PLAN DE SONDAGE DES ENQUÊTES GÉNÉRATION : UTILISATION D'UN CALAGE POUR SURÉCHANTILLONNER LES EXTENSIONS

Christophe BARRET, Mady CISSÉ, Christophe DZIKOWSKI

Céreq, Centre d'Études et de REcherches sur les Qualifications

christophe.barret@cereq.fr

Mots-clés : Échantillonnage, calage, extension

Résumé

Les enquêtes Génération du Céreq permettent d'étudier l'insertion professionnelle des jeunes à l'issue du système éducatif. Il s'agit d'une enquête sur la France entière et sur tous les niveaux et spécialités de formations. De nombreux acteurs publics partenaires sollicitent le Céreq pour mener des extensions d'échantillon sur leur population d'intérêt. Il s'agit par exemple de ministères qui souhaitent une précision accrue sur les formations qui relèvent de leur compétence ou bien des Régions qui souhaitent étudier l'insertion des jeunes sorties de formation sur leur territoire.

Cet article présente les différentes étapes de l'échantillonnage de l'enquête 2016 auprès de la Génération 2013. La méthodologie retenue permet de prendre en compte trois difficultés importantes pour la constitution de l'échantillon.
En effet, le Céreq constitue lui-même la base de sondage des élèves sortants du système éducatif en 2012-2013. Cette base de sondage présente un défaut de sous-couverture lié à la non-réponse de certains établissements et un défaut de sur-couverture lié à la présence dans la base d'individus hors champ de l'enquête. Enfin, le plan de sondage tient compte des demandes d'extension des partenaires. La difficulté principale provient du fait que certaines extensions se croisent. Une solution innovante, par calage sur marges sur les cibles d'extension, est proposée pour éviter d'aboutir à des échantillons dans lesquels le poids des intersections entre extensions soit trop important.

La méthodologie peut se décomposer en deux phases. La première phase consiste à effectuer le calcul des probabilités individuelles de tirage qui tiennent compte du taux de couverture de la base de sondage, des probabilités estimées d'être dans le champ de l'enquête et des probabilités anticipés de réponse. Dans un premier temps, les probabilités de tirage sont déterminées en l'absence d'extension pour atteindre le nombre de questionnaires financés par le Céreq. Pour satisfaire les besoins des extensions, des coefficients de dilatation sont calculés pour atteindre les cibles des extensions et appliqués pour obtenir les probabilités de tirage finales. La deuxième phase consiste à effectuer un unique tirage équilibré de l'échantillon sur la base de ces dernières probabilités de tirage.

En l'absence d'intersection entre les extensions, le calcul des coefficients de dilatation pourraient se faire séquentiellement extension par extension. En revanche, en présence d'intersections entre les extensions, le traitement séquentiel des calculs des coefficients de dilatation posent plusieurs problèmes. Le calcul serait en effet dépendant de l'ordre dans lesquels seraient considérés chaque extension et les intersections seraient de plus artificiellement sur-échantillonnées. La méthode de calage pour calculer les coefficients de dilatation proposée ici permet à la fois de corriger le problème de dépendance à l'ordre et le problème de sur-échantillonnage des intersections.

Cette stratégie d'échantillonnage, en un unique tirage, sera comparée à une stratégie d'échantillonnage plus classique qui consiste à réaliser un premier tirage de l'échantillon national (sans extension) puis un second tirage dans le complémentaire du premier échantillon pour satisfaire les besoins des extensions. Les avantages et inconvénients des deux méthodes seront présentés.

Bibliographie

[1] Jean-Claude Deville and Yves Tillé. « Efficient balanced sampling: the cube method». *Biometrika*, vol 91, n°4, pp 893-912, 2004.
[2] Deville, J.-C., Särndal, C.-E. et Sautory, O. . « Generalized raking procedures in survey sampling ». *Journal of the American Statistical Association*, vol 88, n°423, pp. 1013-1020, 1993.