Direction des Statistiques Démographiques et Sociales
Unité "Méthodes statistiques"
Division "Méthodologie d'Elaboration et d'Analyse des Données"
Timbre F410
Olivier SAUTORY

INSEE
DIRECTION GÉNÉRALE
INSTITUT NATIONAL
DE LA STATISTIQUE
ET DES ÉTUDES
ÉCONOMIQUES

Partial English translation[1] of the document

La macro CALMAR
Redressement d'un échantillon par calage sur marges

Document de travail n° F 9310

Novembre 1994

[1] The main part of this translation has been made by the Australian Bureau of Statistics.

II. Implementing the CALMAR macro

Note : throughout this document the formulations "data set &DATAMAR", "variable &POIDSFIN" etc, signify : the data set specified by the DATAMAR parameter of the macro, variable specified by the POIDSFIN parameter, etc.

II.1. Input data sets

II.1.1. SAS data set containing input data

You need a SAS data set containing the sample data :

. variables to be used for the reweighting, or *calibration variables*

. the *initial weighting variable*

. eventually an i*dentifying variable*.

This data set can of course contain any other variable [2] which does not appear directly in the reweighting.

The name of this SAS data set is given in the mandatory DATA parameter of the macro.

II.1.1.1. Categorical calibration variables

A categorical or qualitative variable in the calibration sense can be either a SAS character or a SAS numeric variable.

- For a **character** variable its p categories need to take values :

1,2, ..., p	if $p \leq 9$
01, 02, ..., p	if $10 \leq p \leq 99$
001, 002, ..., 010, ..., p	if $100 \leq p \leq 999$

- For a **numeric** variable, its p categories need to take values 1, 2, ..., p

These constraints may often necessitate prior recoding of these variables.

The names of the categorical calibration variables, as well as their number of categories are specified in the data set of marginals &DATAMAR (see II.1.2.).

[2] including an additional "weighting "variable &PONDK (see II.2.1.)

The macro carries out the following controls :

- no category of a categorical variable is empty
- a categorical variable which is shown as having p categories takes only the p permitted category values (as specified above).

II.1.1.2. Numerical calibration variables

A numerical (or quantitative) variable in the calibration sense must be numeric in the SAS sense.

The names of the numerical calibration variables are specified in the data set of marginals &DATAMAR (see II.1.2.).

II.1.1.3. The initial weighting variable

This variable takes for each observation k its intial weight d_k. This value is equal to the inverse of the probability of inclusion of the observation in the sample.

For example, in the case of a simple random sample, or a multi-stage "self-weighting" survey, each population unit has the same probability of belonging to the sample, equal to n/N, where n is the sample size and N the population size. In this case, the initial weight applied to each sample observation is constant, and has value N/n, a quantity sometimes called the "inflation factor".

The initial weighting variable needs to be numeric in the SAS sense. It is specified by the POIDS parameter of the macro.

Choice of initial weights

When at least one of the calibration variables is categorical, the initial weight can be defined up to a mutiplicative coefficient : while the weight ratios [3] and the parameter λ (see 1) depend on the choice of initial weights, final weights w_k do not change if you multiply each initial weight by a constant [4].

For example, in the case of a survey where all units have the same probability of selection, the initial weighting variable can take any constant value, for example 1, 1000, N/n...

Nevertheless, there are at least two good reasons to specify the "proper" initial weight (N/n in the above example) :

- from a theoretical point of view, the initial weight is defined as the inverse of the inclusion probability, and the weight ratios measure the deviation of this weighting from the calibration; in particular, these weight ratios have mean 1 in the case of a constant "inflation factor", equal to N/n.
- from a practical point of view, begining with an initial weight far removed from the final weight (for example an initial weight equal to 1 in a sample of 3000 observations for a population of size 21 million) can exceed capacity limits during program calculations : then the macro will send the message "*Le calage ne peut être réalisé*" (i.e. "the calibration cannot be achieved") ... incorrectly since this is only an accidental impossibility due to "poor" specification of the calibration problem.

[3] For an observation weight ratio is the ratio defined as : final weighting / initial weighting.

[4] Provided that bounds are consequently modified, when a bounded method is used.

Remark : in this case the macro prints after this message, the size of the weighted sample (with the initial weights) and the size of the population.

Generated initial weighting

When there is at least one categorical variable among the calibration variables, and if the initial weighting variable is not specified in the POIDS parameter, the macro generates a constant weighting variable, equal to the ratio of population size to number of retained (non-eliminated) observations in the input data set (the population total is calculated from the table of marginals, or is given in the EFFPOP parameter).

If there is no categorical calibration variable, the POIDS parameter must be specified.

II.1.1.4. Other variables in the input data set

The data set &DATA can contain other variables than those defined above. In particular the following can appear :

- a variable which identifies observations, specified by the IDENT parameter
- a variable which defines a supplementary weighting of the observations given by the PONDK parameter (its use is only justified in very particular cases, see reference [2]).

II.1.1.5. Excluded observations

Any observation in the input data set with a missing value for one of the calibration variables or one of the weighting variables, or with a negative or zero value for one of the weighting variables, is excluded from the calibration (and consequently from the output data set eventually created by the macro).

II.1.1.6. Calibration in the presence of nonresponse

In principle, the calibration procedures (as presented in chapter I) are only valid in the absence of total non-response [5] in the sample of size n, or alternatively after adjustment of this non-response. If these conditions are not met, one can work directly on the sample of the m respondents without altering initial weights : it can be shown that this method reduces to applying two corrections simultaneously, one for non-response and the other to improve the estimation [6]. However, two things are worth noting when the current practice, which consists to multiplyithe initial weights by n/m, is used :

- in the case of a simple random sample, the sum of initial weights reproduces the population total N ;
- the weighting algorithm can operate under optimal conditions (see Choice of initial weights above).

II.1.2. SAS data set containing calibration variables and marginals

The names of calibration variables, the number of categories and the associated marginals must be provided in a SAS data set, whose name is specified in the mandatory DATAMAR parameter of the macro.

This data set contains an observation for each calibration variable. Data set variables are : VAR, N, MAR1, MAR2, ..., MARh, taking the following values :

VAR variable name (upper or lower case).

N number of categories of the variable.

This is a strictly positive integer for a categorical variable, and 0 for a numerical variable. A negative value for N is replaced by 0, and a positive non-integer is replaced by its integer portion.

MAR1 for a categorical variable : marginal count for the category 1

for a numerical variable : marginal count

...

MARj for a categorical variable with at least j categories :marginal count for the category j

for a categorical variable with less than j categories or for a numerical variable : missing value (.)

...

MARh for a categorical variable with h categories, where h is the maximum number of categories (i.e. the maximal value for N) : :marginal count for the category h

---

[5] There is total non-response when a sampled unit did not respond to the survey.

[6] Correction for non-response uses a response model based on the same variables as for the calibration (see F. DUPONT : "Calage et redressement de la non-réponse totale", Journées de méthodologie statistique 1993).

for a categorical variable with less than h categories or for a numerical variable : missing value (.)

The macro performs the following controls :

- each variable specified in variable VAR exists in the data set &DATA
- a variable with N = 0 is a numerical variable of data set &DATA.
- for a variable with N = p > 0, the marginal counts MAR1 to MARp are provided.
- totals of categorical marginal counts for the categrical variables are all equal (in principle, these totals equal the size of the population).

Marginal counts for categorical variables given in percentages

For categorical variables, the user can give marginals counts in percentages, provided the macro parameter PCT is set to OUI ("yes"). In this case the marginal totals need to be all equal to 100, and the user must indicate the population size in the parameter EFFPOP.

The reader can refer to III.1.1. and III.2.3. to see examples of correctly set out &DATAMAR data sets, and to IV for examples of errors to avoid.

II.2. Syntax of the macro

II.2.1. Parameters specifying the input SAS datasets

DATA = name of SAS data set

name of the SAS data set containing input data (**mandatory**).

This data set contains for each sample observation the categorical and numerical calibration variables, and eventually an identifying variable. It also contains the **initial weighting** variable (except in the case where the latter is generated).

See detailed contents of this data set at II.1.1.

DATAMAR = name of SAS data set

name of the SAS data set containing the calibration variable names, the numbers of categories and the marginal counts (**mandatory**).

See detailed contents of this data set at II.1.2.

Remark : one can use the WHERE clause, FIRSTOBS, OBS, KEEP options... to define these two data sets. For example, one can write :

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

By using these options with the &DATAMAR data set, one can select or change calibration variables from among a set of potential variables.

POIDS = variable

numerical variable containing initial weights of sample observations (it belongs to the &DATA data set).

This parameter is mandatory when there is no categorical calibration variable (see II.1.1.3.).

PONDQK = variable

numerical weighting variable for sample observations, not tied to the specified variable in the parameter POIDS (it belongs to the &DATA data set) : it allows the calibration function to vary according to observations (see reference [2]).

Default setting : PONDQK = _ UN, a generated variable which takes constant value of 1.

IDENT = variable

variable which identifies observations in printed outputs ; it appears in the output data set (DATAPOI parameter) containing final weights.

PCT = OUI or NON

If PCT equals OUI, the marginal counts for calibration variables are given in percentages in the data set &DATAMAR .

By default : PCT = NON

EFFPOP = value

If PCT equals OUI, this variable is used to specify the population size (which is mandatory to calculate calibration marginal counts).

This parameter is mandatory if PCT = OUI

II.2.2. Parameters specifying the method to be used

M = 1, 2, 3 or 4

method number, i.e. the distance function used to measure the distance from the original weights to the final weights :

1. linear method.
2. raking ratio method.
3. logit method.
4. truncated linear method.

LO = value

lower bound for weight ratios [7], when using a "bounded" method (logit or truncated linear).

This parameter is mandatory when M = 3 or 4

UP = value

upper bound for weight ratios when using a "bounded" method.

This parameter is mandatory when M = 3 or 4

SEUIL = value

threshold ε for stopping rule in the Newton algorithm : there is convergence when the maximum absolute value of differences between the weight ratios calculated from successive iterations is less than the threshold.

By default : SEUIL = 0.0001

MAXITER = n

maximum number of iterations for the Newton algorithm : if the algorithm has not converged after n iterations it stops.

By default : MAXITER = 15

[7] For an observation weight ratio is the ratio defined as final weight / initial weight.

II.2.3. Output data sets parameters

DATAPOI = name of SAS data set

name of the SAS data set containing final weights.

If this data set does not exist, it is created by the macro : it has as many observations as non-excluded observations in the data set &DATA ; it contains the &POIDSFIN variable (see below) and the &IDENT variable (if specified)

If this data set exists, the next parameter indicates how the macro updates the data set.

By default : no data set is created

MISAJOUR = OUI or NON

This parameter specifies how the macro updates the &DATAPOI data set when it already exists :

- if MISAJOUR = OUI, the weighting variable &POIDSFIN, and the variable &IDENT (if specified), is added to the data set
- if MISAJOUR = NON, the macro creates a new data set containing variables &POIDSFIN (and &IDENT), the previous data set with the same name is overwritten.

By default : MISAJOUR = OUI

POIDSFIN = variable

name of the variable containing the final weights for non-excluded observations from the sample (it belongs to the data set &DATAPOI).
This parameter is mandatory when the parameter DATAPOI is provided.

LABELPOI = label

label given to the &POIDSFIN variable.

Remark : this label must not contain commas.

By default : no label is given.

OBSELI = OUI or NON

If OBSELI = OUI, the macro creates a SAS data set, called _OBSELI, which contains the eliminated observations, calibration variables, weighting variables and the &IDENT variable (if specified). The user can print or use this data set after calling the macro.

By default : OBSELI =NON

II.2.4. Parameters specifying printed outputs

CONT = OUI or NON

If CONT = OUI, the macro performs a number of controls :

- on the parameters of the macro (presence of mandatory parameters, coherence of the parameters...)
- on the values given to these parameters (existence of SAS data sets, of weighting variables...)
- on the data given in data set &DATAMAR (existence or variables, presence of all marginal counts...)
- on the calibration variables in data set &DATA.

The whole list of these controls, as well as examples of messages printed by the macro, is given at IV.

By default : CONT = OUI

EDITPOI = OUI or NON

If EDITPOI = OUI, the macro prints values of weight ratios obtained for each combination of values of the calibration variables (both categorical and numerical).

Remark : this data set can be very large, especially where numerical variables are included.

By default : EDITPOI = NON

STAT = OUI or NON

If STAT = OUI, the macro prints statistics (mean, standard deviation, quantiles, extreme values...) and graphics for the distributions of variables **weight ratio** and **final weight** [8].

By default : STAT = OUI

CONTPOI = OUI or NON

If CONTPOI = OUI, the macro summarises content of data set &DATAPOI [9].

By default : CONTPOI = OUI

NOTES = OUI or NON

If NOTES = NON, notes generated by SAS are not printed.

By default : NOTES = NON

II.3. Printed output

The macro prints :

[8] Printed output from an UNIVARLATE procedure.
[9] Printed output from a CONTENTS procedure.

- a table giving parameter values
- a table comparing marginals calculated from the sample with initial weights and population marginals (calibration marginals)
- a table giving the value of the stopping criterion and the number of negative weights after each iteration
- a table giving coefficients of vector of Lagrangian multipliers λ after each iteration
- a table comparing marginals calculated from the sample with final weights and population marginals (calibration marginals) : these marginals must be the same
- **if EDITPOI = OUI** : a table giving values of weight ratios obtained for each combination of values of calibration variables
- **if STAT = OUI** : the output of the UNIVARIATE procedure (mean, median, standard deviation, quantiles, extreme values, stem-and-leaf plot...) on the weight ratio variable and on the final weight variable
- **if CONTPOI = OUI** : the output of the CONTENTS procedure on the data set containing final weights.
- a summary of the calibration :
  - the name of the input data set.
  - the number of (non weighted) observations in this data set.
  - the number of observations excluded, the number of observations retained
  - the name of the initial weighting variable, or, when it is generated by the program, the (constant) value of this variable : the size of population divided by the number of observations
  - the number of categorical variables, their names and the numbers of categories
  - the size of the weighted sample, i.e. the sum of initial weights calculated on retained observations, provided there is at least one categorical variable among the calibration variables
  - the size of population, calculated using the marginals or else given in the parameter EFFPOP, provided there is at least one categorical variable among the calibration variables
  - the number and the list of numerical variables
  - the method used
  - the number of iterations
  - optionnally the name of the variable containing final weights and the name of the data set containing this variable.

When an error has occurred the preceeding outputs are not fully provided, and the macro usually prints a message to explain why the program stopped.

II.4. Output data set

The output data set specified in the DATAPOI parameter can be temporary or permanent. Its observations are the retained observations of the data set &DATA ; it contains the final weighting variable &POIDSFIN and optionnally the identifying variable &IDENT. **Observations are ranked in the same order in the input data set &DATA and the output data set &DATAPOI.**

- If the data set does not exist it is created by the macro.
- If the data set exists, and if MISAJOUR = OUI, it is updated by the macro, i.e. the new variable(s) is added to the existing data set :
  - if one variable with the name of the added variable already exists in the data set, it is overwritten.
  - if the number of (retained) observations is greater than the number of observations in the data set before executing the macro, this data set is "completed" by addition of missing values for the pre-existing variables.
  - if the number of (retained) observations is less than the number of observations in the data set before executing the macro, the new variables are "completed" by the addition of missing values.

  Remark : it is preferable in practice to avoid situations described in the last two cases, for example by creating a number of output data sets. In particular, in such situations, if the variable &IDENT does not change name, the identifiers no longer correspond to the weighting values...
- If the data set exists and if MISAJOUR = NON, the previous version of the data set is overwritten and replaced by a data set containing the new variable &POIDSFIN (and possibly the &IDENT variable).

III. Examples

III.1. Example 1 : a brief annotated example

III.1.1. The program

```
LIBNAME COMPIL 'GR90.MACROS.COMPIL' DISP=SHR;
OPTIONS SASMSTORE=COMPIL MSTORED NODATE;

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 input data set is the data set DON.

POIDS the variable containing the initial weights is the numerical variable POND in the data set DON.

IDENT the variable NOM is used to identify observations in printed output and in the output data set.

DATAMAR the data set containing marginals is the data set MARGES.

The contents of this data set indicate that the calibration will use 3 variables : X and Y are categorical variables having 2 categories each (N takes value 2) and Z is a numerical variable (N takes value 0).
These three variables appear in the data set DON.

The calibration marginals for the variable X are respectively 20 and 60 : this signifies that the weighted total after calibration for category 1 (resp. cat 2) of X should be equal to 20 (resp. 60). Similarly the marginals for Y are 30 and 50. The Z marginal is 140 : this means that the weighted sum for Z after calibration should equal 140.

M the raking ratio method is used.

EDITPOI editing of weight ratios for the different combinations of the values of the variables is requested.

OBSELI the macro creates a data set, called _OBSELI, containing the excluded observations (if any)

DATAPOI the output data set SORTIE contains, all going well, the final weights.

POIDSFIN the variable in the data set SORTIE, containing final weights, is called PONDFIN.

LABELPOI the label "raking ratio weighting" is given to the variable PONDFIN

.
The other parameters take their default-values, namely :

PONDQK _UN : no supplementary weighting.

PCT NON : the marginals are not given in percentages.

SEUIL 0.0001 : threshold for stopping rule.

MAXITER 15 : maximum number of iterations.

STAT OUI : statistics for weight ratios and final weights will be produced.

CONTPOI OUI : the OUTPUT data set content will be produced.

CONT OUI : controls will be run.

NOTES NON : no printing of SAS notes.

III.1.2. The log

```
NOTE: SAS system options specified are:
       SORT=4

 NOTE: The initialization phase used 0.13 CPU seconds and 1149K.
 1          LIBNAME COMPIL 'GR90.MACROS.COMPIL' DISP=SHR;
 NOTE: Libref COMPIL was successfully assigned as follows:
       Engine: V607
       Physical Name: GR90.MACROS.COMPIL
 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: The DATA statement used 0.03 CPU seconds and 1336K.

 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: The DATA statement used 0.02 CPU seconds and 1336K.

 25         ;
 WARNING: The COMPIL.SASMACR catalog is opened for read only.
 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
```

SAS notes are not printed during the execution of the macro. The printing in the log output of successive values of the stopping criterion for the algorithm (0.56651, 0.17766, 0.04198...) allow the user to follow the development of the algorithm in real time.

III.1.3. The output

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)

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

The weighted total for the category 1 of the X variable in the sample takes a value of 18, which represents 22.5 % of the weighted sample total [10] : this category is therefore lightly under-represented, since its proportion in the population is 25 %.
The total of the numerical variable Z in the sample (152) exceeds the population total (140).

Note : the observations B, C and J have been excluded for they take respectively value zero for the weighting POND, a missing value for POND and a missing value for variable X.

[10] Which equals the sum of the initial weight variable using retained observations.

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

The stopping criterion has dropped below the threshold 0.00001 after 5 iterations ; there are no negative weights (which is as expected using raking ratio method). Examining the table of lambda vectors is useful when there is no convergence : often in this case the components of lambda become very large, showing in a certain way the impossibility the algorithm to reach the corresponding marginals.

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	.	.

This table is analogous to the first, but here its sampling marginals are calculated using the final weights : they should be equal to the population marginals ; if this is not the case divergences are indicated using an *.

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

This table is printed because EDITPOI = OUI.
There is one observation in the input table for which X = 1, Y = 1 and Z = 1 ; the ratio of final to initial weights equals 1.01683 for this observation...
The 3 observations for which X = 2, Y = 2 and Z = 2 have a ratio equal to 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)

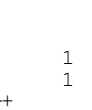
100% Max	1.385113	99%	1.385113
75% Q3	1.385113	95%	1.385113
50% Med	1.187493	90%	1.385113
25% Q1	0.755692	10%	0.213423
0% Min	0.213423	5%	0.213423
		1%	0.213423
Range	1.17169		
Q3-Q1	0.629421		
Mode	1.385113		

Extremes

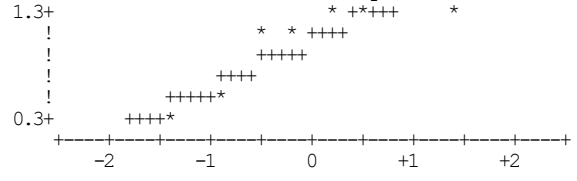
Lowest	ID	Highest	ID
0.213423	(E)	1.14602	(D)
0.494557	(I)	1.228966	(H)
1.016827	(A)	1.385113	(F)
1.14602	(D)	1.385113	(G)
1.228966	(H)	1.385113	(K)

```
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

N	8	Sum Wgts	8
Mean	10	Sum	80
Std Dev	5.061209	Variance	25.61584
Skewness	0.439584	Kurtosis	1.109319
USS	979.3109	CSS	179.3109
CV	50.61209	Std Mean	1.789408
T:Mean=0	5.588441	Pr>!T!	0.0008
Num ¬ =0	8	Num > 0	8
M(Sign)	4	Pr>=!M!	0.0078
Sgn Rank	18	Pr>=!S!	0.0078
W:Normal	0.926636	Pr<W	0.4908

Quantiles(Def=5)

100% Max	19.39158	99%	19.39158
75% Q3	11.84356	95%	19.39158
50% Med	10	90%	19.39158
25% Q1	7.073401	10%	2.774494
0% Min	2.774494	5%	2.774494
		1%	2.774494
Range	16.61709		
Q3-Q1	4.770161		
Mode	2.774494		

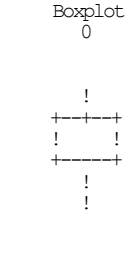
Extremes

Lowest	ID	Highest	ID
2.774494	(E)	9.831729	(H)
4.451013	(I)	10.16827	(A)
9.695789	(F)	11.0809	(G)
9.831729	(H)	12.60622	(D)
10.16827	(A)	19.39158	(K)

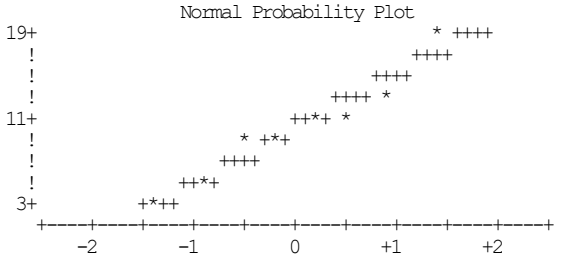
Stem Leaf

```
 18 4
 16
 14
 12 6
 10 21
  8 78
  6
  4 5
  2 8
    ----+----+----+----+
```

Boxplot



Normal Probability Plot



These outputs are printed because STAT = OUI.
The mean of the 8 weight ratios take value 1.031891, their standard deviation 0.444812, the extreme values are 1.385113 (observations F, G and K) and 0.213423 (obs E), etc.
The total for final weights takes value 80, which is as expected since this is the population total. These weights vary from 2.774494 (observation E) to 19.39158 (obs K), that is a range of 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

These outputs are printed because CONTPOI= OUI.
The macro has created the dataset SORTIE, which has 8 observations (those retained from the data set DON) and 2 variables : the final weighting variable PONDFIN with the label "pondération raking ratio", and the identifying variable NOM wihich appears in the data set 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

No comment.

III.2.Example 2 : Survey of food consumption 1991

III.2.1. Calibration variables

INSEE periodically carries out household food consumption surveys with object to estimate consumption by product / commodity in value and volume according to different household types, and to build up a series over time.

The unit of observation is the **household**, but information is also collected at **individual** level concerning meals taken away from home. Hence the re-weighting of this survey occurs at two levels : household and individual.

At household level

The household sample is constrained to the same structure as the population for the following **categorical variables** :

- size of household
- professional group of head of household
- age group of head of household
- type of agglomeration where the dwelling is located

At individual level

- number of males/females 0-14 years.
- number of males/females 15-34 years.
- number of males/females 35-64 years.
- number of males/females 65 + years.

i.e. age group by sex.

The procedure used will **simultaneously** meet both sets of constraints by operating on a single file, that of households. It is sufficient to calculate for each household the number of males less than 15, the number of males 15-34 years etc, and to treat these variables in the calibration as **numerical variables**. The weight of an individual is thus equal to the weight of the household to which he or she belongs.

The list of calibration variables, as well as the categories of categorical variables, is given in the table below.

CATEGORICAL VARIABLES

NBPERS = Number of persons in the household

1 = 1 person, 2 = 2 persons, ..., 6 = 6 persons or more

CS = professional group of head of household

1 = farmer
2 = trades, retail, employers
3 = professional/administrative
4 = para professionals
5 = salaried employees
6 = labourers
7 = retired, inactive, not stated

AGE : group age of head of household

1 = 25 years or less 2 = 25-34 years 3 = 35-44 years 4 = 45-54 years
5 = 55-64 years 6 = 65-74 years 7 = 75 years and over

CCOM : type of agglomeration

1 = rural communes
2 = urban centres < 10 k persons
3 = urban centres 10-50k persons
4 = urban centres 50-200k persons
5 = urban centre > 200k persons
6 = Paris urban centre

NUMERICAL VARIABLES

H14 No. males less than 15 yrs
H34 No. males 15-34 yrs
H64 No. males 35-64 yrs
H65 No. males 65 yrs +
F14 No females less than 15 yrs
F34 No females 15-34 yrs
F64 No females 35-64 yrs
F65 No females 65 yrs +

III.2.2. The program

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

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)
```

The CALMAR macro is first used to implement the linear method, then the logit method LO = 0.64 UP = 1.27. These values of LO and UP lead to a minimal range in the weight ratios, and it is this weighting which has been chosen by the survey administrators. The reader can compare the stem-and-leaf plots in the following listings produced by both methods.

III.2.3. The output

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

N	6357	Sum Wgts	6357
Mean	1	Sum	6357
Std Dev	0.226527	Variance	0.051315
Skewness	0.164995	Kurtosis	0.063889
USS	6683.155	CSS	326.155
CV	22.6527	Std Mean	0.002841
T:Mean=0	351.9703	Pr>!T!	0.0001
Num ¬ =0	6357	Num > 0	6356
M(Sign)	3177.5	Pr>=!M!	0.0001
Sgn Rank	10104450	Pr>=!S!	0.0001
D:Normal	0.043643	Pr>D	<.01

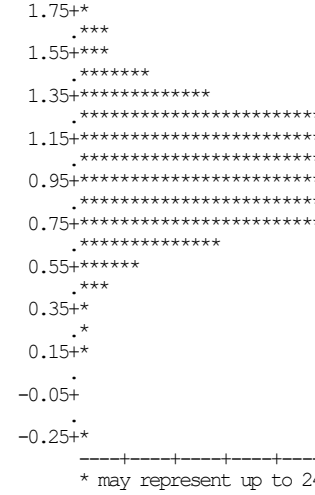
Quantiles(Def=5)

100% Max	1.757045	99%	1.575844
75% Q3	1.144918	95%	1.390131
50% Med	0.986824	90%	1.297802
25% Q1	0.85271	10%	0.728145
0% Min	-0.20337	5%	0.639925
		1%	0.494616
Range	1.960418		
Q3-Q1	0.292209		
Mode	0.893587		

Extremes

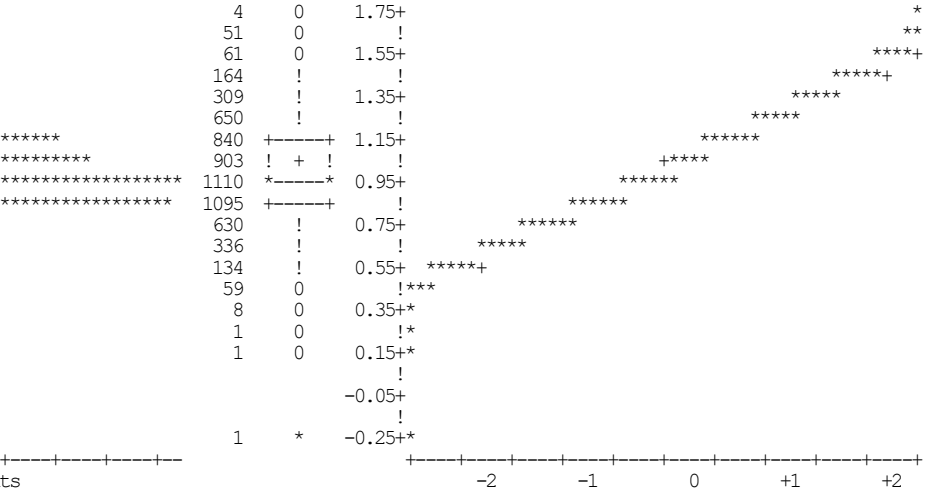
Lowest	Obs	Highest	Obs
-0.20337(	2505)	1.683897(	579)
0.143981(	2071)	1.706895(	125)
0.223909(	1260)	1.712514(	126)
0.321227(	3094)	1.712514(	325)
0.327756(	3626)	1.757045(	241)

Histogram



Boxplot

Normal Probability Plot



Variable=__WFIN Pondération finale

Moments

N	6357	Sum Wgts	6357
Mean	3458.19	Sum	21983714
Std Dev	783.3736	Variance	613674.1
Skewness	0.164995	Kurtosis	0.063889
USS	7.992E10	CSS	3.9005E9
CV	22.6527	Std Mean	9.825232
T:Mean=0	351.9703	Pr>!T!	0.0001
Num ¬ =0	6357	Num > 0	6356
M(Sign)	3177.5	Pr>=!M!	0.0001
Sgn Rank	10104450	Pr>=!S!	0.0001
D:Normal	0.043643	Pr>D	<.01

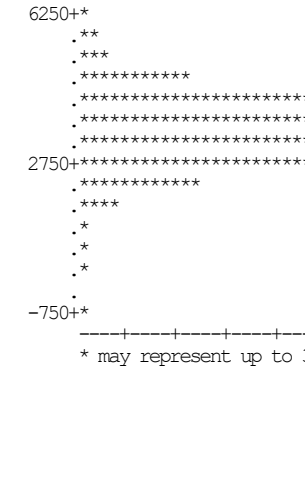
Quantiles(Def=5)

100% Max	6076.196	99%	5449.567
75% Q3	3959.344	95%	4807.338
50% Med	3412.625	90%	4488.046
25% Q1	2948.832	10%	2518.065
0% Min	-703.3	5%	2212.982
		1%	1710.475
Range	6779.496		
Q3-Q1	1010.513		
Mode	3090.192		

Extremes

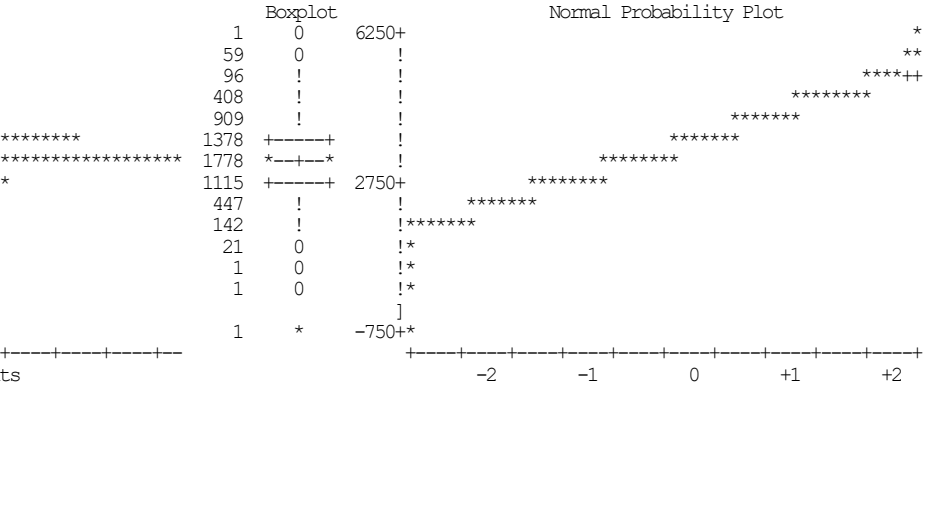
Lowest	Obs	Highest	Obs
-703.3(	2505)	5823.237(	579)
497.9127(	2071)	5902.766(	125)
774.3213(	1260)	5922.199(	126)
1110.863(	3094)	5922.199(	325)
1133.443(	3626)	6076.196(	241)

Histogram



Boxplot

Normal Probability Plot



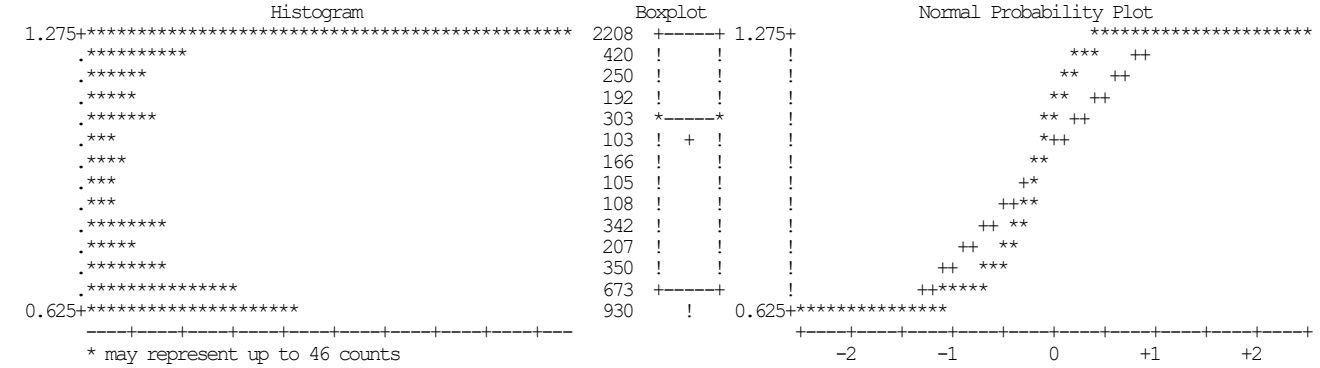
```
                    **********
                    ***    BILAN    ***
                    **********
*
*   Date : 16 AOUT 1993              Heure : 14:35
*
*   Table en entrée : LIB.DONNEES
*
*   Nombre d'observations dans la table en entrée  : 6357
*   Nombre d'observations éliminées                : 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				Quantiles(Def=5)			Extremes			
N	6357	Sum Wgts	6357	100% Max	1.27	99%	1.27	Lowest Obs	Highest	Obs
Mean	1	Sum	6357	75% Q3	1.26904	95%	1.27	0.64(2505)	1.27(	693)
Std Dev	0.26178	Variance	0.068529	50% Med	1.085355	90%	1.269995	0.64(2071)	1.27(	791)
Skewness	-0.25609	Kurtosis	-1.69297	25% Q1	0.69779	10%	0.641969	0.64(1512)	1.27(	806)
USS	6792.57	CSS	435.5696	0% Min	0.64	5%	0.640174	0.64(1260)	1.27(	961)
CV	26.17802	Std Mean	0.003283			1%	0.64	0.64(1043)	1.27(	125)
T:Mean=0	304.5715	Pr>|T|	0.0001	Range	0.63					
Num ¬ =0	6357	Num > 0	6357	Q3-Q1	0.57125					
M(Sign)	3178.5	Pr>=|M|	0.0001	Mode	0.816983					
Sgn Rank	10104452	Pr>=|S|	0.0001							
D:Normal	0.198246	Pr>D	<.01							



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