# Décomposition de la variance due à l'imputation au niveau du non-répondant

Keven Bosa[1], Serge Godbout[2], Frédéric Picard[3] et Fraser Mills[4]

*[1] Statistique Canada, 100, promenade Tunney's Pasture, Ottawa ON K1A 0T6 et keven.bosa@canada.ca*
*[2] Statistique Canada, 100, promenade Tunney's Pasture, Ottawa ON K1A 0T6 et serge.godbout.canada.ca*
*[3] Statistique Canada, 100, promenade Tunney's Pasture, Ottawa ON K1A 0T6 et frederic.picard@canada.ca*
*[4] Statistique Canada, 100, promenade Tunney's Pasture, Ottawa ON K1A 0T6 et fraser.mills@canada.ca*

**Mots-clés.** Variance, variance due à la non-réponse, variance due à l'imputation, imputation, non réponse,

## Résumé

Lorsqu'une méthode d'imputation linéaire est utilisée pour corriger la non-réponse, et sous certaines hypothèses, on peut attribuer au niveau des unités non-répondantes la variance due à la non-réponse. L'imputation linéaire n'est pas aussi restrictive qu'il n'y paraît car les méthodes les plus populaires comme l'imputation par ratio ; donneur ; moyenne et valeur auxiliaire sont toutes des méthodes d'imputation linéaires. Le cadre théorique ainsi que l'expression donnant la décomposition de la variance due à la non-réponse au niveau de l'unité seront présentés. Des résultats par simulation seront aussi présentés. Cette décomposition peut être utilisée pour prioriser le suivi de non-réponse, prioriser les corrections manuelles ou simplement orienter l'analyse des données.

## Description de la présentation

À Statistique Canada, le Programme intégré de la statistique des entreprises (PISE) est un véhicule utilisé par plusieurs enquêtes économiques pour produire les statistiques nécessaires à leur programme. Les principaux objectifs du PISE sont : l'efficacité, l'amélioration de la qualité et être plus réactif aux changements. Le développement du PISE a permis d'élaborer un nouveau modèle de traitement appelé « estimation en continue » qui est basé sur un processus itératif. Le traitement, l'estimation et l'analyse sont répétés à plusieurs reprises pendant la période de collecte sous le modèle d'estimation en continue. Des indicateurs de qualité, tels que proposés par Godbout, Beaucage et Turmelle (2011), peuvent être calculés tout au cours de la collecte. De cette façon, on peut avoir une stratégie de suivi de collecte plus réactive et mettre de l'emphase aux endroits où la qualité est problématique. Cette présentation a été motivée par les besoins du PISE d'avoir une mesure d'impact des unités non-répondantes sur le coefficient de variation et en particulier sur la variance due à la non-réponse.

Dans le contexte du PISE, la non-réponse est habituellement traitée en utilisant l'imputation. Beaumont et Bissonnette (2011) présentent un cadre théorique dans lequel il est possible d'estimer la

variance due à l'imputation en présence d'une méthode d'imputation linéaire. Le point de départ du travail qui a mené à cette présentation est la recherche de l'impact d'une unité non-répondante sur cette variance due à l'imputation. En d'autres termes, le but est d'attribuer la contribution d'une unité non-répondante à la variance due à l'imputation estimée. Dans cette communication, il est supposé que le but de l'enquête est de produire une estimation du total pour une unique variable *y* ainsi que l'estimation de la variance associée à cette dernière. On peut aisément généraliser cette approche dans un cadre multivarié, mais la présentation aurait été grandement alourdie et c'est pour cette raison qu'il a été choisi de se limiter à un contexte simple.

Dans la première partie de la présentation, le cadre théorique tel que présenté par Beaumont et Bissonnette (2011) sera brièvement abordé. Il sera supposé, comme dans la plupart des articles traitant ce sujet, que la non-réponse est ignorable. La définition de l'imputation linéaire ainsi que le modèle d'imputation seront présentés. Il est important de souligner que le cadre théorique de cette communication repose sur l'hypothèse que le modèle d'imputation est linéaire. Plusieurs méthodes d'imputation fréquemment utilisées sont linéaires comme l'imputation par valeur auxiliaire, par régression linéaire et par donneur. Les expressions des différentes composantes de la variance seront également montrées et expliquées à l'auditoire. La variance se décompose principalement en trois termes : la variance due à l'échantillonnage, la variance due à la non-réponse et un terme de covariance. Les deux derniers termes sont parfois regroupés ensemble pour être appelés la composante de la variance due à la non-réponse et cette composante est nulle en l'absence de non-réponse.

Durant la deuxième partie de la présentation, les hypothèses menant à la décomposition de la variance due à l'imputation seront expliquées. De façon grossière, il est supposé que la valeur imputée pour la variable *y* et la vraie valeur sont suffisamment proches et que le modèle d'imputation demeure inchangé lorsqu'une unité devient répondante. Ces hypothèses mèneront aux expressions estimant la contribution d'une unité aux deux termes formant la composante de la variance due à la non-réponse.

Par la suite, des résultats pour plusieurs simulations seront présentés dans le contexte d'imputation par ratio en présence de non-réponse. Plusieurs tailles de population, tailles d'échantillon, taux de réponse et taux de conversion ont été simulés. Dans le cadre de ce travail, le taux de conversion indique le nombre d'unités qui sont converties de non-répondantes à répondantes.

Finalement, la présentation va conclure en montrant la puissance potentielle d'une telle décomposition et en indiquant les limitations de cette dernière. Un parallèle sera aussi fait avec le PISE de façon à montrer comment certains objectifs de ce dernier peuvent être atteints en utilisant cette nouvelle approche.

Il est aussi à noter que les travaux reliés à cette présentation font l'objet d'un article qui est en court de rédaction. Le plan est de soumettre l'article en question à une revue statistique avant l'été 2016.

## Bibliographie

[1] Godbout, S., Beaucage, Y. et Turmelle, C. (2011), *Achieving Quality and Efficiency Using a Top-Down Approach in the Canadian Integrated Business Statistics Program,* Work session on Statistical Data Editing, Ljubljana, Slovenie, du 9 au 11 mai 2011.

[2] Beaumont, J.-F. et Bissonnette, J. (2011), *Variance estimation under composite imputation : The methodology behind SEVANI*, Technique d'enquête, Vol. 37, No.2, 171-179.