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

Techniques d’enquête

Réponse de l’auteur aux commentaires sur l’article « Inférence statistique avec des échantillons d’enquête non probabiliste »

par Changbao Wu

Date de diffusion : le 15 décembre 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2022
Vol. 48, nº 2, pp. 399-406
Statistique Canada, nº 12-001-X au catalogue

Réponse de l'auteur aux commentaires sur l'article « Inférence statistique avec des échantillons d'enquête non probabiliste »

Changbao Wu[1]

Résumé

La présente réponse contient des remarques supplémentaires sur certaines questions soulevées par les participants à la discussion.

Mots-clés : Corrélation due à un défaut des données; double robustesse; pondération de probabilité inverse; hypothèses de modèle; prédiction fondée sur un modèle; échantillon de validation.

Permettez-moi tout d'abord de remercier le rédacteur en chef de *Techniques d'enquête*, Jean-François Beaumont, d'avoir organisé les discussions et d'avoir rassemblé un flamboyant ensemble d'intervenants. Chaque participant à la discussion s'est penché sur la question des échantillons d'enquête non probabilistes et de façon plus générale, sur les sujets de l'intégration des données et de la combinaison de données provenant de sources multiples, avec des points de vue uniques. Ces discussions stimulantes contribuent, selon moi, de façon importante au traitement des échantillons non probabilistes et d'autres types d'échantillons présentant un biais de sélection. Dans les lignes qui suivent, je ferai quelques observations supplémentaires sur certaines questions soulevées par les participants à la discussion.

Michael A. Bailey

Michael A. Bailey s'est concentré sur les limites des méthodes d'estimation que j'ai présentées dans le cadre des hypothèses A1 à A4, et il a appelé à poursuivre le développement en cas de violation de ces hypothèses ainsi que de l'hypothèse dite « de données manquantes au hasard » A1 en particulier. Bailey s'est servi de l'exemple du sondage non probabiliste pour soutenir que « la non-réponse dépend (peut en effet dépendre) de la variable étudiée » et que le danger de violation de l'hypothèse A1 est réel.

Bien que les critiques sur les limites des méthodes examinées dans mon article soient justes et honnêtes, les énoncés « (Wu) pêche dans un coin très précis de l'étang » et il « s'éloigne des modèles de données manquantes non au hasard » semblent montrer une sous-appréciation marquée de l'importance de l'élaboration de méthodes selon les hypothèses types A1 à A4, employées par plusieurs auteurs sur des échantillons d'enquête non probabilistes. Premièrement, l'hypothèse A1 porte sur le mécanisme de participation (ou d'inclusion ou de sélection) pour des échantillons non probabilistes, ce qui n'est pas la même chose que la « non-réponse ». Ces hypothèses peuvent se justifier dans de nombreux scénarios, surtout pour les enquêtes reposant sur des panels Web ou téléphoniques, dans lesquels la participation initiale dépend fortement de certaines variables démographiques. Deuxièmement, le comportement de

1. Changbao Wu, Département des sciences statistiques et actuarielles, Université de Waterloo, Waterloo (Ontario), N2L 3G1. Courriel : cbwu@uwaterloo.ca.

participation aux enquêtes non probabilistes peut être influencé par des facteurs de confusion, tels que certaines variables de l'étude pendant la collecte des données, comme ce que nous constatons dans les enquêtes probabilistes sur la non-réponse, ce qui correspond à la façon dont la littérature actuelle sur les enquêtes non probabilistes a évolué dans le traitement de ces questions. Troisièmement, toute avancée méthodologique dans le traitement des « modèles dits de données manquantes non au hasard » pour les enquêtes non probabilistes nécessiterait les fondements et la compréhension approfondie établis selon les hypothèses A1 à A4.

Michael A. Bailey a par ailleurs affirmé que « bien que les violations de l'hypothèse des données manquantes au hasard constituent un problème dans l'échantillonnage probabiliste (découlant de la non-réponse chez les personnes avec lesquelles on a communiqué au hasard), les violations de l'hypothèse des données manquantes au hasard sont plus graves dans un monde non probabiliste ». Je suis tout à fait d'accord avec cette analyse. De fait, les violations de l'hypothèse de positivité A2 sont aussi graves que les violations de l'« hypothèse de données manquantes au hasard » A1, et les deux sont interreliées. Les violations de l'hypothèse A2 impliquent que $\pi_i^A = P(i \in S_A \,|\, \mathbf{x}_i, y_i) = 0$ pour certaines unités de la population cible, ce qui entraîne un problème de sous-dénombrement qui est aussi connu que la non-réponse. S'il y a violation d'A2, mais qu'A1 se vérifie, on croit souvent que les estimateurs de prédiction fondés sur un modèle peuvent atténuer les biais dus au sous-dénombrement. Dans l'hypothèse A1, la variable de l'indicateur d'inclusion de l'échantillon R et la variable étudiée y sont conditionnellement indépendantes étant donné $\mathbf{x}$, ce qui signifie que :

$$E(y_i \,|\, \mathbf{x}_i, R_i = 1) \;=\; E(y_i \,|\, \mathbf{x}_i). \tag{1}$$

Il s'ensuit qu'un modèle de prédiction valide $y \,|\, \mathbf{x}$ peut être construit au moyen des données observées $\{(y_i, \mathbf{x}_i), i \in S_A\}$ (c'est-à-dire des unités avec $R_i = 1$). Malheureusement, l'équation (1) exige implicitement $P(R_i = 1 \,|\, \mathbf{x}_i) > 0,$ et les estimateurs fondés sur des prédictions ne sont pas à l'abri des biais potentiels dus au sous-dénombrement. L'appel de Michael A. Bailey en faveur d'un « cadre qui englobe la possibilité de violations de l'hypothèse de données manquantes au hasard » est conforme à certains travaux recherche actuels sur le traitement du sous-dénombrement et des mécanismes de participation « non ignorables » pour les échantillons d'enquête non probabilistes. Voir notamment Chen, Li et Wu (2023), Cho, Kim et Qiu (2022) et Yuan, Li et Wu (2022). En bref, pour obtenir des inférences statistiques valides à partir de ces scénarios, il faut des données externes, comme un échantillon de validation, ou des hypothèses supplémentaires, comme l'existence de variables instrumentales.

Je suis exactement sur la même longueur d'onde que Michael A. Bailey à propos de l'étiquette « manquant au hasard », puisque le terme pourrait être confondu avec « manquant aléatoirement » (Wu et Thompson, 2020, page 195). Le terme « ignorable » est également un mauvais choix de terminologie pour les données manquantes et la littérature sur l'inférence causale, car l'analyste des données ne peut certainement pas les ignorer (Rivers, 2007). J'utilise le terme courant « scores de propension » pour les échantillons non probabilistes, tandis que plusieurs autres auteurs lui préfèrent « probabilités de participation », y compris Beaumont (2020) et Rao (2021).

Michael R. Elliott

Michael R. Elliott a traité de plusieurs questions en utilisant des documents supplémentaires et une liste de références plus longue. Il s'agit d'ajouts importants au sujet actuel, en particulier les examens « d'autres approches pour combiner les données tirées d'enquêtes probabilistes et celles d'enquêtes non probabilistes » et l'analyse de sensibilité sur des « hypothèses non vérifiables ».

Les discussions d'Elliott concernant les distinctions entre les paramètres descriptifs et les paramètres analytiques ainsi que la pondération par rapport à la modélisation ont soulevé la question délicate de l'efficacité des estimateurs de pondération par l'inverse de la propension (PIP) dans la pratique. On sait, pour les échantillons d'enquête probabiliste, que l'estimateur de Horvitz-Thompson pondéré par la probabilité inverse du total de population T_y est extrêmement inefficace (en termes de grande variance) quand les probabilités de sélection de l'échantillon π_i sont inégales, mais ont une très faible corrélation avec la variable étudiée y, bien que l'estimateur demeure sans biais dans de tels scénarios. L'exemple de l'éléphant de Basu (Basu, 1971) montrait un « cas convaincant » dans lequel l'estimateur de Horvitz-Thompson pondéré par la probabilité inverse et sans biais a échoué lamentablement, ce qui a entraîné le congédiement du statisticien du cirque. Les discussions sur la pondération par rapport à la modélisation, c'est-à-dire les estimateurs pondérés par la probabilité inverse comparativement aux estimateurs de la prédiction fondés sur un modèle pour les paramètres descriptifs de la population, sont très pertinentes pour les développements théoriques et les applications pratiques. En tant que statisticiens, notre travail de traitement des échantillons d'enquête non probabilistes pourrait être très incertain si nous n'élaborons pas des lignes directrices et des outils de diagnostic solides nous permettant de choisir des méthodes adéquates selon l'ensemble de données à notre disposition et les problèmes d'inférence.

Michael R. Elliott fait écho à mon appel à réaliser quelques enquêtes probabilistes à grande échelle comportant une information riche sur les variables auxiliaires en affirmant « qu'il est de plus en plus essentiel de mettre en place des enquêtes probabilistes structurées et idéalement financées par le gouvernement pour les collectes de données courantes ». Ses commentaires à propos des nouveaux domaines de recherche sur les questions de protection de la vie privée et de confidentialité en raison du besoin de microdonnées dans le contexte de l'analyse d'échantillons d'enquêtes non probabilistes constituent un appel visionnaire, qui mérite une plus grande attention de la part du milieu de la recherche.

Zhonglei Wang et Jae Kwang Kim

Zhonglei Wang et Jae Kwang Kim présentent deux nouvelles approches de l'estimation fondée sur le score de propension : l'une utilise ce qu'on appelle la projection de renseignements au moyen d'un modèle de ratio de densité et l'autre emploie le calage uniforme de fonctions dans un espace de Hilbert à noyau reproduisant (RKHS, de l'appellation anglaise *Reproducing Kernel Hilbert Space*). Ces méthodes sont de nouvelles aventures dans notre domaine, et Kim et ses collaborateurs ont l'expérience et la puissance analytique nécessaires pour faire avancer la recherche dans ce sens.

Le point de départ des deux approches est l'équation suivante, qui relie les scores de propension aux ratios de densité :

$$\frac{1}{P(R_i=1\,|\,\mathbf{x}_i, y_i)} = 1 + \frac{P(R_i=0)}{P(R_i=1)}\frac{f_0(\mathbf{x}_i, y_i)}{f_1(\mathbf{x}_i, y_i)}.$$

Les scores de propension $\pi_i^A = P(R_i=1\,|\,\mathbf{x}_i, y_i)$ nécessitent seulement le modèle sur $R_i=1$ étant donné $\mathbf{x}_i$ et y_i. Toutefois, la justification de l'équation donnée ci-dessus exige un cadre de randomisation conjoint comprenant à la fois le modèle q pour les scores de propension et le modèle de superpopulation ξ sur $(\mathbf{x}, y)$. Du point de vue de la cohérence concernant l'estimateur final de la moyenne de population finie de y, le cadre conjoint impose très peu de restrictions si les ratios de densité sont modélisés de façon non paramétrique. Cette méthode a une incidence considérable sur la variance et l'estimation de la variance. La variance d'un estimateur dans un cadre de randomisation conjoint comporte plus d'une composante, et l'estimation de la variance entraîne d'autres complications en cas de procédures non paramétriques. Les comparaisons d'efficacité entre les méthodes proposées et certaines des méthodes existantes doivent être effectuées dans des configurations appropriées. J'ai hâte de suivre les progrès à venir à partir des méthodes qui ont été avancées.

Sharon L. Lohr

La discussion approfondie de Sharon L. Lohr sur les outils de diagnostic aux fins d'évaluation des hypothèses du modèle est très précieuse pour le sujet qui nous intéresse. Ses analyses des idées et des méthodes existantes et les adaptations au contexte actuel mettent en évidence les questions apparemment différentes, mais profondément liées, auxquelles font face les échantillons d'enquête probabilistes et non probabilistes. L'une de ces questions est le problème du sous-dénombrement (c'est-à-dire les violations de l'hypothèse A2) et la conjugaison des hypothèses A1 et A2. Sharon L. Lohr était à juste titre préoccupée par les estimateurs fondés sur les prédictions dans lesquels le modèle de prédiction de y étant donné $\mathbf{x}$ est construit à partir de l'échantillon non probabiliste S_A et l'estimateur d'imputation de masse est calculé au moyen du $\mathbf{x}$ observé dans l'échantillon probabiliste de référence, un scénario dans lequel chacune des deux hypothèses A1 et A2 n'est pas autonome. Le problème de sous-dénombrement est un exemple où « les procédures de l'ère spatiale ne sauveront pas les données de l'âge de pierre ». Sharon L. Lohr a préconisé de « prendre un petit échantillon probabiliste pour analyser les hypothèses », ce qui est nécessaire en théorie puisqu'il faut des échantillons de validation pour des méthodes rigoureusement défendables dans certains scénarios. Cependant, l'élaboration de stratégies de compromis avec les sources de données disponibles, bien qu'il s'agisse d'une approche plus attrayante, est plus difficile à mettre en pratique.

L'observation de Sharon L. Lohr selon laquelle « les échantillons non probabilistes peuvent améliorer l'équité des données » est importante, car l'inclusion d'unités de groupes susceptibles d'être invisibles dans les échantillons probabilistes peut être favorisée relativement facilement pour les échantillons non

probabilistes. Elle observe par ailleurs que « les groupes historiquement défavorisés pourraient être sous-représentés dans toutes les sources de données, y compris dans les échantillons non probabilistes ». Aborder la question de l'équité des données en présence d'échantillons d'enquête non probabiliste présente à la fois des possibilités et des défis.

Il est difficile de répondre à la question de Sharon L. Lohr « quand faut-il utiliser des échantillons non probabilistes ? ». Cette même question peut être posée à propos de toute méthode statistique. Il semblerait que nous ne remettions pas toujours en question la validité des méthodes et l'utilité des résultats dans de nombreux autres scénarios, car nous sommes convaincus que les hypothèses requises semblent raisonnables. Pour les échantillons non probabilistes, nous nous retrouvons dans une situation de vulnérabilité plus importante en ce qui a trait aux hypothèses, et les évaluations et les diagnostics de ces hypothèses sont plus difficiles que les cas présentant des expériences contrôlées ou des données plus structurées. De ce point de vue, l'analyse approfondie de Sharon L. Lohr sur l'évaluation des hypothèses doit être lue avec attention et reconnaissance. Dans la pratique, un examen scrupuleux de « l'étape d'élaboration du plan » est crucial pour renforcer la confiance à l'égard des hypothèses, si cette étape peut être conçue avant la collecte des données sur des variables qui pourraient être liées au comportement de participation et que l'on inclut ces variables dans l'échantillon en étudiant davantage les sources de données existantes contenant ces variables.

Xiao-Li Meng

L'exposé de Xiao-Li Meng, intitulé « La miniaturisation de la corrélation due à un défaut des données : une stratégie polyvalente de traitement des échantillons non probabilistes »), devrait être un document de travail autonome. Xiao-Li Meng a passé en revue un certain nombre de problèmes dans l'estimation d'une moyenne de population finie en cas d'échantillon non probabiliste, et a examiné des stratégies et des orientations permettant de construire un estimateur approximativement sans biais au moyen d'un concept central dit de *corrélation due à un défaut des données (cdd)*. Ses éléments de discussion fascinants invitent à la réflexion et ils susciteront certainement davantage de discussions et de projets de recherche sur les implications de la *cdd*. J'aimerais saisir l'occasion pour commenter brièvement la relation de la *cdd* avec trois concepts de base de l'échantillonnage probabiliste, soit la *stratégie d'échantillonnage*, le *sous-dénombrement* et *l'estimation assistée par modèle*. Il ne s'agit pas de nostalgie du bon vieux temps où l'échantillonnage probabiliste était la norme par excellence, mais plutôt d'une appréciation de l'évolution de la recherche sur l'échantillonnage d'enquête et de l'utilité potentielle de la *cdd* dans le traitement des échantillons d'enquête non probabilistes.

Le terme *stratégie d'échantillonnage* désigne la paire constituée par le plan de sondage et la méthode d'estimation (Thompson, 1997, section 2.4; Rao, 2005, section 3.1). Ces deux composantes qui vont de pair constituent la colonne vertébrale de la théorie classique de l'échantillonnage probabiliste. Aux fins d'estimation du total de population T_y de la variable étudiée y au moyen d'un échantillon probabiliste S avec les probabilités d'inclusion de premier ordre π_i, l'estimateur de Horvitz-Thompson $\hat{T}_{\text{унг}} = \sum_{i \in S} d_i y_i$

avec le poids $d_i = \pi_i^{-1}$ est l'estimateur sans biais unique au sein d'une classe d'estimateurs linéaires (Wu et Thompson, 2020). L'argument théorique du résultat est simple en raison des probabilités d'inclusion connues π_i selon le plan d'échantillonnage donné. Si l'on utilise la notation de Meng, la *cdd* comprend trois variables, à savoir la variable étudiée G, la variable de poids W, l'indicateur d'inclusion de l'échantillon R, et elle est définie comme le coefficient de corrélation de la population finie entre $\tilde{R} = RW$ et G. La *cdd* pose implicitement R et W comme une paire inséparable pour toute *stratégie d'inférence*, R correspondant au « plan » inconnu et W à la « méthode d'estimation ». En ayant recours au « plan » inconnu caractérisé par les « probabilités divines » inconnues π_I pour l'échantillon non probabiliste, Meng a montré par son équation (3.3), que $W_I \propto \pi_I^{-1}$ est essentiellement une condition requise pour une estimation sans biais de $\bar{G}$ si rien n'est supposé sur le modèle de régression des résultats. Le résultat permet de justifier l'utilisation de l'estimateur de pondération par l'inverse de la propension (PIP) pour les échantillons non probabilistes comme seul choix raisonnable s'il n'y a pas de modèle de superpopulation de la variable étudiée.

Le problème du *sous-dénombrement* a été largement traité dans la littérature portant sur l'échantillonnage probabiliste. Pour les échantillons non probabilistes, la question est étroitement liée à la violation de l'hypothèse de positivité A2, abordée dans la section 7.2 de mon article et mes commentaires sur les discussions de Bailey, Elliott et Lohr. Des précisions supplémentaires sont données dans Chen et coll. (2023). Soit $U = U_0 \cup U_1$, où U_1 est la sous-population non couverte avec $\pi_i^A = P(R_i = 1 \,|\, \mathbf{x}_i, y_i) = 0$. Soit $N = N_0 + N_1$, où N_0 et N_1 sont respectivement les tailles des deux sous-populations U_0 et U_1. Supposons que Cov_I et $\mathrm{Cov}_I^{(0)}$ désignent respectivement la covariance par rapport à la distribution uniforme discrète sur U et U_0. On peut montrer que :

$$\mathrm{Cov}_I(\tilde{R}_I, G_I) = \omega_0 \left\{ \mathrm{Cov}_I^{(0)}(\tilde{R}_I, G_I) - \omega_1 (\bar{G}_1 - \bar{G}_0)\, \hat{N}_0 / N_0 \right\}, \tag{2}$$

où $\omega_k = N_k / N$ pour $k = 0, 1$, $\hat{N}_0 = \sum_{i \in S} W_i$, S est l'ensemble d'unités pour l'échantillon non probabiliste, et $\bar{G}_0$ et $\bar{G}_1$ sont respectivement les moyennes de population de U_0 et U_1 pour la variable étudiée G. L'équation (2) a deux conséquences immédiates. Premièrement, si la méthode d'estimation est valide dans le sens où la valeur de $\mathrm{Cov}_I^{(0)}(\tilde{R}_I, G_I)$ est petite, alors le biais de l'estimateur $\bar{G}_W$ dû au sous-dénombrement dépend de ω_1 (c'est-à-dire la taille de la sous-population non couverte U_1) et $\bar{G}_1 - \bar{G}_0$ (c'est-à-dire la différence entre U_0 et U_1), un énoncé qui a déjà été établi selon l'échantillonnage probabiliste. Deuxièmement, l'équation révèle un scénario de *contrepoids* potentiel : Un estimateur biaisé $\bar{G}_W$ pour la « moyenne de la population échantillonnée » $\bar{G}_0$ peut être moins biaisé pour la moyenne de la population cible $\bar{G}$ si $\mathrm{Cov}_I^{(0)}(\tilde{R}_I, G_I)$ et $\bar{G}_1 - \bar{G}_0$ ont le même signe plus ou moins.

Les discussions de Meng sur la quasi-randomisation ou la superpopulation au moyen de la *cdd* apportent une compréhension nettement plus fine de l'estimation doublement robuste. Historiquement, *l'estimation assistée par un modèle* a commencé à apparaître dans l'échantillonnage d'enquête au début des années 1970; cette méthode est dans le même esprit que la double robustesse. L'estimateur par la

différence généralisée de la moyenne de population $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$, comme le présentent Cassel, Särndal et Wretman (1976), est donné par :

$$\hat{\mu}_{y\text{GD}} = \frac{1}{N}\left\{\sum_{i\in S}\frac{y_i - c_i}{\pi_i} + \sum_{i=1}^{N} c_i\right\}, \tag{3}$$

où S est un échantillon probabiliste, les π_i sont les probabilités d'inclusion de premier ordre et $\{c_1, c_2, \ldots, c_N\}$ est une séquence arbitraire de nombres connus. L'estimateur $\hat{\mu}_{y\text{GD}}$ est exactement sans biais pour μ_y selon le plan d'échantillonnage probabiliste p pour toute séquence c_i; il est également sans biais par rapport au modèle si nous choisissons $c_i = m_i = E_\xi(y_i \mid \mathbf{x}_i)$. Cassel et coll. (1976) ont présenté un résultat théorique principal selon lequel le choix $c_i = m_i$ est optimal, ce qui conduit à une espérance minimale fondée sur un modèle de la variance fondée sur le plan $E_\xi\{V_p(\hat{\mu}_{y\text{GD}})\}$ quand le modèle a une certaine structure de variance. La première partie des résultats sur l'absence de biais est sous (p ou ξ); la deuxième partie sur l'optimalité se trouve sous (p et ξ). Il convient de souligner que l'estimateur $\hat{\mu}_{y\text{GD}}$ avec le choix $c_i = \hat{m}_i$ a exactement la même structure que l'estimateur doublement robuste dont il est abondamment question dans la littérature sur les données manquantes et l'inférence causale depuis les années 1990, les « probabilités divines » π_i étant inconnues et estimées dans ces derniers cas.

Dans la pratique, l'emploi de la *cdd* nécessite des renseignements supplémentaires tirés de la population. La proposition de Meng de créer une miniature représentative à partir d'un échantillon biaisé fait écho à la demande d'un échantillon de validation de petite taille, puisqu'un tel échantillon « peut (aussi) éliminer l'anxiété de nombreux praticiens et les erreurs qu'ils sont susceptibles de commettre parce qu'ils ne maîtrisent pas l'utilisation des poids ».

« Il n'existe pas d'échantillon probabiliste dans le monde réel » est probablement un énoncé défendable pour les populations humaines. Néanmoins, les échantillons probabilistes existent dans d'autres domaines, comme les enquêtes-entreprises et les enquêtes auprès des établissements, les enquêtes relatives à l'agriculture et à l'inventaire de ressources naturelles; voir Wu et Thompson (2020) pour en savoir plus. En revanche, pour les êtres humains, toutes les règles rigoureuses ou les procédures précises sont *presque certainement* une aspiration, mais non une prescription.

Bibliographie

Basu, D. (1971). An essay on the logical foundations of survey sampling. Part One. Dans *Foundations of Statistical Inference*, (Éds., V.P. Godambe et D.A. Sprott), Toronto, 203-242.

Beaumont, J.-F. (2020). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Techniques d'enquête*, 46, 1, 1-30. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2020001/article/00001-fra.pdf.

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

Chen, Y., Li, P. et Wu, C. (2023). Dealing with undercoverage for non-probability survey samples. *Techniques d'enquête*, à l'étude.

Cho, S., Kim, J.K. et Qiu, Y. (2022). *Multiple Bias Calibration for Valid Statistical Inference with Selection Bias*. Document de travail.

Rao, J.N.K. (2005). Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête*, 31, 2, 127-151. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005002/article/9040-fra.pdf.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rivers, D. (2007). Sampling for web surveys. Dans *Proceedings of the Survey Research Methods Section*, Joint Statistical Meetings, American Statistical Association, Alexandrie, Virginie, 1-26.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman & Hall, Londres.

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

Yuan, M., Li, P. et Wu, C. (2022). *Inference with Non-Ignorable Sample Inclusion for Non-Probability Survey Samples*. Document de travail.