LE LOGICIEL POULPE : ASPECTS MÉTHODOLOGIQUES

Nathalie Caron

Sommaire

Introduction 174

1ère partie - Traitement des sondages « directs » 176
I. Notations 176
II. Principaux types de sondage directs 177

2ème partie - Traitement des sondages décomposés 180
I. La stratification 180
II. Le tirage à plusieurs degrés 180
III. Les sondages en plusieurs phases 181
IV. Le calcul du Design Effect 188

3ème partie - Traitement des statistiques complexes 191
I. Notations 191
II. Linéarisation de θ 191
III. Applications 193

4ème partie - La prise en compte de la non-réponse totale 194
I. Traitement de la non-réponse dans une enquête en une phase au niveau de l'échantillonnage 194
II. Traitement de la non-réponse dans une enquête en deux phases au niveau de l'échantillonnage 195

5ème partie - La prise en compte du redressement par CALMAR 196
I. Estimateurs en présence de redressement par calage 196
II. Estimation de variance en l'absence de non-réponse 197
III. Estimation de variance en présence de non-réponse corrigée explicitement 198
IV. Estimation de variance en présence de non-réponse corrigée implicitement par CALMAR 198

Bibliographie 199

Introduction

Le logiciel POULPE (Programme Optimal et Universel pour la Livraison de la Précision des Enquêtes) écrit en langage macro SAS, permet d'évaluer la précision de statistiques issues d'enquêtes par sondage complexes, en particulier les enquêtes auprès des ménages ou des entreprises réalisées par l'Insee. Son utilisation suppose de pouvoir décrire rigoureusement le plan de sondage et de disposer des données permettant de calculer les probabilités d'inclusion.

Outre le traitement des données issues de plans de sondage classiques comportant un nombre arbitraire de degrés de tirage, de strates et des procédures de tirage diverses (probabilités inégales, tirages systématiques,...), le logiciel POULPE intègre aussi le traitement des enquêtes en plusieurs phases, la prise en compte de la correction de la non-réponse ainsi que celle du redressement par le logiciel CALMAR.

Le type de statistique dont le logiciel POULPE est capable de chiffrer la variabilité est général. Ainsi, la statistique peut être le total d'une variable ou une statistique complexe, fonction de plusieurs totaux de variables (moyennes, ratios,...). Dans ce dernier cas, la méthode de linéarisation, programmée dans le logiciel, permet de se ramener à l'estimation de la variance d'un total d'une variable synthétique.

Ce papier constitue une version courte d'un document plus ambitieux qui recense l'ensemble des principaux éléments méthodologiques utilisés ou développés par l'UMS (Unité Méthodes Statistiques) pour mettre au point la première version du logiciel. Ce document paraîtra prochainement dans la série « Méthodologie Statistique » des documents de travail de l'Insee.

Par rapport aux autres logiciels présents sur le marché permettant d'évaluer la précision des enquêtes par sondage, le logiciel POULPE présente les spécificités suivantes :

- des formules d'estimation de la variance utilisables dans le cas d'un sondage « direct » : sondage aléatoire simple à probabilités égales, sondage systématique, sondage équilibré, sondage aléatoire simple à probabilités inégales. Dans ce dernier cas, on utilise une formule d'estimation de variance qui ne nécessite pas la connaissance des probabilités d'inclusion double ;

- *l'utilisation de la récursivité* : les formules récursives d'estimation de variance dans les plans de sondage complexes permettent d'estimer la variance d'un estimateur en appliquant successivement les formules d'estimation de variance

dans le cas d'un sondage « direct » aux différents « branches » de l'arbre décrivant le plan de sondage ;

- *le traitement des enquêtes en deux phases* dans le cas où la première phase est quelconque et la seconde phase est un sondage poissonnien ou un sondage stratifié. Cette particularité du logiciel est très importante pour l'INSEE ; en effet, beaucoup d'enquêtes ménages ont un plan de sondage à deux phases avec une sur ou sous représentation d'une partie des logements au moment du tirage. De plus, en assimilant le traitement de la non-réponse totale par repondération à une phase supplémentaire, cette spécificité nous permet de prendre en compte la correction de la non-réponse ;

- *le traitement des enquêtes en trois phases* dans le cas où la première phase est quelconque, la seconde phase est un sondage stratifié et la troisième est un sondage poissonnien.

1ère partie - Traitement des sondages « directs »

On appelle sondage « direct », les algorithmes qui extraient d'un fichier un échantillon ayant certaines propriétés. Ce terme s'oppose à celui de « sondage décomposé » où l'échantillonnage est décomposé en plusieurs sous-échantillonnages. Dans un ultime sous-échantillonnage (c'est-à-dire dans le sous-échantillonnage conduisant à la sélection des unités enquêtées), on réalise un sondage « direct ». La sélection d'unités d'échantillonnage non ultimes se fait également par ce type de sondage.

Après avoir défini les notations utilisées, nous décrirons les différents types de sondage « directs » disponibles dans le logiciel.

I. Notations

La variable d'intérêt est notée *Y* et nous souhaitons estimer son total $Y = \sum_1^N Y_i$.

On notera $\pi_i = P(i \in \text{échantillon})$, $\pi_{ij} = P(i \text{ et } j \in \text{échantillon})$, les probabilités d'inclusion simple et double.

L'estimateur classique d'Horvitz Thompson défini par $\hat{Y}_\pi = \sum_{i \in s} \frac{y_i}{\pi_i}$ est un estimateur sans biais de Y (c'est-à-dire que la moyenne pondérée des valeurs de cet estimateur obtenues sur tous les échantillons possibles de taille n correspond à la vraie valeur Y).

Sa variance et son estimation de variance sont respectivement :

$$V\left(\hat{Y}_\pi\right) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \text{ et}$$

$$\hat{V}\left(\hat{Y}_\pi\right) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

II. Principaux types de sondage directs

Trois principaux types de sondage directs sont disponibles dans le logiciel POULPE : le sondage aléatoire simple, le sondage systématique et le sondage à probabilités inégales.

Sondage aléatoire simple

Dans le cas d'un sondage aléatoire simple sans remise, les formules de variance et d'estimation de variance se simplifient et deviennent :

$$V\left(\hat{Y}_\pi\right) = N^2(1-f)\frac{S^2}{n} \text{ où } S^2 = \frac{\sum_{i\in U}\left(y_i - \overline{Y}\right)^2}{N-1}$$

$$\hat{V}\left(\hat{Y}_\pi\right) = N^2(1-f)\frac{s^2}{n} \text{ où } s^2 = \frac{\sum_{i\in s}\left(y_i - \overline{y}\right)^2}{n-1} \text{ et } f = \frac{n}{N}.$$

Sondage systématique

D'après la théorie de l'échantillonnage, comme un sondage systématique est la réalisation d'un sondage en grappe où l'on ne sélectionne qu'une seule grappe, il est impossible d'estimer la variance pour ce type de tirage. Cependant, sous diverses hypothèses liées à la modélisation, il est possible d'obtenir des estimateurs corrects de la variance. Ainsi, en supposant que les données sont rangées dans le même ordre que celui précédant le tirage systématique, la formule retenue dans le logiciel POULPE est :

$$\hat{V}\left(\hat{Y}_\pi\right) = N^2(1-f)\frac{t^2}{n}$$

où $t^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}\left(y_i - y_{i+1}\right)^2$ avec y_i qui représente la valeur de la variable Y pour le ième individu du fichier .

Sondage à probabilités inégales

Le cas des tirages à probabilités inégales pose un problème beaucoup plus délicat que les types de sondage directs présentés ci-dessus. En effet, les formules

d'estimations de variance ne se simplifient pas et par conséquent présentent deux inconvénients :

- le premier est de recourir à des sommes doubles qui sont numériquement lourdes à calculer et sans doute assez instables. Ainsi, lorsque l'échantillon compte 100 individus, les sommes doubles comptent environ 5 000 termes. Cette difficulté est toutefois surmontable, surtout si on réalise qu'en pratique, on ne mettra en œuvre le tirage à probabilités inégales que pour des échantillons qui n'excèdent pas une ou deux dizaines d'unités ;

- le second inconvénient est plus délicat à résoudre. Les formules en question font en effet appel aux probabilités d'inclusion d'ordre 2 $\left(\text{les } \pi_{ij}\right)$, probabilités pour que les couples (i, j) soient dans l'échantillon. Or, sauf dans le cas de sondage à probabilités égales, les probabilités doubles π_{ij} ne sont pas connues et ne sont pas calculables pratiquement ou au prix de calculs importants.

Dans le cadre des sondages à probabilités inégales, plusieurs formules approximatives d'estimation de variance se ramenant à des sommes de carrés et ne faisant pas intervenir les probabilités d'inclusion doubles ont été comparées :

① La première formule d'approximation de la variance est due à B. ROSEN (1991)

$$\hat{\mathrm{Var}}_{\mathrm{Rosen}}\left(\hat{Y}\right)=\frac{n}{n-1}\sum_{s}\left(\frac{y_k}{\pi_k}-D\left(\frac{y}{\pi}\right)\right)^2\left(1-\pi_k\right)$$

avec $$D\left(\frac{y}{\pi}\right)=\sum_{s} a_k\frac{y_k}{\pi_k}\Big/\sum_{s} a_k$$

$$a_k=\left(1-\pi_k\right)\log\left(1-\pi_k\right)\ /\pi_k$$

② Deux autres approximations peuvent être obtenues à partir de la théorie décrite par J.-C. DEVILLE (1993) :

$$\hat{V}_1\left(\hat{Y}_\pi\right)=\frac{n}{n-1}\sum_{s}\left(1-\pi_i\right)\left(\frac{y_i}{\pi_i}-D_2\left(\frac{y}{\pi}\right)\right)^2$$

D'où $D_2\left(y/\pi\right)$ est la moyenne pondérée par les $\left(1-\pi_i\right)$ des quantités $\frac{y_i}{\pi_i}$.

$$\hat{V}_2\left(\hat{Y}_\pi\right) = \frac{1}{1-\sum_{i=1}^{s} a_i^2} \sum_{i\in s} (1-\pi_i)\left(\frac{y_i}{\pi_i} - D_2\left(\frac{y}{\pi}\right)\right)^2$$

avec $a_i = (1-\pi_i) \Big/ \sum_{i\in s} (1-\pi_i)$.

Les simulations décrites dans le document de J.-C. DEVILLE et C. VITE SAN-PEDRO (1993) indiquent que quelle que soit la taille de la population, les résultats sont satisfaisants dès que la taille de l'échantillon dépasse 8. Par contre, pour de très petits échantillons et de très petites populations, on obtient une sous-estimation de la variance pouvant atteindre 20%.

La formule retenue dans le cadre du logiciel est $\hat{V}_1\left(\hat{Y}_\pi\right)$.

2ème partie - Traitement des sondages décomposés

I. La stratification

La population est scindée en H parties (appelées strates) à partir d'informations auxiliaires. On réalise un tirage indépendamment dans chacune de ces strates. Un estimateur sans biais du total de la variable Y est $\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h$ où $\hat{Y}_h$ est un estimateur du total de la variable au sein de la strate h. Les tirages étant indépendants d'une strate à une autre, on a $V(\hat{Y}) = \sum_{h=1}^{H} V(\hat{Y}_h)$

II. Le tirage à plusieurs degrés

Dans le cas de sondage à plusieurs degrés, on utilise un système récursif dû à J. DURBIN (1955) et amélioré par D. RAJ (1966) puis par J.N.K. RAO (1975). Nous nous contenterons dans ce document d'en exposer le principe général. Plaçons-nous tout d'abord dans le cas d'un sondage en grappe. L'estimateur du total de la variable d'intérêt est : $\hat{Y} = \sum_{k \in s} y_k / \pi_k$ où s est l'échantillon des grappes et π_k la probabilité d'inclusion d'ordre 1 de la grappe k. Pour les trois types de sondage exposés dans la partie précédente, on connaît un estimateur de la variance $f(y_s)$, où $f(y_s)$ désigne une forme quadratique des y_k de s. Dans le cas d'un sondage aléatoire simple, par exemple, on a en notant N la taille de la population, n celle de l'échantillon et $\bar{y}$ la moyenne dans l'échantillon :

$$f(y_s) = N^2\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{s} (y_k - \bar{y})^2 / (n-1)$$

Plaçons-nous maintenant dans le cas d'un sondage à deux degrés. Dans ce cas, le total de la variable Y dans l'unité primaire k noté y_k est remplacé par un estimateur $\hat{y}_k$ bâti sur un échantillon s_k d'unités secondaires de l'unité primaire considérée. Un estimateur du total de la variable d'intérêt Y est donc :

$$\hat{y}_2 = \sum_{k \in s} \hat{y}_k / \pi_k .$$

D'après RAJ (1966), un estimateur de variance sans biais pour cet estimateur est

$$\hat{V}(\hat{y}_2) = f(\hat{y}_s) + \sum_s \frac{v_k}{\pi_k}.$$

où v_k est un estimateur sans biais de la variance de $\hat{y}_k$ et $f(y_s)$ est l'estimateur de variance correspondant au premier degré.

Ce principe, exposé dans le cas d'un plan de sondage à deux degrés, se généralise à un plan de sondage à n degrés à condition de le décrire par un arbre dont les nœuds sont l'Univers, ***les unités primaires, les unités secondaires*** ... A chaque nœud est associé un type de sondage noté TS (Sondage aléatoire simple à probabilités égales, inégales,...) indiquant la façon dont sont échantillonnées les données au niveau du nœud considéré (se reporter à l'article de J.-N. Petit).

Pour obtenir une estimation de la variance, il faut disposer pour chaque TS mis en œuvre dans le plan de sondage :

- d'une formule donnant un estimateur sans biais de la forme :

$$\hat{t} = \sum \frac{y_k}{\pi_k} \qquad \text{(linéaire)}$$

- d'une formule donnant un estimateur de la variance de $\hat{t}$:

$$f^{TS}(\ldots y_k \ldots) \qquad \text{(quadratique)}$$

On "remonte" l'arbre depuis les éléments terminaux jusqu'à l'Univers. Ainsi dans tout plan de sondage complexe, la variance d'un total estimé par l'estimateur d'Horvitz-Thompson peut être évaluée à partir des fonctions f relatives à chaque forme de sondage direct et de procédés récursifs qui viennent d'être décrits.

III. Les sondages en plusieurs phases

Nous détaillons dans cette partie le traitement dans POULPE des enquêtes en deux phases. Celui des enquêtes en trois phases (correspondant à un plan de sondage en deux phases dont la seconde phase est elle-même composée de deux phases) repose sur le même principe.

III.1. Principe

Un sondage en 2 phases correspond à 2 sondages successifs qui s'appliquent aux mêmes unités :

① un premier échantillon s_1 est sélectionné à l'aide des techniques proposées précédemment. On suppose qu'une information supplémentaire est disponible pour toutes les unités de s_1 ;

② un échantillon s_2 est alors tiré de s_1 en utilisant les techniques précédentes.

La réalisation d'une enquête en deux phases permet en particulier de recueillir une information auxiliaire auprès de l'échantillon 1ère phase et de l'utiliser pour optimiser le tirage de l'échantillon deuxième phase. De plus, la correction de la non-réponse globale est généralement appréhendée par une modélisation d'une phase de sondage supplémentaire. Ainsi, les enquêtes de l'Insee se modélisent en général en trois phases : la seconde phase correspond à une sur-représentation au moment du tirage de l'échantillon et la troisième correspond à une correction de la non-réponse totale.

III.2. Sondages en 2 phases

III.2.1. Cas général

1ère phase :

Un échantillon s_1, de taille n_1, est tiré d'une population U selon un plan de sondage PS_1. On note π_i la probabilité d'inclusion de l'unité i , π_{ij} la probabilité d'inclusion double des unités i et j et $\Delta^1_{ij} = \pi_{ij} - \pi_i \, \pi_j$.

2ème phase :

Un échantillon s_2, de taille n_2, est tiré de s_1 selon un plan de sondage PS_2. On note les "probabilités d'inclusion" liées à ce tirage : p_i, p_{ij} et $\Delta^2_{ij} = p_{ij} - p_i \, p_j$.

Estimateurs

On montre que (voir par exemple C.-E. SÄRNDAL, B. SWENSSON et J. WRETMAN (1992), p. 347 et suivantes) :

- $\hat{Y} = \sum_{i \in s_2} \frac{y_i}{\pi_i p_i}$ est un estimateur sans biais du total $Y = \sum_{i \in U} Y_i$

$\cdot \hat{Y} = \sum_{i \in s_2} \frac{y_i}{\pi_i p_i}$ est un estimateur sans biais du total $Y = \sum_{i \in U} Y_i$

$$\cdot V(\hat{Y}) = \sum_{i \in U} \sum_{j \in U} \Delta^1_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} + E_{s_1}\left[\sum_{i \in s_1} \sum_{j \in s_1} \Delta^2_{ij} \frac{y_i}{\pi_i \, p_i} \frac{y_j}{\pi_j \, p_j} \right]$$

= variance 1ère phase + variance 2ème phase

où E_{s_1} désigne l'espérance relativement à la loi de probabilité de s_1.

$$\cdot \hat{V}(\hat{Y}) = \sum_{i \in s_2} \sum_{j \in s_2} \frac{\Delta^1_{ij}}{\pi_{ij} \, p_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} + \sum_{i \in s_2} \sum_{j \in s_2} \frac{\Delta^2_{ij}}{p_{ij}} \frac{y_i}{\pi_i \, p_i} \frac{y_j}{\pi_j \, p_j} \quad (1)$$

= variance estimée 1ère phase + variance estimée 2ème phase

$\hat{V}(\hat{Y})$ est un estimateur sans biais de $V(\hat{Y})$.

Programmation de la formule (1) dans le logiciel

- Sous réserve d'une description complète du plan de sondage PS_1, le logiciel sait calculer des quantités de la forme :

$$\sum_{i \in s_1} \sum_{j \in s_1} \frac{\Delta^1_{ij}}{\pi_{ij} \, \pi_i \, \pi_j} z_i \, z_j = \sum_{i \in s_1} \sum_{j \in s_1} A_{ij} \, z_i \, z_j$$

$$\text{avec } A_{ij} = \frac{\pi_{ij} - \pi_i \, \pi_j}{\pi_{ij} \, \pi_i \, \pi_j}, \; A_{ii} = \frac{1 - \pi_i}{\pi_i^2}$$

Cette quantité correspond à la variance estimée du total de la variable ***Z***, pour le sondage en une phase PS_1. On rappelle que les ***termes*** A_{ij} ***ne sont en général pas explicites***, mais calculés récursivement et implicitement dans le logiciel (car les π_{ij} ne sont pas connues), à l'aide de formules, exactes ou approchées, permettant d'estimer la variance à chaque degré de tirage.

- De même, sous réserve d'une description de PS_2, le logiciel sait calculer des quantités de la forme :

$$\sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{ij}^2}{p_{ij}\,p_i\,p_j}z_i\,z_j=\sum_{i\in s_2}\sum_{j\in s_2}B_{ij}\,z_i\,z_j$$

$$\text{avec } B_{ij}=\frac{p_{ij}-p_i\,p_j}{p_{ij}\,p_i\,p_j}\quad,\quad B_{ii}=\frac{1-p_i}{p_i^2}$$

C'est donc la 1ère partie de la formule (1) qui pose problème, car elle s'écrit :

$$\sum_{i\in s_2}\sum_{j\in s_2}\frac{A_{ij}}{p_{ij}}y_i\,y_j \quad \text{ou encore :} \sum_{i\in s_1}\sum_{j\in s_1}\frac{A_{ij}}{p_{ij}}z_i\,z_j \quad (1')$$

$$\text{en posant } z_i=\begin{cases}y_i \text{ si } i\in s_2\\ 0 \text{ sinon}\end{cases}$$

Même si les p_{ij} étaient connues (ce qui n'est pas le cas dès que PS_2 est un peu complexe), l'expression (1') n'est pas calculable telle quelle par le logiciel, ni en dehors du logiciel puisque les $A_{ij}\ (i\neq j)$ **ne sont en général pas connues.**

J.C. DEVILLE (1993) a proposé des "éléments pour une solution générale" de ce problème. La complexité de la programmation à mettre en œuvre dans ce cadre a conduit à écarter temporairement cette solution, étant donné que les cas de sondage en 2 phases rencontrés dans la pratique conduisent à des simplifications dans la formule (1). Ces "cas simples" sont obtenus lorsque le sondage 2ème phase est :

- poissonnien (c'est-à-dire que chaque unité i de l'échantillon 1ère phase est sélectionnée de façon indépendante avec une probabilité connue p_i), ce qui va permettre la prise en compte par le logiciel de certains traitements de la non-réponse

- stratifié, avec un sondage aléatoire simple dans chaque strate, ce qui va permettre de traiter les enquêtes en 2 phases tirées dans l'échantillon-maître ainsi que le traitement de la non-réponse par groupes de réponse homogène.

Dans ces deux cas, le principe consiste à décomposer le terme (1) en différents termes qui peuvent être calculés soit par récursivité par le logiciel soit directement car ils ne font intervenir que des probabilités d'inclusion simples. Nous examinons successivement ces deux exemples.

III.2.2. Sondage 2ème phase Poissonnien

Un plan de sondage poissonnien conduit à des probabilités d'inclusion doubles qui vérifient :

$$p_{ij} = p_i\ p_j \text{ si } i \neq j \Rightarrow \Delta_{ij}^2 = 0 \quad \text{si } i \neq j$$

Avec cette hypothèse, la formule (1) s'écrit :

$$\hat{V}\left(\hat{Y}\right) = \sum_{i\in s_2} \sum_{j\in s_2} \frac{A_{ij}}{p_{ij}}\, y_i\, y_j + \sum_{i\in s_2} \frac{p_i(1-p_i)}{p_i} \left(\frac{y_i}{\pi_i\ p_i}\right)^2$$

$$= \sum_{i\in s_2} \sum_{j\in s_2} \frac{A_{ij}}{p_i\ p_j}\, y_i\, y_j + \sum_{i\in s_2} A_{ii}\, y_i^2 \left(\frac{1}{p_i} - \frac{1}{p_i^2}\right) + \sum_{i\in s_2} \frac{1-p_i}{p_i^2} \frac{y_i^2}{\pi_i^2}$$

$$= \text{(1a)} + \text{(1b)} + \text{(2)}$$

$$\text{(1a)} = \sum_{i\in s_1} \sum_{j\in s_1} A_{ij}\, y_i^*\, y_j^*$$

où y^* est une nouvelle variable définie sur s_1 par : $y_i^* = \begin{cases} \dfrac{y_i}{p_i} & \text{si } i \in s_2 \\ 0 \text{ sinon} \end{cases}$

Cette quantité se calcule dans la "fonction arbre" du logiciel.

(1b) et (2) qui sont des sommes simples se calculent directement à partir du fichier de données d'enquête.

La somme des deux termes 1a et 1b correspond à l'estimation de la variance 1ère phase. Le terme 2 correspond à l'estimation de la variance seconde phase. Le logiciel POULPE génère automatiquement la variable y^* à partir de la variable d'intérêt Y et calcule les trois termes.

III.2.3. Sondage 2ème phase = sondage aléatoire simple stratifié

L'échantillon 1ère phase s_1 est scindé en H strates s_{1h} $(h = 1 \cdots H)$, de tailles N_h. Dans chaque strate s_{1h} on réalise un sondage aléatoire simple sans remise avec le taux de sondage f_h, ce qui détermine un échantillon s_{2h} de taille $n_h = f_h\ N_h$. L'échantillon 2ème phase s_2 est formé par la réunion des s_{2h}.

La quasi totalité des enquêtes en 2 phases tirées dans l'échantillon-maître sont réalisées selon cette méthode, à des fins de sur-représentation de certaines catégories de logements ou de ménages (par exemple sur-représentation de certaines catégories socioprofessionnelles ou de certaines catégories de logements comme dans l'enquête "Conditions de Vie").

Probabilités d'inclusion

$$p_i = f_h \quad \text{si} \quad i \in s_{1h}$$

$$\begin{cases} p_{ii} = f_h & \text{si} \quad i \in s_{1h} \\ p_{ij} = f_h \, f_{h'} & \text{si} \quad i \in s_{1h} \text{ et } j \in s_{1h'} \,, h \neq h' \\ p_{ij} = f_h \dfrac{n_h - 1}{N_h - 1} = f_h^2 \Big/ \left(1 + \dfrac{1 - f_h}{n_h - 1} \right) & \text{si} \quad i \text{ et } j \in s_{1h} \,, i \neq j \end{cases}$$

Le premier type de probabilité d'inclusion double correspond au cas où les unités i et j appartiennent à des strates de seconde phase différentes ; le second au cas où i et j appartiennent à la même strate.

Estimateurs

$$\hat{Y} = \sum_h \sum_{i \in s_{2h}} \frac{y_i}{\pi_i \, f_h} = \sum_h N_h \left[\frac{1}{n_h} \sum_{i \in s_{2h}} \underbrace{\frac{y_i}{\pi_i}}_{= \ell_i} \right] = \sum_h N_h \, \bar{\ell}_h$$

La variance estimée de cet estimateur s'écrit :

$$\hat{V}\left(\hat{Y}\right) = \sum_{i \in s_2} \sum_{j \in s_2} \frac{A_{ij}}{p_{ij}} y_i \, y_j + \text{variance estimée pour un SAS stratifié de la variable } \frac{y_i}{\pi_i} = \ell_i$$

$$= \sum_{i \in s_2} \sum_{j \in s_2} \frac{A_{ij}}{p_{ij}} y_i \, y_j + \sum_h N_h^2 \frac{1 - f_h}{n_h} \left(\frac{1}{n_h - 1} \sum_{i \in s_{2h}} \left(\ell_i - \bar{\ell}_h \right)^2 \right)$$

$$\textcircled{1} + \textcircled{2}$$

Calcul dans le logiciel

$$\left.\begin{aligned} (1) &= \sum_{i\in s_2}\sum_{j\in s_2} A_{ij}\,\frac{y_i}{p_i}\,\frac{y_j}{p_j} \\ &+ \sum_{h}\sum_{i\in s_{2h}}\sum_{j\in s_{2h}} A_{ij}\,\frac{y_i}{p_i}\,\frac{y_j}{p_j}\,\frac{1-f_h}{n_h-1} \end{aligned}\right\} \quad (1a)$$

$$+\sum_{h}\sum_{i\in s_{2h}} A_{ii}\; y_i^2\,\frac{n_h-N_h}{f_h(n_h-1)} \qquad (1b)$$

$$(1a) = \sum_{i\in s_1}\sum_{j\in s_1} A_{ij}\; z_i^0\; z_j^0 + \sum_{h}\sum_{i\in s_1}\sum_{j\in s_1} A_{ij}\; z_i^h\; z_j^h$$

où $z^0, z^1 \ldots z^h \ldots z^H$ sont des nouvelles variables **définies** sur s_1 par :

$$z_i^0 = \begin{cases} \dfrac{y_i}{p_i} \text{ si } i \in s_2 \\ 0 \quad \text{sinon} \end{cases}$$

$$z_i^h = \begin{cases} \dfrac{y_i}{p_i}\sqrt{\dfrac{1-f_h}{n_h-1}} \text{ si } i \in s_{2h} \quad h = 1 \ldots H. \\ 0 \quad \text{sinon} \end{cases}$$

(1b) et (2) se calculent directement à partir du fichier de données d'enquête.

La somme des deux termes 1a et 1b correspond à l'estimation de la variance 1ère phase. Le terme 2 correspond à l'estimation de la variance seconde phase. Le logiciel POULPE génère automatiquement les variables $z^0, z^1 \ldots z^h \ldots z^H$ à partir de la variable d'intérêt Y et calcule les trois termes dont seul le terme 1a s'obtient avec le principe de récursivité.

III.3. Remarques

D'après plusieurs études menées à partir de l'enquête "Conditions de Vie" réalisée par l'Insee en 1993, qui est une enquête en deux phases sur-représentant au moment du tirage les logements "défavorisés" (N. CARON (1996)), nous avons obtenu des résultats surprenants. En effet, les variances des estimateurs de paramètres d'intérêt à

faible coefficient de variation sont estimées à partir du logiciel par des quantités fortement négatives ou anormalement faibles.

Ce phénomène s'explique par le fait que pour les enquêtes en deux phases liées à une sur-représentation au niveau du tirage de l'échantillon dont la seconde phase a été réalisée par un tirage systématique tout en conservant l'ordre du tirage (c'est-à-dire unité primaire après unité primaire), nous nous plaçons dans le cas le plus défavorable au sens où l'estimation de la variance première phase est négative. En multipliant le nombre d'unités primaires, le phénomène ne fait que s'accentuer comme c'est le cas sur l'exemple de l'enquête "Situations Défavorisées". En effet, le plan de sondage de 1ère phase de cette enquête est à plusieurs degrés et son plan de sondage de 2de phase est stratifié avec réalisation d'un tirage systématique dans chaque strate. Ainsi, les variances estimées de la première phase par le logiciel pour l'estimation du nombre de femmes ainsi que pour celle du nombre d'actifs sont respectivement **$-17\ 10^9$ et $-3\ 10^9$.**

Dans la version actuelle du logiciel, la solution au problème des estimations de variance négatives est la suivante : si le logiciel POULPE a obtenu une estimation de variance (de la première phase) négative, l'utilisateur en est averti par un message.

La solution de remplacement qui sera développée dans une version ultérieure du logiciel sera la méthode de réplication. Celle-ci permettra de reconstituer artificiellement un échantillon première phase et d'évaluer sur ce dernier l'estimation de variance due à la première phase.

IV. Le calcul du Design Effect

Le logiciel POULPE calcule la précision obtenue en considérant que les données sont issues d'un plan de sondage aléatoire simple. Ce calcul permet de comparer la variance obtenue par le plan de sondage complexe à celle qu'on aurait obtenue si le plan de sondage avait été celui d'un plan de sondage aléatoire simple. Le rapport des deux estimations de variance est appelé en théorie de sondage le « design effect » (deff) ; il permet en particulier d'apprécier un effet grappe si le plan de sondage est à plusieurs degrés.

IV.1. Définition

Soit $\hat{Y}$ $\left(\text{respectivement } \hat{Y}_{SAS}\right)$ l'estimateur d'un total Y d'une variable *Y* pour un plan de sondage quelconque (respectivement un plan de sondage aléatoire simple sans remise de taille fixe égale à celle de l'échantillon effectivement disponible) ;

Par définition,

$$\text{Deff} = \frac{V\left(\hat{Y}\right)}{V_{SAS}\left(\hat{Y}_{SAS}\right)}$$

où $V\left(\hat{Y}\right)$ est la variance de $\hat{Y}$ et $V_{SAS}\left(\hat{Y}_{SAS}\right)$ est la variance de $\hat{Y}_{SAS}$ en considérant un plan de sondage aléatoire simple.

IV.2. Estimation de l'effet de sondage pour les enquêtes en une phase

Le "Design Effect" est estimé par :

$$\hat{\text{Deff}} = \frac{\hat{V}\left(\hat{Y}\right)}{\hat{V}_{SAS}\left(\hat{Y}_{SAS}\right)}$$

où $\hat{V}\left(\hat{Y}\right)$ est estimé à partir du logiciel POULPE avec le "vrai" plan de sondage et où $\hat{V}_{SAS}\left(\hat{Y}_{SAS}\right)$ est un estimateur sans biais de $V_{SAS}\left(\hat{Y}_{SAS}\right)$. On démontre qu'un « bon » estimateur de $V_{SAS}\left(\hat{Y}_{SAS}\right)$ sous un plan de sondage complexe est :

$$\hat{V}_{SAS}\left(\hat{Y}_{SAS}\right) = \left[\frac{1}{n}\left(1-\frac{n}{N}\right)\left\{N\sum_{S}\frac{y_k}{\pi_k} - \left(\sum_{S}\frac{y_k}{\pi_k}\right)^2 + \hat{V}\left(\hat{Y}\right)\right\}\right]$$

qui peut différer nettement de l'estimateur traditionnel $N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n}$ si les poids sont très différents.

Or :

$$\hat{V}\left(\hat{Y}\right) = \hat{\text{Deff}}\ \hat{V}_{SAS}\left(\hat{Y}_{SAS}\right)$$

En approximant N par $\hat{N} = \sum_s \frac{1}{\pi_k}$, et $1 - \frac{\hat{Deff}}{n}\left(1 - \frac{n-1}{N-1}\right)$ par 1, nous obtenons :

$$\hat{Deff} \approx \frac{\hat{V}(\hat{Y})}{\frac{1}{n}\left(1 - \frac{n-1}{\hat{N}-1}\right)\hat{N}\sum_s \frac{1}{\pi_k}(y_k - \bar{y})^2}$$

avec : $\bar{y} = \sum_s \frac{y_k}{\pi_k} \Big/ \sum_s \frac{1}{\pi_k}$

IV.3. Estimation de l'effet de sondage pour les enquêtes en plusieurs phases.

Pour les enquêtes en plusieurs phases, $\hat{V}_{SAS}(\hat{Y}_{SAS})$ est estimée par :

$$\frac{1}{r}\left(1 - \frac{r-1}{\hat{N}-1}\right)\hat{N}\sum_s \frac{1}{\pi_k^*}(y_k - \bar{y})^2$$

où r est le nombre d'individus dans la dernière phase et π_k^* représente la probabilité d'inclusion "totale" de l'unité k (c'est-à-dire le produit des probabilités d'inclusion qui correspondent aux différentes phases).

3ème partie - Traitement des statistiques complexes

On appelle statistique complexe toute fonction non linéaire de totaux de variables, comme un ratio, c'est-à-dire le rapport de deux totaux. Pour le traitement de leur estimateur, le logiciel procède comme le logiciel de calcul de la précision suédois CLAN par linéarisation des fonctions à estimer, à l'aide de la formule de Taylor de développement en série (approche formalisée par Woodruff en 1971). Ainsi, le principe consiste à remplacer le calcul de la précision d'une statistique complexe, par celui d'un estimateur d'un total d'une variable artificielle qui est une fonction linéaire des variables observées dont on sait estimer la variance. Ce principe est détaillé ci dessous.

I. Notations

On s'intéresse à l'estimation d'une fonction de q totaux sur la population définie par :

$\theta = f\left(Y_1 \dots Y_k \dots Y_q\right)$ où $Y_k = \sum_{i \in U} y_{ki}$ = total de la variable Y_k

On note $\hat{Y}_k = \sum_{i \in s} \frac{y_{ki}}{\pi_i}$ l'estimateur de Horvitz-Thompson de Y_k

Dès que f est une fonction non linéaire, le paramètre θ est dit "complexe". Celui-ci est estimé en remplaçant chaque total Y_k par son estimateur $\hat{Y}_k$: on obtient ainsi l'**estimateur par substitution** $\hat{\theta} = f\left(\hat{Y}_1 \dots \hat{Y}_k \dots \hat{Y}_q\right)$.

II. Linéarisation de $\hat{\theta}$

On suppose que f est une fonction dérivable, à dérivées partielles continues, et que $\hat{Y}_k - Y_k$ est une "petite" variation (en $O_P\left(1/\sqrt{n}\right)$).

On peut alors écrire :

$$\hat{\theta} - \theta = \sum_{k=1}^{q} \left(\hat{Y}_k - Y_k\right) \frac{\partial f}{\partial Y_k}\left(Y_1 \dots Y_k \dots Y_q\right) + O_p(1/n) = \sum_{k=1}^{q} a_k \left(\hat{Y}_k - Y_k\right) + O_p(1/n)$$

avec $a_k = \frac{\partial f}{\partial Y_k}\left(Y_1 \dots Y_k \dots Y_q\right)$

Le 1er terme de l'expression de droite est d'espérance nulle : $\hat{\theta}$ est donc un estimateur "approximativement sans biais" de θ (le biais est négligeable si n est assez grand).

On peut donc confondre variance et erreur quadratique moyenne, et écrire :

$$V(\hat{\theta}) \approx EQM(\hat{\theta}) = E(\hat{\theta}-\theta)^2 \approx E\left[\sum_{k=1}^{q} a_k(\hat{Y}_k - Y_k)\right]^2 = V\left(\sum_{k=1}^{q} a_k \hat{Y}_k\right) \qquad \text{puisque } E\hat{Y}_k = Y_k$$

$$= V\left[\sum_{k=1}^{q} a_k\left(\sum_{i\in s}\frac{y_{ki}}{\pi_i}\right)\right] = V\left[\sum_{i\in s}\left(\frac{1}{\pi_i}\sum_{k=1}^{q} a_k y_{ki}\right)\right] = V\left(\sum_{i\in s}\frac{z_i}{\pi_i}\right)$$

en introduisant la **variable linéarisée Z** définie par :

$$z_i = \sum_{k=1}^{q}\frac{\partial f}{\partial Y_k}(Y_1 \ldots Y_k \ldots Y_q)\; y_{ki}$$

On obtient donc : $V(\hat{\theta}) \approx \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\frac{z_i}{\pi_i}\frac{z_j}{\pi_j}$

qui, si les a_k étaient connus, serait estimée par :

$$\hat{V}(\hat{\theta}) \approx \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{z_i}{\pi_i}\frac{z_j}{\pi_j}$$

Les a_k dépendent des totaux Y_k inconnus ; on remplace ces totaux par leurs estimateurs $\hat{Y}_k$, soit :

$$\hat{a}_k = \frac{\partial f}{\partial Y_k}(\hat{Y}_1 \ldots \hat{Y}_k \ldots \hat{Y}_q)$$

et on pose $\hat{z}_i = \sum_{k=1}^{q}\hat{a}_k y_{ki}$, ce qui conduit à l'estimation de $V(\hat{\theta})$:

$$\hat{V}(\hat{\theta}) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{\hat{z}_i}{\pi_i}\frac{\hat{z}_j}{\pi_j}$$

III. Applications

• *ratio*

$$R = \frac{Y}{X} = f(X,Y), \qquad \hat{R} = \frac{\hat{Y}}{\hat{X}} = f\left(\hat{X},\hat{Y}\right)$$

$$\frac{\partial f}{\partial X}\left(\hat{X},\hat{Y}\right) = -\frac{\hat{Y}}{\hat{X}^2} = -\frac{\hat{R}}{\hat{X}}, \qquad \frac{\partial f}{\partial Y}\left(\hat{X},\hat{Y}\right) = \frac{1}{\hat{X}} \;\Rightarrow\; \hat{z}_i = \frac{1}{\hat{X}}\left(y_k - \hat{R}x_k\right)$$

• *moyenne* = cas particulier d'un ratio, où X est la variable constante égale à 1

$$\hat{z}_i = \frac{1}{\hat{N}}\left(y_k - \hat{\bar{Y}}\right)$$

4ème partie - La prise en compte de la non-réponse totale

Les enquêtes par sondage sont généralement confrontées au problème de la non-réponse. On distingue deux grands types de non-réponse: la **non-réponse totale** lorsqu'un individu échantillonné ne fournit aucune réponse à l'ensemble du questionnaire et la **non-réponse partielle** lorsqu'un individu échantillonné ne répond pas à une partie plus ou moins importante du questionnaire. Chaque type de non-réponse nécessite une technique particulière de correction. Les méthodes de repondération, principalement utilisées pour compenser la non-réponse totale, consistent à augmenter judicieusement le poids d'échantillonnage des répondants. Par contre, dans les méthodes d'imputation employées pour la non-réponse partielle, les réponses manquantes sont remplacées par une (ou plusieurs) valeur(s) "plausible(s)".

Dans la première version du logiciel, seule la correction de la non-réponse totale est prise en compte pour l'évaluation de la variance. L'idée consiste à modéliser le mécanisme de la non-réponse et à introduire une phase de sondage supplémentaire.

D'après la théorie développée par J.C. Deville dans le polycopié du cours pour les statisticiens européens (1997), les techniques usuelles de correction de la non-réponse totale peuvent être traduites par des équations de calage particulières. Réciproquement, l'utilisation de CALMAR sur un fichier dont la non-réponse n'a pas été traitée par une méthode classique de repondération permet non seulement de limiter les fluctuations dues à l'échantillonnage mais aussi de corriger la non-réponse totale. Nous parlerons dans ce cas de non-réponse corrigée implicitement par CALMAR. Ce cas sera développé dans la 5ème partie.

I. Traitement de la non-réponse dans une enquête en une phase au niveau de l'échantillonnage

Comme la présence de non-réponse est modélisée comme une phase de sondage supplémentaire, on se retrouve dans le contexte d'un ***sondage en 2 phases*** :

- ***1ère phase*** : tirage de l'échantillon total s (répondants + non-répondants) $\left(s_1 = s = r \cup \bar{r}\right)$
- ***2ème phase*** : tirage de l'échantillon des répondants $\left(s_2 = r\right)$, selon un "plan de sondage" défini par les probabilités de réponse.

  p_i = probabilité de réponse de l'unité i

  p_{ij} = probabilité pour que les unités i et j répondent.

En général, les comportements de réponse d'unités différentes sont indépendants, c'est-à-dire $p_{ij} = p_i\ p_j$ si $i \neq j$. On se trouve donc dans le cadre d'une enquête en deux phases, à cette nuance près que les probabilités de réponse p_i ne sont pas connues. L'estimation de ces probabilités par des quantités $\hat{p}_i$ peut être obtenue selon différentes méthodes qui introduisent des contraintes supplémentaires (autrement dit, la relation $p_{ij} = p_i\ p_j$ si $i \neq j$ n'est plus forcément valable au niveau des estimations) :

☆ modélisation de la forme $p_i = G\left(z_i' c\right)$, en particulier à l'aide d'un modèle logit où z_i est un vecteur de variables explicatives du comportement de réponse. Les estimations de p_i sont $\hat{p}_i = G\left(z_i' \hat{c}\right)$ où $\hat{c}$ est un estimateur convergent de c.

☆ **groupes de réponses homogènes** : La population est divisée en sous-populations supposées homogènes au sens de la non-réponse. Elles sont constituées après réalisation de l'enquête en examinant en général le critère répond / ne répond pas en fonction de variables connues pour les répondants et les non-répondants. On peut par exemple réaliser une régression logistique sur l'échantillon pour choisir les variables auxiliaires les plus explicatives de la non-réponse ainsi que pour effectuer des regroupements adéquats de modalités pour définir les sous-populations. Le taux de réponse est estimé par le rapport $f_h = \frac{r_h}{n_h}$ où r_h est le nombre de répondants dans la sous-population h et n_h le nombre d'individus de l'échantillon de la sous-population h. Ce mode d'estimation introduit des contraintes supplémentaires qui conduisent à utiliser les formules de la seconde partie en considérant que l'on a réalisé un **SAS stratifié pour la seconde phase.**

☆ redressement direct par CALMAR sur l'échantillon des répondants. Cette méthode est détaillée dans la 5ème partie.

Il suffit donc de remplacer dans les formules de la seconde partie les p_i par les $\hat{p}_i$.

II. Traitement de la non-réponse dans une enquête en deux phases au niveau de l'échantillonnage

On se place dans le cadre d'une enquête en 2 phases, où le sondage 2ème phase est un SAS stratifié, et où la non-réponse totale a été corrigée par une méthode de repondération (autre que CALMAR). La non-réponse génère une 3ème phase de sondage. Les formules à appliquer correspondent à celles d'une enquête en trois phases en estimant les probabilités d'inclusion de la 3ème phase.

5ème partie - La prise en compte du redressement par CALMAR

I. Estimateurs en présence de redressement par calage

On ne considère ici que le cas d'un calage "simple" de l'échantillon "final" sur des totaux connus sur l'ensemble de la population. CALMAR transforme les poids "initiaux" des unités, notés d_i, en poids "finaux", notés w_i, tels que les w_i soient les plus proches des d_i tout en vérifiant les "équations de calage" :

$$\sum_{i \in s} w_i \, x_i = X$$

où

$\begin{cases} x_i \text{ est un vecteur de variables auxiliaires } \left(x_i^1 \ldots x_i^k\right) \\ X \text{ est le vecteur connu des totaux de ces variables sur la population} \end{cases}$

Le rapport des poids $\frac{w_i}{d_i}$, appelé aussi facteur de calage, est de la forme :

$$\frac{w_i}{d_i} = F\left(x_i' \, b\right) = g_i$$

où $\begin{cases} F \text{ est une fonction dépendant de la méthode de calage utilisée} \\ b \text{ est un vecteur de multiplicateurs de Lagrange} \end{cases}$

L'estimateur par calage du total Y d'une variable d'intérêt est :

$$\hat{Y}_w = \sum_{i \in s} w_i Y_i$$

Selon les cas, les poids initiaux d_i sont :

- les poids de sondage $\frac{1}{\pi_i}$, où les π_i sont les probabilités d'inclusion
- les poids de sondage corrigés de la non réponse $\frac{1}{\pi_i \, p_i}$

D'après l'article de J.-C. DEVILLE et de C.-E. SÄRNDAL (1992), l'estimateur par calage est équivalent (asymptotiquement) à l'estimateur par régression, dont la variance vaut :

$$V\left(\sum_{i\in s} g_i \frac{u_i}{\pi_i}\right)$$

où u_i est le "vrai" résidu de la régression de Y sur les variables de calage, sur U.

Pour estimer la variance de $\hat{Y}_W$, il suffit donc de disposer d'une formule donnant la variance estimée du total d'une variable d'intérêt *Y*, de remplacer les y_i par les $g_i\hat{u}_i$ dans cette formule, où les $\hat{u}_i$ sont les résidus de la régression, *dans s*, de Y sur les variables de calage (où les unités i sont pondérés par les w_i). Les variables résidus sont calculées directement dans le logiciel POULPE.

Une autre variance estimée possible consiste à faire $g_i = 1$. Même si les auteurs recommandent plutôt la première formule, c'est la seconde qui est programmée dans le logiciel puisqu'elle ne requiert pas en entrée les pondérations d'extrapolation finales.

II. Estimation de variance en l'absence de non-réponse

En l'absence de non-réponse (ou en présence de non-réponse mais non corrigée, ni explicitement, ni implicitement), l'application des résultats du paragraphe précédent donne :

II.1. Sondage en une phase

$$\hat{V}\left(\hat{Y}_w\right) = \sum_{i\in s}\sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}\,\pi_i\,\pi_j}\left(g_i\,\hat{u}_i\right)\left(g_j\,\hat{u}_j\right)$$

où $g_i = w_i\,\pi_i$

II.2. Sondage en 2 phases, avec sondage 2ème phase SAS stratifié

On applique la formule donnée dans la partie 2 en remplaçant les y_i par les $g_i\,\hat{u}_i$, où $g_i = w_i\,\pi_i$.

III. Estimation de variance en présence de non-réponse corrigée explicitement

CALMAR transforme les poids $d_i = \frac{1}{\pi_i \hat{p}_i}$ (poids de sondage corrigés de la non-réponse, où les $\hat{p}_i$ sont les probabilités de réponse estimées) en w_i. On peut appliquer les résultats précédents aux cas suivants.

III.1. Sondage en une phase

On applique les formules de la partie 2, en remplaçant les y_i par les $g_i \hat{u}_i$, et les p_i par les $\hat{p}_i$.

III.2. Sondage en 2 phases, avec sondage 2ème phase SAS stratifié

On applique les formules d'une enquête en trois phases en remplaçant les y_i par les $g_i \hat{u}_i$ et en estimant les probabilités d'inclusion de la 2ème et de la 3ème phase.

IV. Estimation de variance en présence de non-réponse corrigée implicitement par CALMAR

CALMAR transforme les poids $d_i = \frac{1}{\pi_i}$ en w_i. Ce calage "direct" réalise simultanément une correction de non-réponse (et génère donc une phase supplémentaire) et une réduction de la variance (calage sur des données externes). Si on considère que les probabilités de réponse sont estimées par $\hat{p}_i^{-1} = g_i = F\left(x_i' b\right) = \frac{w_i}{d_i}$ (voir F. DUPONT (1993)), on risque d'obtenir des probabilités supérieures à 1. Une étude particulière est nécessaire pour savoir si les formules développées précédemment s'appliquent encore. Dans le cas de données corrigées implicitement de la non-réponse par CALMAR, voici la démarche proposée dans la première version du logiciel POULPE :

- considérer que la probabilité de réponse est constante et égale à $p = \frac{r}{n}$ où r est le nombre de répondants et n la taille de l'échantillon.
- traiter le fichier avec une phase supplémentaire en considérant que la dernière phase est un sondage poissonnien de paramètre p et que les poids ont été modifiés par CALMAR.

Bibliographie

CARON, N. : « Calcul de l'effet de sondage (Design Effect) dans le logiciel POULPE », note interne n°981/F410, 1996.

CARON, N. : « Estimations de variance négatives obtenues à partir de l'enquête Situations Défavorisées », notes internes n°976 et 983/F410, 1996.

DEVILLE, J.-C. : « Estimation de précision de données d'enquêtes », *document de travail Insee de la Direction des Statistiques Démographiques et Sociales* n°F9211, 1992.

DEVILLE, J.-C. : Support de cours TES (DAT 202-F) sur la non-réponse et le calage, 1997.

DEVILLE, J.-C. : « Estimation de la variance pour les enquêtes en deux phases », note interne manuscrite, Insee, 1993.

DEVILLE, J.-C, SÄRNDAL, C.-E. :« Calibration estimators in survey sampling », *JASA*, vol 87, n° 418, 1992.

DEVILLE, J.-C, SÄRNDAL, C.-E., SAUTORY, O. :« Generalized raking procedures in survey sampling », *JASA*, vol 88, n° 423, 1993.

DEVILLE, J.-C., VITE SAN-PEDRO C. : Rapport de recherche, Insee, 1993.

DUPONT, F. : « Calage et redressement de la non-réponse totale : validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989 », *Insee Méthodes,* n° 56-57-58, (Actes de journées de méthodologie Statistique de décembre 1993).

DUPONT, F. : « Éléments de spécifications pour la prise en compte de la non-réponse et des sondages en plusieurs phases dans le logiciel de calcul de précision des enquêtes effectuées par sondage », notes internes n°687/F010 et 4/F420, Insee, 1994.

DURBIN, J. : « Some results in sampling theory when the units are selected with unequal probabilities », *JRSS, serie B,* n°15, 1955.

RAJ, D. : « Some remarks on a simple procedure of sampling without replacement », *JASA*, n°61, 1966.

RAO, J.N.K. : « Unbiased variance estimation for multistage designs », *Sankhya,* C n°37, 1975.

ROSEN, B. : « Variance estimation for systematic pps-sampling », *rapport n°1991:15* de Statistique Suède, 1991

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J. : *Model Assisted Survey Sampling*, Springer-Verlag, 1992.