LE PROBLEME DE LA NON-REPONSE

J.G. Bethlehem et H.M.P. Kersten[1]

Ce document donne un aperçu des recherches effectuées au sujet des non-réponses au Bureau central de statistique (BCS) des Pays-Bas. Le phénomène de la non-réponse est replacé dans un cadre général. Des indications sont données, au sujet de l'ampleur de la non-réponse, au moyen de chiffres tirés d'un certain nombre d'enquêtes effectuées par le BCS. Il est question de l'utilisation de variables auxiliaires comme moyen d'obtenir des renseignements au sujet des non-répondants. Ces variables peuvent soit servir à dégager les caractéristiques des non-répondants, soit être utilisées pour une stratification dans les procédures d'ajustement.

Il est question en plus grand détail de l'ajustement destiné à éliminer la distorsion imputable aux non-réponses au moyen d'une pondération par sous-groupes. Enfin, la dernière section énumère un certain nombre d'autres méthodes qui visent également à réduire cette distorsion.

1. INTRODUCTION

La non-réponse devient depuis quelque temps un sujet de préoccupation croissante dans la recherche relative aux enquêtes. Le phénomène de la non-réponse , qui est dû à ce que certaines personnes ne sont pas en mesure ou ne sont pas désireuses de répondre aux questions posées par l'enquêteur, peut apparaître dans les enquêtes par sondage aussi bien que dans les recensements. Il affecte la qualité de l'enquête de deux manières. Tout d'abord, par suite de la réduction de la quantité de

---

[1] J.G. Bethlehem et H.M.P. Kersten, Bureau central de statistique des Pays-Bas.

Les vues exprimées dans ce document sont celles de l'auteur et ne traduisent pas nécessairement la politique du Bureau central de statistique des Pays-Bas.

renseignements qui est obtenue, les estimations relatives aux paramètres démographiques sont moins précises. En second lieu, s'il existe une relation entre la variable étudiée et l'obtention de réponses, les calculs effectués sur la base des réponses ne sont pas valables pour l'ensemble de la population. Par exemple, si la demande de logements des répondants est supérieure à celle des non-répondants, les estimations de la demande de logements dans l'ensemble de la population seront sensiblement trop élevées.

Il est évident que le taux de non-réponse doit être maintenu à un niveau aussi bas que possible. Si, malgré ces efforts, il subsiste encore une quantité considérable de non-réponses, il faut alors prendre des dispositions pour éviter de formuler des conclusions erronées au sujet de la population. La combinaison de procédures d'ajustement et des techniques habituelles d'estimation doit aboutir à des estimations valables concernant la population.

Deux départements du BCS (Bureau central de statistique des Pays-Bas) participent aux recherches sur la non-réponse. Le Département des enquêtes sociales est responsable du travail effectué sur le terrain à l'occasion des enquêtes. Il s'attache à supprimer les non-réponses lors de la collecte des données. Il effectue des recherches sur le nombre optimal de rappels et le meilleur moment pour l'entrevue (voir Widdershoven et Van den Berg (1980)). Il procède à des expériences pour trouver la meilleure manière d'entrer en contact avec les personnes et les ménages au moyen de lettres introductives. Il s'efforce de mesurer l'impact de la lassitude engendrée par des entrevues trop longues ou trop fréquentes. En fin de compte, malgré ces efforts, il y a toujours une certaine quantité de non-réponses. Le Département des méthodes statistiques entreprend des recherches au sujet de l'effet des non-réponses sur l'exactitude des résultats de l'enquête. Il met au point des méthodes permettant d'ajuster les estimations démographiques pour éliminer la distorsion imputable aux non-réponses. La suite du présent document est essentiellement consacrée aux travaux de ce dernier département.

Les sections ci-après donnent un aperçu des travaux d'analyse effectués par le BCS au sujet des non-réponses. La section 2 présente des définitions, ainsi que les problèmes relatifs. Elle donne des chiffres concernant le taux de non-réponse enregistré à l'occasion d'enquêtes effectuées par le BCS. Dans la section 3 il sera question de méthodes graphiques utilisées pour choisir des variables auxiliaires. Ces méthodes jettent quelque lumière sur la non-réponse et peuvent être utilisées dans les procédures d'ajustement. La section 4 présente les méthodes d'ajustement fondées sur la pondération par sous-groupes et la section 5 donne des indications sur un certain nombre d'autres méthodes.

2. LE PHENOMENE DE LA NON-REPONSE

Dans la présente section le problème de la non-réponse est replacé dans un cadre général, où un certain nombre d'autres problèmes propres aux enquêtes par sondage jouent également un rôle. Des chiffres sont donnés au sujet des non-réponses constatées à l'occasion d'enquêtes effectuées par le BCS. Il sera question aussi de certains cas où une relation existe entre la variable étudiée et l'obtention de réponses. Dans la dernière partie de cette section seront examinés deux modèles concernant le mécanisme de la non-réponse.

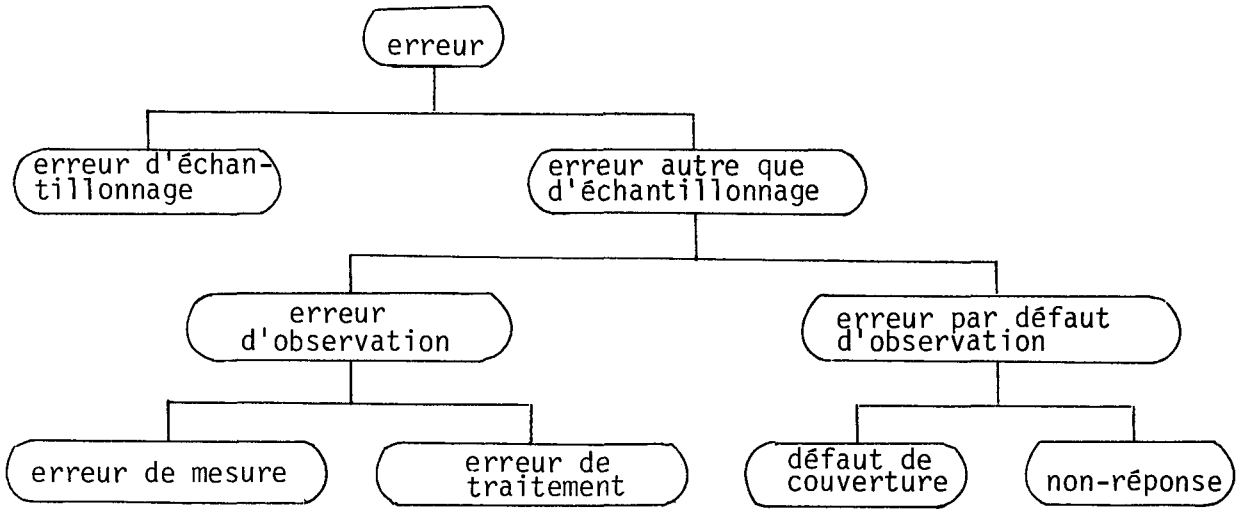
2.1 Terminologie

L'objectif de toute enquête est de déterminer certaines caractéristiques de la population. Par suite de toutes sortes d'erreurs, la véritable valeur de ces caractéristiques ne sera généralement jamais obtenue. Une typologie des sources d'erreur est présentée dans la figure 1, qui est un schéma dû à Kish (1967).

Les deux sources d'erreur, dans les enquêtes, sont les erreurs d'échantillonnage et les erreurs autres que d'échantillonnage.

Les erreurs d'échantillonnage sont constituées par la partie de l'erreur qui est due au fait qu'on observe uniquement un échantillon de valeurs, et non la population totale. L'erreur d'échantillonnage a une distribution de fréquences probables qui est constituée par la totalité des erreurs d'échantillonnage de tous les échantillons possibles de même dimension. On se sert de cette distribution pour estimer la caractéristique de la population que l'on veut étudier.

FIGURE 1. TYPOLOGIE DES ERREURS QUI AFFECTENT LES ENQUETES



Les erreurs autres que d'échantillonnage sont celles que, dans les estimations portant sur un échantillon, on ne peut attribuer aux fluctuations de l'échantillonnage. Elles posent souvent un problème plus sérieux que les erreurs d'échantillonnage. On peut les répartir en erreurs d'observation et erreurs par défaut d'observation.

Les erreurs d'observation sont dues à ce que certaines observations sont obtenues et enregistrées de manière incorrecte. On peut les subdiviser encore en erreurs de mesure et erreurs de traitement.

Les erreurs de mesure sont dues soit à l'enquêteur, soit au répondant. L'enquêteur lui-même peut être une source d'erreur. Il peut influencer la réponse par sa seule présence, par le fait qu'il appartient à tel ou tel des deux sexes, par sa peau, sa couleur, son âge ou sa façon de s'habiller. La manière dont il pose les questions et explicite les réponses affecte également les résultats. La réponse fournie par une personne peut dépendre du type de question (selon qu'il s'agit d'une question mesurant un fait, comme l'année de naissance, ou une opinion). Des erreurs peuvent aussi être introduites par d'autres facteurs, selon par exemple que la personne interrogée comprend ou non la question, qu'elle connaît ou non la réponse, qu'elle tient à garder la réponse pour elle-même, ou qu'elle désire donner d'elle une certaine image. En outre, la mémoire n'est pas toujours exempte d'erreurs.

Les erreurs de traitement se produisent lors du traitement des données par le service statistique, lors du codage, de la mise en tableaux et des calculs statistiques.

Les erreurs par défaut d'observation sont dues à l'impossibilité d'obtenir des observations concernant certaines parties de la population. On peut les subdiviser en erreurs dues à un défaut de couverture et à la non-réponse.

Appelons population-objectif la population sur laquelle l'enquête est censée porter. Des difficultés pratiques qui se présentent dans le cas de certaines parties de la population peuvent mener à les éliminer du champ de l'enquête. Il est possible aussi que la population effectivement sondée contienne des éléments qui ne font pas partie du champ de l'enquête.

Par défaut de couverture on entend toutes les erreurs qui résultent des différences existant entre la population-objectif et la population sondée. Les éléments qui font partie à la fois de la population-objectif et de la population sondée sont les éléments "corrects". Lorsque des éléments de la population-objectif ne figurent pas dans la population sondée, on a un taux de couverture réduit. Ces éléments n'ont aucune chance (probabilité nulle) de figurer dans l'échantillon. Lorsque des éléments de la population sondée ne font pas partie de la population-objectif, on a un taux de couverture excessif (comptages ou observations en trop). Il faut exclure ces éléments de l'échantillon avant de procéder à l'analyse. Si cet excès de couverture est inattendu, la taille de l'échantillon définitif peut être inférieure à celle de l'échantillon prévu.

La non-réponse est le fait de ne pas obtenir d'observations au sujet de certains éléments choisis et désignés pour faire partie de l'échantillon. Une bonne classification des erreurs dues à la non-réponse dépend des conditions dans lesquelles se déroule l'enquête. La classification donnée ci-après est axée essentiellement sur les problèmes rencontrés lors des interrogatoires directs. Un traitement analogue peut être appliqué dans d'autres cas où l'enquête se déroule de manière différente. On peut distinguer entre les diverses catégories ci-après de non réponses:

(1) Absents du foyer. Pour réduire l'ampleur de cette catégorie, on peut procéder à des "rappels". Il y a lieu de faire des recherches sur le nombre optimal de rappels. Il serait utile d'utiliser pour cette catégorie l'expression temporairement indisponible, signifiant que l'entrevue est retardée plutôt que refusée. Le répondant peut être trop occupé, fatigué ou malade lors de la première entrevue, mais il se montrera coopératif à l'occasion d'une nouvelle visite.

(2) Refus. Certaines des causes de refus sont temporaires et peuvent disparaître. Il est possible qu'une personne refuse parce qu'elle est mal disposée ou qu'on la contacte à une heure qui ne lui convient pas. Elle peut fort bien se montrer coopérative lors d'un autre essai ou d'une autre visite. Etant donné cependant qu'un nombre assez important de refus peuvent être considérés comme définitifs, il vaut mieux dans ce cas parler de réponses impossibles à obtenir, pour montrer qu'il s'agit d'un refus catégorique plutôt qu'un retard dans l'observation. Des tentatives répétées n'auraient aucun succès. De ce point de vue on classe dans cette catégorie, plutôt que dans celle des absents du foyer, les répondants que l'on sait être absents pendant toute la période de l'enquête.

(3) Incapacité. Ce terme peut être utilisé dans le cas où la non-réponse est due à une maladie mentale ou physique qui empêche l'enquêté de répondre pendant toute la période de l'enquête. On range aussi dans cette catégorie les non-réponses dues à la méconnaissance de la langue. En généralisant on pourrait grouper cette catégorie avec celle, définie ci-dessus, des non-réponses dues à l'impossibilité d'obtenir des renseignements. Il peut cependant être utile, dans certains cas, de faire la distinction entre les enquêtés qui ne veulent pas répondre et ceux qui le voudraient, mais qui en sont incapables.

(4) Introuvables. Cette catégorie peut être importante, par exemple dans le cas des personnes qui se déplacent constamment. On s'abstient alors de les identifier ou de les suivre parce que cela serait trop coûteux. Dans cette même catégorie générale sont rangés les cas où aucune tentative d'entrevue n'est faite, par exemple pour des raisons d'inaccessibilité (gardien de phare, berger) ou de danger possible (chien de garde, taudis).

(5) Renseignements égarés. Des renseignements peuvent se perdre après une enquête sur le terrain. Certains questionnaires sont parfois inutilisables à cause de leur mauvaise qualité ou parce que le répondant a triché. Il arrive aussi que d'autres aient été perdus ou oubliés.

La typologie décrite ci-dessus est applicable dans la plupart des types d'enquête, mais il faut prendre des précautions dans le cas des plans de sondage complexes. Lorsqu'on effectue par exemple un sondage à plusieurs degrés, cette typologie peut être utilisée à chacun des différents degrés. On peut classer la même source d'erreur différemment aux différents degrés. Donnons un exemple à titre d'illustration: dans une enquête sur les ménages, un échantillon de ménages est d'abord choisi; l'enquêteur dresse la liste de toutes les personnes faisant partie des ménages ainsi choisis et, à partir de cette liste, sélectionne un échantillon. Lors d'une telle énumération l'étudiant qui vit dans une chambre de bonne est souvent "oublié". Au premier degré de la procédure de sondage, cette situation serait considérée comme une erreur de mesure, et au deuxième degré comme une erreur par défaut de couverture.

Pour certaines sources d'erreur, la classification peut dépendre d'autres facteurs. Si une personne appelée à être interrogée meurt avant que l'entrevue ne puisse avoir lieu, la classification dépend de la date du décès. Lorsque le décès est intervenu avant le jour où l'on a choisi l'échantillon, on a alors un taux de couverture excessif, mais s'il s'est produit entre le jour où l'échantillon a été constitué et le jour de l'entrevue, il s'agit alors en fait d'une non-réponse.

Avant de choisir l'échantillon, il faut subdiviser la population en différentes parties que l'on appelle les unités d'échantillonnage. A chaque élément de la population doit correspondre une unité d'échantillonnage et une seule. L'établissement de la liste d'unités d'échantillonnage, appelée base de sondage, soulève souvent un problème

majeur sur le plan pratique. La nature des bases de sondage disponibles est une considération importante dans la constitution de l'échantillon. Parmi les facteurs qui entrent en jeu il y a le type d'unité d'échantillonnage, le taux de couverture, l'exactitude et l'exhaustivité de la liste, ainsi que la quantité de renseignements auxiliaires figurant dans la liste et leur qualité.

Pour les bases de sondage dans lesquelles l'unité d'échantillonnage est une personne, le BCS ne peut disposer que des registres administratifs des autorités locales (municipalités). Pour les enquêtes sur les ménages il peut établir sa propre base de sondage, mais pour le moment il juge approprié d'utiliser la liste des points de distribution de l'administration postale.

2.2 L'ampleur de la non-réponse

Il est assez difficile de comparer les taux de non-réponse enregistrés à l'occasion de différentes enquêtes. Le taux de non-réponse dépend d'un certain nombre de circonstances: but de l'enquête, type d'unité d'échantillonnage, plan de sondage, efficacité des travaux effectués sur le terrain, efficience des enquêteurs, mesures visant à réduire la non-réponse, période au cours de laquelle se déroule l'enquête, population prise comme objectif, longueur du questionnaire, libellé des questions, etc. Même la définition de la non-réponse peut varier. Il est nécessaire d'établir un cadre qui permette de comparer correctement différentes enquêtes. En tenant dûment compte des facteurs qui influent sur le nombre de non-réponses, on peut juger de la qualité des différentes enquêtes. Un tel cadre offre également la possibilité de comparer des enquêtes effectuées dans des pays différents.

Le tableau 2 présente les pourcentages de non-réponse enregistrés à l'occasion d'un certain nombre d'enquêtes effectuées par le BCS. On peut y constater que ce pourcentage manifeste une nette tendance à augmenter.

Tableau 2

Pourcentages de non-réponse dans certaines enquêtes du BSC

Année	EMO		ESC		ECV		ENV		EV	
	tn	rn	tn	rn	tn	rn	tn	rn	tn	rn
1973	13.2									
1974					28.2	15.6				
1975	15.3	9.0	30.1	18.3					14.5	
1976			28.1	18.6	23.0[1]	15.6			12.9	
1977	13.1	6.6	30.9	20.5	29.7	16.9			17.6	9.3
1978			36.1	23.9			33.0	26.2	21.9	12.5
1979	19.7		36.6	24.4	33.7[2]		30.6	23.9	25.5	
1980			36.8	24.7	35.6	19.7	31.1	24.5		

1) personnes âgrées seulement

2) personnes jeunes seulement

tn = pourcentage de non-réponses toutes catégories

rn = pourcentage de refus

EMO = Enquête sur la main-d'oeuvre

ESC = Enquête sur les sentiments des consommateurs

ECV = Enquête sur les conditions de vie

ENV = Enquête nationale sur les voyages

EV = Enquête sur les vacances

Ainsi qu'il a été mentionné plus haut, l'existence d'une relation entre la variable étudiée et l'obtention de réponses réduit la valeur des conclusions de l'enquête. Il n'est pas rare qu'une telle relation existe, comme le montreront les exemples ci-après. Si l'enquête a pour objet de déterminer de quelle manière les gens occupent leurs loisirs, la raison des non-réponses imputables à une "absence du foyer" est alors assez difficile à déterminer, du fait que les personnes

en question sont probablement en train de passer leur temps de loisir quelque part ailleurs. La même chose vaut pour l'enquête sur le nombre d'heures que les gens passent à regarder la télévision: les absents du foyer (le soir) ne sont probablement pas en train de regarder la télévision. L'un des objectifs de l'enquête sur la demande de logements est de mesurer la fréquence avec laquelle les gens déménagent. Comme il y a une quantité considérable de non-réponse dues aux déménagements (l'unité d'échantillonnage est une personne), l'estimation concernant la population totale sera entachée de distorsion. Un certain nombre d'enquêtes montrent que les personnes non mariées ont un taux de réponse plus faible. S'il existe une relation entre la situation matrimoniale et la variable étudiée, alors les estimations seront faussées dans ce cas également.

2.3 Modèles relatifs à l'obtention de réponses

La première chose à faire, lorsqu'on veut mettre au point des théories pour le traitement de la non-réponse, est d'élaborer un modèle mathématique qui décrit la manière dont fonctionne le mécanisme de la non-réponse. Il est souvent question dans les ouvrages techniques de deux modèles de ce genre, que nous appellerons ici le "modèle à taux de réponse aléatoire" et le modèle "à taux de réponse fixe".

D'après le modèle à taux de réponse aléatoire, chaque élément de la population est caractérisé par une certaine probabilité (inconnue) de réponse. Ces probabilités de réponse ne sont pas nécessairement les mêmes pour tous les éléments. Lorsque l'enquêteur se met en rapport avec la personne à questionner, le mécanisme des probabilités entre alors en jeu, ce qui détermine si la personne répondra ou non.

Le modèle à taux de réponse fixe suppose l'existence de deux strates dans la population: une strate de répondants potentiels et une strate de non-répondants potentiels. La dimension et la composition de chaque strate ne sont pas connues d'avance. Elles sont déterminées par les caracté-

ristiques de l'enquête (but, type de questions, techniques d'interrogatoire, enquêteurs, période où s'effectue le travail sur le terrain, etc.). On choisit l'échantillon parmi la population sans tenir compte de ces deux strates. Par conséquent, le nombre de répondants est une variable aléatoire dans l'un et l'autre modèles.

Si, au lieu d'un sondage, on procède à un dénombrement complet, la détermination des répondants reste alors un processus aléatoire dans le cas du modèle à taux de réponse aléatoire, tandis qu'elle serait un processus fixe dans le cas de l'autre modèle. Il y a cependant une certaine ressemblance entre les deux modèles. Si l'on suppose qu'il existe deux mécanismes stochastiques, le mécanisme de l'échantillonnage et le mécanisme de la réponse, les deux modèles ne diffèrent que par l'ordre dans lequel les mécanismes sont appliqués: dans le modèle à taux de réponse fixe, c'est d'abord le mécanisme de la réponse qui entre en jeu pour chaque élément de la population. Cela détermine les deux strates. L'échantillon est choisi ensuite. Dans le modèle à taux de réponse aléatoire, on choisit d'abord l'échantillon, puis le mécanisme de la réponse entre en jeu pour chacun des éléments choisis. Le point de savoir quel modèle il faut utiliser est plus ou moins une question de préférence personnelle. Le modèle à taux de réponse aléatoire donne la possibilité d'estimer les probabilités de réponse. On peut se servir de ces probabilités dans les procédures d'ajustement, ou bien les rattacher à des caractéristiques personnelles. Le modèle à taux de réponse fixe aboutit généralement à des formules plus simples. La théorie fondée sur ce modèle est conditionnée par la composition effective des strates de répondants et de non-répondants. Par conséquent, on peut calculer le degré d'exactitude des estimations, mais non déterminer l'exactitude de la méthode d'estimation. A cause de ce dernier argument, les recherches portent essentiellement sur le modèle à taux de réponse aléatoire.

3. CHOIX DES VARIABLES AUXILIAIRES

3.1 Variables auxiliaires

Il importe de découvrir s'il existe éventuellement une relation entre la variable étudiée et l'obtention de réponses. Il n'est cependant pas possible de dégager cette relation au moyen des données de sondage, car les valeurs de la variable étudiée ne sont pas connues dans le cas des non-répondants. Pour pouvoir dire quelque chose au sujet des non-répondants, on doit avoir des renseignements à leur sujet. Une source d'information concernant les non-réponses est constituée par les variables auxiliaires. Il s'agit de variables que l'on peut mesurer pour les répondants et pour les non-répondants. Il existe deux types de renseignements auxiliaires:

(1) les renseignements que l'enquêteur peut recueillir sans procéder à un interrogatoire direct, comme par exemple le type de ville, le type de logement, l'année (approximative) de construction du logement et le statut social des personnes vivant dans le voisinage;

(2) les renseignements qu'il est possible d'obtenir dans les archives administratives, concernant notamment l'âge, le sexe et la situation matrimoniale.

L'analyse de la relation entre les variables auxiliaires et le taux de réponse jette quelque lumière sur le groupe des personnes qui ne répondent pas. Elle peut renseigner aussi sur la relation existant entre la variable étudiée et le taux de réponse. Les variables auxiliaires qui font ressortir une relation nette avec le taux de réponse jouent un rôle important dans les procédures d'ajustement, ainsi qu'il en sera question plus loin.

Il est admis que les variables auxiliaires sont des variables nominales, c'est-à-dire que des valeurs différentes de ces variables servent uniquement à faire une distinction entre différents groupes. Il n'est pas permis de faire entrer ces valeurs, qui ne sont en fait que des étiquettes dans des opérations arithmétiques.

L'hypothèse selon laquelle les variables sont nominales n'est en pratique pas une restriction. De nombreuses variables sont nominales et d'autres types de variables peuvent facilement être réexprimées sous forme de variables nominales. A titre d'exemple illustrant la quantité de renseignements auxiliaires disponibles, nous donnons ci-après la liste des variables auxiliaires utilisées dans l'enquête de 1977/78 sur la demande de logements:

1) année de naissance
2) sexe
3) situation matrimoniale
4) dimension de la famille
5) structure de la famille
6) type de logement
7) nombre d'étages dans le logement
8) année de construction du logement
9) municipalité
10) quartier de la ville
11) degré d'urbanisation

3.2 Méthodes graphiques

Comme instrument préliminaire pour la sélection des variables auxiliaires, on a mis au point des méthodes graphiques. L'avantage de ces méthodes est évident. Elles mettent en lumière des faits et des relations qui ne sont pas immédiatement visibles et elles peuvent stimuler et faciliter l'analyse. Elles permettent souvent de comprendre plus complètement et d'une manière mieux équilibrée que des tableaux ou des textes explicatifs. En outre, les relations qui s'en dégagent sautent plus clairement aux yeux et sont plus facilement inscrites dans la mémoire (voir Schmid, 1954). Deux méthodes graphiques simples sont présentées dans les sections qui suivent: le diagramme à cases et le diagramme en ailes de moulin.

3.2.1. Le diagramme à cases

Le diagramme à cases peut être considéré comme une généralisation de l'histogramme ou du diagramme en bâtons. Ce nom lui a été donné à cause de sa forme (voir la figure 2).

Un rectangle de largeur standard et de hauteur proportionnelle à la taille de l'échantillon représente l'échantillon. Le rectangle est divisé en un certain nombre de bandes (correspondant aux catégories dans lesquelles peut entrer la variable auxiliaire). La hauteur d'une bande particulière est proportionnelle au nombre d'éléments de l'échantillon qui figurent dans la catégorie correspondante. Chaque bande est divisée par une ligne verticale en une case gauche (la réponse) et une case droite (la non-réponse). La grandeur de ces deux cases est proportionnelle au nombre de réponses et de non-réponses, respectivement dans la catégorie considérée. La figure 3 donne un exemple de diagrammes à cases. Les données proviennent de l'enquête de 1977/78 sur la demande de logements à Amsterdam. La variable auxiliaire est la situation matrimoniale de la personne figurant dans l'échantillon.

FIGURE 2. LE DIAGREMME A CASES

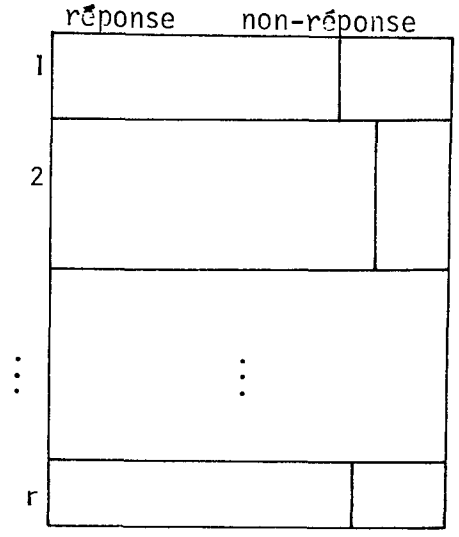
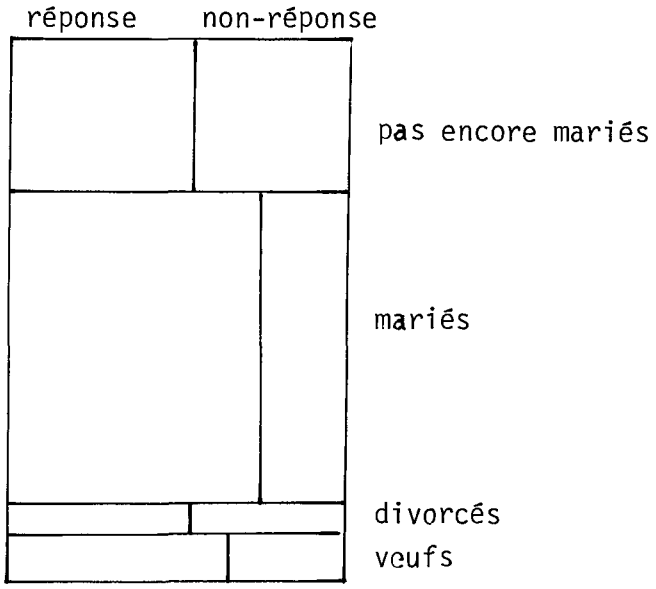


Figure 3. DIAGRAMME A CASES INDIQUANT LA SITUATION MATRIMONIALE DES ENQUETES LORS DE L'ENQUETE DE 1977/78 SUR LA DEMANDE DE LOGEMENTS A AMSTERDAM



Il peut être bon d'appeler l'attention sur un certain nombre de points:

(1) La hauteur des bandes indique dans quelle mesure les diverses catégories contribuent à l'échantillon. De toute évidence, une forte proportion de la population est mariée. La catégorie la plus petite est celle des personnes divorcées.

(2) L'ampleur de la non-réponse peut être évaluée d'après la distance entre les lignes de partage verticales et le côté droit du rectangle. Dans le présent exemple, il y a manifestement une quantité considérable de non-réponses.

(3) Si toutes les lignes de partage constituent approximativement une droite, il n'y a pas de relation entre le taux de réponse et la variable auxiliaire. Dans le présent cas, il y a manifestement une relation: les personnes mariées répondent mieux que les autres. Le taux de réponse est plus faible dans le groupe des personnes non encore mariées et dans celui des divorcés.

On trouvera de plus amples détails au sujet du diagramme à cases dans Bethlehem et Kersten (1981).

3.2.2. Le diagramme en ailes de moulin

Le diagramme en ailes de moulin est une représentation graphique des résultats de l'analyse des correspondances. L'analyse des correspondances est une technique d'analyse des associations existant dans les tableaux à double entrée (voir par exemple Benzécri, 1976). Les lignes du tableau (catégories dans lesquelles entre la variable faisant l'objet d'une tabulation verticale) et les colonnes (catégories dans lesquelles entre la variable faisant l'objet d'une tabulation horizontale) sont représentées géométriquement. Cette représentation géométrique contient toutes les informations disponibles au sujet des associations existant dans le tableau. Par un système de mise à échelle, on affecte aux lignes et aux colonnes des valeurs réduites telles que le coefficient de corrélation calculé au moyen de ces valeurs soit porté à son maximum. A chacune des cases du tableau correspondent deux valeurs réduites: une pour la ligne et une pour la colonne. Si l'on conçoit ces valeurs comme des coordonnées, on peut établir un graphique correspondant au tableau. Cela donne une

grille de points irrégulièrement espacés, qu'il n'est pas toujours facile d'interpréter. Pour simplifier l'interprétation, on porte sur le graphique des droites de régression au lieu des points eux-mêmes. En raison des propriétés particulières des valeurs réduites, la droite de régression expliquant les valeurs de y par rapport aux valeurs de x, dans le graphique, a la forme simple suivante:

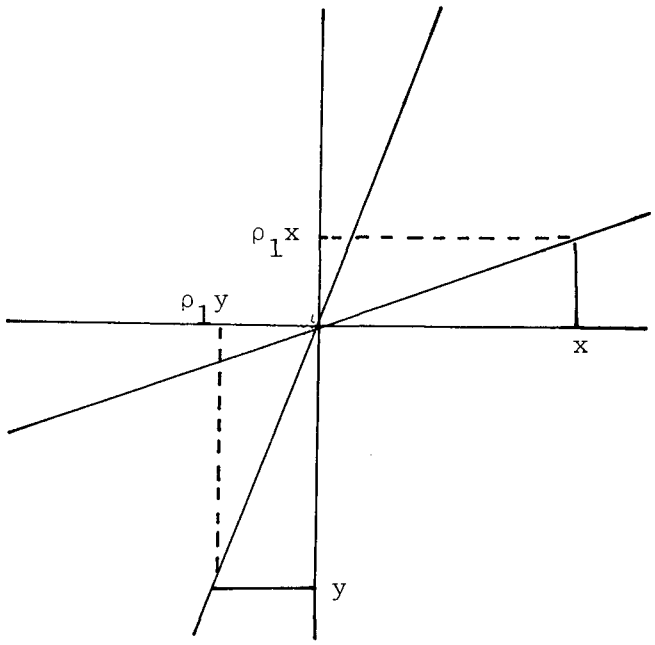
$$y = \rho_1 x \tag{1}$$

et la droite de régression expliquant les valeurs de x par rapport aux valeurs de y a la forme

$$x = \rho_1 y \tag{2}$$

ρ_1 étant le coefficient de corrélation maximalisé. En portant sur le graphique les deux droites de régression, on obtient le diagramme en ailes de moulin - voir la figure 4.

FIGURE 4. LE DIAGRAMME EN AILES DE MOULIN



Il peut être bon de faire un certain nombre de remarques:

(1) L'origine représente les deux distributions marginales du tableau.

(2) Les valeurs réduites proches de l'origine représentent les catégories qui ressemblent à la distribution marginale et qui ont donc un comportement régulier. Les valeurs éloignées de l'origine représentent des catégories ayant un comportement différent.

(3) La relation entre les deux variables est forte si les deux droites de régression sont voisines de la droite à coefficient angulaire de 45°.

(4) La projection d'une variable appartenant à une catégorie à comportement différent sur l'axe de l'autre variable, par l'intermédiaire de la droite de régression, donne une idée des associations existant entre les diverses catégories dans lesquelles entrent les variables.

Le diagramme tel qu'il est décrit ci-dessus ne peut rendre compte de tous les renseignements figurant dans le tableau. Il les explique dans toute la mesure où cela est possible avec un graphique à deux dimensions. Sous certaines conditions on peut superposer au premier diagramme un second graphique qui rend compte autant que possible des renseignements non encore expliqués. Si besoin est on peut même en construire davantage, mais de préférence un seul suffit pour expliquer la plus grande partie des associations.

On peut établir au total s graphiques de ce genre, s étant inférieur d'une unité au nombre de lignes, ou au nombre de colonnes si ce dernier est moins élevé. Désignons par $\rho_1, \rho_2, \ldots, \rho_s$ les coefficients de corrélation maximalisés. Etant donné que

$$\sum_{i=1}^{s} \rho_i^2 = X^2/N, \qquad (3)$$

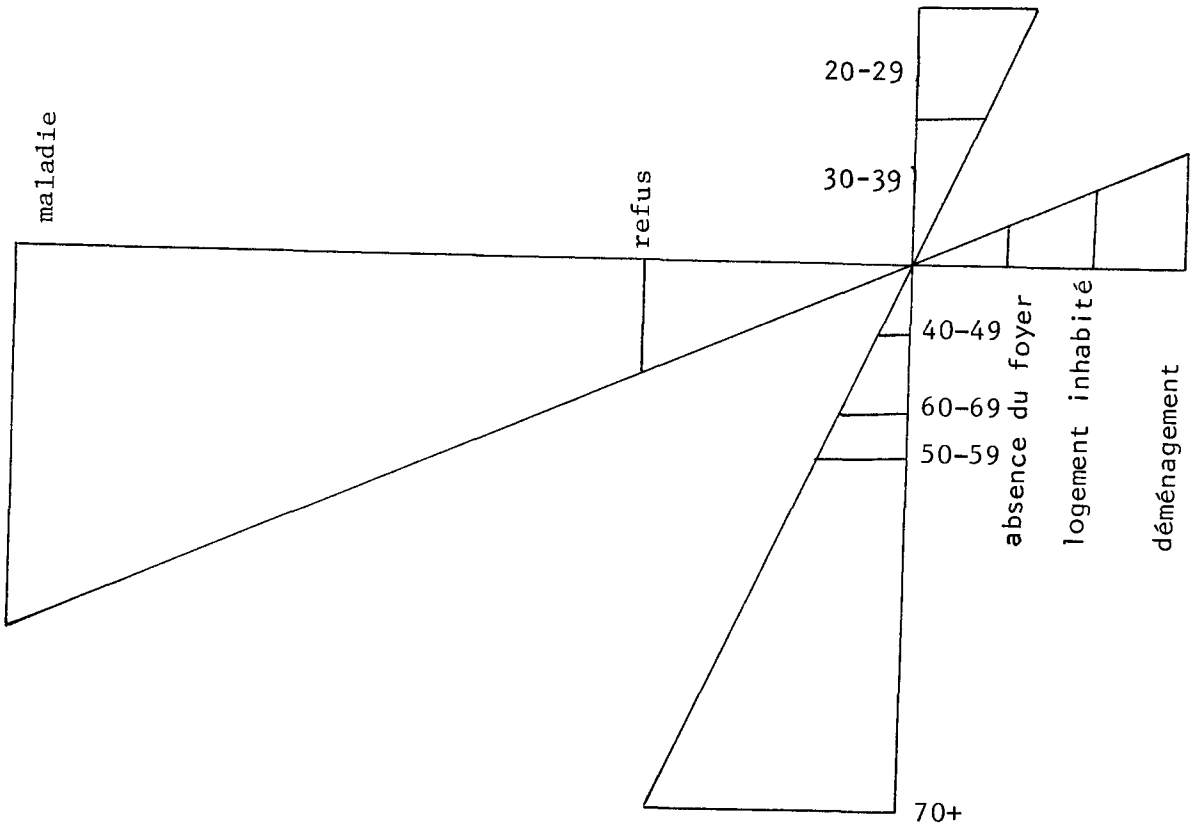
où X^2 est la valeur de chi-carré pour le tableau et N le total général,

$$\tau_i = N\rho_i^2/X^2 \qquad (4)$$

mesure la quantité d'information expliquée par le i-ème graphique ($i = 1,2,\ldots,s$).

La figure 5 représente le premier diagramme en ailes de moulin établi pour deux variables: l'âge (six catégories) et le type de non-réponse (cinq catégories) au titre de l'Enquête de 1977/78 sur la demande de logements à Amsterdam.

FIGURE 5. DIAGRAMME EN AILES DE MOULIN INDIQUANT L'ASSOCIATION ENTRE L'ÂGE ET LE TYPE DE NON-REPONSE DANS L'ENQUÊTE DE 1977/78 SUR LA DEMANDE DE LOGEMENTS A AMSTERDAM



Ce graphique contient environ 88% des informations existant au sujet des associations dans le tableau ($\tau_1 = 0,88$). Les principales raisons qui expliquent la non-réponse, dans le cas des personnes âgées, sont le refus et la maladie. Dans le cas des jeunes, la non-réponse est due à l'impossibilité d'entrer en rapport avec les enquêtés: logement inhabité, absence du foyer et déménagement. On trouvera de plus amples détails sur l'application de l'analyse des correspondances dans Bethlehem et Kersten (1980).

3.3. Autres méthodes de sélection

Il y a de nombreuses autres méthodes, essentiellement non graphiques, qui servent à déterminer l'association existant entre les variables auxiliaires et l'obtention de réponses. On peut trouver dans Bishop, Fienberg & Holland (1975), par exemple, d'amples indications au sujet de l'association dans les tableaux de contingence.

Une méthode couramment utilisée pour la sélection des variables auxiliaires les plus importantes est la méthode AID (Automatic Interaction Detection), décrite par Morgan & Sonquist (1963). Par étapes successives on détermine les variables auxiliaires qui peuvent expliquer une aussi grande partie que possible de la variance caractérisant la variable binaire réponse/non-réponse. Il y a à cette méthode des inconvénients qui rendent sa fiabilité douteuse. Comme le processus de sélection se fait par étapes, on n'a aucune garantie de trouver la solution optimale. Du fait qu'il n'y a aucune règle fondée sur un modèle statistique qui indique à quel moment il faut arrêter le processus, le résultat est, en ce sens, également, assez arbitraire. De plus amples recherches dans ce domaine sont nécessaires; voir par exemple Kass (1980).

4. REDUCTION DE LA DISTORSION DUE A LA NON-REPONSE AU MOYEN D'UNE PONDERATION PAR SOUS-GROUPES

Lorsque l'on constate ou que l'on soupçonne l'existence d'une relation entre la variable étudiée (Y) et l'obtention de réponses (R), il faut prendre des mesures pour réduire la distorsion imputable à la non-réponse. Il sera question dans la présente section d'un certain nombre de procédures d'ajustement qui sont fondées sur la pondération par sous-groupes. L'attention se portera essentiellement sur l'estimation de la moyenne de Y pour l'ensemble de la population.

On peut montrer que la distorsion introduite par le fait de n'utiliser que les valeurs tirées des réponses est proportionnelle à la covariance entre Y et R. S'il est possible de diviser la population en un certain nombre de sous-groupes pour lesquels la covariance est dans chaque cas négligeable, on peut alors combiner les estimations (pratiquement sans distorsion) des moyennes des différents sous-groupes en une estimation (pratiquement sans distorsion) de la moyenne pour l'ensemble de la population.

Considérons que la population finie se compose de N éléments U_1, U_2, .., U_N pour lesquels les valeurs de Y sont Y_1, Y_2, ...,Y_N . Dans cette population on choisit, sans remise, un échantillon aléatoire simple $\underline{u}_1$, $\underline{u}_2$, .., $\underline{u}_n$ (les variables aléatoires sont soulignées de dimension n. Les valeurs correspondantes de Y sont $\underline{y}_1$, $\underline{y}_2$, .., $\underline{y}_n$ et l'obtention ou la non-obtention d'une réponse est indiquée par $\underline{r}_1$, $\underline{r}_2$, ..., $\underline{r}_n$ (r_i = 1 indiquant qu'il y a une réponse et $\underline{r}_i$ = 0 indiquant une non-réponse). En fait, on ne peut observer $\underline{y}_i$ que dans le cas des éléments $\underline{u}_i$ de l'échantillon pour lesquels $\underline{r}_i$ = 1. Les $\underline{m}$ éléments qui répondent sont dénotés $\underline{u}_1^*$, $\underline{u}_2^*$, .., $\underline{u}_m^*$ ($\underline{m}$ = $\underline{r}_1$ + $\underline{r}_2$ + .., + $\underline{r}_n$), et pour ces éléments les valeurs de Y sont $\underline{y}_1^*$, $\underline{y}_2^*$, .., $\underline{y}_m^*$. Soit X une variable auxiliaire entraînant une division de la population en H sous-groupes de dimensions N_1, N_2 .., N_H. Dans la pondération par sous-groupes, on calcule tout d'abord dans chaque sous-groupe h un estimateur $\overline{\underline{y}}_h^*$ de la moyenne pour le sous-groupe:

$$\bar{y}_h^* = \frac{1}{m_h} \sum_{i=1}^{m_h} y_{hi}^*, \qquad (h = 1, 2, .., H) \qquad (5)$$

où $y_{h1}^*, y_{h2}^*, .., y_{hm_h}^*$ sont les valeurs de m_h éléments de sous-groupe h qui répondent. On combine ensuite les estimateurs des différents sous groupes, $\bar{y}_1^*, \bar{y}_2^*, .., \bar{y}_H^*$ en un seul estimateur $\bar{y}^*$ pour l'ensemble de la population.

$$\bar{y}^* = \sum_{h=1}^{H} w_h \bar{y}_h^* \qquad (6)$$

Le type d'estimateur est déterminé par la quantité d'informations qui est disponible au sujet des poids $w_1, w_2, \ldots, w_H$.

Si l'on connaît les dimensions $N_1, N_2, \ldots, N_H$ des sous-groupes, la situation équivaut à une post-stratification (voir par exemple Holt & Smith, 1979). Les poids ne sont pas des quantités aléatoires, mais des quantités bien déterminées:

$$w_h = \frac{N_h}{N} \qquad (h = 1, 2, .., H) \qquad (7)$$

Si ces dimensions ne sont pas connues, on peut les estimer en appliquant la formule

$$w_h = \frac{n_h}{n}, \qquad (h = 1,2,..,H) \qquad (8)$$

dans laquelle n_h est le nombre d'éléments du sous-groupe h qui figurent dans l'échantillon ($n = n_1 + n_2 + \ldots + n_H$).

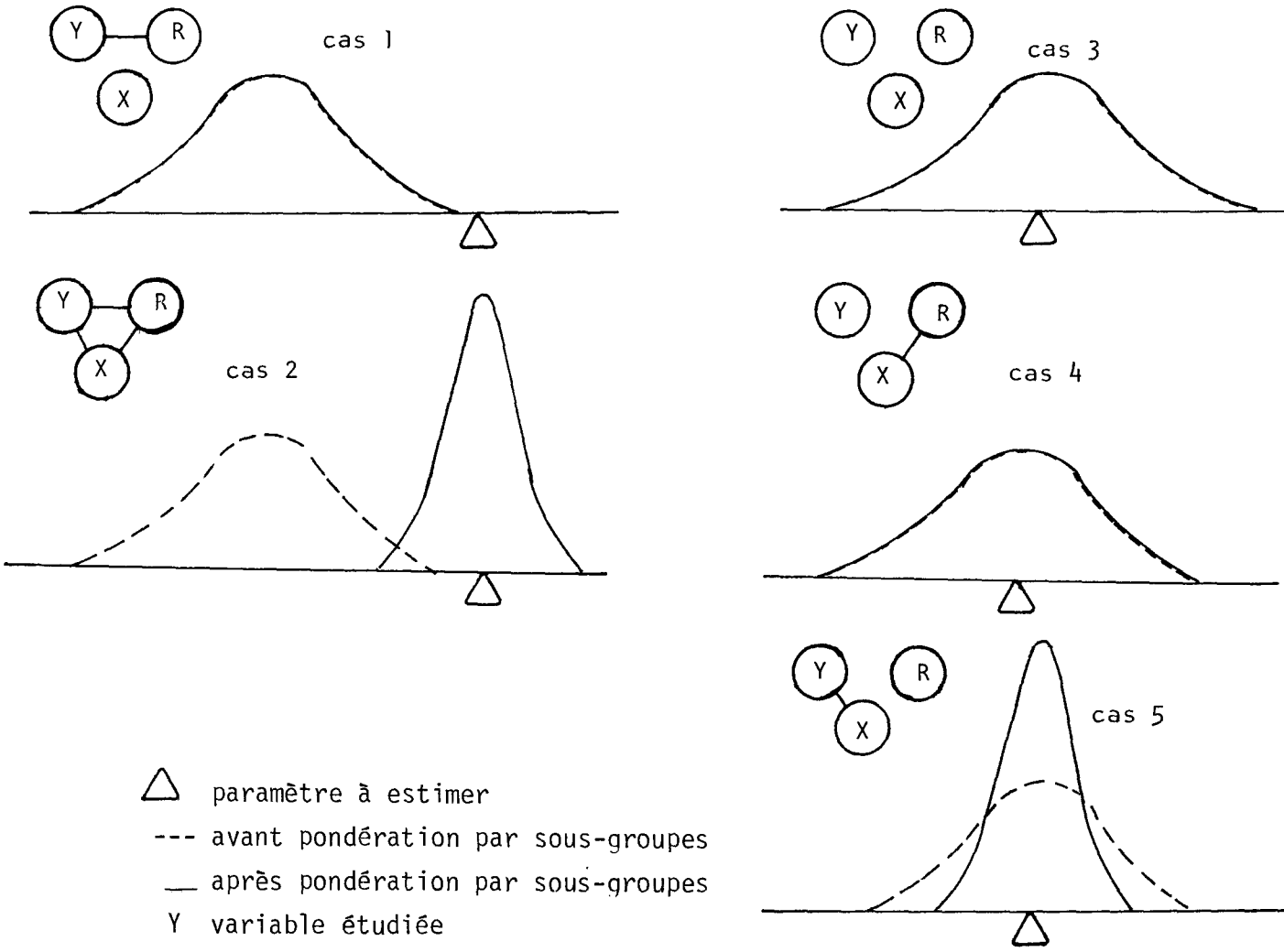
Dans une situation intermédiaire où l'on utilise deux variables auxiliaires X_1 et X_2 et où l'on ne connaît que les totaux marginaux des deux variables, on peut appliquer une autre méthode pour estimer les poids (voir par exemple Chapman, 1976). Supposons que X_1 donne G

groupes et que X_2 donne H groupes. En combinant X_1 et X_2 on obtient une subdivision en G x H groupes. Si l'on ne connaît que les totaux marginaux N_{g+} (g=1, 2, ..., G) de X_1 et N_{+h} (h = 1,2,...,H) de X_2, on peut alors calculer de bonnes estimations $\underline{N}_{gh}$ de N_{gh} en utilisant les renseignements que donne l'échantillon. Les poids sont alors égaux à:

$$\underline{w}_{gh} = \frac{\underline{N}_{gh}}{N} \qquad (g=1,2,..,G;\ h=1,2,..,H) \qquad (9)$$

Quand on utilise ces trois estimateurs avec la même répartition en sous-groupes, ils ont tous la même distorsion, mais plus on dispose de renseignements au sujet de la dimension des sous-groupes, plus la variance de l'estimation sera faible. La pondération par sous-groupes a deux avantages: elle réduit la variance de l'estimation et elle réduit la distorsion imputable à la non-réponse. La figure 6 illustre les cas extrêmes. Si deux variables sont reliées entre elles, cela signifie qu'elles ont une forte corrélation.

FIGURE 6. VARIANCE ET DISTORSION DES ESTIMATEURS AVANT ET APRES PONDERATION PAR SOUS-GROUPES



On peut en tirer un certain nombre de conclusions:

(1) S'il existe une distorsion imputable à la non-réponse, la pondération par sous-groupes est significative lorsqu'il y a corrélation entre X et R (cas 2). La distorsion et la variance sont toutes deux réduites .

(2) S'il n'y a pas de distorsion imputable à la non-réponse, une corrélation entre X et R n'a pas d'effet (cas 4). Seule une corrélation entre X et Y réduit la variance (cas 5).

Faute d'avoir des données sur les non-répondants, il est impossible d'utiliser les données restantes pour trouver une variable auxiliaire X qui soit en forte corrélation avec Y. On peut cependant s'en servir pour rechercher des variables auxiliaires qui soient en forte corrélation avec le taux de réponse, R. Si l'on peut trouver une variable de ce genre, son utilisation pour la pondération par sous-groupes réduira la distorsion imputable à la non-réponse (lorsqu'elle existe), mais pas toujours la variance.

5. AUTRES METHODES D'AJUSTEMENT

Plusieurs autres méthodes d'ajustement sont exposées dans les ouvrages spécialisés. Il sera question de certaines d'entre elles dans la présente section. Il faut de plus amples recherches, dans le cas de telle ou telle d'entre elles, pour en dégager le véritable intérêt.

5.1. Pas d'ajustement

Il arrive qu'aucun ajustement ne soit nécessaire. Si l'on est certain qu'il n'existe aucune relation entre la variable étudiée et l'obtention de réponses, on peut considérer les réponses comme un échantillon aléatoire de la population. Dans le cas également où le résultat indiqué est réputé valoir uniquement pour la population de répondants potentiels, aucune correction n'est nécessaire. Dans toutes les autres situations, l'absence d'ajustement ne se justifie que si la catégorie "non-réponse" figure dans tous les tableaux publiés.

5.2. Imputation

L'imputation permet de résoudre le problème de la non-réponse en remplaçant les observations manquantes par des valeurs tirées des données provenant des réponses obtenues à l'occasion de l'enquête en cours, ou bien de données provenant d'une enquête antérieure. Si la structure des réponses, dans l'enquête en cours, est analogue à celle de l'enquête antérieure, les résultats des deux modes d'imputation seront à peu près les mêmes. L'imputation peut se faire de différentes manières. En voici quelques-unes:

(1) imputation de la valeur observée pour un répondant pris au hasard

(2) imputation de la valeur moyenne observée pour l'ensemble des répondants

(3) imputation de la valeur observée pour un répondant pris au hasard dans le même sous-groupe

(4) imputation de la valeur moyenne observée pour les répondants du même sous-groupe

(5) imputation d'une valeur obtenue par ajustement d'un modèle

(6) imputation de limites supérieures ou inférieures.

Les méthodes (1) et (2) ne réduisent pas la distorsion. Les méthodes (3) et (4) ressemblent à la pondération par sous-groupes. L'effet de la méthode (5) dépend pour beaucoup de la précision du modèle et de la qualité des hypothèses sur lesquelles il est fondé. La méthode (6) donne une idée de ce que pourrait être l'erreur si aucun ajustement n'était effectué.

5.3. Ajustement pour tenir compte des "absents du foyer"

La méthode bien connue de Politz et Simmons (1949) tente de remédier à la distorsion imputable aux absents du foyer en estimant la probabilité qu'il y a de trouver une personne chez elle. Pour cela on demande par exemple aux répondants combien de fois ils étaient chez eux à l'heure de l'entrevue au cours des jours précédents. La probabilité qu'ils soient chez eux, calculée de cette manière, peut construire un modèle expliquant la relation entre la variable étudiée et la probabilité de trouver les répondants chez eux. L'extrapolation de ce modèle au groupe des absents du foyer donne parfois davantage de renseignements au sujet de ce groupe.

5.4. Ajustement pour tenir compte des refus

Il est possible de déterminer dans quelle mesure les gens sont disposés à coopérer à l'enquête (voir Van Tulder, 1977). En utilisant ce renseignement, on peut appliquer une méthode d'ajustement analogue à celle qui est utilisée pour les absents du foyer. De plus, la volonté de coopérer est une mesure du climat dans lequel se déroule l'enquête. La construction d'une échelle de valeurs qui puisse renseigner à ce sujet est probablement plus difficile que dans le cas de l'ajustement appliqué aux absents du foyer.

5.5. Double échantillonnage

Pour recueillir davantage de renseignements au sujet des non-répondants, Hansen et Hurwitz (1946) ont proposé de choisir un échantillon parmi les non-répondants. Des enquêteurs spécialement formés essayent d'obtenir malgré tout les renseignements (ou une partie des renseignements) qui manquent. Les contraintes imposées par le manque de temps et d'argent empêchent souvent de recourir au double échantillonnage.

5.6. La principale question

Si la méthode de Hansen et Hurwitz est trop coûteuse, on peut recourir en remplacement à la méthode de la principale question. Dans la plupart des enquêtes il y a souvent une question de base importante autour de laquelle l'enquête a été centrée. Si au cours des travaux sur le terrain on a des problèmes pour faire remplir le questionnaire, l'enquêteur peut essayer d'obtenir une réponse portant uniquement sur la principale question. On peut même essayer de recourir à cette méthode ultérieurement par lettre ou par téléphone. Cette technique sera testée sous peu dans une des enquêtes du BCS.

6. CONCLUSIONS

Par suite de l'augmentation des taux de non-réponse qui a été constatée ces dernières années, il importe d'effectuer des recherches approfondies en ce qui concerne l'incidence de la non-réponse sur la qualité de l'enquête.

Les ouvrages techniques exposent un assez grand nombre de méthodes d'ajustement qui visent toutes à réduire la distorsion imputable à la non-réponse. Une étude comparative de ces méthodes doit fournir des réponses décisives au sujet de leurs avantages.

Les grandes différences qui existent en ce qui concerne l'objectif, la conception et l'exécution des enquêtes empêchent d'interpréter comme il convient les différences constatées dans les chiffres de non-réponse. Il est par conséquent nécessaire de mettre en place un cadre théorique qui permette d'effectuer une comparaison appropriée.

Bien entendu, la suppression de la non-réponse au cours des travaux sur le terrain restera un sujet important.

Références

Benzécri, J.P., 1976, L'analyse des Données (Dunod, Paris)

Bethlehem, J.G. & H.M.P. Kersten, 1981, Graphical Methods in Non-response Analysis and Sample Estimation (Staatsuitgeverij, The Hague)

Bishop, Y.M.M., S.E. Fienberg & P.W. Holland, 1975, Discrete Multivariate Analysis (MIT Press, Cambridge)

Champman, D.W., 1976, A survey of Non-response Imputation Procedures, Proceedings of the American Statistical Association, social statistics section, pp 245-251

Hansen, M.H. & W.N. Hurwitz, 1946, The Problem of Non-response in Sample Surveys, Journal of the American Statistician 41, pp 517-529

Holt, D. & T.M.F. Smith, 1979, Post Stratification, Journal of the Royal Statistical Society, series A, 142, pp 33-46

Kass, G.V., 1980, An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics 29, pp 119-127

Kish, L., 1967, Survey Sampling (Wiley, New York)

Morgan, J.N. & J.A. Sonquist, 1963, Problems in the Analysis of Survey Data, Journal of the American Statistical Association 58, pp 415-434

Politz, A. & W. Simmons, 1949, An Attempt to get the Not-at-homes into the Sample without Callbacks, Journal of the American Statistical Association 44, pp 9-31

Schmid, C.F., 1954, Handbook of Graphic Presentation (Ronald Press, New York)

Références (suite)

Tulder, J.J.M. van, 1977, Op de grens van non-response, Jaarboek van de Nederlandse Vereniging van Marktonderzoekers 1977, pp 43-52

Widdershoven, M. & J. van den Berg, 1980, Non-respons bij twee "persoons- en gezinsenquêtes", in: CBS Select 1 (Staatsuitgeverij, The Hague).