No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du symposium 2005 de Statistique Canada
Défis méthodologiques pour les besoins futurs d'information

ÉVALUATION DE L'INTERACTION ENTRE LA THÉORIE ET LA PRATIQUE DES ENQUÊTES PAR SONDAGE[1]

J.N.K. Rao[2]

RÉSUMÉ

Une grande partie de la théorie des enquêtes par sondage a été motivée directement par des problèmes d'ordre pratique survenus au moment de la conception et de l'analyse des enquêtes. En revanche, la théorie des enquêtes par sondage a influencé la pratique, ce qui a souvent donné lieu à des améliorations importantes. Dans le présent article, nous examinons cette interaction au cours des 60 dernières années. Nous présentons également des exemples où une nouvelle théorie est nécessaire ou encore où la théorie existe sans être utilisée.

MOTS CLÉS : Analyse des données d'enquête; apports antérieurs; question d'inférence; méthodes de rééchantillonnage; estimation sur petits domaines.

1. INTRODUCTION

Dans cet article, je vais examiner l'inter-relation entre la théorie des sondages et la pratique dans les quelques 60 dernières années. Je vais couvrir une grande variété de sujets: les premières contributions significatives qui ont grandement influencé la pratique, les questions d'inférence, l'estimation par calage qui assure la cohérence aux totaux établis de variables auxiliaires, l'échantillonnage à probabilités inégales sans remplacement, l'analyse de données d'enquêtes, le rôle des méthodes de ré-échantillonnage, et l'estimation pour petits domaines. Je vais aussi présenter quelques exemples où il y a soit besoin d'une nouvelle théorie soit une théorie existante qui n'est pas tellement utilisée.

2. QUELQUES APPORTS MARQUANTS : 1920 – 1970

La présente section rend compte de certains apports marquants à la théorie et aux méthodes des enquêtes par sondage, apports qui ont grandement influencé la pratique. Le statisticien norvégien A.N. Kiaer (1897) fut sans doute le premier à promouvoir l'échantillonnage (appelé « méthode représentative » à l'époque) plutôt qu'un dénombrement complet, quoique la plus ancienne référence à l'échantillonnage remonte au grand récit épique indien Mahabharata (Hacking 1975, page 7). Dans la méthode représentative, l'échantillon doit refléter la population mère finie; à cette fin, on procède par échantillonnage équilibré au moyen de la sélection raisonnée ou par échantillonnage aléatoire. On utilisa la méthode représentative en Russie dès 1900 (Zarkovic 1956) et, vers la même époque, Wright l'employa pour mener des enquêtes par sondage aux États-Unis. Dans les années 1920, on utilisait abondamment la méthode représentative, et l'Institut international de statistique joua un rôle de premier plan en créant en 1924 un comité chargé de produire un rapport sur cette méthode. Le rapport de ce comité portait sur des aspects théoriques et pratiques de la méthode d'échantillonnage aléatoire. Bowley (1926) contribua à ce rapport par ses travaux fondamentaux sur l'échantillonnage aléatoire stratifié avec répartition proportionnelle, qui permit de tirer un échantillon représentatif avec probabilités d'inclusion égales. Hubback (1927) prit

[1] Cet article a initialement paru dans la livraison de décembre 2005 de *Techniques d'enquêtes* (Volume 31, No 2, pp 127-151). Il est republié ici dans ce recueil avec la permission des éditeurs.

[2] J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, (Ontario), Canada, K1S 5B6.

conscience de la nécessité d'un échantillonnage aléatoire dans les enquêtes sur les cultures : « La seule façon d'arriver à une estimation satisfaisante consiste à établir une approximation de l'échantillonnage aléatoire aussi proche que les circonstances le permettent, car ainsi, non seulement on élimine les limites personnelles de l'expérimentateur, mais il devient possible de déterminer la probabilité avec laquelle les résultats d'un nombre donné d'échantillons se situeront à l'intérieur d'une étendue donnée par rapport à la moyenne arithmétique. Concrètement, il s'agit de trouver combien d'échantillons sont nécessaires pour assurer que la probabilité soit d'au moins 20:1 par rapport à la moyenne des échantillons à l'intérieur d'un maund de la vraie moyenne. » Cet énoncé contient deux observations importantes concernant l'échantillonnage aléatoire : 1) il évite les biais personnels dans la sélection d'un échantillon; 2) on peut déterminer la taille de l'échantillon pour satisfaire une marge d'erreur spécifiée par rapport à une chance de 1 sur 20. Mahalanobis (1946b) a observé que les travaux fondamentaux de R.A. Fisher sur la conception des expériences, menés à la Rothamsted Experimental Station, furent directement influencés par Hubback (1927).

Dans un article marquant, devenu un classique, Neyman (1934) a jeté les bases théoriques de l'échantillonnage probabiliste (ou fondé sur le plan de sondage) en ce qui concerne l'inférence à partir d'échantillons d'enquête. Il a montré, avec des arguments théoriques et des exemples pratiques, que l'échantillonnage aléatoire stratifié était préférable à l'échantillonnage équilibré, car ce dernier peut donner de mauvais résultats si les hypothèses sous-jacentes du modèle sont violées. Neyman a également avancé, dans sa théorie de l'échantillonnage aléatoire stratifié sans remise, les notions d'efficience et de répartition optimale en assouplissant la condition des probabilités d'inclusion égales. En généralisant le théorème de Markov sur l'estimation par les moindres carrés, Neyman a prouvé que la moyenne stratifiée, $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, était le meilleur estimateur de la moyenne de population, $\bar{Y} = \sum_h W_h \bar{Y}_h$, dans la classe linéaire d'estimateurs sans biais de forme $\bar{y}_b = \sum_h W_h \sum_i b_{hi} y_{hi}$, où $W_h, \bar{y}_h$ et $\bar{Y}_h$ sont le poids, la moyenne d'échantillon et la moyenne de population de la h^{e} strate $(h = 1, ..., L)$, et b_{hi} est une constante associée à la valeur de l'élément y'_{hi} observée au moment du i^{e} tirage d'échantillon $(i = 1, ..., n_h)$ dans la h^{e} strate. On a obtenu la répartition optimale $(n_1, ..., n_L)$ de la taille de l'échantillon total, n, en minimisant la variance de $\bar{y}_{st}$ sous réserve de $\sum_h n_h = n$; on a découvert plus tard une preuve antérieure de la répartition de Neyman par Tschuprow (1923). Neyman a également proposé une inférence à partir de grands échantillons en fonction d'intervalles de confiance selon la théorie normale, de manière que la fréquence des erreurs dans les énoncés de confiance en fonction de tous les échantillons aléatoires stratifiés qu'il est possible de tirer n'excède pas la limite prescrite à l'avance « quelles que soient les propriétés inconnues de la population ». Une méthode d'échantillonnage qui satisfait l'énoncé de fréquence susmentionné est dite « représentative ». Il est à noter que Hubback (1927) avait déjà fait allusion à l'énoncé de fréquence associé à l'intervalle de confiance. Dans son dernier apport à la théorie des enquêtes par sondage, Neyman (1938) a étudié l'échantillonnage à deux phases de stratification et calculé la taille optimale des échantillons de première phase et de deuxième phase, n' et n, en minimisant la variance de l'estimateur sous réserve d'un coût donné $C = n'c' + nc$, où le coût par unité de deuxième phase, c, est élevé par rapport au coût par unité de la première phase, c'.

Au cours des années 1930, la demande d'information a connu une croissance rapide et l'on a pris conscience des avantages de l'échantillonnage probabiliste – portée accrue, réduction de coût, plus grande vitesse et caractéristiques indépendantes d'un modèle –, d'où une augmentation du nombre et du type d'enquêtes menées par échantillonnage probabiliste et couvrant de grandes populations. La presque totalité des statisticiens d'enquête ont adopté l'approche de Neyman. En outre, cette dernière a inspiré divers ajouts importants, motivés surtout par des critères d'ordre pratique et d'efficience. L'article marquant de Cochran (1939) présente plusieurs résultats importants : le recours à l'analyse de variance pour estimer l'amélioration de l'efficience due à la stratification, l'estimation des composantes de la variance dans l'échantillonnage à deux degrés en vue d'études futures sur un sujet semblable, le choix de l'unité d'échantillonnage, l'estimation par régression sous échantillonnage à deux phases et l'effet des erreurs dans la taille des strates. Dans cet article, Neyman a également proposé le concept de superpopulation : « La population finie doit être considérée comme un échantillon aléatoire d'une population infinie. » Il est intéressant de noter qu'à l'époque, Cochran n'était pas d'accord avec le concept traditionnel de population fixe : « En outre, il est loin d'être réaliste de considérer la population comme un lot fixe de nombres connus. » Cochran (1940) a proposé l'estimation par quotient pour les enquêtes par sondage, mais Laplace (1820) avait déjà utilisé l'estimateur par quotient. Dans un autre article marquant, Cochran (1942) a formulé la théorie de l'estimation par régression. Il a calculé la variance conditionnelle de l'estimateur par régression habituel pour un

échantillon fixe ainsi qu'un estimateur échantillon de cette variance, en supposant un modèle de régression linéaire $y = \alpha + \beta x + e$, où e a une moyenne nulle et une variance constante dans les séries statistiques dans lesquelles x est fixe. Il a également noté que l'estimateur par régression restait sans biais (par rapport au modèle) sous échantillonnage non aléatoire, à condition que le modèle de régression linéaire hypothétique soit correct. Il a calculé le biais moyen en présence d'écarts par rapport au modèle (notamment dans le cas de la régression quadratique) pour l'échantillonnage aléatoire simple à mesure que la taille de l'échantillon n augmente. Cochran a ensuite étendu ses résultats à la régression pondérée et calculé le résultat d'optimalité, aujourd'hui bien connu, pour l'estimateur par quotient; selon lui, il s'agit de « la meilleure estimation linéaire sans biais si la valeur moyenne et la variance changent proportionnellement à x ». Dans les travaux récents, ce dernier modèle est appelé modèle par quotient. Madow et Madow (1944) et Cochran (1946) ont comparé la variance prévue sous un modèle de superpopulation pour étudier analytiquement l'efficience relative de l'échantillonnage systématique et de l'échantillonnage aléatoire stratifié. Cet article a incité d'autres chercheurs à mener des travaux sur l'utilisation de modèles de superpopulation dans le choix de stratégies d'échantillonnage probabiliste, ainsi que sur l'inférence dépendante d'un modèle et l'inférence assistée par un modèle (voir la section 3).

En Inde, Mahalanobis a fait un apport innovateur à la théorie de l'échantillonnage en formulant des fonctions de coût et de variance pour la conception d'enquêtes. Son article marquant (Mahalanobis 1944) présente des résultats théoriques probants sur la conception efficace d'enquêtes par sondage et leurs applications pratiques, notamment dans le cas d'enquêtes sur les surfaces cultivées et le rendement des cultures. Maintenant bien connue, la répartition optimale sous échantillonnage aléatoire stratifié où le coût par unité varie d'une strate à l'autre est obtenue sous forme de cas particulier de sa théorie générale. Dès 1937, Mahalanobis a utilisé des plans de sondage à plusieurs degrés pour les enquêtes sur le rendement des cultures avec, comme unités d'échantillonnage aux quatre degrés d'échantillonnage, des villages, des grilles à l'intérieur des villages, des parcelles à l'intérieur des grilles et des coupes de tailles et de formes différentes (Murthy 1964). Il a également utilisé un plan de sondage à deux phases pour estimer le rendement de l'écorce de quinquina. Il a joué un rôle de premier plan dans l'établissement de la National Sample Survey (NSS) de l'Inde, la plus vaste enquête polyvalente permanente : un personnel à temps plein effectue des interviews sur place pour des enquêtes socioéconomiques et des mesures physiques pour des enquêtes sur les cultures. Plusieurs éminents statisticiens d'enquête, dont D.B. Lahiri et M.N. Murthy, ont collaboré à la NSS.

P.V. Sukhatme, qui a étudié avec Neyman, a également fait un apport innovateur à la conception et à l'analyse d'enquêtes agricoles à grande échelle en Inde, en utilisant l'échantillonnage stratifié à plusieurs degrés. À partir de 1942 – 1943, il a mis au point des plans de sondage efficients pour mener des enquêtes nationales sur les cultures de blé et de riz et a obtenu un degré élevé de précision pour les estimations nationales ainsi qu'une marge d'erreur raisonnable pour les estimations par district. L'approche de Sukhatme différait de celle de Mahalanobis, qui utilisait des parcelles de très petite taille pour les coupes-témoins et employait des enquêteurs *ad hoc*. Sukhatme (1947) et Sukhatme et Panse (1951) ont démontré que l'utilisation d'une petite parcelle pourrait donner des estimations biaisées à cause de la tendance à placer des plantes de bornage à l'intérieur de la parcelle lorsqu'il y a un doute. Ils ont également souligné que le recours à des enquêteurs *ad hoc*, qui se déplacent rapidement d'un endroit à l'autre, obligeait à mesurer uniquement les parcelles de champs échantillonnés qui sont prêts à moissonner à la date de la visite, ce qui est contraire au principe de l'échantillonnage aléatoire. La solution de Sukhatme consistait à utiliser de grandes parcelles pour éviter les biais liés au bornage et à confier les coupes-témoins à l'organisme public local chargé du revenu ou de l'agriculture.

De 1940 à 1970, les statisticiens d'enquête du U.S. Census Bureau, sous la direction de Morris Hansen, William Hurwitz, William Madow et Joseph Waksberg, ont fait des apports fondamentaux à la théorie et à la pratique des enquêtes par sondage, et bon nombre de leurs méthodes sont encore largement utilisées dans la pratique. Hansen et Hurwitz (1943) ont formulé la théorie de base de l'échantillonnage stratifié à deux degrés, une seule unité primaire d'échantillonnage (UPÉ) à l'intérieur de chaque strate étant tirée avec probabilité proportionnelle à la taille (échantillonnage PPT) puis sous-échantillonnée à un rythme qui assure l'autopondération (probabilités de sélection globales égales) à l'intérieur des strates. Cette approche permet de confier aux intervieweurs des charges de travail à peu près égales, ce qui est souhaitable dans le contexte des enquêtes sur le terrain. Elle permet aussi de réduire considérablement la variance en neutralisant la variabilité due à la taille inégale des UPÉ sans vraiment stratifier selon la taille, ce qui permet la stratification selon d'autres variables

pour réduire la variance. Par contre, les charges de travail peuvent varier considérablement si les UPÉ sont sélectionnées par échantillonnage aléatoire simple, puis sous-échantillonnées au même rythme à l'intérieur de chaque strate. Aujourd'hui, on utilise abondamment l'échantillonnage PPT des UPÉ dans la conception d'enquêtes à grande échelle, mais on sélectionne dans chaque strate deux ou plusieurs UPÉ sans remise, de sorte que les probabilités d'inclusion des UPÉ sont proportionnelles à la taille (voir la section 5).

Bon nombre d'enquêtes à grande échelle sont répétées au fil du temps, comme l'Enquête sur la population active (EPA) du Canada, menée chaque mois, et la Current Population Survey (CPS) des États-Unis, avec remise partielle des unités finales (appelée aussi échantillonnage par renouvellement). Dans le cas de l'EPA, par exemple, l'échantillon de ménages est divisé en six groupes de renouvellement (échantillons constants ou panels) et un groupe de renouvellement reste dans l'échantillon pendant six mois consécutifs, puis est retiré de l'échantillon, ce qui donne un chevauchement de cinq sixièmes entre deux mois consécutifs. Dans la foulée des travaux initiaux de Jessen (1942) sur l'échantillonnage à deux reprises avec remise partielle des unités, Yates (1949) et Patterson (1950) ont jeté les bases théoriques de la conception et de l'estimation d'enquêtes à passages répétés et démontré qu'on pouvait améliorer l'efficience de l'estimation de niveau et de changement en tirant parti des données antérieures. Hansen, Hurwitz, Nisseslson et Steinberg (1955) ont mis au point des estimateurs plus simples, appelés estimateurs composites K, applicables aux plans d'échantillonnage stratifié à plusieurs degrés avec échantillonnage PPT au premier degré. Rao et Graham (1964) ont étudié des politiques de remise optimale pour les estimateurs composites K. On a également proposé divers ajouts. On a utilisé des estimateurs composites dans le cas de la CPS et d'autres enquêtes permanentes à grande échelle. Encore récemment, l'EPA du Canada a adopté l'estimation composite, appelée estimation composite par régression, qui utilise l'information sur l'échantillon obtenue au cours des mois précédents et qui peut être mise en œuvre avec un logiciel de poids de régression (voir la section 4).

Keyfitz (1951) a proposé une méthode ingénieuse pour obtenir de meilleures mesures de la taille des UPÉ dans les enquêtes permanentes fondées sur les plus récents dénombrements censitaires. Sa méthode permet de maximiser la probabilité de chevauchement avec l'échantillon antérieur d'une UPÉ par strate, ce qui réduit les coûts d'opération sur le terrain tout en améliorant l'efficience grâce aux meilleures mesures de la taille dans l'échantillonnage PPT. L'EPA du Canada et d'autres enquêtes permanentes ont utilisé la méthode de Keyfitz. Raj (1956) a formulé le problème de l'optimisation comme un « problème de transport » dans la programmation linéaire. Kish et Scott (1971) ont étendu la méthode de Keyfitz aux mesures changeantes des strates et de la taille. Ernst (1999) a brossé un excellent tableau de l'évolution, au cours des 50 dernières années, de la coordination d'échantillons (qui consiste à maximiser ou minimiser le chevauchement des échantillons) au moyen d'algorithmes de transport et de méthodes connexes; voir aussi Mach, Reiss et Schiopu-Kratina (2005) en ce qui concerne les applications aux enquêtes-entreprises avec création et suppression d'entreprises.

Dalenius (1957, chapitre 7) a étudié le problème de la stratification optimale d'un nombre donné de strates, L, dans le cadre de la répartition de Neyman. Dalenius et Hodges (1959) ont obtenu une approximation simple de la stratification optimale, appelée méthode de la fonction cumulative de la racine carrée des fréquences (cum $\sqrt{f}$), qui est abondamment utilisée dans la pratique. Pour les populations très asymétriques dont un petit nombre d'unités comptent pour une forte proportion du total Y, comme les populations d'entreprises, une stratification efficace nécessite une strate à tirage complet ($n_1 = N_1$) de grandes unités et des strates à tirage partiel d'unités moyennes et petites. Lavallée et Hidiroglou (1988) et Rivest (2002) ont mis au point des algorithmes pour déterminer les bornes de stratification en utilisant la méthode puissance (Fellegi 1981; Bankier 1988) et la répartition de Neyman pour les strates à tirage partiel. Aujourd'hui, Statistique Canada et d'autres organismes utilisent ces algorithmes pour les enquêtes-entreprises.

Avant 1950, la recherche portait sur l'estimation de totaux et de moyennes de population pour la population entière et de grandes sous-populations planifiées, comme des États ou des provinces. Or, les utilisateurs s'intéressent également aux totaux et aux moyennes de sous-populations non planifiées (appelées aussi domaines), comme les groupes d'âge-sexe à l'intérieur d'une province, ainsi qu'à des paramètres autres que les totaux et les moyennes, comme les médianes et d'autres quantiles, par exemple le revenu médian. Hartley (1959) a formulé une théorie simple et unifiée de l'estimation par domaine, applicable à n'importe quel plan de sondage et nécessitant uniquement les formules-types pour l'estimateur du

total et son estimateur de variance, dénotés respectivement $\hat{Y}(y)$ et $v(y)$ dans la notation d'opérateur. Il a introduit deux variables synthétiques ${}_j y_i$ et ${}_j a_i$ qui prennent respectivement les valeurs y_i et 1 si l'unité i appartient au domaine j et qui sont nulles dans le cas contraire. Alors, on obtient simplement les estimateurs du total de domaine ${}_j Y = Y({}_j y)$ et de la taille de domaine ${}_j N = Y({}_j a)$ à l'aide des formules pour $\hat{Y}(y)$ et $v(y)$ en remplaçant respectivement y_i par ${}_j y_i$ et ${}_j a_i$. De même, on obtient les estimateurs des moyennes de domaine et des différences de domaine ainsi que leurs estimateurs de variance à l'aide des formules de base pour $\hat{Y}(y)$ et $v(y)$. Durbin (1968) a également obtenu des résultats semblables. Aujourd'hui, on pratique couramment l'estimation par domaine en utilisant l'ingénieuse méthode de Hartley.

Pour l'inférence concernant des quantiles, Woodruff (1952) a proposé une méthode simple et ingénieuse pour obtenir un intervalle de confiance de niveau $(1-\alpha)$ sous des plans d'échantillonnage généraux, en utilisant uniquement la fonction de distribution estimative et son erreur-type (voir l'ouvrage de Lohr (1999), pages 311 à 313). Il est à noter qu'on obtient simplement ces dernières à l'aide des formules pour un total en remplaçant y par une variable indicatrice. En mettant sur le même pied l'intervalle de Woodruff et un intervalle selon la théorie normale à l'égard du quantile, on peut aussi obtenir une formule simple pour l'erreur-type du p^{e} estimateur de quantile, soit la moitié de la longueur de l'intervalle divisé par le point supérieur $\alpha/2$ de la distribution normalisée $N(0, 1)$ qui égale 1,96 si $\alpha = 0{,}05$ (Rao et Wu 1987; Francisco et Fuller 1991). L'intervalle de Woodruff possède une propriété étonnante : il donne de bons résultats même lorsque p est petit ou grand et que la taille de l'échantillon est moyenne (Sitter et Wu 2001).

On s'est rendu compte de l'importance des erreurs de mesure dès les années 1940. Dans un article influent, Mahalanobis (1946a) a mis au point la technique des sous-échantillons superposés (appelée échantillonnage répété par Deming 1960). En Inde, on a beaucoup utilisé cette méthode dans les enquêtes par sondage à grande échelle pour évaluer les erreurs d'échantillonnage et les erreurs de mesure. L'échantillon est tiré sous forme de deux ou plusieurs sous-échantillons indépendants selon le même plan de sondage, de sorte que chaque sous-échantillon fournit une estimation valide du total ou de la moyenne. Les sous-échantillons sont attribués à des intervieweurs différents (ou à des équipes différentes), ce qui produit une estimation valide de la variance totale qui tient compte de la variance de réponse corrélée due aux intervieweurs. Les sous-échantillons superposés entraînent une augmentation des frais de déplacement des intervieweurs, mais on peut les réduire en modifiant les affectations des intervieweurs. Hansen, Hurwitz, Marks et Mauldin (1951), Sukhatme et Seth (1952) et Hansen, Hurwitz et Bershad (1961) ont formulé des théories de base sous des modèles d'erreur de mesure additive et décomposé la variance totale en trois éléments : la variance d'échantillonnage, la variance de réponse simple et la variance de réponse corrélée. On a montré que la variance de réponse corrélée due aux intervieweurs était de l'ordre de k^{-1} sans égard à la taille de l'échantillon, k étant le nombre d'intervieweurs. Par conséquent, elle peut dominer la variance totale si k n'est pas un nombre élevé. Lors du recensement de 1950 aux États-Unis, l'étude de la variance due aux intervieweurs a montré que cette composante était effectivement grande pour les petits domaines. C'est en partie pour cette raison que lors du recensement de 1960, on a adopté l'autodénombrement par la poste pour réduire cette composante de la variance (Waksberg 1998). Il s'agit d'un exemple éloquent de l'influence de la théorie sur la pratique. Fellegi (1964) a proposé de combiner la superposition et la répétition pour estimer la covariance entre l'écart d'échantillonnage et l'écart de réponse. Cette composante est souvent négligée dans la décomposition de la variance totale, mais elle pourrait être appréciable dans la pratique.

Le concept de l'effet du plan de sondage (EPS), dû à Leslie Kish (voir Kish 1965, section 8.2), constitue un autre jalon de la méthodologie des enquêtes par sondage. L'effet du plan de sondage est le ratio de la variance réelle d'une statistique sous le plan de sondage spécifié à la variance qui serait obtenue sous échantillonnage aléatoire simple de même taille. Ce concept est particulièrement utile dans la présentation et la modélisation des erreurs d'échantillonnage, ainsi que dans l'analyse des données d'enquête complexes faisant intervenir la mise en grappes et les probabilités de sélection inégales (voir la section 6).

Le lecteur trouvera dans Kish (1995), Kruskal et Mosteller (1980), Hansen, Dalenius et Tepping (1985) et O'Muircheartaigh et Wong (1981) un examen des apports marquants à la théorie et aux méthodes des enquêtes par sondage.

3. QUESTIONS D'INFÉRENCE

3.1 Cadre unifié fondé sur le plan de sondage

Au départ, l'élaboration de la théorie de l'échantillonnage a progressé de manière plus ou moins inductive, quoique Neyman (1934) ait étudié la meilleure estimation linéaire sans biais pour l'échantillonnage aléatoire stratifié. On a envisagé des stratégies (plan de sondage et estimation) qui semblaient raisonnables et l'on a soigneusement étudié des propriétés relatives au moyen de méthodes analytiques ou empiriques, en comparant surtout des erreurs quadratiques moyennes, et parfois aussi des erreurs quadratiques moyennes ou des variances prévues sous des modèles de superpopulation plausibles, comme nous le mentionnons dans la section 2. On n'a pas insisté sur une estimation sans biais sous un plan de sondage donné, car elle « entraîne souvent une erreur quadratique moyenne beaucoup plus grande que nécessaire » (Hansen, Hurwitz et Tepping 1983). On a plutôt jugé que la cohérence avec le plan de sondage était nécessaire pour les grands échantillons. Les ouvrages classiques de Cochran (1953), Deming (1950), Hansen, Hurwitz et Madow (1953), Sukhatme (1954) et Yates (1949), fondés sur l'approche susmentionnée, ont grandement influencé la pratique des enquêtes. Pourtant, les statisticiens universitaires accordaient peu d'attention à la théorie de l'échantillonnage traditionnelle, peut-être parce qu'il lui manquait un cadre théorique formel et qu'elle n'était pas intégrée à la théorie statistique courante. Plusieurs départements de statistique nord-américains de prestige n'offraient pas de cours supérieurs en théorie de l'échantillonnage.

Dans les années 1950, on a élaboré des cadres et des approches théoriques formels pour intégrer la théorie de l'échantillonnage à l'inférence statistique courante dans des conditions quelque peu idéalistes axées sur les erreurs d'échantillonnage, en supposant l'absence d'erreurs de mesure ou de réponse ainsi que de non-réponse. Horvitz et Thompson (1952) ont apporté une contribution de base à l'échantillonnage avec probabilités de sélection arbitraires en formulant trois sous-classes d'estimateurs linéaires sans biais d'un total Y, dont la classe de Markov étudiée par Neyman. Une autre sous-classe avec poids de sondage d_i lié à une unité d'échantillonnage i et dépendant uniquement de i admettait l'estimateur bien connu avec poids inversement proportionnel à la probabilité d'inclusion π_i comme seul estimateur sans biais. Narain (1951) ayant également découvert cet estimateur, on devrait l'appeler l'estimateur de Narain-Horvitz-Thompson (NHT) au lieu de l'estimateur HT comme on l'appelle couramment. Pour l'échantillonnage aléatoire simple, la moyenne d'échantillon est le meilleur estimateur linéaire sans biais (best linear unbiased estimator ou BLUE) de la moyenne de population dans les trois sous-classes, mais ce n'est pas suffisant pour prétendre que la moyenne d'échantillon est le meilleur de tous les estimateurs linéaires sans biais. Godambe (1955) a proposé une classe générale d'estimateurs linéaires sans biais d'un total Y en supposant des données-échantillons $\{(i, y_i), i \in s\}$ et un poids dépendant de l'unité d'échantillonnage i ainsi que des autres unités échantillonnées s, c'est-à-dire un poids de forme $d_i(s)$. Il a alors établi que l'estimateur BLUE n'existait pas dans la classe générale

$$\hat{Y} = \sum_{i \in s} d_i(s)\, y_i, \qquad (1)$$

même sous échantillonnage aléatoire simple. Ce résultat théorique négatif important a été, dans une grande mesure, négligé pendant une dizaine d'années. Godambe a également établi un résultat positif en liant y à une mesure de taille x au moyen d'un modèle de régression de superpopulation passant par l'origine avec variance d'erreur proportionnelle à x^2, puis en montrant que l'estimateur NHT sous un plan de sondage à taille fixe où π_i est proportionnel à x_i minimisait la variance prévue de la classe sans biais (1). Ce résultat montre clairement les conditions du plan pour l'utilisation de l'estimateur NHT. Rao (1966) a constaté les limites de l'estimateur NHT dans le contexte d'enquêtes avec échantillonnage PPT et caractéristiques multiples. Ici, l'estimateur NHT s'avère très inefficient lorsqu'une caractéristique y n'est pas liée (ou qu'elle est faiblement liée) à la mesure de taille x (comme le dénombrement de volailles y et la taille de la ferme x dans une enquête sur les fermes). Rao a proposé pour ces cas d'autres estimateurs efficients qui font abstraction des poids NHT. En faisant abstraction des résultats susmentionnés, des spécialistes de l'échantillonnage ont avancé plus tard certains critères théoriques pour affirmer qu'il fallait utiliser l'estimateur NHT pour tout plan de sondage. En prenant l'exemple amusant des éléphants d'un cirque, Basu (1971) a illustré la futilité de ces critères. Il a construit un « mauvais » plan dans

lequel π_i n'était pas lié à y_i pour démontrer que l'estimateur NHT produisait des estimations absurdes, ce qui a incité le célèbre statisticien bayésien Dennis Lindley à conclure que ce contre-exemple détruisait la théorie des enquêtes par sondage fondées sur le plan de sondage (Lindley 1996). Cette conclusion est plutôt malheureuse, car NHT et Godambe ont clairement énoncé les conditions du plan pour une utilisation appropriée de l'estimateur NHT, et Rao (1966) et Hajek (1971) ont proposé d'autres estimateurs pour composer respectivement avec les caractéristiques multiples et les mauvais plans. Il est intéressant de noter que les mêmes critères théoriques ont abouti à un mauvais estimateur de variance de l'estimateur NHT comme choix « optimal » (Rao et Singh 1973).

On a aussi tenté d'intégrer la théorie des enquêtes par sondage à l'inférence statistique courante au moyen de la fonction de vraisemblance. Godambe (1966) a montré que la fonction de vraisemblance d'après les données-échantillons $\{(i, y_i),\ i \in s\}$, en considérant comme paramètre le vecteur N des valeurs y inconnues, ne fournissait pas d'information sur les valeurs non observées de l'échantillon ni, par conséquent, sur le total Y. Cette caractéristique non informative de la fonction de vraisemblance est due à la propriété d'étiquette qui traite les unités de population N essentiellement comme des post-strates N. On peut contourner cette difficulté en employant la méthode bayésienne et en supposant des valeurs antérieures informatives (échangeables) sur le vecteur paramètre (Ericson 1969). Une autre solution (fondée sur le plan de sondage) consiste à faire abstraction de certains aspects des données-échantillons pour rendre l'échantillon non unique et arriver ainsi à une fonction de vraisemblance informative (Hartley et Rao 1968; Royall 1968). Par exemple, sous échantillonnage aléatoire simple, en supprimant les étiquettes i et en considérant les données $\{y_i,\ i \in s\}$ en l'absence d'information liant i à y_i, on obtient la moyenne d'échantillon comme estimateur du maximum de vraisemblance de la moyenne de population. En supposant des distributions antérieures non informatives, l'estimation bayésienne produit des résultats semblables à ceux obtenus par Ericson (1969) mais, contrairement à l'estimation d'Ericson, elle dépend du plan de sondage. Dans le cas où y_i est un vecteur qui comprend des variables auxiliaires avec totaux connus, Hartley et Rao (1968) ont montré que sous échantillonnage aléatoire simple, l'estimateur du maximum de vraisemblance était à peu près égal à l'estimateur par régression traditionnel du total. Cet article a été le premier à montrer comment intégrer des totaux de population auxiliaire connus à un cadre de vraisemblance. Pour l'échantillonnage aléatoire stratifié, on fait abstraction des étiquettes à l'intérieur des strates, mais pas des étiquettes de strate, à cause des différences connues entre les strates. L'estimateur du maximum de vraisemblance ainsi obtenu est à peu près égal à un pseudo-estimateur par régression linéaire optimal lorsqu'on dispose de variables auxiliaires avec totaux connus. Ce dernier estimateur possède de bonnes propriétés conditionnelles fondées sur le plan de sondage (voir la section 3.4). L'article de Hartley et Rao (1968) portait sur l'estimation d'un total, mais l'approche de la vraisemblance a une portée beaucoup plus vaste en échantillonnage, dont l'estimation de fonctions de distribution et de quantiles et la construction d'intervalles de confiance fondés sur des rapports de vraisemblance (voir la section 8.1). L'approche de la vraisemblance non paramétrique de Hartley-Rao a été découverte indépendamment vingt ans plus tard (Owen 1988) dans l'inférence statistique courante, sous le nom de « vraisemblance empirique », et a attiré passablement d'attention, notamment pour son application à divers problèmes d'échantillonnage. Dans un certain sens, les efforts d'intégration à la statistique courante ont donc partiellement réussi. L'ouvrage d'Owen (2002) présente une description complète de la théorie de la vraisemblance empirique et de ses applications.

3.2 Approche dépendante d'un modèle

En matière d'inférence, l'approche dépendante d'un modèle suppose que la structure de population obéit à un modèle de superpopulation spécifié. La distribution induite par le modèle hypothétique produit des inférences qui renvoient à l'échantillon donné d'unités s qui a été tiré. Ces inférences conditionnelles peuvent s'avérer plus pertinentes et plus attrayantes que les inférences établies par échantillonnage répété. Par contre, lorsque le modèle n'est pas spécifié correctement, les stratégies dépendantes d'un modèle peuvent donner de mauvais résultats dans le cas de grands échantillons; même de faibles écarts par rapport au modèle hypothétique, difficiles à déceler au moyen de méthodes de vérification de modèle, peuvent causer de graves problèmes. Par exemple, prenons le modèle par quotient souvent utilisé lorsqu'une variable auxiliaire x au total connu X est aussi mesurée dans l'échantillon :

$$y_i = \beta x_i + \varepsilon_i;\ i = 1, \ldots, N \qquad (2)$$

où les ε_i sont des variables aléatoires indépendantes avec moyenne nulle et variance proportionnelle à x_i. En supposant que le modèle soit valable pour l'échantillon, c'est-à-dire sans biais d'échantillonnage, le meilleur prédicteur sans biais par rapport au modèle linéaire du total Y est donné par l'estimateur par quotient $(\bar{y}/\bar{x})X$ sans égard au plan de sondage. Cet estimateur n'est pas convergent selon le plan de sondage, sauf si le plan est autopondéré, par exemple sous échantillonnage aléatoire stratifié avec répartition proportionnelle. Par conséquent, sous des plans non autopondérés, il peut donner de très mauvais résultats dans le cas de grands échantillons, même si les écarts par rapport au modèle sont faibles. Hansen et coll. (1983) ont démontré les mauvais résultats obtenus dans des conditions d'échantillonnage répété, en utilisant un plan d'échantillonnage aléatoire stratifié avec une répartition de l'échantillon presque optimale (couramment utilisé en présence de populations très asymétriques). Rao (1996) a utilisé le même plan pour démontrer les mauvais résultats obtenus dans le contexte d'un cadre conditionnel pertinent à l'approche dépendante d'un modèle (Royall et Cumberland 1981). Néanmoins, les approches dépendantes d'un modèle peuvent jouer un rôle capital dans l'estimation sur petits domaines où la taille de l'échantillon dans un petit domaine peut être infime, voire nulle (voir la section 7).

Brewer (1963) a été le premier à proposer l'approche dépendante d'un modèle dans le contexte du modèle par quotient (2). Royall (1970) et ses collaborateurs ont mené une étude systématique de cette approche. Valliant, Dorfman et Royall (2000) donnent une description complète de la théorie, dont l'estimation de la variance (conditionnelle) par rapport au modèle de l'estimateur qui varie avec s; par exemple, sous le modèle par quotient (2), la variance par rapport au modèle dépend de la moyenne d'échantillon $\bar{x}_s$. Il est intéressant de noter que l'échantillonnage équilibré au moyen de la sélection raisonnée figure dans l'approche dépendante d'un modèle dans le contexte de la protection contre la spécification incorrecte du modèle (Royall et Herson 1973).

3.3 Approche assistée par un modèle

L'approche assistée par un modèle cherche à combiner les caractéristiques positives de la méthode fondée sur le plan de sondage et de la méthode dépendante d'un modèle. Elle considère uniquement les estimateurs convergents selon le plan de sondage du total Y qui sont aussi sans biais par rapport au modèle sous le modèle « de travail » hypothétique. Par exemple, sous le modèle par quotient (2), un estimateur assisté par un modèle de Y pour un plan d'échantillonnage probabiliste spécifié est donné par l'estimateur par quotient $\hat{Y}_r = (\hat{Y}_{\mathrm{NHT}}/\hat{X}_{\mathrm{NHT}})X$ qui est convergent selon le plan de sondage sans égard au modèle hypothétique. Hansen et coll. (1983) ont utilisé cet estimateur dans leur plan d'échantillonnage stratifié pour démontrer que ses résultats étaient supérieurs à ceux de l'estimateur dépendant d'un modèle $(\bar{y}/\bar{x})X$. Pour l'estimation de la variance, l'approche assistée par un modèle utilise des estimateurs convergents pour la variance de l'estimateur par rapport au plan tout en étant exactement ou asymptotiquement sans biais par rapport au modèle pour la variance par rapport au modèle. Toutefois, les inférences sont fondées sur le plan de sondage, car le modèle est utilisé uniquement comme modèle « de travail ».

Pour l'estimateur par quotient $\hat{Y}_r$ l'estimateur de variance est donné par

$$\mathrm{Var}(\hat{Y}_r) = (X/\hat{X}_{\mathrm{NHT}})^2 v(e), \qquad (3)$$

où, dans la notation d'opérateur, $v(e)$ est obtenu à partir de $v(y)$ en remplaçant y_i par les résidus $e_i = y_i - (\hat{Y}_{\mathrm{NHT}}/\hat{X}_{\mathrm{NHT}})x_i$. Cet estimateur de variance est asymptotiquement équivalent à un estimateur de linéarisation courant de la variance $v(e)$, mais il reflète le fait que l'information contenue dans l'échantillon varie avec $\hat{X}_{\mathrm{NHT}}$: les valeurs élevées produisent une faible variabilité, et les valeurs faibles, une grande variabilité. Le pivot normal ainsi obtenu produit des inférences dépendantes d'un modèle qui sont valides sous le modèle hypothétique (contrairement à l'utilisation de $v(e)$ dans le pivot) tout en protégeant contre les écarts par rapport au modèle, en ce sens qu'il produit des inférences asymptotiquement valides fondées sur le plan de sondage. Il est à noter que le pivot est asymptotiquement équivalent à $\hat{Y}(\tilde{e})/[v(\tilde{e})]^{1/2}$ avec $\tilde{e}_i = y_i - (Y/X)x_i$. Si les écarts par rapport au modèle sont faibles, l'asymétrie dans les résidus $\tilde{e}_i$ est faible même si y_i et x_i sont très asymétriques, et les intervalles de confiance normaux donnent de bons résultats. Par contre, pour des populations très asymétriques, les intervalles normaux fondés sur $\hat{Y}_{\mathrm{NHT}}$ et son erreur-type peuvent donner de mauvais résultats sous échantillonnage répété, même pour des échantillons assez grands, car le pivot dépend de

l'asymétrie des y_i. La structure de population joue donc un rôle dans les inférences fondées sur le plan de sondage, contrairement à ce qu'affirment Neyman (1934), Hansen et coll. (1983) et d'autres auteurs. Rao, Jocelyn et Hidiroglou (2003) ont considéré l'estimateur par régression linéaire simple sous échantillonnage aléatoire simple à deux phases avec seulement x observé dans la première phase. Ils ont démontré que le rendement de couverture des intervalles normaux associés pouvait être faible même pour des échantillons de deuxième phase passablement grands si le vrai modèle sous-jacent qui produisait la population s'écartait considérablement du modèle de régression linéaire (par exemple, une régression quadratique de y sur x) et si l'asymétrie de x est grande. Dans ce cas, les valeurs x de la première phase sont observées et une approche assistée par un modèle approprié utiliserait un estimateur par régression linéaire multiple avec x et $z = x^2$ comme variables auxiliaires. Il est à noter que pour l'échantillonnage à une seule phase, on ne peut mettre en œuvre un tel estimateur assisté par un modèle si l'on connaît uniquement les X, puisque l'estimateur dépend du total de population de z.

Särndal, Swenson et Wretman (1992) proposent une description complète de l'approche assistée par un modèle pour estimer le total Y d'une variable y sous le modèle de régression linéaire de travail

$$y_i = x_i'\beta + \varepsilon_i \; ; \; i = 1, \ldots, N \qquad (4)$$

avec moyenne nulle, erreurs non corrélées ε_i et variance par rapport au modèle $V_m(\varepsilon_i) = \sigma^2 q_i = \sigma_i^2$ où les q_i sont des constantes connues et les vecteurs x ont des totaux connus X (les valeurs de population $x_1, \ldots, x_N$ ne sont pas nécessairement connues). Dans ces conditions, l'approche assistée par un modèle produit l'estimateur de régression généralisée (generalized regression ou GREG)

$$\hat{Y}_{gr} \quad \hat{Y}_{\mathrm{NHT}} \quad \hat{B}\ (X \quad \hat{X}_{\mathrm{NHT}}) \ : \sum_{i \in s} w_i(s)\, y_i,$$

travail de $y^{(j)}$ prenne la forme (4) mais nécessite un vecteur x peut-être différent $x^{(j)}$ avec total connu $X^{(j)}$ pour chaque $j = 1, ..., p$:

$$y_i^{(j)} = x_i^{(j)'}\beta^{(j)} + \varepsilon_i^{(j)},\ i = 1, ..., N. \qquad (7)$$

Dans ce cas, les poids g dépendent de j et, à leur tour, les poids finaux $w_i(s)$ dépendent aussi de j. Dans la pratique, il est souvent souhaitable d'utiliser un seul ensemble de poids finaux pour toutes les variables p afin d'assurer la cohérence interne des chiffres lorsqu'ils sont agrégés à partir de variables différentes. On ne peut réaliser cette propriété qu'en élargissant le vecteur x dans le modèle (7) pour recevoir toutes les variables $y^{(j)}$, par exemple $\tilde{x}$ avec total connu $\tilde{X}$, puis en utilisant le modèle de travail

$$y_i^{(j)} = \tilde{x}_i'\beta^{(j)} + \varepsilon_i^{(j)},\ i = 1, ..., N. \qquad (8)$$

Toutefois, les coefficients de régression pondérés ainsi obtenus pourraient devenir instables à cause du risque de multicolinéarité dans l'ensemble élargi de variables auxiliaires. Par conséquent, l'estimateur GREG de $Y^{(j)}$ sous le modèle (8) est moins efficient que l'estimateur GREG sous le modèle (7). En outre, certains poids finaux ainsi obtenus, par exemple $\tilde{w}_i(s)$, risquent de ne pas satisfaire les restrictions relatives à l'étendue en prenant des valeurs inférieures à 1 (dont des valeurs négatives) ou de très grandes valeurs positives. Il est possible de résoudre ce problème en utilisant un estimateur par régression ridge généralisée de $Y^{(j)}$ qui est assisté par un modèle sous le modèle élargi (Chambers 1996; Rao et Singh 1997).

Pour l'estimation de la variance, l'approche assistée par un modèle cherche à utiliser des estimateurs de variance convergents selon le plan de sondage qui sont aussi sans biais par rapport au modèle (du moins pour les grands échantillons) en ce qui concerne la variance conditionnelle par rapport au modèle de l'estimateur GREG. Dénotant l'estimateur de variance de l'estimateur NHT de Y par $v(y)$ dans une notation d'opérateur, un estimateur de variance par linéarisation de Taylor simple satisfaisant la propriété susmentionnée est donné par $v(ge)$, où l'on obtient $v(ge)$ en remplaçant y_i par $g_i(s)e_i$ dans la formule de $v(y)$; voir Hidiroglou, Fuller et Hickman (1976) et Särndal, Swenson et Wretman (1989).

Dans l'exposé qui précède, nous avons supposé un modèle de régression linéaire de travail pour toutes les variables $y^{(j)}$. Dans la pratique, cependant, un modèle de régression linéaire n'est pas nécessairement bien adapté à certaines variables d'intérêt y, par exemple, une variable binaire. Dans ce dernier cas, la régression logistique offre un modèle de travail approprié. Un modèle de travail général qui couvre la régression logistique prend la forme $E_m(y_i) = h(x_i'\beta) = \mu_i$, où $h(.)$ pourrait être non linéaire; le modèle (5) est un cas particulier avec $h(a) = a$. Un estimateur assisté par un modèle du total sous le modèle de travail général est l'estimateur par la différence $\hat{Y}_{\text{NHT}} + \sum_U \hat{\mu}_i - \sum_s \pi_i^{-1}\hat{\mu}_i$, où $\hat{\mu}_i = h(x_i'\hat{\beta})$ et $\hat{\beta}$ est un estimateur du paramètre de modélisation β. Il se réduit à l'estimateur GREG (5) si $h(a) = a$. Cet estimateur par la différence est presque optimal si la probabilité d'inclusion π_i est proportionnelle à σ_i, où σ_i^2 dénote la variance par rapport au modèle, $V_m(y_i)$.

Les estimateurs GREG sont très appréciés par les utilisateurs parce que bon nombre d'estimateurs couramment utilisés peuvent être obtenus sous forme de cas particuliers de (5) par des spécifications appropriées de x_i et q_i. Statistique Canada a mis au point un Système généralisé d'estimation (SGE) fondé sur l'estimateur GREG.

Kott (2005) a proposé un autre paradigme de l'inférence, appelé approche fondée sur un modèle et assistée par randomisation, qui est axé sur l'inférence fondée sur un modèle et assistée par randomisation (ou échantillonnage répété). La définition de la variance prévue est inversée pour devenir la variance prévue par rapport au modèle à randomisation d'un estimateur, mais elle est identique à la variance prévue habituelle lorsque le modèle de travail est valable pour l'échantillon, comme on le suppose dans l'article. Par conséquent, les choix de l'estimateur et de l'estimateur de variance sont souvent semblables à ceux qui sont faits sous l'approche assistée par un modèle. Toutefois, Kott soutient que la motivation est plus

claire et que « l'approche proposée ici pour l'estimation de la variance mène, au besoin, à un traitement logiquement cohérent de rajustements d'une population finie et d'un petit échantillon ».

3.4 Approche conditionnelle fondée sur le plan de sondage

On a également proposé une approche conditionnelle fondée sur le plan de sondage. Cette approche cherche à combiner les caractéristiques conditionnelles de l'approche dépendante d'un modèle avec les caractéristiques indépendantes de l'approche fondée sur le plan de sondage. Elle permet de restreindre l'ensemble d'échantillons de référence à un sous-ensemble « pertinent » de tous les échantillons possibles spécifiés par le plan de sondage. On obtient des inférences conditionnellement valides en ce sens que le ratio de biais conditionnel (soit le ratio du biais conditionnel à l'erreur-type conditionnelle) devient nul à mesure que la taille de l'échantillon augmente. Environ $100(1-\alpha)\%$ des intervalles de confiance réalisés dans l'échantillonnage répété à partir de l'ensemble conditionnel contiennent le total inconnu Y.

Holt et Smith (1979) fournissent des arguments convaincants en faveur de l'inférence conditionnelle fondée sur le plan, même si leur analyse est limitée à la post-stratification simple d'un échantillon aléatoire simple, auquel cas il est naturel de faire des inférences conditionnelles à la taille des strates de l'échantillon réalisé. Rao (1992, 1994) et Casady et Valliant (1993) ont étudié l'inférence conditionnelle lorsque seul le total auxiliaire X est connu d'après des sources externes. Dans ce dernier cas, la subordination à l'estimateur NHT $\hat{X}_{\text{NHT}}$ peut s'avérer raisonnable parce qu'il s'agit « à peu près » d'une statistique auxiliaire lorsque X est connu et que la différence $\hat{X}_{\text{NHT}} - X$ fournit une mesure du déséquilibre de l'échantillon réalisé. La subordination à $\hat{X}_{\text{NHT}}$ permet de calculer l'estimateur par régression linéaire « optimal », de même forme que l'estimateur GREG (5), dans lequel $\hat{B}$ donné par (6) est remplacé par la valeur optimale estimative $\hat{B}_{\text{opt}}$ du coefficient de régression qui fait intervenir la covariance estimative de $\hat{Y}_{\text{NHT}}$ et $\hat{X}_{\text{NHT}}$ et la variance estimative de $\hat{X}_{\text{NHT}}$. Cet estimateur optimal permet d'établir des inférences conditionnellement valides fondées sur le plan de sondage et est sans biais par rapport au modèle sous le modèle de travail (4). Il s'agit également d'un estimateur par calage dépendant uniquement du total X et il peut être exprimé comme suit : $\sum_{i \in s} \tilde{w}_i(s) y_i$ avec poids $\tilde{w}_i(s) = d_i \tilde{g}_i(s)$ et le facteur de calage $\tilde{g}_i(s)$ dépendant uniquement du total X et les valeurs x de l'échantillon. Il fonctionne bien pour l'échantillonnage aléatoire stratifié (couramment utilisé dans les enquêtes-établissements). Toutefois, $\hat{B}_{\text{opt}}$ peut devenir instable dans le cas de l'échantillonnage stratifié à plusieurs degrés, sauf si la différence entre le nombre de grappes d'échantillon et le nombre de strates est passablement élevée. L'estimateur GREG n'exige pas cette dernière condition, mais il peut donner de mauvais résultats en ce qui concerne le ratio de biais conditionnel et les taux de couverture conditionnels, comme l'a montré Rao (1996). L'estimateur NHT sans biais peut être conditionnellement très mauvais, sauf si le plan assure que la mesure du déséquilibre définie plus haut est faible. Par exemple, dans le plan de sondage fondé sur la stratification x efficiente et proposé par Hansen et coll. (1983), le déséquilibre est faible et l'estimateur NHT a donné conditionnellement de bons résultats.

Tillé (1998) a proposé un estimateur NHT du total Y fondé sur des probabilités d'inclusion conditionnelles approximatives en présence de $\hat{X}_{\text{NHT}}$. Sa méthode permet également d'établir des inférences conditionnellement valides, mais l'estimateur n'est pas calé en fonction de X, contrairement à l'estimateur par régression linéaire « optimal ». Park et Fuller (2005) ont proposé une version calée de l'estimateur GREG fondée sur l'estimateur de Tillé qui donne des poids non négatifs plus souvent que l'estimateur GREG.

Je crois que les praticiens devraient accorder une plus grande attention aux aspects conditionnels de l'inférence fondée sur le plan de sondage et envisager sérieusement les nouvelles méthodes qui ont été proposées.

Kalton (2002) a donné des arguments convaincants pour favoriser des approches fondées sur le plan de sondage (et peut-être conditionnelles ou assistées par un modèle) de l'inférence en fonction des paramètres descriptifs d'une population finie. Smith (1994) a nommé « inférence procédurale » l'inférence fondée sur le plan de sondage et a soutenu qu'il s'agissait de l'approche à adopter pour les enquêtes du domaine public. Le lecteur trouvera dans Smith (1976) et Rao et Bellhouse (1990) des études des questions d'inférence dans la théorie des enquêtes par sondage.

4. Estimateurs par calage

On obtient les poids de calage $w_i(s)$ qui assurent la cohérence avec les totaux auxiliaires X spécifiés par l'utilisateur en rajustant les poids de sondage $d_i = \pi_i^{-1}$ pour satisfaire les contraintes d'étalonnage $\sum_{i \in s} w_i(s) x_i = X$. Les estimateurs qui utilisent des poids de calage sont appelés estimateurs par calage et utilisent un seul ensemble de poids $\{w_i(s)\}$ pour toutes les variables d'intérêt. Nous avons mentionné dans la section 3.4 que l'estimateur GREG assisté par un modèle était un estimateur par calage, mais un estimateur par calage n'est pas nécessairement assisté par un modèle, en ce sens qu'il risque d'être biaisé par rapport au modèle sous un modèle de travail (4), sauf si les variables x du modèle coïncident exactement avec les variables correspondant aux totaux spécifiés par l'utilisateur. Par exemple, supposons que le modèle de travail suggéré par les données soit un modèle quadratique dans une variable scalaire x alors que le total spécifié par l'utilisateur est uniquement son total X. L'estimateur par calage ainsi obtenu peut donner de mauvais résultats même dans des échantillons assez grands, comme nous l'avons mentionné dans la section 3.3, contrairement à l'estimateur GREG assisté par un modèle fondé sur le modèle quadratique de travail qui nécessite le total de population des variables quadratiques x_i^2 en plus de X.

Dans la pratique, on utilise abondamment la post-stratification pour assurer la cohérence avec les valeurs connues de la cellule correspondant à une variable de post-stratification, par exemple des valeurs dans différents groupes d'âge vérifiées d'après des sources externes comme des projections démographiques. L'estimateur post-stratifié ainsi obtenu est un estimateur par calage. On a également utilisé dans la pratique des estimateurs par calage qui assurent la cohérence avec les valeurs marginales connues de deux ou plusieurs variables de post-stratification, notamment les estimateurs de la méthode itérative du quotient, qu'on obtient par étalonnage répété des valeurs marginales jusqu'à ce que la convergence soit approximativement réalisée, habituellement en quatre itérations ou moins. Les poids obtenus par la méthode itérative du quotient $w_i(s)$ sont toujours positifs. Dans le cadre du Recensement du Canada, Statistique Canada a déjà utilisé les estimateurs de la méthode itérative du quotient pour assurer la cohérence des estimateurs de données-échantillon (2B) avec les valeurs connues des données intégrales (2A). Toujours dans le contexte du Recensement du Canada, Brackstone et Rao (1979) ont étudié l'efficience des estimateurs de la méthode itérative du quotient et ont aussi calculé des estimateurs de variance par linéarisation de Taylor lorsque le nombre d'itérations était de quatre ou moins. On a également employé les estimateurs de la méthode itérative du quotient dans la Current Population Survey (CPS) des États-Unis. Il convient de noter que la méthode de rajustement des valeurs de la cellule en fonction des valeurs marginales données dans un tableau à double entrée a d'abord été proposée dans l'article marquant de Deming et Stephan (1940).

Des approches unifiées du calage, fondées sur la minimisation d'une mesure appropriée de la distance entre les poids de calage et les poids de sondage sous réserve des contraintes d'étalonnage, ont attiré l'attention des utilisateurs en raison de leur capacité de recevoir un nombre arbitraire de contraintes d'étalonnage spécifiées par l'utilisateur, par exemple, le calage en fonction des valeurs marginales de plusieurs variables de post-stratification. Des logiciels de calage sont également disponibles, dont le SGE (Statistique Canada), LIN WEIGHT (Bureau national de la statistique des Pays-Bas), CALMAR (INSEE, France) et CLAN97 (Bureau de la statistique de Suède).

Une distance de chi carré, $\sum_{i \in s} q_i (d_i - w_i)^2 / d_i$, permet de calculer l'estimateur GREG (5), où le vecteur x – correspond aux contraintes d'étalonnage spécifiées par l'utilisateur et $w_i(s)$ est dénoté w_i par souci de simplicité (Huang et Fuller 1978; Deville et Särndal 1992). Toutefois, les poids de calage ainsi obtenus ne satisfont pas nécessairement les restrictions relatives à l'étendue souhaitable; par exemple, certains poids peuvent être négatifs ou trop grands, surtout lorsque le nombre de contraintes est élevé et que la variabilité des poids de sondage est élevée. Huang et Fuller (1978) ont proposé une mesure de distance de chi carré modifiée à l'échelle et obtenu les poids de calage au moyen d'une solution itérative qui satisfait les contraintes d'étalonnage à chaque itération. Toutefois, il n'existe peut-être pas de solution qui satisfait à la fois les contraintes d'étalonnage et les contraintes relatives à l'étendue. Une autre méthode, appelée minimisation par rétrécissement (Singh et Mohl 1996), se heurte à la même difficulté. On a également proposé des méthodes de programmation quadratique qui minimisent la distance de chi carré sous réserve des contraintes d'étalonnage et des contraintes relatives à l'étendue (Hussain 1969), mais l'ensemble de solutions réalisables satisfaisant les deux types de contrainte peut être vide. D'autres méthodes proposées consistent à modifier la fonction de distance (Deville et Särndal

1992) ou à abandonner certaines contraintes d'étalonnage (Bankier, Rathwell et Majkowski 1992). Par exemple, une distance d'information de forme $\sum_{i\in s} q_i \{w_i \log(w_i/d_i) - w_i + d_i\}$ donne des estimateurs de la méthode itérative du quotient avec poids non négatifs w_i, mais certains poids peuvent être beaucoup trop grands. On a également proposé des poids « ridge » obtenus en minimisant une distance de chi carré pénalisée (Chambers 1996), mais rien ne garantit qu'ils satisfont les contraintes d'étalonnage ou les contraintes relatives à l'étendue, quoique les poids soient plus stables que les poids GREG. Rao et Singh (1997) ont proposé une méthode itérative de « rétrécissement ridge » qui assure la convergence pour un nombre spécifié d'itérations en utilisant une spécification de tolérance intégrée pour assouplir certaines contraintes d'étalonnage tout en satisfaisant les contraintes relatives à l'étendue. Chen, Sitter et Wu (2002) ont proposé une méthode semblable.

On a utilisé les poids de calage GREG dans l'Enquête sur la population active du Canada qui, tout récemment, a fait appel à des estimateurs composites qui utilisent l'information des mois antérieurs sur l'échantillon, comme nous l'avons mentionné dans la section 2 (Fuller et Rao 2001; Gambino, Kennedy et Singh 2001; Singh, Kennedy et Wu 2001). On a également utilisé des estimateurs par calage de type GREG pour intégrer deux ou plusieurs enquêtes indépendantes portant sur la même population. Ces estimateurs assurent la cohérence entre les enquêtes, en ce sens que les estimateurs de variables communes aux deux enquêtes sont identiques, ainsi que l'étalonnage en fonction de totaux de population connus (Renssen et Nieuwenbroek 1997; Singh et Wu 1996; Merkouris 2004). Pour le Recensement du Canada de 2001, Bankier (2003) a étudié des poids de calage correspondant à l'estimateur par régression linéaire « optimal » (section 3.3) sous échantillonnage aléatoire stratifié. Il a montré que la méthode de calage « optimale » donnait de meilleurs résultats que l'estimateur par calage GREG, utilisé lors du recensement précédent, dans la mesure où elle permettait de conserver plus de contraintes d'étalonnage tout en permettant aux poids de calage d'être au moins un. On peut obtenir le poids de calage « optimal » à l'aide du logiciel SGE en précisant dans les contraintes d'étalonnage la taille connue des strates et en définissant comme il se doit la constante de réglage q_i. Il est à noter que l'estimateur par calage « optimal » possède également des propriétés conditionnelles souhaitables par rapport au plan (section 3.4). Pour la pondération des données du Recensement du Canada de 2001, la méthode de la régression linéaire « optimale » a remplacé celle de l'estimateur GREG par projection (utilisée lors du Recensement de 1996).

Demnati et Rao (2004) ont calculé des estimateurs de variance par linéarisation de Taylor pour une classe générale d'estimateurs par calage avec poids $w_i = d_i F(x_i'\hat{\lambda})$, où l'on détermine le multiplicateur de LaGrange $\hat{\lambda}$ en résolvant les contraintes de calage. Le choix $F(a) = 1 + a$ donne des poids GREG et $F(a) = e^a$ permet de calculer des poids obtenus par la méthode itérative du quotient. Dans le cas particulier des poids GREG, l'estimateur de variance se réduit à $v(ge)$ donné dans la section 3.3.

Le lecteur trouvera dans l'article de Fuller (2002), récipiendaire du prix Waksberg, un aperçu et une évaluation très éloquents de l'estimation par régression dans l'échantillonnage d'enquête, y compris l'estimation par calage.

5. ÉCHANTILLONNAGE AVEC PROBABILITÉS INÉGALES SANS REMISE

Nous avons mentionné dans la section 2 que l'échantillonnage PPT d'UPÉ à l'intérieur de strates dans les enquêtes à grande échelle était motivé par des considérations pratiques, soit la volonté de répartir des charges de travail à peu près égales. L'échantillonnage PPT permet également de réduire considérablement la variance en neutralisant la variabilité découlant de la taille inégale des UPÉ sans vraiment stratifier par taille. Les UPÉ sont habituellement échantillonnées sans remise, de manière que la probabilité d'inclusion des UPÉ, π_i, soit proportionnelle à la mesure de la taille des UPÉ x_i. Par exemple, l'échantillonnage PPT systématique, avec ou sans randomisation initiale des étiquettes UPÉ, est un plan avec probabilité d'inclusion proportionnelle à la taille (PIPT) (appelé aussi plan πPT) utilisé dans un grand nombre d'enquêtes complexes, dont l'EPA du Canada. L'estimateur d'un total associé à un plan PIPT est l'estimateur NHT.

L'élaboration de stratégies appropriées (PIPT, NHT) soulève des problèmes sur le plan théorique, dont l'évaluation de probabilités d'inclusion conjointes exactes, π_{ij}, ou des approximations exactes de π_{ij} nécessitant uniquement les π_i

individuels, qui sont nécessaires pour obtenir un estimateur de variance sans biais ou presque sans biais. J'ai étudié ce dernier problème dans la thèse de doctorat que j'ai présentée en 1961 à la Iowa State University. D'éminents statisticiens-mathématiciens ont publié depuis plusieurs solutions nécessitant des outils théoriques perfectionnés. Toutefois, ces travaux théoriques sont souvent qualifiés de « théorie sans application » puisque, dans la pratique, il est courant de traiter les UPÉ comme si elles étaient échantillonnées avec remise, d'où une grande simplification. L'estimateur de variance est obtenu simplement à partir des totaux estimatifs d'UPÉ; cette hypothèse est d'ailleurs à la base des méthodes de rééchantillonnage (section 6). Cet estimateur de variance peut entraîner une surestimation substantielle, sauf si la fraction d'échantillonnage des UPÉ globales est faible, ce qui peut être vrai dans bon nombre d'enquêtes à grande échelle. Dans les paragraphes qui suivent, je tenterai de démontrer que les travaux théoriques portant sur certaines stratégies (PIPT, NHT) et sur des plans de sondage sans PIPT ont une grande applicabilité dans la pratique.

J'aborderai d'abord certaines stratégies (PIPT, NHT). En Suède et dans d'autres pays européens, on utilise souvent l'échantillonnage stratifié à un seul degré en raison de la disponibilité de listes et les plans PIPT sont des options attrayantes, mais les fractions d'échantillonnage sont souvent grandes. Par exemple, Rosén (1991) mentionne que le baromètre de la population active du Bureau de la statistique de Suède échantillonne une centaine de populations différentes en utilisant l'échantillonnage PPT systématique et que les taux d'échantillonnage peuvent dépasser 50 %. Aires et Rosén (2005) ont étudié l'échantillonnage πPT de Pareto pour les enquêtes suédoises. Cette méthode possède des propriétés attrayantes, dont la taille fixe de l'échantillon, l'échantillonnage simple, une bonne précision d'estimation, et une estimation convergente de la variance sans égard aux taux d'échantillonnage. En outre, elle permet de coordonner les échantillons au moyen de nombres aléatoires permanents (NAP), comme dans l'échantillonnage de Poisson, mais cette dernière méthode produit des échantillons de taille variable. En raison de ces mérites, on a mis en œuvre l'échantillonnage πPT de Pareto dans un certain nombre d'enquêtes du Bureau de la statistique de Suède, notamment dans les enquêtes sur l'indice des prix. Ohlsson (1995) a décrit les techniques des NAP qui sont couramment utilisées dans la pratique.

La méthode de Rao-Sampford (voir Brewer et Hanif 1983, page 28) produit des plans PIPT exacts et des estimateurs de variance non négatifs sans biais pour des échantillons de taille fixe arbitraire. Elle a été mise en œuvre dans la nouvelle version du SAS. Stehman et Overton (1994) notent que la structure de la probabilité variable se manifeste naturellement dans les enquêtes environnementales au lieu d'être sélectionnée uniquement pour l'efficience accrue et que les π_i sont connus uniquement pour les unités i de l'échantillon s. En traitant le plan de sondage selon la méthode d'échantillonnage systématique aléatoire avec PPT, Stehman et Overton ont obtenu des approximations des π_{ij} qui dépendent uniquement des π_i, $i \in s$, contrairement aux approximations initiales de Hartley et Rao (1962) qui nécessitent la somme des carrés de tous les π_i de la population. Dans les applications de Stehman et Overton, les taux d'échantillonnage sont assez substantiels pour justifier l'évaluation des probabilités d'inclusion conjointes.

Je vais maintenant aborder les plans sans PIPT utilisant des estimateurs différents de l'estimateur NHT qui assure une variance nulle lorsque y est exactement proportionnel à x. La méthode des groupes aléatoires de Rao, Hartley et Cochran (1962) permet de calculer un estimateur de variance non négatif simple pour n'importe quelle taille fixe de l'échantillon; pourtant, elle se compare favorablement aux stratégies (PIPT, NHT) sur le plan de l'efficience et elle est toujours plus efficiente que la stratégie PPT avec remise. Schabenberger et Grégoire (1994) ont constaté que les stratégies (PIPT, NHT) n'avaient pas trouvé beaucoup d'applications en foresterie à cause de la difficulté de mise en œuvre et ont recommandé la stratégie de Rao-Hartley-Cochran en raison de sa remarquable simplicité et de son efficience. Il est intéressant de constater que cette stratégie a été utilisée dans l'EPA du Canada parce qu'elle permettait d'adopter de nouvelles mesures de la taille en utilisant la méthode de Keyfitz à l'intérieur de chaque groupe aléatoire. Par contre, les stratégies (PIPT, NHT) ne conviennent pas tellement à cette fin (Fellegi 1966). Je crois savoir qu'on utilise souvent la stratégie de Rao-Hartley-Cochran en contrôle par sondage et dans d'autres applications comptables.

Murthy (1957) a utilisé un plan sans PIPT fondé sur le tirage d'unités successives avec probabilités $p_i, p_j / (1 - p_i)$, $p_k / (1 - p_i - p_j)$ et ainsi de suite, et l'estimateur suivant :

$$\hat{Y}_M = \sum_{i \in s} y_i \frac{p(s \mid i)}{p(s)}, \qquad (9)$$

où $p(s|i)$ est la probabilité conditionnelle d'obtenir l'échantillon s lorsque l'unité i a été sélectionnée en premier. Il a également proposé un estimateur de variance non négatif nécessitant les probabilités conditionnelles, $p(s|i, j)$, d'obtenir s lorsque i et j sont sélectionnés dans les deux premiers tirages. Pendant plusieurs années, les praticiens ont accordé peu d'attention à cette méthode à cause de la complexité des calculs mais, plus récemment, on l'a appliquée dans des domaines inattendus, dont la découverte de pétrole (Andreatta et Kaufmann 1986) et l'échantillonnage séquentiel, dont l'échantillonnage inverse et certains schémas d'échantillonnage adaptable (Salehi et Seber 1997). Il convient de noter qu'au cours des dernières années, on s'est beaucoup intéressé à l'échantillonnage adaptable puisqu'il s'agit d'une méthode d'échantillonnage efficiente pour estimer des totaux ou des moyennes de populations rares (Thompson et Seber 1996). Dans son application à la découverte de pétrole, le schéma d'échantillonnage successif est une caractérisation de la découverte et l'ordre dans lequel les champs de pétrole sont découverts est déterminé par l'échantillonnage proportionnel à la taille des champs et sans remise, selon un vieux principe de l'industrie : « en moyenne, on trouve d'abord les grands champs ». Ici, $p_i = y_i / Y$ et la réserve de pétrole totale Y est présumée connue d'après des critères géologiques. Dans cette application, les géologues s'intéressent à la distribution par taille de tous les champs du bassin et, après l'exploration partielle d'un bassin, l'échantillon est composé de grandeurs y_i de dépôts découverts. On peut estimer la fonction de distribution par taille $F(a)$ en utilisant l'estimateur de Murthy (9) dans lequel y_i est remplacé par la variable indicatrice $I(y_i \le a)$. Le calcul de $p(s|i)$ et $p(s)$, toutefois, est très complexe, même pour des échantillons de taille moyenne. Afin de surmonter cette difficulté de calcul, Andreatta et Kaufman (1986) ont utilisé des représentations intégrales de ces quantités pour formuler des expressions asymptotiques de l'estimateur de Murthy, dont les premiers termes sont aisés à calculer. De même, ils obtiennent des approximations calculables de l'estimateur de variance de Murthy. Il est à noter qu'on ne peut employer ici l'estimateur NHT de $F(a)$, car les probabilités d'inclusion sont des fonctions de toutes les valeurs y de la population.

L'exposé qui précède vise à démontrer qu'une théorie donnée peut avoir des applications dans divers secteurs pratiques même si elle n'est pas nécessaire dans une situation donnée, comme les enquêtes à grande échelle avec fractions d'échantillonnage du premier degré négligeables. Il montre également que les plans d'échantillonnage avec probabilités inégales jouent un rôle essentiel dans l'échantillonnage d'enquête, malgré l'affirmation de Särndal (1996) selon laquelle des plans simples, comme l'ÉAS stratifié et l'échantillonnage stratifié de Bernoulli, ainsi que les estimateurs GREG, devraient remplacer les stratégies fondées sur l'échantillonnage avec probabilités inégales sans remise.

6. ANALYSE DES DONNÉES D'ENQUÊTE ET DES MÉTHODES DE RÉÉCHANTILLONNAGE

Les méthodes-types d'analyse des données sont généralement fondées sur l'hypothèse de l'échantillonnage aléatoire simple, quoique certains progiciels tiennent compte des poids d'échantillonnage et fournissent des estimations ponctuelles correctes. Toutefois, l'application de méthodes-types aux données d'enquête, abstraction faite de l'effet du plan de sondage dû à la mise en grappes et aux probabilités de sélection inégales, risque de produire des inférences erronées, même pour de grands échantillons. En particulier, les erreurs-types des estimations de paramètres et des intervalles de confiance associés peuvent être lourdement sous-estimées, les taux d'erreur de type I des tests d'hypothèses peuvent être beaucoup plus élevés que les niveaux nominaux et les diagnostics de modèles-types, comme l'analyse des résidus pour déceler les écarts par rapport au modèle, sont aussi influencés. Kish et Frankel (1974) et d'autres auteurs se sont penchés sur certains de ces problèmes et ont souligné la nécessité de nouvelles méthodes qui tiennent suffisamment compte de la complexité des données provenant d'enquêtes à grande échelle. Fuller (1975) a mis au point des méthodes asymptotiquement valides d'analyse par régression linéaire, fondées sur des estimateurs de variance par linéarisation de Taylor. Au cours des vingt dernières années, on a fait des progrès rapides en mettant au point des méthodes appropriées. Les méthodes de rééchantillonnage jouent un rôle capital dans la mise au point de méthodes qui tiennent compte du plan d'enquête dans l'analyse des données. On a simplement besoin d'un fichier de données contenant les données observées, des poids d'échantillonnage finaux et des poids finaux correspondant à chaque pseudo-répétition produit par la méthode de rééchantillonnage. On peut alors utiliser des progiciels qui tiennent compte des poids d'échantillonnage dans l'estimation ponctuelle des paramètres d'intérêt pour calculer les bons estimateurs et les erreurs-types, comme nous le démontrons ci-

dessous. Les méthodes d'inférence par rééchantillonnage ont donc attiré l'attention des utilisateurs, qui peuvent très facilement effectuer les analyses eux-mêmes à l'aide de progiciels standard. Toutefois, la mise en circulation de fichiers de données à grande diffusion avec poids de rééchantillonnage risque d'entraîner des problèmes de confidentialité, comme l'identification des grappes à partir des poids de rééchantillonnage. Les théoriciens ont d'ailleurs un défi à relever : celui de mettre au point des méthodes appropriées qui préservent la confidentialité des données. Lu, Brick et Sitter (2004) ont proposé de regrouper les strates et de former des pseudo-répétitions en utilisant les strates combinées pour l'estimation de la variance, limitant ainsi le risque d'identification des grappes à partir du fichier de données à grande diffusion ainsi obtenu. Le groupement des strates ou des UPÉ à l'intérieur des strates simplifie l'estimation de la variance en réduisant le nombre de pseudo-répétitions utilisés, comparativement à la méthode jackknife avec suppression d'une grappe, qu'on utilise couramment et que nous abordons ci-dessous. Une méthode d'échantillonnage inverse servant à défaire la structure complexe des données d'enquête tout en offrant une protection contre la révélation des étiquettes de grappe (Hinkins, Oh et Scheuren 1997; Rao, Scott et Benhin 2003) semble prometteuse, mais il reste beaucoup de travail à accomplir sur les méthodes d'échantillonnage inverse avant qu'elle n'intéresse l'utilisateur.

Rao et Scott (1981, 1984) ont mené une étude systématique de l'effet du plan de sondage sur le test chi carré et le test du rapport des vraisemblances, tests standardisés associés à un tableau multiple de comptes estimatifs ou de proportions. Ils ont montré que la variable à tester était asymptotiquement distribuée sous forme de somme pondérée de variables χ_1^2 indépendantes, les poids étant les valeurs propres d'une matrice d'« effets généralisés du plan de sondage ». Ce résultat général montre que le plan d'enquête peut avoir un effet important sur le taux d'erreur de type I. Rao et Scott ont proposé des corrections simples du premier ordre aux statistiques chi carré standardisées, qu'on peut calculer à partir de tableaux publiés comprenant des estimations des effets du plan de sondage pour les cellules d'estimations et leurs totaux marginaux, ce qui facilite les analyses secondaires à partir de tableaux publiés. Ils ont également calculé des corrections du deuxième ordre qui sont plus exactes, mais qui nécessitent la connaissance d'une matrice complète des covariances estimatives des cellules d'estimations, comme dans le cas des tests de Wald, bien connus. Toutefois, les tests de Wald peuvent devenir très instables lorsque le nombre de cellules d'un tableau multiple augmente et que le nombre de grappes d'échantillon diminue, ce qui entraîne des taux d'erreur de type I démesurément élevés par rapport aux niveaux nominaux, contrairement aux corrections du deuxième ordre de Rao-Scott (Thomas et Rao 1987). Les corrections du premier et du deuxième ordre sont maintenant appelées corrections de Rao-Scott et constituent des options par défaut dans la nouvelle version du SAS. Roberts, Rao et Kumar (1987) ont mis au point des corrections du type Rao-Scott pour les tests d'analyse de régression logistique des proportions estimatives des cellules associées à une variable de réponse binaire. Ils ont appliqué les méthodes à un tableau à double entrée de taux d'emploi provenant de l'EPA du Canada de 1977, obtenus en recoupant des groupes d'âge et de niveau de scolarité. Bellhouse et Rao (2002) ont étendu les travaux de Roberts et coll. à l'analyse des moyennes de domaine à l'aide de modèles linéaires généralisés. Ils ont appliqué les méthodes aux moyennes de domaine provenant d'une enquête sur la fécondité menée au Fidji, recoupées par niveau de scolarité et par nombre d'années depuis le premier mariage de la femme, une moyenne de domaine étant le nombre moyen d'enfants nés de femmes de race indienne appartenant au domaine.

Dans le contexte des enquêtes à grande échelle utilisant des plans d'échantillonnage stratifié à plusieurs degrés, les méthodes de rééchantillonnage ont fait l'objet de nombreuses études. Pour les besoins de l'inférence, les UPÉ de l'échantillon sont traitées comme si elles étaient tirées avec remise à l'intérieur des strates. Les variances s'en trouvent surestimées, mais cette surestimation est faible si la fraction d'échantillonnage globale des UPÉ est négligeable. Soit $\hat{\theta}$ l'estimateur pondéré d'un paramètre « de recensement » d'intérêt, calculé d'après les poids finaux w_i, et soient les poids correspondant à chaque pseudo-répétition r produits par la méthode de rééchantillonnage dénotés par $w_i^{(r)}$. L'estimateur fondé sur les pseudo-poids de rééchantillonnage $w_i^{(r)}$ est dénoté $\hat{\theta}^{(r)}$ pour chaque $r = 1, ..., R$. Un estimateur de variance par rééchantillonnage de $\hat{\theta}$ prend alors la forme

$$v(\hat{\theta}) = \sum_{r=1}^{R} c_r (\hat{\theta}^{(r)} - \hat{\theta})(\hat{\theta}^{(r)} - \hat{\theta})' \qquad (10)$$

pour les coefficients spécifiés c_r dans (10) déterminés par la méthode de rééchantillonnage.

Les méthodes de rééchantillonnage couramment utilisées comprennent : a) le jackknife avec suppression d'une grappe (ou d'une UPÉ), b) la répétition compensée (*balanced repeated replicate* ou BRR), notamment pour $n_h = 2$ UPÉ dans chaque strate h, et c) le bootstrap de Rao et Wu (1988). On obtient les pseudo-répétitions jackknife en supprimant tour à tour chaque grappe d'échantillon $r = (hj)$, et les poids de sondage jackknife $d_i^{(r)}$ prennent la valeur 0 si l'unité d'échantillonnage i est dans la grappe supprimée, $n_h d_i /(n_h - 1)$ si i n'est pas dans la grappe supprimée mais dans la même strate, et restent inchangés si i est dans une strate différente. Les poids de sondage jackknife sont alors rajustés pour la non-réponse totale et la post-stratification, ce qui donne les poids jackknife finaux $w_i^{(r)}$. L'estimateur jackknife de la variance est donné par (10) avec $c_r = (n_h - 1)/n_h$ pour $r = (hj)$. La méthode du jackknife avec suppression d'une grappe peut présenter deux inconvénients : 1) lorsque le nombre total d'UPÉ échantillonnées, $n = \sum n_h$, est très élevé, R est aussi très élevé parce que $R = n$; 2) on ignore si l'estimateur jackknife de la variance avec suppression d'une grappe est convergent selon le plan de sondage dans le cas d'estimateurs non lisses $\hat{\theta}$, par exemple, l'estimateur pondéré de la médiane. Pour l'échantillonnage aléatoire simple, on sait que le jackknife est non convergent pour la médiane ou d'autres quantiles. Il serait stimulant (sur le plan théorique) et pertinent (sur le plan pratique) de trouver les conditions de convergence de l'estimateur jackknife de la variance avec suppression d'une grappe d'un estimateur non lisse $\hat{\theta}$.

La méthode BRR peut convenir à l'estimateur non lisse $\hat{\theta}$, mais elle ne s'applique aisément qu'à un cas particulier important, celui de $n_h = 2$ UPÉ par strate. On peut construire un ensemble minimal de demi-échantillons équilibrés à partir d'une matrice Hadamard $R \times R$ en sélectionnant H colonnes, à l'exclusion de la colonne des $+1$, où $H + 1 \leq R \leq H + 4$ (McCarthy 1969). Les poids de sondage BRR $d_i^{(r)}$ égalent $2d_i$ ou 0 selon que i se trouve ou non dans le demi-échantillon. Contrairement à la méthode BRR, une méthode BRR modifiée, due à Bob Fay, utilise toutes les unités échantillonnées lors de chaque répétition en définissant les poids de rééchantillonnage comme suit : $d_i^{(r)}(\varepsilon) = (1 + \varepsilon)d_i$ ou $(1 - \varepsilon)d_i$ selon que i se trouve ou non dans le demi-échantillon, où $0 < \varepsilon < 1$; un bon choix de ε est $1/2$. Les poids BRR modifiés sont alors rajustés pour la non-réponse et la post-stratification, ce qui donne les poids finaux $w_i^{(r)}(\varepsilon)$ et l'estimateur $\hat{\theta}^{(r)}(\varepsilon)$. L'estimateur BRR modifié de la variance est donné par (10) divisé par ε^2, $\hat{\theta}^{(r)}$ étant remplacé par $\hat{\theta}^{(r)}(\varepsilon)$ (voir Rao et Shao (1999)). L'estimateur BRR modifié est particulièrement utile dans le cas d'une réimputation indépendante de réponses manquantes lors de chaque répétition, car il peut utiliser les donneurs de l'échantillon complet pour réaliser l'imputation, contrairement à l'estimateur BRR, qui utilise uniquement les donneurs du demi-échantillon.

Contrairement à la méthode BRR, le bootstrap de Rao-Wu est valide pour les $n_h (\geq 2)$ arbitraires ainsi que pour les estimateurs non lisses $\hat{\theta}$. On construit chaque répétition bootstrap en tirant un échantillon aléatoire simple d'UPÉ de taille $n_h - 1$ à partir des grappes d'échantillon n_h, indépendamment d'une strate à l'autre. Les poids de sondage bootstrap $d_i^{(r)}$ sont donnés par $[n_h /(n_h - 1)]m_{hi}^{(r)} d_i$ si i est dans la strate h et la répétition r, où $m_{hi}^{(r)}$ est le nombre de fois que l'UPÉ échantillonnée (hi) est sélectionnée, $\sum_i m_{hi}^{(r)} = n_h - 1$. Les poids $d_i^{(r)}$ sont alors rajustés pour la non-réponse totale et la post-stratification, ce qui donne les poids bootstrap finaux et l'estimateur $\hat{\theta}^{(r)}$. Souvent, on utilise $R = 500$ répétitions dans l'estimateur bootstrap de la variance (10). Plusieurs enquêtes récentes de Statistique Canada ont adopté la méthode bootstrap d'estimation de la variance en raison de sa souplesse dans le choix de R et de sa grande applicabilité. Les utilisateurs des fichiers de microdonnées d'enquête de Statistique Canada semblent très satisfaits de la méthode bootstrap d'analyse des données.

Les premiers travaux sur le jackknife et le BRR étaient en grande partie empiriques (cf. Kish et Frankel 1974). Krewski et Rao (1981) ont élaboré un cadre asymptotique formel approprié pour l'échantillonnage stratifié à plusieurs degrés et ont établi la convergence selon le plan de sondage des estimateurs jackknife et BRR de la variance lorsque $\hat{\theta}$ peut être exprimé comme une fonction lisse des moyennes estimatives. Plusieurs ajouts à ces travaux de base ont été signalés dans la documentation récente, comme en témoigne l'ouvrage de Shao et Tu (1995, chapitre 6). Le soutien théorique des méthodes de rééchantillonnage est essentiel pour leur utilisation dans la pratique.

Dans l'exposé qui précède, $\hat{\theta}$ dénote l'estimateur d'un paramètre « de recensement ». Ordinairement, le paramètre de recensement θ_C est motivé par un modèle sous-jacent de superpopulation et le recensement est considéré comme un échantillon produit par le modèle, ce qui donne des équations d'estimation de recensement dont la solution est θ_C. Les fonctions d'estimation de recensement $U_C(\theta)$ sont simplement des totaux de population des fonctions $u_i(\theta)$ avec

espérance nulle sous le modèle hypothétique, et les équations d'estimation de recensement sont données par $U_C(\theta) = 0$ (Godambe et Thompson 1986). Kish et Frankel (1974) ont soutenu que le paramètre de recensement est valable même si le modèle n'est pas correctement spécifié. Par exemple, dans le cas de la régression linéaire, le coefficient de régression de recensement pourrait expliquer dans quelle mesure la relation entre la variable réponse et les variables indépendantes est prise en compte par un modèle de régression linéaire. Comme les fonctions d'estimation de recensement sont simplement des totaux de population, on obtient les estimateurs pondérés $\hat{U}(\theta)$ à partir de l'échantillon complet et $\hat{U}^{(r)}(\theta)$ à partir de chaque pseudo-répétition. Les solutions des équations d'estimation correspondantes $\hat{U}(\theta) = 0$ et $\hat{U}^{(r)}(\theta) = 0$ donnent respectivement $\hat{\theta}$ et $\hat{\theta}^{(r)}$. Il est à noter que les estimateurs de variance par rééchantillonnage ont pour objet d'estimer la variance de $\hat{\theta}$ comme un estimateur des paramètres de recensement, mais non des paramètres de modélisation. Dans certaines conditions, on peut faire abstraction de la différence mais, en général, on est en présence d'une situation d'échantillonnage à deux phases où le recensement est l'échantillon de première phase tiré de la superpopulation et l'échantillon est un échantillon probabiliste tiré de la population de recensement. Récemment, on a mené des travaux utiles sur l'estimation de la variance à deux phases lorsque les paramètres de modélisation sont les paramètres cibles (Graubard et Korn 2002; Rubin-Bleuer et Schiopu Kratina 2005), mais il faudrait approfondir ces travaux pour surmonter la difficulté de spécifier la structure de covariance des erreurs de modèle.

Le bootstrap présente une difficulté : la solution $\hat{\theta}^{(r)}$ n'existe pas nécessairement pour certaines répétitions bootstrap r (Binder, Kovacevic et Roberts 2004). Rao et Tausi (2004) ont utilisé la méthode du bootstrap avec fonction d'estimation, qui évite la difficulté. Selon cette méthode, on résout $\hat{U}(\theta) = \hat{U}^{(r)}(\hat{\theta})$ pour θ en utilisant une seule étape de l'itération de Newton-Raphson avec $\hat{\theta}$ comme valeur de départ. On utilise alors dans (10) l'estimateur $\tilde{\theta}^{(r)}$ ainsi obtenu pour calculer l'estimateur bootstrap avec fonction d'estimation de la variance de $\hat{\theta}$ qu'on peut facilement mettre en œuvre à partir du fichier de données qui fournit les poids de rééchantillonnage, en modifiant légèrement un progiciel qui tient compte des poids d'échantillonnage. Il est intéressant de noter que l'estimateur bootstrap avec fonction d'estimation de la variance équivaut à un estimateur de sandwich par linéarisation de Taylor de la variance qui utilise l'estimateur bootstrap de la variance de $\hat{U}(\theta)$ et l'inverse de la matrice d'information observée (dérivée de $-\hat{U}(\theta)$), tous deux évalués à $\theta = \hat{\theta}$ (Binder et coll. 2004).

Contrairement aux méthodes de rééchantillonnage, les méthodes de linéarisation de Taylor produisent des estimateurs de variance asymptotiquement valides pour les plans d'échantillonnage généraux, mais elles nécessitent une formule distincte pour chaque estimateur $\hat{\theta}$. Binder (1983), Rao, Yung et Hidiroglou (2002) et Demnati et Rao (2004) ont fourni des formules unifiées d'estimation de variance par linéarisation pour des estimateurs définis comme des solutions aux équations d'estimation.

Pfeffermann (1993) a étudié le rôle des poids de sondage dans l'analyse des données d'enquête. Si le modèle de population est valable pour l'échantillon (c'est-à-dire s'il est sans biais d'échantillonnage), les estimateurs non pondérés fondés sur un modèle sont alors plus efficaces que les estimateurs pondérés et donnent des inférences valides, notamment pour des données où la taille des échantillons est faible et la variation des poids est élevée. Toutefois, pour les données ordinaires provenant d'enquêtes à grande échelle, le plan d'enquête est informatif et le modèle de population n'est pas nécessairement valable pour l'échantillon. Par conséquent, les estimateurs fondés sur un modèle peuvent être fortement biaisés et les inférences risquent d'être erronées. Pfeffermann et ses collègues ont proposé une nouvelle approche de l'inférence sous échantillonnage informatif (voir Pfeffermann et Sverchkov 2003), qui semble donner des inférences plus efficientes que l'approche pondérée et mérite certainement l'attention des utilisateurs de données d'enquête. Toutefois, il reste beaucoup de travail à accomplir, surtout en ce qui concerne le traitement de données fondées sur l'échantillonnage à plusieurs degrés. Skinner, Holt et Smith (1989), Chambers et Skinner (2003) et Lehtonen et Pahkinen (2004) donnent d'excellentes descriptions des méthodes d'analyse de données d'enquête complexes.

7. ESTIMATION SUR PETITS DOMAINES

Dans les sections précédentes, nous avons abordé les méthodes traditionnelles qui utilisent des estimateurs directs par domaine fondés sur des observations d'échantillons spécifiques aux domaines et sur des données auxiliaires sur la

population. Toutefois, ces méthodes ne donnent pas nécessairement des inférences fiables lorsque la taille des échantillons du domaine est infime, voire nulle pour certains domaines. Dans la documentation, les domaines ou sous-populations dont la taille est infime ou nulle sont appelés petits domaines. Au cours des dernières années, la demande de statistiques fiables sur les petits domaines a grandement augmenté en raison du recours croissant à la statistique des petits domaines dans la formulation des politiques et des programmes, la répartition des fonds et la planification régionale. Manifestement, il est rarement possible d'obtenir des échantillons dont la taille globale est assez grande pour soutenir des estimations directes fiables pour tous les domaines d'intérêt. De plus, dans la pratique, il n'est pas possible de prévoir toutes les utilisations des données d'enquête et « le client exige toujours plus qu'il n'est spécifié à l'étape de l'élaboration du plan de sondage » (Fuller 1999, page 344). Pour faire des estimations sur petits domaines avec un niveau suffisant de précision, il faut souvent utiliser des estimateurs « indirects » qui empruntent de l'information à des domaines connexes par le biais de données auxiliaires, comme celles du recensement et les données administratives courantes, pour accroître la taille « effective » des échantillons à l'intérieur des petits domaines.

Aujourd'hui, on s'entend à reconnaître que des modèles explicites liant les petits domaines par le biais de données auxiliaires et tenant compte de la variation résiduelle entre domaines par le biais des effets aléatoires des petits domaines sont nécessaires pour calculer des estimateurs indirects. Le succès des méthodes fondées sur un modèle dépend fortement de la disponibilité de données auxiliaires fiables et d'une validation complète des modèles au moyen d'évaluations internes et externes. Bon nombre de méthodes axées sur les effets aléatoires et utilisées dans la théorie statistique courante sont pertinentes à l'estimation sur petits domaines, dont la méthode du meilleur prédicteur empirique (ou méthode de Bayes), celle du meilleur prédicteur linéaire sans biais empirique et celle du modèle hiérarchique bayésien fondé sur des lois de distribution a priori des paramètres de modélisation. Rao (2003) donne une description complète de ces méthodes. La pertinence (sur le plan pratique) et l'intérêt (sur le plan théorique) de l'estimation sur petits domaines ont attiré l'attention de nombreux chercheurs, d'où la réalisation de progrès importants dans l'estimation ponctuelle et celle de l'erreur quadratique moyenne. Dans le monde entier, les « nouvelles » méthodes ont été appliquées avec succès à divers problèmes liés aux petits domaines. Aux États-Unis, on a utilisé récemment des méthodes fondées sur un modèle pour produire des estimations par comté et par district scolaire relativement aux enfants pauvres d'âge scolaire. Chaque année, le département de l'Éducation des États-Unis accorde aux comtés des fonds de plus de sept milliards de dollars sur la base d'estimations par comté fondées sur un modèle. Les fonds alloués soutiennent des programmes d'éducation compensatoire pour répondre aux besoins des enfants défavorisés sur le plan scolaire. Le lecteur trouvera dans Rao (2003, exemple 7.1.2) des renseignements sur cette application. Au Royaume-Uni, le Office of National Statistics a mis sur pied un projet d'estimation sur petits domaines pour établir des estimations fondées sur un modèle au niveau des sections électorales (quelque 2 000 ménages). Schaible (1996) décrit la pratique et les méthodes d'estimation des programmes statistiques fédéraux des États-Unis qui utilisent des estimateurs indirects pour produire des estimations publiées. Singh, Gambino et Mantel (1994) et Brackstone (2002) traitent de certains aspects pratiques et stratégiques de la statistique des petits domaines.

L'estimation sur petits domaines constitue un exemple frappant de l'interaction entre la théorie et la pratique. Les progrès de la théorie sont impressionnants, mais bon nombre de questions d'ordre pratique nécessitent une plus grande attention de la part des théoriciens, notamment les suivantes : a) des estimateurs d'étalonnage fondés sur un modèle pour concorder avec des estimateurs directs fiables au niveau des grands domaines; b) l'établissement et la validation de modèles de liaison appropriés et l'étude de questions comme les erreurs dans les variables, la spécification incorrecte du modèle de liaison et les variables omises; c) la mise au point de méthodes qui satisfont plusieurs objectifs : de bonnes estimations spécifiques au domaine, de bons rangs et un bon histogramme des petits domaines.

8. CERTAINS ASPECTS THÉORIQUES MÉRITANT L'ATTENTION DES PRATICIENS ET VICE VERSA

Dans la présente section, j'aborde brièvement quelques exemples d'aspects théoriques importants qui existent mais qui sont peu utilisés dans la pratique.

8.1 Inférence par la vraisemblance empirique

La théorie traditionnelle de l'échantillonnage portait dans une large mesure sur l'estimation ponctuelle et les erreurs-types associées, faisant appel à des approximations normales pour déterminer des intervalles de confiance à l'égard des paramètres d'intérêt. En statistique courante, l'approche de la vraisemblance empirique (VE) (Owen 1988) a beaucoup attiré l'attention en raison de plusieurs propriétés souhaitables. Elle offre une vraisemblance non paramétrique, ce qui donne des intervalles de confiance de VE semblables aux intervalles de vraisemblance paramétrique. La forme et l'orientation des intervalles de VE sont entièrement déterminées par les données; les intervalles préservent l'étendue tout en respectant la transformation et, contrairement aux intervalles symétriques de la théorie normale, ils sont particulièrement utiles puisqu'ils donnent des taux d'erreur équilibrés de la queue. Comme je l'ai mentionné dans la section 3.1, Hartley et Rao (1968) ont été les premiers à proposer l'approche de la VE dans le contexte des enquêtes par sondage, mais leur démarche était axée sur des questions d'inférence liées à l'estimation ponctuelle. Chen, Chen et Rao (2003) ont obtenu des intervalles de VE sur la moyenne de population sous échantillonnage aléatoire simple et sous échantillonnage aléatoire stratifié pour des populations contenant bien des zéros. On trouve ces populations dans le contrôle par sondage, où y dénote le montant d'argent dû à l'État et la moyenne arithmétique $\overline{Y}$ correspond au montant moyen des créances excessives. Des travaux antérieurs sur le contrôle par sondage ont utilisé des intervalles de vraisemblance paramétrique fondés sur des distributions de mélanges paramétriques pour la variable y. Ces intervalles donnent de meilleurs résultats que les intervalles-types de la théorie normale, mais les intervalles de VE donnent de meilleurs résultats en présence d'écarts par rapport au modèle hypothétique de mélanges, en donnant un taux de non-couverture inférieur à la borne inférieure plus proche du taux d'erreur nominal, ainsi qu'une borne inférieure plus grande. Pour les plans généraux, Wu et Rao (2004) ont utilisé une pseudo-vraisemblance empirique (Chen et Sitter 1999) pour obtenir des intervalles de pseudo-VE rajustés sur la moyenne arithmétique et la fonction de distribution qui tiennent compte des caractéristiques du plan, et ils ont montré que les intervalles donnaient des taux d'erreur de la queue plus équilibrés que dans le cas des intervalles de la théorie normale. La méthode VE offre également une approche systématique de l'estimation par calage et de l'intégration des enquêtes. Le lecteur est invité à consulter les articles de Rao (2004) et Wu et Rao (2005).

Il reste encore à perfectionner ces notions, notamment en ce qui concerne la pseudo-vraisemblance empirique, mais la théorie de la VE dans le contexte des enquêtes mérite l'attention des praticiens.

8.2 Analyses exploratoires des données d'enquête

Dans la section 6, nous avons abordé les méthodes d'analyse confirmative de données d'enquête tenant compte du plan de sondage, comme l'estimation ponctuelle des paramètres de modélisation (ou de recensement) et des erreurs-types associées, ainsi que les tests formels d'hypothèses. Les graphiques et les analyses exploratoires des données d'enquête sont aussi très utiles. Ces méthodes ont fait l'objet d'une foule d'études dans la documentation courante. Encore récemment, certains ajouts à ces méthodes modernes ont été signalés dans la documentation sur les enquêtes et ils méritent l'attention des praticiens. J'en aborde brièvement un certain nombre. Premièrement, on utilise couramment des estimations non paramétriques de densité du noyau pour présenter la forme d'un ensemble de données sans recourir à des modèles paramétriques. On peut aussi les utiliser pour comparer différentes sous-populations.

Bellhouse et Stafford (1999) ont proposé des estimateurs de densité du noyau qui tiennent compte du plan d'enquête, en ont étudié les propriétés et ont appliqué les méthodes aux données de l'Enquête sur la santé en Ontario. Buskirk et Lohr (2005) ont étudié les propriétés asymptotiques et les propriétés de population finie des estimateurs de densité du noyau et ont obtenu des bandes de confiance. Ils ont appliqué les méthodes aux données de deux enquêtes américaines, la National Crime Victimization Survey et la National Health and Nutrition Examination Survey.

Deuxièmement, Bellhouse et Stafford (2001) ont mis au point des méthodes de régression polynomiale locale qui tiennent compte du plan de sondage et qu'on peut utiliser pour étudier la relation entre une variable réponse et des variables prédictives sans faire d'hypothèses audacieuses au sujet d'un modèle paramétrique. Les graphiques ainsi obtenus sont utiles pour comprendre les relations ainsi que pour comparer différentes sous-populations. À l'aide des données de l'Enquête sur la santé en Ontario, les auteurs ont illustré la régression polynomiale locale en montrant, par exemple, la

relation entre l'indice de masse corporelle des femmes et leur âge. Bellhouse, Chipman et Stafford (2004) ont étudié des modèles additifs de données d'enquête au moyen de la méthode des moindres carrés pénalisée pour traiter plus d'une variable prédictive, et ont illustré les méthodes à l'aide des données de l'Enquête sur la santé en Ontario. Cette approche offre de nombreux avantages en ce qui concerne les graphiques, l'estimation, les tests et la sélection de paramètres « de lissage » pour ajuster les modèles.

8.3 Erreurs de mesure

Habituellement, on suppose que les erreurs de mesure sont additives et que leur moyenne est nulle. Par conséquent, les estimateurs habituels du total et des moyennes restent sans biais ou convergents. Toutefois, cette caractéristique positive n'est pas nécessairement valable pour des paramètres plus complexes comme la fonction de distribution, les quantiles et les coefficients de régression. Dans ce dernier cas, les estimateurs habituels sont biaisés, même pour de grands échantillons, et peuvent donc produire des inférences erronées (Fuller 1995). Il est possible d'obtenir des estimateurs corrigés pour le biais si l'on dispose d'estimations des variances de l'erreur de mesure. On peut obtenir ces dernières en affectant des ressources, à l'étape de l'élaboration du plan de sondage, pour faire des observations répétées sur un sous-échantillon. Fuller (1975, 1995) a préconisé l'utilisation de méthodes appropriées en présence d'erreurs de mesure, et les méthodes corrigées pour le biais méritent l'attention des praticiens.

Hartley et Rao (1978) et Hartley et Biemer (1978) ont établi des conditions d'affectation des intervieweurs et des codeurs qui permettent d'estimer les variances d'échantillonnage et de réponse pour la moyenne arithmétique ou le total à partir d'enquêtes courantes. Malheureusement, le plan de sondage des enquêtes d'aujourd'hui satisfait rarement ces conditions et, même si c'était le cas, on dispose rarement de l'information requise sur les affectations des intervieweurs et des codeurs à l'étape de l'estimation.

On utilise souvent les composantes linéaires des modèles de variance pour estimer la variabilité des intervieweurs. Ces modèles sont appropriés pour la réponse continue, mais pas pour les réponses binaires. L'approche du modèle linéaire pour les réponses binaires peut entraîner une sous-estimation des corrélations intra-intervieweurs. Scott et Davis (2001) ont proposé des modèles hiérarchiques pour les réponses binaires afin d'estimer la variabilité due aux intervieweurs. Comme les réponses sont souvent binaires dans bon nombre d'enquêtes, les praticiens doivent prêter attention à ces modèles pour effectuer des analyses pertinentes des données d'enquête avec réponses binaires.

8.4 Imputation des données d'enquête manquantes

Dans la pratique, on utilise couramment l'imputation pour remplacer des éléments manquants. On s'assure ainsi que les résultats d'analyses différentes de l'ensemble de données complété sont cohérents entre eux en utilisant le même poids d'échantillonnage pour tous les éléments. Bon nombre d'organismes statistiques utilisent des méthodes d'imputation marginale comme celles du ratio, du plus proche voisin et du donneur aléatoire à l'intérieur des classes d'imputation. Malheureusement, on traite souvent les valeurs imputées comme s'il s'agissait de valeurs vraies, puis on calcule des estimations et des estimations de la variance. Les estimations ponctuelles imputées de paramètres marginaux sont généralement valides en présence d'un mécanisme de réponse ou d'un modèle d'imputation hypothétique. Mais les estimateurs « naïfs » de la variance peuvent produire des inférences erronées, même pour de grands échantillons, notamment une forte sous-estimation de la variance de l'estimateur imputé, faute de prendre en compte la variabilité additionnelle due à l'estimation des valeurs manquantes. Les partisans de l'imputation multiple de Rubin (1987) soutiennent que l'estimateur de variance à imputation multiple peut régler ce problème parce qu'une somme des carrés entre estimateurs imputés est ajoutée à la moyenne des estimateurs naïfs de la variance obtenus au moyen des imputations multiples. Malheureusement, les estimateurs de variance à imputation multiple comportent certaines difficultés, comme en font état Kott (1995), Fay (1996), Binder et Sun (1996), Wang et Robins (1998), Kim, Brick, Fuller et Kalton (2004) et d'autres auteurs. En outre, on préfère souvent l'imputation simple pour des raisons d'efficacité opérationnelle et de rentabilité. Au cours des dernières années, on a fait des progrès impressionnants en réalisant des inférences efficientes et asymptotiquement valides à partir d'ensembles de données imputées une seule fois. Le lecteur est invité à consulter les articles de Shao (2002) et Rao (2000, 2005) sur les méthodes d'estimation de la variance au moyen de l'imputation simple.

Kim et Fuller (2004) ont étudié l'imputation partielle en utilisant plus d'une valeur imputée au hasard et ont montré que cette méthode donnait également des inférences asymptotiquement valides; voir aussi Kalton et Kish (1984) et Fay (1996). L'imputation partielle offre l'avantage de réduire la variance due à l'imputation par rapport à l'imputation unique utilisant une seule valeur imputée au hasard. Les méthodes d'estimation de la variance susmentionnées méritent l'attention des praticiens.

8.5 Enquêtes à bases multiples

Les enquêtes à bases multiples emploient deux ou plusieurs bases chevauchantes pour couvrir entièrement la population cible. Hartley (1962) a étudié le cas particulier d'une base complète B, d'une base incomplète A et d'un échantillonnage aléatoire simple mené indépendamment dans les deux bases. Il a montré que par rapport à l'estimateur à base unique complète, un estimateur à double base « optimal » pouvait donner lieu à d'importants gains d'efficience pour le même coût, à condition que le coût par unité pour la base A soit nettement inférieur au coût par unité pour la base B. Les enquêtes à bases multiples conviennent particulièrement à l'échantillonnage de populations rares ou difficiles à joindre, comme les populations de sans-abri et de personnes atteintes du SIDA, lorsque des listes incomplètes contiennent de fortes proportions de personnes appartenant à la population cible. Dans un article marquant, Hartley (1974) a calculé des estimateurs à double base « optimaux » pour des plans d'échantillonnage généraux et des unités d'observation pouvant être différentes dans les deux bases. Fuller et Burmeister (1972) ont proposé des estimateurs « optimaux » améliorés. Toutefois, les estimateurs optimaux utilisent des ensembles de poids différents pour chaque élément y, ce qui n'est pas souhaitable dans la pratique. Skinner et Rao (1996) ont calculé, pour les enquêtes à double base, des estimateurs du pseudo-maximum de vraisemblance (PMV) qui utilisent le même ensemble de poids pour tous les éléments y, comme dans le cas des estimateurs « à base unique » (Kalton et Anderson 1986), et qui maintiennent l'efficience. Lohr et Rao (2005) ont formulé une théorie unifiée des conditions des enquêtes à bases multiples en prolongeant les estimateurs optimal, PMV et à base unique. Lohr et Rao (2000, 2005) ont obtenu des estimateurs de variance jackknife asymptotiquement valides. Ces résultats généraux méritent l'attention des praticiens lorsqu'on travaille avec deux ou plusieurs bases. Les enquêtes téléphoniques à double base (téléphones cellulaires et téléphones fixes) nécessitent l'attention des théoriciens, car on ignore comment pondérer dans le cas de l'enquête menée par téléphone cellulaire : certaines familles partagent un téléphone cellulaire, d'autres en possèdent un pour chaque personne.

8.6 Échantillonnage indirect

On peut utiliser la méthode de l'échantillonnage indirect lorsqu'on ne dispose pas de la base d'une population cible U^B mais qu'on emploie la base d'une autre population U^A, liée à U^B, pour tirer un échantillon probabiliste. On utilise les liens entre les deux populations pour établir des poids appropriés qui peuvent donner des estimateurs sans biais et des estimateurs de variance. Lavallée (2002) a mis au point une méthode unifiée, appelée méthode généralisée du partage des poids (MGPP), inspirée de plusieurs méthodes connues : la méthode du partage des poids d'Ernst (1989) pour l'estimation transversale à partir d'enquêtes-ménages longitudinales, l'échantillonnage par réseau et l'estimation de la multiplicité (Sirken 1970), ainsi que l'échantillonnage en grappes adaptatif (Thompson et Seber 1996). La théorie de Rao (1968) sur l'échantillonnage à partir d'une base contenant une quantité inconnue de doubles comptes peut être considérée comme un cas particulier de la MGPP. On peut aussi employer la MGPP pour travailler avec des bases multiples; les estimateurs ainsi obtenus sont simples, mais pas nécessairement efficients par rapport aux estimateurs optimaux de Hartley (1974) ou aux estimateurs du PMV. La méthode MGPP a une grande applicabilité et mérite l'attention des praticiens.

9. CONCLUSION

L'apport de Joe Waksberg à la théorie et aux méthodes des enquêtes par sondage reflète bien l'interaction entre la théorie et la pratique. Dans le cadre de son travail au Census Bureau des États-Unis, puis à Westat, il a fait face à de réels problèmes d'ordre pratique et a souvent trouvé des solutions théoriques judicieuses. Par exemple, dans un article marquant (Waksberg 1978), il a décrit une ingénieuse méthode de composition aléatoire (CA) qui réduit considérablement les coûts

d'enquête par rapport à la composition de numéros entièrement au hasard. Il a présenté des arguments théoriques solides pour en démontrer l'efficience. L'utilisation généralisée des enquêtes par CA est due pour une bonne part à l'argumentation théorique de Waksberg (1978) et à des perfectionnements ultérieurs. Joe Waksberg est un spécialiste de l'échantillonnage d'enquête que j'admire énormément et je suis très honoré d'avoir reçu le prix Waksberg 2005 pour les techniques d'enquête.

REMERCIEMENTS

Je tiens à remercier David Bellhouse, Wayne Fuller, Jack Gambino, Graham Kalton, Fritz Scheuren et Sharon Lohr, dont les observations et les suggestions m'ont été très utiles.

RÉFÉRENCES

Aires, N., et Rosén, B. (2005), On inclusion probabilities and relative estimator bias for Pareto πps sampling. *Journal of Statistical Planning and Inference*, 128, 543-567.

Andreatta, G., et Kaufmann, G.M. (1986), Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association*, 81, 657-666.

Bankier, M.D. (1988), Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Bankier, M.D. (2003), 2001 Canadian Census weighting: switch from projection GREG to pseudo-optimal regression estimation. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Rapport technique no. 386, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.

Bankier, M.D., Rathwell, S. et Majkowski, M. (1992), Two step generalized least squares estimation in the 1991 Canadian Census. Document de travail, direction de la méthodologie, division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa.

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

Bellhouse, D.R., et Rao, J.N.K. (2002), Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, 102, 47-58.

Bellhouse, D.R., et Stafford, J.E. (1999), Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

Bellhouse, D.R., et Stafford, J.E. (2001), Régression polynomiale locale dans le cas des enquêtes complexes. *Techniques d'enquête*, 27, 219-226.

Bellhouse, D.R., Chipman, H.A. et Stafford, J.E. (2004), Additive models for survey data via penalized least squares. Rapport technique.

Binder, D.A. (1983), On the variance of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D.A., et Sun, W. (1996), Frequency valid multiple imputation for surveys with a complex design. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281-286.

Binder, D.A., Kovacevic, M. et Roberts, G. (2004), Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Bowley, A.L. (1926), Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv. 1, 6-62.

Brackstone, G. (2002), Stratégies et approches relatives aux statistiques régionales. *Techniques d'enquête*, 28, 125-133.

Brackstone, G., et Rao, J.N.K. (1979), An investigation of raking ratio estimators. *Sankhyā*, Series C, 42, 97-114.

Brewer, K.R.W. (1963), Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

Brewer, K.R.W., et Hanif, M. (1983), *Sampling With Unequal Probabilities*. New York: Springer-Verlag.

Buskirk, T.D., et Lohr, S.L. (2005), Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.

Casady, R.J., et Valliant, R. (1993), Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.

Chambers, R.L. (1996), Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chambers, R.L., et Skinner, C.J. (Éds.) (2003), *Analysis of Survey Data*. Chichester: Wiley.

Chen, J., et Sitter, R.R. (1999), A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 12, 1223-1239.

Chen, J., Chen, S.Y. et Rao, J.N.K. (2003), Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics*, 31, 53-68.

Chen, J., Sitter, R.R. et Wu, C. (2002), Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

Cochran, W.G. (1939), The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.

Cochran, W.G. (1940), The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science*, 30, 262-275.

Cochran, W.G. (1942), Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 191-212.

Cochran, W.G. (1946), Relative accuracy of systematic and stratified random samples from a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

Cochran, W.G. (1953), *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Dalenius, T. (1957), *Sampling in Sweden*. Stockholm: Almquist and Wicksell.

Dalenius, T., et Hodges, J.L. (1959), Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Deming, W.E. (1950), *Some Theory of Sampling*. New York: John Wiley & Sons, Inc.

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

Deming, W.E., et Stephan, F.F. (1940), On a least squares adjustment of a sampled frequency table when the expected margins are known. *The Annals of Mathematical Statistics*, 11, 427-444.

Demnati, A., et Rao, J.N.K. (2004), Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête,* 30, 17-27.

Deville, J., et Särndal, C.-E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Durbin, J. (1968), Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute*, 36, No. 3, 113-119.

Ericson, W.A. (1969), Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, Series B, 31, 195-224.

Ernst, L.R. (1989), Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys* (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh), New York: John Wiley & Sons, Inc., 135-169.

Ernst, L.R. (1999), The maximization and minimization of sample overlap problem: A half century of results. *Bulletin of the International Statistical Institute*, Vol. LVII, Book 2, 293-296.

Fay, R.E. (1996), Alternative peradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Fellegi, I.P. (1964), Response variance and its estimation. *Journal of the American Statistical Association,* 59, 1016-1041.

Fellegi, I.P. (1966), Changing the probabilities of selection when two units are selected with PPS sampling without replacement. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington DC, 434-442.

Fellegi, I.P. (1981), Should the census counts be adjusted for allocation purposes? – Equity considerations. Dans *Current Topics in Survey Sampling* (Éds. D. Krewski, R. Platek et J.N.K. Rao), New York: Academic Press, 47-76.

Francisco, C.A., et Fuller, W.A. (1991), Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

Fuller, W.A. (1975), Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.

Fuller, W.A. (1995), Estimation in the presence of measurement error. *Revue Internationale de Statistique*, 63, 121-147.

Fuller, W.A. (1999), Environmental surveys over time, *Journal of Agricultural, Biological and Environmental Statistics*, 4, 331-345.

Fuller, W.A. (2002), Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

Fuller, W.A., et Burmeister, L.F. (1972), Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section*, American Statistical Association, 245-249.

Fuller, W.A., et Rao, J.N.K. (2001), Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête* 27, 49-56.

Gambino, J., Kennedy, B. et Singh, M.P. (2001), Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application. *Techniques d'enquête* 27, 69-79.

Godambe, V.P. (1955), A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, Series B, 17, 269-278.

Godambe, V.P. (1966), A new approach to sampling from finite populations. *Journal of the Royal Statistical Society*, Series B, 28, 310-328.

Godambe, V.P., et Thompson, M.E. (1986), Parameters of superpopulation and survey population: Their relationship and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Graubard, B.I., et Korn, E.L. (2002), Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73-96.

Hacking, I. (1975), *The Emergence of Probability*. Cambridge: Cambridge University Press.

Hajék, J. (1971), Comments on a paper by Basu, D. Dans *Foundations of Statistical Inference* (Éds. V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston,

Hansen, M.H., et Hurwitz, W.N. (1943), On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Hansen, M.H., Dalenius, T. et Tepping, B.J. (1985), The development of sample surveys of finite populations. Chapter 13 in *A Celebration of Statistics*. The ISI Centenary Volume, Berlin: Springer-Verlag.

Hansen, M.H., Hurwitz, W.N. et Bershad, M. (1961), Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953), *Sample Survey Methods and Theory*, Vols. I et II. New York: John Wiley & Sons, Inc.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983), An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Hansen, M.H., Hurwitz, W.N., Marks, E.S. et Mauldin, W.P. (1951), Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190

Hansen, M.H., Hurwitz, W.N., Nisseslson, H. et Steinberg, J. (1955), The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.

Hartley, H.O. (1959), Analytical studies of survey data. In Volume in Honour of Corrado Gini, Instituto di Statistica, Rome, 1-32.

Hartley, H.O. (1962), Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Hartley, H.O. (1974), Mutiple frame methdology and selected applications. *Sankhyā*, Series C, 36, 99-118.

Hartley, H.O., et Biemer, P. (1978), The estimation of nonsampling variances in current surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 257-262.

Hartley, H.O., et Rao, J.N.K. (1962), Sampling with unequal probability and without replacement. *The Annals of Mathematical Statistics*, 33, 350-374.

Hartley, H.O., et Rao, J.N.K. (1968), A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Hartley, H.O., et Rao, J.N.K. (1978), The estimation of nonsampling variance components in sample surveys. Dans *Survey Measurement* (Éd. N.K. Namboodiri), New York: Academic Press, 35-43.

Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D. (1976), SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa, États-Unis.

Hinkins, S., Oh, H.L. et Scheuren, F. (1997), Algorithmes de plan de sondage inverses. *Techniques d'enquête*, 23, 13-24.

Holt, D., et Smith, T.M.F. (1979), Post-stratification. *Journal of the Royal Statistical Society*, Series A, 142, 33-46.

Horvitz, D.G., et Thompson, D.J. (1952), A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Huang, E.T., et Fuller, W.A. (1978), Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section*, American Statistical Association, 300-305.

Hubback, J.A. (1927), Sampling for rice yield in Bihar and Orissa. Imperial Agricultural Research Institute, Pusa, Bulletin No. 166 (représenté dans *Sankhyā*, 1946, vol. 7, 281-294),

Hussain, M. (1969), Construction of regression weights for estimation in sample surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.

Jessen, R.J. (1942), Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.

Kalton, G. (2002), Models in the practice of survey sampling (revisited), *Journal of Official Statistics*, 18, 129-154.

Kalton, G., et Anderson, D.W. (1986), Sampling rare populations. *Journal of the Royal Statistical Society*, Series A, 149, 65-82.

Kalton, G., et Kish, L. (1984), Some efficient random imputation methods. *Communications in Statistics*, A13, 1919-1939.

Keyfitz, N. (1951), Sampling with probabilities proportional to size: adjustment for changing in the probabilities. *Journal of the American Statistical Association*, 46, 105-109.

Kiaer, A. (1897), The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistics of Norway.

Kim, J., et Fuller, W.A. (2004), Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Brick, J.M., Fuller, W.A. et Kalton, G. (2004), On the bias of the multiple imputation variance estimator in survey sampling. Rapport technique.

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

Kish, L. (1995), The hundred year's wars of survey sampling. *Statistics in Transition*, 2, 813-830.

Kish, L., et Scott, A.J. (1971), Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

Kish, L., et Frankel, M.R. (1974), Inference from complex samples. *Journal of the Royal Statistical Society*, Series B, 36, 1-37.

Kott, P.S. (1995), A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389.

Kott, P.S. (2005), Randomized-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129, 263-277.

Krewski, D., et Rao, J.N.K. (1981), Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Kruskal, W.H., et Mosteller, F. (1980), Representative sampling IV: The history of the concept in Statistics, 1895-1939. *Revue Internationale de Statistique*, 48, 169-195.

Laplace, P.S. (1820), A philosophical essay on probabilities. English translation, Dover, 1951.

Lavallée, P. (2002), *Le Sondage indirect, ou la Méthode généralisée du partage des poids.* Éditions de l'Université de Burxelles, Belgique, Éditions Ellipse, France.

Lavallée, P., et Hidiroglou, M. (1988), Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.
Lehtonen, R., et Pahkinen, E. (2004), *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.

Lindley, D.V. (1996), Letter to the editor. *American Statistician*, 50, 197.

Lohr, S.L. (1999), *Sampling*: *Design and Analysis*. Pacific Grove: Duxbury.

Lohr, S.L., et Rao, J.N.K. (2000), Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 2710280.

Lohr, S.L., et Rao, J.N.K. (2005), Multiple frame surveys: point estimation and inference. *Journal of the American Statistical Association* (en révision),

Lu, W.W., Brick, M. et Sitter, R.R. (2004), Algorithms for constructing combined strata grouped jackknife and balanced repeated replication with domains. Rapport technique, Westat, Rockville, Maryland.

Mach, L, Reiss, P.T. et Schiopu-Kratina, I. (2005), The use of the transportation problem in co-ordinating the selection of samples for business surveys. Rapport technique HSMD-2005-006E, Statistique Canada, Ottawa.

Madow, W.G., et Madow, L.L. (1944), On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.

Mahalanobis, P.C. (1944), On large scale sample surveys. *Philosophical Transactions of the Royal Society*, London, Series B, 231, 329-451.

Mahalanobis, P.C. (1946a), Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mahalanobis, P.C. (1946b), Sample surveys of crop yields in India. *Sankhyā*, 7, 269-280.

McCarthy, P.J. (1969), Pseudo-replication: Half samples. *Review of the International Statistical Institute*, 37, 239-264.

Merkouris, T. (2004), Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Murthy, M.N. (1957), Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

Murthy, M.N. (1964), On Mahalanobis' contributions to the development of sample survey theory and methods. Dans *Contributions to Statistics*: Présenté au professeur P.C. Mahalanobis à l'occasion de son 70^{ième} anniversaire, Calcutta, Statistical Publishing Society: 283-316.

Narain, R.D. (1951), On sampling without replacement with varying probabilities. *Journal of the Indian Society of Aricultural Statistics*, 3, 169-174.

Neyman, J. (1934), On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Neyman, J. (1938), Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

Ohlsson, E. (1995), Coordination of samples using permanent random members. Dans *Business Survey Methods* (Éds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott), New York: John Wiley & Sons, Inc., 153-169.

O'Muircheartaigh, C.A., et Wong, S.T. (1981), The impact of sampling theory on survey sampling practice: A review. *Bulletin of the International Statistical Institute*, Article, 49, No. 1, 465-493.

Owen, A.B. (1988), Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A.B. (2002), *Empirical Likelihood*. New York: Chapman & Hall/CRC.

Park, M., et Fuller, W.A. (2005), Vers des poids de régression non négatifs pour les échantillons d'enquête. *Techniques d'enquête*, 31, 93-101.

Patterson, H.D. (1950), Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B, 12, 241-255.

Pfeffermann, D. (1993), The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

Pfeffermann, D., et Sverchkov, M. (2003), Fitting generalized linear models under informative sampling. Dans *Analysis of Survey Data* (Éds. R.L. Chambers et C.J. Skinner), Chichester: Wiley, 175-195.

Raj, D. (1956), On the method of overlapping maps in sample surveys. *Sankhyā*, 17, 89-98.

Rao, J.N.K. (1966), Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, Series A, 28, 47-60.

Rao, J.N.K. (1968), Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

Rao, J.N.K. (1992), Estimating totals and distribution functions using auxiliary information at the estimation stage. *Proceedings of the workshop on uses of auxiliary information in surveys*, Bureau de la statistique de Suède.

Rao, J.N.K. (1994), Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Rao, J.N.K. (1996), Developments in sample survey theory: An appraisal. *The Canadian Journal of Statistics*, 25, 1-21.

Rao, J.N.K. (2000), Variance estimation in the presence of imputation for missing data. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 599-608.

Rao, J.N.K. (2003), *Small Area Estimation*. Hoboken: Wiley.

Rao, J.N.K. (2004), Empirical likelihood methods for sample survey data: An overview. *Proceedings of the Survey Methods Section*, SSC Annual Meeting, sous presse.

Rao, J.N.K. (2005), Re-sampling variance estimation with imputed survey data: overview. *Bulletin of the International Statistical Institute*.

Rao, J.N.K., et Bellhouse, D.R. (1990), Genèse et évolution des fondements théoriques de l'estimation et de l'analyse fondées sur les sondages. *Techniques d'enquête*, 16, 3-26.

Rao, J.N.K., et Graham, J.E. (1964), Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

Rao, J.N.K., et Scott, A.J. (1981), The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K., et Scott, A.J. (1984), On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.

Rao, J.N.K., et Shao, J. (1999), Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

Rao, J.N.K., et Singh, A.C. (1997), A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57-64.

Rao, J.N.K., et Singh, M.P. (1973), On the choice of estimators in survey sampling. *Australian Journal of Statistics*, 15, 95-104.

Rao, J.N.K., et Tausi, M. (2004), Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics – Theory and Methods*, 33, 2087-2095.

Rao, J.N.K., et Wu, C.F.J. (1987), Methods for standard errors and confidence intervals from sample survey data: Some recent work. *Bulletin of the International Statistical Institute*.

Rao, J.N.K., et Wu, C.F.J. (1988), Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Hartley, H.O. et Cochran, W.G. (1962), On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, Series B, 24, 482-491.

Rao, J.N.K., Jocelyn, W. et Hidiroglou, M.A. (2003), Confidence interval coverage properties for regression estimators in uni-phase and two-phase sampling. *Journal of Official Statistics*, 19.

Rao, J.N.K., Scott, A.J. et Benhin, E. (2003), Défaire les structures des données d'enquête complexes : Théorie élémentaire et applications de l'échantillonnage inverse. *Techniques d'enquête*, 29, 119-131.

Rao, J.N.K., Yung, W. et Hidiroglou, M. (2002), Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, Series A, 64, 364-378.

Renssen, R.H., et Nieuwenbroek, N.J. (1997), Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-375.

Rivest, L.-P. (2002), Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 207-214.

Roberts, G., Rao, J.N.K. et Kumar, S. (1987), Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

Rosén, B. (1991), Variance estimation for systematic pps-sampling. Rapport technique, Bureau de la statistique de Suède.

Royall, R.M. (1968), An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.

Royall, R.M. (1970), On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R.M., et Cumberland, W.G. (1981), An empirical study of the ratio estimate and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

Royall, R.M., et Herson, J.H. (1973), Robust estimation in finite populations, I et II. *Journal of the American Statistical Association*, 68, 880-889 et 890-893.

Rubin, D.B. (1987), *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons, Inc.

Rubin-Bleuer, S., et Schiopu-Kratina, I. (2005), On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, (à paraître),

Salehi, M., et Seber, G.A.F. (1997), Adaptive cluster sampling with networks selected without replacements, *Biometrika*, 84, 209-219.

Särndal, C.-E. (1996), Efficient estimators with variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

Särndal, C.-E., Swenson, B. et Wretman, J.H. (1989), The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

Särndal, C.-E., Swenson, B. et Wretman, J.H. (1992), *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schabenberger, O., et Gregoire, T.G. (1994), Solutions de remplacement pour les plans πpt authentiques : Une étude comparative. *Techniques d'enquête*, 20, 193-200.

Schaible, W.L. (Ed.) (1996), Indirect Estimation in U.S. Federal Programs. New York: Springer

Scott, A., et Davis, P. (2001), Estimating interviewer effects for survey responses. *Proceedings of Statistics Canada Symposium 2001*.

Shao, J. (2002), Resampling methods for variance estimation in complex surveys with a complex design. Dans *Survey Nonresponse* (Éds. R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), New York: John Wiley & Sons, Inc., 303-314.

Shao, J., et Tu, D. (1995), *The Jackknife and the Bootstrap.* New York: Springer Verlag.

Singh, A.C., Kennedy, B. et Wu, S. (2001), Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête*, 27, 35-48.

Singh, A.C., et Mohl, C.A. (1996), Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.

Singh, A.C., et Wu, S. (1996), Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 69-77.

Singh, M.P., Gambino, J. et Mantel, H.J. (1994), Les petites régions : Problèmes et solutions. *Techniques d'enquête*, 20, 3-15.

Sirken, M.G. (1970), Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

Sitter, R.R., et Wu, C. (2001), A note on Woodruff confidence interval for quantiles. *Statistics & Probability Letters*, 55, 353-358.

Skinner, C.J., et Rao, J.N.K. (1996), Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Skinner, C.J., Holt, D. et Smith, T.M.F. (Éds.) (1989), *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Smith, T.M.F. (1976), The foundations of survey sampling: A review. *Journal of the Royal Statistical Society*, Series A, 139, 183-204.

Smith, T.M.F. (1994), Sample surveys 1975-1990; an age of reconciliation? *Revue Internationale de Statistique*, 62, 5-34.

Stehman, S.V., et Overton, W.S. (1994), Comparison of variance estimators of the Horvitz Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.

Sukhatme, P.V. (1947), The problem of plot size in large-scale yield surveys. *Journal of the American Statistical Association*, 42, 297-310.

Sukhatme, P.V. (1954), *Sampling Theory of Surveys, with Applications.* Ames: Iowa State College Press.

Sukhatme, P.V., et Panse, V.G. (1951), Crop surveys in India – II. *Journal of the Indian Society of Agricultural Statistics*, 3, 97-168.

Sukhatme, P.V., et Seth, G.R. (1952), Non-sampling errors in surveys. *Journal of the Indian Society of Agricultural Statistics*, 4, 5-41.

Thomas, D.R., et Rao, J.N.K. (1987), Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

Thompson, S.K., et Seber, G.A.F. (1996), *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

Tillé, Y. (1998), Estimation in surveys using conditional inclusion probabilities: simple random sampling. *Revue Internationale de Statistique*, 66, 303-322.

Tschuprow, A.A. (1923), On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2, 461-493, 646-683.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000), *Finite Population Sampling and Inference*: *A Prediction Approach*. New York: John Wiley & Sons, Inc.

Waksberg, J. (1978), Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Waksberg, J. (1998), The Hansen era: Statistical research and its implementation at the U.S. Census Bureau. *Journal of Official Statistics*, 14, 119-135.

Wang, N., et Robins, J.M. (1998), Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

Woodruff, R.S. (1952), Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C., et Rao, J.N.K. (2004), Empirical likelihood ratio confidence intervals for complex surveys. Soumis pour publication.

Wu, C., et Rao, J.N.K. (2005), Empirical likelihood approach to calibration using survey data. Article présenté à la réunion 2005 International Statistical Institute, Sydney, Australie.

Yates, F. (1949), *Sampling Methods for Censuses and Surveys*. London: Griffin.

Zarkovic, S.S. (1956), Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society,* Series A, 119, 336-338.