Ateliers de formation – 17 et 21 novembre 2014

8ème colloque francophone sur les sondages - Dijon

Echantillonnage adaptatif de réseaux : le cas des probabilités inégales

Jean-Claude Deville – Vendredi 21 novembre de 14h à 17h30

Ce petit cours résulte d'une envie de mettre mon nez dans l'échantillonnage adaptatif pour y mettre un peu d'ordre et vérifier une intuition : ce que je savais sur les propriétés conditionnelles du plan de Poisson pouvait résoudre des problèmes réputés insoluble. Bingo ! Profitant de ma retraite, qui me permet d'exister sans être obligé de publier n'importe quoi, j'ai pu mettre au point quelques petites choses assez sympas. Evidemment, ça dépasse le cadre d'une communication de colloque. Je remercie Anne de m'avoir donné l'occasion de présenter ce travail à quelques potes à Toulouse, il y a deux ans. Pierre m'a pris une version raccourcie à six heures pour le symposium de statcan l'an dernier. Comme je commence à comprendre un peu mieux ce que je fais, j'espère que la version de trois heures offerte à Dijon sera à la fois claire et relativement indolore. Malheureusement, comme disait Laurent Schwartz, il n'y a pas de mathématique sans peine.

Le but est de traiter une information portée par une population rare, mais pas rare partout !
Exemples :

-Statistique écologique -forestry- : répartition de certaines espèces animales ou végétales. Les sapins XXL dans une région montagneuse.

-Statistique sociale : usagers de drogues, porteurs du VIH, sans-logis, immigrés, rupins etc.

Thompson (1990) jette les bases d'une procédure d'échantillonnage originale : on ajoute à un échantillon standard « de départ », les unités « voisines » de certaines unités sélectionnées porteuses d'une « marque » particulière ; et on recommence avec les nouvelles unités marquées ainsi détectées. Les « blocs » d'unités marquées voisines sont appelées les réseaux.
La population possède donc une structure plus riche que dans les sondages habituels avec, d'une part, une sous population marquée susceptible de porter la plus grosse partie de l'information, groupée en réseaux selon une notion de voisinage, et donc de distance entre unités.
Dans ce contexte, l'usage de probabilités inégales dans l'échantillon initial, pourtant très naturel dans les applications, restait à élucider (mais il va l'être dans ce court cours). On s'intéressera en particulier à l'échantillonnage POISScon (pour poissonnien conditionnel) dont les propriétés seront exposées (je n'ose pas dire rappelées !) dans une partie spécifique.

Pour bien utiliser les données recueillies, on est conduit à des problèmes d'estimation pas faciles même quand l'échantillonnage initial est très simple !
Grosso modo, il y a deux catégories de problème :
-Que faire quand la même unité est attrapée de plusieurs façons différentes ?

-Que faire des unités « marginales » hors réseaux attrapées dans la phase d'extension de l'échantillon ?

Des solutions complètes vont être données, ainsi que des solutions approximatives quand on a un peu peur des calculs…
Elles sont basées sur le théorème de Rao-Blackwell qui améliore un estimateur quand on dispose d'une statistique exhaustive (éventuellement minimale). Vu le massacre de ce matériel généralement observé dans la littérature, une rédaction quasi-complète figure dans ces notes, mais sera seulement survolée dans l'exposé oral.

Plan du cours (révisable):
1- Position du problème
Structure de la population, voisinage, distance ; Sous population caractéristique, réseaux, grappes.
Echantillonnage initial ; complétude adaptative ; échantillon de grappes. Que faire ?
Estimateurs déjà disponibles.

2- Rappels sur les fondements
Population, observation, données, paramètre d'intérêt, données cohérentes.
Plan de sondage, échantillon, structure statistique sur les données.
Statistique, partition et tribu exhaustive ; exhaustive minimale, complète.
Estimation « euclidienne », Théorème de Rao-Blackwell et raoblackwellisation (RB).
Estimateurs linéaires sans biais « de base » : Hansen-Hurvitz (HH) et Horvitz-Thompson (HT).

3- Echantillonnages de base
Ceux de la littérature sur le domaine : SAS et Sondage avec remise (!).
Mais aussi les autres : Bernoullien, Poissonnien, Poissonnien conditionnel.
Les sondages stratifiés, multidegrés et autres deux phases ou plus réclameraient une analyse complexe que nous éluderons. Comme d'habitude le tirage systématique est à mettre à part.
En plus : diverses variantes d'échantillonnage indirect, bases multiples, échantillonnage de grappes, échantillonnage en grappes. On peut y attraper plusieurs fois le même client !

4- Echantillonnage de grappes et questions d'estimation
Les unités échantillonnées attrapent des grappes (réseaux+singletons). On regarde ce qui se passe pour les deux estimateurs HH et HT dans le contexte de chaque plan initial.
HT est son propre RB. Comment faire avec le HH ? Petits exemples…

5- Echantillonnage adaptatif de réseaux
Les estimateurs concurrents et la structure de l'échantillon final. Les unités frontières.
La raoblackwellisation (RB) dans le cas des deux types d'estimateurs et des divers types d'échantillonnage initial. Les avantages du plan Poissonnien. Petits exemples.
L'algorithme « merveilleux » et quelques unes de ses utilisations.

6- Deux alternatives si on veut faire des calculs moins lourds

Ateliers de formation – 17 et 21 novembre 2014

8ème colloque francophone sur les sondages - Dijon

SOCIETE FRANÇAISE DE STATISTIQUE

-Sous raoblackwellisation : Utilisation de statistiques exhaustives non-minimale. Quelques idées.
-Estimation par simulation : Que faire quand on ne sait pas (peut pas) calculer les espérances mais qu'on sait échantillonner ?

7- Diverses choses et conclusion. D'autres sujets de méditation possibles :
- Certains assez simples :
  - Les questions de calage et d'usage d'informations auxiliaires.
  - Variance et estimation de variance.
  - On n'a pas de hiérarchie bien nette entre les estimateurs (HH ou HT ?)
- D'autres franchement du domaine de la recherche :
  - Lien entre RB et bootstrap.
  - Les échantillonnages initiaux à plusieurs degrés ou en deux phases.
  - Les questions liées à la non-réponse.
  - L'échantillonnage où les voisins visités sont sélectionnés de façon aléatoire.
  - Le domaine des sous-RB est-il très ouvert ?
  - Approche par le modèle (y'a d'la littérature, mais ça me tente pas trop).

Bibliographie (sommaire!)

- Deville, J.C., « Sondages approfondis », *Notes de cours ENSAE,* 2005 à 2012.
- Deville, J.C. et Qualité, L., « Échantillonnage multi-dimensionnel (de plusieurs échantillons à la fois) à entropie maximum : définition, propriétés, algorithmes et programmes », *JMS Insee,* 2005.
- Dryver, A.L. and Chao, CT, « Ratio estimators in adaptive cluster sampling», *Environmetrics*, 18, pp 607-620, 2007.
- Dryver A.L., Thompson S.K., « Improving Unbiased estimators in adaptive cluster sampling», *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol 67, Issue 1, pp 157–166, February 2005.
- Lavallée,P., « Le sondage indirect ou la méthode généralisée du **partage des poids** », Éditions Ellipses, 2002.
- Levy D, « Une méthode pour adapter l'échantillonnage au cours de la collecte afin de mieux répondre à un objectif », *JMS Insee,* 2009.
- Thompson S.K., «Adaptive Cluster Sampling», *Journal of the American Statistical Association*, Vol. 85, n° 412, pp. 1050-1059, December 1990.
- Thompson S.K., « Sampling », Wiley Series in Probability and Statistics, 2002.
- Thompson S.K., « Adaptive network and spatial sampling », *Survey Methodology*, Vol. 37, n° 2, pp. 183-196, December 2011.
- Thompson S.K., Seber G.A.F, « Adaptive Sampling », Wiley 1996.
- Tillé, Y., « Sampling algorithms », *Springer,* 2006.