ÉCHANTILLONNAGES POUR LE CONTRÔLE DE QUALITÉ DU RECENSEMENT DE 1990

Jean-Claude Deville

Introduction digressive

Bien que cela puisse sembler paradoxal, le Recensement de la Population (RP) engendre un nombre considérable d'opérations de sondage au cours de sa réalisation.

L'opération la plus visible est la réalisation, parallèlement au recensement, d'une enquête spéciale sur les familles sur la base d'un échantillon aréolaire au 1/50. Pour les éditions précédentes, la méthodologie de ces enquêtes est décrite dans J.-C. Deville (1972) pour les enquêtes de 1954 et de 1962, et dans G. Desplanques (1981 et 1987) pour les enquêtes les plus récentes.

En 1990, l'Insee a, de plus, réalisé une enquête de vérification de l'exhaustivité du RP. Réalisée également sur un plan de sondage aréolaire, ses principaux résultats et sa méthodologie sont exposés dans N. Coeffic (1992).

L'exploitation du recensement fait elle-même appel à la technique des sondages. En effet, pour réduire les coûts et les délais de la publication, un quart seulement des bulletins sont codifiés entièrement, notamment en ce qui concerne la détermination de la catégorie socioprofessionnelle (CSP), de l'activité économique et des liens entre personnes d'un même ménage. L'échantillon codifié résulte d'une stratification par commune et par taille de ménage. Il est utilisé pour produire des résultats détaillés au niveau d'unités comptant au moins 5 000 habitants.

Cet échantillon au quart est précédé par la publication des résultats sur échantillon au 1/20 de la population.

L'utilité de ce dernier sondage est d'obtenir rapidement des résultats valides au niveau de la France entière et éventuellement au niveau des régions. Il est réalisé en tirant avec probabilités égales un district sur cinq, dont on codifie le quart des bulletins. Ainsi l'échantillon au 1/20 apparaît-il comme le premier cinquième de l'échantillon au quart.

Rappelons qu'un district est une unité de collecte du RP correspondant, en ville, à un pâté de maisons et, à la campagne, à un village ou une réunion de hameaux. Il peut comporter de quelques logements à plus de mille, et compter une population pouvant

varier de 0 à environ 2 000 habitants. La moyenne est de 150 logements pour 350 habitants environ.

En raison de cette dispersion de taille, le tirage d'un cinquième des districts pouvait produire un sondage très imprécis. On a pu établir (Chartier, 1979) qu'un sondage aléatoire simple de districts multipliait par un facteur variant de deux à cinq environ la variance d'un sondage aléatoire simple de logements. Une technique d'échantillonnage équilibré (Deville, *et Alii*, 1988) a permis d'atténuer très largement cet effet.

Contrôle de qualité

La technique des sondages a aussi été très utile dans le processus de contrôle de fabrication du RP.

Résumons le processus qui a été suivi. À mesure de l'achèvement de la collecte et de sa vérification, les différents bulletins du recensement, notamment les bulletins individuels (BI) et les feuilles de logement (FL), sont soigneusement comptés pour chaque district. Les données récapitulatives des districts sont saisies sur support informatique, tandis que les bulletins groupés dans une chemise de district, partent pour la saisie.

Celle-ci se réalise en deux étapes :

- dans la première étape, dite de l'"exhaustif", des ensembles de districts groupant environ 100 000 logements sont constitués. Ce sont les unités de traitement (UT). Chaque UT est saisie par une entreprise à façon pour le compte de l'Insee. L'Insee, le "client" en termes de théorie du contrôle, vérifie la qualité du travail de chaque façonnier en contrôlant par sondage un certain nombre de bulletins dans chaque UT. Les principes permettant d'optimiser le coût de ce contrôle sont exposés dans cet article ;

- la seconde étape d'élaboration des données est dite opération COLIBRI (pour Codification en Ligne des Bulletins du Recensement des Individus). Recevant des bulletins toujours groupés en districts, les opérateurs et opératrices des Directions Régionales (DR) de l'Insee, procédaient à leur codification pour constituer le sondage au quart.

Physiquement, chaque opérateur(trice) travaille devant un écran qui lui indique l'identifiant du prochain logement à inclure dans l'échantillon au quart dont il doit codifier tous les BI.

Le contrôle de la qualité de la codification est également réalisé par sondage. L'unité de contrôle est l'ensemble du travail réalisé en une semaine dans une Direction Régionale. L'opération dure un peu plus d'un an dans les 22 Directions Régionales soit plus de mille sondages. L'unité à contrôler est le ménage (c'est-à-dire l'ensemble des BI d'un ménage tiré pour figurer dans l'échantillon de contrôle). L'objectif est d'estimer la proportion de bulletins comportant une erreur. Pour cela, on détecte automatiquement

ceux pour lesquels apparaît une divergence entre les deux codifications. Une opération de réconciliation permet de chiffrer le nombre d'erreurs. La modalité pratique et les enseignements tirés de ces contrôles sont détaillés dans G. Badeyan (1992).

Données générales sur les contrôles exhaustif et Colibri : la fonction de coût

Il se trouve que dans les deux problèmes la fonction de coût est la même, exprimée en temps de travail. Le contrôle consiste à aller chercher un district là où il est rangé, à l'amener au poste de travail de l'opérateur (trice) chargé du contrôle. Celui-ci (ou celle-ci) recherche les bulletins à contrôler dont il (elle) dispose des identifiants, en saisit leur codification. Les bulletins sont remis à leur place dans la chemise de district. Celle-ci est ramenée dans l'étagère où elle est habituellement stockée.

Globalement, donc, le temps passé au contrôle d'un district se décompose en un temps de manipulation de la chemise de district, qui ne dépend pas du district (taille ou localisation), et un temps de traitement de chaque bulletin à contrôler dans celui-ci. Ce temps de traitement ne dépend que du type de bulletin (BI ou FL), mais est considéré, à type de bulletin donné, comme indépendant du bulletin.

La fonction de coût du contrôle peut donc être approximée par une formule simple. Supposons que le contrôle porte sur m districts et regroupe n_g bulletins du type g au total. Si C_o est le coût en temps associé à la manipulation d'une chemise de district et C_g le coût en temps associé au contrôle d'un bulletin de type g, le coût du contrôle vaudra :

$$C_t = m\, C_o + \sum_g n_g\, c_g$$

Cette expression est une généralisation naturelle d'une fonction de coût couramment utilisée décomposant les coûts d'une collecte en un coût d'approche C_o et un coût unitaire C_1 (voir par exemple Desabie (1965) ou Cochran (1977)).

Le contrôle de l'exhaustif

Première forme : on ne contrôle que les bulletins individuels

Les districts k d'une UT (notée U) comportent chacun un nombre connu N_k de BI. Parmi ceux-ci D_k comportent une "erreur" (assimilée à une différence entre ce qui est codé à l'Insee et ce qui a été codé chez le façonnier). Le but est donc d'estimer :

$$P = \sum_U D_k / \sum_U N_k$$

Le sondage consistera à tirer un échantillon s de districts avec des probabilités d'inclusion π_k au premier ordre et π_{kl} au second ordre à déterminer. Ensuite, si le district k est tiré dans s on vérifiera n_k BI tirés par sondage aléatoire simple sans remise (SASSR). Soit d_k le nombre de BI erronés qu'on relèvera.

L'estimateur $\hat{P}_k$ de $P_k = D_k/N_k$ sera $\hat{P}_k = d_k/n_k$ et $\hat{D}_k = N_k \hat{P}_k$ estimera D_k sans biais. L'estimateur de P sera :

$$\hat{P} = \frac{\Sigma_s \hat{D}_k/\pi_k}{\Sigma_s N_k/\pi_k} \qquad (4-1)$$

C'est le ratio des estimateurs sans biais de D et de N, le nombre total de bulletins. Bien que ce nombre soit connu, il est bien évident que l'estimateur (4-1) est plus précis que $\frac{1}{N}\sum_s \hat{D}_k / \pi_k$.

On a :

$$VAR(\hat{P}) = E\,Var(\hat{P}/s) + Var(E\,\hat{P}/s) \qquad (4-2)$$

Or :
$$Var(\hat{P}/s) = \hat{N}^{-2} \sum_s \frac{N_k^2}{\pi_k^2} \frac{P_k(1-P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right)$$
$$\text{avec } \hat{N} = \sum_s N_k / \pi_k$$

D'où :
$$E\,Var(\hat{P}/s) \approx N^{-2} \sum_U \frac{N_k^2}{\pi_k} \frac{P_k(1-P_k)N_k}{N_k - 1}\left[\frac{1}{n_k} - \frac{1}{N_k}\right] \qquad (4-3)$$

Par ailleurs :
$$E(\hat{P}/s) = \frac{\sum_s D_k / \pi_k}{\sum_s N_k / \pi_k}$$

La variance de cette quantité s'obtient par linéarisation en introduisant la variable $Z_k = D_k - PN_k = N_k (P_k - P)$.

On obtient : $Var\, E\,(\hat{P}/s) \approx N^{-2}\, Var \left(\sum_s \frac{Z_k}{\pi_k} \right)$

Soit, compte tenu de ce que $\sum_U Z_k = 0$:

$$Var\, E\,(\hat{P}/s) = N^{-2} \left(\sum_k \frac{Z_k^2}{\pi_k} + \sum \sum_{k \neq l} \frac{z_k\, z_l}{\pi_k\, \pi_l}\, \pi_{kl} \right) \qquad (4-4)$$

La somme des quantités (4-3) et (4-4) nous donne la variance de l'estimateur (4-1).

Introduction d'un modèle

La variance de $\hat{P}$ est difficile à manipuler et, de plus, contient des paramètres inconnus. On se tire de la difficulté en faisant de nécessaires hypothèses qui se traduisent par un modèle de superpopulation. On verra plus loin que les paramètres de ce modèle sont susceptibles d'être estimés à partir d'un essai préliminaire de recensement portant sur une toute petite partie du territoire. On note E_ξ l'espérance sous le modèle (resp Var_ξ pour la variance) dont tous les aléas sont supposés indépendants du processus d'échantillonnage.

Le modèle suit les spécifications suivantes :

a) D_k suit une loi binomiale $(N_k\ ,\ p_k)$. P_k est donc, *sous le modèle*, un estimateur de p_k.

b) p_k est lui même aléatoire. On suppose les p_k indépendantes et de même loi avec :

$$E_\xi\ p_k = P$$

$$Var_\xi\ p_k = \sigma^2$$

pour tout k, quelle que soit, en particulier, la valeur de N_k.

En conditionnant, dans le modèle, par les p_k on a évidemment :

$$E_\xi\ (D_k \mid p_k) = N_k\, p_k$$

$$Var_\xi\ D_k = N_k\, p_k\, (1 - p_k)$$

La *variance anticipée* de $\hat{P}$ est la quantité $E_\xi\ Var\, \hat{P}$. C'est à elle que nous allons nous intéresser désormais. Pour l'évaluer on remarque que :

$$a - \quad E_\xi\ (P_k - P)^2 = E_\xi\ (E_\xi\ P_k + p_k - p_k - P)^2 \mid p_k) = \frac{P\,(1-P) - \sigma^2}{N_k} + \sigma^2$$

$$b - \quad E_\xi\ P_k\ (1-P_k) = E_\xi\ (E_\xi\ ((p_k - P_k)^2 \mid p_k)) = E_\xi\ p_k\ (1-p_k)\ \frac{N_k - 1}{N_k}$$

$$= (P\ (1-P) - \sigma^2)\ \frac{N_k - 1}{N_k}$$

$c - \quad E_\xi\ Z_k\ Z_l = 0$ à cause de l'indépendance des Z_k et des Z_l, ce qui nous débarrasse d'un terme bien encombrant en même temps que des π_{kl}.

En recollant tous les morceaux de (4-3) et (4-4) un petit miracle algébrique se produit et nous avons l'expression :

$$E_\xi\ Var\ \hat{P} \approx N^{-2} \sum_U \frac{N_k^2}{\pi_k} \left(\sigma^2 + \frac{\tau^2}{n_k}\right) \qquad (5\text{-}1)$$

avec $\tau^2 = P\ (1-P) - \sigma^2 \geq 0$

Remarque :

Le miracle algébrique s'explique bien si on ne cherche pas à obtenir la variance sous le plan de sondage uniquement. Elle est d'ailleurs la conséquence d'un modèle un peu plus général que celui que nous avons posé.

Supposons que nous voulions estimer le total $N\,\overline{Y} = \sum_U Y_i$ d'une variable Y et que pour cela nous réalisions un tirage à deux degrés : un premier degré où des unités primaires *(UP)* k sont tirées avec des probabilités π_k, un second où n_k unités finales sont tirées par sondage aléatoire simple.

Nous posons un modèle où :

$$Y_i = \overline{Y} + \alpha_k + \varepsilon_i$$

avec α_k variable liée à l'*UP* d'indice k. Les α_k sont indépendantes de même loi d'espérance nulle de variance σ^2.

Les ε_i sont également indépendantes centrées de variance égale à τ^2. Avec $\pi_i^* = \pi_k\ n_k / N_k$ (N_k taille de l'*UP* numéro k), l'estimateur de Horvitz-Thompson du total vaut $\hat{Y} = \Sigma\ \hat{Y}_i / \pi_i^*$ la somme étant étendue à l'échantillon.

Sous le modèle, et conditionnellement à l'échantillon on a :

$$Var_\xi \hat{Y} = \sum_s \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right)$$

L'espérance sous le plan de cette expression redonne la formule (5-1).

Optimisation du sondage

La variance maximum de $\hat{P}$ est fixée par les critères retenus pour le contrôle de qualité. Le sondage étant répété pour chacune des unités de traitement il est tout à fait naturel de chercher à minimiser l'espérance du coût du sondage donné en (3-1) soit :

$$E \sum_s (C_o + n_k\ C_1) = \sum_U \pi_k\ (C_o + n_k\ C_1) \qquad (6\text{ - }1)$$

avec C_o temps de manipulation des chemises de district et C_1 temps requis pour la codification d'un BI.

Le problème d'optimisation s'écrit donc :

$$\text{Minimiser} \sum_U \pi_k\ (C_o + n_k\ C_1)$$

sous les contraintes :

$$N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \leq V_o$$

$$\text{et pour tout } k,\ n_k \leq N_k.$$

Associons un multiplicateur de Lagrange λ à la première contrainte - qui sera évidemment saturée - et des multiplicateurs μ_k aux autres. On obtient les solutions :

$$C_o + n_k\ C_1 = \lambda \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \qquad (6\text{ - }2)$$

$$\text{et, pour tout k : } C_1\ \pi_k = \lambda\ \frac{N_k^2}{\pi_k} \cdot \frac{\tau^2}{n_k^2} + \mu_k \qquad (6\text{ - }3)$$

avec $\mu_k = 0$ si $n_k < N_k$ et μ_k >~0~roman size 10 si~ n sub k~=~N sub k

Pour tous les districts où $\mu_k = 0$ (les plus gros) on obtient :

$$n_k = \frac{\tau}{\sigma}\left(\frac{C_o}{C_1}\right)^{1/2} = n^* \qquad (6\text{ - }4)$$

Chaque district reçoit donc la même allocation, ce qui correspond à l'idée qu'on a besoin de la même précision dans chacun d'eux. Retournons à l'équation (6-3). On constate alors que, toujours pour ces districts, les probabilités d'inclusion π_k doivent être proportionnelles aux tailles N_k soit :

$$\pi_k = \lambda^{1/2}\ C_1^{-1/2}\ \frac{\tau}{n^*}\ N_k \qquad (6-5)$$

C'est la justification habituelle d'un sondage autopondéré avec un premier degré tiré avec des probabilités proportionnelles à une mesure de taille.

Comme n_k ne dépend pas de N_k on ne pourra avoir $n_k = N_k$ *et* $\mu_k > 0$ que si $N_k \leq n^*$. L'équation (6-1) nous permet alors d'obtenir les probabilités d'inclusion à un facteur près :

$$n_k = \lambda^{1/2}\ N_k \left(\frac{\sigma^2 + \tau^2/N_k}{C_o + C_1 N_k}\right) = \lambda^{1/2}\ N_k^{1/2} \left(\frac{N_k \sigma^2 + \tau^2}{N_k\ C_1 + C_o}\right)^{1/2} \qquad (6-6)$$

Les relations (6-5), valide si N_k >=~n sup * et (6-6) valide si $N_k \leq n^*$, établissent que π_k est proportionnelle à une variable connue $T_k = f(N_k)$ dont le graphique est donné à la figure 1.

Pour spécifier entièrement le sondage il reste à trouver le nombre d'unités primaires à tirer. Or, $T = \sum_U T_k$ est aussi une quantité connue.

En se restreignant à un échantillonnage de taille fixe on aura donc $\pi_k = m\ T_k/T$. On trouve m en portant cette valeur dans la contrainte de variance soit :

$$N^2\ V_o\ m = T \sum_U \frac{N_k^2}{T_k}\ (\sigma^2 + \tau^2/n_k)$$

Si, en première approximation, on prend $T_k = N_k$, on obtient la formule simplifiée :

$$m\ V_o = \sigma^2 + \tau^2/n^*$$

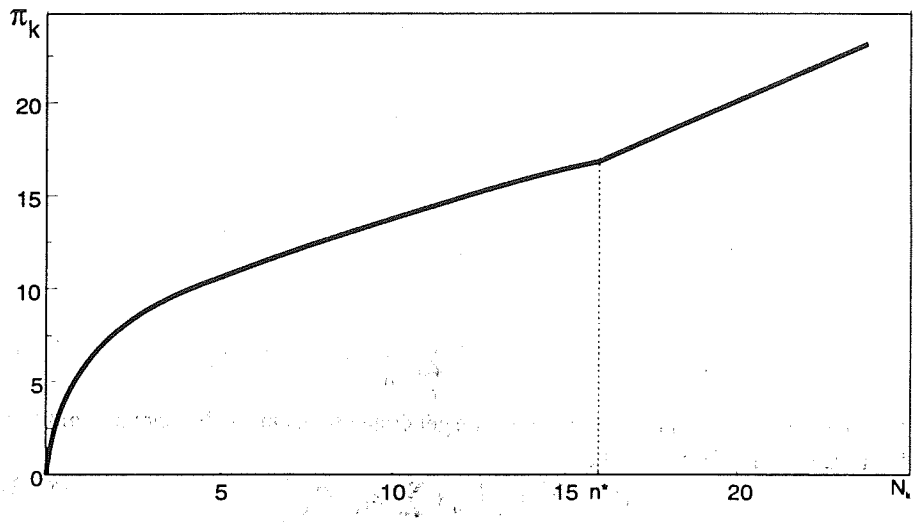
Les données recueillies sur le test de recensement ont permis de retenir les valeurs numériques approximatives suivantes :

$$\tau^2 \approx P^2 \approx 14 \,.\, 10^{-4}$$

$$\sigma^2 \approx P \approx 4 \,.\, 10^{-2}$$

D'autre part, des mesures faites dans les ateliers ont permis d'estimer à 5 minutes le temps de manipulation d'une chemise de district et à 30 secondes le temps de saisie d'un BI. Le rapport C_o/C_1 vaut donc environ 10 d'où on tire $n^* = 16$ puis $m = 40$.
Le contrôle porte donc sur 40 districts dont on extrait au total 640 bulletins.

Figure 1 : graphique de π_k en fonction de N_k



Le contrôle de l'exhaustif, deuxième forme

Ou comment optimiser un sondage à deux degrés où les unités primaires sont elles mêmes stratifiées.

La dure réalité des choses nous amène à compliquer un peu le problème. En fait, il s'agit de contrôler deux types de documents : les bulletins individuels (BI) et les feuilles de logement (FL). On avait été conduit à négliger ces dernières, en première approximation, parce qu'elles sont moins susceptibles de recéler des erreurs et que leur temps de codification est plus court (la moitié environ) que celui nécessaire pour un BI. Toutefois, dans certains districts, par exemple dans les communes très touristiques, on trouve une forte majorité de résidences secondaires et donc beaucoup de FL pour très peu de BI. Cette situation demande une étude particulière.

Il s'avère qu'elle correspond, de plus, à un problème assez général qui est le suivant.

Pour chaque unité primaire (ici les districts d'une UT) on connaît les effectifs N_{kg} d'unités secondaires appartenant à G groupes (ici g = 1, 2) selon qu'on s'intéresse aux BI ou aux FL). La "population" de l'UP numéro k vaut $N_{i+} = \sum_g N_{ig}$. celle du groupe g vaut $N_{+g} = \sum_k N_{k,g}$. Comme dans ce qui précède on cherche avec quelle probabilité

d'inclusion π_k échantillonner l'UP numéro k, le nombre d'UP à tirer et l'allocation n_{kg} de l'échantillon parmi les différents groupes dans l'UP k, sachant que ces n_{kg} unités sont tirées par un SASSR parmi les N_{kg} unités tirables.

Optimisation sous modèle

On postule, dans chacun des groupes, un modèle identique à celui formulé à la section précédente (ou sous une forme plus générale dans la remarque qui la termine).

Pour $g = 1$ à G on aura donc :

$$V_g = E_\xi \; Var(\hat{P}_g) = N_{+g}^{-2} \sum_U \frac{N_{kg}^2}{\pi_k} (\sigma_g^2 + \tau_g^2 / n_{kg}) \qquad (8 - 1)$$

La fonction de coût est donnée par la forme générale (3-1). On va chercher à minimiser l'espérance du coût de sondage :

$$C_T = \sum_U \pi_k \left(c_o + \sum_g n_{kg} \; C_g \right) \qquad (8 - 2)$$

sous les contraintes $V_g \leq \vartheta_g$ où les quantités ϑ_g sont fixées de façon extérieure, par exemple par la qualité des données qu'on veut obtenir et la rigueur du contrôle.

Sous cette forme, le problème peut s'avérer assez complexe. Nous allons écrire un "Lagrangien" général :

$$L = \lambda \, C_T + \sum_g \lambda_g \, V_g$$

Le problème posé fixe $\lambda = 1$ et les λ_g sont des multiplicateurs à déterminer. Une variante simple consiste à fixer les λ_g : on désire alors minimiser une combinaison linéaire donnée des variances sous une contrainte de coût. Dans toutes les hypothèses, on obtient par dérivation par rapport aux n_{kg} (considérées comme des variables réelles) :

$$\lambda \, \pi_k^2 \, C_g = \lambda_g \, N_{+g}^{-2} \, N_{kg}^2 \, \tau_g^2 / n_{kg}^2 \qquad (8 - 3)$$

Les π_k étant, pour l'instant, destinées à être connues à un facteur près, on peut écrire :

$$\pi_k \; n_{kg} = \left(\frac{\lambda_g}{C_g} \right)^{1/2} \tau_g \frac{N_{kg}}{N_{+g}} \qquad (8 - 4)$$

Par sommation sur k on en déduit que :

$$E\, n_{+g} = \sum_U \pi_k\, n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g \qquad (8\text{ - }5)$$

La taille totale de l'échantillon dans chaque groupe est donc directement liée au multiplicateur λ_g.

La dérivation du Lagrangien par rapport aux π_k nous donne de nouvelles relations qui se simplifient miraculeusement si on utilise aussi (8-4). On obtient :

$$C_o = \sum_g C_g \left(\frac{\sigma_g}{\tau_g}\right)^2 n_{kg}^2 \qquad (8\text{ - }6)$$

où encore, si on introduit les nombres

$$n_g^* = \left(\frac{C_o}{C_g}\right)^{1/2} \frac{\tau_g}{\sigma_g}, \text{ on écrit :}$$

$$\sum_g \left(\frac{n_{kg}}{n_g^*}\right)^2 = 1 \qquad (8\text{ - }7)$$

Les n_g^* sont les nombres d'unités secondaires à tirer par *UP* s'il n'y avait qu'un seul groupe ; n_{kg} sera toujours inférieur à n_g^*.

De (8-4), (8-5) et (8-7) on tire les relations :

$$\pi_k^2 = \frac{1}{C_o} \sum_g \lambda_g\, \tau_g^2 \left(\frac{N_{kg}}{N_{+g}}\right)^2$$

Ainsi, les π_k sont proportionnelles aux quantités T_k telles que $T_k^2 = \sum_g \lambda_g\, \tau_g^2 \frac{N_{kg}^2}{N_{+g}^2}$ qui apparaissent comme la mesure de taille adéquate. Les relations (8-4) montrent que, à k fixé, les n_{kg} sont proportionnelles à $n_g^*\, \lambda_g^{1/2}\, \sigma_g \frac{N_{kg}}{N_{+g}}$, ce qui, compte tenu de (8-7) conduit à :

$$n_{kg} = n_g^*\, \lambda_g^{1/2}\, \sigma_g \frac{N_{kg}}{N_{+g}}\, T_k^{-1}$$

Des solutions possibles au problème précédent

a) Si les λ_g étaient connus, c'est-à-dire si on minimisait $\sum_g \lambda_g\, V_g$ sous une contrainte de coût, alors (8-8) nous permettrait de calculer les T_k.

En reportant $\pi_k = m\,T_k/T$ $\quad (T = \sum_U T_k$, *m nombre de districts à tirer*)

dans la contrainte de budget $C_T \leq C_T^*$, on trouve :

$$C_T^* = \frac{m}{T}\left(C_o \sum_U T_k + \sum_g C_g\, n_g^*\, \lambda_g^{1/2}\, \sigma_g\, \frac{N_{kg}}{N_{+g}}\right)$$

$$\text{soit : } m = C_T^* \Big/ \left(C_o + \sum_g n_g^* \frac{\lambda_g^{1/2}\,\sigma_g}{T}\right)$$

Si un seul des λ_g est différent de zéro, on trouve avec satisfaction le résultat donné à la fin de la section 6.

b) Le problème initial (*min* C_T/*sous* $V_g \leq \mathcal{V}_g$) se résoud assez facilement dans deux cas particuliers :

b1 - Dispersion maximale **des groupes. Pour toute** UP_k, **on a** $N_{kg} = N_k$ **pour un certain** k. **Le problème est décomposé en** G **problèmes distincts, chacun d'eux étant du type étudié aux sections 4, 5 et 6.**

b2 - Dispersion minimale **: la répartition est la même dans toutes les** *UP* **; autrement dit on a pour tout** k **et** g

$$N_{kg} = N_{k+}\,\frac{N_{+g}}{N}\left(\text{avec } N = \sum_g N_{+g}\right)$$

T_k est alors proportionnelle à N_{k+}, et les n_{kg} sont des quantités $n_g^*\, u_g$ indépendantes de k.

Avec $\pi_k = mN_{k+}/N$, on obtient en écrivant - $Vg = \mathcal{V}_g$:

$$m\,\mathcal{V}_g = \sigma_g^2 + \tau_g^2 / n_g^*\, u_g$$

$$\text{soit : } m = \frac{\sigma_g^2}{V_g} + u_g^{-1}\,\frac{\tau_g^2}{n_g^*\, V_g}$$

On obtient ainsi G-1 relations linéaires entre les u_g^{-1} ce qui permet, en principe, de résoudre complètement le problème sachant que la somme des u_g^2 vaut 1.

Pour G = 2 (les BI, g = 1, les FL, g = 2) on obtient m = 73, $n_1^* u_1$ = 15 BI, $n_2^* u_2$ = 1 FL pour les données :

. $P_1 = 0{,}04$
. $P_2 = 0{,}01$
. $C_o = 10\,C_1 + 20\,C_2$
. $\mathcal{V}_1 = (0{,}0075)^2$

$\sigma_1 = P_1$
$\sigma_2 = P_2$

$\tau_1^2 = P_1\,(1 - P) - \tau_1^2 = P_1 - 2P_1^2$
$\tau_2^2 = P_2 - 2P_2^2$
$\mathcal{V}_2 = (0{,}0150)^2$

c) Une résolution numérique itérative du problème peut se faire de la façon suivante :

Étape 1 : On fixe une allocation approximative de l'échantillon dans chaque groupe, soit n_{+g} unités dans le groupe g. Pour y arriver on peut, par exemple, se servir de la solution approximative avec les hypothèses du point a) ou du point b).

Étape 2 : La valeur des λ_g est déterminée par les relations (8-5) :

$$\lambda_g = C_g \, n_{+g}^2 / \tau_g^2$$

Étape 3 : Les π_k sont déterminées par les relations (8-8). La somme des π_k fixe, en particulier , le nombre d'UP à tirer.

Étape 4 : Les n_{gk} sont déterminés par les relations (8-4). On peut ensuite itérer par retour à l'étape 2 en espérant que cet algorithme converge vers la solution d'optimisation.

d) Remarque : La probabilité de tirer une unité de type g vaut :

$$\pi_k \, n_{kg} / N_{kg} = \left(\frac{\lambda_g}{C_g} \right)^{1/2} \tau_g / \, N_{+g}$$

Elle est donc uniforme et on en déduit la taille n_g^+ de l'échantillon. Pratiquement, il arrive que l'on fixe "autoritairement" les tailles des échantillons. Ceci revient à déterminer les λ_g ou, implicitement, des variances ϑ_g Ce résultat est assez naturel lui aussi.

Sondage pour le contrôle de qualité "Colibri"

Position du problème

Il s'agit d'estimer la proportion de bulletins présentant une erreur de codification dans l'"univers" U de tous les bulletins codifiés une semaine donnée dans une direction régionale. Le caractère particulier du problème est le suivant : tous les bulletins i sont déjà précodifiés ce qui permet, grâce à des informations tirées de l'essai de recensement, d'attribuer à chacun d'eux une variable numérique positive X_i qui traduit sa "difficulté". Cette variable a été calibrée de façon à ce que Y_i (qui vaut 1 en cas d'erreur et 0 sinon) ait une "espérance" proportionnelle à X_i.

Toujours pour les mêmes raisons de coût du contrôle, on est amené à envisager un sondage à deux degrés :

- au premier degré de sondage on tirera un échantillon s_I de districts k à probabilités inégales π_k à déterminer. On notera π_{kl} les probabilités d'inclusion double pour cet échantillonnage,

– au second degré de sondage, on tirera un échantillon s_k des bulletins (BI) dans le district échantillon k. On notera $\pi_{i|k}$ la probabilité d'inclusion du bulletin i dans le district k, $\pi_{ij|k}$ la probabilité d'inclusion du couple (i, j) dans les districts

$s = \bigcup_{k \in s_1} s_k$ l'échantillon de BI.

On notera $X_k = \sum_{i \in k} X_i$ le total des X_i dans le district k,

$X = \sum_{k \in U} X_k = \sum_{U} X_i$ et on adoptera des notations analogues pour toutes les variables.

Le but est d'estimer une quantité de la forme $R = \dfrac{\sum_U Y_i}{\sum_U W_i}$ où W_i est une variable connue pour chaque bulletin. Cela pourra être $W_i = 1$ ou $W_i = X_i$ selon la mesure qui semble la plus adéquate du taux d'erreur.

Choix d'estimateur et variance

a) Au niveau d'un district (UP numéro k) il est naturel d'estimer le total Y_k des Y_i pour $i \in k$ par le ratio :

$$\hat{Y}_k = X_k \left(\sum_{s_k} Y_i \,/\, \pi_{i \mid k} \right) / \left(\sum_{s_k} X_i \,/\, \pi_{i \mid k} \right) = X_k \, \hat{a}_k$$

Ici $\hat{a}_k$ estime $a_k = Y_k / X_k$ avec un faible biais.

b) Pour estimer le ratio Y/X on utilisera :

$$\hat{a} = \frac{\sum_{s_1} \hat{Y}_k / \pi_k}{\sum_{s_1} X_k / \pi_k} = \frac{\sum_{s_1} \hat{a}_k \, X_k / \pi_k}{\sum_{s_1} X_k / \pi_k}$$

c) Si on veut estimer R, on remarquera que :

$$R = \frac{Y}{X} \cdot \frac{X}{W}$$

où X et W sont des totaux connus (difficulté totale et nombre total de bulletins, par exemple). Comme la variable X_i a été choisie pour sa bonne corrélation avec Y_i, un estimateur *a priori* intéressant de R sera :

$$\hat{R} = \hat{a}\,\frac{X}{W}$$

de sorte que la véritable question semble porter sur l'estimation de $a = \sum_k a_k X_k / X$

d) on aura :

$$Var\,(\hat{a}) = Var\,(E\,\hat{a} \mid s_1) + E\,Var(\hat{a} \mid s_1)$$

Pour le premier terme, compte tenu du fait que $\hat{a}_k$ estime (à peu près) sans biais a_k, on peut écrire :

$$Var\,(E\,\hat{a} \mid s_1) \approx \frac{1}{X^2}\,Var\left(\sum_{s_1} \frac{(a_k - a)\,X_k}{\pi_k}\right)$$

$$= \frac{1}{X^2}\left(\sum_k \frac{(a_k - a)^2\,X_k^2}{\pi_k^2} + \sum\sum_{k \neq l} (a_k - a)\,(a_l - a)\,\frac{X_k\,X_l\,\pi_{kl}}{\pi_k\,\pi_l}\right) \qquad (11 - 1)$$

Pour le second terme, on a, *conditionnellement à s_1* :

$$Var\left(\frac{\sum_{s_1} \hat{a}_k\,X_k/\pi_k}{\sum_{s_1} X_k/\pi_k}\right) = \left(\sum_{s_1} X_k/\pi_k\right)^{-2} \cdot \sum_{s_1} Var\,(\hat{a}_k)\,\frac{X_k^2}{\pi_k^2}$$

L'espérance de cette quantité vaut approximativement :

$$X^{-2} \sum_k E\,Var\,(\hat{a}_k \mid s_1)\,\frac{X_k^2}{\pi_k} \qquad (11 - 2)$$

avec :

$$Var\,(\hat{a}_k \mid s_1) = Var\,\frac{\sum_{s_k} Y_i/\pi_{i\,|\,k}}{\sum_{s_k} X_i/\pi_{i\,|\,k}} \approx \frac{1}{X_k^2}\,Var \sum_{s_k} \frac{Y_i - a_k\,X_i}{\pi_{i\,|\,k}}$$

$$= \frac{1}{X_k^2}\left(\sum_{i \in k} \frac{(Y_i - a_k\,X_i)^2}{\pi_{i\,|\,k}} + \sum\sum_{i \neq j} \frac{(Y_i - a_k\,X_i)\,(Y_j - a_k\,X_j)\,\pi_{ij\,|\,k}}{\pi_{i\,|\,k}\;\;\pi_{j\,|\,k}}\right)$$

Comme dans les parties précédentes, nous arrivons à des formules complexes et, finalement, inutilisables. Un modèle va nous simplifier un peu l'existence.

Intervention d'un modèle

Il aura la même structure que ceux qui ont déjà servi antérieurement :

a) Les a_k seront des variables aléatoires indépendantes de même espérance et de même variance :

$$E_\xi\, a_k = a \qquad Var_\xi\, a_k = \sigma^2$$

La variance prend en compte l'influence de l'opérateur ou de l'opératrice, qu'on renonce à isoler, mais aussi celle du jour de la semaine, de l'heure dans la journée, de certains jours du mois etc.

En revanche, par référence au modèle utilisé pour l'exhaustif, il ne contient plus l'effet lié à la difficulté du district. Celui-ci est pris en compte, du moins on l'espère, par la variable auxiliaire X_k qui intervient au point suivant.

b) Conditionnellement à a_k, les Y_i du district k sont des variables de Bernoulli indépendantes avec $E_\xi\,(Y_i \mid k) = a_k X_i$ et donc :

$$Var_\xi\ (Y_1 \mid k) = a_k X_i - a_k^2 X_i^2$$

Remarque : La variable X_i n'a pas de véritable sens concret, et n'est d'ailleurs définie qu'à un facteur d'échelle près. En revanche $a\,X_i$ et $\sigma\,X_i$ ont une interprétation physique invariante, car ce sont des probabilités. Dans tout ce qui suit il faudra toujours garder en tête que les résultats devront être invariants si les X_i sont multipliés par un facteur arbitraire à condition que a et σ soient divisés par le même facteur. En particulier Var $(\hat{a})$ n'a pas de sens "concret". Seule Var $(\hat{a}\,X)$ en a un.

Comme précédemment nous allons étudier la variance anticipée, espérance sous modèle de la somme de (11 - 1) et (11 - 2).

Pour le premier terme, l'espérance des produits croisés est nulle, comme de bien entendu. L'espérance sous modèle de ce terme est donc :

$$X^{-2}\,\sigma^2 \sum_k X_k^2/\pi_k$$

Pour le second terme on trouve : (vu la définition du 11 a)

$$X^{-2} \sum_k \frac{X_k^2}{\pi_k} \cdot \frac{1}{X_k^2} \sum_i \frac{a_k X_i - a_k^2 X_i^2}{\pi_{i\,|\,k}} = X^{-2} \sum_k \frac{1}{\pi_k} \sum_i \frac{aX_i - (a^2+\sigma^2)\,X_i^2}{\pi_{i\,|\,k}}$$

Globalement donc :

$$E_\xi\, Var(\hat{a}\,X) = \sigma^2 \sum_{k\,\in\,U} X_k^2 / \pi_k + \sum_{k\,\in\,U} \frac{1}{\pi_k} \sum_{i\,\in\,k} \frac{aX_i - (a^2+\sigma^2)\,X_i^2}{\pi_{i\,|\,k}}$$

Ici pas de miracle algébrique. *Pour simplifier* nous admettons que $(a^2+\sigma^2)\,X_i^2$ est négligeable devant aX_i. Numériquement on peut attendre aX_i = 2 à 5 x 10^{-2} et $(a^2+\sigma^2)\,X_i^2 = 3$ à 30×10^{-4}.

D'où notre approximation :

$$E_{\xi} Var(\hat{a}\, X) \approx \sigma^2 \sum_{k \in U} X_k^2/\pi_k + a \sum_{k \in U} \frac{1}{\pi_k} \sum_{i \in k} \frac{X_i}{\pi_{i \mid k}}$$

$$= \sigma^2 \sum_{k \in U} \frac{X_k^2}{\pi_k} + a \sum_{I} \frac{X_i}{\pi_i}$$

Optimisation

Nous utiliserons la fonction de coût suivante :

$$C = \sum_{s_1} C_o + C_1\, n_k$$

Ici, $n_k = \sum_{i \in k} \pi_{i \mid k}$ est la taille de l'échantillon tiré dans le district k (supposé de taille fixe à s_1 fixé). Son espérance vaut :

$$C_T = \sum_{k \in U} \pi_k\, (C_o + C_1\, n_k)$$

$$\text{Posons } \pi_{i \mid k} = n_k\, P_i \text{ (avec } \sum_{i \in k} P_i = 1) \text{ et } Q_k = \pi_k\, n_k$$

Le problème d'optimisation est maintenant :

$$\text{Min} : C_o \sum_k \pi_k + C_1 \sum_k Q_k$$

$$\text{sous} : \sigma^2 \sum_k \frac{X_k^2}{\pi_k} + a \sum_k \frac{1}{Q_k} \sum_{i \in k} \frac{X_i}{P_i} \leq V_o$$

Sous cette forme, on constate avec plaisir qu'on peut minimiser les termes en $\sum_i X_i/P_i$ indépendamment du reste. Autrement dit, n_k n'a pas d'incidence sur ce terme.

Laissons l'optimisation du second degré de tirage pour plus tard et notons seulement S_k^{*2} la valeur optimisée. Avec un multiplicateur de Lagrange λ on obtient par dérivation par rapport aux π_k puis au Q_k :

$$*C_o = \lambda\, \sigma^2 \frac{X_k^2}{\pi_k^2} \quad \text{soit} \quad \pi_k \text{ proportionnel à } X_k \qquad (13 - 1)$$

$$*C_1 = \lambda\, a \frac{S_k^{*2}}{Q_k^2} \quad \text{d'où} \quad n_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{S_k^*}{X} \qquad (13 - 2)$$

En particulier on tirera les districts avec des probabilités proportionnelles à leur difficulté totale.

Passons maintenant au tirage infra-district (deuxième degré de sondage).

Commençons par un cas simple et naïf : on tire les bulletins individuellement. La minimisation conduit à P_i proportionnelle à $\sqrt{X_i} = S_i$

Un calcul simple nous montre qu'alors $S_k^* = S_k = \sum_{i \in k} S_i$. Ceci nous permet de calculer n_k grâce à (13 - 2) et notre problème est entièrement résolu.

En fait les choses sont plus compliquées. Pour des raisons assez naturelles, on ne sélectionnera les BI que par ménages entiers. Autrement dit le sondage au second degré est un sondage en *grappes*. Les valeurs de P_i seront les mêmes, soit Pm, pour tous les membres d'un même ménage m.

Notons par X_m la somme des X_i des individus i du ménage m.

Le problème est donc de : Minimiser $\sum \frac{X_m}{P_m}$ *sous* $\sum n_m P_m = 1$ avec n_m taille du ménage m. On trouve facilement la solution :

$$P_m = \sqrt{\overline{X}_M} \Big/ \sum n_m \sqrt{\overline{X}_m}$$

Avec $\overline{X}_m = X_m / n_m$, *difficulté moyenne du* BI *du ménage m.*

Par suite, on trouve :

$$S_k^* = \sum n_m \sqrt{\overline{X}_m}$$

Cette solution nous permet de déterminer le nombre n_k de BI à tirer grâce à (13 - 2). Le nombre de *ménages*, en revanche n'est pas déterminé. Cette difficulté était prévisible. La fonction de coût, en effet, ne fait pas intervenir cette contrainte. Pour obtenir le nombre m_k de ménages à tirer, on s'arrangera de façon à ce que l'espérance du nombre de BI soit égale à n_k. Elle vaut :

$$m_k \left(\sum n_m \sqrt{\overline{X}_m}\right) \Big/ \sum \sqrt{\overline{X}_m}$$

d'où $m_k = n_k \dfrac{\sum \sqrt{X_m}}{\sum n_m \sqrt{X_m}}$

Compte tenu de (13 - 2) on a aussi :

$$m_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a_{1/2}}{\sigma} \frac{\sum \sqrt{X_m}}{X_k}$$

et la probabilité de tirer un ménage vaut alors :

$$m_k \frac{\sqrt{X_m}}{\sum \sqrt{X_m}}$$

Nous avons obtenu une solution complète du problème.

Remarque 1 : Dans les deux cas qui ont été traités, S_k^* est multiplié par $C^{1/2}$ si les X_i sont multipliés par C. La formule qui donne n_k est donc bien invariante à l'échelle de mesure.

Remarque 2 : La solution du cas 2 privilégie le tirage de petits ménages compliqués.

Remarque 3 : Ici comme dans les parties précédentes nous déterminons des probabilités d'inclusion simple mais pas des probabilités d'inclusion double. L'algorithme de tirage, qui fixe ces dernières, est donc sans influence. Ceci est relativement naturel si nous nous disons que l'information auxiliaire utilisée pour optimiser le tirage déterminera les π_k *et* $\pi_{i|k}$ mais ne peut pas avoir d'influence sur les probabilités doubles.

Remarque 4 : Ce problème fait apparaître des résultats un peu surprenants sur lesquels il est utile de réfléchir un peu.

Dans un premier cas, nous avons supposé qu'on pouvait isoler chaque bulletin. On tirait alors ceux-ci avec des probabilités proportionnelles à leur difficulté individuelle. On supposait, dans une certaine mesure, que le coût d'utilisation de l'information individuelle était nul.

Dans le second cas, la réalité du contrôle, ce coût était considéré comme infini et la seule information ayant un coût négligeable était celle relative à l'ensemble du ménage. La solution fait alors apparaître des probabilités de tirage des individus (BI*)* qui est fonction de la difficulté moyenne de codification des bulletins de l'ensemble du ménage auquel appartient cet individu.

Il en irait de même en ce qui concerne le tirage des districts. Si on sait y distinguer les BI, on les tire avec des probabilités proportionnelles à la difficulté totale ; à l'intérieur des districts, on tirera des BI difficiles avec une plus grosse probabilité. Supposons, au contraire, qu'on ne sache pas distinguer les BI à l'intérieur des districts. Ce serait le

cas, par exemple, si la désignation des BI à contrôler ne pouvait pas se faire en temps réel par suite d'une organisation du traitement inadéquate. On tirerait alors les districts proportionnellement à leur difficulté moyenne : à l'intérieur des districts, on serait obligé de réaliser des sondages aléatoires simples.

Dans le premier cas le sondage privilégiera les gros districts à l'intérieur desquels on tirera plutôt les BI difficiles. Dans le second cas, on privilégiera les petits districts difficiles à l'intérieur desquels on tirera des bulletins à probabilités égales. *Dans les deux cas*, on cherchera à augmenter la probabilité de sonder des BI difficiles. La différence réside simplement dans la possibilité (c'est-à-dire le coût) de mobiliser l'information au moment où on en a besoin.

Variance anticipée optimale et application

Après quelques manipulations algébriques, on trouve la valeur de la variance optimale :

$$E_\xi \, Var(\hat{a}\, X)_{OPT} = \frac{(\sigma X)^2}{m}\left[1 + \frac{a}{\sigma}\frac{a^{-1/2} S^*}{X}\left(\frac{C_1}{C_o}\right)^{1/2}\right]$$

Cette forme respecte le caractère homogène des différents facteurs. On a, en particulier $\frac{a^{-1/2} S^*}{X} = \frac{a^{1/2} S^*}{aX}$ le dénominateur est interprétable comme un nombre total d'erreurs dans un lot, tandis que le numérateur est homogène à une taille.

L'application pratique et numérique de cette théorie repose sur des hypothèses concernant les ordres de grandeurs des différents paramètres (ce qui demande qu'on puisse les raccrocher à une interprétation physique simple). Dans la phase de préparation du recensement, sans mesures préalables très précises, on a utilisé les valeurs $\sigma/a = 0{,}5$ et $C_1 / C_0 = 0{,}1$

A la suite de diverses hypothèses sur les autres paramètres et de discussion entre experts, il a été décidé un contrôle portant sur 50 districts chacun d'eux étant contrôlé pour environ 20 BI (par région et par semaine). Cet ordre de grandeur initial pouvait, évidemment, être modulé dans la suite du contrôle, les paramètres du modèle pouvant être réestimés après chacun d'eux.

Vue d'ensemble

Cette étude est basée sur l'utilisation de modèles de superpopulation pour anticiper la variance d'une mesure par sondage *avant* le sondage. On arrive, en utilisant des modèles simples qu'on voudrait néanmoins assez réalistes, à des expressions plus ou moins complexes qu'on parvient à optimiser, parfois rigoureusement, quelquefois de façon approximative.
La solution du dernier de ces problèmes fait apparaître un facteur assez peu étudié dans les problèmes d'optimisation de plan de sondage : le coût lié à la mobilisation d'une information individuelle.

BIBLIOGRAPHIE

BADEYAN Gérard : Communication aux secondes Journées de Méthodologie Statistique, 17 et 18 Juin 1992, Insee, Paris (1992).

CHARTIER Fernand : Échantillonnage au 1/20^{e} du prochain recensement, *Note interne* 648/470, Insee, Département Population Ménages, Paris (1979).

COCHRAN William : *Sampling Techniques*, 3^{e} édition, Wiley, New-york (1977).

COEFFIC Nicole : L'enquête de mesure de degré d'exhaustivité du recensement de 1990. Insee, *document de travail F9201*, sDirection des Statistiques Démographiques et Sociales, Paris (1992).

DESABIE Jacques : Théorie et Pratique des Sondages, Dunod, Paris (1965).

DESPLANQUES Guy : Une nouvelle enquête sur la constitution des familles - *Courrier des Statistiques*, n° 20, Octobre 1981, pp. 51-52, Paris (1981).

DESPLANQUES Guy : Cycle de Vie et Milieu Social - *Les collections de l'Insee* , D117, Insee, Paris (1987).

DEVILLE Jean Claude : Structure des Familles, *Les collections de l'Insee*, D13-14, Insee, Paris (1972).

DEVILLE Jean Claude, GROBRAS Jean Marie, ROTH Nicole : Efficient Sampling Algorithms and Balanced Samples. *Compstat 1988*, pp 255-266, *Physica Verlag*. Heidelberg (1988).