CONSTRUCTION D'UN ÉCHANTILLON

DISPERSÉ GEOGRAPHIQUEMENT

Marc CHRISTINE (*)

(*) INSEE, DSDS, Unité Méthodes Statistiques

Introduction

Dans la pratique des tirages d'échantillons d'enquêtes ménages tels que les effectue l'Insee, il est d'usage de chercher à concentrer l'échantillon sur certaines zones géographiques de manière à minimiser le déplacement des enquêteurs lors des enquêtes réalisées en face à face.

Dans le même temps, on a souvent le souhait d'assurer une certaine dispersion des points d'enquête, partant du présupposé intuitif qu'une bonne dispersion géographique de l'échantillon permettra d'assurer une couverture adéquate de la diversité du territoire sur lequel on échantillonne.

La réconciliation de ces deux souhaits apparemment contradictoires conduit à la mise en place du système *d'unités primaires* qui sert de substrat aux échantillons d'enquêtes ménages réalisées par l'Insee : les échantillons d'enquête sont concentrés à l'intérieur des unités primaires sélectionnées mais ces dernières sont-elles mêmes tirées avec l'optique de constituer un bon modèle « réduit » de la population de référence et de ses caractéristiques démographiques, économiques et géographiques.

Dans ce papier, on s'intéresse plus particulièrement à la prise en compte de critères géographiques, voire géométriques, en essayant de résoudre le problème de la construction et de la sélection d'un échantillon d'unités en général localisées (communes, adresses, logements...), *assurant une dispersion géographique adéquate des unités tirées.*

On verra que ce problème peut se résoudre en utilisant la technique de l'équilibrage pour la sélection d'échantillons, mise en œuvre grâce à la méthode du CUBE [1].

1. Cadre de référence et notations

Considérons une population de référence, notée U, de taille N, dont les unités sont notées i.

On sélectionne un échantillon sans remise s ; le plan de sondage est quelconque, de taille fixe ou non. Les unités sont sélectionnées avec des probabilités d'inclusion Π_i. Les poids de sondage, qui serviront dans la suite, sont notés : $d_i = \frac{1}{\Pi_i}$. On suppose définies sur la population différentes variables d'intérêt dont on connaît les totaux.

2. Cas de l'ajustement des moyennes

Supposons que l'on souhaite assurer l'égalité entre la moyenne de la variable d'intérêt Y dans l'échantillon, pondérée par les poids de sondage, et la moyenne empirique de cette variable sur la population.

Cette égalité s'écrit :

$$\frac{\sum_{i\in s} d_i Y_i}{\sum_{i\in s} d_i} = \frac{1}{N}\sum_{i\in U} Y_i \qquad \text{(E)}$$

En notant $\overline{Y}$ la moyenne empirique « vraie » de Y (sur la population), cette égalité s'écrit aussi :

$$\sum_{i\in s} d_i Y_i = \overline{Y}\sum_{i\in s} d_i$$

soit

$$\boxed{\sum_{i\in s} d_i \left(Y_i - \overline{Y}\right) = 0}$$

Cette dernière égalité s'interprète comme une *équation d'équilibrage sur le total de la variable* $Y_i - \overline{Y}$ qui est en effet égal à 0 par définition de $\overline{Y}$. Par suite, il sera possible de tirer un échantillon respectant cette condition au moyen de la méthode du CUBE avec condition d'équilibrage relative aux variables $Y_i - \overline{Y}$.

Problème pratique

Le vrai total de la nouvelle variable d'intérêt $Y_i - \overline{Y}$ vaut 0 et l'algorithme du CUBE peut devenir instable en présence de cette particularité.

On peut toutefois contourner cette difficulté de la manière suivante :

L'égalité initiale (E) s'écrit en effet :

$$\sum_{i\in s} d_i Y_i = \frac{\sum_{i\in s} d_i}{N}\sum_{i\in U} Y_i$$

Par suite, si l'on fait en sorte de satisfaire simultanément les deux conditions :

$$\begin{cases} \sum_{i\in s} d_i = N \\ \sum_{i\in s} d_i Y_i = \sum_{i\in U} Y_i \end{cases}$$

alors l'égalité initiale sera vérifiée.

Or, ces deux conditions que l'on a introduites s'interprètent, respectivement, comme une condition d'équilibrage sur la variable 1 et sur la variable Y_i. ***On en conclut qu'on peut finalement satisfaire l'équation (E) au moyen d'un tirage équilibré avec <u>deux</u> conditions d'équilibrage, l'une sur 1 et l'autre sur*** Y_i ***(au lieu d'une seule variable dans l'approche précédente).***

On peut d'ailleurs remarquer de façon évidente que, parmi les trois conditions suivantes :

- équilibrage sur le total d'une variable Y
- équilibrage sur la taille de la population
- identité des moyenne de Y sur l'échantillon et sur la population,

deux d'entre elles entraînent la 3ème.

Nota : naturellement, le tirage de l'échantillon peut être fait en introduisant d'autres conditions d'équilibrage que celle résultant de la contrainte d'identité des moyennes imposée pour une variable d'intérêt donnée Y.

3. Conformité des variances

Pour améliorer la qualité de l'échantillon, assurer une meilleure similitude avec la population mère et diminuer la variance d'échantillonnage pour l'estimation de différentes variables d'intérêt, on peut aussi souhaiter réaliser un tirage qui assure l'identité entre la variance « vraie » d'une variable donnée Y dans la population, et sa variance dans l'échantillon (toujours pondérée par les poids de sondage).

Exemple : si l'on veut qu'un échantillon donne une bonne image de la dispersion des revenus, on sera amené à imposer que la variance des revenus dans l'échantillon soit identique à celle observée dans la population.

L'égalité à assurer s'écrit alors, dans ce cas :

$$\frac{\sum_{i\in s} d_i \left(Y_i - \bar{y}\right)^2}{\sum_{i\in s} d_i} = \frac{1}{N}\sum_{i\in U}\left(Y_i - \bar{Y}\right)^2$$

où $\bar{Y}$ est la moyenne de la variable Y dans la population et $\bar{y}$ la moyenne pondérée dans l'échantillon, soit :

$$\bar{y} = \frac{\sum_{i\in s} d_i Y_i}{\sum_{i\in s} d_i}$$

Or, on a l'égalité classique :

$$\frac{\sum_{i\in s} d_i \left(Y_i - \bar{y}\right)^2}{\sum_{i\in s} d_i} = \frac{\sum_{i\in s} d_i Y_i^2}{\sum_{i\in s} d_i} - \bar{y}^2$$

Par suite, la condition à satisfaire devient :

$$\frac{\sum_{i\in s} d_i Y_i^2}{\sum_{i\in s} d_i} = \frac{1}{N}\sum_{i\in U}\left(Y_i - \bar{Y}\right)^2 + \bar{y}^2$$

$$= \frac{1}{N}\sum_{i\in U} Y_i^2 - \bar{Y}^2 + \bar{y}^2$$

On voit que cette condition sera satisfaite dès lors que le tirage de l'échantillon assure simultanément les égalités :

$$\begin{cases} \bar{y} = \bar{Y} \\ \dfrac{\sum_{i\in s} d_i Y_i^2}{\sum_{i\in s} d_i} = \dfrac{1}{N}\sum_{i\in U} Y_i^2 \end{cases}$$

Ces deux égalités s'interprètent chacune comme une contrainte d'identité entre des moyennes pondérées dans l'échantillon et les moyennes vraies correspondantes dans la population, respectivement pour les variables Y_i et Y_i^2.

Moyennant donc un tirage utilisant deux conditions d'équilibrage (plus d'autres éventuelles sur d'autres variables), on parvient ainsi à assurer l'identité entre les variances d'une variable donnée, dans l'échantillon et dans la population, grâce à la méthode exposée au § 2.

Il est équivalent d'imposer <u>trois</u> conditions d'équilibrage : sur le total de la constante 1, de la variable Y_i et de la variable Y_i^2, en utilisant la remarque pratique du § 2.

4. Application à la construction d'un échantillon dispersé géographiquement

Les techniques exposées ci-dessus vont permettre de résoudre le problème initialement posé.

Considérons une population U dont les unités peuvent être repérées par une position géographique. Concrètement, chaque unité i sera supposée associée à un point géographique A_i dans un plan euclidien.

La dispersion géographique des unités de la population pourra être mesurée par la quantité $\frac{1}{N}\sum_{i\in U} GA_i^2$ où G est l'isobarycentre de tous les points A_i. Ceci correspond à la définition usuelle de l'inertie d'un nuage de points par rapport à un point de référence.

Si O est un point fixe du plan, on a évidemment la relation usuelle relative à la fonction scalaire de LEIBNIZ :

$$\frac{1}{N}\sum_{i\in U} OA_i^2 = OG^2 + \frac{1}{N}\sum_{i\in U} GA_i^2$$

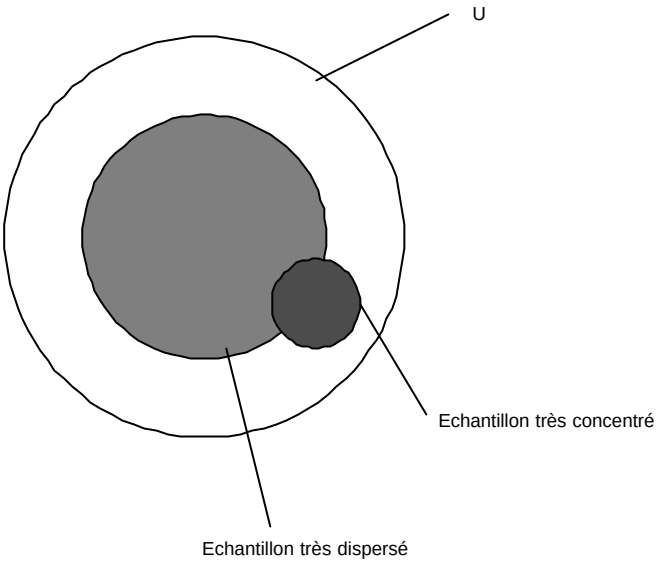
Cette dispersion peut aussi s'exprimer en fonction de la somme des carrés des distances $A_i A_j^2$ pour tous les couples de points (A_i, A_j) correspondant aux unités de la population (théorème de HUYGHENS).

De même, la dispersion géographique des unités de l'échantillon sera définie par la quantité : $\frac{\sum_{i\in s} d_i\, gA_i^2}{\sum_{i\in s} d_i}$, où g est le centre de gravité des points A_i de l'échantillon, pondérés par les poids de sondage d_i.

Dans ces conditions, on cherchera à sélectionner des échantillons assurant la contrainte : $\frac{\sum_{i\in s} d_i\, gA_i^2}{\sum_{i\in s} d_i} = \frac{1}{N}\sum_{i\in U} GA_i^2$, c'est-à-dire tels que les deux dispersions, dans la population et dans l'échantillon, des points représentant les unités soient identiques. Intuitivement, si cette condition est réalisée, l'échantillon ne sera ni plus ni moins « concentré » que la population.

Le fait d'assurer l'identité des moyennes des carrés des distances et non de leurs totaux permet évidemment de tenir compte de la «réduction d'échelle » opérée par le passage de la population à l'échantillon tout en maintenant une similitude des « structures ». Les notions de « concentration » ou de «dispersion » de l'échantillon peuvent ainsi s'entendre indépendamment de leur taille, et deux échantillons de tailles différentes peuvent être comparés intrinsèquement selon cette caractéristique de dispersion.

On peut illustrer ceci par un schéma :



Or :

$$\sum_{i\in s} d_i\, GA_i^2 = \sum_{i\in s} d_i \left(\vec{Gg} + \vec{gA_i}\right)^2$$
$$= (Gg)^2 \left(\sum_{i\in s} d_i\right) + 2\vec{Gg}\cdot\left(\sum_{i\in s} d_i\, \vec{gA_i}\right) + \sum_{i\in s} d_i\, gA_i^2$$

Or : $\sum_{i\in s} d_i\, \vec{gA_i} = \vec{O}$ puisque g est le barycentre des points A_i pour $i \in s$, affectés des poids d_i.

On en déduit que :

$$\boxed{\frac{\sum_{i\in s} d_i\, gA_i^2}{\sum_{i\in s} d_i} = \frac{\sum_{i\in s} d_i\, GA_i^2}{\sum_{i\in s} d_i} - Gg^2}$$

Or, on sait réaliser le tirage d'un échantillon équilibré satisfaisant la condition :

$$\frac{\sum_{i\in s} d_i\, GA_i^2}{\sum_{i\in s} d_i} = \frac{1}{N}\sum_{i\in U} GA_i^2$$

Il s'agit en effet d'un tirage dans lequel on astreint la moyenne pondérée, sur l'échantillon, des variables $Z_i = GA_i^2$ à être identique à la vraie moyenne empirique de cette même variable, sur la population. On est donc dans le cadre d'application de la démarche décrite au § 2.

Par suite, si l'on veut réaliser le tirage d'un échantillon astreint à la condition assurant l'identité des dispersions moyennes, dans l'échantillon et dans la population, des points sélectionnés :

$$\frac{\sum_{i\in s} d_i \, gA_i^2}{\sum_{i\in s} d_i} = \frac{1}{N}\sum_{i\in U} GA_i^2$$

il suffit d'assurer les deux égalités :

$$\begin{cases} gG^2 = O \\ \dfrac{\sum_{i\in s} d_i \, GA_i^2}{\sum_{i\in s} d_i} = \dfrac{1}{N}\sum_{i\in U} GA_i^2 \end{cases}$$

La première ce ces égalités signifie que le centre de gravité g des points sélectionnés dans l'échantillon, pondérés par les poids d_i, est identique à l'isobarycentre de l'ensemble des points de la population.

Si O est un point fixe (origine du repère orthonormé du plan, par exemple), on a :

$\vec{Og} = \dfrac{\sum_{i\in s} d_i \, \vec{OA_i}}{\sum_{i\in s} d_i}$ et $\vec{OG} = \dfrac{1}{N}\sum_{i\in U} \vec{OA_i}$.

L'égalité $g = G$ se traduit alors par les conditions :

$$\frac{\sum_{i\in s} d_i \, \vec{OA_i}}{\sum_{i\in s} d_i} = \frac{1}{N}\sum_{i\in U} \vec{OA_i}$$

Transcrite en passant en composantes sur un repère donné, cette égalité vectorielle est équivalente à deux égalités scalaires :

$$\begin{cases} \dfrac{\sum_{i\in s} d_i \, A_{Xi}}{\sum_{i\in s} d_i} = \dfrac{1}{N}\sum_{i\in U} A_{Xi} \\ \dfrac{\sum_{i\in s} d_i \, A_{Yi}}{\sum_{i\in s} d_i} = \dfrac{1}{N}\sum_{i\in U} A_{Yi} \end{cases}$$

où A_{Xi} et A_{Yi} sont les deux composantes (abscisse et ordonnée) du point A_i.

On obtient à nouveau des égalités qui s'interprètent comme des identités entre moyenne pondérée calculée sur l'échantillon et moyenne empirique sur la population de deux variables numériques A_X et A_Y, toujours justiciables de la technique du § 2.

$\rightarrow$ ***Au final, pour assurer l'identité des dispersions moyennes du nuage de points échantillonnés, pondérés par les poids de sondage*** d_i ***, et de la population tout entière (avec équipondération des unités), il suffit de réaliser un tirage équilibré assurant l'égalité des moyennes pondérées dans l'échantillon et des moyennes empiriques dans la population des trois variables :*** A_{Xi} **,** A_{Yi} ***et*** GA_i^2 ***, respectivement abscisse, ordonnée et carré de la distance à l'isobarycentre de l'ensemble des points dans la population.***

Ceci est possible dans le cadre de l'adaptation de la méthode du tirage équilibré décrite au § 2, au besoin en remplaçant les trois conditions précédentes par quatre conditions d'équilibrage sur des totaux, respectivement de la constante 1 et des variables A_{Xi} , A_{Yi} et GA_i^2 .

5. Application

Pour mettre en œuvre cette méthode, il faut bien entendu connaître les coordonnées des points géométriques correspondant aux différentes unités de la population, relativement à un repère orthonomé donné, de façon à pouvoir calculer explicitement les coordonnées de l'isobarycentre de ces points ainsi que la distance de chacun d'entre eux à cet isobarycentre.

Le système de coordonnées LAMBERT associées à chaque logement et disponibles dans les fichiers du RP permet de résoudre la question dans le cadre du tirage d'un échantillon de logements.

Cette méthode a été mise en œuvre pour le tirage des échantillons des enquêtes Budget de Famille et Emploi réalisées en 2005 dans les DOM. Néanmoins, dans le cas des DOM, on n'a pas pu disposer de l'information orthométrique détaillée au niveau de chaque logement. Par défaut, on a dû affecter à chacun des logements de la population les coordonnées de la mairie de la commune d'appartenance du logement et non celles du logement lui-même.

Bibliographie

[1] ROUSSEAU s. et TARDIEU F., la macro SAS CUBE d'échantillonnage équilibré, documentation de l'utilisateur, série des Documents de travail de méthodologie statistique (DSDS / UMS) n° 0402.