INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES

DIRECTION GENERALE

18, boulevard Adolphe Pinard - 75675 PARIS CEDEX 14

C. E. P. E.

La macro CALMAR2

Redressement d'un échantillon par calage sur marges

Josiane LE GUENNEC

Olivier SAUTORY

26 avril 2005

Table des matières

GENERALITES 7

1ERE PARTIE : ASPECTS THEORIQUES DU CALAGE SUR MARGES 11

I. LE PROBLEME 13

II. RESOLUTION THEORIQUE 14

III. APPLICATION AU CALAGE SUR MARGES 15

III.1 CAS DE VARIABLES CATEGORIELLES 15
III.2 CAS DE VARIABLES CATEGORIELLES ET NUMERIQUES 15

IV. CALAGE SIMULTANE POUR DIFFERENTES UNITES D'UNE MEME ENQUETE 17

IV.1 LE PROBLEME 17
IV.2 CALAGE SIMULTANE DANS UN SONDAGE EN GRAPPES 18
IV.3 CALAGE SIMULTANE DANS UN SONDAGE A DEUX DEGRES 18
IV.3.1 Deux niveaux d'observation 18
IV.3.2 Trois niveaux d'observation 20

V. LES FONCTIONS G USUELLES 22

VI. LE CHOIX DE LA METHODE 23

VII. CALAGE ET REDRESSEMENT DE LA NON-REPONSE 24

VII.1 DEFINITIONS ET PROBLEMES 24
VII.2 METHODES HABITUELLES DE REPONDERATION POUR NON-REPONSE TOTALE 24
VII.2.1 Les modèles de réponse 25
VII.2.2 Calage après redressement pour non-réponse 26
VII.2.3 Redressement direct par calage 27
VII.3 CALAGE EN L'ABSENCE D'INFORMATION SUR LES NON-REPONDANTS 28
VII.3.1 Les équations de calage 29

VIII. CALAGE ET COLINEARITES 31

VIII.1 COLINEARITES STRUCTURELLES 31
VIII.2 AUTRES COLINEARITES 31
VIII.3 LES SOLUTIONS 31

2EME PARTIE : MISE EN ŒUVRE DE LA MACRO CALMAR2 33

IX. CALAGE SIMPLE : UN SEUL NIVEAU D'OBSERVATION 35

IX.1 LES DONNEES EN ENTREE DE LA MACRO 35
IX.1.1 La table SAS contenant les données 35
IX.1.2 La table SAS contenant les variables de calage et les marges 38
IX.2 SYNTAXE DE LA MACRO 38
IX.2.1 Paramètres spécifiant les tables SAS en entrée 39
IX.2.2 Paramètres spécifiant la méthode utilisée 40
IX.2.3 Paramètres relatifs aux tables en sortie 41
IX.2.4 Paramètres spécifiant les sorties imprimées 42
IX.3 LA TABLE EN SORTIE 43
IX.4 LES SORTIES IMPRIMEES 44

X. CALAGE SIMULTANE DANS UN SONDAGE EN GRAPPES 46

X.1 LES DONNEES EN ENTREE DE LA MACRO 46
X.1.1 Les tables SAS contenant les données 46
X.1.2 Les tables SAS contenant les variables de calage et les marges 48

X.2 SYNTAXE DE LA MACRO 48
X.2.1 Paramètres spécifiant les tables SAS en entrée *48*
X.2.2 Paramètres spécifiant la méthode utilisée *50*
X.2.3 Paramètres relatifs aux tables en sortie *52*
X.2.4 Paramètres spécifiant les sorties imprimées *53*
X.3 LES TABLES EN SORTIE 54

XI. CALAGE SIMULTANE DANS UN SONDAGE A DEUX DEGRES AVEC TROIS NIVEAUX D'OBSERVATION 56

XI.1 LES DONNEES EN ENTREE DE LA MACRO 56
XI.1.1 Les tables SAS contenant les données *56*
XI.1.2 Les tables SAS contenant les variables de calage et les marges *59*
XI.2 SYNTAXE DE LA MACRO 59
XI.2.1 Paramètres spécifiant les tables SAS en entrée *59*
XI.2.2 Paramètres spécifiant la méthode utilisée *62*
XI.2.3 Paramètres relatifs aux tables en sortie *63*
XI.2.4 Paramètres spécifiant les sorties imprimées *65*
XI.3 LES TABLES EN SORTIE 67

XII. CAS PARTICULIERS DE CALAGE SIMULTANE 68

XII.1 DEUX NIVEAUX D'OBSERVATION DANS UN SONDAGE A DEUX DEGRES 68
XII.1.1 Les données en entrée de la macro *68*
XII.1.2 Syntaxe de la macro *68*
XII.2 CONTRAINTE D'EGALITE DES POIDS DANS LA GRAPPE, SANS DONNEES SUR LES GRAPPES 69

XIII. CALAGE GENERALISE POUR REDRESSEMENT DE LA NON-REPONSE 71

XIII.1 UN NOUVEAU PARAMETRE : NONREP 71
XIII.2 LA STRUCTURE D'UNE TABLE DE MARGES 71
XIII.3 LA STRUCTURE D'UNE TABLE DE DONNEES 72
XIII.4 LES CONTRAINTES SUR LES VARIABLES 72
XIII.4.1 Les vecteurs X et Z doivent avoir même dimension *72*
XIII.4.2 Les vecteurs X et Z doivent être bien corrélés *74*

XIV. EXEMPLES 75

XIV.1 UN PETIT EXEMPLE COMMENTE 75
XIV.1.1 Le programme *75*
XIV.1.2 La log *77*
XIV.1.3 Le listing *78*
XIV.2 CALAGE SIMULTANE DANS UN SONDAGE EN GRAPPES 84
XIV.2.1 Le programme *84*
XIV.2.2 Le listing *86*
XIV.3 CALAGE D'UN ECHANTILLON AVEC EGALITE DES POIDS DE CALAGE DANS LA GRAPPE 94
XIV.3.1 Le programme *94*
XIV.3.2 Les résultats (extrait du listing) *94*
XIV.4 L'ENQUETE PERMANENTE SUR LES CONDITIONS DE VIE DES MENAGES DE 1996 98
XIV.4.1 Les variables de calage *98*
XIV.4.2 Le programme *100*
XIV.4.3 Extraits du listing *101*
XIV.5 CALAGE GENERALISE POUR REDRESSEMENT DE NON-REPONSE 112
XIV.5.1 Les variables de calage *112*
XIV.5.2 Le programme *112*
XIV.5.3 Le listing (extrait) *113*

XV. LES CONTROLES ET LES MESSAGES D'ERREUR 117

XV.1 LES CONTROLES 117
XV.1.1 Contrôles sur les paramètres de la macro *117*
XV.1.2 Contrôles sur le contenu des tables de marges *118*
XV.1.3 Contrôles sur les modalités des variables catégorielles *119*
XV.1.4 Contrôles en cas de calage généralisé pour non-réponse *119*
XV.1.5 Contrôles sur la table contenant les pondérations finales *120*

XV.2 LES MESSAGES D'ERREUR 120
XV.2.1 Pas d'observation pour réaliser le calage 120
XV.2.2 Messages relatifs au déroulement de l'algorithme 120
XV.3 EXEMPLES 122
XV.3.1 Les totaux des marges catégorielles ne sont pas tous égaux 122
XV.3.2 Une variable catégorielle n'a pas le même nombre de modalités dans les tables échantillon et marges 122
XV.3.3 Pas d'observation valide dans la table en entrée 123
XV.3.4 Colinéarité entre les variables du calage 124
XV.3.5 Calage impossible 125
XV.3.6 Convergence imparfaite 127

BIBLIOGRAPHIE 131

Généralités

La macro SAS **CALMAR2** est une nouvelle version de la macro CALMAR (CALage sur MARges) en usage à l'INSEE depuis 1993. Comme la précédente, elle permet de redresser un échantillon, par repondération des individus, en utilisant une information auxiliaire disponible sur un certain nombre de variables, appelées variables de calage. Les pondérations produites par la macro sont telles que :

- pour une variable de calage catégorielle (ou "qualitative"), les effectifs pondérés des modalités de la variable dans l'échantillon, après redressement, seront égaux aux effectifs connus sur la population.
- pour une variable numérique (ou "quantitative"), le total pondéré de la variable dans l'échantillon, après redressement, sera égal au total connu sur la population.

Le redressement consiste à remplacer les pondérations initiales, qui sont en général les "poids de sondage" des individus (égaux aux inverses des probabilités d'inclusion), par des "poids de calage" (appelés aussi pondérations finales par la suite) aussi proches que possible des pondérations initiales au sens d'une certaine distance, et satisfaisant les égalités indiquées plus haut.

Lorsque les variables servant au redressement sont toutes catégorielles, le redressement consiste à "caler" les "marges" du tableau croisant toutes les variables sur des effectifs connus, d'où le nom de la macro.

Le programme CALMAR2 apporte à l'utilisateur les options supplémentaires suivantes par rapport à la version antérieure :

- une nouvelle fonction de distance : le sinus hyperbolique
- le traitement des colinéarités entre variables de calage
- la codification automatique des variables de calage catégorielles, l'utilisateur pouvant désormais spécifier des paramètres à valeurs discontinues ou de type libellé
- le calage simultané entre différents niveaux d'observation d'une même enquête
- le redressement de la non-réponse à l'aide d'une information auxiliaire connue sur les seuls répondants, par la méthode de calage généralisé mise au point par J.C. Deville

Note : la macro CALMAR2 utilise les modules SAS/STAT et SAS/IML du logiciel SAS.

Comment utiliser cette macro

Le programme CALMAR2 est disponible sous deux versions, qui se distinguent par leur ergonomie.

La première, appelée CALMAR2, à laquelle font référence tous les exemples de cette brochure, est exécutable à l'aide de l'instruction classique du langage SAS-macro d'appel d'une macro paramétrée. Avec cette version, l'utilisateur doit saisir, dans la fenêtre Editor d'une session SAS, les noms des paramètres de la macro ainsi que leur valeur.

La seconde version, appelée CALMAR2_GUIDE, offre une saisie inter-active des paramètres. L'utilisateur n'a plus qu'à taper au clavier les valeurs à assigner aux paramètres en renseignant les champs prévus à cet effet dans des menus déroulants. Selon le type de calage souhaité, la liste des paramètres à renseigner s'affiche automatiquement. L'utilisateur est ainsi guidé dans ses choix.

Le gestionnaire d'enquête chargé de redresser son échantillon une fois la collecte réalisée, avec une liste de variables de calage dûment testée, pourra, surtout s'il n'est pas un praticien expérimenté du logiciel SAS, utiliser de préférence la version CALMAR2_GUIDE.

Le chargé d'étude ou le gestionnaire d'enquête, en phase de simulation, soucieux de tester un grand nombre de modèles de calage et souhaitant exécuter des calages de façon répétitive, éventuellement à l'intérieur d'un programme macro propre à l'utilisateur, utilisera de façon classique la version CALMAR2.

A l'INSEE

Sur site central :

Les macros CALMAR2 et CALMAR2_GUIDE sont des membres du catalogue SAS **GR90.MACROS.COMPIL**, qui contient diverses macros SAS sous forme compilée, et se trouve sur les centres informatiques de Paris, Lille, Orléans et Aix. Pour les utiliser, il faut spécifier les instructions d'allocation suivantes dans le programme SAS, avant l'appel de l'une ou l'autre des macros :

```
LIBNAME ddname "GR90.MACROS.COMPIL" DISP=SHR;
OPTIONS SASMSTORE=ddname MSTORED;
```

`ddname` est un nom choisi par l'utilisateur.

L'appel de la version guidée de CALMAR2 se fait par la seule instruction :

```
%CALMAR2_GUIDE
```

La version classique CALMAR2 s'exécute en tapant l'instruction :

```
%CALMAR2(PARAMETRE1=valeur1,PARAMETRE2=valeur2,…)
```

dans laquelle on doit indiquer la liste de tous les paramètres nécessaires avec leur valeur, séparés par des virgules.

Sur micro-ordinateur :

Les macros CALMAR2 et CALMAR2_GUIDE sont utilisables directement par les utilisateurs des versions SAS fournies par le département de l'informatique.

Hors INSEE

Sur micro-ordinateur muni du système d'exploitation WINDOWS, les macros CALMAR2 et CALMAR2_GUIDE seront implantées sous forme compilée dans un répertoire. Elles seront donc contenues dans un catalogue SAS de nom : SASMACR.sas7bcat (version 8 de SAS) à l'intérieur de ce répertoire.

En l'absence d'option de chargement automatique, l'utilisateur devra allouer ce répertoire en début de session à l'aide d'une instruction LIBNAME de la forme :

```
LIBNAME ddname "chemin d'accès\nom du répertoire";
```

et utiliser l'option MSTORED du système SAS comme ci-dessus.

Si par exemple les macros compilées ont été implantées dans le sous-répertoire VERSION2 du répertoire CALAGE placé sur l'unité de disque D, on aura les instructions :

```
LIBNAME base "d:\calage\version2" ;
OPTIONS MSTORED SASMSTORE=base ;
```

Chacune des versions, CALMAR2 ou CALMAR2_GUIDE sera appelée comme ci-dessus.

Comment écrire les paramètres

Version CALMAR2

Voici quelques règles relatives à l'écriture des paramètres :

- l'ordre dans lequel sont données les valeurs des paramètres n'a pas d'importance ;
- les paramètres doivent être séparés par des virgules (et non des points-virgules) ;
- certains paramètres prennent des valeurs par défaut (ces valeurs sont spécifiées dans la documentation) : ils peuvent donc être omis ;
- certains paramètres sont indiqués dans la documentation comme étant obligatoires : leur absence provoque l'arrêt de la macro ;
- les paramètres mis explicitement à valeur manquante lors de l'appel de la macro sont à prohiber ;
- l'écriture des paramètres est en format "libre" (on peut mettre des blancs où l'on veut), mais il ne faut jamais mettre de virgule dans la valeur d'un paramètre ;
- on peut utiliser indifféremment minuscules et majuscules.

Version CALMAR2_GUIDE

L'utilisateur indique les valeurs des paramètres en face de leurs noms dans les champs prévus à cet effet au fur et à mesure du déroulement des menus de saisie. Les règles ci-dessus s'imposent également. La longueur des champs de saisie est adaptée aux conditions de la version 8 de SAS : un nom de table ou un nom de variable dans une table peut comporter jusqu'à 32 caractères. Le nom alloué à une bibliothèque SAS (ddname) ne peut excéder 8 caractères. Une exception a été faite pour les noms des variables à introduire dans les paramètres PONDQK et POIDKISH, qui peuvent avoir une longueur maximale de 16 caractères.

Le premier menu demande si l'on souhaite réaliser un calage simultané (voir plus loin chapitre IV) entre plusieurs niveaux d'observation, et si oui, la liste de ces niveaux. Les réponses à ces questions déterminent la liste des paramètres à renseigner, qui va s'afficher automatiquement.

La réponse à la première question de ce premier menu est requise de façon obligatoire. On ne peut ni passer à l'écran suivant, ni quitter l'application, sans y avoir d'abord répondu. C'est le seul champ de saisie qu'on ne peut laisser à blanc. Le contrôle de la cohérence des autres paramètres n'est pas inter-actif mais différé, une fois renseigné l'ensemble de tous les paramètres. En cas d'anomalie, les messages d'erreur s'affichent dans la fenêtre OUTPUT de la session SAS.

Certains menus de saisie sont partitionnés et s'affichent en plusieurs étapes. Il faut alors utiliser la commande R du menu de défilement des écrans pour revenir aux champs du haut de l'écran. Selon les cas, cette commande repositionne sur la sous-fenêtre précédente du même écran, ou sur l'écran de saisie précédent. De façon symétrique, la commande Entrée fera progresser vers la sous-fenêtre suivante du même écran ou vers l'écran de saisie suivant.

Au cours d'une même session SAS, à chaque appel de la macro, le programme conserve en mémoire les paramètres saisis lors de l'exécution précédente, à l'exception :

- des réponses au premier écran de saisie, concernant le type de calage souhaité. Ces champs sont à ressaisir à chaque appel de la macro.
- des paramètres initialisés par défaut dans la définition de la macro, dont la liste est donnée dans la suite de ce manuel. Leur valeur est ré-initialisée avec la valeur par défaut à chaque appel de la macro.

Les titres

Si l'on veut faire apparaître des titres en haut des pages contenant les sorties d'une macro, ces titres doivent **précéder** l'appel de la macro. D'autre part, les titres de niveaux 3 et suivants sont utilisés par les macros. L'utilisateur ne peut donc spécifier que des instructions TITLE ou TITLE2 : plus exactement, les titres figurant dans des instructions de type TITLE3, TITLE4... risquent à un moment d'être "écrasés" par ceux figurant dans les macros.

Olivier SAUTORY (Direction Générale de l'INSEE, tél. : 01.41.17.57.46) remercie d'avance les utilisateurs de la macro qui l'informeront des erreurs qui peuvent subsister, ainsi que ceux qui voudront bien lui faire part d'éventuelles suggestions pour améliorer la macro.

1ère partie

Aspects théoriques du calage sur marges

I. Le problème

On considère une population U = {1...k...N} de N individus, dans laquelle on a tiré un échantillon s de taille n. Pour tout individu k de U, on note π_k sa probabilité d'inclusion dans s (elle vaut n/N pour tout k dans le cas d'un sondage aléatoire simple).

Soit y une variable d'intérêt, pour laquelle on désire estimer le total sur la population : $Y = \sum_{k \in U} y_k$

L'estimateur de Y utilisé classiquement est l'estimateur de Horvitz-Thompson :

$$\hat{Y}_\pi = \sum_{k \in s} \frac{1}{\pi_k} y_k = \sum_{k \in s} d_k \, y_k \, .$$

Utiliser cet estimateur sans biais de Y revient à affecter à chaque individu de l'échantillon un poids d_k égal à l'inverse de sa probabilité d'inclusion (c'est le "coefficient d'extrapolation" N/n dans le cas d'un sondage aléatoire simple).

Information auxiliaire

Soit $x_1 \cdots x_j \cdots x_J$ J variables auxiliaires connues sur l'échantillon s, et dont **on connaît les totaux sur la population** : $X_j = \sum_{k \in U} x_{jk}$.

Pour tenir compte de cette information, on va chercher à estimer le total Y de y à l'aide d'un estimateur de la forme : $\hat{Y}_w = \sum_{k \in s} w_k \, y_k$,

où les poids w_k affectés aux individus sont "proches" (dans un sens à préciser) des poids de sondage d_k , et vérifient les **équations de calage** :

$$\boxed{\forall j = 1 \cdots J \quad \sum_{k \in s} w_k \, x_{jk} = X_j}$$

On cherche donc un estimateur "peu différent" de l'estimateur de Horvitz-Thompson qui "cale" l'échantillon sur les totaux des variables auxiliaires.

II. Résolution théorique

On choisit une "fonction de distance" G, d'argument $x = w_k / d_k$, pour mesurer les distances entre les w_k et les d_k ; G doit vérifier les conditions suivantes : elle est positive et convexe, et G(1) = G'(1) = 0.

Une fois la fonction G choisie (voir § V), le problème consiste à déterminer les poids w_k ($k \in s$) solutions du programme suivant (en notant les vecteurs $x_k' = (x_{1k} \dots x_{Jk})$ et $X' = (X_1 \dots X_J)$) :

$$\underset{w_k}{\text{Min}} \sum_{k \in s} d_k \; G(w_k / d_k) \text{ sous la contrainte } \sum_{k \in s} w_k \; x_k = X$$

i.e. on minimise une somme pondérée (par les d_k) des "distances" entre les poids de sondage d_k et les pondérations cherchées w_k , sous les contraintes du calage.

On résout ce problème en introduisant un vecteur de multiplicateurs de Lagrange $\lambda' = (\lambda_1 \cdots \lambda_J)$; le Lagrangien vaut :

$$£ = \sum_{k \in s} d_k \; G(w_k / d_k) - \lambda' \left(\sum_{k \in s} w_k \, x_k - X \right)$$

Les conditions du 1er ordre conduisent à :

$$\boxed{w_k = d_k \; F(x_k' \lambda)}$$

où F est la fonction réciproque de la dérivée de la fonction G.

Le vecteur λ est déterminé par la résolution du système non linéaire de J équations à J inconnues déterminé par les équations de calage :

$$\boxed{\sum_{k \in s} d_k \; F(x_k' \; \lambda) \, x_k = X} \quad \text{(E)}$$

On peut résoudre numériquement ce système par la méthode itérative de Newton ; on calcule une suite de vecteurs $\lambda^{(i)}$ définie par une relation de récurrence, en initialisant l'algorithme avec le vecteur $\lambda^{(0)} = 0$. La convergence est obtenue lorsque les rapports de poids w_k / d_k obtenus lors de deux itérations successives "ne bougent presque plus" :

$$\text{Max} \quad \left| \frac{w_k^{(i+1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| \quad < \varepsilon$$

III. Application au calage sur marges

III.1 Cas de variables catégorielles

Soit $V^1 \dots V^q \dots V^Q$ Q variables catégorielles, dont on note les modalités respectivement $1 \dots i_1 \dots I_1$, $1 \dots i_q \dots I_q$, $1 \dots i_Q \dots I_Q$.

Ces variables sont connues sur l'échantillon *s*, et on connaît les effectifs des modalités (les "marges") dans la population.

Les variables indicatrices associées à ces modalités vont jouer le rôle des variables x_j du § I, sur les totaux desquelles on désire se caler.

On note $\delta^q_{i_q}$ la variable indicatrice associée à la modalité i_q de la variable $\mathcal{V}^q$, définie par :

$$\forall\, k \in U \quad \delta^q_{i_q} = \begin{cases} 1 & \text{si} \quad V^q(k) = i_q \\ 0 & \text{sinon} \end{cases}$$

Le vecteur x_k a donc ici la forme suivante (il est constitué d'une suite de 1 et de 0) :

$$x'_k = \left(\dots \left(\delta^q_1(k) \dots \delta^q_{i_q}(k) \dots \delta^q_{I_q}(k) \right) \dots \right)$$

Le vecteur X des totaux des variables auxiliaires (ici les variables indicatrices) a la forme suivante :

$$X' = \left((N^1_1 \dots N^1_{I_1}) \ \dots \ (N^q_1 \dots N^q_{I_q}) \ \dots \ (N^Q_1 \dots N^Q_{I_Q}) \right)$$

où $N^q_{i_q} = \sum_{k \in U} \delta^q_{i_q}(k) =$ nombre d'individus k de U prenant la modalité i_q de $\mathcal{V}^q$.

Il est facile de vérifier que le système d'équations (E) est dans ce cas surdéterminé : la somme des I_q équations relatives aux modalités d'une variable $\mathcal{V}^q$ vaut toujours N, taille de la population U.

Il y a donc Q-1 équations redondantes : on supprime la dernière équation (relative à la modalité I_q) de chaque variable $V^2 \dots V^Q$, et on pose $\lambda^2_{I_2} = \dots = \lambda^q_{I_q} = \dots = \lambda^Q_{I_Q} = 0$.

Le système non linéaire à résoudre a donc $P = I_1 + (I_2 - 1) + \dots + (I_Q - 1)$ équations et P inconnues

III.2 Cas de variables catégorielles et numériques

Le cas où le redressement utilise simultanément des variables auxiliaires catégorielles et numériques est une simple extension du cas précédent.

Supposons que, outre les Q variables catégorielles définies précédemment, on dispose également de R variables numériques $\mathcal{Z}^1 \dots \mathcal{Z}^r \dots \mathcal{Z}^R$ dont on connaît les totaux sur la population.

On a alors :

$$x'_k = \left(\dots \delta^q_{i_q}(k) \dots \;\middle|\; z^1_k \dots z^R_k \right)$$

$$X' = (\dots N^q_{i_q} \dots \;|\; Z^1 \dots Z^R) \quad \text{où} \quad Z^r = \sum_{k \in U} z^r_k \;\; \text{connu}$$

$$\lambda' = (\dots \lambda^q_{i_q} \dots \;|\; \mu^1 \dots \mu^R)$$

IV. Calage simultané pour différentes unités d'une même enquête

IV.1 Le problème

L'enquête permanente sur les conditions de vie des ménages (PCV) de l'INSEE est réalisée selon un plan de sondage à deux degrés, à partir d'une base de sondage constituée du dernier recensement et stratifiée par taille de commune. Le premier degré consiste à sélectionner un échantillon de logements, par sondage aléatoire simple dans les strates. Au second degré, dans chaque logement sélectionné, on tire un individu parmi les membres du ménage âgés de 15 ans ou plus y résidant. Ce second tirage est aussi un sondage aléatoire simple, effectué selon la méthode Kish.

Le questionnaire comporte trois parties. La première concerne le ménage dans son ensemble (type de logement occupé, nombre de personnes du ménage, profession du chef de ménage...). La seconde est renseignée pour chacun des individus du ménage (sexe, âge, profession). La troisième n'est remplie que par l'individu du ménage sélectionné au deuxième degré, dit « individu Kish ».

L'enquête annuelle d'entreprises (EAE) est réalisée auprès d'un échantillon d'entreprises. Elle comprend, en plus du questionnaire sur l'activité globale de l'entreprise, un questionnaire pour chacun de ses établissements.

Dans ces deux exemples, l'enquête repose sur un échantillonnage à plusieurs degrés : deux degrés de sondage aléatoire dans le premier cas, sondage en grappes dans le second. Le questionnaire collecte des informations à des niveaux différents d'observation, correspondant à chacun des degrés de tirage : unité primaire de l'échantillon (le ménage[1]), toutes les unités secondaires appartenant aux unités primaires sélectionnées (individu), unité secondaire appartenant à l'échantillon du second degré (individu Kish) dans la première enquête ; dans la seconde, unité primaire (entreprise) et unité secondaire (établissement) appartenant à la grappe sélectionnée.

Supposons que l'on dispose d'une information auxiliaire pour chaque type d'unité échantillonnée, un calage séparé des données de l'enquête sur les ménages, puis des données sur les individus et enfin des résultats sur les individus de 15 ans ou plus nous donnerait des pondérations différentes pour les individus d'un même ménage. Le dépouillement de l'enquête pourrait conduire à des résultats inégaux entre la somme des unités secondaires et celle des unités primaires.

Dans le cas d'un sondage en grappes ou d'un sondage à deux degrés, le programme CALMAR2 permet de réaliser simultanément le calage des différentes unités observées, produisant des pondérations identiques pour les unités secondaires incluses dans la même grappe. Un tel calage peut inclure deux ou trois niveaux d'observation, nécessairement emboîtés.

[1] Selon la terminologie des enquêtes démographiques, un ménage est une communauté de personnes habitant de façon permanente le même logement. Il y a donc identité, d'un point de vue statistique, entre logement et ménage.

IV.2 Calage simultané dans un sondage en grappes

On réalise le calage au niveau de la grappe : l'entreprise dans le cas de l'EAE, le ménage dans une enquête auprès d'un échantillon de ménages dont on interroge tous les membres.

Soient : x_m = vecteur de variables auxiliaires connues pour toute grappe m de l'échantillon s_m

$X = \sum_{m \in U_M} x_m$ = vecteur des totaux de ces variables sur la population U_M, connus

$y_{m,k}$ = vecteur de variables auxiliaires connues pour toute unité secondaire k de la grappe m appartenant à l'échantillon s_m

$Y = \sum_{m \in U_M} \sum_{k \in m} y_{m,k}$ = vecteur des totaux de ces variables sur U, connus.

Le programme calcule les totaux par unité primaire :

$$y_m = \sum_{k \in m} y_{m,k}$$

d'où : $Y = \sum_{m \in U_M} y_m$

et utilise les équations de calage :

$$\sum_{m \in s_m} \frac{F\left(x_m^{'} \lambda_1 + y_m^{'} \mu_1\right)}{\pi_m} \left(x_m, y_m\right) = \left(X, Y\right)$$

La pondération $w_{m,k}$ de l'unité secondaire k appartenant à la grappe m dans l'échantillon complet des unités secondaires s est égale à la pondération w_m de la grappe m dans l'échantillon de grappes s_m. Les poids de calage respectent l'égalité des poids entre unités secondaires d'une même grappe, conformément au plan de sondage initial.

IV.3 Calage simultané dans un sondage à deux degrés

On note U_M la liste des unités primaires dans la population, U l'ensemble de la population des individus et U_e la population de référence au 2ème tirage. U_e peut recouvrir la population entière des unités secondaires U, ou n'en être qu'un sous-ensemble. On la désignera comme la population éligible au 2ème tirage.

Dans un grand nombre d'enquêtes de l'INSEE auprès des ménages par exemple, les logements constituent les unités primaires, les individus les unités secondaires, les individus de 15 ans ou plus la population éligible au 2ème degré pour répondre à la partie principale du questionnaire.

IV.3.1 Deux niveaux d'observation

Le questionnaire comprend deux volets, l'un concernant les unités primaires échantillonnées au 1er degré, l'autre les unités secondaires sélectionnées au second degré. Par exemple, dans une enquête sur les

structures familiales, on relève des informations sur l'ensemble du ménage, unité primaire, et sur la mère de famille, seul individu (unité secondaire) du ménage interrogé.

Soient :

s_1 l'échantillon des unités primaires au 1er degré

s_2 l'échantillon des unités secondaires au 2ème degré

s_{2m} l'échantillon d'unités secondaires tirées dans l'unité primaire m

$\pi_m = \text{Prob}\{m \in s_1\}$ est la probabilité d'inclusion de l'unité primaire m au 1er degré

n_m est le nombre d'unités secondaires à tirer dans l'unité primaire m

e_m est le nombre d'unités secondaires éligibles dans l'unité primaire m

$v_{k,m}$ = vecteur de variables auxiliaires connues pour toute unité secondaire k de l'échantillon, appartenant à l'unité primaire m

$\mathrm{V} = \sum_{k \in \mathrm{U_e}} v_k$ = vecteur des totaux de ces variables sur $\mathrm{U_e}$ connus.

Par ailleurs, on dispose, comme précédemment, d'un vecteur de variables auxiliaires x_m connues sur chaque unité primaire de l'échantillon et dont les totaux X sont connus dans la population des unités primaires.

L'individu k appartenant à l'unité primaire m a pour probabilité conditionnelle d'inclusion dans l'échantillon au second degré :

$$\pi_{k,m} = \text{Prob}\{k \in s_2 / m \in s_1\}$$

et pour pondération à l'intérieur de l'unité primaire m :

$$d_{k,m} = \frac{1}{\pi_{k,m}}$$

Lorsque le tirage au 2ème degré est un sondage aléatoire simple :

$$\pi_{k,m} = \frac{n_{k,m}}{e_m}$$

Les informations auxiliaires propres aux unités secondaires sont « remontées » au niveau de l'unité primaire par le calcul des nouvelles variables : $v_m = \sum_{k \in s_{2m}} d_{k,m} v_{k,m}$. v_m est le vecteur des estimateurs Horvitz-Thomson des totaux des variables de calage dans l'unité primaire m.

On utilise les équations de calage :

$$\sum_{m \in s_1} \frac{F\left(x_m^{'} \lambda_2 + v_m^{'} \gamma_2\right)}{\pi_m} \left(x_m, v_m\right) = \left(\mathrm{X}, \mathrm{V}\right)$$

Le poids total redressé de l'unité secondaire k est alors égal au produit du poids de calage w_m de l'unité primaire à laquelle il appartient par son poids conditionnel à l'intérieur de son unité primaire :

$$w_{k,m} = w_m \times d_{k,m}$$

IV.3.2 Trois niveaux d'observation

Le questionnaire comprend trois volets : le premier concerne les unités primaires (UP) sélectionnées au 1er degré, le second toutes les unités secondaires incluses dans les UP de l'échantillon, le troisième les unités secondaires de l'échantillon sélectionné au 2ème degré. L'enquête de l'INSEE sur les conditions de vie des ménages citée plus haut rentre dans ce cadre.

Soient :

s_1 l'échantillon des unités primaires au 1er degré

$s_{1,2}$ l'ensemble de toutes les unités secondaires incluses dans les unités primaires de l'échantillon s_1

s_2 l'échantillon des unités secondaires au 2ème degré

s_{2m} l'échantillon d'unités secondaires tirées dans l'unité primaire m

x_m = vecteur de variables auxiliaires connues pour toute unité primaire m de l'échantillon s_1

$X = \sum_{m \in U_M} x_m$ = vecteur des totaux de ces variables sur la population U_M, connus

$y_{m,i}$ = vecteur de variables auxiliaires connues pour toute unité secondaire i incluse dans l'unité primaire m appartenant à l'échantillon s_m

$Y = \sum_{m \in U_M} \sum_{k \in m} y_{m,k}$ = vecteur des totaux de ces variables sur U, connus.

$v_{k,m}$ = vecteur de variables auxiliaires connues pour toute unité secondaire k de l'échantillon s_2, appartenant à l'unité primaire m

$V = \sum_{k \in U_e} v_k$ = vecteur des totaux de ces variables sur U_e, connus.

$\pi_m = \text{Prob}\{m \in s_1\}$ est la probabilité d'inclusion de l'unité primaire m au 1er degré

n_m est le nombre d'unités secondaires à tirer dans l'unité primaire m

e_m est le nombre d'unités secondaires éligibles dans l'unité primaire m

L'individu k appartenant à l'unité primaire m a pour probabilité conditionnelle d'inclusion dans l'échantillon s_2 : $\pi_{k,m} = \text{Prob}\{k \in s_2 / m \in s_1\}$

et pour pondération à l'intérieur de l'unité primaire m :

$$d_{k,m} = \frac{1}{\pi_{k,m}}$$

Il a par ailleurs pour probabilité d'inclusion dans l'ensemble $s_{1,2}$ la probabilité d'inclusion dans s_1 de l'unité primaire m à laquelle il appartient.

Les informations auxiliaires propres à l'ensemble des unités secondaires de $s_{1,2}$ sont « remontées » au niveau 1 par le calcul des totaux par unité primaire :

$$y_m = \sum_{i \in m} y_{m,i}$$

Les informations auxiliaires propres aux unités secondaires de l'échantillon s_2 sont « remontées » au niveau des unités primaires par le calcul des nouvelles variables : $v_m = \sum_{k \in s_{2m}} d_{k,m} v_{k,m}$

On utilise les équations de calage :

$$\sum_{m \in s_1} \frac{F\left(x_m^{'}\lambda_3 + y_m^{'}\mu_3 + v_m^{'}\gamma_3\right)}{\pi_m}\left(x_m, y_m, v_m\right) = \left(\mathrm{X}, \mathrm{Y}, \mathrm{V}\right)$$

Les informations recueillies sur l'ensemble des unités secondaires incluses dans les unités primaires (échantillon $s_{1,2}$) seront pondérées par le poids de calage w_m des unités primaires auxquelles elles appartiennent.

Pour une unité secondaire k appartenant à l'échantillon du second degré, les informations spécifiques les concernant seront pondérées par le produit du poids de calage w_m de l'unité primaire à laquelle elle appartient par son poids conditionnel dans cette UP :

$$w_{k,m} = w_m \times d_{k,m} \qquad \text{avec } k \in s_2$$

V. Les fonctions G usuelles

On indique pour chacune des 5 méthodes usuelles la fonction G(x) (où $x = w_k / d_k$) et la fonction F(u) (où $u = x_k' \lambda$).

a) méthode "linéaire"

- $G(x) = \frac{1}{2}(x-1)^2, \quad x \in \mathbb{R} \quad$ et $\quad F(u) = 1 + u \quad (\in \mathbb{R})$

La forme linéaire de F donne son nom à cette méthode, dont on peut montrer qu'elle est équivalente à une méthode classique d'estimation, appelée **estimation par régression.**

b) méthode "raking ratio"

- $G(x) = x \operatorname{Log} x - x + 1, \quad x > 0 \quad$ et $\quad F(u) = \exp u (> 0)$

Lorsque les variables auxiliaires sont des variables catégorielles pour lesquelles on connaît les effectifs des modalités dans la population (cf § III.1), le choix de cette fonction G conduit à une méthode classique de redressement, proposée par Deming et Stephan [1], sous le nom de **raking ratio** ; elle est aussi connue (dans SAS en particulier) sous le nom I.P.F. ("Iterative Proportionnal Fitting").

c) méthode "logit"

- $G(x) = \left[(x-L) \operatorname{Log} \frac{x-L}{1-L} + (U-x) \operatorname{Log} \frac{U-x}{U-1} \right] \frac{1}{A}$, si L < x < U (∞ sinon)

avec $A = \frac{U-L}{(1-L)(U-1)}$

- $F(u) = \frac{L(U-1) + U(1-L)\exp(Au)}{U-1+(1-L)\exp(Au)} \in]L, U[$

La forme "logistique" de la fonction F donne son nom à cette méthode, que l'on peut aussi caractériser comme étant une méthode "raking ratio" tronquée aux deux extrémités, de façon que les rapports w_k / d_k soient "bornés" inférieurement par L et supérieurement par U.

d) méthode "linéaire tronquée"

- $G(x) = \frac{1}{2}(x-1)^2 \quad$ si $L \le x \le U$ (∞ sinon)
- $F(u) = 1 + u \in [L, U]$

e) méthode "sinus hyperbolique"

- $G(x) = \int_1^x \frac{1}{2}\left(\sinh\left(\alpha t - \frac{1}{\alpha t} \right) \right) dt$ où α est un coefficient positif
- $F(u) = \frac{1}{2}\left[\frac{1}{\alpha} Log\left(2\alpha u + \sqrt{4\alpha^2 u^2 + 1} \right) + \sqrt{\frac{1}{\alpha^2}\left(Log\left(2\alpha u + \sqrt{4\alpha^2 u^2 + 1} \right) \right)^2 + 4} \right]$

VI. Le choix de la méthode

Les principales caractéristiques des différentes méthodes sont les suivantes :

- la méthode **linéaire** est la plus rapide car elle converge toujours après deux itérations ; elle peut conduire à des poids w_k négatifs, ce qui en général ne satisfait pas le responsable d'enquête... Enfin, les poids ne sont pas bornés supérieurement, et les rapports de poids w_k/d_k peuvent prendre des valeurs que le statisticien jugera élevées (par exemple > 4).

- la méthode **raking ratio** conduit à des poids toujours positifs, mais non bornés supérieurement, d'ailleurs en général supérieurs (pour les poids les plus élevés) à ceux de la méthode "linéaire".

- les méthodes **logit** et **linéaire tronquée** présentent l'avantage de pouvoir définir une borne inférieure L et une borne supérieure U aux rapports w_k/d_k . Toutefois, on ne peut pas choisir a priori n'importe quelles valeurs pour L et U : il existe pour L une valeur maximale L_{max} (inférieure à 1), et pour U une valeur minimale U_{min} (supérieure à 1). Ces valeurs dépendent des données et des marges du calage : plus la structure de l'échantillon est différente de celle de la population, plus ces valeurs sont éloignées de 1.

- la méthode **sinus hyperbolique**, comme le raking ratio, donne des poids toujours positifs. Elle présente l'avantage, par rapport à cette méthode, de réduire la limite supérieure des poids obtenus, avec un coefficient α égal à 1. Ce coefficient joue le même rôle que les bornes L et U dans les méthodes logit et linéaire tronquées. Choisir α supérieur à 1 conduit à borner les poids de calage.

Dans la pratique, la détermination de ces valeurs L_{max} et U_{min}, ou de α se fait par "approximations successives" : on fait tourner la procédure de redressement en augmentant progressivement L (valeurs inférieures à 1), et en diminuant progressivement U (valeurs supérieures à 1) avec la méthode logit, en accroissant progressivement α avec le sinus hyperbolique,... jusqu'à ce que le programme manifeste qu'il n'existe pas de solution.

Face à différents "jeux" de pondération possible (on peut en obtenir théoriquement une infinité en faisant varier L et U) qui, on peut le rappeler, satisfont tous aux contraintes de calage, le responsable d'enquête doit en choisir un, et un seul. Les critères pouvant présider au choix de la pondération qui sera finalement utilisée sont les suivants :

- la plus faible dispersion ;
- la plus faible étendue ;
- l'allure générale de la distribution.

Le choix de la méthode ne peut reposer sur un critère de précision des estimateurs, car les méthodes sont toutes équivalentes (asymptotiquement) (cf [2]). C'est à un concept, non formalisé, de "robustesse" que le statisticien fait appel, et le critère qui préside au choix est donc d'une certaine façon affaire de point de vue.

VII. Calage et redressement de la non-réponse

VII.1 Définitions et problèmes

Le défaut de réponse à un questionnaire est une difficulté inhérente à toute enquête, qu'elle soit exhaustive ou par sondage. On distingue deux formes de non-réponse.

- La non-réponse totale : le questionnaire ne revient pas, malgré les relances postales ; l'enquêteur ne réussit pas à joindre les occupants d'un logement, malgré ses passages répétés ; plus rarement, l'enquêté contacté refuse catégoriquement de répondre. Aucune question n'est alors renseignée.
- La non-réponse partielle : l'enquêté répond à une partie du questionnaire, mais une ou plusieurs questions restent sans réponse.

Si les motifs de non-réponse sont multiformes, les conséquences en sont constantes. En premier lieu, les résultats extrapolés à partir des seuls répondants sont biaisés. En second lieu, ils sont moins précis, puisque déduits d'un échantillon plus petit que l'échantillon initial. Le statisticien est donc contraint de redresser ses résultats.

Aux deux catégories de non-réponse citées plus haut correspondent deux classes de redressement :

- on corrige les effets de la non-réponse totale en recalculant les poids d'extrapolation des répondants. On parle alors de repondération ;
- en cas de non-réponse partielle, on cherche à attribuer la réponse la plus probable à la question restée vierge. Il s'agit de l'imputation.

On ne s'intéressera ici qu'aux mécanismes de repondération pour non-réponse totale.

VII.2 Méthodes habituelles de repondération pour non-réponse totale

La repondération repose sur une modélisation des comportements de réponse. Le mécanisme de réponse est assimilé à une deuxième phase de sondage, équivalente au tirage d'un échantillon aléatoire de r répondants à l'intérieur de l'échantillon initial s. Une fois défini un tel modèle, on estime les probabilités individuelles de réponse qui corrigent les probabilités initiales de tirage et permettent de définir les nouveaux poids d'extrapolation.

Si de plus, le statisticien dispose d'une information auxiliaire sur l'ensemble de la population, bien corrélée aux variables d'intérêt, il pourra redresser le biais d'échantillonnage en appliquant le calage à son échantillon déjà repondéré pour non-réponse.

Sous certaines conditions, les deux opérations peuvent être rendues simultanées lorsqu'on entre en paramètres de CALMAR2 des variables de calage explicatives du comportement de réponse.

VII.2.1 Les modèles de réponse

VII.2.1.1 FORMALISATION DU PROBLEME

Une enquête par sondage est réalisée auprès d'un échantillon s de taille *n* tiré dans l'univers U. On ne recueille des réponses que sur le sous-échantillon R des *r* répondants. L'enquête mesure les variables d'intérêt *Y*. On dispose par ailleurs d'une information auxiliaire, les totaux sur la population des variables *X*, observées également dans l'enquête sur chacun des individus interrogés.

Soient :

$\mathcal{X}$ = vecteur de variables corrélées aux variables d'intérêt $\mathcal{Y}$ et dont on connaît les totaux dans la population U

$Xj = \sum_{k \in U} x_{jk}$ = total de la variable X_j connu sur la population

$\mathcal{Z}$ = vecteur de variables explicatives de la non-réponse dont on connaît les totaux dans la population

$\pi_k = \text{Prob}\{k \in s\}$ probabilité d'inclusion dans l'échantillon *s*

$P_k = \text{Prob}\{k \in R / k \in s\}$ probabilité de réponse

$d_k = \frac{1}{\pi_k}$

$\Theta_k = \frac{1}{P_k}$

Si P_k était connu, les estimateurs d'Horwitz-Thomson en présence de non-réponse s'écriraient :

$$\hat{Y}_{HT} = \sum_{k \in R} \frac{y_k}{\pi_k P_k} \qquad \hat{Z}_{HT} = \sum_{k \in R} \frac{z_k}{\pi_k P_k} \qquad \hat{X}_{HT} = \sum_{k \in R} \frac{x_k}{\pi_k P_k}$$

On recherche les nouveaux poids d'extrapolation $d_k^* = \frac{1}{\pi_k P_k}$ que l'on va affecter aux répondants pour corriger la non-réponse.

La définition d'un modèle de réponse comprend :

- le choix d'une fonction de réponse P(Z ; β) c'est-à-dire de la forme de la fonction P et des variables Z explicatives du comportement de réponse ;
- l'estimation des paramètres β de la fonction P.

VII.2.1.2 MODELE DE REPONSE UNIFORME

Une méthode simple de redressement consiste à multiplier le poids initial, égal à l'inverse de la probabilité de tirage, par l'inverse du taux de réponse observé, ce qui donne :

$$\hat{P}_k = \frac{r}{n}$$

$$\hat{Y} = \sum_{k \in R} \frac{1}{\pi_k} \frac{n}{r} y_k$$

d'où : $d_k^* = \frac{1}{\pi_k} \frac{n}{r}$

Aspects théoriques du calage sur marges

Une telle méthode conduit à postuler le caractère purement aléatoire de la non-réponse, chaque individu ayant la même probabilité de réponse, et par conséquent une parfaite homogénéité entre la population des répondants et celle des non-répondants.

Lorsque tel n'est pas le cas, l'estimateur ci-dessus est biaisé.

VII.2.1.3 GROUPES HOMOGENES DE REPONSE

La population se divise en *H* groupes homogènes du point de vue de la non-réponse. Les individus appartenant au même groupe *h* ont des probabilités de réponse identiques. Deux groupes distincts se caractérisent par des probabilités de réponse différentes et indépendantes l'une de l'autre. Les groupes peuvent être déterminés par régression logistique à partir des informations auxiliaires disponibles à la fois sur les répondants et les non-répondants.

L'estimateur repondéré d'Horwitz-Thomson s'écrit alors :

$\hat{Y} = \sum_{h=1}^{H} \sum_{k \in R_h} \frac{1}{\pi_k} \frac{n_h}{r_h} y_k$ où r_h est le nombre de répondants et n_h la taille de l'échantillon initial dans le groupe *h*

d'où : $d_k^* = \frac{1}{\pi_k} \frac{n_h}{r_h}$ où $k \in h$

VII.2.1.4 MODELE LINEAIRE GENERALISE

On dispose d'un vecteur Z de variables auxiliaires que l'on peut mesurer sur chacun des individus de l'échantillon. La probabilité de réponse dépend des valeurs que prend la combinaison linéaire $z'_k\beta$ pour l'individu k :

$P_k = \text{Prob}\ \{k \in R / k \in s\} = P(z'_k\beta)$

La fonction P peut être :

une fonction linéaire : $Pk = 1 - z'_k\beta$

une fonction log-linéaire : $Pk = \exp(-z'_k\beta)$

ou : $Pk = 1 - \exp(-z'_k\beta)$

une fonction logit : $Pk = \frac{\exp(z'_k\beta)}{1 + \exp(z'_k\beta)}$

Dans le cas linéaire, on peut estimer les paramètres β par régression logistique et en déduire un estimateur de P_k. Les probabilités individuelles de réponse P_k peuvent aussi être estimées directement par la méthode du maximum de vraisemblance ou par celle des moments.

L'estimateur d'un total Y sera alors fourni par :

$$\hat{Y}_{HT} = \sum_{k \in R} \frac{1}{\pi_k} \frac{1}{\hat{P}_k} y_k = \sum_{k \in R} d_k \Theta(\hat{\beta}' z_k) y_k \qquad \text{où :} \qquad \Theta\left(\hat{\beta}' z_k\right) = \frac{1}{\hat{P}_k}$$

Les répondants seront donc repondérés par :

$$d_k^* = d_k \hat{\Theta}_k$$

VII.2.2 Calage après redressement pour non-réponse

Les variables auxiliaires X étant également observées dans l'échantillon, on dispose d'estimateurs des totaux $\hat{X}_{HT}$ après redressement de la non-réponse :

$$\hat{X}_{HT} = \sum_{k \in R} \frac{x_k}{\pi_k \hat{P}_k}$$

soit, avec un modèle linéaire généralisé :

$$\hat{X}_{HT} = \sum_{k \in R} d_k \Theta(z'_k \hat{\beta}) x_k$$

Cet estimateur $\hat{X}_{HT}$ est « calé » sur les vrais totaux X dans la population par le programme CALMAR2, avec en entrée non plus les poids initiaux (inverse des probabilités d'inclusion du plan de sondage), mais les poids $d_k^* = \frac{1}{\pi_k \hat{P}_k}$, issus de la repondération pour non-réponse.

Les équations de calage s'écrivent alors :

$$\left.\begin{array}{l} \sum_{k \in R} w_k x_k = X \\ w_k = d_k^* F(x'_k \lambda) \end{array}\right\} \Rightarrow \sum_{k \in R} d_k^* F(x'_k \lambda) x_k = X \qquad \Leftrightarrow \qquad \sum_{k \in R} \frac{1}{\pi_k} \frac{1}{\hat{P}_k} F(x'_k \lambda) x_k = X$$

Avec un modèle linéaire généralisé :

$$\sum_{k \in R} d_k \Theta(z'_k \hat{\beta}) F(x'_k \lambda) x_k = X \qquad (1)$$

VII.2.3 Redressement direct par calage

Lorsque le vecteur X des variables de calage recouvre le vecteur Z des variables explicatives du comportement de réponse, le calage redresse en même temps la non-réponse.

Soit $T = \begin{bmatrix} X \\ Z \end{bmatrix}$ où *T* est un vecteur de variables corrélées aux variables d'intérêt *Y*, incluant les variables *Z* qui influencent la non-réponse.

L'usage direct de CALMAR2 sur les répondants coïncide avec la méthode ci-dessus, dissociant l'estimation des probabilités individuelles de réponse et le calage sur données externes, lorsque la fonction de calage choisie et celle de réponse sont identiques à la fonction exponentielle (voir [14]) :

$$\begin{aligned} F &= \exp(u) \\ P_k &= \exp(-z'_k \beta) \end{aligned}$$

L'équation (1) devient alors :

$$\sum_{k\in R} d_k \exp(t'_k\beta)\exp(t'_k\lambda)t_k = T \qquad \Leftrightarrow \qquad \sum_{k\in R} d_k \exp[t'_k(\beta+\lambda)]t_k = T$$

soit : $$\sum_{k\in R} d_k \exp(t'_k\gamma)t_k = T$$

On retrouve ici l'expression classique du calage sur les variables *T*, en posant : $\gamma = \lambda + \beta$. Le coefficient γ s'interprète comme la somme d'un facteur de non-réponse et d'un facteur de calage proprement dit.

Cette méthode de redressement nécessite de connaître :

- les totaux *T* des variables de calage et des variables explicatives de la non-réponse dans la population ;
- les valeurs individuelles t_k de ces variables pour les seuls répondants.

Cette dernière simplification ne doit pas faire oublier la nécessité d'avoir des valeurs individuelles t_k parfaitement homogènes par rapport aux totaux *T* dans la population.

VII.3 Calage en l'absence d'information sur les non-répondants

La gestion d'une enquête sur le terrain peut conduire à détecter des mécanismes imprévus de non-réponse susceptibles de biaiser fortement les résultats, et qui s'expliquent par des caractéristiques individuelles non mesurables sur l'ensemble de la population. Les variables z_k sont bien observées sur les répondants dans l'enquête, mais aucune source ne nous donne les totaux Z dans la population.

On dispose en revanche d'une information auxiliaire sous forme d'un vecteur X de variables dont on connaît les totaux dans l'univers, ainsi que les valeurs individuelles x_k pour les répondants.

Exemple : une enquête a pour but d'analyser l'insertion des jeunes dans la vie active. Les élèves ou étudiants ayant terminé un cycle d'étude (fin de CAP, de BEP, terminale de lycée, fin de premier ou de second cycle universitaire...) constituent la population de référence, dans laquelle on tire un échantillon.

Une grande part de la non-réponse provient des changements de domicile intervenus entre temps. Or, ceux-ci ne sont pas indépendants des situations d'emploi que l'on cherche à mesurer. Un jeune change d'autant plus de résidence qu'il trouve un emploi lui assurant l'autonomie, qui plus est parfois dans une commune éloignée.

La non-réponse s'explique ainsi - du moins partiellement - par l'obtention d'un emploi. La corriger suppose de connaître le nombre total de jeunes appartenant au champ de l'enquête, ayant obtenu un contrat de travail au cours de la période de référence, information non disponible.

Un cas particulier d'application est aussi fourni par les enquêtes réalisées à l'aide d'une base de sondage déjà ancienne. Les enquêtes de l'INSEE auprès des ménages, dont les échantillons sont tirés dans le dernier recensement de population (RP), rentrent dans ce champ dès que le recensement s'éloigne dans le temps.

Supposons que les principaux facteurs de non-réponse soient des caractéristiques relevées dans le RP : taille du ménage, âge de la personne de référence, nationalité, taille de la commune de résidence, par exemple. A côté d'autres variables telles que la catégorie socio-professionnelle du chef de ménage, on les utiliserait donc pour redresser les résultats par calage, en prenant pour x_k comme pour le total *X* les valeurs disponibles dans la base de sondage.

Quand l'enquêteur passe cinq ans après le recensement, cela revient à redresser la non-réponse par les caractéristiques du ménage qui habitait le logement cinq ans auparavant, et qui a pu changer. Tout se passe comme si les variables z_k , valeurs au moment de l'enquête, n'étaient observées que sur les répondants, et les totaux Z inconnus.

VII.3.1 Les équations de calage

Soit la situation suivante, avec les mêmes notations que précédemment :

$\underset{(J,1)}{X}$ = vecteur des totaux des variables $\mathcal{X}$, connus dans la population $\left(X_j = \sum_{k \in U} x_{jk} \right)$

$\mathcal{Z}$ = variables observées dans l'échantillon des répondants

$\Theta_k = \dfrac{1}{P_k}$ où $P_k(\beta) = \text{Prob}\{k \in R / k \in s\}$

$d_k = \dfrac{1}{\pi_k}$ où $\pi_k = \text{Prob}\{k \in s\}$

Les variables $\mathcal{X}$ et $\mathcal{Z}$ sont corrélées aux variables d'intérêt $\mathcal{Y}$. Les variables $\mathcal{Z}$ expliquent en outre le comportement de réponse. Les vecteurs x_k et z_k ont même dimension.

Si l'on connaissait les probabilités individuelles de réponse $P_k(\beta)$, le calage sur les totaux X, après repondération pour non-réponse, nous donnerait les équations suivantes :

$$\sum_{k \in R} d_k \Theta(z'_k \beta) F(x'_k \lambda) x_k = X \qquad \text{(I)}$$

L'application du calage généralisé au redressement de la non-réponse dissocie les vecteurs X et Z dans les équations de calage suivantes :

$$\boxed{\sum_{k \in R} d_k \Theta(z'_k \gamma) x_k = X} \qquad \text{(II)}$$

Θ est l'une des fonctions habituelles de calage : linéaire, exponentielle, « logit tronquée » ou sinus hyperbolique. Son inverse remplit les conditions d'une fonction de probabilité : $\dfrac{1}{\Theta}$ est monotone croissante et comprise entre 0 et 1.

Soit : $\gamma = \lambda + \beta$ où β est la vraie valeur du paramètre de la fonction de réponse

$\Theta(z'_k \beta) = \dfrac{1}{P_k}$ est l'inverse de la fonction de réponse

On peut toujours écrire :

$$\Theta(z'_k \gamma) = \frac{\Theta(z'_k \beta + z'_k \lambda)}{\Theta(z'_k \beta)} \Theta(z'_k \beta) = \Theta(z'_k \beta) F_k(z'_k \lambda)$$

avec : $$F_k(z'_k \lambda) = \frac{\Theta(z'_k \beta + z'_k \lambda)}{\Theta(z'_k \beta)}$$

En remplaçant dans (II) il vient :

$$\sum_r d_k \Theta(z'_k \beta) F_k(z'_k \lambda) x_k = X \qquad \text{(III)}$$

F_k est une fonction dépendant de l'unité k, qui vérifie :

- $F_k(\lambda = 0) = 1$
- F_k est monotone et croissante en fonction de λ.

F_k a donc les propriétés d'une fonction de calage. Résoudre (II) équivaut à résoudre (III), c'est-à-dire à caler sur les totaux X un échantillon préalablement repondéré pour non-réponse à l'aide des poids : $d_k^* = d_k \Theta(z'_k \beta)$. Le coefficient γ s'interprète comme la somme d'un facteur de non-réponse et d'un facteur de calage.

Comme dans le cas d'un calage habituel, le sytème d'équations (II) se résout par la méthode itérative de Newton, en partant du point γ=0.

Dans le cas où Θ est une fonction linéaire : $\Theta_k = 1 + z'_k \gamma$

La résolution converge dès la première itération.

L'estimateur du total *Y* d'une variable d'intérêt s'écrit alors :

$$\hat{Y} = \sum_{k \in R} w_k y_k = \sum_{k \in R} d_k \Theta(z'_k \gamma) y_k = \sum_{k \in R} d_k y_k + \left(\sum_{k \in R} d_k z'_k y_k \right) T_r^{-1} (X - X^*) = Y^* + \hat{A}(X - X^*)$$

où :

$X^* = \sum_{k \in R} d_k x_k$

$Y^* = \sum_{k \in R} d_k y_k$

$\hat{A} = [Z'X]^{-1}[Z'Y]$ est l'expression des coefficients de la régression instrumentale de *Y* sur *X* utilisant *Z* en instrument.

$\hat{Y}$ peut donc s'interpréter comme l'estimateur par régression du total *Y*, au moyen des variables instrumentales *Z*.

Cette propriété reçoit également une interprétation géométrique : $\hat{Y}$ apparaît comme la projection du vecteur *Y* sur l'hyperplan des variables *X*, parallèlement à une direction orthogonale à l'hyperplan des variables *Z*.

L'écart $\hat{u}$ entre l'estimateur et la vraie valeur sera donc d'autant plus faible que les espaces vectoriels définis respectivement par *X* et *Z* seront plus proches, autrement dit que le groupe des variables x et celui des variables z seront mieux corrélés.

VIII. Calage et colinéarités

La résolution des équations de calage nécessite l'inversion d'une matrice de la forme : $\Phi = \sum_k x_k x'_k$. Pour que cette matrice soit inversible, les variables de calage ne doivent pas être colinéaires.

Cette règle peut être transgressée dans certains cas.

VIII.1 Colinéarités structurelles

Le programme élimine automatiquement les colinéarités structurelles qui apparaissent lorsque plusieurs variables catégorielles figurent dans les variables de calage (voir § III.1). A partir de la deuxième variable catégorielle, le programme ne prend en compte que les (Pj-1) modalités de chaque variable $\mathcal{V}_j$.

La suppression des équations de calage redondantes touche également les variables instrumentales explicatives de la non-réponse en cas de calage généralisé.

L'utilisateur n'a donc pas à s'en préoccuper.

VIII.2 Autres colinéarités

D'autres colinéarités entre variables de calage peuvent apparaître.

Exemple 1 : on peut souhaiter caler les résultats d'une enquête sur les effectifs des ménages selon leur taille et sur leur répartition selon la composition familiale dans la population. La typologie de ménages construite par l'INSEE pour le dépouillement du recensement retient notamment la modalité : « personne vivant seule ». Cette modalité coïncide avec la modalité « ménage d'une personne » d'une nomenclature des ménages selon leur taille. Il ne serait donc pas possible d'effectuer un calage avec ces deux variables sans opérer un regroupement de modalités.

Exemple 2 : on souhaite caler les résultats d'une enquête simultanément sur la répartition des ménages selon leur taille par région et sur leur répartition selon la profession du chef de ménage par région également. Il y aura ici une équation redondante par région, que le programme ne peut éliminer.

VIII.3 Les solutions

On peut résoudre ces problèmes :

- soit en redéfinissant les modalités des variables de calage pour éviter les colinéarités ;
- soit en utilisant la technique des **matrices inverses généralisées**, qui recherche une pseudo-matrice inverse dans le cas où la matrice Φ n'est pas inversible.

Cette dernière option est disponible dans le programme CALMAR2, qui utilise dans ce cas l'option « GINV » de SAS-IML au lieu de « INV » (inverse simple). Il faut alors renseigner le paramètre **COLIN** de la macro CALMAR2 par OUI (voir plus loin la deuxième partie).

L'usage de la matrice inverse généralisée (COLIN=OUI) est néanmoins déconseillé en cas de calage généralisé pour redressement de la non-réponse par des variables instrumentales. Une corrélation forte entre deux variables de calage (vecteur X) ou entre deux variables de redressement de la non-réponse (vecteur Z) conduit en effet à une instabilité des estimateurs, dont la variance s'accroît alors fortement.

Par ailleurs, le programme ne détecte pas la non convergence avec la fonction de calage linéaire tronquée, lorsqu'on utilise la technique des matrice inverses généralisées.

2ème partie

Mise en œuvre de la macro CALMAR2

La macro CALMAR2 réalise, sur option :

- Un calage simple sur une table d'enquête ne comprenant qu'un seul niveau d'observation.
- Un calage simultané à deux niveaux d'observation emboîtés correspondant aux degrés de tirage d'un sondage en grappes. Le programme cale les résultats sur des totaux dans la population des grappes et sur des totaux dans la population des unités secondaires. Ce peut être, dans le cas d'une enquête auprès des ménages, un calage ménages-individus ; pour une enquête auprès des entreprises, un calage simultané entre les niveaux entreprise et établissement ; pour des données géographiques, entre les niveaux région et zone d'emploi, etc...
- Un calage simultané entre les deux niveaux d'observation d'un sondage à deux degrés. Les résultats sont calés sur des totaux dans la population des unités primaires et sur des totaux dans la population de référence des unités secondaires sélectionnées au $2^{ème}$ degré.
- Un calage simultané entre trois niveaux d'observation dans un plan de sondage à deux degrés, avec un questionnaire comprenant des informations sur les unités primaires, sur chaque unité secondaire appartenant à l'unité primaire de l'échantillon, et sur les unités secondaires sélectionnées au $2^{ème}$ degré. (exemple : ménages, tous individus du ménage, individus appartenant au champ du tirage Kish).
- Dans le cas d'un sondage en grappes sans information au niveau de l'unité primaire, on peut également caler les résultats sur des totaux dans la population des unités secondaires seule, en respectant l'égalité des poids entre unités secondaires appartenant à la même unité primaire.

Dans chaque cas, on peut choisir d'utiliser ou non des variables instrumentales pour redresser la non-réponse par calage généralisé.

Note : dans la suite du document, les formulations du type "table &DATAMEN", "variable &POIDSFIN" etc, signifient : table spécifiée dans le paramètre DATAMEN de la macro, variable spécifiée dans le paramètre POIDSFIN, etc.

IX. Calage simple : un seul niveau d'observation

IX.1 Les données en entrée de la macro

Les résultats de l'enquête pour chaque unité de l'échantillon d'une part, les totaux sur la population des variables de calage d'autre part, doivent être préparés dans deux tables SAS.

IX.1.1 La table SAS contenant les données

Les données relatives à l'échantillon doivent se présenter sous la forme d'une table SAS contenant :

- les variables qui vont être utilisées pour le redressement, ou "variables de calage" ;
- la variable de pondération initiale ;
- éventuellement une variable identifiant.

Cette table peut bien sûr contenir toute autre variable[2] n'intervenant pas directement dans le redressement.

Le nom de cette table SAS est spécifié dans le paramètre (obligatoire) DATAMEN de la macro.

IX.1.1.1 LES VARIABLES DE CALAGE CATEGORIELLES

Une variable catégorielle, ou qualitative, au sens du calage, peut être une variable "caractère" ou "numérique" au sens de SAS.

- Si elle est caractère, elle ne peut avoir plus de 999 modalités.
- Le programme **recodifie automatiquement** les variables catégorielles de l'utilisateur dans un codage numérique séquentiel (1,2,...p).

Les noms des variables de calage catégorielles, ainsi que leurs nombres de modalités, sont spécifiés dans la table des marges &MARMEN (voir § IX.2).

La macro réalise le contrôle suivant :

- une variable catégorielle a, dans la table de données, le nombre de modalités annoncé dans la table des marges.

IX.1.1.2 LES VARIABLES DE CALAGE NUMERIQUES

Une variable de calage numérique (ou quantitative) au sens du calage doit être "numérique" au sens de SAS.

Les noms des variables de calage numériques sont spécifiés dans la table des marges &MARMEN.

[2] Elle peut contenir une variable de pondération supplémentaire, la variable &PONDQK (voir § IX.2).

IX.1.1.3 LA VARIABLE DE PONDERATION INITIALE

C'est la variable donnant pour chaque observation k la valeur de la pondération initiale d_k. Cette valeur est égale à l'inverse de la probabilité d'inclusion de l'observation dans l'échantillon.

Par exemple, dans le cas d'un sondage aléatoire simple, ou dans celui d'un sondage à plusieurs degrés "autopondéré", chaque unité de la population a la même probabilité d'appartenir à l'échantillon, égale à n/N, où n est la taille de l'échantillon et N la taille de la population. La pondération initiale attribuée à chaque observation de l'échantillon est dans ce cas constante, et vaut N/n, quantité appelée parfois le "coefficient d'extrapolation".

La variable de pondération initiale doit être "numérique" au sens de SAS. Elle est spécifiée dans le paramètre POIDS de la macro.

Choix de la pondération initiale

Lorsque figure au moins une variable catégorielle parmi les variables de calage, avec toute autre fonction que le sinus hyperbolique, la pondération initiale peut être définie à un coefficient multiplicatif près : si les rapports de poids[3] et le paramètre λ (voir 1ère partie) dépendent du choix de la pondération initiale, en revanche les pondérations finales w_k n'en dépendent pas[4]. Par exemple, dans le cas d'un sondage où toutes les unités ont la même probabilité d'appartenir à l'échantillon, il est équivalent de spécifier une variable de pondération initiale constante égale à 1, à 1000... , ou à N/n.

Cette propriété n'est pas vérifiée avec la fonction sinus hyperbolique.

Toutefois, il y a (au moins) deux bonnes raisons d'utiliser pour le calage la "bonne" pondération initiale (N/n dans l'exemple ci-dessus) :

- d'un point de vue théorique, la pondération initiale est définie comme l'inverse de la probabilité d'inclusion, et les rapports de poids mesurent de combien on s'écarte de cette pondération par le calage ; en particulier, ces rapports de poids ont pour moyenne 1 dans le cas d'un coefficient d'extrapolation constant, égal à N/n.
- d'un point de vue pratique, partir d'une pondération initiale très éloignée de la pondération finale (par exemple une pondération initiale égale à 1 dans un échantillon de 3000 observations, pour une taille de la population égale à 21 millions) conduit souvent à un dépassement de capacité lors des calculs réalisés par le programme : la macro génère alors le message "Le calage ne peut être réalisé" ... à tort puisque ce n'est qu'une impossibilité fortuite due à une "mauvaise" spécification du problème de calage.

Remarque : dans ce cas, la macro édite à la suite de ce message la taille de l'échantillon pondéré (avec la pondération initiale) et la taille de la population.

Il est cependant possible à l'utilisateur, sans modifier la valeur des poids contenus dans sa table de données &DATAMEN et identifiés par le paramètre POIDS, de choisir un coefficient multiplicatif de ses poids initiaux. Ceux-ci seront alors multipliés par ce coefficient dans le programme CALMAR2 avant le calage proprement dit. La valeur de ce coefficient, qui doit être numérique, est spécifiée dans le paramètre optionnel ECHELLE.

Cette possibilité peut être utilisée pour un redressement uniforme de la non-réponse. Il suffit de choisir un coefficient égal au rapport : taille de la population/somme des poids initiaux. Elle est recommandée en cas d'usage de la fonction sinus hyperbolique. La macro CALMAR2 effectue elle-même le calcul de ce ratio sur option.

[3] Pour une observation, le "rapport de poids" est le rapport pondération finale/pondération initiale.

[4] A condition de modifier les bornes en conséquence, lorque l'on utilise une méthode bornée.

Pondération générée

Lorsqu'il y a au moins une variable catégorielle parmi les variables de calage, et si la variable de pondération initiale n'est pas spécifiée dans le paramètre POIDS, la macro génère une variable de pondération constante, égale au rapport : effectif de la population/nombre d'observations de la table en entrée non éliminées (l'effectif de la population est calculé grâce à la table donnant les marges, ou bien donné dans le paramètre POPMEN).

S'il n'y pas de variable catégorielle, le paramètre POIDS doit être obligatoirement renseigné.

IX.1.1.4 AUTRES VARIABLES DE LA TABLE DES DONNEES

La table &DATAMEN peut contenir d'autres variables que celles définies précédemment. En particulier, peuvent y figurer :

- une variable servant à identifier les observations, spécifiée dans le paramètre IDENT.
- une variable définissant une pondération supplémentaire des observations, spécifiée dans le paramètre PONDQK (son utilisation n'est justifiée que dans des cas très particuliers, voir référence[2]).

IX.1.1.5 OBSERVATIONS ELIMINEES

Est éliminée du calage (et donc de la table en sortie éventuelle créée par la macro) toute observation de la table en entrée ayant une valeur manquante sur l'une des variables du calage ou l'une des variables de pondération, ou prenant une valeur négative ou nulle sur l'une des variables de pondération.

IX.1.1.6 CALAGE EN PRESENCE DE NON-REPONSE

Les procédures de redressement, telles qu'elles sont présentées au § II, ne sont en principe valides qu'en absence de non-réponse totale[5] dans l'échantillon de taille n, ou bien après une opération de correction de cette non-réponse. Si ces conditions ne sont pas vérifiées, on peut opérer directement sur l'échantillon des répondants, dont on note m la taille, sans modifier les pondérations initiales : on peut montrer que cette méthode revient à réaliser deux corrections simultanées, l'une pour non-réponse, et l'autre pour amélioration de l'estimation[6].

Toutefois, la pratique courante, qui revient à multiplier les pondérations initiales par n/m (ou taille de la population/somme des poids initiaux en cas de tirage à probabilités inégales), présente deux intérêts :

- dans le cas d'un sondage aléatoire simple, la somme des pondérations initiales redonne l'effectif total N de la population ;
- l'algorithme de détermination de recherche des pondérations peut se dérouler dans de meilleures conditions (voir § IX.1.1.3).

La macro CALMAR2 réalise elle-même cette correction sur option, selon la valeur du paramètre ECHELLE. La valeur 0 de ce paramètre entraîne une multiplication des poids initiaux par le rapport : taille de la population/somme des poids initiaux. Le programme ne peut cependant calculer ce ratio que si l'une des conditions suivantes est vérifiée :

- les variables de calage comprennent au moins une variable catégorielle ;
- l'utilisateur a indiqué la taille de la population dans le paramètre POPMEN.

[5] Il y a non-réponse totale lorsqu'un individu de l'échantillon n'a pas répondu à l'enquête.

[6] La correction pour non-réponse utilisant un modèle de réponse fondé sur les mêmes variables que celles du calage (voir F. DUPONT : "Calage et redressement de la non-réponse totale", Journées de méthodologie statistique 1993).

IX.1.2 La table SAS contenant les variables de calage et les marges

Les noms des variables de calage, leurs nombres de modalités, et les marges associées doivent se présenter sous la forme d'une table SAS, dont le nom est spécifié dans le paramètre (obligatoire) de la macro MARMEN.

Cette table contient une observation par variable de calage. Les variables de la table sont : VAR, N, MAR1, MAR2, ..., MARh. Elles prennent les valeurs suivantes :

VAR nom de la variable[7].

N nombre de modalités de la variable.

C'est un entier strictement positif pour une variable catégorielle, et 0 pour une variable numérique ; une valeur négative de N est remplacée par 0, et une valeur positive non entière est remplacée par sa partie entière.

MAR1 valeur de la marge associée à la modalité 1 pour une variable catégorielle, valeur de la marge associée pour une variable numérique.

...

MARj valeur de la marge associée à la modalité j pour une variable catégorielle ayant au moins j modalités, valeur manquante (.) pour une variable catégorielle ayant moins de j modalités ou pour une variable numérique.

...

MARh valeur de la marge associée à la modalité h pour une variable catégorielle ayant h modalités, où h est le nombre maximal de modalités (i.e. la valeur maximale de N), valeur manquante (.) pour une variable catégorielle ayant moins de h modalités ou pour une variable numérique.

La macro réalise les contrôles suivants :

- toute variable spécifiée dans la variable VAR existe dans la table &DATAMEN ;
- une variable telle que N=0 est une variable numérique de la table &DATAMEN ;
- pour une variable telle que N=p (p >0) les marges MAR1 à MARp sont renseignées ;
- les totaux des marges des variables catégorielles sont tous égaux (la valeur commune de ces totaux est en principe égale à la taille de la population).

Marges des variables catégorielles données en pourcentages

L'utilisateur peut donner les valeurs des marges catégorielles en pourcentages, à condition de spécifier la valeur OUI pour le paramètre PCT de la macro. Dans ce cas, les totaux des marges doivent tous être égaux à 100.

L'utilisateur doit alors indiquer dans le paramètre POPMEN la taille de la population.

Le lecteur peut se reporter au chapitre XIV pour avoir des exemples de tables &MARMEN correctes, et au chapitre XV pour avoir des exemples des erreurs à ne pas commettre.

IX.2 Syntaxe de la macro

[7] en minuscules ou en majuscules

IX.2.1 Paramètres spécifiant les tables SAS en entrée

* DATAMEN = nom de table SAS

nom de la table SAS contenant les données de l'échantillon**.** Ce paramètre est **obligatoire.**

Cette table contient , pour chaque unité de l'échantillon, les variables, catégorielles et numériques, du calage, et éventuellement une variable identifiant. Elle contient également la variable de **pondération initiale** (sauf dans le cas où celle-ci est générée).

Voir le contenu détaillé de cette table au § IX.1.1.

* MARMEN= nom de table SAS

nom de la table SAS contenant les noms des variables de calage, les nombres de modalités, et les marges associées. Ce paramètre est **obligatoire.**

Voir le contenu détaillé de cette table au § IX.1.2.

* POIDS = variable

variable **numérique** contenant les pondérations initiales des observations de l'échantillon. Elle appartient à la table &DATAMEN.

Ce paramètre est obligatoire lorsqu'il n'y pas de variable de calage catégorielle (voir § IX.1.1.3).

* PONDQK = variable

variable **numérique** de pondération des observations de l'échantillon, non liée à la variable spécifiée dans le paramètre POIDS (elle appartient à la table &DATAMEN) : elle permet de moduler la fonction de calage en fonction de l'observation (voir référence [2]).

Par défaut : PONDQK = __UN, variable générée constamment égale à 1.

* IDENT = variable

variable servant à identifier les observations dans les éditions et récupérée dans la table en sortie éventuelle (paramètre DATAPOI) contenant les pondérations finales. Ce paramètre est facultatif dans le cas du calage simple.

* PCT = OUI ou NON

si PCT vaut OUI, les marges des variables catégorielles dans la table &MARMEN sont données en pourcentages.

Par défaut : PCT = NON.

* POPMEN = valeur

si PCT vaut OUI, on spécifie ici l'effectif total de la population (dont la connaissance est nécessaire pour calculer les marges du calage).

Ce paramètre est obligatoire si PCT = OUI, ou si ECHELLE=0 alors que les variables de calage sont toutes numériques.

IX.2.2 Paramètres spécifiant la méthode utilisée

* M = 1, 2, 3, 4 ou 5

numéro de la méthode, i.e. de la fonction de distance utilisée pour calculer les écarts entre les pondérations initiales et les pondérations finales :

1. méthode linéaire
2. méthode raking ratio
3. méthode logit
4. méthode linéaire tronquée
5. méthode sinus hyperbolique.

Par défaut, M=1.

* LO = valeur

borne inférieure des rapports de poids (voir note 3), lorsque l'on utilise une méthode "bornée" (logit ou linéaire tronquée).

Ce paramètre est obligatoire lorsque M = 3 ou 4.

* UP = valeur

borne supérieure des rapports de poids, lorsque l'on utilise une méthode "bornée" (logit ou linéaire tronquée).

Ce paramètre est obligatoire lorsque M = 3 ou 4.

* ALPHA = valeur

coefficient multiplicatif de l'observation, utilisé dans la méthode sinus hyperbolique. Une augmentation de sa valeur entraîne un resserrement de l'éventail des rapports de poids.

Par défaut, ALPHA=1.

* SEUIL = valeur

seuil ε pour le test d'arrêt de l'algorithme de Newton : il y a convergence lorsque le maximum (en valeur absolue) des différences entre les rapports de poids calculés lors de deux itérations successives est inférieur à ce seuil.

Par défaut : SEUIL = 0.0001.

* MAXITER = n

nombre maximum d'itérations au cours de l'algorithme de Newton : si l'algorithme n'a pas convergé en n itérations, il s'arrête.

Par défaut : MAXITER = 15.

* ECHELLE = n

nombre positif par lequel seront multipliés les poids initiaux avant calage.

Si ECHELLE = 0, le programme calcule lui-même le facteur d'échelle, qui sera égal au rapport : taille de la population/somme des poids initiaux, ce qui revient à redresser la non-réponse de façon uniforme. Dans le cas où toutes les variables de calage sont numériques, l'utilisateur devra obligatoirement renseigner le paramètre POPMEN pour utiliser l'option ECHELLE = 0.

Par défaut : ECHELLE = 1.

* COLIN=OUI ou NON

La résolution des équations de calage conduit à inverser une matrice Φ qui est singulière lorsque les variables de calage sont colinéaires. Si COLIN vaut NON, le programme recherchera l'inverse de cette matrice par la méthode classique (option INV de SAS-IML), et s'interrompra en cas de colinéarité. Si COLIN vaut OUI, le programme tiendra compte des colinéarités éventuelles entre variables de calage dans la table de données, en utilisant la technique des matrices inverses généralisées (option GINV dans SAS-IML) pour le calcul de la matrice Φ^{-1}.

Par défaut : COLIN=NON.

IX.2.3 Paramètres relatifs aux tables en sortie

* DATAPOI = nom de table SAS

nom de la table SAS contenant les pondérations finales.

- Si cette table n'existe pas, elle est créée par la macro : elle a autant d'observations que d'observations **non éliminées** de la table &DATAMEN ; elle contient la variable &POIDSFIN (voir plus loin) et le cas échéant la variable &IDENT.

- Si cette table existe, le paramètre MISAJOUR indique comment la macro opère sur elle.

* MISAJOUR = OUI ou NON

ce paramètre spécifie le traitement de la table &DATAPOI lorsque celle-ci existe déjà :

- si MISAJOUR = OUI, la variable de pondération &POIDSFIN, et le cas échéant la variable &IDENT, est ajoutée à la table.

- si MISAJOUR = NON, la macro crée une nouvelle table, contenant les variables &POIDSFIN et &IDENT, l'ancienne table portant le même nom étant détruite.

Par défaut : MISAJOUR = OUI.

* POIDSFIN = variable

nom de la variable contenant les pondérations finales des observations non éliminées de l'échantillon ; elle appartient à la table &DATAPOI.

Ce paramètre est obligatoire lorsque le paramètre DATAPOI est renseigné.

* LABELPOI = label

label (éventuel) attribué à la variable spécifiée dans le paramètre POIDSFIN.

Remarque : ce label ne doit pas contenir de virgule.

* OBSELI = OUI ou NON

si OBSELI = OUI, la macro crée une table SAS, de nom __OBSELI, contenant les observations éliminées, les variables du calage, les variables de pondération et le cas échéant la variable &IDENT. L'utilisateur peut imprimer, ou utiliser, cette table après l'appel de la macro.

Par défaut : OBSELI = NON.

IX.2.4 Paramètres spécifiant les sorties imprimées

* EDITION = 0, 1, 2 ou 3

Paramètre indiquant le détail souhaité des éditions.

0. aucun résultat n'est édité, sauf les statistiques sur les poids de calage, si le paramètre STAT vaut OUI (voir plus loin).
1. la macro édite la liste des paramètres rentrés par l'utilisateur et le bilan final du calage
2. édition de la liste des paramètres, des tables des marges avant et après calage, du bilan final
3. en plus des éditions (2), la macro édite la valeur des coefficients λ après chaque itération et la valeur du critère d'arrêt.

Par défaut, EDITION vaut 3.

* CONT = OUI ou NON

si CONT vaut OUI, un certain nombre de contrôles sont réalisés sur les paramètres de la macro (présence des paramètres obligatoires, cohérence des paramètres...), sur les valeurs données à ces paramètres (existence des tables SAS, des variables de pondération...), sur les données figurant dans les tables de marges (existence des variables, présence de toutes les marges...), ainsi que sur les variables de calage dans les tables de données.

La liste complète de ces contrôles, ainsi que des exemples de messages produits par la macro, sont donnés au chapitre XV.

Par défaut : CONT = OUI.

* CONTPOI = OUI ou NON

si CONTPOI vaut OUI, la macro édite le contenu de la table &DATAPOI[8].

Par défaut : CONTPOI = OUI.

* EDITPOI = OUI ou NON

si EDITPOI vaut OUI, la macro édite les valeurs des rapports de poids obtenus pour chaque case de l'hyper-tableau de contingence croisant toutes les variables, catégorielles et numériques (i.e. pour chaque combinaison de valeurs de ces variables).

***Remarque* :** ce tableau peut être très volumineux, surtout en présence de variables numériques.

Par défaut : EDITPOI = NON.

* STAT = OUI ou NON

si STAT vaut OUI, la macro édite des statistiques (moyenne, écart-type, quantiles, valeurs extrêmes...) et des graphiques[9] relatifs aux distributions des variables "rapport de poids" et "pondération finale".

Elle édite également un tableau des rapports de poids moyens par modalité des variables catégorielles prises une à une.

Par défaut : STAT = OUI.

* NOTES = OUI ou NON

si NOTES = NON, les notes produites par SAS durant l'exécution de la macro ne sont pas éditées.

Par défaut : NOTES = NON.

IX.3 La table en sortie

La table en sortie spécifiée dans le paramètre DATAPOI peut être temporaire ou permanente. Ses observations sont celles de la table &DATAMEN non éliminées ; elle contient la variable de pondération finale &POIDSFIN et, le cas échéant, la variable identifiant &IDENT.

Les observations sont classées de la même façon dans la table en entrée &DATAMEN et dans la table en sortie &DATAPOI.

- si cette table n'existe pas, elle est créée par la macro.
- Si cette table existe, et si MISAJOUR = OUI, elle est mise à jour par la macro, i.e. la (ou les) nouvelle(s) variable(s) est ajoutée à la table existante :
    - si une variable portant le même nom qu'une variable ajoutée existait déjà dans la table, elle est donc "écrasée"

[8] Il s'agit des sorties d'une procédure CONTENTS.

[9] Il s'agit des sorties d'une procédure UNIVARIATE.

- si le nombre d'observations (non éliminées) est supérieur au nombre d'observations de la table avant l'exécution de la macro, cette table est "complétée" par l'ajout de valeurs manquantes aux variables préexistantes
- si le nombre d'observations (non éliminées) est inférieur au nombre d'observations de la table avant l'exécution de la macro, les nouvelles variables sont "complétées" par l'ajout de valeurs manquantes.

  ***Remarque* :** il est préférable dans la pratique d'éviter les situations décrites dans les deux derniers cas, en créant plusieurs tables en sortie par exemple. En particulier, dans de telles situations, si la variable &IDENT ne change pas de nom, les identifiants ne correspondent plus aux valeurs des pondérations...

- Si cette table existe, et si MISAJOUR = NON, l'ancienne version de la table est détruite, et remplacée par une table contenant la nouvelle variable &POIDSFIN (ainsi que &IDENT).

IX.4 Les sorties imprimées

Selon la valeur donnée au paramètre EDITION, la macro édite :

- un tableau donnant les valeurs des paramètres ;
- un tableau permettant la comparaison entre les marges calculées sur l'échantillon avec la pondération initiale et les marges dans la population (marges du calage) ;
- un tableau donnant la valeur du critère d'arrêt et le nombre de poids négatifs après chaque itération ;
- un tableau donnant les coefficients du vecteur lambda des multiplicateurs de Lagrange après chaque itération ;
- un tableau permettant la comparaison entre les marges calculées sur l'échantillon avec la pondération finale et les marges dans la population (marges du calage) : ces marges doivent être les mêmes ;
- si **EDITPOI = OUI** : un tableau donnant les valeurs des rapports de poids obtenus pour chaque combinaison de valeurs des variables de calage ;
- si **STAT = OUI** : les sorties de la procédure UNIVARIATE (moyenne, médiane, écart-type, quantiles, valeurs extrêmes, stem-and-leaf plot...) sur la variable rapport de poids et sur la variable pondération finale ; un tableau donnant le rapport de poids moyen par modalité de chaque variable catégorielle ;
- si **CONTPOI = OUI** : les sorties de la procédure CONTENTS sur la table contenant la pondération finale ;
- un bilan du calage :
  - le nom de la table en entrée ;
  - le nombre d'observations (non pondérées) de cette table ;
  - le nombre d'observations éliminées, et le nombre d'observations conservées ;
  - le nom de la variable de pondération initiale, ou bien, dans le cas où elle est générée, la valeur (constante) de cette variable : taille de la population / nombre d'observations ;
  - le nombre, la liste, et les nombres de modalités des variables catégorielles ;
  - la taille de l'échantillon pondéré : somme des pondérations initiales calculée sur les observations conservées, lorsque figure au moins une variable catégorielle parmi les variables de calage ;
  - la taille de la population, calculée à l'aide des marges ou bien donnée dans le paramètre POPMEN, lorsque figure au moins une variable catégorielle parmi les variables de calage ;
  - le nombre et la liste des variables numériques ;
  - la méthode utilisée ;

- le nombre d'itérations ;
- le cas échéant, le nom de la variable de pondération finale et le nom de la table contenant cette variable.

En cas d'erreur, les sorties précédentes ne sont pas toutes fournies, et la macro édite en général un message donnant la cause de l'arrêt du programme.

X. Calage simultané dans un sondage en grappes

X.1 Les données en entrée de la macro

On se place ici dans le cadre d'un sondage en grappes avec un questionnaire interrogeant à la fois l'unité primaire (niveau 1) et toutes les unités secondaires incluses dans l'unité primaire sélectionnée (niveau 2). Exemples : échantillon de ménages dont on interroge tous les individus ; échantillon d'entreprises dont on observe tous les établissements. Les résultats de l'enquête devront faire l'objet de deux tables de données : l'une pour les unités directement échantillonnées (niveau 1), l'autre pour les unités de niveau 2 rattachées aux premières. Il en sera de même pour les totaux sur la population des variables de calage, qui devront constituer deux tables de marges.

X.1.1 Les tables SAS contenant les données

Les données relatives à l'échantillon doivent se présenter sous la forme de deux tables SAS, correspondant à chacun des niveaux d'observation.

* DATAMEN

Ce paramètre, obligatoire, spécifie le nom de la table de niveau 1. Cette table contient :

- les variables qui vont être utilisées pour le redressement, ou "variables de calage" ;
- la variable de pondération initiale ;
- une variable identifiant, dont le nom est spécifié dans le paramètre (obligatoire) IDENT de la macro.

* DATAIND

Ce paramètre, obligatoire, spécifie le nom de la table de niveau 2. Cette table contient :

- les variables qui vont être utilisées pour le redressement, ou "variables de calage" ;
- une variable identifiant l'observation, dont le nom est spécifié dans le paramètre (obligatoire) IDENT2 de la macro. IDENT2 doit identifier totalement l'observation concernée. Ce ne peut être seulement le rang de l'individu dans le ménage, ni celui de l'établissement dans l'entreprise.
- une variable identifiant l'unité de niveau 1 à laquelle se rattache l'observation de niveau 2. Ce doit être la même variable que celle spécifiée dans le paramètre IDENT.

Chaque table peut bien sûr contenir toute autre variable n'intervenant pas directement dans le redressement.

X.1.1.2 LES VARIABLES DE CALAGE

Elles peuvent être catégorielles ou numériques et obéissent aux mêmes règles que celles décrites plus haut (voir § IX.1.1.1 et IX.1.1.2).

Les noms des variables de calage, ainsi que leurs nombres de modalités, sont spécifiés dans les tables des marges :

- &MARMEN pour les variables de niveau 1 (voir § IX.1.2 et § X.1.2)
- &MARIND pour les variables de niveau 2.

La macro réalise le contrôle suivant :

- une variable catégorielle a, dans la table de données correspondante, le nombre de modalités annoncé dans la table des marges.

X.1.1.3 LA VARIABLE DE PONDERATION INITIALE

C'est la variable donnant, pour chaque observation m de la table de données des unités primaires (spécifiée dans le paramètre DATAMEN), la valeur de la pondération initiale d_m. Cette valeur est égale à l'inverse de la probabilité d'inclusion de la grappe dans l'échantillon.

La variable de pondération initiale doit être "numérique" au sens de SAS. Elle est spécifiée dans le paramètre POIDS de la macro et présente dans la table &DATAMEN.

Choix de la pondération initiale

Toutes les observations faites plus haut (voir § IX.1.1.3) restent valables. La restriction de certaines propriétés à la présence d'au moins une variable catégorielle parmi les variables de calage doit ici être comprise : dans la table de marges de niveau 1.

X.1.1.4 AUTRES VARIABLES DE LA TABLE DES DONNEES

Les tables &DATAMEN et &DATAIND peuvent contenir d'autres variables que celles définies précédemment. En particulier, peut figurer dans la table &DATAMEN :

- une variable définissant une pondération supplémentaire des observations, spécifiée dans le paramètre PONDQK (son utilisation n'est justifiée que dans des cas très particuliers, voir référence[2]).

X.1.1.5 OBSERVATIONS ELIMINEES

Est éliminée du calage (et donc de la table en sortie éventuelle créée par la macro) toute observation de la table en entrée ayant une valeur manquante sur l'une des variables du calage ou de pondération, ou prenant une valeur négative ou nulle sur l'une des variables de pondération.

Toute observation de la table &DATAIND ayant une valeur manquante sur l'une des variables citées ci-dessus entraîne l'élimination de toutes les observations de niveau 2 se rattachant au même identifiant de niveau 1 que l'observation concernée, et par conséquent l'élimination, dans la table &DATAMEN, de l'observation à laquelle elle se rattache. Pour un échantillon de ménages comprenant un questionnaire individus par exemple, toute valeur manquante sur un individu entraînera l'élimination, dans la table &DATAMEN, du ménage auquel appartient cet individu et dans la table &DATAIND, de tous les individus appartenant à ce ménage.

De façon symétrique, toute observation de la table &DATAMEN ayant une valeur manquante ou une pondération non positive entraîne l'élimination, dans la table &DATAIND, de toutes les unités qui lui sont rattachées.

X.1.1.6 CALAGE EN PRESENCE DE NON-REPONSE

Les remarques faites dans le cas du calage simple (voir § IX.1.1.6) restent valables ici. La condition de présence d'au moins une variable catégorielle pour l'utilisation de l'option 0 du paramètre ECHELLE doit s'entendre ici « parmi les variables de calage de niveau 1 ».

X.1.2 Les tables SAS contenant les variables de calage et les marges

Les noms des variables de calage, leurs nombres de modalités, et les marges associées doivent se présenter sous la forme de deux tables SAS, dont les noms sont spécifiés dans les paramètres (obligatoires) de la macro :

- MARMEN, pour les unités de niveau 1 (ménages ou entreprises, dans les exemples cités plus haut) ;
- MARIND, pour les unités de niveau 2 (individus ou établissements, dans ces exemples).

Ces deux tables ont chacune la même structure que celle décrite dans le cas d'un calage simple (§ IX.1.2). Les mêmes contrôles sont appliqués à chacune de ces tables.

Marges des variables catégorielles données en pourcentages

L'utilisateur peut donner les valeurs des marges catégorielles en pourcentages, à condition de spécifier la valeur OUI pour le paramètre PCT de la macro. Dans ce cas, les totaux des marges des variables concernées doivent tous être égaux à 100. Le paramètre PCT s'applique aux deux tables de marges : &MARMEN et &MARIND.

L'utilisateur doit alors indiquer dans le paramètre POPMEN la taille de la population pour les unités de niveau 1, dans le paramètre POPIND la taille de la population des unités de niveau 2.

X.2 Syntaxe de la macro

X.2.1 Paramètres spécifiant les tables SAS en entrée

* DATAMEN = nom de table SAS

nom de la table SAS contenant les données de niveau 1**.** Ce paramètre est **obligatoire.**

Cette table contient , pour chaque unité primaire de l'échantillon, les variables, catégorielles et numériques, du calage, et une variable identifiant l'observation. Elle contient également la variable de **pondération initiale** (sauf dans le cas où celle-ci est générée).

Voir le contenu détaillé de cette table au § IX.1.1.

* DATAIND = nom de table SAS

nom de la table SAS contenant les données de niveau 2**.** Ce paramètre est **obligatoire.**

Cette table contient , pour chaque unité secondaire de l'échantillon, les variables, catégorielles et numériques, du calage. Elle doit aussi contenir obligatoirement une variable **identifiant** l'observation (paramètre IDENT2) ainsi qu'une variable identifiant la grappe à laquelle elle appartient (paramètre IDENT). Ce dernier identifiant a le même nom que celui de la table &DATAMEN. Exemple : dans le cas d'un calage simultané ménages-individus, on doit avoir dans cette table un identifiant de l'individu et un identifiant du ménage auquel il se rattache.

* MARMEN= nom de table SAS

nom de la table SAS contenant les noms des variables de calage, les nombres de modalités, et les marges associées pour les unités de niveau 1. Ce paramètre est **obligatoire.**

Voir le contenu détaillé de cette table au § IX.1.2.

* MARIND= nom de table SAS

nom de la table SAS contenant les noms des variables de calage, les nombres de modalités, et les marges associées, pour les unités de niveau 2. Ce paramètre est **obligatoire**.

Voir le contenu détaillé de cette table au § IX.1.2.

* POIDS = variable

variable **numérique** contenant les pondérations initiales des observations de l'échantillon. Elle appartient à la table &DATAMEN.

Ce paramètre est obligatoire lorsqu'il n'y pas de variable de calage catégorielle (voir § IX.1.1.3).

* PONDQK = variable

variable **numérique** de pondération des observations de niveau 1 de l'échantillon, non liée à la variable spécifiée dans le paramètre POIDS (elle appartient à la table &DATAMEN) : elle permet de moduler la fonction de calage en fonction de l'observation (voir référence [2]).

Par défaut : PONDQK = __UN, variable générée constamment égale à 1.

* IDENT = variable

variable servant à identifier les observations de niveau 1 dans les éditions et récupérée dans la table en sortie éventuelle (paramètre DATAPOI) contenant les pondérations finales. Elle doit figurer dans la table &DATAMEN et dans la table &DATAIND. Ce paramètre est **obligatoire**.

* IDENT2 = variable

variable servant à identifier les observations de niveau 2 dans les éditions et récupérée dans la table en sortie éventuelle (paramètre DATAPOI2) contenant les pondérations finales. Elle doit figurer dans la table

&DATAIND. Cette variable doit identifier totalement l'entité ; elle ne peut être le simple rang de l'observation à l'intérieur de l'unité de niveau 1. Ce paramètre est **obligatoire**.

Exemple : calage simultané entreprises-établissements. Si l'identification utilisée est celle du répertoire SIRENE, l'identifiant de l'établissement doit être le SIRET, et non le NIC. La table &DATAIND contiendra donc à la fois le SIREN de l'entreprise et le SIRET complet de l'établissement.

* PCT = OUI ou NON

si PCT vaut OUI, les marges des variables catégorielles dans la table &MARMEN et dans la table &MARIND sont données en pourcentages.

Par défaut : PCT = NON.

* POPMEN = valeur

si PCT vaut OUI, on spécifie ici l'effectif total de la population des unités de niveau 1 (dont la connaissance est nécessaire pour calculer les marges du calage).

Ce paramètre est obligatoire si PCT = OUI, ou **si ECHELLE=0 lorsque les variables de calage sont toutes numériques**.

* POPIND = valeur

si PCT vaut OUI, on spécifie ici l'effectif total de la population des unités de niveau 2 (dont la connaissance est nécessaire pour calculer les marges du calage).

Ce paramètre est obligatoire si PCT = OUI.

X.2.2 Paramètres spécifiant la méthode utilisée

* M = 1, 2, 3, 4 ou 5

numéro de la méthode, i.e. de la fonction de distance utilisée pour calculer les écarts entre les pondérations initiales et les pondérations finales :

1. méthode linéaire
2. méthode raking ratio
3. méthode logit
4. méthode linéaire tronquée
5. méthode sinus hyperbolique.

Par défaut, M=1.

* LO = valeur

borne inférieure des rapports de poids (voir note 3), lorsque l'on utilise une méthode "bornée" (logit ou linéaire tronquée).

Ce paramètre est obligatoire lorsque M = 3 ou 4.

* UP = valeur

borne supérieure des rapports de poids, lorsque l'on utilise une méthode "bornée" (logit ou linéaire tronquée).

Ce paramètre est obligatoire lorsque M = 3 ou 4.

* ALPHA = valeur

coefficient multiplicatif de l'observation, utilisé dans la méthode sinus hyperbolique. Une augmentation de sa valeur entraîne un resserrement de l'éventail des rapports de poids.

Par défaut, ALPHA=1.

* SEUIL = valeur

seuil ε pour le test d'arrêt de l'algorithme de Newton : il y a convergence lorsque le maximum (en valeur absolue) des différences entre les rapports de poids calculés lors de deux itérations successives est inférieur à ce seuil.

Par défaut : SEUIL = 0.0001.

* MAXITER = n

nombre maximum d'itérations au cours de l'algorithme de Newton : si l'algorithme n'a pas convergé en n itérations, il s'arrête.

Par défaut : MAXITER = 15.

* ECHELLE = n

nombre positif par lequel seront multipliés les poids initiaux avant calage.

Si ECHELLE = 0, le programme calcule lui-même le facteur d'échelle, qui sera égal au rapport : taille de la population/somme des poids initiaux, ce qui revient à redresser la non-réponse de façon uniforme. Dans le cas où toutes les variables de calage sont numériques, l'utilisateur devra obligatoirement renseigner le paramètre POPMEN pour utiliser l'option ECHELLE = 0.

Par défaut : ECHELLE = 1.

* COLIN=OUI ou NON

La résolution des équations de calage conduit à inverser une matrice Φ qui est singulière lorsque les variables de calage sont colinéaires. Si COLIN vaut NON, le programme recherchera l'inverse de cette matrice par la méthode classique (option INV de SAS-IML), et s'interrompra en cas de colinéarité. Si COLIN vaut OUI, le programme tiendra compte des colinéarités éventuelles entre variables de calage dans la table de données, en utilisant la technique des matrices inverses généralisées (option GINV dans SAS-IML) pour le calcul de la matrice Φ^{-1}.

Par défaut : COLIN=NON.

X.2.3 Paramètres relatifs aux tables en sortie

* DATAPOI = nom de table SAS

nom de la table SAS contenant les pondérations finales associées aux unités de niveau 1.

- Si cette table n'existe pas, elle est créée par la macro : elle a autant d'observations que d'observations **non éliminées** de la table &DATAMEN ; elle contient la variable &POIDSFIN (voir plus loin) et la variable &IDENT.

- Si cette table existe, le paramètre MISAJOUR indique comment la macro opère sur elle.

* DATAPOI 2= nom de table SAS

nom de la table SAS contenant les pondérations finales associées aux entités de niveau 2.

- Si cette table n'existe pas, elle est créée par la macro : elle a autant d'observations que d'observations **non éliminées** de la table &DATAIND ; elle contient la variable &POIDSFIN (voir plus loin) ainsi que les variables &IDENT et &IDENT2.
- Si cette table existe, le paramètre MISAJOUR indique comment la macro opère sur elle.

Ce paramètre est obligatoire lorsque DATAPOI est renseigné.

* POIDSFIN = variable

nom de la variable contenant les pondérations finales des observations non éliminées de l'échantillon ; elle appartient à la table &DATAPOI et à la table &DATAPOI2.

Ce paramètre est obligatoire lorsque les paramètres DATAPOI et DATAPOI2 sont renseignés.

* MISAJOUR = OUI ou NON

ce paramètre spécifie le traitement des tables &DATAPOI et &DATAPOI2 lorsque l'une de celles-ci existe déjà :

- si MISAJOUR = OUI, la variable de pondération &POIDSFIN, la variable &IDENT, (et la variable &IDENT2 dans &DATAPOI2) sont ajoutées à la table.

- si MISAJOUR = NON, la macro crée une nouvelle table, contenant les variables &POIDSFIN et &IDENT (et la variable &IDENT2 dans &DATAPOI2), l'ancienne table portant le même nom étant détruite.

Par défaut : MISAJOUR = OUI.

* LABELPOI = label

label (éventuel) attribué à la variable spécifiée dans le paramètre POIDSFIN.

Remarque : ce label ne doit pas contenir de virgule.

* OBSELI = OUI ou NON

si OBSELI = OUI, la macro crée deux tables SAS, de nom :

- __OBSELI contient les observations éliminées dans la table &DATAMEN, les variables du calage de niveau 1, les variables de pondération &POIDS et &PONDQK , la variable &IDENT. Elle contient également une variable ELIMI égale au nombre d'unités secondaires éliminées dans l'unité primaire. Si ELIMI=0, l'observation est éliminée pour cause de valeur manquante parmi les variables de calage de niveau 1 ou de pondération non positive. Si ELIMI est positif, l'observation est éliminée car elle inclut des unités secondaires ayant une variable de calage de niveau 2 non renseignée.

- __INDELI contient les observations éliminées dans la table &DATAIND, les variables du calage de niveau 2, les variables &IDENT et &IDENT2.

L'utilisateur peut imprimer, ou utiliser, chacune de ces tables après l'appel de la macro.

Par défaut : OBSELI = NON.

X.2.4 Paramètres spécifiant les sorties imprimées

* EDITION = 0, 1, 2 ou 3

Paramètre indiquant le détail souhaité des éditions.

0. aucun résultat n'est édité, sauf les statistiques sur les poids de calage, si le paramètre STAT vaut OUI (voir plus loin).
1. la macro édite la liste des paramètres rentrés par l'utilisateur et le bilan final du calage
2. édition de la liste des paramètres, des tables des marges avant et après calage, du bilan final
3. en plus des éditions (2), la macro édite la valeur des coefficients λ après chaque itération et la valeur du critère d'arrêt.

Par défaut, EDITION vaut 3.

* CONT = OUI ou NON

si CONT vaut OUI, un certain nombre de contrôles sont réalisés sur les paramètres de la macro (présence des paramètres obligatoires, cohérence des paramètres...), sur les valeurs données à ces paramètres (existence des tables SAS, des variables de pondération...), sur les données figurant dans les tables de marges (existence des variables, présence de toutes les marges...), ainsi que sur les variables de calage dans les tables de données.

La liste complète de ces contrôles, ainsi que des exemples de messages produits par la macro, sont donnés au chapitre XV.

Par défaut : CONT = OUI.

* CONTPOI = OUI ou NON

si CONTPOI vaut OUI, la macro édite le contenu des tables &DATAPOI et &DATAPOI2[10].

Par défaut : CONTPOI = OUI.

* EDITPOI = OUI ou NON

si EDITPOI vaut OUI, la macro édite les valeurs des rapports de poids obtenus pour chaque case de l'hyper-tableau de contingence croisant toutes les variables de calage, catégorielles et numériques (i.e. pour chaque combinaison de valeurs de ces variables). Ces croisements sont définis par les valeurs prises par ces variables pour chaque unité primaire de l'échantillon. Il en résulte que la valeur d'une variable de calage de niveau 2 coïncide, dans ce tableau, avec le nombre d'unités secondaires, dans l'unité primaire, ayant la modalité j de la variable y, si y est catégorielle, ou avec le total dans l'unité primaire de la variable numérique x.

***Remarque* :** ce tableau peut être très volumineux, surtout en présence de variables numériques.

Par défaut : EDITPOI = NON.

* STAT = OUI ou NON

si STAT vaut OUI, la macro édite des statistiques (moyenne, écart-type, quantiles, valeurs extrêmes...) et des graphiques[11] relatifs aux distributions des variables "rapport de poids" et "pondération finale".

Elle édite également un tableau des rapports de poids moyens par modalité des variables catégorielles prises une à une.

Par défaut : STAT = OUI.

* NOTES = OUI ou NON

si NOTES = NON, les notes produites par SAS durant l'exécution de la macro ne sont pas éditées.

Par défaut : NOTES = NON.

X.3 Les tables en sortie

Chacune des tables en sortie spécifiées dans les paramètres DATAPOI et DATAPOI2 peut être temporaire ou permanente.

Les observations de la table &DATAPOI sont celles de la table &DATAMEN non éliminées. Cette table contient la variable de pondération finale &POIDSFIN et la variable identifiant &IDENT.

Les observations sont classées de la même façon dans la table en entrée &DATAMEN et dans la table en sortie &DATAPOI.

Les observations de la table &DATAPOI2 sont celles, non éliminées, de la table &DATAIND. Elle contient la variable &POIDSFIN ainsi que les identifiants &IDENT et &IDENT2.

[10] Il s'agit des sorties d'une procédure CONTENTS.

[11] Il s'agit des sorties d'une procédure UNIVARIATE.

Les observations sont classées de la même façon dans la table en entrée &DATAIND et dans la table en sortie &DATAPOI2.

Chacune des tables &DATAPOI et &DATAPOI2 est créée ou mise à jour selon la valeur du paramètre MISAJOUR comme indiqué au § X.2.3.

XI. Calage simultané dans un sondage à deux degrés avec trois niveaux d'observation

XI.1 Les données en entrée de la macro

On se place ici dans le cadre d'un sondage à deux degrés, avec un questionnaire à trois volets interrogeant à la fois les unités primaires sélectionnées au premier degré, toutes les unités secondaires qui leur sont rattachées, et les unités secondaires sélectionnées au second degré. Un grand nombre d'enquêtes de l'INSEE auprès des ménages sont organisées sur ce schéma, l'échantillon au second degré étant sélectionné par sondage aléatoire simple avec une méthode Kish. C'est pourquoi par la suite, on désignera par « échantillon Kish » l'échantillon du second degré. La population de référence des unités Kish peut n'être qu'un sous-ensemble de la population des unités secondaires.

Exemple : échantillon de ménages dans lesquels on interroge un individu parmi les membres de 15 ans ou plus, et dans lesquels on relève des informations pour tous les individus du ménage.

Les résultats de l'enquête devront faire l'objet de trois tables de données : l'une pour les unités échantillonnées au premier degré (niveau 1), l'autre pour toutes les unités secondaires incluses dans les premières (niveau 2) et la troisième pour l'échantillon aléatoire du second degré (niveau 3). Il en sera de même pour les totaux sur la population des variables de calage, qui devront constituer trois tables de marges.

XI.1.1 Les tables SAS contenant les données

Les données relatives à l'échantillon doivent se présenter sous la forme de trois tables SAS, correspondant à chacun des niveaux d'observation.

Le nom de la table de niveau 1 est spécifié dans le paramètre (obligatoire) DATAMEN. Cette table contient, pour chaque unité sélectionnée au premier degré de sondage :

- les variables de niveau 1 qui vont être utilisées pour le redressement, ou "variables de calage" ;
- la variable de pondération initiale de l'unité primaire au premier degré de sondage ;
- une variable identifiant l'unité primaire, dont le nom est spécifié dans le paramètre (obligatoire) IDENT de la macro.

Le nom de la table de niveau 2 est spécifié dans le paramètre (obligatoire) DATAIND. Cette table contient toutes les unités secondaires incluses dans les unités primaires de l'échantillon avec :

- les variables de niveau 2 qui vont être utilisées pour le redressement, ou "variables de calage" ;
- une variable identifiant l'observation, dont le nom est spécifié dans le paramètre (obligatoire) IDENT2 de la macro. IDENT2 doit identifier totalement l'observation concernée. Ce ne peut être seulement le rang de l'individu dans le ménage, ni celui de l'établissement dans l'entreprise.
- une variable identifiant l'unité primaire à laquelle appartient l'observation de niveau 2. Ce doit être la même variable que celle spécifiée dans le paramètre IDENT.

Le nom de la table de niveau 3 est spécifié dans le paramètre (obligatoire) DATAKISH. Cette table contient les unités secondaires sélectionnées au 2ème degré de tirage avec :

- les variables de niveau 3 qui vont être utilisées pour le redressement, ou "variables de calage" ;
- une variable identifiant l'unité primaire à laquelle appartient l'unité secondaire. Ce doit être la même variable que celle spécifiée dans le paramètre IDENT ;
- un identifiant de l'observation, identique à celui contenu dans la variable spécifiée dans le paramètre IDENT2 ;
- la variable de pondération de l'unité secondaire à l'intérieur de l'unité primaire dans laquelle elle est tirée. C'est l'inverse de sa probabilité conditionnelle de tirage au second degré. Cette variable est spécifiée dans le paramètre (obligatoire) POIDKISH.

Chaque table peut bien sûr contenir toute autre variable n'intervenant pas directement dans le redressement.

XI.1.1.1 LES VARIABLES DE CALAGE

Elles peuvent être catégorielles ou numériques et obéissent aux mêmes règles que celles décrites plus haut (voir § IX.1.1.1 et IX.1.1. 2).

Les noms des variables de calage, ainsi que leurs nombres de modalités, sont spécifiés dans les tables des marges (voir § IX.1.2 et XI.1.2) :

- &MARMEN pour les variables de niveau 1;
- &MARIND pour les variables de niveau 2 ;
- &MARKISH pour les variables de niveau 3.

La macro réalise le contrôle suivant :

- une variable catégorielle a, dans la table de données correspondante, le nombre de modalités annoncé dans la table des marges.

XI.1.1.2 LA VARIABLE DE PONDERATION DE L'UNITE PRIMAIRE

C'est la variable donnant, pour chaque observation m de la table de données de niveau 1 (spécifiée dans le paramètre DATAMEN), la valeur de la pondération initiale d_m. Cette valeur est égale à l'inverse de la probabilité d'inclusion de l'unité primaire dans l'échantillon au premier degré de tirage.

La variable de pondération initiale doit être "numérique" au sens de SAS. Elle est spécifiée dans le paramètre POIDS de la macro.

Choix de la pondération initiale

Toutes les observations faites plus haut (voir § IX.1.1.3) restent valables. La restriction de certaines propriétés à la présence d'au moins une variable catégorielle parmi les variables de calage doit ici être comprise : dans la table de marges de niveau 1.

XI.1.1.3 LA VARIABLE DE PONDERATION DE L'UNITE SECONDAIRE

C'est la variable donnant, pour chaque unité de l'échantillon tirée au second degré, la valeur de sa pondération à l'intérieur de l'unité primaire dans laquelle elle a été sélectionnée. Elle est égale à l'inverse de sa probabilité conditionnelle de tirage au second degré parmi les individus éligibles de l'unité primaire.

Dans le cas d'une enquête auprès des ménages et d'un sondage équiprobable au second degré par la méthode Kish, si l'on a tiré au maximum un individu par ménage, le « poids Kish » sera égal au nombre de personnes éligibles du ménage. Si l'on a tiré au maximum deux individus par ménage, le poids est égal à 0 dans un ménage sans individus éligibles, à 1 dans un ménage comprenant une personne éligible, à NELIG/2 dans un ménage comprenant NELIG personnes éligibles.

La variable de pondération de l'unité secondaire doit être numérique au sens de SAS. Elle est spécifiée dans le paramètre (obligatoire) POIDKISH de la macro.

XI.1.1.4 AUTRES VARIABLES DE LA TABLE DES DONNEES

Les tables &DATAMEN, &DATAIND et &DATAKISH peuvent contenir d'autres variables que celles définies précédemment. En particulier, peut figurer dans la table &DATAMEN :

- une variable définissant une pondération supplémentaire des observations, spécifiée dans le paramètre PONDQK (son utilisation n'est justifiée que dans des cas très particuliers, voir référence[2]).

XI.1.1.5 OBSERVATIONS ELIMINEES

Est éliminée du calage (et donc de la table en sortie éventuelle créée par la macro) toute observation de la table en entrée ayant une valeur manquante sur l'une des variables du calage ou l'une des variables de pondération, ou prenant une valeur négative ou nulle sur l'une des variables de pondération.

Toute observation de la table &DATAIND ayant une valeur manquante sur l'une des variables citées ci-dessus entraîne l'élimination de toutes les observations de niveau 2 se rattachant au même identifiant de niveau 1 que l'observation concernée, et par conséquent l'élimination, dans la table &DATAMEN, de l'observation à laquelle elle se rattache. Pour un échantillon de ménages comprenant un questionnaire individus par exemple, toute valeur manquante sur un individu entraînera l'élimination, dans la table &DATAMEN, du ménage auquel appartient cet individu et dans la table &DATAIND, de tous les individus appartenant à ce ménage.

Il en est de même des observations de niveau 3 : toute valeur manquante entraîne l'élimination de l'observation concernée et de toutes celles appartenant à la même unité de niveau 1 dans la table &DATAKISH, ainsi que l'élimination de l'unité de niveau 1 concernée dans la table &DATAMEN.

De façon symétrique, une valeur manquante ou un poids non positif dans la table &DATAMEN entraîne l'élimination de l'unité concernée dans cette table et celle de toutes les unités qui lui sont rattachées dans les tables &DATAIND et &DATAKISH.

XI.1.1.6 CALAGE EN PRESENCE DE NON-REPONSE

Les remarques faites dans le cas du calage simple (voir § IX.1.1.6) restent valables ici. La condition de présence d'au moins une variable catégorielle pour l'utilisation de l'option 0 du paramètre ECHELLE doit s'entendre ici « parmi les variables de calage de niveau 1 ».

XI.1.2 Les tables SAS contenant les variables de calage et les marges

Les noms des variables de calage, leurs nombres de modalités, et les marges associées doivent se présenter sous la forme de trois tables SAS, dont les noms sont spécifiés dans les paramètres (obligatoires) de la macro:

- MARMEN, pour les unités de niveau 1 ;
- MARIND, pour les unités de niveau 2 ;
- MARKISH pour les unités de niveau 3.

Ces trois tables ont chacune la même structure que celle décrite dans le cas d'un calage simple (§ IX.1.2). Les mêmes contrôles sont appliqués à chacune de ces tables.

Marges des variables catégorielles données en pourcentages

L'utilisateur peut donner les valeurs des marges catégorielles en pourcentages, à condition de spécifier la valeur OUI pour le paramètre PCT de la macro. Dans ce cas, les totaux des marges des variables concernées doivent tous être égaux à 100. Le paramètre PCT s'applique aux trois tables de marges : &MARMEN, &MARIND et &MARKISH.

L'utilisateur doit alors indiquer dans le paramètre POPMEN la taille de la population pour les unités de niveau 1, dans le paramètre POPIND la taille de la population des unités de niveau 2 et dans le paramètre POPKISH la taille de la population de niveau 3.

XI.2 Syntaxe de la macro

XI.2.1 Paramètres spécifiant les tables SAS en entrée

* DATAMEN = nom de table SAS

nom de la table SAS contenant les données de niveau 1**.** Ce paramètre est **obligatoire.**

Cette table contient , pour chaque unité primaire de l'échantillon, les variables, catégorielles et numériques, du calage, et une variable identifiant. Elle contient également la variable de **pondération initiale** (sauf dans le cas où celle-ci est générée).

Voir le contenu détaillé de cette table au § XI.1.1.

* DATAIND = nom de table SAS

nom de la table SAS contenant les données de niveau 2**.** Ce paramètre est **obligatoire.**

Cette table contient , pour chaque unité de niveau 2 de l'échantillon, les variables, catégorielles et numériques, du calage. Elle doit aussi contenir obligatoirement une variable **identifiant** l'observation (paramètre IDENT2) ainsi qu'une variable identifiant l'unité primaire à laquelle appartient l'observation (paramètre IDENT). Ce dernier identifiant a le même nom que celui de la table &DATAMEN. Exemple : dans le cas d'un calage simultané ménages-individus, on doit avoir dans cette table un identifiant de l'individu et un identifiant du ménage auquel il se rattache.

* DATAKISH = nom de table SAS

nom de la table SAS contenant les données recueillies sur les unités de l'échantillon du second degré**.** Ce paramètre est **obligatoire.**

Cette table contient, pour chaque unité de l'échantillon, les variables, catégorielles et numériques, du calage, et obligatoirement, une variable **identifiant l'unité primaire** (ménage par exemple) à laquelle elle appartient. Cette variable est la même que celle spécifiée dans le paramètre IDENT et présente dans la table &DATAMEN.

Elle contient également une variable **identifiant** totalement **l'unité secondaire** : c'est la même que celle spécifiée dans le paramètre IDENT2.

Elle contient enfin la variable de **pondération conditionnelle du 2ème degré** spécifiée dans le paramètre POIDKISH.

* MARMEN= nom de table SAS

nom de la table SAS contenant les noms des variables de calage, les nombres de modalités, et les marges associées pour les unités de niveau 1. Ce paramètre est **obligatoire.**

Voir le contenu détaillé de cette table aux § XI.1.2 et IX.1.2.

* MARIND= nom de table SAS

nom de la table SAS contenant les noms des variables de calage, les nombres de modalités, et les marges associées, pour les unités de niveau 2. Ce paramètre est **obligatoire**.

Voir le contenu détaillé de cette table aux § XI.1.2 et IX.1.2.

* MARKISH= nom de table SAS

nom de la table SAS contenant les noms des variables de calage, les nombres de modalités, et les marges associées, pour les unités de niveau 3.

Ce paramètre est **obligatoire**.

Voir le contenu détaillé de cette table aux § XI.1.2 et IX.1.2.

* POIDS = variable

variable **numérique** contenant les pondérations initiales des unités primaires de l'échantillon. Elle appartient à la table &DATAMEN.

Ce paramètre est obligatoire lorsqu'il n'y pas de variable de calage catégorielle (voir § IX.1.1.3).

* PONDQK = variable

variable **numérique** de pondération des observations de niveau 1 de l'échantillon, non liée à la variable spécifiée dans le paramètre POIDS (elle appartient à la table &DATAMEN) : elle permet de moduler la fonction de calage en fonction de l'observation (voir référence [2]).

Par défaut : PONDQK = __UN, variable générée constamment égale à 1.

* POIDKISH = variable

variable **numérique** de pondération conditionnelle de l'unité secondaire dans l'unité primaire (ménage par exemple) dans laquelle elle est tirée (voir § XI.1.1.3). Cette variable doit figurer dans la table &DATAKISH.

Ce paramètre est **obligatoire**.

* IDENT = variable

variable servant à identifier les unités primaires dans les éditions et récupérée dans la table en sortie éventuelle (paramètre DATAPOI) contenant les pondérations finales. Elle doit figurer dans chacune des tables &DATAMEN, &DATAIND et &DATAKISH.

Ce paramètre est **obligatoire**.

* IDENT2 = variable

variable servant à identifier les unités secondaires de l'échantillon dans les éditions et récupérée dans les tables en sortie éventuelles (paramètres DATAPOI2 et DATAPOI3) contenant les pondérations finales. Elle doit figurer dans les tables &DATAIND et &DATAKISH. Cette variable doit identifier totalement l'entité ; elle ne peut être le simple rang de l'observation à l'intérieur de l'unité primaire.

Ce paramètre est **obligatoire**.

* PCT = OUI ou NON

si PCT vaut OUI, les marges des variables catégorielles dans les tables &MARMEN, &MARIND et &MARKISH sont données en pourcentages.

Par défaut : PCT = NON.

* POPMEN = valeur

si PCT vaut OUI, on spécifie ici l'effectif total de la population des unités de niveau 1 (dont la connaissance est nécessaire pour calculer les marges du calage).

Ce paramètre est obligatoire si PCT = OUI, ou **si ECHELLE=0 lorsque les variables de calage sont toutes numériques**.

* POPIND = valeur

si PCT vaut OUI, on spécifie ici l'effectif total de la population des unités de niveau 2 (dont la connaissance est nécessaire pour calculer les marges du calage).

Ce paramètre est obligatoire si PCT = OUI.

* POPKISH = valeur

si PCT vaut OUI, on spécifie ici l'effectif total de la population des unités de niveau 3 (dont la connaissance est nécessaire pour calculer les marges du calage). **Ce paramètre est obligatoire si PCT = OUI**.

XI.2.2 Paramètres spécifiant la méthode utilisée

* M = 1, 2, 3, 4 ou 5

numéro de la méthode, i.e. de la fonction de distance utilisée pour calculer les écarts entre les pondérations initiales et les pondérations finales :

1. méthode linéaire
2. méthode raking ratio
3. méthode logit
4. méthode linéaire tronquée
5. méthode sinus hyperbolique.

Par défaut, M=1.

* LO = valeur

borne inférieure des rapports de poids (voir note 3), lorsque l'on utilise une méthode "bornée" (logit ou linéaire tronquée).

Ce paramètre est obligatoire lorsque M = 3 ou 4.

* UP = valeur

borne supérieure des rapports de poids, lorsque l'on utilise une méthode "bornée" (logit ou linéaire tronquée).

Ce paramètre est obligatoire lorsque M = 3 ou 4.

* ALPHA = valeur

coefficient multiplicatif de l'observation, utilisé dans la méthode sinus hyperbolique. Une augmentation de sa valeur entraîne un resserrement de l'éventail des rapports de poids.

Par défaut, ALPHA=1.

* SEUIL = valeur

seuil ε pour le test d'arrêt de l'algorithme de Newton : il y a convergence lorsque le maximum (en valeur absolue) des différences entre les rapports de poids calculés lors de deux itérations successives est inférieur à ce seuil.

Par défaut : SEUIL = 0.0001.

* MAXITER = n

nombre maximum d'itérations au cours de l'algorithme de Newton : si l'algorithme n'a pas convergé en n itérations, il s'arrête.

Par défaut : MAXITER = 15.

* ECHELLE = n

nombre positif par lequel seront multipliés les poids initiaux avant calage.

Si ECHELLE = 0, le programme calcule lui-même le facteur d'échelle, qui sera égal au rapport : nombre d'unités primaires/somme des poids initiaux des UP, ce qui revient à redresser la non-réponse de façon uniforme. Dans le cas où toutes les variables de calage sont numériques, l'utilisateur devra obligatoirement renseigner le paramètre POPMEN pour utiliser l'option ECHELLE = 0.

Par défaut : ECHELLE = 1.

* COLIN=OUI ou NON

La résolution des équations de calage conduit à inverser une matrice Φ qui est singulière lorsque les variables de calage sont colinéaires. Si COLIN vaut NON, le programme recherchera l'inverse de cette matrice par la méthode classique (option INV de SAS-IML), et s'interrompra en cas de colinéarité. Si COLIN vaut OUI, le programme tiendra compte des colinéarités éventuelles entre variables de calage dans la table de données, en utilisant la technique des matrices inverses généralisées (option GINV dans SAS-IML) pour le calcul de la matrice Φ^{-1}.

Par défaut : COLIN=NON.

XI.2.3 Paramètres relatifs aux tables en sortie

* DATAPOI = nom de table SAS

nom de la table SAS contenant les pondérations finales associées aux unités de niveau 1.

- Si cette table n'existe pas, elle est créée par la macro : elle a autant d'observations que d'observations **non éliminées** de la table &DATAMEN ; elle contient la variable &POIDSFIN (voir plus loin) et la variable &IDENT.
- Si cette table existe, le paramètre MISAJOUR indique comment la macro opère sur elle.

* DATAPOI2= nom de table SAS

nom de la table SAS contenant les pondérations finales associées aux entités de niveau 2.

- Si cette table n'existe pas, elle est créée par la macro : elle a autant d'observations que d'observations **non éliminées** de la table &DATAIND ; elle contient la variable &POIDSFIN (voir plus loin) ainsi que les variables &IDENT et &IDENT2.
- Si cette table existe, le paramètre MISAJOUR indique comment la macro opère sur elle.

Ce paramètre est obligatoire si DATAPOI est renseigné.

* DATAPOI3= nom de table SAS

nom de la table SAS contenant les pondérations finales associées aux unités de l'échantillon du second degré. Ce paramètre est obligatoire si DATAPOI est renseigné.

- Si cette table n'existe pas, elle est créée par la macro : elle a autant d'observations que d'observations **non éliminées** de la table &DATAKISH ; elle contient les variables &POIDSFIN et &POIDSKISHFIN (voir plus loin) ainsi que les variables &IDENT et &IDENT2.
- Si cette table existe, le paramètre MISAJOUR indique comment la macro opère sur elle.

* POIDSFIN = variable

nom de la variable contenant les pondérations finales des unités primaires non éliminées de l'échantillon ; elle appartient aux tables &DATAPOI, &DATAPOI2 et &DATAPOI3. Par construction, ce poids de calage est la pondération totale des unités de niveau 1 et des unités de niveau 2.

Ce paramètre est obligatoire si les paramètres DATAPOI, DATAPOI2 et DATAPOI3 sont renseignés.

* LABELPOI = label

label (éventuel) attribué à la variable spécifiée dans le paramètre POIDSFIN.

Remarque : ce label ne doit pas contenir de virgule.

* POIDSKISHFIN = variable

nom de la variable contenant les pondérations finales des unités de niveau 3 non éliminées de l'échantillon ; elle appartient à la table &DATAPOI3. Ce poids de calage est égal au produit de la pondération finale de l'unité primaire (variable &POIDSFIN) par la pondération conditionnelle de l'unité secondaire à l'intérieur de l'unité primaire (variable &POIDKISH).

Ce paramètre est obligatoire lorsque le paramètre DATAPOI3 est renseigné.

* LABELPOIKISH = label

label (éventuel) attribué à la variable spécifiée dans le paramètre POIDSKISHFIN.

Remarque : ce label ne doit pas contenir de virgule.

* MISAJOUR = OUI ou NON

ce paramètre spécifie le traitement de chacune des tables &DATAPOI, &DATAPOI2, &DATAPOI3 lorsque l'une d'elles existe déjà :

- si MISAJOUR = OUI, la variable de pondération &POIDSFIN et la variable &IDENT (&IDENT2 dans &DATAPOI2 et &DATAPOI3, &POIDSKISHFIN dans &DATAPOI3), sont ajoutées à la table.
- si MISAJOUR = NON, la macro crée une nouvelle table, contenant les variables &POIDSFIN et &IDENT (&IDENT2 dans &DATAPOI2 et &DATAPOI3, &POIDSKISHFIN dans &DATAPOI3), l'ancienne table portant le même nom étant détruite.

Par défaut : MISAJOUR = OUI.

* OBSELI = OUI ou NON

si OBSELI = OUI, la macro crée trois tables SAS, de nom :

- __OBSELI contient les observations éliminées dans la table &DATAMEN, les variables du calage de niveau 1, les variables de pondération &POIDS et &PONDQK, la variable &IDENT. Elle contient aussi les variables ELIMI et ELIMK égales respectivement au nombre d'unités de niveau 2 et au nombre d'unités de niveau 3 éliminées dans l'unité primaire. Si ELIMI (respectivement ELIMK) est positif, l'unité primaire est éliminée parce qu'elle contient des unités de niveau 2 (respectivement de niveau 3) pour lesquelles une variable de calage de niveau 2 (ou 3) n'est pas renseignée. Si ELIMI=0 et ELIMK=0, l'unité primaire est éliminée parce qu'une variable de calage de niveau 1 est manquante ou parce que son poids de sondage initial n'est pas positif.
- __INDELI contient les observations éliminées dans la table &DATAIND, les variables du calage de niveau 2, les variables &IDENT et &IDENT2. Elle ne contient que les observations ayant une valeur manquante dans la table &DATAIND, à l'exclusion de celles éliminées en raison de l'élimination de l'unité primaire à laquelle ces unités appartiennent.
- __KISELI contient les observations éliminées dans la table &DATAKISH, les variables de calage de niveau 3, les variables &IDENT et &IDENT2 ainsi que la variable de pondération du second degré &POIDKISH. Elle ne contient que les observations ayant une valeur manquante dans la table &DATAKISH, à l'exclusion de celles éliminées en raison de l'élimination de l'unité primaire à laquelle ces unités appartiennent.

L'utilisateur peut imprimer, ou utiliser, chacune de ces tables après l'appel de la macro.

Par défaut : OBSELI = NON.

XI.2.4 Paramètres spécifiant les sorties imprimées

* EDITION = 0, 1, 2 ou 3

Paramètre indiquant le détail souhaité des éditions.

0. aucun résultat n'est édité, sauf les statistiques sur les poids de calage, si le paramètre STAT vaut OUI (voir plus loin).
1. la macro édite la liste des paramètres rentrés par l'utilisateur et le bilan final du calage
2. édition de la liste des paramètres, des tables des marges avant et après calage, du bilan final
3. en plus des éditions (2), la macro édite la valeur des coefficients λ après chaque itération et la valeur du critère d'arrêt.

Par défaut, EDITION vaut 3.

* CONTPOI = OUI ou NON

si CONTPOI vaut OUI, la macro édite le contenu des tables &DATAPOI, &DATAPOI2 et &DATAPOI3[12].

Par défaut : CONTPOI = OUI.

[12] Il s'agit des sorties d'une procédure CONTENTS.

* EDITPOI = OUI ou NON

si EDITPOI vaut OUI, la macro édite les valeurs des rapports de poids obtenus pour chaque case de l'hyper-tableau de contingence croisant toutes les variables, catégorielles et numériques (i.e. pour chaque combinaison de valeurs de ces variables). Ces croisements sont définis par les valeurs prises par ces variables pour chaque unité primaire de l'échantillon. Il en résulte que la valeur d'une variable de calage de niveau 2 coïncide, dans ce tableau, avec le nombre d'unités secondaires, dans l'unité primaire, ayant la modalité j de la variable y, si y est catégorielle, ou avec le total dans l'unité primaire de la variable numérique x. Les variables de calage de niveau 3 sont traitées de la même façon, les valeurs étant de plus multipliées par la pondération conditionnelle initiale du second degré de sondage (paramètre &POIDKISH).

***Remarque* :** ce tableau peut être très volumineux, surtout en présence de variables numériques.

Par défaut : EDITPOI = NON.

* STAT = OUI ou NON

si STAT vaut OUI, la macro édite des statistiques (moyenne, écart-type, quantiles, valeurs extrêmes...) et des graphiques[13] relatifs aux distributions des variables "rapport de poids" et "pondération finale".

Elle édite également un tableau des rapports de poids moyens par modalité des variables catégorielles prises une à une.

Par défaut : STAT = OUI.

* CONT = OUI ou NON

si CONT vaut OUI, un certain nombre de contrôles sont réalisés sur les paramètres de la macro (présence des paramètres obligatoires, cohérence des paramètres...), sur les valeurs données à ces paramètres (existence des tables SAS, des variables de pondération...), sur les données figurant dans les tables de marges (existence des variables, présence de toutes les marges...), ainsi que sur les variables de calage dans les tables de données.

La liste complète de ces contrôles, ainsi que des exemples de messages produits par la macro, sont donnés au chapitre XV.

Par défaut : CONT = OUI.

* NOTES = OUI ou NON

si NOTES = NON, les notes produites par SAS durant l'exécution de la macro ne sont pas éditées.

Par défaut : NOTES = NON.

[13] Il s'agit des sorties d'une procédure UNIVARIATE.

XI.3 Les tables en sortie

Chacune des tables en sortie spécifiées dans les paramètres DATAPOI, DATAPOI2 et DATAPOI3 peut être temporaire ou permanente.

Les observations de la table &DATAPOI sont celles de la table &DATAMEN non éliminées. Cette table contient la variable de pondération finale &POIDSFIN et la variable identifiant l'unité primaire &IDENT.

Les observations sont classées de la même façon dans la table en entrée &DATAMEN et dans la table en sortie &DATAPOI.

Les observations de la table &DATAPOI2 sont celles, non éliminées, de la table &DATAIND. Elle contient la variable &POIDSFIN ainsi que les identifiants &IDENT et &IDENT2.

Les observations sont classées de la même façon dans la table en entrée &DATAIND et dans la table en sortie &DATAPOI2.

Les observations de la table &DATAPOI3 sont celles, non éliminées, de la table &DATAKISH. Elle contient les variables &POIDSFIN et &POIDSKISHFIN ainsi que les identifiants &IDENT et &IDENT2.

Les observations sont classées de la même façon dans la table en entrée &DATAKISH et dans la table en sortie &DATAPOI3.

Chacune de ces tables est créée ou mise à jour selon la valeur du paramètre MISAJOUR comme indiqué au §XI.2.3.

XII. Cas particuliers de calage simultané

XII.1 Deux niveaux d'observation dans un sondage à deux degrés

On se place ici dans le cadre d'un sondage à deux degrés, dont le questionnaire recueille des informations concernant les unités primaires sélectionnées au premier degré et les unités secondaires sélectionnées au second degré.

Un exemple nous en est fourni par une enquête auprès des ménages, dans laquelle on ne recueillerait d'informations qu'au niveau de chaque ménage échantillonné et des individus de l'échantillon Kish, non pour chacun des individus du ménage. Les résultats seraient alors calés sur les effectifs de ménages dans la population, et sur les effectifs d'individus appartenant au champ de la population éligible au second degré.

XII.1.1 Les données en entrée de la macro

Les données de l'enquête doivent être présentes dans deux tables SAS, dont les noms sont spécifiés dans les paramètres :

- &DATAMEN pour l'échantillon des unités primaires ;
- &DATAKISH pour l'échantillon des unités secondaires.

Ces deux tables ont la même structure que dans le cas d'un calage simultané à trois niveaux d'observation (voir § XI.1.1).

Comme précédemment, les deux tables doivent contenir une variable identifiant l'unité primaire, dont le nom est spécifié dans le paramètre obligatoire IDENT.

La table &DATAKISH doit également contenir une variable identifiant l'unité secondaire, spécifiée dans le paramètre IDENT2.

Les variables de calage et les marges associées doivent être indiquées dans deux tables SAS dont les noms sont spécifiés dans les paramètres :

- &MARMEN pour les marges concernant la population des unités primaires ;
- &MARKISH pour les marges concernant la population des unités secondaires.

Ces tables ont même structure que dans le cas précédent (voir § XI.1.2).

XII.1.2 Syntaxe de la macro

Les **paramètres obligatoires** sont les suivants, avec les mêmes contenus et règles d'utisation que décrit plus haut (voir § XI.2) :

* **DATAMEN** = nom de table SAS

* **DATAKISH** = nom de table SAS

* **MARMEN** = nom de table SAS

* **MARKISH** = nom de table SAS
* **POIDS** = variable
* **POIDKISH** = variable
* **IDENT** = variable
* **IDENT2** = variable
* **M** = 1, 2, 3, 4 ou 5

Si PCT = OUI, les paramètres suivants doivent être renseignés :

* **POPMEN**
* **POPKISH**

Si l'on veut conserver les poids dans une table SAS, on doit renseigner les paramètres suivants :

* **DATAPOI** = nom de table SAS
* **POIDSFIN** = variable
* **DATAPOI3** = nom de table SAS
* **POIDSKISHFIN** = variable

Les paramètres suivants ne doivent pas être renseignés :

* **DATAIND**
* **MARIND**
* **DATAPOI2**
* **POPIND**

Les autres paramètres obéissent aux mêmes règles que celles déjà décrites (voir § XI.2).

XII.2 Contrainte d'égalité des poids dans la grappe, sans données sur les grappes

On suppose ici un sondage en grappes. L'information est recueillie sur chacune des unités secondaires composant la grappe échantillonnée, sans collecte d'information spécifique de l'unité primaire elle-même. On cale les résultats sur des totaux dans la population des unités secondaires, mais en imposant aux membres d'une même unité primaire de conserver des poids identiques, conformément au plan de sondage initial. Le programme assure de plus un calage sur l'effectif total des unités primaires dans la population.

Les résultats de l'enquête se présentent sous forme d'une table SAS unique dont les observations sont chacune des unités secondaires de l'échantillon, identifiées à la fois comme unités distinctes et comme membres de l'unité primaire à laquelle elles appartiennent.

Ce peut être une enquête auprès d'individus : l'unité primaire échantillonnée est le logement, mais l'information collectée concerne chacun des individus du ménage. On peut alors souhaiter caler les résultats sur des sommes d'individus dans la population, et non sur des statistiques de ménages.

Si l'on opère comme pour un calage simple avec une seule table de données, comme indiqué au § IX.2, les poids pourront être différents d'une observation à l'autre à l'intérieur d'une unité primaire. Le programme CALMAR2 réalise le calage sous cette contrainte d'égalité des poids entre unités de niveau 2 dépendant de la même unité de niveau 1, à condition de **renseigner obligatoirement les paramètres suivants** :

* **DATAIND =** nom de table SAS

nom de la table SAS contenant les données de l'enquête, de niveau 2 par rapport au niveau de tirage de l'échantillon. Cette table contient, pour chaque unité observée de l'échantillon, les variables, catégorielles et numériques, du calage, ainsi que deux variables d'identification : un identifiant complet de l'unité secondaire (individu), et une variable identifiant l'unité primaire de tirage (ménage) auquel elle se rattache. Elle contient enfin la variable de pondération initiale.

* **MARIND =** nom de table SAS

nom de la table SAS contenant les noms des variables de calage, leurs nombres de modalités et les marges associées.

* **EGALPOI =** OUI

EGALPOI doit être renseigné par OUI pour obtenir une égalité des poids entre unités secondaires appartenant à une même unité primaire, sans utiliser une table de données de niveau 1.

* **POPMEN**=valeur

effectif des unités primaires dans la population. Ce paramètre est obligatoire même si les marges sont données en effectifs.

* **IDENT =** variable

variable identifiant l'unité primaire à laquelle appartient l'unité secondaire enquêtée. **Sa présence dans la table &DATAIND est obligatoire.**

* **IDENT2=** variable

variable identifiant complètement l'observation de niveau 2. Ce ne peut être seulement son rang dans l'unité primaire. **Sa présence dans la table &DATAIND est obligatoire**.

* **POIDS=** variable

variable numérique contenant les pondérations initiales des unités primaires de l'échantillon. Par construction, le poids doit être le même pour chaque membre d'une même unité primaire. Le programme contrôle cette égalité. **Sa présence dans la table &DATAIND est obligatoire**.

* **M =** 1, 2, 3, 4 ou 5

numéro de la méthode utilisée (voir § XI.2.2).

Les autres paramètres obéissent aux mêmes règles que précédemment (voir § XI.2). En particulier, si les marges sont données en pourcentages (PCT=OUI), le paramètre POPIND doit être renseigné. Si l'on veut conserver les poids de calage dans une table SAS, il faut renseigner les paramètres DATAPOI2 et POIDSFIN.

XIII. Calage généralisé pour redressement de la non-réponse

La mise en oeuvre du programme CALMAR2 avec en option le redressement de la non-réponse au moyen de variables instrumentales distinctes des variables de calage proprement dites (voir § VII.3) nécessite l'utilisation d'un seul paramètre supplémentaire. Néanmoins, les tables de marges auront une structure particulière. Variables de calage et variables instrumentales explicatives du comportement de réponse doivent aussi respecter certaines contraintes.

XIII.1 Un nouveau paramètre : NONREP

Pour mettre en oeuvre l'option de repondération de la non-réponse par calage généralisé dans CALMAR2, il faut obligatoirement renseigner, outre les paramètres habituels décrits plus haut, le paramètre NONREP.

* NONREP=OUI ou NON

Si NONREP=OUI, CALMAR2 utilisera les variables indiquées dans la table des marges par R=1 comme des variables instrumentales destinées à redresser la non-réponse. Le programme ajustera les totaux de l'échantillon aux valeurs des marges des variables vérifiant R=0 dans les tables des marges.

Par défaut, NONREP=NON.

XIII.2 La structure d'une table de marges

L'usage de CALMAR2 en redressement de la non-réponse reste compatible avec le calage simultané. Il faut donc une table de marges par niveau d'observation de l'enquête en cas de calage simultané.

On doit trouver dans une table de marges toutes les variables de calage, qu'elles soient ou non explicatives de la non-réponse, correspondant au même niveau d'observation. Dans la table &MARMEN (respectivement &MARIND, &MARKISH), on aura les variables de calage X et les variables instrumentales Z de niveau 1 (respectivement 2 ou 3).

Chaque table de marges a la même structure et doit comporter les variables suivantes :

VAR nom de la variable.

R indicatrice du type de variable.
1 pour une variable instrumentale explicative de la non-réponse (vecteur Z).
0 pour une variable de calage (vecteur X).

N nombre de modalités de la variable ; 0 pour une variable numérique.
N est renseigné de la même manière pour une variable de calage X ou pour une variable de non-réponse Z.

MAR1 valeur de la marge associée à la modalité 1 pour une variable de calage catégorielle, valeur de la marge associée pour une variable de calage numérique.
valeur manquante (.) pour toutes les variables instrumentales Z.

...

MARj valeur de la marge associée à la modalité j pour une variable de calage catégorielle ayant au moins j modalités, valeur manquante (.) pour une variable catégorielle ayant moins de j modalités ou pour une variable numérique.
valeur manquante (.) pour toutes les variables instrumentales Z.

...

Toutes les règles habituelles s'appliquent ici, en particulier en cas de marges données en pourcentages.

XIII.3 La structure d'une table de données

Une table de données (respectivement &DATAMEN, &DATAIND, &DATAKISH) doit contenir, pour les unités du niveau correspondant :

- Toutes les variables de calage X, numériques et catégorielles, spécifiques du niveau d'observation.
- Toutes les variables instrumentales Z, explicatives de la non-réponse, numériques et catégorielles, spécifiques du niveau d'observation.
- Une variable identifiant.
- La variable de pondération initiale, présente dans la table &DATAMEN (ou &DATAIND si EGALPOI=oui).
- La variable de pondération conditionnelle du 2ème degré (présente dans la table &DATAKISH) en cas de calage simultané dans un sondage à deux degrés.

Toutes les règles énoncées plus haut (voir § XI.1.1) continuent de s'appliquer ici.

XIII.4 Les contraintes sur les variables

XIII.4.1 Les vecteurs X et Z doivent avoir même dimension

La résolution des équations de calage implique la construction puis l'inversion d'une matrice : $Z'X$, où X est le vecteur des variables de calage et Z celui des variables instrumentales explicatives du comportement de réponse. Les vecteurs X et Z doivent donc être compatibles pour ce produit. La somme du nombre de variables numériques et du nombre total de modalités des variables catégorielles doit donc être identique des deux côtés.

Pour des raisons techniques, il est également nécessaire d'avoir le même nombre de variables catégorielles (indépendamment de leur nombre de modalités) dans les deux vecteurs X et Z.

Dans le cas d'un calage simultané entre plusieurs niveaux d'observation, cette dernière égalité doit être vérifiée dans chacune des tables de marges et de données. Autrement dit, pour un calage simultané ménages-individus par exemple, on devra avoir autant de variables catégorielles de calage que de variables catégorielles de non-réponse au niveau ménages, et autant de variables catégorielles de calage que de non-réponse au niveau individus.

En revanche, la concordance entre les dimensions des deux vecteurs X et Z est nécessaire globalement et non à chacun des niveaux de calage.

Exemple 1 :

Niveau	Variables de calage (vecteur X)			Variables de non-réponse (vecteur Z)		
	Variable	Type	Nbre de modalités	Variable	Type	Nbre de modalités
Ménages	X1	catégorielle	2	ZX1	catégorielle	3
	X2	catégorielle	3	ZX2	catégorielle	4
	X3	numérique	1			
Individus	Y1	catégorielle	4	ZY1	catégorielle	2
				ZY2	numérique	1

Les variables numériques, déclarées dans les tables de marges avec un nombre de modalités nul, apportent chacune ici une coordonnée aux vecteurs X ou Z. Dans cet exemple, le vecteur X comme le vecteur Z a 10 modalités.

Si toutes les variables explicatives de la non-réponse sont observées au niveau du ménage (taille du ménage, catégorie socio-professionnelle du chef de ménage, taille de la commune de résidence...), alors que l'on souhaite caler les résultats de l'enquête également à des effectifs de population, les contraintes précédentes pourront conduire à la nécessité de compléter le vecteur Z par des variables de niveau individus, par exemple en reproduisant les variables de calage de niveau individus dans le vecteur des variables de non-réponse.

Exemple 2 :

Niveau	Variables de calage (vecteur X)			Variables de non-réponse (vecteur Z)		
	Variable	Type	Nbre de modalités	Variable	Type	Nbre de modalités
Ménages	X1	catégorielle	2	ZX1	catégorielle	3
	X2	catégorielle	3	ZX2	catégorielle	4
	X3	numérique	1			
	X4	numérique	1			
Individus	Y1	catégorielle	3	ZY1=Y1	catégorielle	3

On devra alors trouver dans la table de marges &MARIND :

- une ligne avec la variable Y1, ses marges MAR1 à MAR3, et la valeur 0 pour la variable R ;
- une ligne avec un nom de variable différent de Y1, (soit ZY1) la valeur 1 pour la variable R, et la même valeur de N que pour Y1. Sur cette ligne, les variables MAR1 à MAR3 ne seront pas renseignées.

Dans la table de données &DATAIND, l'on devra trouver, outre la variable Y1, la variable de nom ZY1, qui aura la même valeur que Y1 pour chaque individu de la table.

On doit donc avoir autant de variables catégorielles de calage que de variables catégorielles expliquant le comportement de réponse dans chaque niveau d'observation en cas de calage simultané, et le même nombre total de modalités dans les deux vecteurs.

XIII.4.2 Les vecteurs X et Z doivent être bien corrélés

L'interprétation géométrique du procédé avec une fonction de calage linéaire (Cf § VII.3.1) incite à retenir des groupes X et Z de variables corrélés entre eux. Dans la pratique, on déterminera d'abord les variables Z explicatives du comportement de réponse et corrélées aux variable d'intérêt de l'enquête, puis l'on construira le vecteur X par des variables dont on connaît les totaux sur la population et liées aux variables Z.

Deux vecteurs X et Z totalement indépendants rendraient le calage impossible ou conduiraient à des poids négatifs.

En revanche, les variables de calage X doivent être aussi indépendantes que possible entre elles. Il en va de même des variables de redressement de la non-réponse Z. Cette condition est nécessaire pour la stabilité des estimateurs. Une forte corrélation entre deux variables de calage ou entre deux variables de non-réponse entraîne une augmentation des variances des estimateurs.

XIV. Exemples

XIV.1 Un petit exemple commenté

Toutes les observations de l'échantillon se rattachent à un seul type d'unité : il s'agit d'un calage simple. Les variables catégorielles peuvent prendre des valeurs alphanumériques, ou des valeurs numériques disjointes.

XIV.1.1 Le programme

```
LIBNAME COMPIL 'D:\calmar2';
OPTIONS SASMSTORE=COMPIL MSTORED NODATE;

DATA don;
INPUT nom $ x $ y $ z pond;
CARDS;
A 1 f 1 10
B 1 h 2  0
C 1 h 3  .
D 5 f 1 11
E 5 f 3 13
F 5 h 2  7
G 5 h 2  8
H 1 h 2  8
I 5 f 2  9
J . h 2 10
K 5 h 2 14
;
DATA marges;
INPUT var $ n mar1 mar2;
CARDS;
X 2  20  60
y 2  30  50
z 0 140   .
;
TITLE "Un petit exemple commenté de calage sur marges";
%CALMAR2(DATAMEN=don,POIDS=pond,IDENT=nom,
        MARMEN=marges,M=2,EDITPOI=oui,OBSELI=oui,
        DATAPOI=sortie,POIDSFIN=pondfin,LABELPOI=pondération raking ratio)

PROC PRINT DATA=__OBSELI;
TITLE2 "Liste des observations éliminées";
RUN ;
```

DATAMEN la table en entrée est la table DON

POIDS la variable contenant les pondérations initiales, qui ici ne sont pas toutes égales, est la variable numérique POND de la table DON

IDENT la variable NOM servira d'identifiant pour les observations dans les sorties imprimées et dans la table en sortie

MARMEN la table contenant les marges est la table MARGES.

Le contenu de cette table indique que le calage va utiliser 3 variables : X et Y sont des variables catégorielles ayant deux modalités chacune (N vaut 2) et Z est une variable numérique (N vaut 0). Ces 3 variables figurent dans la table DON.

Les marges du calage pour la variable X sont respectivement 20 et 60 : cela signifie que l'effectif pondéré, après calage, de la modalité 1 (resp. de la modalité 2) de X doit être égal à 20 (resp. 60). De même les marges pour Y sont 30 et 50. La marge relative à Z est 140 : cela signifie que la somme pondérée, après calage, de Z doit être égale à 140.

M la méthode utilisée est le raking ratio

EDITPOI on demande l'édition des rapports de poids par combinaison de valeurs des variables

OBSELI la macro va créer une table, de nom __OBSELI, contenant les observations éliminées (s'il y en a)

DATAPOI la table SORTIE contiendra, si tout s'est bien passé..., les pondérations finales

POIDSFIN la variable de la table SORTIE contenant les pondérations finales s'appellera PONDFIN

LABELPOI l'étiquette "pondération raking ratio" sera attribué à la variable PONDFIN.

Les autres paramètres prennent leurs valeurs par défaut, à savoir :

PONDQK __UN : pas de pondération supplémentaire

PCT NON : les marges ne sont pas données en pourcentages

SEUIL 0.0001 : seuil pour le test d'arrêt

MAXITER 15 : nombre maximum d'itérations

ECHELLE 1 : les poids initiaux ne sont pas multipliés par un facteur d'échelle au préalable

COLIN NON : les variables ne sont pas colinéaires

EDITION 3 : tous les tableaux récapitulatifs seront édités

STAT OUI : des statistiques sur les rapports de poids et les pondérations finales seront éditées

CONTPOI OUI : le contenu de la table SORTIE sera édité

CONT OUI : des contrôles seront effectués

NOTES NON : pas d'édition des notes SAS

ATTENTION à l'ordre dans lequel sont renseignées les marges des variables catégorielles dans la table des marges !

La numérotation des variables MAR1, MAR2, ...MARp dans la table des marges doit correspondre à l'ordre alphanumérique des modalités des variables de calage présentes dans la table de données. Le programme va en effet attribuer la marge MAR1 à la première modalité, dans l'ordre alphanumérique, de la variable de calage concernée, MAR2 à la deuxième, etc...

Dans l'exemple ci-dessus, la variable X, dont la codification est numérique, prend des valeurs discontinues qui vont être triées dans l'ordre suivant : 1ère modalité = 1 2ème modalité = 5
La variable Y prend les valeurs « h » et « f », d'où : 1ère modalité = f 2ème modalité = h
Sous MAR1, on doit donc avoir les marges correspondants respectivement aux valeurs 1 de X et f de Y, sous MAR2 les marges correspondant aux valeurs 5 de X et h de Y.

XIV.1.2 La log

```
1    LIBNAME COMPIL 'd:\calmar2' ;
NOTE: Libref COMPIL was successfully assigned as follows:
      Engine:        V8
      Physical Name: d:\calmar2
2    OPTIONS SASMSTORE=COMPIL MSTORED NODATE;
3
4    DATA DON;
5    INPUT nom $ x $ y $ z pond;
6    CARDS;

NOTE: The data set WORK.DON has 11 observations and 5 variables.
NOTE: DATA statement used:
      real time           0.93 seconds
      cpu time            0.03 seconds

18   ;
19   DATA MARGES;
20   INPUT VAR $ N MAR1 MAR2;
21   CARDS;

NOTE: The data set WORK.MARGES has 3 observations and 4 variables.
NOTE: DATA statement used:
      real time           0.09 seconds
      cpu time            0.00 seconds

25   ;
26   TITLE "Un petit exemple de calage sur marges";
27   %CALMAR2(DATAMEN=don,POIDS=pond,IDENT=nom,
28            MARMEN=marges,M=2,EDITPOI=oui,OBSELI=oui,
29            DATAPOI=sortie,POIDSFIN=pondfin,LABELPOI=pondération raking ratio)


          ************************************************************
          ***   VALEUR DU CRITÈRE D'ARRÊT À L'ITÉRATION 1 : 0.56651   ***
          ************************************************************

          ************************************************************
          ***   VALEUR DU CRITÈRE D'ARRÊT À L'ITÉRATION 2 : 0.17766   ***
          ************************************************************

          ************************************************************
          ***   VALEUR DU CRITÈRE D'ARRÊT À L'ITÉRATION 3 : 0.04198   ***
          ************************************************************

          ************************************************************
          ***   VALEUR DU CRITÈRE D'ARRÊT À L'ITÉRATION 4 : 0.00322   ***
          ************************************************************

          ************************************************************
          ***   VALEUR DU CRITÈRE D'ARRÊT À L'ITÉRATION 5 : 0.00002   ***
          ************************************************************
30
31   PROC PRINT DATA=__OBSELI;
32   TITLE2 "Liste des observations éliminées";
33   RUN;

NOTE: There were 3 observations read from the data set WORK.__OBSELI.
NOTE: PROCEDURE PRINT used:
      real time           0.02 seconds
             cpu time            0.01 seconds
```

Les notes SAS ne sont pas éditées durant l'exécution de la macro. L'impression sur la log des valeurs successives du critère d'arrêt de l'algorithme (0.56651, 0.17766, 0.04198...) permet à l'utilisateur, s'il le désire, de suivre le déroulement de l'algorithme "en temps continu".

XIV.1.3 Le listing

```
                    Un petit exemple commenté de calage sur marges

                        *********************************
                        ***   PARAMÈTRES DE LA MACRO   ***
                        *********************************


TABLE(S) EN ENTRÉE :
 TABLE DE DONNÉES DE NIVEAU 1             DATAMEN   =  DON
    IDENTIFIANT DU NIVEAU 1               IDENT     =  NOM
 TABLE DE DONNÉES DE NIVEAU 2             DATAIND   =
    IDENTIFIANT DU NIVEAU 2               IDENT2    =
 TABLE DES INDIVIDUS KISH                 DATAKISH  =
 PONDÉRATION INITIALE                     POIDS     =  POND
 FACTEUR D'ÉCHELLE                        ECHELLE   =  1
 PONDÉRATION QK                           PONDQK    =  __UN
 PONDÉRATION KISH                         POIDKISH  =
 ÉGALITÉ DES POIDS DANS UN MÉNAGE         EGALPOI   =  NON

TABLE(S) DES MARGES :
 DE NIVEAU 1                              MARMEN    =  MARGES
 DE NIVEAU 2                              MARIND    =
 DE NIVEAU KISH                           MARKISH   =
 MARGES EN POURCENTAGES                   PCT       =  NON
 EFFECTIF DANS LA POPULATION :
  DES ÉLÉMENTS DE NIVEAU 1                POPMEN    =
  DES ÉLÉMENTS DE NIVEAU 2                POPIND    =
  DES ÉLÉMENTS KISH                       POPKISH   =

MÉTHODE UTILISÉE                          M         =  2
 BORNE INFÉRIEURE                         LO        =
 BORNE SUPÉRIEURE                         UP        =
 COEFFICIENT DU SINUS HYPERBOLIQUE        ALPHA     =  1
 SEUIL D'ARRÊT                            SEUIL     =  0.0001
 NOMBRE MAXIMUM D'ITÉRATIONS              MAXITER   =  15
 TRAITEMENT DES COLINÉARITÉS              COLIN     =  NON

TABLE(S) CONTENANT LA POND. FINALE
 DE NIVEAU 1                              DATAPOI       =  SORTIE
 DE NIVEAU 2                              DATAPOI2      =
 DE NIVEAU KISH                           DATAPOI3      =
 MISE À JOUR DE(S) TABLE(S) DATAPOI(2)(3) MISAJOUR      =  OUI
 PONDÉRATION FINALE                       POIDSFIN      =  PONDFIN
 LABEL DE LA PONDÉRATION FINALE           LABELPOI      =  PONDÉRATION RAKING RATIO
 PONDÉRATION FINALE DES UNITES KISH       POIDSKISHFIN  =
 LABEL DE LA PONDÉRATION KISH             LABELPOIKISH  =
 CONTENU DE(S) TABLE(S) DATAPOI(2)(3)     CONTPOI       =  OUI

ÉDITION DES RÉSULTATS                     EDITION       =  3
 ÉDITION DES POIDS                        EDITPOI       =  OUI
 STATISTIQUES SUR LES POIDS               STAT          =  OUI

CONTRÔLES                                 CONT          =  OUI
TABLE CONTENANT LES OBS. ÉLIMINÉES        OBSELI        =  OUI
NOTES SAS                                 NOTES         =  NON
```

```
                    Un petit exemple commenté de calage sur marges

    COMPARAISON ENTRE LES MARGES TIRÉES DE L'ÉCHANTILLON (AVEC LA PONDÉRATION INITIALE)
                  ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)

                                    MARGE          MARGE      POURCENTAGE    POURCENTAGE
      VARIABLE    MODALITÉ    ÉCHANTILLON    POPULATION    ÉCHANTILLON    POPULATION

         X           1             18             20          22.50          25.00
                     5             62             60          77.50          75.00

         Y           f             43             30          53.75          37.50
                     h             37             50          46.25          62.50

         Z                        152            140            .              .
```

L'effectif pondéré de la modalité 1 de la variable X dans l'échantillon vaut 18, ce qui représente 22,5% de l'effectif pondéré total[14] de l'échantillon : cette modalité est donc légèrement sous-représentée, puisque sa fréquence dans la population est de 25%.

Le total pondéré de la variable numérique Z dans l'échantillon (152) est supérieur au total de Z dans la population (140).

Note : les observations B, C et J ont été éliminées, car elles prennent respectivement une valeur nulle pour la pondération POND, une valeur manquante pour POND, une valeur manquante pour la variable X.

Un petit exemple commenté de calage sur marges

Méthode : raking ratio
Premier tableau récapitulatif de l'algorithme :
la valeur du critère d'arrêt et le nombre de poids négatifs après chaque itération

Itération	Critère d'arrêt	Poids négatifs
1	0.56651	0
2	0.17766	0
3	0.04198	0
4	0.00322	0
5	0.00002	0

Un petit exemple commenté de calage sur marges

Méthode : raking ratio
Deuxième tableau récapitulatif de l'algorithme :
les coefficients du vecteur lambda de multiplicateurs de Lagrange après chaque itération

Variable	Modalité	LAMBDA1	LAMBDA2	LAMBDA3	LAMBDA4	LAMBDA5
X	1	1.20511	1.70361	1.87331	1.88687	1.88695
X	5	1.32247	1.81959	1.99270	2.00648	2.00656
Y	f	-0.73974	-0.94297	-1.02331	-1.02984	-1.02987
Y	h	.	.	.	.	.
Z		-0.47287	-0.74661	-0.83348	-0.84035	-0.84039

Le critère d'arrêt est devenu inférieur au seuil de 0.0001 au bout de 5 itérations ; il n'y aucun poids négatif (ce qui est normal puisque c'est la méthode du raking ratio qui est utilisée). L'examen du tableau des vecteurs lambda peut se révéler utile lorsqu'il n'y pas convergence : il arrive en effet souvent dans ce cas que des composantes de lambda deviennent très élevées, "traduisant" d'une certaine façon l'impossibilité pour l'algorithme d'atteindre les marges correspondantes.

Un petit exemple commenté de calage sur marges

Méthode : raking ratio
Comparaison entre les marges finales dans l'échantillon (avec la pondération finale)
et les marges dans la population (marges du calage)

Variable	Modalité	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
X	1	20.000	20	25.00	25.00
	5	60.000	60	75.00	75.00
Y	f	30.000	30	37.50	37.50
	h	50.000	50	62.50	62.50
Z		140.000	140	.	.

Ce tableau est analogue au premier tableau, mais les marges sur l'échantillon sont calculées ici avec la pondération finale : elles doivent donc en principe être égales aux marges dans la population ; si ce n'est pas le cas, les divergences sont signalées par des *.

[14] égal à la somme de la variable de pondération initiale calculée sur les observations non éliminées

Un petit exemple commenté de calage sur marges

Méthode : raking ratio
Rapports de poids (pondérations finales / pondérations initiales)
pour chaque combinaison de valeurs des variables

OBS	X	Y	Z	Effectif combinaison	Rapport de poids
1	1	f	1	1	1.01683
2	1	h	2	1	1.22897
3	5	f	1	1	1.14602
4	5	f	2	1	0.49456
5	5	f	3	1	0.21342
6	5	h	2	3	1.38511

Ce tableau est édité car EDITPOI = OUI. Il y a dans la table en entrée une observation pour laquelle X=1, Y=f et Z=1 ; le rapport pondération finale/pondération initiale vaut 1.01683 pour cette observation... Les 3 observations pour lesquelles X=5, Y=h et Z=2 ont un rapport de poids égal à 1.38511.

Un petit exemple commenté de calage sur marges

MÉTHODE : RAKING RATIO
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: _F_ (RAPPORT DE POIDS)

Basic Statistical Measures

Location		Variability	
Mean	1.031891	Std Deviation	0.44481
Median	1.187493	Variance	0.19786
Mode	1.385113	Range	1.17169
		Interquartile Range	0.62942

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.385113
99%	1.385113
95%	1.385113
90%	1.385113
75% Q3	1.385113
50% Median	1.187493
25% Q1	0.755692
10%	0.213423
5%	0.213423
1%	0.213423
0% Min	0.213423

Extreme Observations

Lowest			Highest		
Value	nom	Obs	Value	nom	Obs
0.213423	E	3	1.14602	D	2
0.494557	I	7	1.22897	H	6
1.016827	A	1	1.38511	F	4
1.146020	D	2	1.38511	G	5
1.228966	H	6	1.38511	K	8

```
Stem Leaf                     #      Boxplot                  Normal Probability Plot
  12 3999                     4     +-----+   1.3+                          * +*+++     *
  10 25                       2     *--+--*      |                      *  * ++++
   8                                |     |      |                      +++++
   6                                +-----+      |                  ++++
   4 9                         1       |         |              +++++*
   2 1                         1       |      0.3+         ++++*
     ----+----+----+----+                          +----+----+----+----+----+----+----+----+----+----+----+
Multiply Stem.Leaf by 10**-1                           -2        -1         0        +1        +2
```

Un petit exemple commenté de calage sur marges

MÉTHODE : RAKING RATIO
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: __WFIN (PONDÉRATION FINALE)

Basic Statistical Measures

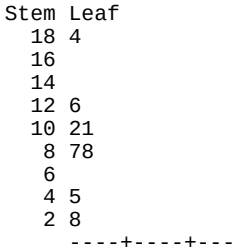
Location		Variability	
Mean	10.00000	Std Deviation	5.06121
Median	10.00000	Variance	25.61584
Mode	.	Range	16.61709
		Interquartile Range	4.77016

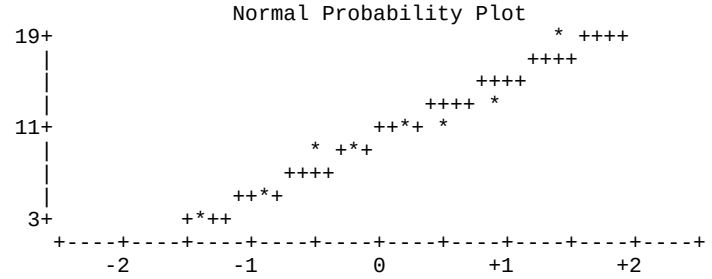
Quantiles (Definition 5)

Quantile	Estimate
100% Max	19.39158
99%	19.39158
95%	19.39158
90%	19.39158
75% Q3	11.84356
50% Median	10.00000
25% Q1	7.07340
10%	2.77449
5%	2.77449
1%	2.77449
0% Min	2.77449

Extreme Observations

Lowest			Highest		
Value	nom	Obs	Value	nom	Obs
2.77449	E	3	9.83173	H	6
4.45101	I	7	10.16827	A	1
9.69579	F	4	11.08090	G	5
9.83173	H	6	12.60622	D	2
10.16827	A	1	19.39158	K	8





Un petit exemple commenté de calage sur marges

MÉTHODE : RAKING RATIO
RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
POUR CHAQUE VALEUR DES VARIABLES

VARIABLE	MODALITE	NOMBRE D'OBSERVATIONS DE NIVEAU 1	RAPPORT DE POIDS
X	1	2	1.12290
X	5	6	1.00156
Y	f	4	0.71771
Y	h	4	1.34608
ENSEMBLE		8	1.03189

Ces sorties sont éditées car STAT = OUI.

La moyenne des 8 rapports de poids vaut 1.031891, leur écart-type 0.444812, le plus grand vaut 1.385113 (observations F G et K), le plus petit vaut 0.213423 (observation E), etc.

Le total de la pondération finale vaut 80, ce qui est normal puisque c'est l'effectif de la population. Cette pondération varie de 2.774494 (observation E) à 19.39158 (observation K), soit une étendue de 16.61709, etc.

Un petit exemple commenté de calage sur marges 40

MÉTHODE : RAKING RATIO
CONTENU DE LA TABLE sortie CONTENANT LA NOUVELLE PONDÉRATION pondfin

The CONTENTS Procedure

Data Set Name:	WORK.SORTIE	Observations:	8
Member Type:	DATA	Variables:	2
Engine:	V8	Indexes:	0
Created:	11:25 Tuesday, October 21, 2003	Observation Length:	16
Last Modified:	11:25 Tuesday, October 21, 2003	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	NO
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	4096
Number of Data Set Pages:	1
First Data Page:	1
Max Obs per Page:	252
Obs in First Data Page:	8
Number of Data Set Repairs:	0
File Name:	d:\saswork_TD157\sortie.sas7bdat
Release Created:	8.0101M0
Host Created:	WIN_NT

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Label
1	nom	Char	8	8	
2	pondfin	Num	8	0	pondération raking ratio

Ces sorties sont éditées car CONTPOI = OUI.

La macro a créé la table SORTIE, qui a 8 observations (les observations de la table DON non éliminées) et 2 variables : la variable de pondération finale PONDFIN, de label "pondération raking ratio", et la variable identifiant NOM qui figurait dans la table DON.

Un petit exemple commenté de calage sur marges

```
                ********************
                ***     BILAN    ***
                ********************
*
*   DATE : 16 JUIN 2000               HEURE : 14:03
*
*   ***********************************
*   TABLE EN ENTRÉE : DON
*   ***********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  : 11
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                : 3
*   NOMBRE D'OBSERVATIONS CONSERVÉES               : 8
*
*   VARIABLE DE PONDÉRATION : POND
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 2
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
      X (2) y (2)
*
*   TAILLE DE L'ÉCHANTILLON (PONDÉRÉ) :              80
*   TAILLE DE LA POPULATION           :              80
*
*   NOMBRE DE VARIABLES NUMÉRIQUES : 1
*   LISTE DES VARIABLES NUMÉRIQUES :
      z
*
*
*   MÉTHODE UTILISÉE : RAKING RATIO
*   LE CALAGE A ETE RÉALISÉ EN 5 ITÉRATIONS
*   LES POIDS ONT ÉTÉ STOCKÉS DANS LA VARIABLE PONDFIN DE LA TABLE SORTIE
```

Un petit exemple commenté de calage sur marges
Liste des observations éliminées

Obs	nom	X	y	z	pond	__UN
1	B	1	h	2	0	1
2	C	1	h	3	.	1
3	J		h	2	10	1

Sans commentaire.

XIV.2 Calage simultané dans un sondage en grappes

On a réalisé une enquête auprès d'un échantillon d'entreprises. Le plan de sondage est stratifié selon le chiffre d'affaires, avec tirage aléatoire simple dans les strates. En plus des résultats comptables de l'entreprise, le questionnaire recueille aussi des informations sur chaque site de l'entreprise échantillonnée. Il s'agit donc d'un sondage en grappes, la grappe étant l'entreprise et l'unité secondaire chacun de ses établissements.

XIV.2.1 Le programme

/* Données concernant les grappes */

```
/* Table échantillon */


DATA ent;
    INPUT ident $ x1ent $ x2ent $ x3ent $ x4ent $ y1ent y2ent pond;
    x3entold=x3ent;
    x4entold=x4ent;
    CARDS;

a 1 1 11 a 1 1 10
b 2 2 12 a 2 2 11
c 1 2 11 b 2 3 12
d 2 3 14 c 1 0 10
e 2 3 13 b 4 1  9
f 2 3 11 c 0 1 10
g 1 1 12 c 5 2 10
h 1 2 11 a 1 1 12
i 1 3 14 c 2 0 10
j 1 2 12 b 2 4  9
k 2 2 13 a 1 2 10
l 1 3 14 a 2 0 11
m 2 1 11 a 0 3 10
n 2 2 12 b 2 2 13
o 1 2 13 c 5 1  8
p 2 3 14 b 6 2 10
q 1 1 13 a 2 5 11
r 2 2 11 a 1 3  9
s 1 3 12 b 2 2 11
t 1 3 13 a 1 1 10
u 2 2 11 b 4 2  9
v 2 2 12 c 0 1 12
w 1 1 14 a 1 2  9
;


/* Table des marges de niveau 1 (entreprises) */


DATA margent;
     INPUT var $ r n mar1-mar4;
     CARDS;

x1ent 0 2 120 116  .  .
x2ent 0 3  60 100 76  .
x3ent 0 4  70  60 50 56
x4ent 0 3 100  70 66  .
y1ent 0 0 480   .  .  .
y2ent 0 0 410   .  .  .
;
```

/* Données concernant les établissements */

```
DATA etab;
     INPUT ident $ num $ x1etab $ x2etab $ y1etab y2etab pond;
     x1etabold=x1etab;
     x2etabold=x2etab;
     identetab=COMPRESS(ident!!num);
     CARDS;

a 1 1 a 1 1 10
a 2 2 a 1 0 10
a 3 2 b 1 3 10
b 1 2 b 1 3 11
b 2 3 a 4 2 11
c 1 3 a 2 0 12
c 2 1 b 3 1 12
d 1 2 a 0 1 10
e 1 3 a 4 5  9
e 2 2 a 1 2  9
f 1 1 b 0 3 10
g 1 1 a 2 1 10
g 2 3 b 1 0 10
g 3 2 a 2 3 10
g 4 3 a 4 1 10
h 1 1 b 1 2 12
i 1 2 b 4 2 10
i 2 3 a 1 2 10
j 1 3 a 0 2  9
k 1 2 b 1 2 10
k 2 1 a 1 4 10
l 1 2 a 2 0 11
l 2 3 a 4 0 11
m 1 1 a 0 3 10
n 1 1 b 4 2 13
n 2 3 a 1 5 13
n 3 2 a 0 1 13
o 1 1 b 5 1  8
p 1 2 b 6 2 10
q 1 3 a 1 5 11
q 2 1 a 2 3 11
r 1 1 b 0 6  9
s 1 3 b 2 4 11
t 1 1 a 0 1 10
t 2 2 b 4 1 10
u 1 1 b 4 3  9
u 2 2 a 2 5  9
v 1 2 a 0 1 12
w 1 1 a 5 2  9
w 2 2 a 1 0  9
;

/* Table des marges de niveau 2 (individus)*/

DATA margetab;
     INPUT var $ R n mar1-mar3;
     CARDS;

x1etab 0 3 140 160 114
x2etab 0 2 270 144   .
y1etab 0 0 820   .   .
y2etab 0 0 850   .   .
;


%CALMAR2(DATAMEN=ent,
         MARMEN=margent,
         IDENT=ident,
         DATAIND=etab,
         MARIND=margetab,
         IDENT2=identetab,
         POIDS=pond,
         DATAPOI=poidsm,
         DATAPOI2=poidsi,
         POIDSFIN=pond,
         CONTPOI=non)
RUN ;
```

XIV.2.2 Le listing

```
                    **********************************
                    ***   PARAMÈTRES DE LA MACRO   ***
                    **********************************


TABLE(S) EN ENTRÉE :
 TABLE DE DONNÉES DE NIVEAU 1             DATAMEN   =  ENT
    IDENTIFIANT DU NIVEAU 1               IDENT     =  IDENT
 TABLE DE DONNÉES DE NIVEAU 2             DATAIND   =  ETAB
    IDENTIFIANT DU NIVEAU 2               IDENT2    =  IDENTETAB
 TABLE DES INDIVIDUS KISH                 DATAKISH  =
 PONDÉRATION INITIALE                     POIDS     =  POND
 FACTEUR D'ÉCHELLE                        ECHELLE   =  1
 PONDÉRATION QK                           PONDQK    =  __UN
 PONDÉRATION KISH                         POIDKISH  =

TABLE(S) DES MARGES :
 DE NIVEAU 1                              MARMEN    =  MARGENT
 DE NIVEAU 2                              MARIND    =  MARGETAB
 DE NIVEAU KISH                           MARKISH   =
 MARGES EN POURCENTAGES                   PCT       =  NON
 EFFECTIF DANS LA POPULATION :
  DES ÉLÉMENTS DE NIVEAU 1                POPMEN    =
  DES ÉLÉMENTS DE NIVEAU 2                POPIND    =
  DES ÉLÉMENTS KISH                       POPKISH   =

MÉTHODE UTILISÉE                          M         =  1
 BORNE INFÉRIEURE                         LO        =
 BORNE SUPÉRIEURE                         UP        =
 COEFFICIENT DU SINUS HYPERBOLIQUE        ALPHA     =  1
 SEUIL D'ARRÊT                            SEUIL     =  0.0001
 NOMBRE MAXIMUM D'ITÉRATIONS              MAXITER   =  15
 TRAITEMENT DES COLINÉARITÉS              COLIN     =  NON

TABLE(S) CONTENANT LA POND. FINALE
 DE NIVEAU 1                              DATAPOI       =  POIDSM
 DE NIVEAU 2                              DATAPOI2      =  POIDSI
 DE NIVEAU KISH                           DATAPOI3      =
 MISE À JOUR DE(S) TABLE(S) DATAPOI(2)(3) MISAJOUR      =  OUI
 PONDÉRATION FINALE                       POIDSFIN      =  POND
 LABEL DE LA PONDÉRATION FINALE           LABELPOI      =
 PONDÉRATION FINALE DES UNITES KISH       POIDSKISHFIN  =
 LABEL DE LA PONDÉRATION KISH             LABELPOIKISH  =
 CONTENU DE(S) TABLE(S) DATAPOI(2)(3)     CONTPOI       =  NON

ÉDITION DES RÉSULTATS                     EDITION       =  3
 ÉDITION DES POIDS                        EDITPOI       =  NON
 STATISTIQUES SUR LES POIDS               STAT          =  OUI

CONTRÔLES                                 CONT          =  OUI
TABLE CONTENANT LES OBS. ÉLIMINÉES        OBSELI        =  NON
NOTES SAS                                 NOTES         =  NON
```

COMPARAISON ENTRE LES MARGES TIRÉES DE L'ÉCHANTILLON (AVEC LA PONDÉRATION INITIALE)
ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
X1ENT	1	123	120	52.12	50.85
	2	113	116	47.88	49.15
X2ENT	1	50	60	21.19	25.42
	2	105	100	44.49	42.37
	3	81	76	34.32	32.20
X3ENT	11	72	70	30.51	29.66
	12	66	60	27.97	25.42
	13	48	50	20.34	21.19
	14	50	56	21.19	23.73
X4ENT	a	103	100	43.64	42.37
	b	73	70	30.93	29.66
	c	60	66	25.42	27.97
Y1ENT		468	480	.	.
Y2ENT		421	410	.	.
X1ETAB	1	143	140	34.54	33.82
	2	154	160	37.20	38.65
	3	117	114	28.26	27.54
X2ETAB	a	259	270	62.56	65.22
	b	155	144	37.44	34.78
Y1ETAB		796	820	.	.
Y2ETAB		872	850	.	.

MÉTHODE : LINÉAIRE
PREMIER TABLEAU RÉCAPITULATIF DE L'ALGORITHME :
LA VALEUR DU CRITÈRE D'ARRÊT ET LE NOMBRE DE POIDS NÉGATIFS APRÈS CHAQUE ITÉRATION

ITÉRATION	CRITÈRE D'ARRÊT	POIDS NÉGATIFS
1	0.62318	0
2	0.00000	0

MÉTHODE : LINÉAIRE
DEUXIÈME TABLEAU RÉCAPITULATIF DE L'ALGORITHME :
LES COEFFICIENTS DU VECTEUR LAMBDA DE MULTIPLICATEURS DE LAGRANGE APRÈS CHAQUE ITÉRATION

VARIABLE	MODALITÉ	LAMBDA1	LAMBDA2
X1ENT	1	0.41877	0.41877
X1ENT	2	0.85971	0.85971
X2ENT	1	2.15329	2.15329
X2ENT	2	0.88797	0.88797
X2ENT	3	.	.
X3ENT	11	-0.32669	-0.32669
X3ENT	12	-0.67101	-0.67101
X3ENT	13	0.41118	0.41118
X3ENT	14	.	.
X4ENT	a	-0.17246	-0.17246
X4ENT	b	0.59136	0.59136
X4ENT	c	.	.
Y1ENT		-0.15113	-0.15113
Y2ENT		-0.33256	-0.33256
X1ETAB	1	-0.41023	-0.41023
X1ETAB	2	-0.08278	-0.08278
X1ETAB	3	0.51082	0.51082
X2ETAB	a	-0.43974	-0.43974
X2ETAB	b	.	.
Y1ETAB		0.05254	0.05254
Y2ETAB		-0.03720	-0.03720

MÉTHODE : LINÉAIRE
COMPARAISON ENTRE LES MARGES FINALES DANS L'ÉCHANTILLON (AVEC LA PONDÉRATION FINALE)
ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
X1ENT	1	120	120	50.85	50.85
	2	116	116	49.15	49.15
X2ENT	1	60	60	25.42	25.42
	2	100	100	42.37	42.37
	3	76	76	32.20	32.20
X3ENT	11	70	70	29.66	29.66
	12	60	60	25.42	25.42
	13	50	50	21.19	21.19
	14	56	56	23.73	23.73
X4ENT	a	100	100	42.37	42.37
	b	70	70	29.66	29.66
	c	66	66	27.97	27.97
Y1ENT		480	480	.	.
Y2ENT		410	410	.	.
X1ETAB	1	140	140	33.82	33.82
	2	160	160	38.65	38.65
	3	114	114	27.54	27.54
X2ETAB	a	270	270	65.22	65.22
	b	144	144	34.78	34.78
Y1ETAB		820	820	.	.
Y2ETAB		850	850	.	.

MÉTHODE : LINÉAIRE
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: _F_ (RAPPORT DE POIDS)

Basic Statistical Measures

Location		Variability	
Mean	1.005422	Std Deviation	0.31570
Median	1.001835	Variance	0.09967
Mode	.	Range	1.17777
		Interquartile Range	0.37652

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.554592
99%	1.554592
95%	1.476010
90%	1.451724
75% Q3	1.184393
50% Median	1.001835
25% Q1	0.807874
10%	0.591266
5%	0.466265
1%	0.376817
0% Min	0.376817

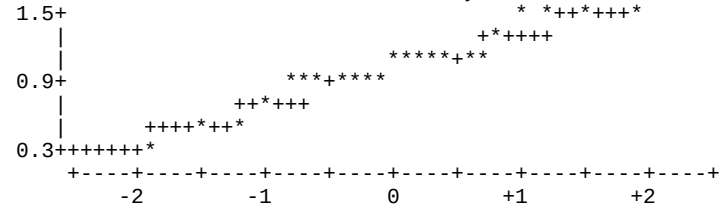
Extreme Observations

Lowest			Highest		
Value	ident	Obs	Value	ident	Obs
0.376817	t	20	1.21871	i	9
0.466265	r	18	1.44498	o	15
0.591266	j	10	1.45172	w	23
0.678614	f	6	1.47601	e	5
0.804071	n	14	1.55459	m	13

Stem	Leaf	#	Boxplot
14	4585	4	|
12	2	1	|
10	0424568	7	+--+--+
8	0134938	7	+-----+
6	8	1	|
4	79	2	|
2	8	1	|

Multiply Stem.Leaf by 10**-1

Normal Probability Plot



MÉTHODE : LINÉAIRE
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: __WFIN (PONDÉRATION FINALE)

Basic Statistical Measures

Location		Variability	
Mean	10.26087	Std Deviation	3.09128
Median	10.70164	Variance	9.55601
Mode	.	Range	11.77775
		Interquartile Range	3.30051

Quantiles (Definition 5)

Quantile	Estimate
100% Max	15.54592
99%	15.54592
95%	14.21272
90%	13.89401
75% Q3	12.18712
50% Median	10.70164
25% Q1	8.88661
10%	5.32140
5%	4.19639
1%	3.76817
0% Min	3.76817

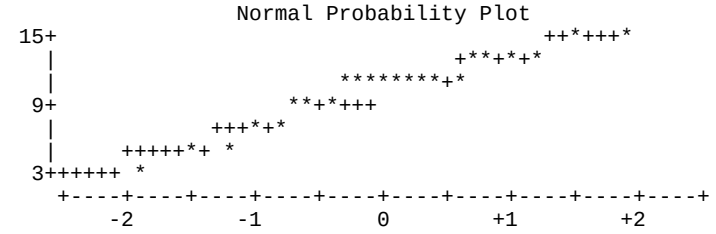
Extreme Observations

Lowest			Highest		
Value	ident	Obs	Value	ident	Obs
3.76817	t	20	13.0655	w	23
4.19639	r	18	13.2841	e	5
5.32140	j	10	13.8940	c	3
6.78614	f	6	14.2127	v	22
7.44823	u	21	15.5459	m	13

Stem	Leaf	#	Boxplot
14	25	2	
12	2139	4	+-----+
10	245702456	9	*--+--*
8	928	3	+-----+
6	84	2	
4	23	2	
2	8	1	0

----+----+----+----+

Normal Probability Plot



MÉTHODE : LINÉAIRE
RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
POUR CHAQUE VALEUR DES VARIABLES

VARIABLE	MODALITE	NOMBRE D'OBSERVATIONS DE NIVEAU 1	RAPPORT DE POIDS
X1ENT	1	12	0.98550
X1ENT	2	11	1.02715
X2ENT	1	5	1.21050
X2ENT	2	10	0.94893
X2ENT	3	8	0.94787
X3ENT	11	7	0.95992
X3ENT	12	6	0.89949
X3ENT	13	5	1.06878
X3ENT	14	5	1.13288
X4ENT	a	10	0.97398
X4ENT	b	7	0.96183
X4ENT	c	6	1.10869
ENSEMBLE		23	1.00542

MÉTHODE : LINÉAIRE
RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
POUR CHAQUE VALEUR DES VARIABLES

VARIABLE	MODALITE	NOMBRE D'OBSERVATIONS DE NIVEAU 2	RAPPORT DE POIDS
X1ETAB	1	14	0.98712
X1ETAB	2	15	1.04776
X1ETAB	3	11	0.97967
X2ETAB	a	25	1.05318
X2ETAB	b	15	0.93221
ENSEMBLE		40	1.00781

```
                  **********************
                  ***     BILAN     ***
                  **********************
*
*   DATE : 21 OCTOBRE 2003            HEURE : 14:55
*
*   ***********************************
*   TABLE EN ENTRÉE : ENT
*   ***********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  :            23
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                :            0
*   NOMBRE D'OBSERVATIONS CONSERVÉES               :            23
*
*   VARIABLE DE PONDÉRATION : POND
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 4
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
      x1ent (2) x2ent (3) x3ent (4) x4ent (3)
*
*   SOMME DES POIDS INITIAUX                                  : 236
*   TAILLE DE LA POPULATION                                   : 236
*
*   NOMBRE DE VARIABLES NUMÉRIQUES : 2
*   LISTE DES VARIABLES NUMÉRIQUES :
      y1ent y2ent
*
*
*   ***********************************
*   TABLE EN ENTRÉE : ETAB
*   ***********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  :            40
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                :            0
*   NOMBRE D'OBSERVATIONS CONSERVÉES               :            40
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 2
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
*     x1etab (3) x2etab (2)

*   SOMME DES POIDS INITIAUX                                  : 414
*   TAILLE DE LA POPULATION                                   : 414
*
*   NOMBRE DE VARIABLES NUMÉRIQUES : 2
*   LISTE DES VARIABLES NUMÉRIQUES :
      y1etab y2etab
*
*
*   MÉTHODE UTILISÉE : LINÉAIRE
*   LE CALAGE A ÉTÉ RÉALISÉ EN 2 ITÉRATIONS
*   LES POIDS ONT ÉTÉ STOCKÉS DANS LA VARIABLE POND DE LA TABLE POIDSM
*   ET DE LA TABLE POIDSI
```

Table des poids des entreprises (extrait)

OBS	IDENT	POND30
1	a	11.4277
2	b	11.0202
3	c	13.8940
4	d	11.4885
5	e	13.2841
6	f	6.7861
7	g	9.7657
8	h	10.7016
9	i	12.1871
10	j	5.3214
11	k	11.1927
12	l	8.8866
13	m	15.5459
14	n	10.4529

Table des poids des établissements (extrait)

OBS	IDENT	IDENTETAB	POND30
1	a	a1	11.4277
2	a	a2	11.4277
3	a	a3	11.4277
4	b	b1	11.0202
5	b	b2	11.0202
6	c	c1	13.8940
7	c	c2	13.8940
8	d	d1	11.4885
9	e	e1	13.2841
10	e	e2	13.2841
11	f	f1	6.7861
12	g	g1	9.7657
13	g	g2	9.7657
14	g	g3	9.7657
15	g	g4	9.7657
16	h	h1	10.7016
17	i	i1	12.1871
18	i	i2	12.1871
19	j	j1	5.3214
20	k	k1	11.1927
21	k	k2	11.1927
22	l	l1	8.8866
23	l	l2	8.8866
24	m	m1	15.5459
25	n	n1	10.4529
26	n	n2	10.4529
27	n	n3	10.4529

XIV.3 Calage d'un échantillon avec égalité des poids de calage dans la grappe

On suppose cette fois-ci un sondage en grappes dans lequel on n'interroge que les unités secondaires. Les résultats sont calés sur des totaux dans la population des individus constituant les grappes, en respectant l'égalité des poids à l'intérieur d'une même unité primaire. L'exemple ci-dessous reprend les mêmes données de niveau 2 que dans l'exemple précédent.

XIV.3.1 Le programme

```
TITLE "Calage de niveau 2 avec contrainte d'égalité sur les poids";
TITLE2;

%CALMAR2(DATAIND=etab,
         MARIND=margetab,
         POIDS=pond,
         IDENT=ident,
         IDENT2=identetab,
         POPMEN=236,
         DATAPOI2=poidegal,
         POIDSFIN=pondetab,
         CONTPOI=non,
         EGALPOI=oui,
         EDITION=2)

PROC PRINT DATA=poidegal;
     TITLE2 "table des poids des établissements";
RUN;
```

Le calage est réalisé avec la valeur par défaut du paramètre M, c'est-à-dire avec une fonction linéaire.

XIV.3.2 Les résultats (extrait du listing)

Calage de niveau 2 avec contrainte d'égalité sur les poids

COMPARAISON ENTRE LES MARGES TIRÉES DE L'ÉCHANTILLON (AVEC LA PONDÉRATION INITIALE)
ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
X1ETAB	1	143	140	34.54	33.82
	2	154	160	37.20	38.65
	3	117	114	28.26	27.54
X2ETAB	a	259	270	62.56	65.22
	b	155	144	37.44	34.78
Y1ETAB		796	820	.	.
Y2ETAB		872	850	.	.

Calage de niveau 2 avec contrainte d'égalité sur les poids

MÉTHODE : LINÉAIRE
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: _F_ (RAPPORT DE POIDS)

Basic Statistical Measures

Location		Variability	
Mean	1.004482	Std Deviation	0.14288
Median	0.961740	Variance	0.02042
Mode	1.102480	Range	0.58083
		Interquartile Range	0.19863

Quantiles (Definition 5)

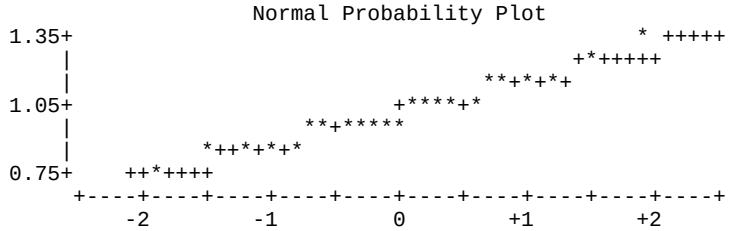
Quantile	Estimate
100% Max	1.363572
99%	1.363572
95%	1.239578
90%	1.170284
75% Q3	1.102480
50% Median	0.961740
25% Q1	0.903848
10%	0.825541
5%	0.815413
1%	0.782745
0% Min	0.782745

Extreme Observations

Lowest			Highest		
Value	ident	Obs	Value	ident	Obs
0.782745	s	19	1.10248	v	22
0.815413	r	18	1.10951	p	16
0.825541	f	6	1.17028	e	5
0.874579	h	8	1.23958	l	12
0.893829	n	14	1.36357	w	23

```
  Stem Leaf                     #             Boxplot
    13 6                        1                |
    12 4                        1                |
    11 0017                     4             +-----+
    10 16789                    5             |  +  |
     9 0122556                  7             *-----*
     8 2379                     4                |
     7 8                        1                |
       ----+----+----+----+
  Multiply Stem.Leaf by 10**-1
```

Normal Probability Plot



Calage de niveau 2 avec contrainte d'égalité sur les poids

MÉTHODE : LINÉAIRE
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: __WFIN (PONDÉRATION FINALE)

Basic Statistical Measures

Location		Variability	
Mean	10.26087	Std Deviation	1.59538
Median	10.08677	Variance	2.54522
Mode	.	Range	6.29664
		Interquartile Range	2.05664

Quantiles (Definition 5)

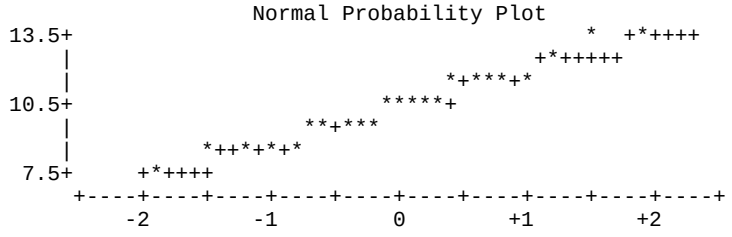
Quantile	Estimate
100% Max	13.63536
99%	13.63536
95%	13.22976
90%	12.27215
75% Q3	11.09512
50% Median	10.08677
25% Q1	9.03848
10%	8.48485
5%	8.25541
1%	7.33872
0% Min	7.33872

Extreme Observations

Lowest			Highest		
Value	ident	Obs	Value	ident	Obs
7.33872	r	18	11.6198	n	14
8.25541	f	6	11.7586	q	17
8.48485	o	15	12.2721	w	23
8.61019	s	19	13.2298	v	22
8.65566	j	10	13.6354	l	12

Stem	Leaf	#	Boxplot
13	26	2	
12	3	1	
11	00168	5	+-----+
10	01559	5	*--+--*
9	02557	5	+-----+
8	3567	4	
7	3	1	

Normal Probability Plot



```
                Calage de niveau 2 avec contrainte d'égalité sur les poids

                                    MÉTHODE : LINÉAIRE
          RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
                              POUR CHAQUE VALEUR DES VARIABLES

                                                  NOMBRE
                                               D'OBSERVATIONS     RAPPORT
                     VARIABLE    MODALITE      DE NIVEAU 2     DE POIDS

                     X1ETAB         1              14           0.98600
                     X1ETAB         2              15           1.04383
                     X1ETAB         3              11           0.97823
                     X2ETAB         a              25           1.04868
                     X2ETAB         b              15           0.93366
                     ENSEMBLE                      40           1.00555



                    **********************
                    ***     BILAN     ***
                    **********************
*
*   DATE : 21 OCTOBRE 2003            HEURE : 14:55
*
*
*   ***********************************
*   TABLE EN ENTRÉE : ETAB
*   ***********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  :            40
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                :            0
*   NOMBRE D'OBSERVATIONS CONSERVÉES               :            40
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 2
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
*     x1etab (3) x2etab (2)

*   SOMME DES POIDS INITIAUX                                    : 414
*   TAILLE DE LA POPULATION                                     : 414
*
*   NOMBRE DE VARIABLES NUMÉRIQUES : 2
*   LISTE DES VARIABLES NUMÉRIQUES :
      y1etab y2etab
*
*
*   MÉTHODE UTILISÉE : LINÉAIRE
*   LE CALAGE A ÉTÉ RÉALISÉ EN 2 ITÉRATIONS
*   LES POIDS ONT ÉTÉ STOCKÉS DANS LA VARIABLE PONDETAB DE LA TABLE POIDEGAL
```

Table des poids (extrait)

OBS	IDENT	IDENTETAB	PONDETAB
1	a	a1	9.5337
2	a	a2	9.5337
3	a	a3	9.5337
4	b	b1	10.0481
5	b	b2	10.0481
6	c	c1	11.0442
7	c	c2	11.0442
8	d	d1	11.0248
9	e	e1	10.5326
10	e	e2	10.5326
11	f	f1	8.2554
12	g	g1	9.4635
13	g	g2	9.4635
14	g	g3	9.4635
15	g	g4	9.4635
16	h	h1	10.4950
17	i	i1	9.1684
18	i	i2	9.1684
19	j	j1	8.6557
20	k	k1	9.0385
21	k	k2	9.0385
22	l	l1	13.6354
23	l	l2	13.6354

XIV.4 L'enquête permanente sur les conditions de vie des ménages de 1996

XIV.4.1 Les variables de calage

L'INSEE réalise chaque année une enquête auprès d'un échantillon de la population sur ses conditions de vie appréhendées par les conditions de travail, la formation, les relations sociales, les pratiques de loisir.

Le plan de sondage comporte deux degrés. Au premier degré est sélectionné un échantillon de logements dans la base du dernier recensement de population. Au second degré, on tire, dans chaque ménage[15] échantillonné, un individu parmi les membres de 15 ans ou plus, de façon équiprobable selon la méthode Kish.

Le questionnaire recueille des informations sur le logement et le chef de **ménage** ainsi que sur chaque **individu** du ménage (sexe, âge, profession). Par ailleurs, le volet principal du questionnaire n'est renseigné que par le membre du ménage sélectionné au second degré, appelé **individu Kish**. C'est pourquoi le redressement de cette enquête s'effectue à trois niveaux : ménage, individu et Kish, autrement dit unités primaires (UP), total des unités secondaires appartenant aux UP sélectionnées, unités secondaires de l'échantillon du second degré. Ici, la population de référence pour le calage de l'échantillon du second degré est un sous-ensemble de la population des unités secondaires : les individus de 15 ans ou plus.

Au niveau ménage

On impose à l'échantillon de ménages d'avoir la même structure que la population pour les **variables catégorielles** suivantes :

- nombre de personnes du ménage
- catégorie socioprofessionnelle du chef de ménage
- catégorie de commune (variable de stratification de l'échantillon).

Au niveau individu

On impose à l'échantillon d'individus d'avoir la même structure par sexe et par âge que la population. On construit donc une variable croisant ces deux critères selon la grille suivante :

- nombre d'hommes de 0 à 14 ans
- nombre d'hommes de 15 à 24 ans
- nombre d'hommes de 25 à 34 ans
- nombre d'hommes de 35 à 44 ans
- nombre d'hommes de 45 à 54 ans
- nombre d'hommes de 55 à 64 ans
- nombre d'hommes de 65 ans et plus

- nombre de femmes de 0 à 14 ans
- nombre de femmes de 15 à 24 ans
- nombre de femmes de 25 à 34 ans
- nombre de femmes de 35 à 44 ans
- nombre de femmes de 45 à 54 ans
- nombre de femmes de 55 à 64 ans
- nombre de femmes de 65 ans et plus

Au niveau Kish

[15] Voir note 1.

L’échantillon Kish est calé sur trois variables catégorielles :

- sexe
- âge
- type d’occupation

Le programme permet de réaliser **simultanément** ces trois redressements, en fusionnant en un seul fichier de ménages les trois niveaux d’observation, puis en opérant le calage sur ce fichier. Le poids d'un individu est alors égal au poids du ménage auquel il appartient. Celui de l’individu Kish est égal au produit de son poids « Kish » à l’intérieur du ménage auquel il appartient par le poids de calage de ce ménage.

Les variables du calage et leurs modalités sont les suivantes.

```
VARIABLES CONCERNANT LES MENAGES

Nombre de personnes du ménage : NBIND

 01 = 1 personne, 02 = 2 personnes, ... , 06 = 6 personnes et plus

Catégorie socioprofessionnelle du chef de ménage : CSPR8

 1 = agriculteurs exploitants          2 = artisans,commerç.,chefs d'entreprise
 3 = cadres et prof. intellect. sup.   4 = professions intermédiaires
 5 = employés                          6 = ouvriers
 7 =retraités                          8 = autres inactifs, non déclarés


Catégorie de commune : STRATE

 0 = communes rurales                      1 = unités urb. de moins de 20 000 h
 2 = unités urb. de 20 000 à 100 000 h     3 = unités urb. de plus de 100 000 h (sauf Paris)
 4 = unité urbaine de Paris


VARIABLES CONCERNANT LES INDIVIDUS

Sexe et âge : AGESEX2

 H-00 = hommes de moins de 15 ans    F-00 = femmes de moins de 15 ans
 H-15 = hommes de 15 à 24 ans        F-15 = femmes de 15 à 24 ans
 H-25 = hommes de 25 à 34 ans        F-25 = femmes de 25 à 34 ans
 H-35 = hommes de 35 à 44 ans        F-35 = femmes de 35 à 44 ans
 H-45 = hommes de 45 à 54 ans        F-45 = femmes de 45 à 54 ans
 H-55 = hommes de 55 à 64 ans        F-55 = femmes de 55 à 64 ans
 H-65 = hommes de 65 ans et plus     F-65 = femmes de 65 ans et plus


VARIABLES CONCERNANT LES INDIVIDUS KISH

 Sexe : SEXEK

 1 = hommes
 2 = femmes

Âge : AGEK

 A15 = 15 à 24 ans
 A25 = 25 à 59 ans
 A60 = 60 ans et plus

Type d’occupation : OCCUPA

 1 = actifs occupés, y c. militaires du contingent
 2 = chômeurs
 5 = retraités
 6 = retirés des affaires
 7 = autres inactifs
```

XIV.4.2 Le programme

```
LIBNAME compil 'd:\calmar2';
LIBNAME pcv    'd:\pcv';
OPTIONS MSTORED SASMSTORE=compil NODATE;


PROC PRINT DATA=marge1;
     TITLE 'Marges de niveau 1 (ménages)';

PROC PRINT DATA=marge2;
     TITLE 'Marges de niveau 2 (individus)';

PROC PRINT DATA=marge3;
     TITLE 'Marges de niveau Kish';
RUN ;


  /* Méthode raking-ratio */

TITLE "ENQUËTE PCV 1996 : CALAGE SIMULTANÉ MÉNAGES-INDIVIDUS-INDIVIDUS KISH";
%CALMAR2(DATAMEN=pcv.men,
         MARMEN=marge1,
         IDENT=ident,
         POIDS=poidsr,
         DATAIND=ind,
         MARIND=marge2,
         IDENT2=id,
         DATAKISH=indkish,
         MARKISH=marge3,
         POIDKISH=nbelig,
         DATAPOI=poidsm,
         DATAPOI2=poidsi,
         DATAPOI3=poidsk,
         POIDSFIN=w1,
         LABELPOI=poids grappe - M=2,
         POIDSKISHFIN=wk1,
         LABELPOIKISH=poids Kish - M=2,
         CONT=NON,
         M=2,
         PCT=oui,
         POPMEN=23450826,
         POPIND=57221469,
         POPKISH=46678524)


  /* Méthode logit */

%CALMAR2(DATAMEN=pcv.men,
         MARMEN=marge1,
         POIDS=poidsr,
         IDENT=ident,
         DATAIND=ind,
         MARIND=marge2,
         IDENT2=id,
         DATAKISH=indkish,
         MARKISH=marge3,
         POIDKISH=nbelig,
         DATAPOI=poidsm,
         DATAPOI2=poidsi,
         DATAPOI3=poidsk,
         POIDSFIN=w2,
         LABELPOI=poids grappe - M=3,
         POIDSKISHFIN=wk2,
         LABELPOIKISH=poids Kish - M=3,
         M=3,
         LO=0.2,
         UP=3.5,
         PCT=oui,
         POPMEN=23450826,
         POPKISH=46678524,
         POPIND=57221469)

PROC PRINT DATA=POIDSM(OBS=10);
     TITLE "POIDS DES MÉNAGES (TABLE DATAPOI)";

PROC PRINT DATA=POIDSI(OBS=20);
     TITLE "POIDS DES INDIVIDUS (TABLE DATAPOI2)";
RUN;
```

La macro CALMAR2 est utilisée d'abord avec la méthode raking-ratio, puis avec la méthode logit LO=0.5 UP=3.0. Ces valeurs de LO et UP conduisent à une étendue des rapports de poids minimale. Le lecteur comparera sur les listings suivants les graphiques stem-and-leaf produits respectivement par les deux méthodes.

XIV.4.3 Extraits du listing

MARGES DE NIVEAU 1 (MÉNAGES)

OBS	VAR	N	MAR1	MAR2	MAR3	MAR4	MAR5	MAR6	MAR7	MAR8
1	strate	5	24.2330	16.1398	13.0386	29.5040	17.0846	.	.	.
2	nbind	6	29.5543	31.3550	16.5229	14.4103	5.7578	2.3997	.	.
3	pcspr8	8	1.8241	5.1599	9.6317	13.1188	11.1593	20.2326	29.9571	8.9165

MARGES DE NIVEAU 2 (INDIVIDUS)

OBS	VAR	N	MAR1	MAR2	MAR3	MAR4	MAR5	MAR6	MAR7
1	agsex2	14	9.00460	6.70735	7.49493	7.50256	6.51283	4.95019	9.16460

OBS	MAR8	MAR9	MAR10	MAR11	MAR12	MAR13	MAR14
1	9.42021	6.89352	7.42412	7.32678	6.50067	4.64641	6.45123

MARGES DE NIVEAU KISH

OBS	VAR	N	MAR1	MAR2	MAR3	MAR4	MAR5
1	occupa	5	48.0809	7.4228	17.0848	4.1170	23.2945
2	sexek	2	48.1062	51.8938	.	.	.
3	agek	3	16.6728	58.1525	25.1747	.	.

ENQUÊTE PCV 1996 : CALAGE SIMULTANÉ MÉNAGES-INDIVIDUS-INDIVIDUS KISH

COMPARAISON ENTRE LES MARGES TIRÉES DE L'ÉCHANTILLON (AVEC LA PONDÉRATION INITIALE)
ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
NBIND	01	6256014.60	6930727.47	26.68	29.55
	02	7561307.76	7353006.49	32.24	31.36
	03	4031002.88	3874756.53	17.19	16.52
	04	3547735.99	3379334.38	15.13	14.41
	05	1442899.30	1350251.66	6.15	5.76
	06	611865.47	562749.47	2.61	2.40
PCSPR8	1	513694.32	427766.52	2.19	1.82
	2	1145725.13	1210039.17	4.89	5.16
	3	2504203.76	2258713.21	10.68	9.63
	4	3541518.75	3076466.96	15.10	13.12
	5	2787565.91	2616948.03	11.89	11.16
	6	4973208.26	4744711.82	21.21	20.23
	7	7377530.13	7025187.40	31.46	29.96
	8	607379.75	2090992.90	2.59	8.92
STRATE	0	5981356.74	5682838.66	25.51	24.23
	1	3862976.53	3784916.41	16.47	16.14
	2	3215164.87	3057659.40	13.71	13.04
	3	6822010.97	6918931.70	29.09	29.50
	4	3569316.89	4006479.82	15.22	17.08
AGESEX2	F_00	5730308.43	5152564.40	9.72	9.00
	F_15	3900380.57	3838044.20	6.62	6.71
	F_25	4009306.78	4288709.05	6.80	7.49
	F_35	4590143.84	4293075.04	7.79	7.50
	F_45	3937470.59	3726737.00	6.68	6.51
	F_55	3049475.34	2832571.44	5.17	4.95
	F_65	5082159.98	5244118.75	8.62	9.16
	H_00	5797653.28	5390382.54	9.84	9.42
	H_15	4301973.86	3944573.41	7.30	6.89
	H_25	3943826.07	4248190.52	6.69	7.42
	H_35	4097223.20	4192491.15	6.95	7.33
	H_45	3817688.63	3719778.87	6.48	6.50
	H_55	2893109.85	2658744.06	4.91	4.65
	H_65	3789674.42	3691488.57	6.43	6.45
AGEK	A15	7669492.99	7782616.95	16.21	16.67
	A25	27726588.67	27144728.67	58.62	58.15
	A60	11906771.08	11751178.38	25.17	25.17
OCCUPA	1	23299326.13	22443454.45	49.26	48.08
	2	3348671.72	3464853.48	7.08	7.42
	5	8278043.56	7974932.47	17.50	17.08
	6	2155718.40	1921754.83	4.56	4.12
	7	10221092.94	10873528.77	21.61	23.29
SEXEK	1	22762805.81	22455264.11	48.12	48.11
	2	24540046.92	24223259.89	51.88	51.89

MÉTHODE : RAKING RATIO
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: _F_ (RAPPORT DE POIDS)

Basic Statistical Measures

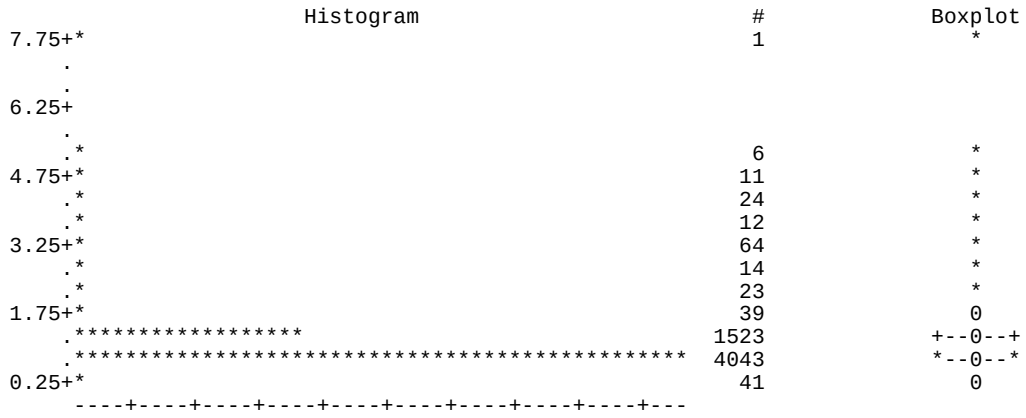
Location		Variability	
Mean	0.998358	Std Deviation	0.45182
Median	0.933594	Variance	0.20414
Mode	0.933594	Range	7.70848
		Interquartile Range	0.16476

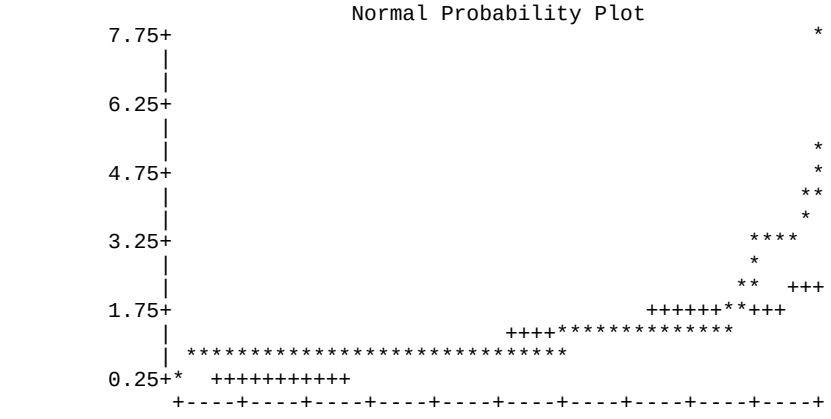
Quantiles (Definition 5)

Quantile	Estimate
100% Max	7.7310494
99%	3.3298480
95%	1.2716225
90%	1.1443734
75% Q3	1.0231171
50% Median	0.9335943
25% Q1	0.8583599
10%	0.7492241
5%	0.6737461
1%	0.5352126
0% Min	0.0225694

Extreme Observations

Lowest			Highest		
Value	IDENT	Obs	Value	IDENT	Obs
0.0225694	7369016140	4317	5.13208	2463009210	1417
0.0377668	7363016270	4167	5.25358	2163007440	763
0.0719417	7363016360	4168	5.25358	2169007040	953
0.0754055	7263010880	3840	5.33847	1169015040	444
0.0824585	7269008290	3964	7.73105	1169014560	439





```
                          -2        -1         0        +1        +2
```

ENQUÊTE PCV 1996 : CALAGE SIMULTANÉ MÉNAGES-INDIVIDUS-INDIVIDUS KISH

MÉTHODE : RAKING RATIO

STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: __WFIN (PONDÉRATION FINALE)

Basic Statistical Measures

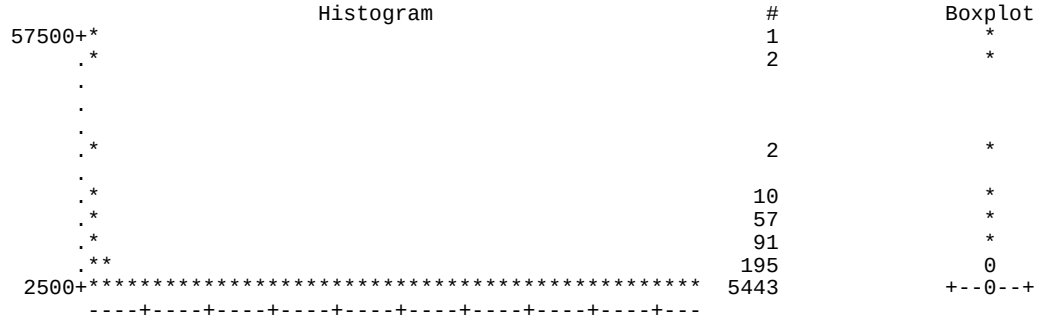
Location		Variability	
Mean	4042.549	Std Deviation	2357
Median	3661.903	Variance	5553740
Mode	3668.828	Range	57150
		Interquartile Range	671.99474

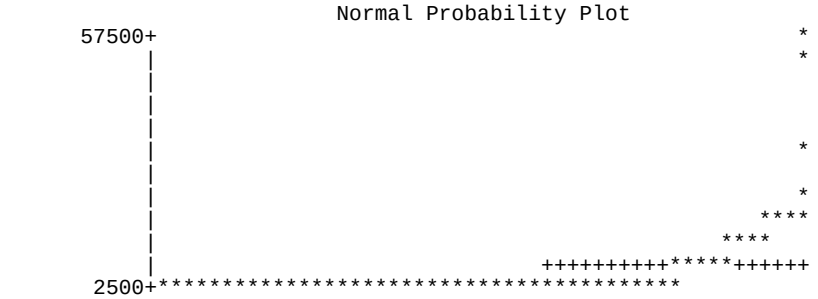
Quantiles (Definition 5)

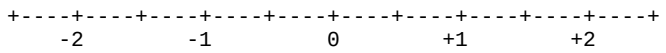
Quantile	Estimate
100% Max	57238.6216
99%	16337.1992
95%	5382.3489
90%	4559.5354
75% Q3	4043.1610
50% Median	3661.9027
25% Q1	3371.1662
10%	2943.9559
5%	2648.4147
1%	2100.5165
0% Min	88.6932

Extreme Observations

Lowest			Highest		
Value	IDENT	Obs	Value	IDENT	Obs
88.6932	7369016140	4317	30381.4	1169014560	439
147.7987	7363016270	4167	31776.8	9363022800	5482
281.5406	7363016360	4168	50676.8	1163001300	6
295.0960	7263010880	3840	53909.8	4263006680	2720
324.0445	7269008290	3964	57238.6	7369004460	4226






-2
-1
0
+1
+2

MÉTHODE : RAKING RATIO
RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
POUR CHAQUE VALEUR DES VARIABLES

VARIABLE	MODALITE	NOMBRE D'OBSERVATIONS DE NIVEAU 1	RAPPORT DE POIDS
NBIND	01	1539	1.10177
NBIND	02	1860	0.97015
NBIND	03	1000	0.96375
NBIND	04	885	0.95573
NBIND	05	361	0.93717
NBIND	06	156	0.91975
PCSPR8	1	124	0.83206
PCSPR8	2	290	1.05635
PCSPR8	3	624	0.89951
PCSPR8	4	870	0.87082
PCSPR8	5	682	0.93740
PCSPR8	6	1237	0.95619
PCSPR8	7	1831	0.95318
PCSPR8	8	143	3.46615
STRATE	0	1453	0.94977
STRATE	1	966	0.98264
STRATE	2	805	0.95330
STRATE	3	1689	1.00755
STRATE	4	888	1.11834
ENSEMBLE		5801	0.99836

MÉTHODE : RAKING RATIO
RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
POUR CHAQUE VALEUR DES VARIABLES

VARIABLE	MODALITE	NOMBRE D'OBSERVATIONS DE NIVEAU 2	RAPPORT DE POIDS
AGESEX2	F_00	1422	0.90139
AGESEX2	F_15	971	0.96530
AGESEX2	F_25	987	1.07336
AGESEX2	F_35	1138	0.93754
AGESEX2	F_45	988	0.94319
AGESEX2	F_55	758	0.93067
AGESEX2	F_65	1268	1.03289
AGESEX2	H_00	1435	0.93446
AGESEX2	H_15	1073	0.91257
AGESEX2	H_25	967	1.08474
AGESEX2	H_35	1025	1.02314
AGESEX2	H_45	950	0.97356
AGESEX2	H_55	718	0.92031
AGESEX2	H_65	940	0.97453
ENSEMBLE		14640	0.97076

MÉTHODE : RAKING RATIO
RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
POUR CHAQUE VALEUR DES VARIABLES

VARIABLE	MODALITE	NOMBRE D'INDIVIDUS KISH	RAPPORT DE POIDS
AGEK	A15	700	1.10799
AGEK	A25	3326	0.98400
AGEK	A60	1775	0.98204
OCCUPA	1	2799	0.95295
OCCUPA	2	399	1.04805
OCCUPA	5	1248	0.95166
OCCUPA	6	306	0.88869
OCCUPA	7	1049	1.18818
SEXEK	1	2633	0.99675
SEXEK	2	3168	0.99970
ENSEMBLE		5801	0.99836

```
          ENQUËTE PCV 1996 : CALAGE SIMULTANÉ MÉNAGES-INDIVIDUS-INDIVIDUS KISH


                  **********************
                  ***     BILAN     ***
                  **********************
*
*   DATE : 22 OCTOBRE 2003            HEURE : 09:06
*
*   ***********************************
*   TABLE EN ENTRÉE : PCV.MEN
*   ***********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  :            5801
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                :            0
*   NOMBRE D'OBSERVATIONS CONSERVÉES               :            5801
*
*   VARIABLE DE PONDÉRATION : POIDSR
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 3
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
      nbind (6) pcspr8 (8) strate (5)
*
*   SOMME DES POIDS INITIAUX                                  : 23450826
*   TAILLE DE LA POPULATION                                   : 23450826
*
*
*   ***********************************
*   TABLE EN ENTRÉE : IND
*   ***********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  :            14640
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                :            0
*   NOMBRE D'OBSERVATIONS CONSERVÉES               :            14640
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 1
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
*     agesex2 (14)

*   SOMME DES POIDS INITIAUX                                  : 58940395
*   TAILLE DE LA POPULATION                                   : 57221469
*
*   ***********************************
*   TABLE EN ENTRÉE : INDKISH
*   ***********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  :            5801
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                :            0
*   NOMBRE D'OBSERVATIONS CONSERVÉES               :            5801
*
*   VARIABLE DE PONDÉRATION KISH :              NBELIG
*   NOMBRE MAXIMUM D'UNITES SECONDAIRES PAR UP : 1
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 3
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
      agek (3) occupa (5) sexek (2)
*
```

MÉTHODE : LOGIT, INF=0.2, SUP=3.5

STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)

The UNIVARIATE Procedure

Basic Statistical Measures

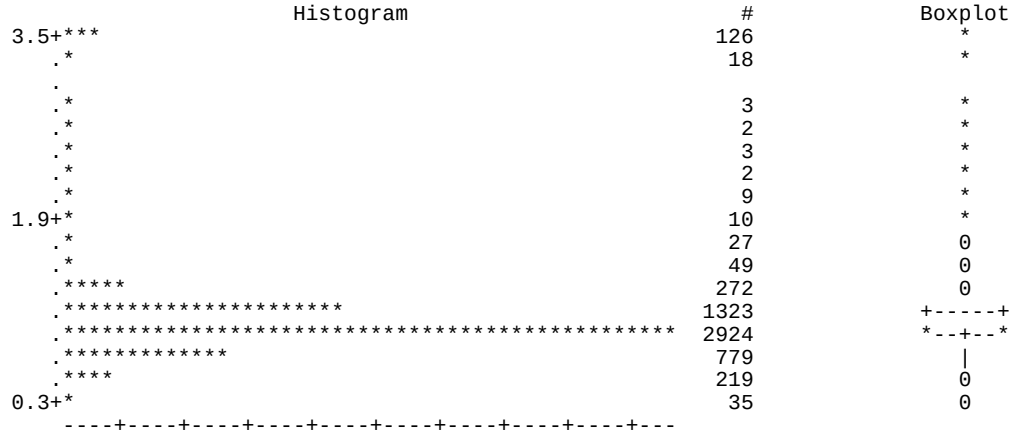
Location		Variability	
Mean	0.997902	Std Deviation	0.44085
Median	0.930899	Variance	0.19435
Mode	0.929633	Range	3.29142
		Interquartile Range	0.19495

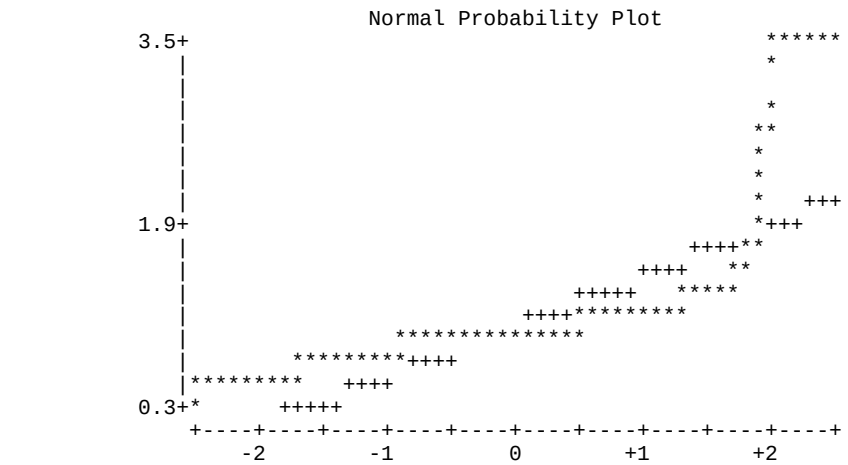
Quantiles (Definition 5)

Quantile	Estimate
100% Max	3.491424
99%	3.448394
95%	1.342990
90%	1.183708
75% Q3	1.040015
50% Median	0.930899
25% Q1	0.845060
10%	0.705367
5%	0.610017
1%	0.437957
0% Min	0.200005

Extreme Observations

Lowest			Highest		
Value	IDENT	Obs	Value	IDENT	Obs
0.200005	7369016140	4317	3.48030	2169007040	953
0.200016	7363016270	4167	3.48108	1169015040	444
0.200233	7363016360	4168	3.48221	4269003100	2786
0.200256	7263010880	3840	3.48250	2463009210	1417
0.200307	7269008290	3964	3.49142	1169014560	439





ENQUËTE PCV 1996 : CALAGE SIMULTANÉ MÉNAGES-INDIVIDUS-INDIVIDUS KISH

MÉTHODE : LOGIT, INF=0.2, SUP=3.5
CONTENU DE LA TABLE poidsm CONTENANT LA NOUVELLE PONDÉRATION w2

The CONTENTS Procedure

Data Set Name:	WORK.POIDSM	Observations:	5801
Member Type:	DATA	Variables:	3
Engine:	V8	Indexes:	0
Created:	10:11 Wednesday, October 22, 2003	Observation Length:	32
Last Modified:	10:11 Wednesday, October 22, 2003	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	NO
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	4096
Number of Data Set Pages:	47
First Data Page:	1
Max Obs per Page:	126
Obs in First Data Page:	85
Number of Data Set Repairs:	0
File Name:	d:\saswork_TD109\poidsm.sas7bdat
Release Created:	8.0101M0
Host Created:	WIN_NT

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Label
1	IDENT	Char	10	16	
2	w1	Num	8	0	poids grappe - M=2
3	w2	Num	8	8	poids grappe - M=3

CONTENU DE LA TABLE poidsi CONTENANT LA NOUVELLE PONDÉRATION w2

The CONTENTS Procedure

Data Set Name:	WORK.POIDSI	Observations:	14640
Member Type:	DATA	Variables:	4
Engine:	V8	Indexes:	0
Created:	10:11 Wednesday, October 22, 2003	Observation Length:	40
Last Modified:	10:11 Wednesday, October 22, 2003	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	NO
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	4096
Number of Data Set Pages:	146
First Data Page:	1
Max Obs per Page:	101
Obs in First Data Page:	65
Number of Data Set Repairs:	0
File Name:	d:\saswork_TD109\poidsi.sas7bdat
Release Created:	8.0101M0
Host Created:	WIN_NT

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Label
2	ID	Char	12	26	
1	IDENT	Char	10	16	
3	w1	Num	8	0	poids grappe - M=2
4	w2	Num	8	8	poids grappe - M=3

ENQUËTE PCV 1996 : CALAGE SIMULTANÉ MÉNAGES-INDIVIDUS-INDIVIDUS KISH

MÉTHODE : LOGIT, INF=0.2, SUP=3.5
CONTENU DE LA TABLE poidsk CONTENANT LA NOUVELLE PONDÉRATION w2

The CONTENTS Procedure

Data Set Name:	WORK.POIDSK	Observations:	5801
Member Type:	DATA	Variables:	6
Engine:	V8	Indexes:	0
Created:	10:11 Wednesday, October 22, 2003	Observation Length:	56
Last Modified:	10:11 Wednesday, October 22, 2003	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	NO
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	8192
Number of Data Set Pages:	41
First Data Page:	1
Max Obs per Page:	145
Obs in First Data Page:	116
Number of Data Set Repairs:	0
File Name:	d:\saswork_TD109\poidsk.sas7bdat
Release Created:	8.0101M0
Host Created:	WIN_NT

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Label
1	ID	Char	12	32	
2	IDENT	Char	10	44	
3	w1	Num	8	0	poids grappe - M=2
5	w2	Num	8	16	poids grappe - M=3
4	wk1	Num	8	8	poids Kish – M=2
6	wk2	Num	8	24	poids Kish – M=3

POIDS DES MÉNAGES (TABLE DATAPOI)

Obs	IDENT	poids grappe M=2	poids grappe M=3
1	1163000280	4807.41	4884.17
2	1163000820	4189.01	4147.90
3	1163000990	4293.89	4396.11
4	1163001140	4034.50	4134.50
5	1163001210	4343.79	4346.50
6	1163001300	50676.84	57864.80
7	1163001580	3570.97	3533.73
8	1163001760	3977.23	3942.02
9	1163001850	4303.95	4353.40
10	1163002150	3679.30	3862.76

POIDS DES INDIVIDUS (TABLE DATAPOI2)

Obs	IDENT	ID	poids grappe M=2	poids grappe M=3
1	1163000280	116300028001	4807.41	4884.17
2	1163000820	116300082001	4189.01	4147.90
3	1163000990	116300099001	4293.89	4396.11
4	1163001140	116300114001	4034.50	4134.50
5	1163001210	116300121001	4343.79	4346.50
6	1163001210	116300121002	4343.79	4346.50
7	1163001210	116300121003	4343.79	4346.50
8	1163001300	116300130001	50676.84	57864.80
9	1163001300	116300130002	50676.84	57864.80
10	1163001580	116300158001	3570.97	3533.73
11	1163001580	116300158002	3570.97	3533.73
12	1163001580	116300158003	3570.97	3533.73
13	1163001760	116300176001	3977.23	3942.02
14	1163001760	116300176002	3977.23	3942.02
15	1163001850	116300185001	4303.95	4353.40
16	1163001850	116300185002	4303.95	4353.40
17	1163002150	116300215001	3679.30	3862.76
18	1163002310	116300231001	2930.91	2850.99
19	1163002310	116300231002	2930.91	2850.99
20	1163002310	116300231003	2930.91	2850.99

POIDS DES INDIVIDUS KISH (TABLE DATAPOI3)

Obs	ID	IDENT	poids grappe M=2	poids unité Kish M=2	poids grappe M=3	poids unité Kish M=3
1	1163000280	1163000280	4807.41	4807.41	4884.17	4884.17
2	1163000820	1163000820	4189.01	4189.01	4147.90	4147.90
3	1163000990	1163000990	4293.89	4293.89	4396.11	4396.11
4	1163001140	1163001140	4034.50	4034.50	4134.50	4134.50
5	116300121001	1163001210	4343.79	8687.57	4346.50	8693.01
6	116300130002	1163001300	50676.84	101353.69	57864.80	115729.60
7	116300158002	1163001580	3570.97	10712.92	3533.73	10601.18
8	116300176001	1163001760	3977.23	7954.47	3942.02	7884.03
9	116300185001	1163001850	4303.95	8607.91	4353.40	8706.79
10	1163002150	1163002150	3679.30	3679.30	3862.76	3862.76
11	116300231001	1163002310	2930.91	8792.72	2850.99	8552.97
12	116300248002	1163002480	3771.96	15087.82	3528.31	14113.25
13	1163002590	1163002590	4769.07	4769.07	4827.34	4827.34
14	116300260002	1163002600	4326.45	8652.90	4365.89	8731.77
15	1163002860	1163002860	4034.56	4034.56	4204.93	4204.93
16	116300293002	1163002930	3034.05	6068.11	3030.50	6060.99
17	1163003010	1163003010	4293.89	4293.89	4396.11	4396.11
18	116300312001	1163003120	4579.18	9158.35	4691.35	9382.70
19	116300367001	1163003670	3707.56	14830.22	3446.82	13787.30
20	1163003740	1163003740	4003.92	4003.92	4102.84	4102.84

IDENT est l'identifiant du ménage et ID celui de l'individu.

XIV.5 Calage généralisé pour redressement de non-réponse

Les données proviennent de la simulation d'un échantillon de non-répondants dans le fichier de l'enquête sur les conditions de vie des ménages de 1996. Bien que le questionnaire comporte plusieurs niveaux d'observation (ménages, individus, Kish), le programme suivant ajuste les résultats sur une structure de ménages : il s'agit d'un calage à un seul niveau.

XIV.5.1 Les variables de calage

Le vecteur *X* des variables de calage est constitué des quatre variables catégorielles suivantes, aux modalités binaires, mesurées au moment de la constitution de la base de sondage, c'est-à-dire en 1990 :

- taille du ménage : SEUL90
  1=personnes seules 0=autres ménages
- activité du chef de ménage : INACT90
  1=inactifs 0=actifs
- nationalité du chef de ménage : ETRPR90
  1=étranger 0=français
- commune de résidence : PARIS90
  1=Paris 0=autre commune

Le vecteur *Z* des variables instrumentales explicatives de la non-réponse est constitué des mêmes variables, mais mesurées au moment de l'enquête auprès des répondants, c'est-à-dire en 1996 :

- taille du ménage : SEUL96
- activité du chef de ménage : INACT96
- nationalité du chef de ménage : ETRPR96
- commune de résidence : PARIS96.

XIV.5.2 Le programme

```
LIBNAME base   'd:\calage\simul3';
LIBNAME compil 'd:\calmar2';
OPTIONS MSTORED SASMSTORE=compil ;


PROC PRINT DATA=base.marge2b;
            TITLE 'Table des marges';
RUN ;

%CALMAR2(DATAMEN=echant,
         MARMEN=base.marge2b,
         IDENT=ident,
         POIDS=dk,
         NONREP=oui,
         DATAPOI=poids,
         POIDSFIN=wcal,
         EDITION=2)
```

XIV.5.3 Le listing (extrait)

TABLE DES MARGES

OBS	VAR	N	R	MAR1	MAR2
1	seul90	2	0	3933	1172
2	inact90	2	0	3153	1952
3	paris90	2	0	4315	790
4	etrpr90	2	0	4865	240
5	seul96	2	1	.	.
6	inact96	2	1	.	.
7	paris96	2	1	.	.
8	etrpr96	2	1	.	.

```
**********************************
***   PARAMÈTRES DE LA MACRO   ***
**********************************


TABLE(S) EN ENTRÉE :
 TABLE DE DONNÉES DE NIVEAU 1              DATAMEN   =  ECHANT
    IDENTIFIANT DU NIVEAU 1                IDENT     =  IDENT
 TABLE DE DONNÉES DE NIVEAU 2              DATAIND   =
    IDENTIFIANT DU NIVEAU 2                IDENT2    =
 TABLE DES INDIVIDUS KISH                  DATAKISH  =
 PONDÉRATION INITIALE                      POIDS     =  DK
 FACTEUR D'ÉCHELLE                         ECHELLE   =  1
 PONDÉRATION QK                            PONDQK    =  __UN
 PONDÉRATION KISH                          POIDKISH  =
 ÉGALITÉ DES POIDS DANS UN MÉNAGE          EGALPOI   =  NON

TABLE(S) DES MARGES :
 DE NIVEAU 1                               MARMEN    =  BASE.MARGE2B
 DE NIVEAU 2                               MARIND    =
 DE NIVEAU KISH                            MARKISH   =
 MARGES EN POURCENTAGES                    PCT       =  NON
 EFFECTIF DANS LA POPULATION :
  DES ÉLÉMENTS DE NIVEAU 1                 POPMEN    =
  DES ÉLÉMENTS DE NIVEAU 2                 POPIND    =
  DES ÉLÉMENTS KISH                        POPKISH   =

REDRESSEMENT DE LA NON-RÉPONSE DEMANDÉ :   NONREP    =  oui

MÉTHODE UTILISÉE                           M         =  1
 BORNE INFÉRIEURE                          LO        =
 BORNE SUPÉRIEURE                          UP        =
 COEFFICIENT DU SINUS HYPERBOLIQUE         ALPHA     =  1
 SEUIL D'ARRÊT                             SEUIL     =  0.0001
 NOMBRE MAXIMUM D'ITÉRATIONS               MAXITER   =  15
 TRAITEMENT DES COLINÉARITÉS               COLIN     =  NON

TABLE(S) CONTENANT LA POND. FINALE
 DE NIVEAU 1                               DATAPOI       =  POIDS
 DE NIVEAU 2                               DATAPOI2      =
 DE NIVEAU KISH                            DATAPOI3      =
 MISE À JOUR DE(S) TABLE(S) DATAPOI(2)(3)  MISAJOUR      =  OUI
 PONDÉRATION FINALE                        POIDSFIN      =  WCAL
 LABEL DE LA PONDÉRATION FINALE            LABELPOI      =
 PONDÉRATION FINALE DES UNITES KISH        POIDSKISHFIN  =
 LABEL DE LA PONDÉRATION KISH              LABELPOIKISH  =
 CONTENU DE(S) TABLE(S) DATAPOI(2)(3)      CONTPOI       =  OUI

ÉDITION DES RÉSULTATS                      EDITION       =  2
 ÉDITION DES POIDS                         EDITPOI       =  NON
 STATISTIQUES SUR LES POIDS                STAT          =  OUI

CONTRÔLES                                  CONT          =  OUI
```

COMPARAISON ENTRE LES MARGES TIRÉES DE L'ÉCHANTILLON (AVEC LA PONDÉRATION INITIALE)
ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
ETRPR90	0	3745	4865	95.78	95.30
	1	165	240	4.22	4.70
INACT90	0	2484	3153	63.53	61.76
	1	1426	1952	36.47	38.24
PARIS90	0	3376	4315	86.34	84.52
	1	534	790	13.66	15.48
SEUL90	0	3064	3933	78.36	77.04
	1	846	1172	21.64	22.96

COMPARAISON ENTRE LES MARGES FINALES DANS L'ÉCHANTILLON (AVEC LA PONDÉRATION FINALE)
ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
ETRPR90	0	4865	4865	95.30	95.30
	1	240	240	4.70	4.70
INACT90	0	3153	3153	61.76	61.76
	1	1952	1952	38.24	38.24
PARIS90	0	4315	4315	84.52	84.52
	1	790	790	15.48	15.48
SEUL90	0	3933	3933	77.04	77.04
	1	1172	1172	22.96	22.96

MÉTHODE : LINÉAIRE
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: _F_ (RAPPORT DE POIDS)

Basic Statistical Measures

Location		Variability	
Mean	1.305627	Std Deviation	0.13123
Median	1.315799	Variance	0.01722
Mode	1.182827	Range	0.70301
		Interquartile Range	0.19532

Quantiles (Definition 5)

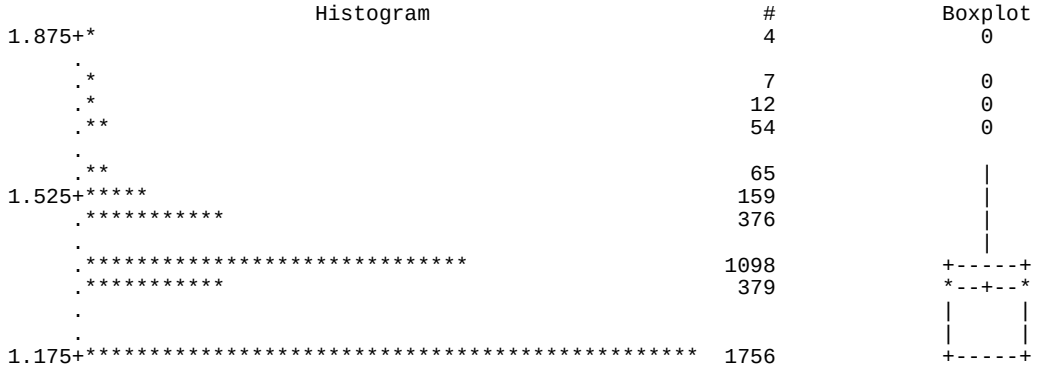
Quantile	Estimate
100% Max	1.88584
99%	1.68054
95%	1.54756
90%	1.48522
75% Q3	1.37815
50% Median	1.31580
25% Q1	1.18283
10%	1.18283
5%	1.18283
1%	1.18283
0% Min	1.18283

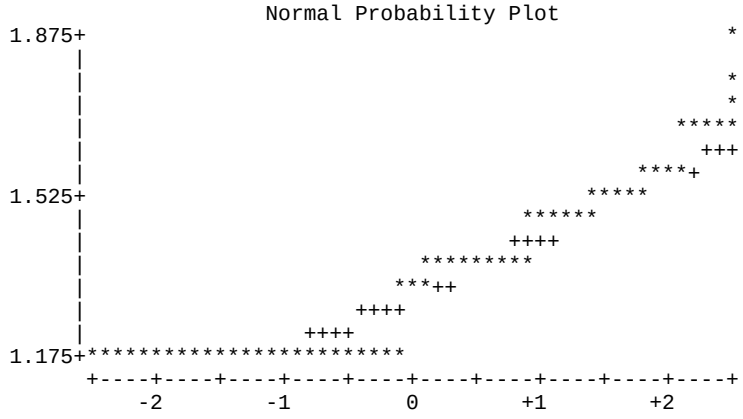
Extreme Observations

Lowest			Highest		
Value	IDENT	Obs	Value	IDENT	Obs
1.18283	9369034500	3910	1.75287	1169054670	691
1.18283	9369034490	3909	1.88584	1163045490	1
1.18283	9369034160	3908	1.88584	1163037600	72
1.18283	9369033930	3907	1.88584	2363000740	83
1.18283	9369033480	3906	1.88584	2163007440	532

MÉTHODE : LINÉAIRE
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: _F_ (RAPPORT DE POIDS)





MÉTHODE : LINÉAIRE
RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
POUR CHAQUE VALEUR DES VARIABLES

VARIABLE	MODALITE	NOMBRE D'OBSERVATIONS DE NIVEAU 1	RAPPORT DE POIDS
ETRPR90	0	3745	1.29907
ETRPR90	1	165	1.45455
INACT90	0	2484	1.26932
INACT90	1	1426	1.36886
PARIS90	0	3376	1.27814
PARIS90	1	534	1.47940
SEUL90	0	3064	1.28362
SEUL90	1	846	1.38534
ENSEMBLE		3910	1.30563

```
                                MÉTHODE : LINÉAIRE
            CONTENU DE LA TABLE poids CONTENANT LA NOUVELLE PONDÉRATION wcal

                                The CONTENTS Procedure

   Data Set Name: WORK.POIDS                              Observations:         3910
   Member Type:   DATA                                    Variables:            2
   Engine:        V8                                      Indexes:              0
   Created:       13:58 Wednesday, October 22, 2003       Observation Length:   24
   Last Modified: 13:58 Wednesday, October 22, 2003       Deleted Observations: 0
   Protection:                                            Compressed:           NO
   Data Set Type:                                         Sorted:               NO
   Label:


                     -----Engine/Host Dependent Information-----

              Data Set Page Size:         4096
              Number of Data Set Pages:   24
              First Data Page:            1
              Max Obs per Page:           168
              Obs in First Data Page:     118
              Number of Data Set Repairs: 0
              File Name:                  d:\saswork\_TD221\poids.sas7bdat
              Release Created:            8.0101M0
              Host Created:               WIN_NT


                  -----Alphabetic List of Variables and Attributes-----

                           #    Variable    Type    Len    Pos
                           ffffffffffffffffffffffffffffffff
                           1    IDENT       Char     10      8
                           2    wcal        Num       8      0




                  **********************
                  ***     BILAN      ***
                  **********************
*
*   DATE : 22 OCTOBRE 2003            HEURE : 13:44
*
*   *********************************
*   TABLE EN ENTRÉE : ECHANT
*   *********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  :            3910
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                :            0
*   NOMBRE D'OBSERVATIONS CONSERVÉES               :            3910
*
*   VARIABLE DE PONDÉRATION : DK
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 4
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
      ETRPR90 (2) INACT90 (2) PARIS90 (2) SEUL90 (2)
*
*   SOMME DES POIDS INITIAUX                                    : 3910
*   TAILLE DE LA POPULATION                                     : 5105
*
* VARIABLES DE NON-REPONSE
*    NOMBRE DE VARIABLES CATEGORIELLES : 4
*    LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
      etrpr96 (2)  inact96 (2)  paris96 (2)  seul96 (2)
*
*   MÉTHODE UTILISÉE : LINÉAIRE
*   LE CALAGE A ÉTÉ RÉALISÉ EN 2 ITÉRATIONS
*   LES POIDS ONT ÉTÉ STOCKÉS DANS LA VARIABLE WCAL DE LA TABLE POIDS
```

XV. Les contrôles et les messages d'erreur

Sauf mention contraire, les résultats des contrôles et les messages indiqués ci-dessous s'affichent dans la fenêtre OUTPUT.

XV.1 Les contrôles

XV.1.1 Contrôles sur les paramètres de la macro

- l'un au moins des paramètres **DATAMEN, DATAIND, DATAKISH** est renseigné ;
- si DATAKISH est renseigné, **DATAMEN** doit l'être aussi ;
- si DATAMEN est renseigné, le paramètre **MARMEN** doit l'être aussi ;
- si DATAIND est renseigné, le paramètre **MARIND** doit l'être aussi ;
- si DATAKISH est renseigné, les paramètres **MARKISH** et **POIDKISH** doivent l'être aussi ;
- si DATAMEN est renseigné ou si EGALPOI = OUI, **IDENT** doit être renseigné ;
- si DATAIND est renseigné, **IDENT2** doit l'être aussi ;
- si DATAKISH est renseigné, **IDENT2** doit l'être aussi ;
- si **IDENT** et **IDENT2** sont renseignés, ils doivent être différents ;
- le paramètre **ECHELLE** est numérique ;
- si NONREP=OUI, le paramètre **ECHELLE** doit être égal à 1 ou non renseigné ;
- si EGALPOI=OUI, les paramètres **DATAIND, MARIND, POIDS, IDENT, IDENT2** et **POPMEN** doivent être renseignés.

Lorsque le paramètre CONT vaut OUI, les contrôles suivants sont effectués.

- la table **&DATAMEN**[16] (respectivement **&DATAIND, &DATAKISH**) existe ;
- la table **&MARMEN** (respectivement **&MARIND, &MARKISH**) existe ;
- la variable **&IDENT** existe dans la table &DATAMEN. En cas de calage simultané, ou si EGALPOI=OUI, elle existe dans les tables &DATAIND et &DATAKISH ;
- la variable **&IDENT2** existe dans la table &DATAIND ;
- la variable **&IDENT2** existe dans la table &DATAKISH ;
- le paramètre **POPMEN** (respectivement **POPIND, POPKISH**) est renseigné lorsque PCT vaut OUI ;
- le paramètre **POIDS** est renseigné lorsque aucune variable catégorielle ne figure dans les variables du calage spécifiées dans la table &MARMEN ou si EGALPOI=OUI ;

[16] i.e. la table spécifiée dans le paramètre DATAMEN

- si **EGALPOI** = OUI, le nombre de valeurs différentes de l'identifiant **IDENT** doit être inférieur au nombre d'observations de la table DATAIND, autrement dit, IDENT doit être un identifiant de l'unité primaire et non de l'unité secondaire ;
- si **EGALPOI** = OUI, les poids initiaux doivent être égaux entre observations appartenant à la même entité de niveau 1 ;
- la variable **&POIDS** existe dans la table &DATAMEN, et elle est numérique ; si le paramètre DATAIND est seul renseigné, la variable &POIDS existe dans la table &DATAIND ;
- la variable **&POIDKISH** existe dans la table &DATAKISH et elle est numérique ;
- si DATAKISH est renseigné, la table &DATAKISH ne doit pas contenir moins d'observations que la table &DATAMEN ;
- la variable **&PONDQK** existe dans la table &DATAMEN, et elle est numérique ;
- le paramètre **M** vaut 1, 2, 3, 4 ou 5;
- les paramètres **LO** et **UP** sont renseignés lorsque M vaut 3 ou 4 ;
- le paramètre **ALPHA** doit être numérique si M=5 ;
- si **ECHELLE**=0 et si toutes les variables de calage de niveau 1 sont numériques, le paramètre POPMEN doit être renseigné ;
- le paramètre **EDITION** prend l'une des valeurs 0,1,2 ou 3.

XV.1.2 Contrôles sur le contenu des tables de marges

- la variable **VAR** existe dans la table &MARMEN (respectivement &MARIND, &MARKISH) ;
- la variable **N** existe dans la table &MARMEN (respectivement &MARIND, &MARKISH), et elle est numérique ;
- les variables **MAR1, MAR2**... de la table &MARMEN (respectivement &MARIND, &MARKISH) sont numériques ;
- les variables de calage nommées dans la variable VAR de la table &MARMEN (respectivement &MARIND, &MARKISH) existent dans la table &DATAMEN (respectivement &DATAIND, &DATAKISH) ;
- les variables de calage "numériques" (i.e. pour lesquelles N=0) nommées dans la variable VAR de la table &MARMEN (respectivement &MARIND, &MARKISH) sont des variables numériques (au sens de SAS) de la table &DATAMEN (respectivement &DATAIND, &DATAKISH) ;
- pour une variable catégorielle à p modalités (i.e. pour laquelle N=p), les marges MAR1 à MARp sont renseignées ;
- pour une variable numérique, la marge MAR1 est renseignée ;
- les totaux des marges des variables catégorielles sont tous égaux (à 100 si le paramètre PCT vaut OUI) (voir exemple XV.3.1).

XV.1.3 Contrôles sur les modalités des variables catégorielles

Pour une variable catégorielle à p modalités (i.e. pour laquelle N=p dans la table de marges), on vérifie que dans la table de données :

- aucune des modalités 1, 2 ... p (ou 01, 02 ... p si p >9) n'a un effectif nul[17] (voir exemple XV.3.2) ;
- la somme des effectifs (pondérés) des modalités 1, 2 ... p est égale à l'effectif (pondéré) de l'échantillon : cet effectif est la somme des pondérations initiales des observations non éliminées de la table &DATA ;

  Lorsque ce contrôle fait apparaître des erreurs, la macro imprime la liste de toutes les modalités (avec leurs effectifs pondérés) de la (ou des) variable(s) en erreur.
- il doit y avoir égalité entre le nombre p de modalités indiqué dans la table de marges et le nombre de modalités différentes effectivement présentes dans la table de données pour cette variable.

XV.1.4 Contrôles en cas de calage généralisé pour non-réponse

Si NONREP=OUI, les contrôles suivants sont réalisés :

- La variable R doit être présente dans l'une au moins des tables de marges.
- Si R existe dans l'une des table de marges, on doit avoir au moins une variable vérifiant R=1.
- Le nombre de variables catégorielles telles que R=0 (variables de calage X) doit être égal au nombre de variables catégorielles telles que R=1 (variables de non-réponse Z) dans chaque table de marges.
- Le total du nombre de variables numériques et du nombre cumulé de modalités des variables catégorielles telles que R=0 doit être égal au total du nombre de variables numériques et du nombre cumulé de modalités des variables catégorielles telles que R=1 dans l'ensemble des tables de marges concaténées.
- Les variables indiquées dans chaque table de marges (variables de calage et variables instrumentales) doivent être présentes dans la table de données correspondante.

Si NONREP=NON, le contrôle suivant est réalisé :

- Si la variable R est présente dans l'une au moins des tables de marges, ce doit être non renseignée ou avec la valeur 0.

La présence dans une table de marges d'une variable ayant la valeur 1 pour R entraîne l'arrêt du programme et un message d'erreur.

[17] Ce contrôle est effectué même si CONT vaut NON.

XV.1.5 Contrôles sur la table contenant les pondérations finales

Ces contrôles sont réalisés même si CONT vaut NON.

- le paramètre **POIDSFIN** est renseigné lorsque l'un des paramètres DATAPOI, DATAPOI2, DATAPOI3 l'est ;
- l'un des paramètres **DATAPOI, DATAPOI2, DATAPOI3** doit être renseigné si POIDSFIN l'est. En cas de calage simultané, lorsque POIDSFIN est renseigné, une table de poids correspondant à chaque niveau d'observation doit être spécifiée ;
- si DATAPOI3 est renseigné, **POIDSKISHFIN** doit l'être aussi ;
- si la table &DATAPOI (respectivement &DATAPOI2, &DATAPOI3) est une table permanente, de la forme XYZ.ABC, une base SAS est allouée **en écriture** au DDNAME XYZ ;

Si une base SAS est allouée, mais seulement en lecture, le message d'erreur est édité **dans la fenêtre Log**, et non dans la fenêtre Output.

Toute anomalie provoque l'arrêt du programme et l'impression d'un message d'erreur.

XV.2 Les messages d'erreur

Outre les messages d'erreur générés par la macro en cas de contrôles négatifs, des messages apparaissent dans les cas suivants.

XV.2.1 Pas d'observation pour réaliser le calage

Cette erreur peut se produire dans les cas suivants :

- la table &DATAMEN (respectivement __&DATAIND ou __&DATAKISH) n'a pas d'observation; ceci peut en particulier arriver lorsque l'on construit préalablement ses tables de données à partir d'un fichier en spécifiant une clause WHERE conduisant à ne sélectionner aucune observation...
- la table &DATAMEN (respectivement __&DATAIND ou __&DATAKISH) n'est pas vide, mais toutes ses observations ont été éliminées car elles ont des valeurs manquantes sur les variables du calage ou sur les variables de pondération (voir exemple XV.3.3).

XV.2.2 Messages relatifs au déroulement de l'algorithme

Dans un certain nombre de cas, l'algorithme ne peut arriver à son terme. Voici les principales raisons d'arrêt de l'algorithme.

XV.2.2.1 LES VARIABLES DU CALAGE SONT COLINEAIRES

Dans le cas où COLIN=NON, lorsque les variables du calage sont colinéaires, le calage est impossible, car le système d'équations (E) du chapitre II est indéterminé. Le programme élimine automatiquement les colinéarités structurelles qui apparaissent lorsque plusieurs variables catégorielles figurent dans les variables de calage (voir § III.1 et § VIII). Les autres colinéarités empêchent le fonctionnement de l'algorithme : elles provoquent en effet la non-inversibilité d'une certaine matrice, ce qui génère un message d'erreur SAS. La macro édite alors les coefficients de la (ou des) combinaison(s) linéaire(s) nulle(s) des variables du

calage[18] : ceci peut permettre à l'utilisateur d'identifier plus facilement l'origine de ces colinéarités (voir exemple XV.3.4).

XV.2.2.2 LE CALAGE NE PEUT ETRE REALISE

Lorsque l'on utilise une méthode autre que la méthode linéaire (M=1), il peut arriver que le système d'équations (E) n'ait pas de solution, parce que les bornes LO et UP[19] (si M=3 ou 4), ou ALPHA (si M=5) imposées aux rapports de poids sont trop "contraignantes". Ceci se traduit lors de l'algorithme par un message d'erreur SAS (édité sur la Log) indiquant un dépassement de capacité, une non-inversibilité de matrice, etc. La macro édite dans ce cas le message suivant dans la fenêtre Output : "Le calage ne peut être réalisé" (voir exemple XV.3.5).

Pour réaliser le calage, l'utilisateur peut alors opérer de plusieurs façons :

- opérer des regroupements de modalités de variables catégorielles rendant les marges du calage plus faciles à atteindre ;
- "relâcher" les contraintes sur les rapports de poids, en diminuant la valeur de LO ou en augmentant la valeur de UP (si M=3 ou 4), ou en diminuant la valeur de ALPHA (si M=5) ;
- choisir un facteur d'échelle supérieur à 1 si l'effectif de l'échantillon, pondéré par la variable de pondération initiale, n'est pas égal à l'effectif de la population (lorsqu'une variable catégorielle figure parmi les variables de calage) ;
- ... ou bien changer les variables du calage.

XV.2.2.3 LE NOMBRE MAXIMUM D'ITERATIONS EST ATTEINT

Le paramètre MAXITER permet à l'utilisateur de fixer un nombre maximum d'itérations pour l'algorithme de Newton, ceci pour éviter que le programme tourne pendant une durée jugée trop longue (ou trop coûteuse...). Au bout de &MAXITER itérations, l'algorithme s'arrête et un message est édité.

XV.2.2.4 CONVERGENCE IMPARFAITE

Il peut arriver que l'algorithme converge (i.e. le critère d'arrêt est satisfait) sans que le calage soit parfaitement réalisé : dans ce cas, la macro édite un message, et les divergences entre les marges de l'échantillon et les marges du calage sont signalées dans le tableau qui permet la comparaison de ces marges (voir exemple XV.3.6).

Ce phénomène peut se produire lorsque les contraintes imposées aux rapports de poids sont "à la limite de ce qu'ils peuvent supporter".

[18] Les variables catégorielles sont "éclatées" en variables indicatrices des modalités.

[19] Lorsque l'on utilise la méthode du raking ratio (M=2), il y a une borne implicite LO=0.

XV.3 Exemples

Pour chacun des exemples suivants, sont donnés le programme et le listing (ou un extrait) produit par la macro.

XV.3.1 Les totaux des marges catégorielles ne sont pas tous égaux

```
DATA don;
        INPUT nom $  X $ Y $ Z T;
        POND=id;
        CARDS;
A 1 1 1 1
B 1 2 2 3
C 1 2 3 1
D 2 1 1 1
E 2 1 3 2
F 2 2 2 3
;
DATA marges;
        INPUT VAR $ N MAR1 MAR2 MAR3;
        CARDS;
X 2  30  60   .
Y 2  60  20   .
Z 0 140   .   .
T 3  10  50  30
;
TITLE "Contrôle sur les totaux des marges catégorielles";
%CALMAR2(DATAMEN=don,POIDS=pond,IDENT=nom,MARMEN=marges)
```

Contrôle sur les totaux des marges catégorielles

ERREUR : les totaux des marges des variables catégorielles ne sont pas tous égaux

VAR	N	MAR1	MAR2	MAR3	TOT_MARG
X	2	30	60	.	90
Y	2	60	20	.	80
T	3	10	50	30	90

XV.3.2 Une variable catégorielle n'a pas le même nombre de modalités dans les tables échantillon et marges

```
DATA don;
        INPUT nom $  X $ Y $ Z T;
        CARDS;
A 1 1 1 1
B 1 2 2 3
C 1 2 3 1
D 2 1 1 3
E 2 1 9 3
;
DATA marges;
        INPUT VAR $ N MAR1 MAR2 MAR3;
        CARDS;
X 2   40  60   .
Y 2   50  50   .
Z 0 2400   .   .
T 3   20  50  30
;
TITLE "Contrôle sur les modalités des variables catégorielles :"
      " pas d'effectif nul";
%CALMAR2(DATAMEN=don,IDENT=nom,MARMEN=marges,M=1,PCT=oui,POPMEN=1000)
```

```
   Contrôle sur les modalités des variables catégorielles : pas d'effectif nul


** ERREUR : LA VARIABLE T A 2 MODALITÉS DANS LA TABLE DON      **
** MAIS EST DÉCLARÉE AVEC 3 MODALITÉS DANS LA TABLE DES MARGES **

   LISTE DES MODALITÉS DE LA VARIABLE T DANS LA TABLE DE DONNÉES

                          EFFECTIF          %
                        ECHANTILLON    ECHANTILLON
                OBS    T  NON PONDÉRÉ    NON PONDÉRÉ

                 1     1        2              40
                 2     3        3              60


                    TABLE DES MARGES : MARGES

               OBS    VAR    N    MAR1    MAR2    MAR3

                1      X     2     40      60       .
                2      Y     2     50      50       .
                4      T     3     20      50      30
```

XV.3.3 Pas d'observation valide dans la table en entrée

```
DATA don;
        INPUT nom $  X $ Y $ Z T pond;
        CARDS;
A . 1 1 1 10
B 1 2 5 3  0
C 1 2 3 . 10
D 2 . 4 3 10
E 2 1 9 2  0
;
DATA marges;
        INPUT VAR $ N MAR1 MAR2 MAR3;
        CARDS;
X 2  40  60   .
Y 2  60  40   .
Z 0 140   .   .
T 3  20  50  30
;
TITLE "Aucune observation valide";
%CALMAR2(DATAMEN=don,POIDS=pond,IDENT=nom,MARMEN=marges,M=1,PCT=oui,POPMEN=1000)
```

```
                                   Aucune observation valide

**************************************************************************
***   ERREUR : LA TABLE  DON                                           ***
***            SPÉCIFIÉE DANS LE PARAMÈTRE DATAMEN A 5 OBSERVATIONS... ***
***            MAIS ELLES SONT TOUTES ÉLIMINÉES !                       ***
***                                                                     ***
***   UNE OBSERVATION DE LA TABLE EN ENTRÉE EST ÉLIMINÉE DÈS QUE :      ***
***   - ELLE A UNE VALEUR MANQUANTE SUR L'UNE DES VARIABLES DU CALAGE   ***
***   - ELLE A UNE VALEUR MANQUANTE, NÉGATIVE OU NULLE SUR L'UNE        ***
***     DES VARIABLES DE PONDÉRATION.                                   ***
**************************************************************************
```

XV.3.4 Colinéarité entre les variables du calage

```
DATA don;
        INPUT nom $  X $ Y $ Z T;
        U=Z+3*T;
        CARDS;
A 2 2 1 1
B 2 1 5 3
C 1 2 3 3
D 1 1 4 3
E 1 2 9 2
;
DATA marges;
        INPUT VAR $ N MAR1 MAR2;
        CARDS;
X 2    40   60
U 0 14000    .
Y 2    60   40
Z 0  5000    .
T 0  3000    .
;
TITLE "Les variables du calage sont colinéaires";
%CALMAR2(DATAMEN=don,IDENT=nom,MARMEN=marges,M=1,PCT=oui,POPMEN=1000)
```

```
 IML Ready
 ERROR: (execution) Matrix should be non-singular.
+ERROR: (execution) Matrix should be non-singular.
+ERROR: (execution) Matrix should be non-singular.

  operation : INV      at line  3422 column 136
  operands  : PHIPRIM

 PHIPRIM       6 rows      6 cols    (numeric)

        600         0       200      8000      3200      1600
          0       400       200      3600      1200       800
        200       200       400      5400      1800      1200
       8000      3600      5400    150000     59400     30200
       3200      1200      1800     59400     26400     11000
       1600       800      1200     30200     11000      6400

  statement : ASSIGN          at line  3422 column 125
 Exiting IML.
```

Les variables du calage sont colinéaires

Comparaison entre les marges tirées de l'échantillon (avec la pondération initiale) et les marges dans la population (marges du calage)

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
X	1	600	400	60.00	40.00
	2	400	600	40.00	60.00
Y	1	400	600	40.00	60.00
	2	600	400	60.00	40.00
U		11600	14000	.	.
Z		4400	5000	.	.
T		2400	3000	.	.

```
Méthode : linéaire

          ****************************************************
          ***   Les variables analysées sont colinéaires :   ***
          ***   le calage ne peut être réalisé               ***
          ***                                                ***
          ***   Pour rendre le calage possible, vous pouvez  ***
          ***   utiliser l'option : COLIN=OUI                ***
          ****************************************************
```

```
                    Les variables du calage sont colinéaires

          Coefficients de la (ou des) combinaison(s) linéaire(s) nulle des variables du calage
    (une variable de nom WXY 2 désigne la variables indicatrice associée à la modalité 2 de la variable catégorielle WXY)

                                  X 1    X 2    Y 1     U    Z    T

                                   0      0      0     -1    1    3
```

XV.3.5 Calage impossible

```
DATA don;
        INPUT nom $ X Y Z;
        POND=10;
        CARDS;
A 1 1 1
B 2 2 2
C 1 2 3
D 2 1 1
E 1 2 3
F 2 1 2
G 1 3 3
H 2 3 2
;
DATA marges;
        INPUT VAR $ N MAR1-MAR3;
        CARDS;
X 2  30  50  .
Y 3  10  50 20
Z 0 250   .  .
;
TITLE "Calage impossible";
%CALMAR2(DATAMEN=don,IDENT=nom,MARMEN=marges,M=2,POIDS=pond)



 IML Ready
 Exiting IML.
        ************************************************************
        ***   Valeur du critère d'arrêt à l'itération 1 : 20.3809   ***
        ************************************************************
 IML Ready
 Exiting IML.
        ************************************************************
        ***   Valeur du critère d'arrêt à l'itération 2 : 10.7807   ***
        ************************************************************
 IML Ready
 Exiting IML.
        ************************************************************
        ***   Valeur du critère d'arrêt à l'itération 3 : 3.05255   ***
        ************************************************************
 IML Ready
 ERROR: (execution) Matrix should be non-singular.
+ERROR: (execution) Matrix should be non-singular.
+ERROR: (execution) Matrix should be non-singular.

  operation : INV      at line  1820 column 136
  operands  : PHIPRIM

 PHIPRIM       5 rows      5 cols    (numeric)

  32.694736         0         0 25.704334 98.084207
         0 119.99402 75.476511 28.834293 239.98804
         0 75.476511 75.476511         0 150.95302
  25.704334 28.834293         0 54.538626 134.78159
  98.084207 239.98804 150.95302 134.78159  774.2287

  statement : ASSIGN          at line  1820 column 125
 Exiting IML.
```

Calage impossible

Comparaison entre les marges tirées de l'échantillon (avec la pondération initiale)
et les marges dans la population (marges du calage)

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
X	1	40	30	50.00	37.50
	2	40	50	50.00	62.50
Y	1	30	10	37.50	12.50
	2	30	50	37.50	62.50
	3	20	20	25.00	25.00
Z		170	250	.	.

Calage impossible

Méthode : raking ratio

```
**********************************************************************
***   Le calage ne peut être réalisé. Pour rendre le calage        ***
***   possible, vous pouvez :                                      ***
***                                                                ***
***   - utiliser la méthode linéaire (M=1)                         ***
***   - opérer des regroupements de modalités de variables         ***
***     catégorielles                                              ***
**********************************************************************
```

Calage impossible

Méthode : raking ratio
Premier tableau récapitulatif de l'algorithme :
la valeur du critère d'arrêt et le nombre de poids négatifs après chaque itération

Itération	Critère d'arrêt	Poids négatifs
1	20.3809	0
2	10.7807	0
3	3.0526	0

Calage impossible

Méthode : raking ratio
Deuxième tableau récapitulatif de l'algorithme :
les coefficients du vecteur lambda de multiplicateurs de Lagrange après chaque itération

Variable	Modalité	LAMBDA1	LAMBDA2	LAMBDA3	LAMBDA4
X	1	-11.8750	-57.3896	-350235652.77	.
X	2	-9.0625	-37.4102	-233490434.49	.
Y	1	3.7500	1.4819	1.57	.
Y	2	0.4375	0.5533	0.61	.
Y	3	.	.	.	.
Z		4.1875	19.1446	116745217.47	.

XV.3.6 Convergence imparfaite

```
DATA don;
        INPUT X Y Z;
        pond=10;
        CARDS;
1 1 1
1 2 2
1 2 3
2 1 1
2 1 3
2 2 2
;
DATA marges;
        INPUT VAR $ N MAR1 MAR2;
        LIST;
        CARDS;
X 2 10  50
Y 2 10  50
Z 0 110  .
;
TITLE "Convergence imparfaite";
%CALMAR2(DATAMEN=don,MARMEN=marges,M=2,SEUIL=0.0001,MAXITER=50,POIDS=pond)
```

Convergence imparfaite

Comparaison entre les marges tirées de l'échantillon (avec la pondération initiale)
et les marges dans la population (marges du calage)

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
X	1	30	10	50.00	16.67
	2	30	50	50.00	83.33
Y	1	30	10	50.00	16.67
	2	30	50	50.00	83.33
Z		120	110	.	.

Convergence imparfaite

Méthode : raking ratio
Premier tableau récapitulatif de l'algorithme :
la valeur du critère d'arrêt et le nombre de poids négatifs après chaque itération

Itération	Critère d'arrêt	Poids négatifs
1	10.7228	0
2	5.5802	0
3	1.6965	0
4	0.2730	0
5	0.0336	0
6	0.0147	0
7	0.0058	0
8	0.0022	0
9	0.0008	0
10	0.0003	0
11	0.0001	0
12	0.0000	0

Convergence imparfaite

Méthode : raking ratio
Deuxième tableau récapitulatif de l'algorithme :
les coefficients du vecteur lambda de multiplicateurs de Lagrange après chaque itération

Variable	Modalité	LAMBDA1	LAMBDA2	LAMBDA3	LAMBDA4	LAMBDA5	LAMBDA6	LAMBDA7	LAMBDA8	LAMBDA9	LAMBDA10	LAMBDA11	LAMBDA12
X	1	2.07692	2.31175	3.41178	5.06007	6.88902	8.8011	10.7616	12.7456	14.7395	16.7372	18.7363	20.7360
X	2	4.30769	4.31801	5.17090	6.70472	8.50611	10.4119	12.3712	14.3550	16.3489	18.3466	20.3458	22.3455
Y	1	-2.69231	-2.91598	-3.42185	-4.22577	-5.14389	-6.1009	-7.0813	-8.0734	-9.0703	-10.0692	-11.0688	-12.0686
Y	2	.	.	.	.	.	.	.	.	.	.	.	.
Z		-0.92308	-1.25138	-1.83942	-2.63801	-3.53955	-4.4924	-5.4720	-6.4640	-7.4609	-8.4598	-9.4593	-10.4592

Convergence imparfaite

Méthode : raking ratio

```
************************************************************
***   ATTENTION : l'algorithme a convergé, mais le calage   ***
***               n'est pas parfaitement réalisé            ***
************************************************************
```

Convergence imparfaite

Méthode : raking ratio
Comparaison entre les marges finales dans l'échantillon (avec la pondération finale)
et les marges dans la population (marges du calage)

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population	Erreur
X	1	10.000	10	16.67	16.67	*
	2	50.000	50	83.33	83.33	
Y	1	10.000	10	16.67	16.67	
	2	50.000	50	83.33	83.33	*
Z		110.001	110	.	.	*

Convergence imparfaite

MÉTHODE : RAKING RATIO
STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
ET SUR LES PONDÉRATIONS FINALES

The UNIVARIATE Procedure
Variable: _F_ (RAPPORT DE POIDS)

Basic Statistical Measures

Location		Variability	
Mean	1.000004	Std Deviation	1.59861
Median	0.500000	Variance	2.55555
Mode	.	Range	4.16667
		Interquartile Range	0.83331

Quantiles (Definition 5)

Quantile	Estimate
100% Max	4.16667E+00
99%	4.16667E+00
95%	4.16667E+00
90%	4.16667E+00
75% Q3	8.33333E-01
50% Median	5.00000E-01
25% Q1	2.39033E-05
10%	6.85639E-10
5%	6.85639E-10
1%	6.85639E-10
0% Min	6.85639E-10

Extreme Observations

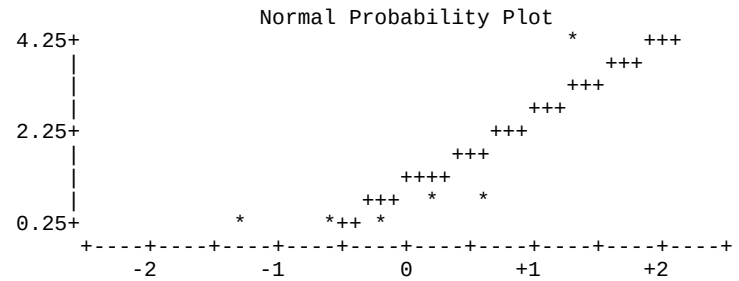
Lowest			Highest		
Value	id	Obs	Value	id	Obs
6.85639E-10	e	5	2.39033E-05	c	3
2.39033E-05	c	3	1.66667E-01	a	1
1.66667E-01	a	1	8.33333E-01	b	2
8.33333E-01	b	2	8.33333E-01	d	4
8.33333E-01	d	4	4.16667E+00	f	6

```
                                Convergence imparfaite

                                MÉTHODE : RAKING RATIO
 STATISTIQUES SUR LES RAPPORTS DE POIDS (= PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
                             ET SUR LES PONDÉRATIONS FINALES

                                The UNIVARIATE Procedure
                             Variable:  _F_  (RAPPORT DE POIDS)

                      Stem Leaf                     #             Boxplot
                         4 2                        1                *
                         3
                         3
                         2
                         2
                         1
                         1
                         0 88                       2             +-----+
                         0 002                      3             +--+--+
                           ----+----+----+----+
```




```
                                MÉTHODE : RAKING RATIO
              RAPPORTS DE POIDS MOYENS (PONDÉRATIONS FINALES / PONDÉRATIONS INITIALES)
                                POUR CHAQUE VALEUR DES VARIABLES

                                                 NOMBRE
                                              D'OBSERVATIONS     RAPPORT
                      VARIABLE    MODALITE      DE NIVEAU 1     DE POIDS

                      X              1               3           0.33334
                      X              2               3           1.66667
                      Y              1               3           0.33333
                      Y              2               3           1.66667
                      ENSEMBLE                       6           1.00000


                    **********
                     ***     BILAN     ***
                     **********
*
*   DATE : 22 OCTOBRE 2003            HEURE : 13:44
*
*   *********************************
*   TABLE EN ENTRÉE : DON
*   *********************************
*
*   NOMBRE D'OBSERVATIONS DANS LA TABLE EN ENTRÉE  :            6
*   NOMBRE D'OBSERVATIONS ÉLIMINÉES                :            0
*   NOMBRE D'OBSERVATIONS CONSERVÉES               :            6
*
*   VARIABLE DE PONDÉRATION : POND
*
*   NOMBRE DE VARIABLES CATÉGORIELLES : 2
*   LISTE DES VARIABLES CATÉGORIELLES ET DE LEURS NOMBRES DE MODALITÉS :
      X (2) Y (2)
*
*   SOMME DES POIDS INITIAUX                                    : 60
*   TAILLE DE LA POPULATION                                     : 60
*
*   NOMBRE DE VARIABLES NUMÉRIQUES : 1
*   LISTE DES VARIABLES NUMÉRIQUES :
      Z
*
*
*   MÉTHODE UTILISÉE : RAKING RATIO
```

```
*   LE CALAGE N'A PU ETRE RÉALISÉ QU'APPROXIMATIVEMENT EN 12 ITÉRATIONS
```

Bibliographie

Sur la théorie du calage :

[1] **W.E. Deming, F.F. Stephan** (1940). On a least squares adjustment of a samples frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics, 11, 427-444*.

[2] **J.-C. Deville et C.-E. Sarndal** (1992). Calibration estimators in survey sampling.
Journal of the American Statistical Association, vol 87, n°418, 376-382.

[3] **J.-C. Deville, C.-E. Sarndal & Sautory O.** (1993). Generalized raking procedures in survey sampling.
Journal of the American Statistical Association, vol 88, n°423, 1013-1020.

[4] **O. Sautory** (1991). Redressement d'échantillons d'enquêtes auprès des ménages par calage sur marges.
Document de travail de la Direction des Statistiques Démographiques et sociales n°F9103, INSEE.

[5] **G. Roy, A. Vanheuverzwyn** (2001). Redressement par la macro CALMAR : applications et pistes d'amélioration. *Traitements des fichiers d'enquête, pp. 31–46, Presses Universitaires de Grenoble*.

Sur le calage simultané :

[6] **V.M. Estevao, C.E. Särndal** (2003). A new perspective on calibration estimators.

[7] **N. Caron, O. Sautory** (1998). Calages simultanés pour différentes unités d'une même enquête. *INSEE, note interne.*

[8] **E. Crenner** (1998). Une expérience de calage sur marges dans une enquête conditions de vie : Calmar simultané ménages-individus-kish sur l'enquête PCV. *Document de travail de la Direction des Statistiques Démographiques et sociales n°F9804, INSEE.*

Sur le calage généralisé :

[9] **J.C. Deville** (2002). La correction de la non-réponse par calage généralisé. *Journées de méthodologie statistique, INSEE.*

[10] **J.C. Deville** (1998). La correction de la non-réponse par calage généralisé ou par échantillonnage équilibré. *Actes de la société de Statistique du Canada, Université de Sherbrooke*.

[11] **J.C. Deville, J. Le Guennec, O. Sautory** (2002). Application du calage généralisé à la correction de la non-réponse : une expérimentation. *Journées de méthodologie statistique, INSEE.*

[12] **J. Bardaji, J. Le Guennec** (2003). Non-réponse suscitée par le questionnaire : le redressement de l'enquête sur les contrats « emploi consolidé » par calage généralisé. *Document de travail à paraître, INSEE.*

Sur le redressement de la non-réponse :

[13] **J.C. Deville et F. Dupont** (1996). Non-réponse : principes et méthodes.
Actes des journées de méthodologie statistique de décembre 1993, INSEE-Méthodes n°56-57-58.

[14] **F. Dupont** (1996).Calage et redressement de la non-réponse totale.
Actes des journées de méthodologie statistique de décembre 1993, INSEE-Méthodes n°56-57-58.

[15] **N. Caron** (1996). Les principales techniques de correction de la non-réponse et les modèles associés.

Document de travail de l'Unité Méthodes Statistiques n°9604, INSEE.
Sur le calcul de précision :

[16] **N. Caron** (1998). Le logiciel de calcul de précision POULPE. *Actes des journées de méthodologie statistique de mars 1998, INSEE-Méthodes n°.84-85-86.*

[17] **C.-E. Sarndal, B. Swensson & J. Wretman** (1992) Model assisted survey sampling. *Springer-Verlag, New-York.*