Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

# Une méthode non paramétrique de production de populations synthétiques qui tient compte des caractéristiques des plans de sondage complexes

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


Juin 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 n° 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 ».



**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, juin 2014
Vol. 40, N° 1, pp. 33-52
Statistique Canada, N° 12-001-X au catalogue


# Une méthode non paramétrique de production de populations synthétiques qui tient compte des caractéristiques des plans de sondage complexes

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

Résumé

Dans la littérature n'ayant pas trait aux sondages, il est fréquent de supposer que l'échantillonnage est effectué selon un processus aléatoire simple qui produit des échantillons indépendants et identiquement distribués (IID). De nombreuses méthodes statistiques sont élaborées en grande partie dans cet univers IID. Or, l'application de ces méthodes aux données provenant de sondages complexes en omettant de tenir compte des caractéristiques du plan de sondage peut donner lieu à des inférences erronées. Donc, beaucoup de temps et d'effort ont été consacrés à l'élaboration de méthodes statistiques permettant d'analyser les données d'enquêtes complexes en tenant compte du plan de sondage. Ce problème est particulièrement important lorsqu'on génère des populations synthétiques en faisant appel à l'inférence bayésienne en population finie, comme cela se fait souvent dans un contexte de données manquantes ou de risque de divulgation, ou lorsqu'on combine des données provenant de plusieurs enquêtes. En étendant les travaux antérieurs décrits dans la littérature sur le bootstrap bayésien en population finie, nous proposons une méthode pour produire des populations synthétiques à partir d'une loi prédictive a posteriori d'une façon qui inverse les caractéristiques du plan de sondage complexe et génère des échantillons aléatoires simples dans une optique de superpopulation, en ajustant les données complexes afin qu'elles puissent être analysées comme des échantillons aléatoires simples. Nous considérons une étude par simulation sous un plan de sondage en grappes stratifié avec probabilités inégales de sélection, et nous appliquons la méthode non paramétrique proposée pour produire des populations synthétiques pour la National Health Interview Survey (NHIS) et la Medical Expenditure Panel Survey (MEPS) de 2006, qui sont des enquêtes à plan de sondage en grappes stratifié avec probabilités inégales de sélection.

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

# 1 Introduction

Hors du contexte des techniques d'enquête, les méthodes statistiques ont habituellement été élaborées sans beaucoup se soucier du plan d'échantillonnage, souvent en supposant implicitement avoir affaire à des échantillons aléatoires simples ou, parfois, à des échantillons en grappes à un degré. En statistique d'enquête contemporaine, d'importants travaux ont pour objectif d'étendre les méthodes à l'analyse de données d'enquêtes complexes (Skinner, Holt et Smith, 1989), en tenant compte de problèmes tels que la stratification, les probabilités inégales de sélection, le biais de non-réponse ou le calage. Hinkins, Oh et Scheuren (1997) ont proposé un algorithme de plan de sondage inverse qui relie la statistique d'enquête et la statistique classique sous un autre angle. Leur idée fondamentale consiste à choisir un sous-échantillon qui possède inconditionnellement une structure d'échantillon aléatoire simple. Le sous-échantillon est souvent nettement plus petit que l'échantillon original, de sorte qu'ils proposent de répéter le processus indépendamment un grand nombre de fois et de prendre la moyenne des résultats pour augmenter la précision. Ils décrivent aussi des schémas d'échantillonnage inverse exacts ou approximatifs pour l'échantillonnage aléatoire simple stratifié, l'échantillonnage en grappes à un degré et l'échantillonnage en grappes à deux degrés. Cependant, l'application de cette nouvelle idée n'est pas très répandue en pratique,

1. Qi Dong, Netflix Inc., 100, Winchester Cir, Los Gatos (CA) 95032, courriel : qidong@umich.edu; Michael R. Elliott, Department of Biostatistics, University of Michigan, 1420, Washington Heights, Ann Arbor (MI) 48109, 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, Survey Methodology Program, Institute for Social Research, University of Michigan, 426, Thompson St., Ann Arbor (MI) 48106. courriel : teraghu@umich.edu.

peut-être parce qu'elle est très gourmande en temps de calcul et que les pertes de précision sont souvent considérables. En outre, produire des populations synthétiques à partir d'une loi prédictive a posteriori de population conditionnellement aux données d'enquêtes complexes en tenant compte du plan de sondage complexe n'est pas chose simple (Little, 1991). Néanmoins, ces dernières années, la demande de populations synthétiques s'est accrue en vue de pouvoir traiter les problèmes de troncation des pondérations ou de windsorisation (Lazzeroni et Little, 1998; Elliott et Little, 2000; Elliott, 2007; Chen, Elliott et Little, 2010), de risque de divulgation (Little, 1993; Raghunathan, Reiter et Rubin, 2003; Reiter, 2004, 2005) ou de combinaison de données provenant de plusieurs enquêtes (Raghunathan, Xie, Schenker, Parsons, Davis, Dodd et Feuer, 2007; Dong, 2012). Les populations synthétiques sont souvent générées sous une hypothèse distributionnelle (normale, binomiale, Poisson) en approximant la loi a posteriori des paramètres du modèle par la loi normale asymptotique. La moyenne et la matrice de covariance de la loi normale sont estimées après avoir tenu compte des caractéristiques du plan de sondage complexe (Raghunathan et coll., 2007).

Une grande faiblesse des méthodes fondées sur un modèle tient au fait que, si le modèle est très mal spécifié, il donnera lieu à des inférences invalides (Little, 2004). Dans un contexte multivarié, nous devons prendre en considération les liens qui existent entre les variables d'intérêt et déterminer un modèle approprié qui est ajusté aux données, ce qui peut être difficile si les données contiennent différents types de variables. Dans le présent article, nous proposons une méthode non paramétrique qui fait pendant aux méthodes fondées sur un modèle pour générer des populations synthétiques. Les travaux que nous présentons étendent le bootstrap bayésien en population finie et les modèles a posteriori de Pólya connexes de Lo (1988), Ghosh et Meeden (1983) et Cohen (1997) en vue de tenir compte des plans de sondage complexes. Puisqu'elle atteint le même objectif que la méthode d'échantillonnage inverse, elle peut être traitée comme la version bayésienne en population finie de l'échantillonnage inverse. Pour faire des inférences en utilisant ce bootstrap bayésien en population finie pondéré, nous pouvons soit nous servir directement des tirages, soit, par souci d'efficacité des calculs, utiliser les résultats établis antérieurement dans la littérature sur le risque de divulgation et l'imputation multiple, puisque ces populations produites non paramétriquement peuvent être considérées comme des imputations multiples des éléments non observés de la population.

Le plan de l'article est le suivant. À la section 2, nous discutons brièvement des populations synthétiques dans le contexte de l'inférence bayésienne en population finie. À la section 3, nous passons en revue et résumons la méthode du bootstrap bayésien et son extension en population finie, et montrons que, pour un échantillonnage avec probabilités inégales, la loi de probabilité des populations synthétiques générées sous une variante du modèle de l'urne de Pólya concorde avec la loi prédictive a posteriori d'un bootstrap bayésien en population finie. À la section 4, nous présentons la méthode proposée sous échantillonnage en grappes stratifié avec probabilités de sélection inégales. À la section 5, nous montrons que l'inférence à partir de ces populations synthétiques générées non paramétriquement peut être obtenue en utilisant les résultats tirés de la littérature sur le risque de divulgation et l'imputation multiple, où chaque population synthétique possède une variance « intra-imputation » nulle. À la section 6, nous décrivons une étude par simulation réalisée pour évaluer la performance de la méthode non paramétrique dans un contexte de rééchantillonnage. À la section 7, nous appliquons la méthode pour générer des populations synthétiques qui peuvent être utilisées pour estimer les taux de couverture par une assurance maladie en utilisant les données de la NHIS et de la MEPS de 2006, et nous comparons le résultat à celui

d'une approche de modélisation paramétrique (log-linéaire). Enfin, nous présentons nos conclusions à la section 8.

# 2 Production de populations synthétiques à partir de données d'enquête

Le concept fondamental de l'inférence bayésienne en population finie consiste à imputer les valeurs non échantillonnées de la population à partir de la loi prédictive a posteriori basée sur les données observées. Supposons que les valeurs de population sont $Y = (Y_1, \ldots, Y_N)$ et que les données observées, $Y_{\text{obs}} = (y_1, \ldots, y_n)$, sont obtenues dans un sondage dont les indicatrices d'échantillonnage sont $I = (I_1, \ldots, I_N)$. L'inférence bayésienne sur la population permet d'utiliser le modèle paramétrique $\Pr(Y \mid \theta)$ pour les données de population basé sur la loi prédictive a posteriori pour les éléments non observés de la population $\Pr(Y_{\text{nob}} \mid Y_{\text{obs}})$ :

$$\Pr(Y_{\text{nob}} \mid Y_{\text{obs}}) = \int \Pr(Y_{\text{nob}} \mid Y_{\text{obs}}, \theta) \Pr(\theta \mid Y_{\text{obs}}) \, d\theta$$

(Ericson, 1969; Little, 1993; Rubin, 1987; Scott, 1977; Skinner et coll., 1989). Ici, nous utilisons le modèle $\Pr(Y \mid \theta)$ pour approximer la distribution de la population complète $\Pr(Y)$ et prenons la moyenne sur la distribution a posteriori basée sur les données d'échantillon $\Pr(\theta \mid Y_{\text{obs}})$. S'il existe des variables de plan de sondage connues pour la population entière, le modèle susmentionné peut être étendu naturellement en conditionnant sur ces variables.

Le fait implicite dans la dérivation susmentionnée est que l'indicatrice d'échantillonnage $I$ ne doit pas être modélisée. Cela exige que l'échantillonnage soit ignorable (Rubin, 1987) (la distribution de $I$ ne doit pas dépendre des données observées), et que le modèle $\Pr(Y \mid \theta)$ utilisé pour les données tienne compte des caractéristiques du plan de sondage et soit suffisamment robuste pour saisir comme il convient tous les aspects pertinents de la distribution de la variable $Y$ d'intérêt. Notre objectif ici est d'élaborer une méthode pour générer des tirages à partir de $\Pr(Y_{\text{nob}} \mid Y_{\text{obs}})$ qui tiennent compte de toutes les caractéristiques du plan de sondage dans $Y_{\text{obs}}$, de manière que les tirages à partir de la loi a posteriori de $Y_{\text{nob}} \mid Y_{\text{obs}}$ puissent être traités comme un échantillon aléatoire simple dans l'analyse.

# 3 Bootstrap bayésien en population finie pondéré

## 3.1 Bootstrap bayésien en population finie (BBPF)

Supposons que les éléments (scalaires) de population $Y_i, i = 1, \ldots, N$ soient échangeables et puissent prendre $K \leq N$ valeurs possibles $(b_1, \ldots, b_K)$; donc, $Y_i \mid \theta \sim \text{MULTI}(1; \theta_1, \ldots, \theta_K)$. Supposons aussi qu'une loi a priori conjuguée de Dirichlet pour $\theta \sim \text{DIR}(\alpha_1, \ldots, \alpha_K)$ donne (Ghosh et Meeden, 1983)

$$
\begin{aligned}
P(Y_{\text{nob}} \mid y) &= P\left(b_1^{\text{nob}} = N_1 - n_1, \ldots, b_K^{\text{nob}} = N_K - n_K \mid b_1^{\text{obs}} = n_1, \ldots, b_K^{\text{obs}} = n_K\right) \\
&= \frac{\int_0^1 \ldots \int_0^1 p(Y_{\text{nob}} \mid y, \theta)\, p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \ldots \int_0^1 p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K} \\
&= \frac{\int_0^1 \ldots \int_0^1 p(Y_{\text{nob}} \mid \theta)\, p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \ldots \int_0^1 p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K} \\
&= \frac{\int_0^1 \ldots \int_0^1 \prod_{i=1}^K \theta_i^{N_i - n_i} \prod_{i=1}^K \theta_i^{n_i} \prod_{i=1}^K \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_K}{\int_0^1 \ldots \int_0^1 \prod_{i=1}^K \theta_i^{n_i} \prod_{i=1}^K \theta_i^{\alpha_i - 1} d\theta_1 \ldots d\theta_K} \\
&= \frac{\left(\prod_{i=1}^K \Gamma(N_i + \alpha_i)/\Gamma(\alpha_i)\right) \Big/ \left(\Gamma(N + \alpha_0)/\Gamma(\alpha_0)\right)}{\prod_{i=1}^K \Gamma(n_i + \alpha_i) \Big/ \Gamma(n + \alpha_0)}
\end{aligned}
\tag{3.1}
$$

où $\alpha_0 = \sum_{i=1}^K \alpha_i$, $\sum_{i=1}^K N_i = N$, et $n_1, \ldots, n_K$ désigne le nombre de valeurs distinctes que nous observons à partir de notre échantillon $y = (y_1, \ldots, y_n)$, $\sum_{i=1}^K n_i = n$. Si $\alpha_i \equiv 0$, alors $p(Y_{\text{nob}} \mid y)$ se réduit à

$$
\left(\prod_{i=1}^K \Gamma(N_i)/\Gamma(n_i)\right) \Big/ \left(\Gamma(N)/\Gamma(n)\right).
$$

Pour faciliter la mise en œuvre, Lo (1988) a proposé de faire des tirages à partir de la loi prédictive a posteriori du BBPF en utilisant une procédure fondée sur le « modèle de l'urne de Pólya». Supposons qu'une urne contient $n$ boules possédant chacune comme étiquette un nombre réel distinct $b_i, i = 1, \ldots, K$. Nous tirons un échantillon de Pólya de taille $m$ en sélectionnant d'abord une boule au hasard dans l'urne et en remettant la boule sélectionnée dans l'urne, puis en plaçant une boule identique dans l'urne et en répétant ce processus jusqu'à ce que $m$ boules aient été sélectionnées. On peut montrer que la probabilité d'obtenir $m_i$ boules de type $b_i$ est donnée par

$$
p(b_1 = m_1, \ldots, b_K = m_K) = \frac{\prod_{i=1}^k \Gamma(n_i + m_i) \Big/ \Gamma(n_i)}{\Gamma(n + m) \Big/ \Gamma(n)} \tag{3.2}
$$

où $n_i$ est le nombre de boules de type $b_i$ se trouvant au départ dans l'urne. La distribution des nombres de boules de type $b_i$ est invariante sous n'importe quelle permutation des tirages. Notons que cela correspond directement à la probabilité a posteriori d'un total de $(m_1, \ldots, m_K)$ éléments de type $(b_1, \ldots, b_K)$ dans une population, sachant que $(n_1, \ldots, n_K)$ éléments ont été observés dans un échantillon (aléatoire simple) de taille $\sum_{i=1}^K n_i = n$. Donc, nous pouvons tirer un échantillon réplique de cette loi a posteriori de Pólya en procédant aux étapes suivantes :

*Étape* 1. Tirer un échantillon de Pólya de taille $m = N - n$, noté $\left(y_1^*, \ldots, y_{N-n}^*\right)$ à partir de l'urne $\{y_1, \ldots, y_n\}$; en vertu de (3.2), avec $m_k = N_k - n_k$ tirages de la valeur $b_k^{\text{obs}}$ pour $k = 1, \ldots, K$, cela correspond à un tirage de $P(Y_{\text{nob}} \mid y)$ à partir de (3.1).

*Étape* 2. Former la population BBPF $y_1,\ldots,y_n,\ y_1^*,\ldots,y_{N-n}^*$.

## 3.2 BBPF avec probabilités de sélection inégales

Cohen (1997) a étendu la procédure du BBPF afin de faire un ajustement pour les probabilités de sélection inégales. Supposons que $(y_1,\ldots,y_n)$ est un échantillon tiré d'une population finie $(Y_1,\ldots,Y_N)$ avec les poids de sondage $(w_1,\ldots,w_n)$, où

$$w_i = \frac{1}{P(I_i = 1)}$$

et $I$ est l'indicatrice d'échantillonnage. La procédure comprend deux étapes :

*Étape* 1. Tirer un échantillon de taille $N-n$, noté $(y_1^*,\ldots,y_{N-n}^*)$, en tirant $y_k^*$ à partir de $(y_1,\ldots,y_n)$ de manière que $y_i$ soit sélectionné avec la probabilité

$$\frac{w_i - 1 + l_{i,k-1}*(N-n)/n}{N-n+(k-1)*(N-n)/n},$$

où $w_i$ est le poids de l'unité $i$ et $l_{i,k-1}$ est le nombre de sélections bootstrap de $y_i$ parmi les $y_1^*,\ldots,y_{k-1}^*$. (La fonction wtpolyap du module R *polypost* peut être utilisée pour obtenir des tirages à partir d'une urne de Pólya pondérée.)

*Étape* 2. Former la population BBPF $y_1,\ldots,y_n,\ y_1^*,\ldots,y_{N-n}^*$.

Cohen (1997) n'a pas fourni la preuve théorique de cette procédure, mais elle peut être obtenue comme une extension simple de l'équivalence du BBPF et de l'urne de Pólya classique décrite à la section 3.1. Premièrement, nous déterminons la loi a posteriori de l'échantillon BBPF avec probabilités de sélection inégales qu'implique la procédure BBPF pondérée. La vraisemblance multinomiale fondée sur notre échantillon pondéré est donnée par

$$p(y_{\text{obs}}\,|\,\theta) = \prod_{i=1}^{K}\theta_i^{w_i^*},$$

où

$$w_i^* = \left(\frac{n}{N-n}\right)\sum_{j=1}^{n} I\left(y_j = b_i\right)\left(w_j - 1\right)$$

est la somme des poids de sondage moins une unité sur l'ensemble des éléments échantillonnés ayant la valeur $b_i, i = 1,\ldots,K$, normalisée pour qu'elle soit égale à $n$. (Soulignons que cela élimine de la vraisemblance les sujets échantillonnés dont le poids est égal à un, c'est-à-dire les éléments de l'« échantillon sélectionné avec certitude », car ils n'ont aucune chance de se trouver dans la partie non observée de la population, et donc n'apportent aucune information au sujet des éléments non observés.) En émettant l'hypothèse d'une loi a priori de Dirichlet impropre $p(\theta) = \prod_{i=1}^{k}\theta_i^{-1}$, la loi a posteriori du bootstrap bayésien en population finie pondéré est donnée par

$$
\begin{aligned}
P(Y_{\text{nob}} \mid y, w) &= P\left(b_1^{\text{nob}} = r_1, \ldots, b_K^{\text{nob}} = r_K \mid w_1^*, \ldots, w_K^*\right) \\
&= \frac{\int_0^1 \ldots \int_0^1 p(Y_{\text{nob}} \mid \theta)\, p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K}{\int_0^1 \cdots \int_0^1 p(y \mid \theta)\, p(\theta)\, d\theta_1 \ldots d\theta_K} \\
&= \frac{\int_0^1 \cdots \int_0^1 \prod_{i=1}^K \theta_i^{r_i} \prod_{i=1}^K \theta_i^{w_i^*} \prod_{i=1}^K \theta_i^{-1} d\theta_1 \cdots d\theta_K}{\int_0^1 \cdots \int_0^1 \prod_{i=1}^K \theta_i^{w_i^*} \prod_{i=1}^K \theta_i^{-1} d\theta_1 \cdots d\theta_K} \\
&= \prod_{i=1}^K \frac{\Gamma\left(w_i^* + r_i\right)}{\Gamma\left(w_i^*\right)} \Big/ \frac{\Gamma(N)}{\Gamma(n)}
\end{aligned}
\tag{3.3}
$$

puisque $\sum_{j=1}^n r_i = N - n$ et $\sum_{j=1}^n w_i^* = n$.

Ensuite, nous montrons que la distribution des échantillons obtenus à partir du modèle d'urne de Pólya sous probabilités inégales de sélection de Cohen (1997) est égale à la loi a posteriori de l'échantillon BBPF avec probabilités de sélection inégales. Sachant les données observées, la probabilité de tirer $N - n$ boules et que les premières boules $r_1$ aient la valeur $b_1$ , et ainsi de suite, et que les dernières, $r_k$ , aient la valeur $b_k$ est :

$$
\begin{aligned}
P(b_1 = r_1, \ldots, b_K = r_K) &= \frac{w_1^*}{n} \times \frac{w_1^* + 1}{n+1} \ldots \times \frac{w_1^* + r_1 - 1}{n + r_1 - 1} \times \ldots \times \frac{w_K^*}{n + \sum_{i=1}^{k-1} r_i} \times \ldots \times \frac{w_K^* + r_K - 1}{n + \sum_{i=1}^{k} r_i - 1} \\
&= \prod_{i=1}^K \frac{\Gamma\left(w_i^* + r_i\right)}{\Gamma\left(w_i^*\right)} \Big/ \frac{\Gamma(N)}{\Gamma(n)}
\end{aligned}
$$

où la première égalité découle du fait que la distribution des nombres de boules de type $b_i$ est invariante sous toute permutation des tirages, comme dans le cas non pondéré, et la deuxième égalité découle de l'identité $\Gamma(x) = (x-1)\Gamma(x)$ pour $x > 0$. Donc, en notant que

$$
\frac{w_i - 1 + l_{i,k-1} * (N-n)/n}{N - n + (k-1) * (N-n)/n} = \frac{w_i^* + l_{i,k-1}}{n + (k-1)},
$$

un tirage à partir du modèle de l'urne de Pólya avec probabilités de sélection inégales donne un tirage à partir de $P(Y_{\text{nob}} \mid y, w)$ dans (3.3).

# 4 Méthode non paramétrique de production de populations synthétiques

À la présente section, nous étendons les méthodes du bootstrap bayésien en population finie au cas d'un plan de sondage en grappes stratifié avec probabilités de sélection inégales en vue d'élaborer une méthode non paramétrique de production de populations synthétiques qui comporte un ajustement pour tenir compte des caractéristiques du plan de sondage complexe. L'idée est de traiter la partie non observée de la population comme des données manquantes et de l'imputer en effectuant des tirages à partir des données réelles. Nous faisons l'imputation de manière que les tirages résultants à partir de la loi

a posteriori de la population reflètent les caractéristiques du plan de sondage complexe et puissent être utilisés de la façon classique pour calculer les lois a posteriori des quantités d'intérêt de la population.

## 4.1 Utilisation du bootstrap bayésien pour l'ajustement pour la stratification et la mise en grappe

Dans le cas d'un échantillonnage en grappes stratifié, nous devons d'abord rééchantillonner les grappes à l'intérieur des strates. Notons $c$ le nombre total de grappes dans les données réelles, $c = \sum_{h=1}^{H} c_h$, et $C$ le nombre de grappes dans la population, $C = \sum_{h=1}^{H} C_h$. Une approche consiste à appliquer d'abord le modèle de l'urne de Pólya avec le BBPF pour imputer les grappes non observées à l'intérieur de chaque strate, $c_1^*, \ldots, c_{C_h - c_h}^*$, qui, avec les grappes observées, fournissent les grappes dans la strate $h$ de la population. Cependant, les données à grande diffusion disponibles ne nous permettent habituellement pas de savoir quel est le nombre de grappes dans une strate. Donc, comme alternative au tirage d'un échantillon BBPF, nous proposons le tirage d'un échantillon bootstrap bayésien classique de grappes dans chaque strate. En tenant compte de l'équivalence entre le bootstrap classique et le bootstrap bayésien, nous procédons comme Rao et Wu (1988), qui ont proposé de tirer un échantillon aléatoire simple avec remise (EASAR) de taille $m_h$ à partir des $c_h$ grappes et, à l'intérieur de chaque strate $h$, de calculer les poids de rééchantillonnage pour chaque échantillon bootstrap comme

$$w^{*(l)} = \left\{ w_{hik}^{*(l)},\ h = 1, \ldots,\ H,\ i = 1, \ldots,\ c_h,\ k = 1, \ldots,\ N_{hi} \right\},$$

où

$$w_{hik}^{*} = w_{hik} \left( \left( 1 - \sqrt{\frac{m_h}{c_h - 1}} \right) + \sqrt{\frac{m_h}{c_h - 1}} \frac{c_h}{m_h}\ m_{hi}^{*} \right)$$

et $m_{hi}^{*}$ désigne le nombre de fois que la grappe $i, i = 1, \ldots, c_h$ est sélectionnée. Pour être certain que tous les poids de rééchantillonnage soient non négatifs, il faut que $m_h \leq (c_h - 1)$; ici et plus loin, nous prenons $m_h = (c_h - 1)$.

Notons qu'en l'absence de grappes, nous tirons simplement un échantillon bootstrap bayésien classique à partir des données échantillonnées dans chaque strate (s'il existe une stratification) ou à partir de l'échantillon complet (en l'absence de stratification, de sorte que $H = 1$) et nous calculons les poids de rééchantillonnage comme étant $w_{hik}^{*} = w_{hik} m_{hi}^{*}$.

Nous répétons cette procédure $L$ fois pour produire $L$ échantillons bootstrap bayésiens (BB) notés $S_1, \ldots, S_L$. Cette étape génère $L$ échantillons bootstrap bayésiens qui sont essentiellement $L$ tirages à partir de la loi prédictive a posteriori des grappes non observées sachant les données réelles. Cependant, les unités formant les $L$ échantillons bootstrap bayésiens possèdent encore des poids et ne peuvent être analysées comme s'il s'agissait d'échantillons aléatoires simples.

## 4.2 Utilisation du modèle de l'urne de Pólya avec le BBPF pondéré pour faire un ajustement pour la pondération

Une fois que nous avons obtenu $L$ échantillons BB avec les poids de rééchantillonnage, la deuxième étape consiste à imputer les unités non observées en utilisant le modèle de l'urne de Pólya avec le BBPF

pondéré. En pratique, la probabilité de sélection de la $k^{e}$ unité, $y_k^*$, dépend de la sélection des $k-1$ premières unités, $y_1^*,\ldots,y_{k-1}^*$. Autrement dit, pour déterminer la probabilité de sélection d'une nouvelle unité, nous devons compter le nombre de fois que chaque unité présente dans l'échantillon a été sélectionnée parmi les sélections antérieures. Si la taille de la population est très grande, il nous suffit de générer des populations synthétiques de taille $T * n$, où $T$ est suffisamment grand pour dominer la taille d'échantillon (p. ex. 20-100). Pour accroître encore davantage l'efficacité des calculs, nous pourrions aussi tirer $F > 1$ fois une population de taille modérée, puis regrouper ces $F$ populations pour produire une population synthétique, $S_l$. La taille de $S_l$ est alors $F * T * n$.

Notons que, sous notre méthode, il ne faut connaître que les poids finaux dans les échantillons en grappes à plusieurs degrés, puisque tous les degrés d'échantillonnage avec probabilités de sélection inégales sont corrigés par l'utilisation du modèle de l'urne de Pólya avec le BBPF pondéré. Cette caractéristique de la méthode proposée est particulièrement utile, car dans de nombreux jeux de données à grande diffusion, les composantes des probabilités de sélection (p. ex. probabilités de sélection au niveau de la grappe, poids de non-réponse) ne sont pas disponibles.

# 5 Inférence à partir de multiples populations synthétiques non paramétriques

Supposons que nous produisons $L$ populations synthétiques, $S_l, l = 1,\ldots,L$ en utilisant la méthode non paramétrique décrite à la section 4, et que notre cible d'inférence est $Q \equiv Q(Y)$, une fonction des données de population (p. ex. moyenne de population, corrélation, estimateur du maximum de vraisemblance de population d'un paramètre de régression). Nous pouvons calculer $Q_l$ comme étant l'estimation de $Q$ obtenue en regroupant les $F$ populations synthétiques utilisées pour imputer les unités non observées de $S_l$; puisqu'il s'agit de tirages directs à partir de la loi prédictive a posteriori de la population, nous pouvons calculer les moyennes, les quantiles et les intervalles de crédibilité a posteriori d'après les estimations empiriques correspondantes à partir des tirages, si $L$ est suffisamment grand.

Cependant, dans de nombreuses situations, le temps de calcul nécessaire pour imputer la population peut être très important, même s'il ne faut pas synthétiser la population complète. D'où, une autre approche de l'inférence consiste à utiliser la loi $t$ comme approximation de la distribution prédictive a posteriori d'une statistique de population scalaire $Q$ :

$$Q \mid S_1,\ldots,S_L \overset{\cdot}{\sim} t_{L-1}\left(\bar{Q}_L,\left(1+L^{-1}\right)V_L\right)$$

où

$$\bar{Q}_L = \frac{\sum_{l=1}^{L} Q_l}{L} = \frac{\sum_{l=1}^{L}\sum_{f=1}^{F} Q_{lf}}{LF} \quad \text{et} \quad V_L = \frac{1}{L}\sum_{l=1}^{L}(Q_l - \bar{Q}_L)^2 .$$

Le résultat découle directement de la section 4.1 dans Raghunathan et coll. 2003, et est fondé sur les règles de combinaison classiques pour l'imputation multiple de Rubin (1987), en traitant les unités non observées de $S_l$ comme des données manquantes et les unités échantillonnées, comme des données observées. La variance « intra-imputation » moyenne est nulle, puisque la population est entièrement synthétisée; d'où,

la variance a posteriori de $Q$ est entièrement une fonction de la variance inter-imputations, et le nombre de degrés de liberté est simplement donné par le nombre d'échantillons BBPF. (Lorsque la population est très grande, il nous suffit de synthétiser un tirage assez grand pour que la variance « intra-imputation » moyenne soit négligeable comparativement à la variance inter-imputations $V_L$.) Le résultat suppose que $E(Q_{lf}) = Q$, ce que notre estimateur BBPF pondéré garantit, et que la taille de l'échantillon est suffisamment grande pour permettre d'appliquer la théorie asymptotique bayésienne.

# 6 Études par simulation

À la présente section, nous décrivons deux études par simulation réalisées pour évaluer les propriétés de rééchantillonnage des estimateurs de population construits en utilisant la méthode non paramétrique qui produit des populations synthétiques en effectuant un ajustement pour tenir compte des caractéristiques du plan de sondage complexe. La première simulation porte sur un plan de sondage à un degré avec probabilités de sélection inégales dans lequel nous faisons varier le nombre de tirages BBPF pondérés pour chaque population synthétique, ainsi que le nombre de populations synthétiques pour évaluer l'effet sur l'inférence. La deuxième simulation a pour objectif de comparer les propriétés inférentielles à partir des données observées et à partir de la loi a posteriori obtenue d'après la population synthétique sous un plan de sondage stratifié à plusieurs degrés avec probabilités de sélection inégales, cette fois-ci en fixant la taille de l'échantillon a posteriori, tout en considérant la moyenne de population ainsi que les paramètres de régression de population comme les cibles des inférences.

## 6.1 Plan de sondage à un degré avec probabilités de sélection inégales

Nous avons généré les données pour la variable de résultat $Y$ dans une population de $N$ sujets provenant d'une loi Gamma modérément asymétrique, conditionnellement à la covariable $X$ qui suit une loi uniforme :

$$X_i \sim \text{UNI}(0,05; 0,65), i = 1, \ldots, N$$

$$Y_i \mid X_i = x_i \sim \text{GAMMA}(10 * x_i, 1)$$

Nous supposons que $X$ est entièrement observé pour la population, et que la probabilité de sélection $\pi$ est proportionnelle à $X$, de sorte que $\pi_i \doteq n x_i / \sum_i x_i$ dans un plan de sondage sans remise à condition que $n << N$. La quantité à estimer est la moyenne de population $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i = 3{,}564$. Notons que $\text{corr}(Y_i, X_i) = 0{,}6794$, de sorte que les moyennes d'échantillon non pondérées présentent un biais positif, et que l'utilisation des poids de sondage $w_i = 1/\pi_i$ est nécessaire pour obtenir des estimations sans biais de $\bar{Y}$. Nous générons une population de taille $N = 1\,000$ à partir de laquelle nous tirons $n = 100$ échantillons; nous estimons ensuite le biais, la variance empirique et la variance estimée, la longueur de l'intervalle de confiance à 95 % et la couverture au niveau de confiance nominal de 95 % au moyen de 200 échantillons indépendants tirés de la population. Nous faisons varier le nombre total de populations simulées $L$ qui prend les valeurs de 5, 20, 100 et 1 000, ainsi que le nombre $F$ de tirages BBPF de taille $N - n$ (de manière que $K = 9$) qui prend les valeurs de 1, 20 et 100, dans un plan factoriel complet. Nous obtenons la variance, la longueur de l'intervalle et la couverture de l'intervalle au moyen de l'approximation normale; pour $L = 100$ et 1 000, nous obtenons également la variance, la longueur de

l'intervalle et la couverture de l'intervalle en utilisant les tirages directs à partir de la loi prédictive a posteriori, puisque nous disposons d'un nombre suffisant de tirages à partir de cette loi pour produire ces estimations.

Le tableau 6.1 donne les résultats de l'étude par simulation. Dans tous les cas, l'estimation ponctuelle $\bar{Q}_L$ de la moyenne de population est approximativement sans biais, ce qui témoigne de la capacité du BBPF pondéré à « défaire » les poids de sondage pour produire la population synthétique. Sous l'approximation normale, l'augmentation du nombre de populations synthétiques est associée à de plus petites variances et des intervalles plus étroits, comme il fallait s'y attendre sous un plus grand nombre de degrés de liberté, quoique la différence entre les résultats obtenus pour 20 et 100 populations soit minime, juste quand la loi $t_{20}$ commence à s'approcher d'une loi normale standard. Enfin, l'utilisation d'un seul tirage BBPF de taille $N-n$ semble donner lieu à une surestimation de la variance et à un surdénombrement, surtout pour les petites valeurs de $L$. Les valeurs de $L$ et de $F$ égales ou supérieures à 20 semblent donner des résultats raisonnables. L'utilisation des tirages directs pour $L=100$ et 1 000 produit des estimations de variance et d'intervalle de crédibilité qui sont fort semblables à celles données par l'approximation normale, les longueurs d'intervalle étant toutefois légèrement plus courtes et les couvertures un peu moins conservatrices.

**Tableau 6.1**
**Biais, variance empirique, moyenne de la variance estimée, longueur de l'intervalle et couverture de l'intervalle de confiance au niveau nominal de 95 % d'une moyenne de population en fonction du nombre de populations synthétiques $(L)$ et du nombre de tirages par bootstrap bayésien en population finie pondéré qui constituent la population synthétique $(F)$. Longueur et couverture de l'intervalle obtenues par approximation par la loi $t$ et empiriquement par simulation directe. Plan de sondage à un degré avec probabilités de sélection inégales. Résultats pour 200 simulations.**

| L | 5 | | | 20 | | | 100 | | | 1 000 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F | 1 | 20 | 100 | 1 | 20 | 100 | 1 | 20 | 100 | 1 | 20 | 100 |
| Biais | -0,020 | 0,009 | -0,026 | 0,021 | -0,030 | 0,010 | -0,031 | 0,024 | -0,028 | -0,045 | -0,070 | 0,079 |
| Variance emp. | 0,126 | 0,099 | 0,106 | 0,088 | 0,092 | 0,120 | 0,093 | 0,079 | 0,085 | 0,084 | 0,093 | 0,078 |
| Variance est. : $t$ | 0,172 | 0,119 | 0,105 | 0,156 | 0,098 | 0,099 | 0,109 | 0,097 | 0,095 | 0,147 | 0,104 | 0,094 |
| Longueur de l'intervalle : $t$ | 2,20 | 1,78 | 1,71 | 1,63 | 1,30 | 1,32 | 1,52 | 1,21 | 1,20 | 1,50 | 1,26 | 1,20 |
| Couverture IC à 95 % : $t$ | 97 | 95 | 96 | 99 | 94 | 92 | 98 | 96 | 95 | 98 | 96 | 98 |
| Variance est. : Empirique | 0,138 | 0,095 | 0,084 | 0,148 | 0,093 | 0,094 | 0,108 | 0,096 | 0,094 | 0,084 | 0,093 | 0,078 |
| Longueur de l'intervalle : Empirique | s.o. | s.o. | s.o. | s.o. | s.o. | s.o. | 1,50 | 1,19 | 1,18 | 1,49 | 1,25 | 1,19 |
| Couverture IC à 95 % : Empirique | s.o. | s.o. | s.o. | s.o. | s.o. | s.o. | 96 | 93 | 94 | 98 | 96 | 97 |

## 6.2 Plan de sondage stratifié à plusieurs degrés avec probabilités de sélection inégales

Nous avons généré une population comprenant des strates et des grappes dans chaque strate à partir de la loi normale bivariée suivante :

$$\begin{pmatrix} X_{1ijk} \\ X_{2ijk} \end{pmatrix} \sim N\left(\begin{pmatrix} 500 + 4,5*i + u_{ij} \\ 500 + 4,5*i + u_{ij} \end{pmatrix}, \begin{pmatrix} 100 & 50 \\ 50 & 100 \end{pmatrix}\right),$$

où

$i = 1 / 150$ désigne l'effet de strate;

$u_{ij} \sim N(0, 10)$ désigne l'effet de grappe aléatoire;

$a_i \sim \text{uniforme}(2, 52)$ désigne le nombre de grappes dans la strate $i$;

$b_{ij} \sim \text{uniforme}(10, 20)$ désigne le nombre d'unités dans la grappe $j$ de la strate $i$.

La population utilisée pour l'étude par simulation compte 61 324 sujets. Nous tirons un échantillon en grappes stratifié avec probabilités de sélection inégales. Plus précisément, nous sélectionnons deux grappes dans chaque strate avec probabilités proportionnelles à la taille de grappe (PPT) données par $b_{i\bullet} = \sum_{j=1}^{a_i} b_{ij}$. Dans chaque grappe sélectionnée, nous sélectionnons environ un cinquième $(1/5)$ de la population. Donc, la probabilité que l'unité $ij$ soit sélectionnée est donnée par

$$\pi_{ij} = \frac{2b_{i\bullet}}{\sum_{j=1}^{a_i} b_{ij}} \times \frac{\lfloor b_{ij}/5 \rfloor}{b_{ij}}$$

pour tout élément $j$ dans la grappe $i$ avec les poids correspondants

$$w_{ij} = \frac{b_{ij} \sum_{j=1}^{a_i} b_{ij}}{2b_{i\bullet} \lfloor b_{ij}/5 \rfloor}.$$

Puisque les nombres de grappes et d'unités sont aléatoires, la taille de l'échantillon complexe diffère légèrement d'une réplique à l'autre, la moyenne étant approximativement de 770.

Comme l'échantillon et la population sont de grande taille, nous nous concentrons sur l'inférence en utilisant les approximations $t$. Nous générons $L = 100$ populations synthétiques en utilisant $F$ échantillons BBPF pondérés de taille $K = 100n$. Les quantités à estimer sont la moyenne marginale de population pour $x_1$

$$\bar{X}_1 = N^{-1} \sum_{i=1}^{N} X_{1i}$$

et celle pour $x_2$, obtenue de manière similaire, ainsi que les coefficients de régression de $x_1$ sur $x_2$, donnés par

$$B_0 = \bar{X}_1 - B_1 \bar{X}_2, B_1 = \frac{\sum_{i=1}^{N} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{N} (X_{2i} - \bar{X}_2)^2}.$$

Nous avons tiré 200 échantillons indépendants de la population et utilisé les données d'échantillon pour calculer directement les moyennes et les coefficients de régression linéaire pour l'échantillon pondéré, ainsi que les estimations correspondantes des variances et des intervalles de confiance au niveau nominal de 95 % en utilisant des approximations par développement en série de Taylor, et les avons comparées aux estimations équivalentes obtenues en utilisant les données synthétiques non paramétriques. Les résultats sont présentés au tableau 6.2. (Puisque les moyennes marginales ont la même valeur de superpopulation, nous combinons les résultats dans le tableau 6.2.) La figure 6.1 donne le diagramme de dispersion des paires de moyennes, d'ordonnées à l'origine et de pentes estimées d'après les échantillons réels et les

populations synthétiques correspondantes, ainsi qu'une droite à 45 degrés. Les distributions d'échantillonnage des estimations d'après les échantillons réels et les populations synthétiques sont très proches. Les estimations ponctuelles et les erreurs-types des moyennes ainsi que des paramètres de régression sont très proches. Les taux de couverture des intervalles de confiance à 95 % sont très proches pour les trois statistiques, et sont proches des valeurs nominales.

**Tableau 6.2**
**Statistiques descriptives et analytiques estimées d'après les données réelles et les populations synthétiques dans une évaluation par simulation de la méthode non paramétrique. Plan de sondage à deux degrés stratifié avec probabilités de sélection inégales. Résultats d'après 200 simulations.**

| Type | Données réelles | | | | Populations synthétiques | | | |
|---|---|---|---|---|---|---|---|---|
| | Estimation | e.-t. | É.-T. | Couverture (%) | Estimation | e.-t. | É.-T. | Couverture (%) |
| Moyenne $\bar{X}$ | 836,701 | 0,461 | 0,491 | 93 | 836,793 | 0,476 | 0,493 | 94 |
| Ordonnée à l'origine $B_0$ | 1,013 | 1,768 | 1,848 | 94 | 1,014 | 1,775 | 1,846 | 92 |
| Pente $B_1$ | 0,999 | 0,002 | 0,002 | 92 | 0,999 | 0,002 | 0,002 | 92 |

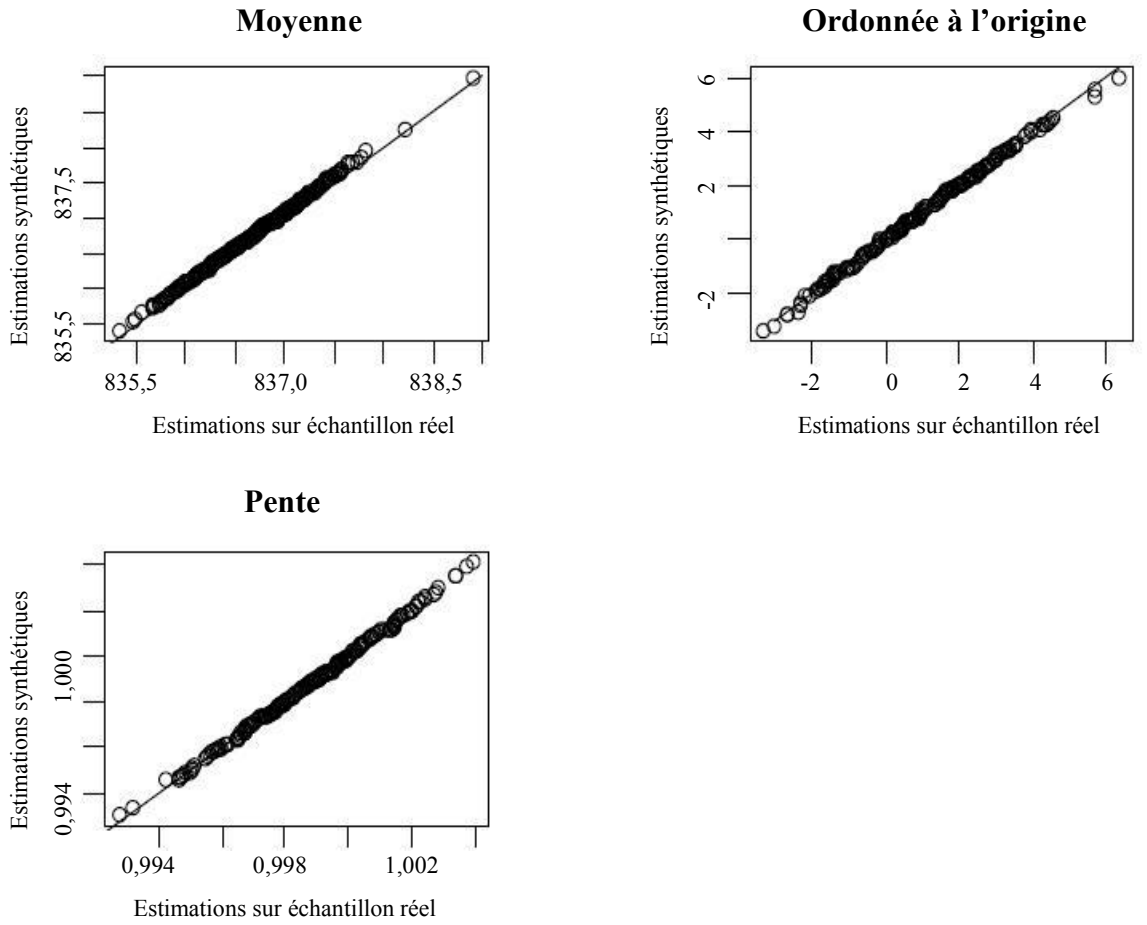


**Figure 6.1 Diagramme de dispersion des statistiques descriptives et analytiques pour les populations réelles et synthétiques**

# 7 Application

À la présente section, nous utilisons des données tirées de la National Health Interview Survey (NHIS) de 2006 et de la Medical Expenditure Panel Survey (MEPS) de 2006 pour évaluer la performance de la méthode non paramétrique sous un plan de sondage en grappes stratifié. La National Health Interview Survey (NHIS) est une enquête sur la santé de portée nationale, réalisée par interview en personne selon un plan stratifié à plusieurs degrés avec suréchantillonnage des populations noires, hispaniques et âgées. Pour des raisons de confidentialité, la stratification et les variables au niveau de l'unité d'échantillonnage (UPE) réelles ne sont pas communiquées dans les fichiers de données à grande diffusion; elles sont remplacées par des pseudo-strates et UPE (deux par strate). La MEPS est réalisée auprès d'un sous-échantillon de l'échantillon de la NHIS de l'année précédente, selon le même plan stratifié à plusieurs degrés.

Tant dans la NHIS que dans la MEPS, on demande aux participants à l'enquête s'ils sont couverts par une assurance maladie et, dans l'affirmative, quel régime d'assurance maladie ils utilisent (privé par opposition à public tel que Medicare ou Medicaid). Nous estimons les taux globaux de couverture par une assurance maladie, ainsi que les taux de couverture dans des sous-populations définies en fonction de variables démographiques telles que le sexe, la race, le niveau de revenu ou des combinaisons de ces variables; en particulier, nous estimons la couverture par une assurance maladie des hommes, des Blancs non hispaniques et des Blancs non hispaniques dont le revenu du ménage est compris entre 25 000 $ et 35 000 $ par année. Nous supprimons les cas pour lesquels les valeurs manquent pour certaines questions et nous axons notre simulation sur les cas complets. Nous obtenons ainsi 20 147 et 20 893 cas pour les données de la NHIS et de la MEPS, respectivement.

## 7.1 Estimation de la couverture par une assurance maladie d'après la NHIS et la MEPS

Dans la présente étude par simulation, nous utilisons la méthode non paramétrique pour apporter un ajustement pour tenir compte de l'échantillonnage en grappes stratifié utilisé dans la NHIS et la MEPS de 2006, et pour produire des populations synthétiques qui peuvent être analysées comme des échantillons aléatoires simples. Nous considérons également une approche fondée sur un modèle pour produire des populations synthétiques en utilisant un modèle log-linéaire pour la situation de couverture par une assurance maladie en fonction de six variables démographiques indépendantes : sexe, race, région de recensement, niveau de scolarité, âge (catégorique) et revenu du ménage (catégorique). Ensuite, nous évaluons la méthode en comparant les estimations du taux de couverture par une assurance maladie pour l'ensemble de la population et pour les sous-domaines choisis pour les populations synthétiques obtenues par la méthode non paramétrique et par celle du modèle log-linéaire à celles obtenues au moyen des données réelles.

### 7.1.1 Production de populations synthétiques non paramétriques

En utilisant la méthode non paramétrique élaborée à la section 3, nous produisons 200 populations synthétiques pour chaque enquête. Plus précisément, nous générons $B = 200$ échantillons BB et, pour chacun de ces échantillons, nous générons $F = 10$ échantillons BBPF de taille $5n\,(K = 5)$. Donc, chaque

population synthétique est 50 fois plus grande que l'échantillon réel (1 007 350 pour la NHIS, 1 044 650 pour la MEPS). Chaque population synthétique est analysée comme un échantillon aléatoire simple et les estimations sont combinées comme il est décrit à la section 5.

### 7.1.2 Production de populations synthétiques au moyen de modèles log-linéaires

Dans la situation fréquente où les données d'enquête d'intérêt prennent la forme d'un tableau de contingence multidimensionnel, un modèle log-linéaire pourrait être considéré comme une approche paramétrique pour générer des tirages à partir d'une loi prédictive a posteriori. Pour simplifier l'exposé, supposons que $Y$ est la variable d'intérêt comprenant $m$ niveaux, et que $Z$ est une variable de plan comprenant $n$ niveaux (p. ex. sexe ou race) dont la loi de probabilité marginale est connue pour la population. Supposons que $\pi_{ij}, i = 1,\ldots,m, j = 1,\ldots,n,$ représente la proportion dans la $ij^{\text{e}}$ cellule, $\sum_{i=1}^{m}\sum_{j=1}^{n}\pi_{ij} = 1.$ Un modèle log-linéaire entièrement saturé est donné par (Agresti, 2002) :

$$\log\left(\pi_{ij}\right) = \lambda_0 + \lambda_i^Z + \lambda_j^Y + \lambda_{ij}^{ZY},\ i = 1,\ldots,m, j = 1,\ldots,n,$$

où $\log\left(\pi_{ij}\right)$ est le logarithme de la probabilité qu'une observation se trouve dans la cellule $ij$ du tableau de contingence, $\lambda_i^Z$ est l'effet principal pour $Z, \lambda_j^Y$ est l'effet principal pour $Y$ et $\lambda_{ij}^{ZY}$ est l'effet d'interaction pour $Z$ et $Y$. Ce modèle comprend tous les effets unidimensionnels et bidimensionnels possibles, et est donc saturé, car il contient le même nombre d'effets que de cellules dans le tableau de contingence. Pour éviter de surajuster les données dans l'exemple, nous pouvons considérer des modèles non saturés dont sont exclus certains termes d'interaction, voire tous, en choisissant le modèle en nous basant sur des tests de rapport de vraisemblance, ou sur le critère AIC ou BIC.

Les populations synthétiques peuvent être générées à partir de la distribution prédictive a posteriori issue du modèle. Toutefois, si les données sont recueillies selon un plan de sondage complexe, nous ne connaissons aucun logiciel statistique standard capable de produire à la fois l'estimation ponctuelle et l'estimation de covariance des coefficients de régression. Nous avons donc choisi d'utiliser une méthode de rééchantillonnage jackknife pour tenir compte de la stratification, de la mise en grappe et de la pondération. Plus précisément, les populations synthétiques paramétriques peuvent être générées selon les étapes suivantes :

1. Estimer les coefficients et la matrice de covariance :

Sous le modèle choisi (supposé être le modèle saturé bidimensionnel ici, simplement pour l'illustration), estimer les coefficients $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1,\ldots,m-1, j = 1,\ldots,n-1$ et la matrice de covariance des estimations $\hat{\lambda} = \left(\hat{\lambda}_0, \hat{\lambda}_i^Z, \hat{\lambda}_j^Y, \hat{\lambda}_{ij}^{ZY}\right)'$ après avoir tenu compte des caractéristiques du plan complexe en utilisant la méthode des répliques équilibrées jackknife (REJ) :

- Pour chaque réplique, retirer une grappe et augmenter les poids de sondage des unités des autres grappes à l'intérieur de la même strate d'un facteur $c_h/(c_h - 1)$ (poids de rééchantillonnage), où $c_h$ désigne le nombre de grappes dans la strate $h$. En supposant que nous avons un total de

$\sum_{h=1}^{H} c_h = C$ grappes, nous avons alors $C$ répliques. Pour chaque réplique, nous ajustons le modèle log-linéaire et obtenons les estimations du maximum de vraisemblance (EMV) des coefficients $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1,\ \ j = 1, \ldots, n-1$.

- Pour chaque réplique, utiliser les poids de rééchantillonnage pour ajuster le modèle log-linéaire. Plus précisément, utiliser les poids de rééchantillonnage pour calculer la taille de chaque cellule du tableau de contingence, qui est utilisé pour ajuster le modèle log-linéaire. Nous notons l'EMV pour la $r^{\text{e}}$ réplique comme un vecteur colonne, $\hat{\lambda}_r, r = 1, \ldots, c_h$ pour la strate $h$. Soulignons que $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1,\ \ j = 1, \ldots, n-1$ est un vecteur colonne de dimension $mn$ par 1. Nous le notons $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)' = (\lambda_0, \lambda_1, \ldots, \lambda_{mn})'$. De même, $\hat{\lambda}_r, r = 1, \ldots, c_h, h = 1, \ldots, H$ sont aussi des vecteurs colonnes de dimensions $mn$ par 1 que nous notons $\left(\hat{\lambda}_0^{(r)}, \hat{\lambda}_1^{(r)}, \ldots, \hat{\lambda}_{mn}^{(r)}\right)'$.

L'EMV des coefficients $\lambda = \left(\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY}\right)'$, $i = 1, \ldots, m-1, j = 1, \ldots, n-1$ peut être obtenu comme $\hat{\lambda}_{\text{EMV}} = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_r / C$. Pour la matrice de covariance de dimensions $mn$ par $mn$, l'estimation par rééchantillonnage jackknife du $pq^{\text{e}}$ $(p, q = 1, \ldots, mn)$ élément est la covariance entre les $p^{\text{e}}$ et $q^{\text{e}}$ coefficients, qui est donnée par :

$$\sum_{h=1}^{H} \frac{c_h - 1}{c_h} \sum_{r=1}^{c_h} \left(\hat{\lambda}_p^{(r)} - \bar{\hat{\lambda}}_p\right)\left(\hat{\lambda}_q^{(r)} - \bar{\hat{\lambda}}_q\right),$$

où $\bar{\hat{\lambda}}_p = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_p^{(r)} / C$ et $\bar{\hat{\lambda}}_q = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \hat{\lambda}_q^{(r)} / C$. Cela nous donne l'estimation de variance correcte de $\hat{\lambda}_{\text{EMV}}$.

2. Obtenir une approximation de la loi a posteriori des coefficients :

Soit $T$ la décomposition de Cholesky telle que $TT^t = \text{cov}\left(\hat{\lambda}_{\text{EMV}}\right)$. Générer un vecteur $z$ de variables aléatoires normales standardisées et définir $\Lambda_* = \hat{\lambda}_{\text{EMV}} + Tz$.

3. Imputer les valeurs non observées de la population :

Supposons que l'on procède à $L$ tirages, $\Lambda_1, \ldots, \Lambda_L$, à partir de la loi a posteriori approximative de $\lambda$. Pour chaque

$$l = 1, \ldots, L, \Lambda_l = \left(\Lambda_0^{(l)}, \Lambda_i^{X(l)}, \Lambda_j^{Y(l)}, \Lambda_{ij}^{XY(l)}\right)',\ i = 1, \ldots, m-1, j = 1, \ldots, n-1,$$

nous pouvons générer un tableau synthétique en utilisant le modèle supposé :

$$\log\left(\pi_{ij}^{(l)}\right) = \Lambda_0^{(l)} + \Lambda_i^{X(l)} + \Lambda_j^{Y(l)} + \Lambda_{ij}^{XY(l)}, i = 1, \ldots, m-1, j = 1, \ldots, n-1.$$

Une fois que les proportions sont déterminées pour chaque cellule, nous pouvons générer un tableau synthétique de n'importe quelle taille.

Les résultats qui suivent sont fondés sur un tableau de contingence à sept dimensions (voir le tableau 7.1 pour les catégories particulières de covariables). Les mesures du BIC indiquent qu'un modèle contenant toutes les interactions bidimensionnelles mais ne contenant aucune interaction tridimensionnelle est celui qui donne l'ajustement le plus parcimonieux.

**Tableau 7.1**
**Variables et catégories de réponse de la NHIS et de la MEPS de 2006 utilisées dans le modèle log-linéaire**

| Variables d'intérêt | Catégories de réponse |
|---|---|
| Âge | 1 : [18; 24]; 2 : [25; 34]; 3 : [35; 44]; 4 : [45; 54]; 5 : [55; 64]; 6 : >= 65 |
| Région de recensement | 1 : Nord-Est; 2 : Mid-Ouest; 3 : Sud; 4 : Ouest |
| Scolarité | 1 : Études secondaires partielles; 2 : Diplôme d'études secondaires; 3 : Études collégiales partielles; 4 : Diplôme d'études collégiales |
| Sexe | 1 : Masculin; 2 : Féminin |
| Couverture par une assurance maladie | 1 : N'importe quel régime privé; 2 : Régime public; 3 : Non assuré |
| Revenu | 1 : (0; 10 000); 2 : [10 000; 15 000); 3 : [15 000; 20 000); 4 : [20 000; 25 000); 5 : [25 000; 35 000); 6 : [35 000; 75 000); 7 : >= 75 000 |
| Race | 1 : Hispanique; 2 : Blanche non hispanique; 3 : Noire non hispanique; 4 : Tous les autres groupes non hispaniques confondus |

## 7.2 Résultats

Les résultats sont résumés au tableau 7.2. Pour la population totale et les sous-populations les plus grandes, nous voyons que les estimations ponctuelles (moyenne a posteriori) des taux de couverture par une assurance médicale sont les mêmes sous les approches non paramétrique et log-linéaire, et qu'elles sont presque identiques à celles obtenues au moyen des données réelles après avoir tenu compte des caractéristiques du plan de sondage complexe. Les deux méthodes donnent des populations synthétiques dont les variances (a posteriori) sont légèrement plus élevées que dans le cas des données réelles, ce qui reflète la perte d'information dans la synthèse. Dans le cas de la NHIS, la perte pour l'estimateur non paramétrique est égale, en moyenne, à un peu plus de 20 % et est légèrement supérieure à celle observée pour le modèle log-linéaire, pour lequel la perte est, en moyenne, de l'ordre de 10 %. Dans le cas de la MEPS, les estimateurs affichent tous deux une perte d'environ 10 % par rapport aux données réelles. Cependant, pour les sous-populations plus petites (Blancs non hispaniques gagnant de 25 000 $ à 35 000 $ par année), le modèle log-linéaire produit des résultats biaisés, dus au fait que le modèle log-linéaire ne contient pas toutes les interactions possibles. La méthode non paramétrique produit des estimations presque identiques à celles obtenues au moyen des données réelles après avoir tenu compte des caractéristiques du plan de sondage complexe. Le modèle log-linéaire donne également lieu à une sous-estimation importante, de l'ordre de 30 % à 40 %, de la variance de la couverture par une assurance médicale pour ces sous-populations, par opposition à une surestimation de l'ordre de 10 % à 40 % dans le cas de l'approche non paramétrique.

**Tableau 7.2**
**Estimations d'après les données réelles et d'après les populations synthétiques (modèles non paramétrique et log-linéaire) pour la NHIS et la MEPS de 2006**

| | | Données réelles (plan complexe) | | Populations synthétiques | | | |
|---|---|---|---|---|---|---|---|
| | | | | Non paramétrique | | Modèle log-linéaire | |
| **Domaine** | **Type** | **NHIS** | **MEPS** | **NHIS** | **MEPS** | **NHIS** | **MEPS** |
| Population complète | | | | Proportion | | | |
| | Régime privé | 0,746 | 0,735 | 0,746 | 0,736 | 0,746 | 0,734 |
| | Régime public | 0,075 | 0,133 | 0,075 | 0,132 | 0,076 | 0,133 |
| | Non assuré | 0,179 | 0,132 | 0,179 | 0,132 | 0,178 | 0,132 |
| | | | | Variance | | | |
| | Régime privé | 2,46E-05 | 2,78E-05 | 3,15E-05 | 3,31E-05 | 2,66E-05 | 2,86E-05 |
| | Régime public | 6,29E-06 | 1,44E-05 | 8,06E-06 | 1,59E-05 | 7,99E-06 | 1,77E-05 |
| | Non assuré | 1,84E-05 | 1,41E-05 | 2,29E-05 | 1,71E-05 | 1,81E-05 | 1,56E-05 |
| Hommes | | | | Proportion | | | |
| | Régime privé | 0,740 | 0,735 | 0,740 | 0,736 | 0,740 | 0,735 |
| | Régime public | 0,060 | 0,101 | 0,060 | 0,100 | 0,060 | 0,102 |
| | Non assuré | 0,200 | 0,164 | 0,200 | 0,164 | 0,200 | 0,164 |
| | | | | Variance | | | |
| | Régime privé | 3,32E-05 | 3,87E-05 | 3,93E-05 | 4,31E-05 | 3,70E-05 | 3,52E-05 |
| | Régime public | 6,82E-06 | 1,53E-05 | 8,81E-06 | 1,63E-05 | 7,91E-06 | 1,91E-05 |
| | Non assuré | 2,94E-05 | 2,64E-05 | 3,29E-05 | 2,79E-05 | 3,19E-05 | 2,56E-05 |
| Race blanche non hispanique | | | | Proportion | | | |
| | Régime privé | 0,805 | 0,788 | 0,804 | 0,788 | 0,804 | 0,788 |
| | Régime public | 0,062 | 0,116 | 0,062 | 0,116 | 0,062 | 0,117 |
| | Non assuré | 0,134 | 0,096 | 0,134 | 0,096 | 0,134 | 0,096 |
| | | | | Variance | | | |
| | Régime privé | 2,99E-05 | 3,35E-05 | 3,79E-05 | 4,12E-05 | 3,07E-05 | 3,98E-05 |
| | Régime public | 8,20E-06 | 1,81E-05 | 1,04E-05 | 2,00E-05 | 1,10E-05 | 2,45E-05 |
| | Non assuré | 2,02E-05 | 1,51E-05 | 2,35E-05 | 1,80E-05 | 1,82E-05 | 1,82E-05 |
| Race blanche non hispanique et revenu [25 000 $; 35 000 $) | | | | Proportion | | | |
| | Régime privé | 0,827 | 0,813 | 0,827 | 0,814 | 0,840 | 0,838 |
| | Régime public | 0,039 | 0,079 | 0,039 | 0,079 | 0,037 | 0,067 |
| | Non assuré | 0,134 | 0,108 | 0,134 | 0,107 | 0,122 | 0,096 |
| | | | | Variance | | | |
| | Régime privé | 1,00E-04 | 1,39E-04 | 1,48E-04 | 1,63E-04 | 6,80E-05 | 8,59E-05 |
| | Régime public | 2,82E-05 | 6,31E-05 | 3,86E-05 | 7,28E-05 | 1,79E-05 | 4,25E-05 |
| | Non assuré | 7,24E-05 | 8,92E-05 | 9,55E-05 | 1,11E-04 | 4,38E-05 | 5,79E-05 |

# 8 Discussion

Dans le présent article, nous proposons et évaluons une méthode non paramétrique pour produire des populations synthétiques. Cette méthode permet de tenir compte des caractéristiques du plan de sondage complexe sans utiliser de modèles hypothétiques pour les données observées, de sorte qu'elle est robuste aux erreurs de spécifications du modèle. En outre, contrairement aux méthodes fondées sur un modèle qui nécessitent l'élaboration de modèles d'imputation distincts pour les diverses variables d'intérêt, la méthode non paramétrique n'utilise que les variables de plan de sondage pour générer les populations synthétiques et n'est donc pas particulière à une variable.

Nous avons considéré les propriétés de rééchantillonnage de nos estimateurs synthétiques non paramétriques sous une loi Gamma univariée et sous une loi normale bivariée, en estimant les moyennes, les pentes et les ordonnées à l'origine. Les estimations ponctuelles étaient sans biais, les intervalles avaient une couverture correspondant approximativement au niveau nominal et les pertes d'efficacité comparativement aux données réelles étaient négligeables. Nous avons également considéré des

conditions « réelles » en générant une loi prédictive pour les données de la NHIS et de la MEPS de 2006 et en estimant les taux de couverture par une assurance médicale et les variances associées par la méthode non paramétrique ainsi que par une approche de modélisation log-linéaire entièrement paramétrique. Lorsque les modèles sont bien ajustés aux données, la méthode fondée sur un modèle est plus efficace que la méthode non paramétrique. Cependant, lorsque le modèle hypothétique n'est pas bien ajusté aux données, comme cela est le cas pour certains petits domaines, la méthode fondée sur un modèle peut produire des inférences non valides. Dans ces situations, la méthode non paramétrique est robuste à l'erreur de spécification du modèle.

Outre la robustesse à l'erreur de spécification du modèle, un autre avantage de la méthode non paramétrique tient au fait qu'elle n'utilise que les variables de plan de sondage, comme la strate, la grappe et le poids, pour imputer la partie non observée de la population. Contrairement aux méthodes fondées sur un modèle, elle ne requiert donc pas la modélisation de relations compliquées entre les variables d'intérêt, laquelle devient impossible si des valeurs manquent pour certains items dans les données réelles. La méthode non paramétrique préserve ces valeurs d'item manquantes dans les populations synthétiques produites. Cette propriété pourrait combler une lacune dans le domaine de l'imputation multiple en ce sens que les méthodes existantes consistent habituellement à imputer les valeurs manquantes dans les données comme si ces dernières avaient été obtenues par échantillonnage aléatoire simple, sans tenir compte des caractéristiques du plan de sondage complexe. Un avantage apparenté est que, même si les populations synthétiques sont produites non paramétriquement en se servant des variables de plan, il n'est pas nécessaire qu'elles contiennent elles-mêmes ces variables, puisqu'elles peuvent être analysées comme des échantillons aléatoires simples. Cela permet donc d'éliminer le risque de divulgation associé à la diffusion des variables du plan de sondage (De Waal et Willenborg, 1997; Mitra et Reiter, 2006; Reiter et Mitra, 2009).

Un quatrième avantage pratique de la méthode non paramétrique est qu'elle est plus facile à mettre en œuvre dans les progiciels statistiques existants, parce qu'elle est axée sur les variables du plan de sondage; de ce fait, il n'est pas nécessaire d'élaborer des stratégies particulières pour les divers types de variables et de structures de données.

Comme l'application du bootstrap bayésien en population finie (BBPF) pondéré ne requiert pas que l'on connaisse le nombre de grappes dans la population ni les probabilités conditionnelles de sélection à chaque degré de sélection dans le cas d'un échantillonnage à plusieurs degrés, nous utilisons un bootstrap bayésien approximatif pour tenir compte de la stratification et de la mise en grappes. Selon nous, cette approche est avantageuse à de nombreux égards, puisqu'habituellement, les ensembles de données à grande diffusion ne contiennent pas la ventilation des poids pour chaque degré d'échantillonnage. Toutefois, l'inconvénient est que, afin de s'assurer que les poids de rééchantillonnage soient positifs, le bootstrap bayésien produit moins de grappes dans les strates qu'il n'y en a dans les données réelles. Quand les probabilités de sélection sont connues pour tous les degrés d'échantillonnage, il semble probable que le BBPF pondéré puisse être mis en œuvre à chaque degré, en imputant la population de grappes non observées et la population d'éléments dans chaque grappe en deux étapes, à l'exemple de Meeden (1999), tout comme le BBPF à un degré s'inspire de Ghosh et Meeden (1983). Il s'agit d'un domaine dans lequel la recherche doit se poursuivre.

# Remerciements

La présente étude a été financée par la subvention R01CA129101 du NCI. Les auteurs remercient le rédacteur, le rédacteur associé et deux examinateurs anonymes de leurs commentaires. Nous sommes tout particulièrement redevables à l'examinateur qui nous a aidés à mieux comprendre et expliquer les liens entre le bootstrap bayésien en population finie et la loi a posteriori de Pólya exposés à la section 3.


# Bibliographie


Agresti, A. (2002). *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Chen, Q., Elliott, M.R. et Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 25-37.

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.

de Waal, A.G., et Willenborg, L.C.R.J. (1997). Statistical disclosure control and sampling weights. *Journal of Official Statistics*, 13, 417-434.

Dong, Q. (2012). Combining Information from Multiple Complex Surveys. Unpublished Thesis.

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 1, 27-40.

Elliott, M.R., et Little, R.J.A. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Ericson, W.A. (1969). Subjective Bayesian modeling in sampling finite populations. *Journal of the Royal Statistical Society*, B31, 195-234.

Ghosh, M., et Meeden, G. (1983). Estimation of the variance in finite population sampling. *Sankhyā: The Indian Journal of Statistics*, B45, 362-375.

Hinkins, S., Oh, H.L. et Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 1, 13-24.

Lazzeroni, L.C., et Little, R.J.A. (1998). Random effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Lo, A.Y. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.

Meeden, G. (1999). A noninformative Bayesian approach for two-stage cluster sampling. *Sankhyā: The Indian Journal of Statistics*, B61, 133-144.

Mitra, R., et Reiter J.P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. Privacy in statistical databases: Lecture Notes in Computer Science, 4302, 177-188.

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

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 2, 235-242.

Reiter, J.P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*, A168, 185-205.

Reiter, J.P., et Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1, 1, Article 6.

Rubin, D.B (1987). *Multiple Imputation for Non-Response in Surveys*, New York: John Wiley & Sons, Inc.

Scott, A.J. (1977). Large sample posterior distributions in finite populations. *The Annals of Mathematical Statistics*, 42, 1113-1117.

Skinner, C., Holt, D. et Smith, T. (1989). *Analysis of Complex Surveys*, New York: John Wiley & Sons, Inc.