L'APPARIEMENT D'ÉCHANTILLONS : VERS UNE APPROCHE PROBABILISTE POUR LES ENQUÊTES SUR LE WEB?

Pierre Lavallée

Statistique Canada, Ottawa (Ontario), K1A 0T6, Canada, pierre.lavallee@statcan.gc.ca

Résumé. De plus en plus, les firmes de sondage utilisent des panels de répondants sur le Web pour effectuer leurs enquêtes. La plupart du temps, ces panels ne peuvent être considérés comme des enquêtes probabilistes et, par conséquent, l'inférence à partir de ces derniers est sujette à caution. Dans le but d'introduire une composante probabiliste aux panels sur le Web, Rivers (2007) a proposé l'*Appariement d'échantillons* (*Sample Matching*). Cette méthode consiste à tirer un échantillon probabiliste à partir d'une base de sondage, puis à relier cet échantillon aux répondants du panel sur le Web en utilisant l'appariement statistique : chaque individu de l'échantillon probabiliste est ainsi apparié à un des répondants du panel selon des caractéristiques données, sans toutefois chercher à faire correspondre exactement les personnes. On décrira tout d'abord en détail de l'Appariement d'échantillons. Cette méthode sera aussitôt comparée au Sondage indirect (Lavallée, 2002). On présentera par la suite les hypothèses et justifications théoriques reliées à l'Appariement d'échantillons. On terminera par des exemples d'application de la méthode.

Mots-clés. Panel sur le Web, appariement statistique, Sondage indirect, sélection ignorable.

1. Introduction

Toute enquête, que ce soit par sondage ou par recensement, possède une phase de collecte de l'information. Jusqu'aux années 70, la collecte se faisait en grande partie par la poste ou en personne. Par la suite, avec l'accessibilité du téléphone à la plupart des ménages, on a vu de plus en plus d'enquêtes avec *collecte téléphonique*. Ceci a conduit, entre autres, au développement de la *composition aléatoire de numéros de téléphone* (ou *Random Digit Dialing*, RDD). Depuis une dizaine d'années, l'avènement du téléphone portable a fait en sorte d'élargir considérablement la couverture des personnes possédant individuellement un téléphone. Malheureusement, malgré cette bonne couverture, les taux de réponse aux enquêtes téléphoniques ne cessent de décroître. Certaines enquêtes qui avaient des niveaux acceptables de taux de réponse (disons, 70%) n'ont plus que de faibles taux (20%, par exemple) et même avec un suivi intense des non-répondants, on ne peut aspirer qu'à des taux encore inacceptables (40%).

L'avènement du Web dans les années 90 a ouvert la porte à de nouvelles formes de collecte de l'information pour les enquêtes. Une de ces formes est la réponse par Internet où on demande aux individus sélectionnés aléatoirement dans l'échantillon d'inscrire directement leur réponse

sur un site Web. Ceci a été le cas du Recensement de la population canadienne de 2006. Il s'agit donc ici de sélectionner un échantillon conventionnel de façon probabiliste et de demander aux individus de répondre sur le Web, plutôt qu'à l'aide d'un questionnaire papier ou téléphonique.

Une autre forme de collecte d'information sur le Web est l'utilisation de grands panels d'internautes à participation volontaire. Ces panels sont beaucoup utilisés, entre autres, dans le domaine du marketing. Ces panels sont constitués d'individus utilisant régulièrement le Web et à qui on pose des questions reliées à différents sujets. Les résultats tirés de leurs réponses sont alors ajustés ou non —par exemple, en effectuant un calage sur marges— afin de fournir des résultats aux chercheurs. L'un des aspects intéressants de ces panels est qu'ils sont généralement très grands et peuvent alors fournir jusqu'à des centaines de milliers de répondants. Il est cependant important de noter que ces panels sur le Web sont généralement très loin de l'échantillonnage probabiliste. En effet, si on désire une couverture complète de la population, il y a premièrement un problème parce que ce n'est pas toute la population —bien que ce nombre augmente de jour en jour— qui a accès à Internet. Deuxièmement, seuls les internautes inscrits aux panels sont consultés. Et finalement, ces internautes sont des volontaires qui n'ont pas été, proprement dit, choisis au hasard.

Rivers (2007) s'est intéressé à l'utilisation des panels sur le Web et notamment au problème consistant à conférer un caractère probabiliste à ces enquêtes. Pour ce faire, il a proposé l'*Appariement d'échantillons* (*Sample Matching*). Cette méthode consiste à tirer un échantillon probabiliste à partir d'une base de sondage, puis à relier cet échantillon aux répondants du panel sur le Web en utilisant l'appariement statistique. Avec l'appariement statistique, chaque individu de l'échantillon probabiliste est apparié à l'un des répondants du panel selon des caractéristiques données, sans toutefois chercher à faire correspondre exactement les personnes (voir Bernier et Nobrega, 1999).

Dans cet article, on décrira tout d'abord en détail l'Appariement d'échantillons. Cette méthode sera aussitôt comparée au Sondage indirect (Lavallée, 2002). On présentera par la suite les hypothèses et justifications reliées à l'Appariement d'échantillons. On terminera par des exemples d'application de la méthode.

2. L'Appariement d'échantillons

On présente ici l'Appariement d'échantillons selon l'article de Rivers (2007). On suppose que l'on s'intéresse à une population *U* de taille *N*. On sélectionne alors un échantillon *s* probabiliste de taille *n* selon un plan de sondage $p(s)$. Soit $\pi_i > 0$, la probabilité de sélection de l'unité *i*. Pour chaque individu *i* de cet échantillon, on désire mesurer une variable d'intérêt y_i, i=1,...,n. On suppose par ailleurs que l'on dispose d'un vecteur de *p* variables auxiliaires $\mathbf{x}_i = (x_{1i}, ..., x_{pi})'$ disponible, soit pour toute la population *U*, soit pour l'échantillon *s*, ou sinon seulement pour des totaux marginaux de $\mathbf{x}_i$. Par exemple, si le vecteur $\mathbf{x}_i$ identifie le groupe âge-sexe de

l'individu *i*, les totaux marginaux de $\mathbf{x}_i$ correspondent aux effectifs des groupes âge-sexe au sein de la population.

Différentes variables sont mesurées à partir d'un panel Γ sur le Web, non probabiliste, de taille $\hat{n}$ [1]. On suppose en général que $\hat{n}$ est grand, par exemple, une centaine de milliers de répondants. Pour chaque individu *k* du panel, on suppose que l'on mesure la variable d'intérêt $\hat{y}_k$ (c'est-à-dire, la variable d'intérêt y de la population *U*, mais mesurée à partir du panel) ainsi que les variables auxiliaires $\hat{\mathbf{x}}_k = (\hat{x}_{1k},...,\hat{x}_{pk})'$. Soit δ_k, la variable indicatrice qui vaut 1 si l'individu *k* est dans le panel, et 0 sinon.

On suppose que le panel n'est constitué que de répondants. Il n'y a donc pas de non-réponse dans la mesure de $\hat{y}_k$ et $\hat{\mathbf{x}}_k$, $k = 1,...,\hat{n}$. Comme le mentionne Rivers (2007), les taux de réponse à l'intérieur des panels sur le Web sont en général très élevés. Notons qu'étant donné son caractère non probabiliste, cela est équivalent à supposer de la non-réponse, mais à se concentrer seulement sur le sous-panel des répondants. Un modèle de non-réponse pourrait être utilisé pour le sous-panel, mais ce dernier serait de forme identique au modèle de participation au panel.

L'Appariement d'échantillons fonctionne comme suit. Pour chaque individu *i* de l'échantillon *s*, au lieu de tenter de mesurer directement à la variable d'intérêt y_i, on détermine l'individu *k* du panel Γ le « plus près » de l'individu *i* de la population *U*, *i*=1,...,*n*. On utilise alors la valeur de la variable $\hat{y}$ de l'individu *k* du panel Γ comme substitut à celle de l'individu *i* de *U*. On procède par appariement statistique, c'est-à-dire qu'on crée un fichier combiné où les enregistrements ne doivent pas nécessairement correspondre à la même entité physique comme, par exemple, le même individu (voir Bernier et Nobrega, 1999). Donc, si l'individu *i* de *U* est apparié avec l'individu *k* de Γ, ces deux individus n'ont pas besoin d'être le même individu. L'appariement se fait en utilisant les variables auxiliaires $\mathbf{x}$ et $\hat{\mathbf{x}}$. L'ensemble des unités appariées est appelé l'*échantillon apparié* (ou *matched sample*). Selon Rivers (2007), si le bassin de répondants du panel Γ est suffisamment grand et diversifié, l'échantillon apparié devrait avoir la même distribution des variables auxiliaires que l'échantillon *s*.

Notons que l'*imputation par le plus proche voisin* s'apparente à l'Appariement d'échantillons. En effet, ces deux méthodes utilisent une certaine forme d'appariement comme substitut à la « vraie » réponse. Si l'appariement se fait seulement sur des valeurs marginales des variables auxiliaires $\mathbf{x}$ et $\hat{\mathbf{x}}$, on se retrouve pratiquement dans le contexte du *Sondage par quotas* (voir Rivers, 2007).

On désire estimer le total $Y = \sum_{i=1}^{N} y_i$ de la population *U*. Pour ce faire, on pourrait utiliser l'estimateur de Horvitz-Thompson $\hat{Y} = \sum_{i=1}^{n} y_i / \pi_i$. Cependant, parce qu'on ne mesure pas la

---

[1] On s'inspirera ici de la notation de Rivers (2007) en utilisant le «ˆ» pour indiquer ce qui a trait au panel sur le Web.

variable *y*, on utilise alors la variable $\widehat{y}_k$ de l'individu *k* de Γ « le plus près » de l'individu *i* de *U*. Cet individu *k* est déterminé selon

$$M(i) = k \quad \text{ssi} \quad \left|\widehat{\mathbf{x}}_k - \mathbf{x}_i\right| \leq \left|\widehat{\mathbf{x}}_\ell - \mathbf{x}_i\right| \quad \text{pour} \quad \ell = 1, ..., \widehat{n}$$

Notons qu'il peut y avoir un très grand nombre de comparaisons pour déterminer $M(i)$ pour une valeur de *i* donnée. Sariyar et Borg (2010) proposent une approche par les techniques de classification pour résoudre ce problème.

Soit $\mathbf{x}_i^* = \widehat{\mathbf{x}}_{M(i)}$, la valeur de $\widehat{\mathbf{x}}$ du panel la plus près de $\mathbf{x}_i$ de la population U [2]. De même, $y_i^* = \widehat{y}_{M(i)}$ est la valeur de *y* pour l'individu *i* de *U* mesurée à partir du panel Γ. On peut alors définir l'*estimateur par appariement* selon

$$\tilde{Y} = \sum_{i=1}^{n} \frac{y_i^*}{\pi_i}$$

Pour chaque individu *i*, on peut facilement mesurer la distance entre $\mathbf{x}_i$ et $\mathbf{x}_i^*$. Si cette distance est petite — ce qui correspond à un appariement relativement proche —, on espère alors que la distribution conditionelle des y_i^* étant donné $\mathbf{x}_i^*$ sera également près de la distribution conditionelle des y_i étant donné $\mathbf{x}_i$. Tel que le mentionne Rivers (2007), on ne s'attend pas à ce que y_i et y_i^* soit hautement corrélées, vu que la variance conditionnelle de *y* étant donné **x** peut être grande, mais les distributions devraient être similaires.

3. Relation entre l'Appariement d'échantillons et le Sondage indirect

Lorsque la base de sondage ne coïncide pas avec la population cible, mais qu'elle est quand même reliée à cette dernière, on se trouve dans une situation de *Sondage indirect* (voir Lavallée, 2002). Plus formellement, on souhaite produire une estimation pour une population cible U^B, mais on ne dispose que d'une base de sondage U^A, qui est reliée d'une certaine façon à U^B. On tire alors un échantillon s^A de U^A pour produire une estimation pour U^B en utilisant les liens existants entre les deux populations. Afin de produire des estimations sans biais des quantités d'intérêt (par exemple, totaux ou moyennes) pour la population cible U^B en utilisant s^A, on construit des poids d'estimation par la *Méthode généralisée du partage des poids*.

Il existe une certaine relation entre l'Appariement d'échantillons et le Sondage indirect. En effet, dans le cas de l'Appariement d'échantillons, on dispose d'une base de sondage *U* —tout comme la base de sondage U^A du Sondage indirect— de laquelle on tire un échantillon *s*. Avec l'appariement statistique, on établit alors un pont entre la population *U* et le panel Γ, ce qui

[2] En cas d'*ex aequo*, on peut toujours choisir une valeur au hasard parmi les indices *k* identifiés.

correspond à l'établissement de liens entre U^A et U^B dans le cas du Sondage indirect. Dans le cas de l'Appariement d'échantillons, les liens entre la population U et le panel Γ sont généralement de un à un, c'est-à-dire d'un individu de U à un seul individu de Γ, tandis qu'avec le Sondage indirect, les liens peuvent être de plusieurs à plusieurs. Cependant, il est clair que l'Appariement d'échantillons pourrait se généraliser au cas de liens de plusieurs à plusieurs.

La grande différence entre l'Appariement d'échantillons et le Sondage indirect réside dans la population cible. Dans le cas du Sondage indirect, la population cible est la population U^B pour laquelle on cherche à produire des estimations (par exemple, le total Y^B). Dans le cas de l'Appariement d'échantillons, la population cible est la population U qui, dans le contexte du Sondage indirect est la population U^A. Heureusement, la théorie relative au Sondage indirect admet une forme à l'estimateur de Y^B qui s'écrit en termes de l'échantillon s^A tiré de U^A (voir Théorème 4.1 de Lavallée, 2002). Ainsi, on peut adapter l'estimation relative au Sondage indirect au contexte de l'Appariement d'échantillons, ce qui donnera ici un estimateur similaire à la forme de l'estimateur $\tilde{Y}$. Ajoutons que Lavallée et Caron (2001) ont traité du cas de Sondage indirect où il y a couplage d'enregistrements pour établir les liens entre U^A et U^B. L'appariement statistique étant une forme de couplage d'enregistrements, ceci renforce le lien entre le Sondage indirect et l'Appariement d'échantillons.

4. Hypothèses et justifications reliées à l'Appariement d'échantillons

Tout d'abord, Rivers (2007) fait le constat qu'il n'y a pas de différence importante entre la non-réponse et la participation volontaire. Dans les deux cas, les données peuvent être générées selon une distribution inconnue qui peut être différente de celle de la population. Pour toute enquête où le taux de réponse est bas, l'utilisation d'échantillons non probabilistes peut donc devenir une approche raisonnable. En effet, pour corriger une non-réponse substantielle, on a typiquement recours à des modèles de non-réponse qui mènent alors à une approche d'estimation hautement axée sur le modèle. Toute inférence dépend donc de la validité des modèles employés. Il n'y a cependant pas de différence marquante entre les hypothèses reliées aux modèles d'ajustement de la non-réponse et celles reliées à l'utilisation d'échantillons de volontaires, c'est-à-dire non probabilistes.

Rivers (2007) mentionne plusieurs conditions techniques pour que l'Appariement d'échantillons fonctionne. On reprend ici les grandes lignes de certaines de ces conditions.

i) Les observations $(y_i, \mathbf{x}_i, \delta_i)$ sont indépendantes et identiquement distribuées.

ii) La sélection du panel est aléatoire, c'est-à-dire que la variable indicatrice δ est indépendante du vecteur $(y, \mathbf{x})$. Cette condition correspond au concept de MCAR (*missing completely at random*) défini par Little et Rubin (2002). Cette condition est extrêmement forte, mais elle peut être remplacée par la condition plus faible qui suit.

iii) La sélection du panel est *ignorable*, c'est-à-dire que la variable indicatrice δ est

indépendante de *y*, étant donné les variables auxiliaires $\mathbf{x}$.

iv) Les distributions des $\mathbf{x}$ sur *U* et des $\hat{\mathbf{x}}$ sur Γ devraient être pratiquement les mêmes. Ceci est nécessaire afin que le panel Γ couvre toutes les portions pertinentes de la population *U*.

La condition (iii) est cruciale pour que l'Appariement d'échantillons donne des résultats comparables à un sondage conventionnel. Tel que noté par Brick (2011), l'Appariement d'échantillons est hautement dépendant des hypothèses de modélisation qui sous-tendent l'appariement de l'échantillon *s* aux individus du panel Γ. Il est donc essentiel d'avoir de bonnes variables auxiliaires pour l'appariement et une bonne compréhension de la relation entre ces variables auxiliaires et la variable d'intérêt. Il faut, en effet, choisir les variables auxiliaires de sorte que, pour un ensemble de variables auxiliaires donné, le fait de participer ou non au panel n'ait aucun rapport avec la variable d'intérêt mesurée. Étant donné les variables auxiliaires $\mathbf{x}$, la non-indépendance entre la participation au panel et la variable d'intérêt *y* peut introduire de sérieux biais dans les estimations.

En se concentrant sur le cas où le vecteur de variables auxiliaires **x** est un scalaire, Rivers (2007) a démontré sous certaines conditions que la distribution conditionnelle de x^*, étant donné *x*, tend vers la distribution de $\hat{x}$ au fur et à mesure que la taille $\hat{n}$ du panel augmente. Ce résultat lui a permis de démontrer que si $n \to \infty$, $\hat{n} \to \infty$ et $n/\hat{n} \to 0$, et que de plus la sélection est ignorable, on obtient que $\tilde{Y}$ converge faiblement en loi vers une variable normale d'espérance *Y*. Ainsi, si le panel Γ est suffisamment grand et que la sélection est ignorable, l'estimateur par appariement $\tilde{Y}$ devrait s'approcher du total *Y* pour une taille d'échantillon *n* suffisamment grande, mais restant petite par rapport à la taille $\hat{n}$ du panel.

5. Exemples d'application

Rivers (2007) a employé l'Appariement d'échantillons pour prédire le vote démocrate à l'élection du Congrès américain de 2006. Les résultats ont montré que l'Appariement d'échantillons donnait de meilleurs résultats que l'utilisation traditionnelle du RDD. Plus récemment, Rivers et Bailey (2009) ont utilisé l'Appariement d'échantillons pour prédire les résultats de l'élection présidentielle américaine de 2008. De nouveau, l'Appariement d'échantillon s'est avéré supérieur aux méthodes plus traditionnelles : le RDD et l'utilisation directe de panels de volontaires sur le Web.

Bibliographie

[1] Bernier, J., Nobrega, K. (1999). Aperçu du couplage d'enregistrements. *Recueil du Symposium 99 de Statistique Canada : Combiner des données de sources différentes,* Ottawa, mai 1999.

[2] Brick, M.J. (2011). The Future of Survey Sampling. *Public Opinion Quarterly,* Oxford

University Press, 2011, pp. 872-888.
[3] Rivers, D. (2007). Sampling for Web Surveys. *Proceeding of the Joint Statistical Meeting*, Salt Lake City, Utah, 2007.
[4] Rivers, D., Bailey, D. (2009). Inference from Matched Samples in the 2008 U.S. National Elections. *Procceding of the Joint Statistical Meeting*, Washington, D.C., 2009.
[5] Lavallée, P. (2002). *Le Sondage indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Belgique, Éditions Ellipse, France.
[6] Lavallée, P., Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements. *Techniques d'enquête*, Vol. 27, No. 2, décembre 2001, pp. 171-187.
[7] Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd ed*. Wiley, New York, 2002.
[8] Sariyar, M., Borg, A. (2010). The RecordLinkage Package: Detecting Errors in Data. *The R Journal*, Vol. 2, No. 2, décembre 2010, pp. 61-67.