

ESTIMATION DES EFFETS DE L'INTERVIEWEUR SUR DES RÉPONSES BINAIRES

Alastair Scott¹ et Peter Davis²

RÉSUMÉ

Dans les enquêtes où l'on a besoin d'intervieweurs spécialisés et chevronnés, on doit souvent se contenter d'un petit nombre de personnes possédant une solide formation et attribuer à chacune une charge de travail importante. On sait très bien que cette contrainte peut entraîner une variabilité de l'intervieweur qui a une incidence relativement grande sur l'erreur totale, notamment pour l'estimation de quantités simples comme les moyennes et les proportions. Dans un article antérieur (Davis et Scott (1995)), nous en avons examiné l'incidence sur des réponses continues à l'aide d'un modèle d'éléments linéaires de la variance. Toutefois, la plupart des réponses aux questionnaires sur la santé sont binaires et l'on sait que cette démarche entraîne une sous-estimation des corrélations intra-grappe et intra-intervieweur à l'égard de réponses binaires. Dans la présente communication, nous utilisons un modèle binaire à plusieurs niveaux pour étudier l'incidence de la variabilité de l'intervieweur sur l'estimateur de proportions.

MOTS-CLÉS : Variance de l'intervieweur; Effet du plan d'échantillonnage; Modèles à plusieurs niveaux.

1. INTRODUCTION

Bon nombre d'enquêtes, notamment dans le domaine de la santé, nécessitent des intervieweurs ou des observateurs possédant une solide formation spécialisée. Comme il est difficile de trouver des personnes suffisamment qualifiées, les enquêteurs sont souvent obligés de faire appel à un nombre relativement restreint d'intervieweurs, chacun ayant, par conséquent, une charge de travail importante. À l'heure actuelle, par exemple, nous menons une étude des événements indésirables survenus dans les hôpitaux de la Nouvelle-Zélande. Notre étude utilise un plan à deux degrés selon lequel, au premier degré, on prélève un échantillon stratifié d'hôpitaux avec probabilité proportionnelle à la taille (PPT) et, au deuxième degré, un échantillon systématique de dossiers de malades dans chacun des hôpitaux choisis. Puis, on examine attentivement chaque dossier choisi pour y déceler toute trace d'erreur médicale. Cet examen exige de l'intervieweur un niveau élevé de connaissances médicales et une période intensive de formation pour qu'il se familiarise avec les protocoles d'étude. Comme il était très difficile de trouver des personnes possédant l'expérience nécessaire et qui étaient libres pour participer durant la période de l'étude, nous avons dû nous contenter d'un petit nombre d'examineurs et leur attribuer une charge de travail moyenne d'environ 300 dossiers, ce qui est assez courant dans ce genre d'étude.

On sait très bien que l'incidence d'une charge de travail élevée de l'intervieweur sur la variance des estimations des moyennes de la population peut être importante, même lorsque la variabilité de l'intervieweur est faible (voir Groves (1982, chapitre 8) ou Lessler et Kalsbeek (1992, section 11.3), par exemple). Sur des statistiques plus complexes, comme des coefficients de régression ou des rapports des probabilités, l'incidence est habituellement moins forte (Kish et Frankel (1974)), mais elle peut aussi être importante dans certains cas. Dans un article antérieur (Davis et Scott (1995)), nous avons examiné les

¹ Alastair Scott, Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1, New Zealand

² Peter Davis, Department of Public Health & General Practice, Christchurch School of Medicine, Christchurch, New Zealand

effets de la variabilité de l'intervieweur sur les moyennes et, plus particulièrement, sur les comparaisons entre moyennes de domaines, à l'aide d'un modèle d'éléments de la variance du type publié dans l'étude novatrice classique de Hartley et Rao (1969). Nous avons ensuite appliqué les résultats aux données d'une enquête à grande échelle sur les soins dentaires, dont les intervieweurs spécialisés avaient des charges de travail élevées. Dans notre étude empirique, nous avons, pour une bonne part, fait abstraction du fait qu'un grand nombre de variables étaient binaires, observant simplement : « On sait très bien qu'une telle méthode entraîne une sous-estimation des éléments de la variance pour les données binaires. Les effets du plan d'échantillonnage sur les proportions que nous avons estimées devraient donc être considérées comme la limite inférieure de la fourchette ». On trouve des observations semblables dans Anderson et Aitken (1985). Comme les variables binaires constituent la norme dans la plupart des enquêtes sur la santé, il importe d'étudier la nature et l'étendue de cette sous-estimation. Tel est le sujet de la présente communication. Dans la prochaine section, nous proposons une théorie fondée sur un modèle logistique normal simple. Dans la section 3, nous examinons certains résultats empiriques de l'étude des événements indésirables mentionnée plus haut. Dans la dernière section, nous formulons quelques conclusions provisoires, notamment en nuanciant notre observation antérieure concernant la sous-estimation des effets du plan d'échantillonnage.

2. THÉORIE DE BASE

2.1 Échantillonnage aléatoire simple

Pour plus de simplicité, nous prenons d'abord le cas de l'échantillonnage aléatoire simple, qui permet de considérer les effets de l'intervieweur sans la complication supplémentaire des unités d'échantillonnage. Nous commençons par passer en revue les résultats courants pour la moyenne d'une variable continue (voir Kish (1962), par exemple). Si nous disposons de mesures d'une variable continue, alors un modèle naturel pour l'observation à partir du i^e répondant avec le j^e intervieweur est le modèle mixte linéaire simple

$$Y_{ir} = \mu + a_i + e_{ij} \quad (1)$$

où a_i représente l'effet du i^e intervieweur et e_{ij} représente l'effet du répondant individuel. Supposons que les valeurs de a_i et de e_{ij} sont des variables aléatoires non corrélées ayant respectivement des moyennes nulles et des variances σ_a^2 et σ^2 . Selon ce modèle, la variance de la moyenne de l'échantillon est alors

$$\begin{aligned} \text{Var}\{\bar{Y}\} &= \text{Var}\left\{\frac{1}{n} \sum \sum Y_{ir}\right\} = \frac{1}{n^2} \sum_i \text{Var}\{n_i a_i + \bar{e}_i\} \\ &= \frac{1}{n^2} \left\{ \left(\sum_i n_i^2 \right) \sigma_a^2 + n \sigma^2 \right\} = \frac{1}{n} (\sigma^2 + \tilde{n}_1 \sigma_a^2), \end{aligned}$$

où n_i est le nombre de réponses obtenues par le i^e intervieweur et $\tilde{n}_1 = \frac{\sum n_i^2}{\sum n_i}$. Voyons maintenant ce qui se passerait si les observations étaient tirées indépendamment du modèle (1). La variance de la moyenne de l'échantillon serait alors

$$\text{Var}\{\bar{Y}\} = \text{Var}\left\{\frac{1}{n} \sum \sum Y_{ir}\right\} = \frac{1}{n} (\sigma^2 + \sigma_a^2) = V_0,$$

par exemple. Ainsi, l'effet des interviews multiples est de gonfler cette variance par un facteur D_0 donné par

$$D_0 = \frac{\text{Var}\{\bar{Y}\}}{V_0} = 1 + (\tilde{n}_1 - 1) \rho_1$$

où $\rho_I = \frac{\sigma_a^2}{\sigma^2 + \sigma_a^2}$ est le coefficient de corrélation intra-intervieweur. Nous appellerons D_0 « effet du plan

d'échantillonnage », quoiqu'il représente l'inflation de la variance selon le modèle des effets aléatoires (1) plutôt que pour l'échantillonnage répété d'une population finie fixe, comme le veut la définition habituelle. Nous voyons que l'incidence de la variabilité de l'intervieweur peut être élevée lorsque la charge de travail moyenne est élevée, même si la corrélation intra-intervieweur, ρ_I , est assez faible.

Voyons maintenant ce qui se passe lorsque la réponse est binaire. Le modèle (1) ne convient plus, mais nous pouvons l'étendre de manière naturelle en supposant qu'il existe une variable latente continue sous-jacente, par exemple Y^* , obéissant au modèle (1). Nous n'observons pas directement Y^* , mais seulement la réponse binaire Y , qui prend la valeur 1 seulement lorsque Y^* se situe en deçà d'une valeur limite (inconnue), par exemple c . Ainsi, Y_{ir} , la valeur du i^e répondant avec le r^e intervieweur, est égale à 1 seulement lorsque

$$Y_{ir}^* - c = \theta + a_i^* + e_{ir}^* \leq 0, \quad (2)$$

où $\theta = \mu^* - c$. Il convient de noter que le modèle (2) n'est défini qu'en fonction d'une constante multiplicative, puisque la multiplication du membre gauche de l'équation (2) par une constante positive arbitraire n'a pas d'incidence sur son signe ni, par conséquent, sur la valeur de Y_{ir} . Cela signifie que nous pouvons uniquement estimer le ratio, $\frac{\sigma_a^2}{\sigma^2}$, plutôt que les éléments individuels de la variance, σ_a^{*2} et σ^{*2} .

Heureusement, la plupart des quantités qui nous intéressent (telles que ρ_I^*) sont des fonctions de ce ratio; on peut donc les estimer à partir des données binaires observées, si les intervieweurs sont bien déployés.

Pour obtenir la valeur de $P\{Y_{ir}=1\}$ à partir du modèle (2), nous devons préciser les distributions de a_i^* et e_{ir}^* . Supposons que a_i^* présente une distribution normale et que e_{ir}^* présente une distribution logistique. (Il serait plus conforme à l'esprit du modèle initial de Hartley-Rao de supposer que e_{ir}^* présente aussi une distribution normale. Toutefois, les résultats sont presque identiques selon l'un ou l'autre modèle, à moins de se situer à l'extrémité des distributions, et l'hypothèse logistique permet de simplifier quelque peu la théorie et les calculs numériques.) On trouvera dans De Leeuw et Kreft (2001) un bon examen des logiciels permettant d'ajuster ce modèle de distribution normale logistique. Comme nous pouvons effectuer un ré-échelonnement arbitraire sans modifier la valeur de Y_{ir} , nous pouvons supposer de manière tout aussi générale que e_{ir}^* présente une distribution logistique *courante*, laquelle présente la variance $\pi^2/3$. Cela

signifie que $\text{Var}\{a_i^*\}$ en (2) correspond à $\sigma_a^2 / \frac{\pi^2}{3} = \frac{\pi^2}{3} \cdot \frac{\sigma_a^2}{\sigma^2}$ en (1). Ainsi, notre estimation de σ_a^{*2} est,

en réalité, une estimation de $\frac{\pi^2}{3} \cdot \frac{\sigma_a^2}{\sigma^2}$, ce qui permet d'établir $\hat{\rho}_I^* = \frac{\hat{\sigma}_a^{*2}}{\hat{\sigma}_a^{*2} + \pi^2/3}$ comme notre estimation de ρ_I^* , la corrélation intra-intervieweur pour la variable latente sous-jacente.

Ce n'est pas la variabilité de l'intervieweur proprement dite qui nous intéresse avant tout, mais plutôt son effet sur la variance de l'estimateur de la proportion de la population. Nous pouvons énoncer cette estimateur, \hat{p} , sous la forme suivante :

$$\hat{p} = \frac{1}{n} \sum_i \sum_r Y_{ir} = \frac{1}{n} \sum_i n_i \hat{p}_i,$$

où \hat{p}_i représente la proportion observée de réponses « Oui » parmi les n_i réponses enregistrées par le i^e intervieweur. Comme les réponses obtenues par différents intervieweurs sont indépendantes, il s'ensuit que

$Var\{\hat{p}\} = \frac{1}{n^2} \sum_i n_i^2 Var\{\hat{p}_i\}$. Or, d'après les résultats courants pour les moyennes et les variances conditionnelles, nous pouvons écrire

$$Var\{\hat{p}_i\} = E\{Var\{\hat{p}_i | a_i^*\}\} + Var\{E\{\hat{p}_i | a_i^*\}\}$$

où, d'après le modèle (2),

$$E\{\hat{p}_i | a_i^*\} = P\{Y_{ir} = 1 | a_i^*\} = e^{\theta+a_i^*} / (1 + e^{\theta+a_i^*}) = p(a_i^*), \quad (3)$$

par exemple, et

$$Var\{\hat{p}_i | a_i^*\} = \frac{p(a_i^*)(1 - p(a_i^*))}{n_i}.$$

Il s'ensuit que

$$Var\{\hat{p}\} = \frac{1}{n^2} \sum_i n_i^2 Var\{\hat{p}_i\} = \frac{1}{n} [E^* + \tilde{n}V^*],$$

où $\tilde{n}_i = \sum_i n_i^2 / n$, comme avant, $E^* = E^*(\theta, \sigma_a^{*2}) = E\{p(a_i^*)(1 - p(a_i^*))\}$ et $V^* = V^*(\theta, \sigma_a^{*2}) = Var\{p(a_i^*)\}$.

Si les observations avaient été faites indépendamment (c'est-à-dire par un intervieweur différent pour chaque répondant), la variance de l'estimateur de la proportion \hat{p} aurait alors été

$$V_0 = \frac{1}{n} [E^* + V^*].$$

Ainsi, l'effet des intervieweurs qui mènent plusieurs interviews est de gonfler la variance par un facteur

$$D_0^* = \frac{Var\{\hat{p}\}}{V_0} = 1 + (\tilde{n}_i - 1)\phi$$

et $\phi = \frac{V^*}{E^* + V^*}$. L'évaluation exacte de E^* et V^* nécessite une intégration numérique, mais nous

pouvons calculer des valeurs approximatives à l'aide de méthodes de linéarisation courantes lorsque les effets de l'intervieweur sont faibles. En extrapolant $p(a) = e^{\theta+a} / (1 + e^{\theta+a})$ en fonction de $a=0$, on obtient

$$p(a) \approx p_0 + p_0(1 - p_0)a + \frac{1}{2} p_0(1 - p_0)(1 - 2p_0)a^2$$

où $p_0 = p(0)$. Cela nous mène à

$$V^* = Var\{p(a_i^*)\} \approx p_0^2(1 - p_0)^2 \sigma_a^{*2}$$

et, après une intervention algébrique, à

$$E^* = E\{p(a_i^*)[1 - p(a_i^*)]\} \approx p_0(1 - p_0) \left\{ 1 + \left[\frac{(1 - 2p_0)^2}{2} - p_0(1 - p_0) \right] \sigma_a^{*2} \right\}.$$

En intégrant ces expressions à l'expression de ϕ et en simplifiant, on obtient

$$\phi \approx \frac{\sigma_a^{*2}}{\frac{(1-2p_0)^2}{2}\sigma_a^{*2} + \frac{1}{p_0(1-p_0)}} = \frac{\rho^*}{\frac{(1-2p_0)^2}{2}\rho^* + \frac{3(1-\rho^*)}{\pi^2 p_0(1-p_0)}} \approx \frac{\pi^2 p_0(1-p_0)\rho^*}{3}.$$

pour une faible valeur de ρ^* . Il convient de noter que ϕ est toujours inférieur à ρ^* et a tendance à être très inférieur, à moins que p_0 ne soit proche de 0,5. Par exemple, si $p_0 = .1$, comme c'est habituellement le cas dans nos études, alors $\phi \approx .3\rho$. (Par contre, si $p_0 = 0.5$, alors $\phi \approx .82\rho$.) Cela signifie que la variabilité de l'intervieweur a habituellement une incidence beaucoup plus faible sur la variance de la variable binaire catégorisée que sur celle de la variable latente continue sous-jacente.

Qu'arrive-t-il si nous tentons d'ajuster le modèle des effets aléatoires linéaires (1), qui ne convient pas, aux données binaires produites par le modèle (2)? L'estimation de l'élément de la variance de l'intervieweur, par exemple $\hat{\sigma}_a^2$, devient alors une estimation de $\text{Var}\{E\{Y | a\}\} = \text{Var}\{p(a)\} = V^*$. De même, la somme des éléments de l'estimateur de la variance, $\hat{\sigma}^2 + \hat{\sigma}_a^2$, devient

$$\text{Var}\{Y\} = E^* + V^*. \text{ Il s'ensuit que l'estimation de la corrélation intra-intervieweur, } \hat{\rho}_I = \frac{\hat{\sigma}_a^2}{\hat{\sigma}^2 + \hat{\sigma}_a^2},$$

est en réalité une estimation de $\phi = \frac{V^*}{E^* + V^*}$ plutôt que la corrélation sous-jacente, ρ^* . Comme nous

l'avons vu plus haut, ϕ est habituellement inférieur – et de loin – à ρ^* , de sorte qu'en ajustant un modèle linéaire qui ne convient pas, on obtient effectivement une sous-estimation de la véritable corrélation intra-intervieweur. Toutefois, l'estimation résultante de l'effet du plan d'échantillonnage donne encore une estimation convergente de l'effet réel du plan d'échantillonnage, contrairement à ce que nous avançons dans notre article antérieur.

2.2 Échantillonnage à deux degrés

Supposons maintenant que nos observations sont tirées à l'aide d'un plan auto pondéré à deux degrés. Davis et Scott (1995) ont considéré un modèle des éléments additifs de la variance pour des réponses continues ayant la forme

$$Y_{ipr} = \mu + a_i + b_p + e_{ipr}, \quad (4)$$

où b_p représente le p^e effet de l'unité primaire d'échantillonnage (UPÉ). Supposons que les valeurs de b_p s présentent une moyenne nulle et une variance σ_b^2 et qu'elles ne sont pas corrélées entre elles ni avec les autres effets aléatoires. À l'aide des mêmes arguments que dans la section précédente, nous constatons que la variance de la moyenne de l'échantillon est gonflée par un facteur

$$D_1 = 1 + (\tilde{n}_I - 1)\rho_I + (\tilde{m}_C - 1)\rho_C \quad (5)$$

comparé à la variance que l'on obtiendrait avec le même nombre d'observations indépendantes. Ici, $\tilde{m}_C = \sum_p m_p^2 / n$, où m_p représente le nombre d'observations dans la p^e UPÉ, $\rho_C = \frac{\sigma_c^2}{\sigma^2 + \sigma_I^2 + \sigma_c^2}$ et ρ_I

est maintenant défini comme suit : $\rho_I = \frac{\sigma_I^2}{\sigma^2 + \sigma_I^2 + \sigma_c^2}$. Ainsi, le modèle linéaire (4) donne une belle décomposition additive de l'effet du plan d'échantillonnage, où un élément est attribuable uniquement à la variance de l'intervieweur et l'autre, uniquement au groupement.

Nous pouvons étendre ce modèle aux réponses binaires exactement de la même façon que dans la section précédente en supposant qu'il existe une variable latente sous-jacente, par exemple Y_{ipr}^* , conforme au modèle (6) et que notre variable binaire, Y_{ipr} , prend la valeur un seulement lorsque $Y_{ipr}^* \leq c$ pour une valeur limite inconnue c . Tout comme dans la section précédente, nous supposons que a_i et b_p sont des variables aléatoires normales indépendantes et que e_{ipr} présente une distribution logistique. Encore une fois, le modèle n'est défini qu'en fonction d'un paramètre à échelle arbitraire, de sorte que nous ne puissions identifier que les ratios $\frac{\sigma_a^{*2}}{\sigma^{*2}}$ et $\frac{\sigma_c^{*2}}{\sigma^{*2}}$ plutôt que les éléments individuels de la variance. Comme ρ_I^* et ρ_p^* sont des fonctions de ces ratios, nous pouvons estimer les deux coefficients de corrélation à partir des données binaires observées.

Nous nous intéressons avant tout à l'effet de la variabilité de l'intervieweur et de l'UPÉ sur la variance de l'estimateur de la proportion de la population, \hat{p} , qui est simplement la proportion de l'échantillon puisque nous supposons que nous avons un plan à autopondération. Supposons que k_{ip} représente le nombre d'interviews menées par le i^e intervieweur dans la j^e UPÉ et que $\tilde{k}_{IC} = \sum \sum k_{ip}^2 / n$. Alors, en appliquant l'expression courante de la variance en fonction de la moyenne et de la variance conditionnelles et de la variance, compte tenu des vecteurs des effets aléatoires, \underline{a} et \underline{b} , on obtient (après une certaine manipulation algébrique) :

$$\begin{aligned} \text{Var}\{\hat{p}\} &= E\{\text{Var}\{\hat{p} | \underline{a}^*, \underline{b}^*\}\} + \text{Var}\{E\{\hat{p} | \underline{a}^*, \underline{b}^*\}\} \\ &= \frac{1}{n} \{E^* + \tilde{k}_{IP} V_{12}^* + \tilde{n}_I V_1^* + \tilde{m}_P V_2^*\} \end{aligned}$$

où $E^* = E\{p(a^* + b^*)(1 - p(a^* + b^*))\}$ et $V_{12}^* = \text{Var}\{p(a^* + b^*)\} - V_1^* - V_2^*$, avec $V_1^* = \text{Var}\{p_1(a^*)\}$ et $V_2^* = \text{Var}\{p_2(b^*)\}$, où $p(a^* + b^*) = e^{\theta + a^* + b^*} / (1 + e^{\theta + a^* + b^*})$ comme dans l'expression (3) et $p_1(a^*)$ et $p_2(b^*)$ représentent les valeurs prévues de $p(a^* + b^*)$ à l'égard, respectivement, de a^* et b^* . À l'aide des mêmes arguments que dans les sections précédentes, nous constatons qu'il existe alors un effet du plan d'échantillonnage selon l'équation

$$D_1^* = 1 + (\tilde{k}_{IC} - 1)\phi_{IC} + (\tilde{n}_I - 1)\phi_I + (\tilde{m}_P - 1)\phi_C.$$

Il convient de noter que, comme $p(a^* + b^*)$ n'est pas une fonction additive de a^* et b^* , on ne peut plus décomposer simplement l'effet du plan d'échantillonnage en deux effets : celui de l'intervieweur et celui de la grappe. Il existe plutôt un terme supplémentaire qui dépend de la manière dont les intervieweurs sont déployés d'une grappe à l'autre.

Encore une fois, pour trouver des valeurs exactes, il faudrait procéder à une intégration numérique; nous nous limitons donc au cas où l'effet de l'intervieweur et celui de l'UPÉ sont faibles. En extrapolant

$p(a+b) = e^{\theta+a+b}/(1+e^{\theta+a+b})$ en fonction de $a=0$, $b=0$ comme dans la section précédente, nous constatons que $\phi_{IC} \approx 0$ à cet ordre-là et que l'effet du plan d'échantillonnage peut être exprimé comme la somme de deux éléments,

$$D_1^* \approx (1 + (\tilde{n}_I - 1)\phi_I + (\tilde{n}_C - 1)\phi_C)$$

où

$$\phi_I \approx \frac{\sigma_a^{*2}}{\frac{(1-2p_0)^2}{2}(\sigma_a^{*2} + \sigma_b^{*2}) + \frac{1}{p_0(1-p_0)}}, \text{ and } \phi_C \approx \frac{\sigma_b^{*2}}{\frac{(1-2p_0)^2}{2}(\sigma_a^{*2} + \sigma_b^{*2}) + \frac{1}{p_0(1-p_0)}}.$$

Dans l'ensemble, du moins lorsque l'effet de l'intervieweur et celui de la grappe sont faibles, les conclusions sont très semblables à celles de la section précédente. Les valeurs de $\phi_I(\phi_C)$ sont plus faibles que les valeurs correspondantes de $\rho_I^*(\rho_C^*)$, de sorte que l'effet du plan d'échantillonnage sur la variable binaire est inférieur à ce qu'il aurait été sur la variable continue sous-jacente. De plus, les estimations des corrélations intra-intervieweur et intra-grappe obtenues en ajustant le modèle des effets aléatoires linéaires (2), qui ne convient pas, sont, en réalité, des estimations des valeurs ϕ_I et ϕ_C correspondantes qui sont inférieures à ρ_I et ρ_C . Cela signifie qu'en calculant ces estimations, on obtient effectivement une estimation de l'effet précis du plan d'échantillonnage, malgré le fait que ρ_I et ρ_C soient des sous-estimations des corrélations correspondantes selon le modèle continu sous-jacent. Par ailleurs, nous avons entrepris d'étudier ce qui se passe lorsque l'effet de l'intervieweur et celui de la grappe ne sont pas assez faibles pour que l'approximation linéaire convienne à nos calculs.

3. EXEMPLE

L'étude de la qualité des soins de santé en Nouvelle-Zélande porte sur les événements indésirables (définis comme des « blessures non intentionnelles à l'origine d'une invalidité temporaire ou permanente, causées par les fournisseurs de soins de santé plutôt que par le processus morbide sous-jacent ») dans les hôpitaux néo-zélandais. (On trouvera dans Davis et coll. (2001) une description complète de l'étude et du plan d'échantillonnage.)

Au premier degré, on a prélevé un échantillon d'hôpitaux avec probabilité proportionnelle à la taille (PPT) parmi tous les hôpitaux publics de la Nouvelle-Zélande comptant au moins 300 lits. Au deuxième degré, on a prélevé un échantillon systématique de 575 dossiers médicaux parmi toutes les hospitalisations de 1998. On a ensuite vérifié chaque dossier échantillonné pour y trouver la trace d'un événement indésirable. L'étude a fait appel à treize experts en médecine, à raison d'au moins trois par hôpital. D'un hôpital à l'autre, la composition des équipes d'examineurs permettait de distinguer formellement les éléments de la variance. Toutefois, comme l'affectation des examineurs aux hôpitaux n'a pas été faite au hasard, nous devons traiter les résultats avec précaution puisqu'ils dépendent grandement du modèle. Nous sommes reconnaissants à Eliza Chan, qui a procédé à l'analyse sur SAS à l'aide des logiciels PROC MIXED pour ajuster les modèles des effets aléatoires linéaires et PROC NLMIXED pour ajuster les modèles de distribution normale logistique.

L'estimation de la principale variable qui nous intéressait était la proportion d'hospitalisations liées à un événement indésirable. La corrélation intra-examineur /était élevée, soit $\hat{\rho}_I^* = .13$, alors que la corrélation intra-hôpital était nettement inférieure, soit $\hat{\rho}_C^* = .007$. Comme nous l'avions prévu, les

valeurs ϕ correspondantes étaient nettement inférieures, soit $\phi_I = .04$ et $\phi_C = .002$. Toutefois, l'incidence de la variabilité des examinateurs, notamment sur la variance de la proportion de l'échantillon, est encore importante, à cause de la charge de travail moyenne très élevée. Nous avons $\tilde{n} = 687$, ce qui donne un effet global du plan d'échantillonnage $D^* = 29,0$, dont la presque totalité est attribuable à la variabilité des examinateurs.

Si nous ajustons un modèle courant des effets aléatoires linéaires, nous obtenons les estimations des coefficients de corrélation $\hat{\rho}_I = .03$ et $\hat{\rho}_C = .0045$, qui sont de loin inférieurs à $\hat{\rho}_I^*$ et $\hat{\rho}_C^*$ mais raisonnablement proches, respectivement, de $\phi_I = .04$ et $\phi_C = .002$. L'estimation de l'effet du plan d'échantillonnage est $D = 21,4$, ce qui est légèrement inférieur à la valeur de $D^* = 29,0$ obtenue à partir du modèle binaire. Tous ces résultats sont raisonnablement cohérents avec la théorie énoncée à la section 2.2.

Au début, les hôpitaux participants s'inquiétaient de la publicité qui risquait d'entourer la variation des taux parmi les hôpitaux; pourtant, tous les hôpitaux échantillonnés ont accepté de participer à l'étude, ce qui est tout à leur honneur. Nous constatons, d'après les deux démarches décrites plus haut, que les différences entre les hôpitaux sont vraiment infimes. Si nous avons utilisé l'estimation usuelle de la variance pour une enquête basée sur un échantillon à deux degrés, en faisant complètement abstraction des examinateurs, nous aurions obtenu une estimation de 0,01 pour le coefficient de corrélation intra-hôpital, ρ_C , et une estimation de l'effet du plan d'échantillonnage de $D = 6,4$. La variabilité manquante des examinateurs se répercute en partie sur le terme inter-hôpital, gonflant l'estimation de ρ_C par un facteur de plus de deux, quoique l'estimation résultante est encore relativement faible dans le cas qui nous occupe. Ce phénomène accroît dans une certaine mesure l'estimation de l'effet du plan d'échantillonnage, mais il s'agit encore d'une sous-estimation très considérable de la véritable valeur.

4. CONCLUSIONS

Nous avons montré que, même si la variabilité de l'intervieweur a un effet plus faible sur la variance des proportions de l'échantillon que sur les moyennes des variables continues, elle peut encore avoir une incidence importante sur l'erreur totale lorsque la charge de travail moyenne est élevée. Nous montrons également que l'ajustement d'un modèle ordinaire d'éléments linéaires de la variance donne une estimation raisonnable de l'effet du plan d'échantillonnage lorsque les effets de l'intervieweur ou de la grappe sont faibles et possiblement de manière plus générale, quoique ce point nécessite une recherche approfondie.

Lorsqu'on analyse les grandes enquêtes sur la santé, on fait habituellement abstraction de l'effet de la variabilité de l'intervieweur. (Nous n'avons d'ailleurs trouvé aucun rapport d'enquête de ce genre dans lequel on tient compte des effets de l'intervieweur.) Lorsqu'on fait abstraction des effets de l'intervieweur, on gonfle les estimations de la corrélation intra-grappe et (sauf dans le cas extrême où les intervieweurs sont complètement emboîtés dans les grappes) on sous-estime l'effet réel du plan d'échantillonnage. Comme il est très coûteux de former de nouveaux intervieweurs dans ces cas particuliers, il nous faut prélever un échantillon de plus petite taille. Or, n'importe quel chercheur du domaine de la santé est conscient de l'effet d'un échantillon de plus petite taille sur l'exactitude des estimations relatives à l'échantillon. Toutefois, comme il est rare que les chercheurs tiennent compte explicitement de l'effet de l'intervieweur dans leurs estimations des erreurs-types, la plupart d'entre eux ne sont toujours pas conscients des véritables conséquences de la décision de former moins d'intervieweurs et de faire plus d'observations. À cet égard, un vaste effort d'information semble s'imposer.

BIBLIOGRAPHIE

- Anderson, D.A. et Aitken, M. (1985). "Variance component models with binary response; interviewer variability", *Journal of the Royal Statistical Society*, Ser B, 47, pp. 203-210.
- Davis, P.D. Lay-Yee, R., Briant, R., Schug, S. et Scott, A.J. (2001), *Adverse Events in New Zealand Public Hospitals*. Wellington: NZ Ministry of Health.
- Davis, P.D. et Scott, A.J. (1995), "La variance de l'interviewer et ses effets sur les comparaisons de domaines", *Techniques d'enquetes*, 21, pp. 111-118.
- De Leeuw, J. et Kreft, I.G.G. (2001), "Software for Multilevel Modelling", dans A.H. Leyland et H. Goldstein (eds.) *Multilevel Modelling of Health Statistics*. New York: Wiley, pp. 206-223.
- Groves, R. (1989). *Survey Errors et Survey Costs*. New York: Wiley.
- Hartley, H.O. et Rao, J.N.K. (1968), "Estimation of nonsampling variance components in sample surveys", dans N.K. Namboodiri, (éds), *Survey Sampling and Measurement*, New York: Academic Press, pp. 35-44.
- Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, pp. 92-115.
- Kish, L. et Frankel. M.R. (1974), "Inferences from complex samples", *Journal of the Royal Statistical Society*, B, 36, pp. 1-37.
- Lessler, J.T. et Kalsbeek, W.D. (1992). *Nonsampling Errors in Surveys*. New York: Wiley
- Pannekoek, J. (1988), "Interviewer variance in a telephone survey", *Journal of Official Statistics*, 4, pp. 375-384.