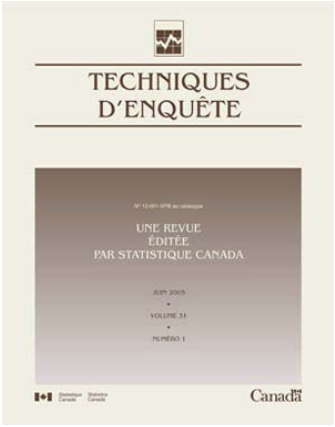
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Commentaires à propos de l'article de Rao et Fuller (2017)



par Chris Skinner

Date de diffusion : le 21 décembre 2017

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 STATCAN.infostats-infostats.STATCAN@canada.ca

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

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

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.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2017
Vol. 43, nº 2, pp. 197-199
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article de Rao et Fuller (2017)

Chris Skinner[1]

Résumé

Cette note de Chris Skinner présente une discussion de l'article « Théorie et méthodologie des enquêtes par sondage : orientations passées, présentes et futures » où J.N.K. Rao et Wayne A. Fuller partagent leur vision quant à l'évolution de la théorie et de la méthodologie des enquêtes par sondage au cours des 100 dernières années.

Mots-clés : Collecte des données; histoire de l'échantillonnage; échantillonnage probabiliste; inférence à partir d'enquêtes.

Cet article donne un excellent compte rendu de la théorie et des méthodes de sondage, en distillant de manière concise et élégante une énorme quantité de savoir sur la théorie et la pratique de la discipline. Je ne commenterai ni le passé ni le présent, mais présenterai, comme j'y ai été invité, certaines réflexions sur l'avenir. Je suis entièrement d'accord avec la dernière partie de l'article au sujet de l'avenir et considère que mes idées se superposent à celles des auteurs.

La présente discussion mettra en relief la perspective d'un Institut national de statistique (INS), et je prévois (et j'espère) que les INS joueront un rôle déterminant dans l'évolution de la méthodologie, même si l'environnement statistique connaît des changements, comme la gamme croissante d'organismes qui fournissent des données aux INS et/ou produisent eux-mêmes des statistiques.

Cibles inférentielles : À mon avis, les mêmes types de populations finies cibles descriptives (vues sous l'angle méthodologique) continueront de revêtir un intérêt fondamental. Les besoins analytiques retiendront également leur importance, mais la façon dont ils seront satisfaits dépendra de l'évolution des modalités d'accès aux données, dans le contexte, par exemple, des préoccupations concernant la confidentialité, ainsi que de l'impact de l'évolution de la pratique de la science des données, telle une plus grande importance accordée à la modélisation prédictive.

Sondages et autres sources de données : La nature et la portée des sources de données pertinentes représenteront un domaine de développement critique. Je ne pense pas que les enquêtes disparaîtront, car il existera toujours un très grand nombre de variables d'intérêt nécessitant une collecte directe de données. Mais je m'attends à ce que le sondage fasse de plus en plus partie intégrante d'un ensemble plus vaste de sources de données qui incluent les recensements, les données administratives et les « mégadonnées » (par exemple, Lohr et Raghunathan, 2017; Zhang, 2012). Le défi méthodologique consistera à intégrer efficacement une telle gamme de sources. Les différentes sources peuvent avoir plusieurs propriétaires et les modalités d'accès influeront considérablement sur la façon dont ces sources peuvent être intégrées. Je ne pense pas non plus que l'échantillonnage disparaîtra, car il sera nécessaire non seulement pour la collecte principale de données, mais aussi pour les enquêtes supplémentaires (voir plus bas) et la gestion des sources de mégadonnées.

1. Chris Skinner, London School of Economics and Political Science. Courriel : C.J.Skinner@lse.ac.uk.

Enquêtes supplémentaires : Le besoin d'échantillons d'enquête supplémentaires pour vérifier la validité ou améliorer l'inférence augmentera vraisemblablement. Les « enquêtes de référence » pourraient accroître le nombre d'échantillons non probabilistes (Elliot et Valliant, 2017); des sondages de couverture pourraient être nécessaires pour vérifier la présence à la fois de sous-dénombrement ou de surdénombrement, par exemple, dans les sources de données administratives, et pour corriger ce genre d'erreurs (Zhang, 2015); des enquêtes appariées au niveau de l'unité pourraient être nécessaires pour vérifier la présence d'une erreur de mesure dans les sources de données.

Non-réponse et échantillonnage : La non-réponse totale posera davantage problème et l'inférence devra inévitablement tenir compte de l'erreur de non-réponse, ainsi que de l'erreur d'échantillonnage. La principale difficulté consistera à éviter (réduire) le biais de sélection. Le recours à la randomisation en échantillonnage pour atteindre cet objectif et pour justifier certaines hypothèses de modélisation pourrait devenir une propriété au moins aussi importante de l'échantillonnage probabiliste que son utilisation pour l'inférence fondée sur le plan de sondage. Il pourrait devenir sensé de considérer des protocoles d'échantillonnage et de gestion de la non-réponse d'une manière plus intégrée, et l'étude de ce genre d'options pourrait permettre d'opter pour des protocoles d'échantillonnage qui comprennent des caractéristiques non probabilistes, à condition que la réduction du biais de sélection demeure l'objectif central. En ce qui concerne la réponse, les options multimode flexibles semblent vraisemblablement être les options naturelles à prendre en considération. La nature des sources de données auxiliaires et les considérations concernant l'estimation doivent aussi, évidemment, être prises en compte sérieusement lorsqu'on évalue simultanément les options d'échantillonnage et de gestion de la non-réponse.

Méthodes et théorie d'estimation : Les méthodes d'estimation évolueront afin de tirer parti de nouvelles sortes de relations statistiques dans les sources de données et entre celles-ci, à la fois pour tenir compte des effets de biais de sélection possibles et pour accroître l'efficacité. De nombreux problèmes d'estimation peuvent être formulés en fonction des variables étudiées Y, qui ne peuvent être obtenues que sur des échantillons sélectifs, et des variables explicatives X, pour lesquelles peuvent être construites de grandes bases de données dont la couverture est proche de 100 %. La construction de ce genre de bases de données pourrait être un objectif clé des INS dans le contexte tant des statistiques sur les entreprises que des statistiques sociales. Dans le second cas, cet objectif pourrait être aligné avec les développements concernant le recensement de la population, lesquels comprennent, par exemple, l'utilisation de sources de données administratives (Skinner, 2017). Dans de telles conditions, une approche générale de l'estimation pourrait combiner les distributions de X au niveau de la population et les distributions conditionnelles de Y sachant X tirées des sources d'échantillons sélectifs sous des hypothèses similaires à celles des données manquant au hasard. L'importance des considérations temporelles, dont les avantages de l'emprunt d'information à diverses périodes, augmentera vraisemblablement, et il sera possible d'exploiter les possibilités qu'offrent les sources de données administratives qui sont habituellement longitudinales. Les méthodes de calage existantes et l'estimation sur petits domaines fondée sur le modèle de Fay-Herriot continueront d'être utilisées en tant que sources appariées à un niveau agrégé. L'appariement au niveau de l'individu ou de l'unité fondée sur les coordonnées GPS (par exemple, immeuble ou adresse) pourrait aboutir à d'autres méthodes (par exemple, Lohr et Raghunathan, 2017). La théorie des sondages, y compris les méthodes de prédiction fondées sur un modèle et les méthodes d'estimation sur petits domaines, continuera de jouer un

rôle essentiel. La théorie des données manquantes offre un cadre naturel pour le traitement des sources de données intégrées, et je m'attends à une confluence croissante entre la théorie de l'échantillonnage et celle des données manquantes. Le traitement des erreurs d'appariement et des erreurs de mesure, par exemple attribuables à des différences de mesure entre les sources et les modes de collecte des données, sera également important.

Évaluation de la qualité et estimation de l'exactitude : Vu que les contraintes budgétaires persisteront, il sera essentiel de faire valoir l'importance de normes élevées de qualité auprès des utilisateurs des produits statistiques si l'on veut éviter que les sondages de haute qualité soient remplacés par des solutions bon marché et non fiables. À cet égard, il serait utile de renforcer le rôle d'évaluateurs de la qualité des organes nationaux établis pour surveiller et accroître la confiance du public dans les produits statistiques, surtout si le nombre et la diversité des fournisseurs de ces produits augmentent. Plus précisément, l'évaluation de l'exactitude sera essentielle. Les méthodes classiques d'estimation de la variance pourraient jouer un rôle et être étendues afin de saisir des sources plus vastes de variation, par exemple, en élargissant la définition des répliques dans les méthodes de rééchantillonnage. Cependant, étant donné l'usage croissant de l'inférence fondée sur un modèle, l'évaluation de l'exactitude devra aussi englober l'évaluation de l'effet des écarts par rapport aux hypothèses dans les méthodes d'estimation. Des approches telles que la vérification du modèle, l'application de diagnostics et l'analyse de sensibilité prendront vraisemblablement de l'importance.

Bibliographie

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

Lohr, S.L., et Raghunathan, T.E. (2017). Combining survey data and other data sources. *Statistical Science*, 32, 293-312.

Skinner, C.J. (2018). Issues and challenges in census taking. *Annual Review of Statistics and its Application*, Volume 5.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.

Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, 31, 381-396.