}}$, nous pouvons simplifier cette statistique de Wald pour obtenir

$$W = \sum_{k=1}^{K} \frac{\hat{\bar{Y}}_{k;greg}^{\alpha 2}}{\hat{d}_k} - \left(\sum_{k=1}^{K} \frac{1}{\hat{d}_k} \right)^{-1} \left(\sum_{k=1}^{K} \frac{\hat{\bar{Y}}_{k;greg}^{\alpha}}{\hat{d}_k} \right)^2. \tag{3.15}$$

Si *s* est tiré par échantillonnage aléatoire simple sans remise et que le plan d'expérience est un DRT, alors les conditions données par Lehmann (1975, annexe 8) peuvent être appliquées pour montrer que $\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}} \to N(\mathbf{C}\boldsymbol{\beta}, \mathbf{CVC}^t)$. Sous les plans d'échantillonnage complexes généraux, lors de l'analyse des données d'enquête, on suppose généralement qu'un théorème limite tient, si bien que $\mathbf{C}\hat{\bar{\mathbf{Y}}}^{\alpha}_{\mathbf{GREG}} \to N(\mathbf{C}\boldsymbol{\beta}, \mathbf{CVC}^t)$. Alors, il s'ensuit, sous l'hypothèse nulle, que *W* suit asymptotiquement une loi du chi carré avec *K-1* degrés de liberté pour le calcul des valeurs *p* ou des régions critiques pour *W*.

3.6 Estimateurs groupés de la variance

Dans le cas d'un DBR, on pourrait améliorer la méthode d'estimation de la variance en groupant les estimateurs des variances de population S^2_{bk}, qui sont définis implicitement dans (3.11), à l'intérieur de chaque bloc. L'estimateur groupé de la variance de population dans le bloc *b* est donné par

$$\hat{S}_b = \frac{1}{n_{b+}-1}\sum_{k=1}^{K}\sum_{i\in s_{bk}}\left(\frac{n_{b+}(y^{\alpha}_{ik} - \hat{\mathbf{b}}^{\alpha t}_k \mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{b+}}\sum_{k=1}^{K}\sum_{i'\in s_{bk}}\frac{n_{b+}(y^{\alpha}_{i'k} - \hat{\mathbf{b}}^{\alpha t}_k \mathbf{x}_{i'})}{N\pi_{i'}}\right)^2, \tag{3.16}$$

ou alternativement par

$$\hat{S}_b = \frac{1}{n_{b+}-K}\sum_{k=1}^{K}\sum_{i\in s_{bk}}\left(\frac{n_{b+}(y^{\alpha}_{ik} - \hat{\mathbf{b}}^{\alpha t}_k \mathbf{x}_i)}{N\pi_i} - \frac{1}{n_{bk}}\sum_{i'\in s_{bk}}\frac{n_{b+}(y^{\alpha}_{i'k} - \hat{\mathbf{b}}^{\alpha t}_k \mathbf{x}_{i'})}{N\pi_{i'}}\right)^2. \tag{3.17}$$

Dans certaines situations, la statistique de Wald calculée coïncide avec la statistique *F* utilisée dans les méthodes d'analyse fondées sur un modèle plus conventionnelles. Considérons un DBR intégré dans un plan d'échantillonnage autopondéré où la répartition des unités d'échantillonnage entre les traitements est proportionnelle sur les blocs, c'est-à-dire $\pi_i = n_{++}/N$ et $n_{bk}/n_{b+} = n_{+k}/n_{++}$ pour tout *b*. Soit $\bar{y}^{\alpha}_k = (1/n_{+k})\sum_{i\in s_k} y^{\alpha}_{ik}$, $\bar{y}^{\alpha}_b = (1/n_{b+})\sum_{i\in s_b} y^{\alpha}_{ik}$ et $\bar{y}^{\alpha} = (1/n_{++})\sum_{i\in s} y^{\alpha}_{ik}$. Si $n_{b+} \approx n_{b+}-1$, alors il s'ensuit pour l'estimateur étendu d'Horvitz-Thompson et l'estimateur groupé de la variance (3.16) que

$$\hat{d}_k = \frac{1}{n_{+k}}\frac{1}{n_{++}}\sum_{b=1}^{B}\sum_{k=1}^{K}\sum_{i\in s_{bk}}(y^{\alpha}_{ik} - \bar{y}^{\alpha}_k - \bar{y}^{\alpha}_b + \bar{y}^{\alpha})^2 \equiv \frac{\hat{d}}{n_{+k}}. \tag{3.18}$$

Soit $\bar{y}^{\alpha}_{bk} = (1/n_{bk})\sum_{i\in s_{bk}} y^{\alpha}_{ik}$. Sous l'estimateur groupé de la variance (3.17), il s'ensuit que

$$\hat{d}_k = \frac{1}{n_{+k}}\frac{1}{n_{++}}\sum_{b=1}^{B}\sum_{k=1}^{K}\sum_{i\in s_{bk}}(y^{\alpha}_{ik} - \bar{y}^{\alpha}_{bk})^2 \equiv \frac{\hat{d}}{n_{+k}}. \tag{3.19}$$

Si ces estimateurs groupés de la variance sont introduits par substitution dans la statistique de Wald (3.15), nous obtenons

$$W = \frac{1}{\hat{d}}\left(\sum_{k=1}^{K} n_{+k}(\bar{y}^{\alpha}_k)^2 - n_{++}(\bar{y}^{\alpha 2})\right).$$

Il s'ensuit que $W/(K-1)$ avec $\hat{d}$ défini en (3.18) correspond à la statistique *F* d'une analyse de variance à deux critères de classification sans interactions. Si on applique $\hat{d}$ défini par (3.19), alors $W/(K-1)$ correspond à la statistique *F* d'une analyse de variance à deux critères de classification avec interactions. (Voir, p. ex., Scheffé, 1959, ch. 4.) Pour un DRT, nous obtenons un estimateur groupé de la variance correspondant à un cas particulier de

(3.13) ou (3.14) en prenant b= 1, $n_{b+} = n_+$ et $n_{bk} = n_k$. Sous les deux estimateurs, $W/(K-1)$ correspond à la statistique F d'une analyse de variance à un critère de classification. (Voir, par exemple, Scheffé, 1959, ch. 3.) Si $n_{++} \to \infty$, alors la statistique F d'une analyse de variance à un et à deux critères de classification tend vers $\chi^2_{[K-1]}/(K-1)$, si bien que la statistique de Wald et la statistique F ont la même loi limite.

3.7 Avantages des dispositifs en blocs randomisés

Le principal avantage des DBR est l'élimination de la variation entre les blocs dans l'analyse des effets du traitement. La réduction de la variance réalisée par la mise en blocs au moyen d'expériences intégrées s'ensuit si l'on compare les composantes de la variance (3.10) et (3.11). Ceci suggère d'utiliser des structures d'échantillonnage, telles que les strates, les UPE, les grappes et les intervieweurs, comme variables de bloc dans un DBR. Considérons, par exemple, une expérience intégrée dans un plan d'échantillonnage stratifié. Il découle des expressions de la variance (3.10) et (3.11) que l'efficacité d'un plan d'échantillonnage stratifié est rendue nulle dans le cas d'un dispositif en randomisation totale, mais qu'elle est préservée sous un dispositif en blocs randomisés où les strates sont utilisées comme variables de bloc. En outre, un DBR avec les strates comme variable de bloc assurent que chaque strate soit suffisamment représentée à l'intérieur de chaque sous-échantillon. Dans le cas des expériences intégrées dans des plans d'échantillonnage à deux degrés ou des échantillons en grappes, il pourrait être efficace d'utiliser les UPE ou les grappes comme variables de bloc, puisque dans la plupart des situations pratiques, les unités d'échantillonnage provenant de la même UPE ou grappe ont un haut degré d'homogénéité comparativement aux unités d'échantillonnage provenant d'UPE ou de grappes différentes. Si la collecte des données est réalisée par ITAO ou par IPAO, alors les intervieweurs sont des variables de bloc éventuelles. Un DBR avec les intervieweurs comme variables de bloc élimine 1) la variation dans les observations due aux effets fixes et aléatoires d'intervieweur spécifiés dans le modèle de l'erreur de mesure (3.1) et 2) la variation entre intervieweurs du paramètre cible des unités d'échantillonnage. La deuxième composante pourrait être importante dans les enquêtes où les données sont recueillies par IPAO, parce que, dans ces conditions, les intervieweurs travaillent dans des régions distinctes assez petites. En général, les taux de réponse pourraient varier considérablement d'un intervieweur à l'autre. Ceci implique que la mise en blocs sur les intervieweurs augmente la possibilité de préserver les propriétés d'orthogonalité du plan d'expérience, par exemple une répartition proportionnelle des unités d'échantillonnage entre les traitements sur les blocs, qui améliore la puissance de l'expérience (voir Van den Brakel et Van Berkel, 2002). L'inconvénient principal de l'utilisation des intervieweurs comme variables de bloc est que ce genre de plan peut compliquer considérablement la collecte des données, puisque chaque intervieweur doit recueillir des données sous chacun des K traitements.

4. EXPÉRIENCES INTÉGRÉES AVEC DIVERS NIVEAUX DE RANDOMISATION

Nous allons maintenant examiner l'analyse d'expériences intégrées où les niveaux de randomisation des unités expérimentales et des unités d'échantillonnage diffèrent. Considérons, par exemple, une expérience intégrée dans un plan d'échantillonnage à deux degrés, où les UPE sont les ménages et où les unités secondaires d'échantillonnage (USE) sont les personnes. Conformément aux plans d'expérience considérés à la section précédente, les USE sont randomisées sur les traitements au moyen d'un dispositif en randomisation totale (DRT) ou d'un dispositif en blocs aléatoires (DBR). Toutefois, dans de nombreuses situations, en pratique, il pourrait être impossible d'appliquer des traitements différents à l'intérieur de la même UPE (ménage). Le cas échéant, les UPE sont randomisées sur les traitements et, conséquemment, les unités expérimentales ne coïncident pas avec les unités d'échantillonnage ultimes du plan d'échantillonnage.

Considérons une population finie U qui comprend M UPE. La j^{e} UPE comprend N_j USE. Soit $N = \sum_{j=1}^{M} N_j$ la taille de population. Pour tester l'hypothèse (3.2) nous disposons d'un échantillon à deux degrés s tiré à partir de U. Soit π_j^I la probabilité d'inclusion de premier ordre de la j^{e} UPE au premier degré du plan d'échantillonnage et $\pi_{i|j}^{II}$ la probabilité d'inclusion de premier ordre de la i^{e} USE au deuxième degré d'échantillonnage, sachant que la j^{e} UPE a été sélectionnée au premier degré. Dans le cas d'un DRT, l'échantillon d'UPE est randomisé sur les K traitements. Soit m_k le nombre d'UPE assignées au sous-échantillon s_k. Alors, $m_+ = \sum_{k=1}^{K} m_k$ représente le nombre total d'UPE dans s. La probabilité conditionnelle que l'UPE j soit assignée au traitement k, sachant la réalisation de premier

degré, est égale à m_k / m_+. Dans le cas d'un DBR, les UPE sont subdivisées de façon déterministe en B blocs s_b. Dans chaque bloc, les UPE sont randomisées sur les *K* traitements. Dans cette situation, les intervieweurs ou les strates du plan d'échantillonnage de premier degré sont des variables de bloc éventuelles. Soit m_{bk} le nombre d'UPE assignées au traitement *k* dans le bloc *b*. Alors $m_{b+} = \sum_{k=1}^{K} m_{bk}$ représente le nombre d'UPE dans le bloc *b*, $m_{+k} = \sum_{b=1}^{B} m_{bk}$ représente le nombre d'UPE dans le sous-échantillon s_k et $m_{++} = \sum_{b=1}^{B}\sum_{k=1}^{K} m_{bk}$ représente le nombre total d'UPE dans *s*. La probabilité conditionnelle que l'UPE *j* soit assignée au traitement *k*, sachant la réalisation de premier degré et que l'UPE $j \in s_b$, est égale à m_{bk} / m_{b+}. Si nous considérons chaque sous-échantillon comme étant une réalisation d'un échantillon à deux phases (voir la section 3.3), il s'ensuit que la probabilité d'inclusion de premier ordre de la j^{e} UPE au premier degré de s_k est égale à $\pi_j^{*I} = (m_k / m_+)\pi_j$ dans le cas d'un DRT ou à $\pi_j^{*I} = (m_{bk} / m_{b+})\pi_j$ dans le cas d'un DBR. La probabilité d'inclusion de premier ordre de la i^{e} USE dans le sous-échantillon s_k est donnée par $\pi_i^* = \pi_j^{*I}\pi_{i|j}^{II}$. Soit n_j le nombre d'USE tirées à partir de chacune des *j* UPE au deuxième degré d'échantillonnage et soit y_{ijk}^{α} l'observation obtenue à partir de la i^{e} USE, tirée à partir de la j^{e} UPE assignée au k^{e} traitement. L'estimateur d'Horvitz-Thompson de $\bar{Y}_k$ fondé sur les observations dans s_k est donné par

$$\hat{\bar{Y}}_k^{\alpha} = \frac{1}{N}\sum_{j\in s_k}\sum_{i=1}^{n_j}\frac{y_{ijk}^{\alpha}}{\pi_j^{*I}\pi_{i|j}^{II}} \equiv \frac{1}{N}\sum_{j\in s_k}\frac{\hat{y}_{jk}^{\alpha}}{\pi_j^{*I}}, \tag{4.1}$$

où $\hat{y}_{jk}^{\alpha}$ représente l'estimateur d'Horvitz-Thompson pour le total de population de la j^{e} UPE assignée au k^{e} traitement. L'estimateur par la régression généralisée de $\bar{Y}_k$ fondé sur les observations dans s_k est donné par l'expression (3.6), où l'estimateur d'Horvitz-Thompson pour les coefficients de régression $\hat{\mathbf{b}}_k^{\alpha}$ et les moyennes de population de l'information auxiliaire $\hat{\bar{\mathbf{X}}}_k$ sont définis par (3.4) et (3.5), respectivement en utilisant les probabilités d'inclusion de premier ordre calculées à la présente section.

La matrice des covariances des contrastes entre la linéarisation de Taylor de premier ordre des moyennes de sous-échantillon est de nouveau donnée par $\mathbf{CVC}^t = E_{\alpha}E_s\mathbf{CDC}^t$. Dans le cas d'un DBR, les éléments diagonaux de $\mathbf{D}$ sont donnés par

$$d_k = \sum_{b=1}^{B}\frac{1}{m_{bk}}\frac{1}{m_{b+}-1}\sum_{j\in s_b}\left(\frac{m_{b+}(\hat{y}_{jk}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\hat{\mathbf{x}}_j)}{N\pi_j^I} - \frac{1}{m_{b+}}\sum_{j'\in s_b}\frac{m_{b+}(\hat{y}_{j'k}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\hat{\mathbf{x}}_{j'})}{N\pi_{j'}^I}\right)^2, \tag{4.2}$$

où

$$\hat{y}_{jk}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\hat{\mathbf{x}}_j = \sum_{i=1}^{n_j}\frac{y_{ijk}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\mathbf{x}_{ij}}{\pi_{i|j}^{II}}. \tag{4.3}$$

Ici $\mathbf{x}_{ij}$ représente le vecteur de l'information auxiliaire de la i^{e} USE tirée à partir de la j^{e} UPE. Un estimateur approximativement sans biais par rapport au plan d'échantillonnage de la matrice des covariances des contrastes entre les moyennes d'échantillon est donné par $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$. Dans le cas d'un DBR, les éléments diagonaux de $\hat{\mathbf{D}}$ sont donnés par

$$\hat{d}_k = \sum_{b=1}^{B}\frac{1}{m_{bk}}\frac{1}{m_{bk}-1}\sum_{j\in s_{bk}}\left(\frac{m_{b+}(\hat{y}_{jk}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\hat{\mathbf{x}}_j)}{N\pi_j^I} - \frac{1}{m_{bk}}\sum_{j'\in s_{bk}}\frac{m_{b+}(\hat{y}_{j'k}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\hat{\mathbf{x}}_{j'})}{N\pi_{j'}^I}\right)^2, \tag{4.4}$$

où $\hat{y}_{jk}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\hat{\mathbf{x}}_j$ est donné par (4.3). Dans le cas d'un DRT, une expression des éléments diagonaux de $\mathbf{D}$ et $\hat{\mathbf{D}}$ est donnée par (4.2), (4.3) et (4.4) en prenant B=1, $m_{bk} = m_k$ et $m_{b+} = m_+$.

Une expression de l'estimateur étendu d'Horvitz-Thompson est définie par (3.14) en utilisant les probabilités d'inclusion de premier ordre calculées à la présente section. Une expression des composantes de la variance $\hat{d}_k$ est donnée par (4.3) où

$$\hat{y}_{jk}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha t}\hat{\mathbf{x}}_j = \sum_{i=1}^{n_j} \frac{y_{ijk}^{\alpha} - \tilde{Y}_k^{\alpha}}{\pi_{i|j}^{II}} = \hat{y}_{jk}^{\alpha} - \tilde{Y}_k^{\alpha}\hat{N}_j .$$

Pour améliorer la méthode d'estimation de la variance, on peut appliquer les estimateurs groupés de la variance proposée à la section 3.6 de façon équivalente. L'hypothèse (3.2) peut être testée au moyen de la statistique de Wald (3.15), en utilisant les estimations à partir des sous-échantillons et les composantes de la variance calculées à la présente section.

Une expérience intégrée dans un plan d'échantillonnage à deux degrés peut être conçue comme un DBR où les UPE sont les variables de bloc et les USE sont les unités expérimentales, ou comme une expérience où les UPE sont les unités expérimentales. Si la variation entre les UPE est grande et que la variation entre les USE à l'intérieur des UPE est faible, alors il est préférable d'utiliser un DBR où les UPE sont les variables de bloc et les USE, les unités expérimentales. Si la variation entre les UPE est faible et que la variation entre les USE à l'intérieur des UPE est grande, alors sous les deux plans d'expérience, une composante importante de la variance résultera de la variation entre les USE à l'intérieur des UPE. Dans cette situation, un DBR où les UPE sont les variables de bloc et les USE, les unités expérimentales a l'avantage d'offrir un plus grand nombre de degrés de liberté dans la méthode d'estimation de la variance. Toutefois, si les UPE sont complètement observées, autrement dit que le plan d'échantillonnage est un échantillonnage en grappes, alors il pourrait être efficace d'appliquer un plan d'expérience où les UPE sont les unités expérimentales. L'utilisation des UPE comme unités expérimentales dans cette situation évite l'introduction d'une composante importante de la variance provenant de la variation entre les USE à l'intérieur des UPE, puisque ces dernières sont observées complètement sous l'un des K traitements. Comme nous le soulignons dans l'introduction de la présente section, un plan expérimental où les UPE sont les unités expérimentales sera généralement pris en considération pour des raisons pratiques s'il n'est pas possible d'appliquer des traitements différents à l'intérieur d'une UPE.

5. TEST D'HYPOTHÈSES AU SUJET DES RATIOS DE TOTAUX DE POPULATION

La méthode d'analyse élaborée aux sections qui précèdent peut être appliquée pour tester les hypothèses au sujet des estimations des moyennes ou des totaux de population. Néanmoins, dans le cas de nombreuses enquêtes par sondage, les paramètres cibles sont définis comme étant le ratio de deux totaux de population. Par conséquent, nous étendrons la méthode des sections précédentes à l'analyse des ratios. Soit $R_k = Y_k / Z_k$ le ratio de deux totaux de population observé sous le traitement k. Alors $\mathbf{R} = (R_1, \ldots, R_K)^t$ représente le vecteur de dimension K des ratios observés sous les divers traitements de l'expérience. Le but de l'expérience est de tester les hypothèses suivantes :

$$\begin{aligned} H_0 &: \mathbf{CR} = \mathbf{0} \\ H_1 &: \mathbf{CR} \neq \mathbf{0} \end{aligned} . \qquad (5.1)$$

Soit y_{ikl}^{α} et z_{ikl}^{α} les observations obtenues à partir de l'unité expérimentale i assignée à l'intervieweur l et au traitement k à la α^{e} occasion. Nous supposons que les deux variables sont une réalisation du modèle de l'erreur de mesure (3.1). L'estimateur par la régression généralisée R_k est donnée par $\hat{R}_{k;greg}^{\alpha} = \hat{Y}_{k;greg}^{\alpha} / \hat{Z}_{k;greg}^{\alpha}$, où $\hat{Y}_{k;greg}^{\alpha}$ et $\hat{Z}_{k;greg}^{\alpha}$ sont les estimateurs par la régression généralisée des totaux de population Y_k et Z_k fondés sur les observations obtenues dans le sous-échantillon s_k. Enfin, $\hat{\mathbf{R}}_{\mathbf{GREG}}^{\alpha} = (\hat{R}_{1;greg}^{\alpha}, \ldots, \hat{R}_{K;greg}^{\alpha})^t$ représente l'estimateur par la régression généralisée de $\mathbf{R}$.

Soit $\mathbf{V}$ la matrice de covariance de $\hat{\mathbf{R}}^{\alpha}_{\mathbf{GREG}}$. Pour établir une expression de la matrice des covariances des *K*-1 contrastes entre les $\hat{\mathbf{R}}^{\alpha}_{\mathbf{GREG}}$, il faut d'abord obtenir une approximation linéaire de $\hat{R}^{\alpha}_{k;greg}$. La linéarisation de Taylor de premier ordre de $\hat{R}^{\alpha}_{k;greg}$ au voisinage du point R_k est donnée par

$$\hat{R}^{\alpha}_{k;greg} \approx R_k + \frac{1}{Z_k}(\hat{Y}^{\alpha}_{k;greg} - R_k \hat{Z}^{\alpha}_{k;greg}) . \tag{5.2}$$

Subséquemment, $\hat{Y}^{\alpha}_{k;greg}$ et $\hat{Z}^{\alpha}_{k;greg}$ dans (5.2) sont linéarisés par une approximation de Taylor de premier ordre au voisinage de $(Y_k, \mathbf{b}_k, \mathbf{X})$ et de $(Z_k, \mathbf{d}_k, \mathbf{X})$, respectivement, où $\mathbf{d}_k$ représente le vecteur de dimension *H* des coefficients de régression de la fonction de régression de z_{ikl} sur $\mathbf{x}_i$. Alors, il s'ensuit que

$$\hat{R}^{\alpha}_{k;greg} \approx R_k + \frac{1}{Z_k}(\hat{Y}^{\alpha}_k + \mathbf{b}^t_k(\mathbf{X} - \hat{\mathbf{X}}_k) - R_k(\hat{Z}^{\alpha}_k + \mathbf{d}^t_k(\mathbf{X} - \hat{\mathbf{X}}_k))) \equiv R_k + \hat{E}^{\alpha}_k + \frac{1}{Z_k}(\mathbf{b}^t_k\mathbf{X} - R_k\mathbf{d}^t_k\mathbf{X}) ,$$

où

$$\hat{E}^{\alpha}_k = \frac{1}{Z_k}\sum_{i\in s_k}\frac{y^{\alpha}_{ik} - \mathbf{b}^t_k\mathbf{x}_i - R_k(z^{\alpha}_{ik} - \mathbf{d}^t_k\mathbf{x}_i)}{\pi^*_i} .$$

La matrice des covariances des contrastes de $\hat{\mathbf{R}}^{\alpha}_{\mathbf{GREG}}$ est approximée par la matrice des covariances de la linéarisation de Taylor de premier ordre des contrastes de $\hat{\mathbf{R}}^{\alpha}_{\mathbf{GREG}}$, qui est obtenue par celle de $\hat{\mathbf{E}}^{\alpha} = (\hat{E}^{\alpha}_1, \ldots, \hat{E}^{\alpha}_K)^t$. La même analyse que celle utilisée pour établir la matrice des covariances des contrastes entre les moyennes de sous-échantillon à la section 3 peut être appliquée pour montrer que $\mathbf{CVC}^t = E_{\alpha}E_s\mathbf{CDC}^t$. Dans le cas d'un DBR, $\mathbf{D}$ est une matrice diagonale dont les éléments sont

$$d_k = \frac{1}{Z^2_k}\sum_{b=1}^{B}\frac{1}{n_{bk}}\frac{1}{n_{b+}-1}\sum_{i\in s_b}\left(\frac{n_{b+}e^{\alpha}_{ik}}{\pi_i} - \frac{1}{n_{b+}}\sum_{i'\in s_b}\frac{n_{b+}e^{\alpha}_{i'k}}{\pi_{i'}}\right)^2 , \tag{5.3}$$

où $e^{\alpha}_{ik} = y^{\alpha}_{ik} - \mathbf{b}^t_k\mathbf{x}_i - R_k(z^{\alpha}_{ik} - \mathbf{d}^t_k\mathbf{x}_i)$. Un estimateur de $\mathbf{CVC}^t$ est donné par $\mathbf{C}\hat{\mathbf{D}}\mathbf{C}^t$ où les éléments diagonaux de $\hat{\mathbf{D}}$ sont donnés par

$$\hat{d}_k = \frac{1}{\hat{Z}^{\alpha^2}_{k;greg}}\sum_{b=1}^{B}\frac{1}{n_{bk}}\frac{1}{n_{bk}-1}\sum_{i\in s_{bk}}\left(\frac{n_{b+}\hat{e}^{\alpha}_{ik}}{\pi_i} - \frac{1}{n_{bk}}\sum_{i'\in s_{bk}}\frac{n_{b+}\hat{e}^{\alpha}_{i'k}}{\pi_{i'}}\right)^2 \tag{5.4}$$

où $\hat{e}^{\alpha}_{ik} = y^{\alpha}_{ik} - \hat{\mathbf{b}}^{\alpha^t}_k\mathbf{x}_i - \hat{R}^{\alpha}_{k;greg}(z^{\alpha}_{ik} - \hat{\mathbf{d}}^{\alpha^t}_k\mathbf{x}_i)$. Pour les expériences intégrées dans des plans d'échantillonnage à deux degrés ou plus, où les UPE correspondent aux unités expérimentales (section 4), une expression des éléments diagonaux de $\mathbf{D}$ sous un DBR est donnée par

$$d_k = \frac{1}{Z^2_k}\sum_{b=1}^{B}\frac{1}{m_{bk}}\frac{1}{m_{b+}-1}\sum_{j\in s_b}\left(\frac{m_{b+}\hat{e}^{\alpha}_{jk}}{\pi^I_j} - \frac{1}{m_{b+}}\sum_{j'\in s_b}\frac{m_{b+}\hat{e}^{\alpha}_{j'k}}{\pi^I_{j'}}\right)^2 , \tag{5.5}$$

où

$$\hat{e}_{jk}^{\alpha} = \sum_{i=1}^{n_j} \frac{y_{ijk}^{\alpha} - \mathbf{b}_k^t \mathbf{x}_{ij} - R_k (z_{ijk}^{\alpha} - \mathbf{d}_k^t \mathbf{x}_{ij})}{\pi_{i|j}^{II}}.$$

Une expression des éléments diagonaux de $\hat{\mathbf{D}}$ sous un DBR est donnée par

$$\hat{d}_k = \frac{1}{\hat{Z}_{k;greg}^{\alpha^2}} \sum_{b=1}^{B} \frac{1}{m_{bk}} \frac{1}{m_{bk} - 1} \sum_{j \in s_{bk}} \left(\frac{m_{b+} \hat{\hat{e}}_{jk}^{\alpha}}{\pi_j^I} - \frac{1}{m_{bk}} \sum_{i' \in s_{bk}} \frac{m_{b+} \hat{\hat{e}}_{j'k}^{\alpha}}{\pi_{j'}^I} \right)^2, \tag{5.6}$$

où

$$\hat{\hat{e}}_{jk}^{\alpha} = \sum_{i=1}^{n_j} \frac{y_{ijk}^{\alpha} - \hat{\mathbf{b}}_k^{\alpha^t} \mathbf{x}_{ij} - \hat{R}_{k;greg}^{\alpha} (z_{ijk}^{\alpha} - \hat{\mathbf{d}}_k^{\alpha^t} \mathbf{x}_{ij})}{\pi_{i|j}^{II}}.$$

Une expression des éléments diagonaux d_k et $\hat{d}_k$ sous un DRT s'obtient comme un cas spécial de (5.3) et (5.4) ou de (5.5) et (5.6), respectivement en prenant $B = 1$, $n_{bk} = n_k$ et $n_{b+} = n_+$ ou $m_{bk} = m_k$ et $m_{b+} = m_+$. Les expressions pour l'estimateur étendu d'Horvitz-Thompson découlent directement d'un cas spécial de l'estimateur par la régression généralisée avec le modèle commun de la moyenne comme scénario de pondération. Les estimateurs groupés de la variance proposés à la section 3.6 peuvent être appliqués de la même façon aux estimateurs de la variance des contrastes entre les ratios. Enfin, l'hypothèse (5.1) peut être testée au moyen de la statistique de Wald définie par (3.15) où $\hat{Y}_{k;greg}^{\alpha}$ est remplacé par $\hat{R}_{k;greg}^{\alpha}$ et où les expressions appropriées des composantes de la variance $\hat{d}_k$ sont appliquées.

6. DISCUSSION

Pour tester les hypothèses au sujet des estimations des paramètres d'une population finie d'une enquête par sondage observés sous différentes approches d'enquête, ou traitements, nous établissons une statistique de Wald fondée sur le plan d'échantillonnage pour l'analyse des dispositifs en randomisation totale (DRT) ou des dispositifs en blocs randomisés (DBR) intégrés dans des plans d'échantillonnage généralement complexes. Les estimateurs des paramètres et de la variance sont fondés sur l'estimateur d'Horvitz-Thompson ou sur l'estimateur par la régression généralisée, qui permettent de tenir compte du mécanisme de randomisation du plan d'échantillonnage, du plan expérimental et de la méthode d'estimation de l'enquête régulière dans l'analyse de l'expérience. L'application d'une méthode d'analyse fondée sur le plan de sondage conjuguée à la sélection des unités expérimentales à partir d'une population finie par échantillonnage aléatoire nous permet de généraliser les résultats observés sur les échantillons particuliers de l'expérience à la population d'enquête complète.

À première vue, une méthode d'analyse qui tient compte du plan d'échantillonnage pourrait être considérée comme une alternative pour tester les hypothèses au sujet des coefficients de régression d'un modèle linéaire ou de combinaisons linéaires des estimations sur les sous-échantillons, qui reflètent les effets de traitement de l'expérience. Cependant, la superposition du plan expérimental au plan d'échantillonnage détermine quelles caractéristiques particulières du plan d'échantillonnage seront annulées ou préservées. Par exemple, l'effet de l'échantillonnage stratifié ou de l'échantillonnage à deux degrés sur la variance des effets de traitement est annulé sous un dispositif en randomisation totale. Par conséquent, l'application de ce genre de méthode d'analyse de rechange peut encore donner lieu à des résultats incorrects, puisqu'il n'est pas tenu compte de la superposition du plan expérimental au plan d'échantillonnage dans la variance des effets du traitement.

Puisque les sous-échantillons sont tirés sans remise à partir d'une population finie, conformément à un plan d'échantillonnage complexe général, il faut s'attendre à obtenir une expression assez compliquée de la matrice des covariances avec des termes hors diagonale non nul. Bien que nous tenions compte des plans d'échantillonnage complexes généraux, ainsi que de la dépendance entre les estimations sur les sous-échantillons, l'estimateur établi

pour cette matrice de covariances a la même structure que si les sous-échantillons étaient tirés indépendamment les uns des autres et que les unités d'échantillonnage étaient sélectionnées avec probabilités inégales avec remise. Ni les probabilités d'inclusion de deuxième ordre ni les covariances par rapport au plan de sondage entre les estimations de sous-échantillon ne doivent être connues, ce qui simplifie considérablement l'analyse. Par conséquent, on obtient une statistique de Wald, établie dans une perspective fondée sur le plan d'échantillonnage sous plans d'échantillonnage complexes généraux, qui a la structure assez simple et attirante des méthodes types d'analyse fondées sur un modèle.

La méthode d'analyse proposée dans le présent article est implémentée à l'heure actuelle dans le progiciel Bascula de Statistique Pays-Bas pour faciliter l'application de ces méthodes. La méthode est étendue à l'analyse des plans d'analyse factorielle intégrés dans les enquêtes par sondage pour tester simultanément l'effet de deux facteurs ou plus sur les estimations d'enquête (Van den Brakel, 2003).

RÉFÉRENCES

Cochran,W. (1977), "*Sampling Techniques*", New York: Wiley.

Fienberg, S.E. and J.M. Tanur (1987), "Experimental and Sampling Structures: Parallels Diverging and Meeting", *International Statistical Review*, 55, pp. 75-96.

Fienberg, S.E. and J.M. Tanur (1988), "From the inside out and the outside in: Combining experimental and sampling structures", *The Canadian Journal of Statistics*, 16, pp. 135-151.

Lehmann, E.L. (1975), "*Nonparametrics: Statistical Methods Based on Ranks*", New York: McGraw-hill.

Montgomery, D.C. (1997), "*Design and Analysis of Experiments*", New York: Wiley.

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

Scheffé, H. (1959), "*The Analysis of Variance*", New York: Wiley.

Van den Brakel, J.A. (2001), "*Design and Analysis of Experiments Embedded in Complex Sample Surveys*" Ph.D. Thesis, Erasmus University Rotterdam.

Van den Brakel, J.A. (2004). "Analysis of factorial designs embedded in sample surveys", Unpublished research paper, Heerlen, The Netherlands, Statistics Netherlands.

Van den Brakel, J.A. and D. Binder (2000), "Variance Estimation for Experiments Embedded in Complex Sampling Schemes", *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section, American Statistical Association*, pp. 805-810.

Van den Brakel, J.A. and R.H. Renssen (1998), "Design and Analysis of Experiments Embedded in Sample Surveys", *Journal of Official Statistics*, 14, pp. 277-295.

Van den Brakel, J.A. and C.A.M. van Berkel (2002), "A Design-Based Analysis Procedure for Two-treatment Experiments Embedded in Sample Surveys", *Journal of Official Statistics*, 18, pp. 217-231.