Redressements d'échantillons d'enquêtes auprès des ménages par calage sur marges

Olivier SAUTORY
Insee

Dans toutes (ou presque toutes) les enquêtes par sondage, le responsable d'enquête doit faire face, à un certain point de l'exploitation (en gros après la phase d'apurement des données et une éventuelle correction des non-réponses), à la redoutable question : "Comment faire pour redresser mon échantillon ?".

Ce problème surgit lorsque le statisticien dresse les premiers tableaux issus de son enquête, concernant généralement quelques variables structurelles "de base". Ainsi le responsable d'une enquête INSEE auprès des ménages examine les répartitions des ménages de son échantillon selon la catégorie de commune, l'âge du chef de ménage, la catégorie socioprofessionnelle du chef de ménage ... pour les confronter aux répartitions issues de sources "dignes de confiance" comme le recensement de la population ou l'enquête-emploi. Des divergences entre ces structures sont considérées comme gênantes, en particulier au niveau des publications : l'INSEE craint en effet que des incohérences entre des enquêtes réalisées à la même époque ne jettent le trouble dans les esprits de ses plus fidèles lecteurs.

Le redressement de l'échantillon va donc consister à modifier les poids de sondage des individus (égaux aux inverses des probabilités d'inclusion, cf §I), de façon à "caler" l'échantillon, pour un certain nombre de variables jugées importantes (celles qui sont supposées avoir un fort caractère explicatif), sur des structures reconnues fiables.

Il semble naturel de chercher une nouvelle pondération aussi proche que possible de la pondération initiale. Ce problème de "calage sur marges" rentre dans un cadre plus général d'estimation du total d'une variable dans une population en présence d'information auxiliaire, qui fait l'objet de la communication présentée au cours de ces journées par J.-C. Deville (2).

Nous commencerons donc par rappeler quelques résultats figurant dans (2), afin de voir comment ils s'appliquent dans le cas d'un calage sur marges, et nous verrons deux exemples d'application des techniques proposées à des enquêtes auprès des ménages.

I. LE CADRE GENERAL

1. Le problème

On considère une population $U = \{ 1 \dots k \dots N \}$ 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,

i.e. la probabilité, avant tirage, que l'individu appartienne à s. Dans le cas d'un sondage aléatoire simple avec remise, cette probablilité vaut n/N pour tout k.

Soit $\mathcal{Y}$ une variable d'intérêt, pour laquelle on désire estimer le total sur la population :

$$Y = \sum_U \mathcal{Y}(k) = \sum_U y_k$$

Remarque : le signe $\sum_U$, $\sum_s$ ... signifie que l'on effectue la sommation sur tous les individus k de U, de s ...

L'estimateur de Y utilisé classiquement est l'estimateur de Horvitz-Thompson, dont les propriétés sont connues (cf par exemple (3)) :

$$\hat{Y}_\pi = \sum_s \frac{1}{\pi_k} y_k = \sum_s d_k \, y_k$$

Utiliser cet estimateur sans biais de Y, appelé aussi estimateur par les valeurs dilatées, revient à affecter à chaque individu de l'échantillon un poids d_k égal à l'inverse de sa probabilité d'inclusion. Ainsi, dans le cas d'un sondage aléatoire simple, ce poids n'est rien d'autre que le "coefficient d'extrapolation" N/n, et chaque individu de s "représente" le même nombre d'individus N/n.

Information auxiliaire

Soit $\mathcal{X}_1 \ldots \mathcal{X}_j \ldots \mathcal{X}_J$ J variables auxiliaires connues sur l'échantillon s. On note :

$$\forall\, j = 1 \ldots J \qquad \forall\, k \in U \qquad \mathcal{X}_j(k) = x_{jk}$$

Les totaux de ces variables sur la population entière $X_j = \sum_U x_{jk}$ sont supposés connus.

On pose :

$$x_k = \begin{pmatrix} x_{1k} \\ \vdots \\ x_{Jk} \end{pmatrix} \qquad X = \begin{pmatrix} X_1 \\ \vdots \\ X_J \end{pmatrix}$$

On cherche à estimer le total Y de $\mathcal{Y}$ à l'aide d'un estimateur linéaire (par rapport aux y_k), de la forme :

$$\hat{Y}_w = \sum_s w_k \, y_k$$

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 :

$$\forall\, j = 1 \ldots J \qquad \sum_s w_k \, x_{jk} = X_j$$

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

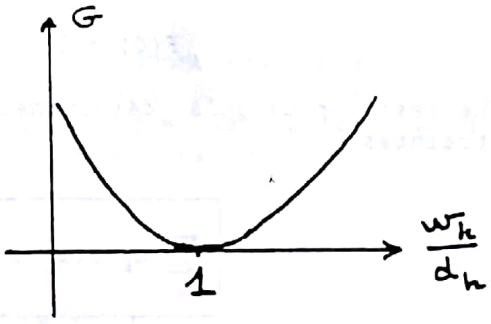
2. Résolution théorique

Il faut commencer par choisir une "fonction de distance" G pour mesurer les proximités entre les poids cherchés w_k et les poids de sondage d_k. La fonction G choisie, dont l'argument est $x = w_k / d_k$, vérifie les conditions suivantes :

(1) G est positive et convexe

(2) $G(1) = G'(1) = 0$

(3) $G''(1) = 1$



Les conditions (1) et (2) assurent des propriétés souhaitables pour G (cf graphe de G ci-dessus), alors que la condition (3) est purement technique (son utilité apparaîtra ultérieurement).

Une fois la fonction G choisie (voir §3 suivant), le problème consiste à déterminer les poids w_k ($k \in s$) solutions du programme suivant :

$$\min_{w_k} \sum_s d_k \, G\left(\frac{w_k}{d_k}\right)$$

$$\text{sous} \sum_s w_k \, x_k = X$$

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.

Le Lagrangien vaut :

$$\mathcal{L} = \sum_s d_k \, G\left(\frac{w_k}{d_k}\right) - \lambda' \left(\sum_s w_k \, x_k - X\right)$$

où $\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_J \end{pmatrix}$ est un vecteur de multiplicateurs de Lagrange.

On écrit les conditions du 1er ordre :

$$\forall\, k \in s \quad \frac{\partial \mathcal{L}}{\partial w_k} = G'\left(\frac{w_k}{d_k}\right) - \lambda' \, x_k = 0$$

d'où :

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

où F est la fonction réciproque de la fonction G' (dont l'existence est assurée par la condition (1)). Les conditions (2) et (3) assurent par ailleurs que :

$$F(0) = 1 \quad , \quad F'(0) = 1$$

Il ne reste plus qu'à déterminer le vecteur λ, en l'introduisant dans les contraintes :

$$\boxed{\sum_s d_k \, F(x_k' \, \lambda) \, x_k = X} \qquad (E)$$

Le problème est donc résolu dès lors que l'on sait résoudre en λ l'égalité ci-dessus, qui est un système non linéaire de J équations à J inconnues.

3. Les fonctions G usuelles

On indique ici les méthodes qui seront utilisées dans les exemples qui suivront. Pour chacune d'elles, on indique la fonction G(x) (où $x = w_k/d_k$) et la fonction F(u) (où $u = x_k' \, \lambda$).

a) méthode "linéaire"

. $G(x) = \frac{1}{2}(x - 1)^2 \quad , x \in \mathbb{R}$

. $F(u) = 1 + u \quad \in \mathbb{R}$

La forme linéaire de F donne son nom à cette méthode, fondée sur la "distance" G la plus "naturelle" (le coefficient 1/2 assure G"(1)=1).

b) méthode "raking ratio"

. $G(x) = x \text{ Log } x - x + 1 \quad , \quad x > 0$

. $F(u) = \exp u > 0$

Cette méthode classique de redressement, proposée par Deming et Stephan (1), est aussi connue sous le nom de méthode R.A.S., ou encore I.P.F. ("Iterative Proportionnal Fitting").

c) méthode "logit"

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

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

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

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

d) méthode "linéaire tronquée"

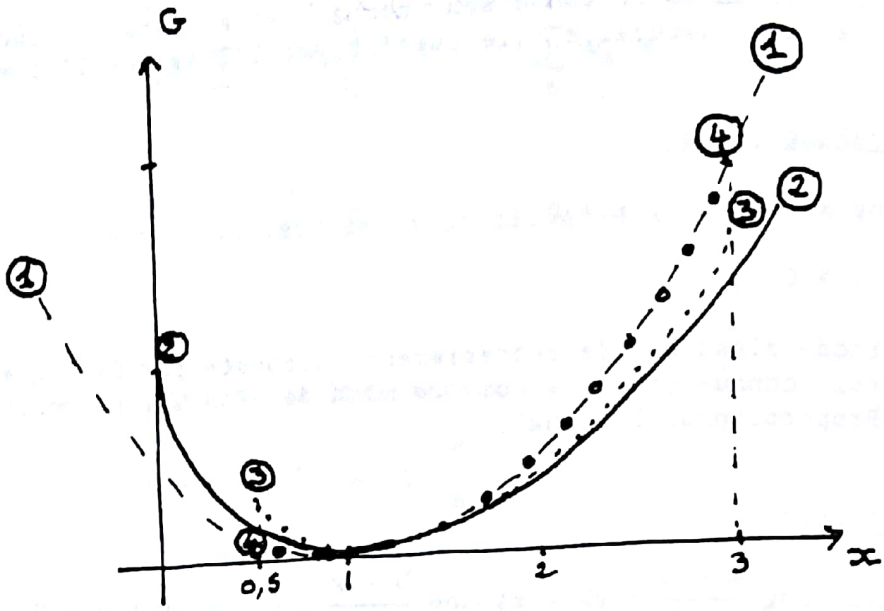
. $G(x) = \frac{1}{2} (x - 1)^2$ si $L \leqslant x \leqslant U$ (∞ sinon)

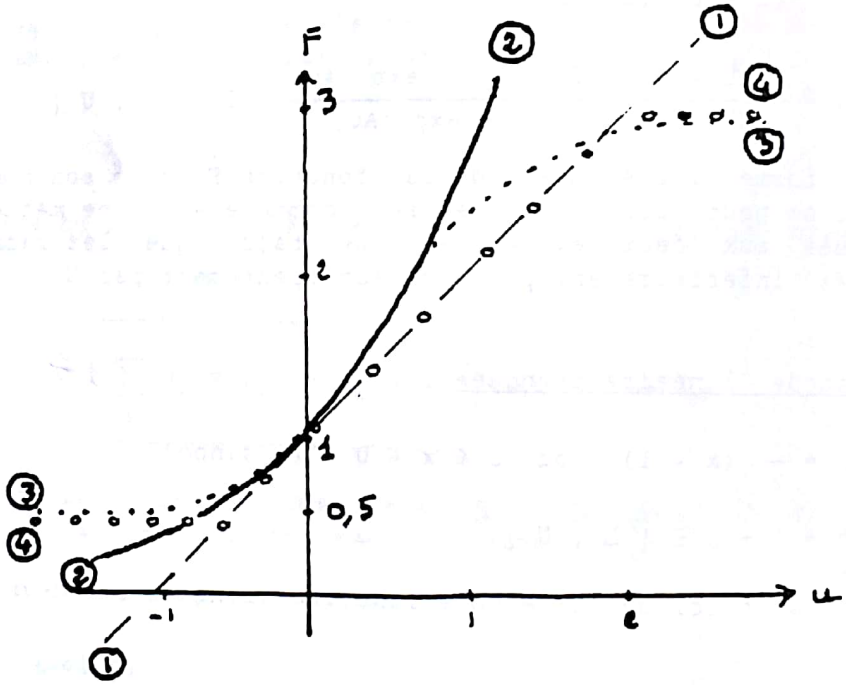
. $F(u) = 1 + u \in [L , U]$

Il s'agit ici de la méthode linéaire tronquée aux deux extrémités.

On trouvera à la page suivante les représentations graphiques des fonctions G(x) et F(u) pour chacune des 4 méthodes :

1. linéaire
2. raking ratio
3. logit (L=0.5 , U=3)
4. linéaire tronquée (L=0.5 , U=3)





II. RESOLUTION NUMERIQUE

Le système d'équations (E) peut se réécrire :

$$X = \sum_s d_k \left(F(x_k' \lambda) - 1 \right) x_k + \sum_s d_k x_k$$
$$= \Phi_s(\lambda) + \hat{X}_\pi$$

où $\Phi_s(\lambda)$ est un vecteur de $\mathbb{R}^J$, fonction de λ et des valeurs x_k de s, et

$$\hat{X}_\pi = \begin{pmatrix} \hat{X}_{1\pi} \\ \vdots \\ \hat{X}_{J\pi} \end{pmatrix}, \text{ où } \hat{X}_{j\pi} = \sum_s \frac{1}{\pi_k} x_{jk} .$$

$\hat{X}_{j\pi}$ est l'estimateur de Horvitz-Thompson du total X_j de $\mathcal{X}_j$.

Le système (E) s'écrit donc maintenant :

$$\boxed{\Phi_s(\lambda) = X - \hat{X}_\pi}$$

Si l'on a $X = \hat{X}_\pi$, i.e. si les contraintes sont vérifiées avec les poids d_k (l'échantillon était donc déjà "calé"), alors la solution évidente du système est $\lambda = 0$ (puisque l'on a $F(0) = 1$).

On peut résoudre numériquement ce système par la méthode de Newton ; on calcule une suite de vecteurs $\lambda^{(i)}$ par la relation de récurrence :

$$\lambda^{(i+1)} = \lambda^{(i)} - \left(\Phi_s'(\lambda^{(i)}) \right)^{-1} \left(\Phi_s(\lambda^{(i)}) - X + \hat{X}_\pi \right)$$

où $\Phi_s'(\lambda) = \sum_s d_k F'(x_k' \lambda) x_k x_k'$ est la matrice (carrée de taille J) des dérivées partielles de Φ.

On initialise l'algorithme avec un vecteur $\lambda^{(0)}$ quelconque. La convergence est obtenue lorsque les poids w_k ($= d_k F(x_k' \lambda)$) obtenus lors de deux itérations successives "ne bougent presque plus" (i.e. $\text{Max} \mid w_k^{(i+1)} - w_k^{(i)} \mid < \epsilon$).

Choix de $\lambda^{(0)}$

Dans la pratique, on prendra toujours $\lambda^{(0)} = \vec{0}$. On a alors :

. $\Phi_s(\vec{0}) = 0$, car $F(0) = 1$

. $\Phi_s'(\vec{0}) = \sum_s d_k x_k x_k' = T_s$, car $F'(0) = 1$

d'où : $\lambda^{(1)} = T_s^{-1} (X - \hat{X}_\pi)$, quelle que soit la méthode utilisée.

Cas de la méthode linéaire

Si l'on choisit la première méthode, on a $F(u) = 1 + u$, $F'(u) = 1$, d'où :

$$\Phi_s(\lambda^{(1)}) = \left(\sum_s d_k \, x_k \, x_k' \right) \lambda^{(1)} = T_s \quad T_s^{-1} \quad (X - \hat{X}_\pi) = X - \hat{X}_\pi$$

$$\Longrightarrow \qquad \lambda^{(2)} = \lambda^{(1)}$$

Il y a donc convergence dès la 1ère itération.

De plus, l'estimateur de Y obtenu a la forme suivante :

$$\hat{Y}_w = \sum_s w_k \, y_k = \sum_s d_k \, F(x_k' \, \lambda) \, y_k = \sum_s d_k \, (1 + x_k' \, \lambda) \, y_k$$

$$= \sum_s d_k \, y_k + \left(\sum_s d_k \, x_k' \, y_k \right) T_s' \, (X - \hat{X}_\pi)$$

$$= \hat{Y}_\pi + \hat{B}' \, (X - \hat{X}_\pi)$$

où $\hat{B} = \left(\sum_s \frac{1}{\pi_k} x_k \, x_k' \right)^{-1} \left(\sum_s \frac{1}{\pi_k} x_k' \, y_k \right)$ est l'estimateur des coefficients de la régression de $\mathcal{Y}$ sur $\mathcal{X}_1 \ldots \mathcal{X}_J$. On reconnaît donc dans $\hat{Y}_w$ la forme classique de l'"estimateur par régression" du total Y (cf (3)).

Conclusion

Si l'on initialise l'algorithme avec $\lambda_0 = \vec{0}$, on obtient toujours à la première itération, quelle que soit la méthode utilisée, le vecteur $\lambda^{(1)}$ correspondant à la solution de la méthode linéaire, solution conduisant à l'estimateur par la régression.

Remarque : le programme REDRE (4) implanté à l'INSEE utilise cette méthode d'estimation par régression pour redresser un échantillon.

III. APPLICATION AU CALAGE SUR MARGES

1. Cas de variables qualitatives

Soit $\mathcal{V}^1 \ldots \mathcal{V}^q \ldots \mathcal{V}^Q$ Q variables qualitatives, dont on note les modalités respectivement $1 \ldots i_1 \ldots I_1$, $1 \ldots i_q \ldots I_q$, $1 \ldots i_Q \ldots I_Q$. On note $c = (i_1 \ldots i_q \ldots i_Q)$ une case de l'hyper-tableau de contingence obtenu en croisant toutes ces variables.

Ces variables sont supposées connues sur l'échantillon s. En général, on se contente de se caler sur les marges, i.e. sur les distributions dans la

population des variables $\mathcal{V}^1 \ldots \mathcal{V}^Q$, et non pas sur les effectifs des cellules c, pour différentes raisons :

. les distributions de ces variables peuvent provenir de sources différentes, et les effectifs des cellules dans la population U ne sont alors pas connus

. lorsque l'on cale sur une "grosse enquête" (par exemple l'enquête-emploi), si les effectifs marginaux sont fiables, il n'en est pas de même pour les effectifs obtenus en croisant plusieurs variables

. même dans le cas où les effectifs des cellules dans la population sont connus, souvent il ne peut être question de réaliser un redressement par cellule, en raison de la faiblesse des effectifs des cellules dans l'échantillon ; ainsi, dans le premier exemple du §IΩ., l'hyper tableau de contingence contient 258048 cases ... alors que l'échantillon n'a que 6807 ménages.

Pour se ramener au cadre général, on introduit les variables indicatrices associées aux modalités des différentes variables qualitatives. Ce sont ces variables qui joueront le rôle des variables $\mathcal{X}_j$ du §I, sur les totaux desquels on désire se caler. .

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

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

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

$$x_k' = \left(\ldots \left(\delta^q_1(k) \ldots \delta^q_{i_q}(k) \ldots \delta^q_{I_q}(k) \right) \ldots \right)$$

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

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

où $N^q_{i_q} = \sum_U \delta^q_{i_q}(k)$ = nombre d'individus k de U tels que $\mathcal{V}^q(k) = i_q$.

Le vecteur λ des multiplicateurs de Lagrange a la forme suivante :

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

L'argument $x_k' \lambda$ de la fonction F, pour un individu k de la cellule $c = (i_1 \ldots i_q \ldots i_Q)$, a pour expression :

$$x_k' \lambda = \lambda^1_{i_1} + \ldots + \lambda^q_{i_q} + \ldots + \lambda^Q_{i_Q} \overset{\text{def}}{=} \eta_c$$

Les équations de calage (E) s'écrivent ici, pour tout q = 1...Q :

$$\forall\, i_q = 1 \ldots I_q \qquad N^q_{i_q} = \sum_{c \ni i_q} \hat{N}^c_\pi \; F(\lambda^1_{i_1} + \ldots + \lambda^q_{i_q} + \ldots + \lambda^Q_{i_Q})$$

où $\hat{N}^c_\pi = \sum_{k \in c} d_k$ est l'effectif estimé (par Horvitz-Thompson) dans la cellule c. Dans cette expression, on voit clairement apparaître les $F(x'_k \lambda)$ comme des coefficients permettant d'ajuster les effectifs des cases $\hat{N}^c_\pi$ sur les effectifs marginaux $N^q_{i_q}$; ces coefficients seront d'autant plus éloignés de 1 que les effectifs estimés et les effectifs marginaux diffèrent.

Le systéme d'équations ci-dessus est en fait sur-déterminé ; en effet, si l'on somme les I_Q équations relatives aux modalités d'une variable $\mathcal{V}^q$ donnée, on obtient :

$$N = \sum_{i_q=1}^{I_Q} N^q_{i_q} = \sum_c \hat{N}^c_\pi \; F(\eta_c)$$

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

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

Exemple : Q=3

On pose : $I_1 = I$, $I_2 = J$, $I_3 = K$, $\lambda' = (a_1 \ldots a_I \; b_1 \ldots b_{J-1} \ldots c_1 \ldots c_{L-1})$.

On obtient alors les expressions suivantes, pour le vecteur $\Phi_s(\lambda)$ et la matrice $\Phi'_s(\lambda)$:

$$\sum_s d_k \, F(x_k' \, \lambda) \, x_k \quad = \quad \begin{matrix} \\ (i) \\ \\ (j) \\ \\ (l) \\ \\ \end{matrix} \begin{pmatrix} \vdots \\ \sum_{j,l} \hat{N}_\pi^{ijl} \, F(a_i + b_j + c_l) \\ \vdots \\ \text{----------} \\ \vdots \\ \sum_{i,l} \hat{N}_\pi^{ijl} \, F(a_i + b_j + c_l) \\ \vdots \\ \text{----------} \\ \vdots \\ \sum_{i,j} \hat{N}_\pi^{ijl} \, F(a_i + b_j + c_l) \\ \vdots \end{pmatrix} \begin{matrix} \uparrow \\ I \\ \downarrow \\ \uparrow \\ J-1 \\ \downarrow \\ \uparrow \\ L-1 \\ \downarrow \end{matrix}$$

$$(= \Phi_s(\lambda) + \hat{X}_\pi)$$

$$\underset{(P,P)}{\Phi_s'(\lambda)} \quad = \quad \begin{matrix} \\ (i) \\ \\ (j) \\ \\ (l) \\ \\ \end{matrix} \begin{pmatrix} \ddots & 0 & | & \vdots & & \\ \sum_{j,l} \hat{N}_\pi^{ijl} \, F'(a_i+b_j+c_l) & & | & \sum_l \cdots & & \\ 0 & \ddots & | & \vdots & & \\ \text{----------} & & | & \text{------} & & \\ & \vdots & | & \ddots & & \\ \cdots \sum_l \hat{N}_\pi^{ijl} \, F'(a_i+b_j+c_l) & \cdots & | & \sum_{i,l} \cdots & \cdots & \\ & \vdots & | & 0 & & \\ \text{----------} & & | & \text{------} & & \\ & \vdots & | & \vdots & & \\ \cdots \sum_j \hat{N}_\pi^{ijl} \, F'(a_i+b_j+c_l) & \cdots & | & \cdots \sum_i \cdots & & \\ & \vdots & | & \vdots & & \end{pmatrix} \begin{matrix} \uparrow \\ I \\ \downarrow \\ \uparrow \\ J-1 \\ \downarrow \\ \uparrow \\ L-1 \\ \downarrow \end{matrix}$$

$$\leftarrow \quad I \quad \rightarrow \quad \leftarrow \quad J-1 \quad \rightarrow \quad \cdots$$

2. Cas de variables qualitatives et quantitatives

Il arrive que le responsable d'une enquête auprès des ménages ne se contente pas d'un redressement au niveau "ménage", mais qu'il soit aussi préoccupé des effectifs et des caractéristiques des individus qui composent ces ménages. Ainsi, dans le cas de l'enquête sur la consommation alimentaire, le statisticien produit des tableaux du type "structure de la consommation alimentaire des ménages selon les caractéristiques des ménages", mais aussi des tableaux sur le nombre de repas pris à l'extérieur par les individus. Il peut donc aussi souhaiter caler son échantillon sur quelques structures de la population, par sexe et âge par exemple.

Ce problème entre également dans le cadre général du §I : il suffit de définir, pour chaque ménage, un certain nombre de variables quantitatives (nombre de personnes de sexe masculin de moins de 15 ans dans le ménage, nombre de veufs dans le ménage ...) : le redressement consistera à se caler sur les totaux sur U de ces variables, supposés connus.

Ce type de redressement est donc une extension de celui présenté dans le paragraphe 1 : outre les Q variables qualitatives, on dispose maintenant de R variables quantitatives $\mathcal{Z}^1 \ldots \mathcal{Z}^r \ldots \mathcal{Z}^R$.

On note : $\forall\, k \in U \quad \forall\, r = 1 \ldots R \quad \mathcal{Z}^r(k) = z_k^r$.

On a alors :

$$x_k' = (\ldots \delta_{i_q}^q(k) \ldots \mid z_k^1 \ldots z_k^R)$$

$$X' = (\ldots N_{i_q}^q \ldots \mid Z^1 \ldots Z^R) \quad \text{où} \quad Z^r = \sum_U z_k^r$$

$$\lambda' = (\ldots \lambda_{i_q}^q \ldots \mid \mu^1 \ldots \mu^R)$$

Pour un individu k (dans l'exemple k désigne un ménage) de la cellule $c = (i_1 \ldots i_q \ldots i_Q)$:

$$x_k' \lambda = \lambda_{i_1}^1 + \ldots + \lambda_{i_q}^q + \ldots + \lambda_{i_Q}^Q + \mu^1 z_k^1 + \ldots + \mu^r z_k^r + \ldots \mu^R z_k^R$$

Le vecteur $\Phi_s(\lambda) + \hat{X}_\pi$ est alors complété par des quantités de la forme :

$$\sum_{k \in s} d_k \, F(x_k' \lambda) \, z_k^r$$

et la matrice $\Phi_s'(\lambda)$ est bordée "latéralement" par des termes de la forme :

$$\sum_{k \in c^{(i_q)}} d_k \, z_k^r \, F'(x_k' \lambda)$$

(où $c^{(i_q)}$ désigne une case de la forme (... i_q ...)),

et bordée "diagonalement par des termes de la forme :

$$\sum_s d_k \, z_k^r \, z_k^{r'} \, F'(x_k' \lambda).$$

IV. DEUX EXEMPLES D'APPLICATION

1. Quelques considérations générales

Ces méthodes de redressement sont utilisées régulièrement à l'INSEE depuis février 1990 pour redresser les enquêtes auprés des ménages, grâce à une procédure informatique utilisant le logiciel SAS.

Le programme permet actuellement d'utiliser les "distances" G présentées au §I.3.

A la lumière des premières expériences, on peut faire les quelques commentaires suivants :

. la méthode "linéaire", qui est évidemment la plus rapide puisqu'elle converge aprés deux itérations, présente la particularité, que les responsables d'enquête considèrent en général comme un inconvénient grave, de pouvoir conduire à des poids w_k négatifs. Par ailleurs, les poids ne sont pas bornés supérieurement, et peuvent prendre des valeurs "indésirables" (du genre $w_k/d_k > 3$ ou 4).

. la méthode "raking ratio" conduit à des poids toujours positifs, mais également 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), valeurs qui dépendent des données et des marges du calage. Ceci peut se comprendre facilement dans le cas d'un tableau 2 × 2.

On considère le tableau suivant, et les marges de calage indiquées:

1	2	2
2	1	4
4	2	

On suppose que $d_k = 1$ pour chacun des 6 individus de ce tableau.

On note a le poids w_k de l'individu de la case (1,1), b la valeur commune des poids des individus de la case (1,2), et de même c pour (2,1)

et d pour (2,2). Quelle que soit la méthode utilisée pour déterminer ces poids, ils doivent vérifier les égalités suivantes :

$$a + 2b = 2 \quad \Longrightarrow \quad b = 1 - a/2$$

$$a + 2c = 4 \quad \Longrightarrow \quad c = 2 - a/2$$

$$2c + d = 4 \quad \Longrightarrow \quad d = a$$

On vérifie aisément que la valeur inférieure de ces poids est a (ou d) si $a < 2/3$, et $b = 1-a/2$ si $a \geqslant 2/3$. Par conséquent, $L_{max} = 2/3$. De même, on trouve que $U_{min} = 4/3$.

L'éloignement par rapport à 1 de L_{max} et U_{min} traduit d'une certaine façon les divergences entre les marges tirées de l'échantillon et les marges du calage.

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 inférieurs 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. On a d'ailleurs pu constater, sur les exemples "grandeur nature" traités jusqu'à présent, que ces deux valeurs limites étaient "indépendantes" (alors que dans le petit exemple précédent, si $L = L_{max} = 2/3$, alors $U_{min} = 5/3$).

On peut indiquer enfin que plus les valeurs L et U sont proches de L_{max} et U_{min}, plus lente est la convergence de l'algorithme de Newton, en particulier pour la méthode "logit".

2. Enquête "modes de vie"

Il s'agit d'une enquête auprès des ménages, réalisée de novembre 1988 à novembre 1989, portant sur le domaine des activités domestiques. Elle s'est déroulée en 8 vagues de 6 semaines chacune, de façon à supprimer les effets saisonniers des comportements que l'on cherche à mesurer.

On trouvera à la page 19 la liste des variables (toutes qualitatives) utilisées pour le redressement, ainsi que les distributions (en pourcentages) de ces variables, dans la population entière des ménages (la référence étant l'enquête-emploi de mars 1989) et dans l'échantillon (comprenant 6807 ménages).

Les "poids de sondage" d_k sont tous égaux au coefficient d'extrapolation 21062416/6807), et tous les tableaux qui suivent concernent les rapports w_k/d_k, que l'on désignera plus simplement par "poids".

Les méthodes "bornées" conduisent ici aux valeurs suivantes :

$$L_{max} = 0.62 \qquad U_{min} = 1.85$$

Les résultats présentés dans les pages suivantes concernent les méthodes "linéaire", "raking ratio", "logit" et "linéaire tronquée" associées à des

couples (L,U) égaux ou légèrement différents de (L_{max}, U_{min}).

Les tableaux et graphiques des pages 20 à 22 décrivent les distributions des poids correspondant à chacune des méthodes. Il est particulièrement intéressant d'examiner les quantités suivantes :

. STD DEV = écart-type

. SKEWNESS = coefficient d'asymétrie

. QUANTILES = quantiles d'ordre 1%, 5%, 10%, 25% ... 99%, dont la médiane

. EXTREMES = les 5 plus petites et les 5 plus grandes valeurs

. RANGE = étendue

. HISTOGRAM = histogramme (horizontal)

Quelques commentaires

a) La méthode "linéaire" est celle qui donne le plus petit écart-type : cela résulte de sa définition même, plus précisément de l'expression de la fonction $G(x) = 1/2\ (x-1)^2$ (1 étant toujours la moyenne de la distribution). Par ailleurs, cette méthode conduit à la distribution qui ressemble le plus à une distibution normale.

b) La méthode "raking ratio" donne la distribution la plus dissymétrique (du côté des valeurs faibles, comme toutes les autres distributions). Son étendue est plus élevée que l'autre méthode non bornée ("linéaire"), en raison des forts poids (3.5), largement supérieurs ; en revanche les valeurs minimales sont plus élevées que celles de la méthode "linéaire".

c) La méthode "logit" (0.60,1.88) provoque une concentration de poids près des deux bornes. Si l'on examine individuellement les poids, on constate que tout se passe comme si, partant de la distribution non bornée" raking ratio", l'introduction d'une borne inférieure L ramenait tous les poids inférieurs à L à des valeurs très proches de L, mais en même temps les poids légèrement supérieurs à L sont également rapprochés de L. On observe un phénomène analogue autour de la borne supérieure.

La distribution associée à la méthode "logit" (0.62,1.85), d'étendue minimale, a les mêmes caractéristiques que la précédente, mais encore plus accentuées : ainsi, le 1er quartile vaut 0.620, contre 0.629 précédemment (avec L = 0.60) , et le quantile d'ordre 0.90 vaut 1.680, contre 1.602.

d) Les méthodes "linéaire tronquée" (0.60,1.86) et (0.62,1.85) conduisent à des distributions ressemblant aux précédentes, avec, partant de la distribution "linéaire", une concentration encore plus forte autour des bornes (L notamment), due au fait que la méthode "linéaire tronquée" permet aux poids d'atteindre effectivement ces bornes, alors que la méthode "logit" ne permet que de s'en approcher asymptotiquement.

Le 1er tableau de la page 23 donne la matrice des coefficients de corrélation linéaire entre les différentes distributions. Ces coefficients sont particulièrement élevés (de l'ordre de 0.99 voire plus) au sein de la famille des distributions "bornées". Le choix de l'une ou l'autre des méthodes de cette famille modifie donc peu la distribution des poids. Les deux méthodes non bornées donnent un coefficient de 0.98. Les coefficients les plus faibles sont obtenus pour les "croisements" entre une méthode bornée et une méthode non bornée, les corrélations demeurant toujours supérieures à 0.90.

Le 2ème tableau de la page 23 et le graphique de la page suivante permettent de visualiser, sous la forme de "courbes de régression" cette fois, les liaisons entre les différentes distributions. Le tableau a été construit de la façon suivante :

. on a classé les 6307 ménages selon les valeurs croissantes de la variable de poids de la méthode "linéaire"

. à partir de ce classement, on a défini 15 groupes d'effectifs (presque) égaux, par découpage en "tranches" de cette variable

. pour chacun de ces groupes, on a calculé la moyenne de différentes variables de pondération, ainsi que l'écart-type, la valeur minimale, la valeur maximale.

Le graphique de la page 24 représente les courbes : "moyenne dans un groupe d'une variable poids quelconque" en fonction de "moyenne dans le groupe de la variable poids linéaire" (la courbe représentative de la méthode "linéaire" est bien entendu une droite, qui serait la bissectrice si les échelles horizontale et verticale étaient identiques). Ce graphique permet de visualiser un certain nombre des commentaires faits précédemment.

Le choix de la méthode

Face à différents "jeux" de pondération, 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 (i.e. sa "bonne tête").

Jusqu'à présent, le critère choisi a toujours été le deuxième, l'idée étant qu'"il ne faut pas avoir de poids trop petits ou trop gros", même si ces poids extrêmes sont peu nombreux.

Dans le cas présent, c'est donc la méthode "logit" (0.62,1.85) qui a été retenue, plutôt que la méthode "linéaire tronquée" (0.62,1.85) encore plus "brutale" aux extrémités.

La méthode "raking ratio" a été rejetée en raison de ses poids

supérieurs à 3, bien que ceux-ci soient peu nombreux, et que le quantile d'ordre 0.90 soit égal à 1.39 (contre 1.68 avec "logit"), le quantile d'ordre 0.10 valant lui 0.60 (contre 0.62).

Quant à la méthode linéaire, ce sont ses poids très faibles (de l'ordre de 0.20) qui ont en particulier conduit à la rejeter.

Remarque :

Dans un premier temps, on avait réalisé le redressement en considérant les classes d'âge "70-79 ans" et "80 ans et plus" (classes regroupées par la suite). La méthode linéaire donnait alors 5 poids négatifs, qui s'expliquaient de la façon suivante : les "80 ans et plus" étant fortement sous-représentés dans l'échantillon, le redressement leur accordait un fort poids, et concommitamment diminuait tellement les poids de certaines catégories sur-représentées que certains ménages, cumulant toutes les sur-représentations, se sont vus affectés des poids négatifs (provoquant la désolation du statisticien ...). Ces poids négatifs ont disparu lorsque l'on a regroupé les deux modalités, car la sous-représentation de la classe d'âge "70 ans et plus", bien qu'importante, est moins forte relativement que celle de la classe "80 ans et plus".

3. L'enquête "Budgets de famille"

Il s'agit également d'une enquête auprès des ménages, réalisée en 1988-1989 en 8 vagues.

On trouvera à la page 25 la liste des variables (toutes qualitatives) utilisées pour le redressement, ainsi que les distributions (en pourcentages) de ces variables, dans la population entière des ménages (la référence étant l'enquête-emploi de mars 1989) et dans l'échantillon . Le choix de la méthode de redressement a été fait, pour des raisons de calendrier, à partir d'un demi-échantillon de 4516 ménages, correspondant aux 4 premières vagues, et les tableaux qui suivent portent sur ce demi-échantillon.

Les méthodes "bornées" conduisent ici aux valeurs suivantes :

$$L_{max} = 0.84 \qquad U_{min} = 1.55$$

Ces valeurs sont plus proches de 1 que dans l'exemple précédent car les contraintes sont moins fortes, ce que l'on peut constater au vu du tableau de la page 25.

On trouvera dans les pages 26 à 28 les distributions correspondants aux différentes méthodes. Comme dans le cas précédent, c'est la méthode "logit" d'étendue minimale qui a finalement été retenue.

CONCLUSION

Les méthodes de redressement d'un échantillon par les méthodes "logit" et "linéaire tronquée" proposées ici en alternative aux méthodes classiques de calage sur marges, "linéaire"(estimation par régression, ou REDRE) ou

"raking ratio" (alias RAS, alias IPF), présentent l'intérêt de conduire à des poids bornés inférieurement et supérieurement : cette propriété semble fondamentale pour les responsables d'enquête qui redoutent que des poids trop dispersés rendent "instables" certains des effectifs publiés.

Le choix de la méthode ne peut reposer sur un critère de précision des estimateurs, puisque 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.

REFERENCES

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

(2) Deville J.-C. et Sarndal C.-E. (1990). Calibration estimators and generalized raking techniques in survey sampling. Journées de méthodologie statistique INSEE, 13-14 mars 1991.

(3) Gourieroux C. (1981). Théorie des sondages. Economica.

(4) Lemel Y. (1976). Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondage. Annales de l'INSEE n°22-23, 272-282.

Enquête "Modes de vie" (1988-1989)

	Population (1)	Echantillon
Age du chef de ménage (AGECH)	%	%
1. moins de 30 ans	12.5	12.6
2. de 30 à 39 ans	20.4	24.7
3. de 40 à 49 ans	17.9	19.0
4. de 50 à 59 ans	16.1	17.8
5. de 60 à 69 ans	16.4	16.3
6. 70 ans et plus	16.6	9.7
CS du chef de ménage (PCSCH)		
1. agriculteurs exploitants	3.2	4.3
2. artisans, commerçants	5.6	6.1
3. cadres et professions intel. supérieures	8.3	9.3
4. professions intermédiaires	13.5	14.7
5. employés	10.9	11.4
6. ouvriers	21.7	24.1
7. retraités	28.6	24.5
8. autres personnes sans activité prof.	8.2	5.8
Taille du ménage (NBPERS)		
1. 1 personne	26.7	20.1
2. 2 personnes	30.2	28.0
3. 3 personnes	17.3	19.3
4. 4 personnes	15.8	19.8
5. 5 personnes	6.7	8.9
6. 6 personnes et plus	3.3	3.9
Strate (STR)		
1. commmune rurale dans canton ent. rural	10.7	13.9
2. commmune rurale dans canton part. urbain	14.2	20.4
3. unité urbaine de - de 20000 h	15.6	15.6
4. unité urbaine de 20000 à 100000 h	13.6	11.9
5. unité urbaine de + de 100000 h	28.2	24.7
6. agglomération parisienne (sauf Paris)	12.3	9.8
7. Paris	5.4	3.7
Type d'immeuble (TYPIM1)		
1. maison individuelle	57.0	62.8
2. immeuble	43.0	37.2
ZEAT		
1. Ile-de-France (1)	20.1	16.5
2. Bassin parisien (2)	17.8	18.6
3. Nord (3)	6.8	6.9
4. Est (4)	8.9	9.3
5. Ouest (5)	13.2	14.6
6. Sud-Ouest (7)	10.2	10.7
7. Centre-Est (8)	11.5	11.6
8. Méditerranée (9)	11.6	11.8
Vague (VA)		
1	12.5	12.2
2	12.5	12.1
3	12.5	12.3
4	12.5	13.1
5	12.5	12.8
6	12.5	11.4
7	12.5	12.4
8	12.5	13.7

(1) Enquête emploi de mars 1989

UNIVARIATE

VARIABLE=POIDS1

MOMENTS

N	6807	SUM WGTS	6807
MEAN	1	SUM	6807
STD DEV	0.365218	VARIANCE	0.133384
SKEWNESS	0.831444	KURTOSIS	1.10685
USS	7714.81	CSS	907.811
CV	36.5218	STD MEAN	0.00442664
T:MEAN=0	225.905	PROB>|T|	0.0001
SGN RANK	11585514	PROB>|S|	0.0001
NUM ^= 0	6807		
D:NORMAL	0.0588803	PROB>D	<.01

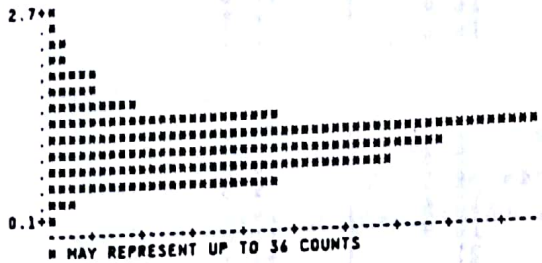
QUANTILES(DEF=4)

100%	MAX	2.63908	99%	2.09905
75%	Q3	1.18294	95%	1.74018
50%	MED	0.985476	90%	1.43255
25%	Q1	0.728764	10%	0.558965
0%	MIN	0.191383	5%	0.487261
			1%	0.382507
RANGE		2.44769		
Q3-Q1		0.454171		
MODE		0.672905		

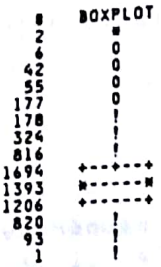
EXTREMES

LOWEST	HIGHEST
0.191383	2.51205
0.235303	2.51205
0.235303	2.57464
0.235303	2.57678
0.241279	2.62085
	2.63908

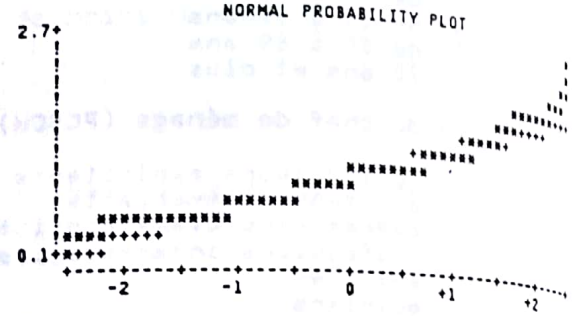
HISTOGRAM



BOXPLOT



NORMAL PROBABILITY PLOT



Linéaire

UNIVARIATE

VARIABLE=POIDS2

MOMENTS

N	6807	SUM WGTS	6807
MEAN	1	SUM	6807
STD DEV	0.373224	VARIANCE	0.139296
SKEWNESS	1.70774	KURTOSIS	4.95925
USS	7755.05	CSS	948.051
CV	37.3224	STD MEAN	0.00452368
T:MEAN=0	221.059	PROB>|T|	0.0001
SGN RANK	11585514	PROB>|S|	0.0001
NUM ^= 0	6807		
D:NORMAL	0.0925331	PROB>D	<.01

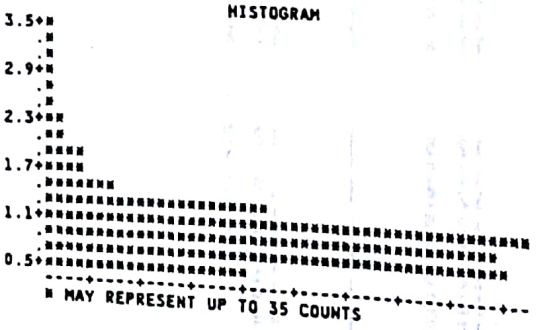
QUANTILES(DEF=4)

100%	MAX	3.53443	99%	2.36332
75%	Q3	1.15188	95%	1.73205
50%	MED	0.958486	90%	1.39199
25%	Q1	0.72957	10%	0.601078
0%	MIN	0.418489	5%	0.556576
			1%	0.497958
RANGE		3.11594		
Q3-Q1		0.422308		
MODE		0.691547		

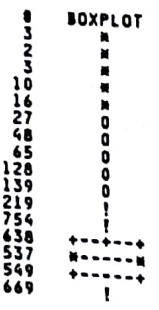
EXTREMES

LOWEST	HIGHEST
0.418489	3.3063
0.438279	3.32822
0.43994	3.47768
0.442964	3.50425
0.444025	3.53443

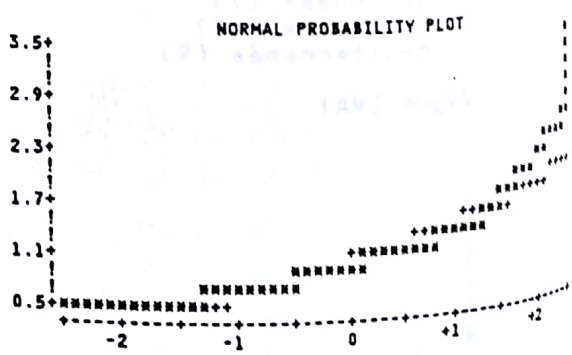
HISTOGRAM



BOXPLOT



NORMAL PROBABILITY PLOT



Raking ratio

SAS

UNIVARIATE

10:06 MONDAY, NOVEMBER 19, 1990

VARIABLE=POIDS3

MOMENTS

N	6807	SUM WGTS	6807
MEAN	1	SUM	6807
STD DEV	0.382799	VARIANCE	0.146535
SKEWNESS	0.829094	KURTOSIS	-0.31343
USS	7804.32	CSS	997.316
CV	38.2799	STD MEAN	0.00463973
T:MEAN=0	215.53	PROB>|T|	0.0001
SGN RANK	11585514	PROB>|S|	0.0001
NUM ^= 0	6807		
W:NORMAL	0.14821	PROB<W	<.01

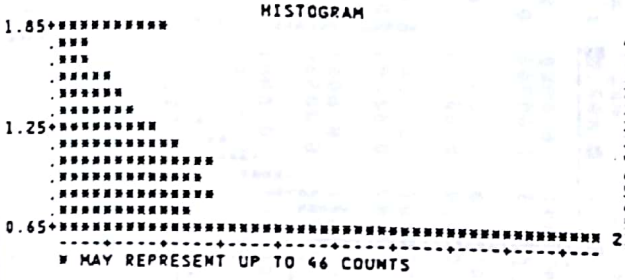
QUANTILES(DEF=4)

100% MAX	1.87993	99%	1.87772
75% Q3	1.22714	95%	1.86216
50% MED	0.915713	90%	1.6021
25% Q1	0.628767	10%	0.606514
0% MIN	0.600304	5%	0.603812
		1%	0.601734
RANGE	1.27963		
Q3-Q1	0.598371		
MODE	0.61007		

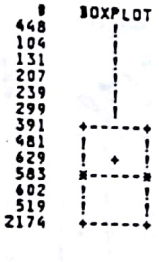
EXTREMES

LOWEST	HIGHEST
0.600304	1.8794
0.600449	1.8796
0.600452	1.8796
0.600452	1.8796
0.600452	1.8796

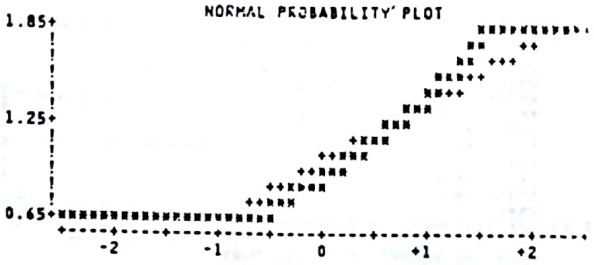
HISTOGRAM



BOXPLOT



NORMAL PROBABILITY PLOT



Logit

0.60 1.88

SAS

UNIVARIATE

10:06 MONDAY, NOVEMBER 19, 1990

VARIABLE=POIDS3

MOMENTS

N	6807	SUM WGTS	6807
MEAN	1	SUM	6807
STD DEV	0.393332	VARIANCE	0.15471
SKEWNESS	0.834845	KURTOSIS	-0.466329
USS	7859.96	CSS	1052.96
CV	39.3332	STD MEAN	0.0047674
T:MEAN=0	209.758	PROB>|T|	0.0001
SGN RANK	11585514	PROB>|S|	0.0001
NUM ^= 0	6807		
W:NORMAL	0.166996	PROB<W	<.01

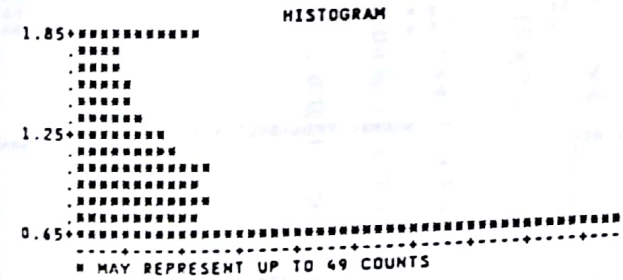
QUANTILES(DEF=4)

100% MAX	1.85	99%	1.85
75% Q3	1.23638	95%	1.84998
50% MED	0.893747	90%	1.68055
25% Q1	0.620073	10%	0.620006
0% MIN	0.62	5%	0.620003
		1%	0.620001
RANGE	1.23		
Q3-Q1	0.61631		
MODE	1.85		

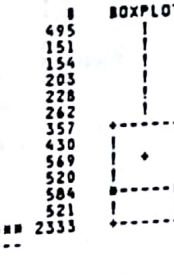
EXTREMES

LOWEST	HIGHE
0.62	1.
0.62	1.
0.62	1.
0.62	1.
0.62	1.

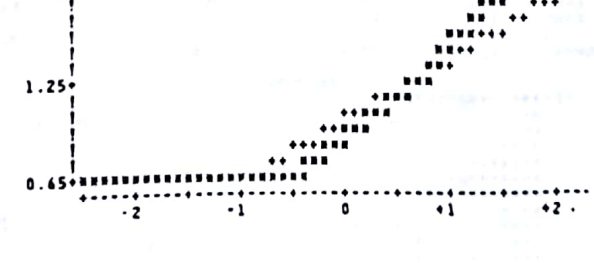
HISTOGRAM



BOXPLOT



NORMAL PROBABILITY PLOT



Logit

0.62 1.85

O. Sautory, Insee Méthodes n° 29-30-31

SAS 10:06 MONDAY, NOVEMBER 19, 1990

UNIVARIATE

VARIABLE=POIDS4

MOMENTS

N	6807	SUM WGTS	6807
MEAN	1	SUM	6807.03
STD DEV	0.391476	VARIANCE	0.153253
SKEWNESS	0.802361	KURTOSIS	-0.408746
USS	7850.11	CSS	1043.04
CV	39.1474	STD MEAN	0.0047449
T:MEAN=0	210.754	PROB>|T|	0.0001
SGN RANK	11585514	PROB>|S|	0.0001
NUM ^= 0	6807		
D:NORMAL	0.177621	PROB>D	<.01

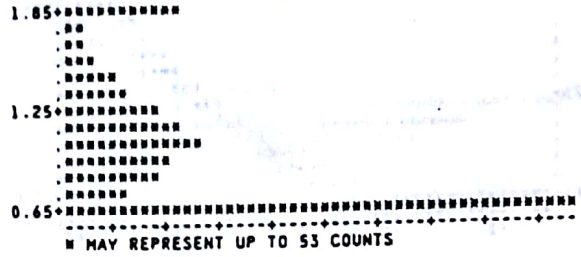
QUANTILES(DEF=4)

100% MAX	1.85	99%	1.85
75% Q3	1.2289	95%	1.85
50% MED	0.931972	90%	1.65024
25% Q1	0.62	10%	0.62
0% MIN	0.62	5%	0.62
		1%	0.62
RANGE	1.23		
Q3-Q1	0.608895		
MODE	0.62		

EXTREMES

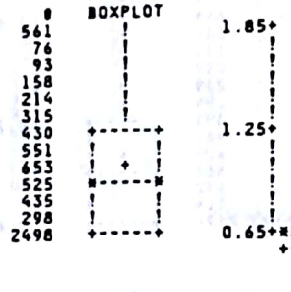
LOWEST	HIGHEST
0.62	1.85
0.62	1.85
0.62	1.85
0.62	1.85
0.62	1.85

HISTOGRAM

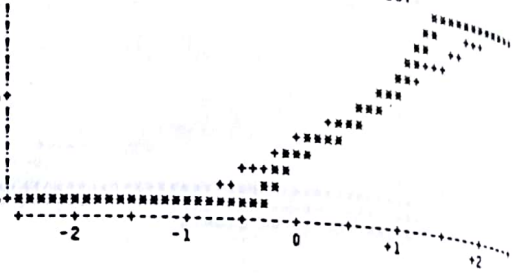


MAY REPRESENT UP TO 53 COUNTS

BOXPLOT



NORMAL PROBABILITY PLOT



Linéaire tronquée

0.62 1.85

SAS 10:06 MONDAY, NOVEMBER 19, 1990 8

UNIVARIATE

VARIABLE=POIDS4

MOMENTS

N	6807	SUM WGTS	6807
MEAN	1	SUM	6807
STD DEV	0.380091	VARIANCE	0.144469
SKEWNESS	0.738731	KURTOSIS	-0.329927
USS	7790.26	CSS	983.258
CV	38.0091	STD MEAN	0.00460691
T:MEAN=0	217.065	PROB>|T|	0.0001
SGN RANK	11585514	PROB>|S|	0.0001
NUM ^= 0	6807		
D:NORMAL	0.146313	PROB>D	<.01

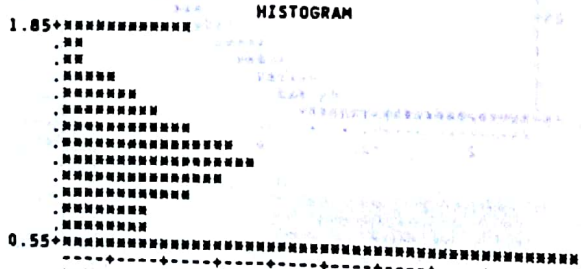
QUANTILES(DEF=4)

100% MAX	1.86	99%	1.86
75% Q3	1.22162	95%	1.86
50% MED	0.958475	90%	1.56342
25% Q1	0.6	10%	0.6
0% MIN	0.6	5%	0.6
		1%	0.6
RANGE	1.26		
Q3-Q1	0.621616		
MODE	0.6		

EXTREMES

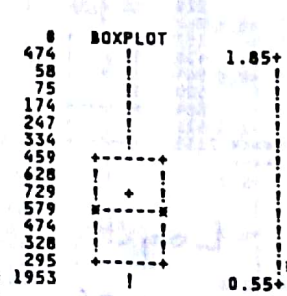
LOWEST	HIGHEST
0.6	1.86
0.6	1.86
0.6	1.86
0.6	1.86
0.6	1.86

HISTOGRAM

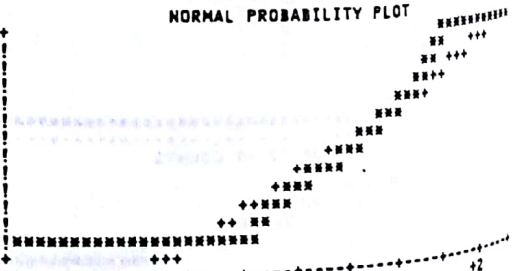


MAY REPRESENT UP TO 41 COUNTS

BOXPLOT



NORMAL PROBABILITY PLOT



Linéaire tronquée

0.60 1.86

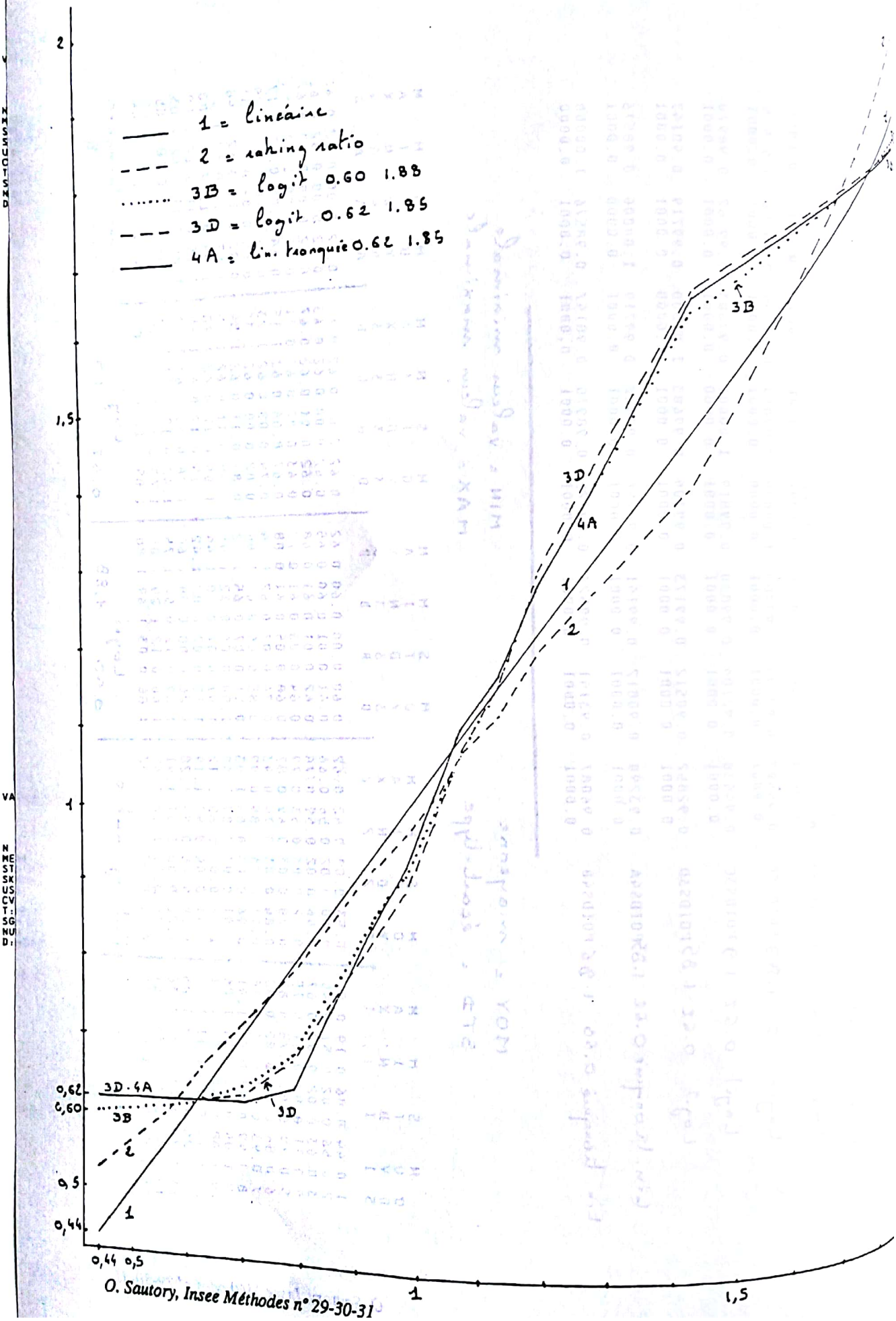
PEARSON CORRELATION COEFFICIENTS / PROB > |R| UNDER H0:RHO=0 / N = 6337

		POIDS1	POIDS2	POIDS3A	POIDS3B	POIDS3C	POIDS3D	POIDS4A	POIDS4B
Linéaire	POIDS1	1.00000	0.97853	0.95643	0.95407	0.94438	0.92852	0.93290	0.96087
		0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Rahing ratio	POIDS2	0.97853	1.00000	0.93491	0.92999	0.92704	0.90512	0.90812	0.93141
		0.0001	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Logit 0.6 1.9	POIDS3A	0.95643	0.93491	1.00000	0.99980	0.99830	0.99173	0.99141	0.99523
		0.0001	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001
Logit 0.6 1.88	POIDS3B	0.95407	0.92999	0.99980	1.00000	0.99819	0.99304	0.99256	0.99525
		0.0001	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001
Logit 0.62 1.9	POIDS3C	0.94438	0.92704	0.99830	0.99819	1.00000	0.99603	0.99462	0.98970
		0.0001	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001
Logit 0.62 1.85	POIDS3D	0.92852	0.90512	0.99173	0.99304	0.99603	1.00000	0.99719	0.98147
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001
Lin. tronquée 0.62 1.85	POIDS4A	0.93290	0.90812	0.99141	0.99256	0.99462	0.99719	1.00000	0.98576
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001
Lin. tronquée 0.60 1.85	POIDS4B	0.96087	0.93141	0.99523	0.99525	0.98970	0.98147	0.98576	1.00000
		0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000

MOY = moyenne
STD = écart-type

MIN = valeur minimale
MAX = valeur maximale

OBS	MOY1	STD1	MIN1	MAX1	MOY2	STD2	MIN2	MAX2	MOY3B	STD3B	MIN3B	MAX3B	MOY3D	STD3D	MIN3D	MAX3D	MOY4A	STD4A	MIN4A	MAX4A
1	0.44	0.06	0.19	0.51	0.53	0.03	0.42	0.62	0.60	0.00	0.60	0.62	0.62	0.00	0.62	0.62	0.62	0.00	0.62	0.62
2	0.56	0.02	0.51	0.60	0.60	0.02	0.56	0.66	0.61	0.01	0.60	0.65	0.62	0.00	0.62	0.65	0.62	0.00	0.62	0.62
3	0.63	0.02	0.60	0.67	0.66	0.02	0.59	0.74	0.62	0.01	0.61	0.68	0.62	0.01	0.62	0.67	0.62	0.00	0.62	0.62
4	0.71	0.02	0.67	0.75	0.72	0.03	0.63	0.79	0.64	0.03	0.61	0.72	0.63	0.02	0.62	0.71	0.62	0.00	0.62	0.63
5	0.79	0.02	0.75	0.82	0.78	0.03	0.69	0.87	0.68	0.05	0.61	0.88	0.67	0.06	0.62	1.15	0.64	0.05	0.62	1.17
6	0.86	0.02	0.82	0.90	0.84	0.03	0.74	0.93	0.76	0.07	0.62	0.94	0.74	0.08	0.62	1.23	0.73	0.08	0.62	1.24
7	0.93	0.02	0.90	0.96	0.91	0.03	0.80	0.99	0.85	0.06	0.64	1.04	0.82	0.08	0.62	1.36	0.84	0.09	0.62	1.33
8	0.99	0.02	0.96	1.01	0.96	0.03	0.83	1.03	0.91	0.07	0.65	1.22	0.89	0.11	0.62	1.54	0.92	0.11	0.62	1.48
9	1.04	0.01	1.01	1.06	1.01	0.03	0.89	1.08	1.00	0.06	0.72	1.46	0.99	0.11	0.62	1.79	1.02	0.11	0.62	1.83
10	1.08	0.01	1.06	1.11	1.06	0.03	0.88	1.12	1.07	0.07	0.70	1.44	1.06	0.12	0.62	1.78	1.09	0.11	0.62	1.80
11	1.14	0.02	1.11	1.17	1.11	0.04	0.92	1.19	1.15	0.07	0.82	1.42	1.15	0.12	0.62	1.65	1.16	0.10	0.62	1.59
12	1.20	0.02	1.17	1.25	1.18	0.05	0.94	1.28	1.26	0.08	0.88	1.64	1.28	0.13	0.62	1.82	1.26	0.11	0.62	1.85
13	1.30	0.03	1.25	1.36	1.26	0.08	1.01	1.41	1.39	0.12	0.97	1.75	1.42	0.19	0.62	1.84	1.39	0.17	0.62	1.85
14	1.45	0.07	1.36	1.61	1.38	0.11	1.12	1.68	1.60	0.12	1.29	1.85	1.63	0.14	1.23	1.85	1.62	0.17	1.26	1.85
15	1.89	0.21	1.61	2.64	2.00	0.37	1.43	3.53	1.86	0.03	1.69	1.88	1.85	0.02	1.69	1.85	1.85	0.01	1.71	1.85
	Linéaire				Rahing ratio				Logit 0.60 1.88				Logit 0.62 1.85				Linéaire tronquée 0.62 1.85			



O. Sautory, Insee Méthodes n° 29-30-31

Enquête "Budgets de famille" (1988-1989)

	Echantillon (1)		Population
Age du chef de ménage (TRAGECH)	Eff.	%	%
1. moins de 35 ans	1017	22.5	22.5
2. de 35 à 54 ans	1740	38.5	36.3
3. de 55 à 74 ans	1293	28.6	28.4
4. 75 ans et plus	466	10.3	12.8
CS du chef de ménage (PCSC2CH1)			
1. agriculteurs exploitants	152	3.4	3.2
2. artisans, commerçants	229	5.1	5.6
3. cadres et professions intell. supérieures	415	9.2	8.4
4. professions intermédiaires	634	14.0	13.5
5. employés	506	11.2	10.9
6. ouvriers	1053	23.3	21.7
7. retraités	1291	28.6	28.6
8. autres personnes sans activité prof.	236	5.2	8.1
Taille du ménage (NBPERS)			
1. 1 personne	972	21.5	26.6
2. 2 personnes	1363	30.2	30.1
3. 3 personnes	861	19.1	17.4
4. 4 personnes	818	18.1	15.9
5. 5 personnes	353	7.8	6.7
6. 6 personnes et plus	149	3.3	3.3
Catégorie de commune (CC5)			
1. commune rurale	1231	27.3	25.0
2. unité urbaine de - de 20000 h	780	17.3	15.6
3. unité urbaine de 20000 à 100000 h	591	13.1	13.6
4. unité urbaine de + de 100000 h	1266	28.0	28.1
5. agglomération parisienne (sauf Paris)	459	10.2	12.3
6. Paris	189	4.2	5.4
Vague (VA)			
1	1187	26.3	25.0
2	1172	26.0	25.0
3	1059	23.4	25.0
4	1098	24.3	25.0

(1) Enquête emploi de mars 1989

UNIVARIATE

VARIABLE=POIDS1

MOMENTS

N	4516	SUM WGTS	4516
MEAN	1	SUM	4516
STD DEV	0.218118	VARIANCE	0.0475753
SKEWNESS	1.26457	KURTOSIS	2.24586
USS	4730.8	CSS	214.803
CV	21.8118	STD MEAN	0.00324574
T:MEAN=0	308.096	PROB>|T|	0.0001
SGN RANK	5099693	PROB>|S|	0.0001
NUM ^= 0	4516		
D:NORMAL	0.0943722	PROB>D	<.01

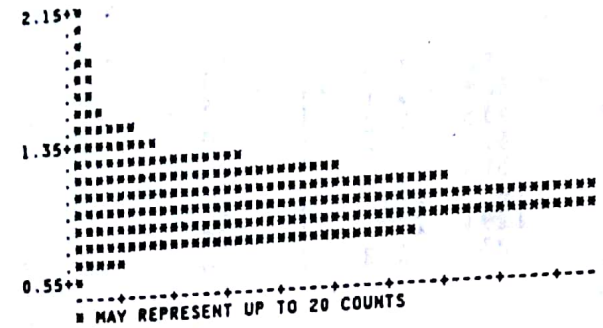
QUANTILES(DEF=4)

100%	MAX	2.16942	99%	1.75941
75%	Q3	1.1126	95%	1.4137
50%	MED	0.94863	90%	1.275
25%	Q1	0.846134	10%	0.771375
0%	MIN	0.583479	5%	0.731327
			1%	0.67897
RANGE		1.58594		
Q3-Q1		0.266471		
MODE		0.909822		

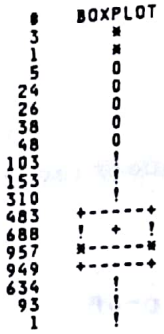
EXTREMES

LOWEST	HIGHEST
0.583479	1.9
0.608568	1.9
0.608568	2.0
0.608568	2.1
0.608568	2.1

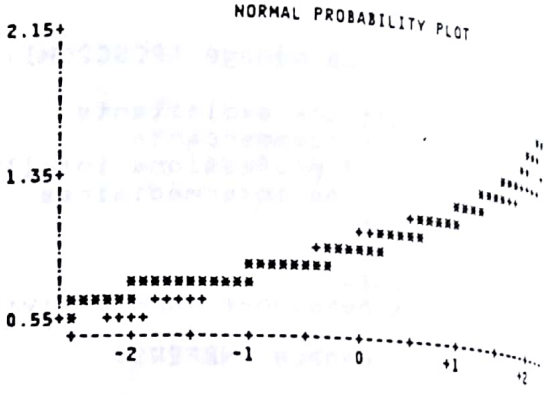
HISTOGRAM



BOXPLOT



NORMAL PROBABILITY PLOT



Linéaire

UNIVARIATE

VARIABLE=POIDS2

MOMENTS

N	4516	SUM WGTS	4516
MEAN	1	SUM	4516
STD DEV	0.220035	VARIANCE	0.0484156
SKEWNESS	1.84881	KURTOSIS	5.59799
USS	4734.6	CSS	218.596
CV	22.0035	STD MEAN	0.00327428
T:MEAN=0	305.411	PROB>|T|	0.0001
SGN RANK	5099693	PROB>|S|	0.0001
NUM ^= 0	4516		
D:NORMAL	0.111804	PROB>D	<.01

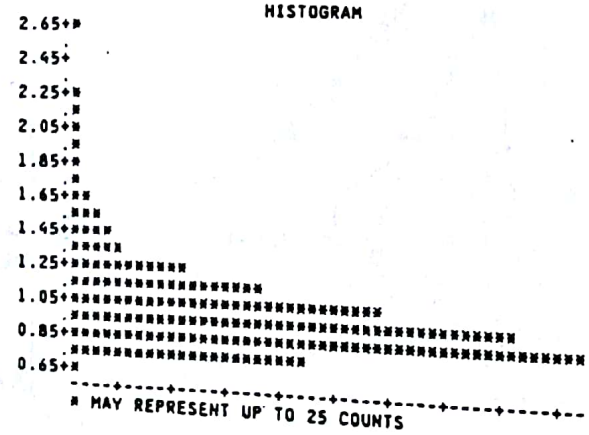
QUANTILES(DEF=4)

100%	MAX	2.65504	99%	1.83689
75%	Q3	1.09269	95%	1.42772
50%	MED	0.944807	90%	1.25878
25%	Q1	0.850378	10%	0.786873
0%	MIN	0.648483	5%	0.753235
			1%	0.715682
RANGE		2.00656		
Q3-Q1		0.242312		
MODE		0.912562		

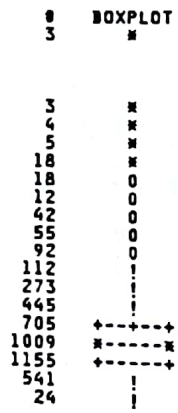
EXTREMES

LOWEST	HIGHEST
0.648483	2.2
0.663426	2.2
0.663426	2.2
0.663426	2.2
0.663426	2.3

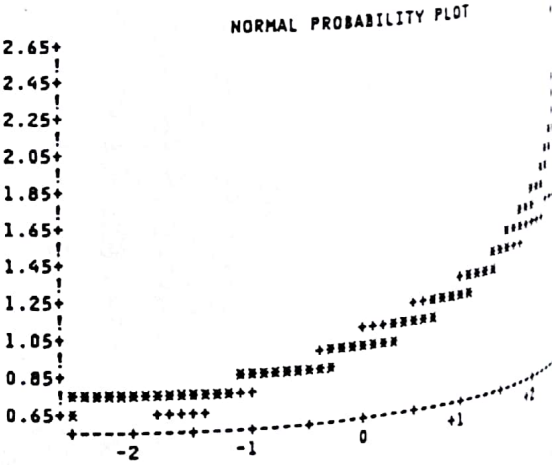
HISTOGRAM



BOXPLOT



NORMAL PROBABILITY PLOT



UNIVARIATE

VARIABLE=POIDS3

MOMENTS

N	4516	SUM WGTS	4516
MEAN	1	SUM	4516
STD DEV	0.250922	VARIANCE	0.062962
SKEWNESS	1.34134	KURTOSIS	0.125518
USS	4800.27	CSS	284.273
CV	25.0922	STD MEAN	0.0037339
T:MEAN=0	267.817	PROB>|T|	0.0001
SGN RANK	5099693	PROB>|S|	0.0001
NUM ^= 0	4516		
D:NORMAL	0.302993	PROB>D	<.01

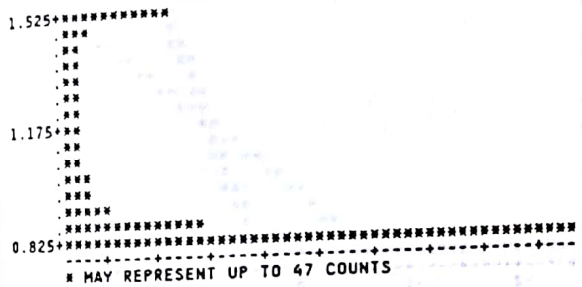
QUANTILES(DEF=4)

100% MAX	1.55	99%	1.55
75% Q3	1.05247	95%	1.55
50% MED	0.850576	90%	1.51345
25% Q1	0.840409	10%	0.840023
0% MIN	0.84	5%	0.840006
		1%	0.84
RANGE	0.71		
Q3-Q1	0.212057		
MODE	1.55		

EXTREMES

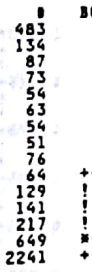
LOWEST	HIGHEST
0.84	1.55
0.84	1.55
0.84	1.55
0.84	1.55
0.84	1.55

HISTOGRAM

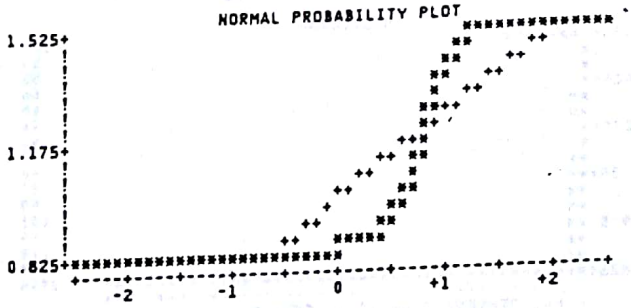


* MAY REPRESENT UP TO 47 COUNTS

BOXPLOT



NORMAL PROBABILITY PLOT



Logit

0.84 1.55

17:06 MONDAY, APRIL 2, 1990 90

SAS

UNIVARIATE

VARIABLE=POIDS3

MOMENTS

N	4516	SUM WGTS	4516
MEAN	1	SUM	4516
STD DEV	0.224172	VARIANCE	0.050253
SKEWNESS	1.25242	KURTOSIS	0.694177
USS	4742.89	CSS	226.892
CV	22.4172	STD MEAN	0.00333583
T:MEAN=0	299.775	PROB>|T|	0.0001
SGN RANK	5099693	PROB>|S|	0.0001
NUM ^= 0	4516		
D:NORMAL	0.157125	PROB>D	<.01

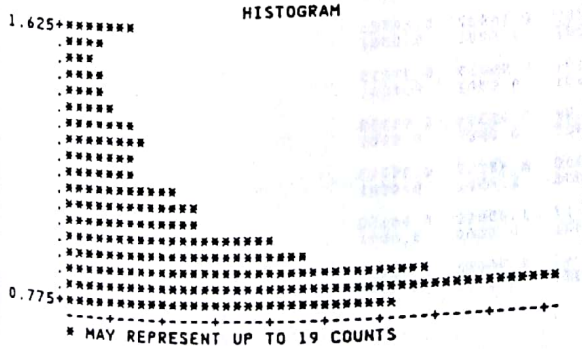
QUANTILES(DEF=4)

100% MAX	1.64929	99%	1.63844
75% Q3	1.10493	95%	1.51664
50% MED	0.91419	90%	1.34743
25% Q1	0.829354	10%	0.796053
0% MIN	0.759051	5%	0.781939
		1%	0.771019
RANGE	0.890241		
Q3-Q1	0.275575		
MODE	0.876053		

EXTREMES

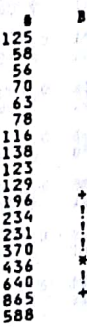
LOWEST	HIGHEST
0.759051	1.64774
0.761269	1.64774
0.761269	1.6492
0.761269	1.6492
0.761269	1.64929

HISTOGRAM

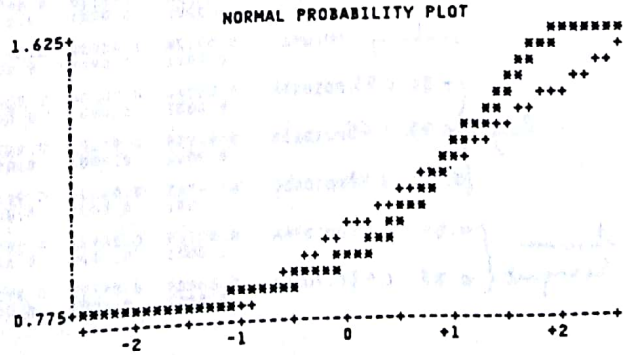


* MAY REPRESENT UP TO 19 COUNTS

BOXPLOT



NORMAL PROBABILITY PLOT



Logit

0.75 1.65

O. Sautory, *Insee Méthodes n° 29-30-31*

UNIVARIATE

VARIABLE=POIDS4

MOMENTS

N	4516	SUM WGTS	4516
MEAN	0.256945	SUM	4516
STD DEV	1.4289	VARIANCE	0.0660208
SKEWNESS	4814.08	KURTOSIS	0.362893
USS	25.6945	CSS	298.084
CV	261.539	STD MEAN	0.00382352
T:MEAN=0	5099693	PROB>|T|	0.0001
SGN RANK	4516	PROB>|S|	0.0001
NUM ^= 0	4516		
D:NORMAL	0.383293	PROB>D	<.01

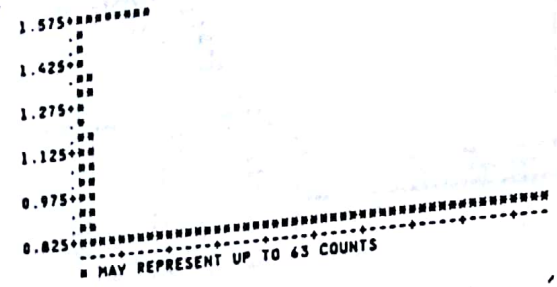
QUANTILES(DEF=4)

100% MAX	1.57	99%	1.57
75% Q3	1.05182	95%	1.57
50% MED	0.85	90%	1.57
25% Q1	0.85	10%	0.85
0% MIN	0.72	5%	0.85
RANGE	0.201824	1%	0.85
Q3-Q1	0.85		
MODE			

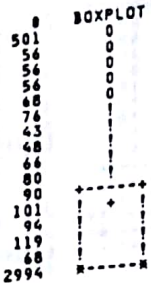
EXTREMES

LOWEST	HIGHEST
0.85	1.57
0.85	1.57
0.85	1.57
0.85	1.57
0.85	1.57

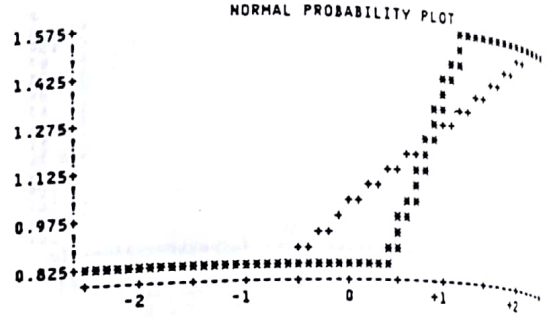
HISTOGRAM



BOXPLOT



NORMAL PROBABILITY PLOT



Linéaire tronquée
0.85 1.57

SAS 17:06 MONDAY, APRIL 2, 1990 3

VARIABLE	N	MEAN	STD DEV	SUM	MINIMUM	MAXIMUM
POIDS1	4516	1.00000000	0.21811764	4516.00000000	0.58347881	2.16942107
POIDS2	4516	1.00000000	0.22003540	4516.00000028	0.64848280	2.65504398
POIDS3A	4516	1.00000000	0.25092231	4515.99999060	0.84000000	1.55000000
POIDS3B	4516	1.00000000	0.22417185	4516.00000000	0.75905126	1.64929212
POIDS3C	4516	1.00000000	0.25858828	4516.00000000	0.85000000	1.57000000
POIDS4A	4516	1.00000000	0.24866764	4516.00000000	0.84000000	1.55000000
POIDS4B	4516	1.00000000	0.25694505	4516.00000000	0.85000000	1.57000000

PEARSON CORRELATION COEFFICIENTS / PROB > !R! UNDER HO:RHO=0 / N = 4516

		POIDS1	POIDS2	POIDS3A	POIDS3B	POIDS3C	POIDS4A	POIDS4B
linéaire	POIDS1	1.00000 0.0000	0.99128 0.0001	0.86926 0.0001	0.97299 0.0001	0.84349 0.0001	0.87715 0.0001	0.84889 0.0001
cubique	POIDS2	0.99128 0.0001	1.00000 0.0000	0.86204 0.0001	0.96071 0.0001	0.84205 0.0001	0.86932 0.0001	0.84723 0.0001
logit 0.84 1.55	POIDS3A	0.86926 0.0001	0.86204 0.0001	1.00000 0.0000	0.94736 0.0001	0.99101 0.0001	0.99619 0.0001	0.99218 0.0001
logit 0.75 1.65	POIDS3B	0.97299 0.0001	0.96071 0.0001	0.94736 0.0001	1.00000 0.0000	0.92608 0.0001	0.95398 0.0001	0.93138 0.0001
logit 0.85 1.57	POIDS3C	0.84349 0.0001	0.84205 0.0001	0.99101 0.0001	0.92608 0.0001	1.00000 0.0000	0.98117 0.0001	0.99743 0.0001
linéaire tronquée 0.84 1.55	POIDS4A	0.87715 0.0001	0.86932 0.0001	0.99619 0.0001	0.95398 0.0001	0.98117 0.0001	1.00000 0.0000	0.98652 0.0001
linéaire tronquée 0.85 1.57	POIDS4B	0.84889 0.0001	0.84723 0.0001	0.99218 0.0001	0.93138 0.0001	0.99743 0.0001	0.98652 0.0001	1.00000 0.0000