

Échantillonnage représentatif optimum à probabilités inégales

Pascal ARDILLY
Insee

- ECHANTILLONNAGE REPRESENTATIF OPTIMUM A PROBABILITES INEGALES -

RESUME: Nous proposons des algorithmes de tirage à probabilités inégales faciles à mettre en oeuvre, approchant la représentativité de l'échantillon pour certaines variables connues pour chaque individu, et assurant l'optimalité des probabilités de tirage $p(s)$ des échantillons. On voit, sur des données cantonales de 1982, qu'il est possible, dans le cas d'échantillons de petite taille (quelques unités) dans des populations elles mêmes de taille limitée (inférieure à 100), de remplacer l'estimateur complexe de la régression par l'estimateur très simple des sommes dilatées de précision comparable, la prise en compte des variables auxiliaires se faisant au niveau du tirage et non de l'estimateur.

I) INTRODUCTION :

Dans la pratique des sondages, lorsqu'on cherche à estimer dans une population finie un paramètre fonction linéaire des données tel que total, moyenne ou pourcentage, on utilise communément l'estimateurs sans biais classique des sommes dilatées (dit encore estimateur de Horvitz-Tompson) (1). Si on note Y_i la valeur de la variable d'intérêt pour l'individu i ($1 \leq i \leq N$), et P_i la probabilité qu'a ce même individu d'appartenir à l'échantillon s , cet estimateur, dans le cas de l'estimation du vrai total

inconnu Y , a pour expression :

$$\hat{Y} = \sum_{i \in s} \frac{Y_i}{P_i}$$

Dans la plupart des cas pratiques, la probabilité P_i est choisie proportionnelle à une variable de taille connue sur l'ensemble de la population. La théorie montre cependant qu'il est préférable, lorsqu'on connaît sur l'ensemble de la population une variable auxiliaire autre que la taille qui soit bien corrélée avec la variable d'intérêt, d'utiliser cette information pour améliorer la qualité des résultats, qualité mesurée par l'erreur quadratique moyenne de l'estimateur dans la population finie. On peut tenir compte de l'information, soit au niveau du tirage grâce à une technique de stratification, soit au niveau de l'estimation, auquel cas on utilise essentiellement des estimateurs par le ratio ou par la régression (2).

Dans la première optique, lorsqu'on choisit un plan de sondage stratifié avec tirage proportionnel à la taille dans chaque strate, on se heurte au problème du calcul de la variance de l'estimateur, soit :

$$V(\hat{Y}) = \sum_{1 \leq i < j \leq N} (P_i \cdot P_j - P_{ij}) \cdot \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2$$

au travers des probabilités d'inclusion doubles P_{ij} , où, P_{ij} est la probabilité que les individus i et j soient simultanément dans l'échantillon s . Les P_{ij} sont, hélas, fonction de l'algorithme de tirage utilisé, et on peut dire qu'il n'existe encore que très peu de procédures de tirage de taille fixe donnant lieu à des expressions analytiques exactes, valables quelque soit la population et la taille de l'échantillon, et donnant lieu à une programmation qui reste accessible (3). Le tirage systématique sur fichier aléatoire (4)-(5) et l'algorithme de Sunter (6) sont parmi les plus satisfaisants, encore que, pour le premier les expressions soient très complexes, et pour le second, outre le fait que la taille d'échantillon initialement prévue ne soit pas nécessairement respectée, les P_{ij} ne sont calculables que par récurrence. La stratification pose en outre le problème du choix de la variable de stratification, du nombre de strates, et de la limite des strates (7); il existe certes de bonnes approximations des limites optimales de strates dans le cas où on utilise un sondage aléatoire simple avec allocation optimale de Neymann dans chaque strate, mais le problème devient insoluble si on souhaite tirer les unités proportionnellement à leur taille dans chaque strate.

Dans la seconde optique, si on souhaite agir sur l'expression de l'estimateur en utilisant un estimateur par le ratio ou par la régression, il existe un biais qui peut-être important si l'échantillon est de petite taille. Comme dans le plan de sondage alternatif stratifié, les variances font encore intervenir P_{ij} . Il est donc impossible de fournir une valeur exacte de l'erreur quadratique moyenne de l'estimateur du total d'une variable

d'intérêt partout connue, de même qu'il est impossible d'estimer sans biais l'erreur quadratique moyenne d'un estimateur du total lorsque la taille de l'échantillon est faible (Par exemple inférieure à 10). Lorsqu'on dispose de plusieurs informations auxiliaires, celles-ci seront probablement perdues dans la plupart des cas si on opte pour le ratio, car la méthode théorique du multiratio (1) fait intervenir des pondérations pratiquement incalculables. Quant à l'estimateur par la régression, il ne serait pas raisonnable de l'utiliser si la taille de l'échantillon est inférieure à 10 (et encore, dans le cas où on ne dispose que d'un seul régresseur).

Du point de vue pratique, ratio et régression, dont on sait qu'ils peuvent réduire considérablement la variance si la variable d'intérêt est bien expliquée par les variables auxiliaires, ont l'énorme inconvénient de nécessiter un traitement spécifique à chaque variable d'intérêt : l'estimateur par la régression impose un calcul des coefficients de régression propre à chaque variable, contrainte fortement dissuasive qui limite largement leur utilisation; sauf pour l'enquête de conjoncture auprès des entreprises où, typiquement, la seule connaissance de l'investissement est recherchée, cette méthode n'a jamais été utilisée à notre connaissance dans les enquêtes à grande échelle (encore fait-il nuancer, puisqu'on sait depuis peu (8) que les méthodes de calage sur marge sont asymptotiquement équivalentes à des estimations par régression). Dans les arguments qui pourraient privilégier l'utilisation de l'information auxiliaire au niveau du tirage sur sa prise en compte dans l'expression de l'estimateur, on trouve l'influence de l'ordre chronologique dans lequel s'effectuent les opérations : notre propos a son origine dans la constitution de l'échantillon-maître français de 1992.

L'échantillon-maître est une réserve de logements destinée à alimenter les échantillons de toutes les enquêtes-ménages de l'INSEE sur la période inter-censitaire (sauf l'enquête emploi) (13). Il est constitué sur la base d'un tirage stratifié à plusieurs degrés. Dans la strate des cantons ruraux ou des fractions rurales de cantons partiellement urbains, on souhaite tirer les cantons proportionnellement à leur taille car on estime que, sur la plupart des variables d'intérêt, les totaux par canton sont bien corrélés avec le nombre total de logements du canton. Il est cependant facile de comprendre, en zone rurale comme en fait partout ailleurs, que l'on ne doit pas imposer l'utilisation en "aval" de certains estimateurs plus compliqués que l'estimateur des sommes dilatées. Dans ces conditions, cela ne nous engage à rien d'utiliser en "amont" l'information au niveau du tirage : libre aux utilisateurs de n'employer que les sommes dilatées, sachant que la qualité devrait être améliorée par rapport à la situation actuelle (voir résultats), et libre à eux de programmer quand même un ratio ou une régression en sus si la taille de l'échantillon et le budget le permettent.

Ce qui suit a pour but de proposer une méthode de tirage et des variantes permettant un calcul rigoureux exact de variance vraie à partir de simples estimateurs des sommes dilatées, et cela dans

le cadre d'un tirage "aussi représentatif que possible" au sens de Hajek (9).

Dans un premier temps, après avoir précisé le type d'approche choisie pour répondre à ces préoccupations, on recense les contraintes imposées, on formalise simplement la question, et on décrit les algorithmes utilisés pour y répondre. Dans un deuxième temps, on présente les résultats issus de nos méthodes, et on les compare avec ceux qui résultent des pratiques classiques. Enfin, on cite sommairement quelques extensions possibles de la méthode.

II) LES ALGORITHMES DE TIRAGE D' ECHANTILLONS :

Les algorithmes proposés sont au nombre de trois. Tous résultent d'une approche du problème qui consiste à fabriquer une liste d'échantillons potentiels en tenant compte de l'information auxiliaire, échantillons auxquels on affecte une probabilité de tirage optimale selon un certain critère. En effet, les expressions originelles de l'espérance et de la variance d'un estimateur \hat{Y} d'un paramètre de total Y font intervenir l'ensemble des échantillons s réalisables (noté S) à partir de la population finie dont on dispose. Ainsi :

$$E(\hat{Y}) = \sum_{s \in S} p(s) \cdot \hat{Y}(s)$$

$$V(\hat{Y}) = \sum_{s \in S} p(s) \cdot (\hat{Y}(s) - E(\hat{Y}))^2$$

, où $\hat{Y}(s)$ est l'estimateur de Y si on tire l'échantillon s , et $p(s)$ la probabilité de tirer l'échantillon s .

Ici, l'idée est d'agir à la source, soit directement, au niveau des probabilités d'échantillon $p(s)$, au lieu de déterminer des probabilités d'inclusion P_i pertinentes comme dans les approches traditionnelles.

Notre recherche d'optimalité doit tenir compte d'un certain nombre de contraintes, soit imposées par le sondeur pour améliorer l'efficacité des estimateurs, soit imposées "mathématiquement" par la théorie des probabilités.

En ce qui concerne le premier type de contraintes, on cherche à adapter au niveau du tirage la propriété fondamentale des estimateurs par le ratio ou la régression : estimer avec une variance nulle le total d'une variable auxiliaire connue sur l'ensemble de la population. L'idéal serait de réaliser le tirage de l'échantillon (s) en imposant :

$$(1) \quad \sum_{i \in s} \frac{X_i}{P_i} = \sum_{i=1}^N X_i, \quad \forall s \in S$$

où X_i est la variable auxiliaire attachée à l'individu i , P_i la probabilité d'inclusion de l'individu i , et N la taille de la population. Cette idée de l'échantillon équilibré qui apparaît chez Hajek (9) et Royall (10), a été exploitée à l'INSEE dans le cas d'un tirage aléatoire simple lorsque l'échantillon est de grande taille (11). Dans notre cadre, surtout si la taille de l'échantillon, soit n , est faible, il est probable que (1) soit numériquement irréalisable quelque soit s - Par conséquent, on impose :

$$(2) \quad \left| \frac{\sum_{i \in s} \frac{X_i}{P_i} - \sum_{i=1}^N X_i}{\sum_{i=1}^N X_i} \right| \leq 1(x)$$

où $l(x)$ est un seuil maximum admissible, considéré comme paramètre du problème. Cette contrainte nous assure une estimation du total des variables connues aussi bonne que possible compte tenu de la structure des données, ce que ne permet pas systématiquement le critère classique de la variance. Elle correspond à la véritable notion de représentativité approchée d'un échantillon (que l'on confond très souvent avec l'allocation proportionnelle du tirage stratifié). La contrainte (2) peut être vue comme une manière de tronquer la loi de l'estimateur \hat{X} pour diminuer sa variance. Si on dispose à l'origine de la liste complète S de tous les échantillons de taille fixe n réalisables, l'application de la contrainte (2) sélectionne un sous-ensemble de S , noté S_1 , d'échantillons admissibles. On ne contrôle évidemment en rien la constitution de S_1 , et il est malheureusement possible que certains individus atypiques de la population Y soient très peu représentés, voire pas du tout.

Le second type de contrainte s'obtient en écrivant la définition de la probabilité d'inclusion P_i de chacun des individus de la population. Ainsi :

$$(3) \quad P_i = \sum_{\substack{\Delta \ni i \\ \Delta \in S_1}} p(s) \quad , \quad \forall i \in \{1, 2, \dots, N\}$$

On remarquera que la somme s'effectue sur les échantillons de S_1 seulement, les autres éléments de S étant définitivement écartés. Le tirage étant de taille fixe n , proportionnel à une taille T_i , on impose :

$$(4) \quad P_i = n \cdot \frac{T_i}{\sum_{i=1}^N T_i} \quad , \quad \forall i \in \{1, 2, \dots, N\}$$

Comme $p(s)$ est une probabilité, on doit avoir :

$$0 \leq p(s) \leq 1 \quad , \quad \forall s \in S_1$$

et

$$\sum_{s \in S_1} p(s) = 1$$

On peut cependant écrire :

$$\begin{aligned} \sum_{s \in S_1} p(s) &= \sum_{s \in S_1} \left(\frac{\sum_{i=1}^N \mathbb{1}_{i \in \Delta}}{n} \right) p(s) \\ &= \frac{1}{n} \cdot \sum_{i=1}^N \sum_{s \in S_1} p(s) \cdot \mathbb{1}_{i \in \Delta} \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{n} \sum_{s \in S_1} \sum_{\substack{1 \leq i \leq N \\ \Delta \theta \leq 1}} p(s) \\
 &= \frac{1}{n} \cdot \sum_{1 \leq i \leq N} P_i \\
 &= 1
 \end{aligned}$$

La dernière contrainte est par conséquent impliquée par (3) et (4). Par ailleurs, si le tirage proportionnel à la taille est possible, c'est à dire si on a :

$$P_i \leq 1, \quad \forall i \in \{1, \dots, N\}$$

, alors automatiquement :

$$p(s) \leq 1, \quad \forall s \in S_1$$

, d'après (3).

a) Si on s'intéresse prioritairement à l'estimateur $\hat{X}(s)$ d'un total X bien particulier, alors on peut chercher à obtenir une qualité maximale sur cet estimateur, compte tenu des deux types de contraintes que l'on s'est imposées. La formalisation de cet objectif va conduire au premier algorithme :
Chercher la famille $p(s)$, s décrivant S_1 , vérifiant le programme P :

$$\begin{aligned}
 (P) \quad & \text{Min} \sum_{s \in S_1} p(s) \cdot (\hat{X}(s) - X)^2 \\
 & \text{s c } \begin{cases} \sum_{\substack{\Delta \theta \leq 1 \\ s \in S_1}} p(s) = P_i, \quad \forall i \in \{1, \dots, N\} \\ 0 \leq p(s), \quad \forall s \in S_1 \end{cases}
 \end{aligned}$$

Si on s'intéresse à un ensemble de variables auxiliaires x^k , $1 \leq k \leq K$, on peut choisir un critère $C(s)$ construit à partir des valeurs $\hat{X}(s)$ reflétant la qualité d'ensemble du tirage, puis résoudre le programme P_0 :

$$\begin{aligned}
 (P_0) \quad & \text{Min} \sum_{s \in S_1} p(s) \cdot C(s) \\
 & \text{s c } \begin{cases} \sum_{\substack{\Delta \theta \leq 1 \\ s \in S_1}} p(s) = P_i, \quad \forall i \in \{1, \dots, N\} \\ 0 \leq p(s), \quad \forall s \in S_1 \end{cases}
 \end{aligned}$$

Par exemple :

$$(5) \quad C(s) = \sum_{k=1}^K \left(\bar{X}^k(\Delta) - X^k \right)^2$$

ou

$$(6) \quad C(s) = \sum_{k=1}^K \frac{\left(\bar{X}^k(\Delta) - X^k \right)^2}{X^k}$$

ou

$$(7) \quad C(s) = \sum_{k=1}^K \left(\frac{\bar{X}^k(\Delta) - X^k}{X^k} \right)^2$$

Le critère (5) conduit à une minimisation de la somme des variances des estimateurs construits à partir des variables auxiliaires. En un certain sens, il donne bien une importance plus grande aux variables dont les valeurs numériques sont les plus fortes, et s'utilise donc si on accorde une importance moindre aux populations rares. Le critère (6) est du type chi-deux, et représente donc une distance classique. Le critère (7) est parfaitement homogène, et s'apparente à une minimisation de coefficient de variation.

La résolution numérique du problème s'effectue de la manière qui suit :

La première étape est une étape de génération d'échantillons qui conduit à la constitution de la liste S. On rappelle que tirer n individus parmi N nécessite, avec ce procédé, de constituer un fichier de C_N^n enregistrements : tirer 5 individus parmi 100 va créer 75 millions de possibilités ! En tirer 5 parmi 50 ne va plus en créer que 2 millions (ce qui est à la portée des ordinateurs actuels si on n'est pas trop pressé...). Plus contraignant est le traitement de la minimisation, car les programmes adaptés ne peuvent pas manipuler un trop grand nombre de contraintes ni de variables (aucun problème cependant en dessous de 2000 variables pour une centaine de contraintes). Pour ces raisons, il est nécessaire, avec l'informatique dont nous disposons actuellement, de nous limiter à des petites tailles d'échantillons dans une population restreinte.

La seconde étape, après test d'admissibilité sur chacun des échantillons de la liste S, conduit à la liste S1 vérifiant (2). Partant de là, si on s'attache à la résolution de P0, on sait que, sauf configuration très particulière des données, on obtient toujours une solution si le domaine défini par les contraintes n'est pas vide. En théorie, dans une population de taille N quelconque, si on se donne à priori un ensemble de réels P_i ($1 \leq i \leq N$) compris entre 0 et 1 qui vérifient :

$$\sum_{i=1}^N P_i = n, \quad n \in \mathbb{N}^*$$

, il est toujours possible de déterminer un système de probabilités $p(s)$ tel que :

$$\sum_{\substack{\Delta \ni i \\ \Delta \in S}} p(s) = P_i, \quad \forall i \in \{1, \dots, N\}$$

Ayant éliminé les échantillons ne satisfaisant pas à (2), le théorème ne s'applique plus, et le domaine des contraintes de P_0 peut être vide. Qualitativement, pour minimiser ce risque, nous avons cherché à constituer des échantillons tels que chaque individu apparaisse dans un nombre d'échantillons à peu près proportionnel à sa taille. Cette technique permet de limiter les inconvénients introduits par l'effet de grappe que l'on génère : supposons par exemple qu'un individu i de taille T_i importante prenne des valeurs X_i atypiques, et qu'il ne soit présent que dans le seul échantillon \bar{s} . Pour que \bar{s} soit à peu près équilibré au sens de (2), il est fort probable que i sera associé à au moins un individu j de taille inférieure. Dans ce cas, on aura :

$$P_j = \sum_{s \ni j} p(s) \geq p(\bar{s}) = P_i$$

ce qui est contredit par l'ordre des tailles de i et j . Ne contrôlant pas la structure des données, et compte tenu de l'absence de liaison entre la taille et le caractère plus ou moins atypique des variables auxiliaires, on peut constater dans la liste S_1 de forts écarts à la proportionnalité recherchée pour certains individus.

A la troisième étape, on est par conséquent amené à sélectionner un sous-ensemble S_2 de S_1 qui résulte de l'algorithme suivant :

- on considère tous les échantillons admissibles S_1
- on les classe par valeur croissante du critère $C(s)$
- on calcule, pour chaque individu i , le nombre théorique $N_i(1)$ d'échantillons où il doit être présent, compte tenu de la taille prévue de S_2
- A l'étape u , on retient l'échantillon s classé en position u si pour tout i de s , on a $N_i(u) > 0$, et on met à jour selon : $N_i(u+1) = N_i(u) - 1$.
Sinon on passe directement à l'échantillon classé $u+1$.

On arrête la procédure lorsque S_2 comprend le nombre d'échantillons que l'on s'était fixé comme limite à priori (mettons de l'ordre de 2000). En réalité, on obtient systématiquement moins d'échantillons que ce qui était demandé à cause de l'épuisement plus ou moins rapide des individus dans les échantillons sélectionnés. Pour cette raison, la méthode n'est pas optimale, et la question de son amélioration reste ouverte.

Le problème devient finalement P_1 :

$$\begin{array}{l}
 \text{Min} \sum_{s \in S_2} p(s) \cdot C(s) \\
 \text{(P1)} \quad \text{s c} \left\{ \begin{array}{l}
 \sum_{\substack{s \in S_2 \\ i \in S_1}} p(s) = P_i, \forall i \in \{1, \dots, N\} \\
 0 \leq p(s), \forall s \in S_2
 \end{array} \right.
 \end{array}$$

P1 est un programme de minimisation de fonction linéaire sous contraintes linéaires dans le domaine convexe des variables positives. L'algorithme du simplexe est adapté à un tel problème (12), et son utilisation à partir de bibliothèques scientifiques standards (IMSL ou HARWELL) fournit directement en sortie les valeurs des $p(s)$ optimisées. Les variances des estimateurs des totaux des variables auxiliaires peuvent alors être calculés de manière exacte, et fournir une référence numérique pour les calculs de précision ultérieurs à partir d'autres variables qui leurs seraient corrélées. Les probabilités P_{ij} d'inclusion doubles sont immédiatement calculables à partir de :

$$P_{ij} = \sum_{\substack{s \in S_2 \\ i \in S_1 \\ j \in S_1}} p(s), \forall i, j \in \{1, \dots, N\}$$

b) Dans la méthode précédente, par construction les probabilités $p(s)$ de tirage sont fonction des variables auxiliaires. Si les variables d'intérêt sont bien corrélées avec les variables auxiliaires, alors on peut s'attendre à une bonne précision - Dans le cas contraire, on prend un risque, que l'on peut tenter de minimiser en résolvant un programme de type minimax envisageant toutes les valeurs à priori possibles des variables auxiliaires. Dans un premier temps, on peut, sans fixer les probabilités d'inclusion et en ajoutant une contrainte évidente de normalisation, formuler la question ainsi :

$$\begin{array}{l}
 \text{Min} (\text{Max}_{\substack{p(s) \\ C(s)}} \sum_{s \in S_2} p(s) \cdot C(s)) \\
 \text{(P2)} \quad \text{s c} \left\{ \begin{array}{l}
 \sum_{s \in S_2} p(s) = 1 \\
 0 \leq p(s) \leq 1, \forall s \in S_2 \\
 \sum_{s \in S_2} C^2(s) = 1
 \end{array} \right.
 \end{array}$$

Par la méthode du Lagrangien, cette optique minimax nous permet de conclure qu'à critère donné quelconque il est nécessaire de choisir des probabilités $p(s)$ constantes. On est donc tenté de rechercher un jeu de probabilités $p(s)$ présentant une aussi faible variabilité que possible. On peut utiliser, à cette fin, deux critères concurrents qui vont mener aux deux algorithmes simples que voici :

Le premier utilise un critère quadratique :

$$(P2) \quad \begin{array}{l} \text{Min} \sum_{s \in S2} \left(p(s) - \frac{1}{|S2|} \right)^2 \\ \text{s.c.} \left[\begin{array}{l} \sum_{s \in S2} p(s) = P_i, \forall i \in \{1, \dots, N\} \\ 0 \leq p(s), \forall s \in S2 \end{array} \right. \end{array}$$

où $|S2|$ représente la cardinalité de $S2$.

Si on ne tient pas compte des contraintes de positivité, on obtient :

$$(8) \quad \vec{v} = \vec{e} + A' \cdot B^{-1} \cdot \vec{R}, \quad \vec{v} = (p(s))_{1 \leq s \leq |S2|}$$

où

$$\vec{e} = \left(\frac{1}{|S2|} \right)$$

, vecteur de coordonnées constantes de taille $|S2|$

$$A = (A_{ij}), \quad 1 \leq i \leq N, \quad 1 \leq j \leq N,$$

avec $A_{ij} = 1$ si l'individu (i) appartient à l'échantillon (j)

$$B = (B_{ij}), \quad 1 \leq i \leq N, \quad 1 \leq j \leq N,$$

avec $B_{ij} = \text{Card} \{ s \in S2 \mid i \in s \text{ et } j \in s \}$

$$\vec{R} = (R_i), \quad 1 \leq i \leq N,$$

avec

$$R_i = P_i - \frac{B_{ii}}{|S2|}$$

Compte tenu de la relation :

$$\sum_{j=1}^N B_{ij} = n \cdot B_{ii}, \quad \forall i \in \{1, \dots, N\}$$

on peut aussi écrire :

$$\vec{v} = \vec{e} + A' \cdot \left(B^{-1} \cdot \vec{P} - \frac{\vec{e}}{n} \right)$$

, où \vec{P} est le vecteur des probabilités d'inclusion P_i , $1 \leq i \leq N$
De plus :

$$(9) \quad \begin{aligned} P_{ij} &= \frac{B_{ij}}{|S2|} + \vec{C}_{ij} \cdot A' \cdot \left(B^{-1} \cdot \vec{P} - \frac{\vec{e}}{n} \right) \\ &= \frac{B_{ij}}{|S2|} + \vec{C}_{ij} \cdot (\vec{v} - \vec{e}) \end{aligned}$$

où \vec{C}_{ij} est un vecteur ligne de taille $|S2|$, avec $C_{ij,k} = 1$ si l'échantillon (k) contient (i) et (j).

On voit bien que pour minimiser notre critère, il est avantageux de construire S_2 de façon à obtenir :

$$P_i \approx \frac{B_{ii}}{|S_2|}$$

c'est à dire B_{ii} autant que possible proportionnel à la taille T_i de l'individu i .

On retrouve la condition qui conduisait à la constitution de S_2 pour P_1 , consistant à placer chaque individu dans un nombre d'échantillons autant que possible proportionnel à sa taille - Dans ce cas, les probabilités doubles données par (9) se réduisent à la proportion empirique d'échantillons contenant le couple (i, j) .

(8) ayant été établie sans tenir compte des contraintes de positivité (ce qui entraînerait une trop grande complication), la pratique a consisté à rendre nulles les quelques probabilités négatives (voir résultats), et à normer celles qui étaient positives.

c) Le second algorithme permettant de réduire la dispersion des probabilités $p(s)$ utilise un critère de type "entropie" :

$$(P_3) \quad \begin{array}{l} \text{Min} \quad \sum_{s \in S_2} p(s) \cdot \text{Log}(p(s)) \\ \text{s c} \quad \left[\begin{array}{l} \sum_{s \in S_2} p(s) = P_i, \quad \forall i \in \{1, \dots, N\} \\ 0 < p(s), \quad \forall s \in S_2 \end{array} \right. \end{array}$$

On sait en effet qu'une minimisation d'entropie sous contrainte de normalisation conduit à un minimum où toutes les variables prennent la même valeur. Outre le fait de fournir une solution numérique différente de celle qui résulte du critère quadratique et donc conduisant peut-être à des variances plus faibles (les applications numériques fourniront des réponses partielles sur ce point), l'avantage de ce critère réside dans l'économie des corrections à posteriori pour rendre les $p(s)$ positives. Il n'est en effet plus nécessaire de tenir compte des contraintes d'inégalité qui sont automatiquement vérifiées puisque :

$$p(s) = \exp\left(-1 - \sum_{i=1}^N L_i \cdot A_{is}\right)$$

où L_i est le multiplicateur associé à l'individu i . Celui-ci s'obtient en résolvant le système :

$$P_i = \exp(-1) \cdot \sum_{\substack{s \ni i \\ s \in S_2}} \prod_{j \in s} \exp(-L_j), \quad \forall i \in \{1, \dots, N\}$$

grâce à un algorithme approprié traitant les systèmes non linéaires (et il s'agit bien, là encore, de recherche opérationnelle standard).

III) RESULTATS NUMERIQUES :

Les algorithmes précédents ont été appliqués à des données issues de la Banque de Données Locales 1982, et constituées par des effectifs cantonaux en zone rurale. Les variables suivantes ont été retenues :

- Nombre d'agriculteurs exploitants hommes et femmes dans le canton (Noté AGRI)
- Nombre d'ouvriers (y.c. agricoles) hommes et femmes dans le canton (Noté OUVR)
- Nombre d'hommes et femmes dans le canton classés dans une des CSP suivante : Artisans, commerçants, chefs d'entreprise, cadres, professions intellectuelles supérieures (Note CADR)
- Nombre total de jeunes de 19 ans et moins résidant dans le canton (Noté JEUNES)
- Population du canton ayant un emploi résidant et travaillant dans la même commune (Noté MIGR)
- Population totale active française dans le canton (Noté ACTIF)
- Nombre de résidences secondaires dans le canton (Y.C. logements meublés loués ou à louer) (Noté NRESSEC)
- Population totale du canton (Noté PENQ)
- Population totale ayant un emploi dans l'industrie (Noté INDUS)
- Population de 15 ans ou plus ayant un bac général ou technique (Noté POPBAC)
- Population totale de 60 ans et plus (Noté VIEUX)
- Variable constante valant 1 dans chaque canton (Noté CSTE)

Le contrôle (2) a porté sur les 8 premières. Nous nous sommes limités aux cantons du Limousin (RG=74) et de la Provence - Côte d'Azur (RG=93) où, sur des populations respectives de 55 et 61 cantons, nous avons fixé la taille de l'échantillon à 3 : cette valeur est exactement le nombre de cantons présents dans la strate correspondante de l'échantillon-maître issu du Recensement 1982.

Le tableau I fournit les coefficients de variation des variables de la liste pour 11 plans de sondage concurrents :

- . LIGNE (1) : Simplexe P1 avec, pour les variables considérées dans l'ordre de la liste, les seuils $l(x)$ respectifs :
- 15%, 20%, 20%, 15%; 20%; 15%; 15%; 15%; pour RG93
- et 15% partout pour RG74
- Critère selon (5); limite à souhaiter 2 000 échantillons pour S2.
Nombre d'échantillons finalement présents dans S2: 1724 pour RG 74 et 1072 pour RG 93
- " (2) : Idem avec le critère (6)
- " (3) : Idem avec le critère (7)
- " (4) : Idem avec le critère : $\sum_{s \in \Omega} p(s) \cdot (\hat{X}(s) - X)^2$
où X_i représente la population totale du canton i
- " (5) : Minimax P2, critère C(s) selon (5)
- " (6) : Minimax P3, critère C(s) selon (5)
- " (7) : Tirage systématique sur fichier trié selon la taille.
Estimateur des sommes dilatées
- " (8) : Tirage selon la méthode de Sunter, estimateur des sommes dilatées
- . (9) : Stratification en 3 strates, selon la 1ère composante principale, avec limites de strates "optimales" à la suite de diverses simulations - Tirage proportionnel à la taille d'un individu par strate
- . (10) : Tirage systématique (fichier trié) et estimateur par le ratio.
N' est retenu que le meilleur coefficient de variation parmi les 11 ratios possibles pour chaque variable
- . LIGNE (11) : Tirage systématique (fichier trié) et estimation par la régression.
Les paramètres de la régression sont estimés sur l'ensemble des variables auxiliaires dont la significativité n'est pas manifestement nulle.

Le jeu de seuils utilisé n'est évidemment pas optimum. Il résulte d'un balayage qui constitue un compromis entre le désir de "serrer" les seuils $l(x)$ au maximum pour obtenir des échantillons aussi représentatifs que possible, et la nécessité de ne pas engager des coûts informatiques sans commune mesure avec le gain qui pourrait résulter d'un abaissement marginal de ceux-ci. Dans notre cas, une recherche optimale de seuils équivaldrait à une estimation de densité d'un vecteur aléatoire à 8 dimensions !

Les seuils imposés dans les deux régions ont été fixés à la suite de quelques tentatives plus ou moins heureuses. Une valeur uniformément égale à 15% en région 74 permettait à chaque canton d'être présent au moins une fois dans la liste S2. Cela ne fonctionnait plus avec la région 93, principalement à cause de la présence d'un tout petit canton atypique : d'où le choix, dans cette région, de seuils différenciés.

Nous avons utilisé l'algorithme du simplexe ZX3LP programmé dans la librairie IMSL, ainsi que le programme ZSPOW de résolution des systèmes d'équations non linéaires. Les variances des tirages stratifiés et systématiques sont issues de simulations (280 tirages indépendants).

Le tableau IV donne les coefficients de corrélation linéaire entre la variable de taille et les variables d'intérêt.

Les principales conclusions sont les suivantes :

a) Les coefficients de variation de P1 ont des ordres de grandeur comparables à ceux des estimateurs par régression : même si certaines variables sont nettement améliorées par l'une ou l'autre des méthodes, l'amélioration d'ensemble apportée par la régression reste globalement peu importante.

La comparaison directe des coefficients de variation doit être tout de même tempérée par les remarques suivantes :

Les résultats de la méthode du simplexe sont fournis pour des variables qui sont intégrées dans le critère d'optimisation, ce qui, par construction les "avantage" par rapport à l'estimation de la régression. En réalité, il faudrait exclure la variable d'intérêt du critère d'optimisation pour obtenir des chiffres directement comparables. Cependant, à part le cas de VIEUX en RG93, les variables absentes du critère INDUS, POPBAC et VIEUX ont des comportements comparables dans les deux méthodes. Par ailleurs et surtout, les variables les plus nombreuses seront des variables non contrôlées directement, mais assez bien corrélées avec l'une au moins des variables contrôlées, auquel cas le programme P1 devrait pouvoir approcher les précisions dues à la régression. Cependant, et cela est à l'avantage de P1, les coefficients de régression utilisés ici ont été calculés sur l'ensemble des 55 ou des 61 cantons de la région. Il s'agit donc de paramètres vrais optimaux issus d'une connaissance exhaustive que l'on a jamais en pratique (par définition). L'argument majeur à ce niveau est encore une fois qu'avec nos 3 cantons

échantillonnés nous n'aurions jamais pu obtenir la moindre estimation de ces coefficients de régression ! Au niveau théorique, pour de gros échantillons et sur les variables auxiliaires bien connues pour chaque individu, il n'en reste pas moins que l'estimateur par régression demeure imbattable de par sa construction même.

b) Par rapport aux tirages systématiques et selon la méthode de Sunter, algorithmes n'utilisant pas d'information auxiliaire autre que les tailles (lignes 7 et 8), ainsi que par rapport au tirage stratifié qui intègre une information synthétique au travers de la première composante principale (ligne 9), l'amélioration apportée par les algorithmes d'optimisation est considérable.

On remarquera que le tirage systématique sur un fichier trié selon la taille donne tout de même de bons résultats : cela n'est pas surprenant, car ce type de tirage est à peu près équivalent à un tirage stratifié selon la taille.

c) Le critère (5) favorise les variables à fort total, et on voit que pour celles qui ont un total plutôt faible, on peut avoir intérêt à utiliser le critère (6) du type chi-deux pour obtenir des gains (surtout sensible sur INDUS). Par contre, le critère (7) qui accorde la même importance à chaque variable n'est presque jamais meilleur que le critère (6).

La ligne 4 permet de voir que, sans dégrader nettement la qualité des estimations des variables autres que INDUS, on peut compenser presque totalement la complication du ratio ou de la régression par un tirage d'échantillon auquel on affecte directement des probabilités bien choisies : dans les 2 cas, la population totale PENQ est estimée de façon parfaite ou presque parfaite.

d) Une variable constante, ou peu dispersée par "continuité" est aussi bien voire mieux estimée par P1, P2, P3 que par un tirage systématique simple sur fichier trié ou par un tirage selon l'algorithme de Sunter (sur un fichier lui aussi trié selon la taille). Le schéma optimum est, dans ce cas, une stratification selon un critère corrélé à la taille. Il n'est pas surprenant d'obtenir des précisions très médiocres sur des variables à faible variance dans la population avec les programmes P1, P2 et P3 puisque les critères déterminant la composition des échantillons ne font intervenir que des variables assez corrélées à la taille. On peut constater qu'il n'est pas rare que l'échantillon comprenne une petite unité, une moyenne et une grosse. Un tirage parfaitement équilibré sur la constante vérifierait :

$$\frac{1}{T} = \frac{1}{n} \cdot \sum_{i \in S} \frac{1}{T_i} \quad , \quad \forall S \in S_2$$

ce qui se produit plus facilement avec une telle structure d'échantillon. Cette condition pourrait d'ailleurs être utilisée comme critère de sélection dans (2), sachant qu'alors, d'après

l'inégalité de Jansen, l'échantillon serait constitué d'unités dont la taille moyenne deviendrait supérieure à celle que l'on rencontre dans l'ensemble de la population.

e) les probabilités de tirage $p(s)$ issues de P1, P2 et P3 sont sans corrélation avec la valeur du critère $C(s)$. Cela tient encore à la composition des échantillons : les $p(s)$ sont soumis à de fortes contraintes pour respecter les probabilités d'inclusion, alors que la qualité des échantillons a peu de rapport avec la taille des unités qui les constituent.

f) les programmes P2 et P3 sont équivalents, avec une légère supériorité pour P2 (malgré la mise à zéro d'autorité des $p(s)$ négatifs). Ces 2 programmes donnent des résultats moins satisfaisants que ceux de la méthode du simplexe; en ce qui concerne la constante, l'amélioration qu'ils apportent est négligeable. Ce résultat est logique pour les variables liées à la taille, mais plus surprenant pour les variables peu dispersées.

g) Pour les variables non prises en compte dans le critère et médiocrement corrélées avec la taille, le critère (5) du simplexe est presque toujours à conseiller.

h) Pour les programmes P2 et P3, les estimateurs $\hat{X}(s)$ ont une distribution d'allure symétrique (voir figure II). Pour P1, cette question est moins pertinente : le nombre d'échantillons vérifiant $p(s) > 0$ est égal à la taille de la population, car on se situe sur un sommet du convexe formé par les contraintes. Le tableau (III) fournit un exemple d'échantillons possibles et de leurs estimateurs associés. Les valeurs extrêmes sont soulignées. En un certain sens, on peut dire que P1 conduit à la constitution d'échantillons d'unités-types selon un plan de sondage rigoureusement aléatoire.

i) L'influence des seuils $l(x)$ et du nombre maximum d'échantillons demandés est assez faible. On peut prendre conscience de l'imprécision attachée aux tirages non équilibrés en constatant que la sélection d'échantillons constituant S2 devient impossible si tous les seuils sont amenés à 10% en RG74, ou à 15% uniformément en RG93. Si, par contre, les seuils sont augmentés, on ne gagne rien, et la liste S2 n'est pas modifiée : en effet, les échantillons correspondant à des $C(s)$ faibles sont aussi ceux qui ont des biais relatifs faibles au sens de (2). On peut même constater qu'en limitant la taille de S2 à 500, on ne perd globalement presque rien sur une taille limite de 2 000, et que, pour certaines variables, la première situation est même préférable. Lorsqu'on utilise P2, on obtient presque systématiquement une dégradation des coefficients de variations lorsqu'on admet davantage d'échantillons (seuils fixés).

j) Du point de vue pratique, la constitution des échantillons représentant les individus proportionnellement à leur taille a l'intérêt de faire ressortir ceux qui contribuent le plus à la variance : il suffit pour cela de mesurer l'écart entre le nombre

d'échantillons prévus contenant l'individu et le nombre effectivement présent. En RG93, par exemple, le canton 0404 devrait être présent dans 311 échantillons sur les 2 000 demandés - En réalité, seulement 21 le contiennent; il est facile de voir que ce décalage est dû au taux anormalement fort de résidences secondaires, et, dans ce cas, le tirage proportionnel à la taille en nombre de logements n'est pas recommandé. Outre l'aspect opérationnel du système, nous avons constaté que, si on se limite à une simple liste d'échantillons triés par critère $C(s)$ croissant, l'ensemble des $p(s)$ vérifiant les contraintes est très fréquemment vide : Il en ressort qu'il est impossible dans ce cas de réaliser un tirage proportionnel à la taille (quelque soit l'algorithme utilisé) qui garantisse une représentativité "optimale" au sens de (2); autrement dit, il est nécessaire d'autoriser un certain nombre d'échantillons peu satisfaisants si on s'en tient à cette méthode de tirage.

k) Il n'y a pas de relation entre la taille de l'individu et la qualité des échantillons dans lesquels il apparaît; en RG93, où les tailles sont fortement dispersées, les petites unités se trouvent même en majorité dans les échantillons les plus équilibrés.

l) Pour de nombreux couples (i, j) , on a :

$$P_{ij} = 0,$$

ce qui interdit malheureusement toute estimation sans biais de la variance. Ce résultat est directement issu du mode de constitution des échantillons, où on génère volontairement des effets de grappe. La mise en oeuvre d'algorithmes autorisant toutes les associations 2 à 2 d'individus ne permet pas d'éviter des "catastrophes" avec l'estimateur des sommes dilatées. On peut donc voir l'absence d'estimateur de variance comme le prix à payer pour se protéger de regroupements malencontreux au sein du même échantillon. Cependant, le problème de la présence des P_{ij} nuls devrait s'atténuer lorsque la taille de l'échantillon augmente.

On peut enfin noter que la connaissance que nous avons des $p(s)$ nous permet de calculer les probabilités d'inclusion à n'importe quel ordre, ce qu'infiniment peu d'algorithmes autorise.

On peut donc dire, en résumé, qu'au sein d'une classe de méthodes d'échantillonnage a priori favorables à l'estimation de totaux de variables assez bien corrélées à la probabilité d'inclusion, on peut réaliser, avec les trois techniques précédentes, des gains de variance importants approchant ceux qui résultent d'une estimation par régression. L'avantage est aussi de pouvoir utiliser un estimateur sans biais d'une grande simplicité et des méthodes de programmation standards, sans dégrader les estimations de totaux de variables peu liées à la taille, mais en contrepartie sans pouvoir estimer la variance. Sauf si la configuration initiale des données se caractérise par la présence d'unités très atypiques, auquel cas les estimateurs seront de

toutes façon très imprécis compte tenu des tailles d'échantillons envisagées ici, le calcul des probabilités optimales attachées à des échantillons représentatifs est possible à moindre coût, et donne lieu à des estimateurs d'allure Gaussienne.

Pour pouvoir réaliser automatiquement la méthode du simplexe P1, il existe une procédure informatique sur l'ordinateur 3090 de l'INSEE qui traite l'ensemble des opérations en un temps inférieur à 3 minutes dès lors qu'il y a moins de 426010 combinaisons à examiner si la taille de l'échantillon est 4 ou moins, ou qu'il y a moins de 294930 combinaisons à examiner si la taille de l'échantillon est supérieure à 4. Au delà de ces limites, on retient les 65540 meilleurs échantillons parmi les 426010 (ou 294930) premiers générés, et, grâce à une recherche dichotomique, on insère éventuellement les échantillons de rang supérieur parmi les 65540 meilleurs. On a donc, à chaque instant, les meilleurs échantillons en mémoire centrale en limitant constamment la taille des tableaux manipulés.

La contrepartie réside dans le temps CPU consommé, qui augmente très vite dès lors qu'il y a davantage d'échantillons à insérer (la génération proprement dite est très peu coûteuse, mais la recherche dichotomique l'est beaucoup). On relève, par exemple :

- $n=4$, $N=58$ ou $n=6$, $N=26$, ou $n=7$, $N=23$: quelques secondes CPU
- $n=7$, $N=25$ avec 20905 insertions : 8 minutes et 2 secondes CPU
- $n=7$, $N=26$ avec 42675 insertions : 16 minutes , 27 secondes CPU

IV) EXTENSIONS :

a) Connaissant les coefficients de régression B_k de Y sur les X^k , on peut tenir compte de cette information dans le critère lui-même. Si :

$$Y_i = \sum_{k=1}^K B_k \cdot X_i^k + U_i, \quad E(U_i) = 0, \quad V(U_i) = \sigma^2$$

on résoud P_1 avec :

$$C(s) = \sum_{k=1}^K B_k^2 \cdot (\hat{X}_{(s)}^k - X^k)^2 + 2 \cdot \sum_{k < k'}^K B_k \cdot B_{k'} \cdot (\hat{X}_{(s)}^k - X^k) \cdot (\hat{X}_{(s)}^{k'} - X^{k'})$$

En RG93, sur les 4 variables JEUNES, CADR, PENQ et VIEUX, on obtient les coefficients de variation respectifs : 3,93 % , 9,50% , 1,54% et 11,57%. Il peut donc y avoir amélioration du critère, mais la connaissance des B_k nous fait toujours privilégier l'estimateur par la régression.

b) Pour limiter le cas des P_{ij} nuls, nous avons essayé d'augmenter les seuils et le nombre d'échantillons autorisés. Ce type de tentative est un échec car il existe encore trop d'associations dont le biais selon (2) est fort, voire très fort (Du moins pour une faible taille d'échantillon). En modifiant l'algorithme de façon à sélectionner un maximum de couples (i, j) , on peut diminuer assez nettement le nombre de P_{ij} nuls. Dans ce cas, sans détériorer beaucoup les coefficients de variation, de plus en plus d'échantillons fournissent des estimateurs de variance positifs proches des vraies variances. Il subsiste encore malgré tout de nombreux cas d'estimateurs négatifs, même avec des seuils $l(x)$ de 50% :

Comme :

$$V = \sum_{\{i,j/P_{ij}>0\}} (P_i P_j - P_{ij}) \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2 + \sum_{\{i,j/P_{ij}=0\}} P_i P_j \cdot \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2$$

on estime sans biais :

$$V = \sum_{\{i,j/P_{ij}=0\}} P_i P_j \cdot \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2$$

La présence d'estimateurs négatifs peut s'expliquer par le fait que les couples (i, j) pour lesquels P_{ij} est nul sont précisément ceux pour lesquels

$$\left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2$$

est grand.

c) Si la taille de la population est trop importante, il est envisageable de former, par k tirages systématiques de PAS égal à k sur le fichier trié par taille, et dont les premières unités respectives seraient $1, 2, \dots, k$, k sous-populations de taille N/k qui aient une structure voisine par construction. Les procédés précédents pourraient être appliqués sur chacune des sous-populations au niveau seulement de la génération des échantillons et du calcul du critère. On contournerait ainsi la principale difficulté actuelle qui réside dans l'impossibilité de traiter en une seule fois un trop grand nombre d'échantillons. Les k listes seraient ensuite fusionnées, et l'optimisation réalisée normalement avec les contraintes habituelles. Il ne fait aucun doute que l'on obtienne ainsi une liste de $p(s)$ optimisés, mais la perte éventuelle d'efficacité par rapport à la génération d'échantillons sur l'ensemble de la population reste à mesurer.

d) Les critères P_1, P_2 et P_3 pourraient être utilisés pour déterminer un plan de sondage lié, par exemple, au problème suivant : on veut déterminer 2 échantillons-maître disjoints dans une même population, de façon à pouvoir considérer que les 2 échantillons de tailles respectives n et m sont tirés proportionnellement à la taille. On aimerait, autant que possible, que les unités respectives des 2 échantillons ne soient pas trop distantes géographiquement, à la fois au sein du premier échantillon et d'un échantillon à l'autre.

On veut déterminer $p(s_1)$, probabilité de tirage de l'échantillon s_1 constituant le premier échantillon -maître parmi une liste d'échantillons possibles $S_2(a)$ contrôlés selon les critères déjà vus auxquels on rajoute une mesure de proximité géographique entre les unités primaires de s_1 . Pour chaque s_1 envisageable, on décide de retenir, pour le 2ème échantillon-maître, l'unité i n'appartenant pas à s_1 avec la probabilité conditionnelle $P(i|s_1)$. Si X_i est la taille de l'unité primaire i , les contraintes sont :

$$\left\{ \begin{array}{l} \sum_{s_1 \in S_2(a)} p(s_1) \cdot P(i|s_1) = m \cdot \frac{X_i}{X} \quad , \forall i \in \{1, \dots, N\} \\ \sum_{s_1 \in S_2(a)} p(s_1) = n \cdot \frac{X_i}{X} \quad , \forall i \in \{1, \dots, N\} \\ \sum_{1 \leq i \leq N} P(i|s_1) = m \quad , \forall s_1 \in S_2(a) \end{array} \right.$$

Etant donnée une mesure de proximité $d(i, s_1)$ de chaque unité primaire i à s_1 , on pose :

$$\forall i, \forall s_1 : P(i|s_1) = \frac{\lambda(s_1)}{d(i, s_1)} \cdot \mathbb{1}[i \notin s_1]$$

Si $C(s_1)$ est un des critères précédemment utilisés, on peut résoudre :

$$\begin{array}{l}
 \text{Min } \sum_{s_1 \in S_2(a)} p(s_1) \cdot C(s_1) \\
 \text{s c } \left\{ \begin{array}{l}
 \sum_{\substack{j \neq i \\ s_1 \in S_2}} p(s_1) = n \cdot \frac{X_i}{X}, \quad \forall i \in \{1, \dots, N\} \\
 \sum_{\substack{s_1 \in S_2(a) \\ j \neq i}} p(s_1) \cdot \frac{\lambda(j, s_1)}{d(i, s_1)} = m \cdot \frac{X_i}{X}, \quad \forall i \in \{1, \dots, N\} \\
 \sum_{\substack{1 \leq i \leq N \\ i \neq s_1}} \frac{\lambda(i, s_1)}{d(i, s_1)} = m, \quad \forall s_1 \in S_2(a) \\
 0 \leq p(s_1), \quad \forall s_1 \in S_2(a)
 \end{array} \right.
 \end{array}$$

La 3ème contrainte d'égalité ne sert qu'à déterminer $\lambda(s_1)$, et l'ensemble des contraintes d'égalité reste linéaire en $p(s_1)$ avec $2N$ équations.

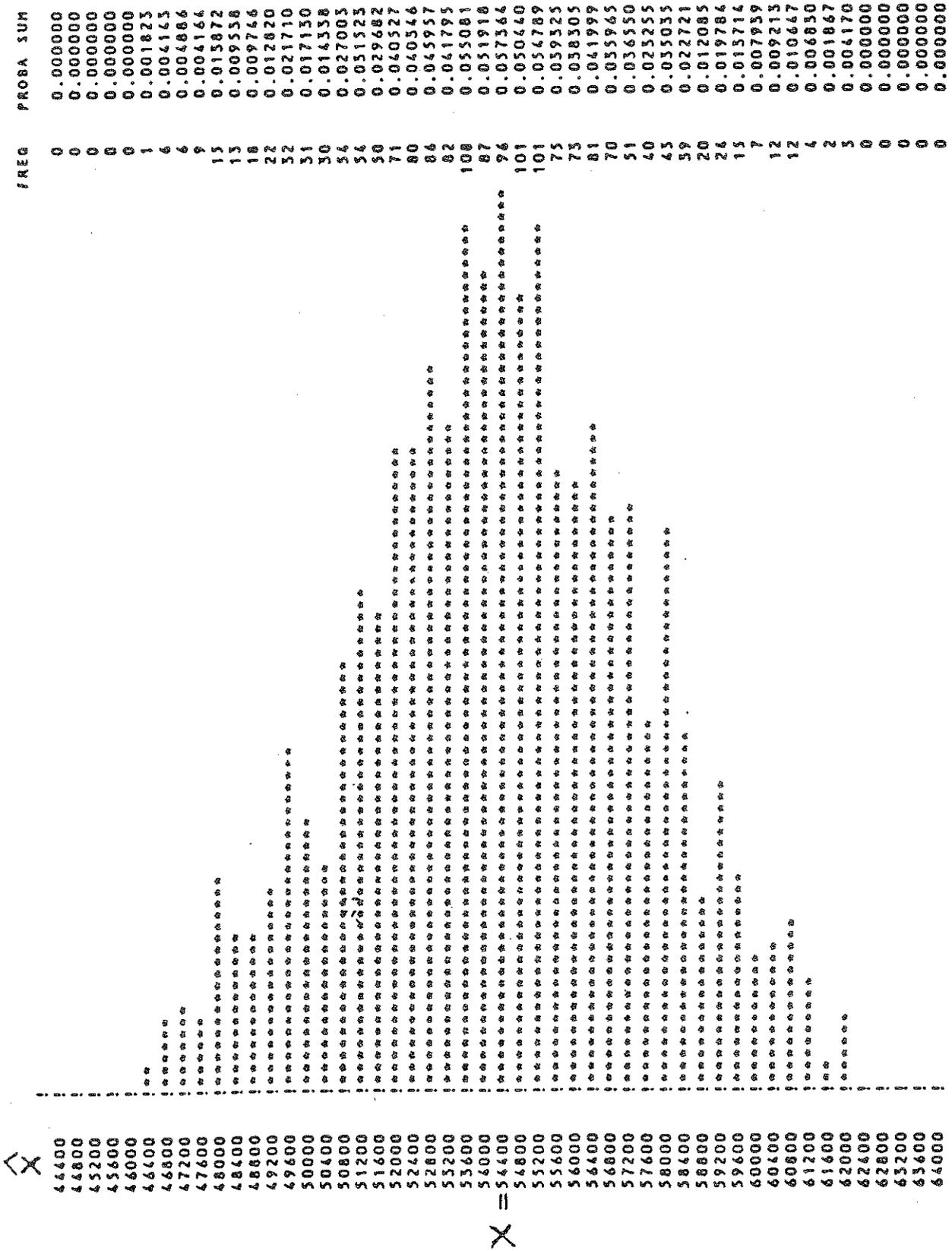
BIBLIOGRAPHIE

- (1) COCHRAN, W (1977), Sampling Techniques, J.WILEY, New York
- (2) GROSBAS, JM (1987), Méthodes Statistiques des Sondages, Economica
- (3) BREWER, K.R.W-HANIF, N (1983), Unequal Probability sampling, Springer-Verlag
- (4) HARTLEY, H.O - RAO, J.N.K, (1962), "Sampling with unequal probabilities and without replacement"- Annals of Mathematical Statistics (33)
- (5) CONNOR, W.S (1966), An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement, JASA (61)
- (6) SUNTER, A.B (1986), Solutions to the problem of unequal probability sampling without replacement, Inst.Statist. Revue (54)
- (7) GOURIEROUX, Ch(1981), Théorie des Sondages, Economica
- (8) DEVILLE J.C-SARNDAL C.E., (1990), Soumis à JASA
- (9) HAJEK K.J (1981), Sampling from a finite population, Dekker
- (10) ROYALL, R - HERSON, J (1973), Robust Estimation in Finite Populations, JASA (68)
- (11) DEVILLE J.C - GROSBAS J.M - ROTH N., (1988), Efficient Sampling Algorithms and Balanced Samples, Compstat
- (12) CIARLET, P.G (1985), Introduction à l'analyse numérique matricielle et à l'optimisation, MASSON
- (13) ARDILLY P (1989), Principe de tirage des enquêtes-ménages en France entre 1984 et 1992 , STATECO

	ACTIF	JEUNES	AGRI	CADR	INDUS	MIGR	NBRE SSCE	OUVR	PENQ	POP/BAC	VIEUX	CSTE
TABLEAU 1												
SIMPLEXE (5)	3,01	5,98	9,88	10,15	20,40	8,18	9,37	4,37	1,38	28,43	11,57	85,56
SIMPLEXE (6)	3,75	6,35	9,72	8,61	11,49	6,95	9,25	4,10	2,12	24,58	12,12	82,71
SIMPLEXE (7)	4,08	8,09	10,32	8,91	14,46	8,29	9,70	6,03	2,95	24,25	11,48	83,40
SIMPLEXE	4,19	7,38	12,37	11,59	19,87	11,27	9,35	7,25	0,79	26,26	12,97	86,09
P2 (5)	4,62	6,71	9,49	8,82	18,95	8,47	8,81	6,36	2,62	27,42	11,37	85,80
P3 (5)	4,80	6,76	11,82	10,86	20,25	10,19	9,83	6,46	2,96	25,25	12,07	81,05
SYSTEM	12,58	15,94	26,60	16,39	25,20	13,12	26,31	15,00	11,72	23,59	13,35	152,35
SUNTER	18,04	22,47	37,59	20,08	33,08	22,12	41,53	20,91	18,50	28,23	17,80	81,60
STRATES	18,33	23,46	39,42	20,65	34,37	23,01	41,62	21,34	18,77	28,48	17,93	50,12
RATIO	5,04	6,90	28,90	11,50	17,00	15,10	26,30	8,50	5,10	21,40	12,30	179,14
REGRESSION	3,05	2,91	11,20	7,18	20,85	7,06	36,26	3,97	1,21	26,71	4,10	

TABLEAU 1 REGION 74	ACTIF	JEUNES	AGRI	CADR	INDUS	MIGR	NBRE SSCE	OUVR	PENQ	POPBAC	VIEUX	CSTE
SIMPLEXE (5)	2,15	3,44	4,70	8,25	17,28	4,21	5,26	3,42	0,75	26,16	4,48	13,01
SIMPLEXE (6)	2,59	3,32	4,24	6,29	8,21	4,43	4,56	3,22	1,50	29,06	4,39	15,03
SIMPLEXE (7)	3,48	4,81	5,57	6,49	10,73	6,03	6,51	4,99	3,18	26,54	6,20	16,26
SIMPLEXE	4,14	5,37	8,02	6,60	21,77	6,39	7,38	6,97	0,09	26,50	5,01	17,30
P2 (5)	3,15	4,56	5,90	7,16	18,10	5,18	7,03	4,86	1,55	30,24	5,38	15,81
P3 (5)	3,43	5,05	6,46	7,30	18,69	5,39	7,36	5,49	1,85	30,12	5,42	15,00
SYSTEM	6,98	10,63	11,38	9,94	21,73	10,46	15,87	9,63	5,75	31,43	5,86	16,96
SUNTER	10,15	13,21	16,27	12,62	28,55	9,91	17,03	15,89	7,36	34,31	5,97	16,87
STRATES	8,74	18,33	16,85	11,43	27,09	10,17	13,25	13,97	6,29	33,75	6,27	8,11
RATIO	3,74	6,75	10,69	8,55	17,12	8,77	15,87	5,10	3,93	26,82	5,86	18,09
REGRESSION	1,45	3,11	4,89	5,06	14,51	6,78	17,30	3,22	1,35	37,85	2,57	

FIGURE II



- LOI DE L'ESTIMATEUR "JEUNES" AVEC P2 - (R6 = 74) -

Tableau III (suite) :

Valeurs des vrais totaux par variable en région 94

AGRI	JEUNES	CADR	ACTIF	INDUS	MIGR	NBRE SSCE	OUVR	POP8AC	VIEUX	PENQ
32 064	54 352	14 300	102 964	17 580	64 680	28 990	45 904	1 072	85 044	259 684

Tableau IV

	ACTIF	JEUNES	AGRI	CADR	INDUS	MIGR	NBRE SSCE	OUVR	PENQ	POPBAC	VIEUX
RG 74	0,87	0,80	0,67	0,86	0,65	0,85	0,56	0,78	0,93	0,38	0,95
RG 93	0,80	0,73	0,60	0,81	0,58	0,78	0,68	0,76	0,78	0,71	0,78

Corrélation entre la taille

et les variables d'intérêt