Imputation pour le traitement de la non-réponse partielle dans les enquêtes

David Haziza

Université de Montréal

Dans les enquêtes, il est de coutume de distinguer la non-réponse totale de la non-réponse partielle. La première est caractérisée par une absence totale d'information sur une unité échantillonnée alors que la seconde correspond au cas où une partie des variables de l'enquête ne sont pas renseignées. Dans ce cours, l'accent sera mis sur le traitement de la non-réponse partielle qui est habituellement traitée au moyen de l'imputation. On distingue l'imputation simple de l'imputation multiple. Dans le cas de l'imputation simple, une valeur manquante est remplacée par une seule valeur de remplacement, appelée valeur imputée, conduisant à la création d'un fichier de données complété. Dans le cas de l'imputation multiple, une valeur manquante est imputée au moyen de plus $M \geq 2$ valeurs imputées conduisant à $M \geq 2$ fichiers de données complétés.

L'avantage de l'imputation simple comme méthode de traitement de la non-réponse partielle sont doubles : (i) Elle conduit à la création d'un seul fichier de données complété, ce qui facilite le travail des utilisateurs de données. (ii) Les estimations ponctuelles après imputation peuvent être obtenues au moyen des procédures d'estimation utilisées dans un cas de données complètes.

La littérature portant sur l'imputation dans les enquêtes est riche. L'objectif du cours est de dresser un état de l'art sur plusieurs aspects de l'inférence en présence données imputées.

Plan du cours :

Chapitre 1 : **Introduction**

- Population finie
- Paramètres de population finies (total et moyenne, total d'un domaine et moyenne d'un domaine, fonction de répartition et quantile), équations d'estimation.
- Plan de sondage
- Estimation dans un cas de données complètes.
- Vérification (editing)
- Estimateurs imputés

Chapitre 2 : **Méthodes d'imputation déterministes**

- Méthodes d'imputation paramétriques ou semi-paramétriques : imputation par la régression linéaire et non-linéaire, variables d'enquête continue et catégorielle.
- Méthode d'imputation non-paramétrique : imputation par la plus proche voisin et imputation par le plus proche voisin de la valeur prédite (*predictive mean matching*),

imputation par la moyenne dans des classes d'imputation construites au moyen de la méthode des scores, méthodes par noyaux, fléau de la dimension (*curse of dimensionality*)
- Méthodes d'imputation composite
- Propriétés des estimateurs : total et moyenne, total d'un domaine et moyenne d'un domaine (concept de *congeniality*), fonction de répartition et quantile.
- Choix de la méthode d'imputation : nature de la variable à imputées, paramètres que l'on cherche à estimer, micro-données (méthodes par donneur vs. méthodes par valeur prédites), modèle d'imputation et diagnostiques.

Chapitre 3 : **Méthodes d'imputation aléatoires**

- Définition de l'imputation aléatoire, imputation par hot-deck aléatoire dans les classes
- Variables d'enquête continue et catégorielle.
- Propriétés des estimateurs : total et moyenne, total d'un domaine et moyenne d'un domaine, fonction de répartition et quantile.
- Imputation fractionnelle
- Imputation équilibrée

Chapitre 4 : **Estimation de la variance en présence de données imputées**

- Cadre de travail pour l'estimation de la variance : cadre à deux phases et cadre renversé
- Approche par modèle d'imputation
- Méthode d'estimation de la variance : méthode de Särdnal (1992) et méthode de Shao-Steel (1999), liens avec les méthodes de rééchantillonnage (bootstrap, jackknife)

Chapitre 5 : **Imputation multiple**

- Définition, estimation ponctuelle et estimation de la variance
- Propriétés de l'imputation multiple dans les approches fréquentistes (modèle de non-réponse et modèle d'imputation) et Bayésienne
- Imputation propre (*proper imputation*)
- Concept de *self-efficiency* et *congeniality*
- Imputation multiple dans les enquêtes