NON-RÉPONSE : PRINCIPES ET MÉTHODES

Jean-Claude Deville, Françoise Dupont

1 - Nature et ampleur de la non-réponse

1.1 - Classification des non-réponses

Toutes les enquêtes, que ce soit auprès des ménages ou des entreprises, sont soumises au phénomène de non-réponse. Celui-ci se manifeste de deux façons :

- la non-réponse totale (unit nonresponse in english) où aucune des variables d'intérêt n'est collectée. En général, on n'a pas recueilli de questionnaire et on ne dispose pour l'unité en question que de données issues de la base de sondage ou collectées par l'enquêteur sans contact avec l'unité enquêtée ;

- la non-réponse partielle, ou donnée manquante (item nonresponse) où certaines variables d'intérêt seulement manquent dans un questionnaire.

Les raisons suivantes sont généralement invoquées pour expliquer la non-réponse totale :

- unité non contactée (absence, adresse ou "coordonnées" mauvaises, autres raisons) ;

- refus ;

- abandon en cours de collecte ;

- incapacité ;

- défauts du processus de production (perte ou vol de documents, documents inexploitables) ;

- négligence du répondant (enquêtes postales surtout).

Pour la non-réponse partielle on peut ajouter les causes suivantes :

- incompréhension où impossibilité de répondre à la question ;
- refus ;
- négligence de l'enquêteur ;
- invalidation d'une réponse (par exemple pour incohérence).

Cette classification est bien sûr aussi incomplète qu'arbitraire et nous n'insisterons pas dessus. On se gardera bien, toutefois, de confondre la non-réponse totale avec le concept d'unité hors champ. Dans les enquêtes auprès des ménages, on échantillonne des logements pour atteindre des résidences principales, lieu où, par définition, on peut trouver des ménages. Une résidence secondaire ou un logement vacant est une unité hors champ et pas une non-réponse.

De même, quand un questionnaire prévoit une modalité "ne sait pas" ou "sans opinion", cela ne constitue pas une non-réponse. La *figure ci-dessous* résume cette discussion.

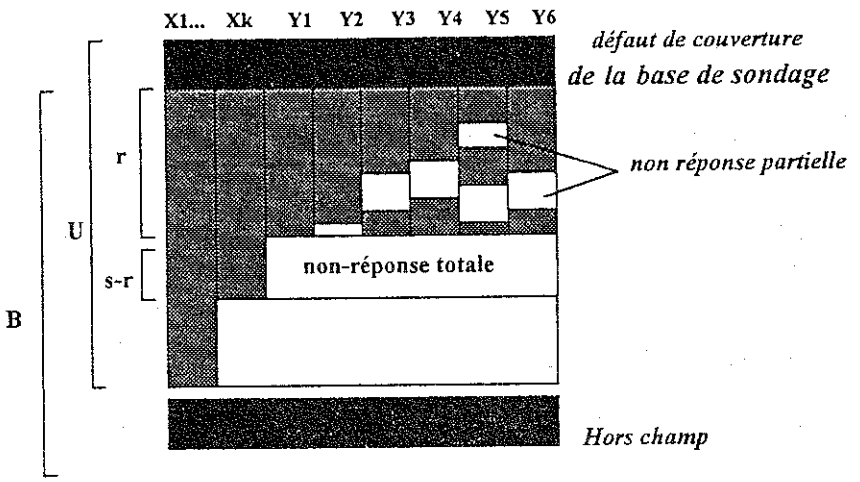

Tableau 1 - Non-réponse de l'enquête sur l'emploi de 1992

1992	Nombre de Logements	% acceptés	% refus	% ALD, impossibles à joindre
Échantillon spécial	2 749	91,9	2,4	5,7
Échantillon aérolaire	72 112	92,9	2,2	4,9
- dont :				
- Communes rurales	18 672	95,8	1,6	2,6
- Unités urbaines de moins de 20 000 habitants	11 643	94,7	1,8	3,5
- Unités urbaines de 20 000 à moins de 200 000 habitants	15 028	93,2	2,2	4,6
- Unités urbaines de 200 000 habitants ou plus (sauf agglomération parisienne)	14 752	92,1	2,2	5,7
- Agglomération parisienne	12 017	87,5	3,8	8,7
Ensemble	**74 861**	**92,9**	**2,3**	**4,8**

Tableau 2 - Non-réponse à diverses enquêtes sur les conditions de vie

Enquête	Non Contacté (%)	Contacté + Refus (%)	Accepté (%)		Total (%)
Loyers et charges (Janvier 1989)	6,5	4,5	89		100
Conjoncture (mai 1989)	7,6	8,3	84,1		100
Consommation Alimentaire (1991)	7,5	8,6	83,9		100
			complet 77	abandon 7,0	100
Emploi du temps (1985-1986)	7,5	8	84,5		100
			complet 77	abandon 7,5	
Budget de Famille (1989)	8,3	12,5	79,4		100
			complet 67,8	abandon 11,6	
Logement (1984)	5	7	88		100
Actifs Financiers (1986)	7	14	79		100
Situations défavorisées (1986-1987)	6	8	86		100
			complet 83	abandon 3	
Loisirs (1987-1988)	6	7	87		100
			complet 86	abandon 1	
Biens durables (1988)	6,5	10	83,1		100
			complet 82	abandon 1,5	

1.2 - Ampleur du problème

Dans les enquêtes auprès des ménages, la proportion de non-réponses totale varie de 7 % environ à l'enquête emploi jusqu'à 30 % pour des enquêtes complexes, impliquant par exemple l'usage d'un carnet de compte et plusieurs visites (enquêtes sur les budgets de famille en particulier). Si on particularise certaines catégories, ces taux peuvent atteindre des valeurs encore plus grandes (ménage de 1 personne, vivant à Paris, cadre supérieur). Les *tableaux 1 et 2* tirés de [PA] et de [EE], donnent une idée des ordres de grandeur.

Pour ce qui concerne les enquêtes auprès des entreprises, il est plus difficile de se faire une idée. Ces enquêtes sont en effet réalisées le plus généralement par voie postale. On a du mal alors à distinguer les unités hors champ (disparues) des non-réponses : une entreprise peut ne plus exister ou ne pas répondre. On arrive toutefois à des niveaux assez comparables, à ce qu'on observe dans les enquêtes de population : 10 à 15 % selon les secteurs ou les tranches de taille dans les enquêtes annuelles d'entreprise (EAE), 8 % dans les enquêtes trimestrielles sur les stocks. Ces enquêtes ont un statut obligatoire ; à l'opposé l'enquête mensuelle complémentaire sur les services atteint environ 40 % de non-réponses, chiffre vraisemblablement encore moins élevé que dans les enquêtes qualitatives de conjoncture.

2 - Quelques principes

Des chiffres d'une telle ampleur ne peuvent pas laisser indifférent, et deux principes doivent guider le responsable d'enquête qui veut parer aux méfaits de la non-réponse.

Premier principe : Faire en sorte d'avoir le moins possible de non-réponse.

Il existe tout un ensemble de techniques et de méthodes pour limiter le phénomène. Elles sont l'objet d'expérimentations et de chiffrages dont on ne parlera pas dans cet exposé. Nous nous contenterons d'examiner quelques possibilités :

- pratique de lettre-avis avant enquête ;
- plan de sondage adapté (permettant de limiter les déplacements et favorisant les repérages ; plan surreprésentant les catégories mal répondantes, etc.) ;
- pratique systématique des rappels ;
- utilisation de techniques d'enquête appropriées (face à face, CAPI, CATI,...) ;

- choix judicieux des enquêteurs (exemple : des femmes pour des enquêtes sur la contraception) ;

- formation rigoureuse des enquêteurs ;

- conception du questionnaire et formulation des questions ;

- incitatifs (cadeaux, récompenses, explications, éventuellement menaces liées à l'obligation de répondre).

Second principe : Bien être conscient, qu'en dépit du premier principe, la non-réponse fait partie de l'enquête et est inévitable.

D'où le corrollaire :

On doit établir une stratégie de correction des défauts liés à la non-réponse dès la conception de l'enquête. En particulier :

"Ne rien faire c'est faire quelque chose".

Supposons par exemple qu'on veuille estimer l'effectif d'un groupe. On ne peut pas appliquer aux répondants des pondérations issues directement du plan de sondage : l'estimation qui en résulterait serait sous-estimée, en première approximation, du taux moyen de non-réponse. Au minimum on se sent obligé de multiplier ce résultat par l'inverse de ce taux. Si on le fait (ce qui est déjà une décision !), on fait l'hypothèse implicite que les non-répondants sont tirés au hasard dans l'échantillon. Le caractère simpliste de cette hypothèse saute aux yeux et, généralement, on voit bien, pour le moins, que diverses parties de la population sondée ont des comportements de réponse différents.

De façon analogue, on doit prévoir la collecte de données qui faciliteront l'analyse et la correction des non-réponses. Ceci peut, en particulier, se traduire par des consignes aux enquêteurs signalant l'importance du recueil de données relatives à l'environnement de l'unité enquêtée.

On en arrive à une autre conséquence du second principe :

On doit procéder à une analyse statistique poussée de la non-réponse de façon à bien la décrire. Ceci permet, si tout se passe bien, de comprendre certains facteurs qui déterminent le phénomène. On est alors amené à modéliser le mécanisme de réponse. Le modèle est une formulation/formalisation des hypothèses que suscite l'analyse descriptive. La modélisation du mécanisme de réponse a pour but unique de corriger les données pour compenser certains effets indésirables.

Elle doit éviter un autre écueil qui est la complexification abusive : les modèles de réponses trop complexes deviennent incontrôlables et peuvent engendrer des estimations de paramètres qui conduisent à des corrections instables. Ont peut parfois introduire plus d'imprécision qu'on en corrige. Il faut veiller à faire simple : le modèle n'est pas destiné à refléter la vérité des choses ; le but n'est pas d'estimer parfaitement des paramètres mais d'apporter une correction honorable pour l'analyse des données de l'enquête en atténuant des imperfections évidentes.

Dernier principe, plus difficile à appliquer : une fois mise au point une stratégie de correction pour non-réponse, on doit essayer d'évaluer son impact sur la précision des analyses issues de l'enquête. Si les biais sont, par nature, difficiles à évaluer, on doit essayer, pour le moins, de chiffrer la part de variance induite par la correction pour non-réponse. Nous n'insisterons pas non plus sur ce point dans cette communication.

3 - Possibilités d'action

3.1 - À quoi veut-on arriver ?

Le but formel de la correction pour non-réponse est d'obtenir un fichier "rectangulaire", organisé en individus et variables, utilisable à toutes fins statistiques : tabulations, analyses descriptives de données, calculs d'indicateurs résumés ou d'indices, ajustement de modèles paramètriques, tests d'hypothèse, estimation de variance.

Si on voit les choses de façon qualitative, la non-réponse doit être considérée comme une modalité admissible au même titre que d'autres. Supposons qu'on pose par exemple, une question sur l'appartenance religieuse dans une enquête. On peut utiliser la réponse à cette question comme cofacteur, comme critère de ventilation d'une analyse statistique. La modalité "non réponse" doit alors, en général, être utilisée au même titre qu'une réponse positive. Même si une modalité "sans religion" est prévue, on ne sait pas si l'absence de réponse doit être interprétée comme un signe d'absence de religion, de rattachement à un groupe majoritaire, ou, au contraire, d'appartenance à un groupe plus confidentiel et peu affiché.

En revanche, si on voit les choses sous un aspect quantitatif, le problème change de nature : comment estimer le nombre ou la proportion de personnes qui se réclament par exemple de la religion catholique ? On ne peut pas aisément ventiler les non-répondants au prorata des déclarations positives ni les affecter allègrement au groupe majoritaire. On se trouve alors devant deux possibilités méthodologiques.

La première consiste à remplacer la non-réponse par une valeur plausible. On impute une valeur à une question non répondue et on parlera de technique d'imputation. L'autre

optique consiste à ne s'intéresser qu'à la population des répondants et à jouer sur les pondérations accordées aux unités pour compenser la non-réponse et faire des extrapolations. On parlera de technique de repondération (les anglo-saxons utilise volontiers le terme de quasi randomisation pour rappeler que la non-réponse peut être vue comme une forme non contrôlée d'échantillonnage).

3.2 - Principes généraux des techniques d'imputation

Nous ne parlerons plus dans ce papier ni dans cette session des techniques d'imputations, aussi allons-nous leur dire au revoir (et à bientôt !) en indiquant quelques axes de réflexions :

- l'idée derrière toute imputation est celle d'un modèle de prévision de la ou des variables manquantes à partir de variables présentes ;

- explicitement ou implicitement donc, on base une imputation sur une estimation de la loi de probabilité suivie par la (ou les) variable (s) à imputer en fonction de cofacteurs observés ;

- selon le cas on peut baser l'imputation sur la valeur prédite en espérance (imputation par la moyenne de classe, par ratio, par régression...) où sur une valeur aléatoire dans la loi estimée (souvent implicitement). A cette catégorie peuvent être rattachée les imputations par donneur (hot-deck en français) selon des techniques multiples.

3.3 - Choix entre imputation et repondération

Supposons que l'enquête ne porte que sur une seule variable d'intérêt. C'est rare mais ça se rencontre : certaines enquêtes de conjoncture auprès des entreprises vérifient à peu près ce critère ; il en va de même du sondage servant à élaborer l'indice du coût de la construction. La différence entre non-réponse totale et non-réponse partielle n'a pas de sens et le critère de choix entre imputation et repondération se ramène à la limitation de l'erreur quadratique moyenne (biais carré + variance). Les méthodes de correction par imputation ont généralement des défauts génants : une imputation par espérance change les distributions des variables (accumulation sur la moyenne pour parler vite). À l'opposé les méthodes aléatoires par donneur créent en supplément de variance artificiel lié au tirage aléatoire du donneur. On préfèrera donc généralement appliquer une méthode par repondération.

S'il y a plusieurs variables d'intérêt des difficultés apparaissent : chaque variable possède son propre ensemble de répondants, cumul de la non-réponse totale et de la

non-réponse spécifique aux questions permettent d'élaborer cette variable. Une optique de repondération conduit donc à un système de poids par variable. Cette condition est déjà lourde à admettre mais fait naître de nouvelles difficultés.

Dès qu'une statistique utilise plusieurs variables, on s'aperçoit qu'une pondération spécifique à cet ensemble de variables doit être utilisée, et que, de ce fait, on n'est généralement pas assuré de la cohérence entre deux statistiques. Par exemple l'estimation du total d'un tableau peut cesser d'être la somme des estimations des totaux de ses cases.

En pratique, donc, on utilisera la repondération pour la correction de la non-réponse totale et on corrigera la non-réponse par item, si nécessaire, par des imputations.

4- Théorie sommaire de la repondération

4.1 - Mécanisme de réponse et modèle de réponse

Commençons par formaliser le "mécanisme de réponse". Le plan de sondage initial est décrit comme une loi de probabilité sur l'ensemble de tous les échantillons possibles. Cette loi tient compte d'une information auxiliaire contenue dans la base de sondage et décrite par $z_U = \{z_k$ pour k dans $U\}$ où z_U est le vecteur des informations contenues dans la base de sondage. Le plan est donc décrit par les nombres $p\ (s\ ;\ z_U)$.

L'échantillon de répondants, r, est issu d'un mécanisme inconnu décrit par une loi de probabilité conditionnelle $q\ (r \mid s\ ; x_U, y_U)$ où x_U est un ensemble de variables présentes dans la population (et pouvant contenir celles qui figurent dans les z) et y_U la (ou les) variable (s) d'intérêt. C'est la présence de cette variable qui pose des problèmes épineux. En effet, si nous cherchons à mesurer y_k et que le fait que l'unité k réponde dépend explicitement de la valeur de y_k, on sent bien que les choses ne vont pas être simples.

Nous dirons, dans le cadre de ce papier, qu'un mécanisme de réponse est ignorable (on prononce en général iguenorabeulle mais ça n'a rien d'obligatoire) si on a les propriétés suivantes :

- y_k peut être mis sous la forme $f_k\ (x_K, \varepsilon_k)$ où f_k est une fonction connue et ε_K une variable nouvelle telle que la transformation $(x_k\ ,\ y_k) \rightarrow (x_k\ ,\ \varepsilon_k)$ soit régulière (biunivoque en particulier).

- le mécanisme de réponse ne dépend pas de ε_k :

$$q(r|s ; x_U, y_U) = q(r|s ; x_U, f_U(x_U, \varepsilon_U)) = q(r|s ; x_U).$$

Autrement dit le mécanisme de réponse ne dépend des y_U que par ce que les x_U en "expliquent". Autrement dit encore si on interprète la relation entre y et x comme un modèle probabiliste (par exemple de régression de y sur x), les réponses et les "résidus" ε_k sont indépendants conditionnellement à x.

Toute la "philosophie" de la correction pour non-réponse (ignorable) consiste à découvrir et à utiliser des variables x telles qu'on puisse considérer comme indépendantes réponses et variables d'intérêt.

4.2 - *Non-réponse et enquête en deux phases*

Si le mécanisme de réponse (ignorable) est entièrement connu (en particulier x_U est connu), l'enquête avec non-réponse peut être considérée comme une enquête en deux phases :

- phase 1 : tirage de l'échantillon s par la loi $p(s ; z_U)$ dans la population U ;
- phase 2 : tirage de l'échantillon r de répondants dans l'échantillon s par la loi $q(r|s ; x_U)$

On peut alors appliquer la théorie de ce type d'enquête ([COCHRAN], [SSW]) qui conduit aux estimateurs dits par expansion.

Le mécanisme de réponse étant supposé connu, on est capable de calculer la probabilité d'inclusion conditionnelle (probabilité de réponse si on est dans s) :

$$P_k = \sum_{r \ni k} q(r|s)$$

L'estimateur par expansion du total d'une variable y sera alors égal à :

$$\hat{Y} = \sum_r y_k/\pi_k \; P_k$$

où π_k est la probabilité d'inclusion dans l'échantillon s. On voit que l'espérance conditionnelle $E(\hat{Y} \mid s) = \sum_s y_k/\pi_k$ n'est autre que l'estimateur de HORVITZ-THOMPSON qu'on obtiendrait si on observait y sur tout l'échantillon s. Ce conditionnement permet aussi de décomposer la variance de $\hat{Y}$ en un terme dû à la première phase du sondage et un terme dû à la seconde :

$$Var(\hat{Y}) = = Var_p\left(\sum_s y_k/\pi_k\right) + E_p\ Var\left(\sum_r \frac{y_k/\pi_k}{P_k} \mid s\right)$$

On peut alors, si on connaît les probabilités d'inclusion doubles $P_{kl} = Pr(k \text{ et } l \in r \mid s)$ et un estimateur de variance de l'échantillon s, former un estimateur de la variance de $\hat{Y}$ (voire par exemple [SS] ou [SSW]).

En conclusion, si le mécanisme est ignorable et que nous le connaissons, le problème de la correction pour non-réponse est résolu : on utilise l'estimateur $\hat{Y}$ et donc on repondère les observations en mulpliant les poids de sondage $1/\pi_k$ par les poids de deuxième phase $1/P_k$.

4.3 - Modèles de réponse

Le problème est que nous ne connaissons qu'imparfaitement le mécanisme de réponse. Nous avons à choisir les variables x qui le déterminent et à modéliser la façon dont elles le font. En général nous ne serons pas capables (et nous n'aurons pas besoin !) de formaliser ce modèle dans tous ses détails. Nous commencerons par y introduire des paramètres mis dans un vecteur β. Ensuite nous nous contenterons de donner une forme analytique aux probabilités P_k en fonction du vecteur x_k. On notera la restriction de nos ambitions par rapport à une explicitation d'un plan $q(r \mid s)$. On écrira donc :

$$P_k = P_k(x_k ; \beta).$$

Pour que ce modèle soit identifiable il faudra bien entendu que la dimension des vecteurs x_k soit supérieure ou égale à celle du vecteur β. En pratique c'est l'égalité qui sera de règle.

Le modèle le plus simple est celui des groupes homogènes de réponses : x_k est un vecteur dont toutes les coordonnées sont nulles sauf une qui vaut un, indiquant l'appartenance de k à une catégorie ("le groupe homogène"). Le modèle postule que la probabilité de réponse dépend de la catégorie et est la même pour chaque individu de cette catégorie ; il contient donc un paramètre par groupe de réponse.

Bien que très utilisé, ce type de modèle est souvent avantageusement remplacé par des modèles linéaires généralisés où on pose $P_k = P(x'_k \beta)$. On peut utiliser :

- un modèle linéaire : $P_k = 1 - x'_k \beta$. Comme x_k est généralement un vecteur de nombres positifs, cette formulation permet d'avoir β composé de nombres positifs. De plus P_k est généralement voisin de 1 de sorte que cette formulation permet d'avoir des β proches de zéro ;

- le modèle log-linéaire-1 $P_k = \exp(-x'_k \beta)$;

- le modèle log-linéaire-0 $P_k = 1 - \exp(-x'_k \beta)$;

- le modèle Logit : $P_k = \exp(-x'_k \beta) / (1 + \exp(-x'_k \beta))$.

Nous verrons dans l'exposé suivant que certains autres modèles dérivés s'introduisent assez naturellement.

4.4 - Techniques d'estimation et but de l'enquête

La question est maintenant de savoir comment estimer les paramètres du modèle. Dans ce problème, on ne doit pas oublier que ce sont les $\hat{P}_k = P_k(X_k ; \hat{\beta})$ qui nous importent plutôt que les $\hat{\beta}$, et que même, on est indifférent, à la limite, aux valeurs de $\hat{P}_k$ si l'estimateur "estimé" :

$$\hat{Y}(\hat{\beta}) = \sum_r y_k / \pi_k \hat{P}_k$$

est de bonne qualité.

Ici encore nous allons avoir des principes.

4.4.1 - Premier principe

Toute procédure d'estimation raisonnable doit nous fournir un estimateur tel que $\delta = \hat{\beta} - \beta$ ait une variance de l'ordre de 1/n où n est la taille de l'échantillon s. D'autre part, les probabilités de réponses sont des quantités finies, insensibles, en particulier, à la taille de l'échantillon s sur lequel elles opèrent. La taille m de l'échantillon de répondants est donc telle que 1/m sera considéré comme un ordre de grandeur (probabiliste) équivalent à 1/n. On remarque qu'il en va de même de la taille n-m

de l'échantillon de non-répondants (échantillon 0). La conséquence de ces considérations peut se formuler sous forme de principe :

La technique d'estimation choisie n'a que peu d'influence sur l'estimation des variables d'intérêt du sondage.

On peut en effet remarquer que :

$$\hat{Y}(\hat{\beta}) = \sum_r \frac{y_k}{\pi_k} \frac{1}{P_k(\hat{\beta})} = \sum_r \frac{y_k}{\pi_k} \frac{1}{P_k(\beta)} - (\hat{\beta} - \beta)' \sum_r \frac{y_k}{\pi_k} \frac{P^{\bullet}(\beta)}{P^2{}_k(\beta)} = + \underset{reste}{}$$

Dans cette égalité $P^{\bullet}{}_k(\beta)$ est le vecteur des dérivés partielles de $P_k(s)$ par rapport aux coordonnées de β et *"Reste"* est une quantité dont l'ordre en probabilité est inférieur $1/\sqrt{n}$.

L'expression $\sum_r \frac{y_k}{\pi_k} \frac{P_{\bullet k}(\beta)}{P_k(\beta)} \frac{1}{P_k(\beta)} = \hat{A}$ peut être vue comme l'estimateur (conditionnel à s) d'une certaine quantité finie U. La variance de $\hat{A}$ est en 1/m de sorte que $\hat{A} - A$ est en probabilité de l'ordre de $1/\sqrt{m} = 1/\sqrt{n}$. On peut donc affirmer que (si le modèle est vrai), $\hat{Y}(\hat{\beta})$ est une approximation de faible biais de $\hat{Y}(\beta)$. Si celui-ci est en 1/n, son carré sera négligeable devant la variance. On aura donc :

$$Var(\hat{Y}(\hat{\beta})) - \hat{Y}(\beta)) \simeq Var[(\hat{\beta} - \beta)'(\hat{A} - A)] = 0(1/n^2)$$

car $(\hat{b} - b) = O_p(1/\sqrt{n})$ et $(\hat{A} - A) = O_p(1/\sqrt{n})$. Les estimateurs $\hat{Y}(\hat{\beta})$ *et* $\hat{Y}(\beta)$ auront donc la même valeur à des termes en $1/n^2$ près.

4.4.2 - *Un autre principe : On ne change plus un estimateur qui gagne*

La précision de l'estimation de $(\hat{\beta})$ est indifférente dès lors qu'on applique une méthode convergente en $n^{-1/2}$. On est donc en droit de chercher d'autres règles que celle de l'efficacité asymptotique de la statistique inférentielle classique.

Commençons par un bilan de l'information nécessaire à l'estimation de β. Nous avons besoin de connaître, sur l'échantillon s, les cofacteurs x_k et la variable de réponse R_k = 1 ou 0 selon que l'unité répond ou pas. Nous devons aussi, bien sûr, connaître la fonction de réponse P_k (x_k ; β). On doit, de plus, mettre en œuvre un principe d'estimation qui requiert éventuellement une information supplémentaire.

4.4.2.1 - Optique classique

Remplis de respect pour les canons de la statistique classique, on peut vouloir utiliser la méthode du maximum de vraisemblance. Plus ou moins implicitement on admet alors que les R_k sont des variables indépendantes et donc que l'échantillonnage de deuxième phase est POISSONIEN.

La maximisation de la vraisemblance conduit à une système d'équations dites du score :

$$\sum_r \frac{P^{\bullet}k}{P_k} = \sum_o \frac{P^{\bullet}k}{1-P_k} \qquad (1)$$

qu'on peut aussi écrire sous la forme :

$$\sum_r \frac{P^{\bullet}k}{P_k(1-P_k)} = \sum_s \frac{P^{\bullet}k}{1-P_k} \qquad (2)$$

Dans le cas d'un modèle linéaire généralisé où $P_k = P(x'_k \beta)$ ces équations s'écrivent (avec p dérivée de P) :

$$\sum_r x_k \frac{p_k}{P_k(1-P_k)} = \sum_s x_k \frac{p_k}{1-P_k} \qquad (2\ bis)$$

On pourrait trouver d'autres équations en appliquant d'autres principes de minimisation de l'écart entre les données et le modèle : moindre carrés, ou chi-2 minimum par exemple. Toutes nous conduisent à des équations estimantes qui ressemblent à (1) ou (2) et qui ne s'avèrent "parlantes" que dans des cas particuliers. Nous préfèrerons baser l'estimation sur des principes plus robustes et plus faciles à manipuler basés sur les équations estimantes elles-mêmes.

4.4.2.2 - Principes des moments

Considérons une variable Z_k, de même dimension que x_k, est observée sur s tout entier. On peut, sous le modèle, calculer l'espérance de $\sum_s Z_k R_k$ et l'égaler à la valeur observée. Ceci conduit aux équations estimantes :

$$\sum_s Z_k P_k = \sum_r Z_k \quad (3)$$

L'équation $(2bis)$ peut-être vue comme une équation aux moments pour $Z_k = x_k \frac{P_k}{P_k(1-P_k)}$. Les Z_k seront parfaitement définies si p/P (1-P) est une constante c'est-à-dire si la fonction donnant la probabilité de réponse est le Logit. C'est le seul cas simple où les équations de score s'identifient à des équations de moment.

4.4.2.3 - Optiques "calage"

Le principe des moments est encore très imprégné d'une idée d'ajustement et d'estimation de paramètres. Gardons les mêmes hypothèses : Z_k connu sur s. En l'absence de non-réponse – c'est le cas pour la variable Z –, on dispose d'un estimateur $\hat{Z}$ du total de Z qu'on pense être idéal compte tenu de l'information dont on dispose :

$$\hat{Z} = \sum_s Z_k W_k \qquad avec\ W_k = 1/\pi_k$$

si on se sert de l'estimateur de HORWITZ-THOMPSON.

En situation de non-réponse, on utilisera un estimateur "deux fois dilaté" :

$$\hat{\hat{Z}} = \sum_r Z_k W_k g_k \qquad \text{avec } g_k = 1/P_k$$

On aura, en général, $\hat{\hat{Z}} \neq \hat{Z}$. Si l'estimateur $\hat{Z}$ a été convenablement choisi, il est vraisemblable que $\hat{\hat{Z}}$ aura une variance plus grande que celle de $\hat{Z}$. Si, en particulier, la variable **Z** a un bon pouvoir explicatif sur **y** et qu'on l'ait choisie pour améliorer par régression l'estimation de **y**, il est naturel de ne pas vouloir changer l'estimateur , ne pas vouloir "changer l'estimateur $\hat{Z}$ qui gagne". Si on pose $v_k = w_k Z_k$, ceci conduit aux équations estimantes (avec H = 1/P) :

$$\sum_s v_k = \sum_r v_k \, H_k(x_k ; \beta) \quad (4)$$

Si la fonction $P_k = P(x'_k \beta)$ est liée à un modèle linéaire généralisé on aura avec $H = 1/P$

$$\sum_s v_k = \sum_r v_k \, H(x'_k \beta) \quad (4\,bis)$$

Ces équations sont typiquement des équations de calage ([D.S] et [DSS]) dans le cas où $x_k = Z_k$ et ou les poids W_k sont tous égaux. C'est pourquoi on peut appeler les équations (*4bis*) équations de calage.

Si nous revenons aux équations (2) ou (*2bis*) nous voyons que les équations du score se ramènent à des équations de calage dans le cas où $p/(1-P)$ est une constante c'est-à-dire dans le cas du modèle log-linéaire-0 : $P(x'_k\beta) = 1 - \exp(-x'_k \beta)$.

Remarque 1 : *Les équations (2) peuvent aussi s'écrire :* $\sum_s \frac{P^\bullet_k}{P_k} = \sum_o \frac{P^\bullet_k}{P_k(1-P_k)}$.

Pour un modèle log-linéaire 1 ces équations seraient celles d'un calage de l'échantillon de non-répondants.

Remarque 2 : *Dans le cas d'un modèle logit le principe de calage conduit aux équations :*

$$\sum_r v_k (1 + \exp(x'_k \beta)) = \sum_s v_k$$

soit

$$\sum_r v_k \exp(x'_k \beta) = \sum_o v_k$$

C'est donc, pratiquement, une variante du raking-ratio.

Remarque 3 *: Le principe de calage présente un avantage assez extraordinaire : les variables* x_k *n'ont pas besoin d'être connues sur s pour que l'estimation soit possible et on a le choix des* v_k *sur s. Ceci fonctionne même dans des cas élémentaires. Supposons un modèle à groupes homogènes de réponse, disons par catégories sociales (CS). Celles-ci sont observées sur r, mais pas sur s. Sur s on ne connait que le type de quartier, certes lié à la CS mais différent. Le principe de calage permet d'estimer les probabilités de réponse, ce que ne permettent ni le principe de vraisemblance ni le principe des moments.*

5 - Conclusions provisoires

L'élaboration d'un modèle de réponse pour repondération demande une analyse poussée du mécanisme de réponse. Celle-ci doit conduire à la spécification d'un modèle de réponse. Néanmoins on doit se garder de complications trop grandes. La façon d'estimer les paramètres du modèle influe peu sur la qualité les estimations finales. De ce fait, on recommande d'utiliser un critère de calage basé sur des équations estimantes ayant un sens concret évident.

BIBLIOGRAPHIE

[PA] : Pascal ARDILLY (1993) : Les Techniques de Sondage –Technip

[EE] : INSEE (1993) : Résultats détaillés de l'enquête sur l'emploi de 1992.

[COCHRAN] : Willian COCHRAN (1977) : Sampling Techniques, Third edition, WILEY

[SSW] : Carl Erik SÄRNDAL, Bengt SWENSSON, Jan WRETMAN (1992) : Model Assited Survey Sampling, Springer.

[SSW] : Carl Erik SÄRNDAL, Bengt SWENSSON (1987) : A General View of Estimation for Two-Phases of Selection with Application to Two-Phase Sampling and Non-response, International Statistical Review- Vol 55, pp : 279-294.

[DS] : Jean-Claude DEVILLE, Carl Erik SÄRNDAL, (1992) : Calibration Estimators in Survey Sampling, JASA, Vol 87 pp : 376-382.

[DS] : Jean-Claude DEVILLE, Carl Erik SÄRNDAL, Olivier SAUTORY (1992) : Generalized Raking Procedures in Survey Sampling, JASA, Vol 88, pp : 1013-1020.

Erratum à l'intervention de Jean-Claude Deville et Françoise Dupont

NON-RÉPONSE : PRINCIPES ET MÉTHODES

(Session 2, page 53 du volume principal)

Page 61, lire :

L'estimateur par expansion du total d'une variable y sera alors égal à :

$$\hat{Y} = \sum_r y_k / \pi_k P_k$$

Page 63, lire :

- le modèle log-linéaire-1 : $P_k = \exp (- x'_k \beta)$;
- le modèle log-linéaire-0 : $P_k = 1 - \exp (- x'_k \beta)$;
- le modèle Logit : $P_k = \exp (- x'_k \beta) \ / \ (1 + \exp (- x'_k \beta))$.

À partir de la page 64 :

l'expression P^*_k ***est à lire :*** $P^\bullet_k$

Page 64, lire :

$$\hat{Y} (\hat{\beta}) = \sum_r \frac{y_k}{\pi_k} \frac{1}{P_k (\hat{\beta})} =$$

$$\sum_r \frac{y_k}{\pi_k} \frac{1}{P_k (\beta)} - (\hat{\beta} - \beta)' \sum_r \frac{y_k}{\pi_k} \frac{P^\bullet_k (\beta)}{P^2_k (\beta)} = + \textit{reste}$$

Dans cette égalité $\overset{\bullet}{P}_k$ (β) est le vecteur des dérivés partielles de P_k (β) par rapport aux coordonnées de β et ***"Reste"*** est une quantité dont l'ordre en probabilité est inférieur $1/\sqrt{n}$.

Au lieu de : $(\hat{b} - b) = O_p(1/\sqrt{n})$ et $(\hat{A} - A) = O_p(1/\sqrt{n})$.

lire : $(\hat{\beta} - \beta) = O_p(1/\sqrt{n})$ et $(\hat{A} - A) = O_p(1/\sqrt{n})$.

Page 67, au lieu de : Si on pose $v_k = w_k Z_k$

lire : Si on pose $v_k = W_k Z_k$

x'_k ***est à lire*** x_k ***dans toutes les formules.***