No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Défis méthodologiques reliés à l'analyse des données nutritionnelles de l'enquête sur la santé dans les collectivités canadiennes sur la nutrition

François Verret[1]

Résumé

Statistique Canada a mené en 2004 l'Enquête sur la santé dans les collectivités canadiennes sur la nutrition. L'objectif principal de l'enquête était d'estimer les distributions d'apports alimentaires habituels des Canadiens au niveau provincial pour 15 groupes d'âge et de sexe. En général, on estime de telles distributions à l'aide du logiciel SIDE, mais obtenir ces estimations en tenant compte des choix qui ont été faits en termes de plan d'échantillonnage et de méthode d'estimation de la variabilité d'échantillonnage n'est pas chose facile. Cet article traite des défis méthodologiques reliés à l'estimation de distributions d'apports habituels à l'aide de SIDE avec les données de l'enquête.

MOTS CLÉS: Enquête sur la nutrition; apport habituel; apport quotidien; modèles d'erreur de mesure; bootstrap.

1. Introduction

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une série d'enquêtes transversales comportant deux cycles : les enquêtes des cycles .1 ont pour but d'obtenir des estimations sur la santé générale de la population au niveau régional, tandis que les enquêtes des cycles .2 ont pour but d'obtenir des estimations sur la santé de la population pour un domaine particulier de la santé au niveau provincial. Durant le développement du contenu des enquêtes de l'ESCC, on a jugé que la nutrition était un sujet très important à traiter étant donné les préoccupations des Canadiens et des Canadiennes au sujet de l'embonpoint et de l'obésité. De plus, il y avait très longtemps qu'une enquête nationale sur la nutrition avait eu lieu, la dernière enquête de ce type remontant à 1972. Il a donc été décidé que le cycle 2.2 de l'ESCC porterait sur la nutrition.

Plus spécifiquement, l'objectif principal de l'ESCC sur la nutrition est d'obtenir des estimations des distributions de l'apport nutritionnel habituel dans la population des provinces en termes de nutriments, d'aliments et de groupes alimentaires au niveau provincial pour quinze groupes d'âge et de sexe durant toute l'année 2004. Ces groupes d'âge et de sexe ciblés sont : les bébés âgés de moins d'un an (estimations requises au niveau national seulement); les enfants de 1 à 3 ans et les enfants de 4 à 8 ans sexes confondus; les personnes de 9 à 13 ans, de 14 à 18 ans, de 19 à 30 ans, de 31 à 50 ans, de 51 à 70 ans et les personnes de 71 ans et plus avec distinction du sexe. Les personnes vivant dans les réserves indiennes et sur les terres de la couronne, les résidents d'institutions, les membres à temps plein des Forces canadiennes et les résidents de certaines régions éloignées sont tous exclus de la population cible (les exclusions représentent environ 2 % de la population des provinces canadiennes).

L'apport alimentaire habituel d'une personne peut être décrit comme l'espérance de la distribution théorique générant ses apports alimentaires quotidiens. On peut aussi voir l'apport habituel d'une personne comme la moyenne de ses apports quotidiens sur une longue période de temps. Il est donc impossible de mesurer directement l'apport habituel d'une personne. Cependant, il est possible de mesurer l'apport quotidien d'une personne en effectuant un rappel de ce qu'elle a mangé ou bu pendant une période de 24 heures. De tels rappels ont été effectués pour le cycle 2.2 de l'ESCC où la période couverte par un rappel était les 24 heures de la journée précédant l'interview (de minuit à minuit). Il aurait été impensable de tenter d'estimer l'apport habituel de chacun des répondants étant donné le nombre de rappels nécessaires par répondant, le fardeau du répondant que cela engendre et le coût de la collecte

[1]François Verret, Statistique Canada, 16e étage, Immeuble R.H.-Coats, 100 promenade du pré Tunney, Ottawa, Ontario, Canada, K1A 0T6 (Francois.Verret@statcan.ca)

d'un rappel. Toutefois, l'objectif de l'enquête est d'avoir une estimation de la distribution de l'apport habituel dans la population et il suffit pour ce faire d'un rappel par répondant, d'un second rappel pour un sous-ensemble de ces répondants et d'une méthodologie sophistiquée faisant intervenir la théorie des modèles d'erreur de mesure.

Cet article présente les enjeux rencontrés lors de l'application de cette théorie dans le contexte de l'ESCC. La section suivante présente très brièvement certains aspects du plan de sondage et de l'estimation. Ensuite, la section trois décrit la méthodologie du logiciel Software for Intake Distribution Estimation (SIDE) utilisé pour l'analyse de données nutritionnelles. Finalement, la section quatre discute des enjeux, problèmes et solutions reliés à l'estimation de la variance échantillonnale des estimations produites par SIDE.

2. L'enquête sur la santé dans les collectivités canadiennes sur la nutrition

Il a été déterminé qu'un échantillon de 35 000 répondants (donc autant de premiers rappels) était nécessaire pour que les estimateurs aient la précision désirée. Il s'agit d'un échantillon très important pour une enquête sur la nutrition. On a d'ailleurs dû utiliser plusieurs bases de sondage pour générer cet échantillon qui a été réparti également aux quatre trimestres de l'année 2004 pour éliminer l'effet saisonnier dans les données. Un sous-échantillon de 10 000 seconds rappels a également été recueilli pour tenir compte de la variabilité dans les habitudes alimentaires des répondants d'un jour à l'autre. On peut trouver plus de détails sur le plan de sondage dans Béland et coll. (2003) et dans Junkins et Vigneault (2003). Le plan donne une adresse qu'un intervieweur doit visiter afin d'entrer en contact avec le ménage qui s'y trouve. Les interviews étaient assistées par un ordinateur qui lui faisait sur place la sélection d'une personne du ménage. Le premier rappel était fait en personne au domicile du répondant tandis que le second rappel était fait au téléphone. On considère que les données du premier rappel sont plus fiables que celles du second. En effet, les répondants pouvaient avoir tendance à biaiser leurs réponses au deuxième rappel soit parce qu'ils s'étaient efforcés de mieux manger suite à la première interview, soit afin d'accélérer la seconde interview.

Le processus de pondération de l'échantillon est complexe en raison du plan d'échantillonnage qui a dû être utilisé pour atteindre les objectifs de l'enquête. Pour plus de renseignements sur les étapes habituelles de pondération de l'ESCC, on peut consulter Brisebois et Thivierge (2001). La méthode d'estimation de la variabilité d'échantillonnage préconisée avec les données de l'ESCC est la méthode du bootstrap. Pour le cycle 2.2, on a jugé que le bootstrap était la technique la plus appropriée étant donné la complexité du plan de sondage, le temps requis pour produire les estimations et les bonnes propriétés des estimateurs de variance qu'elle produit (Rao et Wu , 1988). Pour chacune des 500 répliques bootstrap, on échantillonne n_h-1 grappes avec remise parmi les n_h grappes de la strate h. La version du bootstrap qui a été implantée suppose que les fractions de sondage au premier degré sont négligeables.

3. L'analyse des données nutritionnelles à l'aide de software for intake distribution estimation (SIDE)

3.1 Le logiciel SIDE

Le logiciel Software for intake distribution estimation (SIDE) est utilisé par la vaste majorité des analystes lorsque les données d'enquête sont sous la forme de rappels de 24 heures. Sa grande popularité vient du fait qu'il est très complet : il applique le modèle d'erreur de mesure nécessaire pour estimer la distribution de l'apport habituel dans une population en plus de faire a priori et a posteriori des ajustements sophistiqués aux données. La méthodologie de SIDE est décrite en détails dans Nusser et coll. (1996) et est résumée dans la prochaine sous-section.

3.2 Les étapes de SIDE

La première étape consiste à ajuster les données de l'apport quotidien pour améliorer leur qualité et pour simplifier les calculs dans les étapes subséquentes. Le logiciel force d'abord la moyenne et la variance des seconds rappels à être égaux à la moyenne et à la variance des premiers rappels. Cet ajustement est basé sur l'hypothèse que les données du premier rappel sont de meilleure qualité que celles du second. De plus, il est possible d'appliquer un

ajustement par le ratio aux données pour éliminer l'effet de variables nuisibles (continues ou discrètes). La première étape se termine par un lissage des données produisant un échantillon à poids égaux, ce qui simplifie les calculs dans les étapes suivantes. Pour ce faire, on calcule la fonction de répartition empirique des apports quotidiens à l'aide de la formule $\hat{F}_Y(a) = \sum_{i=1}^{n} w_i \sum_{j=1}^{k_i} I_{Y_{ij}}(a)$, où Y_{ij} est la mesure de l'apport quotidien (à laquelle on a appliqué les transformations décrites précédemment) du rappel *j* du répondant *i* et $I_{Y_{ij}}(a)$ est la fonction indicatrice égale à un si Y_{ij} est plus petit ou égal à *a*. La fonction est rendue continue en joignant par des droites les points milieux des marches de $\hat{F}_Y(a)$. $\widetilde{F}_Y$ est la fonction résultante. Un échantillon à poids égaux Z_{ij} est créé en appliquant la formule $Z_{ij} = \widetilde{F}_Y^{-1}\left[\left(\sum_{i=1}^{n} k_i\right)^{-1}\left(s_{ij} - 0.5\right)\right]$ pour *i*=1, 2, ..., *n* et *j*=1, 2, ..., k_i, où s_{ij} est le rang de Y_{ij}.

Le modèle d'erreur de mesure qui est utilisé par SIDE requiert la normalité des données. Pour cette raison, la seconde étape est une transformation complexe à la normalité. Pour ce faire, une transformation puissance est d'abord appliquée. Ensuite, on ordonne les données et on les apparie avec les percentiles de la loi normale centrée réduite. Finalement, on divise les données en intervalles égaux et on trouve les meilleures transformations cubiques sur ces intervalles en effectuant des régressions. On obtient les données X_{ij} qui sont centrées en zéro, contrairement aux Z_{ij} et aux Y_{ij} qui sont centrées à la moyenne des apports quotidiens des premiers rappels non transformés.

À la troisième étape, SIDE ajuste le modèle d'erreur de mesure suivant :

$$\begin{aligned} X_{ij} &= x_i + u_{ij}, \\ u_{ij} &= \sigma_i e_{ij}, \end{aligned} \qquad i = 1,...,n, \qquad j = 1,...,k_i, \tag{1}$$

où X_{ij} est la j^e mesure de l'apport quotidien transformé du répondant *i*, x_i est l'apport habituel du répondant *i* et u_{ij} est l'erreur de mesure. On suppose également les distributions suivantes : $x_i \sim NI(\mu_x, \sigma_x^2)$, $e_{ij} \sim NI(0,1)$ et $\sigma_i^2 \sim (\mu_A, \sigma_A^2)$. Tel que mentionné précédemment, une estimation précise de x_i n'est pas possible avec les données de l'enquête. Les paramètres d'intérêts sont plutôt μ_x et σ_x^2 parce qu'ils définissent entièrement la distribution de l'apport habituel dans la population dans l'échelle normale. Soient *n* le nombre de répondants (ou le nombre de premiers rappels), *N* le nombre total de rappels, k_i le nombre de rappels du répondant *i* (égal à un ou à deux), $\overline{X}_{i\cdot}$ la moyenne des rappels du répondant *i* et $n_0 = N - N^{-1}(4N - 3n)$. Le paramètre de centralité μ_x est alors estimé par $n^{-1}\sum \overline{X}_{i\cdot}$ et σ_x^2 est estimé par la méthode des moments (voir le Tableau 1) par :

$$\hat{\sigma}_x^2 = \frac{1}{n_0}\left[\sum_{i=1}^{n} k_i\left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2 - \frac{n-1}{N-n-1}\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2\right]. \tag{2}$$

Tableau 1
Tableau d'analyse de la variance (tableau d'ANOVA) de SIDE

Source	DL	SC	E[CM]
Individuelle	*n-1*	$\sum_{i=1}^{n} k_i\left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2$	$\frac{n_0}{n-1}\sigma_x^2 + \mu_A$
Résiduelle	*N-n-1*	$\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2$	μ_A
Totale	*N-2*		

La dernière étape de SIDE consiste à transformer la distribution estimée de l'apport habituel de l'échelle normale à l'échelle originale. Cette étape est fort importante parce que les normes de consommation sont définies dans l'échelle originale. La transformation est donnée par l'espérance de l'apport quotidien *Y* dans l'échelle originale sachant la

valeur de l'apport habituel dans l'échelle normale $\ddot{x}_i$: $E\left[Y \middle| x = \ddot{x}_i\right] = E\left[g^{-1}(x+u) \middle| x = \ddot{x}_i\right] = h(\ddot{x}_i)$, où g est la transformation à la normalité de la seconde étape de SIDE, u est l'erreur de mesure et h est la transformation recherchée.

Une fois la distribution de l'apport habituel dans la population estimée, l'analyste s'intéresse en général aux statistiques suivantes : l'apport habituel moyen, les percentiles de l'apport habituel et la proportion de la population sous ou au-dessus d'une limite de consommation donnée. Cependant, il est important d'adapter l'utilisation du logiciel au plan de sondage de l'enquête et à la méthode utilisée pour estimer la variabilité d'échantillonnage. Cette tâche n'est pas simple en raison du nombre et de la complexité des étapes de SIDE. La prochaine section porte sur un des aspects importants de cette adaptation.

4. Problème dans l'estimation de la variance de l'apport habituel

4.1 Le problème

L'estimateur de la variance de l'apport habituel $\hat{\sigma}_x^2$ donné à l'équation (2) peut algébriquement prendre des valeurs négatives. Quand il est négatif, la distribution de l'apport habituel estimée n'est pas définie. Lorsque cela survient, le logiciel SIDE s'arrête et l'utilisateur n'a aucune estimation en sortie. Les valeurs négatives surviennent parce que la méthode des moments est utilisée pour estimer le paramètre σ_x^2. Ainsi, on pourrait penser qu'obtenir un estimateur par une autre méthode ne donnant que des valeurs positives serait une meilleure alternative, mais développer un tel estimateur est une tâche ardue sous le modèle d'erreur de mesure en (1).

La probabilité d'avoir une estimation négative du paramètre de variance lorsque la méthode des moments est utilisée n'est pas nulle et est discutée à la sous-section 4.3. Il faut tenir compte de cet aspect dans l'estimation ponctuelle de la distribution de l'apport habituel. D'autre part, quand la réplication par la méthode du bootstrap est utilisée pour mesurer la variabilité d'échantillonnage, il est possible d'avoir des estimations de variance négatives pour les répliques individuelles. En théorie, une certaine proportion des répliques devraient avoir une estimation de variance négative. Il est donc important lorsque l'on estime la distribution de l'apport habituel d'une population de tenir compte du phénomène des valeurs négatives du paramètre de variance pour l'estimation ponctuelle comme pour l'estimation de la variance échantillonnale.

4.2 Les solutions

4.2.1 Premier cas : L'estimation ponctuelle de la variance des apports habituels est négative

Première option : Utiliser un estimateur de la variance tronqué à zéro

Les valeurs admissibles de la variance de l'apport habituel σ_x^2 sont les valeurs positives ou zéro, tandis que l'estimateur $\hat{\sigma}_x^2$ peut prendre des valeurs négatives. La première solution au problème d'estimation de variance négative consiste à utiliser $\hat{\sigma}_x^{2*} = \max\left(\hat{\sigma}_x^2, 0\right)$ plutôt que $\hat{\sigma}_x^2$ comme estimateur de σ_x^2. Pour les répliques bootstrap on utilisera $\hat{\sigma}_x^{2*(b)} = \max\left(\hat{\sigma}_x^{2(b)}, 0\right)$. Il faut noter que $\hat{\sigma}_x^{2*}$ est un estimateur biaisé vers les valeurs positives car $\hat{\sigma}_x^2$ est sans biais. Lorsque $\hat{\sigma}_x^2$ est négatif, $\hat{\sigma}_x^{2*}$ est nul. On estime alors que l'apport habituel ne varie pas d'une personne à l'autre et donc que toute la population étudiée a exactement le même apport habituel μ_x. La distribution de l'apport habituel estimée est alors discrète avec toute la masse de probabilité concentrée en la moyenne. À partir d'une telle distribution, il est possible d'estimer l'apport habituel moyen ainsi que la proportion de la population sous ou au-dessus d'un certain seuil de consommation (la probabilité estimée est alors soit de 0%, soit de 100%). Cependant, les percentiles ne sont pas définis pour une telle distribution. Il est alors impossible pour l'analyste d'obtenir une estimation d'un percentile à partir de la distribution estimée.

Deuxième option : Imposer une valeur de la variance des apports habituels à partir d'une source externe

Il pourrait être plus raisonnable du point de vue de l'analyste d'utiliser une source externe plus fiable et plus stable pour estimer la variance des apports habituels de la population étudiée. En effet, une estimation de variance négative ou nulle pour l'estimation ponctuelle pourrait être signe que les données ne sont pas suffisamment précises pour estimer le paramètre de variance. Pour justifier cette approche, il faut faire l'hypothèse que la variance des apports habituels du domaine étudié et celle de la source externe sont égales. Cette approche est facile à mettre en oeuvre car on dispose en général de sources externes pour lesquelles l'hypothèse est raisonnable et parce que le logiciel SIDE permet d'imposer une valeur de σ_x^2 par le biais de paramètres d'entrée.

Au niveau de l'estimation de la variance échantillonnale, il faut répliquer la même méthodologie que lors de l'estimation ponctuelle. Il faut donc imposer des valeurs de σ_x^2 aux répliques individuelles. Si on impose la même valeur à toutes les répliques, alors la variabilité d'échantillonnage qui sera mesurée par la réplication bootstrap sera uniquement due à la variation d'une réplique à l'autre dans l'estimation de la moyenne et des transformations. Ceci revient à supposer que la valeur imposée à partir de la source externe est connue exactement, sans erreur d'échantillonnage, ce qui est une hypothèse très forte. Si possible, on ne fera pas cette hypothèse et on palliera à ce problème en faisant varier les valeurs imposées d'une réplique à l'autre pour imiter la variation échantillonnale due à l'estimation de la variance de l'apport habituel à partir de la source. Il est relativement simple d'appliquer une telle méthode lorsque les données de la source proviennent de l'enquête. Par exemple, lors de l'estimation pour une province de l'Atlantique comme l'Île-du-Prince-Édouard, on pourrait imposer les valeurs de variance de l'estimation ponctuelle et des 500 estimations bootstrap obtenues à partir de l'échantillon de l'Atlantique au grand complet plutôt que d'utiliser les valeurs plus instables de l'échantillon provincial.

4.2.2 Second cas : L'estimation ponctuelle de la variance des apports habituels est positive (mais la variance de certaines répliques bootstrap est négative)

Première option : Supprimer les répliques problématiques lors du calcul de la variance échantillonnale

Un réflexe qu'il est naturel d'avoir lorsque l'estimation principale donne une estimation de variance de l'apport habituel positive mais que quelques répliques ont une estimation négative est de supprimer ces répliques dans le calcul de la variance échantillonnale. Cette approche vient biaiser l'estimation de la variance échantillonnale car certaines combinaisons d'unités primaires d'échantillonnage sont alors éliminées du calcul de la variance. Ce biais sera faible si le nombre de répliques éliminées est petit. En conséquence, on devrait se tourner vers cette approche si le nombre de répliques à supprimer du calcul de la variance échantillonnale est faible.

Deuxième option : Utiliser un estimateur de la variance tronqué à zéro

Une alternative qui permet d'utiliser l'information de toutes les répliques est d'adopter l'estimateur de la variance des apports habituels $\hat{\sigma}_x^{2*(b)} = \max\left(\hat{\sigma}_x^{2(b)}, 0\right)$. Cet estimateur est biaisé vers les valeurs positives. En revanche, l'estimateur de la variance échantillonnale qui en découle est moins biaisé qu'avec la solution précédente parce que toutes les combinaisons d'unité primaires d'échantillonnage sont incluses dans le calcul.

Une conséquence de cette approche est que la forme de la distribution des apports habituels estimée de façon ponctuelle sera continue tandis que la forme des distributions obtenues à partir des répliques pour lesquelles $\hat{\sigma}_x^{2*(b)}$ est nulle sera discrète avec toute la masse de probabilité concentrée en la moyenne. Lorsque les statistiques obtenues à partir de ces distributions sont la moyenne de l'apport habituel ou la proportion de la population sous ou au-dessus d'un certain seuil, le calcul de la variabilité d'échantillonnage se fait aisément parce que les statistiques sont bien définies pour l'estimation ponctuelle comme pour toutes les répliques. Toutefois, lorsque la statistique d'intérêt est un percentile, elle n'est pas définie pour les répliques avec une estimation de variance nulle. Une solution à ce problème consiste à utiliser la méthode de Woodruff (1952) adaptée au contexte du bootstrap. En effet, cette

méthode est applicable ici parce que, bien qu'il soit impossible d'obtenir des estimations de percentiles pour toutes les répliques, il est possible d'estimer des probabilités (des aires sous la courbe) pour chacune de ces dernières.

La construction de l'intervalle de confiance dans la suite s'inspire de la section 9.5.2 de Lohr (1999) où on applique la méthode de Woodruff (1952). Soient F_x la fonction de répartition de l'apport habituel dans la population, $\hat{F}_x$ l'estimation ponctuelle de cette distribution et le percentile d'intérêt $\theta_q = F_x^{-1}(q)$. L'intervalle de confiance bootstrap de niveau $1-\alpha$ pour $F_x(y)$ est alors donné par $\hat{F}_x(y) \pm z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(y)]}$, où l'indice B indique l'estimateur de la variance obtenu par la réplication bootstrap et $z_{\alpha/2}$ est le $100\times\alpha/2$-percentile de la loi normale centrée réduite. On peut calculer la variance bootstrap en incluant l'information de toutes les répliques parce qu'il s'agit d'un intervalle pour la proportion de la population avec un apport habituel sous le seuil *y*. On peut donc construire un intervalle de confiance pour $F_x(\theta_q)$, duquel découlera l'intervalle de confiance pour θ_q. Si la distribution échantillonnale de $\hat{F}_x(\theta_q)$ est approximativement normale et puisque F_x et $\hat{F}_x$ sont continues, on a :

$$
\begin{aligned}
0.95 &\approx P\left(\hat{F}_x(\theta_q) - z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(\theta_q)]} \le q \le \hat{F}_x(\theta_q) + z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(\theta_q)]}\right) \\
&= P\left(\hat{F}_x^{-1}\left\{q - z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(\theta_q)]}\right\} \le \theta_q \le \hat{F}_x^{-1}\left\{q + z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(\theta_q)]}\right\}\right),
\end{aligned}
$$

Par conséquent, l'intervalle approximatif de niveau $1-\alpha$ pour le quantile θ_q est donné par :

$$
\left[\hat{F}_x^{-1}\left\{q - z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(\hat{\theta}_q)]}\right\}, \hat{F}_x^{-1}\left\{q + z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(\hat{\theta}_q)]}\right\}\right]. \tag{3}
$$

La variance bootstrap calculée ici fait intervenir des calculs de probabilités avec les répliques plutôt que des calculs de percentiles, ce qu'il est possible de faire pour toutes les répliques, même quand la variance de l'apport habituel estimée est nulle. En pratique, on suivra les étapes suivantes pour appliquer la méthode :

1. Estimer le percentile θ_q de façon ponctuelle $\left(\hat{\theta}_q\right)$;
2. Estimer la proportion de la population avec un apport habituel inférieur à la valeur calculée à la première étape pour chacune des répliques bootstrap $\left(\hat{F}_x^{(1)}(\hat{\theta}_q),\ \hat{F}_x^{(2)}(\hat{\theta}_q),\ \ldots,\ \hat{F}_x^{(500)}(\hat{\theta}_q)\right)$;
3. Calculer l'estimateur de variance bootstrap de $\hat{F}_x(\hat{\theta}_q)$ et construire l'intervalle de confiance correspondant

$$
\left[\max\left\{0, q - z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(\hat{\theta}_q)]}\right\}, \min\left\{1, q + z_{\alpha/2}\sqrt{\hat{V}_B[\hat{F}_x(\hat{\theta}_q)]}\right\}\right]; \tag{4}
$$

4. Estimer de façon ponctuelle les percentiles de l'intervalle de confiance correspondant aux bornes de l'intervalle calculé à l'étape trois. L'intervalle obtenu correspond à celui donné à l'équation (4).

L'intervalle de confiance donné en (4) peut être large si les données ne sont pas suffisamment précises. En particulier, la borne inférieure peut être zéro et la borne supérieure peut être un. Ceci résultera en la perte d'une ou des deux bornes de l'intervalle de confiance construit à l'étape quatre. Cette « perte » d'information est due au manque de précision des données et non à la méthode utilisée. En effet, les données sont alors trop peu précises pour évaluer avec précision l'intervalle de confiance sur une proportion donné en (4).

Troisième option : Imposer une valeur de la variance des apports habituels à partir d'une source externe

Pour augmenter la précision des résultats par rapport à la deuxième option, il est nécessaire de faire des hypothèses supplémentaires sur la distribution de l'apport habituel que l'on veut estimer. Une solution possible est de faire l'hypothèse que la variance des apports habituels de la population étudiée est égale à celle d'une autre population (pour laquelle on a plus de précision dans l'estimation de cette variance). Cette approche a aussi l'avantage de corriger pour les estimations de variance négatives dans les répliques bootstrap. On imposera donc les valeurs de la variance des apports habituels de l'échantillon de la population plus précis lors de l'estimation de la distribution des apports habituels de la population d'intérêt. On se ramène alors à la seconde option décrite à la sous-section 4.2.1. Il

est important de noter que dans ce cas-ci on imposera les valeurs de variances non seulement aux répliques bootstrap qui ont initialement une estimation de variance négative, mais aussi aux autres répliques bootstrap et à l'estimation principale. Effectivement, il est plus logique de profiter au maximum de la précision apportée par l'hypothèse supplémentaire qui est faite.

4.3 Évaluation de l'ampleur du problème

Il peut être intéressant d'estimer l'ampleur du problème du point de vue de la probabilité que le problème survienne dans plusieurs situations. Premièrement, le calcul de cette probabilité peut servir lors de la construction du plan de sondage à l'étape de la détermination des tailles d'échantillon. Pour des valeurs données des paramètres du modèle décrit en (1), on peut calculer les tailles d'échantillons nécessaires pour que la probabilité d'avoir une valeur négative de $\hat{\sigma}_x^2$ soit raisonnablement petite. En effet, plus les nombres de premiers et de seconds rappels sont grands, plus cette probabilité est petite. C'est d'ailleurs l'approche qui a été adoptée dans cette enquête (Junkins et Vigneault, 2003). Une seconde application est qu'on peut calculer cette probabilité une fois l'échantillon recueilli afin de savoir à quelle fréquence le problème surviendra pour différentes valeurs des paramètres du modèle. Ceci peut donc servir à qualifier un échantillon. Finalement, la situation où il est le plus intéressant de calculer une telle probabilité dans le contexte de cet article est au niveau de la réplication bootstrap. On peut calculer la probabilité qu'une réplique donne une estimation de variance négative en fonction des valeurs ponctuelles estimées des paramètres du modèle d'erreur de mesure. Ceci donne le nombre de répliques qui devraient « échouer » avant de faire les calculs associés à chacune des répliques. Cette évaluation est intéressante parce que les calculs de l'ensemble des répliques prennent un temps considérable à s'exécuter dans SIDE et parce qu'elle permet de déterminer à l'avance quelle solution aux répliques qui « échouent » on choisira. En particulier, on pourra décider à l'avance si on aura besoin de renforcer la précision de l'échantillon en faisant une hypothèse supplémentaire sur le paramètre de variance des apports habituels. D'autre part, on peut préférer choisir une option de la sous-section 4.2 en se basant sur la probabilité d'avoir une estimation ponctuelle et des estimations bootstrap négatives plutôt qu'en se basant sur le fait que l'estimation ponctuelle soit négative ou non. Pour calculer la probabilité, on utilisera alors les valeurs espérées des paramètres de variance des apports habituels.

Il n'est pas aisé de calculer la probabilité d'obtenir une estimation de variance négative étant donné le nombre et la complexité des étapes de SIDE. De plus, le plan complexe de l'enquête vient augmenter le niveau de difficulté de cette tâche. Ici on fera donc ce calcul en faisant plusieurs hypothèses. Premièrement, on supposera un plan aléatoire simple, ce qui permettra de contourner l'étape d'égalisation des poids. En deuxième lieu, on se placera dans le cas où aucun ajustement initial n'est requis et où la transformation de normalité g et la transformation « inverse » h sont simplement linéaires. Finalement, on simplifiera le modèle d'erreur de mesure donné en (1) en supposant que la variance intra-individuelle est constante d'une personne à l'autre, c'est-à-dire que le paramètre σ_A^2 est nul. Cette dernière hypothèse vient transformer le modèle d'erreur de mesure en un modèle conventionnel où les X_{ij} sont normaux. Sous ces conditions, on peut utiliser la loi de Fisher pour calculer la probabilité d'intérêt :

$$
\begin{aligned}
P\left(\hat{\sigma}_x^2 < 0\right) &= P\left(\frac{1}{n_0}\left[\sum_{i=1}^{n} k_i \left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2 - \frac{n-1}{N-n-1} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2 \right] < 0 \right) \\
&= P\left(\frac{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} k_i \left(\overline{X}_{i\cdot} - \hat{\mu}_x\right)^2}{\dfrac{n_0}{n-1}\sigma_x^2 + \mu_A} \Bigg/ \frac{\dfrac{1}{N-n-1}\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{k_i} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2}{\mu_A} < \mu_A \Bigg/ \left(\frac{n_0}{n-1}\sigma_x^2 + \mu_A \right) \right) \\
&= P\left(F_{N-n-1}^{n-1} < \mu_A \Big/ \left(\frac{n_0}{n-1}\sigma_x^2 + \mu_A \right) \right).
\end{aligned}
\tag{5}
$$

Si n est grand, alors le côté droit de l'inéquation de la dernière ligne de l'équation (5) est approximativement égal à $1/\left[1 + (1+\gamma)\sigma_x^2 / \mu_A\right]$ (une constante plus petite que un), où $\gamma = (N-n)/n$ est le nombre de seconds rappels divisé

par le nombre de premiers rappels. De l'autre côté de l'inéquation, la statistique de Fisher a une distribution qui est de plus en plus concentrée autour de un à mesure que les tailles d'échantillon augmentent. Par conséquent, la probabilité diminue si n est fixe et γ augmente vers un (si le nombre de seconds rappels augmente vers n) ou si γ est fixe et que n augmente (le nombre de répondants augmente en gardant la proportion de seconds rappels constante). D'autre part, l'estimateur $\hat{\sigma}_x^2$ en (2) est formé d'une combinaison quadratique d'estimateurs de la forme Horvitz-Thompson. Ainsi, puisque les fractions de sondage sont négligeables, la majeure partie de la variabilité de l'estimateur sera due à l'échantillonnage à partir de la population finie et non au mécanisme qui génère cette population (Binder et Roberts, 2003). La réplication bootstrap dans ce cas-ci donne donc une estimation de la variabilité d'échantillonnage et de la variabilité totale. La formule (5) représente donc l'espérance de la proportion de répliques qui donneront une estimation de variance négative. Cependant, en pratique on ne dispose pas des valeurs des paramètres. On remplacera donc ces valeurs dans la formule par les estimations ponctuelles correspondantes (ou par leurs valeurs espérées) pour estimer la probabilité.

Dans le cas général où le plan est complexe et où toutes les étapes de SIDE sont nécessaires, il faut adapter la formule (5) pour obtenir la probabilité exacte. Néanmoins, dans les applications énumérées précédemment on veut une mesure de l'ampleur du problème plutôt qu'un chiffre exact. On peut donc à cette fin se servir de la probabilité donnée en (5) dans le cas général. D'autant plus qu'en pratique la valeur obtenue avec la formule semble être très proche de la proportion observée de répliques qui « échouent ».

Références

Béland, Y., Dufour, J., MacNabb, L., et Pierre, F. (2003), "Sample Design of the 2004 Canadian Nutrition Survey", *Proceedings of the Survey Methods Section, Statistical Society of Canada,* pp. 91-96.

Binder, D. A., et Roberts, G. R. (2003), "Design-based and Model-based Methods for Estimating Model Parameters", dans R. L. Chambers et C. J. Skinner (eds.) *Analysis of survey data*, New York: Wiley, pp. 29-48.

Brisebois, F., et Thivierge, S. (2001), "The Weighting Strategy of the Canadian Community Health Survey", *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Junkins, B., et Vigneault, M. (2003), "Number of Repeat Recalls for CCHS Nutrition Focus Survey", rapport non publié, Ottawa, Canada: Health Canada.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, Duxbury Press, New York.

Nusser S. M., Carriquiry A. L., Dodd K. W., et Fuller W. A. (1996), "A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions", *Journal of American Statistical Association,* 91, pp. 1440-1449.

Rao, J. N. K., et Wu, C. F. J. (1988), "Resampling Inference with Complex Survey Data", *Journal of the American Statistical Association*, 83, pp. 231-241.

Woodruff R. S. (1952), "Confidence Intervals for Medians and Other Position Measures", *Journal of American Statistical Association*, 47, pp. 635-646.