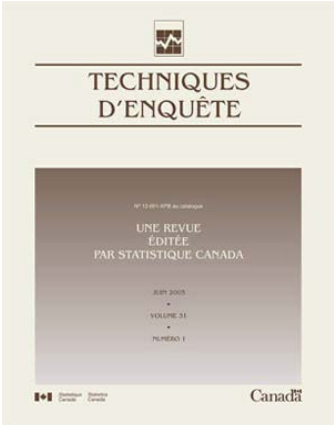
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

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



par Danny Pfeffermann

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. 177-182
Statistique Canada, nº 12-001-X au catalogue

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

Danny Pfeffermann[1]

Résumé

Cette note de Danny Pfeffermann 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.

C'est avec plaisir que je reconnais avoir invité J.N.K. Rao et Wayne Fuller à faire un exposé lors d'une séance spéciale parrainée par l'Association internationale des statisticiens d'enquêtes (AISE) à l'assemblée de l'Institut international de statistique (IIS) qui s'est tenue à Rio de Janeiro en 2015. Toutefois, le mérite revient au professeur Vijay Nair, président de l'IIS à l'époque, qui a instauré ce genre de séance sollicitée pour toutes les sections de l'IIS. Ainsi, quand j'ai su qu'il était possible d'organiser une telle séance, j'ai tout de suite pensé qu'il serait naturel de demander à Rao et Fuller, les deux rois incontestés de la théorie et de la méthodologie modernes des sondages, de donner un exposé sur les orientations passées, présentes et futures de cette discipline, et par bonheur, ils ont accepté d'emblée. Honnêtement, je ne m'attendais pas à leur accord immédiat; en effet, avant de les inviter, j'avais dressé une liste d'arguments en vue de les convaincre, mais comme nous le savons tous, les rois ont des devoirs envers leurs concitoyens. Ce que j'ignorais à l'époque, c'est que, deux ans plus tard, je serais invité à discuter d'un article fondé sur leur exposé, une invitation que j'ai acceptée avec gratitude.

Dans la discussion qui suit, je me concentrerai principalement sur la troisième partie de l'article, l'avenir des sondages, sujet dont je me préoccupe de plus en plus, depuis le tournant pris par ma carrière il y a quatre ans lorsque je suis devenu statisticien national et directeur du bureau central des statistiques d'Israël (ICBS). Naturellement, ma discussion s'appuie sur mon expérience en Israël, mais j'ai de bonnes raisons de croire qu'elle représente en grande partie ce qui se passe également dans d'autres pays.

À la section intitulée « L'avenir », Rao et Fuller (ci-après désignés « RF », comme mon autre roi, Roger Federer) donnent une longue liste d'éléments, qui, selon eux, domineront l'avenir des sondages. Je suis entièrement d'accord avec cette liste, mais avec une réserve. Nombre des éléments de cette liste sont déjà prédominants aujourd'hui. Les budgets sont déjà restreints et les demandes de produits augmentent constamment, non seulement les demandes intérieures faites dans les pays, mais aussi celles des organisations internationales, dont les Nations Unies, l'OCDE, Eurostat, le FMI et la Banque mondiale. Les demandes de statistiques de ces organisations sont souvent similaires et parfois même chevauchantes, mais elles doivent être satisfaites sous différentes formes et à différents moments, couvrant différentes périodes, créant ainsi une surcharge de travail pour les bureaux nationaux de la statistique (BNS). L'utilisation de

1. Danny Pfeffermann, Statisticien national et Directeur du Central Bureau of Statistics, Israël, Hebrew University of Jerusalem, Israël; Southampton Statistical Sciences Research Institute, Royaume-Uni. Courriel : MsDanny@cbs.gov.il.

données administratives fait déjà l'objet de travaux dans les BNS du monde entier. Dans les pays scandinaves, les recensements de population sont fondés uniquement sur des données administratives, et de nombreux autres pays européens ainsi qu'Israël investissent beaucoup de ressources dans la création de bases de données administratives fiables qui remplaceront leurs recensements dans la prochaine génération de recensements, qui seront réalisés autour de 2030. Dans ces pays, les recensements actuels s'appuient déjà fortement sur des données administratives, mais requièrent encore des échantillons pour corriger le sous- ou le surdénombrement. L'usage de données administratives soulève une question juridique importante ayant trait à l'accès à ces données. Les organismes publics et privés refusent simplement de transférer les données. En Israël, la personne occupant le poste de statisticien national est autorisée par la loi à obtenir tout ensemble de données qu'il ou qu'elle juge important pour la production des statistiques officielles, mais certains organismes maintiennent qu'ils se sont engagés à l'égard de leurs clients à ne transférer aucune donnée privée. Je sais que d'autres pays sont confrontés à un conflit similaire. Arriverons-nous à favoriser une culture de partage des données dans l'avenir ? À l'heure actuelle, le défi semble colossal.

Comme il est mentionné dans l'article, la collecte de données par interview téléphonique ou sur place devient de plus en plus difficile, ce qui se traduit par des taux plus élevés de non-réponse. La situation est même pire pour les enquêtes auprès des entreprises, parce que, très souvent, on demande aux mêmes établissements (grands pour la plupart) de participer annuellement à de nombreuses enquêtes (parfois sept et plus). C'est ce qui se passe également en Israël, mais contrairement à d'autres pays, nous arrivons encore à maintenir des taux de réponse raisonnables (autour de 70 %), car la participation à la plupart de nos enquêtes est obligatoire. Diverses approches peuvent être adoptées pour faire face à ce problème. La première consiste à élaborer de nouvelles procédures d'échantillonnage pour alléger le fardeau de réponse. Plusieurs procédures fondées sur les nombres aléatoires permanents sont déjà appliquées, mais des méthodes plus pointues doivent être établies, même si les établissements ou entreprises uniques, ainsi que les gros établissements ou entreprises, n'en bénéficieront qu'à peine, voire jamais. Cela nous pousse à trouver de meilleures méthodes d'imputation au moyen de données antérieures et de données observées pour d'autres unités échantillonnées, quoique l'imputation pour les unités relativement grandes puisse être pratiquement impossible. Un moyen idéal d'alléger le fardeau de réponse consisterait à recueillir les microdonnées telles quelles et les métadonnées pertinentes auprès des entreprises, puis à les traiter au BNS. Cette approche nécessitera évidemment la coopération des entreprises et de plus grandes compétences en science des données dans leurs bureaux. Les chances d'une telle coopération sont-elles raisonnables ? À mon âge, j'ai le droit d'être sceptique, mais en y réfléchissant, moyennant des accords de protection des données appropriés, pourquoi pas ?

À la section 3.3, RF discutent de plusieurs méthodes d'imputation, en énumérant des références clés. Autant que je sache, une hypothèse sous-jacente à toutes ces méthodes est l'accès à des données externes ou internes (antérieures et autres données observées pour les mêmes unités échantillonnées) qui expliquent entièrement les données manquantes. Cependant, selon mon expérience à l'ICBS, cela n'est souvent pas le cas; en outre, dans de nombreuses enquêtes, la non-réponse ne correspond pas à des données manquant au hasard et est donc une non-réponse NMAR (pour *not missing at random*). Le traitement de la non-réponse NMAR est un problème très difficile pour la simple raison que les variables d'intérêt cibles ne sont pas

observées pour les non-répondants, ce qui oblige à formuler des hypothèses au sujet du mécanisme de non-réponse sous forme de modèles statistiques, en n'ayant que des possibilités limitées de les tester. La présente discussion n'est pas censée mettre en relief mes propres travaux de recherche et ceux de mes collègues, mais je voudrais mentionner une approche proposée dans Pfeffermann et Sikov (2011), Feder et Pfeffermann (2015) et Pfeffermann (2017) pour traiter la non-réponse NMAR et qui permet de tester le modèle de réponse, du moins partiellement. L'idée qui sous-tend cette approche consiste à supposer un modèle (paramétrique ou non paramétrique) pour les données de population, et un modèle paramétrique pour le mécanisme de réponse, puis à vérifier si le modèle implicite tient pour les données observées, en utilisant des procédures de vérification de modèle classiques. Voir aussi Riddles, Kim et Im (2016) pour une condition d'identifiabilité importante pour le modèle vérifié pour les données observées, et Sverchkov et Pfeffermann (2017) pour l'application de cette condition dans le contexte de l'estimation sur petits domaines sous non-réponse NMAR.

Deux autres approches pour traiter la non-réponse nécessiteront une étude plus approfondie au cours des années à venir, à savoir l'utilisation des plans de collecte de données adaptatifs et l'acceptation de modes de collecte de données de rechange. L'idée générale qui sous-tend le plan de collecte de données adaptatif est l'utilisation d'information auxiliaire provenant de données de registre et/ou des observations faites par les intervieweurs, afin de personnaliser le plan de collecte de manière à optimiser les taux de réponse et, conséquemment, à réduire le biais de non-réponse. Pour obtenir des précisions et des exemples, consulter l'ouvrage publié récemment par Schouten, Peytchev et Wagner (2017), ainsi que les nombreuses références qu'il contient. RF discutent brièvement des méthodes de collecte des données et de leurs effets sur l'inférence. Nous connaissons tous bien les modes traditionnels ainsi que plus modernes de collecte des données, à savoir l'interview sur place, l'interview par téléphone (téléphone mobile), par la poste (courrier électronique) et, aujourd'hui, par Internet, qui est le mode le moins coûteux et donc, privilégié de collecte des données, même s'il requiert encore l'envoi d'une adresse Internet et d'un mot de passe aux unités échantillonnées. Afin d'accroître les taux de réponse, les organismes d'enquêtes ont tendance à attribuer différents modes de collecte à différentes unités échantillonnées ou, plus généralement, à laisser les unités échantillonnées choisir leur mode de réponse préféré. On parle alors d'enquêtes à mode mixte. Une autre variante des enquêtes à mode mixte, souvent appliquée en pratique, consiste à offrir séquentiellement différents modes de réponse aux unités qui n'ont pas répondu par le mode précédent. Cependant, ces procédures posent un problème d'inférence en raison d'effets de mode éventuels découlant d'un effet de sélection (l'effet de différences entre les caractéristiques des répondants privilégiant différents modes de réponse et, par conséquent, de différences éventuelles dans les valeurs des variables étudiées), et un effet de mesure (l'effet de réponses différentes fournies par un même membre de l'échantillon selon le mode de réponse). Dans Pfeffermann (2015), je passe brièvement en revue les méthodes disponibles pour résoudre ce problème et propose une autre approche, mais les travaux de recherche doivent se poursuivre à ce sujet. Notons qu'un effet de mode (de mesure) peut exister même si un seul mode de réponse est offert.

Un autre aspect de la collecte des données, qui n'est pas évoqué dans l'article, mais qui est très fréquent en pratique, par exemple dans les enquêtes sur la population active, les enquêtes sur la santé et même les recensements modernes, est le recours à la réponse par personne interposée, où un membre du ménage

fournit l'information pour tous les autres membres du ménage. Ce genre d'enquête peut poser un problème éthique en ce sens que l'on ne demande pas aux membres du ménage non interrogés de consentir à ce que l'information à leur sujet soit fournie par le membre qui est interrogé. D'un point de vue statistique, l'utilisation d'enquêtes par personne interposée peut accroître la possibilité d'erreurs de mesure (corrélées). Le membre du ménage interrogé connaît-il l'état de santé de tous les autres membres du ménage, ou sait-il s'ils étaient à la recherche de travail durant la semaine qui a précédé l'interview ? Et s'il se trompe au sujet d'un membre du ménage, ne pourrait-il pas se tromper au sujet d'autres membres ? Ce problème a-t-il été étudié de manière systématique dans la littérature et des solutions ont-elles été proposées ? Pour empirer encore la situation, qu'en est-il de l'interaction entre les erreurs de mesure et la non-réponse ?

La dernière question, mais sans doute la principale lorsqu'on discute de l'avenir des enquêtes par sondage, est celle de l'utilisation éventuelle de mégadonnées en tant qu'alternative aux enquêtes traditionnelles fondées sur des échantillons probabilistes. Ces dernières années, j'ai moi-même donné des exposés sur cette question et mes principales réflexions sont déjà résumées dans Pfeffermann (2015). J'en répéterai certaines ici, dans la perspective d'un statisticien national, préoccupé par la production de statistiques officielles.

RF déclarent correctement que le terme mégadonnées n'est pas bien défini, mais mentionnent les données des médias sociaux comme exemple de ce genre de données. Je pense effectivement que c'est un bon exemple, qui, cependant, illustre aussi les problèmes que pose le traitement de ce genre de données. Ces données sont généralement diverses et non structurées, apparaissent irrégulièrement et peuvent même cesser d'exister. (Néanmoins, tous les types de mégadonnées ne se comportent pas comme cela, et d'autres types pourraient être considérés simplement comme des mégaversions de ce qui est habituellement appelé données administratives.) RF ajoutent que les données provenant des réseaux sociaux intéressent les spécialistes des sciences sociales. De nouveau, je suis d'accord avec eux, mais les BNS devraient-ils publier des estimations des indices sociaux obtenus à partir des réseaux sociaux ? Peut-être que oui, mais quelle population ces indices représenteront-ils ? RF mentionnent aussi que les mégadonnées devraient servir de données auxiliaires. Ils n'entrent pas dans les détails, mais je suppose qu'ils pensent à leur utilisation comme covariables dans l'inférence assistée par modèle ou dépendante d'un modèle, de la même manière que les données administratives. C'est une évidence et de nombreux exemples ont été publiés dans la littérature. Mais la question n'est pas aussi simple qu'elle le paraît, en raison de tous les obstacles informatiques qu'il faudra surmonter avant que les données soient prêtes à l'emploi, y compris le couplage d'enregistrements, s'il faut apparier des mégadonnées et des microdonnées d'enquêtes. Manifestement, il faudra ajuster et tester des modèles appropriés.

Néanmoins, le vrai défi, selon moi, tient à l'usage possible des mégadonnées pour remplacer des enquêtes par sondage classiques. Pour calculer l'IPC, il est indiscutable qu'il est bien plus efficace et moins coûteux d'obtenir les prix de vente (et les quantités) électroniquement, directement auprès des magasins, au lieu d'envoyer des enquêteurs pour faire le relevé des prix. Mais des indices des prix sont souvent demandés pour des sous-groupes de la population, définis selon l'âge, l'origine, etc., et les mégadonnées obtenues électroniquement ne contiennent en général pas d'information démographique. Arriverons-nous à utiliser

l'information des cartes de crédit pour relier les acheteurs aux achats ? Les sociétés de cartes de crédit fourniront-elles l'information requise ? Comment cela se fera-t-il ? Et qu'en est-il des problèmes de couverture des mégadonnées ? Les opinions exprimées sur les réseaux sociaux représentent-elles les opinions du grand public ? Les annonces d'offre d'emploi sur Internet peuvent-elles remplacer les enquêtes auprès des entreprises au sujet des emplois vacants ? Lors d'une conférence qui a eu lieu plus tôt cette année pour célébrer le 80e anniversaire de Jon Rao, le professeur Jae Kim a envisagé trois procédures possibles pour corriger le biais de couverture possible des mégadonnées. À l'assemblée de l'IIS tenue à Marrakech, j'ai proposé une quatrième procédure. Les quatre procédures paraissent raisonnables, mais il est clair que beaucoup d'autres études théoriques et appliquées seront nécessaires avant que l'on puisse effectivement recommander l'usage de n'importe laquelle d'entre elles.

Une des procédures proposées par Kim requiert l'appariement des mégadonnées à un échantillon approprié, afin d'estimer les probabilités d'inclusion dans les mégadonnées. Il s'agit là d'un très bel exemple de combinaison de mégadonnées avec des données d'enquêtes. Di Zio, Zhang et De Waal (2017) discutent d'une autre utilisation de l'échantillonnage classique, pour « assister » la construction et la validation de modèles. À cet égard, une question importante est de savoir comment intégrer les erreurs d'échantillonnage avec les erreurs de modélisation dans l'inférence subséquente. Devrait-on évaluer l'estimateur pour mégadonnées dans une perspective fondée sur un modèle uniquement, quand la construction du modèle fait appel à des données d'échantillon susceptibles de présenter des erreurs d'échantillonnage ? Mise en garde : lorsque l'on combine des mégadonnées et des données d'enquêtes, on devrait vérifier soigneusement la définition des variables qui figurent dans les deux ensembles de données, même si elles semblent mesurer le même phénomène. Le chômage dans un ensemble de mégadonnées pourrait avoir une définition très différente de celle de l'Oganisation internationale du Travail (OIT) adoptée dans les enquêtes sur la population active. Un autre aspect de la combinaison éventuelle de mégadonnées avec des données d'enquêtes, et encore davantage, de l'usage de mégadonnées seulement, est l'élaboration de nouveaux algorithmes d'échantillonnage à appliquer aux mégadonnées. L'échantillonnage des mégadonnées permet de réduire l'espace de stockage, aide à protéger les renseignements personnels et à prévenir la divulgation, et produit des ensembles de données maniables sur lesquels des algorithmes peuvent être exécutés pour ajuster des modèles et produire des estimations. Toutefois, l'échantillonnage aléatoire applicable à des mégadonnées versatiles et dynamiques est manifestement différent de l'échantillonnage pour les populations finies et requiert de nouveaux algorithmes. L'échantillonnage par réseau des mégadonnées remplacera-t-il l'échantillonnage en population finie classique ?

Pour conclure cette brève discussion, je dirai qu'à mon avis, les enquêtes par sondage classiques continueront de jouer un rôle essentiel dans l'avenir prévisible. Cependant, comme le souligne Marker (2017), l'existence de mégadonnées a modifié les attentes en matière d'actualité des données et les BNS devront trouver un moyen de réaliser les enquêtes et les recensements plus rapidement, sinon les utilisateurs s'appuieront sur les mégadonnées disponibles sans comprendre ce qu'ils y perdent.

Bibliographie

Di Zio, M., Zhang, L.C. et De Waal, T. (2017). Statistical methods for combining multiple sources of administrative and survey data. *The Survey Statistician*, 17-26.

Feder, M., et Pfeffermann, D. (2015). Statistical inference under non-ignorable sampling and nonresponse-an empirical likelihood approach. Southampton Statistical Sciences Research Institute, University of Southampton, Royaume-Uni. http://eprints.soton.ac.uk/id/eprint/378245.

Marker, D. (2017). How have National Statistical Institutes improved quality in the last 25 years? *Statistical Journal of the IAOS*, 1, 1-11.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *The Journal of Survey Statistics and Methodology* (*JSSAM*), 3, 425-483.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples. A unified approach. *Calcutta Statistical Association (CSA) Bulletin*, 69, 35-63.

Pfeffermann, D., et Sikov, A. (2011). Estimation and imputation under non-ignorable non-response with missing covariate information. *Journal of Official Statistics*, 27, 181-209.

Riddles, K.M., Kim, J.K. et Im, J. (2016). A propensity-scoreadjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4, 215-245.

Schouten, B., Peytchev, A. et Wagner, J. (2017). *Adaptive Survey Design*. Chapman&Hall/CRC Statistics in the Social and Behavioral Sciences.

Sverchkov, M., et Pfeffermann, D. (2017). Small area estimation under informative sampling and not missing at random nonresponse. (À l'étude).