Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Combinaison de l'information de plusieurs enquêtes complexes

par Qi Dong, Michael R. Elliott et Trivellore E. Raghunathan

Date de diffusion : 19 décembre 2014

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Comment accéder à ce produit

Le produit no 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, décembre 2014
Vol. 40, n° 2, pp. 379-387
Statistique Canada, n° 12-001-X au catalogue

Combinaison de l'information de plusieurs enquêtes complexes

Qi Dong, Michael R. Elliott et Trivellore E. Raghunathan[1]

Résumé

Le présent document décrit l'utilisation de l'imputation multiple pour combiner l'information de plusieurs enquêtes de la même population sous-jacente. Nous utilisons une nouvelle méthode pour générer des populations synthétiques de façon non paramétrique à partir d'un bootstrap bayésien fondé sur une population finie qui tient systématiquement compte des plans d'échantillonnage complexes. Nous analysons ensuite chaque population synthétique au moyen d'un logiciel standard de données complètes pour les échantillons aléatoires simples et obtenons une inférence valide en combinant les estimations ponctuelles et de variance au moyen des extensions de règles de combinaison existantes pour les données synthétiques. Nous illustrons l'approche en combinant les données de la *National Health Interview Survey* (NHIS) de 2006 et de la *Medical Expenditure Panel Survey* (MEPS) de 2006.

Mots-clés : Populations synthétiques; répartition prédictive a posteriori; bootstrap bayésien; échantillonnage inverse.

1 Introduction

Il arrive souvent que les organismes d'enquête tirent de multiples échantillons à partir de populations similaires et recueillent des variables semblables, parfois même en utilisant la même base de sondage. Par exemple, la *National Health Interview Survey* (NHIS) et la *National Health and Nutrition Examination Survey* (NHANES) sont toutes deux réalisées par le *National Center for Health Statistics* des États-Unis. Ces deux enquêtes ciblent la population non institutionnalisée des États-Unis et leurs questions se recoupent considérablement. En combinant l'information provenant de diverses enquêtes, nous espérons obtenir une inférence plus exacte pour la population que si nous utilisions les données d'une seule enquête.

L'une des plus grandes difficultés liées à une telle combinaison d'information concerne la compatibilité de diverses sources de données. Les enquêtes peuvent utiliser différents plans d'échantillonnage ou modes de collecte des données, ce qui peut donner lieu à diverses propriétés d'erreur d'échantillonnage et d'erreur non due à l'échantillonnage. Au lieu de compiler directement les données de plusieurs enquêtes à partir d'une analyse simple, nous devons corriger pour tenir compte des écarts entre les données et les rendre comparables.

Différentes méthodes pour la combinaison de données tirées de deux enquêtes différentes ont été proposées dans la documentation sur les méthodes d'enquête (Hartley 1974; Skinner et Rao 1996; Lohr et Rao 2000; Elliott et Davis 2005; Raghunathan, Xie, Schenker, Parsons, Davis, Dodd et Feuer 2007; Schenker, Gentleman, Rose, Hing et Shimizu 2002; Schenker et Raghunathan 2007; Schenker, Raghunathan et Bondarenko 2009). Les derniers travaux de Raghunathan et coll. (2007) et Schenker et coll. (2009) appliquaient des approches fondées sur un modèle. Le principe pour l'approche fondée sur un

1. Qi, Dong, Google, Inc., 1R4A, Quad 5, Google Inc, 399 N. Whisman Road, Mountain View, CA 94043. Courriel : qdong@google.com; Michael R. Elliott, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109 et Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. Courriel : mrelliot@umich.edu; Trivellore E. Raghunathan, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109 et Survey Methodology Program, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48106. Courriel : teraghu@umich.edu.

modèle est d'intégrer un modèle d'imputation aux données de meilleure qualité et d'utiliser le modèle intégré pour imputer les valeurs dans les autres échantillons de qualité inférieure. Dans la mesure où le modèle d'imputation est correctement défini, cette approche peut tirer profit des forces des différentes sources de données et améliorer l'inférence statistique. Cependant, comme suggéré par Reiter, Raghunathan et Kinney (2006), lorsque l'échantillon est recueilli au moyen de plans d'échantillonnage complexes, le fait de ne pas tenir compte de ces caractéristiques peut entraîner des estimations biaisées de la perspective fondée sur le plan. Cependant, il est très difficile de tenir pleinement compte des caractéristiques du plan d'échantillonnage complexe en pratique. Par exemple, Raghunathan et coll. (2007) et Schenker et coll. (2009) ont utilisé une méthode simplifiée pour tenir compte de la stratification et de la mise en grappe. Raghunathan et coll. (2007) ont utilisé un concept rudimentaire de l'effet de plan et Schenker et coll. (2009) ont utilisé des scores de propension pour créer des sous-groupes de correction pour la modélisation.

Ici, nous proposons une nouvelle méthode pour combiner plusieurs enquêtes; cette méthode tient compte des caractéristiques du plan d'échantillonnage complexe dans chaque enquête. La population non observée dans chaque enquête sera traitée comme des données manquantes à imputer. Le modèle d'imputation tiendra compte des caractéristiques du plan complexe au moyen d'une nouvelle méthode de génération de la population synthétique non paramétrique (Dong, Elliott et Raghunathan 2014). Pour chaque enquête, les données observées et la population non observée imputée produisent des populations synthétiques multiples. Une fois la population complète obtenue, les caractéristiques du plan d'échantillonnage complexe, comme la stratification, la mise en grappe et la pondération, seront inutiles dans l'analyse, et les populations synthétiques peuvent être traitées comme des échantillons aléatoires simples équivalents. Enfin, l'estimation de la quantité de population d'intérêt sera calculée à partir de chaque population synthétique et sera combinée d'abord à chaque enquête individuelle, puis à plusieurs enquêtes.

Le présent document procède comme suit : la section 2 résume la génération de la population synthétique tout en tenant compte des caractéristiques du plan d'échantillonnage complexe au moyen de l'approche non paramétrique. La section 3 décrit la méthodologie pour produire les estimations combinées à partir de ces populations synthétiques multiples. À la section 4, nous appliquons la méthode proposée pour combiner la NHIS de 2006 et la *Medical Expenditure Panel Survey* (MEPS) afin d'estimer les taux de couverture de l'assurance-maladie de la population des États-Unis. La section 5 conclut par une discussion et des orientations pour les recherches à venir.

2 Production de populations synthétiques à partir des données d'une seule enquête en tenant compte des plans d'échantillonnage complexes

Dong et coll. (2014) ont poursuivi les travaux relatifs au bootstrap bayésien pour la population finie afin d'élaborer une approche non paramétrique de la génération de distributions prévisionnelles postérieures. Voici un résumé de l'algorithme pour tirer les populations synthétiques l, $l = 1,...,L$ pour les plans d'échantillonnage en grappes stratifiés ayant des probabilités inégales de sélection :

1. Utilisez le bootstrap bayésien (BB) (Rubin 1981) pour tenir compte de la stratification et de la mise en grappe. Tirez un échantillon aléatoire simple avec remise (EASAR) de taille m_h à partir des grappes c_h dans chaque strate $h=1,...,H$ et calculez les poids de rééchantillonnage bootstrap pour chacune des observations n_{hi} dans chaque grappe comme suit : $w^{*(l)}=\left\{w_{hi}^{*(l)},\ h=1,\ldots,H,\ i=1,\ldots,c_h,\ k=1,\ldots,n_{hi}\right\}$, où $w_{hik}^{*}=w_{hik}\left(\left(1-\sqrt{\left(m_h/c_h-1\right)}\right)+\sqrt{\left(m_h/c_h-1\right)}\left(c_h/m_h\right)m_{hi}^{*}\right)$ et m_{hi}^{*} indique le nombre de fois qu'une grappe i, $i=1,\ldots,c_h$ est sélectionnée. Pour que tous les poids de rééchantillonnage soient non négatifs, $m_h\le\left(c_h-1\right)$; ici et ci-après, nous supposons que $m_h=\left(c_h-1\right)$.

2. Utilisez le bootstrap bayésien pour la population finie (BBPF) (Lo 1986; Cohen 1997) pour des probabilités de sélection inégales pour tenir compte des probabilités inégales de sélection. Pour chaque grappe i dans la strate h où la taille de la population est de N_{hi}, tirez un échantillon de taille $N_{hi}-n_{hi}$, indiquée par $\left(y_1^{*},\ldots,y_{N_{hi}-n_{hi}}^{*}\right)$, en tirant y_{hik}^{*} des données relatives à une grappe $\left(y_1,\ldots,y_{n_{hi}}\right)$ avec la probabilité $\dfrac{w_{hik}^{*}-1+l_{hik,j-1}*\left(N_{hi}-n_{hi}\right)/n_{hi}}{N_{c_H}-n_{c_H}+\left(j-1\right)*\left(N_{hi}-n_{hi}\right)/n_{hi}}$, où w_{hik}^{*} est le poids de rééchantillonnage de l'unité k dans la grappe i de la strate h, et $l_{hik,j-1}$ est le nombre de sélections bootstrap de y_{hik} dans $y_1^{*},\ldots,y_{j-1}^{*}$. Créez la population BBPF $y_1,\ldots,y_{n_{hi}},y_1^{*},\ldots,y_{N_{hi}-n_{hi}}^{*}$.

3. Produisez F échantillons BBPF pour chaque échantillon BB, représenté par $S_{l1},\ldots,S_{lF},l=1,\ldots,L$. Rassemblez les F échantillons BBPF afin de produire une population synthétique, S_l. (parce que $N=\sum_h\sum_i N_{hi}$ pourrait avoir une taille déraisonnablement grande, la production d'un échantillon de taille $k*n$ pour une grande valeur de k est suffisante).

3 Règle de combinaison pour les populations synthétiques d'enquêtes multiples

Supposons que $Q=Q\left(Y\right)$ est la quantité de population d'intérêt selon l'ensemble de variables Y qui sont recueillies au cours de plusieurs enquêtes : par exemple, une moyenne de population, une proportion ou un total, un vecteur des coefficients de régression, etc. Par souci de simplicité de l'exposition, nous supposons que Q est scalaire. Supposons qu'au moyen des données d'une seule enquête s, nous créons L populations synthétiques, $S_l^{(s)}$, $l=1,\ldots,L$, à partir de la méthodologie résumée à la section 2. Désignons $Q_l^{(s)}$ comme l'estimation correspondante de la quantité de population Q obtenue de la population synthétique l générée au moyen des données de l'enquête s (soulignons que cette estimation peut être obtenue à partir d'une hypothèse d'échantillonnage aléatoire simple). Dong et coll. (2014)

démontrent que conformément à des suppositions asymptotiques raisonnables (taille d'échantillon suffisante pour la quantité d'échantillons d'intérêt répartie suivant une distribution normale, populations synthétiques conformes au plan d'enquête),

$$Q \mid S_1^{(s)},...,S_L^{(s)} \dot{\sim} t_{L-1}\left(\bar{Q}_L^{(s)},\left(1+L^{-1}\right)B_L^{(s)}\right) \tag{3.1}$$

où $\bar{Q}_L^{(s)} = L^{-1}\sum_{l=1}^{L} Q_l^{(s)}$ est la moyenne de Q à l'étendue des populations synthétiques L et $B_L^{(s)} = (L-1)^{-1}\sum_{l=1}^{L}\left(Q_l^{(s)} - \bar{Q}_L^{(s)}\right)^2$ est la variance entre les étapes d'imputation. Le résultat suit immédiatement à partir de la section 4.1 de Raghunathan, Reiter et Rubin (2003), et est fondé sur les règles standard de combinaison de l'imputation multiple de Rubin (1987). La variance moyenne de l'imputation « interne » est de zéro, puisque la population complète est synthétisée; par conséquent, la variance a posteriori de Q dépend entièrement de la variance entre les étapes d'imputation.

La règle de combinaison obtenue à (3.1) ne donnera pas nécessairement une inférence valide pour les paramètres d'intérêt pour plusieurs enquêtes, puisque les modèles pour générer les populations synthétiques pour les enquêtes multiples pourraient être différents. Par conséquent, une nouvelle règle pour combiner les estimations dans plusieurs enquêtes doit être élaborée.

3.1 Approximation normale lorsque L est grand

Supposons que $\bar{Q}_L^{(s)}$ et $B_L^{(s)}$ soient l'estimateur combiné de la quantité de population d'intérêt et sa variance pour l'enquête s obtenue au moyen des formules de combinaison pour les populations synthétiques $S_{syn}^{(s)} = \left\{S_l^{(s)}, l = 1,\ldots,L\right\}$, $s = 1,\ldots,S$ dans le contexte d'une seule enquête. Lorsque L est grand, nous avons

$$Q \mid S_{syn}^{(1)},...,S_{syn}^{(S)} \dot{\sim} N\left(\bar{Q}_L, B_L\right) \tag{3.2}$$

où $\bar{Q}_L = \sum_{s=1}^{S}\left(\bar{Q}_L^{(s)} / B_L^{(s)}\right) \Big/ \sum_{s=1}^{S}\left(1/B_L^{(s)}\right)$ et $B_L = 1 \Big/ \sum_{s=1}^{S}\left(1/B_L^{(s)}\right)$. L'équation (3.2) suit immédiatement à partir des résultats bayésiens standard, en supposant que 1) la vraie variance de $\bar{Q}_L^{(s)}$, B_s, peut être estimée par $B_L^{(s)}$ obtenue à partir des populations synthétiques comme à la section 3, c.-à-d. $\left(\bar{Q}_L^{(s)} \mid Q, B_s\right) = \left(\bar{Q}_L^{(s)} \mid Q, B_L^{(s)}\right) \sim N\left(Q, B_L^{(s)}\right)$, 2) chaque enquête est indépendante et 3) Q a une répartition a priori non informative $\pi\left(Q \mid B_L^{(s)}\right) \propto 1$.

3.2 Répartition corrigée en fonction de T pour un L petit ou modéré

Pour un L, petit ou modéré, la répartition a posteriori de Q s'estime mieux comme suit

$$Q \mid S_{syn}^{(1)},...,S_{syn}^{(S)} \dot{\sim} t_{\nu_L}\left(\bar{Q}_L, \left(1+L^{-1}\right)B_L\right) \tag{3.3}$$

où $\bar{Q}_L$ et B_L sont définis comme à 3.1, et les degrés de liberté $\vartheta_L = (L-1) \Big/ \sum_{s=1}^{S} \left(\left(1/b_L^{(s)}\right) \Big/ \sum_{s=1}^{S} \left(1/b_L^{(s)}\right) \right)^2$. Dong (2012) fournit davantage de détails, qui suivent les recherches de Raghuanthan et coll. (2003) ayant servi à déterminer indirectement les résultats pour L grand.

4 Estimations combinées de la couverture d'assurance-maladie de la NHIS, la MEPS et la BRFSS

Les données de la NHIS et de la MEPS de 2006 sont des échantillons probabilistes à plusieurs degrés qui comprennent la stratification, la mise en grappe et le suréchantillonnage de certaines sous-populations (p. ex. les Noirs, les Hispaniques et les Asiatiques au cours des dernières années). Pour des motifs de confidentialité, les vraies strates et les UPE sont supprimées. La NHIS est publiée avec 300 pseudo-strates et deux pseudo-UPE par strate; la MEPS, qui est un sous-échantillon de ménages qui participent à la NHIS, est publiée avec 203 pseudo-strates et jusqu'à trois pseudo-UPE par strate (Ezzati-Rice, Rohde et Greenblatt 2008; National Center for Health Statistics 2007). Dans le cadre de la NHIS et de la MEPS, on demande à un adulte sélectionné au hasard dans chaque ménage s'il a une assurance-maladie et, dans l'affirmative, s'il s'agit d'une assurance privée ou publique. Nous considérons cette répartition trinomiale de la couverture d'assurance dans l'ensemble de la population adulte, ainsi que dans les sous-populations composées d'hommes, d'Hispaniques, de Blancs non hispaniques et de Blancs non hispaniques gagnant de 25 000 $ à 35 000 $ par année. Nous supprimons les cas comportant des données manquantes et nous étudions les cas complets. Nous obtenons ainsi 20 147 et 20 893 cas pour les données de la NHIS et de la MEPS, respectivement.

La BRFSS de 2006 est obtenue au moyen de la composition aléatoire de numéros à partir d'un échantillonnage par liste, stratifié par état. Bien que ce genre de plans évite la mise en grappe, une probabilité inégale de sélection est introduite, parce que la taille de l'échantillon est à peu près égale dans chaque état; en outre, un seul adulte est échantillonné par ménage. Contrairement à la NHIS et à la MEPS, la BRFSS se contente de demander si la personne a une assurance ou pas, ce qui nous permet de calculer uniquement la proportion de répondants qui ne sont pas couverts par une assurance. Nous supprimons les cas ayant des valeurs manquantes aux questions et nous nous concentrons sur notre simulation des cas complets. Il y a 294 559 cas complets dans les données de la BRFSS de 2006.

Nous générons les populations synthétiques pour les trois enquêtes à partir de 200 échantillons BB, chacun comportant 10 échantillons BBPF de taille $5n$ $(B = 200,\ F = 10,\ k = 5)$. Nous produisons alors les estimations combinées des taux de couverture par une assurance-maladie des personnes au moyen de la méthode de combinaison d'enquêtes susmentionnée. Étant donné que pour les trois enquêtes, nous savons si les personnes ont une assurance ou pas, nous pouvons combiner la NHIS, la BRFSS et la MEPS afin d'estimer la proportion de personnes non assurées. Cependant, la BRFSS ne demande pas aux personnes quel type d'assurance elles ont (publique ou privée). Pour ces proportions, nous pouvons seulement combiner la NHIS et la MEPS. Les résultats sont résumés au tableau 4.1. Les estimations de la variance pour l'estimateur combiné sont bien plus petites que celles qui ont été obtenues à partir des données réelles. Plus précisément, la précision des estimations obtenues de la NHIS est accrue de 43 % en

moyenne, l'augmentation la plus prononcée de 98 % étant obtenue par la combinaison de la NHIS et de la MEPS. Les gains de précision pour la MEPS sont encore plus importants. L'augmentation moyenne de la précision pour la MEPS est de 101 %, l'augmentation la plus prononcée étant de 202 %. La précision est accrue davantage lorsque nous combinons les trois enquêtes. Par exemple, pour la proportion de personnes qui ne sont pas assurées, en moyenne, la précision est quintuplée pour la NHIS, multipliée par 1,5 pour la BRFSS et par 4,2 pour la MEPS. Autrement dit, les gains de précision grâce à l'utilisation de l'information de plusieurs enquêtes peuvent être considérables, et plus nous combinons d'information, plus les gains de précision seront importants.

5 Discussion

Dans cet article, nous proposons une nouvelle façon de combiner de l'information de plusieurs enquêtes complexes. Nous appliquons la nouvelle méthode pour combiner de l'information au sujet de la couverture d'assurance-maladie dans le cadre de la NHIS, de la MEPS et de la BRFSS de 2006. Les résultats indiquent que l'estimation combinée est plus précise que les estimations des enquêtes individuelles. Comme l'ont démontré les travaux précédents (Dong et coll. 2014), peu d'information se perd en ce que les propriétés d'échantillonnage des inférences de la population synthétique et de l'échantillon réel sont très semblables. Par conséquent, lorsque nous combinons les estimations de trois échantillons, l'estimation combinée est considérablement plus efficace que les estimations des enquêtes individuelles. (Soulignons que cette application sert principalement à titre d'exemple; des inférences semblables pourraient être faites en calculant les estimations fondées sur le plan et les variances pour chacune des enquêtes, puis en appliquant la règle de combinaison dans (3.2) dans les estimations fondées sur le plan.)

Cette nouvelle méthode de combinaison d'enquêtes offre deux avantages par rapport à la méthodologie existante. D'abord, l'approche utilisée ici pour générer des populations synthétiques, décrite en détail dans Dong et coll. (2014), tient compte du plan de sondage complexe de façon non paramétrique en extrapolant la méthodologie du bootstrap bayésien de la population finie. Étant donné que les populations synthétiques obtenues peuvent être analysées comme des échantillons aléatoires simples, l'information d'autres enquêtes peut être utilisée pour tenir compte des erreurs non dues à l'échantillonnage et/ou pour imputer les variables manquantes. Un autre avantage de cette méthode est qu'elle n'a pas de limite du nombre d'enquêtes à combiner, dans la mesure où les enquêtes ont la même population sous-jacente. La méthode proposée qui tient compte des caractéristiques du plan d'échantillonnage complexe peut être appliquée à chaque enquête indépendamment. Une fois l'information manquante imputée, quel que soit le nombre d'enquêtes à combiner, il nous suffit de combiner les estimations de chaque enquête au moyen de la règle de combinaison décrite dans le présent document. Un dernier avantage de l'approche proposée est la capacité des populations synthétiques générées par la méthode non paramétrique de conserver les valeurs manquantes aux questions dans les données réelles. Cette méthode pourrait combler une lacune dans la zone visée par l'imputation multiple, parce que les méthodes d'imputation existantes ne prennent pas en compte habituellement les caractéristiques du plan d'échantillonnage complexe dans les données et imputent les valeurs manquantes comme s'il s'agissait d'échantillons aléatoires simples. Nous envisageons cette application dans les travaux à venir.

Tableau 4.1
Estimations individuelles et combinées pour la NHIS, la MEPS et la BRFSS de 2006

Domaine		Données réelles (plan complexe)			Estimations combinées	
	Types	NHIS	BRFSS	MEPS	NHIS et MEPS	NHIS, BRFSS et MEPS
Population complète	**Proportion**					
	Régime privé	0,746		0,735	0,741	
	Régime public	0,075		0,133	0,094	
	Non assuré	0,179	0,154	0,132	0,152	0,153
	Variance					
	Régime privé	2,46E-05		2,78E-05	1,61E-05	
	Régime public	6,29E-06		1,44E-05	5,35E-06	
	Non assuré	1,84E-05	3,32E-06	1,41E-05	9,80E-06	2,55E-06
Hommes	**Proportion**					
	Régime privé	0,740		0,735	0,738	
	Régime public	0,060		0,101	0,074	
	Non assuré	0,200	0,167	0,164	0,181	0,172
	Variance					
	Régime privé	3,32E-05		3,87E-05	2,06E-05	
	Régime public	6,82E-06		1,53E-05	5,72E-06	
	Non assuré	2,94E-05	8,88E-06	2,64E-05	1,51E-05	5,61E-06
Hispaniques	**Proportion**					
	Régime privé	0,494		0,506	0,5014	
	Régime public	0,096		0,161	0,1157	
	Non assuré	0,410	0,371	0,334	0,3684	0,3689
	Variance					
	Régime privé	1,24E-04		1,73E-04	9,76E-05	
	Régime public	2,57E-05		8,03E-05	2,66E-05	
	Non assuré	1,23E-04	7,18E-05	1,19E-04	8,71E-05	3,79E-05
Blancs non hispaniques	**Proportion**					
	Régime privé	0,805		0,788	0,796	
	Régime public	0,062		0,116	0,081	
	Non assuré	0,134	0,1059	0,096	0,113	0,107
	Variance					
	Régime privé	2,99E-05		3,35E-05	1,97E-05	
	Régime public	8,20E-06		1,81E-05	6,86E-06	
	Non assuré	2,02E-05	2,15E-06	1,51E-05	1,02E-05	1,90E-06
Blancs non hispaniques ayant un revenu de [25 000 $ à 35 000 $)	**Proportion**					
	Régime privé	0,827		0,813	0,821	
	Régime public	0,039		0,079	0,053	
	Non assuré	0,134	0,173	0,108	0,122	0,154
	Variance					
	Régime privé	1,0E-04		1,39E-04	7,74E-05	
	Régime public	2,82E-05		6,31E-05	2,52E-05	
	Non assuré	7,24E-05	2,78E-05	8,92E-05	5,14E-05	1,93E-05

Bibliographie

Cohen, M.P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 635-638.

Dong, Q. (2012). Unpublished PhD thesis, University of Michigan.

Dong Q., Elliott, M.R. et Raghunathan T.E. (2014). Une méthode non paramétrique de production de populations synthétiques qui tient compte des caractéristiques des plans de sondage complexes. *Techniques d'enquête*, 40 (1), 33-52.

Elliott, M.R. et Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Journal of the Royal Statistical Society C*: Applied Statistics, 54, 595-609.

Ezzati-Rice, T.M., Rohde, F. et Greenblatt, J. (2008). Sample design of the medical expenditure panel survey household component, 1998–2007. *Methodology Report No. 22*. Agency for Healthcare Research and Quality, Rockville, MD. Consulté au http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr22/mr22.pdf, février 2014.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *The Indian Journal of Statistics*, C, 38, 99-118.

Lo, A.Y. (1986). Bayesian statistical inference for sampling a finite population. *Annals of Statistics*, 14, 1226-1233.

Lohr, S.L. et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

National Center for Health Statistics (2007). Data file documentation, National Health Interview Survey, 2006 (machine readable data file and documentation). *National Center for Health Statistics*, Centers for Disease Control and Prevention, Hyattsville, Maryland. Consulté au: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2006/srvydesc.pdf, février 2014.

Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., Xie, D.W., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W. et Feuer, D.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102, 474-486.

Reiter, J.P., Raghunathan, T.E. et Kinney, S.K. (2006). L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes. *Techniques d'enquête*, vol. 32, 161-168.

Rubin, D.B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9, 131-134.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schenker, N., Gentleman, J.F., Rose, D, Hing, E. et Shimizu, I.M. (2002). Combining estimates from complementary surveys: A case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*, 117, 393-407.

Schenker, N. et Raghunathan, T.E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, 26, 1802-1811.

Schenker, N., Raghunathan, T.E. et Bondarenko, I. (2009). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine*, 29, 533-545.

Skinner, C.J. et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.