LA FORMATION DE POOLS DE SÉRUM DE SANG POUR L'ANALYSE DANS L'ENQUÊTE CANADIENNE SUR LES MESURES DE LA SANTÉ

François Verret et Suzelle Giroux[1]

RÉSUMÉ

L'Enquête canadienne sur les mesures de la santé recueille des renseignements sous forme de mesures physiques directes auprès d'un échantillon de Canadiens qui permettent d'obtenir des indicateurs de la santé. Certains tests à effectuer sur les prélèvements recueillis sont coûteux et nécessitent une quantité minimum de spécimen pour détecter les composés à analyser. À la première réalisation de cette enquête biennale, soit au cycle 1, on a formé des pools à partir de sérum de sang résiduel pour mesurer les niveaux de composés organiques halogénés dans dix groupes d'âge-sexe à l'échelle nationale. Cet article décrit la méthodologie utilisée pour la création des pools et l'estimation. Ceux-ci ont été créés en combinant le sérum de répondants ayant des poids de sondage similaires pour minimiser le biais. Des groupes aléatoires dépendants d'unités primaires d'échantillonnage ont été formés à travers les strates pour estimer la variance.

MOTS CLÉS : Groupes aléatoires dépendants; inférence basée sur le plan de sondage; pooling de spécimens de sang.

ABSTRACT

The Canadian Health Measures Survey collects directly physical measures from a sample of Canadians to produce indicators of health. Some tests on collected specimens are costly and require a minimum amount of specimen to detect the components analyzed. In cycle 1 of the biennial survey, residual blood serum was pooled to estimate average levels of organohalogens in ten age-sex groups at the national level. This article describes the pooling and estimation methodologies. Pools were created by grouping serum of individuals having similar sampling weights to minimise the bias of the estimates and dependent random groups of primary sampling units were formed across the strata in order to estimate the variance.

KEY WORDS: Blood specimens pooling, Dependent random groups, Design-based inference.

1. INTRODUCTION

Il existe un besoin pour des données sur les niveaux de contaminants tels que les éthers diphényliques polybromés (EDP), les dioxines, les furanes et les biphényles polychlorés (BPC) contenus dans le sang des Canadiens. Au cycle 1 de l'Enquête sur les mesures de la santé (ECMS), on a mesuré à partir de spécimens sanguins individuels provenant d'une partie de l'échantillon de l'enquête les niveaux d'une quarantaine d'EDP et de BPC. L'ECMS recueille de l'information sur la santé des Canadiens à partir de mesures directes auprès d'un échantillon représentatif de la population canadienne âgé de 6 à 79 ans, composé de 5 604 personnes et choisi à l'aide d'un plan de sondage complexe. Ce plan comprend trois degrés d'échantillonnage qui comportent chacun une stratification (Giroux, 2007). Plus de détails sur le plan de sondage de l'ECMS sont donnés à la section 2.2. Parmi l'échantillon, 1 696 adultes âgés entre 20 et 79 ans faisaient partie du sous-échantillon destiné à mesurer les niveaux de contaminants. Plusieurs EDP, dioxines et furanes n'ont pu être mesurés à cette occasion en raison des volumes de spécimens nécessaires et du coût élevé relié aux mesures de ces contaminants. De plus, certains des contaminants étudiés avaient une concentration sous les seuils de détection pour un grand nombre de spécimens individuels.

Pour combler les besoins en information que le sous-échantillon ne peut satisfaire, Haines et coll. (2009) ont amorcé une étude consistant à mesurer environ 80 contaminants en combinant en pools le sérum résiduel des répondants à l'ECMS. Le pooling consiste à grouper et à mélanger les spécimens de plusieurs répondants pour former des pools et à effectuer

[1] François Verret et Suzelle Giroux, Statistique Canada, 100, promenade du pré Tunney, Ottawa, Ontario, Canada, K1A 0T6, Francois.Verret@statcan.gc.ca, Suzelle.Giroux@statcan.gc.ca.

les mesures sur ces derniers. Une telle analyse a l'avantage par rapport à une analyse par mesures individuelles d'avoir potentiellement une plus grande proportion de mesures dépassant la barre du niveau détectable car elle est faite sur de plus grands volumes de spécimen (Caudill et coll. 2007, Needham et coll. 2007, Caudill 2008). D'autre part, étant donné que le coût des mesures en laboratoire est élevé, le pooling réduit considérablement le nombre de mesures à effectuer et donc les coûts d'analyse. Finalement, étant donné le faible nombre de mesures à prendre, il a été jugé raisonnable dans cette étude d'avoir recours à des analyses plus coûteuses mais plus sensibles et spécifiques que lors de l'analyse par mesures individuelles dans le but de faire diminuer les seuils de détectabilité.

Cet article traite plus spécifiquement de la méthodologie entourant la formation des pools pour l'étude de Haines et coll. (2009) et de l'estimation s'y rattachant. L'objectif principal à atteindre dans la construction de cette méthodologie était d'estimer la concentration de contaminants par quantité de sang pour les dix groupes d'âge et de sexe cibles de l'ECMS au niveau national. Un objectif secondaire, qu'il fallait tenter d'atteindre, était de développer une méthode d'estimation de la variance qui donnerait au moins un ordre de grandeur de la précision des estimateurs de concentration. La section 2 se penche sur l'estimation fondée sur le plan de sondage dans les études de pooling et discute du plan de sondage de l'ECMS. La section 3 présente la stratégie de création des pools et la méthode d'estimation ponctuelle s'y rattachant. Suit la section 4 qui donne la méthode d'estimation de la variance adoptée et les raffinements à la stratégie de création des pools qu'il est nécessaire de faire pour l'appliquer. Une conclusion termine l'article et présente quelques recommandations relativement à l'analyse des données. On y discute également de la méthodologie à adopter dans des études semblables où l'on tient compte de l'interaction entre le plan de sondage et le pooling.

2. PLANS DE SONDAGE ET POOLING

2.1 Importance du plan de sondage dans le pooling

Plusieurs études de pooling ont été menées ces dernières années à travers le monde. Parmi les plus récentes, on en dénombre aux États-Unis dans la National Health and Nutrition Examination Survey (NHANES) (Patterson et coll., 2008); en Australie dans le National Dioxins Program (Harden et coll., 2004); en Nouvelle-Zélande (Bates et coll., 2004) et dans la province canadienne de l'Alberta auprès de femmes enceintes dans l'Alberta Biomonitoring Program (2008). Dans ces exemples, les estimations ne sont pas basées purement sur le plan de sondage. Un tel genre d'estimation dans des études de pooling est un domaine de recherche relativement nouveau et inexploré. Il se pourrait qu'on puisse ignorer le plan de sondage dans l'analyse de pools, et qu'une étude basée purement sur des modèles soit appropriée. Cependant, a priori il faudrait justifier une telle analyse, comme on le fait en général pour d'autres études utilisant les données de l'enquête à plan complexe considérée.

Voici quelques conditions théoriques permettant une telle justification dans le cas d'une variable *y* fixe pour chaque individu. Soient $U=\{1,2,...,k,...,N\}$ la population finie d'intérêt et *s* l'échantillon de taille fixe *n* qui en est tiré avec probabilités de sélection $\pi_k>0,\ k=1,2,...,N$. Si on s'intéresse à estimer une moyenne, un estimateur simple ignorant le plan de sondage est la moyenne échantillonnale $\bar{y}=\sum_s y_k/n$. Cet estimateur estime la moyenne de la population finie $\bar{Y}$ sans biais sous le plan de sondage si ce dernier est autopondéré ou si la relation suivante existe : $\frac{1}{n}\sum_U \pi_k y_k=\frac{1}{N}\sum_U y_k=\frac{t_y}{N}=\bar{Y}$. Il estime sans biais la moyenne μ sous un modèle de superpopulation et sous le plan si la probabilité de sélection π_k est indépendante de la variable d'intérêt y_k sous le modèle. Dans le cas de l'étude par pools de l'ECMS, le plan n'est définitivement pas autopondéré et les deux autres conditions pour l'absence de biais sont difficiles à respecter et à vérifier. C'est particulièrement difficile, et probablement impossible à vérifier, dans une étude de pooling où les mesures individuelles ne sont pas disponibles par définition. La problématique est similaire pour l'estimation de la variance. En effet, l'estimateur de la variance de l'estimateur $\bar{y}$ utilisant les mesures individuelles et ignorant le plan de sondage est $\frac{1}{n(n-1)}\sum_s (y_k-\bar{y})^2=\frac{s^2}{n}$. Ce dernier est un bon estimateur de la variance sous le plan en présence d'un plan d'échantillonnage aléatoire simple avec remise ou sans remise avec faible fraction de sondage, mais ce n'est pas un bon estimateur dans les autres plans autopondérés ayant un effet de plan très différent de l'unité.

Finalement, il s'agit d'un bon estimateur de la variance sous le modèle et sous le plan si l'effet de plan est près de un et si la probabilité de sélection et la variable d'intérêt sont indépendantes sous le modèle. Dans le cas de l'ECMS, le plan n'est pas autopondéré et les effets de plan sont en général supérieurs à un.

En comparaison, l'utilisation d'estimateurs basés sur le plan de sondage est plus facile à justifier. Un estimateur sans biais de la moyenne sous le plan $\overline{Y}$ est la moyenne pondérée $\overline{y}_w = \sum_s w_k y_k / N = \hat{t}_{y\pi} / N$, où w_k est le poids de sondage et

$$\hat{t}_{y\pi} = \sum_s w_k y_k \tag{1}$$

est l'estimateur d'Horvitz-Thompson. Cet estimateur est également sans biais sous le modèle et sous le plan pour la moyenne de la superpopulation μ. D'autre part, un bon estimateur de la variance sous le plan de cet estimateur ponctuel sera aussi un bon estimateur de sa variance sous le modèle et sous le plan si la taille d'échantillon est grande, et si la fraction de sondage dans l'ensemble $f = n/N$ est petite (Binder et Roberts, 2009). Ces deux conditions sont satisfaites dans l'ECMS et ce même quand on répartit l'échantillon parmi les dix groupes cibles.

Pour se convaincre davantage qu'il ne faut pas ignorer le plan de sondage lorsqu'on étudie les variables d'intérêt, on peut utiliser les données récoltées dans le cadre du sous-échantillon de mesures individuelles de contaminants. Ce sous-échantillon de 1 696 adultes représente six des dix groupes d'âge et de sexe ciblés par l'étude par pooling, soient les groupes d'âge 20 à 39 ans, 40 à 59 ans et 60 à 79 ans croisés avec le sexe. Les personnes qui composent ce sous-échantillon étaient à jeun lors de la récolte des spécimens. Les trois contaminants ayant le moins de mesures individuelles sous les niveaux de détection sont étudiés ici : EDP 47, BPC 118 et BPC 156 (les numéros suivant les acronymes étant ceux de la classification élaborée par l'Union internationale de chimie pure et appliquée). Pour les besoins de cette étude, lorsque les niveaux des mesures individuelles étaient en dessous du seuil de détection, soient 0,03 ; 0,01 et 0,01 µg/L respectivement, le niveau était imputé à la moitié de ce seuil. La statistique étudiée est l'effet de plan, soit le ratio de la variance de $\overline{y}_w$ sous le plan et la variance du même estimateur sous un échantillonnage hypothétique aléatoire simple. Si ce ratio n'est pas près de l'unité, on ne pourra ignorer le plan de sondage dans l'estimation de la variance de l'estimateur de moyenne. La première variance est estimée en appliquant le bootstrap de Rao, Wu et Yue (1992) à l'estimateur de moyenne pondéré pour tenir compte du plan stratifié à plusieurs degrés de l'ECMS. La seconde variance est estimée en utilisant la formule mentionnée dans Rust et Rao (1996) et Gambino (2009) :

$$\hat{V}_{\text{SRS}} = \left(1 - \frac{n}{N}\right) \frac{1}{n(N-1)} \left[\sum_s w_k y_k^2 - \frac{\hat{t}_{\pi y}{}^2}{N} \right]. \tag{2}$$

Les effets de plan estimés sont donnés dans le tableau 1. On y remarque que la majorité des effets de plan sont plus grands que un, ce qui indique qu'on devrait tenir compte du plan de sondage dans l'étude de pooling. Cependant, comme on le verra à la prochaine section, qui décrit le plan de sondage de l'ECMS, le numérateur est une estimation qui pourrait être instable étant donné le faible nombre d'unités primaires d'échantillonnage (UPE) du plan. De plus, il pourrait être biaisé pour plusieurs raisons telles que la probabilité de sélection au premier degré qui peut être grande pour certaines UPE. En l'absence de données de meilleure qualité, par prudence et parce que tous les effets de plan pour tous les contaminants ne peuvent être étudiés, la méthode de pooling développée et présentée dans cet article repose sur une approche fondée sur le plan de sondage.

Tableau 1 – Effets de plan estimés pour la moyenne de concentration de trois contaminants

Contaminant	Groupe d'âge et de sexe					
	Hommes 20-39	Femmes 20-39	Hommes 40-59	Femmes 40-59	Hommes 60-79	Femmes 60-79
EDP 47	1,39	1,67	0,84	1,55	1,64	0,98
BPC 118	4,53	4,57	1,33	2,52	3,81	3,05
BPC 156	2,01	1,94	4,15	1,65	2,63	1,65

2.2 Plan de sondage de l'ECMS et estimation fondée sur ce plan à l'aide de microdonnées

L'ECMS fait appel à un plan de sondage comprenant trois degrés comportant tous une stratification. Tout d'abord, le Canada est divisé en cinq strates selon les régions géographiques : les provinces de l'Atlantique, le Québec, l'Ontario, les Prairies et Yellowknife, et la Colombie-Britannique et la ville de Whitehorse. À l'intérieur de chacune des régions, des sites de collecte sont sélectionnés. Un site de collecte est une région géographique comptant au moins 10 000 habitants et définie telle que la distance à parcourir par les participants de l'enquête n'excède pas une certaine distance puisque les répondants auront à se rendre à une clinique. On compte 257 sites de collecte au Canada, et 15 d'entre eux ont été échantillonnés systématiquement avec une probabilité de sélection proportionnelle à la taille de leur population. Le nombre de sites choisis varie de 1 à 6 selon la région géographique. Au second degré, des logements sont choisis au hasard parmi les sites de collecte retenus. Finalement, on choisit parmi les logements échantillonnés, une ou deux personnes par ménage.

Au premier degré, la probabilité de sélection des sites de collecte (ou UPE) peut aller jusqu'à 30 %. En général, on souhaite que la probabilité de sélection au premier degré soit faible afin de pouvoir ignorer les degrés subséquents du plan dans l'estimation de la variance (voir Särndal, Swensson et Wretman, 1992, p.140). D'ailleurs, en raison du coût élevé associé à l'échantillonnage d'un site, le nombre d'UPE sélectionnées est faible : 1, 4, 6, 2 et 2 respectivement dans chacune des strates. Afin de rentabiliser la sélection du faible nombre de sites et pour minimiser la variance des estimateurs, une stratification implicite des UPE est faite à l'intérieur des strates. Dans chaque strate, les sites sont triés selon un indicateur de région métropolitaine puis par taille de population du site. Un échantillon systématique proportionnel à cette taille est ensuite tiré. Ceci fait en sorte que l'échantillon final contient une bonne répartition de sites métropolitains et non métropolitains. Par exemple, dans la cinquième strate la moitié de la population réside dans des sites métropolitains de la ville de Vancouver. La stratification implicite fait en sorte que l'échantillon final contient assurément un site de Vancouver et un site hors de Vancouver. Si les deux types de sites présentent des caractéristiques différentes, ceci amène une certaine stabilité au niveau de l'estimation ponctuelle, mais en contrepartie la stratification implicite complique l'estimation de la variance.

Pour l'estimation, des poids d'enquête sont produits pour tenir compte des trois degrés d'échantillonnage et de la stratification. Ces poids sont également corrigés pour tenir compte des non-réponses et poststratifiés selon des projections démographiques. La méthode du bootstrap de Rao-Wu-Yue (1992), où l'on suppose une petite probabilité de sélection au premier degré, est utilisée pour estimer la variance due au plan à l'aide de poids de répliques. Finalement, pour compenser l'unique UPE pour la région de l'Atlantique (un seul site de collecte ayant été sélectionné), les strates du Québec et de l'Atlantique sont combinées en une seule aux fins de l'estimation de la variance.

3. STRATÉGIE DE POOLING ET D'ESTIMATION PONCTUELLE

3.1 Nombre de pools possibles

La première étape dans l'élaboration d'une stratégie de pooling consiste à déterminer combien de pools peuvent être formés à partir des spécimens de sang échantillonnés. Afin de simplifier les manipulations en laboratoire et pour minimiser les erreurs possibles, la même quantité de spécimen est utilisée dans la formation des pools pour chacun des répondants. Une telle restriction vient limiter la quantité de spécimen disponible pour les pools. Sous cette contrainte, on a initialement déterminé que 0,35 ml de sang était disponible par répondant. D'autre part, 25 ml de sang par pool est nécessaire pour effectuer les mesures de concentration de contaminants. Ce qui implique que les pools devraient être composés du sang d'au minimum 71 individus. On a choisi de combiner le sang de 80 répondants pour compenser toutes

les formes de non-réponse possibles. En effet, les répondants de l'ECMS peuvent refuser que leur sang soit utilisé pour l'étude de pooling. De plus, le volume de spécimen peut être sous la limite de 0,35 ml. Finalement, il pourrait y avoir des accidents lors des manipulations et on pourrait perdre quelques spécimens. Ces trois types de non-réponse ont été observés dans l'étude. Sous ces contraintes, 59 pools ont été formés. Le tableau 2 donne le nombre de répondants participant à l'étude et le nombre de pools formés pour chaque groupe cible de l'enquête.

Tableau 2 – Nombre de répondants participant à l'étude sur le pooling et nombre de pools formés par groupe d'âge et de sexe cible

	Groupe d'âge et de sexe									
	6-11		12-19		20-39		40-59		60-79	
	M	F	M	F	M	F	M	F	M	F
Nombre de répondants	444	436	457	436	486	600	547	611	522	520
Nombre de pools formés	5	5	5	5	6	7	7	7	6	6

L'objectif est de produire des estimations pour chacun des dix groupes d'âge et de sexe ciblés. La même méthodologie est appliquée de façon indépendante pour chacun. Les prochaines sections la décrivent pour un groupe cible donné.

3.2 Formation des pools pour minimiser le biais des moyennes estimées

Dans l'analyse fondée sur le plan de sondage, l'estimateur d'Horvitz-Thompson est généralement l'estimateur de référence à cause de sa simplicité et du fait qu'il est sans biais. La stratégie de formation des pools développée tente de produire un estimateur aussi proche que possible de cet estimateur, et un estimateur de variance proche d'un estimateur de variance qui serait valable.

Les mesures individuelles de concentration y_k ne sont pas disponibles dans une étude par pooling. Cependant, on peut reproduire l'estimateur de Horvitz-Thompson lors de la formation des pools en permettant le mélange d'un volume de sang différent pour chaque répondant. En effet, c'est le cas lorsque le volume de sang mélangé est proportionnel au poids de sondage. Soient G le nombre de total de pools pour le groupe cible considéré, s_g l'échantillon composant un pool donné g, v_k la quantité de sang mélangé pour l'individu k, x_k la masse de contaminant dans ce liquide et $y_k = x_k / v_k$ qui est mesurée en unités de masse par unité de volume. Si le volume est proportionnel au poids de sondage $\left(\text{c'est-à-dire } v_k = 25 w_k \Big/ \sum_{s_g} w_k\right)$, alors on a l'identité suivante :

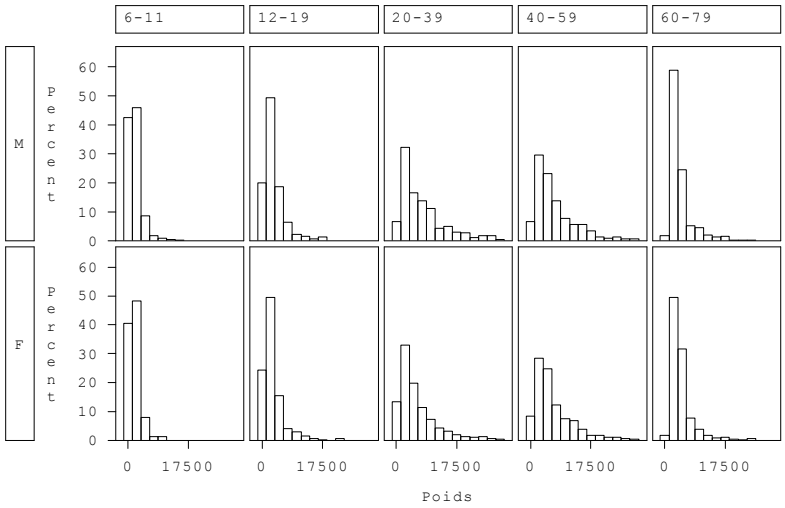
$$\tilde{t}_{y\pi} = \sum_{g=1}^{G}\left(\sum_{s_g} w_k\right) y_g = \sum_{g=1}^{G}\left(\sum_{s_g} w_k\right)\left(\sum_{s_g} x_k/25\right) = \sum_{g=1}^{G}\left(\sum_{s_g} w_k\right)\left(\sum_{s_g} v_k y_k/25\right) = \sum_{g=1}^{G}\left(\sum_{s_g} w_k y_k\right) = \hat{t}_{y\pi}\,.$$

Néanmoins, on ne peut utiliser cette approche sous la restriction d'un volume de sang constant à mélanger par répondant énoncée à la section 3.1. Sous cette contrainte, le même estimateur devient:

$$\tilde{t}_{y\pi} = \sum_{g=1}^{G}\left(\sum_{s_g} w_k\right) y_g = \sum_{g=1}^{G}\left(\sum_{s_g} w_k\right)\left(\sum_{s_g} y_k/n_g\right) = \sum_{g=1}^{G}\overline{w}_g\left(\sum_{s_g} y_k\right) = \sum_{g=1}^{G}\left(\sum_{s_g} \overline{w}_g y_k\right). \qquad (3)$$

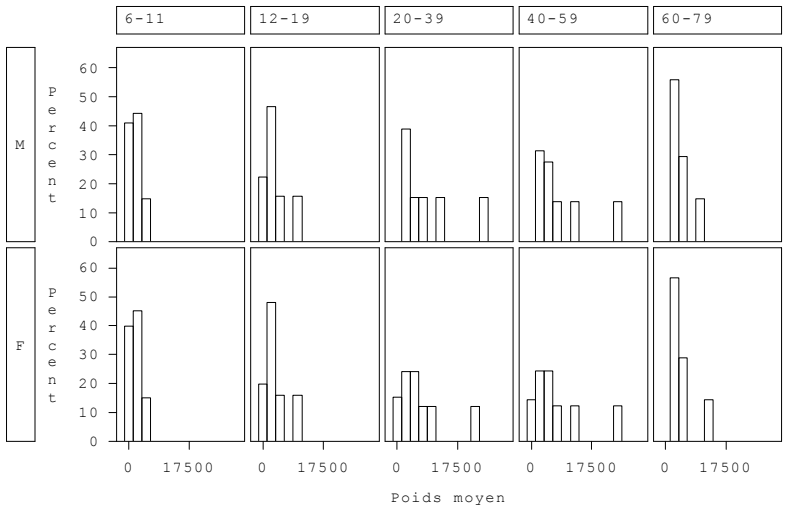
L'estimateur d'Horvitz-Thompson et l'estimateur ignorant le plan de sondage $\left(\sum_s w_k\right)\overline{y}$ sont des cas particuliers de l'estimateur (3) correspondants respectivement à $G = n$ pools et à $G = 1$ pool. Pour avoir la meilleure approximation possible de l'estimateur d'Horvitz-Thompson avec un nombre de pools G fixe, on doit combiner le sang des répondants ayant des poids de sondage similaires dans l'estimateur (3).

Figure 1 – Histogrammes du poids de sondage des dix groupes d'âge et de sexe ciblés



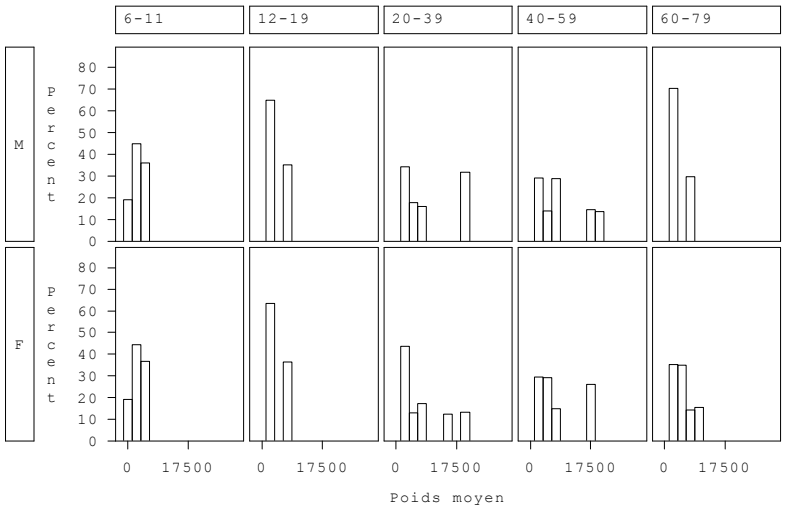
La figure 1 donne les histogrammes du poids de sondage pour chacun des dix groupes cibles. On y remarque que la distribution des poids est asymétrique à droite. De plus, le nombre de répondants par groupe n'est jamais un multiple de 80, ce qui fait en sorte qu'on pourrait choisir d'avoir certains pools plus nombreux que les autres. Pour minimiser la modification aux grands poids de sondage, on a donc avantage à former un pool plus nombreux à partir du sang des individus avec les poids de sondage les moins élevés. On peut se représenter ceci graphiquement en se disant que les pools sont créés en formant des groupes de 80 à partir de la droite des graphiques en allant vers la gauche, avec un groupe résiduel plus nombreux contenant les individus les plus à gauche. La figure 2 permet de visualiser la modification au poids de sondage résultante et présente les histogrammes du poids moyen des répondants équivalent à la formation des pools. Ces derniers sont aux nombres de 5 à 7 selon le groupe cible.

Figure 2 – Histogrammes du poids de sondage moyen des dix groupes d'âge et de sexe ciblés (5 à 7 pools par groupe)



On obtient des histogrammes assez similaires à ceux de la figure 1 avec en général une barre isolée à la droite représentant l'ensemble des poids les plus élevés. Cependant, il sera nécessaire d'approximer davantage le poids, comme il le sera présenté à la prochaine section, pour les besoins d'estimation de la variance. En effet, une approche par réplique doit être appliquée, ce qui vient nuire à l'approximation du poids. La figure 3 donne les histogrammes du poids de sondage moyen des répondants correspondant aux pools finaux. On y remarque que l'approximation du poids est plus grossière, mais qu'on est toujours éloigné du cas extrême ignorant le plan de sondage, ce qui correspondrait à n'avoir qu'une seule barre avec toute la masse par histogramme.

Figure 3 – Histogrammes du poids de sondage moyen des dix groupes d'âge et de sexe ciblés (2 répliques disjointes basées sur le plan de 2 à 4 pools)



4. STRATÉGIE DE POOLING ET D'ESTIMATION DE LA VARIANCE

La section précédente présentait une méthode pour combiner les spécimens en pools de manière à bien approximer l'estimateur d'Horvitz-Thompson sous la restriction d'avoir 0,35 ml de spécimen pour chaque répondant. La présente section aborde une stratégie de pooling permettant l'estimation de la variance due au plan utilisant une approche par répliques autre que le bootstrap de Rao-Wu-Yue (1992). La sous-section 4.1 présente cette stratégie dans le cas d'un plan de sondage ne nécessitant pas le recours au regroupement de strates. Ceci correspondrait au plan de sondage de l'ECMS si on excluait la strate de l'Atlantique. La sous-section 4.2 étend la stratégie aux plans avec regroupement de strates, soit le plan actuel de l'ECMS.

4.1 Stratégie de pooling et d'estimation de la variance sans regroupement de strates

Pour simplifier cette section, on présente ici une méthode pour estimer la variance de l'estimateur d'Horvitz-Thompson (1). Un estimateur de variance pour l'estimateur (3) n'a pas été développé par les auteurs, mais l'estimateur étudié dans cet article devrait être bon dans la mesure où l'estimateur (3) approxime bien l'estimateur (1). Lorsque les microdonnées sont disponibles, on utilise le bootstrap de Rao-Wu-Yue (1992) pour estimer la variance dans l'ECMS, et on le fait en sélectionnant des grappes au hasard dans les strates pour former les répliques. Il serait donc tentant d'estimer la concentration par grappe à partir de pools et d'appliquer le même bootstrap. Cependant, les pools sont trop peu nombreux pour le faire. De plus, même si 15 pools avaient pu être formés par groupe cible, la méthode d'approximation du poids de la section 3 nécessiterait plusieurs pools par grappe pour avoir une bonne estimation de la concentration du contaminant dans la grappe. On pourrait aussi envisager d'estimer la variance qu'on aurait obtenue sous un plan aléatoire simple sans remise en utilisant la formule (2) et de multiplier par un effet de plan estimé. Ceci n'est pas facile à faire quand les mesures individuelles ne sont pas disponibles. En effet, ça nécessiterait d'estimer à l'aide de pools une somme pondérée de carrés $\sum_s w_k y_k^2$, ce qui est bien plus complexe que d'estimer à l'aide de pools la somme $\sum_s w_k y_k$. De plus, même si les poids étaient égaux, la somme $\sum_s y_k^2$ serait difficile à estimer à l'aide de pools car former un pool est équivalent à sommer les quantités y_k avant d'avoir pu les mettre au carré. Finalement, il faudrait avoir une bonne idée de l'effet de plan à utiliser sous cette approche.

À la place, la théorie des groupes aléatoires dépendants (GAD) est utilisée pour estimer la variance de l'estimateur ponctuel (voir Särndal, Swensson et Wretman, 1992, p.426-430). Cette méthode consiste à diviser l'échantillon *s* en *A* parties disjointes : $s = \bigcup_{a=1}^{A} s^{(a)}$, de manière à ce que chaque partition $s^{(a)}$ ait un plan de sondage similaire à *s*. Pour

chacune de ces partitions, on calcule l'estimateur d'intérêt $\hat{\theta}^{(a)}$, dans ce cas-ci l'estimateur $\hat{t}_{y\pi}^{(a)}$. L'estimateur ponctuel $\hat{\theta}_{\text{GAD}}$ est la moyenne des A estimateurs. Dans le cas particulier de l'estimateur $\hat{t}_{y\pi}^{(a)}$, $\hat{\theta}_{\text{GAD}}$ correspond alors à l'estimateur sans biais (1). L'estimateur de la variance par répliques est donné par $\hat{V}\left(\hat{\theta}_{\text{GAD}}\right)=\frac{1}{A\left(A-1\right)}\sum_{a=1}^{A}\left(\hat{\theta}^{(a)}-\hat{\theta}_{\text{GAD}}\right)^2$. Comme les répliques sont disjointes, un avantage de cette approche par rapport à bien d'autres méthodes d'estimation de la variance est que l'information de chaque individu n'est utilisée qu'une seule fois dans la somme de carrés, tout comme le spécimen d'un individu ne fait partie que d'un seul pool. De plus, un GAD peut couvrir plusieurs strates et grappes à la fois, ce qui signifie qu'un plus faible nombre de pools est nécessaire pour son application comparativement aux autres approches par répliques comme le bootstrap de Rao-Wu-Yue (1992).

En raison du faible nombre de pools qu'il est possible de former par groupe d'âge et de sexe cible, la stratégie finale adoptée consiste à créer le nombre minimal de GAD possibles, soit $A = 2$. Effectivement, comme l'estimateur final n'est pas (1), mais (3), on veut recourir à la mise à la moyenne des poids indépendamment à l'intérieur de chacune des répliques GAD. Il est possible de former entre cinq et sept pools par groupe d'âge et de sexe cible, ce qui correspond à entre deux et quatre pools par groupe cible par réplique GAD. Selon la définition des strates de l'ECMS, d'est en ouest au Canada, le nombre d'UPE par strate est de un, quatre, six, deux et deux respectivement. Pour imiter le plan de sondage de *s*, les répliques GAD sont formées en leur assignant au hasard et sans remise des UPE au travers des strates de manière à être représentative de toutes les strates. Si on exclut la région de l'Atlantique, une réplique est formée de la moitié des UPE de chaque strate, soit deux UPE du Québec, trois UPE de l'Ontario, une UPE des Prairies et une UPE de la Colombie-Britannique. L'autre réplique est formée de l'autre moitié des UPE.

Dans le développement de l'estimateur de variance, le but premier était d'obtenir une estimation proche de celle que l'on obtiendrait avec la méthode d'estimation de la variance de l'ECMS si on avait disposé des données individuelles, ce qui correspond à faire le bootstrap de Rao-Wu-Yue (1992). Pour ce faire, on a assigné les UPE dans les strates en tirant un échantillon aléatoire simple sans remise de la moitié des UPE dans chacune des strates. Il faut cependant noter que le bootstrap de Rao-Wu-Yue (1992) dans l'ECMS ignore la stratification implicite des UPE qui a été faite à l'intérieur des strates. Un autre objectif que l'on pourrait ainsi chercher à atteindre serait d'estimer la variance en tentant de tenir compte de cette stratification implicite. On peut alors tirer un échantillon systématique de sites dans chaque strate en respectant l'ordre de tri de la sélection initiale. Il faut noter qu'il y a une différence entre les deux approches seulement dans les strates du Québec et de l'Ontario car ce sont celles qui contiennent plus de deux UPE. Bien que la stratégie de pooling adoptée visait à reproduire le bootstrap de Rao-Wu-Yue (1992) ignorant la stratification implicite, par un heureux hasard, les répliques résultantes sont très proches de répliques qu'on aurait obtenues par sous-échantillonnage systématique de sites dans les strates.

Un désavantage de l'estimateur de variance GAD est qu'il est biaisé. Sous un plan stratifié à plusieurs degrés, son biais est donné par :

$$B\left(\hat{V}_{\text{GAD}}\right)=\sum_{h=1}^{H}\frac{n_{hI}}{n_{hI}-1}\left[V_{\substack{\text{Hyp: Avec remise}\\ \text{1er degré}}}\left(\hat{t}_{hy\pi}\right)-V_{\text{1er degré}}\left(\hat{t}_{hy\pi}\right)\right], \tag{4}$$

où *h* représente la strate, n_{hI} est le nombre d'UPE sélectionnées dans la strate *h*, la première variance à droite de l'égalité est la variance due au premier degré de l'estimateur de Horvitz-Thompson dans la strate si ce degré avait été avec remise, et la deuxième est la véritable variance due au premier degré du plan (voir Särndal, Swensson et Wretman, 1992, p. 429-430). Le biais est fonction de la différence entre ces deux variances. La seconde variance devrait être plus petite que la première puisque la stratification implicite devrait produire un plan plus efficace au premier degré qu'un simple tirage avec remise. Le biais devrait donc être positif et par conséquent conservateur, ce qui n'est pas une mauvaise chose en pratique. De plus, le biais n'est pas fonction du faible nombre de répliques *A*. Finalement, il est une fonction croissante du nombre d'UPE sélectionnées dans les strates. Ces nombres sont faibles dans toutes les strates, ce qui peut gonfler le biais de façon considérable si la différence entre les variances est grande.

Bien que le petit nombre de répliques ne vienne pas biaiser davantage l'estimateur de variance, il vient par contre pénaliser l'estimateur au niveau de sa stabilité. Le lien entre la variance d'un estimateur de variance par groupe aléatoire indépendant et le nombre de répliques est discuté dans Wolter (2007, p.57-64). Cette variance diminue quand le nombre de répliques augmente et quand le quatrième moment centré de l'estimateur ponctuel diminue. Pour l'étude de pooling, des groupes aléatoires dépendants sont plutôt utilisés et la théorie de Wolter ne peut être appliquée directement. Une étude par simulation a plutôt été effectuée pour cet article pour donner une idée de l'instabilité que peut avoir l'estimateur de variance GAD. La stratégie de pooling énoncée plus haut a été simulée à partir du sous-échantillon de mesures individuelles. Les 16 paires de répliques GAD qu'il est possible d'obtenir par échantillonnage systématique ont été simulées. Pour chaque paire, et pour chacun des contaminants n'ayant pas un nombre de mesures individuelles sous les niveaux de détection frôlant les 100%, on a estimé le coefficient de variation (CV) de l'estimateur ponctuel (1). Le tableau 3 donne quelques statistiques pour les six groupes d'âge et de sexe couverts par le sous-échantillon et pour le contaminant EDP 47 qui est le contaminant le plus prévalent du lot.

Tableau 3 – Moyennes, coefficients de variation, minimums et maximums du coefficient de variation estimé de l'estimateur d'Horvitz-Thompson à partir du sous-échantillon par groupe cible exprimés en pourcentage, EDP 47

	Groupe d'âge et de sexe					
	20-39		40-59		60-79	
	M	F	M	F	M	F
Moyenne	6,8	8,4	11,9	8,0	6,5	9,0
CV	78,0	35,3	73,0	47,1	65,5	44,1
Minimum	0,2	4,3	1,4	2,1	1,3	3,4
Maximum	14,7	12,5	26,0	13,9	14,4	14,5

On y remarque qu'en moyenne le CV se situe aux alentours de 8 %, ce qui indique que l'estimateur ponctuel est précis. Cependant, le CV du CV est très élevé (au-dessus de 30 %) et les minimums et maximums sont très éloignés, ce qui indique que l'estimateur de variance est très instable. D'ailleurs, c'est le même scénario pour tous les autres contaminants de cette simulation, à la différence que leurs CV moyens et maximums sont plus petits. À la lumière de ceci, il semble qu'on devrait utiliser les estimations de variance des GAD au mieux pour se donner une idée de la précision de l'estimateur ponctuel. En effet, comme les CV moyens sont petits, on peut conclure que les estimations ponctuelles seront précises, mais on ne peut pas dire la même chose des estimateurs de variance. On pourra donc se fier aux estimations ponctuelles, mais des estimations par intervalle de confiance ou des tests d'hypothèse devraient être évités autant que possible.

4.2 Stratégie de pooling et d'estimation de la variance avec regroupement de strates

Le problème se complique davantage quand la région de l'Atlantique est prise en compte par regroupement de strates. Pour effectuer le regroupement, on a assigné l'UPE de l'Atlantique au premier GAD (ce qui est équivalent à l'assigner au hasard) et on a utilisé le couple d'estimateurs suivant :

$$\hat{t}_{y\pi}^{(1)} = \frac{\hat{t}_{y\pi}^{\text{Atl.}} + \sum_{s^{(1)} \cap \text{Qc}} w_k y_k}{1/A'} + A\hat{t}_{y\pi}^{(1)\ \text{ouest}}$$
$$\hat{t}_{y\pi}^{(2)} = \frac{\sum_{s^{(2)} \cap \text{Qc}} w_k y_k}{1-1/A'} + A\hat{t}_{y\pi}^{(2)\ \text{ouest}}. \qquad (5)$$

Pour répliquer l'estimateur (3), on approxime le poids comme à la section 3 indépendamment dans chaque réplique. Il est primordial de déterminer quel poids sera utilisé à l'intérieur des répliques. En fait, ceci revient à déterminer la constante

A' dans les estimateurs (5). Effectivement, les pools formés dépendent du poids choisi et il n'est donc pas possible de revenir en arrière et de changer d'idée quant au poids à utiliser une fois que les pools sont formés. Pour faire ce choix, il est bon d'étudier les biais des estimateurs ponctuel $\hat{\theta}_{\text{GAD}} = \left(\hat{t}_{y\pi}^{(1)} + \hat{t}_{y\pi}^{(2)}\right)\Big/2$ et de la variance :

$$\begin{aligned} B\left(\hat{\theta}_{\text{GAD}}\right) &= \frac{A'-2}{2}\left[t_y^{\text{Atl}} + \frac{A'-2}{2\left(A'-1\right)}t_y^{\text{Qc}}\right] \\ B\left(\hat{V}_{\text{GAD}}\right) &= \frac{A'^2}{4}\left[t_y^{\text{Atl}} + \frac{A'-2}{2\left(A'-1\right)}t_y^{\text{Qc}}\right] + B_{\text{regroupement}}^{\text{sans}}\left(\hat{V}_{\text{GAD}}\right) \end{aligned} \tag{6}$$

où $B_{\text{regroupement}}^{\text{sans}}\left(\hat{V}_{\text{GAD}}\right)$ est le biais de l'estimateur de variance qui n'est pas dû au regroupement, soit le biais donné en (4). On note que le regroupement de strates vient ajouter un biais additionnel aux deux estimateurs. Si on choisit $A'>1$, on remarque également que le biais additionnel à l'estimateur de variance sera plus grand que le carré du biais de l'estimateur ponctuel. Ceci fait en sorte que l'estimateur de variance pourra être utilisé comme un estimateur conservateur de l'erreur quadratique moyenne de l'estimateur ponctuel et que les intervalles de confiance seraient aussi conservateurs. Le choix simpliste de $A'=2$ élimine le biais de l'estimateur ponctuel, mais produit un énorme biais pour l'estimateur de variance. D'autre part, en général le regroupement de strates produit des résultats moins biaisés quand les moyennes des strates regroupées sont semblables.

Si on fait l'hypothèse que la moyenne de concentration de contaminant est la même au Québec et en Atlantique, on élimine les biais additionnels en (6) en choisissant $A' = \left(N^{\text{Atl}} + N^{\text{Qc}}\right)\Big/\left(N^{\text{Atl}} + N^{\text{Qc}}/2\right)$, où les quantités *N* sont les tailles de population dans le groupe cible considéré. En raison des délais de production serrés, le choix qui a été retenu est plutôt une fonction du nombre d'UPE échantillonnées : $A' = \left(n_I^{\text{Atl}} + n_I^{\text{Qc}}\right)\Big/\left(n_I^{\text{Atl}} + n_I^{\text{Qc}}/2\right) = 5/3 = 1,\overline{6}$. La logique derrière ce choix était qu'en Atlantique et au Québec, la première réplique représentait les cinq UPE de l'échantillon original à l'aide de trois UPE. Comme la distribution par groupe d'âge et de sexe est similaire dans les deux strates et comme les UPE sont de tailles assez semblables, choisir A' en fonction de la taille de la population ou du nombre d'UPE tirées donne des résultats assez similaires. En effet, utiliser les tailles de population donne des valeurs de A' entre 1,613 et 1,645 selon le groupe cible.

5. CONCLUSION

Le projet de développement de la méthodologie du pooling et de l'estimation s'y rattachant de l'étude de Haines et coll. (2009) avait pour but la production d'estimateurs ponctuels de concentration de contaminants pour dix groupes d'âge et de sexe au niveau national, et si possible la production d'indicateurs de qualité de ces estimateurs. Une approche fondée sur le plan de sondage a été adoptée où l'estimateur ponctuel approxime l'estimateur de Horvitz-Thompson en combinant en pools les répondants avec des poids similaires. Des groupes aléatoires dépendants ont été formés pour estimer la variance avec un biais conservateur. Le regroupement de la strate de l'Atlantique avec celle du Québec pour les fins de l'estimation de la variance et le choix de l'estimateur par réplique utilisé produisent un estimateur ponctuel biaisé. Cependant, ce biais est compensé par un biais additionnel à l'estimateur de la variance qui est lui aussi causé par le regroupement. Lorsque les estimateurs (5) sont utilisés, un choix judicieux de la constante A' éliminerait ces biais additionnels sous l'hypothèse d'une moyenne de concentration égale dans les deux strates regroupées. Finalement, la stratégie adoptée produit peu de répliques, ce qui rend l'estimateur de variance instable.

Les mesures faites en laboratoire seront terminées et prêtes pour l'analyse dans le courant de l'année 2011. Dans ces analyses, il est recommandé de ne pas utiliser les estimations de variance instables du pooling pour produire des intervalles de confiance ou pour faire des tests d'hypothèses, par exemple pour comparer les résultats à d'autres pays ou pour les étudier dans le temps. On pourrait au mieux utiliser ces estimations de variance pour donner une idée générale de la qualité de l'estimateur ponctuel, par exemple pour déterminer s'il est publiable ou non. En effet, dans le tableau 3 on a vu que l'estimation du CV avait tendance à être petite en moyenne, ce qui indique que les estimateurs de moyennes sont

précis et donc publiables. Les tendances étaient les mêmes pour les autres contaminants mesurables de l'étude par mesures individuelles.

Pour d'autres études par pooling semblables, on pourrait considérer d'autres formes d'estimateurs que (5). On pourrait considérer ajuster les poids par strate et par groupe d'âge et de sexe plutôt que globalement au travers de la constante A'. De plus, si (5) est utilisé, le choix de A' pourrait être fonction du groupe cible comme avec $A' = \left(N^{\text{Atl}} + N^{\text{Qc}}\right) \Big/ \left(N^{\text{Atl}} + N^{\text{Qc}}/2\right)$. D'autre part, augmenter le nombre d'UPE sélectionnées dans le plan de sondage aurait plusieurs avantages. Dans le cas où des microdonnées sont disponibles, ceci permettrait d'améliorer la stabilité des estimateurs de variance de Rao-Wu-Yue (1992). Dans l'ECMS, ces estimateurs ne sont pas très stables car seulement 15 UPE sont choisies au total dans les 5 strates, ce qui représente 10 degrés de liberté (11 quand on fait le regroupement de strates). Dans les études de pooling, cela n'aiderait pas à la stabilité de l'estimateur de variance à moins que plus de pools puissent être créés. Cela impliquerait d'avoir une plus grande quantité de spécimens, ce qui serait possible si la quantité de spécimen prélevée par répondant était augmentée ou si le nombre de répondants était augmenté. Cependant, augmenter le nombre d'UPE sans augmenter le nombre de pools et le nombre de répliques aurait eu d'autres avantages dans l'étude de pooling de l'ECMS cycle 1. Ajouter une UPE en Atlantique aurait permis d'éviter le recours au regroupement de strates. Le biais (4) de l'estimateur de variance non dû au regroupement aurait également été réduit. Finalement, le biais dû à la stratification implicite pourrait être réduit. Il est à noter que pour le cycle 2 de l'ECMS, une 16$^{\text{e}}$ UPE a été ajoutée pour la région de l'Atlantique ce qui facilitera l'estimation de la variance.

En conclusion, la variable d'intérêt étudiée dans cet article était la concentration de contaminant mesurée en unités de masse de contaminant par unité de volume de sang. Il pourrait être intéressant du point de vue analytique d'étudier la concentration de contaminant *y* mesurée en unités de masse de contaminant par unité de masse de lipide dans le sang. On pourrait alors imiter les deux approches de la section 3.2 dans le pooling pour reproduire exactement ou de façon approximative l'estimateur d'Horvitz-Thompson $\hat{t}_{y\pi}$ avec la nouvelle définition de la variable *y*. Si on envisage de former des pools avec des volumes de sang différents pour chaque répondant, alors utiliser l'estimateur $\tilde{t}_{y\pi}$ en prenant des volumes de sang proportionnels au poids de sondage divisé par la concentration de lipide dans le sang sera équivalent à utiliser l'estimateur d'Horvitz-Thompson. D'un autre côté, si le volume de sang par répondant dans les pools doit être fixe, $\tilde{t}_{y\pi}$ approximera l'estimateur d'Horvitz-Thompson si on combine en pools le sang des répondants ayant des valeurs similaires du ratio du poids de sondage et de la concentration de lipide dans le sang.

REMERCIEMENTS

Les auteurs aimeraient remercier Harold Mantel et Georgia Roberts pour leur aide précieuse dans ce projet de recherche.

RÉFÉRENCES

Alberta Health and Welness (2008). Alberta biomonitoring program: chemicals in serum of pregnant women in Alberta. Edmonton: Alberta health and wellness.

Bates, M., Buckland, S., Garrett, N., Ellis, H., Needham, L., Patterson, D., Wayman, T. et Russell, D. (2004). « Persistent Organochlorines in the serum of the non-occupationally exposed New Zealand population ». *Chemosphere*, 54, p.1431-1443

Binder, D.A. et Roberts, G. (2009). « Design- and model-based inference for model parameters ». *Sample Surveys : Inference and Analysis*, Vol. 29B, Elsevier B.V.

Caudill, S.P., Turner, W.E. et Patterson Jr., D.G. (2007). « Geometric mean estimation from pooled samples ». *Chemosphere*, 69, p.371-380.

Caudill, S.P. (2008). « Characterizing populations of individuals using pooled samples ». *Journal of exposure science and environmental epidemiology*, p.1-9.

Gambino, J.G. (2009). « Design effect caveats ». *The American Statistician*, **63**, p.141-146.

Giroux, S. (2007). « Enquête canadienne sur les mesures de la santé : aperçu de la stratégie d'échantillonnage ». *Rapports sur la santé*, 18 (supplément), no 82-003-XWF au catalogue de Statistique Canada, p. 35-40.

Haines, D., Rawn, T., Macey, K., Van Oostdam, J., Ryan, J. et Lévesque, J. (2009). Organohalogens in Pooled Serum Specimens from the Canadian Health Measures Survey. Étude menée par Santé Canada et Statistique Canada.

Harden, F., Müller, J., Toms, L., Gaus, C., Moore, M., Päpke, O., Ryan, J., Hobson, P., Symons, R., Horsley, K., Sim M., Van den Berg, M. et Fürst, P. (2004). *Dioxins in the Australian Population: Levels in Blood, National Dioxins*. Program Technical Report No. 9, Australian Government Department of the Environment and Heritage, Canberra.

Needham, L.L., Naiman, D.Q., Patterson Jr., D.G. et LaKind, J.S. (2007). « Assigning concentration values for dioxin and furan congeners in human serum when measurements are below limits of detection: An observational approach ». *Chemosphere*, 67, p.439-447.

Patterson, D., Turner, W., Samuel, C. et Needham, L. (2008). « Total TEQ reference range (PCDDs, PCDFs, cPCBs, mono-PCBs) for the US population 2001-2002 ». *Chemosphere*. **73**, p.261-277.

Rao, J.N.K., Wu, C.F.J. et Yue, K. (1992). « Some Recent work on Resampling Methods for Complex Surveys ». *Techniques d'enquête*, Vol. **18**, No. 2, p.209-217.

Rust, K.F. et Rao, J.N.K. (1996). « Variance estimation for complex surveys using replication techniques ». *Statistical Methods in Medical Research*, **5**, p.283-310.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag, Inc.

Wolter, K.M. (2007). *Introduction to variance estimation*. New York: Springer.