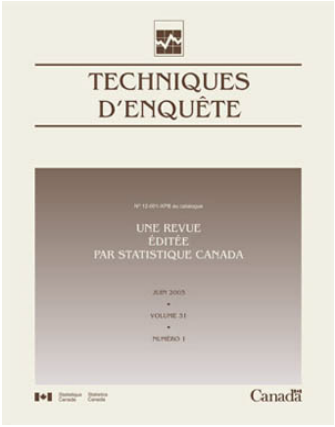
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Hypothèse de l’échangeabilité dans des méthodes d’ajustement fondées sur le score de propension aux fins d’estimation de la moyenne de population au moyen d’échantillons non probabilistes



par Yan Li

Date de diffusion : le 25 juin 2024

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.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 45-65
Statistique Canada, nº 12-001-X au catalogue

Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes

Yan Li[1]

Résumé

Des échantillons non probabilistes émergent rapidement pour aborder des sujets prioritaires urgents dans différents domaines. Ces données sont actuelles, mais sujettes à un biais de sélection. Afin de réduire le biais de sélection, une littérature abondante portant sur la recherche sur les enquêtes a étudié l'utilisation de méthodes d'ajustement par le score de propension (SP) pour améliorer la représentativité de la population des échantillons non probabilistes, au moyen d'échantillons d'enquête probabilistes utilisés comme références externes. L'hypothèse d'échangeabilité conditionnelle (EC) est l'une des principales hypothèses requises par les méthodes d'ajustement fondées sur le SP. Dans le présent article, j'examine d'abord la validité de l'hypothèse de l'EC conditionnellement à plusieurs estimations de scores d'équilibrage qui sont utilisées dans les méthodes d'ajustement fondées sur le SP existantes. Un score d'équilibrage adaptatif est proposé aux fins d'estimation sans biais des moyennes de population. Les estimateurs de la moyenne de population selon les trois hypothèses de l'EC sont évalués au moyen d'études de simulation de Monte Carlo et illustrés au moyen de l'étude sur la séroprévalence du SRAS-CoV-2 des National Institutes of Health pour estimer la proportion d'adultes aux États-Unis qui présentaient des anticorps de la COVID-19 du 1er avril au 4 août 2020.

Mots-clés : Appariement par scores de propension; échantillon par quota; estimateur de la variance par linéarisation en séries de Taylor; étude de séroprévalence du SRAS-CoV-2; pondération par scores de propension; score d'équilibrage.

1. Introduction

Des échantillons non probabilistes émergent rapidement pour aborder des sujets prioritaires urgents des sujets prioritaires urgents dans différents domaines (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile et Tourangeau, 2013; Kennedy, Mercer, Keeter, Hatley, McGeeney et Gimenez, 2016). Ces données sont actuelles, mais sujettes à un biais de sélection. Les participants sont souvent autosélectionnés et se portent volontaires pour participer à une étude sans probabilités de sélection préétablies. Les exemples comprennent des échantillons épidémiologiques composés de volontaires qui ne sont pas choisis aléatoirement et qui ne sont donc généralement pas représentatifs d'une population. De plus, les volontaires sont souvent sujets à des « effets de volontaires sains » (Pinsky, Miller, Kramer, Church, Reding, Prorok, Gelmann, Schoen, Buys, Hayes et Berg, 2007), ce qui se traduit habituellement par des estimations plus faibles de l'incidence de la maladie et de la mortalité chez les volontaires que dans la population générale. Un autre exemple concerne les données recueillies à partir de panels Web à échantillonnage probabiliste, qui peuvent donner un taux d'attrition élevé et dans lesquels les taux de non-réponse sont souvent de 90 % ou plus (Baker et coll., 2013). Bien qu'une non-réponse élevée ne soit pas nécessairement un indice du biais dans les

1. Yan Li, Joint Program in Survey Methodology and Department of Epidemiology and Biostatistics, University of Maryland, College Park. Courriel : yli6@umd.edu.

réponses (Groves et Peytcheva, 2008; Brick et Tourangeau, 2017), le biais de sélection est très préoccupant parce que la composition des panels Web diffère souvent de celle de la population sous-jacente.

Contrairement aux échantillons non probabilistes, les enquêtes probabilistes basées sur la population sont conçues pour produire des estimations presque sans biais des caractéristiques de la population. Elles s'appuient sur des plans d'échantillonnage probabiliste, comme l'échantillonnage stratifié par grappes à plusieurs degrés, pour sélectionner des échantillons. Les échantillons qui en résultent, s'ils sont correctement pondérés par les poids d'enquête, peuvent représenter fidèlement la population cible; ils sont par conséquent moins sujets au biais de sélection.

Afin de réduire le biais de sélection des échantillons non probabilistes, une littérature abondante portant sur la recherche sur les enquêtes a étudié l'utilisation de méthodes d'ajustement par le score de propension (SP) pour améliorer la représentativité de la population des échantillons non probabilistes, au moyen des échantillons d'enquête probabiliste comme références externes (Elliott et Valliant, 2017). Différentes méthodes d'ajustement basées sur le SP ont été élaborées; elles peuvent être regroupées en deux catégories : 1) les méthodes de pondération par l'inverse du SP (par exemple Chen, Li et Wu, 2020; Elliott, 2013; Valliant et Dever, 2011) ou les méthodes de pondération par cotes inverses (par exemple Wang, Valliant et Li, 2021) (*pondération* par SP); 2) les méthodes d'appariement par SP ou par logarithme du risque du SP (*appariement* par SP) (par exemple Lee et Valliant, 2009; Wang, Graubard, Katki et Li, 2022; Rivers, 2007).

Les méthodes de pondération par SP établissent une pseudo-pondération pour chaque unité d'échantillon non probabiliste comme étant l'inverse de sa propension à la participation. Elles corrigent le biais de sélection selon les vrais modèles de propension, bien qu'elles puissent être sensibles à la spécification erronée du modèle de propension (Valliant, 2020) et produire des estimations présentant de grandes variances en raison des poids extrêmes (Stuart, 2010). En revanche, les méthodes d'appariement par SP s'appuient sur le score de propension pour mesurer la similarité dans les distributions des covariables comprises dans le modèle de propension entre l'enquête probabiliste et l'échantillon non probabiliste; elles tendent par conséquent à être moins sensibles à la spécification erronée du modèle de propension. De plus, comme les méthodes d'appariement par SP évitent les poids extrêmes, elles produisent des estimations présentant des variances plus petites. Pour obtenir une synthèse exhaustive sur les autres méthodes d'analyse des échantillons non probabilistes et d'intégration des données, consulter Beaumont (2020), Rao (2021) et Valliant (2020).

Les méthodes d'ajustement fondées sur le SP (par exemple Chen et coll., 2020) nécessitent les hypothèses clés suivantes pour faire des inférences d'échantillons non probabilistes. Premièrement, l'échantillon de l'enquête de référence, par la pondération, représente correctement la population finie (PF) d'intérêt. Deuxièmement, toutes les unités de la PF ont une propension à la participation positive (c'est-à-dire que tous les membres de la population ont une propension positive à participer à des échantillons non probabilistes). Troisièmement, l'échangeabilité conditionnelle (EC) se vérifie sans covariables non mesurées, c'est-à-dire que la probabilité que chaque membre de la PF participe à l'échantillon non probabiliste n'est pas liée à son résultat, conditionnellement à toutes les covariables

mesurées. Quatrièmement, le fait d'être échantillonné dans le cadre de l'enquête de référence et de participer à l'échantillon non probabiliste sont indépendants. Toutes ces hypothèses sont essentielles. Dans le présent article, nous nous intéressons à l'hypothèse de l'EC et nous examinons plusieurs scores d'équilibrage (c'est-à-dire des fonctions de covariables) qui satisfont à l'hypothèse de l'EC.

Dans les études observationnelles pour les inférences causales, les chercheurs tentent généralement d'ajuster pour toutes les covariables mesurées, afin d'imiter une expérience complètement randomisée et supposent que de tels ajustements sont suffisants pour des estimations sans biais des effets du traitement. Cette hypothèse est connue sous le nom d'« échangeabilité des assignations de traitement » (Rubin, 1978). Toutefois, la recherche sur les enquêtes vise à faire une inférence au sujet des paramètres de la PF, et il existe peu de recherches sur la suffisance de l'hypothèse mentionnée ci-dessus. Certaines études (par exemple celle de Wang et coll., 2021) ont mentionné la nécessité de faire des hypothèses sur le fait que la propension à la participation est ignorable étant donné un ensemble de variables d'ajustement. Cependant, dans ces études, on s'est contenté d'observer la présence ou l'absence d'estimations biaisées, et on a rarement examiné si et dans quelle mesure on va à l'encontre de l'hypothèse de l'EC quand on fait une inférence à propos des paramètres de la population finie.

La contribution du présent article consiste à : 1) étudier la validité de l'hypothèse de l'EC qui est conditionnelle à diverses estimations de scores d'équilibrage qui sont utilisées dans les méthodes actuelles d'ajustement par le SP, y compris les méthodes de pondération par le SP et d'appariement par SP, pour les inférences d'échantillons non probabilistes; 2) élaborer un score d'équilibrage adaptatif pour l'hypothèse de l'EC afin d'en améliorer l'efficacité. Dans l'article, nous n'élaborons pas de nouvelles méthodes d'ajustement fondées sur le SP, mais nous étudions divers scores d'équilibrage qui satisfont à l'hypothèse de l'EC. La pondération par SP en utilisant la méthode de la propension logistique ajustée (PLA) est utilisée à des fins d'illustration. Le score d'équilibrage établi peut également être utilisé dans des méthodes d'appariement par SP comme le lissage par la méthode du noyau (Wang et coll., 2022). Les estimateurs PLA, supposant l'échangeabilité des résultats conditionnelle à divers scores d'équilibrage, sont évalués via des études de simulation de Monte Carlo et illustrés au moyen de l'étude sur la séroprévalence du SARS-CoV-2 des National Institutes of Health pour estimer la proportion d'adultes des États-Unis présentant des anticorps contre la COVID-19 du 1er avril au 4 août 2020.

2. Hypothèse de l'échangeabilité conditionnelle

2.1 Notation

Considérons une population finie (PF) cible comme un échantillon aléatoire de N personnes tirées d'un modèle de superpopulation, indexé par $U = \{1, 2, \ldots, N\}$, avec des observations sur une variable d'étude y et un vecteur de covariables $\mathbf{x}$. Soit $\{y_i, \mathbf{x}_i : i \in C\}$ les observations dans l'échantillon non probabiliste de personnes, où $C \subset U$ de taille n_c. Nous cherchons à estimer la moyenne de la PF $\overline{Y}_N = \frac{1}{N}\sum_{i \in U} y_i$ au moyen

de l'échantillon non probabiliste C. Le problème est que nous observons C, qui, cependant, peut ne pas être un échantillon représentatif tiré de U. Par conséquent, $E_C(y \mid U) \neq \overline{Y}_N$, où l'indice C désigne le caractère aléatoire en raison du processus de participation à l'échantillon non probabiliste inconnu tiré de U. Soit $E(y \mid C) = E_U(E_C(y \mid U))$ et $E(y \mid U) = E_U(\overline{Y}_N)$, où l'indice U désigne l'espérance sous le modèle de superpopulation. L'espérance de y dans C peut différer de celle dans U, à savoir $E(y \mid C) \neq E(y \mid U)$ en raison du biais de sélection de l'échantillon non probabiliste C.

2.2 Hypothèse de l'échangeabilité conditionnelle et score d'équilibrage

Pour obtenir un estimateur convergent par rapport au plan de sondage de $\overline{Y}_N$ au moyen de C, l'EC suppose

$$E\{y \mid b(\mathbf{x}), C\} = E\{y \mid b(\mathbf{x}), U\}, \tag{2.1}$$

où $b(\mathbf{x})$ est une fonction des covariables $\mathbf{x}$, qu'on appelle le score d'équilibrage.

L'hypothèse de l'EC (2.1) indique que conditionnellement au score d'équilibrage $b(\mathbf{x})$, c'est-à-dire une fonction de covariables mesurées, le résultat a la même espérance dans C que dans U. Autrement dit, les unités de l'échantillon non probabiliste ayant la même valeur de score d'équilibrage $b(\mathbf{x})$ représenteraient le même nombre d'unités de la PF. Intuitivement, si deux personnes avaient la même propension à la participation, elles représenteraient le même nombre d'unités de la PF. Par conséquent, un choix naturel de $b(\mathbf{x})$ est la propension à la participation $P(i \in C \mid \mathbf{x}, U)$, c'est-à-dire la probabilité que l'unité de la PF i participe à C conditionnellement à la valeur de $\mathbf{x}$.

De façon plus générale, le *critère de base* pour choisir un score d'équilibrage est que $b(\mathbf{x})$ est plus fin que, sinon égal à, $P(i \in C \mid \mathbf{x}, U)$ pour que l'hypothèse de l'EC soit valide (2.1). Par conséquent, le choix le plus fin de score d'équilibrage est $b(\mathbf{x}) = \mathbf{x}$ et le moins fin est $b(\mathbf{x}) = P(i \in C \mid \mathbf{x}, U)$ ou sa fonction monotone. Par conséquent, les $b(\mathbf{x})$ choisis doivent pouvoir permettre de distinguer les unités C ayant des propensions à la participation différentes.

Dans l'inférence causale (Rosenbaum et Rubin, 1983), l'hypothèse de l'échangeabilité conditionnelle indique que le résultat est échangeable entre le groupe *traité* et le groupe *témoin*, conditionnellement à toutes les covariables mesurées. La distribution des covariables dans le groupe traité est appariée à celle dans le groupe témoin au moyen de méthodes de pondération par SP ou d'appariement par SP, selon un modèle de *propension de l'attribution du traitement (treatment assignment propensity model)*. On estime ensuite l'effet du traitement en comparant les moyennes des deux groupes après pondération ou appariement. De façon analogue, dans les inférences d'échantillons non probabilistes, la distribution des covariables dans l'échantillon non probabiliste est appariée à celle dans la PF selon un modèle de *propension à la participation* (*échantillon non probabiliste*). Au lieu d'estimer l'effet du traitement, on estime la moyenne de la PF en supposant l'échangeabilité du résultat entre l'échantillon non probabiliste et la PF après la pondération par SP ou l'appariement par SP. Pour en savoir plus, les lecteurs trouveront dans Mercer, Kreuter, Keeter et Stuart (2017) des précisions sur les parallèles entre l'inférence causale et l'inférence d'échantillon non probabiliste.

3. Scores d'équilibrage existants

3.1 Estimation de $P(i \in C \mid \mathbf{x}, U)$

On peut estimer directement la propension à la participation $P(i \in C \mid \mathbf{x}, U)$ si les covariables $\mathbf{x}$ sont connues pour toutes les personnes dans U. Malheureusement, nous n'avons pas la mesure de $\mathbf{x}$ pour l'ensemble de U, mais on peut estimer sa distribution à partir d'un échantillon probabiliste S de taille n_s, $\{\mathbf{x}_i : i \in S\}$. Différentes méthodes de modélisation de la propension où S est utilisé comme enquête de référence ont été proposées (Chen et coll., 2020; Kern, Li et Wang, 2021). À titre d'illustration, nous supposons un modèle de régression logistique

$$\log\left\{\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right\} = \mathbf{B}^T g(\mathbf{x}_i), \quad \text{pour } i \in U, \tag{3.1}$$

où le score de propension $p(\mathbf{x}_i)$ est la propension de l'unité i à faire partie de l'échantillon non probabiliste par rapport à la population finie, selon une approximation de l'échantillon d'enquête pondéré, désigné par S_w. De même, $\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} = P(i \in C \mid \mathbf{x}_i, U)$. $g(\mathbf{x}_i)$ est une fonction connue des covariables observées, et $\mathbf{B}$ sont les coefficients de régression inconnus qu'il faut estimer; voir dans Wang et coll. (2021, section 2.3) la justification du modèle de propension (3.1). Nous définissions w_i comme le poids d'échantillon de l'unité $i \in S$. Quand on résout $S(\mathbf{B}) = \left\{\sum_{i' \in C}(1-p(\mathbf{x}_{i'}))\, g(\mathbf{x}_{i'}) - \sum_{i \in S} w_i p(\mathbf{x}_i)\, g(\mathbf{x}_i)\right\} = 0$ pour $\mathbf{B}$, l'estimation est désignée par $\hat{\mathbf{B}}_w$. L'indice w indique que les poids de l'enquête de référence servent à estimer $\mathbf{B}$ dans le modèle de propension (3.1). La propension à la participation $P(i \in C \mid \mathbf{x}, U)$ pour $i \in C \cup S$ peut être estimée par $\exp(\mathbf{x}_i \hat{\mathbf{B}}_w) = \frac{\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)}{1-\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)}$, $\hat{p}(\mathbf{x}_i, \hat{\mathbf{B}}_w)$ étant l'estimation du score de propension $p(\mathbf{x}_i)$.

3.2 Hypothèse de l'échangeabilité conditionnelle à $b(\mathbf{x}; \hat{\mathbf{B}}_w)$

Pour satisfaire à l'hypothèse de l'EC (2.1), le score d'équilibrage doit être aussi fin ou plus fin que le taux de participation estimé. D'après les conclusions de Wang et coll. (2022), le prédicteur linéaire, c'est-à-dire une transformation logarithmique naturelle de la propension à la participation estimée, est utilisé comme score d'équilibrage, à savoir $b(\mathbf{x}_i; \hat{\mathbf{B}}_w) = \hat{\mathbf{B}}_w^T g(\mathbf{x}_i) = \log \hat{p}(i \in C \mid \mathbf{x}_i, U)$. Alors, selon le modèle de propension (3.1), on suppose que $b(\mathbf{x}) = b(\mathbf{x}; \hat{\mathbf{B}}_w)$ dans (2.1), c'est-à-dire que

$$E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_w), C\} = E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_w), U\}$$

se vérifie approximativement. Comme dans ce qui suit, nous estimons la moyenne de la population au moyen de plusieurs méthodes existantes d'ajustement fondées sur le SP. Par exemple, la méthode de pondération par SP par la propension logistique ajustée ou PLA (Wang et coll., 2021) pondère l'unité i dans C par l'inverse de $\hat{p}(i \in C \mid \mathbf{x}_i, U) = \exp(b(\mathbf{x}_i; \hat{\mathbf{B}}_w))$. Un autre exemple est la méthode d'appariement par SP par pondération du noyau (Wang et coll. 2022), qui apparie les unités dans C et S selon la similarité dans $b(\mathbf{x}; \hat{\mathbf{B}}_w)$. Il a été prouvé que les estimations par la PLA et par pondération du noyau sont approximativement sans biais selon l'hypothèse de l'EC conditionnellement à $b(\mathbf{x}; \hat{\mathbf{B}}_w)$.

Toutefois, un inconvénient important de cette méthode est l'inflation de la variance potentiellement importante dans $b(\mathbf{x};\hat{\mathbf{B}}_w)$ en raison de la variabilité (Scott et Wild, 2001; Li, Graubard et DiGaetano, 2011) des poids de l'enquête de référence différentiels par rapport aux poids de l'échantillon non probabiliste (= 1) dans l'estimation du paramètre du modèle $\mathbf{B}$. Aux fins de réduction de la variance, les poids de l'enquête n'ont pas été pris en compte dans l'estimation de $\mathbf{B}$ (Wang, Graubard, Katki et Li, 2020; Lee et Valliant, 2009).

3.3 Hypothèse de l'échangeabilité conditionnelle à $b(\mathbf{x};\hat{\mathbf{B}}_0)$

Supposons que le modèle de propension (3.1) est ajusté aux données combinées de l'échantillon non probabiliste et de l'enquête non pondérées $(C \cup S)$ et que le score d'équilibrage qui en résulte est $b(\mathbf{x}_i,\hat{\mathbf{B}}_0)=\hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$, $\hat{\mathbf{B}}_0$ ayant été obtenu en résolvant l'équation d'estimation $S(\mathbf{B})=\left\{\sum_{i'\in C}(1-p(\mathbf{x}_{i'}))\,g(\mathbf{x}_{i'})-\sum_{i\in S}p(\mathbf{x}_i)\,g(\mathbf{x}_i)\right\}=0$ pour $\mathbf{B}$, sans tenir compte des poids de l'échantillon probabiliste. Par conséquent, l'EC fondée sur $\hat{\mathbf{B}}_0$, supposée par les méthodes existantes de pondération par SP ou d'appariement par SP (Wang et coll., 2020; Lee et Valliant, 2009; Kern et coll., 2021), est

$$E\{y \mid b(\mathbf{x};\hat{\mathbf{B}}_0), C\} = E\{y \mid b(\mathbf{x};\hat{\mathbf{B}}_0), U\}.$$

Lorsque les poids d'enquête ne sont pas utilisés, le score d'équilibrage estimé $b(\mathbf{x},\hat{\mathbf{B}}_0)$ peut être plus stable que $b(\mathbf{x},\hat{\mathbf{B}}_w)$. La question consiste à déterminer dans quelle mesure l'hypothèse de l'EC conditionnellement à $b(\mathbf{x},\hat{\mathbf{B}}_0)$ est plausible dans des problèmes réels.

Mentionnons que $b(\mathbf{x},\hat{\mathbf{B}}_0)$ produit une distribution de $\mathbf{x}$ équilibrée entre C et S, et que par conséquent, l'échangeabilité de la distribution de y (tous les $\mathbf{x}$ étant équilibrés) se vérifie entre C et S, ce qui toutefois ne suffit pas pour obtenir une estimation sans biais de la moyenne de la PF. En effet, il est nécessaire d'avoir l'échangeabilité de la distribution de y entre C et U conditionnellement à $b(\mathbf{x},\hat{\mathbf{B}}_0)$. Nous savons d'après la section 2.2 que $P(i\in C \mid \mathbf{x},U)$ est le score d'équilibrage le plus grossier satisfaisant à (2.1) et que $b(\mathbf{x},\hat{\mathbf{B}}_w)$ produit approximativement une distribution de y équilibrée entre C et U. Selon les critères de base du choix de score d'équilibrage, le score d'équilibrage $b(\mathbf{x},\hat{\mathbf{B}}_0)$ doit être aussi fin ou plus fin que $b(\mathbf{x},\hat{\mathbf{B}}_w)$. Un exemple en est que $b(\mathbf{x}_i;\hat{\mathbf{B}}_0)=\hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$ est une fonction linéaire de $b(\mathbf{x}_i;\hat{\mathbf{B}}_w)=\hat{\mathbf{B}}_w^T g(\mathbf{x}_i)$, c'est-à-dire $\hat{\mathbf{B}}_w^T=\text{const.}\times\hat{\mathbf{B}}_0^T$. Supposons que l'enquête de référence S suréchantillonne, de par son plan de sondage, un groupe minoritaire, par exemple les femmes afro-américaines. Cette relation linéaire exige que la distribution du même groupe minoritaire, défini par la race ou l'origine ethnique et le genre dans l'échantillon non probabiliste, soit proportionnelle à celle de l'enquête de référence. Or, en réalité, nous n'avons aucun contrôle sur l'échantillonnage non probabiliste et, par conséquent, la relation linéaire ne se vérifie que par hasard. L'estimateur fondé sur $b(\mathbf{x};\hat{\mathbf{B}}_0)$ est efficace, mais il peut être biaisé.

Exemple hypothétique

À des fins d'illustration, supposons qu'un échantillon non probabiliste et un échantillon d'enquête sont sélectionnés par échantillonnage avec probabilité proportionnelle à la taille (PPT), avec la mesure de taille

pour l'unité i de la PF définie, respectivement, par $s_{ic} = \exp(x_{i1}B_1 + x_{i2}B_2)$ pour la participation à l'échantillon non probabiliste et $s_{is} = \exp(x_{i1}B_1' + x_{i3}B_3)$ pour la sélection de l'échantillon d'enquête. Supprimons l'indice i et supposons que $B_1 \approx B_1'$. La probabilité qu'une unité de la PF participe à l'échantillon non probabiliste (p_c) par rapport à sélection dans l'enquête (p_s) est

$$\begin{aligned}\log\left(\frac{p_c}{p_s}\right) &= \log\left(\frac{n_c s_c}{\sum_U s_c} \Big/ \frac{n_s s_s}{\sum_U s_s}\right) = \log\left(\frac{n_c \sum_U s_s}{n_s \sum_U s_c} \times \frac{s_c}{s_s}\right), \\ &= \text{const.} + x_1(B_1 - B_1') + x_2 B_2 - x_3 B_3 = \text{const.} + \mathbf{x}^T \mathbf{B}_0,\end{aligned}$$

où $\mathbf{x} = (x_1, x_2, x_3)^T$ et $\mathbf{B}_0 = (B_1 - B_1', B_2, -B_3)^T$. En ajustant un modèle logistique, y compris toutes les variables $\mathbf{x}$, à l'échantillon combiné (enquête non probabiliste et enquête non pondérée), un score d'équilibrage estimé serait

$$b(\mathbf{x}; \hat{\mathbf{B}}_0) = \mathbf{x}^T \hat{\mathbf{B}}_0.$$

Mentionnons que $\hat{\mathbf{B}}_0$ inclut l'effet x_1 atténué dans la construction $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ en raison de la distribution similaire de x_1 dans S et dans C. Par conséquent, les scores d'équilibrage estimés ne peuvent pas distinguer les unités C ayant différentes propensions à la participation par x_1, et donc $E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0), C\} \neq E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0), U\}$, ce qui conduit à une estimation biaisée de $\bar{Y}_N$.

Dans la section suivante, nous proposons un score d'équilibrage adaptatif qui ajuste $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ pour que ce soit une fonction monotone de l'estimation de $P(i \in C \mid \mathbf{x}, U)$ pour l'estimation sans biais de la moyenne de la PF.

4. Score d'équilibrage adaptatif

Nous proposons un score d'équilibrage ajusté en trois étapes. La première étape consiste à adapter un modèle de régression logistique à l'échantillon $C \cup S$ combiné sans poids, donné par (Wang et coll., 2020)

$$\log\left\{\frac{p(i \in C \mid \mathbf{x}_i, U)}{p(i \in S \mid \mathbf{x}_i, U)}\right\} = \log\left\{\frac{p^*(\mathbf{x}_i)}{1 - p^*(\mathbf{x}_i)}\right\} = \mathbf{B}_0^T g(\mathbf{x}_i) \quad \text{pour } i \in U \tag{4.1}$$

et les estimations du paramètre du modèle $\mathbf{B}_0$ sont désignées par $\hat{\mathbf{B}}_0$, où $p^*(\mathbf{x}_i)$ est la propension à être dans C par rapport à être dans S pour l'unité i. Comme nous l'avons vu à la section 3.3, $b(\mathbf{x}; \hat{\mathbf{B}}_0) = \hat{\mathbf{B}}_0^T g(\mathbf{x}_i)$ équilibre la distribution de $\mathbf{x}$ entre C et S. Si l'on ne tient pas compte des poids de l'échantillon dans l'analyse, $\hat{\mathbf{B}}_0$ tend à être plus efficace que $\hat{\mathbf{B}}_w$. Il peut toutefois y avoir violation de l'hypothèse de l'EC (2.1) quand $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ n'arrive pas à équilibrer la distribution dans $\mathbf{x}$ entre C et U.

La deuxième étape vise à élaborer un facteur de correction du biais pour ajuster $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ de façon à ce que la distribution équilibrée dans $\mathbf{x}$ entre C et U (obtenue par approximation par l'enquête de référence pondérée S_w) puisse être atteinte. En tant que dispositif *de calcul*, on construit une pseudo-population de $S^* \cup U$ où S^* est un double de S qui a les mêmes distributions conjointes de covariables $\mathbf{x}$ et le même

résultat y que l'original S. Dans la pseudo-population $S^* \cup U$, S^* et S sont traités comme deux ensembles différents. Nous modélisons $q(\mathbf{x}_i)$ comme étant la probabilité que l'unité i soit incluse dans S à partir de la pseudo-population, c'est-à-dire

$$q(\mathbf{x}_i) = p(i \in S \mid \mathbf{x}_i, S^* \cup U) = \frac{p(i \in S \mid \mathbf{x}_i, U)}{1 + p(i \in S \mid \mathbf{x}_i, U)}.$$

Supposons un modèle logistique

$$\log\{p(i \in S \mid \mathbf{x}_i, U)\} = \log\left\{\frac{q(\mathbf{x}_i)}{1 - q(\mathbf{x}_i)}\right\} = \boldsymbol{\gamma}^T g(\mathbf{x}_i), \quad \text{pour } i \in U \tag{4.2}$$

où $\boldsymbol{\gamma}$ désigne les paramètres du modèle, estimés par la résolution de l'équation estimant $S(\boldsymbol{\gamma}) = \sum_{i \in S}(1 - q(\mathbf{x}_i) - w_i q(\mathbf{x}_i))\, g(\mathbf{x}_i) = 0$ pour $\boldsymbol{\gamma}$. L'estimation est désignée par $\hat{\boldsymbol{\gamma}}_w$ et mesure les effets de $g(\mathbf{x})$ sur la sélection de l'échantillon S. Nous l'utilisons pour corriger les effets déformés ou manquants de $g(\mathbf{x})$ sur la propension à la participation de l'échantillon non probabiliste C dans $b(\mathbf{x}; \hat{\mathbf{B}}_0)$, en particulier pour les variables intervenant à la fois dans les processus d'échantillonnage S et de participation C.

À l'étape 3, le nouveau score d'équilibrage estimé est construit comme étant

$$b(\mathbf{x}_i; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w) = (\hat{\boldsymbol{\gamma}}_w^T + \hat{\mathbf{B}}_0^T)\, g(\mathbf{x}_i) \quad \text{pour } i \in U.$$

Comme nous l'avons mentionné, l'addition des modèles (4.1) et (4.2) donne le modèle (3.1), le premier membre étant égal à

$$\log\left\{\frac{p(i \in C \mid \mathbf{x}_i, U)}{p(i \in S \mid \mathbf{x}_i, U)}\right\} + \log\{p(i \in S \mid \mathbf{x}_i, U)\} = \log\{P(i \in C \mid \mathbf{x}_i, U)\},$$

une fonction monotone de la propension à la participation, et le deuxième membre la même forme fonctionnelle $g(\mathbf{x}_i)$ que dans le modèle (3.1). Nous savons que $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ dans le modèle (3.1), bien que satisfaisant à l'hypothèse de l'EC (2.1), peut être inefficace en raison des poids différentiels dans l'analyse. Au lieu d'ajuster le modèle (3.1) directement aux données combinées de l'échantillon non probabiliste et de l'enquête pondérée $(C \cup S_w)$ pour obtenir $\hat{\mathbf{B}}_w$, nous construisons le score d'équilibrage ajusté $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w)$ basé sur $\hat{\mathbf{B}}_0$ et $\hat{\boldsymbol{\gamma}}_w$ en trois étapes. Ce score d'équilibrage ajusté est une fonction monotone (logarithme naturel) de la propension à la participation de l'échantillon C, et par conséquent la distribution de y est échangeable entre C et U, c'est-à-dire que

$$E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T), C\} = E\{y \mid b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T), U\}$$

se vérifie approximativement.

Comme dans ce qui suit, des méthodes d'ajustement fondées sur le SP peuvent servir à créer des pseudo-pondérations pour les unités dans C en se basant sur le nouveau score d'équilibrage adaptatif $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$. Les méthodes de pondération par SP pondèrent chaque unité dans C par l'inverse du taux de participation

estimé. En revanche, les méthodes d'appariement par SP apparient les unités C et S en s'appuyant sur le score d'équilibrage adaptatif, puis distribuent les poids de sondage dans S aux unités C selon leurs similarités. Par exemple, la méthode de pondération par la PLA (Wang et coll., 2021) crée des pseudo-pondérations.

$$\hat{w}_j^{\text{PLA}} = \exp^{-1}(b(\mathbf{x}_j; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)) \quad \text{pour } j \in C.$$

Le lissage par la méthode du noyau (Wang et coll., 2022) crée des pseudo-pondérations en additionnant les poids fractionnels distribués à partir de chaque unité d'enquête $i \in S$,

$$\hat{w}_j^{\text{KW}} = \sum_{i \in S} w_i K_{ij} \quad \text{avec} \quad K_{ij} = \frac{K\left(\frac{d_{ij}}{h}\right)}{\sum_{l \in C} K\left(\frac{d_{il}}{h}\right)} \quad \text{pour} \quad j \in C,$$

où $K(\cdot)$ est une fonction de noyau arbitraire telle que la fonction de densité normale standard, h est la largeur de bande associée à $K(\cdot)$, et la distance $d_{ij} = b(\mathbf{x}_i; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T) - b(\mathbf{x}_j; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ mesure la similarité dans la distribution de $\mathbf{x}$ entre l'unité d'échantillonnage non probabiliste $j \in C$ et l'unité d'enquête $i \in S$.

La moyenne de population peut ensuite être estimée par

$$\bar{y} = \frac{1}{\sum_{j \in C} \hat{w}_j} \sum_{j \in C} \hat{w}_j y_j, \tag{4.3}$$

où $\hat{w}_j$ peut être $\hat{w}_j^{\text{ALP}}$ ou $\hat{w}_j^{\text{KW}}$.

Pour estimer la variance de $\bar{y}$, nous supposons la taille de PF $N \to \infty$ et considérons le caractère aléatoire attribuable à l'échantillonnage de S et au processus de participation de C tiré de U. On élabore l'estimateur de la variance par linéarisation en séries de Taylor (LT) pour tenir compte de la variabilité attribuable à l'estimation des scores de propension $p^*(\mathbf{x}_i)$ et $q(\mathbf{x}_i)$ aux étapes 1 et 2. La technique de linéarisation en séries de Taylor est couramment utilisée dans les ouvrages publiés portant sur les enquêtes pour calculer des estimateurs de variance convergents par rapport au plan de sondage (Li, Graubard, Huang et Gastwirth, 2015; Li et Graubard, 2012). En supposant l'indépendance entre le fait d'être échantillonné dans l'enquête de référence et la participation à l'échantillon non probabiliste, la variance de $\bar{y}$ peut être estimée par (Korn et Graubard, 1999)

$$\text{var}_{\text{TL}}(\bar{y}) \cong \text{var}\left(\sum_{j \in C} z_j\right) + \text{var}\left(\sum_{i \in S} z_i\right), \tag{4.4}$$

où z_j (ou z_i) est l'écart de Taylor pour la j^{e} (ou i^{e}) unité dans C (ou dans S) que nous avons calculé en prenant la dérivée de $\bar{y}$ par rapport au poids de sondage (Shah, 2004). Par exemple, quand $\hat{w}_j = \hat{w}_j^{\text{PLA}}$, l'écart de Taylor pour l'unité $j \in C$ est

$$z_j = \frac{\partial}{\partial w_j} \bar{y} = \frac{\hat{w}_j (y_j - \bar{y})}{\sum_{l \in C} \hat{w}_l} + \frac{\sum_{l \in C} (y_l - \bar{y})}{\sum_{l \in C} \hat{w}_l} \left(\frac{\partial}{\partial w_j} \hat{w}_l\right)$$

et

$$\frac{\partial}{\partial w_j}\hat{w}_l = \left(\frac{\partial}{\partial \hat{\theta}}\hat{w}_l\right)\left(\frac{\partial}{\partial w_j}\hat{\theta}\right) = -\hat{w}_l x_l \left(\frac{\partial}{\partial w_j}\hat{\theta}\right),$$

où $\hat{\theta}$ désigne les paramètres estimés du modèle, qui peuvent être $\hat{B}_0$, $\hat{B}_w$, ou $\hat{B}_0 + \hat{\gamma}_w$, par exemple

$$\frac{\partial}{\partial w_j}(\hat{B}_0 + \hat{\gamma}_w) = (1-\hat{p}_j^*)\, x_j \left(\sum_{j' \in C \cup S} \hat{p}_{j'}^*(1-\hat{p}_{j'}^*)\, x_{j'}\, x_{j'}^T\right)^{-1},$$

où $\hat{p}_j^*$ pour $j \in C$ est le score de propension estimé pour l'unité j dans le modèle (4.1).

Pour l'unité $i \in S$, l'écart de Taylor est

$$z_i = \frac{\sum_{j \in C}(y_j - \bar{y})}{\sum_{j \in C}\hat{w}_j}\left(\frac{\partial}{\partial w_i}\hat{w}_j\right),$$

et

$$\frac{\partial}{\partial w_i}\hat{w}_j = \frac{\partial}{\partial \hat{\theta}}\hat{w}_j \frac{\partial}{\partial w_i}\hat{\theta} = -\hat{w}_j x_j \left(\frac{\partial}{\partial w_i}\hat{\theta}\right),$$

où $\hat{\theta}$ peut être $\hat{B}_0$, $\hat{B}_w$, ou $\hat{B}_0 + \hat{\gamma}_w$ par exemple

$$\begin{aligned}\frac{\partial}{\partial w_i}(\hat{B}_0 + \hat{\gamma}_w) \quad &= -\hat{p}_i^* x_i \left(\sum_{j' \in C \cup S} \hat{p}_{j'}^*(1-\hat{p}_{j'}^*)\, x_{j'} x_{j'}^T\right)^{-1} \\ &\quad + \left(1-\hat{q}_i - \hat{q}_i w_i\right) x_i \left(\sum_{j' \in S}(1+w_{j'})\,\hat{q}_{j'}\,(1-\hat{q}_{j'})\, x_{j'} x_{j'}^T\right)^{-1},\end{aligned}$$

où $\hat{q}_i$ pour $i \in S$ est le score de propension estimé pour l'unité i dans le modèle (4.2). L'écart de Taylor pour chaque unité mesure la variation de l'estimateur non linéaire, $\bar{y}$ dans notre cas, comme si l'unité avait été supprimée de l'échantillon. L'estimateur de la variance par linéarisation en séries de Taylor de $\bar{y}$ est ensuite calculé approximativement par (4.4), où $\text{var}\left(\sum_{i \in S} z_i\right)$ prend en compte la variabilité attribuable à l'échantillonnage complexe de S. D'après les conclusions de Wang et coll. (2021), on peut prouver que $\bar{y}$ est convergent par rapport au plan de sondage et $\text{var}_{\text{TL}}(\bar{y}) = O\left(\frac{1}{n_c}\right) + O\left(\frac{1}{n_s}\right)$. Les sections 5 et 6 présentent les estimations par la propension logistique ajustée pour illustrer les hypothèses d'échangeabilité conditionnelles à différents scores d'équilibrage. De même, il est possible de dériver les estimateurs de la variance des estimations par pondération de noyau avec les scores d'équilibrage adaptatif, que nous donnerons dans un futur article.

5. Étude par simulations

5.1 Génération de la population

Des études par simulations sont menées pour évaluer les estimations par la PLA basées sur le score d'équilibrage ajusté $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$, ainsi que $b(\mathbf{x}; \hat{\mathbf{B}}_0)$ et $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ à des fins de comparaison. Nous générons

une population finie (PF) de taille $N = 1\ 000\ 000$ avec trois covariables indépendantes $x_1,\ x_2,\ x_3,$ chacune suivant une loi normale standard $N(0,1).$ Le résultat binaire $\mathbf{Y}$ est généré avec la moyenne définie par

$$P(Y=1)=\frac{\exp(\beta_0+x_1\beta_{x_1}+x_2\beta_{x_2}+x_1x_2\beta_{x_1x_2})}{1+\exp(\beta_0+x_1\beta_{x_1}+x_2\beta_{x_2}+x_1x_2\beta_{x_1x_2})}, \tag{5.1}$$

où $\beta_y=(\beta_0,\beta_{x_1},\beta_{x_2},\beta_{x_1x_2})^T$ sont les paramètres du modèle de résultat spécifiés comme étant $\beta_0=-1,\ \beta_{x_1}=0{,}8,\ \beta_{x_2}=0{,}2,\ \beta_{x_1x_2}=0{,}5.$ La moyenne du résultat binaire est d'environ 30 %. Les résultats ont montré une tendance similaire quand $\beta_0=-2$ ou $-3;$ ils ne sont donc pas indiqués.

5.2 Sélection de l'échantillon probabiliste S

Nous sélectionnons un échantillon aléatoire probabiliste S de taille n_s avec remise à partir de la population finie en utilisant un échantillonnage avec probabilité proportionnelle à la taille (PTT), la mesure de la taille de la k^{e} personne de la PF (mos_k) étant définie par

$$\text{mos}_k=\exp\left[a\times(\alpha_0+x_{k1}\alpha_{x_1}+x_{k2}\alpha_{x_2}+x_{k3}\alpha_{x_3}++x_{k1}x_{k2}\alpha_{x_1x_2}+x_{k1}x_{k3}\alpha_{x_1x_3})\right] \tag{5.2}$$

de sorte que la probabilité d'inclusion soit

$$p(k\in S\mid x;U)=\frac{n_s\times\text{mos}_k}{\sum_{k\in U}\text{mos}_k},$$

et que le poids d'échantillon correspondant soit l'inverse de la probabilité d'inclusion, c'est-à-dire $w_k=\frac{\sum_{k\in U}\text{mos}_k}{n_s\times\text{mos}_k}.$ Nous spécifions $\left(\alpha_0,\alpha_{x_1},\alpha_{X_2},\alpha_{x_3},\alpha_{x_1x_2},\alpha_{x_1x_3}\right)=(-1,\ 0{,}5,\ 0,\ 0{,}5,\ 0,\ -0{,}2)$ et supposons $a=0{,}5,$ 1 ou 1,5 pour faire varier le coefficient de variation (CV) des poids de sondage dans S (désignés par w_s), correspondant respectivement à $\text{CV}(w_s)=0{,}38,$ 0,86 ou 1,5. Notons que les variables de sélection dans l'échantillonnage S sont x_1 et $x_3,$ et que l'échantillon probabiliste pondéré par $w_k,\ S,$ se rapproche de la PF.

5.3 Sélection de l'échantillon non probabiliste C

Le processus de sélection sous-jacent pour l'échantillonnage C est inconnu. Nous sélectionnons C de taille $n_c=2\ 500$ dans la PF au moyen de l'échantillonnage PPT avec $\text{mos}_k,$ donné par (5.2) et spécifié pour inclure trois scénarios : 1) un échantillon par quota qui a la même distribution conjointe de x_1 et x_2 que dans la PF, désigné par Quota.$x_1x_2;$ 2) un échantillon par quota qui a la même distribution de x_2 que dans la PF, désigné par Quota.$x_2;$ et 3) un échantillon de volontaires dont les distributions dans x_1 ou x_2 sont différentes de celles dans la PF, désigné par Volontaire. La variable x_3 n'est pas prédictive du résultat et, par conséquent, n'induit aucun biais dans l'estimation de la moyenne de la PF (Li, Irimata, He et Parker, 2022). Le tableau 5.1 résume les paramètres du modèle pour la génération de résultat dans (5.1), la sélection de l'échantillon probabiliste S et la sélection de trois échantillons non probabilistes dans (5.2). Nous changeons la taille de l'échantillon probabiliste $n_s=1\ 250,\ 2\ 500,\ 3\ 750$ et la taille de l'échantillon non probabiliste est fixée à $n_c=2\ 500.$ Les poids de sondage associés aux unités C sont masqués dans l'analyse.

Tableau 5.1
Spécifications des paramètres du modèle pour la génération de résultat, la sélection de l'échantillon probabiliste (*S*) et la sélection de l'échantillon non probabiliste (*C*).

Modèle	Ordonnée à l'origine	x_1	x_2	x_3	x_1x_2	x_1x_3
Résultat	-1	0,8	0,2	0	0,5	0
Sélection de l'échantillon S	-1	0,5	0	0,5	0	-0,2
Participation de l'échantillon C						
Quota. x_1x_2	-1	0	0	0,5	0	0
Quota. x_2	-1	0,5	0	0	0	-0,2
Volontaire	-1	0,5	0,5	0,5	0	-0,2

Les trois estimations par la propension logistique ajustée (4.3) fondées sur le score d'équilibrage adaptatif $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$, celui non pondéré $b(\mathbf{x};\hat{\mathbf{B}}_0)$ et celui pondéré $b(\mathbf{x};\hat{\mathbf{B}}_w)$ sont calculées pour chacune des $R=$ 1 000 exécutions de simulation et évaluées comme suit :

- Biais relatif (RelBias%) = Biais (= moyenne des R moyennes simulées − moyenne de la population) divisé par la moyenne de la population $\times 100\,\%$.
- Variance empirique (VE) = Variance des R moyennes simulées $\times 10^4$.
- Ratio de variance (RV) = variance LT/variance empirique.

Pour construire les scores d'équilibrage estimés, la fonction $g(\mathbf{x}_i)$ dans (3.1) à (4.2) comprend non seulement les effets principaux de x_1, x_2, x_3, mais aussi leurs effets d'interaction par paires. On s'attend à ce que les estimations par la PLA fondées sur $b(\mathbf{x};\hat{\mathbf{B}}_w)$ soient approximativement sans biais, mais avec une variance gonflée en raison des poids différentiels; les estimations par la PLA fondées sur $b(\mathbf{x};\hat{\mathbf{B}}_0)$ ont la plus petite variance, mais elles peuvent être biaisées. En revanche, on s'attend à ce que les estimations par la PLA fondées sur le score d'équilibrage adaptatif $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ soient approximativement sans biais et comportent une plus petite variance selon les vrais modèles de propension.

5.4 Résultats

Le tableau 5.2 présente le biais relatif (%) des estimations par la PLA basées sur les trois scores d'équilibrage au moyen d'échantillons non probabilistes de Quota. x_1x_2, Quota. x_2 et Volontaire. À des fins de comparaison, nous incluons également les estimations non pondérées. Nous pouvons faire trois observations : 1) Comme on pouvait s'y attendre, les estimations non pondérées sont sans biais pour Quota. x_1x_2, mais fortement biaisées pour les échantillons Quota. x_2 et Volontaire. Ce résultat est conforme aux conclusions de Li et coll. (2022). 2) Pour corriger cela, les scores d'équilibrage de $b(\mathbf{x};\hat{\mathbf{B}}_w)$ ou $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ apparient l'échantillon Quota. x_2 ou Volontaire avec la distribution conjointe de x_1 et x_2 dans la PF et, par conséquent, produisent des estimations approximativement sans biais pour les trois échantillons non probabilistes. 3) En revanche, score non pondéré $b(\mathbf{x};\hat{\mathbf{B}}_0)$ donne des estimations biaisées puisqu'il ne s'agit pas d'une fonction monotone ou plus fine de la propension à la participation estimée des trois échantillons non probabilistes.

Tableau 5.2
Biais relatif (%) des estimations par la propension logistique ajustée de la moyenne de la population ($\overline{Y} = 0{,}3$) avec les tailles d'échantillon probabiliste et non probabiliste $n_s = n_c = 2\,500$ et $\mathrm{CV}(w_s) = 0{,}86$.

	Quota. $x_1 x_2$ $\mathrm{CV}(w_c) = 0{,}53$	Quota. x_2 $\mathrm{CV}(w_c) = 0{,}6$	Volontaire $\mathrm{CV}(w_c) = 1{,}10$
Sans pondération	-1,33	24,33	33,67
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	-1,33	-1,33	-1,33
$b(\mathbf{x}; \hat{\mathbf{B}}_0)$	19,33	19,33	19,33
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	-1,33	-1,33	-1,33

Ensuite, nous comparons au tableau 5.3 les deux estimations PLA sans biais avec $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ et $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ pour ce qui est de leur efficacité quand on fait varier les coefficients de variation (CV) des poids de l'échantillon probabiliste $\mathrm{CV}(w_s) = 0{,}38$, 0,86 ou 1,50. Nous pouvons faire trois observations. Premièrement, quand $\mathrm{CV}(w_s)$ augmente, la variance augmente comme nous nous y attendions. Par exemple, quand nous utilisons $b(\mathbf{x}; \hat{\mathbf{B}}_w)$, la variance empirique augmente pour passer de 1,00 à 1,12 puis à 1,30 pour Quota. $x_1 x_2$. Deuxièmement, quand $\mathrm{CV}(w_s)$ augmente, le gain d'efficacité de $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ sur $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ augmente. Par exemple, la différence relative des deux variances empiriques passe de 1 % (= (1 − 0,99) / 1,00) à 4 % (= (1,12 − 1,07) / 1,12) à 12 % (= (1,3 − 1,14) / 1,30) quand $\mathrm{CV}(w_s)$ augmente pour passer de de 0,38 à 0,86 puis à 1,5 pour Quota. $x_1 x_2$. Troisièmement, si l'on compare les trois échantillons non probabilistes, le gain d'efficacité de $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ sur $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ est le plus grand pour Quota. $x_1 x_2$. Intuitivement, les pseudo-pondérations créées pour Quota. $x_1 x_2$ sont non informatives et ajoutent donc une variance supplémentaire en raison de l'estimation de $b(\mathbf{x}; \hat{\mathbf{B}}_w)$.

Tableau 5.3
Variance empirique ($\times 10^4$) de deux estimations par la propension logistique ajustée sans biais selon des coefficients variables de variation de poids d'échantillon probabiliste $\mathrm{CV}(w_s)$, $n_s = n_c = 2\,500$.

	Quota. $x_1 x_2$	Quota. x_2	Volontaire
	$\mathrm{CV}(w_s) = 0{,}38$		
Sans pondération	0,81	0,94	0,81
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,00	0,97	1,44
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	0,99	0,98	1,45
	$\mathrm{CV}(w_s) = 0{,}86$		
Sans pondération	0,85	0,90	0,99
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,12	1,00	1,62
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1,07	1,02	1,64
	$\mathrm{CV}(w_s) = 1{,}50$		
Sans pondération	0,85	0,90	0,99
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,30	1,11	1,72
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1,14	1,07	1,68

Le tableau 5.4 présente la variance empirique (VE) dans le panneau de gauche et le ratio des variances (RV) dans le panneau de droite pour les estimations par la PLA quand nous varions la taille des échantillons probabilistes (n_s = 1 250; 2 500; 3 750) ayant une taille d'échantillon non probabiliste fixe n_c = 2 500. Nous pouvons faire trois observations. Premièrement, la variance empirique diminue quand n_s augmente : par exemple, la VE des estimations par la PLA avec $b(\mathbf{x}; \hat{B}_w)$ pour Quota. $x_1 x_2$ diminue pour passer de 1,33 à 1,08 puis à 0,99. Toutefois, la différence diminue, ce qui signifie une réduction de la VE plus grande de 0,25 (= 1,33 − 1,08) – quand n_s augmente pour passer de 1 250 à 2 500, comparativement à une baisse modérée de 0,09 (= 1,08 − 0,99) quand n_s augmente pour passer de 2 500 à 3 750. Ce résultat est attribuable au fait que $\mathrm{Var}(\bar{y}) = O\left(\frac{1}{n_c}\right) + O\left(\frac{1}{n_s}\right)$ est dominé par $O\left(\frac{1}{n_c}\right)$ quand $n_s > n_c$ et, par conséquent, le gain d'efficacité est modéré si n_s est augmenté une fois que $n_s > n_c$. Deuxièmement, si nous comparons les deux scores d'équilibrage, $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$ est plus efficace que $b(\mathbf{x}; \hat{B}_w)$ quand n_s est petit. Intuitivement, quand $n_s < n_c$, la variance des estimations par la PLA est dominée par l'échantillon probabiliste S, qui a des poids de sondage différentiels w_s et induit donc une grande variabilité lors de l'estimation de $\hat{B}_w$. Cela se produit particulièrement pour les échantillons par quota où les poids de sondage utilisés pour l'estimation de $b(\mathbf{x}; \hat{B}_w)$ sont approximativement non informatifs et ajoutent par conséquent une variance supplémentaire. Troisièmement, l'estimateur de la variance par linéarisation en séries de Taylor (LT) proposé obtient généralement de bons résultats, le ratio des variances se rapprochant de 1 (voir la partie de droite dans le tableau 5.4). Toutefois, la variance LT fondée sur $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$, surestime la variance pour Quota. $x_1 x_2$ quand n_s est petit. On constate que le RV pour $b(\mathbf{x}; \hat{B}_0, \hat{\gamma}_w^T)$ est plus proche de 1 quand n_s augmente ou quand $\mathrm{CV}(w_s)$ est petit (résultats non indiqués).

Tableau 5.4
Variance empirique ($\times 10^4$) et ratio des variances de deux estimations par la propension logistique ajustée sans biais avec des tailles d'échantillon probabiliste variables de n_s, $\mathrm{CV}(w_s)$ = 0,86 et n_c = 2 500.

	Variance empirique (VE)			Ratio des variances (RV)		
	Quota. $x_1 x_2$	Quota. x_2	Volontaire	Quota. $x_1 x_2$	Quota. x_2	Volontaire
Sans pondération	0,81	0,87	1,03	1,02	0,95	0,86
			n_s = 1 250			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,33	1,18	1,77	1,02	0,98	0,96
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1,17	1,08	1,80	1,41	1,35	1,14
			n_s = 2 500			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	1,08	0,98	1,65	1,06	1,01	0,93
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	1,00	0,96	1,69	1,31	1,23	1,03
			n_s = 3 750			
$b(\mathbf{x}; \hat{\mathbf{B}}_w)$	0,99	0,94	1,60	1,08	1,00	0,93
$b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$	0,95	0,94	1,63	1,26	1,15	1,00

En résumé, au moyen d'études par simulations, on a observé que des estimations par la PLA fondées sur $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ et $b(\mathbf{x}; \hat{\mathbf{B}}_w)$ sont approximativement sans biais et d'une efficacité comparable quand l'échantillon probabiliste de référence a une grande taille d'échantillon n_s ou des poids de sondage stables

avec un petit $\mathrm{CV}(w_s)$. En revanche, quand l'échantillon probabiliste de référence a un petit n_s ou des poids de sondage variables, $b(\mathbf{x}; \hat{\mathbf{B}}_0, \hat{\boldsymbol{\gamma}}_w^T)$ a tendance à produire des estimations plus efficaces, en particulier pour les échantillons par quota. Les spécialistes des enquêtes doivent choisir une enquête de référence ayant une taille suffisamment grande et des poids de sondage stables, car le gain d'efficacité obtenu en augmentant n_s est modéré une fois que $n_s > n_c$.

6. Analyse des données de séropositivité au SARS-CoV-2 des National Institutes of Health

Le principal objectif de l'étude sur la séropositivité au SARS-CoV-2 est d'estimer la prévalence de la séropositivité aux anticorps du virus SARS-CoV-2 dans la population cible composée d'adultes de 18 ans et plus vivant aux États-Unis qui n'ont pas reçu de diagnostic de COVID-19 pendant la première phase de la pandémie, d'avril à août 2020. Dans les semaines après l'annonce de l'étude, plus de 460 000 personnes se sont portées volontaires. L'étude ne pouvait toutefois se permettre qu'un sous-ensemble de ces volontaires. Un échantillon par quota a été sélectionné en fonction de 6 variables de quota, à savoir le groupe d'âge, la race, le sexe, l'origine ethnique, la densité de la population et la région géographique, pour qu'il corresponde approximativement à la répartition de ces variables chez les adultes des États-Unis. Quelque 8 058 participants ont répondu au questionnaire sur les facteurs cliniques et fourni des échantillons de sang servant à évaluer la séropositivité. L'échantillon prélevé dans le cadre de l'étude sur la séropositivité au SRAS-CoV-2 a été appelé « échantillon de la COVID ». Bien que l'échantillon de la COVID ait été un échantillon aléatoire ayant des probabilités de sélection connues tiré du bassin de volontaires de Kalish, Klumpp-Thomas, Hunsberger, Baus, Fay, Siripong, Wang, Hicks, Mehalko, Travers, Drew, Pauly, Spathies, Ngo, Adusei, Karkanitsa, Croker, Li, Graubard, Czajkowski, Belliveau, Chairez, Snead, Frank, Shunmugavel, Han, Giurgea, Rosas, Bean, Athota, Cervantes-Medina, Gouzoulis, Heffelfinger, Valenti, Caldararo, Kolberg, Kelly, Simon, Shafiq, Wall, Reed, Ford, Lokwani, Denson, Messing, Michael, Gillette, Kimberly, Reis, Hall, Esposito, Memoli et Sadtler (2021), ce bassin de volontaires est un échantillon non aléatoire de la population cible aux États-Unis et peut présenter un biais de sélection élevé.

Pour aider à corriger le biais de sélection, nous utilisons le Behavioral Risk Factor Surveillance System (BRFSS ou Système de surveillance des facteurs de risque comportementaux, Centers for Disease Control and Prevention, 2022) comme enquête de référence. Le BRFSS se compose d'enquêtes annuelles à l'échelle des États américains, qui sont combinées en une enquête représentative nationale comportant des observations à grande échelle au niveau de l'État. En plus des 6 variables de quota, 10 variables démographiques et liées à la santé sont recueillies dans le BRFSS, lesquelles sont également prédictives de la séropositivité, mais ne sont pas utilisées dans l'échantillonnage par quota. Après avoir supprimé les observations ayant des valeurs manquantes pour une ou plusieurs des 16 variables, n_s = 367 165 participants au total ont été inclus dans l'analyse. Le CV des poids de sondage du BRFSS est $\mathrm{CV}(w_s)$ = 1,92.

Le tableau 6.1 montre la distribution pondérée de l'échantillon pour les 16 variables du BRFSS et l'échantillon de la COVID. Comme on s'y attendait, les distributions des 6 variables de quota dans les deux échantillons sont très proches. Pour les 10 variables démographiques et liées à la santé, la plupart des distributions diffèrent considérablement entre les deux échantillons. En général, les participants de l'échantillon de la COVID ont tendance à être plus scolarisés, propriétaires de leur logement, employés et en meilleure santé. À titre d'exemple, 84 % des participants de l'échantillon de la COVID, comparativement à 29 % pour le BRFSS pondéré, possèdent un diplôme d'études collégiales ou de niveau supérieur. Par conséquent, un biais de sélection existe dans l'échantillon de la COVID, et notre objectif est de réduire le biais de sélection dans l'estimation de la séropositivité non diagnostiquée au SRAS-CoV-2.

Le tableau 6.2 montre les estimations par la PLA de la prévalence de la séropositivité non diagnostiquée fondées sur les trois scores d'équilibrage. Comme nous l'avons indiqué, l'estimation par la PLA fondée sur $b(\mathbf{x};\hat{\mathbf{B}}_w)$ a permis de détecter un taux de séropositivité de 4,65 %, proche du taux de 4,67 % détecté au moyen de $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$. Les deux erreurs-types correspondantes sont également proches (0,78 contre 0,77). En revanche, le $b(\mathbf{x};\hat{\mathbf{B}}_0)$ non pondéré donne un taux de séropositivité de 3,95 %, proche de la moyenne non pondérée de 3,77 %, les deux présentant un biais de sélection. Il est intéressant de mentionner que le score d'équilibrage adaptatif $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ a produit des pseudo-pondérations stables pour l'échantillon de la COVID avec $\mathrm{CV}(\hat{w}_c) = 2{,}24$, proche du 2,25 du $b(\mathbf{x};\hat{\mathbf{B}}_0)$ non pondéré, et que les deux sont inférieurs à $\mathrm{CV}(\hat{w}_c) = 2{,}33$ produit par le $b(\mathbf{x};\hat{\mathbf{B}}_w)$ pondéré.

Tableau 6.1
Distribution des covariables (%) dans l'échantillon de la COVID par rapport à celles du Behavioral Risk Factor Surveillance System (BRFSS).

	Échantillon de la COVID-19	BRFSS pondéré		Échantillon de la COVID-19	BRFSS pondéré		Échantillon de la COVID-19	BRFSS pondéré
Groupe d'âge			Milieu urbain/rural			Vacciné contre la grippe		
18 à 44 ans	41,6	42,9	Urbain	94,7	93,2	Oui	73,8	51,3
45 à 69 ans	42,6	41,8	Rural	5,3	6,8	Non	26,2	48,7
70 à 95 ans	15,8	15,2	Présence d'enfants			Maladie cardiovasculaire		
Sexe			Oui	32,5	34,7	Oui	4,1	9,5
Homme	47,4	47,8	Non	67,5	65,3	Non	95,9	90,5
Femme	52,6	52,2	Scolarité			Maladie pulmonaire		
Race			<= Études secondaires	2,6	39,4	Oui	18,8	18,7
Blanc seulement	77,5	74,8	Études collégiales	13,8	31,5	Non	81,2	81,3
Noir seulement	9,4	12,6	>= Études collégiales	83,6	29,1	Immunisé		
Autres	13,1	12,5	Propriétaire			Oui	23,4	31,1
Origine ethnique			Propriétaire	75,2	68,8	Non	76,6	68,9
Hispanique	15,9	14,1	Locataire	20,2	25,6	Diabète		
Non hispanique	84,1	85,9	Autre	4,7	5,6	Oui	5,5	11,9
Région			Emploi			Non	94,5	88,1
Nord-est	16,7	17,1	A un emploi	71,2	57,4	Assurance maladie		
Midwest	15,8	17,6	Inactif	23,8	32,2	Oui	97,4	89,0
Centre de l'Atlantique	20,8	17,3	Sans emploi	5,0	10,4	Non	2,6	11,0
Sud/Centre	14,2	15,7						
Montagnes/Sud-ouest	15,5	15,3						
Ouest/Pacifique	17,0	16,9						

Tableau 6.2
Taux de séropositivité non diagnostiquée chez les adultes des États-Unis du 1er avril au 4 août 2020.

	$CV(\hat{w}_c)$	Estimations (%)	Erreur-type* ($\times 10^{-2}$)
Sans pondération	0,00	3,77	0,22
$b(\mathbf{x};\hat{\mathbf{B}}_0)$	2,25	3,94	0,52
$b(\mathbf{x};\hat{\mathbf{B}}_w)$	2,33	4,65	0,78
$b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$	2,24	4,67	0,77

* : pour tenir compte de la variabilité attribuable à l'estimation de $\mathbf{B}$, $\mathbf{B}_0$ ou $\boldsymbol{\gamma}$.

7. Conclusion et discussion

Dans le présent article, nous avons examiné l'échangeabilité du résultat conditionnelle au $b(\mathbf{x};\hat{\mathbf{B}}_w)$ pondéré et au $b(\mathbf{x};\hat{\mathbf{B}}_0)$ non pondéré sont utilisés dans les méthodes existantes de pondération et d'appariement fondées sur le score de propension pour les inférences d'échantillons non probabilistes. Nous proposons un score d'équilibrage adaptatif $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ pour corriger le biais potentiel dans $b(\mathbf{x};\hat{\mathbf{B}}_0)$ en trois étapes : 1) estimer le score d'équilibrage non pondéré $b(\mathbf{x};\hat{\mathbf{B}}_0)$; 2) estimer le facteur de correction du biais $b(\mathbf{x};\hat{\gamma}_w)$; 3) construire $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)=b(\mathbf{x};\hat{\mathbf{B}}_0)+b(\mathbf{x};\hat{\boldsymbol{\gamma}}_w)$, qui est une fonction monotone de la propension à la participation estimée.

Le critère de base pour choisir le score d'équilibrage est qu'il doit être plus fin que, sinon égal à la propension à la participation afin d'équilibrer la distribution de x entre l'échantillon non probabiliste et la population finie. $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ et $b(\mathbf{x};\hat{\mathbf{B}}_w)$ produisent tous deux des estimations sans biais d'une efficacité comparable, $b(\mathbf{x};\hat{\mathbf{B}}_0,\hat{\boldsymbol{\gamma}}_w^T)$ étant plus efficace pour les échantillons par quota quand l'enquête de référence est petite ou a des poids d'échantillon variables. Les spécialistes des enquêtes doivent choisir comme enquête de référence un échantillon suffisamment grand ayant des poids d'échantillon stables. Notons que le gain d'efficacité obtenu en augmentant la taille de l'échantillon probabiliste n_s est modéré une fois que $n_s > n_c$.

Nous avons cerné deux limites : 1) le score d'équilibrage adaptatif est construit en supposant l'exactitude du modèle de propension par régression logistique aux étapes 1 et 2 pour obtenir les scores d'équilibrage non pondérés et le facteur de correction du biais; 2) dans les deux étapes, la régression logistique est supposée avoir la même forme fonctionnelle. En conséquence, nous proposons les deux prolongements suivants de l'article dans de futurs travaux de recherche : 1) Permettre une forme fonctionnelle différente à l'étape 2, où nous modélisons la probabilité pour la sélection de l'échantillon de référence, à partir de la forme fonctionnelle supposée à l'étape 1. À l'aide de variables de sélection connues et de la probabilité de sélection pour chaque unité de l'enquête de référence, il est possible de mettre en œuvre des diagnostics de modèle comme une courbe ROC (courbe caractéristique de la performance d'un test) pour faciliter la sélection du modèle. 2) Construire plusieurs modèles de propension. Un modèle de régression logistique a été adapté pour estimer les scores de propension aux étapes 1 et 2 de la section 4. Toutefois, la spécification erronée du modèle de régression logistique pourrait donner des scores de propension mal estimés qui vont à l'encontre de l'hypothèse (2.1) et, par conséquent, donnent des estimations biaisées. Les méthodes non

paramétriques, comme les méthodes d'apprentissage automatique, peuvent fournir des solutions de rechange, qui assouplissent les spécifications du modèle paramétrique supposé concernant la sélection des variables, la forme fonctionnelle et la sélection des termes des polynômes et des interactions multidirectionnelles spécifiées dans la modélisation paramétrique.

Dans l'article, nous avons discuté de la façon de construire des scores d'équilibrage qui satisfont à l'hypothèse de l'EC de sorte que la distribution du résultat soit échangeable entre l'échantillon non probabiliste et la population finie. Notons que le score d'équilibrage est une fonction des covariables observées $\mathbf{x}$ qui sont recueillies à la fois dans l'échantillon non probabiliste C et dans l'enquête de référence S. S'il manque des covariables importantes dans S ou C, alors quel que soit le score d'équilibrage choisi, les estimations moyennes de la PF sont inévitablement biaisées. Il reste des éléments importants à prendre en considération : Quelles variables doivent être recueillies dans C et S? Comment les questions de l'enquête seront-elles harmonisées dans la collecte des données de C et S? Et comment les erreurs de mesure ou de déclaration peuvent-elles être réduites dans la conception du questionnaire? Le mode de traitement de ces questions peut être essentiel pour satisfaire à l'hypothèse de l'EC dans les méthodes d'ajustement fondées sur le score de propension pour l'analyse des échantillons non probabilistes. En résumé, comme l'exige l'hypothèse de l'échangeabilité conditionnelle, il est important d'avoir des enquêtes de référence de grande qualité qui permettent de recueillir des ensembles complets de variables comportant un minimum d'erreurs de mesure et de déclaration, qui ont une taille d'échantillon suffisamment grande et qui sont bien conçues avec des poids d'échantillonnage informatifs et stables.

Remerciements

Nous remercions Barry Graubard pour ses discussions stimulantes et Lingxiao Wang pour son aide en matière de simulation. La présente étude est financée par la subvention NIH R03 CA252782.

Bibliographie

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. et Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1(2), 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/fr/pub/12-001-x/2020001/article/00001-fra.pdf.

Brick, J., et Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33, 735-752. DOI: https://doi.org/10.1515/jos-20170034.

Centers for Disease Control and Prevention (2022). Behavioral Risk Factor Surveillance System: Annual survey data. Atlanta, Géorgie: Centers for Disease Control and Prevention, US Department of Health and Human Services. Récupéré de http://www.cdc. gov/brfss/annual_data/annual_data.htm.

Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021.

Elliott, M. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2.

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

Groves, R., et Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72(2), 167-189. DOI: https://doi.org/10.1093/poq/nfn011.

Kalish, H., Klumpp-Thomas, C., Hunsberger, S., Baus, H.A., Fay, M.P., Siripong, N., Wang, J., Hicks, J., Mehalko, J., Travers, J., Drew, M., Pauly, K., Spathies, J., Ngo, T., Adusei, K.M., Karkanitsa, M., Croker, J.A., Li, Y., Graubard, B.I., Czajkowski, L., Belliveau, O., Chairez, C., Snead, K.R., Frank, P., Shunmugavel, A., Han, A., Giurgea, L.T., Rosas, L.A., Bean, R., Athota, R., Cervantes-Medina, A., Gouzoulis, M., Heffelfinger, B., Valenti, S., Caldararo, R., Kolberg, M.M., Kelly, A., Simon, R., Shafiq, S., Wall, V., Reed, S., Ford, E.W., Lokwani, R., Denson, J.-P., Messing, S., Michael, S.G., Gillette, W., Kimberly, R.P., Reis, S.E., Hall, M.D., Esposito, D., Memoli, M.J. et Sadtler, K. (2021). Undiagnosed SARS-CoV-2 seropositivity during the first six months of the COVID-19 pandemic in the United States. *Sci Transl Med*, 13(601), eabh3826.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K. et Gimenez, A. (2016). *Evaluating Online Nonprobability Surveys.* Washington, DC: Pew Research Center.

Kern, C., Li, Y. et Wang, L. (2021). Boosted kernel weighting – Using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 9(5), 1088-1113.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/9781118032619.

Lee, S., et Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37, 319-343.

Li, Y., et Graubard, B. (2012). Pseudo semiparametric maximum likelihood estimation exploiting gene environment independence for population-based case-control studies with complex samples. *Biostatistics*, 13, 711-723.

Li, Y., Graubard, B. et DiGaetano, R. (2011). Weighting methods for population-based case-control studies with complex sampling. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 60, 165-185.

Li, Y., Graubard, B., Huang, P. et Gastwirth, J. (2015). Extension of the Peters–Belson method to estimate health disparities among multiple groups using logistic regression with survey data. *Statistics in Medicine*, 34, 595-612.

Li, Y., Irimata, K.E., He, Y. et Parker, J. (2022). Variable inclusion strategies through directed acyclic graphs to adjust health surveys subject to selection bias for producing national estimates. *Journal of Official Statistics*, 38(3), 1-27.

Mercer, A.W., Kreuter, F., Keeter, S. et Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, 250-271. DOI: https://doi.org/10.1093/poq/nfw060.

Pinsky, P.F., Miller, A., Kramer, B.S., Church, T., Reding, D., Prorok, P., Gelmann, E., Schoen, R.E., Buys, S., Hayes, R.B. et Berg, C.D. (2007). Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *American Journal of Epidemiology*, 165(8), 874-881.

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

Rivers, D. (2007). Sampling for web surveys. Document présenté aux *Joint Statistical Meetings - Section on Survey Research Methods*.

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

Rubin, D. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6, 34-58.

Scott, A., et Wild, C. (2001). The analysis of clustered case-control studies. *Journal of the Royal Statistical Society Series C*, 50, 389-401.

Shah, B.V. (2004). Commentaires à propos de l'article "Estimateurs de variance par linéarisation pour des données d'enquête" par A. Demnati et J.N.K. Rao. *Techniques d'enquête*, 30, 1, 18. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2004001/article/6991-fra.pdf.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1).

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.

Valliant, R., et Dever, J. (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 analyses: A kernel weighting approach. *Journal of the Royal Statistical Society Series A*, 183, 1293-1311.

Wang, L., Graubard, B.I., Katki, H.A. et Li, Y. (2022). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. *Revue Internationale de Statistique*, 90, 146-164.

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. DOI: https://doi.org/10.1002/sim.9122.