Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

L'échantillonnage pour les statistiques officielles – Certaines réflexions sur les orientations

Ray Chambers[1]

Résumé

Dans la perspective d'un modélisateur, je décris la situation actuelle en matière d'inférence fondée sur les enquêtes pour la production de statistiques officielles. Ce faisant, je tente de dégager les forces et les faiblesses des approches inférentielles fondées sur le plan de sondage, d'une part, et sur un modèle, d'autre part, appliquées aujourd'hui à l'échantillonnage, du moins en ce qui concerne les statistiques officielles. Je termine par un exemple tiré d'un plan de collecte adaptatif qui illustre pourquoi l'adoption d'une perspective fondée sur un modèle (fréquentiste ou bayésien) représente le meilleur moyen pour les statistiques officielles d'éviter la « schizophrénie inférentielle » débilitante qui semble inévitable si l'on applique les méthodes actuelles pour répondre aux nouvelles demandes d'information du monde d'aujourd'hui (et peut-être même de demain).

Mots clés : Inférence fondée sur le plan de sondage; inférence assistée par modèle; inférence fondée sur un modèle; pondération calée; modèle bayésien calé; plan de collecte adaptatif.

1. Introduction

Le présent article donne un aperçu des différentes approches courantes d'inférence pour des enquêtes officielles par échantillonnage (EOE), en mettant l'accent sur leur pertinence en regard de l'évolution continue de l'environnement de production des statistiques officielles qui repose à l'heure actuelle sur des méthodes de sondage. En particulier, je décrirai mon optique de l'inférence pour des EOE passée, ou inférence fondée sur le plan de sondage, de l'inférence pour des EOE actuelle, ou inférence assistée par un modèle, et de l'inférence pour des EOE future, ou inférence fondée sur un modèle. Ce faisant, je soulignerai le changement de paradigme en cours dans le domaine de l'inférence pour les statistiques officielles (pour passer d'une approche fondée sur le plan de sondage et assistée par un modèle à une approche fondée sur un modèle), et les implications pour la méthodologie des statistiques officielles dans le futur.

Afin d'illustrer ce changement, je discute plus bas de l'utilité du raisonnement fondé sur un modèle pour élaborer une stratégie de plan de collecte adaptatif pour tenir compte de la non-réponse informative. Il ne s'agit toutefois que d'un exemple parmi d'autres. Personnellement, j'ai appliqué des idées fondées sur un modèle à l'estimation sur petits domaines, à l'analyse de données appariées de façon probabiliste, à l'inférence au niveau de la population en utilisant de nouvelles sources d'information auxiliaire, p. ex. l'information sur des réseaux de population, ainsi qu'à des situations où l'information provenant de sondages et de registres de population doit être combinée.

Mon intérêt pour ces questions a été suscité en grande partie par certains commentaires de Frauke Kreuter dans l'exposé qu'elle a donné à la Conférence Graybill de 2013, où elle a déclaré que [Traduction] « … ces dernières années, les grands organismes d'enquête ont fait beaucoup d'efforts pour renforcer l'information sur tous les cas d'échantillonnage en se servant de paradonnées, de données de fournisseurs commerciaux et de l'appariement à des données administratives en vue d'améliorer les opérations sur le terrain ou les ajustements pour la non-réponse. » Dans ma lettre du président subséquente parue dans le Bulletin de l'Association internationale des statisticiens d'enquête en juin 2013, en m'appuyant sur ces idées, j'ai souligné que [Traduction] « l'inférence fondée sur l'échantillonnage devra être adaptée à ce nouveau paradigme de collecte des données qui réduit considérablement

[1]National Institute for Applied Statistics Research Australia, University of Wollongong, Australie, 2522 (ray@uow.edu.au)

l'importance de l'erreur d'échantillonnage et requiert vraiment que l'on s'attelle à déterminer comment caractériser des notions fondamentales, telles que l'incertitude, dans le mélange résultant d'erreurs dues à la non-réponse, d'erreurs d'appariement, d'erreurs de mesure et d'erreurs de spécification du modèle. »

Dans le présent article, je tente de définir plus clairement ce que j'entendais par ces commentaires. Cependant, il est utile pour cela d'établir d'abord certains concepts fondamentaux et la notation.

2. Concepts fondamentaux

Le concept d'une population finie est au cœur de la collecte de données pour la production de statistiques officielles. Il s'agit de la population de N unités pour laquelle l'information est requise. Cela implique à son tour l'existence (du moins conceptuellement) d'une liste U de N unités constituant cette population, telle que chaque unité de la population est identifiable sur cette liste au sens où la liste contient un identificateur ou une étiquette unique pour chaque unité de la population. Sans perte de généralité, nous indiçons la population finie par cet identificateur, de sorte qu'il prend les valeurs $i = 1,\ldots,N$.

En ayant cette liste, nous pouvons, sans perte de généralité, supposer aussi que nous connaissons les valeurs de population Z_i d'une variable auxiliaire Z . Cela pourrait n'être rien de plus que l'identificateur d'une unité de population. Cependant, dans la plupart des cas, il s'agira de bien davantage. La chose la plus importante est que chaque valeur de population de Z peut être associée à une valeur unique Z_i de l'identificateur de population. Sans perte de généralité, nous supposons que Z est scalaire. Enfin, nous caractérisons les valeurs d'intérêt inconnues de la population par les valeurs Y_i d'une variable cible scalaire Y . Le principal objectif est alors d'inférer la valeur d'une fonction bien définie Q des valeurs de population de Y et Z . Dans ce contexte, le total de population de Y est souvent pris pour objectif, mais il ne s'agit que d'un parmi de nombreux objectifs potentiels pour un exercice de collecte de données pour la production de statistiques officielles.

Une méthode scientifique fondamentale utilisée pour des EOE est l'échantillonnage aléatoire. En d'autres termes, pour chaque unité i dans la population finie, l'échantillonneur génère une valeur I_i pour une variable indicatrice aléatoire I qui est égale à 1 (0) si cette unité (c.-à-d. étiquette) particulière de la population est échantillonnée (n'est pas échantillonnée). Soit $s = \{i; I_i = 1\}$ les étiquettes des n unités échantillonnées de la population. Nous pouvons alors définir les vecteurs de population suivants : $\mathbf{I}_U$, le vecteur de dimension N des valeurs de population de I ; $\mathbf{Z}_U$, le vecteur de dimension N des valeurs de population de Z ; $\mathbf{Y}_U$, le vecteur de dimension N des valeurs de population de Y ; et $\mathbf{Y}_s$, le vecteur de dimension n des valeurs d'échantillon de Y . Notons que la distribution de $\mathbf{I}_U | \mathbf{Z}_U$ définit ce qui est généralement appelé le plan de sondage. Voir Smith (1983), Sugden et Smith (1984) et Pfeffermann (1993).

L'échantillonnage aléatoire est essentiel aux EOE parce qu'il garantit un échantillonnage non informatif sachant $\mathbf{Z}_U$. Il s'agit de toute méthode d'échantillonnage telle que la distribution de $\mathbf{I}_U$ dépend uniquement des valeurs de population connues dans $\mathbf{Z}_U$. Autrement dit, la distribution de $\mathbf{I}_U$ est spécifiée entièrement sachant $\mathbf{Z}_U$. Il s'ensuit directement que, sous échantillonnage non informatif sachant $\mathbf{Z}_U$, les distributions de $\mathbf{I}_U$ et $\mathbf{Y}_U$ sont conditionnellement indépendantes sachant $\mathbf{Z}_U$.

Nous notons que l'hypothèse d'échantillonnage non informatif n'est pas appropriée s'il peut exister un biais de sélection et, par conséquent, une dépendance entre les distributions de $\mathbf{Y}_U$ et $\mathbf{I}_U$, même après conditionnement sur $\mathbf{Z}_U$. Presque toujours, le biais de sélection n'est préoccupant que si l'échantillonnage est effectué par « quelqu'un d'autre » (le cas d'une analyse secondaire) ou que l'échantillon sélectionné et l'échantillon réalisé ne sont pas identiques, principalement en raison des non-contacts et de la non-réponse. Cette seconde situation est celle qui est habituellement la plus préoccupante pour les EOE.

3. Le rôle de l'inférence fondée sur le plan de sondage dans les EOE

L'inférence fondée sur le plan de sondage est à l'heure actuelle le paradigme d'inférence standard adopté pour les statistiques officielles et représente la philosophie qui sous-tend la plupart des textes classiques sur la théorie des sondages. Ce paradigme, dont l'origine est habituellement associée à Neyman (1934) traite toutes les valeurs de population comme étant générées par la « nature », c.-à-d. que $\mathbf{Y}_U$ et $\mathbf{Z}_U$ sont des vecteurs de paramètres de population finie sur lesquels se fait par conséquent le conditionnement dans toute inférence. Toute fonction inconnue des valeurs de la population (p. ex., Q) est alors également un paramètre de population finie dont la valeur doit être inférée d'après les données d'échantillon. En particulier, on suppose qu'il existe une fonction $\hat{Q}\left(\mathbf{Z}_U, \mathbf{Y}_s\right)$ de ces données qui est un estimateur de Q.

Dans ces conditions, il est clair que la seule variable aléatoire qui peut être utilisée pour l'inférence au sujet de Q sachant $\hat{Q}$ est celle dont la distribution est (du moins en théorie) entièrement connue, c.-à-d. $\mathbf{I}_U$. En fait, l'inférence est faite par rapport aux valeurs potentielles de $\hat{Q}\left(\mathbf{Z}_U, \mathbf{Y}_s\right)$ qui peuvent être générées sachant les valeurs possibles de $\mathbf{I}_U$ (et donc s) et les valeurs fixes de $\mathbf{Y}_U$ et $\mathbf{Z}_U$. Une inférence valide requiert l'absence de biais par rapport au plan de sondage (si possible) ou la cohérence avec le plan de sondage (au moins), au sens où, pour tout choix d'une distribution pour $\mathbf{I}_U$ (le processus d'échantillonnage), la distribution de la variable aléatoire $\hat{Q}$ doit être telle que $E\left(\hat{Q}-Q\middle|\mathbf{Z}_U, \mathbf{Y}_U\right) \approx 0$. Une inférence efficace requiert alors que l'on trouve pour $\mathbf{I}_U$ une distribution qui rend $E\left(\left(\hat{Q}-Q\right)^2\middle|\mathbf{Z}_U, \mathbf{Y}_U\right)$ aussi petite que possible, habituellement sous une contrainte de coût. Si aucune restriction n'est imposée sur $\mathbf{Y}_U$, la tâche est impossible, comme le montre clairement le résultat de non-existence de Godambe (1955). Voir Basu (1971) pour une preuve accessible.

Comme il est mentionné plus haut, la théorie des sondages fondée sur le plan a tendance à se concentrer sur l'estimation du total de population T de Y, et en particulier l'estimation linéaire de ce total, de sorte que l'estimateur privilégié de T est de la forme

$$\hat{T} = \sum\nolimits_{i \in s} w_i Y_i = \sum_{i=1}^{N} w_i I_i Y_i .$$

Ici, w_i est un poids qui dépend uniquement des données auxiliaires $\mathbf{Z}_U$. Puisque les distributions de w_i et I_i sont entièrement déterminées par $\mathbf{Z}_U$, la condition nécessaire pour l'absence de biais sous le plan est

$$E(w_i I_i \middle| \mathbf{Z}_U, \mathbf{Y}_U) = E(w_i I_i \middle| \mathbf{Z}_U) = 1$$

ou la définition plus connue de Horvitz-Thompson $w_i = \pi_i^{-1}$, où $\pi_i = E(I_i \middle| \mathbf{Z}_U)$. Autrement dit, même si ce choix de poids peut ne pas être efficace, il donnera toujours lieu à une inférence valide. Par conséquent, l'estimateur de Horvitz-Thompson et ses variantes ont joué un rôle fondamental dans les EOE fondées sur le plan depuis les années 1940.

À ce stade, il convient d'être réaliste. Les bases de sondage inadéquates, la non-réponse et le ciblage des prises de contact et de l'obtention de réponses par les intervieweurs ainsi que les gestionnaires d'enquête sont des facteurs qui signifient tous que la distribution réelle de $\mathbf{I}_U$ ne concorde presque jamais avec sa distribution « selon le plan ». Ce fait, qui a toujours été connu des statisticiens travaillant dans le domaine des EOE, constitue un énorme défi en ce qui concerne la mise en œuvre appropriée de l'inférence fondée sur le plan de sondage dans le cas des EOE. La divergence de la distribution des indicateurs associés aux unités de population qui fournissent effectivement les données d'enquête par rapport à la distribution des indicateurs d'inclusion dans l'échantillon déterminée par le plan de sondage implique que l'échantillonneur n'exerce pas le plein contrôle sur le mécanisme d'échantillonnage qui sous-tend l'échantillon réalisé (plutôt que sélectionné), et qu'il ne « connaît » donc pas ses propriétés. En particulier, l'hypothèse cruciale d'échantillonnage non informatif $E(I_i \middle| \mathbf{Z}_U, \mathbf{Y}_U) = E(I_i \middle| \mathbf{Z}_U)$ peut alors ne plus être valide. Dans cette situation, une inférence valide fondée sur la pondération de Horvitz-Thompson risque de ne plus être possible. Une autre approche d'inférence est nécessaire.

4. Le rôle de l'inférence assistée par un modèle dans les EOE

Comme l'indique clairement la preuve du résultat de non-existence de Godambe, le principal problème de l'inférence fondée sur le plan de sondage est qu'elle est trop générale, puisque ses propriétés doivent être vérifiées pour tout $\mathbf{Y}_U$. Par conséquent, les praticiens des EOE ont adopté dès le départ, et certainement dès les années 1940, des stratégies d'échantillonnage qui sont, en un certain sens, raisonnables sur l'espace des réalisations possibles de $\mathbf{Y}_U$. De telles valeurs de $\mathbf{Y}_U$ sont habituellement déterminées en émettant l'hypothèse d'un modèle pour la distribution de $\mathbf{Y}_U$ sachant $\mathbf{Z}_U$. Ce modèle est utilisé dans la spécification du plan d'échantillonnage (définition des strates, des grappes, des probabilités d'inclusion) ainsi que dans la spécification de l'estimateur (estimateurs par le ratio, par la régression).

En émettant l'hypothèse supplémentaire d'un modèle pour $\mathbf{Y}_U$ sachant $\mathbf{Z}_U$, il est théoriquement possible de déterminer des plans de sondage optimaux en minimisant l'espérance sous le modèle de l'erreur quadratique moyenne (EQM) sous le plan de sondage. Cependant, cela se fait rarement. Le modèle est plutôt utilisé pour créer des poids calés pour produire des estimations assistées par un modèle, où les poids « de sondage » $w_i = \pi_i^{-1}$ sont modifiés pour rétablir les chiffres de population connus (totaux de contrôle) à partir des données observées de l'échantillon. Comme le choix des totaux de contrôle de la population d'intérêt équivaut à spécifier un modèle linéaire pour $\mathbf{Y}_U$ par rapport aux variables qui définissent les totaux de contrôle, la modélisation de la population oriente nécessairement ce choix. Par conséquent, l'estimateur de T fondé sur ces poids calés est alors un estimateur sans biais de l'espérance de T sous ce modèle.

Cependant, l'inférence demeure fondée sur le plan de sondage sous l'approche assistée par un modèle. En particulier, l'exigence clé de validité est encore en vigueur, de sorte que les exigences d'absence de biais et de convergence par rapport au plan sont considérées comme cruciales. Autrement dit, même si des modèles sont utilisés pour définir des stratégies d'estimation, toute inférence résultante demeure fondée sur le plan de sondage, du moins en principe.

Cette validité perçue est souvent utilisée pour justifier l'affirmation selon laquelle l'inférence assistée par un modèle pour les EOE est robuste à l'erreur de spécification du modèle parce qu'elle requiert une absence de biais sous le plan exacte ou approximative faisant que l'inférence est correcte quelle que soit la « vraie » nature de la distribution de $\mathbf{Y}_U$ sachant $\mathbf{Z}_U$. Comme l'a dit un ami et collègue (Ken Brewer), [Traduction] « (a)dopter l'approche assistée par un modèle équivaut à porter une ceinture et des bretelles pour retenir son pantalon. Si la ceinture (le modèle) casse, on ne sera pas totalement embarrassé, puisque les bretelles (absence de biais sous le plan) devraient encore maintenir le pantalon en place. »

Cette remarque est raisonnable, puisque toute spécification de modèle est fausse. Cependant, elle ne tient pas compte du fait que l'inférence résultante peut être très inefficace. Une légère adaptation de Kendall (1959) me paraît saisir l'essentiel ici. Kendall nous offre une révision pleine d'esprit du poème épique de Longfellow dans laquelle nous notons que Hiawatha mise sur la randomisation plutôt que sur l'entraînement au tir pour essayer de gagner un concours de tir à l'arc, en partant du principe que sa technique de tir est savamment randomisée et que ses flèches sont toutes assistées par modèle, si bien que sa stratégie de tir à l'arc est convergente sous le plan et, par conséquent, très proche d'être sans biais pour la cible réelle. Lorsque l'inévitable se produit et qu'il est classé dernier du concours, Hiawatha se retire dans la forêt, où [Traduction]

« Dans un coin de la forêt
Assis solitaire mon Hiawatha
Ne cesse de réfléchir
À la loi normale des erreurs.
Se demandant dans les moments de répit
S'il se pourrait qu'une plus grande précision
Soit parfois meilleure
Même au prix d'un biais,
Si l'on pouvait ainsi de temps à autre
Atteindre une cible. »

5. Le rôle de l'inférence fondée sur un modèle dans les EOE

Contrairement à l'approche fondée sur le plan de sondage, l'approche fondée sur un modèle dans les EOE n'a pas de point de départ évident. Brewer (1963) est peut-être le premier à avoir tenté de fonder explicitement l'inférence sur un modèle dans un contexte de sondage. Cependant, les travaux de loin les plus influents concernant cette approche ont été effectués indépendamment par Royall et ses collègues à partir de 1969 environ. Voir Royall (1976a) pour un exposé clair et complet des idées de base, étayé par l'observation que l'hypothèse d'échantillonnage non informatif signifie que $\mathbf{I}_U$ est auxiliaire pour l'inférence au sujet de $\mathbf{Y}_U$ sachant $\mathbf{Z}_U$, et donc que l'inférence au sujet de Q doit par conséquent être conditionnelle à $\mathbf{I}_U$ et $\mathbf{Z}_U$, et non $\mathbf{Y}_U$ et $\mathbf{Z}_U$.

Sous l'approche fondée sur un modèle, une exigence naturelle pour un estimateur $\hat{Q}$ linéaire est l'absence de biais sous le modèle. C'est-à-dire $E\left(\hat{Q}-Q\middle|\mathbf{I}_U,\mathbf{Z}_U\right)=0$. En outre, l'incertitude est quantifiée par l'erreur quadratique moyenne de prédiction (EQMP), $E\left(\left(\hat{Q}-Q\right)^2\middle|\mathbf{I}_U,\mathbf{Z}_U\right)$. Des arguments fondés sur la prédiction statistique classique permettent alors de définir les poids optimaux à utiliser pour l'estimation et l'inférence à partir d'un échantillon. En fait, l'inférence en échantillonnage devient un sous-ensemble de la théorie de la prédiction statistique classique.

Dans le contexte des EOE, on a pratiquement ignoré les idées fondées sur un modèle (ou résisté à ces idées). Cela semble en grande partie tenir à l'observation rassurante selon laquelle, étant donné un modèle linéaire pour $\mathbf{Y}_U$, la propriété d'absence de biais sous le modèle pour un estimateur linéaire équivaut à utiliser des poids calés, où le calage est effectué sur les totaux de population des covariables du modèle. Donc, la pondération assistée par un modèle est considérée comme étant, « dans l'esprit », fondée sur un modèle. Cependant, cela découle aussi du fait que, aux premières étapes de l'élaboration de l'approche fondée sur un modèle, on a beaucoup insisté sur le fait qu'en émettant l'hypothèse d'un modèle pour $\mathbf{Y}_U$ impliquait l'existence d'un échantillon optimal qui minimisait l'EQMP d'un estimateur. Cela peut être considéré comme étant sans grande conséquence dans de nombreux modèles de population utilisés fréquemment, p. ex. quand la relation de régression entre Y et Z est linéaire et que la variance de l'erreur est constante, auquel cas l'échantillon optimal pour l'estimateur par la régression de T est un échantillon équilibré où la moyenne d'échantillon de Z est égale à la moyenne de population de Z, propriété qui est vérifiée en espérance sous échantillonnage aléatoire simple. Toutefois, cela est aussi conceptuellement inquiétant, puisque cela semble éliminer le besoin fondamental de randomisation en inférence fondée sur un échantillon. En particulier, sous le modèle très répandu de « régression linéaire passant par l'origine » pour Y en fonction d'une variable de taille Z, où la variance de Y sachant Z est proportionnelle à une puissance non négative de Z, l'échantillon optimal pour l'estimateur fondé sur un modèle optimal de T consiste en les n unités de population ayant les valeurs les plus grandes de Z. Dans la perspective des EOE, cela était généralement considéré comme un pas de trop, quoique Karmel et Jain (1987) offrent une évaluation explicite et positive de cette stratégie pour des EOE.

L'élaboration d'approches bayésiennes de l'inférence fondée sur un modèle en échantillonnage a suivi rapidement une fois que le cadre théorique de prédiction pour cette inférence a été établi. Ici, nous résumons cette approche, en utilisant p pour désigner la distribution des données dans la population sous le modèle supposé, et *prior*, *post* pour désigner les distributions a priori et a posteriori, respectivement, des quantités inconnues dans le modèle. Sous cette spécification, la distribution prédictive a posteriori de $Q=Q(\mathbf{Y}_U,\mathbf{Z}_U)$ s'écrit alors

$$post\left(Q\middle|\mathbf{Y}_s,\mathbf{I}_U,\mathbf{Z}_U\right)\propto\int \underbrace{p\left(Q\middle|\mathbf{Y}_s,\mathbf{I}_U,\mathbf{Z}_U;\theta,\phi\right)}_{\text{densité conditionnelle basée sur le modèle de } Q} post\left(\theta,\phi\middle|\mathbf{Y}_s,\mathbf{I}_U,\mathbf{Z}_U\right)d\theta d\phi\,.$$

Sous échantillonnage non informatif, on peut montrer que $p\left(Q\middle|\mathbf{Y}_s,\mathbf{I}_U,\mathbf{Z}_U;\theta,\phi\right)=p\left(Q\middle|\mathbf{Y}_s,\mathbf{Z}_U;\theta\right)$ et

$$post\left(\theta,\phi\middle|\mathbf{Y}_s,\mathbf{I}_U,\mathbf{Z}_U\right)\propto\underbrace{p\left(\mathbf{Y}_s\middle|\mathbf{Z}_U;\theta\right)prior\left(\theta\middle|\mathbf{Z}_U\right)}_{post\left(\theta\middle|\mathbf{Y}_s,\mathbf{Z}_U\right)}\underbrace{p\left(\mathbf{I}_U\middle|\mathbf{Z}_U;\phi\right)prior\left(\phi\middle|\mathbf{Z}_U\right)}_{post\left(\phi\middle|\mathbf{I}_U,\mathbf{Z}_U\right)}.$$

Par conséquent,

$$post\left(Q\middle|\mathbf{Y}_s,\mathbf{I}_U,\mathbf{Z}_U\right)\propto\int p\left(Q\middle|\mathbf{Y}_s,\mathbf{Z}_U;\theta\right)p\left(\mathbf{Y}_s\middle|\mathbf{Z}_U;\theta\right)prior\left(\theta\middle|\mathbf{Z}_U\right)d\theta.$$

En général, il n'est pas possible de déterminer analytiquement cette fonction. Cependant, on peut la calculer par simulation, en utilisant des logiciels largement disponibles (p. ex., BUGS / BRUGS).

Plus récemment, des signes d'un apparent rapprochement entre l'idéologie fondée sur le plan de sondage, mais assistée par modèle et celle « entièrement » fondée sur un modèle ont commencé à apparaître, sous la forme de ce que Little (2006, 2012) appelle un modèle bayésien calé. Les desiderata de cette approche s'articulent autour de l'exigence que tout modèle utilisé en analyse bayésienne des données d'enquête doit inclure les caractéristiques du plan de sondage. En particulier, l'approche « bayésienne calée » doit utiliser des modèles robustes ayant de bonnes propriétés d'échantillonnage répété, puisque les modèles qui ne tiennent pas compte de caractéristiques telles que les poids de sondage sont vulnérables à l'erreur de spécification. L'objectif fondamental est alors d'orienter la spécification des modèles bayésiens vers des modèles dans lesquels sont intégrées les caractéristiques du plan et qui peuvent donc donner des inférences ayant de bonnes propriétés sous le plan de sondage. Les lignes directrices pour y arriver comprennent l'utilisation des poids de sondage comme covariables dans le modèle de prédiction (p. ex., Gelman, 2007) et la modélisation de plans de sondage à plusieurs degrés en utilisant des modèles à effets aléatoires multiniveaux, une idée qui remonte à Royall (1976b) et qui est vraiment la « norme de l'industrie » en estimation sur petits domaines à l'heure actuelle.

6. Un peu de recul

Arrivé ici, prenons un peu de recul et faisons le point. Il est incontestable que l'échantillonnage probabiliste et la loi des grands nombres font en sorte qu'un estimateur convergent sous le plan de sondage de Q aura une très faible probabilité de s'écarter beaucoup de sa valeur réelle si l'échantillon est grand. Cependant, quelle taille doit avoir un « grand » échantillon pour que l'asymptotique devienne pertinente? J'hasarderais l'idée que les résultats asymptotiques sont sans pertinence pour la plupart des produits des sondages dans le contexte des EOE. En outre, le recours généralisé à l'estimation sur petits domaines fondée sur un modèle soulève la question délicate de savoir pourquoi, si l'inférence fondée sur un modèle est suffisante pour les petits échantillons, ne l'est-elle pas pour les grands échantillons. Cette curieuse dichotomie de la « respectabilité » inférentielle dans l'EOE a été qualifiée de « schizophrénie inférentielle » par Little (2012). Enfin, je souligne que la robustesse de la validité que confère l'utilisation de l'inférence fondée sur le plan de sondage n'est pas la robustesse de l'efficacité. En particulier, j'ai souvent constaté que des méthodes ayant de bonnes propriétés sous le plan de sondage en grand échantillon peuvent être (et sont souvent) inefficaces pour toute taille particulière d'échantillon.

Mais rien n'est gratuit. L'inférence fondée sur un modèle n'est pas valide sous l'erreur de spécification du modèle, de sorte que la recherche d'une bonne spécification du modèle est essentielle. Si le modèle est « invalide », la stratégie optimale d'échantillonnage (et d'estimation) fondée sur ce modèle peut donner lieu à des erreurs importantes. L'élément important ici est de savoir ce que l'on entend par choix d'un modèle invalide. La déclaration souvent citée (habituellement attribuée à George Box) selon laquelle « tous les modèles sont faux, mais certains sont utiles » s'applique ici. Il est trompeur de dire que l'adoption d'une approche inférentielle fondée sur un modèle exige que l'on décide d'une certaine façon avant d'avoir vu les données d'échantillon qu'une spécification particulière d'un modèle pour la relation entre $\mathbf{Y}_U$ et $\mathbf{Z}_U$ devrait être le fondement de toute inférence ultérieure au sujet de Q. Les résultats présentés dans Hansen, Madow et Tepping (1983) illustrent les conséquences de l'adoption d'une stratégie d'échantillonnage qui n'est pas robuste à la spécification du modèle et une spécification d'un modèle de travail qui ne tient pas compte de l'information au sujet de la défaillance du modèle figurant dans les données d'échantillon. La modélisation robuste aux erreurs de spécification requiert de la minutie, ce qui explique pourquoi l'usage prudent de la stratification (pré- et post-) est important, de même que l'inclusion des covariables pertinentes. Dans ce contexte, le conseil habituel au sujet de l'importance des diagnostics du modèle et l'application de stratégies de modélisation adaptatives (y compris la stratification et l'utilisation de multiples covariables, si elles sont disponibles) est essentiel.

La modélisation non paramétrique est une stratégie de modélisation adaptable aux données. Autrement dit, nous utilisons des méthodes non paramétriques pour modéliser $E\left(\mathbf{Y}_U \middle| \mathbf{Z}_U\right)$, en supposant une spécification du modèle flexible de la forme

$$\mathbf{Y}_U = m(\mathbf{Z}_U) + \mathbf{e}_U$$

où $\mathbf{e}_U$ est le vecteur des erreurs de modélisation dans la population et m est une fonction suffisamment lisse des covariables dans $\mathbf{Z}_U$. Les méthodes populaires pour l'ajustement de m comprennent la méthode du noyau et les

méthodes fondées sur les splines, qui permettent les unes et les autres d'écrire l'estimateur final sous une forme linéaire (c.-à-d. pondérée).

L'un des problèmes qui découlent directement de cette approche est que, en général, les poids non paramétriques ne sont pas calés. Ce problème est facile à résoudre, mais il est intéressant de comparer l'approche de ce calage assistée par un modèle et fondée sur un modèle.

L'approche de calage assistée par un modèle est habituellement associée à Deville et Särndal (1992). L'idée ici est de choisir un vecteur de poids de sondage $\mathbf{w}_s^{cal}$ qui est proche du vecteur de poids de sondage non paramétriques $\mathbf{w}_s^{np}$, mais satisfaisant aussi les contraintes de calage (c.-à-d. que les poids $\mathbf{w}_s^{cal}$ sont calés sur les colonnes de $\mathbf{Z}_U$). Cela équivaut naturellement à une correction paramétrique du biais sous le modèle linéaire spécifié par les contraintes de calage. En tout cas, une mesure de « proximité » utilisée habituellement est la mesure euclidienne

$$V = (\mathbf{w}_s^{cal} - \mathbf{w}_s^{np})^T \, diag(\mathbf{v}_s)(\mathbf{w}_s^{cal} - \mathbf{w}_s^{np})$$

où $\mathbf{v}_s$ est un ensemble spécifié de constantes qui est supposé caractériser l'hétéroscédasticité des résidus générés par ce modèle. Minimiser V sous la contrainte du calage donne les poids de sondage calés de la forme

$$\mathbf{w}_s^{cal} = \mathbf{w}_s^{np} + \underbrace{\mathbf{H}^T(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{w}_s^{np})}_{correction\ par\ calage}$$

où $\mathbf{H}$ est une matrice telle que $\mathbf{H}\mathbf{y}_s$ définit l'estimateur du vecteur des coefficients du modèle linéaire sous le modèle. Notons que poser les poids non paramétriques égaux aux poids de type Horvitz-Thompson (c.-à-d. l'inverse des probabilités d'inclusion dans l'échantillon) mène à l'estimateur par la régression généralisée (GREG). En tout cas, l'estimateur calé défini par $\mathbf{w}_s^{cal}$ sera sans biais par rapport au modèle sous le modèle linéaire défini par $\mathbf{Z}_U$.

Dans une perspective fondée sur un modèle, l'interprétation du calage diffère quelque peu. Cette approche a pour point de départ un modèle de travail linéaire défini par les colonnes de $\mathbf{Z}_U$. Cependant, nous reconnaissons que ce modèle n'est qu'une approximation de la réalité, et nous ajoutons donc prudemment une correction non paramétrique du biais à l'estimateur optimal par la régression sous ce modèle de travail (approximatif). Dans ces conditions, Chambers, Dorfman et Wehrly (1993) proposent d'utiliser les résidus d'échantillon pour calculer une estimation non paramétrique du biais, puis de soustraire cette estimation du biais de l'estimation optimale originale fondée sur le modèle. En d'autres termes, étant donné un modèle de travail de la forme $\mathbf{y}_s = \mathbf{Z}_s\boldsymbol{\beta} + \mathbf{e}_s$, où $\mathbf{e}_s \sim (\mathbf{0}, diag(\mathbf{v}_s))$, nous utilisons les moindres carrés ordinaires pour obtenir les valeurs prédites $\hat{\mathbf{y}}_s = \mathbf{Z}_s\mathbf{H}\mathbf{y}$ et les résidus $\mathbf{r}_s = (\mathbf{I}_n - \mathbf{Z}_s\mathbf{H})\mathbf{y}_s$. Puis, nous utilisons le vecteur $\mathbf{m}_s$ de poids non paramétriques pour estimer le total de population de ces résidus, et nous ajoutons cette estimation à l'estimateur optimal par la régression sous le modèle linéaire original. Cela équivaut naturellement à utiliser les poids fondés sur le modèle corrigés non paramétriquement du biais

$$\begin{aligned}
\mathbf{w}_s^{p+np} &= \underbrace{\mathbf{1}_n + \mathbf{H}^T(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{1}_n)}_{poids\ paramétrique\ (BLUP)} + \underbrace{(\mathbf{I}_n - \mathbf{H}^T\mathbf{Z}_s^T)\mathbf{m}_s}_{correction\ non\ paramétrique\ du\ biais} \\
&= \mathbf{1}_n + \mathbf{H}^T(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{1}_n) + (\mathbf{I}_n - \mathbf{H}^T\mathbf{Z}_s^T)(\mathbf{w}_s^{np} - \mathbf{1}_n) \\
&= \mathbf{w}_s^{np} + \underbrace{\mathbf{H}^T(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{w}_s^{np})}_{correction\ par\ calage}.
\end{aligned}$$

Autrement dit, on obtient exactement la même correction que celle donnée par le calage assisté par un modèle. Puisque l'efficacité de l'approche fondée sur un modèle dépend de la correction non paramétrique du biais, nous voyons aussi que l'ensemble de poids que l'on choisit comme point de départ pour le calage importe considérablement.

7. Implications pour le plan et l'estimation dans les EOE

Un problème d'ordre très pratique doit être surmonté avant que l'on puisse espérer que les idées fondées sur un modèle gagnent du terrain dans le domaine des EOE. Cela tient au fait que, dans le contexte des EOE, les enquêtes sont habituellement polyvalentes et recueillent de nombreux éléments d'information différents mesurés sur des échelles distinctes, ce qui implique de nombreux modèles et plans de sondage optimaux associés différents. Il est impossible d'imposer tous ces plans de sondage simultanément. Dans ce contexte, la stratification, la sélection systématique (si la population peut être ordonnée), la sélection aléatoire (quand elle ne peut pas l'être) sont des outils qui sont tous

utiles pour créer des plans de sondage polyvalents donnant des échantillons que l'on peut affirmer être « représentatifs » d'une population.

Autant que je sache, il n'existe aucune mesure de la représentativité, mais une approximation raisonnable consiste à choisir un échantillon équilibré, c.-à-d. un échantillon dont la distribution des valeurs dans $\mathbf{Z}_s$ fait écho à celle dans $\mathbf{Z}_U$. Dans la plupart des cas, cela signifie qu'il faut sélectionner un échantillon qui est au moins calé approximativement sur un modèle de travail fondé sur $\mathbf{Z}_U$, c.-à-d. un modèle tel que les estimations des totaux de contrôle connus sont exactes. Cependant, je souligne aussi les travaux récents sur les plans de sondage robustes par rapport au modèle de Welsh et Wiens (2013), où un plan de sondage (c.-à-d. une valeur pour $\mathbf{Z}_s$) est choisi de manière à minimiser l'erreur de prédiction maximale qui pourrait avoir lieu quand le modèle réel se trouve dans un voisinage du modèle de travail.

À ce stade, il paraît raisonnable de se demander quel est le rôle de la randomisation dans le plan de sondage sous l'approche fondée sur un modèle. Il est clair qu'elle fait partie intégrante de l'approche fondée sur le plan de sondage ainsi que de l'approche assistée par un modèle. Sans sélection aléatoire de l'échantillon, il n'y a pas d'inférence dans ces cas. Par contre, l'inférence fondée sur un modèle n'impose aucune contrainte (outre le caractère non informatif) sur la méthode effectivement utilisée pour sélectionner l'échantillon. Ce sont les caractéristiques de l'échantillon choisi qui importent (p. ex., ses caractéristiques d'équilibre). Mais, à moins d'être strictement contrôlé, l'échantillonnage probabiliste peut aussi donner des échantillons fortement non robustes dans une perspective fondée sur un modèle. Par conséquent, l'utilisation de méthodes d'échantillonnage probabiliste qui permettent de contrôler les caractéristiques de l'échantillon semble être une bonne chose. Dans ce contexte, on peut mentionner que l'échantillonnage aléatoire contraint a joué un rôle important dans l'échantillonnage fondé sur le plan de sondage; voir Tam et Chan (1984) et Deville (1992) pour une discussion de l'échantillonnage par rejet, et Deville et Tillé (2004) pour l'échantillonnage par la méthode du cube, qui permet de sélectionner des échantillons aléatoires qui sont également équilibrés.

Cependant, le rôle approprié de la randomisation sous l'approche fondée sur un modèle n'est pas d'assurer un équilibre exact (ce qu'elle ne peut faire). Elle garantit toutefois le caractère non informatif, ce qui permet de déterminer si le modèle est valide selon les données d'échantillon. Autrement dit, on peut considérer que le rôle approprié de la randomisation n'est pas tant de rendre valide l'inférence fondée sur le plan de sondage, mais en fait de rendre valide l'inférence fondée sur le modèle. En particulier, la randomisation fait en sorte d'avoir un échantillonnage non informatif quand elle est utilisée pour définir la distribution $p\left(\mathbf{I}_U \middle| \mathbf{Z}_U; \phi\right)$. Sa caractéristique de protection découle alors du fait qu'elle établit l'équilibre sur les covariables manquantes dans $p\left(\mathbf{Y}_U \middle| \mathbf{Z}_U; \theta\right)$ « en espérance ».

8. Opérationnaliser l'inférence fondée sur un modèle dans un environnement des EOE

Par définition, l'inférence fondée sur un modèle requiert la spécification d'un modèle, c.-à-d. la définition de $\mathbf{Z}_U$. Dans ce contexte, l'exigence d'un échantillonnage non informatif, c.-à-d. $\mathbf{Y}_U \perp \mathbf{I}_U \middle| \mathbf{Z}_U$, est une déclaration tant au sujet de $\mathbf{Y}_U$ et $\mathbf{Z}_U$ qu'au sujet de $\mathbf{I}_U$. En particulier, elle nécessite que nous définissions $\mathbf{Z}_U$ de manière qu'elle comprenne les covariables qui déterminent $\mathbf{I}_U$. Au minimum, la stratification et la mise en grappes doivent donc être « intégrées » dans $\mathbf{Z}_U$. Les poids de sondage doivent aussi être intégrés dans $\mathbf{Z}_U$ en incluant les covariables qui déterminent ces poids; ou si ces covariables du plan ne sont pas disponibles, les poids eux-mêmes.

L'inconvénient de cette approche est l'obtention d'un modèle de prédiction éventuellement inefficace, contenant un trop grand nombre de prédicteurs non pertinents pour toute variable cible particulière. Des diagnostics du modèle peuvent être utilisés pour décider quels aspects du plan de sondage il faut garder dans le modèle. En outre, il existe aujourd'hui une vaste littérature sur les méthodes de correction de modèles surspécifiés (p. ex., ridging, régularisation) qui visent essentiellement à réduire le « nombre de degrés de liberté » du modèle. Voir Clark et Chambers (2008) pour une exploration de la spécification du modèle dans un contexte de sondage.

En bout de ligne, toutefois, persiste toujours la question de savoir si $\mathbf{Z}_U$ est spécifiée « correctement ». La principale raison de cette question est la non-réponse. Comment le statisticien peut-il être certain que la méthode d'échantillonnage, y compris l'échantillonnage involontaire dû à la non-réponse, est non informative sachant les variables incluses dans $\mathbf{Z}_U$? Cette question est encore plus préoccupante pour les analystes secondaires qui, souvent, n'ont pas accès aux importantes variables du plan de sondage (outre les données sur les strates, les grappes et la pondération). Par conséquent, il appartient aux statisticiens des organismes officiels de statistique de veiller à ce que les données des EOE diffusées pour la modélisation par les membres du public contiennent suffisamment d'information auxiliaire pour qu'elles puissent être analysées comme si elles résultaient d'un processus d'échantillonnage non informatif.

Ce qui nous amène au message à retenir : un bon plan de sondage contrôle ce qui doit être contrôlé (au sens de définir $\mathbf{Z}_U$ et $\mathbf{Z}_s$ de manière à s'assurer d'une prédiction décente de T) et randomise sur ce qui ne peut pas être contrôlé. Il faut espérer que ce qui est contrôlé mais n'a pas besoin de l'être (les composantes de $\mathbf{Z}_U$ dont on peut se passer) a peu d'effet sur l'inférence. Il faut espérer encore davantage que la randomisation fait en sorte que toutes composantes manquantes de $\mathbf{Z}_U$ (celles associées aux non-prises de contact et à la non-réponse, qui sont essentiellement incontrôlable) sont bien équilibrées sur l'échantillon réalisé, et ont donc peu d'effet sur la prédiction de T .

9. Un exemple de raisonnement fondé sur un modèle pour des EOE

La question que nous abordons (brièvement) dans cette dernière section est simplement la suivante : comment adapter les stratégies de suivi utilisées dans une enquête multivague avec prise de contact et interview pour tenir compte de la non-réponse non ignorable possible? Afin de voir ce qui peut être fait dans ce cas, nous supposons que le modèle de travail pour les données d'enquête est le modèle linéaire classique

$$\xi: \quad \mathbf{y}_U = \mathbf{Z}_U\boldsymbol{\beta} + \mathbf{e}_{\xi U}.$$

Ici, $\mathbf{Z}_U$ est une matrice connue des covariables de population, $\mathbf{e}_{\xi U} \sim (\mathbf{0}, \sigma_\xi^2 \mathbf{I}_N)$, et il est bien connu que les poids de sondage optimaux sous ξ sont $\mathbf{w}_{\xi s} = \mathbf{1}_n + \mathbf{Z}_s\left(\mathbf{Z}_s^T\mathbf{Z}_s\right)^{-1}\left(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{1}_n\right)$. L'utilisation de ξ pour l'estimation par sondage est justifiée à condition que l'échantillonnage soit non informatif sachant $\mathbf{Z}_U$, c.-à-d. $\mathbf{e}_{\xi U} \perp \left(\mathbf{I}_U, \mathbf{Z}_U\right)$. Cela est difficile à justifier (du moins en théorie) en présence de non-réponse (non-contacts et/ou refus). Cependant, le suivi des cas de non-réponse donne l'occasion de « cibler » des non-répondants particuliers afin de réduire le biais de non-réponse.

Supposons que m unités de l'échantillon sont observées (c.-à-d. répondent) et que leurs poids fondés sur le modèle de travail sont $\mathbf{w}_{\xi o} = \mathbf{1}_m + \mathbf{Z}_o\left(\mathbf{Z}_o^T\mathbf{Z}_o\right)^{-1}\left(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_o^T\mathbf{1}_m\right)$. Ici, nous utilisons o pour désigner l'échantillon observé. Le problème qui se pose alors est que ces poids ne définissent pas un estimateur sans biais du total de population T , puisque $\mathbf{e}_{\xi U} \perp \left(\mathbf{R}_U, \mathbf{I}_U, \mathbf{Z}_U\right)$ n'est pas vérifiée.

Supposons que ce biais est dû à des covariables manquantes, et qu'il existe une variable X observable seulement sur l'échantillon et possédant les valeurs d'échantillon $\mathbf{X}_s$, telle que le « vrai » modèle pour ces données d'échantillon est

$$\eta: \quad \mathbf{y}_s = \mathbf{Z}_s\boldsymbol{\gamma} + \mathbf{X}_s\boldsymbol{\lambda} + \mathbf{e}_{\eta s}$$

où $\mathbf{e}_{\eta s} \sim (\mathbf{0}, \sigma_\eta^2 \mathbf{I}_n)$, $\sigma_\eta^2 \le \sigma_\xi^2$, et nous avons $\mathbf{e}_{\eta s} \perp \left(\mathbf{R}_S, \mathbf{I}_U, \mathbf{Z}_U\right)$. Puisque

$$E_\eta\left(\mathbf{y}_U \middle| \mathbf{Z}_U\right) = E_\xi\left(\mathbf{y}_U \middle| \mathbf{Z}_U\right) = \mathbf{Z}_U\boldsymbol{\beta}$$

il s'ensuit que

$$E_\eta\left(\mathbf{y}_s \middle| \mathbf{Z}_s\right) = \mathbf{Z}_s\boldsymbol{\gamma} + E_\eta\left(\mathbf{X}_s \middle| \mathbf{Z}_s\right)\boldsymbol{\lambda} = \mathbf{Z}_s\boldsymbol{\beta}$$

donc η peut être réexprimé sous la forme

$$\eta: \quad \mathbf{y}_s = \mathbf{Z}_s\boldsymbol{\gamma} + E_\eta\left(\mathbf{X}_s \middle| \mathbf{Z}_s\right)\boldsymbol{\lambda} + \left\{\mathbf{X}_s - E_\eta\left(\mathbf{X}_s \middle| \mathbf{Z}_s\right)\right\}\boldsymbol{\lambda} + \mathbf{e}_{\eta s} = \mathbf{Z}_s\boldsymbol{\beta} + \tilde{\mathbf{X}}_s\boldsymbol{\lambda} + \mathbf{e}_{\eta s}.$$

Nous pouvons approximer $\tilde{\mathbf{X}}_s$ au moyen de l'orthogonalisation de Gram-Schmidt, c.-à-d. que

$$\tilde{\mathbf{X}}_s = \left[\mathbf{I}_s - \mathbf{Z}_s\left(\mathbf{Z}_s^T\mathbf{Z}_s\right)^{-1}\mathbf{Z}_s^T\right]\mathbf{X}_s$$

d'où il découle que

$$\tilde{\mathbf{X}}_s^T\mathbf{w}_{\xi s} = \tilde{\mathbf{X}}_s^T\mathbf{1}_n + \underbrace{\tilde{\mathbf{X}}_s^T\mathbf{Z}_s}_{=0}\left(\mathbf{Z}_s^T\mathbf{Z}_s\right)^{-1}\left(\mathbf{Z}_U^T\mathbf{1}_N - \mathbf{Z}_s^T\mathbf{1}_n\right) = \tilde{\mathbf{X}}_s^T\mathbf{1}_n.$$

Les poids optimaux pour l'échantillon de répondants sous η sont alors

$$\mathbf{w}_{\eta o} = \mathbf{w}_{\xi o} + \tilde{\mathbf{X}}_o\left(\tilde{\mathbf{X}}_o^T\tilde{\mathbf{X}}_o\right)^{-1}\left(\tilde{\mathbf{X}}_U^T\mathbf{1}_N - \tilde{\mathbf{X}}_o^T\mathbf{1}_m\right).$$

Le remplacement de $\tilde{\mathbf{X}}_U^T\mathbf{1}_N$ par son estimation sur l'échantillon $\tilde{\mathbf{X}}_s^T\mathbf{w}_{\xi s} = \tilde{\mathbf{X}}_s^T\mathbf{1}_n$ donne

$$\mathbf{w}_{\eta o} = \mathbf{w}_{\xi o} + \tilde{\mathbf{X}}_o\left(\tilde{\mathbf{X}}_o^T\tilde{\mathbf{X}}_o\right)^{-1}\tilde{\mathbf{X}}_{s-o}^T\mathbf{1}_{n-m} = \mathbf{w}_{\xi o} + \mathbf{u}_{\eta o}.$$

Nous disons que le sous-échantillon de répondants o est équilibré si $\mathbf{u}_{\eta o} = \mathbf{0}$. Le plan adaptatif correspond alors à la sélection d'un sous-échantillon de non-répondants pour le suivi si o n'est pas suffisamment équilibré (ou s'il n'y a pas assez de répondants).

Le processus de suivi des non-répondants peut être opérationnalisé comme il suit. Soit f un sous-échantillon de suivi possible de taille k. L'objectif est de choisir f afin de minimiser

$$\theta_f = E_\eta\left(\mathbf{w}_{\xi(o+f)}^T\mathbf{y}_{o+f} - \mathbf{w}_{\eta(o+f)}^T\mathbf{y}_{o+f}\right)^2 \propto \mathbf{1}_{n-m-k}^T\tilde{\mathbf{X}}_{s-o-f}\left(\tilde{\mathbf{X}}_o^T\tilde{\mathbf{X}}_o\right)^{-1}\tilde{\mathbf{X}}_{s-o-f}^T\mathbf{1}_{n-m-k}.$$

Posons $f = \{i\}$. Nous avons alors un indice $\theta_{\{i\}}$ qui permet de classer les non-répondants de manière à pouvoir cibler les plus influents. L'extension de cette idée à des populations en grappes et à de multiples vagues de rappels est simple.

Bibliographie

Basu, D. (1971), « An Essay On The Logical Foundations Of Survey Sampling I », dans *Foundations of Statistical Inference*. Toronto: Holt, Rinehart et Winston.

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, p. 93-105.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993), « Bias Robust Estimation In Finite Populations Using Nonparametric Calibration », *Journal of the American Statistical Association*, 88, p. 268-277.

Clark, R.G. et Chambers, R.L. (2008), « Calage adaptatif pour la prédiction de totaux de population finie », *Techniques d'enquête*, 34, p. 181-192.

Deville, J.-C. (1992), « Constrained Samples, Conditional Inference, Weighting: Three Aspects Of The Utilisation Of Auxiliary Information », *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, Statistics Sweden, Örebro, 5 au 7 octobre 1992.

Deville, J.C. et Särndal, C.E. (1992), « Calibration Estimators In Survey Sampling », *Journal of the American Statistical Association*, 87, p. 376-382.

Deville, J.-C. et Tillé, Y. (2004), « Efficient Balanced Sampling: The Cube Method », *Biometrika*, 91, p. 893–912.

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, p. 776-793.

Gelman, A. (2007), « Struggles With Survey Weighting And Regression Modeling », *Statistical Science*, 22, p. 153-164.

Godambe, V.P. (1955), « A Unified Theory Of Sampling From Finite Populations », *Journal of the Royal Statistical Society Series B*, 17, p. 269-278.

Karmel, T.S. et Jain, M. (1987), « Comparison Of Purposive And Random Sampling Schemes For Estimating Capital Expenditure », *Journal of the American Statistical Association*, 82, p. 52-57.

Kendall, M. (1959), « Hiawatha Designs An Experiment », *The American Statistician*, 13, p. 23-24.

Little, R.J.A. (2006), « Calibrated Bayes: A Bayes/Frequentist Roadmap », *American Statistician*, 60, p. 213-223.

Little, R.J.A. (2012), « Calibrated Bayes: An Alternative Inferential Paradigm For Official Statistics », *Journal of Official Statistics*, 28, p. 309-372.

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, p. 558-606.
Pfeffermann, D. (1993), « The Role Of Sampling Weights When Modelling Survey Data », *International Statistical Review*, 61, p. 317-337.
Royall, R.M. (1976a), « Current Advances In Sampling Theory: Implications For Human Observational Studies », *American Journal of Epidemiology*, 104, p. 463-474.
Royall, R.M. (1976b), « The Linear Least Squares Prediction Approach To Two-Stage Sampling », *Journal of the American Statistical Association*, 71, p. 657-664.
Smith, T.M.F. (1983), « On The Validity Of Inferences From Non-Random Samples », *Journal of the Royal Statistical Society Series A*, 146, p. 394-403.
Sugden, R.A. et Smith, T.M.F. (1984), « Ignorable And Informative Designs In Survey Sampling Inference », *Biometrika*, 71, p. 495-506.
Tam, S.M. et Chan, N.N. (1984), « Screening Of Probability Samples », *International Statistical Review*, 52, p. 301-308.
Welsh, A.H. et Wiens, D.P. (2013), « Robust Model-Based Sampling Designs », *Statistics and Computing*, 23, p. 689-701.