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 » : La miniaturisation de la corrélation due à un défaut des données : une stratégie polyvalente de traitement des échantillons non probabilistes

par Xiao-Li Meng

Date de diffusion : le 15 décembre 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2022
Vol. 48, n° 2, pp. 369-392
Statistique Canada, n° 12-001-X au catalogue

Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » : La miniaturisation de la corrélation due à un défaut des données : une stratégie polyvalente de traitement des échantillons non probabilistes

Xiao-Li Meng[1]

Résumé

Il n'est pas possible de tirer parti de la puissante *probabilité du plan* pour établir l'inférence fondée sur la randomisation à partir d'échantillons non probabilistes. Cela nous incite à exploiter une *probabilité divine* naturelle qui accompagne toute population finie. Dans cette perspective, un des paramètres principaux est la *corrélation due à un défaut des données (cdd)*, qui est la corrélation de la population finie sans modèle entre l'indicateur d'inclusion de l'échantillon de la personne et la caractéristique de la personne échantillonnée. Un mécanisme de génération de données équivaut à un échantillonnage probabiliste, en ce qui concerne l'effet de plan, si et seulement si la *cdd* correspondante est de l'ordre (stochastique) $N^{-1/2}$, où N est la taille de la population (Meng, 2018). Par conséquent, les méthodes d'estimation linéaire valides existantes pour les échantillons non probabilistes peuvent être converties en plusieurs stratégies de miniaturisation de la *cdd* jusqu'à l'ordre $N^{-1/2}$. Les méthodes quasi fondées sur le plan permettent d'accomplir cette tâche en réduisant la variabilité entre les N propensions d'inclusion au moyen d'une pondération. L'approche fondée sur un modèle de superpopulation permet d'atteindre le même objectif par la réduction de la variabilité des caractéristiques des N personnes en les remplaçant par leurs résidus issus d'un modèle de régression. Les estimateurs doublement robustes doivent la propriété dont ils portent le nom au fait qu'une corrélation est nulle chaque fois qu'une des variables corrélées est constante, quelle qu'elle soit. Comprendre les points communs de ces méthodes au moyen de la *cdd* nous aide à voir clairement la possibilité d'une « robustesse plus que double », c'est-à-dire une estimation valide qui ne dépend pas de la pleine validité du modèle de régression ni de la propension d'inclusion estimée, qui ne sont garanties ni l'une ni l'autre parce que les deux reposent sur la *probabilité du procédé*. Les renseignements générés par la *cdd* incitent également à un *sous-échantillonnage de contrebalancement*, une stratégie visant à créer une miniature de la population à partir d'un échantillon non probabiliste, et comportant un compromis de qualité et de quantité favorable parce que les erreurs quadratiques moyennes sont beaucoup plus sensibles à la *cdd* qu'à la taille de l'échantillon, en particulier pour les populations de grande taille.

Mots-clés : Biais de non-réponse; estimateurs d'enquête assistés par un modèle; indice de défaut des données; inférence fondée sur le plan de sondage; probabilité divine; probabilité du plan; probabilité du procédé.

1. Distinguer la probabilité du plan, la probabilité divine et la probabilité du procédé

1.1 Que peuvent indiquer les statistiques et les statisticiens au sujet des échantillons non probabilistes ?

Le traitement des échantillons non probabilistes est une affaire délicate, particulièrement pour les statisticiens. Ceux qui estiment que les statistiques reposent sur le raisonnement probabiliste et l'inférence peuvent se demander si les statistiques ont quelque chose d'utile à offrir au monde non probabiliste. Bien

1. Xiao-Li Meng, Département de statistique, Université Harvard, Cambridge, MA 02138. Courriel : meng@stat.harvard.edu.

que ce questionnement puisse refléter une certaine ignorance, voire de l'hostilité à l'égard des statistiques, du point de vue conceptuel, il mérite une introspection et une extrospection de la part des statisticiens. À quel genre de probabilité faisons-nous référence quand l'échantillon est non probabiliste ? Toute la théorie et toutes les méthodes d'échantillonnage probabiliste reposent sur le caractère aléatoire introduit par de puissants mécanismes d'échantillonnage, qui permettent d'obtenir ensuite le magnifique cadre inférentiel fondé sur le plan sans qu'il soit nécessaire de *concevoir* que quoi que ce soit d'autre soit aléatoire (Kish, 1965; Wu et Thompson, 2020; Lohr, 2021). Quand cette puissance et cette beauté nous sont enlevées, que reste-t-il aux statisticiens ?

La réponse philosophique de certains statisticiens serait de rejeter complètement la question en déclarant qu'il n'existe pas d'échantillon probabiliste dans le monde réel. (Andrew Gelman m'a rappelé cette opinion quand je lui ai demandé de commenter la présente étude; voir une analyse sur le sujet dans https://statmodeling.stat.columbia.edu/2014/08/06.) Quand les données arrivent à notre bureau ou dans notre disque dur, même le plan d'échantillonnage probabiliste le mieux conçu est compromis par les imperfections de l'exécution, qui vont des défauts (incontrôlables) dans les bases de sondage aux non-réponses à différentes étapes, en passant par les erreurs de mesure dans les réponses. En ce sens, la notion d'échantillon probabiliste est toujours théorique, de la même manière que l'hypothèse de l'efficience du marché en économie, qui offre un cadre mathématiquement élégant pour l'idéalisation et les approximations, mais qui ne doit jamais être prise au pied de la lettre (par exemple Lo, 2017).

L'article d'actualité du professeur Changbao Wu (Wu, 2022) fournit une réponse plus pratique, en présentant la façon dont les statisticiens ont traité les échantillons non probabilistes dans la longue littérature sur les enquêtes par échantillons et (bien entendu) les études d'observation, en particulier concernant l'inférence causale; voir Elliott et Valliant (2017) et Zhang (2019) pour obtenir deux aperçus complémentaires portant sur le même défi. Pour mieux comprendre l'utilité de la théorie des probabilités en présence d'échantillons non probabilistes, il est important de reconnaître (au moins) trois types de concepts probabilistes aux fins d'inférence statistique, qui sont présentés dans la section 1.2. Les échantillons non probabilistes permettent de n'en retirer qu'un seul des trois, ainsi ils contraignent en général les spécialistes à une plus grande dépendance aux deux autres.

Une fois ces questions conceptuelles clarifiées, les autres sections traitent d'une stratégie unifiée de traitement des échantillons non probabilistes. La section 2 porte sur une identité fondamentale pour l'erreur d'estimation, qui a mené à l'élaboration de la corrélation due à un défaut des données (Meng, 2018). La section 3 traite ensuite de la façon dont ce concept peut inspirer une stratégie unifiée. La section 4 présente la stratégie pour les configurations qp et ξp, respectivement, dans l'étude de Wu (2022). Dans la section 5, on applique ensuite la stratégie aux deux configurations simultanément pour donner un aperçu immédiat de la célèbre double robustesse, laquelle est examinée dans l'étude de Wu (2022). À partir de ce même concept, la section 6 repose sur *l'échantillonnage de contrebalancement* comme stratégie de remplacement de la pondération. La section 7 se termine par un appel général à traiter la théorie de l'échantillonnage probabiliste comme une aspiration plutôt que comme l'élément central de la recherche sur les enquêtes et l'échantillonnage.

1.2 Un trio de concepts de probabilité

Le premier des trois concepts nommés ci-dessous, la probabilité du plan, se passe d'explications. Il est au cœur de la théorie de l'échantillonnage et il a été réifié par la mise en œuvre pratique, aussi imparfaite qu'elle puisse être. Bien que la distinction entre les deux concepts suivants, la probabilité divine et la probabilité du procédé, puisse être plus nuancée, en particulier sur le plan pratique, leurs différences conceptuelles ne sont pas moins importantes que la distinction entre un paramètre et un estimateur. Comme il se doit, l'indicateur d'enregistrement ou d'inclusion des données, lequel représente une quantité importante dans la modélisation des échantillons non probabilistes, fournit une illustration concrète des trois concepts de probabilité; voir le paragraphe principal de la section 4.

Probabilité du plan. Les répliques randomisées (Craiu, Gong et Meng, 2022) sont un concept et un outil primordiaux pour les statistiques et les sciences en général. En concevant et en exécutant un mécanisme probabiliste pour générer des répliques randomisées, nous créons des données probabilistes qui peuvent directement servir à faire des énoncés inférentiels vérifiables. De plus, l'échantillonnage probabiliste dans les enquêtes, la randomisation dans les essais cliniques, les bootstraps aux fins d'évaluation de la variabilité, les tests de permutation pour les vérifications d'hypothèses et les simulations Monte Carlo pour le calcul sont tous des exemples de méthodes statistiques élaborées à partir de la probabilité du plan. Les échantillons non probabilistes, par définition, n'ont pas de probabilité de plan, du moins pas de probabilité de plan définie. Par conséquent, le terme « échantillons non probabilistes » devrait être considéré comme une formulation abrégée d'« échantillons sans concept de probabilité du plan défini ».

Il faut toutefois nous rappeler que les probabilités du plan peuvent reprendre une place importante, surtout pour les grands ensembles de données non probabilistes, comme les données administratives, en raison de l'adoption de la confidentialité différentielle (Dwork, 2008), par exemple par le Bureau du recensement des États-Unis (voir l'éditorial de Gong, Groshen et Vadhan, 2022, et le numéro spécial du *Harvard Data Science Review* qu'il présente). Les méthodes de confidentialité différentielle permettent d'introduire du bruit aléatoire bien conçu dans les données dans le but de protéger la confidentialité des données sans sacrifier indûment l'utilité des données. Comme la probabilité du plan utilisée pour l'échantillonnage probabiliste, le fait que le mécanisme d'introduction de bruit soit conçu par le curateur des données et qu'il soit rendu public permet de fournir la transparence essentielle pour que l'utilisateur des données produise une inférence statistique valide (Gong, 2022). La question de savoir comment analyser correctement les données non probabilistes en cas de protection différentielle de la vie privée est largement ouverte. Il est encore plus fascinant de connaître la façon de tenir compte des défauts existants dans les données non probabilistes lors de la conception de mécanismes de protection probabiliste pour la confidentialité des données, afin d'éviter d'ajouter du bruit inutile. Les lecteurs qui souhaitent avoir une vue d'ensemble des problèmes statistiques découlant de la confidentialité des données devraient consulter l'excellent article de synthèse de Slavkovic et Seeman (2022) sur le domaine de la confidentialité des données en général.

Probabilité divine. En l'absence de probabilité du plan pour procéder à l'inférence fondée sur la randomisation, nous nous appuyons habituellement sur le concept selon lequel les données à notre disposition sont la réalisation d'un mécanisme probabiliste génératif donné par la nature ou par Dieu, afin d'effectuer une inférence statistique (classique). (J'ai entendu le terme « modèle de Dieu » pendant ma formation de doctorat et je l'ai considéré comme une expression de la foi ou comme quelque chose qui échappe au contrôle de l'homme, plutôt que comme le reflet des croyances religieuses d'une personne. L'adjectif « divine » est adopté dans la présente étude avec une connotation similaire.) Nous le faisons que nous croyions ou non que le monde est intrinsèquement déterministe ou stochastique (par exemple voir David Peat, 2002; Li et Meng, 2021). Nous devons supposer cette probabilité divine principalement en raison de la nature restrictive du cadre probabiliste auquel nous sommes si habitués. Par exemple, pour invoquer l'hypothèse de répartition au hasard des données manquantes, nous devons évoquer un mécanisme probabiliste dans lequel le concept de « données manquantes au hasard » (Rubin, 1976) peut être mis en forme. Comme l'ont souligné Elliott et Valliant (2017), la méthode de quasi-randomisation, qui correspond au cadre *qp* de Wu (2022), « suppose que l'échantillon non probabiliste ait en fait un mécanisme d'échantillonnage probabiliste, bien que ce soit un mécanisme comportant des probabilités qui doivent être estimées selon des hypothèses d'identification ». Cela signifie que nous remplaçons la probabilité du plan par une probabilité divine à laquelle nous croyons et qui est alors généralement traitée comme la « vérité » ou du moins comme un paramètre.

Sur le plan conceptuel, nous devons donc reconnaître que l'hypothèse d'une sorte particulière de probabilité divine n'est pas involontaire, car sinon nous n'aurions pas besoin de dépendre de notre foi pour continuer à travailler. Ce n'est pas non plus toujours nécessaire. Toute population finie fournit un histogramme naturel pour toute caractéristique quantifiable, ou un tableau de contingence pour toute caractéristique classable de ses éléments et, par conséquent, elle induit une probabilité divine sans faire référence à un quelconque caractère aléatoire, conceptualisé ou réalisé, *si notre cible inférentielle est la population finie elle-même* (et non une superpopulation qui permet de la générer, par exemple). La méthode de la vraisemblance empirique tire parti de ce cadre de probabilité naturel, qui se révèle également fondamental pour quantifier la qualité des données au moyen de la corrélation due à un défaut des données (voir Meng, 2018). Zhang (2019), pour qui le critère unifié était fondé sur la même identité pour établir la corrélation due à un défaut des données, fait la même constatation; voir la section 2 ci-dessous.

Probabilité du procédé. La plupart des probabilités utilisées dans la modélisation statistique sont, de loin, des procédés servant à exprimer notre croyance, nos connaissances *a priori*, nos hypothèses, nos idéalisations, nos compromis, voire notre désespoir (comme quand on impose une distribution *a priori* pour assurer l'identifiabilité, puisque rien d'autre ne fonctionne). Bien que la littérature statistique ait toujours mis l'accent sur la modélisation de la réalité, nous sommes inévitablement contraints de réaliser une variété de simplifications, d'approximations et, parfois, de distorsions délibérées afin de composer avec des contraintes pratiques (par exemple l'utilisation de l'inférence variationnelle à des fins d'efficacité du calcul; voir Blei, Kucukelbir et McAuliffe, 2017). Par conséquent, pour bon nombre de ces probabilités

du procédé, rien ne contraint à ce qu'elles soient réalisables ni même mathématiquement cohérentes (par exemple l'emploi de distributions de probabilité conditionnelle incompatibles pour l'imputation à chaînes multiples; voir Van Buuren et Oudshoorn, 1999). Il n'est pas non plus facile de valider ces probabilités, ni même possible, comme Zhang (2019) l'a étudié et fait valoir dans un contexte d'échantillonnage non probabiliste, en particulier avec la méthode de modélisation de superpopulation, qui correspond au cadre ξp de Wu (2022). Néanmoins, les probabilités de procédé offrent le meilleur rendement pour procéder à des inférences statistiques. La méthode de quasi-randomisation et la modélisation de superpopulation reposent toutes deux sur des probabilités du procédé pour fonctionner, comme le montre Wu (2022) et comme il est expliqué en détail dans les sections 4 et 5 de la présente étude. L'absence de probabilité du plan ne peut que favoriser la progression des probabilités de procédé. Pour paraphraser une citation célèbre de Box, « tous les modèles sont faux, mais certains sont utiles ». Cela signifie que toutes les probabilités de procédé posent des problèmes, mais que certaines permettent de résoudre des problèmes.

1.3 La réduction des données inexactes menant à des paquets d'intelligence artificielle

Pour résumer, les concepts probabilistes sont plus nécessaires pour les échantillons non probabilistes que pour les échantillons probabilistes, précisément parce qu'ils n'ont pas de probabilité de plan. La difficulté de traitement des échantillons non probabilistes ne représente pas un nouveau défi pour les statisticiens. S'il y a quelque chose de nouveau, il s'agit de la quantité considérable de grands ensembles de données non probabilistes, comme les données administratives et les données provenant des médias sociaux, et du besoin accéléré de combiner plusieurs sources de données, qui sont pour la plupart intrinsèquement non probabilistes parce qu'elles ne sont pas recueillies à des fins d'inférence statistique (par exemple Lohr et Rao, 2006; Meng, 2014; Buelens, Burger et van den Brakel, 2018; Beaumont et Rao, 2021). Contrairement à la croyance courante, la grande taille des « mégadonnées » peut empirer notre inférence, en raison du « paradoxe des mégadonnées » (Meng, 2018; Msaouel, 2022), quand nous ne tenons pas compte de la qualité des données lors de l'évaluation des erreurs et des incertitudes dans nos analyses; voir la section 6.1.

Il est donc plus urgent que jamais de sensibiliser et d'informer à grande échelle à propos de l'importance cruciale de la qualité des données et de la façon dont nous pouvons utiliser les méthodes et les théories statistiques pour contribuer à réduire le défaut des données. Dans le cadre de la présente étude, la préoccupation centrale va au-delà de la mise en garde habituelle au sujet du dicton « à données inexactes, résultats erronés ». En effet, si des données sont reconnues comme étant mauvaises, elles seront probablement traitées comme telles (probablement, mais pas toujours, car, comme Andrew Gelman me l'a rappelé, « de nombreux chercheurs croient fermement à la *procédure* plutôt qu'à la *mesure*, et pour ces personnes, le plus important consiste à respecter les règles, et non pas à s'interroger sur la provenance de leurs données »). L'objectif est d'empêcher que des données inexactes mènent à un paquet d'intelligence artificielle (Meng, 2021), c'est-à-dire que des données de faible qualité sont traitées automatiquement au

moyen de procédures génériques afin de créer un paquet d'intelligence artificielle esthétiquement attrayant qui serait vendu à des consommateurs non informés, ou pire encore, à ceux qui cherchent des « données probantes » pour induire en erreur ou désinformer. De toute évidence, le traitement adéquat des échantillons non probabilistes ne résout pas tous les problèmes de qualité des données, mais il contribue grandement à régler le problème de plus en plus courant et nuisible du manque de contrôle de la qualité des données en science des données.

Je remercie donc le professeur Changbao Wu pour son analyse opportune et complète des « incontournables » de la grande usine de fabrication de saucisses qu'est le traitement d'échantillons non probabilistes. Elle apporte des vues beaucoup plus détaillées et nuancées que celles de l'analyse générale d'Elliott et Valliant (2017), qui illustrait de manière très fine de nombreuses formes d'échantillons non probabilistes ainsi que leurs méfaits. Elle présente également des jalons théoriques et méthodologiques qui nous permettront de mieux apprécier ceux exposés dans l'analyse intellectuelle de Zhang (2019), qui met les statisticiens et les scientifiques des données en général au défi de mieux comprendre la qualité, ou plutôt l'absence de qualité, des produits que nous fabriquons et promouvons. Ensemble, ces trois articles de synthèse donnent un aperçu informé à quiconque souhaiterait se joindre aux spécialistes cherchant à traiter les difficultés de plus en plus grandes causées par les données non probabilistes. Le mieux serait peut-être de commencer par l'étude d'Elliott et de Valliant (2017), qui brosse un portrait général de la situation, puis de lire l'étude de Wu (2022), qui s'attache principalement à exposer les méthodologies, et de terminer par l'étude de Zhang (2019), qui suscite des réflexions plus profondes sur certains défis particuliers. Pour en savoir plus sur d'autres méthodes courantes de traitement des échantillons non probabilistes, comme la modélisation multiniveau et la poststratification, nous invitons les lecteurs à lire Gelman (2007), Wang, Rothschild, Goel et Gelman (2015) et Liu, Gelman et Chen (2021).

En tant que chercheur et enseignant, je me suis penché sur ce domaine, mais j'ai souvent été frustré par le manque de temps ou d'énergie pour l'étudier en profondeur. Je suis donc particulièrement reconnaissant envers le rédacteur en chef, Jean-François Beaumont, de m'avoir invité à contribuer à ce que les messages du professeur Wu soient clairement entendus : les données ne peuvent pas être traitées comme si elles étaient représentatives à moins que les données observées ne soient véritablement des échantillons probabilistes (ce qui est extrêmement rare). De nombreux remèdes ont été proposés et essayés, mais il faut en élaborer et en évaluer bien davantage. Parmi ces remèdes, le concept de corrélation due à un défaut des données est une mesure générale prometteuse qui doit être étudiée et élaborée, comme nous le montrons ci-dessous.

2. Une identité déterministe de population finie pour l'erreur réelle

Pour démontrer la richesse du cadre de population finie, considérons l'estimation de la moyenne de la population, indiquée par $\bar{G}$, de $\{G_i = G(X_i) : i \in \mathcal{N}\}$, où $\mathcal{N} = \{1, \ldots, N\}$ indexe une population finie, et les X_i sont des données recueillies sur une personne i. Pour chaque i, supposons que $R_i = 1$ si G_i (ou plutôt X_i) est enregistré dans notre échantillon, et que $R_i = 0$ sinon. La taille de l'échantillon est alors

$n_R = \sum_{i=1}^{N} R_i$. Nous insistons sur le fait qu'il s'agit d'un indicateur global, qui peut (et devrait) être décomposé en $R_i = r_i^{(1)}, \ldots, r_i^{(J)}$, quand la collecte des données consiste en J étapes (par exemple $r_i^{(1)}$ indique si la i^{e} personne a été échantillonnée et $r_i^{(2)}$, si la personne a répondu ou non une fois qu'elle a été échantillonnée).

Supposons que $\{W_i, i \in S\}$ est un ensemble de poids à déterminer où l'indice est paramétré à $S = \{i : R_i = 1\}$, de sorte que $\sum_{i \in S} W_i > 0$. Supposons que $\bar{G}_W$ est la moyenne pondérée de l'échantillon, qu'on peut exprimer de trois façons :

$$\bar{G}_W = \frac{\sum_{i \in S} W_i G_i}{\sum_{i \in S} W_i} = \frac{\sum_{i=1}^{N} R_i W_i G_i}{\sum_{i=1}^{N} R_i W_i} = \frac{\mathrm{E}_I(\tilde{R}_I G_I)}{\mathrm{E}_I(\tilde{R}_I)}, \tag{2.1}$$

où $\tilde{R}_I = R_I W_I$, et E_I est prise par rapport à la distribution uniforme de l'indice paramétré $\mathcal{N}$. La première expression dans l'équation (2.1) définit simplement une moyenne pondérée de l'échantillon. À l'aide de R_i, la deuxième expression permet de transformer les moyennes de l'échantillon en moyennes de population finie. Cette nouvelle expression banale est fondamentale parce qu'elle explique le rôle de R_i dans l'influence sur le comportement de $\bar{G}_W$ en tant qu'estimateur de $\bar{G}$. La troisième expression révèle une probabilité divine au moyen de I, la variable de l'indice de population finie (IPF), grâce au fait que le calcul de la moyenne revient à prendre en compte l'espérance d'un indice aléatoire uniformément distribué I. Tous les moments de population finie peuvent alors être exprimés au moyen de E_I.

En particulier, nous pouvons exprimer l'erreur réelle de $\bar{G}_W$ par l'identité suivante, dont la première expression remonte à Hartley et Ross (1954), qui l'ont utilisée pour exprimer les biais dans des estimateurs par le ratio. La deuxième expression a été donnée dans Meng (2018), mais elle comportait une expression légèrement différente (mais équivalente) :

$$\bar{G}_W - \bar{G} = \frac{\mathrm{Cov}_I(\tilde{R}_I, G_I)}{\mathrm{E}_I[\tilde{R}_I]} = \rho_{\tilde{R},G} \times \sqrt{\frac{N - n_W}{n_W}} \times \sigma_G. \tag{2.2}$$

Dans cette équation, $\rho_{\tilde{R},G} = \mathrm{Corr}_I(\tilde{R}_I, G_I)$ est la *corrélation de population finie* entre $\tilde{R}_I$ et G_I, σ_G^2 est la variance de la population finie de G_I et n_W est la taille d'échantillon efficace en raison de l'utilisation des poids (Kish, 1965),

$$n_W = \frac{n_R}{1 + \mathrm{CV}_W^2}, \tag{2.3}$$

où CV_W est le coefficient de variation (c'est-à-dire l'écart-type ou la moyenne) de $\{W_i, i \in S\}$.

L'expression de l'équation (2.2) est une identité algébrique parce qu'elle se vérifie pour toute instance de $\{(G_i, R_i W_i), i \in \mathcal{N}\}$. Ainsi, aucune hypothèse de modèle n'est imposée, pas même l'hypothèse que R (ou toute quantité) est aléatoire, ce qui rappelle le commentaire de Mary Thompson cité dans l'étude de Wu (2022), selon lequel « le fait que l'indicateur d'inclusion dans l'échantillon R est une variable aléatoire est en soi une hypothèse ». La seule exigence est que la valeur de G_i enregistrée soit identique à celle de G_i dans la population cible. (Il faut mentionner toutefois que cette exigence comporte deux

éléments : 1) il n'y a pas de surdénombrement, c'est-à-dire que chaque personne dans l'échantillon appartient à la population cible, par exemple aucun électeur non admissible n'est sondé quand la population cible est celle des électeurs admissibles; 2) il n'y a pas d'erreur de mesure. Il peut y avoir des extensions de cas comportant des erreurs de mesure, mais elles ne sont pas examinées dans la présente étude.) Quand nous utilisons des poids égaux, les trois facteurs du membre de droite de l'équation (2.2) représentent, respectivement (de gauche à droite), le défaut des données, l'insuffisance des données et la difficulté du problème, comme l'explique Meng (2018) et comme l'illustrent en détail Bradley, Kuriwaki, Isakov, Sejdinovic, Meng et Flaxman (2021) dans le contexte des enquêtes sur la vaccination contre la COVID-19.

En particulier, quand tous les poids sont égaux, $\rho_{\tilde{R},G}$ est appelée *corrélation due à un défaut des données* (*cdd*) dans Meng (2018) parce qu'elle permet de mesurer le manque de représentativité de l'échantillon en saisissant la dépendance de l'indicateur d'inclusion ou d'enregistrement aux caractéristiques : plus la dépendance est élevée, plus la moyenne de l'échantillon est biaisée quand il faut estimer les moyennes de population. Quand l'on utilise les stratégies de base de l'échantillonnage probabiliste ou de la pondération de probabilité inverse, la *cdd* est nulle en moyenne parce que $\mathrm{E}(W_i R_i) = 1,$ et elle est de l'ordre $O_p(N^{-1/2})$ parce qu'il s'agit essentiellement d'une moyenne de N termes indépendants (Meng, 2018). Notre objectif général est donc de ramener la *cdd* à $O_p(N^{-1/2})$ pour les échantillons non probabilistes, ce que nous appellerons « miniaturiser la *cdd* » parce que $N^{-1/2}$ est généralement un nombre minuscule dans la pratique.

Quand nous utilisons des poids, le premier terme $\rho_{\tilde{R},G}$ saisit le défaut des données qui existe toujours après l'ajustement de la pondération, puisqu'aucun poids n'est parfait en pratique. L'identité dans l'équation (2.2) montre l'incidence des poids sur la qualité et la quantité des données. L'incidence sur la taille d'échantillon efficace *nominale* n_W n'est jamais positive, car $n_W \le n_R$ comme on peut le voir dans l'équation (2.3). Par ailleurs, l'exactitude de l'équation (2.3) révèle qu'en fait, cette expression bien connue n'est pas une approximation (ce qui est souvent attribué à Kish, 1965), mais une formule exacte de réduction de la taille de l'échantillon en raison de la pondération *si la pondération n'a pas d'incidence sur la cdd*. Cependant, la pondération peut avoir une incidence positive importante sur la réduction de l'erreur globale quand on choisit judicieusement des poids pour diminuer considérablement la *cdd*, bien qu'apparemment cela se fasse au prix de $n_W < n_R$. Bien entendu, c'est exactement ce que vise le cadre de quasi-randomisation dont il est question ci-dessous. Plus important encore, l'équation (2.2) donne un aperçu unifié de la variété des méthodes examinées dans l'étude de Wu (2022), notamment une explication intuitive de la propriété doublement robuste, qui fait l'objet d'une attention accrue aux fins d'intégration des sources de données, concernant à la fois des échantillons probabilistes et non probabilistes (par exemple Yang, Kim et Song, 2020).

En effet, Zhang (2019, section 3.1) a utilisé la première expression dans l'équation (2.2) pour définir une hypothèse de non-informativité asymptotique non paramétrique unifiée, qui exige que le numérateur $\mathrm{Cov}_I(\tilde{R}_I, G_I)$ passe à zéro, tout en gardant le dénominateur $\mathrm{E}_I[\tilde{R}_I]$ positif, quand $N \to \infty$. Cette

unification a permis à Zhang (2019) d'évaluer la méthode de quasi-randomisation et la modélisation par la régression au moyen d'un critère commun. Comme le montre la section 3, le cadre de la *cdd* fait écho à cette unification. La section 4 met plutôt l'accent sur le message général de Zhang (2019). La section 5 traite d'un autre avantage simple de la formulation de la *cdd* qui fournit une explication immédiate de la célèbre double robustesse. La section 6 aborde quant à elle le domaine beaucoup plus difficile de l'élaboration d'un sous-échantillon plus représentatif à partir d'un grand échantillon non représentatif, soit un compromis précieux, puisque la qualité des données est beaucoup plus importante que la quantité (Meng, 2018), comme nous le voyons brièvement ci-dessous.

3. Une stratégie unificatrice fondée sur la corrélation due à un défaut des données

Dans la configuration de Wu (2022), pour chaque personne i, nous avons un ensemble de caractéristiques $A_i = \{y_i, \mathbf{x}_i\}$, où y est la caractéristique d'intérêt et $\mathbf{x}$ est une variable auxiliaire, ce qui est utile de deux façons. Premièrement, la réduction du biais d'échantillonnage attribuable à l'échantillonnage non probabiliste devient possible quand le mécanisme non probabiliste peut être (entièrement) expliqué par $\mathbf{x}$. Deuxièmement, en tirant parti des relations entre y_i et $\mathbf{x}_i$, nous pouvons améliorer l'efficacité de notre estimation. Comme point de départ, Wu (2022) suppose que nous avons deux sources de données disponibles, que nous désignons au moyen de deux indicateurs d'enregistrement, R et R^*. La source principale des données est un échantillon non probabiliste dans lequel nous observons à la fois y_i et $\mathbf{x}_i$ lorsque $i \in S \equiv \{i : R_i = 1\}$, mais l'indicateur d'enregistrement R_i est déterminé par un mécanisme non contrôlé par une probabilité de plan (connue). La deuxième source est (supposée être) un échantillon probabiliste dans lequel nous observons seulement $\mathbf{x}_i$ lorsque $i \in S^* \equiv \{i : R_i^* = 1\}$. Ce deuxième échantillon fournit des renseignements pour estimer des renseignements auxiliaires sur la population, qui sont utiles dans l'estimation des quantités de la population de y, par exemple sa moyenne. Par conséquent, cette configuration est étroitement liée à la configuration où $S \cup S^* = \mathcal{N}$; voir Tan (2013).

Pour toute fonction $m(\mathbf{x})$, considérons que $z_i = y - m(\mathbf{x}_i), i \in \mathcal{N}$. Il est clair que nous pouvons estimer la moyenne de la population $\bar{y}_N = \mathrm{E}_I(y_I)$ en estimant $\bar{z} = \mathrm{E}_I(z_I)$ et $\bar{m} = \mathrm{E}_I\left[m(\mathbf{x}_I)\right]$. À partir du deuxième échantillon, $\bar{m}$ peut être estimé sans biais puisqu'il ne concerne que $\mathbf{x}$. Nous pouvons alors nous concentrer sur l'estimation de $\bar{z}$, tout en reconnaissant qu'une méthode plus fondée sur des principes nous amènerait à établir un modèle de probabilité ou un modèle bayésien pour estimer conjointement toutes les quantités inconnues (Pfeffermann, 2017). L'application de l'identité dans l'équation (2.2) où $G = z$ nous indique alors que notre tâche centrale consiste à choisir le poids $\{W_i, i \in S\}$ ou la fonction m pour miniaturiser la *cdd*, $\rho_{\tilde{R},z}$. Dans la présente étude, il est plus facile de tout expliquer au moyen de la covariance

$$c_{\tilde{R},z} \equiv \mathrm{Cov}_I(\tilde{R}_I, z_I) = \mathrm{Cov}_I(W_I R_I, y_I - m(\mathbf{x}_I)) = \frac{1}{N}\sum_{i=1}^{N} W_i R_i (z_i - \bar{z}) \tag{3.1}$$

au lieu de la corrélation $\rho_{\tilde{R},z}$ parce que $\mathrm{Cov}_I(\tilde{R}_I, z_I)$ est une fonction bilinéaire dans R_I et z_I. Toutefois, sur le plan théorique et à des fins de modélisation, $\rho_{\tilde{R},z}$ est plus attrayante en raison de sa normalisation; voir les sections 6 et 7.

L'expression dans l'équation (3.1) nous indique immédiatement la façon de la rendre nulle dans les espérances sur le plan opérationnel, et dans quel sens conceptuel. Quelle que soit la probabilité que nous imposions à R_i (à préciser dans les dernières sections), supposons que $\pi_i = \Pr(R_i = 1 \mid \mathbf{A})$, que nous assumons dépendra de A_i seulement. Alors, la linéarité de l'opérateur de covariance implique que la covariance moyenne pour ce qui est du caractère aléatoire dans R_i est obtenue par

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{A}] = \mathrm{Cov}_I\left(W_I\pi_I, y_I - m(\mathbf{x}_I)\right), \tag{3.2}$$

où $\mathbf{A} = \{A_i, i \in \mathcal{N}\}$. De même, si l'on est prêt à postuler un modèle conjoint pour $\{(R_i, y_i), i \in \mathcal{N}\}$ conditionné sur $\mathbf{X}$ sous forme d'indépendance $\Pi_{i=1}^{N} P\left(R_i, y_i \mid \mathbf{x}_i\right)$, alors

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{X}] = \mathrm{Cov}_I\left(W_I\pi_I, \mathrm{E}(y_I \mid \mathbf{x}_I) - m(\mathbf{x}_I)\right). \tag{3.3}$$

De façon très intuitive, on peut assurer une covariance ou une corrélation nulle entre deux variables en faisant de l'une des deux une constante. Les deux choix mèneraient alors respectivement à la méthode de quasi-randomisation si l'on fait de $W_I\pi_I \propto 1$ et à la méthode de la superpopulation si l'on fait de $\mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)$ une constante (par exemple zéro). La double robustesse naît du fait que l'une ou l'autre suffise à rendre la covariance nulle (dans le modèle conjoint), puisque la variable n'a pas d'importance. Cependant, il est évident que ce ne sont pas les seules méthodes permettant d'obtenir une corrélation ou une covariance nulle, ou une double robustesse, comme le soulignent Kang et Schafer (2007) dans leur volonté de démystifier la double robustesse (Robins, Rotnitzky et Zhao, 1994; Robins, 2000; Scharfstein, Rotnitzky et Robins, 1999). La question est aussi abordée dans l'étude de Tan (2007, 2010), qui porte sur plusieurs estimateurs et leur comparaison, y compris ceux qui correspondent seulement à la méthode de quasi-randomisation ou à la méthode de la superpopulation. Certains estimateurs sont doublement robustes.

En effet, parce que la formule (2.2) est une identité pour l'erreur réelle, tout estimateur asymptotiquement sans biais (linéaire) de la moyenne de population doit impliquer que la *cdd* correspondante est asymptotiquement sans biais pour les valeurs nulles, et vice versa, pour ce qui est du caractère aléatoire dans R ou dans $\{R, y\}$. Cependant, il est possible que la *cdd* soit asymptotiquement sans biais pour les valeurs nulles, sans supposer que le modèle est correctement précisé, comme l'illustre un exemple dans la section 5. (Cette « robustesse plus que double » est différente de la « robustesse multiple » de Han et Wang (2013), qui doit encore supposer la validité d'au moins un des modèles postulés.) Ces deux observations donnent à penser que toute stratégie générale suffisante et nécessaire qui assure des estimateurs asymptotiquement convergents ou sans biais (linéaires) pour la moyenne de la population équivaudrait à miniaturiser la *cdd*.

À titre d'exemple d'aperçu unifié qui autrement ne serait pas aussi intuitif, l'expression de l'équation (3.2) donne à penser que nous devrions inclure notre estimation de π_I comme élément du prédicteur dans le modèle de régression $m(\mathbf{x}_I)$, puisque cela peut aider à réduire la corrélation entre $W_I\pi_I$ et $z_I = y_I - m(\mathbf{x}_I)$, en particulier quand nous utilisons des poids constants W_I. En général, il est difficile de motiver l'utilisation de $\hat{\pi}_I$ comme prédicteur pour y uniquement du point de vue de la régression, surtout quand nous supposons que y et R sont indépendants étant donné $\mathbf{x}$ (ce qui est habituellement une condition nécessaire pour continuer, comme nous l'expliquons dans la section suivante). Cependant, l'expression de l'équation (3.2) nous indique que pour estimer la moyenne de y, il n'est pas absolument nécessaire d'ajuster le bon modèle de régression $m(\mathbf{x})$. En fait, il suffit de s'assurer que le « résidu » z_I est autant non corrélé à $W_I\pi_I$ quand I varie. Cependant, il est extrêmement important de reconnaître qu'il ne suffit pas d'assurer une corrélation nulle ou faible seulement dans les données observées, car $\mathrm{Cov}_I(W_I\pi_I, z_I \mid R_I = 1)$ nous informe peu sur $\mathrm{Cov}_I(W_I\pi_I, z_I \mid R_I = 0)$. Dans la configuration de Wu (2022), notre capacité à extrapoler de $R_I = 1$ à $R_I = 0$ dépend de la disponibilité des données auxiliaires (indépendantes) indexées par $R_I^* = 1$, ce qui nous permet d'observer certains x_I pour lesquels $R_I = 0$.

La littérature montre les avantages présentés par la stratégie consistant à inclure des estimations de la propension comme prédicteur. Par exemple, Little et An (2004) ont inclus le logit de $\hat{\pi}$ dans leur modèle d'imputation et ils ont constaté que cette inclusion a amélioré la robustesse de la moyenne imputée par rapport à la spécification erronée du modèle d'imputation. Zhang et Little (2009) et Tan, Flannagan et Elliott (2019) ont mis au point cette méthode et ils l'ont améliorée davantage; ils ont utilisé l'expression « robuste au carré » pour souligner la robustesse accrue. Dans un article plus récent portant sur une stratégie similaire pour les échantillons non probabilistes, Liu et coll. (2021) ont montré qu'il était important d'inclure la propension estimée $\hat{\pi}_i$ « comme prédicteur » dans $m(x, \hat{\pi})$ (en utilisant la notation de la présente étude). De plus, dans la littérature sur l'estimation par la méthode du maximum de vraisemblance ciblée (EMVC) pour les modèles semi-paramétriques de traitement des données non probabilistes (van der Laan et Rubin, 2006; Luque-Fernandez, Schomaker, Rachet et Schnitzer, 2018) (voir aussi Scharfstein et coll., 1999; Tan, 2010), les variables $R_I / \hat{\pi}_I$ et $(1 - R_I) / (1 - \hat{\pi}_I)$ sont appelées *covariables intelligentes* et sont utilisées dans les modèles de régression pour y_I. Les mises en œuvre et les théories de l'EMVC et celles de l'EMVC collaborative liée (van der Laan et Gruber, 2009 et 2010), sont mathématiquement plus impliquées que celles en contexte de population finie, comme nous le verrons plus bas, mais les résultats tirés des équations (3.2) et (3.3) peuvent nous permettre d'avoir des raisonnements intuitifs utiles sur la compréhension de l'essence de ces méthodes.

4. Quasi-randomisation *ou* mises en œuvre de la méthode de la superpopulation

En bref, la méthode de quasi-randomisation vise à faire de $W_I\pi_I$ une variable constante (induite par l'indice de population finie I). Lorsque notre échantillon est véritablement sélectionné selon un plan

probabiliste intentionnel, alors $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$ lorsque $i \in \mathcal{N}$ est une probabilité du plan sans y_i, mais elle peut dépendre de $\mathbf{x}_i$, par exemple quand $\mathbf{x}_i$ comprend une variable de stratification. Quand la probabilité du plan n'est pas disponible, nous devons d'abord invoquer une probabilité divine. Il pourrait s'agir d'une probabilité naturelle donnée par la population finie, comme la propension $\pi_i = \Pr_I(R_I = 1 \mid A_I = A_i)$ induite par l'IPF, où $A_i = \{y_i, \mathbf{x}_i\}$, ou d'une probabilité imaginaire de superpopulation comme celle où R_i est généré indépendamment de $\text{Ber}(\pi_i)$, où $\pi_i = \Pr(R_i = 1 \mid A_i) > 0$. Cette hypothèse de positivité est nécessaire si la population finie est précisée au préalable, ou si son imposition définit la population finie qui peut être étudiée. (Cette considération est pertinente en pratique, comme dans le cas des sondages électoraux, où la population finie n'est pas toujours précisée au préalable, pas même en théorie.) Étant donné que ces probabilités divines sont inconnues et qu'elles constituent notre paramètre, nous devons supposer certaines probabilités de procédé, comme par l'intermédiaire d'un modèle linéaire généralisé $\pi_i = g(y_i, \mathbf{x}_i)$ pour poursuivre, même si nous ne croyons pas vraiment en un choix particulier de g.

Aux fins de notre analyse actuelle, nous supposons que notre probabilité divine est donnée par le modèle de superpopulation de Bernoulli, soit $n_R = \sum_{i=1}^{N} R_i$, et $\tilde{p}(\mathbf{A}) = \Pr(n_R > 0 \mid \mathbf{A}) = 1 - \Pi_{i \in N}(1 - \pi_i)$, où $\mathbf{A} = \{A_i, i \in \mathcal{N}\}$. Étant donné que, dans ce cas-ci, R_i est contrôlé par une probabilité divine, la taille de l'échantillon n_R n'est plus une variable du plan de sondage qui conditionne notre schéma de réplication; ce n'est généralement plus une statistique ancillaire. Néanmoins, nous devrions imposer comme condition que $n_R > 0$, ce qui représente une exigence universelle pour l'élaboration d'estimations fondées sur des données pour $\bar{G}$. Heureusement, cette condition ne crée pas de complications mathématiques à la simplicité accordée par l'indépendance de $\pi_i, i \in \mathcal{N}$ comme fonctions de A_i. Cela est dû au fait que $\tilde{\pi}_i(\mathbf{A}) \equiv \Pr(R_i = 1 \mid \mathbf{A}, n_R > 0) = \pi_i / \tilde{p}(\mathbf{A})$, mais la constante de normalisation $\tilde{p}(\mathbf{A})$, qui dépend de la totalité de $\mathbf{A}$, n'est pas pertinente pour les besoins de la présente étude, comme l'attribution de poids proportionnels à $\tilde{\pi}_i^{-1}(\mathbf{A})$.

Par conséquent, selon cette probabilité divine, qui correspond au (véritable modèle pour le) scénario du modèle q dans l'article de Wu (2022), nous obtenons, pour chaque W_I choisi selon l'équation (3.1),

$$\begin{aligned} \mathrm{E}(c_{\tilde{R},z} \mid \mathbf{A}, n_R > 0) &= \mathrm{Cov}_I(W_I \mathrm{E}[R_I \mid \mathbf{A}, n_R > 0], y_I - m(\mathbf{x}_I)) \\ &= \tilde{p}^{-1}(\mathbf{A}) \,\mathrm{Cov}_I(W_I \pi_I, y_I - m(\mathbf{x}_I)), \end{aligned} \tag{4.1}$$

où E est choisi par rapport à la probabilité divine (inconnue) en lien avec R_I (pour une valeur I fixe). Ensuite, que nous voulions ou non garantir une espérance nulle dans l'équation (3.2) ou l'équation (4.1), nous imposerons $W_I \pi_I \propto 1$, c'est-à-dire $W_I \propto \pi_I^{-1}$, la pondération de probabilité inverse bien connue. Par conséquent, si notre modèle postulé q nous permet de saisir π_i de façon fiable dans la réalité, alors $c_{\tilde{R},z} = O_p(N^{-1/2})$ parce qu'il a une moyenne nulle (pour ce qui est de la probabilité divine), et qu'il s'agit d'une moyenne pondérée de N variables de Bernoulli essentiellement indépendantes, comme on le voit dans l'équation (3.1).

Il s'agit d'une approche axée sur la randomisation parce qu'elle traite toutes les valeurs des caractéristiques de la population finie $\mathbf{A}$ comme étant fixes, et les répliques hypothétiques sont générées uniquement au moyen de réalisations répétées de l'indicateur d'enregistrement R_I. Bien sûr, les valeurs de $\{\pi_i, i \in \mathcal{N}\}$ sont généralement inconnues, et pire encore, elles ne peuvent pas être estimées à partir d'un échantillon non probabiliste sans que l'on formule d'autres hypothèses. Pour poursuivre, nous posons des hypothèses comme l'hypothèse de répartition au hasard des données manquantes, c'est-à-dire $\Pr(R_i = 1 \mid A_i) = \Pr(R_i = 1 \mid \mathbf{x}_i)$, et l'exigence d'un échantillon auxiliaire afin que nous ayons certaines valeurs de $\mathbf{x}_i$ au moyen de $R_i = 0$. Nous avons également des choix sur la façon d'estimer la propension d'inclusion $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$, de façon paramétrique ou non paramétrique. Ces hypothèses, ces exigences et ces méthodes d'estimation sont toutes essentielles à la mise en œuvre pratique, comme Wu (2022) en a fait l'évaluation et l'examen avec soin; voir aussi, dans l'article de Tan (2010), une comparaison détaillée des différentes stratégies d'estimation. Néanmoins, l'idée générale des méthodes de quasi-randomisation est de choisir W_I pour libérer $\tilde{R}_I = W_I R_I$ de I en espérance selon les répliques hypothétiques postulées, pour regagner la liberté garantie par l'échantillonnage probabiliste.

De façon complémentaire, les méthodes de superpopulation visent à miniaturiser $c_{\tilde{R}, z}$ en rendant l'autre variable dans $c_{\tilde{R}, z}$, c'est-à-dire z_I, exempte de I en espérance, mais selon un schéma de réplication hypothétique différent. Dans ce cas-ci, l'idée est de choisir un $m(\mathbf{x}_i)$ qui soit une bonne approximation de y_i de sorte que le résidu $z_i = y_i - m(\mathbf{x}_i)$ sera nul en espérance conditionnellement à $\mathbf{x}$. En général, on le fait en considérant un modèle conjoint pour $\{R_i, y_i\}$ étant donné $\mathbf{x}_i$ et avec un modèle de régression particulier $\xi(y \mid \mathbf{x})$, au moyen de la notation de Wu (2022). Il est important d'admettre que, bien que nous ne précisions que le modèle de régression y_i étant donné $\mathbf{x}_i$, nous devons inclure R_i dans les répliques afin de saisir la dépendance possible de R_i à la totalité de $A_i = \{y_i, \mathbf{x}_i\}$, ce qui est la principale préoccupation pour les échantillons non probabilistes. En effet, c'est cette spécification conjointe qui permet l'adoption de l'hypothèse de répartition au hasard des données manquantes pour réduire $P(y_i \mid \mathbf{x}_i, R_i) = P(y_i \mid \mathbf{x}_i)$, ce qui nous permet de nous concentrer sur la spécification d'un modèle de régression unique $\xi(y_i \mid \mathbf{x}_i)$ pour les personnes observées et non observées. Par conséquent, quand nous écrivons E_ξ, nous entendons l'espérance pour ce qui est de

$$P(R_i, y_i \mid \mathbf{x}_i) \;=\; P(R_i \mid \mathbf{x}_i)\, P(y_i \mid R_i, \mathbf{x}_i) \;=\; \pi_i^{R_i}\, (1-\pi_i)^{1-R_i}\, \xi(y_i \mid \mathbf{x}_i), \tag{4.2}$$

où $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$ n'est pas précisé, contrairement à ce qui se passe dans la méthode de quasi-randomisation.

Il s'ensuit alors que, conditionnellement à $\mathbf{X} = \{\mathbf{x}_i, i \in \mathcal{N}\}$ et $n_R > 0$, qui ne modifie pas $P(y \mid \mathbf{X})$ parce que y et R sont indépendants étant donné $\mathbf{X}$, nous obtenons

$$\mathrm{E}(c_{\tilde{R}, z} \mid \mathbf{X}, n_R > 0) \;=\; [\tilde{p}(\mathbf{X})]^{-1}\, \mathrm{Cov}_I\, (W_I \pi_I, \mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)). \tag{4.3}$$

Il est clair que l'équation (4.3) devient nulle quand nous choisissons $m(\mathbf{x}_I) = \mathrm{E}_\xi[y_I \mid \mathbf{x}_I]$ et que le modèle ξ (de premier ordre) est correctement précisé, c'est-à-dire que $\mathrm{E}_\xi[y_I \mid \mathbf{x}_I] = \mathrm{E}[y_I \mid \mathbf{x}_I]$. Cela résume la

méthode de la superpopulation et donne $c_{\tilde{R},z} = O_p(N^{-1/2})$ pour des raisons semblables à celles données pour le cadre de quasi-randomisation.

5. Quasi-randomisation *et* mise en œuvre de la méthode de superpopulation

Une fois qu'un modèle conjoint pour $\{R_i, y_i\}$ est configuré, nous pouvons bien sûr l'utiliser pour estimer à la fois π_i et la fonction de régression $m(\mathbf{x})$. Chaque estimation est rendue possible par la disponibilité de l'échantillon probabiliste auxiliaire et l'hypothèse de répartition au hasard des données manquantes. Mais comme nous l'avons montré précédemment, il suffit de préciser et d'estimer correctement l'une des deux pour miniaturiser $c_{\tilde{R},z}$. Cependant, à partir de l'équation (4.3), pour que la covariance ou la corrélation soit nulle, ni la correction multiplicative de π_I au moyen de W_I ni l'ajustement additif de $\mathrm{E}(y_I \,|\, \mathbf{x}_I)$ au moyen de $m(\mathbf{x}_I)$ n'ont besoin d'être corrects. Il faut seulement qu'après la correction ou l'ajustement, les éléments qui restent ne soient pas corrélés entre eux. Le cadre d'EMVC collaborative mentionné précédemment a essentiellement été élaboré à partir de cette perspective (par exemple voir la section 3.1 de van der Laan et Gruber, 2009), bien que les traitements mathématiques complexes présents dans la littérature aient pu décourager les lecteurs de chercher une compréhension aussi intuitive.

Pour donner un exemple simple, prenons une population finie qui est un échantillon indépendant et identiquement distribué d'un modèle de superpopulation :

$$\mathrm{E}[y \,|\, x] = \sum_{k=0}^{3} \beta_k x^k, \quad x \sim N(0,1). \tag{5.1}$$

L'échantillon non probabiliste est généré par un mécanisme R comme $\Pr\left(R=1 \,|\, y, x\right) = \pi\left(|x|\right)$, c'est-à-dire qu'il est déterminé par l'ampleur de x seulement. Supposons que nous avons précisé incorrectement la forme de fonction pour π (par exemple le modèle divin peut ne pas être monotone dans $|x|$, mais le modèle de procédé, comme le lien logistique classique, l'est), ainsi que le modèle de régression en choisissant $m(x) = b_0 + b_1 x + b_2 x^2$. Puisque x^2 n'est corrélé ni avec x ni avec x^3 selon $x \sim N(0,1)$, nous savons que notre estimateur par les moindres carrés pour b_2 serait toujours valide pour β_2 même dans le modèle de régression incorrectement précisé. Cela suffit pour assurer l'absence de biais asymptotique (quand $N \to \infty$) de l'estimateur « doublement robuste » suivant pour $\mu = \bar{y}_N$, la moyenne de population finie,

$$\hat{\mu}_{+} = \frac{\sum_{i=1}^{N} R_i w\left(|x_i|\right)\left(y_i - \hat{m}(x_i)\right)}{\sum_{i=1}^{N} R_i w\left(|x_i|\right)} + \frac{\sum_{i=1}^{N} R_i^{*} \hat{m}(x_i)}{\sum_{i=1}^{N} R_i^{*}}, \tag{5.2}$$

où R^{*} indique l'échantillon auxiliaire (de $\mathbf{x}$ seulement). Ou encore, de façon équivalente,

$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I\left(R_I w(|x_I|), y_I - \hat{m}(x_I)\right)}{\mathrm{E}_I\left(R_I w(|x_I|)\right)} + \frac{\mathrm{Cov}_I\left(R_I^*, \hat{m}(x_I)\right)}{\mathrm{E}_I\left(R_I^*\right)}, \tag{5.3}$$

ce qui rend plus clair le fait que tout biais dans $\hat{\mu}_+$ est contrôlé par la covariance (ou la corrélation) comprenant R, puisque la covariance comprenant R^* est déjà miniaturisée par l'hypothèse selon laquelle l'échantillon auxiliaire est probabiliste (qui, par souci de simplicité, est supposé être un échantillon aléatoire simple).

Dans le cas présent, $w(x)$ représente toute fonction de pondération telle que $\mathrm{E}_\phi\left[|x|^3 w(|x|)\right] < \infty$, où l'espérance est prise par rapport à $x \sim N(0,1)$, et $\hat{m}(x) = b_0 + b_1 x + \hat{\beta}_2 x^2$, où $\hat{\beta}_2$ est l'estimateur par les moindres carrés pour β_2 à partir de l'échantillon biaisé, et où b_0 et b_1 peuvent être choisis arbitrairement. Parce que la covariance ou la corrélation de la population finie entre $\pi(|x_I|)w(|x_I|)$ et x_I^k est $O_p(N^{-1/2})$ quand $k=1$ et quand $k=3$, les parties ajustées de façon erronée pour π ou m ne contribuent pas (asymptotiquement) à la *cdd*, puisqu'elles ne sont pas corrélées les unes avec les autres dans le modèle de la superpopulation, ce qui entraîne une robustesse supplémentaire dépassant la « double robustesse ». Bien entendu, cela ne signifie pas que nous pouvons ajuster de façon erronée un modèle arbitrairement et tout de même obtenir des estimateurs valides, mais cela implique qu'avoir un modèle correct au moins est une condition suffisante, mais pas nécessaire, pour que les estimateurs doublement robustes soient valides.

Il est aussi intéressant de souligner que, dans la mise en forme du modèle de régression, nous n'avons pas nécessairement besoin d'invoquer une probabilité de procédé, par exemple un modèle de régression de la superpopulation, parce que la variable d'IPF fournit une régression de la population finie par l'application de la méthode des moindres carrés pour exécuter une régression de y_i sur $\mathbf{x}_i, i \in \mathcal{N}$. Cette régression qui s'autoajuste ne nous précise pas vraiment si la droite de régression $y = \hat{m}(\mathbf{x})$ qui en résulte est un bon ajustement de $(y_i, \mathbf{x}_i)$ ou non. Cependant, l'exemple ci-dessus indique que pour estimer la moyenne de population de y, l'absence d'ajustement peut ne pas avoir beaucoup d'importance, tant que le « résidu » $z_I = y_I - \hat{m}(\mathbf{x}_I)$ a peu de corrélation avec $W_I \pi_I$, en tant que deux fonctions de la variable d'IPF I. En effet, comme nous l'avons vu à la section 3, nous pouvons envisager d'inclure $\hat{\pi}_I$ dans le modèle de régression $\hat{m}(\mathbf{x}_I, \hat{\pi}_I)$. D'autres recherches devront examiner le degré d'efficacité générale de cette stratégie.

6. Sous-échantillonnage de contrebalancement

6.1 L'effet dévastateur du défaut des données sur la taille d'échantillon efficace

L'étude de la qualité des données aboutit à une constatation importante qui en a surpris plus d'un, à savoir la petite taille de nos « mégadonnées » si l'on tient compte du défaut des données. Pour la prouver mathématiquement, nous pouvons établir un rapport d'égalité entre l'erreur quadratique moyenne (EQM)

de $\bar{G}_W$ dans l'équation (2.1) et l'EQM d'un estimateur d'échantillonnage aléatoire simple de taille n_{eff}. Cela permet d'obtenir (voir le calcul dans l'article de Meng, 2018) :

$$n_{\text{eff}} \approx \frac{f_W}{1-f_W}\frac{1}{\mathrm{E}[\rho^2_{\tilde{R},G}]} \approx \frac{f_W}{1-f_W}\frac{1}{\rho^2_{\tilde{R},G}}, \tag{6.1}$$

où $f_W = n_W/N$ et l'espérance E est choisie par rapport à la distribution conditionnelle de $\tilde{R}$ étant donné n_W. Il convient de mentionner que cette distribution (conditionnelle) peut comprendre les trois types de probabilité abordés à la section 1.2, car les variations dans $\tilde{R}$ peuvent provenir de sources multiples. À titre d'exemple, dans les sondages d'opinion typiques, il y a 1) une probabilité du plan dans l'indicateur d'échantillonnage, 2) une probabilité divine dans la formulation du mécanisme de la non-réponse et 3) une probabilité de procédé pour l'estimation du mécanisme et des poids.

L'expression (6.1) est la version pondérée ou l'extension de l'expression fournie dans l'article de Meng (2018) comportant des poids égaux, ce qui révèle l'effet dévastateur d'une *cdd* apparemment minuscule. Supposons que notre échantillon représente 1 % de la population et qu'il est miné par une *cdd* de 0,5 %. Si nous appliquons (au moyen des poids égaux) f_W = 0,01 et $\rho_{\tilde{R},G}$ = 0,005 à l'expression (6.1), nous obtenons $n_{\text{eff}} \approx 404$, *quelle que soit la taille de l'échantillon* n_R. Dans le cas de l'élection présidentielle de 2020 aux États-Unis, 1 % de la population d'électeurs représente environ 1,55 million de personnes. La perte de la taille d'échantillon en raison d'une *cdd* de 0,5 % est ainsi d'environ 1 - (404/1 550 000) > 99,97 %. Ces pertes apparemment impossibles ont été relevées dans les études sur les élections (Meng, 2018) et les études sur la vaccination contre la COVID-19 (Bradley et coll., 2021). Une des conséquences les plus dévastatrices de ces pertes est le « paradoxe des mégadonnées » : plus la taille (apparente) des données est grande, plus nous nous leurrons parce que notre fausse confiance (au sens technique et littéral) augmente en fonction de la taille erronée des données, alors que la probabilité de couverture réelle des intervalles de confiance incorrectement construits devient extrêmement faible (Meng, 2018; Msaouel, 2022).

Cette réalité a tout de même une conséquence positive : nous pouvons échanger la grande quantité de données contre des données de qualité et finir par obtenir des estimations statistiquement plus exactes. Bien sûr, pour réduire le biais, il nous faudra des renseignements le concernant. Si nous avons des renseignements fiables sur la valeur de la *cdd*, nous pouvons corriger directement le biais dans l'estimation de la moyenne de la population correspondant à la *cdd*, par exemple par une approche bayésienne semblable à celle adoptée par Isakov et Kuriwaki (2020) dans leur analyse par scénario. De plus, si nous avons suffisamment de renseignements pour construire des poids fiables, nous pouvons utiliser les poids pour corriger les biais de sélection comme nous le faisons habituellement. Néanmoins, même dans ces cas, il est parfois utile de créer une miniature représentative de la population à partir d'un échantillon biaisé à des fins générales. Cela peut, par exemple, éliminer l'anxiété de nombreux spécialistes et les erreurs qu'ils sont susceptibles de commettre parce qu'ils ne maîtrisent pas l'utilisation des poids. De fait, peu de gens savent vraiment traiter les poids, parce que « la pondération des enquêtes est une véritable pagaille! » (Gelman, 2007).

Cependant, en général, la création d'une miniature représentative à partir d'un échantillon biaisé est une tâche difficile, surtout parce que la *cdd* peut (et va) varier selon la variable d'intérêt. Néanmoins, tout comme la pondération est un outil populaire, malgré toutes ses imperfections, nous étudierons la miniaturisation représentative pour voir jusqu'où cette idée peut nous mener. L'exemple suivant nous servira donc seulement de piste de réflexion. Nous examinerons un scénario courant, mais difficile, dans lequel nous disposons de renseignements ou d'une compréhension raisonnables sur l'orientation du biais, c'est-à-dire le signe de la *cdd*, mais de renseignements plutôt vagues sur son ampleur. Un bon exemple est la non-représentativité des sondages électoraux parce que les électeurs ont tendance à ne pas vouloir divulguer leurs préférences quand ils prévoient voter pour un candidat impopulaire dans la société. Nous connaissons donc l'orientation du biais, mais peu de choses au sujet de son degré, à part quelques conjectures (par exemple une fourchette de 10 points de pourcentage).

6.2 La création d'un sous-échantillon moins biaisé

L'idée de base consiste à utiliser ces renseignements partiels au sujet du biais de sélection pour concevoir un plan de sous-échantillonnage *biaisé* afin de *contrebalancer* le biais de l'échantillon original, de sorte que les sous-échantillons qui en résultent aient une *forte probabilité* d'être moins biaisés que l'échantillon original de notre population cible. Autrement dit, nous créons un indicateur de sous-échantillonnage S_I de façon à obtenir une probabilité élevée de réduction de la corrélation entre $S_I R_I$ et G_I, comparativement à la valeur originale de $\rho_{R,G}$, à un point tel qu'elle compensera la perte de la taille d'échantillon et réduira donc l'EQM de notre estimateur (par exemple la moyenne de l'échantillon). Nous parlons de *probabilité élevée*, dans un sens non technique, car en l'absence de renseignements complets sur le mécanisme de réponse ou d'enregistrement, nous ne pouvons jamais garantir qu'un sous-échantillonnage de contrebalancement (SECB) soit toujours meilleur. Il reste qu'une exécution judicieuse peut réduire la probabilité de commettre de graves erreurs.

À titre d'exemple, prenons le cas où y est binaire. Soit $\Delta = r_1 - r_0$, où r_y est la propension de réponse ou de déclaration chez les personnes dont les réponses auront la valeur y: $r_y = \Pr_I (R_I = 1 \,|\, y_I = y)$. Si l'échantillon est représentatif, alors comme $\rho_{R,G}$, Δ est miniaturisé, ce qui signifie qu'il est de l'ordre de $N^{-1/2}$. Cela se constate le plus clairement au moyen de l'identité facilement vérifiable (voir la formule [4.1] dans l'article de Meng, 2018)

$$\Delta = \frac{\mathrm{Cov}_I (y_I, R_I)}{p(1-p)} = \rho_{R,y} \sqrt{\frac{f_R (1-f_R)}{p(1-p)}}, \tag{6.2}$$

où $p = \Pr_I (y_I = 1)$ et $f_R = \Pr_I (R_I = 1)$, qui est le taux d'échantillonnage original. Un ingrédient clé du SECB consiste à déterminer $s_y = P_I (S_I = 1 \,|\, y_I = y, R_I = 1)$ pour $y = 0, 1$, c'est-à-dire, les probabilités de sous-échantillonnage des personnes qui ont déclaré $y = 1$ et $y = 0$, respectivement.

Pour déterminer les choix avantageux, supposons que $f_S = \Pr_I (S_I = 1 \,|\, R_I = 1)$ est le taux de sous-échantillonnage et que $\Delta_S = s_1 r_1 - s_0 r_0$. Ensuite, en appliquant l'équation (2.2) (au moyen de poids égaux)

et l'équation (6.2) à la moyenne de l'échantillon et à la moyenne du sous-échantillon, nous constatons que l'erreur de magnitude (réelle) du sous-échantillon est plus faible si, et seulement si

$$\left(\frac{\Delta_S}{f_S f_R}\right)^2 < \left(\frac{\Delta}{f_R}\right)^2 \Leftrightarrow f_S^2 > \left(\frac{\Delta_S}{\Delta}\right)^2. \tag{6.3}$$

Si l'on écrit $r = r_1/r_0$ et $s = s_1/s_0$, le deuxième membre de l'expression (6.3) devient

$$\left[sp^* + (1-p^*)\right]^2 > \left(\frac{rs-1}{r-1}\right)^2, \tag{6.4}$$

où $p^* = \Pr_I(y_I = 1 | R_I = 1)$ est observé dans l'échantillon original, ce qui devrait nous rappeler que p^* peut être plutôt différent du p que nous recherchons, en raison du mécanisme R biaisé.

Un choix immédiat pour satisfaire l'expression (6.4) consiste à établir $s = r^{-1}$, ce qui bien sûr est généralement irréaliste, car si nous connaissons la valeur de r, le problème serait beaucoup plus simple. Pour étudier la marge de manœuvre dont nous disposons pour nous écarter de ce choix idéal, nous supposons $\delta = r - 1$ pour ensuite montrer que l'expression (6.4) équivaut à

$$(s-1)\left\{[1+(1+p^*)\delta](s-1)+2\delta\right\} < 0. \tag{6.5}$$

Cela indique précisément les choix admissibles de s sans correction excessive (dans l'ampleur du biais qui en résulte) :

(i) Quand $r > 1$, c'est-à-dire $\delta > 0$, nous pouvons prendre tout s de sorte que

$$\frac{[1-(1-p^*)\delta]_+}{1+(1+p^*)\delta} \le s < 1; \tag{6.6}$$

(ii) Quand $r < 1$, c'est-à-dire $\delta < 0$, nous pouvons prendre tout s de sorte que

$$1 < s \le \frac{1-(1-p^*)\delta}{[1+(1+p^*)\delta]_+}. \tag{6.7}$$

Cette paire de résultats confirme un certain nombre de nos intuitions, mais offrent également des quantifications moins évidentes. Étant donné que nous sous-échantillonnons pour compenser le biais de l'échantillon original, s et r doivent rester du côté opposé à 1, c'est-à-dire $(s-1)(r-1) = (s-1)\delta < 0$, comme on le voit dans les expressions (6.6) et (6.7). Pour éviter les corrections excessives, il faut certaines limites, mais il est également possible que le biais initial soit si mauvais qu'aucun schéma de sous-échantillonnage ne puisse empirer les choses, ce qui se reflète dans la fonction de positivité $[x]_+$ des deux expressions mentionnées ci-dessus. Toutefois, pour les limites comme pour les seuils d'activation des fonctions de positivité, les expressions ne sont pas si évidentes. Il n'est pas non plus évident que ces expressions dépendent indirectement de la variable inconnue p par l'intermédiaire de la variable observée

p^*. Ainsi, seule la connaissance préalable de r est requise pour la mise en œuvre ou l'évaluation du SECB.

Cette observation donne à penser que nous pouvons mettre en œuvre un SECB avantageux quand nous pouvons emprunter des renseignements provenant d'autres enquêtes (ou études) dans lesquelles les comportements de réponse ou d'enregistrement sont de nature semblable. Par exemple, nous pourrions apprendre que pour une enquête antérieure semblable, $r = 1{,}5$ (par exemple les enquêtes pour lesquelles $y = 1$ avaient 6 % de chance d'être enregistrées et celles pour lesquelles $y = 0$ avaient seulement 4 % de chance). Compte tenu de l'incertitude de la similarité des deux enquêtes, nous pourrions sans hésitation poser (1,2; 1,8) comme fourchette plausible pour r dans la présente étude. Si nous observons que $p^* = 0{,}6$, cela signifie que le maximum, selon la fourchette $r \in (1{,}2;\ 1{,}8)$, de la limite inférieure de la variable s admissible, comme l'illustre l'expression (6.6), est

$$\frac{[1-(1-0{,}6)\,(r-1)]_+}{1+1{,}6(r-1)} = \frac{[1{,}4-0{,}4r]_+}{1{,}6r-0{,}6} \le \frac{1{,}4-0{,}4\times 1{,}2}{1{,}6\times 1{,}2-0{,}6} = 0{,}7. \qquad (6.8)$$

Par conséquent, tant que nous choisissons $s \in [0{,}7;1)$, il est peu probable que notre correction soit excessive. Le prix que nous payons pour cette robustesse est que le sous-échantillon qui en résulte n'est pas aussi bon qu'il pourrait l'être, par exemple quand la variable sous-jacente r de la présente étude est effectivement 1,5 (en espérance). Quel que soit le choix de $s \in [0{,}7;1)$, il ne donnera pas la correction complète fournie par $s = 1/r = 0{,}67$, c'est-à-dire que la moyenne du sous-échantillon aura toujours un biais positif, mais elle aura une EQM plus petite que la moyenne de l'échantillon d'origine. Il faut bien entendu soigneusement étudier la faisabilité et l'efficacité de ce type de SECB avant de pouvoir le recommander pour un usage général, en particulier au-delà d'un y binaire. La littérature sur l'échantillonnage inverse (Hinkins, Oh et Scheuren, 1997; Rao, Scott et Benhin, 2003) est d'une grande pertinence pour de telles investigations, car elle vise également à produire des échantillons aléatoires simples par sous-échantillonnage, bien qu'avec une motivation différente (pour transformer des enquêtes complexes en enquêtes simples afin de faciliter l'analyse).

7. L'échantillonnage probabiliste comme aspiration et non comme prescription

Comme la définition de la *cdd* devrait l'indiquer explicitement, il n'est pas possible de l'estimer directement à partir de l'échantillon biaisé seulement. C'est la raison pour laquelle on pourrait naturellement se demander (et l'on devrait se demander) dans quelle mesure la *cdd* est utile. En fait, la réponse se révèle de plus en plus longue étant donné que la *cdd* est sans modèle et donc, elle constitue une mesure polyvalente de la qualité des données pour les échantillons probabilistes et les échantillons non probabilistes. Son utilité pour produire des renseignements théoriques est démontrée par son rôle dans la quantification du compromis entre la qualité et la quantité des données au moyen de la taille d'échantillon

efficace, comme on l'a vu dans la formule (6.1), par la compréhension des erreurs de simulation dans la méthode de quasi-Monte Carlo, comme l'a étudiée Hickernell (2016), et par l'anticipation du phénomène de « robustesse plus que double » présenté dans la section 5. Ses emplois méthodologiques sont illustrés dans les analyses de scénarios concernant l'élection présidentielle américaine de 2020 (Isakov et Kuriwaki, 2020) et dans les évaluations de la vaccination contre la COVID-19 (Bradley et coll., 2021). Ses effets pratiques se trouvent dans des études épidémiologiques (Dempsey, 2020), dans des ouvrages sur la physique des particules (Courtoy, Houston, Nadolsky, Xie, Yan et Yuan, 2022) et dans des sondages politiques (Bailey, 2023).

Il n'est pas surprenant que, dans le cadre de ces applications pratiques, les spécialistes aient trouvé utiles la notion de *cdd* et la décomposition des erreurs sous-jacentes de l'équation (2.2) en raison des échantillons non probabilistes avec lesquels ils doivent composer, soit en raison de distorsions des échantillons probabilistes, par exemple celles causées par un mécanisme de non-réponse biaisé, ou en raison de biais de sélection, comme dans le cas des tests sélectifs de dépistage de la COVID-19. L'analyse du professeur Wu, ainsi que les nombreuses références indiquées dans son article et dans la présente étude, devrait clairement montrer que les échantillons non probabilistes sont *presque certainement* partout. Je n'ai pas recours à cette expression fortement probabiliste seulement pour sa valeur humoristique. Lorsque nous considérons le nombre inexplicable de valeurs possibles pour la moyenne de la *cdd*, la probabilité (peu importe la façon dont nous la construisons pour saisir le Far West des processus de collecte de données) qu'elle soit précisément de zéro doit être nulle. Cette moyenne nulle est une condition nécessaire pour que l'échantillon soit un échantillon probabiliste, car un échantillon probabiliste laisse entendre que la *cdd* soit de l'ordre $N^{-1/2}$ (Meng, 2018), ce qui est impossible quand sa moyenne n'est pas nulle (asymptotiquement). Cette observation nous donne à penser que nous devrions nous éloigner de la tradition consistant à traiter l'échantillonnage probabiliste comme une pièce maîtresse, puis à essayer de modéliser le monde beaucoup plus vaste des échantillons non probabilistes comme des « déviations » de celui-ci. Nous devrions plutôt commencer par étudier les échantillons à l'aide de mécanismes de collecte généraux et d'outils ou de concepts comme la *cdd*, puis nous devrions traiter les échantillons probabilistes (de plan) comme un cas idéal très particulier, qui doit toujours rester une aspiration, mais jamais la seule prescription à respecter.

Remerciements

Je tiens à remercier le rédacteur en chef, Jean-François Beaumont, de m'avoir invité à discuter de l'analyse fort opportune et stimulante de Changbao Wu. Je remercie aussi James Bailie, Radu Craiu, Adel Daoud, Andrew Gelman, Stas Kolenikov, Rod Little, Cory McCartan, Kelly McConville, James Robins, Zhiqiang Tan et Li-Chun Zhang de leur appui moral et de leurs critiques constructives. Je remercie également la Fondation nationale des sciences pour son soutien financier partiel et Steve Finch pour sa lecture d'épreuves minutieuse.

Bibliographie

Bailey, M.A. (2023). *Polling at a Crossroads – Rethinking Modern Survey Research*. Cambridge University Press.

Beaumont, J.-F., et Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *Survey Statistician*, 83, 11-22.

Blei, D.M., Kucukelbir, A. et McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, Z.-L. et Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890), 695-700.

Buelens, B., Burger, J. et van den Brakel, J.A. (2018). Comparing inference methods for nonprobability samples. *Revue Internationale de Statistique*, 86(2), 322-343.

Courtoy, A., Houston, J., Nadolsky, P., Xie, K., Yan, M. et Yuan, C.-P. (2022). Parton distributions need representative sampling. *arXivpreprint arXiv:2205.10444.*

Craiu, R.V., Gong, R. et Meng, X.-L. (2022). Six statistical senses. *arXivpreprint arXiv:2204.05313.*

David Peat, F. (2002). *From Certainty to Uncertainty: The Story of Science and Ideas in the Twentieth Century*. Joseph Henry Press.

Dempsey, W. (2020). The hypothesis of testing: Paradoxes arising out of reported coronavirus case-counts. *arXiv preprint arXiv:2005.10425*.

Dwork, C. (2008). Differential privacy: A survey of results. Dans *International Conference on Theory and Applications of Models of Computation*, Springer, 1-19.

Elliott, M.R., et Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153-164.

Gong, R. (2022). Transparent privacy is principled privacy. *Harvard Data Science Review*, (Numéro spécial 2), 24 juin 2022. https://hdsr.mitpress.mit.edu/pub/ld4smnnf.

Gong, R., Groshen, E.L. et Vadhan, S. (2022). Harnessing the known unknowns: Differential privacy and the 2020 Census. *Harvard Data Science Review*, (Numéro spécial 2), 24 juin 2022. https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

Han, P., et Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100(2), 417-430.

Hartley, H.O., et Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174(4423), 270-271.

Hickernell, F.J. (2016). The trio identity for Quasi-Monte Carlo error. Dans *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, Springer, 3-27.

Hinkins, S., Oh, H.L. et Scheuren, F. (1997). Algorithmes de plan de sondage inverses. *Techniques d'enquête*, 23, 1, 13-24. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1997001/article/3101-fra.pdf.

Isakov, M., et Kuriwaki, S. (2020). Towards principled unskewing: Viewing 2020 election polls through a corrective Lens from 2016. *Harvard Data Science Review*, 2(4), 3 novembre 2020. https://hdsr.mitpress.mit.edu/pub/cnxbwum6.

Kang, J.D.Y., et Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523-539.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Li, X., et Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-*p*-zero-*n*: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, 116(533), 353-367.

Little, R., et An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14(3), 949-968.

Liu, Y., Gelman, A. et Chen, Q. (2021). Inference from non-random samples using Bayesian machine learning. *arXiv preprint arXiv:2104.05192*.

Lo, A.W. (2017). Adaptive markets. Dans *Adaptive Markets*. Princeton University Press.

Lohr, S., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475), 1019-1030.

Lohr, S.L. (2021). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Luque-Fernandez, M.A., Schomaker, M., Rachet, B. et Schnitzer, M.E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16), 2530-2546.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). Dans *Past, Present, and Future of Statistical Science*, (Éds., Lin et coll.), CRC Press.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2), 685-726.

Meng, X.-L. (2021). Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(4), 1161-1175.

Msaouel, P. (2022). The big data paradox in clinical practice. *Cancer Investigation*, 1-27.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples: A unified approach. *Calcutta Statistical Association Bulletin*, 69(1), 35-63.

Rao, J.N.K., Scott, A.J. et Benhin, E. (2003). Défaire les structures des données d'enquête complexes : Théorie élémentaire et applications de l'échantillonnage inverse. *Techniques d'enquête*, 29, 2, 119-143. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2003002/article/6787-fra.pdf.

Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. Dans *Proceedings of the American Statistical Association*, Indianapolis, IN, 1999, 6-10.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Scharfstein, D.O., Rotnitzky, A. et Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (avec discussions). *Journal of the American Statistical Association*, 94(448), 1096-1146.

Slavkovic, A., et Seeman, J. (2022). Statistical data privacy: A song of privacy and utility. *arXiv preprint arXiv:2205.03336*.

Tan, Y.V., Flannagan, C.A.C. et Elliott, M.R. (2019). "Robust-Squared" imputation models using Bart. *Journal of Survey Statistics and Methodology*, 7(4), 465-497.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22(4), 560-568.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3), 661-682.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika*, 100(2), 399-415.

Van Buuren, S., et Oudshoorn, K. (1999). *Flexible Multivariate Imputation by MICE*. Leiden: TNO.

van der Laan, M.J., et Gruber, S. (2009). Collaborative double robust targeted penalized maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, 246.

van der Laan, M.J., et Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1).

van der Laan, M.J., et Rubin, D. (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(3), 980-991.

Wu, C. (2022). Inférence statistique avec des échantillons d'enquête non probabiliste (avec discussions). *Techniques d'enquête*, 48, 2, 307-338. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2022002/article/00002-fra.pdf.

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

Yang, S., Kim, J.K. et Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), 445-465.

Zhang, G., et Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3), 911-918.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, 3(2), 103-113.