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

Techniques d’enquête

Commentaires à propos de l’article « Inférence statistique avec des échantillons d’enquête non probabiliste »

par Michael R. Elliott

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, nº 2, pp. 347-358
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste »

Michael R. Elliott[1]

Résumé

Cet exposé vise à approfondir l'examen de Wu sur l'inférence à partir d'échantillons non probabilistes, ainsi qu'à mettre en évidence les aspects qui constituent probablement d'autres pistes de recherche utiles. Elle se termine par un appel en faveur d'un registre organisé d'enquêtes probabilistes de grande qualité qui visera à fournir des renseignements utiles à l'ajustement d'enquêtes non probabilistes.

Mots-clés : Analyse de sensibilité; estimation doublement robuste; pseudo-pondération; score de propension.

1. Introduction

Je remercie M. Changbao Wu de son excellent examen des travaux précédents et des questions ouvertes pour l'inférence statistique à partir d'échantillons non probabilistes. Étant donné l'ampleur et l'évolution rapide des travaux dans ce domaine, il est clair que M. Wu n'a pas été en mesure d'être exhaustif. Ma propre analyse comporte elle aussi des limites, mais je vais aborder quelques approches supplémentaires qui se rapportent aux sujets examinés par mon confrère. Je me pencherai également sur la question de la modélisation par rapport à la pondération pour différentes cibles inférentielles, et me servirai de l'analyse et des conclusions de M. Wu pour souligner l'importance cruciale des échantillons probabilistes – notamment les études de grande qualité qui portent sur l'estimation de covariables pertinentes – dans le but d'améliorer l'inférence pour la profusion des échantillons non probabilistes utilisés pour remplacer les échantillons probabilistes traditionnels dans un grand nombre d'études et de contextes statistiques officiels. Pour éviter toute confusion de notation, toutes les notations suivront celles de Wu, sauf lorsque de nouvelles notations seront nécessaires.

La section 2 examine d'autres approches pour combiner les données tirées d'enquêtes probabilistes et celles d'enquêtes non probabilistes. La section 3 examine brièvement la question de la pondération par rapport à la modélisation au moment de rajuster des données d'enquête non probabilistes. La section 4 examine certaines découvertes récentes liées aux analyses de sensibilité des hypothèses types applicables à l'ajustement des données d'enquêtes non probabilistes à l'aide des données d'enquêtes probabilistes. Enfin, la section 5 se termine par un appel à concevoir systématiquement un ensemble d'enquêtes probabilistes dans le but explicite de rajuster les enquêtes non probabilistes.

2. Approches supplémentaires pour combiner les données d'enquêtes probabilistes et les données d'enquêtes non probabilistes

L'article de Wu suit le principe général fondé sur les trois points suivants : 1) appliquer une estimation de modèle suivie d'un calibrage des estimations des distributions de covariables d'échantillons

1. Michael R. Elliott, Département de biostatistique, University of Michigan, M4124 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109. Courriel : mrelliot@umich.edu.

probabilistes; 2) élaborer des estimations de scores de propension selon les écarts entre les données d'échantillons probabilistes et celles d'échantillons non probabilistes; 3) trouver des méthodes doublement robustes qui combinent les points 1 et 2 pour faire en sorte qu'un seul des deux modèles sous-jacents soit exact.

2.1 Estimateurs de score de propension

Rivers (2007) semble avoir été le premier à suggérer d'estimer le score de propension en utilisant la régression logistique avec l'appartenance à l'échantillon non probabiliste comme résultat et en utilisant comme poids d'inclusion la réciproque des scores de propension résultants. Cette approche a été systématisée davantage dans Valliant et Dever (2011). Séparément, en utilisant des résultats simples du théorème de Bayes et de l'analyse discriminante décrite pour la première fois par Elliott et Davis (2005), Elliott, Resler, Flannagan et Rupp (2010) et Elliott (2013) ont élaboré un estimateur quelque peu différent prenant la forme

$$\hat{\pi}_i^A(\mathbf{x}_i, \boldsymbol{\alpha}) \;=\; \hat{P}(i \in S_A) \propto P(i \in S_B) \frac{\hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)}{\hat{P}\left(i \in S_B \,\middle|\, i \in S_A \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)}. \tag{2.1}$$

Le segment $\hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$ peut être obtenu par une régression logistique, ou au moyen d'une série d'approches qui reposent sur l'apprentissage automatique, comme les machines à vecteurs de support (Soentpiet, 1999), l'estimation pondérée du maximum de vraisemblance (Van Der Laan et Rubin, 2006) ou les arbres de régression additifs bayésiens (Chipman, George et McCulloch, 2010), et où le segment $\hat{P}\left(i \in S_A \,\middle|\, i \in S_B \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$ est obtenu en tant que $1 - \hat{P}\left(i \in S_A \,\middle|\, i \in S_A \text{ ou } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha}\right)$. En principe, le segment $P(i \in S_B) = 1/d_i^B$ est connu, puisque les probabilités d'échantillonnage sont connues pour tous les éléments de la population, y compris ceux de l'échantillon non probabiliste, mais dans la pratique, les analystes qui n'ont accès qu'aux données à grande diffusion pourraient aussi avoir besoin d'estimer ce paramètre. (En outre, d_i^B peut inclure des ajustements de calage et de non-réponse qui ne sont pas connus pour les éléments de l'échantillon non probabiliste.) Ce dernier point est essentiel, car l'utilisation de l'échantillon probabiliste pour établir des scores de propension en utilisant uniquement les écarts entre l'échantillon non probabiliste et l'échantillon probabiliste sera faussée à moins que l'échantillon probabiliste ne repose sur un plan à probabilités égales *epsem* pour *equal probability selection methods*, comme l'a souligné Wu.

Par contre, Chen, Li et Wu (2020) montrent que l'utilisation d'une approche de pseudo-vraisemblance pour estimer $\hat{\pi}_i^A(\mathbf{x}_i, \boldsymbol{\alpha})$ directement à partir de la vraisemblance de population pour les indicateurs $I(i \in S_A)$ en fonction de $\mathbf{x}_i$ produit un estimateur qui ne requiert pas le segment $P(i \in S_B)$ pour les éléments de l'échantillon non probabiliste, à condition que $\pi_i^A(\mathbf{x}_i, \boldsymbol{\alpha})$ suive un modèle linéaire généralisé avec un lien canonique, c'est-à-dire une régression logistique.

(Aucune de ces approches ne fournit la bonne ordonnée permettant d'obtenir un score de propension réel. Cependant, comme l'indique l'article de Wu, une estimation pondérée repose généralement sur des estimateurs de type Hájek [qui utilisent des poids pour estimer un total de population pour les

dénominateurs; Hájek, 1971], de sorte que des scores de propension estimés jusqu'à l'obtention d'une constante de normalisation sont suffisants.)

2.2 Estimateurs doublement robustes

Si l'inférence est centrée sur une variable particulière Y offerte seulement dans l'échantillon non probabiliste, on peut alors revenir aux estimateurs assistés par un modèle qui remontent à Cassel, Särndal et Wretman (1976) et qui posent comme postulat un modèle d'espérance $E\left(y_i \mid \mathbf{x}_i\right) = m_i$. La combinaison de ce modèle avec les estimations du score de propension de la probabilité de faire partie de l'échantillon non probabiliste (que nous traiterons comme un « échantillon probabiliste inconnu » et que nous examinerons plus en détail dans la section sur les hypothèses) permet d'obtenir les estimateurs de la formule

$$\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B}\sum_{i\in S_B} d_i^B\hat{m}_i \tag{2.2}$$

correspondant au segment $\hat{\mu}_{\mathrm{DR}2}$ de la formule (4.11) indiquée dans l'article de Wu. On pourrait penser que tout biais découlant d'une erreur de spécification du modèle dans l'estimation de m_i dans $\frac{1}{\hat{N}^B}\sum_{i\in S_B} d_i^B\hat{m}_i$ sera égal et de signe opposé à $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ si le modèle pour π_i^A est correctement spécifié. À l'inverse, si le modèle pour π_i^A est mal spécifié alors que m_i est correctement spécifié, $y_i-\hat{m}_i$ sera iid avec une moyenne nulle et, par conséquent, $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ aura aussi une moyenne de zéro, ce qui donnera un estimateur sans biais. Chen, Valliant et Elliott (2019) ont utilisé la méthode LASSO pour les prévisions, de pair avec des estimateurs par la régression généralisée (McConville, Breidt, Lee et Moisen, 2017) lorsque $\mathbf{X}$ est de grande dimension. Comme Wu le fait remarquer, Wu et Sitter (2001) montrent l'équivalence entre l'estimateur par la régression généralisée appliqué aux valeurs prédites et les estimateurs doublement robustes de la formule dans la section (2.2), qui indique que l'approche adoptée par Chen et coll. (2019) est équivalente à la formule de la section 2.2 avec estimation de la méthode LASSO pour m_i et une hypothèse d'échantillonnage aléatoire simple pour l'échantillon non probabiliste.

Un inconvénient d'utiliser la formule (2.1) par rapport à la méthode de Chen et coll. (2020) comme estimateur de π_i^A, et donc de d_i^A, tient au fait qu'on doive connaître les poids de l'échantillon probabiliste d_i^B, ou qu'il faille au moins les estimer pour l'échantillon non probabiliste. L'un des avantages de la formule (2.1) est que les modèles non linéaires et les méthodes d'apprentissage automatique peuvent être utilisés dans l'estimation. Rafei, Flannagan et Elliott (2020) utilisent les arbres de régression additifs bayésiens pour estimer m_i et π_i^A, ce qui permet de réduire l'incidence d'une erreur de spécification du modèle. Les simulations ont montré une amélioration considérable de la réduction du biais et de la variance par rapport à la méthode de Chen et coll. (2020) en cas de spécification erronée des modèles linéaires. On peut effectuer l'estimation de la variance en adaptant les règles d'imputation multiple de Rubin : à partir de M tirages indépendants selon les arbres de régression additifs bayésiens, la moyenne des variances calculées traitant le tirage de d_i^A comme connu reposant sur des estimateurs classiques d'un

plan d'échantillonnage complexe et ajouté à $\frac{M+1}{M}$ fois la variance des estimations ponctuelles calculées pour les tirages de d_i^A donne un estimateur de variance approximativement sans biais.

Une autre approche de l'estimation doublement robuste consiste à utiliser le score de propension comme « score d'équilibrage » le plus grossier possible qui contient toute l'information sur l'association entre l'indicateur d'échantillonnage et le résultat visé. Cela a mené à la mise au point d'estimateurs de la moyenne qui utilisent des fonctions lisses de pondération pour produire des estimateurs convergents susceptibles d'être plus efficaces lorsque les poids sont très variables ou faiblement liés au résultat (Elliott et Little, 2000; Zheng et Little, 2005). Zhou, Elliott et Little (2019) ont étendu cette idée à l'inférence causale dans les études non randomisées, où la probabilité d'attribution à un traitement ou à une exposition (score de propension) est estimée comme une fonction des covariables $P_Z\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ utilisant la régression logistique, puis les résultats possibles non observés Y^z au volet de traitement $z_i' \neq z_i$ pour le traitement observé z_i sont imputés à partir de

$$Y_i^Z \sim N\left(s\left(\hat{P}_Z^*\left(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}\right) \middle| \boldsymbol{\theta}_Z\right)\right) + g_Z\left(\hat{P}^*\left(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}\right), \mathbf{x}_i \middle| \boldsymbol{\beta}_Z\right), \sigma^2 \tag{2.3}$$

où P^* est la transformation logit de P, $s(\hat{P}_Z^* \,|\, \boldsymbol{\theta}_Z)$ désigne une spline pénalisée avec nœuds fixes de propension (Eilers et Marx, 1996), et $g_Z(\hat{P}^*, \mathbf{x}_i \,|\, \boldsymbol{\beta}_Z)$ représente une fonction générale des covariables, y compris les scores de propension. L'estimateur qui en résulte est doublement robuste en ce sens que si $P_Z\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ ou $E(Y^z) = g_Z(\hat{P}^*, \mathbf{x}_i \,|\, \boldsymbol{\beta}_Z)$ est correctement spécifié, $Y^{(z)}$ sera approximativement sans biais (voir Zhang et Little, 2009). Il peut être mis en œuvre dans l'échantillon non probabiliste en remplaçant $\hat{P}_Z(\mathbf{x}_i, \boldsymbol{\alpha})$ dans le modèle de moyenne pour (2.3) par la valeur $\hat{\pi}_i^A$ estimée à l'aide de (2.1) pour obtenir un tirage de $Y_i^{(b)}$. (Veuillez noter que cela nécessite l'obtention de $\hat{\pi}_i^A$ pour les éléments de l'échantillon probabiliste nécessitant une prédiction.) L'inférence peut se faire en obtenant des tirages $b = 1, \ldots, B$ à partir de la distribution postérieure de la quantité estimée de population étudiée, par exemple pour la moyenne de la population

$$Y^{(b)} = \frac{\sum_{i \in S_R} N_i^{(b)} Y_i^{(b)} + \sum_{i \in S_A} \left(y_i - Y_i^{(b)}\right)}{N}$$

où $N_i^{(b)}$ est maintenant une estimation de la population représentée par le poids d_i^R obtenu à partir du bootstrap bayésien en population finie (Little et Zheng, 2007). Des extensions plus complètes du bootstrap bayésien en population finie à des plans d'échantillonnage complexes qui comprennent des mises en grappes et des stratifications sont consultables dans Dong, Elliott et Raghunathan (2014).

Comme dans l'estimation de (2.1), la composante non paramétrique (spline) de (2.3) peut être remplacée par d'autres estimateurs d'apprentissage automatique; voir le chapitre 4 de Rafei (2021) pour la mise en œuvre à l'aide de processus gaussiens. En outre, les applications aux modèles non normaux sont directes, bien qu'elles ne soient pas nécessairement faciles à calculer.

2.3 Estimateurs poststratifiés

Wu décrit également l'utilisation d'estimateurs poststratifiés dans le contexte d'échantillonnage par quota, qui est non seulement une très vieille forme d'échantillonnage non probabiliste, mais en fait la norme avant que Neyman défende l'idée de l'échantillonnage aléatoire stratifié (Neyman, 1934). La section 5 de l'article de Wu propose une solution de rechange solide aux estimations de score de propension obtenues en classant les observations tirées de l'échantillon probabiliste par $\hat{\pi}_i$, en les stratifiant en K strates fondées sur ce classement, et en calculant la proportion prédite de la population appartenant à la k^{e} strate comme proportion des poids d'échantillonnage W_k de cette strate en utilisant l'échantillon probabiliste à l'aide de la formule

$$\hat{\mu}_{\text{PST}} = \sum_k \hat{W}_k \bar{y}_k \tag{2.4}$$

où $\bar{y}_k$ est la moyenne dans la k^{e} strate de l'échantillon non probabiliste. Wu souligne le compromis entre le choix d'un K suffisamment grand pour conserver l'homogénéité à l'intérieur des unités, mais assez petit pour obtenir des estimations stables de $\bar{y}_k$, en indiquant 30 comme ancienne « règle empirique » pour les « tailles d'échantillon [suffisamment] grandes ». J'ajouterais qu'une approche plus officielle dont il est question dans Little (1986) consiste à adopter une méthode de création de strates (dans un contexte d'ajustement pour la non-réponse) qui limite l'erreur quadratique moyenne en maximisant la variance entre les strates à la variance dans la strate. Il semblerait qu'une telle approche conviendrait aussi à l'estimateur poststratifié des échantillons non probabilistes.

Une approche plus directe pour obtenir des estimations à l'aide d'un estimateur poststratifié est la régression multiniveau et la poststratification (Wang, Rothschild, Goel et Gelman, 2015; Downes et Carlin, 2020). Dans le cas présent, seules les données de l'échantillon non probabiliste sont utilisées dans le modèle de résultat

$$E\left(Y_{k[i]}\right) = \beta_0 + \mathbf{x}_k^T \boldsymbol{\beta} + \sum_j a_{l[k]}^j \tag{2.5}$$

où $k = 1, \ldots, K$ indexe la poststrate créée à partir des variables $j = 1, \ldots, J$, $a_{l[k]}^j \sim N(0, \sigma_j^2)$ pour $l = 1, \ldots, L_j$ et $l[k]$ met en correspondance la cellule k de la postrate avec la catégorie l de la variable j appropriée. L'estimateur poststratifié est toujours donné par (2.4), $\hat{W}_k$ étant maintenant remplacé par les totaux W_k de population connus; l'inférence postérieure est obtenue par les tirages postérieurs de β_0, $\boldsymbol{\beta}$ et $a_{l[k]}^j$ pour obtenir un tirage de

$$\hat{\mu}_{\text{PST}}^{(b)} = \sum_k W_k \left[\frac{1}{n_k} \sum_{i \in k} \left(\beta_0^{(b)} + \mathbf{x}_k^T \boldsymbol{\beta}^{(b)} + \sum_j a_{l[k]}^{j(b)} \right) \right].$$

Bien que cet estimateur ne soit pas doublement robuste au sens strict, il a démontré son bon fonctionnement dans certaines applications où J est suffisamment grand pour saisir tous les écarts importants entre un échantillon probabiliste et un échantillon non probabiliste, et lorsque l'échantillon non

probabiliste est suffisamment grand pour permettre une estimation plutôt exacte de $a_{l[k]}^{j}$. En l'absence de distributions conjointes connues de $\mathbf{X}$ à grande dimensionnalité, cette approche présente la faiblesse de reposer sur des distributions estimées qui sont instables. Une solution de rechange possible consisterait à remplacer $\bar{y}_k$ par (2.5) dans l'estimateur poststratifié de Wu (2.4), en partant du principe que les poids d'échantillonnage d_i^R résument l'information au sujet de $\mathbf{X}$ dans l'échantillon probabiliste semblable à celle du score de propension pour l'échantillon non probabiliste.

3. Pondération ou modélisation pour l'utilisateur général

L'article de Wu et les addendas ci-dessus ont tendance à suivre les sentiers battus au chapitre du choix entre la pondération et la modélisation dans le contexte de l'inférence de population finie, qui remonte au moins à Hansen, Madow et Tepping (1983). En pensant à ce choix, je crois qu'il est important de faire la distinction entre les modèles utilisés pour en déduire les paramètres dits descriptifs – au sens de Kalton (1983) – et les modèles dignes d'intérêt, les paramètres dits d'analyse dans les modèles de régression, l'analyse des classes latentes, etc. Dans le premier cas, distinguer une cible descriptive d'intérêt Y des covariables de modélisation potentielles $\mathbf{X}$ présente l'avantage de créer des estimateurs doublement robustes qui sont axés sur un seul paramètre descriptif. Cela nécessite également des hypothèses comme A1 à la section 2.1 (le score de propension ne dépend pas de la conditionnalité de Y par rapport à $\mathbf{X}$). Lorsque les modèles eux-mêmes sont les cibles d'intérêt, il est possible que l'élaboration de poids à l'aide de scores de propension pour tenir compte du biais de sélection et, comme Wu le fait remarquer, l'utilisation d'équations d'estimation pondérées standards soit le choix le plus sensé, étant donné qu'on peut habituellement envisager un grand nombre de modèles. Ce choix nécessite une double robustesse, puisqu'il n'y a généralement aucune tentative de modéliser directement les paramètres d'analyse. Trouver des façons d'étendre la double robustesse à une vaste catégorie d'estimations de paramètres de modèle est un exercice qui peut s'avérer fructueux.

4. Hypothèses non vérifiables : découvertes récentes au chapitre de l'analyse de sensibilité

Wu pose quatre hypothèses clés nécessaires pour corriger le biais de sélection dans les enquêtes non probabilistes en utilisant des données d'enquêtes probabilistes : Il s'agit essentiellement d'une « sélection aléatoire » (les covariables des échantillons non probabilistes expliquent la probabilité de sélection dans les échantillons non probabilistes); la « positivité » (tous les éléments de la population ont une probabilité non nulle de sélection dans l'échantillon non probabiliste); l'« indépendance » (les éléments sont sélectionnés de façon indépendante dans l'échantillon non probabiliste); et les « covariables courantes » (il existe une enquête probabiliste avec des covariables dont le sous-ensemble correspond aux covariables nécessaires pour que l'hypothèse de données manquantes au hasard reste valable). Il conviendrait de

mentionner que les deux premières hypothèses nécessitent que l'enquête non probabiliste soit une enquête probabiliste « déguisée », c'est-à-dire qu'il y a réellement des probabilités non nulles de sélection, dans l'enquête non probabiliste, de tous les éléments de la population, mais en tant qu'analystes, nous ne connaissons pas la nature de ces éléments.

Dans la pratique, il se peut qu'aucune de ces hypothèses ne soit vérifiable avec précision. Certaines études récentes ont porté sur l'échec de la première hypothèse, soit celle de la « sélection aléatoire ». Certaines mesures existantes tirées de la littérature sur la non-réponse ont été adaptées dans le cas présent, comme la mesure de l'indicateur R (Schouten, Cobben et Bethlehem, 2009), qui dans ce contexte est la mesure de la variabilité des probabilités de sélection dans l'échantillon non probabiliste :

$$\hat{R} = 1 - 2\sqrt{\frac{1}{n_a - 1}\sum_{i=1}^{n_A}\left(\hat{\pi}_i^A - \sum_{j=1}^{n_A}\hat{\pi}_j^A / n_a\right)^2}$$

$\hat{R}$ varie entre 0 et 1, où la valeur 1 est atteinte lorsque la probabilité de sélection est constante, ce qui permet de penser à un simple échantillon aléatoire présentant un risque moindre de biais de sélection, et la valeur 0 indiquant que tous les éléments sont soit inclus dans la probabilité de 1 ou de 0, ce qui augmente le risque de biais de sélection.

Bien entendu, en l'absence du résultat Y dans l'échantillon probabiliste, il n'y a aucun moyen d'évaluer directement le biais de sélection. De ce fait, des travaux récents ont étendu le champ d'action de l'étude de Andridge et Little (2011) qui permet d'élaborer une analyse de sensibilité à l'aide d'un modèle de mélange de schémas d'observation, où la sélection dans l'échantillon non probabiliste a la possibilité de dépendre entièrement d'une réduction scalaire des covariables $\mathbf{X}$, entièrement du résultat Y, ou d'une combinaison convexe de ces éléments. Little, West, Boonstra et Hu (2020), Andridge, West, Little, Boonstra et Alvarado-Leiton (2019), et West, Little, Andridge, Boonstra, Ware, Pandit et Alvarado-Leiton (2021) considèrent la sensibilité en lien avec cette hypothèse dans l'estimation de la moyenne d'une variable normalement distribuée, la moyenne d'un résultat binaire et les paramètres de régression d'un modèle de régression linéaire, respectivement, dans les échantillons non probabilistes. En variant le paramètre de mélange convexe ϕ, il est possible d'évaluer la sensibilité à l'hypothèse de « sélection aléatoire ». Boonstra, Little, West, Andridge et Alvarado-Leiton (2021) constatent que ces « mesures normalisées du biais » soutiennent favorablement la comparaison avec d'autres solutions, comme $\hat{R}$ dans une étude de simulation. Il est important de mentionner que les méthodes qui complètent l'étude de Andridge et Little (2011) ne dépendent pas de l'hypothèse de covariables communes dans un échantillon probabiliste. Cela donne à penser que les méthodes qui utilisent les renseignements disponibles dans l'échantillon probabiliste pour évaluer l'hypothèse de la « sélection aléatoire » sont un domaine laissant place à l'évolution.

La seconde hypothèse, la positivité, est également peu susceptible d'être applicable avec précision dans de nombreux contextes pratiques. Mes propres travaux dans ce domaine ont porté sur des études de conduite en situation réelle tirées le plus souvent d'échantillons de commodité dans une région

géographique limitée. Par exemple, le *Second Strategic Highways Research Program* a recruté des conducteurs dans six régions géographiques précises des États-Unis (*Transportation Research Board* de la *National Academy of Sciences*, 2013). Cela correspond au deuxième scénario énoncé par Wu dans la section 7.2, où seule une sous-population a la possibilité d'être sélectionnée dans l'échantillon non probabiliste, ce qui, comme il le fait remarquer, n'a « pas de solution simple ». En suivant sa notation de D indiquant l'appartenance à une sous-population, il semblerait que si $D_i \perp \mathbf{X}_i, Y_i \mid \pi_i^A$ – c'est-à-dire, si la distribution de $\mathbf{X}, Y$ est la même pour $D = 0$ et $D = 1$ après pondération avec π_i^A à l'intérieur de la strate $D = 1$ –, alors l'absence de positivité n'aurait aucune incidence sur l'inférence. Il s'agit probablement d'un défi de taille dans les contextes les plus généraux, mais la positivité pourrait être plutôt bien estimée si l'analyse d'intérêt comprend un sous-ensemble de $\mathbf{X}, Y$ qui n'est que faiblement associé à D, même avant ajustement.

Enfin, en ce qui concerne la quatrième hypothèse, soit l'existence d'un échantillon probabiliste avec $\mathbf{X}$ disponible, je suis tout à fait d'accord avec l'observation de Wu voulant que les méthodes permettant de tirer parti d'enquêtes probabilistes multiples doivent être développées davantage. Cependant, il reste plus probable qu'un chercheur s'efforce de trouver un seul échantillon probabiliste avec suffisamment de covariables que de se débattre avec une surabondance d'options (ce que Wu appelle le « dilemme de la personne riche »). À cette fin, je terminerai par un appel à l'action lancé par les spécialistes des enquêtes.

5. L'échantillonnage probabiliste au 21e siècle : maintenant plus que jamais

J'ai appris les statistiques, plus particulièrement les statistiques issues d'enquêtes, vers la fin du 20e siècle, lorsque l'échantillonnage probabiliste était la pierre de touche incontestée du plan d'enquête. J'ai été initié pour la première fois au problème d'inférence à partir d'échantillons non probabilistes à la fin des années 2000, dans un contexte d'analyse des blessures faisant appel aux données du *Crash Injury Research Engineering Network* des États-Unis, où les analystes traitaient un échantillon très restreint de personnes accidentées dans des véhicules automobiles comme s'il s'agissait d'un échantillon aléatoire de victimes d'accidents de la route et qu'ils obtenaient par conséquent des résultats non réalistes (Elliott et coll., 2010). Vers la même époque, la popularité des enquêtes en ligne explosait et les statisticiens d'enquête ne savaient pas vraiment comment tirer des conclusions à partir de ces données. J'avoue avoir eu une attitude plutôt paternaliste à l'époque; j'ai presque évité de faire de la recherche dans ce domaine parce que je pensais que cela ne ferait qu'encourager un « mauvais comportement » à l'égard des plans d'échantillonnage. Je ne pensais pas pouvoir arrêter cette tendance à moi tout seul, mais je ne voulais pas participer à ce que j'ai perçu comme la dévalorisation de la science. J'en suis néanmoins venu à reconnaître que bon nombre de ces nouvelles sources de données présentent des avantages qui vont au-delà de ce que l'on peut obtenir au moyen d'un échantillon probabiliste traditionnel, surtout avec des budgets limités. Cela va bien au-delà des défis croissants que représente la mise en œuvre d'enquêtes

probabilistes, notamment auprès de la population générale, en raison des non-réponses, de l'absence de bases de sondage adéquates, etc.

Pour autant, je reste inquiet à l'idée qu'en ayant conçu des méthodes pour faire face aux limites des enquêtes non probabilistes, l'échantillonnage probabiliste est devenu chose du passé chez les scientifiques et les décideurs dont la formation statistique est limitée, et ce, malgré des efforts comme ceux déployés par Bradley, Kuriwaki, Isakov, Sejdinovic, Meng et Flaxman (2021), et Marek, Tervo-Clemmens, Calabro et coll. (2022). Cependant, comme Wu le souligne dans son analyse, l'absence d'échantillons probabilistes rend l'échantillon non probabiliste moins vulnérable à la possibilité d'un calage même partiel ou d'autres méthodes d'ajustement (bien que les analyses de sensibilité comme les approches de « mesures normalisées du biais » mentionnées ci-dessus ne nécessitent pas d'échantillons probabilistes de référence). Par conséquent, je crois qu'il est de plus en plus essentiel de mettre en place des enquêtes probabilistes structurées et idéalement financées par le gouvernement pour les collectes de données courantes. Ce type d'enquêtes existe déjà – l'*American Community Survey* du *United States Census Bureau* et la *National Health Interview Survey* du *National Center for Health Statistics* en font notamment partie –, mais à l'avenir, je crois qu'il serait utile pour les organismes statistiques de s'entendre sur la nécessité de mettre les enquêtes probabilistes de grande qualité à la disposition des enquêtes non probabilistes à des fins d'analyse, et non de les diffuser comme de simples publications autonomes. Cela signifie qu'il faut réfléchir attentivement aux covariables importantes dans diverses fonctions de santé publique et de sciences sociales dans lesquels les données d'enquête jouent un rôle. Il faudra faire des choix compte tenu des contraintes budgétaires limitées et, en même temps, prévoir un financement suffisant pour conserver la qualité nécessaire aux ajustements. Enfin, bien que certaines méthodes ne nécessitent pas de microdonnées et peuvent se servir de mesures agrégées, comme celles qu'on retrouve dans l'*American Communities Survey*, d'autres nécessiteront ce type de données, ce qui signifie que de nouveaux domaines de recherche seront à explorer au chapitre des renseignements personnels et de la confidentialité en ce qui concerne la combinaison de données tirées d'enquêtes probabilistes et non probabilistes.

Bibliographie

Andridge, R.R., et Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 153-180.

Andridge, R.R., West, B.T., Little, R.J., Boonstra, P.S. et Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society*, C68, 1465-1483.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. et Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of Official Statistics*, 37, 751-769.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.L. et Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600, 695-700.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chen, J.K.T., Valliant, R.L. et Elliott, M.R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society*, 68, 657-681.

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.

Chipman, H.A., George, E.I. et McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266-298.

Dong, Q., Elliott, M.R. et Raghunathan, T.E. (2014). Une méthode non paramétrique de production de populations synthétiques qui tient compte des caractéristiques des plans de sondage complexes. *Techniques d'enquête*, 40, 1, 33-52. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014001/article/14003-fra.pdf.

Downes, M., et Carlin, J.B. (2020). Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: A simulation study. *Biometrical Journal*, 62, 479-491.

Eilers, P.H., et Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-121.

Elliott, M.R. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M.R., et Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from the Behavioral Risk Factor Surveillance Survey and the National Health Interview Survey. *Journal of the Royal Statistical Society*, C54, 595-609.

Elliott, M.R., et Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

Elliott, M.R., Resler, A., Flannagan, C.A. et Rupp, J.D. (2010). Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention*, 42, 530-539.

Hájek, J. (1971). Comment on a paper by D. Basu. *Foundations of Statistical Inference*, 236.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Kalton, G. (1983). Models in the practice of survey sampling. *Revue Internationale de Statistique*, 51, 175-188.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54 139-157.

Little, R.J.A., et Zheng, H. (2007). The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics*, 8, 1-20.

Little, R.J.A., West, B.T., Boonstra, P.S. et Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 932-964.

Marek, S., Tervo-Clemmens, B., Calabro, F.J. et coll. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, sous presse.

McConville, K.S., Breidt, F.J., Lee, T. et Moisen, G.G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5, 131-158.

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, 558-625.

Rafei, A. (2021). Robust and efficient Bayesian inference for large-scale non-probability samples. Thèse de l'University of Michigan. Accessible sur https://www.overleaf.com/project/6228db145a47be05f8da3777.

Rafei, A, Flannagan, C.A.C. et Elliott, M.R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8, 148-180.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meetings*. Disponible sur https://static.texastribune.org/media/documents/Rivers_matching4.pdf.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 101-113. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10887-fra.pdf.

Soentpiet, R. (1999). *Advances in Kernel Methods: Support Vector Learning*. Boston: MIT Press.

Transportation Research Board of the National Academy of Sciences (2013). *The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset.*

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

Van Der Laan, M.J., et Rubin, D.R. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Wang, W., Rothschild, D., Goel, S. et Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31, 980-991.

West, B.T., Little, R.J.A., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. et Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The Annals of Applied Statistics*, 15, 1556-1581.

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.

Zhang, G., et Little, R.J.A. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 911-918.

Zheng, H., et Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Zhou, T., Elliott, M.R., et Little, R.J.A. (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114, 1-19.