Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Meilleur ajustement des pondérations des répondants dans les populations asymétriques

Glen Meeden, Zack Almquist et Charles Geyer[1]

Résumé

Sous l'approche classique de traitement des observations manquantes fondée sur le plan de sondage, la construction de classes de pondération et le calage sont utilisés pour ajuster les poids de sondage pour les répondants présents dans l'échantillon. Ici, nous utilisons ces poids ajustés pour définir une loi de Dirichlet qui peut servir à faire des inférences au sujet de la population. Des exemples montrent que les procédures résultantes possèdent de meilleures propriétés de performance que les méthodes classiques quand la population est asymétrique.

1. Introduction

Selon l'approche des sondages fondée sur le plan, l'information au sujet de la population est utilisée pour choisir le plan de sondage et, après avoir observé l'échantillon, pour ajuster les poids de sondage. Sous l'approche bayésienne formelle des sondages, l'information au sujet de la population est incorporée dans une loi a priori. Après avoir observé l'échantillon, 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. La loi a posteriori ne dépend pas du plan de sondage. Dans les enquêtes à grande échelle, les méthodes bayésiennes ont été peu utilisées en pratique parce qu'il est difficile de trouver des lois a priori sensibles. En particulier, on ne sait pas clairement comment intégrer dans une loi a priori le type d'information contenue dans le plan de sondage, afin de l'utiliser ensuite dans le calage pour tenir compte des observations manquantes. Un avantage de l'approche bayésienne tient au fait que l'on peut trouver des estimateurs ponctuels et d'intervalle pour de nombreux paramètres de la population par des simulations à partir de la loi a posteriori. Ici, nous soutiendrons que l'on peut combiner les caractéristiques des deux approches pour aboutir à de meilleures inférences. Les estimations seront fondées sur une loi a posteriori, mais, paradoxalement, sans que l'on doive spécifier une loi a priori. Au lieu de cela, après avoir observé l'échantillon, on sélectionne une loi a posteriori qui dépend du plan de sondage et de toutes les autres informations dont dispose le statisticien.

2. L'estimateur de Horvitz-Thompson

Considérons une population finie de taille N, où y est la variable d'intérêt et x est une variable auxiliaire qui contient de l'information au sujet de y. Étant donné un plan de sondage à taille d'échantillon fixe n, soit π_i la probabilité que l'unité i soit sélectionnée dans l'échantillon. Si l'on suppose que les y_i sont à peu près proportionnelles aux x_i, alors un plan très utilisé est l'échantillonnage aléatoire sans remise, où la probabilité que l'unité i soit sélectionnée est proportionnelle à x_i. Dans ce cas, $\pi_i = n(x_i / T_x)$, où T_x est le total de population des valeurs de x. Si $\mathrm{wt}_i = 1/\pi_i$, alors il s'agit du poids attribué à l'unité i. Une unité obtient un petit poids si la population ne contient que quelques autres unités ayant des valeurs de x similaires et un grand poids s'il existe de nombreuses autres unités ayant une valeur similaire de x. Étant donné une unité dans l'échantillon, son poids représente le nombre d'autres unités dans la population qui lui sont similaires.

[1] School of Statistics, 313 Ford Hall, 224 Church ST S.E., University of Minnesota, Minneapolis, MN 55455-0460

Si s désigne les étiquettes des unités comprises dans l'échantillon, alors $\mathrm{HT} = \sum_{i \in s} \mathrm{wt}_i * y_i$ est l'estimateur de Horvitz-Thompson du total de population. Il est facile de vérifier qu'il s'agit d'un estimateur sans biais et que, pour tout échantillon s, nous avons $\sum_{i \in s} \mathrm{wt}_i x_i = T_x$. Autrement dit, l'estimateur HT est calé sur x (Särndal, 2007). Dans l'approche des sondages fondée sur le plan de sondage, ces poids jouent un rôle important. Non seulement ils définissent l'estimateur, mais ils servent aussi à obtenir une estimation de la variance de l'estimateur.

Si l'on utilise un plan d'échantillonnage aléatoire simple sans remise, alors $\mathrm{wt}_i = N / n$ et, pour chaque échantillon, $\sum_{i \in s} \mathrm{wt}_i = N$. Toutefois, pour la plupart des autres plans, il n'en est pas ainsi. Par conséquent, l'estimateur HT n'est pas robuste à l'hypothèse que $y_i \propto x_i$. Si l'on remplace chaque y_i par $y_i + \lambda$ pour un certain nombre fixé λ, l'estimateur HT présente un comportement beaucoup moins désirable. Il est encore sans biais, mais son erreur absolue peut devenir beaucoup plus grande et ses intervalles de confiance, beaucoup plus larges. Cela tient au fait que la somme des poids dans un échantillon n'est pas égale à la taille de la population. Pour une discussion plus approfondie de cet aspect, voir Strief et Meeden (2014). Si l'on renormalise les poids de manière que leur somme soit égale à la taille de la population, ils ne seront plus calés sur x. Une meilleure solution, à notre avis, consiste à trouver des poids qui sont proches des poids de sondage, qui sont calés sur x et dont la somme est égale à N. Il s'agit d'un problème de programmation quadratique dont la solution peut être trouvée à l'aide de nombreux progiciels.

À la section suivante, nous discuterons de cette approche plus en détail quand il existe des observations manquantes et la comparerons à une approche plus bayésienne.

3. Observations manquantes

3.1 L'approche classique

L'approche classique de traitement des observations manquant au hasard consiste à supposer que, pour chaque i, il existe une probabilité, disons ψ_i, que l'unité i soit observée quand elle est incluse dans l'échantillon. Cette probabilité de réponse est supposée être indépendante du plan de sondage et la probabilité que nous observions effectivement y_i dans notre échantillon est donc $\pi_i \psi_i$ qui, à son tour, donne un poids pour l'unité. Malheureusement, les probabilités ψ_i ne sont presque jamais connues. Pour pallier cette absence de connaissance des ψ_i, le statisticien utilise les valeurs observées de x pour construire des classes d'ajustement de la pondération dans l'espoir que les répondants et les non-répondants compris dans la même classe soient similaires, c'est-à-dire que les ψ_i à l'intérieur de chaque classe soient approximativement constantes. Cela suppose que les valeurs de x soient connues pour chaque unité dans l'échantillon complet. Alors, dans chaque classe, le poids total des unités de l'échantillon qui se trouvent dans la classe en question est réparti uniformément entre les répondants compris dans la classe. Soit s_r les étiquettes des répondants dans l'échantillon. Étant donné un échantillon s, pour $i \in s_r$ soit wt_i son poids ajusté obtenu selon cette procédure.

En général, les wt_i ne sont pas calés et leur somme n'est pas égale à N. Soit $\gamma = \{\gamma_i : i \in s_r\}$ un ensemble possible de poids. Comme nous l'avons indiqué plus haut, nous recommandons de trouver un nouvel ensemble de poids, disons γ^*, qui est une solution du problème

$$\min_{\gamma} f(\gamma) = \sum_{i \in s_r} (x_i / \mathrm{wt}_i)(\gamma_i - \mathrm{wt}_i)^2$$

sous les contraintes

$$\sum_{i \in s_r} x_i \gamma_i = T_x \quad \text{and} \quad \sum_{i \in s_r} \gamma_i = N,$$

où nous supposons que T_x est connu. On peut aussi inclure les contraintes additionnelles voulant que $p_i \leq \text{bd}$, $i \in s_r$, où $0 < \text{bd} < 1$ est un nombre choisi par le statisticien.

La fonction que nous choisissons pour déterminer dans quelle mesure un ensemble de poids s'écarte des wt_i est très souvent utilisée, mais d'autres choix communs ne modifieront pas beaucoup le scénario. Nous désignons par HTC l'estimateur fondé sur l'ensemble de poids trouvé comme solution du problème susmentionné.

3.2 Une approche bayésienne séquentielle

L'approche classique suppose implicitement que les seules valeurs possibles pour les unités de la population sont celles qui ont été observées dans l'échantillon. Étant donné un échantillon, soit $p = \{p_i : i \in s_r\}$, où p_i est la proportion d'unités dans la population que l'on suppose être identiques au répondant i. Nous pouvons voir p comme un paramètre inconnu que nous souhaitons estimer. En fait, tout ensemble de poids pour les répondants peut être converti en une estimation de p simplement en divisant par la somme totale des poids. Si p est un paramètre inconnu, alors étant donné un nombre $0 < \text{bd} < 1$, un espace de paramètre naturel pour p est le polytope qui est l'ensemble des vecteurs p satisfaisant

$$\sum_{i \in s_r} p_i = 1 \quad \text{et} \quad \sum_{i \in s_r} x_i p_i = \mu_x \quad \text{et} \quad 0 \leq p_i \leq \text{bd}, \ \text{pour tout} \ i \in s_r,$$

où μ_x est la moyenne des valeurs de x de la population. Nous désignons ce polytope par Γ_{bd}. Habituellement, pour les $\gamma*$ définissant l'estimateur HTC, γ^* / N sera un point de la frontière relative de Γ_{bd}.

Nous adopterons une approche bayésienne séquentielle (ABS) du problème de l'estimation de p en définissant une loi « a posteriori » pour p après avoir observé l'échantillon. Puisque nous utilisons l'approche ABS, nous ne devons pas définir une loi a priori unique sur laquelle fonder les inférences. Une introduction à cette approche peut être consultée dans Ghosh et Meeden (1997). Faute d'espace, nous omettrons la justification de l'approche ABS pour les méthodes présentées ici.

Notre loi a posteriori dépendra aussi de $\text{wt} = \{\text{wt}_i : i \in s_r\}$. Nous obtenons cette loi a posteriori en deux étapes. Soit $\hat{w} = \{w_i : i \in s_r\}$, où $w_i = n_r \, \text{wt}_i / \sum_{j \in s_r} \text{wt}_j$. À la première étape, nous utilisons $\hat{w}$ comme paramètre pour une loi de Dirichlet restreinte à l'ensemble Γ_{bd}. (Nous expliquerons un peu plus tard comment nous choisissons bd.) Ensuite, nous utilisons le package R (R Core Team, 2016) `polyapost` (Meeden et coll., 2015) pour calculer l'espérance de p, disons $\hat{p}$, sous cette loi. Notons que $\hat{p}$ est toujours dans l'intérieur relatif de Γ_{bd}, contrairement à γ^* / N qui est habituellement sur la frontière relative.

Notons que $\hat{p}$ dépend du plan, mais tient également compte de l'information contenue dans x. Alors, $N\hat{p}_i$ est un poids sensible pour l'unité i et $\sum_{i \in s_r} N\hat{p}_i y_i$ est une estimation du total de population de y. Nous désignons cet estimateur par WD.

Nous devons encore trouver un moyen d'évaluer la variabilité de l'estimateur WD. Au lieu d'utiliser la loi qui nous donne $\hat{p}$, nous nous inspirons de Strief et Meeden (2014) et définissons une deuxième loi de Dirichlet. Le paramètre de cette loi est $\alpha = n_r \cdot \hat{p}$. Dans la perspective bayésienne séquentielle, nous pouvons fonder nos inférences pour le vecteur p sur la loi de Dirichlet de vecteur de paramètre α. Notons que cette loi n'est plus restreinte à Γ_{bd}, mais s'applique au simplex complet (des vecteurs de probabilité dont les composantes sont non négatives et dont la somme est égale à un) de dimension $n_r - 1$. Le fait que cette loi a posteriori permet des vecteurs de p qui ne satisfont la contrainte qu'en moyenne aide à tenir compte de ce que les $\hat{p}_i$ sont des estimations dont les valeurs réelles sont inconnues.

Sachant cette loi a posteriori, il est facile de trouver la variance a posteriori de l'estimation correspondante du total de population. Nous supposerons qu'elle suit approximativement une loi normale et utiliserons l'approximation normale habituelle pour obtenir un intervalle de confiance à 95 % approximatif pour le total de population. Nous appellerons cette loi la loi a posteriori de Dirichlet pondérée. Pour d'autres paramètres de population d'intérêt, on peut simplement exécuter des simulations à partir de cette loi pour trouver des estimations ponctuelles et d'intervalle. Nous insistons sur le fait que cette loi a posteriori n'est fondée sur aucune hypothèse de modélisation de la relation qui existe entre les variables x et y. La seule hypothèse est que les unités qui ont des valeurs de x similaires auront tendance à avoir des valeurs de y similaires et que la probabilité de réponse est une fonction « lisse » de x.

Nous notons que Rao et Wu (2010) fondent aussi leurs inférences sur une loi de Dirichlet, mais leur justification diffère de celle donnée ici.

4. Résultat des simulations

Nous avons construit trois populations comptant $N = 10\,000$ unités fondées chacune sur la même population de valeurs de x. Puisque nous nous intéressons à des populations asymétriques, nous laissons la variable auxiliaire x être un échantillon aléatoire tiré d'une loi log-normale dont la moyenne et l'écart-type du logarithme sont $\sqrt{e}/2$ et $\sqrt{e^2 - 2e}/2$, où e est la base des logarithmes naturels. La valeur minimale, les quantiles correspondant à 25 %, 50 % et 75 %, et la valeur maximale pour cette population valent 0,09, 1,41, 2,26, 3,60 et 26,73, respectivement.

Dans la première population, les y_i sont conditionnellement indépendants sachant les x_i, et la loi conditionnelle de y_i sachant x_i est une loi normale de moyenne $8x_i^2$ et d'écart-type 0,4. La corrélation entre x et y est 0,825.

Dans la deuxième population, nous laissons la fonction de moyenne de y être une fonction de x qui n'est pas la fonction identité. Cette fonction de moyenne, m, est définie comme il suit

$$m(x) = \begin{cases} 1000(2-x)^2, & x \le 2 \\ 4(x-2)^2, & x > 2 \end{cases}$$

Pour cette population, la corrélation entre x et y vaut -0,35.

Dans la troisième population, la loi conditionnelle de y_i sachant x_i est une loi normale de moyenne $1{,}5x_i$ et d'écart-type 3. La corrélation entre x et y vaut 0,75. Dans ces deux dernières populations, la loi des y_i est conditionnellement indépendante sachant les x_i, tout comme dans la première.

Ensuite, nous devons modéliser les observations manquantes. Pour cela, nous devons définir le vecteur des ψ_i. Nous supposerons que les unités dont la valeur de x est grande seront moins susceptibles de répondre que celles dont la valeur de x est petite. Par souci de commodité, la population est étiquetée de manière que x soit une fonction croissante des indices. Nous commençons par considérer le vecteur correspondant à la suite qui va de 0,4 à 0,2 en 9 999 étapes égales. Il s'agit d'une fonction décroissante lisse des étiquettes. En pratique, nous ne nous attendrions pas à ce que le vecteur de réponse soit si lisse. Donc, nous avons ajouté des erreurs aléatoires indépendantes tirées d'une loi normale de moyenne 0 et d'écart-type 0,05 à chaque composante. Puis, nous avons appliqué au vecteur résultant la fonction linéaire qui a rééchelonné ces composantes de retour à l'intervalle $[0{,}2,\ 0{,}4]$. C'est le vecteur ψ que nous avons utilisé pour définir la probabilité qu'une unité réponde dans nos simulations. Naturellement, aucun des estimateurs que nous calculons n'est fondé sur la connaissance de ψ.

Nous avons utilisé deux plans de sondage différents dans nos simulations. Le premier est un échantillonnage aléatoire simple. Soit v le vecteur qui va de 0,3 à 0,8 en 9 999 étapes égales. Le second plan correspond à un échantillonnage

proportionnel à v. Puisque les x_i sont une fonction croissante des étiquettes, cela résultera en un plus grand nombre d'unités ayant de grandes valeurs de x_i dans l'échantillon.

Pour chacune des trois populations, nous avons tiré 500 échantillons de taille $n = 150$ pour chacun des deux plans de sondage. Dans chaque cas, le nombre moyen de répondants était environ 44. Pour chaque ensemble de simulations pour l'estimateur HTC et pour l'estimateur WD, nous avons calculé leur valeur moyenne, leur erreur absolue moyenne, la longueur moyenne de leur intervalle de confiance à 95 % approximatif et la fréquence de leurs intervalles contenant le vrai total de population.

Sous échantillonnage aléatoire simple, les ratios des erreurs absolues moyennes de l'estimateur HTC à l'erreur absolue moyenne de l'estimateur WD pour les trois populations étaient de 1,24, 1,37 et 0,93. Pour le second plan de sondage, ces ratios étaient de 1,12, 1,27 et 0,93. Nous avons réalisé d'autres simulations qui ne seront pas présentées ici où les populations étaient moins asymétriques. Dans ces cas, nous avons constaté que le comportement des deux estimateurs était assez similaire, sauf quand $y_i \propto x_i$, auquel cas l'estimateur HTC donne de légèrement meilleurs résultats. Cela donne à penser que, à moins qu'il existe de fortes preuves que $y_i \propto x_i$, on devrait utiliser l'estimateur WD, surtout quand la population d'intérêt est asymétrique.

Les deux estimateurs étaient presque sans biais. Par exemple, dans les deux premières populations, les deux estimateurs sont biaisés à la baisse d'à peine plus d'un pour cent. La question qui se pose naturellement est celle de savoir en quoi les pondérations HTC diffèrent des pondérations WD. L'estimateur WD donne plus de poids aux unités possédant les plus grandes et les plus petites valeurs de x et, donc, de ce fait, un peu moins aux unités se trouvant au milieu. Pour la première population sous échantillonnage aléatoire simple, pour un échantillon donné, nous avons considéré les unités ayant pour valeur de x la valeur minimale, le premier quartile (0,25), la médiane, le troisième quartile (0,75) et la valeur maximale. Les poids moyens attribués à ces unités sous l'estimateur HTC étaient de 181, 202, 217, 249 et 263, respectivement. Tandis que pour l'estimateur WD, les poids moyens étaient de 201, 201,213, 233 et 331, respectivement.

La fréquence de couverture des intervalles de confiance à 95 % approximatifs de l'estimateur WD pour la première population était de 0,984 et de 0,976 pour les deux plans de sondage. Pour la deuxième population, les chiffres correspondants étaient de 0,948 et 0,896, tandis que pour la troisième, ils étaient de 0,962 et 0,992.

Notre première population est similaire à celle discutée dans Dorfman (1994) et dans Rao et coll. (2003). Ces auteurs ont observé qu'émettre l'hypothèse d'une relation linéaire entre y et x quand en fait elle était quadratique peut mener à de mauvais intervalles de confiance dont les probabilités de couverture réelles s'écartent fortement des niveaux nominaux. Le problème tient au fait que l'on ne peut pas utiliser un modèle quadratique quand le total de population des valeurs de x_i^2 est inconnu. Pour étudier cet aspect plus en profondeur, nous avons sélectionné 500 échantillons aléatoires simples de taille 44 dans la première population où il n'y avait pas d'observations manquantes. Rappelons que 44 est le nombre moyen de répondants que nous avons obtenu dans nos simulations précédentes. Le ratio de l'erreur absolue moyenne pour l'estimateur par la régression issu de cet ensemble de simulations à l'erreur absolue moyenne de l'estimateur WD provenant du premier ensemble de simulations était de 1,06. Les intervalles de confiance à 95 % associés pour l'estimateur par la régression ne contenaient le vrai total de population que dans 67,8 % des cas et le ratio de la longueur moyenne des intervalles pour les deux méthodes était de 0,36.

Nous avons effectué une seconde simulation similaire pour la deuxième population, et le comportement de l'estimateur par la régression sans observations manquantes était similaire à celui de l'estimateur WD avec observations manquantes.

Enfin, nous avons exécuté une simulation similaire pour la troisième population. Ici, le ratio de l'erreur absolue moyenne de l'estimateur par la régression à l'erreur absolue moyenne de l'estimateur WD dans le premier ensemble de simulations était de 0,88. Nous avons également calculé l'erreur absolue moyenne de l'estimateur HT et de sa version HTC qui fait en sorte que les poids soient calés sur x et aussi que leur somme soit égale à N. L'erreur absolue moyenne de l'estimateur HTC était à peine 1 % plus grande que celle de l'estimateur par la régression, tandis

que l'erreur absolue moyenne de l'estimateur HT était 50 % plus grande que celle de l'estimateur par la régression. Donc, même si $y_i \propto x_i$ dans cette population, l'estimateur HT donne de mauvais résultats parce que la population est très asymétrique.

Dans ces simulations, étant donné un échantillon, nous posons que $\mathrm{bd} = 5/n_r$, où n_r est le nombre de répondants dans l'échantillon. Si au lieu de 5, nous avions utilisé dans le numérateur de bd tout nombre variant de 3 à 9, nos résultats n'auraient pas changé beaucoup. Mais le choix de $\mathrm{bd} = 2/n_r$ serait trop petit. Donc, nous voyons que notre méthode est assez robuste au choix de bd.

5. Commentaires

Toutes les simulations décrites ici ont été exécutées avec la package R `polyapost` qui, depuis sa version 1.4-2, permet de calculer les espérances pour toute loi de Dirichlet contrainte à un polytope (l'ensemble satisfaisant une famille finie de contraintes linéaires d'égalité et d'inégalité). Cela permet à toute personne familiarisée avec R de calculer facilement l'estimateur WD. Pour simplifier, nous n'avons considéré que la situation avec une seule variable auxiliaire, mais il est possible d'intégrer plus d'une variable dans l'ensemble de contraintes.

L'approche bayésienne séquentielle décrite ici est une extension de certaines idées exposées dans Strief et Meeden (2014). Ces auteurs soutenaient que le plan de sondage n'avait pas d'importance après que les données avaient été recueillies, et ils ont utilisé des contraintes linéaires sur les variables auxiliaires et la loi uniforme sur un polytope pour obtenir un ensemble de poids. Des travaux subséquents ont montré que l'approche décrite ici, dans laquelle est incorporé le plan de sondage, donne de meilleurs résultats. Nous avons vu que les procédures résultantes ne comportent aucune hypothèse de modélisation au sujet de la relation entre y et x. Elles combinent de manière cohérente et objective l'information avant et après tirage de l'échantillon dont dispose l'échantillonneur et peuvent donner des procédures possédant de bonnes propriétés sous le plan pour les populations asymétriques pour lesquelles les méthodes classiques échouent. Une étude plus approfondie du comportement des intervalles de confiance approximatifs de notre loi a posteriori de Dirichlet pondérée est nécessaire. Alternativement, nous pourrions utiliser nos poids dans des méthodes fondées sur le plan de sondage classiques pour obtenir une estimation de la variance de notre estimateur ponctuel.

Bibliographie

Dorfman, A. H. (1994). A note on variance estimation for the regression estimator in double sampling. *Journal of the American Statistical Association*, 89:137–140.

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

Meeden, G., Lazar, R., and Geyer, C. (2015). R package polyapost: Simulating from the polya posterior, version 1.4-2. http://CRAN.R-project.org/package=polyapost.

R Core Team (2016). *R: A language and evironment for statistical computing*. R Foundation for Statistical Computing. http://R-project.org.

Rao, J. N. K., Jocelyn, W., and Hidiroglou, M. A. (2003). Confidence interval coverage properties for regression estimators in uni-phase and two-phase sampling. *Journal of Official Statistics*, 19:17–30.

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

Särndal, C. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119.

Strief, J., and Meeden, G. (2014). Objective stepwise bayes weights in survey sampling. *Survey Methodology*, 39:1–27.