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

Techniques d’enquête

Inférence statistique avec des échantillons d’enquête non probabiliste

par Changbao Wu

Date de diffusion : le 15 décembre 2022

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 suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2022
Vol. 48, n° 2, pp. 307-338
Statistique Canada, n° 12-001-X au catalogue

Inférence statistique avec des échantillons d'enquête non probabiliste

Changbao Wu[1]

Résumé

Nous offrons un examen critique et quelques discussions approfondies sur des questions théoriques et pratiques à l'aide d'une analyse des échantillons non probabilistes. Nous tentons de présenter des cadres inférentiels rigoureux et des procédures statistiques valides dans le cadre d'hypothèses couramment utilisées et d'aborder les questions relatives à la justification et à la vérification d'hypothèses sur des applications pratiques. Certains progrès méthodologiques actuels sont présentés et nous mentionnons des problèmes qui nécessitent un examen plus approfondi. Alors que l'article porte sur des échantillons non probabilistes, le rôle essentiel des échantillons d'enquête probabilistes comportant des renseignements riches et pertinents sur des variables auxiliaires est mis en évidence.

Mots-clés : Information auxiliaire; estimateur de la variance bootstrap; méthode de calage; estimateur doublement robuste; équations d'estimation; pondération de la probabilité inverse; prédiction fondée sur un modèle; poststratification; pseudo-vraisemblance; score de propension; enquête par quotas; analyse de sensibilité; estimation de la variance.

1. Introduction

Le domaine de l'échantillonnage se distingue des autres domaines de la statistique en raison d'un certain nombre de caractéristiques uniques. La population cible est constituée d'un nombre fini d'unités et les paramètres de population peuvent être déterminés sans erreur, du moins conceptuellement, en menant un recensement. Les contraintes opérationnelles et l'avantage sur le plan administratif pour la collecte des données rendent nécessaire d'envisager la stratification, la catégorisation des résultats et la sélection à probabilités inégales. Depuis le document précurseur de Neyman (1934), les méthodes d'échantillonnage probabiliste sont devenues l'un des outils de collecte de données primaires pour les statistiques officielles et les chercheurs dans les domaines des sciences de la santé, des études sociales et économiques, des affaires et de la commercialisation, des stocks de ressources agricoles et naturelles, et autres domaines. Des échantillons d'enquête probabilistes ont également été utilisés pour effectuer des études analytiques comportant des modèles et des paramètres de modèles; voir, par exemple, Binder (1983), Godambe et Thompson (1986), Thompson (1997), Rao et Molina (2015), entre autres. Des échantillons d'enquête probabilistes et une inférence fondée sur le plan ont constitué une mesure couronnée de succès dans le cadre des sciences statistiques au cours des 80 dernières années.

Toutefois, au cours des dernières années, « on a observé un vent de changement et on considère de plus en plus d'autres sources de données » (Beaumont, 2020). Le succès des échantillons d'enquête probabilistes a mené à des plans d'études ambitieux, à de longs questionnaires compliqués et à un fardeau accru sur les répondants. Les taux de réponse ont diminué et le coût de la collecte des données a grimpé en flèche au fil des ans. Compte tenu des progrès de nouvelles technologies et de l'explosion de l'information

1. Changboa Wu, Département des statistiques et de l'actuariat, Université de Waterloo, Waterloo (Ontario) N2L 3G1. Courriel : cbwu@uwaterloo.ca.

sur Internet, il existe également un fort désir d'accéder à des statistiques en temps réel. Statistique Canada a lancé les initiatives dites de modernisation appelée « Aller au-delà de l'approche fondée sur les données d'enquête pour adopter de nouvelles méthodes et intégrer des données provenant de diverses sources existantes ».

Les échantillons non probabilistes représentent l'une de ces sources des données qui ont gagné en popularité au cours des dernières années. Les échantillons non probabilistes ne sont pas nouveaux dans le domaine de l'échantillonnage. Ils ont été utilisés depuis les débuts de la réalisation d'enquêtes. Par exemple, les enquêtes par quotas ont donné des échantillons non probabilistes et la méthode est largement utilisée et peut réussir dans certaines conditions; voir la section 5 pour consulter d'autres analyses. Les échantillons non probabilistes n'ont jamais pris un véritable élan dans le passé dans la pratique des enquêtes en raison de l'absence d'un cadre de travail théorique mature pour l'analyse des données. Néanmoins, il existe des données accessibles qui sont moins chères et plus rapides à obtenir et qui sont devenues courantes pour la recherche en ligne. Les entreprises commerciales d'enquête créent et tiennent à jour une longue liste de personnes, appelées « panels volontaires », qui ont accepté que l'on communique avec elles pour participer à des enquêtes comme volontaires ou grâce à des incitatifs. Les mécanismes précis d'inclusion des personnes dans le panel sont habituellement inconnus, se traduisant par des échantillons d'enquête non probabilistes fondés sur des panels.

Le principal problème des échantillons non probabilistes est qu'il s'agit d'échantillons biaisés et que ceux-ci ne sont pas représentatifs de la population cible. On peut soutenir que, outre les échantillons à unités indépendantes et identiquement distribuées, la plupart des échantillons sont biaisés, et même les échantillons d'enquête probabilistes le sont. Les probabilités d'inclusion connues du plan d'enquête sont la raison pour laquelle nous ne nous inquiétons pas de la nature biaisée des échantillons d'enquête probabilistes, car elles mènent à des méthodes d'estimation valides par l'entremise de procédures appropriées de pondération. Le principal enjeu véritable des échantillons non probabilistes est donc l'aspect inconnu de l'inclusion dans un échantillon ou des mécanismes de participation. Il ressortira clairement des analyses présentées à la section 4 que la nature biaisée des échantillons non probabilistes ne peut être corrigée au moyen de l'échantillon lui-même. Cela nécessite des renseignements auxiliaires sur la population cible.

Le présent article offre un examen critique et quelques discussions approfondies sur des questions théoriques et pratiques à l'aide d'une analyse d'échantillons non probabilistes. La section 2 brosse un tableau du contexte général, des hypothèses communément utilisées ainsi que des cadres inférentiels employés dans les procédures statistiques traitées dans l'article. La section 3 présente une approche de prédiction des échantillons non probabilistes fondée sur un modèle. La section 4 aborde l'estimation de scores de propension et la construction d'estimateurs fondés sur les scores de propension. La section 5 illustre les liens entre les estimateurs pondérés de probabilité inverse et les enquêtes par quotas avec des extensions à la poststratification. La section 6 met l'accent sur les techniques ainsi que sur les enjeux liés à

l'estimation de la variance. La section 7 aborde l'importante question touchant la manière de contrôler et de vérifier les hypothèses requises dans la pratique. Quelques conclusions sont présentées à la section 8.

2. Hypothèses et cadres inférentiels

Supposons que la population cible $U = \{1, 2, \ldots, N\}$ est constituée de N unités étiquetées. Associées à l'unité i, on retrouve les valeurs $\mathbf{x}_i$ et y_i pour les variables auxiliaires $\mathbf{x}$ et la variable étudiée y. Les discussions portent sur un seul y, mais les ensembles de données sont plus susceptibles de contenir de multiples variables étudiées. Soit $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$, la moyenne de la population, laquelle est le paramètre d'intérêt. Soit $\{(y_i, \mathbf{x}_i), i \in S_A\}$, l'ensemble de données pour l'échantillon non probabiliste S_A avec n_A unités participantes. Pour la plupart des scénarios pratiques, la moyenne simple de l'échantillon $\bar{y}_A = n_A^{-1}\sum_{i \in S_A} y_i$ est un estimateur biaisé de μ_y et, par conséquent, il n'est pas valide.

2.1 Hypothèses

Soit $R_i = I(i \in S_A)$, la variable indicatrice pour l'unité i comprise dans l'échantillon non probabiliste S_A. Il convient de souligner que la variable R_i est définie pour tous les i dans la population cible. Supposons que :

$$\pi_i^A = P\left(i \in S_A \,|\, \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \,|\, \mathbf{x}_i, y_i\right), \quad i = 1, 2, \ldots, N.$$

Nous appelons la π_i^A les scores de propension, un terme emprunté aux ouvrages publiés sur les données manquantes (Rosenbaum et Rubin, 1983). Certains auteurs utilisent le terme « probabilités de participation ». Voir, par exemple, Beaumont (2020) et Rao (2021), entre autres. Les scores de propension π_i^A caractérisent l'inclusion dans un échantillon et les mécanismes de participation. Ils sont inconnus et nécessitent des hypothèses appropriées du modèle pour l'élaboration de méthodes d'estimation valides. Chen, Li, et Wu (2020) ont utilisé les trois hypothèses de base suivantes, lesquelles ont été adaptées à partir de la littérature sur les données manquantes.

A1 L'inclusion dans l'échantillon et l'indicateur de participation R_i, ainsi que la variable étudiée y_i sont indépendants compte tenu de l'ensemble de covariables $\mathbf{x}_i$, c'est-à-dire $\left(R_i \perp y_i\right) \,|\, \mathbf{x}_i$.

A2 Toutes les unités dans la population cible ont des scores de propension non nuls, c'est-à-dire $\pi_i^A > 0,\ i = 1, 2, \ldots, N$.

A3 Les variables indicatrices $R_1, R_2, \ldots, R_N$ sont indépendantes compte tenu de l'ensemble de variables auxiliaires, $\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)$.

L'hypothèse A1 est semblable à l'hypothèse de répartition au hasard des données manquantes pour l'analyse des données manquantes. Selon A1, cela donne $\pi_i^A = P\left(R_i = 1 \,|\, \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \,|\, \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$. L'hypothèse A2 peut être problématique dans la pratique. Voir la section 7 pour obtenir d'autres analyses. L'hypothèse A3 se vérifie habituellement quand les participants sont approchés un à la fois, mais peut être

discutable lors de l'utilisation de sélections en grappes. Il est démontré à la section 4 que l'estimation de $\pi_i^A = \pi(\mathbf{x}_i)$ selon l'hypothèse A1 nécessite une information auxiliaire sur la population cible. Le scénario idéal est que l'information auxiliaire complète $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ soit disponible. Le scénario le plus pratique est que l'information auxiliaire peut être obtenue d'une enquête probabiliste existante.

A4 Il existe un échantillon d'enquête probabiliste S_B de taille n_B avec des renseignements sur les variables auxiliaires $\mathbf{x}$ (mais pas sur y) disponibles dans l'ensemble de données $\left\{(\mathbf{x}_i, d_i^B), i \in S_B\right\}$, où d_i^B sont les poids de sondage pour l'échantillon probabiliste S_B.

Le S_B est appelé « l'échantillon d'enquête probabiliste de référence ». La partie la plus essentielle de l'hypothèse A4 est que l'ensemble de variables auxiliaires $\mathbf{x}$ est observé dans l'échantillon non probabiliste S_A et dans l'objet probabiliste S_B. Un échantillon d'enquête probabiliste de référence est souvent disponible dans la pratique, mais l'ensemble de variables auxiliaires commun peut ne pas contenir toutes les composantes pour satisfaire l'hypothèse A1.

2.2 Cadre inférentiel

Il existe trois sources possibles de variation en vertu du contexte général de deux échantillons S_A et S_B : i) Le modèle q pour le score de propension sur l'inclusion dans l'échantillon et la participation dans l'échantillon d'enquête non probabiliste S_A; ii) le modèle ξ pour la régression des résultats $(y \mid \mathbf{x})$ ou l'imputation; et iii) le plan d'échantillonnage probabiliste p pour l'échantillon d'enquête probabiliste de référence S_B. Pour les trois approches de l'inférence qui seront abordées dans les sections 3 et 4, l'échantillon probabiliste de référence S_B est toujours mis à contribution. Chacune des trois approches nécessite un cadre de randomisation conjoint mettant en jeu p et un de (q, ξ).

a) Approche de prédiction fondée sur un modèle : le cadre ξp sous la randomisation conjointe du modèle de régression des résultats ξ et le plan d'échantillonnage avec probabilités p.

b) Pondération de la probabilité inverse à l'aide de scores de propension estimés : le cadre qp sous la randomisation conjointe du modèle de score de propension q et le plan d'échantillonnage avec probabilités p.

c) Inférence doublement robuste : le cadre qp ou le cadre ξp, sans spécification de l'un ou de l'autre.

Le cadre inférentiel est la base du développement théorique. La convergence des estimateurs ponctuels doit être établie en vertu de la randomisation conjointe appropriée. Les variances théoriques comportent habituellement deux composantes, une de chaque source de variation, et le calcul correct des deux composantes constitue la clé de la construction d'estimateurs de variance convergents selon le cadre inférentiel désigné.

3. Approche de prédiction fondée sur un modèle

Les méthodes de prédiction fondées sur un modèle pour les paramètres de population finie nécessitent deux ingrédients essentiels : la quantité de renseignements auxiliaires qui sont disponibles à l'étape de l'estimation et la fiabilité du modèle théorique d'inférence. En l'absence de toute information auxiliaire, le modèle moyen commun $E_\xi(y_i) = \mu_0,\ V_\xi(y_i) = \sigma^2,\ i = 1, \ldots, N$ peut être considéré comme raisonnable, mais l'estimateur de prédiction fondé sur un modèle $\hat{\mu}_y = \bar{y}_A = n_A^{-1}\sum_{i \in S_A} y_i,$ bien que non biaisé selon le modèle depuis $E_\xi\left(\bar{y}_A - \mu_y\right) = 0,$ n'est généralement pas un estimateur acceptable de μ_y. La variance σ^2 pour le modèle moyen commun est habituellement de grande taille et elle rend l'estimateur $\hat{\mu}_y = \bar{y}_A$ avec une variance de prédiction qui est trop importante pour être d'une utilité pratique.

3.1 Modèles de régression des résultats semi-paramétriques

Sans perte de généralité, nous supposons que $\mathbf{x}$ contient 1 comme sa première composante correspondant à l'ordonnée à l'origine d'un modèle de régression. Dans les conditions décrites dans la section 2, nous considérons le modèle semi-paramétrique suivant pour la population finie, désigné comme ξ :

$$E_\xi\left(y_i \mid \mathbf{x}_i\right) = m\left(\mathbf{x}_i, \boldsymbol{\beta}\right), \text{ et } V_\xi\left(y_i \mid \mathbf{x}_i\right) = v\left(\mathbf{x}_i\right)\sigma^2, \quad i = 1, 2, \ldots, N, \tag{3.1}$$

ou la fonction moyenne $m(\cdot,\cdot)$ et la fonction variance $v(\cdot)$ ont des formes connues et les y_i sont également présumés être conditionnellement indépendants compte tenu des $\mathbf{x}_i$. Soit $\boldsymbol{\beta}_0$ et $\sigma_0^2,$ les valeurs vraies des paramètres $\boldsymbol{\beta}$ et σ^2 du modèle selon le modèle théorique. La première conséquence importante de l'hypothèse A1 est que $E_\xi\left(y_i \mid \mathbf{x}_i, R_i = 1\right) = E_\xi\left(y_i \mid \mathbf{x}_i\right)$ et $V_\xi\left(y_i \mid \mathbf{x}_i, R_i = 1\right) = V_\xi\left(y_i \mid \mathbf{x}_i\right)$. Le modèle (3.1), qui est considéré pour la population finie, se vérifie également pour les unités dans l'échantillon d'enquête non probabiliste S_A. L'estimateur par le maximum de quasi-vraisemblance $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}_0$ est obtenu à l'aide de l'ensemble de données $\left\{(y_i, \mathbf{x}_i), i \in S_A\right\}$ pour l'échantillon d'enquête non probabiliste comme la solution des équations de quasi-score (McCullagh and Nelder, 1989) données par :

$$\mathrm{S}\left(\boldsymbol{\beta}\right) = \sum_{i \in S_A} \frac{\partial m\left(\mathbf{x}_i, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} \left\{v(\mathbf{x}_i)\right\}^{-1} \left\{y_i - m(\mathbf{x}_i, \boldsymbol{\beta})\right\} = \mathbf{0}. \tag{3.2}$$

Le modèle semi-paramétrique (3.1) peut être élargi pour remplacer $v(\mathbf{x}_i)$ par une fonction de variance générale $v(\mu_i)$ où $\mu_i = m(\mathbf{x}_i, \boldsymbol{\beta})$. La théorie de l'estimation par la méthode du maximum de vraisemblance englobe les modèles de régression linéaires et non linéaires comportant des estimateurs par les moindres carrés pondérés, le modèle de régression logistique et d'autres modèles linéaires généralisés. Soit $m_i = m(\mathbf{x}_i, \boldsymbol{\beta}_0)$ et $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}),\ i = 1, 2, \ldots, N.$

3.2 Deux formes générales d'estimateurs de prédiction

Il existe deux estimateurs de prédiction fondée sur un modèle pour μ_y en présence de renseignements auxiliaires complets $\left\{\mathbf{x}_1, \ldots, \mathbf{x}_N\right\}$; voir le chapitre 5 de Wu et Thompson (2020). Notons que $E_\xi(\mu_y) = N^{-1}\sum_{i=1}^N m_i.$ Les deux estimateurs de prédiction sont établis comme suit :

$$\hat{\mu}_{y1} = \frac{1}{N}\sum_{i=1}^{N}\hat{m}_i \quad \text{et} \quad \hat{\mu}_{y2} = \frac{1}{N}\left\{\sum_{i\in S_A} y_i - \sum_{i\in S_A}\hat{m}_i + \sum_{i=1}^{N}\hat{m}_i\right\}. \tag{3.3}$$

L'estimateur $\hat{\mu}_{y2}$ est établi en fonction de $\mu_y = N^{-1}\left\{\sum_{i\in S_A} y_i + \sum_{i\notin S_A} y_i\right\}$ et s'appuie sur $\sum_{i\notin S_A}\hat{m}_i = \sum_{i=1}^{N}\hat{m}_i - \sum_{i\in S_A}\hat{m}_i$ pour prédire le terme non observé $\sum_{i\notin S_A} y_i$. Dans un modèle de régression linéaire où $m(\mathbf{x},\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$, les deux estimateurs donnés dans (3.3) se réduisent à :

$$\hat{\mu}_{y1} = \boldsymbol{\mu}_{\mathbf{x}}'\hat{\boldsymbol{\beta}} \quad \text{et} \quad \hat{\mu}_{y2} = \frac{n_A}{N}\left(\bar{y}_A - \bar{\mathbf{x}}_A'\hat{\boldsymbol{\beta}}\right) + \boldsymbol{\mu}_{\mathbf{x}}'\hat{\boldsymbol{\beta}}, \tag{3.4}$$

où $\boldsymbol{\mu}_{\mathbf{x}} = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i$ est le vecteur des moyennes de population des variables de $\mathbf{x}$ et $\bar{\mathbf{x}}_A = n_A^{-1}\sum_{i\in S_A}\mathbf{x}_i$ est le vecteur de la moyenne simple de l'échantillon de $\mathbf{x}$ de l'échantillon non probabiliste S_A. Si le modèle de régression linéaire contient une ordonnée à l'origine et $\hat{\boldsymbol{\beta}}$ est l'estimateur des moindres carrés ordinaires, nous avons $\hat{\mu}_{y2} = \hat{\mu}_{y1} = \boldsymbol{\mu}_{\mathbf{x}}'\hat{\boldsymbol{\beta}}$, puisque $\bar{y}_A - \bar{\mathbf{x}}_A'\hat{\boldsymbol{\beta}} = 0$ en raison de la somme nulle des résidus prédits. Les estimateurs de prédiction en (3.4) sous le modèle linéaire ne requièrent que la moyenne de population $\boldsymbol{\mu}_{\mathbf{x}}$ en plus de l'échantillon non probabiliste S_A. Dans les conditions décrites à la section 2 comportant des renseignements auxiliaires sur $\mathbf{x}$ fournies par un échantillon probabiliste de référence S_B, nous remplaçons simplement $\sum_{i=1}^{N}\hat{m}_i$ par $\sum_{i\in S_B} d_i^B\hat{m}_i$ pour les estimateurs en (3.3) et substituons $\boldsymbol{\mu}_{\mathbf{x}}$ par $\hat{\boldsymbol{\mu}}_{\mathbf{x}} = \hat{N}_B^{-1}\sum_{i\in S_B} d_i^B\mathbf{x}_i$ pour les estimateurs en (3.4), où $\hat{N}_B = \sum_{i\in S_B} d_i^B$. La taille de la population N apparaissant en (3.3) ou (3.4) doit également être remplacée par $\hat{N}_B$ même si elle est connue.

3.3 Imputation massive

Les estimateurs de prédiction fondée sur un modèle μ_y reposant sur un échantillon d'enquête non probabiliste sur $(y,\mathbf{x})$ et un échantillon d'enquête probabiliste de référence sur $\mathbf{x}$ ont été traditionnellement présentés comme *l'estimateur d'imputation massive*. La variable étudiée y n'est pas observée pour aucune des unités dans l'échantillon de l'enquête de référence S_B et donc peut être perçue comme manquante pour tous les $i\in S_B$. Soit y_i^*, une valeur imputée pour y_i, $i\in S_B$. L'estimateur d'imputation massive de μ_y est ensuite établi comme suit :

$$\hat{\mu}_{y\text{IM}} = \frac{1}{\hat{N}_B}\sum_{i\in S_B} d_i^B y_i^*, \tag{3.5}$$

où $\hat{N}_B$ est défini comme il l'est précédemment et l'indice « IM » signifie « imputation massive » (et non pas « imputation multiple »). Sous l'imputation par la régression déterministe où $y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, l'estimateur $\hat{\mu}_{y\text{IM}}$ se réduit à l'estimateur de prédiction fondée sur un modèle $\hat{\boldsymbol{\mu}}_{\mathbf{x}}'\hat{\boldsymbol{\beta}}$ comme il est mentionné à la section 3.2.

L'approche de l'imputation massive pour l'analyse d'échantillons d'enquête non probabilistes s'inscrit dans le même esprit que les méthodes de prédiction fondées sur un modèle, mais elle ouvre la porte à l'utilisation de modèles plus flexibles et de techniques d'imputation qui ont été mises au point dans la littérature existante sur les problèmes de données manquantes. Cette approche a été examinée la première fois par Rivers (2007) à l'aide de la méthode dite d'*appariement d'échantillons*. Pour chaque $i\in S_B$, le y_i

« manquant » est imputé comme $y_i^* = y_j$ pour quelques valeurs $j \in S_A$, où j est un donneur apparié de S_A sélectionné par la méthode du plus proche voisin, mesuré par la distance entre $\mathbf{x}_i$ et $\mathbf{x}_j$. Le modèle sous-jacent ξ pour la méthode d'imputation du plus proche voisin est non paramétrique, c'est-à-dire $E_\xi\left(y_i \mid \mathbf{x}_i\right) = m\left(\mathbf{x}_i\right)$ pour une fonction inconnue $m(\cdot)$. La valeur d'appariement y_j peut être perçue comme la valeur prédite du y_i manquant selon le modèle. Les propriétés théoriques des estimateurs fondées sur l'imputation du plus proche voisin ont été abordées par Chen et Shao (2000, 2001) pour les problèmes de données d'enquêtes manquantes.

Le modèle semi-paramétrique (3.1) peut être utilisé pour l'imputation massive par la régression déterministe. Dans l'hypothèse A1, un estimateur convergent $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ est d'abord obtenu de l'ensemble de données d'échantillons non probabilistes $\left\{(y_i, \mathbf{x}_i), i \in S_A\right\}$, et l'estimateur $\hat{\boldsymbol{\beta}}$ est ensuite utilisé pour calculer les valeurs imputées $y_i^* = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$ pour $i \in S_B$. En d'autres mots, l'hypothèse A1 met en jeu ladite *transportabilité du modèle* par Kim, Park, Chen et Wu (2021) : le modèle qui est construit pour l'échantillon non probabiliste peut être utilisé pour une prédiction avec l'échantillon probabiliste de référence. L'estimateur d'imputation massive résultant $\hat{\mu}_{y\mathrm{IM}}$ est semblable à l'un des estimateurs de prédiction fondée sur un modèle présentés à la section 3.2. Les propriétés asymptotiques et l'estimation de la variance pour l'estimateur $\hat{\mu}_{y\mathrm{IM}}$ réalisée à l'aide du modèle semi-paramétrique (3.1) ont été décrites par Kim et coll. (2021).

Selon l'approche de l'imputation massive, le seul rôle joué par le y_i observé pour $i \in S_A$ est d'estimer les paramètres du modèle $\boldsymbol{\beta}$. L'estimateur $\hat{\mu}_{y\mathrm{IM}}$ est établi à l'aide du modèle prédit et des renseignements auxiliaires de l'échantillon probabiliste de référence S_B. Il semble que nous n'avons pas pleinement utilisé les renseignements sur le y_i observé étant donné que μ_y est le paramètre principal d'intérêt. Cela a mené à la question de recherche décrite au chapitre 17 de Wu et Thompson (2020) sur *l'appariement inversé d'échantillons*. L'estimateur proposé est construit comme suit : $\hat{\mu}_{y\mathrm{A}} = (\hat{N}^*)^{-1} \sum_{i \in S_A} d_i^* y_i$ en utilisant tous les y_i observés dans l'échantillon non probabiliste, où $\hat{N}^* = \sum_{i \in S_A} d_i^*$. Le d_i^* est un poids d'enquête apparié de S_B de sorte que $d_i^* = d_j^B$ avec $j \in S_B$ soit le plus proche voisin de $i \in S_A$, mesuré par $\left\| \mathbf{x}_i - \mathbf{x}_j \right\|$. Les propriétés théoriques de l'estimateur apparié inverse $\hat{\mu}_{yA}$ reposant sur le plus proche voisin $j \in S_B$ pour coupler d_i^* avec d_j^B n'a pas fait l'objet d'une enquête formelle dans la littérature existante.

Wang, Graubard, Katki et Li (2020) ont proposé une approche de pondération par noyau pour inverser l'appariement des échantillons en utilisant $d_i^* \propto \sum_{j \in S_B} K_{ij} d_j$, où K_{ij} est une distance entre les noyaux $\hat{p}_i$ et $\hat{p}_j$; voir la méthode de pondération par la propension logistique ajustée à la fin de la section 4.1.1 sur le calcul de $\hat{p}_i$. Ils ont montré que l'estimateur $\hat{\mu}_{y\mathrm{A}}$ est convergent dans certaines conditions de régularité. Dans un récent document de travail sur arXiv par Liu et Valliant (2021), les auteurs ont traité des problèmes avec le biais et la variance de l'estimateur à appariement inverse dans différents cadres de randomisation mettant en jeu une source, deux sources ou les trois sources (p, q, ξ). Les auteurs ont également proposé une étape de calage sur les poids appariés, ce qui semble être une idée prometteuse. Des études supplémentaires sur ce sujet sont nécessaires.

L'approche par imputation massive pour l'analyse des échantillons d'enquête non probabilistes mène à une question de recherche intéressante faisant actuellement l'objet d'une étude par un étudiant au doctorat

de l'université de Waterloo : Est-il théoriquement possible et pratiquement utile de créer un ensemble de données obtenues par imputation massive $\left\{(y_i^*, \mathbf{x}_i, d_i^B), i \in S_B\right\}$ selon l'échantillon d'enquête probabiliste de référence qui peut être utilisé pour des inférences statistiques générales ? La réponse dépend clairement des types de problèmes inférentiels à effectuer sur les ensembles de données obtenues par imputation. Une exigence minimale est que la distribution conditionnelle de la variable étudiée y compte tenu des covariables $\mathbf{x}$ est préservée pour l'ensemble de données obtenues par imputation massive. La méthode d'imputation du plus proche voisin et la méthode d'imputation par la régression aléatoire peuvent être utiles à cette fin. L'imputation fractionnaire est une autre possibilité, plus particulièrement pour les variables étudiées binaires ou ordinales. L'imputation multiple est également potentiellement utile dans ce sens pour créer de multiples ensembles de données obtenues par imputation massive. Dans ce cas, l'indice « IM » pourrait devoir être changé pour « IM^2 », signifiant « imputation massive avec imputation multiple ».

4. Approche fondée sur les scores de propension

Les scores de propension $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right)$ pour l'échantillon d'enquête non probabiliste S_A sont théoriquement définis pour toutes les unités dans la population cible. L'estimation des scores de propension pour les unités dans S_A, lequel joue le rôle le plus important pour les méthodes fondées sur les scores de propension, nécessite un modèle théorique sur les scores de propension et des renseignements auxiliaires au niveau de la population. Dans la présente section, nous abordons d'abord les procédures d'estimation pour les scores de propension en vertu des conditions et des hypothèses décrites à la section 2; nous fournirons ensuite un aperçu des méthodes d'estimation proposées dans la littérature récente sur la moyenne de la population μ_y comportant des estimations de scores de propension.

4.1 Estimation des scores de propension

Dans l'hypothèse A1, les scores de propension $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$ sont une fonction des variables auxiliaires $\mathbf{x}_i$, mais la forme fonctionnelle peut être compliquée et est complètement inconnue. Trois formes paramétriques populaires $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ pour traiter une réponse binaire peuvent être considérées : i) la fonction logit inverse $\pi_i^A = 1 - \left\{1 + \exp(\mathbf{x}_i'\boldsymbol{\alpha})\right\}^{-1}$; ii) la fonction probit inverse $\pi_i^A = \Phi(\mathbf{x}_i'\boldsymbol{\alpha})$, où $\Phi(\cdot)$ est la fonction de distribution cumulative de $N(0,1)$; iii) la fonction log-log complémentaire inverse $\pi_i^A = 1 - \exp\left\{-\exp(\mathbf{x}_i'\boldsymbol{\alpha})\right\}$. Des techniques non paramétriques, sans présumer une forme fonctionnelle explicite pour $\pi\left(\mathbf{x}\right)$, constituent des solutions de rechange intéressantes pour l'estimation des scores de propension.

4.1.1 La méthode par pseudo maximum de vraisemblance

Soit $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$, une forme paramétrique précisée avec des paramètres de modèle inconnus $\boldsymbol{\alpha}$. Dans la situation idéale où l'information auxiliaire complète $\left\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right\}$ est disponible avec l'hypothèse d'indépendance A3, la fonction du logarithme du rapport de vraisemblance complète sur $\boldsymbol{\alpha}$ peut être formulée comme suit (Chen et coll., 2020) :

$$\ell(\boldsymbol{\alpha}) = \log\left\{\prod_{i=1}^{N}\left(\pi_i^A\right)^{R_i}\left(1-\pi_i^A\right)^{1-R_i}\right\} = \sum_{i\in S_A}\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right)+\sum_{i=1}^{N}\log\left(1-\pi_i^A\right). \tag{4.1}$$

L'estimateur du maximum de vraisemblance de $\boldsymbol{\alpha}$ est le maximiseur de $\ell(\boldsymbol{\alpha})$. Dans les conditions actuelles où l'information auxiliaire sur la population est fournie par un échantillon d'enquête probabiliste de référence S_B, nous remplaçons $\ell(\boldsymbol{\alpha})$ par une fonction de pseudo-logarithme du rapport de vraisemblance (Chen et coll., 2020) :

$$\ell^*(\boldsymbol{\alpha}) = \sum_{i\in S_A}\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) + \sum_{i\in S_B} d_i^B\log\left(1-\pi_i^A\right). \tag{4.2}$$

L'estimateur du maximum de pseudo-vraisemblance $\hat{\boldsymbol{\alpha}}$ est le maximiseur de $\ell^*(\boldsymbol{\alpha})$ et peut être obtenu comme la solution aux équations de pseudo-score données par $\mathbf{U}(\boldsymbol{\alpha}) = \partial\ell^*(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = \mathbf{0}$. Si la fonction logit inverse est présumée pour π_i^A, les fonctions de pseudo-score sont données par :

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum_{i\in S_A}\mathbf{x}_i - \sum_{i\in S_B} d_i^B\pi(\mathbf{x}_i,\boldsymbol{\alpha})\,\mathbf{x}_i. \tag{4.3}$$

En général, les fonctions de pseudo-score $\mathbf{U}(\boldsymbol{\alpha})$ aux valeurs réelles des paramètres du modèle $\boldsymbol{\alpha}_0$ ne sont pas biaisées dans la randomisation qp conjointe dans le sens que $E_{qp}\{\mathbf{U}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$, laquelle indique que l'estimateur $\hat{\boldsymbol{\alpha}}$ est qp-convergent pour $\boldsymbol{\alpha}_0$ (Tsiatis, 2006).

Valliant et Dever (2011) ont tenté plus tôt d'estimer les scores de propension en regroupant l'échantillon non probabiliste S_A avec l'échantillon probabiliste de référence S_B. Soit $S_{AB} = S_A \cup S_B$, l'échantillon groupé sans retirer toute unité dupliquée potentielle. Soit $R_i^* = 1$, si $i \in S_A$ et $R_i^* = 0$ si $i \in S_B$. Valliant et Dever (2011) ont proposé d'adapter un modèle de régression logistique pondéré d'enquête à l'ensemble des données groupées $\left\{(R_i^*, \mathbf{x}_i, d_i), i \in S_{AB}\right\}$, où les poids sont définis comme $d_i = 1$ si $i \in S_A$ et $d_i = d_i^B\left(1 - n_A/\hat{N}_B\right)$ si $i \in S_B$. La principale motivation derrière la création des poids d_i est que le poids total $\sum_{i\in S_{AB}} d_i = \sum_{i\in S_B} d_i^B = \hat{N}_B$ pour l'échantillon groupé correspond à la taille de la population estimée, et il reste à espérer que le modèle de régression logistique pondéré d'enquête entraînera des estimations valides pour les scores de propension. Chen et coll. (2020) ont démontré que l'approche de l'échantillon groupé de Valliant et Dever (2011) ne produit pas des estimateurs convergents pour les paramètres du modèle de scores de propension à moins que l'échantillon non probabiliste S_A ne soit un simple échantillon aléatoire de la population cible.

La méthode de Valliant et Dever (2011) révèle une difficulté fondamentale avec les approches fondées sur l'échantillon groupé S_{AB}. Si les unités dans l'échantillon non probabiliste S_A sont traitées comme des unités échangeables dans l'échantillon groupé S_{AB}, comme l'indiquaient les poids égaux $d_i = 1$ utilisés dans la méthode de Valliant et Dever (2011), les estimations résultantes des scores de propension ne seront pas valides à moins que S_A soit un simple échantillon aléatoire. Cette observation a des répercussions sur la validité des méthodes non paramétriques ou des méthodes à arborescence de régression qui seront abordées à la section 4.1.3.

Dans un article récent, Wang, Valliant et Li (2021) ont proposé une méthode de pondération par la propension logistique ajustée. La méthode comporte deux étapes de calcul des scores de propension estimés. Les estimations initiales, désignées comme S_{AB} pour $d_i = 1,$ sont obtenues en adaptant le modèle de régression logistique pondéré d'enquête à l'échantillon groupé $\hat{p}_i$ semblable à Valliant et Dever (2011), assorties des poids définis comme suit : $i \in S_A$ si $d_i = d_i^B$ et $i \in S_B$ si $i \in S_A$. Les scores de propension estimés définitifs sont calculés comme suit : $\hat{\pi}_i^A = \hat{p}_i / (1 - \hat{p}_i)$. Le principal argument théorique est que l'équation $\pi_i^A = p_i / (1 - p_i)$ où $\pi_i^A = P(i \in S_A \mid U)$, $p_i = P\left(i \in S_A^* \mid S_A^* \cup U\right)$ et S_A^*, est une copie de S_A, mais est considéré comme un ensemble différent. Toutefois, les arguments soulèvent des questions conceptuelles, car les probabilités $\pi_i^A = P(i \in S_A \mid U)$ sont définies dans le modèle de scores de propension théorique avec la population finie donnée $U,$ et le modèle théorique ne mène pas à une interprétation significative des probabilités $p_i = P\left(i \in S_A^* \mid S_A^* \cup U\right)$. Ces dernières nécessitent un espace de probabilité différent et sont conditionnelles au S_A donné. En fait, on peut facilement faire valoir qu'en vertu du modèle de scores de propension théorique et conditionnel au S_A donné, on observe $p_i = 1$ si $i \in S_A$ et $p_i = 0$ autrement.

4.1.2 Estimation de méthodes fondées sur des équations

Les équations de pseudo-score $\mathbf{U}(\boldsymbol{\alpha}) = \mathbf{0}$ dérivées de la fonction de pseudo-vraisemblance $\ell^*(\boldsymbol{\alpha})$ peuvent être remplacées par un système d'équations d'estimation générales. Soit $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha}),$ un vecteur personnalisé de fonctions avec la même dimension de $\boldsymbol{\alpha}$. Supposons que :

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in S_A} \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\alpha}) \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}). \tag{4.4}$$

Il s'ensuit que $E_{qp}\{\mathbf{G}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$ pour tout $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$ choisi. En principe, un estimateur $\hat{\boldsymbol{\alpha}}$ de $\boldsymbol{\alpha}$ peut être obtenu en résolvant $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ au moyen de la forme paramétrique choisie $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ et des fonctions choisies $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha}),$ et l'estimateur $\hat{\boldsymbol{\alpha}}$ est convergent.

L'estimateur $\hat{\boldsymbol{\alpha}}$ reposant sur des fonctions personnalisées arbitraires $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$ est habituellement moins efficace que celui fondé sur des fonctions de pseudo-score, en raison de l'optimalité de l'estimateur par le maximum de vraisemblance (Godambe, 1960). Quelques résultats empiriques limités montrent également que la solution à $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ peut se révéler instable pour certains choix de $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. Néanmoins, les méthodes fondées sur des équations d'estimation procurent un outil utile pour l'estimation des scores de propension dans des scénarios plus restreints. Par exemple, si nous avons $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{x} / \pi(\mathbf{x}, \boldsymbol{\alpha}),$ les fonctions d'estimation données en (4.4) réduisent à :

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in S_A} \frac{\mathbf{x}_i}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - \sum_{i \in S_B} d_i^B \mathbf{x}_i. \tag{4.5}$$

La forme de $\mathbf{G}(\boldsymbol{\alpha})$ en (4,5) ressemble à la version « déformée » des fonctions de pseudo-score données en (4.3) sous un modèle de régression logistique pour les scores de propension. Toutefois, la différence la

plus importante en pratique entre les deux versions est le fait que le $\mathbf{G}(\boldsymbol{\alpha})$ donné en (4.5) ne nécessite que des totaux de population estimés pour les variables auxiliaires $\mathbf{x}$. Il existe des scénarios où les totaux de population des variables auxiliaires $\mathbf{x}$ peuvent être consultés ou estimés à partir d'une source existante, mais les valeurs de $\mathbf{x}$ au niveau des unités pour l'ensemble de la population ou même juste un échantillon probabiliste ne sont pas disponibles. Grâce à l'utilisation des fonctions d'estimation $\mathbf{G}(\boldsymbol{\alpha})$ données (4.5), il a été possible d'obtenir des estimations valides des scores de propension pour les unités de l'échantillon non probabiliste. La section 6.3 donne un exemple dans lequel l'approche fondée sur des équations d'estimation mène à un estimateur de variance valide pour l'estimateur doublement robuste de la moyenne de la population.

4.1.3 Méthodes non paramétriques et méthodes fondées sur un arbre de régression

Les scores de propension $\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i\right)$ sont la fonction de moyenne $E_q\left(R_i \mid \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$ pour la réponse binaire R_i. Les méthodes non paramétriques pour l'estimation $\pi(\mathbf{x})$ peuvent constituer une solution de rechange intéressante. Le principal défi est d'élaborer des procédures d'estimation qui procurent des estimations valides des scores de propension. Comme il est mentionné à la section 4.1.1, les méthodes d'estimation fondées sur l'échantillon groupé $S_{AB} = S_A \cup S_B$ peuvent mener à des estimations invalides. Des stratégies semblables à celle utilisée par Chen et coll. (2020) peuvent être justifiées théoriquement dans le cadre *qp* conjoint, où les procédures d'estimation sont calculées d'abord à l'aide de données de l'ensemble de la population finie et où les quantités de populations inconnues sont ensuite remplacées par des estimations obtenues de l'échantillon probabiliste de référence.

Nous examinons l'estimateur par la régression par noyau de $\pi_i^A = \pi\left(\mathbf{x}_i\right)$. Supposons que l'ensemble de données $\left\{(R_i, \mathbf{x}_i), i = 1, 2, \ldots, N\right\}$ est disponible pour la population finie. Soit $K_h\left(t\right) = K\left(t/h\right)$, un noyau choisi ayant une largeur de bande h. L'estimateur par la régression par noyau Nadaraya-Watson (Nadaraya, 1964; Watson, 1964) de $\pi\left(\mathbf{x}\right)$ est donné par :

$$\tilde{\pi}\left(\mathbf{x}\right) = \frac{\sum_{j=1}^{N} K_h\left(\mathbf{x} - \mathbf{x}_j\right) R_j}{\sum_{j=1}^{N} K_h\left(\mathbf{x} - \mathbf{x}_j\right)}. \tag{4.6}$$

Un estimateur par noyau dans la forme de $\tilde{\pi}(\mathbf{x})$ donnée en (4.6) ne possède généralement pas de valeurs pratiques, car nous n'avons pas de renseignements auxiliaires complets pour la population finie. Il s'avère que, pour l'estimation des scores de propension, le numérateur en (4.6) n'a besoin que d'observations de l'échantillon non probabiliste provenant de la variable binaire R_j, et le dénominateur est un chiffre de population et peut être estimé en utilisant l'échantillon probabiliste de référence. L'estimateur par la régression par noyau non paramétrique des scores de propension est donné par (Yuan, Li et Wu, 2022) :

$$\hat{\pi}_i^A = \hat{\pi}\left(\mathbf{x}_i\right) = \frac{\sum_{j \in S_A} K_h\left(\mathbf{x}_i - \mathbf{x}_j\right)}{\sum_{j \in S_B} d_j^B K_h\left(\mathbf{x}_i - \mathbf{x}_j\right)}, \quad i \in S_A. \tag{4.7}$$

L'estimateur $\hat{\pi}_i^A$ donné en (4,7) est convergent en vertu du cadre qp conjoint et le modèle q pour les scores de propension est très flexible en raison de l'hypothèse non paramétrique sur $\pi(\mathbf{x})$. Les scores de propension estimés sont faciles à calculer quand la taille de $\mathbf{x}$ n'est pas trop élevée. Les problèmes avec un $\mathbf{x}$ de grandes dimensions et les choix du noyau $K_h(\cdot)$ et de la largeur de bande h subsistent, comme dans les applications générales des méthodes d'estimation à base de noyaux. Les résultats d'une simulation rapportés par Yuan et coll. (2022) montrent que la méthode de l'estimation par la méthode des noyaux donne de solides résultats pour les scores de propension à l'aide du noyau normal et de choix populaires pour la largeur de bande.

Chu et Beaumont (2019) ont examiné les méthodes fondées sur un arbre de régression. Leur méthode TrIPW proposée est une variante de l'algorithme CART (Breiman, Friedman, Olshen et Stone, 1984) et repose sur des données de l'échantillon combiné de l'échantillon non probabiliste et de l'échantillon probabiliste de référence. La méthode vise à construire un arbre de classification, les nœuds terminaux de l'arbre final étant traités comme des groupes homogènes en matière de scores de propension. L'estimateur de μ_y est construit en fonction de l'arbre final et de la poststratification. La section 5 contient d'autres renseignements sur les estimateurs stratifiés *a posteriori*.

Les techniques d'apprentissage statistique comme les arbres de classification et de régression et les forêts aléatoires d'arbres décisionnels ont été élaborées principalement à des fins de prédiction. Leur utilisation pour estimer les scores de propension d'échantillons non probabilistes nécessite des travaux supplémentaires. Il ne s'agit pas d'une approche souhaitable pour naïvement appliquer les méthodes sur l'échantillon groupé S_{AB} sans justification théorique sur la convergence des estimateurs finaux. Il faut encourager d'autres études dans ce sens.

4.2 Pondération de probabilité inverse

Soit $\hat{\pi}_i^A$, une estimation de $\pi_i^A = P\left(i \in S_A \mid \mathbf{x}_i\right)$ selon une méthode choisie pour l'estimation des scores de propension. Deux versions de l'estimateur pondéré de la probabilité inverse (PPI) de μ_y sont établies comme suit :

$$\hat{\mu}_{\text{PPI1}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i}{\hat{\pi}_i^A} \quad \text{et} \quad \hat{\mu}_{\text{PPI2}} = \frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i}{\hat{\pi}_i^A}, \tag{4.8}$$

où N est la taille de la population et $\hat{N}^A = \sum_{i\in S_A}(\hat{\pi}_i^A)^{-1}$ est la taille estimée de la population. L'estimateur $\hat{\mu}_{\text{PPI1}}$ est une version de l'estimateur de Horvitz-Thompson et $\hat{\mu}_{\text{PPI2}}$ correspond à l'estimateur Hájek comme en fait état la théorie de l'estimation fondée sur le plan. Il y a suffisamment de données probantes des justifications théoriques et des observations pratiques que l'estimateur Hájek $\hat{\mu}_{\text{PPI2}}$ donne de meilleurs résultats que l'estimateur Horvitz-Thrompson et qu'il doit être utilisé en pratique même si la taille de la population N est connue.

La validité des estimateurs PPI $\hat{\mu}_{\text{PPI1}}$ et $\hat{\mu}_{\text{PPI2}}$ dépend de la validité des scores de propension estimés. Selon les hypothèses A1 et A2 et le modèle paramétrique $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha}_0)$, la convergence de $\hat{\mu}_{\text{PPI1}}$ suit un

argument standard en deux étapes. Soit $\tilde{\mu}_{\text{PPI}} = N^{-1}\sum_{i\in S_A} y_i/\pi_i^A$, lequel n'est pas un estimateur calculable, mais un outil analytique utile à des fins asymptotiques. Il s'ensuit que $E_q\left(\tilde{\mu}_{\text{PPI}}\right) = \mu_y$ et l'ordre $V_q\left(\tilde{\mu}_{\text{PPI}}\right) = O\left(n_A^{-1}\right)$ se vérifient à la condition que $n_A\pi_i^A/N$ ait une borne de inférieure strictement positive. Par conséquent, nous avons $\tilde{\mu}_{\text{PPI}} \to \mu_y$ en probabilité quand $n_A \to \infty$. Dans le modèle correctement indiqué $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}_0\right)$ pour les scores de propension, l'ordre type n-racine $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = O_p\left(n_A^{-1/2}\right)$ se vérifie pour les scénarios fréquemment rencontrés. Nous pouvons montrer qu'en traitant $\hat{\mu}_{\text{PPI1}}$ comme une fonction de $\hat{\boldsymbol{\alpha}}$ et en utilisant un développement en série de Taylor que $\hat{\mu}_{\text{PPI1}} = \tilde{\mu}_{\text{PPI}} + O_p\left(n_A^{-1/2}\right)$ dans certaines conditions peu contraignantes de moments finis. La convergence de $\hat{\mu}_{\text{PPI2}}$ peut être établie en utilisant des arguments standard pour un estimateur par le ratio (section 5.3, Wu et Thompson, 2020) où $N^{-1}\sum_{i\in S_A}\left(\pi_i^A\right)^{-1} = 1 + o_p(1)$.

4.3 Estimation doublement robuste

La dépendance de l'estimateur PPI sur la validité du modèle de scores de propension théorique est perçue comme une faiblesse de la méthode. Le problème n'est pas propre aux estimateurs PPI et se pose pour de nombreuses autres approches comportant un modèle statistique théorique. Les procédures robustes d'estimation qui procurent un certain degré de protection contre les erreurs de spécification du modèle ont été poursuivies par des chercheurs et lesdits estimateurs doublement robustes ont eu du succès depuis les travaux de Robins, Rotnitzky et Zhao (1994).

L'estimateur doublement robuste (DR) de μ_y est établi à l'aide du modèle de score de propension q et du modèle de régression des résultats ξ. L'estimateur DR avec les scores de propension donnés π_i^A, $i \in S_A$ et les réponses moyennes $m_i = E_\xi\left(y_i \mid \mathbf{x}_i\right)$, $i = 1, 2, \ldots, N$ présente la forme générale suivante :

$$\tilde{\mu}_{\text{DR}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i - m_i}{\pi_i^A} + \frac{1}{N}\sum_{i=1}^{N} m_i. \tag{4.9}$$

Le second terme du côté droit de l'équation (4.9) est la prédiction modéliste de μ_y. Le premier terme est un ajustement fondé sur le score de propension reposant sur les erreurs $\varepsilon_i = y_i - m_i$ du modèle de régression des résultats. L'importance du terme d'ajustement est négativement corrélée à la « qualité de l'ajustement » du modèle de régression des résultats. Il est possible de démontrer que $\tilde{\mu}_{\text{DR}}$ est un estimateur exactement sans biais de μ_y si l'un des deux modèles q et ξ est correctement précisé et ainsi doublement robuste. L'estimateur $\tilde{\mu}_{\text{DR}}$ possède une structure identique à l'estimateur par la différence généralisée de Wu et Sitter (2001). Il convient de souligner que la propriété de double robustesse de $\tilde{\mu}_{\text{DR}}$ ne nécessite pas de savoir lequel des deux modèles est correctement précisé. Il est également évident que l'estimateur $\tilde{\mu}_{\text{DR}}$ donné en (4.9) n'est pas calculable dans des applications pratiques.

Soient $\hat{\pi}_i^A$ et $\hat{m}_i$, les estimateurs de π_i^A et q respectivement, selon les modèles théoriques ξ et m_i. Dans le contexte de deux échantillons décrit à la section 2, les deux estimateurs DR de μ_y proposés par Chen et coll. (2020) sont donnés par :

$$\hat{\mu}_{\text{DR1}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \tag{4.10}$$

et

$$\hat{\mu}_{\text{DR2}} = \frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B}\sum_{i\in S_B} d_i^B \hat{m}_i, \tag{4.11}$$

où d_i^B sont les poids déterminés par le plan d'échantillonnage pour l'échantillon probabiliste S_B, $\hat{N}^A = \sum_{i\in S_A}\left(\hat{\pi}_i^A\right)^{-1}$ et $\hat{N}^B = \sum_{i\in S_B} d_i^B$. L'estimateur $\hat{\mu}_{\text{DR2}}$ fondé sur la taille estimée de la population offre un meilleur rendement en matière de biais et d'erreur quadratique moyenne et devrait être utilisé dans la pratique.

Le plan d'enquête probabiliste p fait partie intégrante du cadre théorique pour évaluer les deux estimateurs $\hat{\mu}_{\text{DR1}}$ et $\hat{\mu}_{\text{DR2}}$. On suppose que S_A et S_B sont sélectionnés indépendamment, ce qui signifie que $E_p\left(\sum_{i\in S_B} d_i^B \hat{m}_i\right) = \sum_{i=1}^N \hat{m}_i$. La convergence des estimateurs $\hat{\mu}_{\text{DR1}}$ et $\hat{\mu}_{\text{DR2}}$ peut être établie en vertu de qp ou du cadre ξp. Il convient de souligner que même l'échantillon non probabiliste S_A est un simple échantillon aléatoire avec $\pi_i^A = n_A/N$, l'estimateur doublement robuste dans la forme de (4.9) ne se réduit pas à l'estimateur de prédiction modéliste $\hat{\mu}_{y2}$ donné en (3.3).

4.4 L'approche de la pseudo-vraisemblance empirique

Les méthodes de la pseudo-vraisemblance empirique (PEL pour *Pseudo Empirical Likelihood*) pour les échantillons d'enquête probabilistes sont en cours d'élaboration depuis deux décennies. Deux articles sur le sujet sont ceux de Chen et Sitter (1999) sur l'estimation ponctuelle intégrant de l'information auxiliaire, et celui de Wu et Rao (2006) sur les intervalles de confiance du rapport de PEL. Les approches de la PEL sont en outre utilisées pour des enquêtes à bases multiples (Rao et Wu, 2010a) et les inférences de Bayesian pour des données d'enquête (Rao et Wu, 2010b; Zhao, Ghosh, Rao et Wu, 2020b). L'utilisation des méthodes de la PEL pour des problèmes inférentiels généraux avec des enquêtes complexes a été étudiée dans deux articles récents (Zhao et Wu, 2019; Zhao, Rao et Wu, 2020a).

Chen, Li, Rao et Wu (2022) ont démontré que la PEL offre une solution de rechange intéressante à l'inférence avec des échantillons d'enquête non probabilistes. Soit $\hat{\pi}_i^A$, $i\in S_A$ étant les scores de propension estimés sous un modèle paramétrique ou non paramétrique théorique, q. La fonction de PEL pour l'échantillon d'enquête non probabiliste S_A est définie comme suit :

$$\ell_{\text{PEL}}(\mathbf{p}) = n_A\sum_{i\in S_A}\tilde{d}_i^A \log\left(p_i\right), \tag{4.12}$$

où $\mathbf{p} = (p_1,\ldots,p_{n_A})$ est une mesure de probabilité discrète sur les unités n_A sélectionnées dans S_A, $\tilde{d}_i^A = (\hat{\pi}_i^A)^{-1}/\hat{N}^A$ et $\hat{N}^A = \sum_{j\in S_A}(\hat{\pi}_j^A)^{-1}$ qui est définie précédemment dans la section 4. Sans utiliser de renseignements supplémentaires, maximiser $\ell_{\text{PEL}}\left(\mathbf{p}\right)$ sous la contrainte de normalisation suivante :

$$\sum_{i \in S_A} p_i = 1 \tag{4.13}$$

donne $\hat{p}_i = \tilde{d}_i^A$, $i \in S_A$. L'estimateur du maximum de PEL de μ_y est donné par $\hat{\mu}_{\text{PEL}} = \sum_{i \in S_A} \hat{p}_i y_i$, lequel est identique à l'estimateur PPI $\hat{\mu}_{\text{PPI2}}$ donné en (4.8).

L'approche de la PEL pour les échantillons d'enquête non probabilistes procure une flexibilité en combinant des renseignements du fait de contraintes supplémentaires et en construisant des intervalles de confiance et en menant des tests d'hypothèses à l'aide de statistiques sur le rapport de PEL. L'estimateur du maximum de PEL $\hat{\mu}_{\text{PEL}} = \sum_{i \in S_A} \hat{p}_i y_i$ est doublement robuste si $\left(\hat{p}_1, \ldots, \hat{p}_{n_A}\right)$ est le maximiseur de $\ell_{\text{PEL}}(\mathbf{p})$ sous la contrainte de normalisation et la contrainte de calage de modèle donné par :

$$\sum_{i \in S_A} p_i \hat{m}_i = \bar{m}^B, \tag{4.14}$$

où $\bar{m}^B = \left(\hat{N}^B\right)^{-1} \sum_{i \in S_B} d_i^B \hat{m}_i$ est calculé à l'aide des valeurs prédites $\hat{m}_i$, $i \in S_B$ d'un modèle de régression des résultats théorique, ξ. L'équation (4.14) est une version modifiée de la contrainte de calage de modèle originale de Wu et Sitter (2001) reposant sur l'échantillon probabiliste S_B. L'étude de Chen et coll. (2022) contient d'autres données détaillées sur les répartitions asymptotiques des statistiques sur le rapport de PEL et des études de simulation sur le rendement des intervalles de confiance du rapport de PEL sur une proportion de population finie.

5. Enquête par quotas et poststratification

Les enquêtes par quotas représentent l'une des plus vieilles méthodes d'échantillonnage non probabiliste qui sont encore utilisées en pratique de nos jours. Pour une taille globale prédéterminée d'échantillon n_A, les quotas des tailles d'échantillons sont établis pour des sous-populations qui sont définies par des variables démographiques et des indicateurs de l'état socioéconomique ou d'autres variables des caractéristiques appropriées pour la population cible. Les processus de collecte des données se poursuivent jusqu'à ce que les quotas pour chacune des sous-populations soient remplis. Les unités de la population sont habituellement abordées par tous les moyens pratiques existants, et il y a peu ou pas de contrôle sur la manière de sélectionner les unités pour l'échantillon définitif autre que des quotas prédéterminés.

La théorie des estimateurs PPI pour les échantillons d'enquêtes non probabiliste offre une occasion d'examiner des scénarios où les enquêtes par quotas peuvent réussir ou échouer. Pour faciliter la notation sans perte de généralité, supposons que S_A est l'échantillon d'enquête par quotas et que $\mathbf{x}$ est l'ensemble de variables catégoriques utilisé pour définir les sous-populations et fixer les quotas. L'échantillon global peut être divisé en $S_A = S_{A1} \cup \cdots \cup S_{AK}$, ce qui correspond au classement recoupé des unités échantillonnées au moyen des combinaisons de niveaux des variables $\mathbf{x}$. Par exemple, si $\mathbf{x} = (x_1, x_2)'$, x_1 ayant deux niveaux et x_2 ayant trois niveaux, nous avons au total $K = 2 \times 3 = 6$ sous-populations définies par $\mathbf{x}$. Supposons que n_k est la taille prédéterminée de S_{Ak} et que N_k est la taille de la sous-population

correspondante. Selon l'hypothèse A1, les scores de propension $\pi_i^A = \pi(\mathbf{x}_i)$ deviennent une constante pour des unités dans la même sous-population et sont donnés par $\pi_i^A = n_k / N_k$ pour la k^{e} sous-population. L'estimateur PPI $\hat{\mu}_{\text{PPI2}}$ donnée en (4.8) se réduit à :

$$\hat{\mu}_{\text{PPI2}} = \frac{1}{\hat{N}^A} \sum_{k=1}^{K} \sum_{i \in S_{Ak}} \frac{y_i}{\hat{\pi}_i^A} = \sum_{k=1}^{K} \hat{W}_k \bar{y}_k, \tag{5.1}$$

où $\bar{y}_k = n_k^{-1} \sum_{i \in S_{Ak}} y_i$. $\hat{W}_k = \hat{N}_k / \hat{N}^A$, $\hat{N}_k$ est la taille de la k^{e} sous-population obtenue ou estimée à partir de sources externes, et $\hat{N}^A = \sum_{k=1}^{K} \hat{N}_k$. Dans les conditions actuelles et compte tenu de la disponibilité d'un échantillon probabiliste de référence S_B, nous formons la même subdivision, recoupée selon les niveaux de $\mathbf{x}$ et obtenons $S_B = S_{B1} \cup \cdots \cup S_{BK}$. Nous pouvons alors utiliser $\hat{N}_k = \sum_{i \in S_{Bk}} d_i^B$.

L'estimateur donné en (5.1) est l'estimateur stratifié *a posteriori* standard de μ_y. Il nécessite l'information sur les « poids de strate » $\hat{W}_k$, $k = 1, \ldots, K$, qui n'est pas disponible à partir des données échantillon elles-mêmes. Les enquêtes par quotas, associées à l'utilisation de l'estimateur stratifié *a posteriori*, peuvent produire avec succès des estimations de population valides pour la variable de l'étude y si les conditions suivantes sont maintenues :

i) Les variables catégoriques $\mathbf{x}$ utilisées pour définir les sous-populations et établir des quotas offrent des caractérisations du comportement de participation des unités pour les enquêtes à participation volontaire.

ii) L'inclusion d'unités dans l'enquête est quelque peu aléatoire au sein de chaque sous-population et aucun groupe précis n'est exclu intentionnellement de l'enquête.

iii) L'information sur les poids de strate correspondant aux classements recoupés dans l'établissement des quotas peut être obtenue de manière fiable de sources externes.

iv) Les non-répondants irréductibles dans la population qui ne remplissent jamais d'enquêtes à participation volontaire présentent des traits semblables à ceux des répondants en ce qui a trait à la variable étudiée y.

Les estimateurs PPI $\hat{\mu}_{\text{PPI1}}$ et $\hat{\mu}_{\text{PPI2}}$ donnés en (4.8) peuvent être sensibles à des petites valeurs de scores de propension estimés. L'estimateur stratifié *a posteriori* dans la forme de (5.1) sert de solution de rechange fiable dans des scénarios généraux où la dimension de $\mathbf{x}$ n'est pas faible et où certaines composantes de $\mathbf{x}$ sont continues. Les strates K sont formées en fonction de groupes homogènes en matière de scores de propension. Supposons que $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}})$, $i \in S_A$ est calculé en fonction d'un modèle paramétrique, q. Supposons également que $n_A = m_A K$ avec le K choisi où m_A est un nombre entier. Supposons que $\hat{\pi}_{(1)}^A \leq \cdots \leq \hat{\pi}_{(n_A)}^A$ représente les scores de propension estimés en ordre croissant. Supposons que S_{A1} est l'ensemble des premières unités m_A dans la séquence, S_{A2} représente les deuxièmes unités m_A dans la séquence, et ainsi de suite. L'estimateur stratifié *a posteriori* de μ_y est calculé comme $\hat{\mu}_{\text{PST}} = \sum_{k=1}^{K} \hat{W}_k \bar{y}_k$, qui a la même forme que l'estimateur donné en (5.1). Les estimations des poids de strate $\hat{W}_k$, $k = 1, 2, \ldots, K$ peuvent être obtenues en utilisant l'échantillon probabiliste de

référence S_B comme suit. Supposons que $b_k = \max\left\{\hat{\pi}_i^A : i \in S_{Ak}\right\}$, $k = 1, 2, \ldots, K-1$. Supposons que $b_0 = 0$ et que $b_K = 1$.

a) Calculer $\hat{\pi}_i = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}})$, $i \in S_B$.

b) Définir $S_{Bk} = \left\{i \,\middle|\, i \in S_B, b_{k-1} < \hat{\pi}_i \le b_k\right\}$, $k = 1, 2, \ldots, K$.

c) Calculer $\hat{N}_k = \sum_{i \in S_{Bk}} d_i^B$, $k = 1, 2, \ldots, K$.

Il est évident que $S_B = S_{B1} \cup \cdots \cup S_{BK}$ et $\sum_{k=1}^K \hat{N}_k = \hat{N}^B = \sum_{i \in S_B} d_i^B$. Les poids de strate estimés sont donnés par $\hat{W}_k = \hat{N}_k \big/ \hat{N}^B$.

Le choix de K doit refléter l'équilibre entre l'homogénéité des unités au sein de chaque strate *a posteriori* (en matière de scores de propension) et la stabilité de l'estimateur stratifié *a posteriori* (en matière de tailles d'échantillons de strate). Quand la taille de l'échantillon n_A est petite ou modérée, il faut utiliser un petit nombre comme $K = 5$. Pour des scénarios où n_A est grand, un plus grand K doit être utilisé de sorte que les unités au sein du même échantillon stratifié *a posteriori* S_{Ak} ont des scores de propension estimés similaires. Un guide pratique pour le choix de K est de s'assurer que $m_A \ge 30$ pour les échantillons stratifiés *a posteriori*. Pour ceux qui sont assez vieux, vous souvenez-vous du bon vieux temps quand « la taille de l'échantillon est grande » signifiait « $n \ge 30$ » ?

6. Estimation de la variance

L'estimation de la variance dans la configuration de deux échantillons S_A et S_B comporte au moins deux sources différentes de variation. Le plan d'échantillonnage avec probabilités pour l'échantillon de référence S_B reste l'une des sources employées, et ce, indépendamment des approches utilisées pour les échantillons d'enquête non probabilistes. L'estimation de la composante de variance provenant de l'utilisation de S_B nécessite des formules appropriées d'approximation de la variance ou des poids de rééchantillonnage comme faisant partie de l'ensemble de données de l'échantillon probabiliste de référence. Dans la présente section, notre analyse suppose que l'estimateur de variance fondé sur le plan pour l'estimateur ponctuel pondéré de l'enquête fondé sur S_B est disponible.

6.1 Estimation de la variance pour les estimateurs d'imputation massive

L'estimation de la variance pour l'estimateur de prédiction basé sur un modèle $\hat{\mu}_y$ comporte d'abord le calcul de la formule de la variance asymptotique pour $\mathrm{Var}\left(\hat{\mu}_y - \mu_y\right)$ selon le modèle de régression des résultats théorique ou le modèle d'imputation ξ et le plan d'échantillonnage avec probabilités p, et ensuite l'utilisation des diverses quantités inconnues de population.

L'estimateur d'imputation massive $\hat{\mu}_{y\mathrm{IM}} = \hat{N}_B^{-1} \sum_{i \in S_B} d_i^B y_i^*$ donné en (3.5) est un type spécial d'estimateur de prédiction basé sur un modèle, où le modèle ξ fait référence à celui utilisé pour l'imputation et n'est pas nécessairement le même que le modèle de régression des résultats. La méthode de

l'imputation joue un rôle clé dans le calcul de la formule de la variance asymptotique et l'estimateur de la variance doit être établi en conséquence. Il convient de souligner que $\hat{\mu}_{y\mathrm{IM}}$ est un estimateur de type Hájek en raison de l'utilisation de la taille de population estimée $\hat{N}_B$ et que les calculs de la formule de la variance asymptotique commencent en mettant la valeur réelle N en premier et en traitant ensuite $\hat{\mu}_{y\mathrm{IM}}$ comme un estimateur par le ratio. Kim et coll. (2021) ont examiné l'estimation de la variance pour $\hat{\mu}_y = N^{-1}\sum_{i\in S_B} d_i^B y_i^*$, où $y_i^* = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$ est la valeur imputée pour y_i selon le modèle semi-paramétrique (3.1). La formule de la variance asymptotique est élaborée en deux étapes. Premièrement, une version linéarisée de $\hat{\mu}_y$ est obtenue en utilisant un développement en série de Taylor à $\boldsymbol{\beta}^*$, où $\boldsymbol{\beta}^*$ est la limite de probabilité de $\hat{\boldsymbol{\beta}}$ de sorte que $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p\left(n_A^{-1/2}\right)$. Deuxièmement, deux composantes de variance sont calculées pour $\mathrm{Var}\left(\hat{\mu}_y - \mu_y\right)$ en fonction de la version linéarisée en utilisant le modèle semi-paramétrique (3.1) et le plan d'échantillonnage pour S_B. Le processus est fastidieux, ce qui est le cas pour la plupart des méthodes d'estimation de la variance basées sur un modèle. Un estimateur de la variance bootstrap se révèle être plus attrayant pour des applications pratiques. Voir Kim et coll. (2021) pour obtenir d'autres renseignements détaillés.

6.2 Estimation de la variance pour les estimateurs PPI

L'estimateur PPI couramment utilisé $\hat{\mu}_{\mathrm{PPI2}}$ donné en (4.8) est valide dans le modèle théorique q pour les scores de propension. Une formule de variance asymptotique pour $\hat{\mu}_{\mathrm{PPI2}}$ peut être obtenue selon le cadre qp conjoint quand les scores de propension sont estimés à l'aide de la méthode du pseudo-maximum de vraisemblance ou d'une méthode fondée sur une équation d'estimation comme en fait état la section 4.1. L'outil théorique est la formule de variance de type sandwich pour les estimateurs ponctuels définis comme la solution à un système combiné d'équations d'estimation pour μ_y et $\boldsymbol{\alpha}_0$.

Examinons la forme paramétrique $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ pour les scores de propension, où les paramètres du modèle $\boldsymbol{\alpha}$ sont estimés à l'aide des équations d'estimation (4.4) avec des fonctions personnalisées $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. La première étape majeure dans le calcul de la formule de la variance asymptotique pour $\hat{\mu}_{\mathrm{PPI2}}$ est de formuler le système des équations d'estimation conjointes pour μ_y et $\boldsymbol{\alpha}_0$. Soit $\boldsymbol{\eta} = \left(\mu, \boldsymbol{\alpha}'\right)'$, le vecteur des paramètres combinés. L'estimateur $\hat{\boldsymbol{\eta}} = \left(\hat{\mu}_{\mathrm{PPI2}}, \boldsymbol{\alpha}'\right)'$ est la solution au système d'équations d'estimation conjointes $\boldsymbol{\Phi}_n\left(\boldsymbol{\eta}\right) = \mathbf{0}$, où :

$$\boldsymbol{\Phi}_n\left(\boldsymbol{\eta}\right) = \begin{pmatrix} N^{-1}\sum_{i=1}^N R_i\left(y_i - \mu\right)/\pi_i^A \\ N^{-1}\sum_{i=1}^N R_i \mathbf{h}\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) - N^{-1}\sum_{i\in S_B} d_i^B \pi_i^A \mathbf{h}\left(\mathbf{x}_i, \boldsymbol{\alpha}\right) \end{pmatrix}. \tag{6.1}$$

Le facteur N^{-1} est redondant, mais utile pour faciliter les ordres asymptotiques. Les fonctions d'estimation définie par (6.1) sont sans biais selon le cadre qp conjoint, c'est-à-dire $E_{qp}\left\{\boldsymbol{\Phi}\left(\boldsymbol{\eta}_0\right)\right\} = \mathbf{0}$, où $\boldsymbol{\eta}_0 = \left(\mu_y, \boldsymbol{\alpha}_0'\right)'$. Il y a deux conséquences majeures à l'absence de biais du système d'équations d'estimation. Premièrement, on peut soutenir la convergence de l'estimateur $\hat{\boldsymbol{\eta}}$ en utilisant la théorie des fonctions d'estimation générales semblables à celles présentées à la section 3.2 de Tsiatis (2006).

Deuxièmement, la matrice de covariance-variance asymptotique de $\hat{\boldsymbol{\eta}}$, désignée comme $\mathrm{AV}(\hat{\boldsymbol{\eta}})$, a la forme standard en sandwich et est donnée par :

$$\mathrm{AV}(\hat{\boldsymbol{\eta}}) = \left[E\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\}\right]^{-1} \mathrm{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\}\left[E\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\}'\right]^{-1},$$

où $\boldsymbol{\phi}_n(\boldsymbol{\eta}) = \partial\boldsymbol{\Phi}_n(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$, laquelle dépend des formes de $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ et $\mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha})$. Le terme $\mathrm{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\}$ comporte deux composantes, une en raison du modèle du score de propension q et l'autre en raison du plan d'échantillonnage avec probabilités pour S_B. Plus particulièrement, nous avons $\mathrm{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\} = V_q(\mathbf{A}_1) + V_p(\mathbf{A}_2)$, où $V_q(\cdot)$ désigne la variance sous le modèle du score de propension q et $V_p(\cdot)$ représente la variance fondée sur le plan sous le plan d'échantillonnage avec probabilités p, et

$$\mathbf{A}_1 = \frac{1}{N}\sum_{i=1}^{N} R_i \begin{pmatrix} (y_i - \mu)/\pi_i^A \\ \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) \end{pmatrix}, \quad \mathbf{A}_2 = \frac{1}{N}\sum_{i\in S_B} d_i^B \begin{pmatrix} 0 \\ \pi_i^A \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) \end{pmatrix}.$$

L'expression analytique pour $V_q(\mathbf{A}_1)$ découle immédiatement de $V_q(R_i) = \pi_i^A(1-\pi_i^A)$ et l'indépendance parmi $R_1, \ldots, R_N$. La composante de variance fondée sur le plan $V_p(\mathbf{A}_2)$ nécessite des renseignements supplémentaires sur le plan d'enquête pour S_B ou une formule d'approximation de la variance appropriée avec le plan donné.

La formule de la variance asymptotique pour l'estimateur PPI $\hat{\mu}_{\mathrm{PPI2}}$ est le premier élément en diagonale de la matrice $\mathrm{AV}(\hat{\boldsymbol{\eta}})$. L'estimateur définitif de la variance pour $\hat{\mu}_{\mathrm{PPI2}}$ peut alors être obtenu en remplaçant diverses quantités de population par des estimateurs des moments fondés sur l'échantillon. Dans leur étude, Chen et coll. (2020) ont présenté l'estimateur de la variance avec des expressions explicites quand $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ sont modélisés par la régression logistique et $\hat{\boldsymbol{\alpha}}$ est obtenu par la méthode du pseudo-maximum de vraisemblance.

6.3 Estimation de la variance pour les estimateurs doublement robustes

Il s'avère que l'estimation de la variance pour l'estimateur doublement robuste est un problème difficile. Alors que la double robustesse est une propriété souhaitable pour l'estimation ponctuelle, elle pose un dilemme pour l'estimation de la variance. L'estimateur $\hat{\mu}_{\mathrm{DR2}}$ donné en (4.11) est convergent si le modèle de score de propension q ou le modèle de régression des résultats ξ est correctement précisé. Il n'existe aucun besoin de savoir quel modèle est correctement précisé, ce qui constitue la partie la plus cruciale derrière la double robustesse. Toutefois, cette caractéristique ambiguë devient un problème pour une estimation de la variance. La formule de la variance asymptotique selon le modèle q est habituellement différente de celle selon le modèle ξ et, par conséquent, il est difficile de concevoir un estimateur de variance cohérent avec des scénarios inconnus sur les spécifications du modèle.

Plusieurs stratégies ont été proposées dans les ouvrages portant sur l'estimation de la variance pour les estimateurs doublement robustes. Une approche naïve consiste à utiliser l'estimateur de la variance dérivé dans le modèle de score de propension théorique q et de prendre le risque qu'un tel estimateur de la

variance puisse avoir des biais non négligeables dans le modèle de régression des résultats. Une bonne nouvelle est que, dans le modèle de score de propension, l'estimation des paramètres $\boldsymbol{\beta}$ pour le modèle de régression des résultats n'a aucune incidence asymptotiquement sur la variance des estimateurs doublement robustes. Nous pouvons observer cela en utilisant $\hat{\mu}_{\text{DR1}}$ de (4.10) comme un exemple. Soit $\hat{m}_i = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$, où $\hat{\boldsymbol{\beta}}$ est obtenu en fonction du modèle de travail (3.1) qui n'est pas nécessairement correct. Soit $\boldsymbol{\beta}^*$, la limite de probabilité de $\hat{\boldsymbol{\beta}}$ de sorte que $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p\left(n_A^{-1/2}\right)$ indépendamment du vrai modèle de régression des résultats (White, 1982). Supposons que $m_i^* = m\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)$ et que $\mathbf{a}(\mathbf{x}, \boldsymbol{\beta}) = \partial m(\mathbf{x}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$. Nous pouvons constater que :

$$\frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \;-\; \frac{1}{N}\sum_{i\in S_A} \frac{\hat{m}_i}{\hat{\pi}_i^A} \;=\; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \;-\; \frac{1}{N}\sum_{i\in S_A} \frac{m_i^*}{\hat{\pi}_i^A} \;+\; \left\{\mathbf{B}\left(\boldsymbol{\beta}^*\right)\right\}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \;+\; o_p\left(n_A^{-1/2}\right),$$

où

$$\mathbf{B}\left(\boldsymbol{\beta}^*\right) \;=\; \frac{1}{N}\sum_{i\in S_B} d_i^B \mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right) \;-\; \frac{1}{N}\sum_{i\in S_A} \frac{\mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)}{\hat{\pi}_i^A}. \tag{6.2}$$

Comme les deux termes sur le côté droit de (6.2) sont tous les deux des estimateurs de $N^{-1}\sum_{i=1}^N \mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)$, nous concluons que $\mathbf{B}\left(\boldsymbol{\beta}^*\right) = o_p(1)$ et

$$\frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \;-\; \frac{1}{N}\sum_{i\in S_A} \frac{\hat{m}_i}{\hat{\pi}_i^A} \;=\; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \;-\; \frac{1}{N}\sum_{i\in S_A} \frac{m_i^*}{\hat{\pi}_i^A} \;+\; o_p\left(n_A^{-1/2}\right).$$

Il s'ensuit que :

$$\hat{\mu}_{\text{DR1}} \;=\; \frac{1}{N}\sum_{i\in S_A} \frac{y_i - m_i^*}{\hat{\pi}_i^A} \;+\; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \;+\; o_p\left(n_A^{-1/2}\right).$$

Les mêmes arguments s'appliquent à $\hat{\mu}_{\text{DR2}}$. Nous pouvons traiter $\hat{\boldsymbol{\beta}}$, car il est fixe dans le calcul de la variance asymptotique pour $\hat{\mu}_{\text{DR1}}$ et $\hat{\mu}_{\text{DR2}}$ sous le modèle de score de propension théorique. Les techniques décrites à la section 6.2 peuvent être utilisées directement lorsque la première fonction d'estimation dans (6.1) est remplacée par celle servant à définir $\hat{\mu}_{\text{DR1}}$ ou $\hat{\mu}_{\text{DR2}}$. Voir le théorème 2 de Chen et coll. (2020) pour obtenir d'autres renseignements détaillés. Toutefois, l'estimateur de la variance calculé selon le modèle du score de propension théorique présente généralement un biais dans le modèle de régression des résultats.

Dans l'étude de Chen et coll. (2020), on a également décrit une technique faisant appel à une idée originale présentée par Kim et Haziza (2014) pour la construction du soi-disant estimateur de variance doublement robuste. Cette technique est délicate et présente un certain attrait théorique, mais pose divers problèmes pour des applications pratiques. Nous utilisons $\hat{\mu}_{\text{DR1}}$ comme un exemple pour illustrer les étapes de la construction de l'estimateur de variance doublement robuste. Supposons que :

$$\hat{\mu}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} R_i \frac{y_i - m(\mathbf{x}_i,\boldsymbol{\beta})}{\pi(\mathbf{x}_i,\boldsymbol{\alpha})} + \frac{1}{N}\sum_{i\in S_B} d_i^B m(\mathbf{x}_i,\boldsymbol{\beta}).$$

Il s'ensuit que $\hat{\mu}_{\text{DR1}} = \hat{\mu}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$ si $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ sont les méthodes d'estimation originales. La première étape consiste à modifier l'estimation de $\boldsymbol{\alpha}$ et de $\boldsymbol{\beta}$ de manière à obtenir $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ comme solutions à :

$$\frac{\partial\hat{\mu}(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\boldsymbol{\alpha}} = \mathbf{0} \quad \text{et} \quad \frac{\partial\hat{\mu}(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \mathbf{0}. \tag{6.3}$$

Dans le modèle de régression logistique q où $\text{logit}\{\pi(\mathbf{x}_i,\boldsymbol{\alpha})\} = \mathbf{x}_i'\boldsymbol{\alpha}$ et le modèle de régression linéaire ξ où $m(\mathbf{x}_i,\boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, le système d'équations (6.3) devient :

$$\frac{1}{N}\sum_{i=1}^{N} R_i \left\{\frac{1}{\pi(\mathbf{x}_i,\boldsymbol{\alpha})} - 1\right\}(y_i - \mathbf{x}_i'\boldsymbol{\beta})\,\mathbf{x}_i = \mathbf{0}, \tag{6.4}$$

$$\frac{1}{N}\sum_{i=1}^{N} \frac{R_i\mathbf{x}_i}{\pi(\mathbf{x}_i,\boldsymbol{\alpha})} - \frac{1}{N}\sum_{i\in S_B} d_i^B\mathbf{x}_i = \mathbf{0}. \tag{6.5}$$

Les équations d'estimation dans (6.5) sont sans biais selon le cadre qp conjoint. Elles sont identiques à l'équation (4.5) abordée à la section 4.1.2. Les équations d'estimation en (6.4) sont également sans biais dans le modèle de régression des résultats, mais elles sont différentes des équations de quasi-score qui figurent à l'équation (3.2). Les estimateurs $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ obtenus comme solutions à (6.4) et (6.5) sont moins stables que ceux des méthodes standard. En outre, le système d'équations (6.4) et (6.5) n'aura pas de solution si $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ n'ont pas la même dimension, car le nombre d'équations en (6.4) est déterminé par la dimension de $\boldsymbol{\alpha}$ et le nombre d'équations en (6.5) est le même que la dimension de $\boldsymbol{\beta}$. L'estimateur définitif $\hat{\mu}_{\text{DR}} = \hat{\mu}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$ subit également des pertes d'efficacité quand $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ sont estimés en résolvant (6.4) et (6.5).

La raison expliquant l'utilisation du système d'équations (6.3) est purement technique. Cela peut être démontré par une expansion de premier ordre de Taylor que les estimateurs $\hat{\boldsymbol{\alpha}}$ et $\hat{\boldsymbol{\beta}}$ tirés de (6.3) n'ont pas d'effet asymptotiquement sur la variance de $\hat{\mu}_{\text{DR}} = \hat{\mu}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$. Cette manœuvre technique permet que des expressions explicites simples pour la variance $V_{qp}(\hat{\mu}_{\text{DR}})$ selon le cadre qp et pour la variance de prédiction $V_{\xi p}(\hat{\mu}_{\text{DR}} - \mu_y)$ selon le cadre ξp puissent aisément être obtenues. La construction d'un estimateur de variance doublement robuste pour $\hat{\mu}_{\text{DR}}$ commence par l'estimateur par substitution pour $V_{qp}(\hat{\mu}_{\text{DR}})$ dans le modèle de scores de propension q. Un terme de correction de biais est ensuite ajouté pour obtenir un estimateur valide pour $V_{\xi p}(\hat{\mu}_{\text{DR}} - \mu_y)$ dans le modèle de régression des résultats ξ. L'aspect positif est que le terme de correction du biais possède une forme analytique $N^{-2}\sum_{i=1}^{N}(R_i/\pi_i^A - 1)\sigma_i^2$ où $\sigma_i^2 = E_\xi(y_i \mid \mathbf{x}_i)$, lequel est négligeable dans le modèle de scores de propension q. L'estimateur de variance corrigé pour le biais est valide dans le modèle de scores de propension ou le modèle de régression des résultats.

Un estimateur de variance doublement robuste pour le $\hat{\mu}_{\mathrm{DR}2}$ couramment utilisé ne figure pas dans la littérature. Une solution pratique est d'utiliser des méthodes bootstrap. Chen et coll. (2022) ont démontré que les procédures bootstrap standard avec remise appliquées séparément à S_A et à S_B procuraient des intervalles de confiance doublement robustes en utilisant une approche fondée sur la pseudo-vraisemblance empirique à des échantillons d'enquête non probabilistes quand l'échantillon de référence est sélectionné par des plans d'échantillonnage avec probabilités inégales à un seul degré. Des complications surviennent quand l'échantillon probabiliste S_B s'appuie sur des méthodes d'échantillonnage à plusieurs degrés, un défi connu pour l'estimation de la variance dans le cadre d'enquêtes complexes. La construction d'estimateurs de variance doublement robustes pour l'estimateur doublement robuste $\hat{\mu}_{\mathrm{DR}2}$ dans des conditions générales mérite que l'on y consacre des efforts dans de futurs travaux de recherche.

7. Hypothèses réexaminées

Nos commentaires portant sur les procédures d'estimation pour des échantillons d'enquête non probabilistes figurent dans la section sur les hypothèses A1 à A4 et l'accent est mis sur la validité et l'efficacité des estimateurs pour la moyenne de population finie selon trois cadres inférentiels. Les résultats théoriques sur la prédiction fondée sur un modèle, la pondération de probabilité inverse et une estimation doublement robuste ont été rigoureusement établis dans le cadre de ces hypothèses. Il semble que les chercheurs ont du succès lorsqu'ils traitent du domaine émergent des sources de données non probabilistes. Toutefois, comme l'a souligné le président de l'ASA 2021 Robert Santos dans sa tribune intitulée « Utiliser nos superpouvoirs pour contribuer à l'intérêt public » (en anglais, Armstat News, mai 2021), « Nos superpouvoirs ne sont aussi fiables que leurs hypothèses sous-jacentes, hypothèses qui sont trop souvent acceptées avec assurance, sans pourtant pouvoir être prouvées » [traduction]. La façon de vérifier les hypothèses A1 à A4 dans des applications pratiques des méthodes est une question à laquelle on ne peut jamais répondre complètement et ils restent des étapes à suivre pour renforcer la confiance en utilisant les résultats théoriques. Il est également important de comprendre les conséquences possibles quand certaines hypothèses deviennent très discutables.

7.1 Hypothèse A1

L'hypothèse A1 stipule que $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \middle| \mathbf{x}_i\right)$. Il s'agit de l'hypothèse la plus importante pour la validité de l'estimateur du pseudo-maximum de vraisemblance de Chen et coll. (2020) et l'estimateur obtenu par lissage par la méthode du noyau non paramétrique présenté à la section 4.1.3 pour les scores de propension, bien que toutes les autres hypothèses soient également en cause. Elle est équivalente à l'hypothèse des données manquantes au hasard pour les ouvrages publiés sur les données manquantes. Il est bien compris que l'hypothèse des données manquantes au hasard ne peut être testée à l'aide des données elles-mêmes de l'échantillon. Le même énoncé vaut pour l'hypothèse A1 comportant des échantillons d'enquête non probabilistes.

En résumé, l'hypothèse A1 indique que les variables auxiliaires $\mathbf{x}$ comprises dans l'échantillon non probabiliste caractérisent complètement le comportement de participation ou le mécanisme d'inclusion des échantillons pour des unités dans la population. Une attention suffisante doit être donnée à l'étape de conception de l'étude avant la collecte des données, si une telle étape existe, afin d'examiner les facteurs et les caractéristiques potentiels des unités qui peuvent être liées à la participation et à l'inclusion des échantillons. Pour des populations humaines, les facteurs et les caractéristiques peuvent comprendre des variables démographiques, des indicateurs sociaux et économiques ainsi que des variables géographiques.

L'hypothèse A1 mène à la conclusion que la répartition conditionnelle de y étant donné $\mathbf{x}$ pour des unités de l'échantillon non probabiliste est la même que la répartition conditionnelle de y étant donné $\mathbf{x}$ pour les unités de la population cible. Cela suppose que les variables auxiliaires $\mathbf{x}$ doivent comprendre des prédicteurs pertinents pour la variable étudiée y. Pour ce qui est des ensembles de données établis S_A et S_B, une analyse de sensibilité par comparaison des répartitions marginales et des modèles conditionnels peut être utile pour renforcer la confiance à l'égard de l'hypothèse A1. Pour les variables qui sont présentées dans S_A et S_B, on peut comparer les fonctions de répartition empirique (ou moments) de S_A aux fonctions de répartition empirique pondérés de l'enquête (ou moments) de S_B. Des différences marquées entre les deux indiquent que S_A est un échantillon non probabiliste avec des scores de propension inégaux. Une analyse de sensibilité possible sur l'hypothèse A1 consiste à sélectionner une variable z qui présente certaines similarités avec y, et un ensemble de variables auxiliaires $\mathbf{u}$ avec z et $\mathbf{u}$ disponibles de S_A et de S_B. Nous avons ajusté un modèle conditionnel $z \mid \mathbf{u}$ en utilisant des données de S_A et un modèle conditionnel pondéré d'enquête $z \mid \mathbf{u}$ en utilisant des données de S_B. Si $\mathbf{u}$ comprend toutes les variables auxiliaires principales pour l'hypothèse A1, nous devrons voir que les deux versions des modèles ajustés sont semblables. D'importantes différences entre les deux modèles ajustés représentent un fort signe que le z est lui-même une variable auxiliaire importante de l'hypothèse A1 ou que l'hypothèse est discutable.

7.2 Hypothèse A2

Un coup d'œil à l'hypothèse A2 peut donner à penser qu'elle pourrait facilement être satisfaite en pratique, car une hypothèse similaire est largement utilisée dans les analyses sur les données manquantes et les inférences causales. Il s'avère que l'hypothèse peut se révéler très problématique et pour des scénarios où l'hypothèse ne peut se vérifier, la population cible est différente de celle supposée pour les méthodes d'estimation. Elle est semblable au sous-dénombrement incomplet de la base de sondage et à des problèmes de non-réponse, lesquels sont longuement étudiés dans l'échantillonnage probabiliste.

L'hypothèse A2 stipule que $\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i, y_i\right) > 0$ pour tous les i. Elle équivaut à déclarer que chaque unité de la population cible présente une probabilité non nulle d'être incluse dans l'échantillon non probabiliste. Si l'échantillon était prélevé par une méthode d'échantillonnage probabiliste, il s'agirait d'un scénario où la base de sondage est complète et où il n'y a pas de non-répondants irréductibles. Pour la plupart des échantillons non probabilistes, le concept de « *base de sondage* » est souvent non pertinent ou

simplement, une liste pratique et le processus de sélection et d'inclusion d'unités pour l'échantillon peut se révéler non structuré. Dans sa présentation à l'atelier CANSSI-NISS de 2021, Mary Thompson mentionnait que « l'énoncé que l'indicateur d'inclusion d'un échantillon R est une variable aléatoire est lui-même une hypothèse » [traduction] pour les échantillons d'enquête non probabilistes.

Supposons que U est l'ensemble de N unités pour la population cible. Soit $U_0 = \left\{ i \,\middle|\, i \in U \text{ et } \pi_i^A > 0 \right\}$. Il est évident que $U_0 \subset U$ et $U_0 \neq U$ quand l'hypothèse A2 n'est pas respectée. On retrouve deux scénarios types en pratique. On peut appeler le premier un *sous-dénombrement stochastique*, où l'échantillon non probabiliste S_A est sélectionné d U_0 et U_0 lui-même peut être perçu comme un échantillon aléatoire de U. Par exemple, la liste de personnes à joindre d'une enquête probabiliste existante est utilisée pour approcher des unités de la population pour une participation dans l'échantillon non probabiliste. Dans ce cas, U_0 est constitué d'unités de l'échantillon probabiliste. Un autre exemple est une enquête auprès de volontaires où la population cible est constituée d'adultes dans une ville ou une région particulière, mais où les participants sont recrutés parmi les visiteurs de grands centres commerciaux de la région au cours d'une période donnée. La sous-population U_0 comprend des visiteurs aux emplacements choisis au cours de la période d'échantillonnage et il est raisonnable de supposer que U_0 est un échantillon aléatoire de la population cible. Soit $D_i = 1$, si $i \in U_0$ et $D_i = 0$ autrement, $i = 1, 2, \ldots, N$. Nous observons :

$$P\left(R_i = 1 \,\middle|\, \mathbf{x}_i, y_i, D_i = 1\right) > 0 \quad \text{et} \quad P\left(R_i = 1 \,\middle|\, \mathbf{x}_i, y_i, D_i = 0\right) = 0$$

pour $i = 1, 2, \ldots, N$. Si la sous-population U_0 est formée avec un mécanisme stochastique sous-jacent tel que $P\left(D_i = 1 \,\middle|\, \mathbf{x}_i, y_i\right) > 0$ pour tous les $i \in U$, nous avons :

$$\pi_i^A = P\left(R_i = 1 \,\middle|\, \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \,\middle|\, \mathbf{x}_i, y_i, D_i = 1\right) P\left(D_i = 1 \,\middle|\, \mathbf{x}_i, y_i\right) > 0$$

pour $i = 1, 2, \ldots, N$. En d'autres mots, l'hypothèse A2 est valide selon le scénario du sous-dénombrement stochastique pour les échantillons non probabilistes.

Le second scénario est appelé *sous-dénombrement déterministe* où les unités présentant certaines caractéristiques ne seront jamais incluses dans l'échantillon non probabiliste. Supposons que la participation dans l'enquête non probabiliste nécessite un accès à Internet et une adresse de courriel valide, et que 20 % de la population n'a pas accès à Internet ni à une adresse de courriel, nous avons un exemple où 20 % de la population aura des scores de propension nuls. Il n'existe pas de solution simple aux procédures inférentielles élaborées sous l'hypothèse A2. La thèse de doctorat de Yilin Chen de l'université de Waterloo (Chen, 2020) contenait un chapitre traitant de certains aspects précis du scénario.

7.3 Hypothèse A3

Parmi toutes les hypothèses, celle-ci est la moins essentielle à la validité des procédures inférentielles proposées. Dans l'hypothèse A3, la fonction de vraisemblance complète pour les scores de propension est

donnée dans (4.1). Pour tout modèle paramétrique sur $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$, la fonction de logarithme du rapport de quasi-vraisemblance $\ell^*(\boldsymbol{\alpha})$ donnée en (4.2) mène aux fonctions de quasi-scores $\mathbf{U}(\boldsymbol{\alpha}) = \partial \ell^*(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$, lesquelles restent sans biais même si l'hypothèse A3 n'est pas respectée. Il pourrait y avoir certaines pertes d'efficacité sans l'hypothèse A3 lors de l'estimation des paramètres du modèle $\boldsymbol{\alpha}$, mais les méthodes d'estimation sont encore valides dans les trois autres hypothèses.

7.4 Hypothèse A4

Il n'est pas difficile de trouver un échantillon probabiliste existant de la même population cible. Toutefois, il peut s'avérer très difficile d'avoir un échantillon d'enquête probabiliste qui contient les variables auxiliaires souhaitables. Des enquêtes probabilistes existantes sont conçues à des fins précises et des objectifs scientifiques, et les variables auxiliaires comprises dans l'enquête ne sont pas nécessairement adaptées à l'analyse d'un échantillon d'enquête non probabiliste particulier. L'objectif ultime pour satisfaire l'hypothèse A4 est d'établir et d'avoir accès à un échantillon d'enquête probabiliste existant avec une riche collecte de variables démographiques, d'indicateurs sociaux et économiques et de variables géographiques.

Un problème des personnes riches (quand une personne a trop d'argent) pour l'hypothèse A4 peut également survenir quand au moins deux échantillons d'enquête probabilistes sont disponibles. La façon de combiner toutes ces données pour réaliser une analyse efficace des échantillons d'enquête non probabilistes est un sujet de recherche qui mérite plus d'attention. Certains conseils pratiques sur le choix d'un échantillon probabiliste de référence de solutions de rechange possibles comprennent les éléments suivants.

i) Vérifier la disponibilité de variables auxiliaires importantes qui présentent un intérêt pour la caractérisation du comportement de participation ou pour donner un pouvoir prédictif à l'étude des variables dans l'échantillon non probabiliste;

ii) Accorder la priorité à l'un d'un plus grand ensemble de variables qui sont communes à l'échantillon non probabiliste;

iii) Attribuer une seconde préférence à l'échantillon probabiliste ayant une plus grande taille d'échantillon;

(iv) Finalement, utiliser l'échantillon probabiliste pour lequel le mode de collecte des données est le même que celui pour l'échantillon non probabiliste.

Chen et coll. (2020) ont démontré que deux échantillons d'enquête probabilistes de référence ayant le même ensemble de variables auxiliaires communes ont tendance à produire des estimateurs PPI très similaires, mais que celui ayant la plus grande taille d'échantillon mène à de meilleurs estimateurs d'imputation massive.

8. Conclusions

Au début du 21e siècle, les enquêtes en ligne ont commencé à être populaires, ce qui a suscité un intérêt important pour la recherche sur ce sujet (Tourangeau, Conrad et Couper, 2013). Les enjeux et les défis posés par les échantillons d'enquête en ligne et autres enquêtes non probabilistes ont mené à la publication de « Summary Report of the AAPOR Task Force on Non-probability Sampling » par Baker, Brick, Bates, Bataglia, Couper, Dever, Gile et Tourangeau (2013). Entre autres, le rapport indiquait que i) contrairement à l'échantillonnage probabiliste, il n'y avait pas de cadre unique qui englobait adéquatement tous les échantillonnages non probabilistes; ii) faire des inférences pour toute enquête probabiliste ou non probabiliste nécessitait de s'appuyer sur des hypothèses de modélisation; iii) si les échantillons non probabilistes devaient être mieux acceptés chez les spécialistes de la recherche sur les enquêtes, il devait y avoir un cadre plus cohérent et des ensembles de mesures d'accompagnement pour évaluer leur qualité.

Les spécialistes de l'échantillonnage ont répondu à l'appel en menant des études accrues sur l'inférence statistique avec des échantillons d'enquête non probabilistes. La configuration actuelle de deux échantillons S_A et S_B, avec l'échantillon non probabiliste S_A ayant des mesures sur la variable étudiée y et les variables auxiliaires $\mathbf{x}$ et l'échantillon probabiliste S_B fournissant des renseignements sur $\mathbf{x}$, a été examiné en premier par Rivers (2007) sur un échantillon correspondant à l'imputation du plus proche voisin, laquelle est l'idée originale menant à la méthode de l'imputation massive (Kim et coll., 2021). La régression logistique pondérée reposant sur l'échantillon groupé pour estimer les scores de propension proposés par Valliant et Dever (2011) fut le premier essai sérieux sur le sujet, lequel sert de motivation pour la méthode du pseudo-maximum de vraisemblance élaborée par Chen et coll. (2020). Brick (2015) a examiné l'inférence du modèle compositionnel dans les mêmes conditions. Elliot et Valliant (2017) ont tenu des débats éclairés sur l'inférence des échantillons non probabilistes. Yang, Kim et Song (2020) ont traité des problèmes associés à des données de grande dimension en combinant des échantillons d'enquête probabilistes et non probabilistes.

L'inférence statistique avec des échantillons d'enquête non probabilistes fait partie du thème plus général sur la combinaison de données de multiples sources. Le terme « intégration des données » est fréquemment utilisé dans ce contexte. Combiner les renseignements d'échantillons d'enquête probabilistes indépendants a été longuement étudié dans les ouvrages publiés portant sur les enquêtes; voir, par exemple, Wu (2004), Kim et Rao (2012) et les références connexes. L'inférence avec des échantillons de l'enquête à base multiple représente un autre sujet qui a été étudié en profondeur par les statisticiens d'enquête; voir Lohr et Rao (2006) et Rao et Wu (2010a) et les références connexes. Dans son récent article sollicité du prix Waksberg, Lohr (2021) a offert un aperçu des enquêtes à cadres multiples et quelques analyses fascinantes de l'utilisation d'une structure à cadres multiples pour servir de principe organisateur pour d'autres méthodes de combinaison de données. Compte tenu des nouvelles sources de données émergentes et des modifications de la façon dont les sources des données traditionnelles comme les dossiers administratifs sont considérés, l'intégration des données est devenue un domaine très vaste qui nécessite la poursuite de travaux de recherche. D'autres analyses sont fournies par Lohr et Raghunathan

(2017) sur la combinaison de données d'enquête avec d'autres sources de données, ainsi que par Thompson (2019) sur la combinaison de sources nouvelles et traditionnelles dans les enquêtes sur les populations. Kim et Tam (2021), ainsi que Yang, Kim et Hwang (2021), ont abordé l'intégration des données en combinant des mégadonnées et des données d'échantillons d'enquête pour l'inférence de population finie. L'article de Yang et Kim (2020) contenait un examen de l'intégration de données statistiques dans un échantillonnage.

Un des messages essentiels que le présent article transmet a trait aux concepts de *validité* et d'*efficacité* dans l'analyse des échantillons d'une enquête non probabiliste. La validité fait référence à la convergence des estimateurs ponctuels, et l'efficacité est mesurée par la variance asymptotique de l'estimateur ponctuel. La validité est la principale préoccupation, et la quête de l'efficacité est un objectif secondaire quand d'autres approches valides sont disponibles. Les analyses sur la validité et l'efficacité nécessitent un cadre inférentiel approprié et la mise au point rigoureuse de procédures statistiques, ce qui constitue un autre message principal du présent article. Les échantillons non probabilistes ne conviennent pas au cadre inférentiel fondé sur un plan ou fondé sur un modèle pour les échantillons d'enquête probabilistes. Toutefois, des concepts statistiques standard et des procédures inférentielles peuvent être conçus dans un cadre approprié pour une inférence valide et efficace avec des échantillons d'enquête non probabilistes.

Les échantillons non probabilistes peuvent avoir une très grande taille. Une grande taille d'échantillon constitue une épée à double tranchant : quand les procédures inférentielles sont valides, une grande taille d'échantillon mène à une inférence plus efficace; quand les estimateurs sont biaisés, une grande taille d'échantillon rend le biais encore plus prononcé. Un échantillon d'enquête non probabiliste présentant une fraction de sondage de 80 % par rapport à la population ne procure pas nécessairement de meilleurs résultats pour l'estimation qu'un petit échantillon probabiliste (Meng, 2018).

Une grande taille d'échantillon produit également des échantillons non probabilistes associés aux problèmes modernes des mégadonnées. Le rôle des méthodes statistiques traditionnelles dans l'ère des mégadonnées a été plaidé de façon convaincante par Richard Lockhart (2018) : « Les ressources massives de calcul n'éliminent pas le besoin d'une modélisation soignée, d'une évaluation honnête de l'incertitude, ou de la conception d'un bon devis expérimental. Les idées classiques de statistique jouent un rôle crucial pour que l'analyse de données demeure honnête, efficace et efficiente » [traduction].

Jean-François Beaumont (2020) a soulevé la question suivante : « Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? ». La réponse courte est que les méthodes d'échantillonnage probabiliste et les échantillons d'enquête probabilistes continueront d'être un important outil de collecte des données pour de nombreux domaines, y compris les statistiques officielles, et l'inférence fondée sur un plan jouera un rôle crucial pour tout cadre inférentiel en évolution. La tendance actuelle qui consiste à utiliser des échantillons non probabilistes et des données d'autres sources se poursuivra. Une inférence statistique valide et efficace avec des échantillons non probabilistes nécessite une information auxiliaire de la population cible. Quelques enquêtes probabilistes nationales de haute

qualité comportant des variables d'enquête soigneusement conçues peuvent jouer un rôle central dans l'analyse des échantillons d'enquête non probabilistes.

Remerciements

La présente étude a reçu des subventions du Conseil de recherches en sciences naturelles et en génie du Canada et de l'Institut canadien des sciences statistiques. Une première version du document a été présentée à la rencontre annuelle de 2021 de la Société statistique du Canada (SSC) en tant qu'allocution spéciale de l'invité du président par le Groupe des méthodes d'enquête de la SSC. L'auteur remercie le rédacteur en chef de *Techniques d'enquête*, Jean-François Beaumont, pour l'invitation et pour l'organisation des discussions sur le sujet émergent de l'inférence statistique avec des échantillons d'enquête non probabilistes. Il y a également lieu de remercier les deux réviseurs anonymes qui ont offert des commentaires constructifs sur la première ébauche, ce qui a permis d'améliorer l'article.

Bibliographie

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. et Tourangeau, R. (2013). Report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Techniques d'enquête*, 46, 1, 1-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2020001/article/00001-fra.pdf.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J. (1984). *Classification and Regression Trees*, second edition. Wadsworth & Brooks/Cole Advanced Books & Software.

Brick, J.M. (2015). Compositional model inference. Dans Proceedings of the Survey Research Methods Section, Joint Statistical Meetings, American Statistical Association, Alexandrie, Virginie, 299-307.

Chen, J., et Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*,16, 113-131.

Chen, J., et Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96, 260-269.

Chen, J., et Sitter, R.R. (1999). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

Chen, Y. (2020). *Statistical Analysis with Non-probability Survey Samples*, Thèse de doctorat, Department of Statistics and Actuarial Science, University of Waterloo.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Chen, Y., Li, P., Rao, J.N.K. et Wu, C. (2022). Pseudo empirical likelihood inference for non-probability survey samples. *The Canadian Journal of Statistics*, accepté.

Chu, K.C.K., et Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. Proceedings of the Survey Methods Section of SSC.

Elliott, M., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 249-264.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208-1212.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Kim, J.K., et Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24, 375-394.

Kim, J.K., et Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99, 85-100.

Kim, J.K., et Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference. *Revue Internationale de Statistique*, 89, 382-401.

Kim, J.K., Park, S., Chen, Y. et Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society, Series A*, 184, 941-963.

Liu, Z., et Valliant, R. (2021). Investigating an alternative for estimation from a nonprobability sample: Matching plus calibration. arXiv:2112.00855v1 [stat.ME]. Déc. 2021.

Lockhart, R. (2018). Special issue on big data and the statistical sciences: Guest editor's introduction. *The Canadian Journal of Statistics*, 46, 4-9.

Lohr, S.L. (2021). Les enquêtes à bases de sondage multiples pour un monde fait de sources de données multiples. *Techniques d'enquête*, 47, 2, 247-285. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-fra.pdf.

Lohr, S.L., et Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Lohr, S.L., et Rao, J.N.K. (2006). Estimation in multiple frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models*, second edition, New York: Chapman and Hall.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12, 685-726.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9, 141-142.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*, second Edition. Hoboken, NJ: Wiley.

Rao, J.N.K., et Wu, C. (2010a). Pseudo empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.

Rao, J.N.K., et Wu, C. (2010b). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B*, 72, 533-544.

Rivers, D. (2007). Sampling for web surveys. Dans *Proceedings of the Survey Research Methods Section*, Joint Statistical Meetings, American Statistical Association, Alexandrie, Virginie, 1-26.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Tourangeau, R., Conrad, F.G. et Couper, M.P. (2013). *The Science of Web Surveys*, first edition. Oxford: Oxford University Press.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Londres : Chapman & Hall.

Thompson, M.E. (2019). Combining data from new and traditional sources in population surveys. *Revue Internationale de Statistique*, 87, S79-89.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Valliant, R., et Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105-137.

Wang, L., Graubard, B.I., Katki, H.A. et Li, Y. (2020). Improving external validity of epidemiologic cohort analysis: A kernel weighting approach. *Journal of the Royal Statistical Society*, *Series A*, 183, 1293-1311.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā A*, 26, 359-372.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *The Canadian Journal of Statistics*, 32, 15-26.

Wu, C., et Rao, J.N.K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, 34, 359-375.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Wu, C., et Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer, Cham.

Yang, S., et Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

Yang, S., Kim, J.K. et Hwang, Y. (2021). Intégration de données d'enquêtes probabilistes et de mégadonnées aux fins d'inférence de population finie au moyen d'une imputation massive. *Techniques d'enquête*, 47, 1, 33-64. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2021001/article/00004-fra.pdf.

Yang, S., Kim, J.K. et Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society, Series B*, 82, 445-465.

Yuan, M., Li, P. et Wu, C. (2022). Nonparametric estimation of propensity scores for non-probability survey samples. Document de travail.

Zhao, P., et Wu, C. (2019). Some theoretical and practical aspects of empirical likelihood methods for complex surveys. *Revue Internationale de Statistique*, 87, S239-256.

Zhao, P., Rao, J.N.K. et Wu, C. (2020a). Empirical likelihood methods for public-use survey data. *Electronic Journal of Statistics*, 14, 2484-2509.

Zhao, P., Ghosh, M., Rao, J.N.K. et Wu, C. (2020b). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society, Series B*, 82, 155-174.