Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Pondérations par une méthode bayésienne séquentielle objective dans l'échantillonnage

par Jeremy Strief et Glen Meeden

Juin 2013

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

• Service de renseignements statistiques	1-800-263-1136
• Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
• Télécopieur	1-877-287-4369

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^{s}	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, juin 2013
Vol. 39, N° 1, pp. 1-32
Statistique Canada, N° 12-001-X au Catalogue

Pondérations par une méthode bayésienne séquentielle objective dans l'échantillonnage

Jeremy Strief et Glen Meeden[1]

Résumé

Bien que l'utilisation de pondérations soit très répandue dans l'échantillonnage, leur justification ultime dans la perspective du plan de sondage pose souvent problème. Ici, nous argumentons en faveur d'une justification bayésienne séquentielle des pondérations qui ne dépend pas explicitement du plan de sondage. Cette approche s'appuie sur le type classique d'information présent dans les variables auxiliaires, mais ne suppose pas qu'un modèle relie les variables auxiliaires aux caractéristiques d'intérêt. La pondération résultante d'une unité de l'échantillon peut être interprétée de la manière habituelle comme étant le nombre d'unités de la population que cette unité représente.

Mots clés : Échantillonage; pondérations; inférence bayésienne.

1 Introduction

Les pondérations jouent un rôle important dans l'approche de l'échantillonnage fondée sur un plan. En théorie, la pondération attribuée à une unité observée dans un échantillon est la réciproque de la probabilité de sélection de cette unité et elle est interprétée comme étant le nombre d'unités de la population que représente l'unité en question. En pratique, après avoir observé un échantillon, on ajuste souvent ses pondérations afin de le rendre plus représentatif de la population. Ces ajustements peuvent être faits pour tenir compte de l'information sur la population non incluse dans le plan de sondage, ainsi que des observations manquantes sur l'échantillon. Bien que ces modifications des pondérations fondées sur le plan de sondage soient sans aucun doute utiles dans certains cas, leur justification théorique n'est, en dernière analyse, pas claire. Selon nous, la confusion nait en partie d'une argumentation sur une base non conditionnelle avant le tirage de l'échantillon; ainsi, l'estimateur de Horvitz-Thompson est sans biais si l'on prend la moyenne sur tous les échantillons possibles, et l'est donc conditionnellement après que l'échantillon a été tiré, en ajustant les pondérations fondées sur le

1. Jeremy Strief, Statisticien principal, Medtronic Energy and Component Center, Brooklyn Center, MN 55430. Courriel : jstrief@gmail.com; Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Courriel : glen@stat.umn.edu.

plan de sondage des unités observées dans l'échantillon. En particulier, accorder une trop grande importance au plan de sondage à la deuxième étape, ou étape conditionnelle, peut compliquer inutilement les choses. Une fois que l'échantillon a été observé, nous pensons qu'une meilleure approche consiste à ignorer formellement le plan d'échantillonnage, mais à utiliser toute l'information disponible, y compris celle intégrée dans le plan, pour trouver un ensemble de pondérations raisonnable. Selon ce mode de pensée, une pondération attribuée à une unité peut encore être interprétée comme le nombre d'unités de la population que l'unité représente, mais elle n'est plus dérivée par ajustement de la probabilité de sélection de cette unité. Comment cela peut-il se faire ?

Dans l'approche bayésienne, l'information au sujet de la population est incorporée dans une loi a priori. En théorie, la loi a priori peut ensuite être utilisée pour sélectionner délibérément un échantillon optimal; cependant, cela ne se fait pratiquement jamais. Une fois que l'échantillon est observé, les inférences sont fondées sur la loi a posteriori des unités non observées dans la population, sachant les valeurs des unités observées dans l'échantillon. Dans la plupart des situations, la loi a posteriori ne dépend pas de la façon dont l'échantillon a été sélectionné, de sorte que le plan de sondage ne joue aucun rôle à l'étape de l'inférence. Les méthodes bayésiennes ont été peu utilisées en pratique, parce qu'il est difficile de trouver des lois a priori qui reflètent les types courants d'information a priori disponible.

Une interprétation bayésienne séquentielle (*stepwise Bayesian*) peut être donnée à de nombreux estimateurs classiques (Ghosh et Meeden 1997). Sous cette approche, étant donné un échantillon, l'inférence est encore fondée sur une loi a posteriori, mais l'ensemble (pour tous les échantillons possibles) des lois a posteriori ne provient pas d'une loi a priori unique, mais d'une famille complète de lois a priori. Dans la situation où l'on pense que les unités observées et les unités non observées sont approximativement interchangeables, la loi a posteriori sous l'approche bayésienne séquentielle est la loi a posteriori de Pólya.

Quand on possède l'information a priori sur les moyennes et sur les quantiles de population des variables auxiliaires, Lazar, Meeden et Nelson (2008) soutiennent que la loi a posteriori de Pólya contrainte, qui est une généralisation de la loi a posteriori de Pólya, est un moyen raisonnable d'intégrer ce genre d'information a priori. Ici, nous montrerons comment la loi a posteriori de Pólya contrainte peut être utilisée pour définir des pondérations pour les unités de l'échantillon. Bien que les pondérations résultantes dépendent des variables auxiliaires, elles ne s'appuient pas explicitement sur le plan de sondage.

À la section 2, nous passons en revue la loi a posteriori de Pólya et, à la section 3, la loi a posteriori de Pólya contrainte. Les deux idées principales qui sous-tendent l'article sont énoncées aux deux sections suivantes. À la section 4, nous montrons comment la loi a posteriori de Pólya contrainte peut être utilisée pour attacher une pondération à chaque unité de l'échantillon de manière que ces pondérations ne dépendent pas directement du plan de sondage. À la section 5, nous introduisons la loi a posteriori de Dirichlet pondérée pour accompagner la loi a posteriori de Pólya contrainte. Elle permet d'utiliser les pondérations définies par la loi a posteriori de Pólya contrainte pour faire des inférences au sujet des paramètres de population au moyen d'une simulation simple. À la section 6, nous comparons les pondérations basées sur la loi a posteriori de Pólya contrainte à celles utilisées dans l'estimateur de Horvitz-Thompson. À la section 7, nous considérons plusieurs exemples afin de voir comment les pondérations résultantes se comportent en pratique et de montrer comment la loi a posteriori de Dirichlet pondérée peut être utilisée pour obtenir une estimation de la variance d'un estimateur sans effectuer d'importants calculs. À la section 8, nous présentons certaines conclusions.

À la première lecture, certains auront l'impression que les méthodes proposées ici sont très bayésiennes, parce que toutes nos inférences sont fondées sur les lois « a posteriori ». Mais comme il est mentionné plus haut, techniquement, nos lois « a posteriori » ne sont pas bayésiennes, mais bayésiennes séquentielles. Cela signifie que, du point de vue de la mise en œuvre, on peut concevoir nos lois a posteriori comme étant construites après que l'échantillon a

été observé. Ces « lois a posteriori » construites ne dépendent pas de l'information a priori subjective ni du plan de sondage, mais s'appuient uniquement sur les valeurs d'échantillons observées et sur l'information objective et publique disponible au sujet des variables auxiliaires. Comme nous le verrons, cela permet de construire des estimateurs des paramètres de population qui sont approximativement sans biais sous divers plans de sondage et qui possèdent de bonnes propriétés fréquentistes. Néanmoins, notre méthode présente des limites importantes. Premièrement, elle n'est applicable qu'à des plans de sondage à une étape et deuxièmement, elle ne permet pas de corriger le biais de sélection.

2 La loi a posteriori de Pólya

Soit s l'ensemble d'étiquettes d'un échantillon de taille n tiré d'une population de taille N. Pour simplifier, nous supposons que les membres de s sont $1, 2, \ldots, n$ et nous supposons aussi que la fraction n / N est très petite. Soit $y = (y_1, y_2, \ldots, y_N)$ la caractéristique d'intérêt et y_s les valeurs d'échantillon observées.

La loi a posteriori de Pólya repose sur l'échantillonnage selon le modèle de l'urne de Pólya dont le principe est le suivant : supposons que les valeurs de n unités observées ou vues sont marquées sur n balles et sont placées dans l'urne 1. Les $N - n$ unités restantes non vues de la population sont représentées par $N - n$ balles non marquées placées dans l'urne 2. De chaque urne on tire une balle avec la même probabilité et on attribue à la balle provenant de l'urne 2 la valeur de la balle provenant de l'urne 1. On retourne ensuite les deux balles dans l'urne 1. Donc, à la deuxième étape de l'échantillonnage de Pólya, l'urne 1 contient $n + 1$ balles et l'urne 2 contient $N - n - 1$ balles. On répète cette procédure jusqu'à ce que l'urne 2 soit vide, moment auquel les N balles contenues dans l'urne 1 constituent une copie simulée complète de la population. Toute quantité de population finie – moyenne, total, quantile, coefficient de régression – peut maintenant être calculée à partir de la copie complète. Nous pouvons simuler K de ces copies complètes et, dans chaque cas, calculer la valeur de la quantité de population

d'intérêt. L'estimation ponctuelle est donnée par la moyenne de ces valeurs simulées et un intervalle de crédibilité bayésien à 95 % approximatif est donné par les quantiles correspondant à 2,5 % et 97,5 % des valeurs.

On peut vérifier que, sous la loi a posteriori de Pólya, l'espérance a posteriori de la moyenne de population est simplement la moyenne d'échantillon et la variance a posteriori est simplement égale à $(n-1)/(n+1)$ fois la variance sous le plan habituelle de la moyenne d'échantillon sous échantillonnage aléatoire simple sans remise. La loi a posteriori de Pólya possède une justification fondée sur la théorie de la décision en raison de sa nature bayésienne séquentielle. En s'appuyant sur ce fait, on peut montrer que de nombreux estimateurs classiques sont admissibles. Des renseignements détaillés figurent dans Ghosh et Meeden (1997). La loi a posteriori de Pólya est le bootstrap bayésien de Rubin (1981) appliqué à l'échantillonnage en population finie. Lo (1988) discute aussi du bootstrap bayésien dans le cas de l'échantillonnage en population finie. Certains des premiers travaux dans ce domaine sont décrits dans Hartley et Rao (1968) et dans Binder (1982).

Pour l'unité échantillonnée i soit p_i la proportion d'unités présentes dans une copie simulée complète de la population qui possédent la valeur y_i. Ghosh et Meeden (1997) ont montré que, sous la loi a posteriori de Pólya, $E(p_i) = 1/n$. Si nous posons que

$$w_i = NE(p_i) = N/n,$$

alors w_i peut être interprété comme la pondération appliquée à l'unité i puisqu'elle est égale au nombre moyen d'unités de la population représentée par l'unité i, sous la loi a posteriori de Pólya. Rappelons que sous échantillonnage aléatoire simple sans remise, la fraction n/N est la probabilité d'inclusion de chaque unité. Donc, dans ce cas, il y a concordance entre la pondération fréquentiste habituelle, qui est la réciproque de la probabilité d'inclusion, et la pondération sous la loi a posteriori de Pólya définie plus haut.

Donc, quand l'information a priori est limitée, la loi a posteriori de Pólya donne des pondérations identiques aux pondérations fréquentistes dérivées du plan d'échantillonnage aléatoire simple sans remise. La justification du choix de la loi a posteriori de Pólya pour ces pondérations ne dépend pas explicitement du plan d'échantillonnage et conviendrait dans toute circonstance où l'échantillonneur considère que les unités observées et non observées de la population sont approximativement interchangeables.

Nous allons maintenant examiner la relation entre la loi a posteriori de Pólya et les méthodes bootstrap habituelles sous échantillonnage en population finie. Les deux approches sont fondées sur une hypothèse d'interchangeabilité. Gross (1980) a présenté l'idée fondamentale du bootstrap. Supposons que l'on procède à un échantillonnage aléatoire simple sans remise et que $N / n = m$ est un entier. Étant donné un échantillon, nous créons une bonne approximation de la population en combinant m répliques de cet échantillon. En tirant des échantillons aléatoires répétés de taille n de cette population créée, nous pouvons étudier le comportement d'un estimateur d'intérêt. Booth, Bulter et Hall (1994) ont étudié les propriétés asymptotiques de ce genre d'estimateurs. Hu, Zhang, Cohen et Salvucci (1997) donnent un exemple d'utilisation de l'échantillon pour construire une population artificielle, suivie du tirage d'échantillons répétés de cette population pour construire une estimation de la variance de leur estimateur et des intervalles de confiance.

Notons que cette situation diffère de celle de la loi a posteriori de Pólya où l'échantillon est considéré fixe et des versions complètes de la population sont générées de manière répétée.

3 La loi a posteriori de Pólya contrainte

Commençons par rappeler une approximation bien connue de la loi a posteriori de Pólya. Si la fraction n / N est petite, sous la loi a posteriori de Pólya, $p = (p_1, \ldots, p_n)$ est alors approximativement une loi de Dirichlet caractérisée par un vecteur de paramètres ne contenant que des valeurs un, c'est-à-dire qu'elle est uniforme sur le simplex de dimension $n - 1$, où

$\sum_{j=1}^{n} p_j = 1$. Il est habituellement plus efficace de produire des copies complètes de la population en utilisant cette approximation plutôt que le modèle de l'urne décrit à la section précédente. En outre, cette approximation sera utile quand nous considérerons la loi a posteriori de Pólya contrainte, une généralisation de la loi a posteriori de Pólya obtenue lorsque l'échantillonneur dispose d'information a priori au sujet des variables auxiliaires.

Dans de nombreux problèmes, en plus de la variable d'intérêt, y, l'échantillonneur possède des variables auxiliaires pour lesquelles de l'information a priori est disponible. Un cas très fréquent est celui où la moyenne de population d'une variable auxiliaire est connue. De manière plus générale, nous supposerons que l'information a priori au sujet de la population peut être exprimée par un ensemble de contraintes linéaires d'égalité et d'inégalité sur une série de variables auxiliaires.

Nous supposons qu'en plus de la caractéristique d'intérêt y, il existe un ensemble de variables auxiliaires $x^1, x^2, \ldots, x^m$. Pour l'unité i, soit

$$(y_i, x_i) = (y_i, x_i^1, x_i^2, \ldots, x_i^m)$$

le vecteur des valeurs de y et des variable auxiliaires. Nous supposons que ce vecteur des valeurs est observé pour toute unité présente dans l'échantillon. Nous supposons aussi que l'information a priori au sujet de la population peut être exprimée au moyen d'un ensemble de contraintes linéaires d'égalité et d'inégalité sur les valeurs de population des variables auxiliaires. Pour l'ensemble de valeurs possibles pour une variable auxiliaire donnée, les coefficients qui définissent une contrainte correspondront aux proportions d'unités dans la population qui prennent ces valeurs. Nous allons maintenant illustrer ceci de manière plus précise en expliquant comment nous transposons cette information a priori au sujet de la population aux valeurs observées dans l'échantillon. Étant donné un échantillon, cela nous permettra de construire des copies simulées de la population compatibles avec l'information a priori.

Étant donné un échantillon s, pour $i = 1, 2, \ldots, n$, soit (y_i, x_i) les valeurs observées que, pour simplifier, nous supposons être distinctes. Soit p_i la proportion d'unités auxquelles est attribuée la valeur (y_i, x_i) dans une copie complète simulée de la population. Toute contrainte linéaire sur la valeur de population d'une variable auxiliaire se traduit de manière évidente par une contrainte linéaire sur ses valeurs observées. Par exemple, si l'on sait que la moyenne de population de x^1 est inférieure ou égale à une certaine valeur, disons b_1, alors pour la population simulée, cette information se traduit en la contrainte

$$\sum_{i=1}^{n} p_i x_i^1 \leq b_1.$$

Si l'on sait que la médiane de population de x^2 est égale à b_2, la contrainte pour la population simulée devient

$$\sum_{i=1}^{n} p_i u_i = 0,5$$

où $u_i = 1$ si $x_i^2 \leq b_2$ et est égal à zéro autrement. Donc, étant donné un ensemble de contraintes de population fondées sur l'information a priori et un échantillon, nous pourrons représenter les contraintes correspondantes sur une valeur simulée de p par deux systèmes d'équations

$$A_{1,s} p = b_1 \tag{3.1}$$

$$A_{2,s} p \leq b_2 \tag{3.2}$$

où $A_{1,s}$ et $A_{2,s}$ sont les matrices de dimensions $m_1 \times n$ et $m_2 \times n$, respectivement, et b_1 et b_2 sont des vecteurs de dimension appropriée.

Soit P le sous-ensemble du simplex de dimension n défini par les équations (3.1) et (3.2). Nous supposons que l'échantillon est tel que P est non vide et donc qu'il s'agit d'un polytope de dimension non pleine. Dans ce cas, la version approximative appropriée de la loi a posteriori de Pólya doit être simplement la loi uniforme sur P. Nous donnons à cette distribution le nom de loi a posteriori de Pólya contrainte (LPPC). S'il était possible de générer des observations indépendantes à partir de la LPPC, on pourrait trouver approximativement l'espérance

a posteriori des paramètres de population d'intérêt et obtenir des intervalles de crédibilité bayésiens séquentiels à 95 % approximatifs. Malheureusement, nous ne savons pas comment le faire. À la place, on peut utiliser des méthodes de Monte Carlo par chaîne de Markov (MCMC) pour trouver approximativement ces estimations. On peut pour cela travailler en R (R Development Core Team 2005) en utilisant le module externe *polypost* qui est disponible dans le CRAN. Des renseignements plus détaillés sur la LPPC et les simulations à partir de cette loi figurent dans Lazar et coll. (2008).

4 Pondérations basées sur la loi a posteriori de Pólya contrainte

Une critique de la loi a posteriori de Pólya et de la loi a posteriori de Pólya contrainte pourrait être que toute copie complète simulée de la population ne contient que les valeurs des caractéristiques qui figurent dans l'échantillon. Cependant, c'est exactement cette propriété qui va nous permettre d'attribuer des pondérations aux membres de l'échantillon.

Nous supposons que nous avons un échantillon fixe pour lequel le sous-ensemble du simplex défini par les équations (3.1) et (3.2) est non vide. Pour $j = 1, \ldots, n$ soit

$$w_j = NE(p_j) = N\mu_j \tag{4.1}$$

où l'espérance est prise par rapport à la LPPC. Notons que la somme des éléments de $w = (w_1, \ldots, w_n)$ est égale à la taille de la population N et que w_j peut être considéré comme la pondération associée au j^{e} membre de l'échantillon. Ces pondérations ne dépendent que des valeurs observées des variables auxiliaires et des contraintes de population connues. Par conséquent, il s'agit d'une méthode bayésienne séquentielle d'attribution des pondérations aux unités de l'échantillon dans laquelle est intégrée l'information a priori présente dans les variables auxiliaires et qui ne dépend pas explicitement du plan de sondage.

Nous supposons ici que la taille de la population N est connue, ce qui n'est pas toujours vrai. Le cas échéant, on pourrait remplacer N par une estimation dans l'équation susmentionnée. Si

l'estimation est bonne, les inférences résultantes pour un total de population devraient être satisfaisantes. Dans le cas de l'estimation d'une moyenne de population, les résultats seraient nettement moins sensibles à la mesure dans laquelle l'estimation est proche de la taille réelle de la population.

De nombreuses données d'enquête utilisées par les chercheurs du domaine de la science sociale sont fournies avec des pondérations appliquées aux unités individuelles. Dans ces situations, les pondérations basées sur la LPPC pourraient être reliées aux unités de la même façon et l'utilisateur n'aurait pas besoin de faire appel à des méthodes MCMC pour calculer les pondérations. Nous allons nous servir des pondérations pour définir la loi a posteriori de Dirichlet pondérée qui peut être utilisée pour trouver les estimations ponctuelles et les estimations des intervalles pour les quantités de population d'intérêt moyennant des calculs relativement modestes. Dans la suite de l'article, nous illustrerons à l'aide d'exemples comment ces poids peuvent être utilisés pour générer des procédures d'inférence ayant de bonnes propriétés fréquentistes.

Mais avant de poursuivre, faisons une simple observation. Supposons que nous disposions de l'échantillon ainsi que d'un ensemble de pondérations. Si N est grand, nous pouvons construire une population dont la proportion d'unités de type (y_i, x_i) est w_i / N pour $i = 1, \ldots, n$. Étant donné l'échantillon et l'ensemble de pondérations, nous pouvons considérer cette population construite comme étant la meilleure approximation de la population inconnue. Alors

$$\overline{y}_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} y_i \text{ et } \sigma^2_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} (y_i - \overline{y}_{bw})^2 \tag{4.2}$$

sont la moyenne et la variance de cette population construite.

5 La loi a posteriori de Dirichlet pondérée

Il arrive souvent que des pondérations soient attachées à des données dans les fichiers de données à grande diffusion. Ces pondérations sont alors utilisées par les chercheurs pour produire

des estimations ponctuelles et des estimations d'intervalles pour les paramètres de population. Nous allons voir que les pondérations fondées sur la méthode bayésienne séquentielle présentée ici peuvent souvent être utilisées dans les formules fréquentistes classiques pour estimer les paramètres d'intérêt tout comme les pondérations habituelles. Nous utiliserons nos pondérations pour définir la loi a posteriori de Dirichlet pondérée et montrer qu'elle offre un autre moyen de calculer les estimations ponctuelles et les estimations d'intervalles pour diverses quantités de population.

Soit les w_j un ensemble de pondérations défini par l'équation (4.1) avec $\mu_j = w_j / N$. Considérons la distribution de Dirichlet sur le simplex défini par le vecteur $n\mu = (n\mu_1, \ldots, n\mu_n)$ comme une loi a posteriori de rechange pour $p = (p_1, \ldots, p_n)$ en utilisant l'échantillon observé pour produire des copies simulées complètes de la population. Nous donnerons à cette loi a posteriori le nom de loi a posteriori de Dirichlet pondérée (LPDP). Soulignons que la LPDP est une version moins contrainte de la LPPC. Sous la LPPC, chaque copie complète de la population satisfait les contraintes; par contre, sous la LPDP, seule la moyenne des populations simulées satisfait les contraintes. Il est facile de voir que sous la LPDP

$$E\left(\sum_{i=1}^{n} p_i y_i\right) = \sum_{i=1}^{n} \mu_i y_i = \bar{y}_{bw} \tag{5.1}$$

et

$$\begin{aligned} V\left(\sum_{i=1}^{n} p_i y_i\right) &= \sum_{i=1}^{n} y_i^2 V(p_i) + \sum\sum_{i<j} y_i y_j \mathrm{Cov}(p_i, p_j) \\ &= \sum_{i=1}^{n} \frac{n\mu_i(n - n\mu_i)y_i^2}{n^2(n+1)} - 2\sum\sum_{i<j} \frac{n\mu_i n\mu_j y_i y_j}{n^2(n+1)} \\ &= \frac{1}{n+1}\left(\sum_{i=1}^{n} \mu_i(1-\mu_i)y_i^2 + 2\sum\sum_{i<j} \mu_i n\mu_j y_i y_j\right) \\ &= \frac{1}{n+1}\left(\sum_{i=1}^{n} \mu_i y_i^2 - \sum_{i=1}^{n}\sum_{i=1}^{n} \mu_i \mu_j y_i y_j\right) = \frac{1}{n+1}\sigma_{bw}^2 \end{aligned} \tag{5.2}$$

où $\bar{y}_{bw}$ et σ_{bw}^2 sont définies par l'équation (4.2).

Partant de cela, nous voyons que, lorsqu'on estime la moyenne de population, la simulation à partir de la LPDP équivaut à utiliser l'échantillon et ses pondérations pour construire la meilleure approximation possible de la population. En particulier, quand les pondérations sont toutes égales, la LPDP est simplement la loi a posteriori de Pólya.

Deux grandes raisons nous poussent à introduire la LPDP. Premièrement, à mesure que le nombre de contraintes utilisées augmente, les intervalles de crédibilité à 95 % approximatifs fondés sur la LPPC deviennent trop courts et contiennent la valeur réelle du paramètre dans moins de 95 % du temps. Il en est ainsi parce que, quand le nombre de contraintes est grand, la LPPC ne permet pas d'obtenir une variabilité suffisante dans les copies complètes simulées de la population qu'elle produit. Deuxièmement, il est nettement plus facile d'exécuter la simulation à partir de la LPDP qu'à partir de la LPPC. Maintenant, il serait possible d'effectuer la simulation à partir de la LPDP contrainte de manière que toutes les contraintes soient satisfaites, mais cela demanderait autant d'effort que d'effectuer la simulation à partir de la LPPC. En outre, nous pensons que cela donnerait des intervalles de crédibilité à 95 % approximatifs ayant de mauvaises propriétés de couverture fréquentistes parce qu'ils seraient trop courts.

Supposons maintenant que notre ensemble de pondérations est constitué des réciproques des probabilités d'inclusion provenant du plan de sondage. Soit $W = \sum_{i=1}^{n} w_i$. Pour la plupart des échantillons, cette valeur n'est pas égale à N, mais s'en approche souvent. De nouveau, nous pouvons construire notre meilleure approximation de la population en nous basant sur les pondérations. La moyenne et la variance de cette population seront données par

$$\bar{y}_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} y_i \text{ et } \sigma^2_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} (y_i - \bar{y}_{dw})^2. \tag{5.3}$$

Si nous utilisons $\bar{y}_{dw}$ comme estimation de la moyenne de population inconnue, une estimation sans biais de cette variance dépend des probabilités d'inclusion conjointes des unités dans l'échantillon. Comme il est souvent difficile d'obtenir ces probabilités, il a été recommandé en pratique (Särndal, Swensson et Wretman 1992) de supposer que l'échantillonnage a été fait avec

remise, même si ce n'est pas le cas. Alors, l'estimation approximative résultante de la variance de $\overline{y}_{dw}$ est

$$\begin{aligned} \hat{V}_d(\overline{y}_{dw}) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(n \frac{w_i}{W} y_i - \overline{y}_{dw} \right)^2 \\ &= \frac{\sigma_{dw}^2 + \gamma_{dw}}{n-1} \end{aligned} \tag{5.4}$$

où la deuxième ligne découle de simples opérations algébriques et où

$$\gamma_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} y_i^2 \left(n \frac{w_i}{W} - 1 \right). \tag{5.5}$$

Notons que, sous échantillonnage aléatoire simple avec ou sans remise et $N = nk$, $\gamma_{dw} = 0$. Dans ce cas, l'estimation de la variance donnée par l'équation (5.4) est essentiellement équivalente à celle donnée par l'équation (5.2).

Dans les situations où il est sensé d'utiliser l'estimateur de Horvitz-Thompson, les calculs ont montré que γ_{dw} a tendance à être négatif, ce qui laisse entendre que les intervalles fondés sur la LPDP ont alors tendance à être prudents. Cependant, les calculs montrent aussi que le terme γ_{dw} a tendance à être positif dans les situations où l'estimateur de Horvitz-Thompson n'est pas approprié. Nous verrons que, dans de tels cas, l'approximation habituelle peut donner de médiocres résultats et que les intervalles fondés sur la LPDP peuvent posséder de meilleures propriétés fréquentistes.

6 Pondérations et estimateurs de Horvitz-Thompson

La pondération attribuée à une unité de l'échantillon est habituellement définie comme étant l'inverse de la probabilité d'inclusion de cette unité. On est donc invité à se représenter la pondération de l'unité comme étant le nombre d'unités de la population qu'elle représente. L'estimateur résultant du total de population est l'estimateur de Horvitz-Thompson (HT) qui est sans biais. Comme nous l'avons déjà mentionné, l'estimation sans biais de sa variance dépend

des probabilités de sélection conjointes de toutes les paires d'unités figurant dans l'échantillon. Comme, en pratique, il peut être impossible de calculer ces probabilités, on utilise souvent l'approximation de l'équation (5.4).

L'estimateur HT donne de bons résultats quand y_i est approximativement proportionnel à sa probabilité de sélection. Pour comparer le comportement de cet estimateur à la méthode basée sur la LPDP, nous avons réalisé une petite expérience par simulation. Nous avons construit la variable x en tirant un échantillon aléatoire de 2 000 unités d'une loi de probabilité gamma de paramètre de forme égal à 5 et de paramètre d'échelle égal à 1, et en ajoutant 20 à chaque valeur. Pour générer y, nous avons postulé que la loi conditionnelle de y_i sachant x_i était une loi normale de moyenne $5x_i$ et d'écart-type 20. La corrélation de la population résultante était de 0,49. Nous avons désigné cette population par A. Nous avons créé une deuxième population, B, en utilisant le même vecteur de valeurs de x, mais en ajoutant 400 à chaque valeur de y_i. Dans notre plan d'échantillonnage, nous nous sommes servis de x pour effectuer un échantillonnage proportionnel à la taille, c'est-à-dire ppt(x). Nous avons utilisé le module externe *sampling* de R, de sorte que les probabilités d'inclusion étaient exactes. Sous ce plan, nous nous attendons à ce que l'estimateur HT donne de bons résultats pour la population A mais qu'il ait de moins bonnes propriétés pour la population B. Nous avons également considéré un troisième estimateur, NHT, obtenu simplement en rééchelonnant les pondérations de l'estimateur HT de manière que leur somme soit égale à la taille N de la population. Nous avons produit 500 échantillons de taille 50. Les résultats sont présentés au tableau 6.1.

Tableau 6.1
Résultats pour les populations A et B basés sur 500 échantillons de taille 50. L'estimateur NHT correspond à l'estimateur HT normalisé de manière que la somme des pondérations soit égale à la taille de la population, $N = 2\,000$. La couverture nominale est égale à 0,95 pour chaque population.

Population	Méthode	Erreur absolue moyenne	Longueur moyenne	Fréquence de couverture
A	HT	4 628	21 898	0,940
B	HT	8 965	43 914	0,960
A et B	LPDP	4 706	24 381	0,960
A	NHT	5 051	21 897	0,896
B	NHT	5 051	43 919	0,998

Bien que cela ne soit pas montré dans le tableau, les estimateurs HT et LPDP sont tous deux sans biais pour les deux populations. Comme prévu, l'estimateur HT est le meilleur pour la population A, mais sa performance baisse considérablement pour la population B. Par ailleurs, les propriétés de l'estimateur LPDP sont exactement les mêmes pour les deux populations. En tant qu'estimateur ponctuel, l'estimateur NHT donne de nettement meilleurs résultats que l'estimateur HT pour la population B, et des résultats moins bons pour la population A. Dans l'ensemble, l'estimateur LPDP est clairement celui qui donne les meilleurs résultats. Comment ces différences s'expliquent-elles ?

Dans la population A, $y_i \propto x_i$ et les calculs montrent que γ_{dw} est presque toujours négatif et que sa valeur absolue est petite comparativement à σ_{dw}. En d'autres termes, quand l'estimateur HT est approprié, c'est la variance de la population construite en se basant sur les pondérations de cet estimateur qui est essentiellement utilisée pour obtenir l'estimation de sa variance.

La seule différence entre les populations A et B est qu'une constante a été ajoutée à la valeur de y de chaque unité. Alors, si les pondérations de l'échantillon nous permettent d'obtenir une bonne approximation de la population dans le premier cas, quel est le problème qui, dans le deuxième cas, fait que l'estimateur HT donne de si mauvais résultats ? Pour le voir, considérons ce qui suit.

Dans l'estimation HT, la somme des pondérations dans l'échantillon n'est presque jamais égale à N, la taille de la population. Étant donné un échantillon de la population B, l'estimation HT est donnée par

$$\sum_{i=1}^{50} w_i y_i = \sum_{i=1}^{50} w_i y_i' + 400 \sum_{i=1}^{50} w_i$$

où y_i' désigne la valeur de l'unité dans la population A et y_i, sa valeur dans la population B. Notons que le deuxième terme de l'équation susmentionnée ajoute une variabilité supplémentaire à l'estimateur HT. Dans la population B, les calculs montrent que le terme γ_{dw} de

l'équation (5.5) est positif et peut être assez grand. Il explique la variabilité excédentaire de l'estimateur HT dans la population B découlant du fait qu'ici, $y_i \propto x_i + 400$ et non x_i.

Notons que Zheng et Little (2003) ont soutenu que, lorsqu'on estime un total de population finie et qu'on utilise un plan d'échantillonnage avec probabilités proportionnelles à la taille, un estimateur fondé sur un modèle non paramétrique à splines pénalisées donne généralement de meilleurs résultats que l'estimateur de Horvitz-Thompson. Zheng et Little (2005) ont élaboré des méthodes pour estimer la variance de leur estimateur. Certains travaux apparentés sont décrits dans Zheng et Little (2004).

Les pondérations basées sur la LPDP ne comprennent que la contrainte voulant que les copies complètes simulées de la population aient la moyenne de population correcte pour x. Cette hypothèse est plus robuste que celle qui sous-tend l'estimateur HT. En toute honnêteté, il faut se rappeler (comme l'a fait remarquer un examinateur) que l'estimateur HT a été élaboré avec l'objectif limité d'obtenir des estimateurs linéaires sans biais du total de population. Toutefois, sa simplicité ne semble plus aussi importante maintenant que des estimateurs plus compliqués et plus efficaces sont devenus plus faciles à calculer. La performance supérieure de la méthode bayésienne séquentielle laisse entendre ici que si l'on pense posséder pour les unités échantillonnées un ensemble de pondérations dont la somme est égale à la taille de la population et qui donne une bonne approximation de la population, il faut utiliser la variance de cette bonne approximation de la population pour construire une estimation de la variance de l'estimation de la moyenne de population au lieu d'utiliser l'équation (5.4). Cela vaut particulièrement pour les grandes enquêtes portant sur plusieurs caractéristiques y d'intérêt. Il serait très surprenant que toutes satisfassent les hypothèses nécessaires pour que l'équation (5.4) soit une bonne estimation de la variance d'une moyenne d'échantillon. Suivant l'observation faite dans Royall et Cumberland (1981) et Royall et Cumberland (1985) voulant que de bons échantillons équilibrés (la moyenne d'échantillon est proche de la moyenne de population) peuvent améliorer la

performance, on devrait fonder les inférences sur les copies complètes simulées de la population auxquelles est intégrée l'information a priori disponible contenue dans les variables auxiliaires.

7 Exemples

Nous pensons que la théorie fondée sur les plans de sondage classiques accorde trop d'importance au rôle que les probabilités de sélection doivent jouer dans les inférences faites après que l'échantillon a été observé. À la présente section, nous présentons des exemples qui montrent comment la LPDP permet d'utiliser de l'information a priori objective après que l'échantillon a été sélectionné.

7.1 Une étude de simulation

Afin de mieux comprendre comment l'utilisation des pondérations obtenues par la méthode bayésienne séquentielle dans la LPDP peut fonctionner, nous avons procédé à une étude en simulation. Nous avons construit une population de 2 000 unités et une variable auxiliaire unique, x. Cette variable consistait en un échantillon aléatoire tiré d'une loi de probabilité gamma dont le paramètre de forme était égal à 5 et le paramètre d'échelle, égal à 1. La loi conditionnelle de y_i sachant x_i était une loi normale de moyenne $100 + (x_i - 8)^2$ et d'écart-type égal à 20. La corrélation pour la population résultante était de -0,38. Nous désignons cette population par quad. Il s'agit clairement d'un petit exemple simple et la forme particulière de la relation entre x et y n'a pas d'importance en ce qui concerne les méthodes basées sur la LPDP, outre le fait que x contient une certaine information au sujet de y. Nous allons maintenant comparer les estimateurs basés sur la LPDP à deux méthodes classiques sous quatre plans de sondage différents.

Pour construire la LPPC, nous avons supposé que les valeurs de x pour la population étaient connues et nous les avons utilisées pour construire trois strates après avoir observé l'échantillon. Ces strates n'ont pas été construites de la manière habituelle. Nous avons procédé ainsi afin de minimiser le rôle habituel du plan de sondage et de mettre l'accent sur la robustesse de notre

approche au choix du plan de sondage. Nous avons tiré un échantillon de taille $n = 60$ et nous avons construit trois strates a posteriori. Soit $x_{[1]} < x_{[2]} < \cdots < x_{[60]}$ la statistique d'ordre des valeurs de x dans l'échantillon. Soit q_{20} et q_{40} les quantiles de population de $x_{[20]}$ et $x_{[40]}$ respectivement. Alors, la LPPC repose sur l'hypothèse que la probabilité totale attribuée aux unités de l'échantillon possédant les 20 plus petites valeurs de x doit être q_{20} et que la probabilité totale attribuée aux unités possédant les 20 plus petites valeurs suivantes doit être $q_{40} - q_{20}$. Autrement dit, nous divisons l'échantillon en trois groupes égaux et utilisons l'information contenue dans les valeurs de x pour obtenir la taille de la population appropriée des strates correspondantes. En outre, la LPPC repose sur l'hypothèse que les probabilités attribuées aux unités de l'échantillon doivent satisfaire la contrainte de la moyenne de population pour x.

La LPDP résultante sera comparée à deux méthodes fréquentistes classiques. La première est l'estimateur poststratifié qui utilise la même information sur les strates que la LPPC. La deuxième est l'estimateur par la régression habituel qui repose sur l'hypothèse que la moyenne de population de x est connue. Bien que l'estimateur par la régression ne soit pas réellement approprié pour la population quad, il a été inclus à titre de comparaison. Les intervalles de confiance à 95 % pour le total de population pour les deux méthodes fréquentistes ont été calculés en supposant un échantillonnage aléatoire simple même si des plans de sondage différents avaient été utilisés. Nous désignons ces deux estimateurs par STR et REG respectivement.

Le premier plan utilisé était un plan d'échantillonnage aléatoire simple sans remise. Pour le deuxième, nous avons produit un ensemble de poids d'échantillonnage en tirant un échantillon aléatoire de 2 000 unités d'une loi gamma dont le paramètre de forme était égal à 5 et le paramètre d'échelle, égal à 1. Puis nous avons ajouté 5 à chaque valeur pour obtenir le vecteur v, disons. Il faut noter que les valeurs de v et de y sont complètement indépendantes. Nous avons ensuite utilisé un plan $\text{ppt}(v)$ approximatif où, à chaque étape, la probabilité qu'une unité soit

sélectionnée est proportionnelle à sa valeur de v et dépend uniquement des unités non sélectionnées qui demeurent dans la population. Nous donnons à ce plan le nom de plan à pondération aléatoire. Pour le troisième plan, nous avons utilisé le plan $\text{ppt}(x)$ approximatif. Pour le quatrième, nous avons trouvé la fonction linéaire, disons l, qui applique le domaine de valeur de y sur l'intervalle $[1, 2]$. Nous avons ensuite utilisé le plan $\text{ppt}(l(y))$ approximatif comme plan de sondage. Nous donnons à ce plan le nom de plan dépendant de y. Dans ce plan, les probabilités de sélection dépendent faiblement des valeurs de y et les unités dont la valeur de y est grande sont plus susceptibles d'être sélectionnées que celles dont la valeur est faible. En particulier, l'unité possédant la valeur de y la plus grande est deux fois plus susceptible d'être sélectionnée que celle possédant la valeur de y la plus petite. De toute évidence, le plan à pondération aléatoire et le plan dépendant de y ne sont pas des plans classiques et ne seraient jamais utilisés en pratique. Nous les avons inclus afin de mettre en relief notre conviction que, dans de nombreux cas, étant donné un échantillon, une bonne estimation ne dépend pas de la façon dont l'échantillon a été sélectionné.

Pour chaque plan, nous avons tiré 500 échantillons de taille 60 et calculé l'estimation ponctuelle, son erreur absolue et la longueur de son intervalle estimé, et nous avons déterminé si celui-ci contenait ou non la valeur réelle du paramètre. Les résultats sont présentés au tableau 7.1.

Rappelons que, dans cet exemple, la LPDP utilise l'information provenant à la fois de la poststratification et du fait que la moyenne de population de x est connue, tandis que l'estimateur STR utilise seulement la première information et l'estimateur REG, seulement la seconde. Sous les plans EAS et à pondération aléatoire, les quatre méthodes donnent à peu près les mêmes résultats. Pour les deux autres plans, la méthode LPDP est celle qui est la meilleure. Sous les quatre plans de sondage, sa fréquence de couverture est celle qui est la plus proche du niveau nominal de 0,95. L'utilisation de la contrainte faisant intervenir la moyenne de population de x permet la correction d'une partie du biais introduit par les plans de sondage, ce que la méthode STR ne peut pas faire. Cependant, cette contrainte a ses limites. Si, sous le plan

dépendant de y, le domaine de valeur de l est $[1, 4]$, l'erreur absolue moyenne de l'estimateur LPDP est 4,5 % meilleure que celle de l'estimateur STR, et la fréquence de couverture sur les intervalles nominaux à 95 % est de 0,86 et 0,80, respectivement. La variable x ne contient tout simplement pas suffisamment d'information pour corriger un biais de sélection de cette importance.

Tableau 7.1
Résultats de la simulation pour la population quad décrite à la section 7.1 pour 500 échantillons aléatoires de taille 60 sous quatre plans de sondage différents. Le total de population réel était égal à 227 923,0. Pour chaque méthode, la couverture nominale était de 0,95.

Méthode	Valeur moyenne	Erreur moyenne	Longueur moyenne	Fréquence de couverture
		EAS		
STR	227 856,1	4 165,0	21 332,1	0,950
REG	227 602,1	4 302,7	21 300,3	0,944
LPDP	227 546,9	4 190,6	23 029,7	0,958
	Les min. et max. moyens des paramètres de la LPDP étaient 0,658 et 1,580.			
		Pondération aléatoire		
STR	227 976,5	4 371,2	21 254,1	0,938
REG	227 715,5	4 462,2	21 305,9	0,934
LPDP	227 721,2	4 420,6	22 901,4	0,950
	Les min. et max. moyens des paramètres de la LPDP étaient 0,651 et 1,583.			
		ppt(x)		
STR	225 295,8	5 228,9	23 008,4	0,916
REG	224 207,2	5 611,2	21 780,3	0,878
LPDP	227 471,1	4 919,2	22 706,6	0,936
	Les min. et max. moyens des paramètres de la LPDP étaient 0,374 et 3,024.			
		Dépendant de y		
STR	231 590,0	5 229,0	21 170,8	0,892
REG	231 424,4	5 143,4	21 127,9	0,902
LPDP	231 139,1	4 967,6	22 867,0	0,938
	Les min. et max. moyens des paramètres de la LPDP étaient 0,660 et 1,643.			

Pour chaque plan de sondage, nous avons inclus la moyenne des valeurs les plus faibles, d'une part, et les plus grandes, d'autre part, des paramètres définissant la LPDP, dont, ici, la somme doit être égale à 60. Nous voyons que c'est pour le plan ppt(x) que l'intervalle est le plus grand.

Nous avons également utilisé la LPDP dans les simulations pour construire des intervalles de crédibilité à 95 % pour la médiane de population de y. Pour les quatre plans de sondage, les fréquences de couverture respectives étaient de 0,956, 0,950, 0,952 et 0,930.

Nous avons effectué une autre étude en simulation où la variable x a été produite de la même façon, mais la loi conditionnelle de y_i sachant x_i était une loi normale de moyenne $60 + x_i$ et d'écart-type $2\sqrt{x_i}$. La corrélation entre x et y était de 0,46. Sous les quatre plans de sondage, les propriétés des estimateurs ponctuels étaient très similaires. Les intervalles avaient tendance à être un peu plus longs pour l'estimateur LPDP que pour les autres, mais, sur les quatre plans de sondage, sa fréquence moyenne de couverture pour le total de population était de 0,949. Sous le plan dépendant de y, sa fréquence de couverture pour le total de population était de 0,934, tandis que pour les estimateurs STR et REG, les couvertures correspondantes étaient de 0,896 et 0,886. Sa fréquence de couverture moyenne pour la médiane de population de y était de 0,942.

Un fréquentiste pourrait soutenir qu'il s'agit d'un exemple injuste, puisque l'estimateur par la régression n'a pas beaucoup de sens pour la population en question et, naturellement, il aurait raison. Si, pour ce problème, on supposait qu'il existe une relation quadratique entre y et x et que les deux premiers moments de population de x étaient connus, alors l'estimateur par la régression résultant donnerait de meilleurs résultats que l'estimateur LPDP. Lazar et coll. (2008) décrivent un exemple de ce genre. En outre, ils montrent qu'inclure une contrainte pour le deuxième moment de la LPPC ne modifie pour ainsi dire pas le comportement des estimations résultantes. Donc, lorsque l'on possède une bonne information a priori au sujet du modèle reliant x et y, elle pourrait être utilisée dans l'analyse. Si ce genre d'information a priori n'est pas disponible, nous pensons que l'estimateur LPDP possède un certain avantage, même s'il ne donne pas nécessairement lieu à une amélioration spectaculaire par rapport aux méthodes classiques. Il s'appuie uniquement sur l'information a priori objective et ne comprend aucune hypothèse de modélisation au sujet des liens entre les caractéristiques d'intérêt et les variables. Il peut faire une correction pour une légère dépendance des probabilités de sélection à l'égard de la caractéristique d'intérêt. Même si le plan de sondage ne joue aucun rôle explicite dans le calcul de cet estimateur, l'information qui y est souvent intégrée peut être reformulée sous forme d'une contrainte et utilisée pour définir la LPPC. Étant donné un échantillon, les inférences fondées sur

la LPDP s'appuient sur de nombreuses copies complètes simulées de la population qui, en moyenne, sont compatibles avec l'information a priori. Par conséquent, il est facile d'estimer d'autres paramètres qu'une moyenne ou un total de population.

7.2 Stratification et estimation de la médiane

Dans de nombreuses applications quelques observations seulement, parfois juste deux, sont tirées de chaque strate. Dans de telles conditions, trouver un bon intervalle de confiance lorsqu'on estime la médiane de population peut être difficile. Nous allons comparer la méthode classique, voir par exemple la section 5.11 de Särndal et coll. (1992), à celle fondée sur la LPDP. Nous émettons l'hypothèse d'un échantillonnage aléatoire simple sans remise dans les strates.

Par souci de définition précise, supposons que nous avons L strates et que la strate j contient N_j unités. Soit $N = \sum_{j=1}^{L} N_j$ la taille totale de la population. Supposons que deux observations sont tirées de chaque strate. Alors, la pondération attribuée à chaque unité échantillonnée est égale à la moitié de la taille de la strate dans laquelle elle a été sélectionnée. Dans la méthode classique, on utilise ces pondérations pour trouver l'intervalle de confiance.

Pour ce scénario, nous appliquons la loi a posteriori de Pólya habituelle dans chaque strate, de manière indépendante d'une strate à l'autre. Ou bien, on peut envisager d'appliquer une loi a posteriori de Pólya contrainte (LPPC) où la quantité de probabilité attribuée aux deux unités échantillonnées dans la strate j doit être égale à N_j / N. Si $p_j = (p_{j,1}, p_{j,2})$ représente la probabilité attribuée aux deux unités échantillonnées dans la strate j, alors sous la LPPC, $E(p_j) = (N_j / (2N), N_j / (2N))$. En reprenant la notation de la section 5, nous voyons que sous la LPDP, la pondération attribuée à chacune des deux unités échantillonnées dans la strate j est $(LN_j) / N$. Rappelons que simuler des copies complètes de la population en utilisant la LPDP signifie qu'il est presque certain que les copies simulées individuelles ne satisferont pas les contraintes, mais que ces dernières seront satisfaites si l'on prend la moyenne sur l'ensemble des copies simulées. À première vue, cela pourrait sembler être une mauvaise idée, mais nous verrons

que, lorsqu'on estime l'intervalle de la médiane de population, les estimations fondées sur la LPDP se comportent mieux que les intervalles classiques, qui sont trop courts. Nous verrons que la variabilité supplémentaire présente dans la LPDP produit des intervalles plus longs ayant de meilleures propriétés fréquentistes.

Les populations stratifiées que nous considérons ont été construites comme il suit. Les tailles de strate correspondaient à un échantillon aléatoire tiré d'une loi de Poisson de paramètre $\lambda = 100$. Les moyennes de strate correspondaient à un échantillon aléatoire tiré d'une loi normale de moyenne $\mu = 150$ et d'écarts-types $\sigma = 10$ ou $\sigma = 20$. Les écarts-types de strate correspondaient à un échantillon aléatoire tiré d'une loi gamma de paramètre d'échelle égal à un et de paramètre de forme $\gamma\sigma$ avec soit $\gamma = 0,10$ ou $\gamma = 0,25$. Nous avons construit deux versions de chacun des quatre types, l'une contenant 20 strates et l'autre, 40 strates. Pour chacune des huit populations, nous avons tiré 500 échantillons, chacun constitué de deux observations sélectionnées au hasard sans remise dans chaque strate. Pour chaque échantillon, nous avons comparé les estimations obtenues par l'approche classique à celles fondées sur la LPDP. Les résultats figurent au tableau 7.2. Nous ne présentons que ceux pour les populations à 20 strates, parce que les résultats pour les populations à 40 strates sont similaires. Les deux méthodes sont approximativement sans biais et l'estimation ponctuelle fondée sur la LPDP paraît être un tout petit peu meilleure, mais les intervalles de confiance produits par la LPDP sont clairement supérieurs. Même si, dans un cas, les intervalles selon la LPDP sont manifestement trop longs, globalement, ces intervalles sont nettement meilleurs que ceux obtenus par la méthode classique.

Quelles sont les causes des mauvaises propriétés des intervalles donnés par la LPDP dans un cas ? Des simulations supplémentaires indiquent que, lorsque les moyennes de strate varient fortement et que les variances de strate ont tendance à être relativement faibles, les intervalles donnés par la LPDP ont tendance à être trop longs. Dans nos simulations, le cas où $\sigma = 20$ et $\gamma = 0,10$ produit une population contenant ce genre de strate. Lorsque la taille d'échantillon a été augmentée pour passer à quatre unités par strate, la différence entre les deux méthodes n'était

plus aussi importante, mais les constatations restaient les mêmes. Les intervalles sous la méthode classique ont tendance à être trop courts et à avoir une couverture trop faible, tandis que les intervalles sous la LPDP sont plus longs et ont tendance à avoir une trop grande couverture.

Tableau 7.2
Résultats de simulation sur 500 échantillons aléatoires stratifiés de taille deux dans chaque strate des populations contenant 20 strates. La couverture nominale est de 0,95 pour chaque méthode.

Méthode	Valeur moyenne	Erreur moyenne	Longueur moyenne	Fréquence de couverture
		$\sigma = 10$ et $\gamma = 0,10$		
Classique	148,40	2,37	8,30	0,808
LPDP	148,39	2,22	12,20	0,95
		$\sigma = 10$ et $\gamma = 0,25$		
Classique	144,28	5,70	20,59	0,834
LPDP	144,18	5,41	28,38	0,950
		$\sigma = 20$ et $\gamma = 0,10$		
Classique	152,75	3,02	10,52	0,828
LPDP	152,61	2,78	22,88	0,996
		$\sigma = 20$ et $\gamma = 0,25$		
Classique	155,94	6,72	23,17	0,826
LPDP	155,89	6,35	34,96	0,962

Clairement, le choix d'une bonne méthode pour construire un intervalle de confiance dépend non seulement de la taille des intervalles qu'elle produit, mais aussi de la probabilité avec laquelle ces intervalles ne contiennent pas la valeur réelle, mais inconnue, du paramètre. Cohen et Strawderman (1973) et Meeden et Vardeman (1985), entre autres, ont étudié la question de l'admissibilité des procédures pour estimer les intervalles de confiance. Bien que les résultats présentés dans ces études ne s'appliquent pas directement ici, le deuxième article montre que, dans certaines situations, certaines procédures bayésiennes peuvent donner des procédures presque admissibles. Ce genre d'arguments et le fait que l'intervalle sous la méthode classique est beaucoup trop court constituent, à notre avis, des preuves circonstancielles que les intervalles sous la LPDP dans le présent exemple ne sont pas outrageusement trop longs. En résumé, nous pensons que, dans le cas particulier important où les tailles d'échantillon sont égales à deux et où

les strates ne diffèrent pas considérablement, les intervalles sous la LPDP semblent être des concurrents sérieux des intervalles sous la méthode classique.

7.3 Série de microdonnées à grande diffusion intégrées

Le Minnesota Population Center (MPC) est un groupe interdépartemental de recherche en démographie à l'Université du Minnesota. Un objectif important du MPC est de créer des bases de données et des outils statistiques qui peuvent être utilisés pour étudier le comportement économique et social. Une base de données d'intérêt est la Integrated Public Use Microdata Series (IPUMS), qui résulte de la consolidation des données des recensements des États-Unis et d'autres enquêtes nationales couvrant la période de 1850 à aujourd'hui (Ruggles, Sobek, Alexander, Fitch, Goeken, Hall, King et Ronnander 2004). Le terme *microdonnées* est utilisé dans ce contexte parce que chaque ligne de l'ensemble de données IPUMS correspond à une personne ou à un ménage; un contraste peut être fait entre ce fin niveau de détail et une publication ou un tableau sommaire en ligne typique du Census Bureau fournissant à l'utilisateur des données une totalisation des microdonnées à un niveau géographique particulier préétabli (qui peut correspondre à l'ensemble du pays, à des États, des comtés, des secteurs de recensement, *etc*.).

Un ensemble de données qui offre un riche tableau de variables numériques est celui de l'American Community Survey (ACS) de 2005. Ce produit du U.S. Census Bureau provient d'une grande enquête sur échantillon, et le Census Bureau ne connaît pas les moyennes de population réelle des variables. Pour procéder à des simulations avec les données de l'ACS de 2005, nous avons attribué à l'échantillon le rôle de la population. Plus précisément, nous avons supposé que la population complète était un ensemble de 3 579 résidents de Minneapolis qui étaient en âge de travailler (de 25 à 75 ans) et dont la rémunération annuelle était comprise entre 20 000 $ et 120 000 $. Pour les besoins de notre étude, les deux variables d'intérêt étaient :

- *inctot* : le revenu total avant impôt en 2004.

- *sei* : l'indice socioéconomique de Duncan. Créé durant les années 1950, cet indice est une variable numérique visant à évaluer le prestige associé à la profession d'une personne. Le domaine de valeur de cette variable est [1,100].

Pour nos simulations, nous avons pris $y = \log(inctot)$ et $x = sei$. La corrélation entre y et x est de 0,398 et nous supposons que la moyenne de x est connue. Pour estimer la moyenne de population de y nous avons considéré l'estimateur fondé sur la LPDP et l'estimateur par la régression. Nous avons utilisé deux plans de sondage différents : le plan d'échantillonnage aléatoire simple et le plan ppt(x) approximatif. Dans chaque cas, nous avons tiré 300 échantillons de taille 30. Les résultats sont présentés au tableau 7.3. Nous voyons que, même si les deux méthodes sont comparables, la LPDP donne clairement de meilleurs intervalles.

Tableau 7.3
Résultats de simulation sur 300 échantillons aléatoires de taille 30 provenant de la population IPUMS. La couverture nominale est de 0,95 pour chaque méthode.

Plan	Méthode	Erreur moyenne	Longueur/2 moyenne	Fréquence de couverture
EAS	Régression	0,052	0,128	0,943
	LPDP	0,052	0,138	0,947
ppt(x)	Régression	0,062	0,132	0,897
	LPDP	0,066	0,133	0,937

8 Dernières remarques

Dans le domaine de l'échantillonnage, la construction de pondérations relève souvent davantage de l'art que de la science. Il s'agit de l'une des conclusions que l'on peut tirer de l'article récent de Gelman (2007) et de la discussion qui l'accompagne. Il argumente en faveur d'une approche bayésienne pour construire les pondérations en utilisant des modèles de régression qui relient la caractéristique d'intérêt aux variables auxiliaires. Ici, nous avons présenté des arguments en faveur d'une approche bayésienne séquentielle qui exploite l'information présente dans les variables auxiliaires sans émettre l'hypothèse d'un modèle reliant

la caractéristique d'intérêt à ces variables auxiliaires. La pondération résultante d'une unité de l'échantillon peut être interprétée de la manière habituelle comme étant le nombre d'unités de la population que l'unité en question représente.

Une pondération fréquentiste, disons w_i, est l'inverse de la probabilité d'inclusion, et ce nombre représente le nombre d'unités de la population représentées par une unité particulière dans l'échantillon. Donc, $w_i \geq 1$ pour tout i et $\sum_{i \in s} w_i \approx N$. À la section 6, nous avons vu que, pour l'estimateur de Horvitz-Thompson, la somme des pondérations des unités n'est habituellement pas égale à la taille de la population, ce qui peut donner lieu à un mauvais estimateur, sauf dans des circonstances très particulières. Un autre problème des pondérations fréquentistes est qu'elles sont souvent ajustées – après avoir observé l'échantillon – pour s'assurer que les estimations fréquentistes concordent avec l'information a priori au sujet de la population (Kostanich et Dippo 2002). Après les ajustements, les pondérations peuvent être rééchelonnées de manière que leur somme soit égale à un total de population. Cependant, les pondérations fréquentistes ajustées ne dépendent plus uniquement du plan de sondage et elles ne représentent plus les inverses des probabilités d'inclusion. Les notions intuitives qui sous-tendent les pondérations fréquentistes portent par conséquent quelque peu à confusion. Avant les ajustements, les pondérations fréquentistes sont des fonctions du plan; par contre, après les ajustements, elles sont des fonctions du plan et d'autres informations a priori qui peuvent ou non être reliées au plan.

Les bayésiens pensent que, dans le contexte de l'échantillonnage, l'estimation est un problème de prédiction. Leurs prédictions sont fondées sur un modèle hypothétique qui peut donner lieu à l'attribution de pondération aux unités de l'échantillon. Voir, par exemple, l'article susmentionné de Gelman (2007) et Little (2004). Comme l'ont fait remarquer un certain nombre d'auteurs (Pfeffermann 1993), effectuer une analyse pondérée pour un modèle utilisant les inverses des probabilités d'inclusion peut protéger l'échantillonneur contre l'erreur de spécification du

modèle. En outre, dans certaines situations, les deux approches peuvent produire des résultats similaires.

Récemment, Rao et Wu (2010) ont élaboré des méthodes faisant appel à une approche de pseudo-vraisemblance empirique et fondent leurs inférences sur des lois a posteriori de Dirichlet. Les procédures résultantes, même si elles sont sur le plan formel quelque peu similaires à celles dont il est question ici, s'appuient sur l'information a priori d'une manière différente. Pour ces auteurs, la majorité de l'information a priori doit être filtrée à travers le plan de sondage, alors que nous pensons que l'information a priori qui est souvent incluse dans le plan de sondage peut être utilisée directement pour produire de bonnes lois a posteriori. Pour le meilleur ou pour le pire, nous sommes plus proches du scénario bayésien classique où la loi a posteriori ne dépend pas du plan de sondage.

Ici, nous nous sommes concentrés sur l'utilisation de la loi a posteriori de Pólya contrainte (LPPC) pour produire un ensemble de pondérations fondé sur l'échantillon et sur l'information a priori, puis nous avons fait nos inférences en utilisant la loi a posteriori de Dirichlet pondérée (LPDP) fondée sur ces pondérations. Strief (2007) a considéré des exemples où les pondérations produites par la LPPC étaient utilisées dans les formules fréquentistes appropriées pour obtenir une estimation de la variance et a constaté que les résultats étaient comparables à ceux donnés par les méthodes classiques. On pourrait aussi imaginer de fonder leurs inférences sur la LPDP, mais en utilisant des poids fréquentistes, obtenus disons par des méthodes de calage (Särndal et Lundström 2005). Bien que cette option mérite d'être étudiée plus en profondeur, nous nous attendons à ce que ce genre d'approche donne lieu à des procédures inférencielles ayant de bonnes propriétés fréquentistes.

Dans l'approche fondée sur le plan de sondage, la convergence est une propriété importante que doit posséder un estimateur. Pour un cas particulier important, sous un plan EAS, les estimateurs selon la LPPC sont convergents. Cela est démontré dans Geyer et Meeden (2013).

Tout comme la LPPC, la LPDP possède une justification bayésienne séquentielle. Pour obtenir plus de détails, voir Strief (2007). Les pondérations utilisées dans la LPDP ont une formulation et une interprétation cohérentes. Il s'agit toujours d'une espérance a posteriori et leur somme est toujours égale à la taille de population. Elles représentent le nombre moyen de fois que chaque unité de l'échantillon apparaît dans une copie complète simulée de la population sous la LPPC. Cette moyenne est calculée par rapport à la loi uniforme sur toutes les copies possibles de la population qui contiennent juste les unités comprises dans l'échantillon et qui satisfont les contraintes données. Ces pondérations ne dépendent que du même type d'information a priori objective au sujet de la population que celle souvent utilisée pour définir et ajuster les pondérations fréquentistes. Il est donc possible d'y intégrer l'information a priori sans devoir spécifier explicitement une loi a priori.

Dans la plupart des cas, la pondération attribuée à une unité de l'échantillon dépend des autres unités de l'échantillon. Nous avons soutenu qu'après avoir sélectionné l'échantillon, on devrait raisonner conditionnellement. Autrement dit, étant donné l'échantillon, les pondérations devraient dépendre de toute l'information a priori disponible au sujet de la population, mais non de la façon dont l'échantillon a été sélectionné. (Nous supposons que la personne qui sélectionne l'échantillon et l'analyste ne font qu'un.) Toute procédure élaborée de cette manière devrait donner de bons résultats pour divers plans de sondage. Pour toute procédure, qu'elle soit fréquentiste, bayésienne ou bayésienne séquentielle, il s'agit du critère décisif : elle doit être évaluée d'après la façon dont elle se comporte sous échantillonnage répété conformément au plan de sondage d'intérêt.

Afin de mettre en œuvre les méthodes décrites ici, on doit d'abord utiliser la LPPC pour calculer les pondérations pour l'échantillon observé. Ensuite, on doit utiliser ces pondérations dans la LPDP pour simuler des copies complètes de la population. La première étape est la plus difficile, quoique le module externe *polyapost* la rend relativement simple pour toute personne familiarisée avec R. Une fois que les pondérations sont connues, il est facile d'exécuter la

simulation à partir de la LPDP au moyen de nombreux progiciels. Cela rend notre approche plus pratique pour les ensembles de données d'enquêtes (comme l'IPUMS) qui sont présentés avec les pondérations connexes et sont utilisés par de multiples chercheurs. Une limite plus sérieuse tient au fait que nous n'avons pris en considération que des plans d'échantillonnage simples à un seul degré. Les travaux doivent se poursuivre afin d'étendre ces méthodes à des plans à plusieurs degrés plus compliqués. Si les contraintes sous-jacentes sont sélectionnées judicieusement, les procédures résultantes peuvent avoir de bonnes propriétés fréquentistes pour divers plans de sondage. Les pondérations fondées sur la méthode bayésienne séquentielle peuvent être vues comme notre meilleure approximation de la population inconnue étant donné les unités échantillonnées et l'information a priori dont nous disposons.

Remerciements

L'étude a été financée en partie par la subvention NSF Grant DMS 0406169.

Bibliographie

Binder, D. (1982). Non-parametric Bayesian models for samples from a finite population. *Journal of the Royal Statistical Society*, Séries B, 44, 388-393.

Booth, J.G., Bulter, R.W. et Hall, P. (1994). Bootstrap methods for finite population sampling. *Journal of the American Statistical Association*, 89, 1282-1289.

Cohen, A., et Strawderman, W. (1973). Admissible confidence interval and point estimation for translation of scale parameters. *Annals of Statistics*, 1, 545-550.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (avec discussion). *Statistical Science*, 22, 153-188.

Geyer, C., et Meeden, G. (2013). Asymptotics for constrained Dirichlet distributions. *Bayesian Analysis*, 8, 89-110.

Ghosh, M., et Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman and Hall, Londres.

Gross, S. (1980). Median estimation in survey sampling. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 181-184.

Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 159-167.

Hu, M., Zhang, F., Cohen, M. et Salvucci, S. (1997). On the performance of replication-based variance estimation methods with small number of psus. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Kostanich, D.L., et Dippo, C.S. (2002). Design and methodology: 63rv. Rapport technique, The U.S. Census Bureau et The Department of Labor Statistics.

Lazar, R., Meeden, G. et Nelson, D. (2008). Une approche bayésienne non informative de l'échantillonnage d'une population finie en utilisant des variables auxiliaires. *Techniques d'enquête*, 34, 1, 55-70.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite poplation sampling. *Journal of the American Statistical Association*, 99, 546-556.

Lo, A. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.

Meeden, G., et Vardeman, S. (1985). Bayes and admissible set estimation. *Journal of the American Statistical Association*, 80, 465-471.

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

Rao, J.N.K., et Wu, C. (2010). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society*, Séries B, 72, 533-544.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, www.R-project.org.

Royall, R., et Cumberland, W. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R., et Cumberland, W. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C.A., Goeken, R., Hall, P.K., King, M. et Ronnander, C. (2004). Integrated public use microdata series: Version 3.0 [machine-readable database]. University of Minnesota.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.

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

Strief, J. (2007). Bayesian Sampling Weights: Toward a Practical Implementation of the Polya Posterior. Thèse de doctorat, University of Minnesota.

Zheng, H., et Little, R. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics,* 19, 99-117.

Zheng, H., et Little, R. (2004). Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés. *Techniques d'enquête*, 30, 2, 233-243.

Zheng, H., et Little, R. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.