Estimation de la variance par linéarisation via l'indicatrice d'échantillonnage avec application à la non-réponse

Audrey-Anne Vallée[1] & Yves Tillé[2]

[1] *Institut de statistique, Université de Neuchâtel, Avenue de Bellevaux 51, 2000 Neuchâtel, Suisse. audrey-anne.vallee@unine.ch*
[2] *Institut de statistique, Université de Neuchâtel, Avenue de Bellevaux 51, 2000 Neuchâtel, Suisse. yves.tille@unine.ch*

Résumé. En présence de non-réponse, la repondération et l'imputation sont couramment utilisées pour l'estimation de paramètres d'intérêts. Un éventail de méthodes est détaillé dans la littérature. À l'étape de l'inférence, plusieurs aspects sont à considérer: le mécanisme de non-réponse, le cadre de travail pour l'inférence (basé sur le modèle de non-réponse ou sur le modèle d'imputation) et la méthode d'imputation. Ces aspects sont importants pour l'estimation de la variance du paramètre d'intérêt. Cette variance est souvent approchée à l'aide d'une linéarisation de Taylor par rapport aux totaux ou aux poids de sondage. Dans cette présentation, une approche est utilisée pour linéariser directement l'estimateur du paramètre d'intérêt par rapport aux éléments aléatoires. Cette approche permet de simplifier les calculs et d'obtenir un estimateur de la variance explicite. Cette technique est appliquée aux deux cadres de travail et à différentes méthodes d'imputation. Elle permet aussi l'estimation de la variance de statistiques non-linéaires dans une base de données complète, par exemple celle de l'estimateur calé.

Mots-clés. estimation de la variance, modèle d'imputation, non-réponse, plan de sondage.

1 Introduction à l'estimation de la variance

La collecte des données lors des sondages est constamment améliorée afin de recueillir une base de données complète. Malgré les continuels efforts et améliorations, la non-réponse reste souvent inévitable. Deux pratiques courantes pour la traiter sont la repondération des unités répondantes et l'imputation des valeurs manquantes. En fonction de l'information auxiliaire disponible et du contexte, il existe un grand nombre de méthodes de repondération et d'imputation détaillées dans la littérature. Pour l'inférence en présence de traitement de la non-réponse, deux cadres de travail sont couramment utilisés. Le premier est celui basé sur le modèle de non-réponse. Dans ce cadre, le plan de sondage et le mécanisme de non-réponse sont considérés comme aléatoires, tandis que

le vecteur des valeurs prises par la variable d'intérêt est vu comme fixe. Il est nécessaire de faire des hypothèses sur le modèle de non-réponse, donc sur la probabilité de réponse. Le second cadre de travail est basé sur le modèle d'imputation. Dans ce contexte, le plan de sondage et le modèle d'imputation sont traités comme étant aléatoires. Des hypothèses sont énoncées sur la distribution de la variable d'intérêt. La seule hypothèse sur le mécanisme de non-réponse est qu'il est ignorable; il ne dépend pas de la variable d'intérêt, en ayant pris en compte l'information auxiliaire.

L'estimation de la variance du paramètre d'intérêt est plus complexe en présence de non-réponse. Deux approches sont distinguées pour décomposer la variance en termes explicites. La première est l'approche deux phases, basée sur le fait qu'un échantillon est sélectionné dans la population et un échantillon de répondants y est ensuite sélectionné. Rao (1990) et Rao et Sitter (1995) ont discuté cette approche dans le cadre de travail sous le modèle de non-réponse. L'approche dans le cadre de travail sous le modèle d'imputation est plutôt détaillée par Särndal (1992). La deuxième approche pour estimer la variance est l'approche renversée suggérée par Fay (1991). Cette fois, le processus de sélection d'individus est vu de façon inverse. Un ensemble de répondants est sélectionné dans la population et ensuite un échantillon est tiré (avec des répondants et des non-répondants). Shao et Steel (1999) ont discuté l'approche sous les deux cadres de travail. Les termes de la variance décomposée peuvent être calculés explicitement. Les paramètres d'intérêts et les méthodes d'estimation étant souvent complexes, il peut être ardu de trouver un estimateur de variance explicite. Des méthodes de ré-échantillonnage ont été développées. Rao et Shao (1992) ont introduit une méthode avec le jackknife adapté pour la présence de données imputées. Shao et Sitter (1996) ont quant à eux proposé un ajustement du bootstrap pour l'estimation de la variance dans ce contexte. Haziza (2005, 2009) a proposé un aperçu des différentes méthodes d'inférence avec des valeurs imputées.

Les paramètres d'intérêt sont souvent des fonctions non linéaires des éléments aléatoires. En effet, l'estimateur peut être complexe et la méthode d'estimation aussi. Pour calculer explicitement la variance, une technique de linéarisation de Taylor est utilisée pour faciliter les calculs. Dans cette présentation, une approche considérant les éléments aléatoires est utilisée pour la linéarisation.

2 Une approche pour la linéarisation de la variance

Afin de décomposer l'estimateur de la variance, la linéarisation de Taylor en fonction des totaux estimés (Binder, 1996) ou en fonction des poids de sondage (Demnati et Rao, 2004) est souvent utilisée pour faciliter les calculs. Graf (2011) suggère de linéariser l'estimateur directement en fonction de ce qui est aléatoire seulement. Par exemple, la fonction de totaux peut être dérivée par rapport à l'indicatrice de présence des individus dans l'échantillon plutôt que par rapport au poids de sondage. Dans cette présentation, ce raisonnement est utilisé afin d'obtenir des estimateurs de la variance. Il est aussi utilisé

pour obtenir des estimateurs de variance en présence de non-réponse. Pour le cadre de travail basé sur la non-réponse, les paramètres d'intérêts sont linéarisés en fonction des indicatrices de présence des individus dans l'échantillon et de réponse. Pour le cadre basé sur le modèle d'imputation, la linéarisation est faite en fonction des indicatrices de présence dans l'échantillon et de la variable d'intérêt. La méthode proposée est simple et applicable à différents plans de sondage, cadres de travail et méthodes d'imputation. Elle est aussi très utile pour l'estimation de la variance dans le cas de bases de données complètes. Par exemple, la variance de l'estimateur calé, voir Deville (1999), est approchée avec cette technique de linéarisation. Des exemples d'applications dans le cas complet et en présence de données imputées seront discutés dans la présentation.

Bibliographie

[1] Binder, D.A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, **22**, 17–22.

[2] Demnati, A. et Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

[3] Deville, J.C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes: linéarisation et techniques des résidus. *Techniques d'enquête*, **25**, 219–230.

[3] Fay, R. E. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference*, 429–440. U.S. Census Bureau.

[4] Graf, M. (2011). Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis: Theory and Applications*, éds. V. Pawlowsky-Glahn et A. Buccianti, 114–127. Chichester: Wiley.

[5] Haziza, D. (2005). Inférence en présence d'imputation simple dans les enquêtes: un survol. *Journal de la Société Française de Statistique*, **146**, 69–118.

[6] Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Sample surveys: Design, Methods and Applications*, éds. D. Pfeffermann et C. R. Rao, 215–246. New York: Elsevier/North-Holland.

[7] Rao, J. N. K. (1990). Variance estimation under imputation for missing data. Tech. rep., Ottawa: Statistics Canada.

[8] Rao, J. N. K. et Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, **79**, 811–822.

[9] Rao, J. N. K. et Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453–460.

[10] Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241–252.

[11] Shao, J. et Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, **91**, 1278–1288.

[12] Shao, J. et Steel, P. (1999). Variance estimation for survey data with composite

imputation and nonneglible sampling fractions. *Journal of the American Statistical Association*, **94**, 254–265.