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

Techniques d'enquête

Réponse de l'auteur aux commentaires sur 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 »

par Yan Li

Date de diffusion : le 25 juin 2024

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

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.

Techniques d'enquête, juin 2024
Vol. 50, nº 1, pp. 81-86
Statistique Canada, nº 12-001-X au catalogue

Réponse de l'auteur aux commentaires sur 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 »

Yan Li[1]

Résumé

Dans cette réplique, je réponds aux commentaires des participants à l'analyse, M. Takumi Saegusa, M. Jae-Kwang Kim et Mme Yonghyun Kwon. Les commentaires de M. Saegusa, qui portent sur les différences entre l'hypothèse d'échangeabilité conditionnelle (EC) pour les inférences causales et l'hypothèse d'EC pour les inférences de population finie au moyen d'échantillons non probabilistes ainsi que sur la distinction entre les méthodes fondées sur le plan et celles fondées sur un modèle pour l'inférence de population finie au moyen d'échantillons non probabilistes, sont examinés et clarifiés dans le contexte de mon article. Je réponds ensuite au cadre exhaustif de M. Kim et de Mme Kwon pour classer les méthodes actuelles d'estimation des scores de propension (SP) en méthodes conditionnelles et inconditionnelles. J'étends leurs études par simulations pour varier les poids de sondage, permettre des modèles de SP incorrectement précisés, et inclure un estimateur supplémentaire, à savoir l'estimateur par la propension logistique ajustée mis à l'échelle (Wang, Valliant et Li [2021], noté sWBS). Dans mes simulations, on observe que l'estimateur sWBS dépasse de façon constante les autres estimateurs ou leur est comparable dans le modèle de SP incorrectement précisé. L'estimateur sWBS, ainsi que les estimateurs WBS ou ABS décrits dans mon article, ne supposent pas que les unités superposées dans les échantillons de référence probabiliste et non probabiliste sont négligeables, et ils n'exigent pas non plus l'identification des unités superposées, comme le nécessitent les estimateurs proposés par M. Kim et Mme Kwon.

Mots-clés : Échangeabilité conditionnelle; inférences causales; score de propension; essais randomisés; études d'observation; étude de séroprévalence du SRAS-CoV-2.

Je tiens à remercier les participants à la discussion pour leurs observations perspicaces sur mon article ainsi que pour les excellentes références supplémentaires qu'ils ont citées. Je commencerai par répondre aux commentaires de M. Saegusa sur deux éléments importants. Le premier compare les différences entre l'hypothèse d'échangeabilité conditionnelle (EC) pour les inférences causales et l'hypothèse d'EC pour les inférences de population finie au moyen d'échantillons non probabilistes. Le deuxième élément porte sur la distinction entre les méthodes fondées sur le plan et celles fondées sur un modèle aux fins d'inférence de population finie au moyen d'échantillons non probabilistes.

1. Réponse aux commentaires de M. Saegusa

L'hypothèse d'EC dans l'inférence causale et l'inférence de population finie

M. Saegusa a fourni une explication détaillée de l'hypothèse d'EC pour l'estimation des effets causaux des traitements dans des essais randomisés et des études d'observation. La condition clé est la satisfaction

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

de l'échangeabilité conditionnelle pour que soient effectuées des inférences causales. Dans les essais randomisés, l'EC des résultats potentiels est obtenue par l'attribution aléatoire de traitements. En revanche, dans l'analyse des études d'observation, l'EC est supposée plutôt que garantie afin que soient tirées les conclusions causales. L'hypothèse d'EC dans les études d'observation affirme que la distribution des résultats potentiels (pour différents traitements), compte tenu de toutes les covariables *observées*, est échangeable entre les groupes de traitement. Dans ce cas, il n'y a pas de covariables *non observées* qui influencent à la fois l'attribution du traitement et le résultat d'intérêt; voir Rubin (2007) pour en savoir plus sur l'estimation de l'effet causal au moyen d'essais randomisés et d'études d'observation. Notons que contrairement aux *essais randomisés*, dans le contexte de l'inférence de population finie (FP) au moyen d'échantillons non probabilistes, l'EC est supposée plutôt que garantie, ce qui est semblable à l'hypothèse d'EC nécessaire aux fins de l'inférence causale dans les études d'observation.

M. Saegusa a correctement établi un lien entre l'autosélection dans les échantillons non probabilistes et l'attribution du traitement dans l'inférence causale, et il a défini que le fait d'être autosélectionné dans l'échantillon non probabiliste C par rapport à dans le reste de la population finie (c.-à-d. U\C) comme étant les deux traitements. Toutefois, dans l'inférence de PF au moyen d'échantillons non probabilistes, nous souhaitons estimer la moyenne de PF pour un seul résultat, plutôt que les effets du traitement, c.-à-d. la différence entre deux résultats potentiels selon les traitements C et U\C. La situation n'est pas celle de plusieurs traitements appliqués à des groupes « semblables » pour obtenir plusieurs résultats. On a plutôt un seul résultat potentiel réalisé. Par conséquent, l'hypothèse d'EC dans l'inférence de PF diffère de l'EC dans les études d'observation, où elle affirme que la distribution du résultat (*unique*) est échangeable entre l'échantillon non probabiliste C et la population finie U, compte tenu de toutes les covariables *observées*. Selon cette hypothèse d'EC, on peut alors inférer la moyenne de PF au moyen de C sans observer le résultat dans U\C.

En résumé, l'EC dans l'inférence de PF est semblable à l'EC dans l'inférence causale quand on utilise des études d'observation. Cependant, contrairement à l'inférence causale, l'échangeabilité se rapporte à un seul résultat entre C et U dans l'inférence de PF.

Méthodes fondées sur un modèle ou sur le plan de sondage aux fins d'inférences de population finie

M. Saegusa a bien exposé les différences fondamentales entre les méthodes fondées sur un modèle et celles fondées sur le plan aux fins de l'inférence de PF. Les méthodes fondées sur un modèle traitent le résultat comme une variable aléatoire, tandis que les méthodes fondées sur le plan considèrent l'indicateur de sélection (dans l'échantillon) comme un indicateur aléatoire (avec un résultat constant). Dans l'article, un ensemble de pseudo-poids fondés sur le plan a été construit pour l'échantillon non probabiliste afin d'estimer la moyenne de la PF selon l'hypothèse d'EC de $E\{y \mid b(\mathbf{x}), C\} = E\{y \mid b(\mathbf{x}), U\}$, où l'espérance $E(.)$ est par rapport à deux niveaux de caractère aléatoire de (1) la réalisation aléatoire de la PF à partir d'une superpopulation et de (2) l'autosélection aléatoire dans C à partir de la population finie U.

M. Saegusa indique ensuite qu'une « définition claire de y dans C et/ou U est souhaitable ». Cependant, les résultats pour obtenir une estimation sans biais de la moyenne de PF s'appliquent tant que la PF est la réalisation aléatoire d'une superpopulation. Seule l'existence d'une fonction de distribution avec des moments finis appropriés de variables est nécessaire pour la superpopulation. Il n'est pas nécessaire de préciser une forme spécifique du modèle paramétrique; voir par exemple Graubard et Korn (2002) pour en savoir plus.

2. Réponse aux commentaires de M. Jae-Kwang Kim et de Mme Yonghyun Kwon

Dans ce qui suit, j'aborde la discussion très judicieuse de M. Kim et de Mme Kwon, dans laquelle ils ont d'abord présenté un cadre exhaustif établi pour classer les méthodes actuelles d'estimation des scores de propension (SP) en méthodes conditionnelles et inconditionnelles. Ils ont ensuite mené des études par simulations comparant différents estimateurs, et je suis heureux de constater que l'estimateur ABS (échantillonnage fondé sur des adresses) que nous avons proposé a bien fonctionné dans leurs simulations.

Les méthodes conditionnelles comprennent deux phases : la première phase consistant à échantillonner $S_1 = A \cup B$ à partir de U et la deuxième phase à échantillonner B à partir de S_1. Les paramètres du modèle ϕ de la probabilité d'inclusion conditionnelle pour la deuxième phase $P(i \in S_2 \mid i \in S_1) = p(x_i;\phi)$ sont estimés par

$$\hat{\phi} = \arg\max \sum_{i \in S_1} \left[\delta_i \log(p(x_i;\phi)) + (1-\delta_i)\log(1-p(x_i;\phi)) \right],$$

où $\delta_i = I(i \in B)$ est la fonction indicatrice de l'événement $i \in B$. Selon un modèle statistique, disons un modèle de régression logistique $\text{logit}\{p(x_i;\phi)\} = x_i'\phi$, on peut obtenir $\hat{\phi}$ en résolvant pour ϕ

$$\sum_{i \in B} \left(1 - p(x_i;\phi)\right) x_i - \sum_{i \in A \text{ et } i \notin B} p(x_i;\phi)\, x_i = 0.$$

Notons que les unités qui se superposent et qui sont sélectionnées dans les deux échantillons de A et de B doivent être identifiées et retirées de l'échantillon A pour la deuxième sommation ci-dessus. Involontairement, « et $i \notin B$ » a été omis dans la deuxième sommation de la discussion. À partir de l'estimation $\hat{\phi}$, M. Kim et Mme Kwon ont proposé le pseudo-poids du score de propension conditionnel ou SPC pour la i^{e} unité dans B, donné par

$$\hat{w}_i^{(B)} = 1 + w_i^{(A)} \left(\frac{1}{p(x_i;\hat{\phi})} - 1 \right),$$

où $w_i^{(A)}$ est souvent inconnu et estimé selon un modèle paramétrique en pratique.

Les méthodes inconditionnelles comprenaient seulement une étape. L'estimateur par le maximum de vraisemblance conditionnelle de ϕ a été estimé à partir de l'échantillon combiné $S_1 = A \cup B$. Tout comme

l'estimateur SPC, la méthode proposée du score de propension inconditionnel (SPIC) repose également sur l'hypothèse que les unités qui appartiennent à l'intersection de A et B peuvent être identifiées.

Le SP conditionnel et le SP inconditionnel proposés ont été évalués au moyen d'études par simulations, où il a été tenu compte de tailles d'échantillon variées dans l'échantillon B sélectionné au moyen d'un échantillonnage aléatoire simple stratifié (EASS) avec une variable de stratification par catégorie. Bien que simple, ce plan est intelligent. Il s'harmonise avec le véritable modèle de SP sous-jacent pour toutes les méthodes examinées, ce qui assure une comparaison équitable. Pour mieux évaluer le rendement des estimations proposées, nous avons étendu les études par simulations en incluant un estimateur supplémentaire selon le même plan d'échantillonnage EASS, mais en faisant varier les poids de sondage. Rappelons que la taille de la population est $N = 5\ 000$, que la taille de l'échantillon A est $n_A = 250$, et celle de l'échantillon B varie entre $n_B = 250$ et 2 500. Nous considérons les trois estimateurs qui ont les plus petites erreurs quadratiques moyennes (EQM) dans le tableau 4.1 de leur discussion : WBS, ABS et SPC. Puisque le SPIC fonctionne de la même façon que le SPC, il n'est pas pris en compte ici. Rappelons que WBS désigne l'estimateur par la propension logistique ajustée proposé par Wang et coll. (2021). Dans le même article, les auteurs ont également proposé l'estimateur WBS mis à l'échelle, noté sWBS, dans lequel les poids mis à l'échelle sont la valeur de 1 pour les unités de l'échantillon B et $n_s w_i^{(A)} \Big/ \sum_{i \in S_A} w_i^{(A)}$ pour l'unité i de l'échantillon A. Les fractions de sondage varient à l'intérieur des strates d'échantillonnage. Dans la strate 1, $n_{B1} = f_1 n_B$ échantillons sont sélectionnés par échantillonnage aléatoire simple. Dans la strate 2, $n_{B2} = (1 - f_1)\, n_B$ échantillons sont sélectionnés par échantillonnage aléatoire simple. La valeur de f_1 varie et est 0,7, 0,8 et 0,9 pour produire des valeurs différentes du coefficient de variation des poids de sondage de l'EASS (CVWT pour l'anglais *coefficient of variation of the weights*). Dans l'analyse du score de propension, nous considérons deux modèles : (M1) les effets principaux (x_1, x_2, x_3) et leurs effets d'interaction par paires; (M2) les effets principaux (x_{1c}, x_2, x_3) et leurs effets d'interaction par paires, où $x_{1c} = I\{x_1 < 0\}$, la fonction indicatrice de l'événement $x_1 < 0$ où x_1 sont générés à partir d'un $N(0,1)$. Notons que M2 est conforme au plan d'EASS tandis que M1 est incorrectement précisé par l'inclusion de la variable continue x_1 dans l'analyse du SP.

Quatre observations sont formulées : (1) les quatre estimateurs sont approximativement sans biais selon le vrai modèle de SP; (2) les estimateurs ABS et SPC ont des performances similaires pour une petite taille d'échantillon de $n_B = 250$ selon les deux modèles; (3) quand la taille de l'échantillon est grande $n_B = 2\ 500$, le SPC a toujours le plus petit ET et la plus petite EQM dans le vrai modèle. Ces résultats correspondent aux attentes puisqu'il y a un fort pourcentage d'unités superposées dans les deux échantillons. Par conséquent, la méthode du SPC gagne en efficacité, car elle suppose que les unités superposées peuvent être identifiées ; (4) dans le modèle de SP incorrectement précisé, sWBS présente toujours le biais le plus petit, particulièrement quand le coefficient de variation des poids de sondage est grand. En revanche, le SPC a le biais le plus grand et l'ET le plus grand quand le coefficient de variation des poids de sondage et n_B sont grands. Le biais et la perte d'efficacité du SPC peuvent être attribués à la spécification erronée du modèle de $P(i \in B \mid i \in A \cup B)$, la taille limitée de l'échantillon obtenue par la suppression des unités superposées

(~50 %) de l'échantillon A, et les poids de sondage variables. Le SPC est sensible à la spécification erronée du modèle, en particulier quand n_B et le coefficient de variation des poids de sondage sont grands.

En résumé, sous le vrai modèle de score de propension, les estimateurs ABS et SPC ont des performances similaires quand n_B est petit. Quand n_B est grand, l'estimateur SPC est plus efficace en raison du nombre croissant d'unités sélectionnées et identifiées dans les deux échantillons, A et B. En cas de spécification erronée du modèle de SP, sWBS (Wang et coll., 2021) a de meilleures performances ou des performances comparables à celles des autres estimateurs. Il faudrait étudier de façon plus approfondie les effets de divers modèles de SP incorrectement précisés ou les valeurs de mise à l'échelle sur les performances de l'estimateur sWBS. Ensuite, on a élaboré les estimateurs ABS, WBS, sWBS, ainsi que SPC, sans supposer que les unités superposées dans les deux échantillons sont négligeables. Pour un échantillon de grande taille n_B = 2 500, le taux d'échantillonnage pour l'échantillon B, n_B/N = 50 %, est non négligeable. Tous les estimateurs, comme le montre le tableau 1, sont approximativement sans biais selon le vrai modèle de SP, qui prouve empiriquement que les quatre méthodes ne nécessitent pas l'hypothèse selon laquelle les unités superposées dans les deux échantillons sont négligeables. Enfin, du point de vue pratique, il importe que le lecteur sache que l'estimateur SPC exige l'identification des unités superposées. Or, cela peut être impossible dans de nombreuses situations. Par exemple, dans l'étude de la NIH sur la séropositivité au SARS-CoV-2 abordée dans mon article, ces renseignements d'identification n'avaient pas été recueillis.

Tableau 1
Biais, erreur type et erreur quadratique moyenne (× 100) selon un EASS avec variation du CV des poids de sondage (CVWT) après 5 000 répétitions.

	Modèle de SP correctement précisé (x_{1c})						**Modèle de SP incorrectement précisé (x_1)**					
	n_B = 250			**n_B = 2 500**			**n_B = 250**			**n_B = 2 500**		
	BIAIS	**ET**	**EQM**	**BIAIS**	**ET**	**EQM**	**BIAIS**	**ET**	**EQM**	**BIAIS**	**ET**	**EQM**
CVWT = 0,44												
Moyenne C	-3,25	2,98	4,41	-3,33	0,94	3,46	4,84	2,87	5,63	4,74	0,93	4,83
WBS	0,04	3,60	3,60	-0,04	1,42	1,42	0,51	3,06	3,10	0,23	1,54	1,56
sWBS	0,02	3,56	3,56	-0,05	1,41	1,41	0,33	3,05	3,07	0,18	1,62	1,63
ABS	0,02	3,54	3,54	-0,04	1,38	1,38	0,56	2,98	3,04	0,56	1,45	1,55
SPC	0,01	3,53	3,53	0,01	1,20	1,20	0,55	2,99	3,04	0,12	1,36	1,36
CVWT = 0,75												
Moyenne C	-4,97	2,89	5,76	-4,99	0,92	5,07	7,10	2,97	7,70	7,10	0,95	7,16
WBS	-0,03	4,16	4,16	0	1,56	1,56	1,03	3,18	3,34	0,47	1,55	1,62
sWBS	-0,08	4,09	4,09	-0,03	1,55	1,55	0,39	3,23	3,25	0,18	1,64	1,65
ABS	-0,08	4,08	4,08	-0,02	1,52	1,52	1,13	3,15	3,34	1,11	1,52	1,88
SPC	-0,08	4,07	4,07	0,05	1,34	1,34	1,10	3,16	3,34	-0,41	1,64	1,69
CVWT = 1,33												
Moyenne C	-6,58	2,88	7,19	-6,66	0,90	6,72	9,49	3,07	9,98	9,45	0,97	9,50
WBS	0,11	5,65	5,65	0,00	1,91	1,91	2,74	3,49	4,44	1,39	1,67	2,17
sWBS	0,06	5,54	5,54	-0,03	1,89	1,89	1,07	3,78	3,93	-0,05	1,85	1,85
ABS	0,03	5,49	5,49	-0,04	1,86	1,86	2,60	3,50	4,36	1,94	1,69	2,58
SPC	0,03	5,49	5,49	0,03	1,71	1,71	2,54	3,54	4,36	-3,00	3,10	4,31

Bibliographie

Graubard, B.I., et Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17(1), 73-96.

Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat Med.*, 26(1), 20-36. Doi: 10.1002/sim.2739. PMID: 17072897.

Wang, L., Valliant, R. et Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.*, 40(24), 5237-5250. Doi: 10.1002/sim.9122. PMID: 34219260; PMCID: PMC8526388.