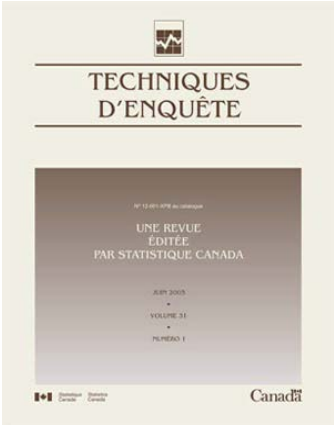
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

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



par Graham Kalton

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)

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2017

Tous droits réservés. 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, décembre 2017
Vol. 43, nº 2, pp. 183-188
Statistique Canada, nº 12-001-X au catalogue

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

Graham Kalton[1]

Résumé

Cette note de Graham Kalton 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.

Le bref article rédigé par Jon Rao et Wayne Fuller est d'une très grande portée. Sa lecture m'a incité à réfléchir à ma propre expérience quant à l'histoire de la recherche par sondage au cours des 50 dernières années ou plus, principalement en tant que spécialiste de la recherche appliquée en enquêtes sociales au Royaume-Uni et aux États-Unis. Dans l'ensemble, d'étonnants progrès ont eu lieu dans toutes les disciplines de la recherche par sondage au cours de ma vie professionnelle, y compris dans les domaines de l'échantillonnage et des méthodes de collecte de données. Jon et Wayne ont, naturellement, contribué considérablement à ces progrès. Ma discussion portera sur deux grands volets des changements qui ont eu lieu.

Évolution du rôle des modèles dans l'inférence par échantillonnage

Au début de ma carrière, le mode d'inférence fondé sur le plan de sondage établi par Neyman primait, mais sa prédominance a diminué au fil du temps, surtout récemment. L'inférence fondée sur le plan de sondage doit son intérêt au fait que la convergence des estimateurs des paramètres de population fondés sur un échantillon probabiliste tiré d'une population finie ne dépend pas de modèles, contrairement à l'estimation dépendante d'un modèle où l'inférence est fonction de la validité des hypothèses de modélisation. Dès le début, l'échantillonnage probabiliste et l'inférence fondée sur le plan de sondage étaient assistés par modèle, par exemple dans le cas de la répartition de l'échantillon sous échantillonnage stratifié et estimation par la régression. Néanmoins, bien que la formulation incorrecte de ces modèles de travail ait une incidence sur la précision des estimateurs de l'enquête, la convergence des estimateurs demeure intacte.

Les conditions à satisfaire pour une inférence purement fondée sur le plan de sondage sont que chaque unité de la population cible possède une probabilité de sélection non nulle connue (ou du moins une probabilité de sélection *relative* connue) et que des données d'enquête valides soient obtenues pour chaque unité échantillonnée. Dans les enquêtes sociales, ces conditions ne sont presque jamais entièrement satisfaites en pratique, parce que des données manquent inévitablement en raison de problèmes de

1. Graham Kalton est un vice-président principal à Westat, 1600 Research Blvd., Rockville, MD, 20850, États-Unis. Courriel : grahamkalton@westat.com.

non-couverture et de non-réponse (totale ainsi que partielle). Les modèles jouent un rôle essentiel dans le traitement des données manquantes, que ce soit par des méthodes de pondération et d'imputation ou en ignorant le problème, c'est-à-dire en employant implicitement un modèle qui traite les données manquantes comme des données manquant entièrement au hasard (MCAR pour *missing completely at random*).

Si je me souviens bien de la situation au début de ma carrière, l'usage de modèles pour produire des estimations au moyen de données d'enquêtes n'était guère reconnu et, en fait, l'opposition au recours à des modèles pour l'inférence à partir d'enquêtes était forte. De nombreux chercheurs ont résisté avec acharnement à l'imputation quand celle-ci a commencé à se répandre autour des années 1980, sous prétexte qu'elle impliquait des « données fabriquées »; les analystes procédaient plutôt à une analyse de cas complète, en émettant implicitement l'hypothèse MCAR. À cette époque, les taux de non-réponse étaient faibles, de sorte que l'appui sur des hypothèses de modélisation n'était pas important; quand de simples ajustements de la pondération pour tenir compte de la non-réponse étaient effectués, peu d'attention leur était accordée. Étant donné la hausse des taux de non-réponse ces dernières années, la situation a changé considérablement et les estimations d'après des données d'enquêtes dépendent maintenant fortement de modèles. Par conséquent, une foule de travaux de recherche ont porté sur les méthodes de modélisation des données manquantes, comme l'ont souligné Rao et Fuller.

Une autre façon dont les modèles jouent un rôle dans la pratique des sondages découle de l'utilisation de méthodes d'échantillonnage non probabiliste ou de méthodes d'échantillonnage qui ne répondent pas strictement à l'exigence que les probabilités de sélection soient connues. Les considérations de coût jouent un rôle important dans l'élaboration de l'échantillon, et elles peuvent mener au choix d'un plan d'échantillonnage avec probabilités de sélection inconnues qui sont ensuite calculées approximativement en se basant sur des hypothèses de modélisation. Un exemple bien connu est l'échantillonnage par quota, une méthode d'échantillonnage non probabiliste qui a été utilisée à grande échelle dans les études de marché et qui a été appliquée, au départ, dans un certain nombre d'études sociales (voir Stephan et McCarthy, 1958). Sudman (1966) décrit un plan de répartition par quota selon lequel des intervieweurs employaient des contrôles de quota pour créer des groupes de personnes que l'on supposait posséder les mêmes probabilités de sélection. Dans un secteur donné, l'intervieweur était alors libre de sélectionner n'importe quel sujet, à la condition que l'échantillon résultant satisfasse les contrôles de quota. Un autre exemple est celui de l'échantillonnage par marche aléatoire (*random walk*) qui permet d'éviter le coût du listage des logements dans un secteur échantillonné; selon cette méthode, l'intervieweur doit commencer à un emplacement spécifié et suivre une route donnée. Bauer (2016) montre comment cette méthode ne fournit pas l'échantillon avec probabilités égales qu'elle est censée produire. Les coûts de listage sont également évités dans le cas des enquêtes d'évaluation rapide du Programme élargi de vaccination (PEV) de l'Organisation mondiale de la santé lesquelles sont conçues pour estimer la couverture de la vaccination des enfants dans une région donnée. À l'intérieur d'une grappe échantillonnée (par exemple, un village), l'intervieweur commence par un ménage « sélectionné au hasard », puis passe au ménage suivant le plus proche, et ainsi de suite, en série, jusqu'à ce que la taille d'échantillon spécifiée soit atteinte (souvent sept enfants admissibles). En plus de supposer que les enfants sont échantillonnés au hasard à l'intérieur de la grappe échantillonnée, la méthode s'appuie sur l'hypothèse incorrecte que les grappes sont échantillonnées avec des probabilités exactement

proportionnelles au nombre d'enfants admissibles dans la grappe au moment de l'enquête. Bennett (1993) et d'autres ont proposé des modifications pour éviter les biais découlant du modèle hypothétique de probabilités de sélection égales pour cette méthode d'échantillonnage du PEV utilisée à très grande échelle.

Ces dernières années, la demande d'enquêtes pour étudier des populations rares, dont certaines sont définies en fonction de caractéristiques sensibles (y compris des comportements illégaux), a augmenté considérablement. Voir, par exemple, Tourangeau, Edwards, Johnson, Wolter et Bates (2014). Le recours à des méthodes d'échantillonnage non probabilistes est nécessaire dans les situations où l'échantillonnage probabiliste est jugé infaisable. Cependant, ces méthodes n'offrent pas la sécurité de l'inférence fondée sur le plan de sondage. Les plans de sondage non probabiliste largement utilisés pour l'étude de populations rares difficiles à échantillonner comprennent l'échantillonnage boule de neige, l'échantillonnage dirigé par les répondants, l'échantillonnage de lieux (fondé sur les lieux de rencontre), et les enquêtes en ligne.

Les enquêtes en ligne sont attrayantes parce qu'elles permettent d'obtenir des réponses aux enquêtes à peu de frais et presque instantanément. Ces enquêtes prennent de nombreuses formes, dont les enquêtes en ligne auto-sélectionnées, les panels volontaires d'utilisateurs d'Internet et les panels en ligne fondés sur des échantillons probabilistes (Couper, 2000). Les tailles d'échantillon des enquêtes en ligne auto-sélectionnées et des panels de volontaires sont souvent très grandes, mais la principale préoccupation tient aux biais possibles dans les estimations d'enquête. Comme l'indique le tristement célèbre sondage de 1936 de la revue *Literacy Digest*, les grands échantillons ne protègent pas contre le biais dans les estimations d'enquête. Le sondage en question était une enquête par la poste dont le questionnaire a été envoyé à environ 10 millions de personnes sélectionnées principalement à partir d'annuaires téléphoniques et de listes d'immatriculations de véhicules, et environ 2 millions de personnes ont répondu. Le sondage prédisait une victoire écrasante d'Alf Landon à l'élection présidentielle de 1936 aux États-Unis, alors que Franklin Roosevelt l'a remportée largement (voir Converse, 1987, pages 456-457 pour des références tentant d'expliquer l'échec du sondage). Il reste à déterminer si les méthodes modernes d'ajustement des pondérations appliquées à une collecte de données en ligne non probabiliste à grande échelle permettent de contourner les problèmes du sondage du *Literary Digest* et, élément plus critique, dans quelles conditions on peut se fier en toute confiance à la qualité des estimations dépendantes d'un modèle qui sont produites. Même quand un panel en ligne est recruté en utilisant un plan d'échantillonnage probabiliste, le taux de réponse global généralement très faible met gravement en question la sécurité de l'inférence fondée sur le plan de sondage.

Après des années d'opposition, les méthodes d'estimation sur petits domaines dépendantes d'un modèle sont aujourd'hui généralement acceptées, comme le font remarquer Rao et Fuller qui ont beaucoup contribué à la littérature sur ce sujet. Cette acceptation est attribuable au fait que la grande demande d'estimations sur petits domaines manifestée par les décideurs et d'autres ne peut pas être satisfaite par des méthodes fondées sur le plan de sondage avec des tailles d'échantillon d'un coût abordable. L'estimation sur petits domaines débute certes par la collecte de données auprès d'un échantillon probabiliste, mais ensuite elle « emprunte de l'information » à des modèles qui utilisent des données administratives, des données de recensements antérieurs, et d'autres données disponibles au niveau des petits domaines. Les modèles de petits domaines sont construits prudemment et évalués dans la mesure du possible, mais les estimations sur petits domaines résultantes dépendent néanmoins du modèle.

En résumé, se fier entièrement à l'inférence fondée sur le plan de sondage n'est pas raisonnable de nos jours pour diverses raisons. Une plus grande attention doit être accordée aux moyens de communiquer l'incertitude au sujet des estimations produites à partir de données hybrides contenant des composantes fondées sur le plan de sondage et d'autres dépendantes de modèles, en tenant compte des niveaux plausibles d'erreur de spécification du modèle.

Évolution des capacités informatiques des dernières décennies

Les progrès concernant les capacités informatiques au cours des dernières décennies ont eu une influence importante sur tous les aspects de la recherche par sondage. Au début de ma carrière, l'analyse des données d'enquêtes se faisait au moyen de cartes perforées sur des compteuses-trieuses et autres appareils de ce genre. La totalisation était presque la seule forme d'analyse. Les erreurs-types reflétant les plans d'échantillonnage complexes étaient rarement calculées; on appliquait plutôt de simples règles empiriques pour modifier les erreurs-types sous échantillonnage aléatoire simple. Dans son ouvrage intitulé *Sample Design in Business Research*, Deming (1960) recommandait que les échantillons soient conçus en 10 répliques pour faciliter l'estimation de la variance, et en outre, il proposait que l'erreur-type d'une estimation soit obtenue par la simple division par 10 de la différence entre la plus grande et la plus petite répliques. Dans *Survey Sampling*, Kish (1965) insistait sur la simplicité des calculs de variance fondés sur un plan avec échantillons appariés selon lequel deux unités primaires d'échantillonnage sont sélectionnées dans chaque strate, et il a exposé la façon d'effectuer ces calculs à la main. Aujourd'hui, les estimations de variance pour des statistiques simples ou complexes fondées sur des plans d'échantillonnage complexes sont calculées facilement à l'aide de progiciels utilisant des techniques telles que les répliques répétées équilibrées, les répliques répétées jackknife, le bootstrap et la linéarisation. En outre, pour les méthodes de rééchantillonnage, il est simple de recalculer des ajustements de pondération, même complexes, pour chaque réplique, ce qui permet d'intégrer dans les estimations de variance la variabilité associée à ces ajustements.

L'impact des ordinateurs sur la statistique d'enquête ne se limite pas à l'estimation de la variance. Les ordinateurs permettent aussi d'appliquer des plans de sondage plus complexes, ce qui a mené à la multiplication des méthodes complexes d'analyse (comme l'expliquent Rao et Fuller). Considérons le cas d'une stratification poussée à titre d'exemple d'un plan plus complexe. Goodman et Kish (1950) décrivent une méthode de stratification poussée, appelée sélection contrôlée, pouvant être effectuée par de simples calculs. Les méthodes d'échantillonnage équilibré plus récentes, à savoir l'échantillonnage par la méthode du cube et la méthode connexe de l'échantillonnage réjectif mentionnées par Rao et Fuller, sont de loin plus compliquées à appliquer.

Les ordinateurs ont également eu des effets importants sur d'autres aspects du processus d'enquête. Il y a 50 ans, les données d'enquêtes étaient recueillies au moyen d'interviews utilisant papier et crayon (IPC) ou de questionnaires envoyés par la poste. La méthode IPC a été en grande partie remplacée par la collecte de données assistée par ordinateur (Couper, Baker, Bethlehem, Clark, Martin, Nicholls et O'Reilly, 1998). Les interviews sur place assistées par ordinateur (IPAO) sont réalisées en utilisant des ordinateurs portables ou, plus fréquemment aujourd'hui, des tablettes. Certaines données – surtout les données sensibles – peuvent être recueillies par auto-interviews assistées par ordinateur avec interface audio (audio-AIAO).

Dans le cas d'une collecte de données par IPAO, la totalité ou des parties d'une interview peuvent être enregistrées (interview enregistrée assistée par ordinateur – IEAO); l'IEAO peut être utile pour les prétests et pour vérifier la performance des intervieweurs tout au long de la période de collecte des données. Les ordinateurs permettent aussi de recueillir les emplacements GPS des interviews, ce qui rend possible la surveillance des falsifications des intervieweurs et fournit des données pour diverses analyses fondées sur la localisation. Plus récemment, la collecte de données en ligne est devenue un mode de collecte rentable intéressant, mais à une époque où les taux de réponse s'effondrent, elle doit souvent être complétée par des interviews sur place ou d'autres méthodes. Les enquêtes à mode de collecte mixte deviennent de plus en plus populaires, et leur utilisation augmentera vraisemblablement dans l'avenir (en accordant l'attention nécessaire aux différences possibles de réponse à certaines questions selon le mode de collecte). Voir Dillman (2017) pour une revue de problèmes liés au fait d'inciter initialement les participants aux enquêtes à mode mixte à répondre sur le Web.

Conclusion

L'histoire de la recherche par sondage est caractérisée par une augmentation rapide du nombre et de la complexité des demandes de données d'enquêtes. Ces demandes ont abouti à l'utilisation de fichiers à grande diffusion (FGD) pour les analyses individuelles et à des préoccupations concernant la protection des données confidentielles des répondants. La diffusion de nombreux tableaux et d'autres analyses suscite des préoccupations comparables. Pour répondre à ces préoccupations, des méthodes de contrôle de la divulgation statistique sont élaborées pour les FGD (par exemple, voir Hundepool, Domingo-Ferrer, Franconi, Giessing, Schulte Nordholt, Spicer et de Wolf, 2012), des fichiers de données à usage restreint sont utilisés et des enclaves statistiques sont créées pour permettre aux analystes d'effectuer leurs analyses sous supervision. En outre, des archives ont été établies pour stocker et gérer les ensembles de données d'enquêtes.

La croissance rapide de la demande de données d'enquêtes constatée au cours des dernières décennies est une tendance qui se poursuivra vraisemblablement, ce qui semble réserver un bel avenir à la recherche par sondage. Toutefois, comme l'a dit Paul Valery, « Le problème avec notre temps, c'est que l'avenir n'est plus ce qu'il était », remarque qui semble particulièrement pertinente dans le cas de la recherche par sondage en ce moment. D'aucuns voient dans les mégadonnées et les dossiers administratifs de sérieux concurrents pour les enquêtes, mais je ne suis pas aussi convaincu. Ces deux types de données peuvent satisfaire certains besoins, mais la nature multivariée des enquêtes et, souvent, la nécessité de recueillir certains éléments de données qui ne peuvent être obtenus qu'auprès des répondants (par exemple, opinions, niveau de littératie des adultes, dépenses des ménages, diabète) signifient que les enquêtes continueront de jouer un rôle important. Les données administratives peuvent produire des estimations pour certaines statistiques officielles (surtout les statistiques économiques), particulièrement quand il est permis de fusionner des ensembles de fichiers administratifs. Cependant, selon moi, dans les enquêtes sociales officielles, les dossiers administratifs seront utilisés principalement comme un supplément susceptible de réduire le fardeau de réponse en remplaçant les questions de l'enquête par des données enregistrées et pouvant fournir des

données longitudinales pour la période qui précède ainsi que celle qui suit la collecte des données de l'enquête. Naturellement, la qualité des données enregistrées et des couplages d'enregistrements doit être évaluée. À mon avis, la menace la plus importante qui pèse sur la recherche par sondage tient à la réticence croissante des membres du public à répondre aux enquêtes. Jusqu'à présent, aucune bonne solution contre cette menace n'a été découverte.

Bibliographie

Bauer, J.J. (2016). Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4, 263-287.

Bennett, S. (1993). Cluster sampling to assess immunization: A critical appraisal. *Bulletin of the International Statistical Institute*, 49e Session, 55(2), 21-35.

Converse, J.M. (1987). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.

Couper, M.P. (2000). Web surveys. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. et O'Reilly, J.M. (Éds.) (1998). *Computer Assisted Survey Information Collection.* New York: John Wiley & Sons, Inc.

Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2017). Inciter les participants aux enquêtes à mode mixte à répondre sur le Web : les promesses et les défis. *Techniques d'enquête*, 43, 1, 3-34. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14836-fra.pdf.

Goodman, R., et Kish, L. (1950). Controlled selection - A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. et de Wolf, P.-P. (2012). *Statistical Disclosure Control*. Chichester, Royaume-Uni: Wiley.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons, Inc.

Stephan, F.F., et McCarthy, P.J. (1958). *Sampling Opinions*. New York: John Wiley & Sons, Inc.

Sudman, S. (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61, 749-771.

Tourangeau, R., Edwards, B., Johnson, T.P., Wolter, K.M. et Bates, N. (Éds.) (2014). *Hard-to-survey populations*. Cambridge, Royaume-Uni: Cambridge University Press.