Construction d'échantillons astreints à des conditions de recouvrement par rapport à un échantillon antérieur et à des conditions d'équilibrage par rapport à des variables courantes :

aspects théoriques et mise en œuvre dans le cadre du renouvellement des échantillons des enquêtes d évaluation des élèves

Marc CHRISTINE[1], Thierry ROCHER[2]

L'objet de ce papier est de fournir un cadre théorique et des méthodes de résolution d'un problème d'échantillonnage qui intervient dès lors qu'on a construit un 1er échantillon d'enquête à une certaine date et que, postérieurement, on souhaite tirer un 2nd échantillon de caractéristiques fixées (taille, probabilités d'inclusion, conditions d'équilibrage) et présentant des conditions d'articulation avec le 1er (en un sens précisé ci-dessous), mais sans pouvoir agir sur le tirage déjà effectué du 1er échantillon.

A l'origine de ce problème repose la problématique des enquêtes PISA[3] : différentes vagues d'enquêtes portent sur une même matière dominante (exemple : mathématiques en 2003 et 2012). Pour faire des comparaisons pertinentes entre les deux périodes, il est préférable d'assurer un recouvrement entre les échantillons d'écoles aux deux dates. Mais il est aussi important d'avoir la meilleure « représentativité» de l'échantillon 2012 par rapport aux caractéristiques actuelles de l'univers, ce que l'on traduira par des conditions d'équilibrage.

Plus généralement, les cas les plus usuels relevant de cette approche sont : les échantillons en deux phases (2ème échantillon inclus dans le 1er), les échantillons avec disjonction lors des tirages successifs (2ème échantillon disjoint du 1er), les échantillons avec recouvrement de l'un par rapport à l'autre, les échantillons où l'on impose des conditions de « représentativité » lorsqu'on travaille soit sur la réunion des deux, soit sur le second seulement…

Le cadre proposé de réflexion et de résolution peut s'appliquer à tous les cas où les unités sont tirées *à probabilités inégales* (souvent conditionnées par un facteur de taille) : enquêtes entreprises, tirage d'unités primaires géographiques dans une enquête ménages, tirage d'établissements scolaires pour les enquêtes d'évaluation des élèves…

L'approche théorique repose sur la notion d'**échantillonnages séquentiels conditionnels** et sur la technique **des échantillons équilibrés**. On montrera dans ce papier que l'ensemble des contraintes astreintes au 2ème échantillon sont en général incompatibles, ce qui conduira à **rechercher des solutions approchées**, obtenues en relâchant la contrainte relative aux probabilités d'inclusion finales du 2ème échantillon. Ces solutions approchées posent des problèmes de calcul explicite des solutions (en général seulement possible de manière numérique) et de propriétés statistiques des estimateurs en dérivant.

[1] Insee, DCSRI
[2] DEPP
[3] Programme for international student assessment

Une mise en œuvre de ces méthodes a été ensuite expérimentée à partir des bases d'établissements scolaires françaises. L'objectif est de simuler le tirage d'un échantillon dans la base 2009 avec des conditions de recouvrement par rapport à un échantillon tiré dans la base 2000 et des conditions d'équilibrage par rapport à l'environnement de 2009. Pour simplifier, les bases ont été apurées de façon que les univers qu'elles représentent soient rigoureusement identiques (les établissements nouveaux ou disparus sont mis hors base ; en revanche, un même établissement peut avoir des caractéristiques différentes entre les deux années de référence, voire appartenir à des strates différentes).

Dans un premier temps, on simule des tirages d'échantillons qui représenteront l'échantillon tiré en 2000. Pour chacun de ces échantillons, on calcule les probabilités approchées d'inclusion de l'échantillon 2009 rendant le problème soluble. On étudie empiriquement les propriétés de ces nouvelles probabilités et, notamment, leur écart par rapport aux valeurs des probabilités de référence que l'on souhaitait imposer a priori à ce 2nd échantillon.

Des variantes seront étudiées, selon les modalités de tirage du 1er échantillon (avec ou sans équilibrage), selon le type de distance utilisée pour définir la proximité entre les probabilités d'inclusion de référence du 2nd échantillon et les probabilités approchées et, enfin, selon les valeurs du taux de recouvrement entre le 1er et le 2nd échantillon.

Dans un second temps, pour chaque tirage du 1er échantillon, on simulera des tirages du 2nd échantillon avec des probabilités conditionnelles de tirages adaptées en fonction du calcul des nouvelles probabilités finales. Des estimateurs d'un certain nombre de totaux de variables d'intérêt seront calculés en utilisant ces nouvelles probabilités d'inclusion. On étudiera empiriquement le biais et la précision de ces estimateurs, en comparant leurs résultats aux vraies valeurs connues dans la base de sondage.