# LES RENCONTRES DE STATISTIQUE APPLIQUÉE

**Institut National d'Études Démographiques**
**(Salle Alfred Sauvy)**

# Traitement des données manquantes

## Vendredi 14 février 2014 (14h – 17h)

Les données manquantes représentent souvent un cauchemar pour tout statisticien confronté à une analyse de ses données. La non-réponse dans les enquêtes peut en effet affecter significativement les estimateurs calculés.

Après avoir défini différents types de non-réponse rencontrés dans les fichiers d'enquêtes, cette séance se propose de présenter dans un premier exposé quelques solutions à mettre en œuvre pour la correction de non-réponse. Le cas des données manquantes en analyse de données sera ensuite présenté dans le cas de variables catégorielles, par adaptation de l'algorithme NIPALS. Enfin, un dernier exposé présentera la mise en œuvre d'une méthodologie visant à corriger les estimateurs des paramètres mesurant les relations entre variables en cas de non réponse partielle, lorsque répondants et non-répondants ne partagent pas les mêmes caractéristiques au regard des variables d'intérêt.

| Résumés des présentations |
|---|
| 14h15 : Guillaume CHAUVET (Ensai -Crest) **Exposé introductif : Méthodes de correction de la non-réponse dans les enquêtes**<br>On rencontre des problèmes de données manquantes dans les enquêtes quand certaines des unités refusent de répondre, ou quand il est impossible de les contacter. On parle de non-réponse partielle lorsqu'un individu échantillonné ne renseigne qu'une partie des questions de l'enquête, et de non-réponse totale lorsqu'aucune réponse n'est observée pour un individu. En situation de non-réponse, la variance des estimateurs augmente car la taille de l'échantillon effectif diminue. La non-réponse peut également conduire à des estimateurs biaisés si les répondants diffèrent des non-répondants au regard des variables étudiées.<br>Dans cet exposé, nous introduirons les différents mécanismes de non-réponse, et nous présenterons des méthodes classiques de correction de la non-réponse dans les enquêtes. |
| 15h00 : Christian DERQUENNE (EDF)<br>**Données manquantes et algorithme NIPALS : le cas des variables catégorielles**<br>Le problème crucial des données manquantes est présent dans de nombreux domaines d'applications et par conséquent pour les différents types de variables : numériques, catégorielles (booléenne, ordinale, nominale), nombres entiers (comptage). Faut-il supprimer les observations manquantes, au risque de se retrouver avec un tableau de données quasiment vide ? De nombreuses méthodes ont été proposées pour tenir compte de ces données absentes, notamment l'algorithme NIPALS, développé par H. Wold et al. (1966, 1973) pour estimer les composantes principales d'une ACP. Nous proposons dans cet exposé de discuter de la généralisation de l'algorithme NIPALS pour le traitement des données |

catégorielles, notamment dans le cadre de l'Analyse des Correspondances Multiples (Chavent et al. - 2009, Derquenne – 2006).

16h00 : Brigitte GELEIN (Ensai)

**Imputation, MIVQUE et préservation des relations entre variables**

La non réponse peut affecter de façon significative la qualité des études statistiques lorsque répondants et non-répondants ne partagent pas les mêmes caractéristiques au regard des variables d'intérêt. On s'intéresse ici aux estimateurs des paramètres mesurant les relations entre variables en cas de non réponse partielle. Shao et Wang (2002) ont proposé une méthode d'imputation jointe par régression qui préserve ces relations. Un inconvénient de cette méthode tient à la variance additionnelle, dite d'imputation. Cette méthode repose sur l'utilisation de coefficients qui peuvent être estimés grâce aux Minimum In Variance QUadratic Estimators (MIVQUE). La méthode MIVQUE est basée sur une étude géométrique de la structure de covariance des variables. Nous proposons une méthode d'imputation où le calcul de MIVQUE est utilisé pour définir des contraintes sur les valeurs imputées finales. Des simulations montrent que la prise en compte de ces contraintes permet d'améliorer les valeurs imputées initiales obtenues avec la méthode de Shao et Wang. Nous observons en effet une réduction supplémentaire de la variance d'imputation.

Chaque trimestre, le Service Méthodes Statistiques de l'Ined propose un séminaire de statistique appliquée. Le séminaire est ouvert à tous, sans frais de participation.
Pour une bonne organisation, nous demandons aux personnes désirant y assister de s'inscrire à l'adresse suivante : http://www.ined.fr/fr/rendez_vous/rencontres_statistique_appliquee/

Pour tout autre renseignement contacter Bénédicte Garnier (benedicte.garnier@ined.fr)
Ined : 133, bd Davout, Paris 20e • Standard: 01 56 06 20 00
Métro : L9 (Porte de Montreuil) ou L3 (Porte de Bagnolet)
Bus 57 ou Tram T3b (Marie de Miribel)