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 » : Échantillons non probabilistes : évaluation et voie à suivre

par Michael A. Bailey

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. 339-345
Statistique Canada, n° 12-001-X au catalogue

Commentaires à propos de l'article « Inférence statistique avec des échantillons d'enquête non probabiliste » : Échantillons non probabilistes : évaluation et voie à suivre

Michael A. Bailey[1]

Résumé

Les enquêtes non probabilistes jouent un rôle croissant dans la recherche par enquête. L'étude de Wu rassemble de façon compétente les nombreux outils disponibles lorsqu'on suppose que la non-réponse est conditionnellement indépendante de la variable étudiée. Dans le présent exposé, j'étudie la façon d'intégrer les idées de Wu dans un cadre plus large qui englobe le cas dans lequel la non-réponse dépend de la variable étudiée, un cas qui est particulièrement dangereux dans les sondages non probabilistes.

Mots-clés : Échantillonnage d'enquête; sondages non probabilistes.

1. Introduction

Les enquêtes sont en train de subir des changements importants. Nous sommes loin de l'époque où une enquête par composition aléatoire de numéros de téléphone pouvait produire de manière fiable des échantillons représentatifs. À l'heure actuelle, peu de gens répondent aux appels téléphoniques ou même aux courriels. Les enquêteurs ont réagi en proposant une myriade de nouvelles façons ingénieuses de générer des réponses aux enquêtes dans un environnement peu accueillant.

L'innovation la plus répandue est sans aucun doute l'utilisation d'échantillons non probabilistes, souvent par Internet. Bien que la mise en œuvre varie, l'approche consiste habituellement à recueillir les coordonnées d'un grand nombre de personnes qui sont disposées à répondre, puis à sélectionner un sous-ensemble de ce bassin pour une enquête donnée. Ces enquêtes se sont révélées rentables et ont souvent, quoique peut-être pas toujours, produit des résultats utilisables.

Cela dit, ces enquêtes sont-elles crédibles ? La plupart des enquêtes n'ont pas de « vérité fondamentale » qui pourrait être utilisée pour évaluer les résultats; l'absence de renseignement de ce genre est après tout la raison qui a poussé quelqu'un à mener l'enquête. Les échantillons probabilistes permettent de surmonter ce problème en s'appuyant sur la théorie, car les propriétés de ce genre d'enquête sont bien comprises. Pour les échantillons non probabilistes, cependant, la pratique a largement dépassé la théorie, ce qui signifie que les raisons de se fier aux résultats sont plutôt spéculatives.

L'étude de Wu est donc une contribution utile à notre compréhension des enquêtes non probabilistes. Wu se concentre sur la classe d'estimateurs qui supposent une non-réponse ignorable; il les met en contexte les uns par rapport aux autres et il détermine des pistes pour les études à venir.

Wu soulève un point important en mentionnant qu'« il devait y avoir un cadre plus cohérent et des ensembles de mesures d'accompagnement pour évaluer leur qualité » (page 332). Je suis tout à fait

1. Michael A. Bailey, Georgetown University. Courriel : baileyma@georgetown.edu.

d'accord. Dans le présent exposé, j'approfondis ce point de trois façons. À la section 2, j'étudie la façon de le faire dans le cadre des travaux sur lesquels portent ses études. À la section 3, je cherche à élargir la portée d'un tel cadre, en soulignant que les conséquences des violations des principales hypothèses sont tellement plus graves dans un contexte non probabiliste que nous devrions construire notre cadre qui comprend les violations des principales hypothèses des données manquantes au hasard (DMH). À la section 4, j'étudie ensuite ce que nous pouvons faire, si jamais quoi que ce soit peut être fait. Enfin, à la section 5, je présente quelques observations finales.

2. Enquêtes non probabilistes lorsque les données sont manquantes au hasard

Wu fonde son analyse sur une présentation claire des quatre hypothèses sous-jacentes aux modèles qu'il examine. L'hypothèse la plus importante est que les données sont manquantes au hasard, ce qui signifie que pour un ensemble donné de covariables, la variable étudiée est indépendante de la décision de répondre. (Bien que la nomenclature soit normalisée dans la littérature, je ne peux pas m'empêcher d'exprimer mon malaise à l'égard du terme « données manquantes au hasard ». Bien sûr, les données sont manquantes au hasard, ce qui est vrai même pour son « contraire », tout aussi mal nommé, « données non manquantes au hasard ». Je rêve d'un jour où la nomenclature correspondra à la définition, peut-être en remplaçant DMH par le terme « indépendance conditionnelle », qui serait un meilleur nom. Cependant, je reconnais à quel point il est difficile de changer les termes acceptés que les gens utilisent.)

Compte tenu de ces hypothèses, Wu divise les approches selon celles basées sur des modèles, celles basées sur la pondération par l'inverse du score de propension (PISP) et celles en modèles doublement robustes. Dans les approches basées sur les modèles, nous voyons la gamme d'efforts à imputer à partir de l'échantillon observé, y compris l'imputation de masse qui, au sens large, comprend des approches souples d'appariement des échantillons qui nous permettent de représenter une plus grande population en fonction des points de données observés qui sont « proches », un mot ayant diverses définitions. La PISP s'appuie sur les mêmes hypothèses. Les estimateurs doublement robustes ont tendance à être plus récents et attrayants en raison de leur capacité à offrir aux spécialistes deux chances de parvenir à des hypothèses correctes. Wu documente habilement les maux de tête que ces modèles provoquent lorsqu'on les utilise pour réaliser une inférence.

Bien que Wu ait montré les différences dans ces approches, il est utile de comprendre qu'après tout, il n'a fait que « pêcher dans un seul coin de la mare ». Tous les modèles utilisent des renseignements semblables de façons semblables : ils supposent tous que les données sont manquantes au hasard et ils fournissent des outils pour modéliser ou imputer le comportement de personnes non observées sous forme d'extrapolations directes tirées des données observées. Si les titulaires d'un diplôme collégial diffèrent des autres diplômés et que nous avons trop de titulaires d'un diplôme collégial, toutes les approches fondées sur l'hypothèse des DMH consisteront à extrapoler à la population générale directement à partir des

données de l'échantillon sur deux groupes au prorata de la présence de ces groupes dans la population cible.

Mon intuition me dit que les modèles envisagés par Wu sont à peu près aussi utiles, et aussi plus ou moins vulnérables aux violations de l'hypothèse des DMH. Ou y a-t-il des contextes dans lesquels nous nous attendons à ce que les différences entre les méthodes soient considérables ? Il n'est pas facile de répondre à cette question, certes, mais je serais fasciné d'apprendre le point de vue de Wu sur les situations où la principale « action » se trouve dans les échantillons non probabilistes et quels modèles il considère les mieux adaptés pour tenir compte de tels problèmes.

Une des approches possibles serait d'étudier la flexibilité entre les modèles. À ce stade, mon intuition me dit que même si, en théorie, ces différences peuvent être substantielles, en pratique, ces différences sont relativement modestes. Cela est particulièrement vrai si un spécialiste expérimenté possédant des connaissances dans le domaine définit un modèle paramétrique avec doigté, y compris les bonnes interactions et ainsi de suite.

3. Enquêtes non probabilistes lorsque les données ne sont pas manquantes au hasard

Nous devrions prendre très au sérieux l'appel de Wu en faveur d'un cadre plus cohérent pour analyser les échantillons non probabilistes. Et nous devrions viser bien haut, parce qu'un paradigme pour les échantillons non probabilistes est, essentiellement, un paradigme pour l'ensemble du champ de recherche étant donné l'importance et la trajectoire des échantillons non probabilistes.

Alors que nous réfléchissons à l'élaboration d'un cadre de sondage, il est utile de se rappeler le célèbre aphorisme de George Box : « Puisque tous les modèles sont faux, le scientifique doit être vigilant face à ce qu'il y a de plus faux. Il ne convient pas de s'inquiéter des souris quand les tigres sont à l'affût » (Box, 1976). Le tigre dans les échantillons non probabilistes ne rôde pas entre l'échantillonnage par quotas et les modèles de PISP. En fait, le tigre se trouve sans nul doute dans l'hypothèse des DMH. La violation de cette hypothèse constitue la faiblesse emblématique des hypothèses de DMH, et tout cadre pour les enquêtes non probabilistes devrait donc commencer là.

Le problème réside dans le fait que, malgré les violations des hypothèses de DMH qui constituent un problème dans l'échantillonnage probabiliste (découlant de la non-réponse chez les personnes avec lesquelles on communique au hasard), les violations des hypothèses de DMH sont plus graves dans un monde non probabiliste. Meng (2018) précise cette idée et définit l'erreur glissée dans une enquête :

$$\bar{Y}_n - \bar{Y}_N = \underbrace{\rho_{R,Y}}_{\text{qualité des données}} \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{quantité des données}} \underbrace{\sigma_Y}_{\text{difficulté des données}}. \tag{3.1}$$

Le premier segment de l'équation est $\rho_{R,Y}$, soit la corrélation dans la population entre R et Y. Cette quantité peut être utilisée pour refléter la qualité de données relatives à l'échantillonnage. Le deuxième segment dans l'équation de Meng, $\sqrt{N-n/n}$, se rapporte à la taille de la population (N majuscule) et à la taille de l'échantillon (n minuscule). Le troisième segment de l'équation de Meng est σ_Y, l'écart type de Y.

Quand $\rho_{R,Y} \neq 0$, la moyenne échantillonnée sera non nulle sauf si $n = N$ (c'est-à-dire que l'échantillon représente la population tout entière) ou si $\sigma_Y = 0$ (c'est-à-dire que la valeur de Y est pareille pour tous les membres de la population). Dans un cas comme dans l'autre, aucun ne constitue un contexte de sondage intéressant.

Il s'agit d'une identité, par conséquent, même lorsque la valeur prévue de $\rho_{R,Y} = 0$, il en résulte un certain nombre d'erreurs (comme dans le cas de l'échantillonnage aléatoire). Cependant, lorsque nous passons à l'échantillonnage non aléatoire, nous pouvons nous attendre à ce que la corrélation réalisée de R et Y s'élargisse. Plus la valeur de $\rho_{R,Y}$ est grande, plus l'erreur d'échantillonnage est grande, et son ampleur exacte interagira avec les autres segments.

L'implication la plus explosive de l'équation de Meng découle de l'interaction des deux premiers segments. Quand des données manquantes non au hasard (DMNH) (c'est-à-dire qu'il y aura une raison précise de s'attendre à ce que $\rho_{R,Y} \neq 0$ parce que R dépend de Y), l'erreur réelle dépend de la population totale. Ce résultat est choquant en ce qui concerne le caractère délicat des sondages dans le monde actuel, mais il est essentiel d'en tenir compte dans le contexte de l'échantillonnage non aléatoire.

Nous pouvons construire un monde simple constitué de deux pays pour expliquer en détail le fonctionnement. Supposons que les taux d'infection à la COVID-19 représentent notre variable étudiée et, pour notre exemple, que les taux d'infection sont les mêmes dans les deux pays. Le premier pays est immense (comme la Chine) et l'autre est petit (comme le Luxembourg). Si nous échantillonnions *au hasard* 1 000 personnes dans chaque pays, nous pourrions produire des estimations ayant la même précision pour chaque pays, malgré leurs énormes différences sur le plan de la population.

Que se passe-t-il si nous avons affaire à un échantillon *non aléatoire* de 1 000 personnes dans chaque pays ? Supposons, par souci de simplicité, que l'empressement des gens à se faire tester ne soit qu'une fonction de leurs symptômes et que les personnes présentant plus de symptômes soient plus susceptibles d'avoir la COVID-19. Cela crée un échantillonnage de DMNH parce que le choix de faire partie de l'échantillon sera associé à des valeurs attendues plus élevées pour notre variable étudiée.

En Chine, nous obtiendrons les 1 000 personnes les plus malades. Ces personnes seront vraiment malades, car elles feront partie du 0,00001^{e} centile supérieur ou quelque chose du genre. Au Luxembourg, nous aurons aussi les 1 000 personnes les plus malades, mais ces personnes n'ont pas à être aussi malades pour faire partie de cet ensemble par rapport à un ensemble d'un pays beaucoup plus grand. Cela signifie que les 1 000 personnes les plus malades au Luxembourg seront dans environ le 0,2^{e} centile supérieur; elles seront encore très malades par rapport à la population, mais elles ne seront pas aussi biaisées qu'en

Chine. En bref, les DMNH produiront une erreur proportionnelle à la taille de la population pour une taille d'échantillon donnée.

(Il convient de noter que les échantillons véritablement aléatoires sont extrêmement rares, étant donné la non-réponse parmi les gens avec lesquels on communique au hasard. La pratique réelle des échantillons probabilistes peut être décrite comme un contact aléatoire, défini comme des enquêtes dans le cadre desquelles on communique avec les gens au hasard, même si la réponse des personnes avec lesquelles on communique peut être non aléatoire. Les enquêtes par communication aléatoire peuvent contrevenir à l'hypothèse des DMH, mais ont néanmoins de fortes vertus. Bradley, Kuriwaki, Isakov, Sejdinovic, Meng et Flaxman [2021] et Bailey [2023] montrent la façon dont l'erreur d'enquête dans les enquêtes par communication aléatoire est proportionnelle au taux de réponse plutôt qu'à la taille de la population.)

Les violations de l'hypothèse des DMH dans l'échantillonnage non probabiliste entraînent des erreurs qui sont proportionnelles à la taille de la population. Pour utiliser la métaphore de Box, c'est là que se trouvent les tigres. C'est pourquoi, alors que nous poursuivons l'exhortation de Wu concernant une plus grande cohérence dans la façon dont nous évaluons les nouvelles formes de sondages, nous devrions nous entendre pour adopter un cadre qui comprend la possibilité de violations de l'hypothèse des DMH plutôt qu'un cadre qui prétend que ce problème n'existe pas.

4. Que faire au sujet des violations de l'hypothèse des données manquantes au hasard ?

Wu suit une grande partie de la littérature en se détournant des modèles des DMNH. Cela est en partie en raison de la perception selon laquelle la non-réponse des DMNH est essentiellement insoluble. Par exemple, Wu fait remarquer de façon quelque peu pessimiste qu'« il est bien compris que l'hypothèse des données manquantes au hasard ne peut être testée à l'aide des données elles-mêmes de l'échantillon » (page 328) et que « la nature biaisée des échantillons non probabilistes ne peut être corrigée au moyen de l'échantillon lui-même » (page 308).

Pour ce qui est des conseils aux spécialistes de la recherche sur les enquêtes qui s'inquiètent des violations de l'hypothèse des DMH, Wu n'offre qu'un test modeste, qui consiste essentiellement à trouver une autre variable semblable à la variable étudiée, mais qui est disponible pour l'ensemble de la population. Si seulement c'était si facile! Des générations d'enquêteurs ont passé les données au peigne fin pour trouver de telles variables et demeurent néanmoins préoccupées par les DMNH, surtout lorsque la réponse est non aléatoire.

Le cadre de Wu sous-estime ce que nous pouvons faire à propos des violations de l'hypothèse des DMH. Ces efforts exigeront des hypothèses, certes, mais au moins nous pourrons assouplir l'hypothèse intransigeante imposée par le concept des DMH. Le lien avec les points précédents est essentiel : comme nous aurons besoin d'hypothèses, il est important que nous disposions d'un cadre pour réfléchir aux hypothèses les plus importantes afin de pouvoir concentrer nos efforts de façon appropriée. L'équation de

Meng montre de quelle façon les violations de l'hypothèse des DMH jouent un rôle central dans la création d'erreurs dans un monde d'échantillonnage non probabiliste et, par conséquent, nous devrions faire tout ce qui est en notre pouvoir pour régler ce problème.

Le modèle de sélection de Heckman (1979) est un exemple bien connu d'un modèle qui peut s'attaquer à l'hypothèse des DMNH. Ce modèle permet de tenir compte des violations de l'hypothèse des DMH et d'estimer même leur ampleur. Il n'est pas sans problème, bien sûr. En pratique, une restriction d'exclusion est exigée (une hypothèse selon laquelle une ou plusieurs variables ont une incidence sur la réponse, mais pas sur la variable étudiée) et de nombreux spécialistes modernes sont naturellement prudents à l'égard de la forte hypothèse paramétrique du modèle de Heckman.

Les spécialistes ont réalisé des progrès considérables au-delà du modèle de Heckman pour traiter les violations de l'hypothèse des DMH (Bailey, 2023). Il est facile d'assouplir l'hypothèse paramétrique grâce aux fonctions de copule (Gomes, Radice, Brenes et Marra, 2019). Si nous voulons étudier les déterminants de Y, il existe une documentation abondante et grandissante qui applique des fonctions de contrôle très souples aux contextes des DMNH (Das, Newey et Vella, 2003; Liu et Yu, 2022). Et si nous pouvions déterminer les variables qui ont une incidence sur la propension à réagir, mais pas sur le résultat d'intérêt, plusieurs méthodes permettraient de modéliser et de neutraliser l'échantillonnage des DMNH (Peress, 2010; Sun, Liu, Miao, Wirth, Robins et Tchetgen-Tchetgen, 2018).

5. Conclusion

Dans son étude, Wu résume de manière compétente et utile l'état de la littérature en ce qui concerne l'analyse des données d'enquête non probabilistes selon l'hypothèse des DMH. Il souligne également qu'il est grandement nécessaire que le champ de recherche se rallie autour d'un cadre plus cohérent pour évaluer ces innovations et d'autres innovations en matière de sondage.

Dans la présente étude, je me base sur le travail de Wu pour proposer un cadre qui comprend non seulement les modèles des DMH qu'il a analysés, mais aussi les modèles des DMNH, car la violation de l'hypothèse des DMH est particulièrement pertinente et nuisible aux enquêtes non probabilistes.

Bibliographie

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

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791-799.

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.

Das, M., Newey, W.K. et Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70, 33-58.

Gomes, M., Radice, R., Brenes, J.C. et Marra, G. (2019). Copula selection models for nongaussian outcomes that are missing not at random. *Statistics in Medicine*, 38, 480-496.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-162.

Liu, R., et Yu, Z. (2022). Sample selection models with monotone control functions. *Journal of Econometrics*, 226, 321-342.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (1): Law of large populations, big data paradox, and the 2016 presidential election. *The Annals of Applied Statistics*, 12, 685-726.

Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. *Journal of the American Statistical Association*, 105, 1418-1430.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. et Tchetgen-Tchetgen, E.J. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28, 1965-1983.