Calage et redressement de la non réponse totale :

Validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enqête sur la consommation alimentaire de 1989

F. Dupont

La pratique courante du ***redressement*** *des enquêtes réalisées auprès des ménages par l'INSEE, consiste à caler la structure des répondants sur la structure de la population française connue à la même période pour des variables qualitatives* $x_1 \ldots x_k$ *du type âge, sexe, CS, catégorie d'agglomération, nombre de personnes dans le ménage (1).*

Cette opération préalable à toute exploitation de l'enquête vise simultanément à éliminer les déformations de structure dues à un comportement de ***non-réponse*** *non uniforme, et à améliorer les performances des estimations futures tirées de l'enquête en amoindrissant les effets de l'****erreur d'échantillonnage****. On intègre à cet effet, la connaissance, sans aléa, de statistiques de même nature (totaux) qui portent sur des variables* $x_1, \ldots, x_k$ *. Lorsque les variables* $x_1, \ldots, x_k$ *sont corrélées aux variables d'enquête, cette opération améliore l'estimation des variables d'enquêtes. On suppose également que lorsque les variables* $x_1, \ldots, x_k$ *expliquent entièrement les disparités dans le mécanisme de réponse, le redressement par calage corrige les biais induit par les déformations de structure dues au comportement de* ***non-réponse*** *non uniforme.*

Or, à l'heure actuelle, la justification théorique rigoureuse de cette méthode appelée ***méthode n°1*** *dans la suite, n'est acquise que lorsque le comportement de réponse est uniforme.*

Une démarche alternative naturelle, appelée ***méthode n°2*** *dans la suite, correcte sur le plan théorique, consiste à corriger la forme des estimateurs tirés de l'enquête pour tenir compte de la non-réponse dans une* ***première étape****, et à améliorer la performance des estimations dans un* ***deuxième temps****, c'est-à-dire à traiter l'erreur d'échantillonnage.*
La première étape requiert une modélisation du comportement de non-réponse pour laquelle les modèles économétriques offrent un cadre général permettant d'utiliser l'information auxiliaire directement sous forme qualitative ou quantitative.
La deuxième étape consiste à caler l'échantillon sur une structure externe considérée connue sans aléa, après avoir divisé les poids de sondage de chaque individu par les probabilités de non-réponse données par la première étape.

Cette deuxième méthode est plus lourde à mettre en oeuvre (2) et requiert plus d'information que la première. Elle nécessite en effet de connaître pour l'ensemble des individus tirés la

valeur du groupe de variables utilisé pour estimer le modèle de non-réponse. La première méthode, en revanche ne requiert aucune information au niveau individuel ou agrégé sur les non-répondants.

En pratique, les deux méthodes conduisent à élaborer de nouvelles pondérations pour les individus qui remplaceront l'inverse des poids de sondage dans l'estimation de statistiques tirées de l'enquête.

Le but de cette étude est de déterminer le domaine de ***validité de la pratique courante,*** *et d'étudier les* ***différences entre les résultats obtenus par les deux méthodes*** *dans un cas concret.*
partie I : *On montre dans cette étude que les deux méthodes* ***coincident*** *exactement dans un cas particulier. Lorsque la fonction utilisée dans la procédure de calage et la forme fonctionnelle du modèle de réponse sont* ***exponentielles, ou*** *lorsque la non-réponse ne dépend que* ***d'une seule variable qualitative*** *qui est prise en compte dans le calage, les deux méthodes sont équivalentes. Dans ce dernier cas, les deux méthodes se confondent avec une* ***postratification.***

partie II : *On étudie ensuite l'ampleur de l'écart entre les deux méthodes lorsqu'elles ne coincident pas exactement pour une enquête réalisée par l'INSEE :* ***l'enquête sur la consommation alimentaire de 1989.***

NOTES :

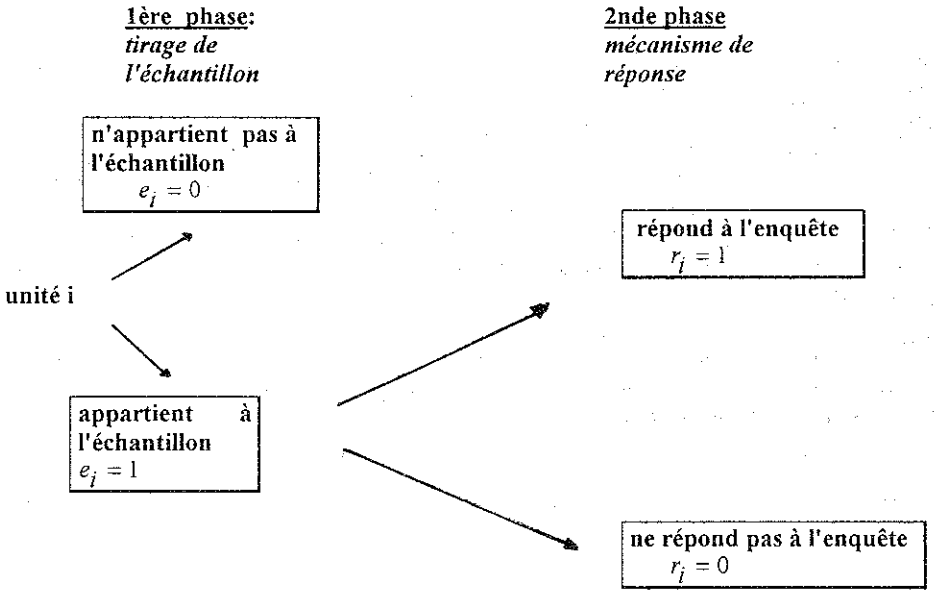
(1) La structure de référence utilisée pour le calage est issue, sauf proximité avec le recensement,de l'enquête emploi de mars de la même année, qui fait alors office de mini-recensement. La taille de l'echantillon de l'enquête emploi ne permet pas d'obtenir des structures croisées stables, on se limite donc au calage sur les distributions marginales pour un vecteur de variables du type: âge, csp, catégorie d'agglomération, réalisé par le logiciel CALMAR.

(2) En effet la méthode n°2 requiert l'estimation par le maximum de vraisemblance du modèle de non-réponse (PROC LOGISTIC) et, surtout, dans l'organisation actuelle du stockage des résultats d'enquête, l'appariement du fichier de saisie de l'enquête et du fichier de sondage constitué au moment du tirage contenant les variables x1...xk et les poids de sondage. Lorsque les pondérations sont simples cet appariement n'est pas nécessaire pour appliquer la méthode n°1.

1-cadre théorique et présentation des deux méthodes de redressement alternatives :

1-1 cadre probabiliste :

Le cadre probabiliste utilisé pour modéliser les aléas d'échantillonnage et du processus de réponse est le suivant:



On note :

$T_i = P(e_i = 1)$ probabilité d'inclusion de l'unité d'échantillonnage i

$p_i = P(r_i = 1)$ probabilité de réponse de l'unité i

U : population cible

s : ensemble des n unités échantillonnées

r : ensemble des m répondants

L'unité échantillonnée est le **logement**, mais l'unité d'observation est le **ménage.** Les **pondérations** seront donc in fine relatives au **ménage.** Les deux méthodes de redressement étudiées consistent à modifier les poids initiaux associés à chaque ménage et découlant du plan de sondage selon deux stratégies utilisant toutes les deux de l'information auxiliaire.

1-2 calage simple en l'absence de non-réponse:

On supposera dans la suite que l'on s'est ramené, ce qui est toujours possible, au cas où les variables qualitatives sont déjà sous la forme de variables indicatrices de modalités.

On dispose d'information auxiliaire sous la forme de k variables $x_1 \ldots x_k$ quantitatives ou qualitatives :

-au niveau individuel sur l'ensemble de l'échantillon interrogé,

-au niveau de la population, sous la forme du vecteur des totaux $X = (X_1, \ldots X_k) = \sum_{i \in U} x_i = (\sum_{i \in U} x_{i1}, \ldots, \sum_{i \in U} x_{ik})$,supposé connu sans aléa où $x_i = (x_{i1}, \ldots x_{ik})$.

On souhaite estimer le total sur la population U d'une variable y. L'estimateur naturel de ce total $Y = \sum_{i \in U} y_i$ en l'absence d'information auxiliaire est donné par $\hat{Y} = \sum_{i \in s} \frac{y_i}{T_i}$, estimateur d'**Horvitz-Thompson**, qui estime sans biais le total Y. Or l'aléa d'échantillonnage se traduit par le fait que $\hat{Y} = \sum_{i \in s} \frac{y_i}{T_i}$ va varier dans un intervalle centré en Y de longueur proportionnelle à son écart-type. De la même façon $\hat{X} = \sum_{i \in s} \frac{x_i}{T_i}$ va varier dans un intervalle centré en X de longueur proportionnelle à son écart-type . Or on connaît avec certitude le total X. Si la variable, (ou plus généralement le vecteur de variables), y est corrélée au vecteur de variables x dans la population, une partie du gain de précision qu'il y a entre, estimer le total X par $\sum_{i \in s} \frac{x_i}{T_i}$ et utiliser la vraie valeur, peut être répercutée sur l'estimation de Y pour en améliorer la précision.

La solution naturelle consiste à utiliser l'**estimateur par régression** de y sur le vecteur de variables $x = (x_1, \ldots x_k)$. Soit:

$$Y_{reg} = \sum_{i \in U} x_i B + \sum_{i \in s} \frac{y_i - x_i B}{T_i} = \sum_{i \in s} \frac{y_i}{T_i} + (X - \sum_{i \in s} \frac{x_i}{T_i}) B$$

où B représente le coefficient de la regression de y sur le vecteur de variables $x = (x_1, \ldots x_k)$ estimé sur l'échantillon s.

En réalité, la solution la plus générale à ce jour, qui recouvre la précédente consiste à construire une famille d'estimateurs (**estimateurs par calage**) aussi proches que possible de l'estimateur d'Horvitz-Thompson (au sens d'une distance sur les poids), qui réconcilie exactement l'estimateur et la vraie valeur pour le vecteur de variables x.

On recherche alors les poids $(w_i)_{i \in s}$ du nouvel estimateur vérifiant $\sum_{i \in s} x_i w_i = X$ les plus proches possibles des poids de sondage initiaux $\left(d_i = \frac{1}{T_i} \right)_{i \in s}$.

Le choix de la distance entre les poids initiaux $\left(d_i = \frac{1}{T_i}\right)_{i\in s}$ et les poids après redressement $(w_i)_{i\in s}$ caractérise la méthode. On résoud ainsi un programme de minimisation sous contrainte:

$$\min \sum_{i\in s} H(d_i, w_i) \quad \text{où } H(.,.) \text{ représente une pseudo distance sur } \Re,$$
$$\text{sous la contrainte} \quad \sum_{i\in s} x_i w_i = X$$

Pour assurer l'existence et l'unicité d'une solution, il est nécessaire d'imposer des conditions de régularité sur la fonction $H(.,.)$.

En pratique, on réduit donc la classe des pseudo distances possibles aux fonctions de la forme:

$$H(d, w) = d.T(\frac{w}{d})$$

en imposant des conditions de régularité supplémentaires :

T est une fonction convexe $\Re \to \Re^+$
$T(1) = T'(1) = 0$
$T''(1) = 1$

Parmi les pseudo distances vérifiant ces conditions, on trouve à une constante près les deux formes du chi deux selon que la référence est w ou d.

La résolution du programme de minimisation à l'aide des multiplicateurs de Lagrange b conduit au système équivalent suivant :

$$(1)' \quad \frac{w_i}{d_i} = F(x_i b) \qquad \text{où F est reliée à T par } F(u) = T'^{-1}(u)$$
$$(2) \quad \sum_{i\in s} x_i w_i = X$$

Les nouveaux poids apparaissent comme une correction multiplicative des poids initiaux au niveau individuel basée sur la valeur du vecteur des variables auxiliaires $(x_1, ...x_k)$ pour chaque individu.

Différents choix sont possibles pour F . Les différentes fonctions F ainsi que les fonctions T et H associées sont:

	F	H	T
linéaire	$F(x) = 1 + x$	$H(w,d) = \frac{(w-d)^2}{2d}$	$T(u) = \frac{1}{2}(u-1)^2$
exponentielle ou raking ratio	$F(x) = \exp(x)$	$H(w,d) = w.\log(\frac{w}{d}) - w + d$	$T(u) = u\log u - u + 1$
linéaire tronquée	$F(x) = 1 + x$ si x∈[L,U] $F(x) = 1 + L$ si x<L $F(x) = 1 + U$ si x>U	$H(w,d) = \frac{(w-d)^2}{2d}$ si $\frac{w}{d} \in [L,U]$ $H(w,d) = \infty$ sinon	$T(u) = \frac{1}{2}(u-1)^2$ si x∈[L,U] $T(u) = \infty$ sinon
logit	$F(x) = \frac{L(U-1)+U(1-L)\exp(Ax)}{U-1+(1-L)\exp(Ax)} \in [L,U]$ avec : $A = \frac{U-L}{(1-L)(U-1)}$	$H(w,d) = d\Pi(\frac{w}{d})$	$T(u) = \left[(u-L)Log\frac{u-L}{1-L} + (U-u)Log\frac{U-u}{U-1}\right]\frac{1}{A}$ si u∈[L,U) $T(u) = \infty$ sinon
chi-deux	$F(x) = (1-2x)^{-1/2}$	$H(w,d) = \frac{(w-d)^2}{2w}$	$T(u) = \frac{(u-1)^2}{2u}$
hellinger	$F(x) = (1-x)^{-2}$	$H(w,d) = (\sqrt{w} - \sqrt{d})^2$	$T(u) = (\sqrt{u} - 1)^2$
entropie	$F(x) = (1+x)^{-1}$	$H(w,d) = dLog(\frac{w}{d}) + (w-d)$	$T(u) = -Log(u) + u - 1$

Lorsque F est **linéaire,** l'estimateur par calage correspond à l'**estimateur par régression** qui apparait dans ce cas comme une présentation duale possible de la méthode linéaire.

Lorsque F est exponentielle, le calage correspond à la technique du **raking ratio**.

La **justification théorique** de l'emploi de l'**estimateur par calage** est asymptotique et repose sur diverses hypothèses, dont la plus importante est la convergence de $\sum_{i \in s} \frac{y_i}{T_i}$ vers Y à la vitesse de $\frac{N}{\sqrt{n}}$ lorsque $n \to \infty$ et $N \to \infty$.

Lorsque les hypothèses sont vérifiées, les différentes méthodes d'estimation par calage sont **asymptotiquement équivalentes** (voir DEVILLE SARNDAL 1992), c'est à dire que lorsque $n \to \infty$ et $N \to \infty$, les résultats obtenus à l'aide de ces différentes méthodes d'estimation se rapprochent. Des comparaisons menées sur des enquêtes INSEE confirment la proximité des résultats obtenus à partir des différentes méthodes. Les variances de ces estimateurs, et donc leurs précisions, sont également équivalentes asymptotiquement.

Ces méthodes peuvent être mises en oeuvre facilement grâce à la macro CALMAR (CALage sur MARges), écrite en langage SAS par O.SAUTORY (voir O.SAUTORY 1993).

1-3 la pratique courante du redressement des enquêtes en présence de non-réponse : calage simple (méthode n°1)

Le calage a pour but de réduire les effets de l'aléa d'échantillonnage. Toutefois, la pratique courante pour le redressement des enquêtes consiste à appliquer cette méthode à l'ensemble des répondants, de façon à réduire les effets de l'aléa d'échantillonnage tout en corrigeant simultanément les déformations de structure induite par la non-réponse :

Le calage effectué grâce à l'information auxiliaire $X = \sum_{i \in U} x_i$ conduit à des poids w_i obtenus à partir de l'une des fonctions F mentionnées précédemment, qui vérifient:

(3) $\sum_{i \in r} x_i w_i = X$

(4) $w_i = \frac{F(x_i b)}{T_i \hat{p}}$ où $\hat{p}$ représente le taux de non-réponse observé (donc estimé), c'est à dire m/n.

1-4 une démarche correcte : le redressement en deux étapes (méthode n°2).

La non réponse est traitée comme une phase additionnelle de tirage. L'estimateur naturel du total Y est l'estimateur sans biais $\sum_{i \in r} \frac{y_i}{T_i p_i}$ qui tient compte du comportement de réponse à travers la probabilité de réponse p_i.

Ayant corrigé l'estimateur pour intégrer la non-réponse, on peut utiliser l'information auxiliaire X pour limiter les effets de l'aléa d'échantillonnage. On modifie donc les pondérations $\frac{1}{T_i p_i}$ des individus en des pondérations w_i * définies par :

$$(5) \qquad \sum_{i \in r} x_i w_i^* = X$$

$$(6) \qquad w_i^* = \frac{F^*(x_i b^*)}{T_i p_i}$$

où F* est une des fonctions possibles pour le calage. La justification asymptotique de l'emploi de l'estimateur par calage est alors acquise, si l'on considère le sondage en deux phases que constitue le tirage des répondants. Il suffit alors de remplacer n par m, d'où $m \to \infty$ et $N \to \infty$.

Cependant, en pratique, p_i est inconnu et doit être estimé dans une première étape. La démarche correcte pour effectuer le redressement consiste donc à :

1- modéliser le comportement de réponse de façon à estimer p_i. dans une 1ère étape, et modifier les pondérations des individus en divisant les poids de sondage par les probabilités de réponse estimées : $\hat{p}_i$.

2- caler l'échantillon des répondants sur une structure connue pour la population totale, en partant des poids de sondage modifiés pour tenir compte de la non-réponse.

Les pondérations résultantes w_i * à utiliser dans les estimations issues de l'enquête sont alors données par:

$$(5) \qquad \sum_{i \in r} x_i w_i^* = X$$

$$(6)' \qquad w_i^* = \frac{F^*(x_i b^*)}{T_i \hat{p}_i}$$ où F* est une des fonctions possibles pour le calage.

En pratique, on utilise pour p_i un modèle paramétrique de type $p_i = G(z_i c)$ où $(z_1 \ldots\ldots z_h)$ représente un vecteur de variables auxiliaires. En général, la condition $p_i \in [0,1]$ conduit à choisir pour G une fonction de répartition. On utilise ainsi couramment un modèle LOGIT qui correspond à la fonction de répartition d'une loi logistique et le modèle PROBIT qui correspond à la fonction de répartition d'une loi normale. On peut également ne pas inclure la contrainte $p_i \in [0,1]$ dans le modèle, et envisager par exemple une modélisation linéaire $p_i = z_i c$ ou exponentielle $p_i = \exp(z_i c)$.

Le paramètre c est alors estimé par une méthode convergente. La plus habituelle est la méthode du maximum de vraisemblance, qui présente l'avantage d'être facilement mise en oeuvre lorsque G est la fonction de répartition de la loi logistique ou de la loi normale ou encore de la loi de Gompertz.

NB : La pratique courante, ou **méthode n°1,** correspond alors par construction, d'après ce qui précède, au cas particulier d'un modèle de non-réponse uniforme pour la **méthode n°2,** où la probabilité de non-réponse constante est estimée par m/n.

1-5 quelques remarques sur la mise en pratique de ces deux méthodes :

La méthode n°1 nécessite de connaître :

- les totaux $X_1 = \sum_{i \in U} x_{i1} \ldots\ldots X_k = \sum_{i \in U} x_{ik}$ pour la population
- les valeurs des variables $x_1 \ldots x_k$ au niveau individuel **pour les répondants seulement**.

La méthode n°2 nécessite de connaître :

♦ *pour le calage :*

- les totaux $X_1 \ldots\ldots X_k$ pour la population
- les valeurs des variables $x_1 \ldots x_k$ au niveau individuel **pour les répondants**

♦ *pour l'estimation du modèle de réponse*

- Simultanément les valeurs des variables $z_1 \ldots\ldots z_h$ au niveau individuel **pour les répondants et les non-répondants** dans le **même fichier**.

résumé de l'information nécessaire:

méthode n°1	**méthode n°2**
$x_{i1} \ldots\ldots x_{ik} \quad \forall i \in \mathbf{r}$	$z_{i1} \ldots\ldots z_{ik} \quad \forall i \in \mathbf{s}$
	$x_{i1} \ldots\ldots x_{ik} \quad \forall i \in \mathbf{r}$
$X_1 \ldots\ldots X_k$	$X_1 \ldots\ldots X_k$

La méthode n°1 présente donc deux avantages:

- elle requiert moins d'information,
- elle est plus légère dans sa mise en oeuvre : il n'y a pas de modèle de réponse à estimer.

De plus, à l'INSEE, compte tenu de l'organisation actuelle du stockage des résultats d'enquête, la méthode n°2 nécessite l'appariement du fichier de saisie de l'enquête et du fichier de sondage constitué au moment du tirage contenant les variables $x_1 \ldots x_k$, ce qui alourdit encore sa mise en oeuvre.

1-6 domaine de validité de la pratique courante et lien entre les résultats obtenus par les deux méthodes :

On a vu précédemment que la pratique courante correspondait par construction à la méthode n°2, appliquée en utilisant un modèle de réponse uniforme. Dans la réalité, l'utilisation qui est faite du calage direct (méthode n°2) est plus large. Elle repose sur l'intuition que si le mécanisme de réponse ne dépend que des variables $x_1 \ldots x_k$ utilisées dans le calage, le calage direct permet une correction simultanée de la non réponse d'une part et des erreurs d'échantillonnage d'autre part.

Le but de cette étude est d'examiner les conditions sous lesquelles l'intuition et donc la pratique courante peuvent être validées.

Nous allons en effet établir que les **deux méthodes coincident** exactement lorsque la fonction utilisée pour le calage et la fonction utilisée dans le modèle de réponse sont exponentielles, et que les variables de calage $x_1 \ldots x_k$ recouvrent, au sens de l'espace vectoriel engendré par ces variables, les variables $z_1 \ldots\ldots z_h$ du modèle de réponse.

On établit également que lorsque les deux techniques sont identiques, elles sont aussi équivalentes à une nouvelle instance de la méthode n°2 appliquée cette fois avec les probabilités de réponse $p_i = \exp(z_i c)$ exactes, (non estimées).

Nous verrons que les deux techniques coincident également lorsque la variable prise en compte dans le calage est une variable qualitative, et qu'elle recouvre les variables du modèle de réponse. Les deux méthodes réalisent alors une **poststratification**.

La méthode usuelle (méthode n°1) n'admettant pas de justification théorique naturelle dans le cas général, l'objet de la **deuxième partie** est d'étudier l'écart entre les deux méthodes lorsqu'elles ne coincident pas en pratique. Nous comparerons ainsi les résultats obtenus pour l'enquête sur la consommation alimentaire réalisée en 1989 par l'INSEE.

1-6-a Lorsque les fonctions de calage F et F* et la forme fonctionnelle du modèle de réponse G sont exponentielles : les deux méthodes coincident exactement.

Méthode n°1: pratique courante, calage simple:

Les poids $w_i = \dfrac{F(x_i b)}{T_i \hat{p}}$ sont donnés par l'équation en l'inconnue b :

$$(7) \qquad \sum_{i \in r} \frac{F(x_i b)}{T_i \hat{p}} x_i = X \quad \text{où} \quad \hat{p} = \frac{m}{n}$$

♦ soit lorsque $F(u) = F^*(u) = G(u) = \exp(u)$:

$$(8) \qquad \sum_{i \in r} \exp(x_i b - \log(\hat{p})) x_i = X$$

Méthode n°2:

Les poids $w_i^* = \dfrac{F^*(x_i b^*)}{T_i \hat{p}_i}$ sont donnés par l'équation en l'inconnue b*:

$$(9) \qquad \sum_{i \in r} \frac{F^*(x_i b^*)}{T_i G(z_i \hat{c})} x_i = X$$

♦ soit lorsque $F(u) = F^*(u) = G(u) = \exp(u)$:

$$(10) \qquad \sum_{i \in r} \frac{\exp(x_i b^* - z_i \hat{c})}{T_i} x_i = X$$

On montre (voir annexe) que lorsque les variables de calage recouvrent les variables $z_1 \ldots\ldots z_h$ du modèle de réponse, la solution de l'équation $\sum_{i \in r} x_i \frac{\exp(x_i a)}{T_i} = X$ lorsqu'elle existe est unique.

Dans cette situation, on peut toujours se ramener au cas où $z = x$.

Par unicité de la solution on obtient alors, lorsque la variable constante figure parmi les variables explicatives,

$b = b^* - c^* - (\log(\hat{p}), 0, \ldots\ldots, 0)$

Les poids $w_i = \frac{F(x_i b)}{T_i \hat{p}}$ et $w_i^* = \frac{F^*(x_i b^*)}{T_i \hat{p}_i}$ sont donc égaux et les deux méthodes coincident exactement.

La **méthode n°2** appliquée avec les **probabilités de réponse exactes** est alors identique aux deux précédentes par un raisonnement analogue. Les poids w_i^{**} sont en effet déterminés à partir de l'équation

$$(9) \qquad \sum_{i \in r} \frac{F^*(x_i b^{**})}{T_i G(z_i c)} x_i = X$$

L'unicité de la solution de l'équation $\sum_{i \in r} x_i \frac{\exp(x_i a)}{T_i} = X$ conduit à l'égalité des poids w_i , w_i^* , w_i^{**}

1-6-b lorsque le calage s'effectue sur la base d'une variable qualitative qui explique entièrement le mécanisme de réponse, les deux méthodes coincident et réalisent une poststratification :

Dans ce cas en effet, les formes fonctionnelles non tronquées donnent toutes des paramétrisations équivalentes. Le choix de F d'une part, de F* d'autre part et enfin de G sont alors indifférents.On peut donc se ramener au cas où F=F*=G=exp et appliquer le résultat précédent (voir annexe).

1-7 Cas de fonctions F F* et G quelconques :

$x_{i1}.....x_{ik}, c^*, X_1.....X_k$ étant donnés, les deux méthodes consistent à résoudre k équations non linéaires en les k inconnues que sont $b=(b_1,....b_k)'$ pour la méthode n°1 ou $b^*=(b_1{}^*,....b_k{}^*)'$ pour la méthode n°2.

Une **interprétation géométrique** va permettre de mieux comprendre le lien entre les deux méthodes .

On note $d_i=\frac{1}{T_i}$, les poids initiaux, corrects en l'absence de non-réponse.

1/ **Calage en l'absence de non-réponse** :

Le vecteur $(w_1,.....w_n)'$ ***des poids, obtenu par la méthode n°1, vérifie*** :

$$(1)\quad \begin{cases} \sum_{i\in s} x_{i1}w_i = X_1 \\ \\ \sum_{i\in s} x_{ik}w_i = X_k \end{cases}$$

soit k conditions affines sur le vecteur $(w_1,.....w_n)'$

$$(2)'\qquad w_i=\frac{F(x_i,b)}{T_i}$$

♦La condition (1) peut se réécrire $\begin{cases} (w|x_1) = X_1 \\ \\ (w|x_k) = X_k \end{cases}$ où (|) représente le produit scalaire usuel sur $\Re^n$ et $x_1 = (x_{11}, \ldots x_{n1})', \ldots x_k = (x_{1k}, \ldots x_{nk})'$ les vecteurs contenant les valeurs des variables auxiliaires $x_1 \ldots x_k$ pour les n inidividus répondants. Elle s'interprète comme l'appartenance de w à un espace affine de dimension n-k dans $\Re^n$.

♦La condition (2), s'interprète elle, comme l'appartenance de w à une courbe paramétrée par b soit k paramètres dans $\Re^n$.

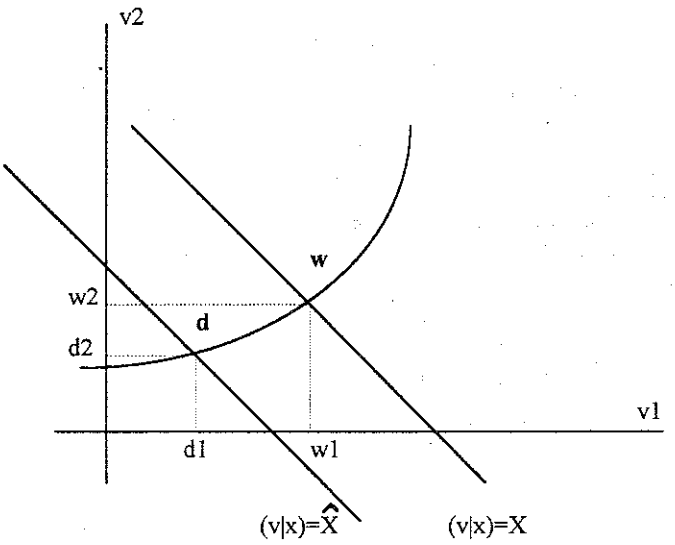
Le vecteur $d = (d_1, \ldots d_n)'$ ***des poids initiaux vérifie lui par définition de l'estimateur d'Horwitz-Thompson*** $\hat{X} = \begin{pmatrix} \hat{X}_1 \\ \\ \hat{X}_k \end{pmatrix}$:

$$(11) \quad \begin{cases} \sum_{i \in s} d_i x_{i1} = \hat{X}_1 \\ \\ \sum_{i \in s} d_i x_{ik} = \hat{X}_k \end{cases}$$

soit : $\begin{cases} (d|x_1) = \hat{X}_1 \\ \\ (d|x_k) = \hat{X}_k \end{cases}$, qui correspond à l'appartenance de d à un espace affine de dimension n-k parallèle à celui défini par (1).

Lorsque l'information auxiliaire consiste en une seule variable (k=1), le vecteur des poids est déterminé par l'intersection d'un hyperplan et d'une courbe paramétrée par un paramètre. Les poids initiaux sont situés eux, sur un hyparplan parallèle.

Une **représentation graphique dans le cas où n=2** permet de mieux comprendre:



2/ <u>Les deux méthodes en présence de non-réponse</u> :

L'ensemble des m répondants étant donné, ainsi que les valeurs de la variable auxiliaire, ***le vecteur des poids*** $w = (w_1,.....,w_m)'$ ***obtenu par la méthode n°1 est déterminé par le système à résoudre en b :***

(3) $\sum_{i \in r} x_i w_i = X$ sous-espace affine de dimension m-k dans $\Re^m$

(4) $w_i = \frac{F(x_i b)}{T_i \hat{p}} = \frac{F(x_i b)}{\hat{p}} d_i$ courbe paramétrée par m paramètres dans $\Re^m$

soit l'intersection d'un sous-espace affine et d'une courbe paramétrée.

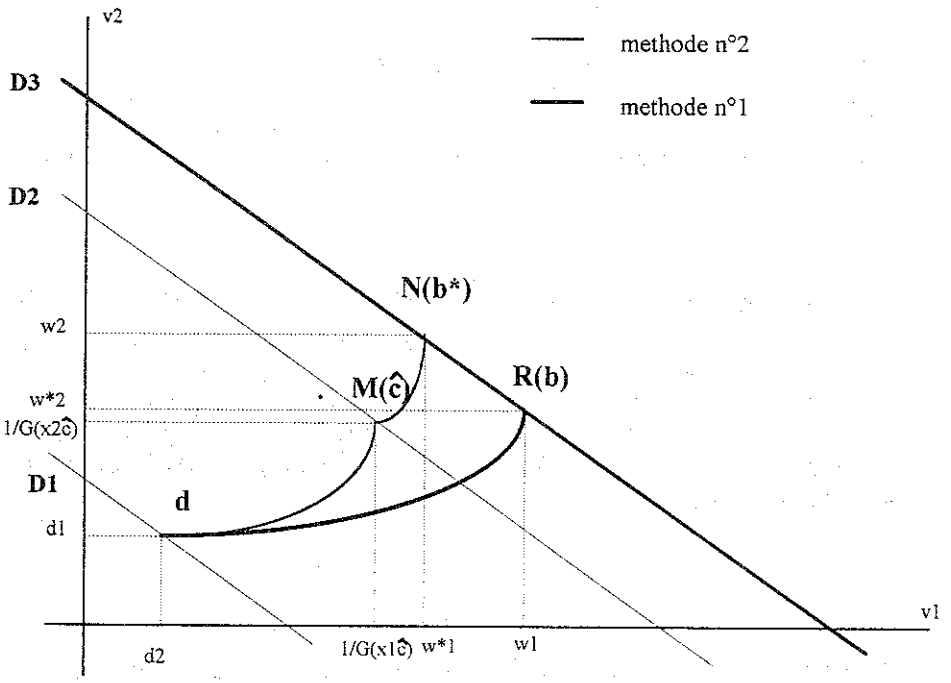
Le vecteur des poids w* obtenus par la méthode n°2 est déterminé par le système d'équations :

(5) $\sum_{i \in r} x_i w_i^* = X$ sous-espace affine de dimension m-k dans $\Re^m$

(6) $w_i^* = \frac{F^*(x_i b^*)}{T_i \hat{p}_i} = \frac{F^*(x_i b^*)}{G(x_i \hat{c})} d_i$ courbe paramétrée par m paramètres dans $\Re^m$

soit l'intersection d'un sous-espace affine et d'une courbe paramétrée par b*, c étant fixé dans une étape précédente.

une **représentation graphique dans le cas où m=2** permet de fixer les idées :



où :

D1 est la droite d'équation $\langle v|x \rangle = \sum_{i \in r} d_i x_i$

D2 est la droite d'équation $\langle v|x \rangle = \sum_{i \in r} \frac{d_i}{G(x_i \hat{c})} x_i$

D3 est la droite d'équation $\langle v|x \rangle = X$

♦ L'égalité des deux méthodes 1 et 2 lorsque $F(u) = F^*(u) = G(u) = \exp(u)$ correspond au fait que les courbes paramétrées **d ,M(c̃)** et **M(c̃)** , **N(b*)** sont les mêmes lorsque cette triple égalité est vérifiée.

♦ La représentation graphique peut être utilisée pour un cas plus général où **l'information auxiliaire se réduit à une variable quantitative qui ne prend que deux valeurs x1 et x2**. On suppose alors en outre que les poids initiaux ne prennent que deux valeurs, d_1 et d_2 , pour la même partition des individus que celle définie pour x_1 et x_2, c'est à dire :

$d_i = d_1 \Leftrightarrow x_i = x_1$ et
$d_i = d_2 \Leftrightarrow x_i = x_2$

Les poids w (resp w^*) ne prendront que deux valeurs notées w_1 et w_2 (resp $w_1{}^*$ et $w_2{}^*$). De même, les probabilités de réponse $G(x_i\hat{c})$ ne prendront que deux valeurs p_1 et p_2. La représentation graphique précédente correspondant à m=2 reste donc valide.

Lorsque m $\to \infty$ et N $\to \infty$ $\sum_{i \in r} \frac{x_i}{G(x_i\hat{c})}$ converge asymptotiquement vers X (cf DEVILLE, SARNDAL 1992). Ainsi la droite D2 se raproche de D3 et asymptotiquement on a : D2=D3. On s'attend donc à ce que la correction de non réponse ait une plus grande influence que le calage et en particulier que le choix de G ait une plus grande influence que le choix de F*. On retrouve ainsi graphiquement le fait que le choix de la fonction de calage devrait peu influer sur les poids finaux dans la méthode en deux étapes.

1-8 retour sur l'utilisation naive de CALMAR dans le cas général : interprétation en termes d'estimation du modèle de réponse de la méthode de redressement en une étape :

Asymptotiquement (m $\to \infty$ et N $\to \infty$), les **corrections pour non réponse** sont **finies** contrairement aux **correction pour calage** qui sont **en 1/√m** (voir DEVILLE, DUPONT 1993).

Les **équations de calage en une étape** (7)' $\sum_{i \in r} d_i x_i F(x_i b) = X$ peuvent de ce fait s'interpreter comme des **équations estimantes** des **probabilités de réponse** dans le cas où les probabilités de réponses sont de la forme $p_i = \frac{1}{F(x_i c)}$

En effet, si c était parfaitement connu, l'estimateur corrigé de la non-réponse s'écrirait $\hat{X} = \sum_{i \in r} d_i x_i F(x_i c)$. L'équation (7)' se réécrirait alors :

(a) $\sum_{i \in} d_i x_i F(x_i b) = \sum_{i \in r} d_i x_i F(x_i c) \frac{F(x_i (c+\lambda))}{F(x_i c)} = X$ où $b = c + \lambda$

Cette équation apparait simplement comme une équation de calage avec une *fonction de calage dépendant de l'unité i.* Tout se passe en effet comme si on partait de poids de sondage $d_i^* = d_i F(x_i c)$ et que l'on réalise un calage pour obtenir des poids de la forme $w_i = d_i * F_i(x_i \lambda)$ avec $F_i(u) = \frac{F(x_i c + u)}{F(x_i c)}$.

La solution de l'équation (7)' reçoit alors une inteprétation naturelle assez simple :

Supposons que nous disposions de la valeur des x_i sur l'échantillon s tiré tout entier, nous pourrions estimer un modèle de réponse postulé sous la forme $p_i = \frac{1}{F(x_i c)}$ son **estimation par le principe du calage** (voir DEVILLE, DUPONT 1993), conduirait à résoudre :

(b) $\hat{X} = \sum_{i \in s} d_i x_i = \sum_{i \in r} d_i x_i F(x_i \hat{c}) = \sum_{i \in r} d_i x_i F(x_i c) \frac{F(x_i (c+\delta))}{F(x_i c)}$ avec $\hat{c} = c + \delta$

La quantité δ, tout comme la quantité λ est d'un ordre infiniment petit $\frac{1}{\sqrt{m}}$.

Introduisons les quantités $f_i = \frac{dLogF}{du}\Big|_{x_i c} = \frac{F'(x_i c)}{F(x_i c)}$ et linéarisons les deux équations (a) et (b).

Si T est la matrice $T = \sum_{i \in r} d_i F(x_i c) f_i x_i' x_i$ et $\hat{X}_1 = \sum_{i \in r} d_i F(x_i c) x_i$, on a

$\lambda = T^{-1}(X - \hat{X}_1)$ et

$\delta = T^{-1}(\hat{X} - \hat{X}_1)$

et par conséquent :

$b = c + \lambda = \hat{c} - \delta + \lambda = \hat{c} + T^{-1}(X - \hat{X}_1)\lambda = \lambda$

Le vecteur b de (7)' apparait donc comme un estimateur de c dont la variance est du même ordre de grandeur que celle de $\hat{c}$ quoique, en principe, plus grande.

Ainsi donc l'**usage naif de CALMAR reçoit une interprétation en terme d'estimation de modèle de réponse**. Toutefois, les fonctions de calage habituelles ne s'interprètent pas à l'exception des fonctions exponentielles et logit comme provenant de modèles de réponse très naturels.

La **prépondérance** de la **correction de non-réponse** sur le **calage,** joue également un rôle essentiel pour le **calcul de variance** des estimations construite à partir du redressement en une étape. En effet, la récriture de l'équation (7)' en (a) permet de calculer la variance comme celle d'un estimateur par calage dont les poids initiaux incluent une correction pour non-réponse. Les poids initiaux conduisent alors à un estimateur sans biais qui converge vers la vraie valeur lorsque $m \to \infty$ et $N \to \infty$. Les hypothèses permettant le calcul de variance pour l'estimateur par calage sont alors vérifiées.

L'estimation de la variance nécessite de connaitre c. Il suffit alors de remplacer c par son estimation convergente $\hat{c}$.

La prédominance de la correction de non-réponse par rapport à la correction pour erreur d'échantillonnage conduit à **réexaminer le choix des données externes** sur lesquelles on cale une enquête. L'interprétation en termes de modèle de réponse conduit à choisir des variables qui expliquent bien le comportement de réponse. Le calage classique, conduit à choisir des variables qui expliquent bien les variables d'interêt. Une voie prometteuse à explorer consisterait à associer les deux idées : si z

explique bien la réponse et x explique bien la variable d'interet et que le total est connu, on peut imaginer reponderer en résolvant les équations :

$$Z = \sum_{i \in r} d_i z_i F(x_i b)$$

1-9 remarques sur la modification des poids de sondage initiaux avant calage dans la méthode en une étape (méthode n°1):

♦La pratique courante consiste comme on l'a vu à corriger d'un facteur n/m les poids de sondage initiaux avant calage. on détermine alors les nouveaux poids $w_i = \frac{F(x_i b)}{T_i \hat{p}}$ par la résolution de l'équation de calage $\sum_{i \in r} \frac{F(x_i b)}{T_i \hat{p}} x_i = X$.

♦L'interprétation en termes d'estimation des probabilités de réponses inclinerait à chercher des poids $(w_i)_{i \in s}$ sous la forme $w_i = \frac{F(x_i b)}{T_i \hat{p}}$ et à résoudre l'équation de calage $\sum_{i \in r} \frac{F(x_i b)}{T_i} x_i = X$.

En réalité on peut montrer que **pour les fonctions de calage F énumérées en 1-2**, ces deux **calages alternatifs** donnent les **mêmes poids finaux** $(w_i)_{i \in s}$ dès lors que la variable constante appartient à l'espace vectoriel engendré par les variables auxiliaires $x_1 .... x_k$. Ceci est le cas, dès qu'il existe une variable qualitative dans les variables auxiliaires. La variable constante est en effet obtenue comme la somme des variables indicatrices associées à la variable qualitative.

On établit en fait que la **multiplication des poids initiaux par une constante** quelconque **ne modifie pas** les résultat du calage en termes de **poids finaux** (à condition de modifier en conséquence les **bornes** qui portent sur les rapports de poids dans les méthodes bornées). Le lien entre les paramètres b des deux méthodes avec et sans modification préalable des poids de sondage s'écrit simplement dans la méthode

exponentielle. Les deux vecteurs b ne diffèrent en effet que sur la direction donnée par la variable constante .

Toutefois, la multiplication des poids initiaux joue un rôle au niveau de la **résolution numérique des équations de calage**. En modifiant les poids initiaux d'un facteur n/m on modifie le point initial de la résolution. Cette modification prend en compte la plus grande partie de l'effet de la variable constante dans $F(x_i b)$. On évite ainsi que la solution en b comporte de trop grandes valeurs dues à une grande correction sur la variable constante lorsque le taux de non réponse moyen est fort et que le comportement de réponse est relativement homogène.(En effet ceci revient à dire que la plus grande partie de $x_i b$ est donnée par le vecteur constant). La **modification des poids** initiaux effectuée en pratique donne alors une **valeur initiale plus favorable**. La résolution sans correction préalable peut en effet se réveler impossible lorsque le taux de non réponse est trop important.

2- comparaison empirique des deux stratégies de redressement:résultats obtenus sur une enquête réalisée par l'INSEE, l'enquête sur la consommation alimentaire de 1989.

L'échantillon est obtenu par un tirage à plusieurs degrés dans la **base de sondage B** constituée de la réunion du fichier du recensement de 1982 **(B1)** et d'une liste des logements construits depuis 1982 tenue à jour **(B2)** : on souhaite enquêter des **ménages ordinaires (population cible)**, pour cela on tire des **logements**.

En 1982, il y a équivalence entre l'ensemble des *résidences principales* occupées par des ménages ordinaires et l'ensemble des ménages ordinaires. A la date de l'enquête le passage entre logement et ménage ordinaire est réalisé en éliminant après constat sur le terrain, les l*ogements détruits* et les *logements vacants* ou occupés à titre de *résidence secondaires* à la date de l'enquête qui sont traités comme des unités hors champ.

année d'enquête → / année du recensement ↓	Résidences Principales	Résidences Secondaires	logements Vacants	logements détruits
Résidences Principales				
residences secondaires				
logements vacants				
logements à construire				

unités appartenant au champ de l'enquête
unités conservées dans l'étude

Comme on l'a vu en 1-5, la méthode de redressement en deux étapes nécessite de connaitre la valeur des variables utilisées dans le modèle de réponse pour les non-répondants. Les variables utilisées pour estimer le modèle de réponse doivent être disponibles pour les répondants et les non répondants, elles proviennent par conséquent nécessairement de la base de sondage. Or l'information sur la partie logements neufs de la base de sondage ne porte que sur la date d'achèvement et le maitre d'oeuvre. La méthode de redressement en deux étapes ne peut donc pas être appliquée aux unités extraites de la base de sondage **logements neufs**. Ces unités seront donc **exclues de l'étude** ainsi que les logements non principaux au moment du recensement. Celle ci ne portera donc que sur les unités extraites du recensement de 1982, correspondant à des résidences principales en 1982.

Ce **problème** apparaît en réalité de manière générale **pour toutes les enquêtes réalisées par l'INSEE à partir de l'échantillon maître**. La procédure de redressement en deux étapes ne peut s'appliquer puisqu'elle requiert l'élimination des logements neufs tirés. On voit donc l'importance de l'enjeu de l'équivalence des deux méthodes de redressement démontrée précedemment qui valide du même coup la seule méthode de redressement applicable qui correspond à la pratique courante.

L'étude qui suit a pour objectif de donner une idée de la **divergence entre les deux techniques de redressement** lorsqu'elles ne coincident pas dans le contexte des enquêtes réalisées par l'INSEE. Le choix de l'enquête sur la consommation alimentaire s'explique en grande partie par l'étude déjà effectuée par O.Sautory sur l'influence du choix des fonctions de calage sur les pondérations au niveau individuel. Le choix s'est

donc porté sur l'enquête **consommation alimentaire de 1989** en dépit de la distance avec le recensement de 1982. En effet, cet écart accroit la proportion de logements neufs dans l'échantillon et diminue donc d'une part la taille de l'échantillon utilisable pour l'étude, il fragilise d'autre part l'estimation d'un modèle de réponse puisque les variables utilisées sont relatives aux ménages occupant le logement au moment du recensement. Les resultats de cette étude ne sont donc qu'indicatifs de la divergence dans le cas de figure le plus défavorable du point de vue de la distance au recensement.

Les variables retenues ici pour le **calage** sont les variables qui ont été utilisées pour le calage effectif de cette enquête, à l'exception, pour des raisons de simplicité de la variable tranche d'âge x sexe relatives aux individus. En effet, le calage de cette enquête comporte un calage du niveau individu et un calage du niveau ménage, ces deux calages pouvant être effectués simultanément en substituant la variable ménage : *nombre d'individus par âge x sexe* à la variable du niveau individu. Cette variable n'étant pas accessible dans la base de sondage, elle a été éliminée dans l'étude.

Le **modèle de réponse** pouvait, quant à lui, inclure a priori toute variable disponible pour l'ensemble des individus recensés (RP82 exhaustif). Les limitations des variables disponibles dans l'échantillon maître 1982, ainsi que des considérations de robustesse et des tests de significativité du modèle de réponse ont finalement conduit à ne retenir que *deux variables supplémentaires par rapport au calage* : la nationalité française et la région de référence.

Ainsi les variables prises en compte dans l'étude sont :

2-1 variables retenues pour le calage (qualitatives):

-nombre de personnes du ménage :
- 1 personne
- 2 personnes
- 3 personnes
- 4 personnes
- 5 personnes
- 6 personnes et plus

-CS du chef de ménage :
- 1- agriculteurs, exploitants
- 2- artisans, commerçants, chefs d'entreprise
- 3- cadres et prof intellectuelles
- 4- professions intermédiaires
- 5-employés
- 6-ouvriers
- 7- inactifs et non déclarés

-âge du chef de ménage :
16 à 24 ans
25 à 34 ans
35 à 44 ans
45 à 54 ans
55 à 64 ans
65 à 74 ans
75 et plus

- catégorie de commune :
commune rurale
moins de 10 000 h
10 000 à 50 000 h
50 000 à 200 000 h
plus de 200 000 h

2-2 variables retenues pour le modèle de non-réponse (qualitatives) :

- variables de calage
+
- region référence
- nationalité:
1 français
2 étranger

2-3 choix des formes fonctionnelles F, F* et G :

Dans CALMAR trois fonctions F sont utilisées sans créer de problèmes ; il s'agit des fonctions **exponentielle** (raking ratio) , **logit** et **linéaire tronquée**. Seules ces trois fonctions seront donc utilisées dans l'étude empirique.

En effet, la résolution pour la fonction F **linéaire** peut déboucher sur des poids négatifs qui ne reçoivent aucune interprétation. Par ailleurs rappellons qu'en cas d'utilisation d'une pondération comportant des poids négatifs, ceux ci seraient éliminés par la suite lors de l'utilisation de procédures SAS telles que freq means etc...

Les fonctions F **chi-deux** et **Hellinger**, quant à elles, posent un problème de domaine de définition : ces fonctions ne sont pas définies sur tout $\Re$ et donc pas pour certaines valeurs de $x_i b$.

Pour le modèle de non réponse, on peut estimer facilement les modèles avec la procédure SAS proc logistic qui autorise trois fonctions G, fonction de répartition des lois logistique, normale et Gompertz. Les deux premières sont trop proches pour induire des différences significatives. On utilisera donc seulement les fonctions de répartition des lois logistiques et Gompertz.

2-4 Résultats des deux méthodes de redressement sur l'enquête consommation alimentaire :

2-4-1 comparaison au niveau individuel :

On a cherché à évaluer la répercussion du choix de la méthode de redressement au **niveau individuel**, c'est-à-dire la répercussion du choix des méthodes sur la valeur des poids. Pour cela on calcule au niveau de chaque ménage le rapport entre les poids obtenus par deux méthodes alternatives. La sensibilité de la pondération à une modification du choix de F ou F* ou G et/ou au choix entre les méthodes 1 et 2 est étudiée à partir des *écarts types* de la *distribution des rapports*.

En effet, soit $\left(w_i^{\,a}\right)_{i\in s}$ et $\left(w_i^{\,b}\right)_{i\in s}$ les poids obtenus par deux procédures de redressement différentes a et b. La moyenne des rapports de poids est égale à un. L'écart type de la distribution du rapport des poids $\left(\frac{w_i^{\,a}}{w_i^{\,b}}\right)_{i\in s}$ mesure l'écart relatif au niveau individuel des résultats obtenus par les deux méthodes. En effet $\sigma\left(\frac{w^a}{w^b}\right)=\sqrt{E\left(\frac{w^a-w^b}{w^b}\right)^2}$. Cette mesure est indépendante des poids initiaux d_i qui sont éliminés dans le rapport. Elle doit être comparée à l'ampleur de la correction des méthodes a ou b. Celle ci peut être mesurée par le coefficient de variation de la distribution des poids $\left(w_i^{\,a}\right)_{i\in s}$ obtenus par la méthode a ou par le coefficient de variation de la distribution des poids $\left(w_i^{\,b}\right)_{i\in s}$ obtenus par la méthode b.

Ainsi, si l'on cherche à mesurer l'influence du choix de F dans la méthode en une étape sur les résultats du redressement au niveau individuel, on utilisera

$$\sigma\left(\frac{w^a}{w^b}\right)=\sigma\left(\frac{F^a(xb^a)}{F^b(xb^b)}\right)$$

De même si l'on cherche à mesurer l'influence du choix de F seule, c'est à dire ayant fixé G, les variables du modèle de réponse et la méthode d'estimation, dans la méthode en deux étapes sur les résultats du redressement au niveau individuel, on utilisera :*

$$\sigma\left(\frac{w^a}{w^b}\right)=\sigma\left(\frac{F^{*a}(xb^a)G(z\hat{c})}{F^{*b}(xb^b)G(z\hat{c})}\right)=\sigma\left(\frac{F^{*a}(xb^a)}{F^{*b}(xb^b)}\right)$$

Les comparaisons ont également été menées à titre indicatif avec une autre mesure qui ne tient pas compte des valeurs extrêmes de la distribution des poids redressés. Il s'agit de l'écart interquartile de la distribution des rapports de poids.

Soulignons que la comparaison entre méthodes de redressements alternatifs a été effectuée sur la base d'une **enquête réelle** et diffère dans son optique d'une comparaison menée sur la base de **simulations**.

Dans le **premier cas**, il s'agit en effet de mesurer l'écart entre un groupe de procédures de redressement admettant une justification théorique et un autre groupe de procédures qui sont celles **effectivement utilisées** pour redresser les enquêtes ménages. On souhaite alors évaluer l'écart entre les résultats que l'on utilise en pratique et les résultats admettant une justification **que l'on aurait pu utiliser**. On est donc amené à envisager dans chaque groupe de procédure les variantes effectivement utilisables à l'heure actuelle.

Dans le **second cas** en revanche, l'angle d'attaque est plus général et consiste à discuter des divergences entre les deux groupes de méthodes d'un **point de vue théorique**. Il s'agit mesurer les divergences en fonction de divers paramètres :

1- mécanisme de réponse vrai : variables influant sur le fait de répondre et forme fonctionnelle du modèle de réponse

2- forme des fonctions de calage F

*3- forme des fonctions de calage F**

4- forme fonctionnelle G utilisée pour l'estimation de la non réponse

5- variables retenues dans l'estimation du modèle de réponse (oubli de variables par rapport au mécanisme de réponse vrai)

6- méthode d'estimation retenue pour le modèle de réponse (maximum de vraisemblance, calage, moments)

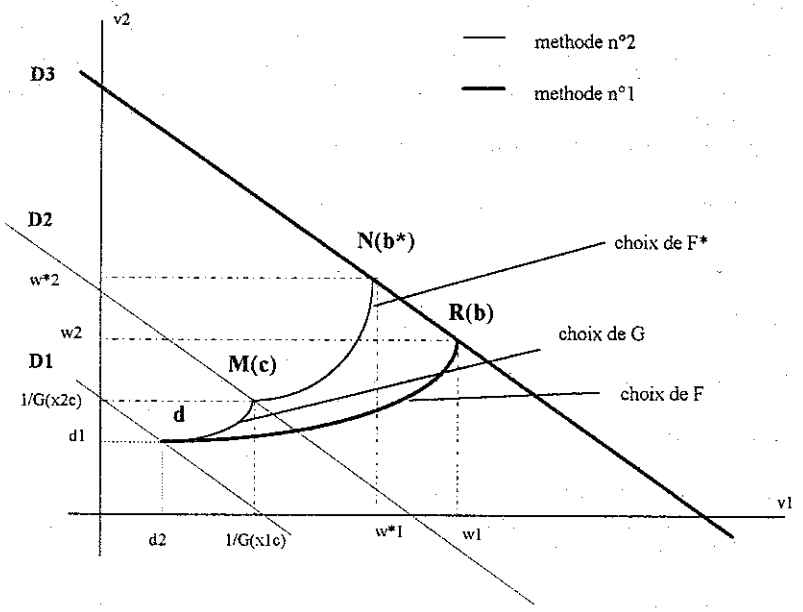
Ainsi, les comparaisons effectuées sur la base de l'**enquête alimentaire 1989** permettent seulement de **replacer les conséquences sur les pondérations finales des choix qui ont été faits** au niveau du redressement entre deux groupes de méthodes d'une part et de leurs variantes d'autres part **par rapport aux choix qui auraient pu**

être faits. Elles n'ont pas de portée suffisamment générale pour infirmer la théorie générale.

les résultats sont les suivants :

Dans cet exemple, les **corrections pour non réponse** sont **deux fois moins importantes** (au sens d'une mesure par le coefficient de variation) que les **corrections pour erreur d'échantillonnage**. Les coefficients sont en effet respectivement de l'ordre de 0.2 et 0.4 quelque soient les choix opérés aux différents niveaux. On ne se situe donc pas dans la configuration attendue en fonction de la théorie asymptotique.

Les résultats que l'on obtient dans les comparaisons découlent directement de cet état de fait et peuvent s'interpreter en relation avec la figure p17 vue en 1-7



Ainsi, par exemple, les effets du choix de F* n'ont pas de raison a priori d'être très petits devant les effets du choix de G. La distance entre les droites D1 et D2 est en effet deux fois "plus petite"que la distance entre les droites D2 et D3. La différence entre deux pondérations alternatives dépend alors essentiellement de l'ampleur de la différence entre les formes fonctionnelles utilisées.

a/ choix de G :

La différence entre les pondérations obtenues avec la fonction de répartition d'une loi logistique et avec la fonction de répartition d'une loi de Gompertz est de l'ordre de 0.07. Elle n'est pas tout à fait négligeable par rapport à l'ampleur des

corrections pour non réponse qui sont de 0.2 et les corrections globales qui sont de 0.4.

b/ choix des variables explicatives :

L'influence de la modification dans le choix des variables explicatives du modèle de réponse est de l'ordre de 0.07 également.

c/ choix de F* :

Les deux fonctions logit et exponentielles (raking ratio) conduisent à des poids assez proches dans la mesure où les bornes choisies sont inactives (1). La différence mesurée en écart type du rapport des poids ne dépasse pas 0.06. En revanche, les poids finaux diffèrent notablement plus lorsqu'on oppose deux versions de la méthode 2 avec un calage réalisé à l'aide de l'une des deux fonctions logit ou exponentielles d'une part, et linéaire tronquée d'autre part. L'écart est de 0.13 . Il n'est pas dû aux bornes qui sont inactives elles aussi dans ce cas. La sensibilité des résultats au choix entre linéaire tronquée d'une part et exponentielle ou logit d'autre part est en rapport direct entre les différences entre ces trois formes fonctionnelles sur les plages de valeurs des quantités xb*. (sur lesquelles varient xb*). Cette différence n'est pas neutre lorsqu'on la compare à l'ordre de grandeur de la correction globale qui est de 0.4

(1) on sait en effet que la fonction exponentielle est obtenue en faisant $U \to 0$ et $L \to \infty$ dans la fonction logit.

d/ choix de F :

Les mêmes remarques s'appliquent pour la fonction F.

e/ choix de méthodes :

Les différences entre les pondérations obtenues en utilisant une version de la méthode 1 et une version de la méthode 2 sont en rapport direct avec les choix de fonctions éffectués dans les deux versions comparées.

L'égalité des deux méthodes dans le cas exponentiel repose en effet sur l'égalité (R) $G(a)+F^*(b)=F(a+b)$ lorsque $G=F^*=F=exp$. Dans ce cas en effet, les deux "trajets" empruntés par les deux méthodes sont les mêmes. Tout se passe comme si les résultats traduisaient cette "distance plus ou moins grande des choix de (G,F^*,F) à la relation (R).

Ainsi,

un choix (logit, ratio, ratio) donne une différence de 0.02.
un choix (logit, ratio, logit calage) donne une différence de 0.05.
un choix (logit, logit calage, ratio) donne une différence de 0.033
un choix (logit, logit calage, logit calage) donne une différence de 0.065

un choix (logit, linéaire tronquée , ratio) donne une différence de 0.166
un choix (logit, linéaire tronquée , logit calage) donne une différence de 0.2
un choix (logit, linéaire tronquée, linéaire tronquée) donne une différence de 0.14.

un choix (logit, ratio, linéaire tronquée) donne une différence de 0.136
un choix (logit, logit calage, linéaire tronquée) donne une différence de 0.129

Dans cet exemple, le choix de méthodes n'est pas neutre dans tous les cas si on le compare à l'ordre de grandeur des corrections appliquées pour l'une quelconque des méthodes : les différences ne sont pas négligeables lorsque la fonction linéaire intervient dans l'un au moins des termes de la comparaison. Elles sont néamoins exactement comparables aux différences que l'on trouve lorsque l'on compare deux versions de la méthode n°2.

Tout se passe comme si le choix de méthode n'avait pas plus d'influence sur les pondérations finales que le choix des fonctions dans l'application de la méthode valide.

Tous ces résultats sont évidemment relatifs à ce cas particulier et découlent du fait que les corrections pour non réponse sont deux fois moins importantes que les corrections pour erreur d'échantillonnage.

2-4-2 comparaison au niveau agrégé :

Nous nous sommes ensuite intéressés à l'influence des choix de méthodes éffectués au niveau de redressement sur l'**estimation** et donc sur les **résultats de l'enquête** proprement dits. Pour les résultats agrégés, nous avons trouvé que l'influence des choix opérés au niveau du redressement étaient négligeables : en effet, l'influence de la procédure de redressement ne dépasse pas 0.1 point sur les pourcentages calculés pour la répartition des variables qualitatives et moins de 0.4% de différence sur les moyennes calculées pour les variables quantitatives. Ainsi, on obtient que l'influence du choix de la méthode usuelle plutôt que de la méthode en deux étapes est tout aussi négligeable que l'influence du choix de la fonction de calage. Les calculs à un niveau moins agrégé restent à poursuivre et pourraient conduire à une conclusion différente.

2-4-3 conclusion :

Toutes ces conclusions restent fragiles et attachées au cas particulier de l'enquête consommation alimentaire pour laquelle la plupart des facteurs explicatifs de la non réponse sont pris en compte dans le calage (ie on n'a pas mis en évidence de facteur explicatif supplémentaire important de non réponse par rapport au facteurs introduits naturellement dans le calage).
Des simulations en cours viendront compléter ces résultats de façon à leur donner une portée plus générale et à les infirmer le cas échéant.

ANNEXE 1 :

information disponible pour l'ensemble des ménages tirés:

a/ logement enquêté en 1982, B1 :
variables de l'exploitation exhaustive du RP82, information relative à la situation de 1982 du ménage qui occupait ce logement en 1982 :

identifiant:
- région
- département
- commune
- arrondissement
- canton
- vague d'enquête
numéro de fiche adressse

type d'habitat:
- catégorie de commune rural/urbain et nombre d'habitants état matrilmonial du chef de ménage
- nombre de logements par catégories en 1982 :principales, secondaires, vacants
- appartenance à une ville nouvelle

*caractéristiques du ménage**:
- nombre de personnes par tranche d'âge
- nombre de personnes actives du ménage
- nationalité du chef de ménage (français/étranger)
- catégorie socio-professionnelle du chef de ménage
- statut du chef de ménage
- âge détaillé du chef de ménage
- sexe du chef de ménage

réalisation de l'enquête:
- service enquêteur DR
- nombre d'enquêtes réalisées dans la commune à chaque
- vague

b/ logement construit depuis 1982, B1 :
- région
- commune
- département
- vague
- date d'achèvement du logement
- maître d'oeuvre
- catégorie de logement

ANNEXE 2 : DEMONSTRATION DE L'UNICITE DE LA SOLUTION DES EQUATIONS DE CALAGE :

L'égalité des deux méthodes de redressement repose sur l'unicité de la solution des équations $\sum_{i\in r} x_i d_i \exp(x_i a) = X$ que nous allons démontrer en reprenant l'interprétation géométrique vue en 1-7.

Il est en effet équivalent de résoudre $\sum_{i\in r} x_i d_i \exp(x_i a) = X$ en a ou de rechercher l'intersection

- de la courbe paramétrée définie dans $\Re^m$ par $w_i = d_i \exp(x_i a)$

et - du sous-espace affine de dimension m-k défini dans $\Re^m$ par $(w|x) = X$

Supposons que ces équations admettent au moins deux solutions c'est à dire qu'il existe au moins deux points d'intersection dans $\Re^m$. Notons A1 et A2 les deux valeurs du paramètre a associé. On a $\sum_{i\in r} x_i \exp(x_i A1) = X = \sum_{i\in r} x_i \exp(x_i A2)$

Soit h la fonction définie de $[0,1]$ dans $\Re$ par : $h(t) = \left[\sum_{i\in r} x_i \exp(x_i A1 + t(A2 - A1))\right](A1 - A2)$

h est continue sur $[0,1]$ et dérivable sur $]0,1[$. Or h(0)=h(1). En appliquant le théorème de Rolle on obtient que h s'annule en un point de l'intervalle ouvert $]0,1[$. Or, la dérivée de h ne peut s'annuler. En effet, exprimons la dérivée de h en un point t quelconque :

$$h'(t) = \left[\sum_{i\in r} x_i (A1 - A2) x_i \exp[x_i A1 + t(A1 - A2)]\right](A1 - A2) = \sum_{i\in r} [x_i (A1 - A2)]^2 \exp[x_i A1 + t(A1 - A2)]$$

Puisque A1≠A2, il existe une composante l sur laquelle A1 et A2 différent c'est à dire $A1_l \neq A2_l$. h'(t) est donc toujours strictement positive puisque les variables auxiliaires sont supposées non nulles et donc en particulier la lème variable est non nulle.

ANNEXE 3 : CAS OU LE CALAGE REPOSE SUR UNE SEULE VARIABLE QUALITATIVE : LES DEUX METHODES COINCIDENT ET REALISENT UNE POSTSTRATIFICATION

Supposons que l'on effectue les redressements sur la base d'une variable qualitative x* à k modalités. Les variables du redressement sont les k variables indicatrices associées aux k modalités. Les paramétrages en b, $F(x_i b)$ sont tous équivalents puisu'ils définissent exactement un paramètre pour chaque modalité. La fonction $F(x_i b)$ vaut en effet $F(b_l)$ lorsque x* prend la l ème modalité. On utilisera donc le paramétrage par groupe équivalent.

le **redressement en deux étapes** devient

1ère étape : correction pour non réponse :

la paramétrisation $p_i = G(x_i c)$ correspond d'après la remarque qui précède à un modèle de réponse homogène par groupe, c'est à dire : $p_i = p_l$ lorsque x* prend la lème modalité. Soit r_l (resp s_l, U_l), l'ensemble des répondants (resp des individus tirés, de la population totale) pour lesquels x* prend la l ème modalité. Les probabilités de réponse vont être estimées par les taux de réponse observés dans chaque groupe s_l. On obtient donc : $\hat{p}_l = \frac{m_l}{n_l}$

2ème étape : correction pour erreur d'échantillonnage : calage

Les équations de calage vont donner une correction multiplicative constante α_l dans chaque groupe r_l d'après la remarque sur l'équivalence des paramétrages. Elles s'écrivent :

$$\sum_{i \in r} d_i \frac{n_l}{m_l} \alpha_l = N_l \quad \text{où } N_l \text{ représente l'effectif de } U_l$$

Les nouveaux poids à l'issue du redressement par la méthode en deux étapes sont donc : $d_i \frac{n_l}{m_l} \alpha_l = \frac{d_i N_l}{\sum_{i \in r_l} d_i}$ dans le groupe r_l . Il est facile de voir qu'il ne dépendent pas de l'étape de correction pour non réponse. En particulier, ils ne dépendent pas de la méthode d'estimation des probabilités p_l.

Le **redressement en une étape** fait intervenir quant à lui une correction multiplicative des poids γ_l déterminée par les équations :

$$\sum_{i \in r} d_i \gamma_l = N_l$$

Les poids après redressement s'écrivent donc $d_i^* = d_i \gamma_l = \dfrac{d_i N_l}{\sum_{i \in r_l} d_i}$ dans le groupe r_l et les deux méthodes coincident.

L'estimateur associé pour une variable y dont on veut estimer le total s'écrit :

$$\hat{Y} = \sum_{i \in r_l} d_i^* y_i = \sum_{i \in r_l} \frac{d_i N_l}{\sum_{i \in r_l} d_i} y_i$$

Lorsque l'échantillon a été obtenu par un sondage aléatoire simple sans remise, on obtient :

$\hat{Y} = \sum_{l \in (1,k)} N_l \bar{y}_l$ c'est à dire l'estimateur poststratifié.

BIBLIOGRAPHIE :

non-réponse :

J.C.DEVILLE, F.DUPONT : non-réponse : principes et méthodes. Journées de méthodologie décembre 1993

C.E.SARNDAL,B.SWENSSON,J.WRETMAN : Model assisted survey sampling(Springer verlag,1991)

J.M.GROSBRAS : Méthodes statistiques des sondages, (économica,1987)

OH et SCHEUREN 1983 : weighting adjustment for unit non response. Incomplete data in sample surveys tome 2, 1983 academic press

estimation par calage :

J.C.DEVILLE,C.E.SARNDAL : Calibration estimators in survey sampling (Journal of the American Statistical Association vol47 n°418,juin 1992)

J.C.DEVILLE, C.E.SARNDAL, O.SAUTORY : Generalized Raking Procedures in survey sampling (Journal of the American Statistical Association, septembre 1993, volume 88 n°423)

O.SAUTORY : Redressement d'échantillons d'enquêtes auprès des ménages par calage sur marges (Document de travail de la Direction des Statistiques Démographiques et Sociales n°F9103).

O.SAUTORY : La macro SAS CALMAR: redressement d'un échantillon par calage sur marges. (Document de travail de la Direction des Statistiques Démographiques et Sociales n°F9108). Le document relatif à la nouvelle version de CALMAR est en cours de rédaction.

F.DUPONT : redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire note n°608/f010 du 10 novembre 1993.

mise en oeuvre des modèle économétriques sur variables qualitatives sous SAS :

O.VERGER, M.MARPSAT : L'économétrie et l'étude des comportements:présentation et mise en oeuvre de modèles de regression qualitatifs (Direction des Statistiques Démographiques et Sociales n°F9110, ouvrage collectif)