SÉMINAIRE INED: LES RENCONTRES DE STATISTIQUE APPLIQUÉE

Analyse de tableaux de contingence et de tableaux disjonctifs complets

Vendredi 5 février 2010 de 14h à 17h30

RÉSUMÉS DES INTERVENTIONS

14h30 – Le PEM, pourcentage de l'écart maximum : principe et utilisation dans Trideux v 5.0 • Philippe Cibois et Alex Alber (Université de Versailles Saint-Quentin Laboratoire Printemps)

Si pour un tableau de contingence il existe plusieurs indices qui permettent de mesurer la dépendance entre lignes et colonnes (coefficients de Cramer, de Pearson, de Tschuprow), le Pourcentage de l'Ecart Maximum permet de donner l'intensité de la liaison qui existe, pour une case donnée d'un tableau de contingence, entre la modalité en ligne et celle en colonne. Cet indice n'est pas sensible à l'effectif et il permet de rechercher automatiquement les cases intéressantes dans tous les tableaux croisés possibles et de construire, pour une modalité donnée, un profil de toutes les modalités en lien avec elle. Son principe est de comparer l'écart à l'indépendance d'une case avec l'écart à l'indépendance qu'il y aurait si, en tenant compte des marges, la liaison était à son maximum, c'est-à-dire si tout l'effectif d'une des marges se trouvait dans la case. La situation d'indépendance est le degré zéro de l'indice et les écarts négatifs peuvent être envisagés.

Il est possible de généraliser l'indice à l'ensemble d'un tableau en le munissant d'un ordre sur les lignes et les colonnes. On peut ainsi, dans une enquête, repérer toutes les questions en liaison avec une variable d'intérêt, manière de faire implémentée dans Modalisa et dans Trideux.

Nous présenterons les diverses utilisations du PEM dans le logiciel Trideux Version 5.0

15h15 – Le modèle log-linéaire d'un tableau de contingence : origine historique, architecture générale et quelques exemples d'application en sociologie • Louis-André Vallet (CNRS-UMR 2773 Crest)

On évoquera l'origine historique du odds ratio, de la définition (Yules, 1900), de l'absence d'interaction entre trois variables dichotomiques (Bartlett, 1935), de l'algorithme créé pour ajuster un tableau de contingence à des marges différentes (Deming et Stephan, 1940) jusqu'à la formalisation complète du modèle multiplicatif (ou log-linéaire) d'un tableau de contingence. On présentera systématiquement le modèle saturé d'un tableau de contingence (à deux, puis à trois dimensions), la « galerie » des modèles non saturés de type hiérarchique qui peuvent en être déduits et leur signification, les procédures de test de la qualité d'ajustement d'un modèle, l'interprétation des paramètres d'un modèle et les tests statistiques associés, enfin la notion utile de modèle de base ou modèle « nul ».

L'intérêt de ce type de modélisation sera illustré par trois exemples : l'analyse des proximités et distances entre catégories sociales que révèle une table de mobilité sociale, via des modèles situés entre l'indépendance statistique et le modèle saturé ; l'analyse de la variation de la fluidité sociale entre générations selon le temps et le type de mobilité ; les avancées récentes vers la modélisation log-multiplicative pour percevoir clairement la transformation historique d'une association statistique dotée d'une forte inertie (le lien entre origine sociale et diplôme le plus élevé obtenu). Des lectures complémentaires seront proposées et on insistera sur le fait que la régression logistique constitue un cas particulier du modèle log-linéaire général.

16h15 – Classification de variables qualitatives• Brigitte Gelein (Ensai)

Si les méthodes de classification les plus connues s'appliquent à la réalisation de typologies d'individus, la classification de variables s'avère également utile. En effet, l'émergence de bases de données toujours plus volumineuses confère à la réduction des dimensions une place cruciale. Or les méthodes factorielles (ACP, AFC et ACM) ne répondent pas à tous les besoins. Par la création de groupes de variables liées, la classification de variables offre la possibilité de construire des

représentants de classes (variables synthétiques) - ou encore de choisir parmi les variables initiales celles qui sont les plus représentatives de leur classe (parangons). Ce nouvel ensemble plus restreint de variables sera plus facilement gérable et interprétable.
Certains domaines s'attachent davantage à la classification de variables qu'à celle d'individus. C'est le cas notamment de l'analyse sensorielle (création de groupes de descripteurs) ou encore de la médecine (élaboration de syndromes à partir d'un ensemble de symptômes). Cette utilisation de la classification de variables existe également dans l'économique, le social et l'environnement.
On exposera, dans une première partie, les grands principes de la classification de variables qualitatives, puis on détaillera l'approche divisive utilisée par la procédure Varclus de Sas. Enfin, la mise en œuvre pratique de cette procédure sera décrite au travers des résultats obtenus par nos travaux sur la base permanente des équipements de l'Insee.

Arnaud Bringé et Bénédicte Garnier[1]

[1] Institut national d'études démographiques • Établissement public scientifique et technologique
133, bd Davout, Paris 20e • Standard : 01 56 06 20 00 • Métro : Porte de Montreuil (ligne 9) ou Porte de Bagnolet (ligne 3)
Le Service Méthodes Statistiques de l'Ined organise un séminaire trimestriel se proposant de couvrir un domaine de Statistique Appliquée traitant de l'application d'une méthodologie. Ce séminaire est ouvert à tous, sans frais de participation.
Programme sur le site de l'Ined : http://www.ined.fr/fr/rendez_vous/rencontres_statistique_appliquee/
Pour tout renseignement contacter Mme Chafika Mekhazni, secrétaire du service, au 01 56 06 20 91