IMPUTATION ÉQUILIBRÉE POUR LA NON-RÉPONSE EN FROMAGE SUISSE

Audrey-Anne VALLÉE, Yves TILLÉ

Université de Neuchâtel, Institut de statistique

audrey-anne.vallee@unine.ch

Mots-clés : calage, échantillonnage équilibré, imputation, non-réponse.

Résumé

La non-réponse en fromage suisse ou la non-réponse non monotone regroupe les cas où toutes les variables d'une enquête contiennent des valeurs manquantes sans schéma particulier. Il est attendu qu'un sous-ensemble des unités échantillonnées soit complètement observé, tandis que toutes les variables du reste de l'échantillon soient sujettes à la non-réponse. Nous développons l'imputation équilibrée par les *k* plus proches voisins (Hasler et Tillé, 2016) pour la non-réponse en fromage suisse. Il s'agit d'une méthode d'imputation par donneurs qui est aléatoire et construite pour répondre à plusieurs exigences. D'abord, un non-répondant peut être imputé par des donneurs qui sont proches de lui. Les distances sont calculées avec les valeurs disponibles des variables. Ensuite, toutes les valeurs manquantes d'un non-répondant sont imputées par le même donneur choisi aléatoirement. Enfin, les donneurs sont choisis de façon à ce que si on imputait les valeurs observées des non-répondants aussi, les estimations des totaux imputés et les totaux connus devraient être les mêmes. Pour imputer en respectant de telles contraintes, une matrice de probabilités d'imputation est construite à l'aide de méthodes de calage. Plus précisément, considérons un échantillon s d'une population U. Nous sommes intéressés aux variables $x_1,\ldots,x_j,\ldots,x_J$. Une partie s_r de l'échantillon est complète et le reste de l'échantillon s_m est sujet à la non-réponse. L'unité k est associée au poids de sondage d_k, qui peut être par exemple l'inverse de la probabilité d'inclusion dans l'échantillon. La variable r_{kj} vaut 1 si l'unité k a répondu à la variable x_j et 0 sinon, $j=1,\ldots,J$. La valeur x_{kj} est l'observation de la variable x_j sur l'individu k. Les donneurs sont sélectionnés de façon à respecter

$$\sum_{k \in s_m} d_k r_{kj} \dot{x}_{kj} = \sum_{k \in s_m} d_k r_{kj} x_{kj},$$

où $\dot{x}_{kj}$ est l'espérance de la valeur qui serait imputée pour la variable j de l'unité k si les valeurs observées étaient imputées aussi. Nous avons

$$\dot{x}_{kj} = \sum_{i \in s_r} \varphi_{ik} x_{kj},$$

où φ_{ik} est la probabilité que le répondant i soit le donneur choisi pour imputer les valeurs manquantes du non-répondant k. Suite au calcul des probabilités φ_{ik}, l'ensemble des donneurs est sélectionné aléatoirement à l'aide de l'échantillonnage équilibré.

Cette méthode d'imputation présente plusieurs avantages. Les valeurs imputées à chaque non-répondant peuvent être qualitatives ou quantitaves et elles sont toujours des valeurs observées. La méthode d'imputation étant aléatoire, la distribution des variables devraient être préservées. De

plus, si la relation entre les variables est linéaire ou si les voisins ont des valeurs très similaires, des estimations sans biais sont assurées. Ces propriétés sont étudiées dans une étude par simulations.

Bibliographie

[1] Hasler, C. et Tillé, Y. (2016). Balanced *k*-nearest neighbour imputation. *Statistics*, 50:1310-1331.