

Techniques d'enquête

Commentaires à propos de l'article « Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes » :

Inférence causale, échantillon non probabiliste et population finie

par Takumi Saegusa

Date de diffusion : le 25 juin 2024



Statistique
Canada

Statistics
Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Publication autorisée par le ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par le ministre de l'Industrie, 2024

L'utilisation de la présente publication est assujettie aux modalités de l'[entente de licence ouverte](#) de Statistique Canada.

Une [version HTML](#) est aussi disponible.

This publication is also available in English.

Commentaires à propos de l'article « Hypothèse de l'échangeabilité dans des méthodes d'ajustement fondées sur le score de propension aux fins d'estimation de la moyenne de population au moyen d'échantillons non probabilistes » :

Inférence causale, échantillon non probabiliste et population finie

Takumi Saegusa¹

Résumé

Dans certains articles sur les échantillons non probabilistes, l'hypothèse de l'échangeabilité conditionnelle est jugée nécessaire pour une inférence statistique valide. Cette hypothèse repose sur une inférence causale, bien que son cadre de résultat potentiel diffère grandement de celui des échantillons non probabilistes. Nous décrivons les similitudes et les différences entre deux cadres et abordons les enjeux à prendre en considération lors de l'adoption de l'hypothèse d'échangeabilité conditionnelle dans les configurations d'échantillons non probabilistes. Nous examinons aussi le rôle de l'inférence de la population finie dans différentes approches de scores de propension et de modélisation de régression des résultats à l'égard des échantillons non probabilistes.

Mots-clés : Inférence causale; population finie; Échantillon non probabiliste.

1. Introduction

Je félicite la professeure Yan Li pour un autre important ajout à sa recherche active sur les échantillons non probabilistes. Dans son article, la professeure Li a classé la recherche existante sur les échantillons non probabilistes dans les catégories suivantes : 1) les méthodes de pondération du score de propension et 2) les méthodes d'appariement par scores de propension, et indiqué que l'hypothèse d'échangeabilité conditionnelle (EC) était requise pour la première catégorie. Après avoir examiné les méthodes existantes en vue de l'hypothèse d'EC, la professeure Li a proposé le nouveau score d'équilibrage adaptatif pour s'assurer que l'hypothèse d'EC tenait. Étant donné la cristallisation de l'abondance d'articles sur les échantillons non probabilistes et l'inférence causale, son article exige une quantité considérable de connaissances pour comprendre les concepts complexes. Le but premier de notre analyse sera d'examiner les concepts de base et les enjeux fondamentaux que la présentation de la professeure Li n'a abordés que légèrement.

La présente analyse est structurée de la manière suivante. À la section 2, nous analysons l'hypothèse d'échangeabilité conditionnelle dans l'inférence causale. Nous décrivons les différences de cadres probabilistes dans l'inférence causale et les échantillons non probabilistes, et abordons les enjeux à prendre en considération lors de l'adoption de l'hypothèse d'échangeabilité conditionnelle dans les échantillons non

1. Takumi Saegusa, Department of Mathematics, University of Maryland, College Park, Maryland 20742, États-Unis d'Amérique. Courriel : tsaegusa@umd.edu.

probabilistes. Dans la section 3, nous décrivons deux approches principales dans les problèmes de données manquantes, y compris l'inférence causale. Nous abordons ensuite les enjeux liés au rôle de l'inférence de la population finie découlant de l'hypothèse d'échangeabilité conditionnelle dans différentes approches.

2. Inférence causale

Premièrement, nous analysons le lien entre l'hypothèse d'EC et l'inférence causale. Dans le présent article, l'hypothèse d'EC est formulée comme l'équation

$$E[y|b(x), C] = E[y|b(x), U] \quad (2.1)$$

où $b(x)$ est une fonction des covariables x que l'on appellera un score d'équilibrage, U est une population finie et $C \subset U$ est un échantillon non probabiliste. Bien que défini simplement, le critère de son choix dans l'article indique que le score d'équilibrage semble être implicitement déterminé pour satisfaire l'hypothèse d'EC. Qui plus est, il est énoncé comme fait sans autre discussion que toute quantité (y compris le score de propension) inférieure au score de propension satisfait l'hypothèse d'EC comme score d'équilibrage. Une importante analyse qui aide à comprendre ces concepts est celle qui a été rédigée conjointement par la professeure Li (Wang, Graubard, Katki et Li, 2022), qui est, autant que nous sachions, le premier article qui a présenté explicitement les scores d'équilibrage et l'échangeabilité conditionnelle dans l'inférence causale dans la littérature sur les échantillons non probabilistes. Dans Wang, Graubard, Katki et Li (2022), toutefois, ces concepts étaient directement empruntés des travaux de Rosenbaum et Rubin (1983) sur l'inférence causale, et l'on y affirmait que les résultats sur les scores de propension tenaient dans l'environnement non probabiliste sans analyse formelle. Comme les définitions de l'hypothèse d'EC et du score d'équilibrage dans l'article sont différentes de celles de Rosenbaum et Rubin (1983), et comme le cadre contrefactuel de Rosenbaum et Rubin (1983) est relativement différent de la configuration des échantillons non probabilistes, il est intéressant d'accorder une attention particulière aux similitudes et aux différences entre l'inférence causale et les échantillons non probabilistes.

Pour ce faire, nous résumons d'abord brièvement Rosenbaum et Rubin (1983) où les variables d'intérêt sont les résultats potentiels ($Y(0), Y(1)$), covariables X et l'attribution de traitement $Z \in \{0,1\}$. Le score d'équilibrage $b(x)$ dans Rosenbaum et Rubin (1983) a été défini comme la fonction des covariables $X = x$ qui satisfait l'indépendance conditionnelle entre X et l'attribution de traitement Z étant donné $b(X)$ (c'est-à-dire $X \perp Z | b(X)$). Il a été démontré que le score de propension en traitement est un score d'équilibrage, et que toute fonction de x qui peut être cadrée dans le score de propension est aussi un score d'équilibrage. Comme le laisse supposer la définition, il n'y a aucune exigence relativement au lien entre les résultats potentiels et les covariables. L'hypothèse qui relie ces variables est l'échangeabilité conditionnelle en ce qui a trait aux covariables (ou forte ignorabilité de Rosenbaum et Rubin (1983)), définie différemment comme l'indépendance conditionnelle entre les résultats potentiels et l'attribution de traitement étant donné les covariables (c'est-à-dire $(Y(0), Y(1)) \perp Z | X$). Le résultat principal est que

l'échangeabilité conditionnelle en ce qui a trait à la covariable X laisse supposer une échangeabilité conditionnelle en ce qui a trait à un score d'équilibrage $b(X)$. Autrement dit, à partir de l'hypothèse d'échangeabilité conditionnelle clé étant donné les covariables x on peut réduire l'information de x à un score d'équilibrage. Les scores d'équilibrage $b(x)$ ne sont significatifs qu'en présence d'échangeabilité conditionnelle en ce qui a trait aux covariables x . Une conséquence de ce résultat est que la différence entre deux résultats potentiels est expliquée uniquement par l'attribution de traitement.

Une façon naturelle d'appliquer ces résultats à la configuration d'échantillon non probabiliste est de considérer la sélection à l'échantillon non probabiliste comme attribution de traitement, et les résultats de l'échantillon non probabiliste C et du reste dans la population finie (c'est-à-dire $U \setminus C$) comme deux résultats potentiels. Dans cette configuration, l'échangeabilité conditionnelle de Rosenbaum et Rubin (1983) suppose l'échangeabilité conditionnelle en ce qui a trait au score de propension de sorte que C et $U \setminus C$ soient comparables étant donné le score de propension. En revanche, la professeure Li suppose immédiatement la comparabilité de C et U étant donné le score de propension. Du point de vue de l'inférence causale, la comparabilité de Rosenbaum et Rubin (1983) est une conséquence d'une hypothèse vérifiable sur le plan conceptuel alors que la professeure Li commence avec la comparabilité désirée en la supposant. Si, au lieu de cela, on commence à partir de l'échangeabilité conditionnelle comme dans Rosenbaum et Rubin (1983), il se pourrait qu'un résultat ne soit tout de même pas satisfaisant, car deux échantillons (soit C et $U \setminus C$) demeurent différents par « traitement » de participation dans un échantillon non probabiliste. Par exemple, si les échantillons non probabilistes sont des dossiers d'hôpitaux ou des participants à un certain programme éducatif, les deux échantillons sont différents en raison de la réception de soins par l'hôpital ou de l'effet éducatif. Même si nous ne trouvons pas un tel « traitement » qui différencie l'échantillon non probabiliste et le reste, la comparabilité conditionnelle entre C et $U \setminus C$ ne correspond pas nécessairement à la population finie U . Pour obtenir la bonne population cible, il faut obtenir une répartition du score de propension dans la population finie U . Cette tâche n'est pas simple à effectuer comme décrit ci-dessous en ce qui a trait à la représentation des probabilités du score de propension.

Une autre approche est de dévier de l'inférence causale en commençant à partir de l'indépendance conditionnelle entre Y et sélection de Z dans C étant donné X au lieu de l'échangeabilité conditionnelle avec des résultats potentiels. Dans ce cas, toutes les dérivations de fait demeurent valides pour conclure le résultat que $Y \perp Z | X$ suppose que $Y \perp Z | b(X)$ comme souhaité. Cependant, une nouvelle hypothèse d'indépendance conditionnelle est simplement l'hypothèse standard des données manquantes au hasard dans le problème de données manquantes, qui est aussi adoptée par Chen, Li et Wu (2020) dans leur recherche sur les échantillons non probabilistes. L'hypothèse des données manquantes au hasard est connue de nombreux statisticiens et est plus facile à examiner que l'hypothèse d'échangeabilité conditionnelle de la professeure Li. Si cette approche est celle qui est implicitement adoptée dans Wang, Graubard, Katki et Li (2022), ainsi que dans le présent article, il convient d'aborder les avantages supplémentaires de cette approche par rapport à l'hypothèse des données manquantes au hasard en plus de l'écart entre $U \setminus C$ et U

pour la comparabilité. Si une approche différente est adoptée, un lien non vérifié entre les scores d'équilibrage et l'hypothèse d'EC (2.1) devrait être dérivé de manière explicite. Par ailleurs, nous aimerions souligner que Chen, Li et Wu (2020) n'est pas le seul article qui n'utilise pas l'hypothèse d'EC de la professeure Li pour les méthodes de pondération du score de propension (voir par exemple Kim et Morikawa [2023] pour le cas de données manquantes non ignorables).

Comme mentionné ci-dessus, la comparabilité de C et $U \setminus C$ permet une estimation fiable du modèle de régression selon C pour les éléments dans $U \setminus C$ mais l'estimation de \bar{Y}_N exige une estimation cohérente des scores de propension pour U pour relier la régression étant donné X à l'ensemble de la population U . Cependant, une simple estimation du score de propension n'est pas possible, car X n'est pas disponible pour tous les éléments dans $U \setminus C$. La variable X est disponible dans un échantillon de référence S de U avec un plan d'échantillonnage connu, mais S n'est pas une solution de rechange simple pour $U \setminus C$ car les éléments dans S peuvent aussi être dans un échantillon non probabiliste C . Pour solutionner cet enjeu complexe, Wang, Valliant et Li (2021) ont découvert le lien entre le score de propension dans C par rapport à U et le score de propension dans C par rapport à l'échantillon empilé de C et U où les mêmes éléments dans C et S sont traités différemment (pour une dérivation rigoureuse, voir Savitsky, Williams, Gershunskaya, Beresovsky et Johnson [2023]). Au moyen de cette relation, la professeure Li a modélisé le dernier score de propension par régression binaire pour estimer le premier. L'événement pour le dernier score de propension pour un échantillon empilé est construit artificiellement et conceptuellement difficile à modéliser. Cet enjeu augmente la possibilité accrue d'erreur de spécification du modèle, qui invaliderait l'estimation cohérente avec le plan d'échantillonnage de \bar{Y}_N . L'événement pour le premier score de propension est l'événement initial et est naturel à modéliser. Cette approche a été adoptée par Savitsky, Williams, Gershunskaya, Beresovsky et Johnson (2023).

3. Inférence de la population finie

Un autre concept que nous voulons aborder est le rôle de la population finie dans les échantillons non probabilistes. Le but de l'article est de mettre au point un estimateur cohérent avec le plan d'échantillonnage de la moyenne de la population finie \bar{Y}_N . Aux fins de convergence par rapport au plan de sondage, l'on suppose une série de conditions dans la séquence des populations finies avec toutes les variables sauf la sélection dans les échantillons traitée de façon non aléatoire. En revanche, l'approche basée sur le modèle traite la population finie comme réalisation aléatoire à partir de la super population, et modèle la relation stochastique parmi les variables. Dans la recherche sur les données manquantes, par ailleurs, deux approches principales (et leurs combinaisons) pour l'estimation sont la modélisation du score de propension et la modélisation de régression du résultat. Une approche plus convenable à l'égard de l'approche fondée sur le plan de sondage est la modélisation du score de propension qui modélise la sélection dans les échantillons étant donné les covariables, car il est possible de considérer des sélections aléatoires alors que toutes les

autres variables peuvent être traitées fixes. En revanche, la modélisation de régression du résultat présume une répartition pour Y étant donné X , et elle convient à l'approche basée sur un modèle.

La professeure Li a fait une tentative difficile de lier l'approche de régression du résultat à l'approche fondée sur le plan de sondage. Il convient de noter que l'attente conditionnelle peut être considérée comme une régression avec des variables conditionnelles comme covariables. De cette perspective, l'approche semble être purement l'approche fondée sur un modèle en fonction de la régression du résultat. Cependant, la professeure Li a tenté de soigneusement mettre au point l'attente conditionnelle étape par étape en commençant par une population finie et un échantillon non probabiliste. Si la condition était purement fondée sur un modèle, la variable y dans la condition (2.1) est simplement une variable aléatoire à partir de la super population. Dans l'approche conditionnelle de l'article, cette variable y devrait être clairement définie par rapport à la population finie U et l'échantillon non probabiliste C au moyen d'indices. Si y est un choix aléatoire d'une variable d'un échantillon S de U , $E_S[y|U] = \sum_{i \in U} \pi_i Y_i$ où π_i désigne la probabilité d'inclusion pour l'unité i . Dans ce cas, l'échantillon autopondéré S satisfait $E_S[y|U] = \bar{Y}_N$ mais un échantillon stratifié S , par exemple, ne satisfait pas cette équation en général. Autrement dit, il se pourrait que l'enjeu de biais allégué ne soit pas unique à un échantillon non probabiliste. Pour mesurer pleinement la condition d'échangeabilité conditionnelle, une définition claire de y dans C ou U est plus que souhaitable. En outre, il est souhaitable d'élucider la manière dont la condition fondée sur un modèle de l'hypothèse d'EC mène au résultat fondé sur le plan de sondage malgré l'écart conceptuel.

Bibliographie

- Chen, Y., Li, P. et Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), 2011-2021. Extrait de <https://doi-org.proxy-um.researchport.umd.edu/10.1080/01621459.2019.1677241>. Doi: 10.1080/01621459.2019.1677241.
- Kim, J., et Morikawa, K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. À paraître dans *Calcutta Statistical Association Bulletin*, 35.
- Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55. Extrait de <https://doi-org.proxy-um.researchport.umd.edu/10.1093/biomet/70.1.41>. Doi: 10.1093/biomet/70.1.41.
- Savitsky, T.D., Williams, M.R., Gershunskaya, J., Beresovsky, V. et Johnson, N.G. (2023). *Methods for Combining Probability and Nonprobability Samples Under Unknown Overlaps*.
- Wang, L., Graubard, B.I., Katki, H.A. et Li, Y. (2022). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. *Revue Internationale de Statistique*, 90, 146-164.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(24), 5237-5250. Extrait de <https://doi-org.proxy-um.researchport.umd.edu/10.1002/sim.9122>. Doi: 10.1002/sim.9122.