INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

DIRECTION GÉNÉRALE

18, boulevard Adolphe Pinard - 75675 PARIS CEDEX 14

Unité Méthodes Statistiques

La macro CALMAR

Redressement d'un échantillon

par calage sur marges

Document n° F 9310

25 novembre 1993

Olivier SAUTORY

Série des documents de travail

de la Direction des Statistiques Démographiques et Sociales

Table des matières

Généralités

La macro SAS **CALMAR** (CALage sur MARges) 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.

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

Comment écrire les paramètres

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.

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 la macro. 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.

I. Aspects théoriques du calage sur marges

I.1 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 $\boldsymbol{X_1} \ldots \boldsymbol{X_j} \ldots \boldsymbol{X_J}$ 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** :

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

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

I.2 Résolution théorique

On choisit une "fonction de distance" G, d'argument $r = w_k / d_k$, 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 § 1.3.), 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 $X' = (X_1 \dots X_J)$) :

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

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 :

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

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 résultant des équations de calage :

$$\boxed{\sum_{\in} d \;\; F(x' \, \lambda) \, x \;\; = X} \quad \text{(E)}$$

On peut résoudre numériquement ce système par la méthode itérative de Newton ; on calcule une suite de vecteurs $\lambda^{(j)}$ définis 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" :

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

I.3 Les fonctions G usuelles

On indique pour chacune des 4 méthodes usuelles la fonction G(r) *(*où $r = w_k / d_k$) et la fonction F(u) (où $u = x_k' \lambda$).

a) méthode "linéaire"

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

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 utilisant de l'information auxiliaire, appelée **estimation par régression.**

b) méthode "raking ratio"

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

Lorsque les variables auxiliaires sont des variables catégorielles pour lesquelles on connaît les effectifs des modalités dans la population (cf § 1.4.1), 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 Proportional Fitting").

c) méthode "logit"

- $G(r) = \left[(r-L)\ \operatorname{Log}\frac{r-L}{1-L} + (U-r) \operatorname{Log}\frac{U-r}{U-1} \right] \frac{1}{A},$ si $L < r < U$ (et $+\infty$ 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"

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

I.4 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 § 1.1, 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 V , définie par :

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

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 V .

Il est facile de vérifier que le système d'équations (E) est dans ce cas surdéterminé : la somme des I

I.5 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.

Dans la pratique, la détermination de ces valeurs L_{max} et U_{min} 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) ... jusqu'à ce que le programme manifeste qu'il n'existe pas de solution.

Face à différents systèmes 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. Des 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.

On peut ainsi souhaiter utiliser une méthode bornée (M = 3 ou 4) sans trop déformer la distribution des rapports de poids obtenus par la méthode du raking ratio par exemple : l'utilisation des valeurs L_{max} et U_{min} conduit en général à une très forte concentration des rapports de poids au voisinage de ces valeurs limites.

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.

Références

[1] **Deming W.E. and Stephan F.F**. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known.
Annals of Mathematical Statistics, 11, 427-444.

[2] **Deville J. -C. and Särndal C. -E**. (1992). Calibration estimators in survey sampling.
Journal of the American Statistical Association, vol 87, n°418, 376-382.

[3] **Deville J. -C., Särndal C. -E. and Sautory O.** (1993). Generalized raking procedures in survey sampling.
Journal of the American Statistical Association, vol 88, n°423, 1013-1020.

II. Mise en œuvre de la macro CALMAR

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

II.1 Les données en entrée de la macro

II.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[1] n'intervenant pas directement dans le redressement.

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

II.1.1.1 Les variables de calage catégorielles

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

- Si elle est caractère, ses p modalités doivent obligatoirement prendre les valeurs :

  1, 2, ... , p si $1 \leq p \leq 9$,

  01, 02, ... , p si $10 \leq p \leq 99$,

  001, 002, ... , 010, ... , p si $100 \leq p \leq 999$.

- Si elle est numérique, ses p modalités doivent obligatoirement prendre les valeurs 1, 2, ..., p.

Ces contraintes nécessitent souvent une recodification préalable de ces variables.

Les noms des variables de calage catégorielles, ainsi que leurs nombres de modalités, sont spécifiés dans la table des marges &DATAMAR (voir § II.2.1).

[1] Elle peut contenir une variable de pondération supplémentaire, la variable &PONDQK (voir § II.2.1).

La macro réalise les contrôles suivants :

- aucune modalité d'une variable catégorielle n'a un effectif nul dans l'échantillon ;
- une variable catégorielle indiquée comme ayant p modalités ne prend aucune autre valeur que les p modalités permises (telles qu'elles ont été précisées plus haut).

II.1.1.2 Les variables de calage numériques

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 &DATAMAR (voir § II.2.1).

II.1.1.3 La variable de pondération initiale

C'est la variable donnant pour chaque observation k la valeur de la pondération initiale d_k. Cette valeur est en général 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.

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, la pondération initiale peut être définie à un coefficient multiplicatif près : on montre en effet que si les rapports de poids[2] et le paramètre λ (voir §1) dépendent du choix de la pondération initiale, en revanche les pondérations finales w_k n'en dépendent pas[3].

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.

Toutefois, il y a (au moins) deux bonnes raisons de spécifier 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 de poids de sondage égaux à N/n (s'il existe au moins une variable de calage catégorielle) ;
- 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

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

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

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.

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 EFFPOP).

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

II.1.1.4 Autres variables de la table des données

La table &DATA 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]).

II.1.1.5 Observations éliminées

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.

II.1.1.6 Calage en présence de non-réponse

Les procédures de redressement, telles qu'elles sont présentées au §1, ne sont en principe valides qu'en absence de non-réponse totale[4] 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[5].

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

[5] 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).

II.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) DATAMAR de la macro.

Cette table contient une observation pour chaque variable de calage. Les variables de la table s'appellent obligatoirement VAR, N, MAR1, MAR2, ..., MARh ; elles prennent les valeurs suivantes :

VAR nom de la variable[6].

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 &DATA ;
- une variable telle que N=0 est une variable numérique de la table &DATA ;
- 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).

[6] en minuscules ou en majuscules

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, et l'utilisateur doit indiquer dans le paramètre EFFPOP la taille de la population.

Le lecteur peut se reporter aux § III.1.1 et III.2.2 pour avoir des exemples de tables &DATAMAR correctes, et au § IV pour avoir des exemples des erreurs à ne pas commettre.

II.2 Syntaxe de la macro

II.2.1 Paramètres spécifiant les tables SAS en entrée

* DATA = nom de table SAS

nom de la table SAS contenant les données **(obligatoire).**

Cette table contient pour chaque observation 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 § II.1.1.

* DATAMAR = nom de table SAS

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

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

***Remarque* :** on peut utiliser la clause WHERE, les options FIRSTOBS, OBS, KEEP... pour définir ces deux tables. Par exemple, on peut écrire :

```
DATA=A(WHERE=(SEXE="2")), DATAMAR=B(OBS=5)
```

Utiliser ces options avec la table &DATAMAR permet de sélectionner ou de changer les variables du calage parmi un ensemble de variables potentielles.

* POIDS = variable

variable **numérique** contenant les pondérations initiales des observations de l'échantillon (elle appartient à la table &DATA).

Ce paramètre est obligatoire lorsqu'il n'y pas de variable de calage catégorielle (voir § II.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 &DATA) : 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.

* PCT= OUI ou NON

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

Par défaut : PCT = NON.

* EFFPOP = 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.

II.2.2 Paramètres spécifiant la méthode utilisée

* M = 1, 2, 3 ou 4

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.

* LO = valeur

borne inférieure des rapports de poids[7], 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.

* 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.

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

II.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 de la table &DATA **non éliminées** ; elle contient la variable &POIDSFIN (voir plus loin) et le cas échéant la variable &IDENT.

Si cette table existe, le paramètre suivant 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.

II.2.4 Paramètres spécifiant les sorties imprimées

* 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 la table &DATAMAR (existence des variables, présence de toutes les marges...), ainsi que sur les variables de calage de la table &DATA.

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

Par défaut : CONT = OUI.

* EDITPOI = OUI ou NON

si EDITPOI vaut OUI, la macro édite les valeurs des rapports de poids obtenus pour chaque combinaison de valeurs des variables de calage[8], catégorielles et numériques.

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".

Par défaut : STAT = OUI.

* CONTPOI = OUI ou NON

si CONTPOI vaut OUI, la macro édite le contenu de la table &DATAPOI[10].

Par défaut : CONTPOI = 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.

[8] Deux observations prenant les mêmes valeurs pour toutes les variables de calage ont en effet le même *rapport de poids*.

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

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

II.3 Les sorties imprimées

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 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 ;
- 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, i.e. la 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 EFFPOP, 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.

II.4 La table en sortie

La table en sortie spécifiée dans le paramètre DATAPOI peut être temporaire ou permanente. Ses observations sont les observations de la table &DATA 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 &DATA 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" ;

  - 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).

III. Exemples

III.1 Exemple 1 : un petit exemple commenté

III.1.1 Le programme

```
DATA DON;
INPUT NOM $ X $ Y $ Z POND;
CARDS;
A 1 1 1 10
B 1 2 2  0
C 1 2 3  .
D 2 1 1 11
E 2 1 3 13
F 2 2 2  7
G 2 2 2  8
H 1 2 2  8
I 2 1 2  9
J . 2 2 10
K 2 2 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 de calage sur marges";
%CALMAR(DATA=DON,POIDS=POND,IDENT=NOM,
        DATAMAR=MARGES,M=2,EDITPOI=OUI,OBSELI=OUI,
        DATAPOI=SORTIE,POIDSFIN=PONDFIN,LABELPOI=pondération raking ratio)

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

DATA 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

DATAMAR 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 la méthode du raking ratio

EDITPOI on demande l'édition des rapports de poids pour toutes les combinaisons de valeurs des variables de calage

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 le label "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

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.

III.1.2 La log

```
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: The DATA statement used 0.03 CPU seconds and 1336K.

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

NOTE: The data set WORK.MARGES has 3 observations and 4 variables.
NOTE: The DATA statement used 0.02 CPU seconds and 1336K.
26         TITLE "Un petit exemple de calage sur marges";
27         %CALMAR(DATA=DON,POIDS=POND,IDENT=NOM,
28                 DATAMAR=MARGES,M=2,EDITPOI=OUI,OBSELI=OUI)
IML Ready
Exiting IML.

         ************************************************************
         ***   Valeur du critère d'arrêt à l'itération 1 : 0.56651   ***
         ************************************************************

IML Ready
Exiting IML.

         ************************************************************
         ***   Valeur du critère d'arrêt à l'itération 2 : 0.17766   ***
         ************************************************************
IML Ready
Exiting IML.

         ************************************************************
         ***   Valeur du critère d'arrêt à l'itération 3 : 0.04198   ***
         ************************************************************
IML Ready
Exiting IML.

         ************************************************************
         ***   Valeur du critère d'arrêt à l'itération 4 : 0.00322   ***
         ************************************************************
IML Ready
Exiting IML.

         ************************************************************
         ***   Valeur du critère d'arrêt à l'itération 5 : 0.00002   ***
         ************************************************************
29
30          PROC PRINT DATA=__OBSELI;
31          TITLE2 "Liste des observations éliminées";
NOTE: The PROCEDURE PRINT printed page 10.
NOTE: The PROCEDURE PRINT used 0.01 CPU seconds and 2940K.

NOTE: The SAS session used 4.60 CPU seconds and 2972K.
NOTE: SAS Software Limited, Wittington House, Marlow, SL7 2EB
```

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".

III.1.3 Le listing

```
                                Un petit exemple commenté de calage sur marges

                            *********************************
                            ***   Paramètres de la macro   ***
                            *********************************

Table en entrée                     DATA      =  DON
 Pondération initiale               POIDS     =  POND
 Pondération Qk                     PONDQK    =  __UN
 Identifiant                        IDENT     =  NOM

Table des marges                    DATAMAR   =  MARGES
 Marges en pourcentages             PCT       =  NON
 Effectif de la population          EFFPOP    =

Méthode utilisée                    M         =  2
 Borne inférieure                   LO        =
 Borne supérieure                   UP        =
 Seuil d'arrêt                      SEUIL     =  0.0001
 Nombre maximum d'itérations        MAXITER   =  99

Table contenant la pond. finale     DATAPOI   =  SORTIE
 Mise à jour de la table DATAPOI    MISAJOUR  =  OUI
 Pondération finale                 POIDSFIN  =  PONDFIN
 Label de la pondération finale     LABELPOI  =  pondération raking ratio

Edition des poids                   EDITPOI   =  OUI
 Statistiques sur les poids         STAT      =  OUI
 Contenu de la table DATAPOI        CONTPOI   =  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)

                          Modalité          Marge          Marge      Pourcentage    Pourcentage
Variable              ou variable     échantillon     population     échantillon     population

X                               1              18             20           22.50          25.00
                                2              62             60           77.50          75.00

Y                               1              43             30           53.75          37.50
                                2              37             50           46.25          62.50

VAR.NUM                         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[11] de l'échantillon : cette modalité est donc légèrement sous-représentée, puisque sa fréquence dans la population est de 25%.
Le total 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.

[11] 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
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	2	1.32247	1.81959	1.99270	2.00648	2.00656
Y	1	-0.73974	-0.94297	-1.02331	-1.02984	-1.02987
Y	2	.	.	.	.	.
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 a 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 ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
X	1	20.000	20	25.00	25.00
	2	60.000	60	75.00	75.00
Y	1	30.000	30	37.50	37.50
	2	50.000	50	62.50	62.50
VAR.NUM	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 *.

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	1	1	1	1.01683
2	1	2	2	1	1.22897
3	2	1	1	1	1.14602
4	2	1	2	1	0.49456
5	2	1	3	1	0.21342
6	2	2	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=1 et Z=1 ; le rapport pondération finale/pondération initiale vaut 1.01683 pour cette observation... Les 3 observations pour lesquelles X=2, Y=2 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

Univariate Procedure

Variable=_F_ Rapport de poids

Moments

N	8	Sum Wgts	8
Mean	1.031891	Sum	8.255131
Std Dev	0.444812	Variance	0.197858
Skewness	-1.21649	Kurtosis	0.18399
USS	9.903406	CSS	1.385006
CV	43.10651	Std Mean	0.157265
T:Mean=0	6.561485	Pr>!T!	0.0003
Num ¬ =0	8	Num > 0	8
M(Sign)	4	Pr>=!M!	0.0078
Sgn Rank	18	Pr>=!S!	0.0078
W:Normal	0.811594	Pr<W	0.0394

Quantiles(Def=5)

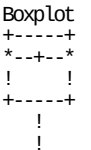
100% Max	1.385113	99%	1.385113
75% Q3	1.385113	95%	1.385113
50% Med	1.187493	90%	1.385113
25% Q1	0.755692	10%	0.213423
0% Min	0.213423	5%	0.213423
		1%	0.213423
Range	1.17169		
Q3-Q1	0.629421		
Mode	1.385113		

Extremes

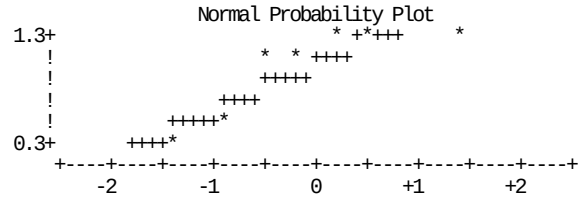
Lowest	ID	Highest	ID
0.213423(E	)	1.14602(D	)
0.494557(I	)	1.228966(H	)
1.016827(A	)	1.385113(F	)
1.14602(D	)	1.385113(G	)
1.228966(H	)	1.385113(K	)

```
Stem Leaf                     #
  12 3999                     4
  10 25                       2
   8                          0
   6                          0
   4 9                        1
   2 1                        1
     ----+----+----+----+
Multiply Stem.Leaf by 10**-1
```

Boxplot



Normal Probability Plot



```
                    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

                                  Univariate Procedure

Variable=__WFIN        Pondération finale

                Moments                                  Quantiles(Def=5)                          Extremes

N                8  Sum Wgts         8      100% Max  19.39158       99%  19.39158        Lowest    ID      Highest    ID
Mean            10  Sum              80       75% Q3   11.84356       95%  19.39158     2.774494(E       ) 9.831729(H       )
Std Dev    5.061209  Variance    25.61584       50% Med        10       90%  19.39158     4.451013(I       ) 10.16827(A       )
Skewness   0.439584  Kurtosis    1.109319       25% Q1   7.073401       10%  2.774494     9.695789(F       )  11.0809(G       )
USS         979.3109 CSS         179.3109        0% Min  2.774494        5%  2.774494     9.831729(H       ) 12.60622(D       )
CV         50.61209  Std Mean    1.789408                                1%  2.774494     10.16827(A       ) 19.39158(K       )
T:Mean=0   5.588441  Pr>!T!       0.0008      Range      16.61709
Num ¬ =0         8  Num > 0            8      Q3-Q1      4.770161
M(Sign)          4  Pr>=!M!      0.0078      Mode       2.774494
Sgn Rank        18  Pr>=!S!      0.0078
W:Normal   0.926636  Pr<W         0.4908

     Stem Leaf                            Boxplot                          Normal Probability Plot
       18 4                       1          0                    19+                                  * ++++
       16                                                         !                                ++++
       14                                                         !                              ++++
       12 6                       1          !                    !                          ++++ *
       10 21                      2         +--+--+              11+                       ++*+ *
        8 78                      2         !     !                !                  * +*+
        6                                   +-----+                !                ++++
        4 5                       1          !                    !             ++*+
        2 8                       1          !                   3+          +*++
          ----+----+----+----+                                      +----+----+----+----+----+----+----+----+----+----+
                                                                       -2        -1         0        +1        +2
```

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

                                  Méthode : raking ratio
                Contenu de la table SORTIE contenant la nouvelle pondération PONDFIN

                                  CONTENTS PROCEDURE

Data Set Name: WORK.SORTIE                          Observations: 8
Member Type: DATA                                   Variables: 2
Engine: V607                                        Indexes: 0
Created: 17:46 Wednesday, August 11, 1993           Observation Length: 16
Last Modified: 17:46 Wednesday, August 11, 1993     Deleted Observations: 0
Protection: Compressed: NO
Data Set Type: Sorted: NO
Label:

                                        ...

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

                     Variable    Type    Len    Pos    Label
                    ------------------------------------------------------------
                    1    NOM         Char      8      0
                    2    PONDFIN     Num       8      8    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 : 11 AOUT 1993              Heure : 17:46
*
*   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 été réalisé en 5 itérations
*   Les poids ont été stockés dans la variable PONDFIN de la table SORTIE
```

Un petit exemple de calage sur marges
Liste des observations éliminées

OBS	NOM	X	Y	Z	POND	__UN
1	B	1	2	2	0	1
2	C	1	2	3	.	1
3	J		2	2	10	1

III.2 Exemple 2 : l'enquête sur la consommation alimentaire de 1991

III.2.1 Les variables de calage

Jusqu'en 1993, l'INSEE a réalisé de façon périodique des enquêtes auprès des ménages sur la consommation alimentaire, permettant d'évaluer la consommation par produit, en valeur et en quantité, selon différentes catégories de ménages, et de disposer d'une information sur l'évolution de cette consommation.

L'unité d'observation est le **ménage**, mais on recueille également des informations au niveau **individuel**, en ce qui concerne les repas pris hors domicile. C'est pourquoi le redressement de cette enquête s'effectue à deux niveaux : ménage et individu.

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
- âge du chef de ménage (en classes)
- catégorie de commune.

Au niveau individu

On impose à l'échantillon d'individus d'avoir la même structure que la population pour les variables suivantes :

- nombre d'hommes de 0 à 14 ans
- nombre d'hommes de 15 à 34 ans
- nombre d'hommes de 35 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 34 ans
- nombre de femmes de 35 à 64 ans
- nombre de femmes de 65 ans et plus

i.e. pour la variable croisée sexe × tranche d'âge.

La procédure utilisée permet de réaliser **simultanément** ces deux redressements, en opérant sur un seul fichier, celui des ménages. Il suffit pour cela de calculer, pour chaque ménage, le nombre d'hommes de moins de 15 ans, le nombre d'hommes de 15 à 34 ans, etc., et de prendre en compte ces variables dans le calage en tant que **variables numériques**. Le poids d'un individu est alors égal au poids du ménage auquel il appartient.

La liste des variables du calage, ainsi que la signification des modalités des variables catégorielles, figurent dans le tableau suivant.

```
VARIABLES CATEGORIELLES

Nombre de personnes du ménage : NBPERS

 1 = 1 personne, 2 = 2 personnes, ... , 6 = 6 personnes et plus

Catégorie socioprofessionnelle du chef de ménage : CS

 1 = agriculteurs exploitants          2 = artisans,commerç.,chefs d'entreprise
 3 = cadres et prof. intellect. sup.   4 = professions intermédiaires
 5 = employés                          6 = ouvriers
 7 = inactifs, retraités, non déclarés

Tranche d'âge du chef de ménage : AGE

 1 = 25 ans ou moins     2 = 25 à 34 ans     3 = 35 à 44 ans     4 = 45 à 54 ans
 5 = 55 à 64 ans         6 = 65 à 74 ans     7 = 75 ans ou plus

Catégorie de commune : CCOM

 1 = communes rurales                    2 = unités urb. de moins de 10 000 h
 3 = unités urb. de 10 000 à 50 000 h    4 = unités urb. de 50 000 à 200 000 h
 5 = unités urb. de plus de 200 000 h    6 = unité urbaine de Paris


VARIABLES NUMERIQUES

 Nombre d'hommes de moins de 15 ans  : H14
 Nombre d'hommes de 15 à 34 ans      : H34
 Nombre d'hommes de 35 à 64 ans      : H64
 Nombre d'hommes de 65 ans et plus   : H65
 Nombre de femmes de moins de 15 ans : F14
 Nombre de femmes de 15 à 34 ans     : F34
 Nombre de femmes de 35 à 64 ans     : F64
 Nombre de femmes de 65 ans et plus  : F65
```

III.2.2 Le programme

```
TITLE "Consommation alimentaire 1991";
PROC PRINT DATA=LIB.MARGES;
TITLE2 "Les marges du calage";
RUN;
TITLE2;
%CALMAR(DATA=LIB.DONNEES,DATAMAR=LIB.MARGES,M=1,
        DATAPOI=TABPOIDS,POIDSFIN=POND1,LABELPOI=méthode linéaire,CONTPOI=NON)
%CALMAR(DATA=LIB.DONNEES,DATAMAR=LIB.MARGES,M=3,LO=0.64,UP=1.27,
        DATAPOI=TABPOIDS,POIDSFIN=POND2,LABELPOI=logit 0.64 1.27)
```

La macro CALMAR est utilisée d'abord pour mettre en œuvre la méthode linéaire, puis la méthode logit LO=0.64 UP=1.27. Ces valeurs de LO et UP conduisent à une étendue des rapports de poids minimale, et c'est cette pondération qui a été choisie par le responsable d'enquête. Le lecteur comparera sur les listings suivants les stem-and-leaf plots produits respectivement par les deux méthodes.

III.2.3 Extraits du listing

Consommation alimentaire 1991
Les marges du calage

OBS	VAR	N	MAR1	MAR2	MAR3	MAR4	MAR5	MAR6	MAR7
1	nbpers	6	5877995	6837628	3837825	3439589	1357160	633517	.
2	cs	7	600974	1238331	2014891	2915746	2237863	4674507	8301402
3	age	7	853360	4042908	4673046	3405158	3428823	2923662	2656757
4	ccom	6	5573103	2390861	2485027	3112787	4545572	3876364	.
5	h14	0	5487252	.	.	.	.	.	.
6	h34	0	8286609	.	.	.	.	.	.
7	h64	0	10033635	.	.	.	.	.	.
8	h65	0	3276351	.	.	.	.	.	.
9	f14	0	5239125	.	.	.	.	.	.
10	f34	0	8263830	.	.	.	.	.	.
11	f64	0	10298373	.	.	.	.	.	.
12	f65	0	4792209	.	.	.	.	.	.

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
NBPERS	1	4855298.80	5877995	22.09	26.74
	2	7120413.27	6837628	32.39	31.10
	3	4080664.23	3837825	18.56	17.46
	4	3617266.77	3439589	16.45	15.65
	5	1566560.08	1357160	7.13	6.17
	6	743510.86	633517	3.38	2.88
CS	1	556768.59	600974	2.53	2.73
	2	1116995.38	1238331	5.08	5.63
	3	1836298.90	2014891	8.35	9.17
	4	3603434.01	2915746	16.39	13.26
	5	2406900.26	2237863	10.95	10.18
	6	4907171.65	4674507	22.32	21.26
	7	7556145.21	8301402	34.37	37.76
AGE	1	1016707.87	853360	4.62	3.88
	2	4077206.04	4042908	18.55	18.39
	3	5024750.11	4673046	22.86	21.26
	4	3212658.53	3405158	14.61	15.49
	5	3627641.34	3428823	16.50	15.60
	6	2835715.82	2923662	12.90	13.30
	7	2189034.29	2656757	9.96	12.09
CCOM	1	6103705.40	5573103	27.76	25.35
	2	2610933.47	2390861	11.88	10.88
	3	2770010.21	2485027	12.60	11.30
	4	2994792.56	3112787	13.62	14.16
	5	4419566.85	4545572	20.10	20.68
	6	3084705.50	3876364	14.03	17.63
VAR.NUM	H14	6421858.88	5487252	.	.
	H34	8368819.86	8286609	.	.
	H64	10322697.23	10033635	.	.
	H65	3250698.63	3276351	.	.
	F14	6193618.34	5239125	.	.
	F34	8828759.14	8263830	.	.
	F64	10882924.01	10298373	.	.
	F65	4243199.16	4792209	.	.

```
                                        Méthode : linéaire
                 Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
                                     et sur les pondérations finales

                                          Univariate Procedure

Variable=_F_          Rapport de poids

              Moments                             Quantiles(Def=5)                          Extremes

 N             6357  Sum Wgts       6357   100% Max  1.757045    99%  1.575844     Lowest    Obs     Highest    Obs
 Mean             1  Sum            6357    75% Q3   1.144918    95%  1.390131   -0.20337(   2505) 1.683897(    579)
 Std Dev   0.226527  Variance   0.051315    50% Med  0.986824    90%  1.297802   0.143981(   2071) 1.706895(    125)
 Skewness  0.164995  Kurtosis   0.063889    25% Q1    0.85271    10%  0.728145   0.223909(   1260) 1.712514(    126)
 USS       6683.155  CSS         326.155     0% Min  -0.20337     5%  0.639925   0.321227(   3094) 1.712514(    325)
 CV         22.6527  Std Mean   0.002841                          1%  0.494616   0.327756(   3626) 1.757045(    241)
 T:Mean=0  351.9703  Pr>!T!       0.0001   Range     1.960418
 Num ¬ =0      6357  Num > 0        6356   Q3-Q1     0.292209
 M(Sign)     3177.5  Pr>=!M!      0.0001   Mode      0.893587
 Sgn Rank  10104450  Pr>=!S!      0.0001
 D:Normal  0.043643  Pr>D           <.01

                   Histogram                          Boxplot                  Normal Probability Plot
   1.75+*                                         4     0     1.75+                                                *
       .***                                      51     0         !                                               **
   1.55+***                                      61     0     1.55+                                             ****+
       .*******                                 164     !         !                                          *****+
   1.35+*************                           309     !     1.35+                                        *****
       .****************************            650     !         !                                     *****
   1.15+***********************************     840  +-----+  1.15+                                 ******
       .**************************************  903  !  +  !      !                              +****
   0.95+*********************************************** 1110  *-----*  0.95+                        ******
       .********************************************** 1095  +-----+      !                     ******
   0.75+***************************             630     !     0.75+                 ******
       .**************                          336     !         !             *****
   0.55+******                                  134     !     0.55+  *****+
       .***                                      59     0         !***
   0.35+*                                         8     0     0.35+*
       .*                                         1     0         !*
   0.15+*                                         1     0     0.15+*
       .                                                          !
  -0.05+                                                     -0.05+
       .                                                          !
  -0.25+*                                         1     *    -0.25+*
        ----+----+----+----+----+----+----+----+--                 +----+----+----+----+----+----+----+----+----+
        * may represent up to 24 counts                                   -2        -1         0        +1        +2

Variable=__WFIN       Pondération finale

              Moments                             Quantiles(Def=5)                          Extremes

 N             6357  Sum Wgts       6357   100% Max  6076.196    99%  5449.567     Lowest    Obs     Highest    Obs
 Mean       3458.19  Sum        21983714    75% Q3   3959.344    95%  4807.338     -703.3(   2505) 5823.237(    579)
 Std Dev   783.3736  Variance   613674.1    50% Med  3412.625    90%  4488.046   497.9127(   2071) 5902.766(    125)
 Skewness  0.164995  Kurtosis   0.063889    25% Q1   2948.832    10%  2518.065   774.3213(   1260) 5922.199(    126)
 USS       7.992E10  CSS        3.9005E9     0% Min    -703.3     5%  2212.982   1110.863(   3094) 5922.199(    325)
 CV         22.6527  Std Mean   9.825232                          1%  1710.475   1133.443(   3626) 6076.196(    241)
 T:Mean=0  351.9703  Pr>!T!       0.0001   Range     6779.496
 Num ¬ =0      6357  Num > 0        6356   Q3-Q1     1010.513
 M(Sign)     3177.5  Pr>=!M!      0.0001   Mode      3090.192
 Sgn Rank  10104450  Pr>=!S!      0.0001
 D:Normal  0.043643  Pr>D           <.01

                   Histogram                          Boxplot                  Normal Probability Plot
   6250+*                                         1     0     6250+                                                *
       .**                                       59     0         !                                               **
       .***                                      96     !         !                                             ****++
       .***********                             408     !         !                                        ********
       .************************                909     !         !                                  *******
       .*************************************  1378  +-----+      !                            *******
       .*********************************************** 1778  *--+--*      !                     ********
   2750+******************************         1115  +-----+  2750+              ********
       .************                            447     !         !      *******
       .****                                    142     !         !*******
       .*                                        21     0         !*
       .*                                         1     0         !*
       .*                                         1     0         !*
       .                                                          ]
   -750+*                                         1     *     -750+*
        ----+----+----+----+----+----+----+----+--                 +----+----+----+----+----+----+----+----+----+
        * may represent up to 38 counts                                   -2        -1         0        +1        +2
```

```
                    **********
                    ***    BILAN    ***
                    **********
*
*   Date : 16 AOUT 1993              Heure : 14:35
*
*   Table en entrée : LIB.DONNEES
*
*   Nombre d'observations dans la table en entrée  : 6357
*   Nombre d'observations éliminées                : 0
*   Nombre d'observations conservées               : 6357
*
*   Variable de pondération : taille de la population (21983714) / nombre d'observations (6357) (générée)
*
*   Nombre de variables catégorielles : 4
*   Liste des variables catégorielles et de leurs nombres de modalités :
      NBPERS (6) CS (7) AGE (7) CCOM (6)
*   Taille de l'échantillon (pondéré) : 21983714
*   Taille de la population           : 21983714
*
*   Nombre de variables numériques : 8
*   Liste des variables numériques :
      H14 H34 H64 H65 F14 F34 F64 F65
*
*   Méthode utilisée : linéaire
*   Le calage a été réalisé en 2 itérations
*   Il y a 1 poids négatifs
*   Les poids ont été stockés dans la variable POND1 de la table TABPOIDS
```

Méthode : logit, inf=0.64, sup=1.27
Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
et sur les pondérations finales

Univariate Procedure

Variable=_F_ Rapport de poids

Moments

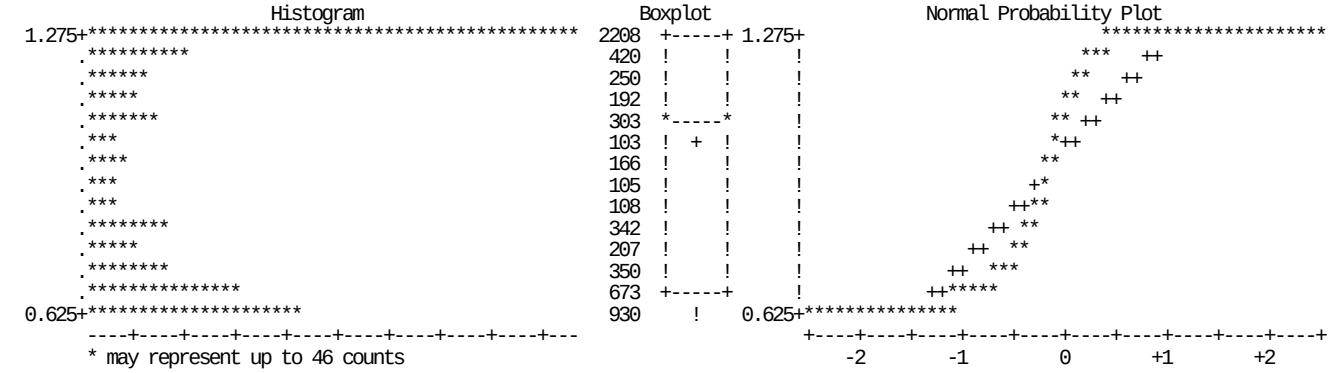
N	6357	Sum Wgts	6357
Mean	1	Sum	6357
Std Dev	0.26178	Variance	0.068529
Skewness	-0.25609	Kurtosis	-1.69297
USS	6792.57	CSS	435.5696
CV	26.17802	Std Mean	0.003283
T:Mean=0	304.5715	Pr>|T|	0.0001
Num ¬ =0	6357	Num > 0	6357
M(Sign)	3178.5	Pr>=|M|	0.0001
Sgn Rank	10104452	Pr>=|S|	0.0001
D:Normal	0.198246	Pr>D	<.01

Quantiles(Def=5)

100% Max	1.27	99%	1.27
75% Q3	1.26904	95%	1.27
50% Med	1.085355	90%	1.269995
25% Q1	0.69779	10%	0.641969
0% Min	0.64	5%	0.640174
		1%	0.64
Range	0.63		
Q3-Q1	0.57125		
Mode	0.816983		

Extremes

Lowest	Obs	Highest	Obs
0.64(	2505)	1.27(	693)
0.64(	2071)	1.27(	791)
0.64(	1512)	1.27(	806)
0.64(	1260)	1.27(	961)
0.64(	1043)	1.27(	125)



Méthode : logit, inf=0.64, sup=1.27
Contenu de la table TABPOIDS contenant la nouvelle pondération POND2

CONTENTS PROCEDURE

Data Set Name: WORK.TABPOIDS	Observations: 6357
Member Type: DATA	Variables: 2
Engine: V607	Indexes: 0
Created: 14:37 Monday, August 16, 1993	Observation Length: 16
Last Modified: 14:37 Monday, August 16, 1993	Deleted Observations: 0
Protection: Compressed: NO	
Data Set Type: Sorted: NO	
Label:	

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

Data Set Page Size: 6144
Number of Data Set Pages: 17
File Format: 607
First Data Page: 1
Max Obs per Page: 380
Obs in First Data Page: 344
Physical Name: SYS93228.T143515.RA000.WWCA91.R0000001
Release Created: 6.07
Release Last Modified: 6.07
Created by: WWCA91
Last Modified by: WWCA91
Subextents: 4
Total Blocks Used: 17

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

	Variable	Type	Len	Pos	Label
1	POND1	Num	8	0	méthode linéaire
2	POND2	Num	8	8	logit 0.64 1.27

****
*** BILAN ***
****

```
*
*   Date : 16 AOUT 1993              Heure : 14:35
*
*   Table en entrée : LIB.DONNEES
*
*   Nombre d'observations dans la table en entrée  : 6357
*   Nombre d'observations éliminées                : 0
*   Nombre d'observations conservées               : 6357
*
*   Variable de pondération : taille de la population (21983714) / nombre d'observations (6357) (générée)
*
*   Nombre de variables catégorielles : 4
*   Liste des variables catégorielles et de leurs nombres de modalités :
      NBPERS (6) CS (7) AGE (7) CCOM (6)
*   Taille de l'échantillon (pondéré) : 21983714
*   Taille de la population           : 21983714
*
*   Nombre de variables numériques : 8
*   Liste des variables numériques :
      H14 H34 H64 H65 F14 F34 F64 F65
*
*   Méthode utilisée : logit, borne inférieure = 0.64, borne supérieure = 1.27
*   Le calage a été réalisé en 9 itérations
*   Les poids ont été stockés dans la variable POND2 de la table TABPOIDS
```

IV. Les contrôles et les messages d'erreur

Sauf mention contraire, les résultats des contrôles et les messages indiqués ci-dessous sont imprimés sur le listing.

IV.1 Les contrôles

Lorsque le paramètre CONT vaut OUI, les contrôles suivants sont effectués.

IV.1.1 Contrôles sur les paramètres de la macro

- le paramètre **DATA** est renseigné ;
- la table **&DATA**[12] existe ;
- le paramètre **DATAMAR** est renseigné ;
- la table **&DATAMAR** existe ;
- le paramètre **M** vaut 1, 2, 3 ou 4 ;
- les paramètres **LO** et **UP** sont renseignés lorsque M vaut 3 ou 4 ;
- le paramètre **EFFPOP** 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 &DATAMAR ;
- la variable **&POIDS** existe dans la table &DATA, et elle est numérique ;
- la variable **&IDENT** existe dans la table &DATA ;
- la variable **&PONDQK** existe dans la table &DATA, et elle est numérique ;
- la variable **&IDENT** existe dans la table &DATA.

IV.1.2 Contrôles sur le contenu de la table &DATAMAR

- la variable **VAR** existe dans la table &DATAMAR ;
- la variable **N** existe dans la table &DATAMAR, et elle est numérique ;

[12] i.e. la table spécifiée dans le paramètre DATA

- les variables **MAR1, MAR2**... de la table &DATAMAR sont numériques ;
- les variables de calage nommées dans la variable VAR de la table &DATAMAR existent dans la table &DATA ;
- les variables de calage "numériques" nommées dans la variable VAR de la table &DATAMAR (i.e. pour lesquelles N=0) sont des variables numériques (au sens de SAS) de la table &DATA ;
- 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 IV.3.1).

IV.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 &DATAMAR), on vérifie que dans la table &DATA :

- aucune des modalités 1, 2 ... p (ou 01, 02 ... p si p >9) n'a un effectif nul[13] (voir exemple IV.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.

Ce contrôle est souvent utile car il permet de vérifier que les seules modalités de la variable catégorielle dans la table &DATA sont 1, 2 ... p, et en particulier que le recodage préalable éventuel de cette variable était correct (par exemple que l'on n'a pas oublié de recoder une modalité "autre" valant 9, ou 99...).

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 (voir exemple IV.3.3).

IV.1.4 Contrôles sur la table contenant les pondérations finales

Ces contrôles sont réalisés même si CONT vaut NON.

- le paramètre **POIDSFIN** est renseigné lorsque le paramètre DATAPOI l'est ;
- si table &DATAPOI 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 **sur la log**, et non sur le listing.

[13] Ce contrôle est effectué même si CONT vaut NON.

IV.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.

IV.2.1 Pas d'observation pour réaliser le calage

Ceci peut se produire dans les cas suivants :

- la table &DATA n'a pas d'observation ; ceci peut en particulier arriver lorsque l'on spécifie une clause WHERE dans la paramètre DATA, conduisant à ne sélectionner aucune observation...
- la table &DATA 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 IV.3.4).

IV.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.

IV.2.2.1 Les variables du calage sont colinéaires

Lorsque les variables du calage sont colinéaires, le calage est impossible, car le système d'équations (E) du § I.2 est indéterminé. Le programme élimine automatiquement les colinéarités structurelles qui apparaissent lorsque plusieurs variables catégorielles figurent dans les variables de calage (voir § I.4.1). 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[14] : ceci peut permettre à l'utilisateur d'identifier plus facilement l'origine de ces colinéarités (voir exemple IV.3.5).

IV.2.2.2 Le calage ne peut être réalisé

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[15] 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 sur le listing : "Le calage ne peut être réalisé" (voir exemple IV.3.6).

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 ;

[14] Les variables catégorielles sont "éclatées" en variables indicatrices des modalités.

[15] Lorsque l'on utilise la méthode du raking ratio (M=2), il y a une borne implicite LO=0.

- "relâcher" les contraintes sur les rapports de poids, en diminuant la valeur de LO ou en augmentant la valeur de UP ;

- rectifier la variable de pondération initiale si l'effectif de l'échantillon, pondéré par cette variable, 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.

IV.2.2.3 Le nombre maximum d'itérations 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.

IV.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 IV.3.7).

Ce phénomène peut se produire lorsque les contraintes imposées aux rapports de poids sont "à la limite de ce qu'ils peuvent supporter".

Ce message peut également apparaître, ainsi que les * indiquant les divergences, alors même que tout semble s'être bien passé. Cela est dû semble-t-il à la précision avec laquelle SAS compare marges de l'échantillon et marges du calage, bien supérieure à celle attendue en général par le statisticien.

IV.3 Exemples

Pour chacun des exemples suivants, sont donnés le programme et le listing (ou un extrait) produit par la macro.

IV.3.1 Les totaux des marges catégorielles ne sont pas tous égaux

```
DATA DON;
INPUT NOM $  X $ Y $ Z T;
POND=10;
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";
%CALMAR(DATA=DON,POIDS=POND,IDENT=NOM,DATAMAR=MARGES,M=1)
```

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

IV.3.2 Modalités de variables catégorielles d'effectif nul

```
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";
%CALMAR(DATA=DON,IDENT=NOM,DATAMAR=MARGES,M=1,PCT=OUI,EFFPOP=1000)
```

Contrôle sur les modalités des variables catégorielles : pas d'effectif nul

Comparaison entre les marges tirées de l'échantillon (avec la pondération initiale)
et les marges dans la population (marges du calage)
ERREUR : l'effectif d'une modalité (au moins) d'une variable catégorielle est nul
alors que la marge correspondante est non nulle : le calage est impossible

Variable	Modalité ou variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population	Effectif nul
X	1	600	400	60.00	40.00	
	2	400	600	40.00	60.00	
Y	1	600	500	60.00	50.00	
	2	400	500	40.00	50.00	
T	1	400	200	40.00	20.00	
	2	0	500	0.00	50.00	*
	3	600	300	60.00	30.00	
VAR.NUM	Z	3200	2400	.	.	

IV.3.3 Modalités de variables catégorielles non permises

```
DATA DON;
INPUT NOM $  X $ Y $ Z T;
POND=200;
CARDS;
A 1 1 1 1
B 1 2 2 3
C 0 2 3 1
D 2 1 1 3
E 2 1 9 2
;
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 "Contrôle sur les modalités des variables catégorielles :"
      " pas de modalités interdites";
%CALMAR(DATA=DON,POIDS=POND,IDENT=NOM,DATAMAR=MARGES,M=1,PCT=OUI,EFFPOP=1000)
```

Contrôle sur les modalités des variables catégorielles : pas de modalités interdites

ERREUR : pour au moins une variable catégorielle, l'effectif cumulé (pondéré) des modalités n'est pas égal à l'effectif (pondéré) de l'échantillon

Variable	Modalité	Marge échantillon	Pourcentage échantillon	Effectif cumulé	Effectif échantillon	Erreur
X	1	400	40	.	.	
X	2	400	40	800	1000	*
Y	1	600	60	.	.	
Y	2	400	40	1000	1000	
T	1	400	40	.	.	
T	2	200	20	.	.	
T	3	400	40	1000	1000	

Contrôle sur les modalités des variables catégorielles : pas de modalités interdites

Les effectifs (pondérés) des modalités des variables catégorielles en erreur

X	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	200	20.0	200	20.0
1	400	40.0	600	60.0
2	400	40.0	1000	100.0

IV.3.4 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";
%CALMAR(DATA=DON,POIDS=POND,IDENT=NOM,DATAMAR=MARGES,M=1,PCT=OUI,EFFPOP=1000)
```

Aucune observation valide

```
**************************************************************************
***   ERREUR : la table DON                                            ***
***            spécifiée dans le paramètre DATA 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.                                  ***
**************************************************************************
```

IV.3.5 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";
%CALMAR(DATA=DON,IDENT=NOM,DATAMAR=MARGES,M=1,PCT=OUI,EFFPOP=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
VAR.NUM	U	11600	14000	.	.
	Z	4400	5000	.	.
	T	2400	3000	.	.

Méthode : linéaire

```
**************************************************
***   Les variables analysées sont colinéaires :   ***
***   le calage ne peut être réalisé               ***
**************************************************
```

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

IV.3.6 Calage impossible

```
DATA DON;
INPUT X Y Z;
POND=10;
CARDS;
1 1 1
2 2 2
1 2 3
2 1 1
1 2 3
2 1 2
1 3 3
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";
%CALMAR(DATA=DON,DATAMAR=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
VAR.NUM	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	.

IV.3.7 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";
%CALMAR(DATA=DON,DATAMAR=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
VAR.NUM	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

La macro CALMAR

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	*
VAR.NUM	Z	110.001	110	.	.	*

Convergence imparfaite

Méthode : raking ratio

Statistiques sur les rapports de poids (= pondérations finales / pondérations initiales)
et sur les pondérations finales

Univariate Procedure

Variable=_F_ Rapport de poids

Moments

N	6	Sum Wgts	6
Mean	1.000004	Sum	6.000024
Std Dev	1.598608	Variance	2.555546
Skewness	2.141812	Kurtosis	4.79421
USS	18.77778	CSS	12.77773
CV	159.8601	Std Mean	0.652629
T:Mean=0	1.532271	Pr>|T|	0.1860
Num ¬= 0	6	Num > 0	6
M(Sign)	3	Pr>=|M|	0.0313
Sgn Rank	10.5	Pr>=|S|	0.0313
W:Normal	0.691145	Pr<W	0.0041

Quantiles(Def=5)

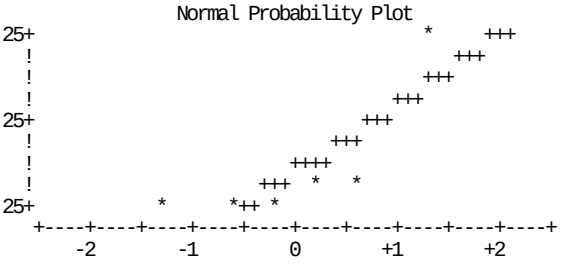
100% Max	4.166667	99%	4.166667
75% Q3	0.833333	95%	4.166667
50% Med	0.5	90%	4.166667
25% Q1	0.000024	10%	6.86E-10
0% Min	6.86E-10	5%	6.86E-10
		1%	6.86E-10
Range	4.166667		
Q3-Q1	0.833309		
Mode	6.86E-10		

Extremes

Lowest	Obs	Highest	Obs
6.86E-10(	5)	0.000024(	3)
0.000024(	3)	0.166667(	1)
0.166667(	1)	0.833333(	2)
0.833333(	2)	0.833333(	4)
0.833333(	4)	4.166667(	6)

```
Stem Leaf                     Boxplot
   4 2                1        *
   3                             |
   3                             |
   2                             |
   2                             |
   1                             |
   1                             |
   0 88               2        +-----+
   0 002              3        +--+--+
     ----+----+----+----+
```

Normal Probability Plot



Convergence imparfaite

```
                  **********
                  ***    BILAN    ***
                  **********
*
*   Date : 18 OCTOBRE 1993          Heure : 10:40
*
*   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)
*   Taille de l'échantillon (pondéré) : 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 être réalisé qu'approximativement en 12 itérations
```