N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2004 : Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

2004

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2004 de Statistique Canada
Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

ENQUÊTES AUPRÈS DE POPULATIONS RARES OU DIFFICILES À REJOINDRE AU MOYEN D'ÉCHANTILLONNAGE PAR RÉSEAU : REVUE HISTORIQUE

Monroe G. Sirken[1]

RÉSUMÉ

Cet article décrit la suite d'événements heureux et imprévisibles qui a débuté vers la fin des années 1960 et qui devait successivement donner naissance à l'échantillonnage par réseau une dizaine d'années plus tard, et à l'apparition d'applications inattendues de cet échantillonnage au milieu des années 1990. L'article décrit le cadre conceptuel de l'échantillonnage par réseau. Il présente également des résultats de recherche dans ce domaine et illustre des applications de l'échantillonnage par réseau avant et après le milieu des années 1990. Il évoque enfin très brièvement les orientations futures et les perspectives d'avenir de l'échantillonnage par réseau.

MOTS CLÉS : Enquêtes-établissements, enquêtes-ménages, multiplicité, patrons d'heureux hasards.

1. INTRODUCTION

1.1 Contexte

La mise au point de méthodes efficaces de conception d'enquêtes par sondage auprès de populations rares représente un défi pour les chercheurs du domaine depuis des décennies. Et c'est là un défi qu'ils doivent continuer à relever. Dans son allocution de 1956 à titre de présidente de l'ASA (Cox, 1957), Gertrude Cox a identifié la conception de méthodes efficaces pour l'échantillonnage d'items rares comme étant l'un des objets premiers de l'investigation statistique. Depuis que la D^{re} Cox a dressé il y a près de 50 ans sa liste des problèmes statistiques à investiguer, plusieurs méthodes novatrices ont été développées pour enquêter les populations rares ou difficiles à joindre. On prévoit parler d'un grand nombre de ces méthodes au symposium. Je suis heureux d'avoir été invité à vous entretenir de l'une d'entre elles, soit l'échantillonnage par réseau, à cette séance d'ouverture du 21^{e} Symposium international sur les questions de méthodologie de Statistique Canada.

Cet article décrit l'évolution de l'échantillonnage par réseau depuis ses origines il y a presque 50 ans, c'est-à-dire à peine deux ans après l'allocution présidentielle de la D^{re} Cox en 1956. Avant d'en faire l'historique, un certain nombre de concepts et de termes techniques sont définis et ma version de l'histoire de l'échantillonnage par réseau est présentée.

1.2 Règles de dénombrement et réseaux

Les règles de dénombrement tiennent une place essentielle parmi les caractéristiques de conception des enquêtes (Sirken, 1973). Elles précisent la nature des relations qui relient les unités d'observation aux unités de sélection où elles sont énumérables dans l'enquête. Elles mettent les unités de sélection en grappes, de sorte que chaque grappe contienne les unités d'observation rattachées à une même unité de sélection et que les mêmes unités d'observation puissent être rattachées à des unités de sélection multiples. Elles répartissent en outre les unités d'observation en réseaux, chacun de ces réseaux contenant les unités d'observation rattachées au même ensemble d'unités de sélection. Le nombre d'unités de sélection reliées à un réseau est sa multiplicité. Les règles de dénombrement ont

[1] Sirken, Monroe G., National Center for Health Statistics, 3311, chemin Toledo, Hyattsville, Maryland, États-Unis, MSirken@cdc.gov.

des effets sur l'échantillonnage, parce qu'elles déterminent les modes de répartition des unités d'observation en grappes et en réseaux. Elles ont aussi des effets sur la mesure des erreurs, parce qu'elles précisent les unités de sélection auxquelles les unités d'observation peuvent être énumérées.

Par leur nature, ces règles varient selon qu'il s'agit d'un échantillonnage par réseau ou d'un échantillonnage conventionnel. Dans un échantillonnage conventionnel, on utilise des règles de dénombrement unitaire qui créent un lien unique entre une unité d'observation et une unité de sélection à laquelle elle est énumérable dans l'enquête. Ainsi, la règle du lieu officiel de résidence dans les enquêtes-ménages conventionnelles est une règle de dénombrement unitaire qui crée un lien unique entre chaque individu et le ménage de son lieu habituel de résidence. Une telle restriction ne s'applique pas à l'échantillonnage par réseau. Des règles de multiplicité permettent de rattacher une unité d'observation à des unités de sélection multiples. Ainsi, la règle de dénombrement de fratrie, qui est une règle de multiplicité souvent appliquée dans l'échantillonnage par réseau dans les enquêtes-ménages, relie chaque fratrie d'un ensemble de fratries à tous les ménages qui sont des lieux officiels de résidence pour chaque fratrie de l'ensemble. Les multiplicités peuvent varier d'un réseau de fratries à un autre en fonction du nombre de ménages rattachés aux fratries de l'ensemble. On peut voir l'échantillonnage conventionnel comme un cas particulier de l'échantillonnage par réseau où la multiplicité de chaque réseau est égale à un.

La variabilité de la taille des réseaux offre à l'échantillonnage par réseau des possibilités de conception propres à résoudre les problèmes de plan de sondage qui se posent souvent dans l'échantillonnage conventionnel. Tel est sûrement le cas lorsque, dans un plan de sondage, on se trouve inopinément ou inévitablement à rattacher les mêmes unités d'observation à des unités de sélection multiples. D'une plus grande importance stratégique, l'échantillonnage par réseau peut être délibérément promu comme une stratégie en vue d'améliorer l'efficacité des enquêtes où l'échantillonnage conventionnel engendre des erreurs d'échantillonnage et autres erreurs de mesure importantes. Ainsi, l'échantillonnage par réseau est habituellement plus efficace que l'échantillonnage conventionnel dans les enquêtes-ménages auprès de populations rares, car les ménages sont plus uniformément répartis par des règles de multiplicité que par des règles de dénombrement unitaire. Il sera aussi probablement plus efficace lorsque les populations rares sont également difficiles à rejoindre ou sont des populations « sensibles » difficilement énumérables dans leurs lieux officiels de résidence dans le cadre d'enquêtes-ménages conventionnelles.

L'échantillonnage par réseau n'est pas sans inconvénients. Pour obtenir des estimations sans biais, il faut recueillir un complément d'information permettant d'établir les valeurs de multiplicité, ce dont on n'a pas besoin en échantillonnage conventionnel, la multiplicité se réduisant alors à l'unité. Comme la multiplicité vise seulement dans l'échantillonnage par réseau les unités d'observation signalées par les unités de sélection échantillonnées, elle est habituellement déclarée dans l'enquête par les unités de sélection qui signalent les unités d'observation. Il reste que l'obtention de ce complément d'information se trouve à ajouter aux frais d'enquête. Peut-être une conséquence plus grave que le surcroît de coût est-elle le risque d'erreur de réponse associé aux multiplicités. L'échantillonnage par réseau n'est donc pas une panacée aux problèmes de plan de sondage qui débordent le cadre de l'échantillonnage conventionnel, mais dans une application judicieuse et sélective, il a le potentiel d'améliorer l'efficacité des plans de sondage des enquêtes-ménages et des enquêtes-établissements (Sirken, 1997).

1.3 Aperçu historique

Pour mon historique de l'échantillonnage par réseau depuis 45 ans et plus, j'ai pensé vous entretenir de mon expérience personnelle de cette technique, l'occasion étant là de décrire de première main le rôle important et intéressant des heureux hasards dans ses origines et son évolution. Dans sa préface au délicieux ouvrage « Travels and Adventures of Serendipity » de Robert Merton et Elinor Bender, Robert Shulman décrit ainsi les heureux hasards :

« On trouvera peut-être quelque chose de précieux en cherchant quelque chose d'entièrement différent ou il peut s'agir de découvrir un objet recherché d'une manière tout à fait inattendue [TRADUCTION]. »

J'ai eu la bonne fortune et le plaisir d'expérimenter ces deux types d'heureux hasards dans mes recherches intermittentes sur l'échantillonnage par réseau durant ces décennies.

Selon le professeur Merton, les patrons d'heureux hasards en science se présentent en trois étapes que je transposerai ici en quatre étapes :

Étape 1. Une recherche dirigée vers un objectif mène à une découverte anormale et inattendue ou à une solution étonnante et imprévue.

Étape 2. Cette découverte ou cette solution inattendue amène le chercheur à insérer ces nouvelles données dans un cadre plus général du savoir.

Étape 3. La recherche suscitée par cette découverte ou cette solution fait naître des idées pour une théorie ou une technologie nouvelle.

Étape 4. Cette théorie ou cette technologie est appliquée et éprouvée expérimentalement.

L'histoire de l'échantillonnage par réseau se divise commodément en deux périodes. La première période, l'émergence de l'échantillonnage par réseau, s'étend sur 35 ans, c'est-à-dire de 1958 à 1983. C'est la période où cet échantillonnage a émergé en tant qu'heureux hasard suite à une découverte inattendue dans une enquête par sondage auprès des fournisseurs de services médicaux sur la prévalence de maladies rares. La seconde période, l'émergence des enquêtes-ménages et enquêtes-établissements reliées s'étend sur 20 ans et plus depuis 1983. C'est une période où de nouvelles applications de l'échantillonnage par réseau ont émergé en tant qu'heureux hasard suite à une solution imprévue à un problème d'estimation dans une enquête-établissements auprès des fournisseurs de services médicaux selon les déclarations dans une enquête-ménages. Pour chacune de ces périodes, je présenterai la chronologie des quatre étapes déjà mentionnées des patrons d'heureux hasards. À la dernière section du article, je reviendrai sur l'histoire de l'échantillonnage par réseau pour me guider dans mes prévisions sur l'avenir de cette technique d'échantillonnage.

2. ORIGINES DE L'ÉCHANTILLONNAGE PAR RÉSEAU : 1960-1983

2.1 Découverte inattendue

En 1959, on a entrepris à l'échelle nationale une enquête pilote sur les médecins et les hôpitaux dans trois États de la Nouvelle-Angleterre en vue d'estimer la prévalence des cas de fibrose kystique ayant fait l'objet d'un diagnostic médical (Kramm et coll., 1962). Pour cette enquête pilote auprès d'environ 1 600 médecins et hôpitaux, on a utilisé un plan d'échantillonnage stratifié dans lequel les pédiatres et les hôpitaux ayant un service spécialisé de pédiatrie ont été sélectionnés avec certitude.

Cette maladie génétique relativement rare de l'enfance a été reconnue comme affection distincte au milieu des années 1930 et, à l'époque de cette enquête vers la fin des années 1950, les tests de diagnostic étaient toujours relativement grossiers et leurs résultats, souvent difficiles à interpréter. Les procédures mises en œuvre dans cette enquête pilote pour tenir compte de ce problème des diagnostics ont révélé l'existence d'un problème imprévu de plan de sondage.

Les sources médicales interrogées devaient déclarer tous les patients atteints de fibrose kystique qu'ils avaient traités depuis 1972. On apprenait ainsi le nom et l'adresse, le lieu de naissance et le sexe des patients avec les constatations médicales à l'appui du diagnostic établi. On a aussi obtenu le nom et l'adresse des sources médicales de renvoi qui, dans les cas échéants, avaient traité chacun des patients. De ces sources, on s'est par la suite enquis d'informations supplémentaires au sujet du diagnostic des patients pour lesquels ils avaient été rapportés comme source médicale de renvoi. Une fois l'enquête terminée, on a examiné et évalué la certitude des diagnostics d'après les indications communiquées par les sources médicales de consultation initiale et de renvoi.

Après appariement des fiches des patients, on a vu que l'échantillon initial de sources médicales avait déclaré environ 650 cas distincts de fibrose kystique et que les patients en question avaient été traités par plus de 1 000 sources médicales différentes. En moyenne, chaque patient avait été pris en charge par environ 1,6 fournisseur de services médicaux. Environ la moitié des patients avaient été traités par une source, le tiers par deux sources, le dixième par trois sources et 3 % par quatre ou cinq sources.

On n'avait aucune difficulté à estimer sans biais la prévalence de la fibrose kystique dans l'enquête pilote, puisque la presque totalité des patients avaient été déclarés par des fournisseurs des strates échantillonnées avec certitude. Le problème de biais se serait autrement posé, l'appariement seul ne suffisant pas à assurer une estimation sans biais lorsque les patients sont déclarés uniquement par des sources appartenant à des strates non échantillonnées avec certitude. La présence d'un biais risquait de poser un grave problème dans l'enquête nationale envisagée auprès des fournisseurs de services médicaux où on prévoyait qu'une proportion appréciable des patients seraient déclarés par des fournisseurs des strates non échantillonnées avec certitude.

2.2 Recherche suscitée par une découverte inattendue

La recherche d'une solution au problème d'estimation qui se présentait dans l'enquête pilote sur la fibrose kystique a fait naître une initiative de recherche afin de développer la théorie d'échantillonnage qui serait appropriée à la conception d'enquêtes par sondage semblables à cette dernière, où les patients sont souvent traités par des fournisseurs multiples. Cette recherche a donné trois estimateurs sans biais du paramètre **N**, soit le nombre total de patients atteints de la maladie. Elle a aussi donné les variances de deux de ces estimateurs (Birnbaum et Sirken, 1965). Les trois estimateurs exploitent l'information sur la multiplicité des sources médicales de traitement des mêmes patients, mais diffèrent selon le type d'information de multiplicité requis.

L'estimateur de multiplicité, qui est le plus simple et le plus robuste des trois, dénombre les fiches de chaque patient dans l'enquête et pondère chacune par l'inverse de la multiplicité du patient. Avec cet estimateur, on n'a pas à apparier les fiches des patients pour relever les doubles. Ainsi, on aurait pu établir les valeurs de multiplicité des patients de l'enquête pilote en Nouvelle-Angleterre à l'aide des données déclarées sur les sources médicales de renvoi par l'échantillon initial de fournisseurs de services médicaux sans effectuer d'opération d'appariement pour l'élimination des fiches en double. Dans cette éventualité, l'estimateur de multiplicité est sans biais si chaque patient de l'univers est rattaché à au moins un fournisseur et déclaré par lui.

L'attention curieuse prêtée à la question de l'efficacité de l'estimateur de multiplicité du paramètre **N** a à son tour suscité une recherche au sujet des effets respectifs sur le plan de sondage de l'échantillonnage par réseau et de l'échantillonnage conventionnel dans les enquêtes auprès des fournisseurs de services médicaux. Voici quelques résultats de recherche dans ce domaine (Sirken, 1970a) : (1) la différence entre les variances des estimateurs issus de l'échantillonnage par réseau et de l'échantillonnage conventionnel dépend de la configuration des liens entre les unités de sélection et les unités d'observation qui sont formés par les règles de dénombrement unitaire et les règles de multiplicité; (2) l'échantillonnage par réseau n'est pas nécessairement plus efficace que l'échantillonnage conventionnel pour toutes les configurations possibles de liens, mais il sera sans doute plus efficace pour la plupart des règles de multiplicité; (3) l'échantillonnage par réseau est nécessairement plus efficace que l'échantillonnage conventionnel toutes les fois que les règles de multiplicité forment des configurations spécifiques de liens.

Cette troisième constatation est stratégiquement la plus importante. Ainsi, l'échantillonnage par réseau sera nécessairement plus efficace que l'échantillonnage conventionnel pour l'estimation de **N** si aucune des unités de sélection n'est rattachée à des unités d'observation multiples par la règle de multiplicité. Dans cette configuration, l'espérance de l'estimateur de multiplicité est $\mathbf{P = N / M}$, où **M** est le nombre total d'unités de sélection dans l'univers. Si **P** est petit, l'effet de plan de l'échantillonnage par réseau, qui est le quotient des variances de multiplicité et de dénombrement unitaire, est moindre que le paramètre **K**. Fraction de l'intervalle 0 à 1, **K** est la moyenne des inverses du nombre d'unités de sélection rattachées aux mêmes unités d'observation (c'est-à-dire les inverses des valeurs de multiplicité des unités d'observation). L'effet de plan s'améliore à mesure que décroît **P** et au gré de l'augmentation et de la diminution respectives des moyennes et des variances de multiplicité. Lorsque la variance de la multiplicité est « ignorable », c'est-à-dire que toutes les valeurs de multiplicité correspondent à l'entier positif s, l'effet de plan de l'échantillonnage par réseau se ramène à l'équation $\mathbf{K < 1/s}$. [Je donne la preuve de cette équation dans mon article (Sirken, Monroe G., 1970b).]

2.3 Les résultats de recherche engendrent l'échantillonnage par réseau

La recherche sur les effets de plan de l'échantillonnage par réseau a mis en évidence le rôle primordial des règles de dénombrement dans la conception d'enquêtes par sondage. Elle a mené à de nouvelles idées sur les moyens d'améliorer les plans de sondage en tirant profit des possibilités de conception offertes par les règles de multiplicité. Les conséquences des résultats de recherche sur l'estimation sont évidentes, à savoir que l'échantillonnage par réseau est sans biais et que l'échantillonnage conventionnel est biaisé toutes les fois que des unités de sélection multiples sont inopinément ou inévitablement rattachées aux mêmes unités d'observation.

Les conséquences des résultats de recherche sont moins transparentes pour les effets de plan que pour l'estimation sans biais, mais elles sont d'un intérêt bien plus stratégique. Elles impliquent que l'échantillonnage par réseau mérite tout particulièrement notre attention comme moyen de conception afin d'améliorer la précision de l'échantillonnage lorsque les unités de sélection d'une enquête conventionnelle tendent à être binomiales avec de petites valeurs de **P**. C'est une conclusion qui a inspiré l'idée d'un échantillonnage par réseau délibéré dans les enquêtes-ménages conventionnelles menées auprès de populations rares. L'idée fondamentale est de promouvoir cette technique dans les enquêtes-ménages en se servant de règles de multiplicité pour former des configurations de liens remplissant simultanément et autant que possible les trois conditions suivantes :

Condition 1. La multiplicité est égale ou supérieure à un pour chaque individu.

Condition 2. La distribution des valeurs de multiplicité se caractérise par une grande moyenne et une petite variance.

Condition 3. Les individus sont rattachés à des ménages capables et désireux de déclarer les valeurs de multiplicité et les variables d'intérêt.

Avec la condition 1, on évite le biais de couverture. On la remplit d'ordinaire grâce à des règles de dénombrement composite qui rattachent les individus à leur propre ménage et à d'autres. Avec la condition 2, on optimise l'efficacité de l'échantillonnage par réseau. On la remplit souvent par des règles de dénombrement qui comportent une multiplicité de liens. Avec la condition 3, on réduit au minimum les effets des erreurs de réponse associées à l'échantillonnage par réseau. On la remplit par des règles de dénombrement comme celles qui rattachent les individus aux ménages de parents, d'amis, de voisins ou de collègues avec lesquels ils entretiennent des relations étroites et bien définies.

On a par la suite étendu l'application de la théorie de l'échantillonnage par réseau aux enquêtes menées auprès de populations « fuyantes » et « sensibles » afin de diminuer les erreurs de couverture et de réponse qui se présentent dans les enquêtes-ménages par échantillonnage conventionnel, les membres de ces populations étant difficiles à joindre et à énumérer dans leur propre ménage.

2.4 Vérification empirique de l'échantillonnage par réseau

Souvent, des règles de dénombrement favorables pour les erreurs d'échantillonnage ne le sont pas pour les erreurs de mesure et vice versa. Ainsi, la règle de multiplicité qui rattache chaque individu à chaque ménage se trouve à éliminer entièrement les erreurs d'échantillonnage, mais s'avère totalement impraticable en ce qui concerne les erreurs de réponse.

À la figure 1, nous résumons les résultats de quatre expériences d'enquête où on a vérifié les effets des erreurs d'échantillonnage et des biais de réponse de l'échantillonnage par réseau dans des enquêtes-ménages portant sur des populations ou des événements rares. Des expériences visant les naissances et les mariages ont été intégrés à l'enquête israélienne sur la population active de 1974 (Nathan, 1976). L'expérience relative au diabète s'est insérée dans l'enquête nationale par interviews sur la santé aux États-Unis en 1976 (Sirken et coll., 1978) et l'expérience relative au cancer a été réalisée en 1981 par le Survey Research Center de l'Université de l'Illinois dans cet État américain (Czaja et coll., 1986). Dans chacune de ces expériences, on a mis en comparaison les erreurs d'échantillonnage et les biais de réponse de la règle de dénombrement conventionnelle (règle du lieu officiel de résidence) et de deux règles de dénombrement composite ou de parenté dans l'estimation du paramètre N. La

figure 1 ordonne les valeurs relatives d'efficacité des trois règles de dénombrement dans les expériences relatives aux naissances et aux mariages; la figure 2 fait de même pour les expériences relatives au diabète et au cancer.

Diapositive 1 : Classement des règles de dénombrement selon les biais de réponse et les erreurs d'échantillonnage

Règles de dénombrement	Biais de réponse	Erreurs d'échantillonnage	Biais de réponse	Erreurs d'échantillonnage
	Mariages		Naissances	
Enquêtes d'événements démographiques				
Règle conventionnelle	**2**	**3**	**1**	**3**
Règle composite/parents	**3**	**2**	**3**	**2**
Règle composite/parents, fratries	**1**	**1**	**2**	**1**
	Diabète		**Cancer**	
Enquêtes de prévalence de la maladie				
Règle conventionnelle	**2**	**3**	**1**	**3**
Règle composite/fratries	**3**	**2**	**3**	**1**
Règle composite/enfants	**1**	**1**	**2**	**2**

Dans chacune de ces expériences, la règle conventionnelle vient au troisième rang pour les erreurs d'échantillonnage. Dans les expériences relatives au diabète et aux mariages, les règles de dénombrement composite d'enfants et composite de parents/fratries, respectivement, occupent le premier rang pour les deux types d'erreurs, ce qui implique que ces règles composites sont uniformément plus efficaces que la règle conventionnelle. Dans le cas des expériences relatives au cancer et aux naissances, les résultats sont inégaux, puisque la règle conventionnelle est première pour les biais de réponse et dernière pour les erreurs d'échantillonnage. Par conséquent, les règles composites sont plus efficaces si le taux de prévalence est bas et que l'échantillon de ménages n'atteint pas une taille spécifiée. Dans les autres cas, l'échantillonnage conventionnel se révèle plus efficace. Ainsi, pour un taux de prévalence de sièges cancéreux de 1 %, la règle composite d'enfants est plus efficace que la règle conventionnelle pour des tailles d'échantillon de ménages allant jusqu'à environ 40 000. Pour un taux de natalité de 3 %, la règle composite parent/fratrie est plus efficace que la règle de dénombrement conventionnelle jusqu'à une taille d'échantillon d'environ 19 000 ménages.

Ces expériences confirment que l'échantillonnage par réseau peut largement accroître l'efficacité de l'échantillonnage et la validité des enquêtes-ménages portant sur des populations ou des événements rares. Même dans les cas où il a un effet négatif sur la validité, l'échantillonnage par réseau peut améliorer les erreurs quadratiques moyennes, surtout lorsque les tailles d'échantillon et les taux de prévalence sont faibles.

3. ORIGINES DES ENQUÊTES RELIÉES : DEPUIS 1983

3.1 Solution inattendue

Vers 1985, le NCHS a lancé un programme de recherche où il a adopté la National Health Interview Survey (Enquête nationale par interviews sur la santé ou NHIS) comme base principale de sondage en vue de l'intégration des plans de sondage de ses familles d'enquêtes indépendantes sur la population et sur les fournisseurs de services de santé (Sirken et Greenberg, 1983). Des listes de ménages et de personnes de la NHIS ont constitué les bases de sondage des enquêtes auprès de la population et des listes d'unités primaires d'échantillonnage de la NHIS ont constitué celles des enquêtes auprès des fournisseurs de services de santé. En 1992, on a chargé un groupe d'experts du Committee on National Statistics d'examiner les plans du NCHS en vue de la restructuration de ses enquêtes auprès des fournisseurs de services de santé (Wunderlich, 1992). Ce groupe d'étude a bien aimé l'idée de l'intégration des plans de sondage, mais jugeait que l'intégration des enquêtes auprès d'établissements de services

de santé serait plus efficace si les liens avec la NHIS se faisaient au niveau des ménages plutôt qu'à celui des unités primaires d'échantillonnage. Il a recommandé au NCHS d'examiner la faisabilité et l'utilité des listes de ces fournisseurs qui sont consultés et déclarés par les ménages de la NHIS comme bases de sondage de ses enquêtes auprès des fournisseurs de services de santé au lieu de bases de sondage autonomes qui énumèrent tous les fournisseurs et en mesurent la taille. Cette recommandation a inspiré un projet de recherche sur les caractéristiques de conception des enquêtes-établissements qui exploitent les bases de sondage des enquêtes-ménages. C'est ce qu'on appelle l'enquête reliée établissements-ménages (« linked establishment/population survey » ou LEPS).

Au départ, la LEPS a été conçue comme enquête-établissements à deux degrés pour l'estimation du paramètre X qui est la somme des valeurs de la variable x sur les transactions des ménages avec les établissements. Les établissements de la base de sondage de l'enquête-ménages sont les unités de sélection du premier degré alors que les transactions des établissements sélectionnés avec les ménages constituent les unités de sélection du second degré. Dans le modèle d'erreur, on a supposé que l'échantillon de ménages de l'enquête-ménages d'où vient la base de sondage est issu d'un échantillonnage aléatoire simple (EAS) avec remise (AR) et que, pour chaque établissement, l'échantillon de transactions est tiré indépendamment d'un EAS sans remise (SR) où la taille d'échantillon est proportionnelle au nombre de transactions des établissements avec les ménages de l'enquête-ménages. Dans une optique d'enquête-établissements, le modèle d'erreur a exprimé l'estimateur LEPS à deux degrés de X asymptotiquement sans biais, mais non pas la variance LEPS (Judkins et coll., 1995).

Il m'est apparu que le problème d'estimation gagnerait en transparence et sa solution en praticabilité si la LEPS était modélisée comme une enquête-ménages à deux degrés où les ménages seraient les unités de sélection du premier degré et où toutes les transactions des établissements avec les ménages seraient les unités de sélection du second degré. (À noter que les unités du second degré sont les transactions des établissements avec tous les ménages, et non pas les transactions avec les ménages de l'échantillon.) Modélisée ainsi, la LEPS est une enquête-ménages à échantillonnage par réseau, puisque les transactions des établissements sont rattachées aux ménages multiples avec lesquels ces transactions ont lieu. Le modèle a exprimé l'estimateur LEPS à deux degrés de X et sa variance (Sirken, Shimizu et Judkins, 1995).

3.2 Recherche suscitée par la solution inattendue

La dérivation de l'expression de l'estimateur LEPS sans biais et de sa variance a ouvert la voie à une recherche où on a comparé, sur le plan de l'efficacité d'échantillonnage, l'estimateur LEPS à deux degrés de X et l'estimateur à deux degrés conventionnel de Hansen-Hurwitz où les établissements sont prélevés sans remise sur une base de sondage autonome avec probabilité proportionnelle à la taille (Sirken et Shimizu, 2002).

Il y a équivalence entre l'estimateur à deux degrés de LEPS et l'estimateur à deux degrés conventionnel des enquête-établissements pour des tailles prévues équivalentes d'échantillon d'établissements et de transactions si et si seulement les transactions de la population qui génèrent la base de sondage LEPS sont distribuées uniformément de sorte que chaque ménage a une seule transaction. Lorsqu'on s'écarte de cette distribution uniforme parce que certains ménages n'ont pas de transactions et/ou que les transactions ne sont pas distribuées uniformément sur la population tronquée de ménages ayant des transactions, la composante du premier degré de la variance LEPS augmente, d'où une perte d'efficacité dans presque tous les cas par rapport à la composante du premier degré de la variance d'une enquête-établissements conventionnelle. Dans un échantillonnage à deux degrés, le résultat est un peu moins favorable à l'enquête-établissements conventionnelle, car la composante intra-établissement de la variance de second degré est moindre dans la LEPS que dans une enquête conventionnelle toutes les fois que, au second degré, les échantillons de transactions sont sélectionnés sans remise et que des établissements multiples sont rattachés aux mêmes ménages.

Malgré ces effets de plan sans doute préjudiciables, la LEPS mérite toute notre attention chaque fois qu'il est peu pratique ou trop coûteux d'élaborer ou de tenir à jour des bases de sondage autonomes avec une couverture raisonnablement complète et de bonnes mesures de taille qui conviennent aux populations cibles et aux thèmes de l'enquête. Du point de vue des coûts, la LEPS devient particulièrement attrayante lorsqu'elle se trouve en intégration avec des enquêtes-ménages et des enquêtes-établissements permanentes.

3.3 Les résultats de recherche suscitent des applications de la LEPS dans les enquêtes-ménages

L'étude des effets de plan de la LEPS dans les enquêtes-établissements a fait naître des idées d'application aux enquêtes-ménages auprès de populations rares ou difficiles à rejoindre, surtout là où les variables d'intérêt sont déclarées avec plus de précision par les établissements que par les ménages (Sirken et Shimizu, sous presse). Assimilée à une enquête-ménages, la LEPS est une enquête-ménages à échantillonnage par réseau où on applique une règle de multiplicité rattachant les transactions de l'établissement à chaque ménage avec lequel il a des transactions. La règle de dénombrement de la LEPS répartit les transactions en réseaux, chacun de ces réseaux contenant les transactions d'un établissement distinct. En réalité, les établissements sont des réseaux et, dans l'enquête, ils sont essentiellement des répondants par substitution qui déclarent les variables d'intérêt au sujet de leurs propres transactions avec les ménages concernés.

Il y a équivalence, sur le plan de l'efficacité d'échantillonnage, de la LEPS à deux degrés et de l'enquête-ménages conventionnelle à un degré pour des tailles équivalentes d'échantillon de ménages si et si seulement la composante intra-établissement de la variance est « ignorable ». Dans les autres cas, l'échantillonnage est presque toujours moins efficace dans l'enquête-ménages conventionnelle que dans la LEPS à un degré de même que dans la LEPS à deux degrés lorsque, au second degré, on sélectionne des échantillons de transactions suffisamment importants. Si aucun des ménages n'a de transactions multiples par exemple, ce qui est une configuration de transactions des plus probables dans les enquêtes menées auprès de populations rares, une taille d'échantillon LEPS qui, au second degré, ne sera pas supérieure au nombre de transactions des ménages dans l'enquête-ménages sera assez grande pour conférer à la LEPS une efficacité d'échantillonnage du moins égale à celle de l'enquête conventionnelle.

Bref, la LEPS a le potentiel d'améliorer substantiellement la qualité des estimations du paramètre X dans les enquête-ménages conventionnelles, surtout là où les unités d'observation appartiennent à des populations rares, « fuyantes » ou « sensibles » qui sont difficiles à joindre ou à énumérer à leur lieu habituel de résidence et où les établissements sont de bonnes sources d'information sur les variables d'intérêt de l'enquête.

3.4 Vérification empirique de la LEPS

Autant que je sache, on n'a pas appliqué la LEPS aux enquêtes de santé afin d'estimer le volume de transactions qui interviennent entre les établissements et les ménages. Il reste que les listes d'établissements énumérés dans les enquêtes-ménages ont servi de base de sondage à des enquêtes économiques et à des enquêtes auprès d'organismes là où des listes complètes et fiables de l'univers des établissements sont difficiles à obtenir, à dresser ou à mettre à jour.

Les bases de sondage d'enquêtes auprès de la population sont utilisées dans les enquêtes auprès des entreprises afin d'estimer les dépenses de la population en biens et services. Ainsi, les bases de sondage de l'enquête sur les prix à la consommation, qui est une enquête-entreprises permanente du U.S. Bureau of Labor Statistics (BLS), sont des listes d'établissements du commerce de détail qui viennent de l'enquête permanente sur les points de vente de l'IPC, une enquête-ménages nationale où on demande aux répondants de rapporter leurs des achats et d'identifier les marchands leur ayant vendu la marchandise. Le BLS ne prend pas l'estimateur LEPS pour estimer le volume des dépenses de consommation (Leaver et Valliant, 1995).

Ajoutons que, dans les enquêtes sur les congrégations religieuses et les enquêtes sur les organismes employeurs et bénévoles, on se sert de bases de sondage d'enquêtes-ménages pour estimer les caractéristiques des organismes et de leur clientèle (Kallenberg et coll.). Ainsi, les bases de sondage des enquêtes sur les congrégations religieuses ont été générées des occasions de 1992 et 1998 de l'enquête sociale générale, qui est une enquête-ménages nationale où le National Opinion Research Center (NORC) demande de préciser la fréquentation d'assemblées religieuses (Chaves, 1999). Les sociologues emploient fréquemment l'estimateur LEPS dans les enquêtes auprès d'organismes et parlent alors d'échantillonnage par multiplicité ou par « hyperréseau ».

4. UN DOUBLE REGARD EN ARRIÈRE ET EN AVANT

Si on revoit l'histoire de l'échantillonnage par réseau dans la recherche sur les enquêtes pour mieux juger de ses perspectives d'avenir, on a l'impression que son évolution dépendra de découvertes anormales et inattendues et de solutions neuves et imprévues qu'on exploitera stratégiquement pour améliorer les plans de sondage. Je ne tenterai sûrement pas de prévoir des événements qui, par définition, se présenteront par accident.

Il me semble que, dans l'avenir comme par le passé, l'échantillonnage par réseau continuera de s'appliquer aux enquêtes où les mêmes unités d'observation sont inévitablement liées à des unités de sélection multiples et où elles sont difficiles à joindre ou à énumérer par les règles de dénombrement unitaire propres à l'échantillonnage conventionnel. Comme je l'ai mentionné, on peut s'attendre à ce que, demain comme hier, de nouvelles applications de l'échantillonnage par réseau naissent d'heureux hasards découlant de solutions et découvertes inattendues aux problèmes de plan de sondage dans des conditions qu'il m'est impossible d'imaginer.

REMERCIEMENTS

Une première version de cet article a été présentée en octobre 2004 au Survey Research Laboratory de l'Université de l'Illinois. Une version corrigée paraîtra dans un futur numéro du bulletin d'information de ce laboratoire. Je remercie Althelia Harris, mon ex-secrétaire, de m'avoir aidé à produire ce manuscrit selon les règles de présentation documentaire du symposium. Les opinions exprimées sont celles de l'auteur et ne correspondent pas nécessairement aux vues du National Center for Health Statistics.

RÉFÉRENCES

Birnbaum, Z.W. et Sirken, M.G. (1965), "Design of Sample Surveys to Estimate the Prevalence of Rare diseases: Three Unbiased Estimates", *Vital and Health Statistics*, PHS Publication, Series 2, No. 11.

Chavis, M. (1999), "Religious Congregations and Welfare, Reform: Who Will Take Advantage of Charitable Choice", *American Sociological Review*, pp. 836-846.

Cox, G.M. (1957), "Statistical Frontiers", *Journal of the American Statistical Association*, 64, pp. 1-12.

Czaja, R. F., Snowden, C.B. et Casady, R. J. (1986), "Reporting Bias and Sampling Errors in Surveys of Rare Populations Using Multiplicity Counting Rules", *Journal of the American Statistical Association*, 81, pp. 411-419.

Kallenberg, A., Knoke, D., Marsden,P., et Spaeth, J. *Organizations In America.*

Judkins, D., Berk, M., Edwards, S., Mohr, P., Stewart, K. et Waksberg, J. (1995), "National Health Care Survey: List Verses Network Sampling", rapport non publié, National Center for Health Statistics.

Kramm, E. R., Crane, M. M., Sirken, M. G. et Brown, M. L. (1962), "A Cystic Fibrosis Pilot Survey in Three New England States", *American Journal of Public Health*, 52, pp. 2041-2057.

Leaves S. et Valiant R. (1995), "Statistical Problems in Estimating the U.S. Consumer Price Index", dans B. G. Cox, et al (eds.), *Business Survey Methods*, New York: Wiley.

Nathan, G. (1976), "Empirical Study of Response and Sampling Errors for Multiplicity Estimators with Different Counting Rules", *Journal of the American Statistical Association*, 71, pp. 803-815.

Sirken, M. G. (1970a), "Household Surveys with Multiplicity", *Journal of the American Statistical Association*, 65, pp. 257-266.

Sirken, M. G. (1070b), "Survey Strategies for Estimating Rare Health Attributes", *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 135-144.

Sirken, M. G. (1973), "The Counting Rule Strategy in Sample Surveys", *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 340-342.

Sirken, M. G. (1977), "Network Sampling", *The Encyclopedia of Biostatistics*, Wiley, pp. 2977-2985.

Sirken, M. G., Graubard, B. I. et McDaniel, M. J. (1978), "National Network Survey of Diabetes", *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA. 631-635.

Sirken, M., Roystin, P., Bercini, D., Czaja, R., Eastman, E. et Warnecke R. (1981), "Completeness of Enumeration of Cancer Cases in Health Surveys", *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Sirken, M. G. et Greenberg, M. S. (1983), "Redesign and Integration of a Population Based Health Survey Program", *Proceedings of the 44th Session of the International Statistical Institute*.

Sirken, M. G., Shimizu, I. et Judkins, D. (1995), "Population Based Establishment Surveys", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 470-475.

Sirken, M. G. et Shimizu (2002), "Effets de plan de sondage dus aux bases de sondage dans les enquêtes auprès des établissements", *Techniques d'enquête*, 28, no. 2, pp. 183-190.

Sirken, M. G. et Shimizu, (circa 2005), "Establishment Based Population Surveys: Design Effects of Linked Population/establishment Surveys Compared to Conventional Surveys", dans P.S.R.S. Rao and M.J. Katzoff (eds.) *A Handbook of Sampling Methods and Analysis*, Chapman Hall/CRC Press.

Wunderlich, G. S. (ed.) (1992), *Toward a National Health Care Survey: A Data System for the 21st Century*, National Academy Press.