Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Une note sur le concept d'invariance dans les plans d'échantillonnage à deux phases

par Jean-François Beaumont et David Haziza

Date de diffusion : le 20 décembre 2016

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 STATCAN.infostats-infostats.STATCAN@canada.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

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 « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

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

Signes conventionnels dans les tableaux

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, décembre 2016
Vol. 42, nº 2, pp. 337-342
Statistique Canada, nº 12-001-X au catalogue

Une note sur le concept d'invariance dans les plans d'échantillonnage à deux phases

Jean-François Beaumont et David Haziza[1]

Résumé

Les plans d'échantillonnage à deux phases sont souvent utilisés dans les enquêtes lorsque la base de sondage ne contient que peu d'information auxiliaire, voire aucune. Dans la présente note, nous apportons certains éclaircissements sur le concept d'invariance souvent mentionné dans le contexte des plans d'échantillonnage à deux phases. Nous définissons deux types de plans d'échantillonnage à deux phases invariants, à savoir les plans fortement invariants et les plans faiblement invariants, et donnons des exemples. Enfin, nous décrivons les implications d'une forte ou d'une faible invariance du point de vue de l'inférence.

Mots-clés : Échantillonnage à deux phases; estimateur d'Horvitz-Thompson; invariance faible; estimateur par double dilatation; invariance forte.

1 Introduction

Les plans d'échantillonnage à deux phases sont souvent utilisés dans les enquêtes lorsque la base de sondage ne contient que peu d'information auxiliaire, voire aucune. L'échantillonnage à deux phases consiste à tirer d'abord un grand échantillon de la population (habituellement selon un plan d'échantillonnage rudimentaire) en vue de recueillir des données sur des variables liées aux caractéristiques d'intérêt pour lesquelles la collecte est peu coûteuse. La notion qui sous-tend l'échantillonnage à deux phases est la création d'une pseudo-base de sondage plus riche en information auxiliaire que la base de sondage originale. Ensuite, en se servant des variables observées à la première phase, on peut appliquer une procédure d'échantillonnage efficace pour tirer un sous-échantillon (habituellement petit) de l'échantillon de première phase en vue de recueillir des données sur les caractéristiques d'intérêt. L'échantillonnage à deux phases peut aussi s'avérer utile dans le contexte de la non-réponse, vu que l'ensemble de répondants est souvent traité comme un échantillon de deuxième phase.

Nous adoptons la notation suivante : considérons une population U de taille N. Un vecteur $\mathbf{I}_1$ est généré conformément au plan d'échantillonnage $F(\mathbf{I}_1)$, où $\mathbf{I}_1 = (I_{11},\ldots,I_{1N})^{\mathrm{T}}$ désigne un vecteur de variables indicatrices telles que I_{1i} est égal à 0 ou à 1. L'échantillon de première phase, désigné par s_1, est l'ensemble d'unités de la population pour lesquelles $I_{1i} = 1$, et $n_1 = \sum_{i\in U} I_{1i}$ est la taille de s_1. Ensuite, un vecteur $\mathbf{I}_2$ est généré conformément au plan d'échantillonnage $F(\mathbf{I}_2 \mid \mathbf{I}_1)$, où $\mathbf{I}_2 = (I_{21},\ldots,I_{2N})^{\mathrm{T}}$ désigne le vecteur de variables indicatrices telles que I_{2i} est égal à 0 ou à 1. L'échantillon de deuxième phase, désigné par s_2, est l'ensemble d'unités de la population pour lesquelles $I_{1i} = 1$ ainsi que $I_{2i} = 1$, et $n_2 = \sum_{i\in U} I_{1i}I_{2i}$ est la taille de s_2. En pratique, notons que les variables indicatrices I_{2i} ne sont pas générées pour les unités de population qui appartiennent à l'ensemble $U - s_1$. Cependant, du moins

1. Jean-François Beaumont, Division de la coopération internationale et des méthodes statistiques institutionnelles, Statistique Canada, 100 promenade Tunney's Pasture, immeuble R.-H.-Coats, 25e étage, Ottawa, Canada, K1A 0T6. Courriel : jean-francois.beaumont@canada.ca; David Haziza, Département de mathématiques et statistique, Université de Montréal, Montréal, Canada, H3C 3J7. Courriel : haziza@dms.umontreal.ca.

conceptuellement, rien n'empêche de définir ces variables indicatrices pour les unités en dehors de l'échantillon de première phase.

Soit $\pi_{1i} = P(I_{1i} = 1)$ et $\pi_{1ij} = P(I_{1i} = 1, I_{1j} = 1)$ les probabilités de sélection d'ordre un et d'ordre deux à la première phase. De même, soit $\pi_{2i}(\mathbf{I}_1) = P(I_{2i} = 1 \mid \mathbf{I}_1)$ et $\pi_{2ij}(\mathbf{I}_1) = P(I_{2i} = 1, I_{2j} = 1 \mid \mathbf{I}_1)$ les probabilités de sélection d'ordre un et d'ordre deux à la deuxième phase. Notons que les probabilités de sélection (d'ordre un et d'ordre deux) à la deuxième phase peuvent dépendre de l'échantillon réalisé s_1.

La présentation de l'article est la suivante. À la section 2, nous définissons les concepts d'invariance faible et forte, et donnons certains exemples. À la section 3, nous discutons des implications de l'invariance faible et de l'invariance forte du point de vue de l'inférence. En particulier, nous discutons de la décomposition inverse de la variance dans le cas d'un plan d'échantillonnage à deux phases fortement invariant.

2 Le concept d'invariance

Nous faisons la distinction entre le concept d'invariance forte, qui peut également être appelée invariance en loi, et celui d'invariance faible, qui peut également être appelée invariance en les deux premiers moments.

Définition 1. *Un plan d'échantillonnage à deux phases est dit fortement invariant (ou invariant en loi) à condition que*

$$F(\mathbf{I}_2 \mid \mathbf{I}_1) = F(\mathbf{I}_2) \tag{2.1}$$

Une implication de la définition 1 est que $F(\mathbf{I}_1, \mathbf{I}_2) = F(\mathbf{I}_1)F(\mathbf{I}_2)$ et donc que, sous un plan d'échantillonnage à deux phases fortement invariant, le vecteur $\mathbf{I}_2$ peut être généré avant le vecteur $\mathbf{I}_1$. En pratique, le concept d'invariance forte n'est satisfait que pour quelques plans d'échantillonnage à deux phases seulement. Un premier exemple est l'échantillonnage de Poisson à la deuxième phase. Cela englobe le cas de la non-réponse, qui est souvent considéré comme un échantillonnage de Poisson de deuxième phase. Un autre exemple est celui de l'échantillonnage à deux degrés. Tous deux sont décrits plus en détail ci-dessous.

Exemple 1. *À la première phase, un échantillon s_1 est sélectionné conformément à un plan d'échantillonnage arbitraire, suivi à la deuxième phase d'un échantillonnage de Poisson, où les probabilités de sélection $\pi_{2i}(\mathbf{I}_1)$ des unités sont fixées avant l'échantillonnage, ce qui signifie que $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ pour $i \in U$. Puisque l'échantillonnage de Poisson est entièrement caractérisé par les probabilités de sélection d'ordre un, nous avons $F(\mathbf{I}_2 \mid \mathbf{I}_1) = F(\mathbf{I}_2)$. Par conséquent, ce plan d'échantillonnage est fortement invariant. Il peut être mis en œuvre comme il suit : premièrement, générer le vecteur $\mathbf{I}_2$ conformément au plan d'échantillonnage de Poisson $F(\mathbf{I}_2)$ et, indépendamment, générer le vecteur $\mathbf{I}_1$ conformément au plan $F(\mathbf{I}_1)$.*

Exemple 2. *L'échantillonnage en grappes à deux degrés peut être décrit comme suit. Au premier degré, on tire un échantillon de grappes de la population de grappes. Puis, au deuxième degré, dans chaque grappe sélectionnée au premier degré, on tire aléatoirement un échantillon d'éléments. Notons que, même dans ce*

cas, le vecteur $\mathbf{I}_1$ *est encore défini au niveau de l'élément, sa taille* N *correspondant au nombre d'éléments dans la population. Dans ces conditions, la variable indicatrice de sélection d'un élément* j *dans la grappe* i, I_{1ij}, *est égale à 1 pour tous les éléments* j *à l'intérieur d'une grappe sélectionnée* i. *Par conséquent, l'échantillonnage à deux degrés est un cas particulier de l'échantillonnage à deux phases décrit à la section 1. Si le tirage dans les grappes est indépendant de la sélection des grappes au premier degré, nous sommes alors en présence d'un plan d'échantillonnage en grappes à deux degrés fortement invariant. Cela est satisfait si le tirage des éléments dans une grappe est indépendant du tirage des éléments dans toute autre grappe. Un plan d'échantillonnage en grappes à deux degrés fortement invariant peut être mise en œuvre en inversant l'acte d'échantillonnage : au lieu d'échantillonner d'abord les grappes, nous commençons par tirer les éléments dans chacune des grappes de la population, puis nous échantillonnons les grappes.*

Notons que notre définition d'invariance forte pour les plans à deux degrés diffère légèrement de celle donnée dans Särndal, Swensson et Wretman (1992, chapitre 4), parce que cette dernière est restreinte aux grappes sélectionnées au premier degré. Cependant, à toute fin pratique, les définitions sont essentiellement équivalentes. Nous avons utilisé la définition 1 plutôt que la définition classique de Särndal et coll. (1992) parce qu'il n'est pas facile d'étendre cette dernière au cas de l'échantillonnage à deux phases.

Définition 2. *Un plan d'échantillonnage à deux phases est dit faiblement invariant (ou invariant en les deux premiers moments) si*

$$\pi_{2i}(\mathbf{I}_1) = \pi_{2i} \quad et \quad \pi_{2ij}(\mathbf{I}_1) = \pi_{2ij} \quad i \in s_1, j \in s_1.$$

Clairement, un plan d'échantillonnage à deux phases fortement invariant est faiblement invariant, mais le contraire n'est pas vrai. L'exemple qui suit décrit un plan d'échantillonnage qui est faiblement invariant, mais non fortement invariant.

Exemple 3. *À la première phase, nous tirons un échantillon,* s_1, *de taille* n_1, *conformément à un plan d'échantillonnage à taille fixe arbitraire. De* s_1, *nous tirons un échantillon aléatoire simple sans remise,* s_2, *de taille* n_2, *où* n_2 *est fixée avant l'échantillonnage. Ce plan d'échantillonnage à deux phases est faiblement invariant puisque* $\pi_{2i} = n_2/n_1$, *et* $\pi_{2ij} = n_2(n_2-1)/n_1(n_1-1)$, *qui restent les mêmes d'une réalisation de* $\mathbf{I}_1$ *à l'autre. Cependant, il n'est pas fortement invariant, puisqu'il est impossible de générer* $\mathbf{I}_2$ *avant* $\mathbf{I}_1$ *et de satisfaire la contrainte de taille d'échantillon fixe pour* n_2. *En fait, cela serait également vrai pour tout plan d'échantillonnage à taille fixe à la deuxième phase satisfaisant* $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ *et* $\pi_{2ij}(\mathbf{I}_1) = \pi_{2ij}$.

Enfin, nous décrivons un plan d'échantillonnage à deux phases non invariant.

Exemple 4. *À la première phase, nous tirons un échantillon aléatoire simple sans remise,* s_1, *de taille* n_1, *conformément à un plan d'échantillonnage à taille fixe arbitraire. Pour chaque* $i \in s_1$, *nous enregistrons une variable auxiliaire* x. *De* s_1, *nous tirons un échantillon de deuxième phase,* s_2, *de taille fixe* n_2, *suivant une procédure de sélection avec probabilité d'inclusion proportionnelle à la taille. Dans ce cas, nous avons*

$$\pi_{2i}(\mathbf{I}_1) = \frac{n_2 x_i}{\sum_{i \in U} x_i I_{1i}}.$$

Clairement, la probabilité d'inclusion de l'unité i dans s_2 varie d'une réalisation de $\mathbf{I}_1$ à l'autre. Puisque $\pi_{2i}(\mathbf{I}_1)$ est une fonction de $\mathbf{I}_1$, elle n'est connue qu'après que l'échantillon de première phase s_1 soit effectivement réalisé.

3 Implications de la propriété d'invariance

3.1 Invariance faible

Pour un plan d'échantillonnage à deux phases arbitraire, la probabilité d'inclusion de l'unité i, $\pi_i, i \in s_1$, est généralement inconnue et définie comme étant

$$\begin{aligned} \pi_i &= \mathrm{E}(I_{1i} I_{2i}) \\ &= \mathrm{E}\{I_{1i} \mathrm{E}(I_{2i} \mid \mathbf{I_1})\} \\ &= \sum_{\mathbf{i}_1:\, i_{1i}=1} \pi_{2i}(\mathbf{I_1}) P(\mathbf{I}_1 = \mathbf{i}_1), \end{aligned} \tag{3.1}$$

où $\mathbf{i}_1$ désigne une réalisation du vecteur aléatoire $\mathbf{I}_1$. Donc, les π_i sont généralement inconnues parce qu'elles nécessitent de connaître non seulement $P(\mathbf{I}_1 = \mathbf{i}_1)$ pour chaque $\mathbf{I}_1$ possible (ce que nous connaissons dans de nombreux cas), mais aussi $\pi_{2i}(\mathbf{I_1})$ pour chaque $\mathbf{I_1}$. Ces dernières sont généralement inconnues parce que $\pi_{2i}(\mathbf{I_1})$ peut dépendre du résultat de la phase 1. Cependant, si le plan d'échantillonnage est faiblement invariant, $\pi_{2i}(\mathbf{I_1}) = \pi_{2i}$ et (3.1) se réduit à

$$\pi_i = \pi_{2i} \sum_{\mathbf{i}_1:\, i_{1i}=1} P(\mathbf{I}_1 = \mathbf{i}_1) = \pi_{1i} \pi_{2i}. \tag{3.2}$$

Supposons que nous souhaitions estimer le total de population $t_y = \sum_{i \in U} y_i$. Puisque les π_i sont généralement inconnues, l'estimateur d'Horvitz-Thompson de t_y,

$$\hat{t}_{HT} = \sum_{i \in s_2} \pi_i^{-1} y_i,$$

ne peut habituellement pas être utilisé. Fréquemment en pratique, on recourt plutôt à l'estimateur par double dilatation (*DE* pour *double expansion*)

$$\hat{t}_{DE} = \sum_{i \in s_2} \pi_{1i}^{-1} \pi_{2i}(\mathbf{I}_1)^{-1} y_i.$$

En général, $\hat{t}_{HT}$ ainsi que $\hat{t}_{DE}$ diffèrent. Toutefois, pour les plans d'échantillonnage à deux phases faiblement invariants, l'expression (3.2) montre clairement que tous deux sont identiques.

3.2 Invariance forte

Soit θ un paramètre de population finie et $\hat{\theta}$, un estimateur de θ. La variance totale de $\hat{\theta}$ peut être exprimée sous la forme

$$V\left(\hat{\theta}\right) = VE\left(\hat{\theta}\,\middle|\,\mathbf{I}_1\right) + EV\left(\hat{\theta}\,\middle|\,\mathbf{I}_1\right). \tag{3.3}$$

La décomposition (3.3) est souvent appelée décomposition à deux phases de la variance; par exemple, Särndal et coll. (1992). Si le plan d'échantillonnage à deux phases est fortement invariant, la variance totale de $\hat{\theta}$ peut aussi être décomposée comme suit

$$V\left(\hat{\theta}\right) = EV\left(\hat{\theta}\,\middle|\,\mathbf{I}_2\right) + VE\left(\hat{\theta}\,\middle|\,\mathbf{I}_2\right). \tag{3.4}$$

La décomposition (3.4) est souvent appelée décomposition inverse de la variance, car l'ordre d'échantillonnage est inversé, ce qui ne se justifie que si le plan à deux phases est fortement invariant. La décomposition (3.4) ne peut pas être utilisée dans le cas d'un plan d'échantillonnage à deux phases faiblement invariant, car le vecteur $\mathbf{I}_2$ ne peut pas être généré avant le vecteur $\mathbf{I}_1$. La décomposition inverse a été étudiée dans le contexte de la non-réponse par Fay (1991), Shao et Steel (1999), et Kim et Rao (2009), entre autres. Dans un contexte de non-réponse, en supposant que les unités répondent indépendamment les unes des autres, l'ensemble de répondants peut être considéré comme un échantillon de deuxième phase sélectionné par échantillonnage de Poisson où les probabilités d'inclusion, appelées probabilités de réponse, sont inconnues. Si ces dernières restent les mêmes d'une réalisation de l'échantillon à l'autre, nous sommes essentiellement en présence d'un plan d'échantillonnage à deux phases fortement invariant. La décomposition (3.4) peut être utilisée pour justifier des estimateurs de variance simplifiés pour les plans d'échantillonnage à deux phases; voir Beaumont, Béliveau et Haziza (2015).

Remerciements

Les auteurs remercient un rédacteur associé et un examinateur de leurs commentaires et suggestions, qui leur ont permis d'améliorer la qualité du présent article. Les travaux de recherche de David Haziza ont été financés par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

Bibliographie

Beaumont, J.-F., Béliveau, A. et Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 3, 524-542.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, US Bureau of the Census, 429-440.

Kim, J.K., et Rao, J.N.K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

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

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.