Statistique Canada Statistics Canada

TECHNIQUES D'ENQUÊTE

décembre 1981

volume 7

numéro 2

Préparé par les méthodologistes de Statistique Canada

Canada

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA
1010200655

DATE DUE
DATE DE RETOUR

FEB 25 1986 FEV			
[redacted]			
FEB 9 1996 FEV			
SEP 12 1996			
OCT 10 1996			
NOV 13 1996			

LOWE-MARTIN No. 1137

TECHNIQUES D'ENQUÊTE

décembre 1981 Vol. 7 No. 2

Preparé par les méthodologistes de Statistique Canada.

TABLE DES MATIÈRES

L'inférence statistique basée sur des plans d'échantillonage complexes
GAD NATHAN 109

Le problème de la non-reponse
J.G. BETHLEHEM et H.M.P. KERSTEN 131

Les variances d'estimateurs asymptotiquement normaux basés sur des enquêtes complexes
DAVID A. BINDER 162

La statistique de la santé au Canada: rétrospective et jalons pour l'avenir
LORNE ROWEBOTTOM 179

Modèles d'estimation des erreurs d'échantillonnage
P.D. GHANGURDE 187

8-3200-501
No. de références
Z-079

ISSN: 0714-0045

TECHNIQUES D'ENQUÊTE

décembre 1981 Vol. 7 No. 2

Preparé par les méthodologistes de Statistique Canada

Comité de rédaction:

R. Platek - Président
M.P. Singh - Rédacteur en chef
P.F. Timmons

J.H. Gough - Rédacteur adjoint

Politique de la rédaction:

La revue "Techniques d'enquête" veut donner aux personnes qin s'intéressent aux aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration de méthodes d'enquête: les problèmes de conception causés par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de documents pour publication

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement, et d'enqûetes ménages Statistique Canada, 6e etage, Edifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi.

TECHNIQUES D'ENQUÊTE 1981, VOL. 7 NO. 2

L'INFÉRENCE STATISTIQUE BASÉE SUR DES PLANS D'ÉCHANTILLONNAGE COMPLEXES

Gad Nathan[1]

Les problèmes d'inférence statistique basée sur des plans d'échantillonnage complexes sont décrits. La définition du paramètre d'intérêt est de grande importance et on doit décider si on veut estimer un paramètre de la population à nombre fini ou un paramètre du modèle de superpopulation. Les méthodes générales basées sur la statistique généralisée de Wald et sa modification, aussi bien que la modification des statistiques classiques sont présentées. Les méthodes spécifiques pour la régression et les modèles linéaires et pour l'analyse des données qualitatives sont décrites en détail.

1. INTRODUCTION

Règle générale, l'application des méthodes classiques d'inférence, telles que la régression, l'analyse de variance ou les tests d'hypothèses, suppose que les observations proviennent d'un échantillon aléatoire simple d'une population infinie ayant une distribution de probabilité particulière à une famille hypothétique. La grande diffusion des programmes informatiques a rendu l'utilisation de ces méthodes particulièrement facile. Toutefois, il est presque impossible de les appliquer telles quelles à des données recueillies au moyen de plans d'échantillonnage complexes.

---

1 G. Nathan, Hebrew University, Jérusalem et Bureau de la statistique d'Israel.

Dans le présent article, nous tentons de donner quelques conseils pratiques sur ce qui peut et ne doit pas être fait dans ces cas. Cet examen est fondé sur des travaux récents (voir bibliographie) portant sur ces questions et comprend plusieurs exemples d'applications possibles.

Quiconque veut entreprendre une analyse statistique doit tout d'abord définir les paramètres qu'il faut estimer. Pour ce faire, on peut s'inspirer des travaux de Brewer et Mellor (1973) et de Smith (1976). De leur côté, Kish et Frankel (1974) soutiennent qu'une inférence valable doit porter sur des paramètres de la population finie, notamment le coefficient de régression d'une population:

$$B = \sum_{i=1}^{N} (X_i-\bar{X})(Y_i-\bar{Y}) \Big/ \sum_{i=1}^{N} (X_i-\bar{X})^2 ,$$

ou les coefficients de corrélation multiple ou partielle ou d'autres mesures qui sont définies par rapport à la population finie seulement, sans utiliser un modèle de superpopulation. Dans ce cas, l'inférence est basée sur un plan d'échantillonnage (Särndal, 1978), c'est-à-dire qu'elle est fondée uniquement sur les propriétés de la distribution d'échantillonnage. Cependant, on peut aussi utiliser l'inférence basée sur un modèle pour estimer un paramètre de la population finie (Hartley et Sielken, 1975).

Il existe une autre possibilité d'estimation dans ce domaine, celle que propose par exemple Fienberg (1980). Selon lui, l'inférence doit porter sur les paramètres d'une distribution de probabilité (superpopulation) dont la population finie constitue la réalisation. On peut trouver des exemples de ce type d'inférence dans les travaux de Konijn (1962), Fuller (1975), Thomsen (1978) et Pfeffermann et Nathan (1981). Si les paramètres se rapportent à un modèle de superpopulation, l'inférence ne peut être basée uniquement sur un plan d'échantillonnage; elle doit être fondée sur un modèle (Särndal, 1978) ou encore, sur un modèle et un plan d'échantillonnage en même temps. Si l'on suppose qu'il y a indépendance entre la distribution du modèle et la distribution d'échantillonnage, l'inférence classique (basée sur un modèle) est alors valide et le plan d'échantillonnage peut seul avoir un effet sur l'efficacité de l'inférence.

Ces deux démarches peuvent soulever de vives objections, car la méthode basée sur un modèle repose en grande partie sur des hypothèses appliquées à un modèle théorique qui ne peuvent généralement pas être garanties; par voie de conséquence, l'inférence ne sera pas robuste s'il y a déviation de ce modèle. Par ailleurs, dans le cas de l'inférence basée sur un plan d'échantillonnage, les paramètres d'une population finie sont habituellement des "copies" des paramètres d'un modèle théorique qui présentent une valeur descriptive très faible, à moins qu'il ne s'agisse d'un modèle de base. Par exemple, le coefficient de corrélation d'une population finie peut être une mesure utile de la relation qui existe entre deux variables, mais seulement si cette relation est approximativement linéaire.

Dans de nombreux cas, il est préférable de chercher un certain équilibre entre ces deux méthodes d'inférence. Il est possible d'y arriver si l'on utilise uniquement les paramètres d'une population finie qui constituent des éléments approximatifs des paramètres de la superpopulation d'un modèle pertinent auquel les données peuvent être appliquées. Par exemple, lorsque des équations de régression distinctes sont ajustées aux sous-ensembles appropriés, on obtient alors un meilleur ajustement linéaire que si l'on procède à une régression générale. Si les sous-ensembles sont de taille assez grande, les coefficients de régression de la population finie pourront mieux s'apparenter aux paramètres de la superpopulation et, ainsi, l'inférence portant sur les paramètres d'une population finie pourra être considérée comme caractérisant les paramètres de la superpopulation.

Pour pouvoir assurer la meilleure correspondance possible entre les paramètres d'un modèle et ceux d'une population finie, il faut procéder à une analyse exploratoire en profondeur du modèle avant d'aborder l'analyse théorique. Cette analyse exploratoire des divers modèles alternatifs peut, dans la plupart des cas, être fondée sur des mesures descriptives pour lesquelles le plan d'échantillonnage peut être pris en compte, ou encore sur des représentations graphiques. Toutefois, les résultats doivent être interprétés avec prudence en fonction du plan d'échantillonnage. Par exemple, un petit nombre de grands résidus ayant des coefficients de pondération de l'échantillon faibles peut être beaucoup moins important

qu'un grand nombre de petits résidus présentant des coefficients de pondération élevés. Un des outils d'analyse particulièrement utile dans le cas de la régression est la différence entre les coefficients de régression pondérée et non pondérée. Un écart important démontre souvent que le modèle ne convient pas.

Une fois les paramètres définis, il faut déterminer le genre d'inférence, soit l'estimation ponctuelle, l'estimation par intervalle ou les tests d'hypothèses. Il apparaît que l'estimation ponctuelle et les intervalles de confiance conviennent très bien aux paramètres d'une population finie, alors que les tests d'hypothèses, plus particulièrement d'hypothèses simples, sont valables uniquement lorsqu'ils sont appliqués à des paramètres d'une superpopulation d'un modèle très bien défini. Par exemple, l'hypothèse que les moyennes d'une paire d'ensembles sont égales peut être admise seulement s'il s'agit des moyennes d'une superpopulation et non des réalisations de leur population finie. Lorsqu'on veut éviter la formulation d'un modèle, il est recommandé de procéder à l'estimation ponctuelle ou de calculer les intervalles de confiance, plutôt que de faire des tests d'hypothèses, pour établir la différence entre les moyennes d'un ensemble. S'il faut appliquer ces tests aux paramètres d'une population finie, il est préférable de tester une hypothèse composée (par exemple tester si la différence entre les moyennes se situe à l'intérieur d'un ordre de valeurs) que de tester une hypothèse simple, c'est-à-dire vérifier si la différence est zéro. Il convient de souligner que, dans le cas d'échantillons assez grands, tout écart plus grand que zéro, peu importe s'il l'est à peine, sera considéré comme différent de zéro de façon significative.

Nous allons maintenant examiner certaines méthodes générales d'analyse des données provenant de plans d'échantillonnage complexes, de même que certaines autres méthodes particulières concernant les modèles linéaires et les tests de validité de l'ajustement et d'indépendance dans les tables de contingence. De façon générale, nous étudierons l'inférence se rapportant aux paramètres d'une population finie. Cependant, nous estimons que cette inférence est pertinente seulement si les paramètres s'apparentent aux paramètres du modèle de superpopulation, ce qui laisse à chacun la liberté

d'appliquer soit la méthode basée sur un plan d'échantillonnage soit celle basée sur un modèle, selon le degré de confiance que chacun a dans la validité d'un modèle sous-jacent.

2. MÉTHODES GÉNÉRALES DE BASE

2.1 Statistique généralisée de Wald

Si l'hypothèse qu'il faut tester est linéaire (ou peut être linéarisée) dans les valeurs prévues de statistiques asymptotiquement normales, pour lesquelles il existe un estimateur convergent de la matrice des variances, on peut alors utiliser la statistique généralisée de Wald (Grizzle, Starmer et Koch (1969), Koch, Freeman et Freeman (1976), Freeman, Freeman, Brock et Koch (1976), Shah, Holt et Folsom (1977) et Koch, Stokes et Brock (1980)).

Nous supposons que nous voulons tester l'hypothèse:

$$H_0: \quad X\beta = \theta_0, \tag{2.1.1}$$

où X est une matrice $r^{\times}p$ connue de plein rang du plan d'échantillonnage. β est un vecteur inconnu $p^{\times}1$ des paramètres (soit des paramètres d'une population finie soit des paramètres d'une superpopulation) et θ_0 est un vecteur connu $r^{\times}1$ des constantes. Si l'hypothèse n'est pas linéaire, on peut procéder à une approximation de premier ordre des séries de Taylor (Nathan, 1972, et Shuster et Downing, 1976).

Nous supposons qu'il y a un estimateur asymptotique normal convergent $\hat{\beta}$ de β ainsi qu'un estimateur convergent $(\hat{V})$ de la matrice des covariances de $\hat{\beta}$, dont la distribution est indépendante de celle de $\hat{\beta}$.

Alors, la statistique généralisée de Wald définie par:

$$X_w^2 = (X\hat{\beta}-\theta_0)' \, (X\hat{V}X')^{-1} \, (X\hat{\beta}-\theta_0) \tag{2.1.2}$$

est distribuée asymptotiquement, selon l'hypothèse nulle, comme χ^2 ayant des degrés de liberté correspondant à la dimension de l'hypothèse (p-r).

La compatibilité de $\hat{\beta}$ et de $\hat{V}$ et les distributions asymptotiques de $\hat{\beta}$ et de X_W^2 peuvent être étudiées en fonction de la distribution de la superpopulation.

La principale difficulté que présente cette démarche a trait au calcul de l'estimateur convergent $\hat{V}$ de la matrice des covariances, lorsque $\hat{\beta}$ n'est pas linéaire dans les données de l'échantillon (ce cas est très fréquent). Rao (1975) a étudié les diverses méthodes d'estimation de la variance qui peuvent être appliquées, notamment la linéarisation (Tepping, 1968), la répétition équilibrée (McCarthy, 1969) et la méthode dite "Jackknife" (Miller, 1974). Il existe également plusieurs programmes informatiques qui peuvent être utilisés dans ces cas, par exemple SUPERCARP (Hidiroglou, Fuller et Hickman, 1980), SUDAAN (Shah, 1978) pour la linéarisation et OSIRIS IV: PSALMS pour la répétition équilibrée. Une liste complète et une comparaison des programmes est présentée par Kaplan, Francis et Sedransk (1979).

Des comparaisons empiriques des estimateurs de la variance ont été établies par Kish et Frankel (1974) et par Richards et Freeman (1980), de même que des comparaisons théoriques, par Krewski et Rao (1981).

Toutefois, il convient de porter une attention particulière à la stabilité de l'estimateur de la variance, surtout lorsqu'il y a un grand nombre de paramètres. De plus, il faut aussi tenir compte des conditions dans lesquelles la convergence et les propriétés asymptotiques se réalisent, dans le cas des plans complexes. Par exemple, avec un plan d'échantillonnage à deux degrés, il est possible que les résultats asymptotiques commandent un grand nombre d'unités primaires d'échantillonnage (UPE) et un grand nombre d'unités du dernier degré par UPE.

2.2 Approximation et élaboration d'un modèle de covariances

Les problèmes courants liés au calcul d'un estimateur convergent stable de la matrice des covariances ont poussé les spécialistes à essayer d'appliquer des approximations simplifiées à ces estimateurs. Le principe de base est que si l'on suppose une certaine structure de la matrice des covariances, on peut alors utiliser des estimateurs plus stables d'un petit nombre de paramètres.

L'approximation peut être calculée à l'aide de la méthode basée sur un plan d'échantillonnage, directement en fonction de la matrice des covariances. Si l'on peut faire des hypothèses sur l'égalité des effets du plan d'échantillonnage sur les variances et covariances à l'intérieur d'un sous-groupe donné de paramètres, les estimateurs d'ensemble de la covariance peuvent alors être utilisés. Cette démarche est préconisée par Nathan (1973), Fuller et Rao (1978), Fellegi (1980) et Lepkowski et Landis (1980).

Par ailleurs, l'élaboration d'un modèle de la structure de la population proprement dite peut favoriser la construction de matrices de covariances simplifiées qui peuvent être facilement estimées (voir Altham (1976), Fuller et Battese (1973), Tomberlin (1979), Holt, Richardson et Mitchell (1980), Imrey, Sobel et Francis (1980) et Pfeffermann et Nathan (1981).

2.3 Modification des tests classiques

L'utilisation très répandue de programmes informatiques types a favorisé l'élaboration de nouveaux tests qui tiennent compte de plans d'échantillonnage complexes. Ces travaux peuvent être évalués sous l'angle d'une extension naturelle de l'utilisation des effets du plan d'échantillonnage comme facteurs multiplicatifs des variances basées sur un échantillon aléatoire simple d'une même taille, en vue d'apporter les modifications appropriées lorsqu'il s'agit de plans d'échantillonnage complexes.

La rectification peut certes être basée sur les effets du plan de divers estimateurs ou sur la moyenne des effets du plan (voir Cowan et Binder (1978), Fay (1979), Fellegi (1980), Rao et Scott (1981) et Scott et Holt (1981).

L'autre solution possible est d'étudier le comportement des fonctions des observations utilisées dans un test selon un certain modèle de superpopulation et de modifier en conséquence les fonctions types (Cohen, 1979, et Campbell, 1977).

3. MÉTHODES PARTICULIÈRES

3.1 Régression et modèles linéraires

La définition préalable du modèle et des paramètres d'intérêt est extrêmement importante dans le cas de l'analyse de la régression et des modèles linéaires. Ainsi, lorsque des relations de régression différentes doivent être calculées pour diverses strates ou UPE dans un plan d'échantillonnage à deux degrés, le paramètre d'intérêt peut être une simple moyenne des coefficients de régression (Konijn, 1962), une moyenne pondérée des coefficients (Pfeffermann et Nathan, 1981) ou leur valeur prévue (selon une certaine distribution préalable - Porter, 1973).

Règle générale, le modèle et les paramètres d'intérêt doivent être définis en fonction de la structure présumée de la population globale et ne doivent pas se rapporter à la structure du plan d'échantillonnage. Toutefois, dans de nombreux cas, le plan d'échantillonnage reproduit la structure de la population de sorte que les variables du plan font partie du modèle. À titre d'exemple, considérons le modèle:

$$E(Y|X_1,X_2) = X_1 \beta_{1.2} + X_2 \beta_{2.1} \qquad (3.1.1)$$

où X_1 comprend uniquement les variables qui ne se rapportent pas au plan d'échantillonnage et X_2 représente toutes les variables incluses dans le

plan d'échantillonnage complexe, c'est-à-dire que la distribution de l'échantillonnage dépend seulement de X_2:

$$P(s|X_1,X_2) = P(s|X_2). \tag{3.1.2}$$

L'estimation de $\beta_{1.2}$ et de $\beta_{2.1}$ dans la définition (3.1.1) et l'inférence qui en est établie peuvent être calculées selon la méthode classique, comme si l'échantillonnage était aléatoire simple, pour autant que le terme (3.1.1) se vérifie.

Cependant, si les variables du plan, X_2, ne sont pas incluses dans l'équation de régression d'intérêt:

$$E(Y|X_1) = X_1\beta_1 \tag{3.1.3}$$

et que la variable du plan, X_2, est en corrélation avec Y (fonction conditionnelle de X_1), alors l'estimation type par la méthode des moindres carrés ordinaires de β_1 n'est pas convergente (voir Nathan et Holt (1980) et Holt et Smith (1979) qui préconisent l'utilisation d'estimations modifiées pondérées et non pondérées de β_1, qui sont convergentes). Holt, Smith et Winter (1980) donnent un exemple d'application de ces estimateurs.

Si le modèle linéaire:

$$E(Y_i|x_i) = x_i'\beta \tag{3.1.4}$$

$$\mathrm{cov}(Y_i,Y_j|x_i,x_j) = \begin{cases} \sigma^2 & i=j \\ 0 & i\neq j \end{cases} \tag{3.1.5}$$

est vrai pour toutes les unités (i, j=1, ..., N) d'une population finie et x_i, le vecteur colonne p^X1, comprend toutes les variables du plan d'échantillonnage, alors l'estimation non pondérée des moindres carrés ordinaires:

$$\hat{\beta} = (X_n' X_n)^{-1} X_n' Y_n \tag{3.1.6}$$

basée sur les valeurs échantillonnées

$$\underset{p\times n}{X'_n} = (x_1, \ldots, x_n) \text{ et } Y'_n = (Y_1, \ldots, Y_1)$$

produit le "meilleur" estimateur de β sans biais d'un modèle linéaire, peu importe le plan d'échantillonnage. On le qualifie de meilleur parce qu'il définit une variance minimum basée sur le modèle. Toutefois, $\hat{\beta}$ n'est pas, de façon générale, un estimateur sans biais basé sur le plan d'échantillonnage, ni même un estimateur convergent (basé sur le plan) du paramètre de la population:

$$B = (X'_N X_N)^{-1} X'_N Y_N, \tag{3.1.7}$$

où $\underset{p\times N}{X'_N} = (x_1, \ldots, x_N)$ et $Y'_N = (Y_1, \ldots, Y_N)$.

L'estimateur convergent de B selon le plan d'échantillonnage est l'estimateur pondéré:

$$\hat{\beta}_W = (X'_n W_n X_n)^{-1} X'_n W_n Y_n, \tag{3.1.8}$$

où la matrice de pondération $W_n = \text{diag}\,(\Pi_1^{-1}, \ldots, \Pi_n^{-1})$, est la matrice diagonale $n\times n$ des inverses des probabilités d'inclusion de l'échantillon $\Pi_i = Pr\,(i\varepsilon s)$

De toute évidence, la compatibilité de $\hat{\beta}_W$, comme estimateur de , ne dépend pas de ce que le modèle (3.1.4) se vérifie; en outre, la pertinence de l'estimation de B, lorsque le modèle ne se vérifie pas, peut être mise en doute. On peut démontrer que, si certaines conditions se réalisent dans un modèle non-linéaire, ce qui suppose que l'espérance mathématique conditionnelle de Y (selon X) est une fonction différenciable de X, l'espérance de B basée sur le modèle peut être exprimée approximativement comme une moyenne pondérée des pentes de cette fonction aux points X_i (les coefficients de pondération dépendant seulement de $X_i - \overline{X}$). Cependant, dans la pratique, cette interprétation a une valeur limitée.

De toute façon, $\hat{\beta}_W$ est un estimateur sans biais de β basé sur le modèle, chaque fois que le modèle (3.1.4) se vérifie. Règle générale, il ne sera pas un estimateur optimal de β selon (3.1.5) lorsqu'il y a échantillonnage à probabilité inégale, mais il le sera si la variance conditionnelle du modèle de Y_i est proportionnelle à Π_i, soit:

$$V(Y_i | x_i) = k\Pi_i \quad . \tag{3.1.9}$$

Comme l'estimateur pondéré $\hat{\beta}_W$ est plus robuste que l'estimateur non pondéré $\hat{\beta}$, en ce sens qu'il est à la fois un estimateur sans biais de β basé sur le modèle, si le modèle est vrai, et un estimateur convergent de B basé sur le plan d'échantillonnage, si le modèle ne se vérifie pas, il est recommandé d'utiliser l'estimateur pondéré $\hat{\beta}_W$ pour estimer B, chaque fois qu'on n'est pas sûr si le modèle (3.1.4) - (3.1.5) se vérifie. Il reste alors aux spécialistes à déterminer si B est un paramètre valable à estimer.

Il convient de souligner que, dans le cas de plans d'échantillonnage autopondérés, $\hat{\beta}$ et $\hat{\beta}_W$ correspondent. L'estimateur $\hat{\beta}_W$ (3.1.8) peut être calculé directement à partir de programmes informatiques types qui définissent la <u>régression pondérée</u> (par exemple BMDP) à l'aide des coefficients de pondération $1/\Pi_i$, ou encore à l'aide d'autres programmes (par exemple SPSS) qui affectent la <u>régression non pondérée</u> aux variables transformées $Y_i/\sqrt{\Pi_i}$ et $x_i/\sqrt{\Pi_i}$, mais pas aux variables pondérées Y_i/Π_i, x_i/Π_i. Cependant, il est bon de noter que, sous l'hypothèse alternative, les <u>variances et covariances des estimateurs sont incorrectes</u> et que <u>les tests de signification usuels (par exemple, les tests F) ne sont pas valides et peuvent fausser grandement les conclusions</u>.

Soit le modèle (3.1.4) - (3.1.5), la variance du modèle de $\hat{\beta}$ est:

$$V(\hat{\beta} | X_n) = \sigma^2 (X'_n X_n)^{-1} , \tag{3.1.10}$$

ce qui est le résultat fourni par les programmes de régresssion non pondérée. Cependant, la variance du modèle $\hat{\beta}_W$ est:

$$V(\hat{\beta}_W | X_n) = \sigma^2 (X'_n W_n X_n)^{-1} X'_n W'_n W_n X_n (X'_n W_n X_n)^{-1} \quad . \tag{3.1.11}$$

Le programme de régression pondérée, dont les coefficients sont $1/\Pi_i$, définit une valeur de $(X_n' W_n X_n)^{-1}$ de la variance du modèle de $\hat{\beta}_W$, ce qui correspond à (3.1.11) seulement si $W_n = I_n$. Par conséquent, aucun résultat concernant les erreurs types ou les tests d'hypothèses n'est exact.

Toutefois, l'estimateur du coefficient de corrélation multiple défini à l'aide de la régression pondérée:

$$\hat{R}^2 = \frac{(Y_n - X_n \hat{\beta}_W)' W_n (Y_n - X_n \hat{\beta}_W)}{(Y_n - \bar{y}_n \underline{1}_n)' W_n (Y_n - \bar{y}_n \underline{1}_n)}, \qquad (3.1.12)$$

où $\bar{y}_n = (\Sigma_s Y_i/\Pi_i) / (\Sigma_s 1/\Pi_i)$ est un estimateur convergent basé sur un plan d'échantillonnage du coefficient de corrélation multiple de la population:

$$R^2 = \frac{(Y_N - X_N B)' (Y_N - X_N B)}{(Y_N - \bar{Y}_N \underline{1}_M)' (Y_N - \bar{Y}_N \underline{1}_N)} \qquad (3.1.13)$$

où $\bar{Y}_N = (1/N) \underline{1}_N' Y_N$.

La variance basée sur un plan de $\hat{\beta}_W$, lequel doit être considéré comme la mesure appropriée de la précision de $\hat{\beta}_W$ en tant qu'estimateur de B, ne peut pas en général être calculée uniquement à l'aide des probabilités d'inclusion de premier ordre Π_i. Pour la plupart des plans d'échantillonnage appliqués couramment, la variance de $\hat{\beta}_W$ basée sur un plan est estimée à l'aide d'une des méthodes d'estimation de la variance mentionnées plus haut, soit la linéarisation, la répétition équilibrée ou le "jackknife" (voir Jonrup et Rennermalm (1976) et Holt et Scott (1981)).

3.2 Analyse des données qualitatives

La méthode la plus simple d'analyse des données qualitatives est celle qui consiste à construire une classification unique de la population en k classes comportant des probabilités (fréquences relatives)

$\underline{p}' = (p_1, \ldots, p_{k-1})$. Lorsqu'on veut tester l'hypothèse nulle de la validité de l'ajustement d'une distribution connue, $\underline{p}_o' = (p_{o1}, \ldots, p_{ok-1})$:

$$H_o: \underline{p} = \underline{p}_o \ , \tag{3.2.1}$$

on peut suivre les démarches expliquées au chapitre 2.

Nous supposons qu'il existe un estimateur d'enquête convergent $\hat{\underline{p}}' = (\hat{p}_1, \ldots, \hat{p}_{k-1})$ de $\underline{p}'$. S'il est asymptotiquement normal:

$$\sqrt{n}\ (\hat{\underline{p}} - \underline{p}) \rightsquigarrow N(\underline{0},V) \tag{3.2.2}$$

et qu'il existe un estimateur convergent $\hat{V}$ de V, alors la statistique généralisée de Wald:

$$X_W = n(\hat{\underline{p}} - \underline{p}_o)'\ \hat{V}^{-1}(\hat{\underline{p}} - \underline{p}_o) \ , \tag{3.2.3}$$

qui est distribuée asymptotiquement par χ^2_{k-1} sous H_o, peut être utilisée pour tester H_o.

Pour bon nombre de plans d'échantillonnage simples, des estimateurs convergents de V sont obtenus directement alors que, pour les plans plus complexes, ils peuvent être calculés à l'aide des méthodes classiques. Cependant, si des tests de validité de l'ajustement doivent être effectués pour une variété de variables et de classifications, il est alors préférable d'utiliser la variable χ^2 type:

$$X^2 = n \sum_{i=1}^{k} (\hat{p}_i - p_{oi})^2/p_{oi} = n(\hat{\underline{p}}-\underline{p}_o)P_o^{-1}(\hat{\underline{p}}-\underline{p}_o), \tag{3.2.4}$$

où $P_o = \text{diag}\ (\underline{p}_o) - \underline{p}_o\underline{p}_o'$, tout en y apportant les modifications qui s'imposent. Rao et Scott (1981) ont démontré que la distribution asymptotique X^2 sous H_o est celle de la somme pondérée de k-1 variables indépendantes χ^2 ayant chacune un degré de liberté.

$$X^2 \rightsquigarrow \sum_{i=1}^{k-1} \lambda_i Z_i^2 ; \quad Z_i^2 \sim N(0,1) \quad \text{indépendantes} \tag{3.2.5}$$

où $\lambda_1, \ldots, \lambda_{k-1}$ sont les valeurs propres de

$$D = P_o^{1} V \; (\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{k-1} > 0). \qquad (3.2.6)$$

On peut alors procéder à un test prudent de (3.2.1) en utilisant la variable X^2/λ_1 et une distribution χ^2_{k-1}. λ_1 peut représenter l'effet maximum du plan sur toutes les combinaisons linéaires des composantes de $\hat{\underline{p}}$. Par exemple, avec un échantillonnage stratifié proportionnel, $\lambda_1 \leq 1$ de sorte que X^2 peut être utilisé comme une fonction conservative des observations dans le test.

Dans les autres cas, l'utilisation de $X^2/\bar{\lambda}$

soit $$\bar{\lambda} = \frac{1}{k-1} \sum_{i=1}^{k-1} \lambda_i = \frac{1}{k-1} \sum_{i=1}^{k} d_i(1-p_i),$$

où $d_i = V[\hat{p}_i]/[p_i(1-p_i)]$ est l'effet du plan pour $\hat{p}_i$, s'est révélée un bon test d'approximation (Hidiroglou et Rao (1981), pour l'Enquête Santé Canada, et Holt, Scott et Ewings (1980), pour des enquêtes à grande échelle au Royaume-Uni). Une

approximation alternative , $X^2/\bar{d}$, où $\bar{d} = k^{-1} \sum_{i=1}^{k} d_i$, a été

proposée par Fellegi (1980).

La formulation directe de modèles pour $\underline{p}$ a été présentée par Altham (1976) et par Cohen (1976), mais leurs modèles comportent de sérieuses limitations puisqu'ils supposent que
$\lambda_1 = \lambda_2 = \ldots = \lambda_{k-1} = \bar{\lambda}$, ce qui équivaut à un effet constant du plan sur les catégories. Cela ne constitue pas généralement une hypothèse réaliste et a pour résultat que $X^2/\bar{\lambda}$ contient une distribution asymptotique χ^2_{k-1}. Pour effectuer un test d'indépendance dans une table de contingence, les hypothèses peuvent être exprimées:

$$H_o: h_{ij}(\underline{p}) = p_{ij} - p_{i+} p_{+j} = 0$$
$$(i=1, \ldots, r-1; \; j=1, \ldots, c-1), \qquad (3.2.7)$$

où p_{ij} est la probabilité de distribution de la population dans la case (i, j), p_{i+}, p_{+j} sont les probabilités marginales et $\underline{p}' = (p_{11}, \ldots, p_{rc-1})$. La statistique généralisée de Wald appliquée au test de H_0 est notée:

$$X^2_{WI} = n[\underline{h}(\hat{\underline{p}})]' \hat{V}_h^{-1} \underline{h}(\hat{\underline{p}}), \qquad (3.2.8)$$

où $[\underline{h}(\hat{\underline{p}})]' = [h_{11}(\hat{\underline{p}}), \ldots, h_{r-1\ c-1}(\hat{p})]$ et $\hat{V}_h/n$ est un estimateur convergent de la matrice des covariances de $\underline{h}(\hat{\underline{p}})$. Des versions de (3.2.8) ont été appliquées à des plans d'échantillonnage particuliers de même que diverses méthodes d'estimation de $\hat{V}_{h/n}$ ont été utilisées par Garza-Hernandez et McCarthy (1962), Nathan (1969, 1975) Shuster et Downing (1976) et Fellegi (1980).

Une variable modifiée s'apparentant à $X^2/\bar{\lambda}$ a été introduite par Rao et Scott (1981):

$$X^2_{CI} = (n/\hat{\delta}) \sum_{i=1}^{r} \sum_{j=1}^{c} (\hat{p}_{ij}-\hat{p}_{i+}\ \hat{p}_{+j})^2 / (\hat{p}_{i+}\ \hat{p}_{+j}), \qquad (3.2.9)$$

où $\hat{\delta} = \frac{1}{(r-1)(c-1)} \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{v}_{ij}(\underline{h})/(\hat{p}_{i+}\ \hat{p}_{+j})$ et

$\hat{v}_{ij}(\underline{h})/n$ est un estimateur convergent de la variance de $h_{ij}(\hat{\underline{p}})$. $\hat{\delta}$ peut être exprimé en fonction des effets du plan estimés de $h_{ij}(\hat{\underline{p}})$:

$$\hat{\delta} = \frac{1}{(r-1)(c-1)} \sum_{i=1}^{r} \sum_{j=1}^{c} (1-\hat{p}_{i+})(1-\hat{p}_{+j})\ \hat{\delta}_{ij}, \qquad (3.2.10)$$

où $\hat{\delta}_{ij}$ est un estimateur de l'effet du plan d'échantillonnage, δ_{ij}, de $h_{ij}(\hat{\underline{p}})$:

$$\delta_{ij} = nV[h_{ij}(\hat{\underline{p}})]/[p_{i+}\ p_{+j}\ (1-p_{i+})(1-p_{+j}) \quad . \qquad (3.2.11)$$

Il peut être plus facile d'estimer les effets du plan d'échantillonnage que d'estimer les variances.

Des travaux de recherche empiriques menés par Holt, Scott et Ewings (1980) et par Hidiroglou et Rao (1981) démontrent que la distribution de X^2_{CI} s'apparente à $\chi^2(r-1)(c-1)$.

3.3 Autres genres d'analyses

Les modèles linéaires, de même que les tests de validité de l'ajustement et d'indépendance se prêtent à de nombreuses applications d'analyses, alors que d'autres genres d'analyses telles que l'analyse en composantes principales, l'analyse factorielle et discriminante, l'analyse de corrélation, la régression logistique, les modèles logarithmiques linéaires, les méthodes non paramétriques, etc. ne peuvent être appliquées aussi facilement. Bien que les méthodes générales décrites au chapitre deux puissent être utilisées, leur application pose certaines difficultés et, de fait, on n'a signalé que très peu de cas possibles.

Étant donné que les coefficients de corrélation sont un élément de base dans la plupart des analyses à plusieurs variables, un certain nombre d'études empiriques sur l'effet du plan d'échantillonnage sur l'estimation de ces coefficients ont été menées par Kish et Frankel (1974), Bebbington et Smith (1977) et Holt, Richardson et Mitchell (1980). Quoiqu'on ne puisse tirer de conclusions générales de leurs travaux, il apparaît que les effets du plan d'échantillonnage sont loin d'être négligeables. Bebbington et Smith (1977) ont également étudié la variabilité échantillonnale des estimateurs des composantes principales.

Dans d'autres domaines également, l'effet du plan d'échantillonnage sur les logits a été examiné par Lepkowski et Landis (1980) et les intervalles de confiance des quantiles, par Woodruff (1952) et par Sedransk et Meyer (1978).

REMERCIEMENTS

L'auteur tient à remercier MM. D. Binder, N. Chinnappa, S.E. Fienberg, M. Hidiroglou, C.J.C. Hole, J.N.K. Rao et A. Scott de lui avoir consacré du temps et fait part de leurs commentaires.

BIBLIOGRAPHIE

[1] Altham, P.M.E. (1976), "Discrete Variable Analysis for Individuals Grouped into Families", Biometrika, 63, 263-269.

[2] Brewer, K.R. and Mellor, R.W. (1973), "The Effect of Sample Structure on Analytical Surveys", Aust. J. Statist., 15, 145-152.

[3] Bebbington, A.C. and Smith, T.M.F. (1977), "The Effect of Survey Design on Multivariate Analysis", The Analysis of Survey Data (C.A. O'Muircheartaich and C. Payne, Editors), Vol. 2. Model Fitting", New York: Wiley, 175-192.

[4] Campbell, C. (1977), "Properties of Ordinary and Weighted Least Squares Estimators for Two Stage Samples", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 800-805.

[5] Cohen, J.E. (1976), "The Distribution of the Chi-Squared Statistic Under Clustered Sampling", J. Amer. Statist. Assoc. 71, 665-670.

[6] Cowan, J. and Binder, D.A. (1978), "L'effet d'un plan d'échantillonnage à deux degrés sur les tests d'indépendance", Techniques d'enquête 4, n° 1, 16-29.

[7] Fay, R.E. (1979), "On Adjusting the Pearson Chi-Square Statistic for Clustered Sampling", Proc. Soc. Statist. Sect., Amer. Statist. Assoc. 402-405.

[8] Fellegi, I.P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples", J. Amer. Statist. Assoc. 75, 261-268.

[9] Fienberg, S.E. (1980), "The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey", The Statistician, 29, 313-350.

[10] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sanklya, 22, 117-132.

[11] Fuller, W.A. and Battese, C.E. (1973), "Transformations for Estimation of Linear Models with Nested-Error Structure", J. Amer. Statist. Assoc. 68, 626-632.

[12] Fuller, W.A. and Rao, J.N.K. (1978), "Estimation for a Linear Regression Model with Unknown Diagonal Covariance Matrix", Ann. Statist. 6., 1149-1158.

[13] Freeman, D.H. Jr., Freeman, J., Brock, D.B. and Koch, G.G., "Strategies in the Multivariate Analysis of Data from Complex Surveys II: An Application to the United States National Health Interview Survey", Inter. Statist. Rev. 44, 317-330.

[14] Garza-Hernandez, T. and McCarthy, P.J. (1962), "A Test of Homogeneity for a Stratified Sample", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 200-202.

[15] Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969), "Analysis for Categorical Data by Linear Models", Biometrics, 25, 489-504.

[16] Hartley, H.O. and Sielken, R.L. (1975), "A Superpopulation Viewpoint for Finite Population Sampling", Biometrics, 31, 411-422.

[17] Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1980), "Super Carp: Sixth Edition", Statistical Laboratory Survey Section, Iowa State University, Ames, Iowa.

[18] Hidiroglou, M.A. and Rao, J.N.K. (1981), "Chi-Square Tests for the Analysis of Categorical Data from the Canada Health Survey", Invited Paper for 43rd Session of I.S.I., Buenos-Aires.

[19] Holt, D. Richardson, S.C. and Mitchell, P.W. (1980), "The Analysis of Correlations In Complex Survey Data", (non publié).

[20] Holt, D. and Scott, A.J. (1981), "Regression Analysis Using Survey Data", The Statistician, 30. (à paraître).

[21] Holt, D. Scott, A.J., and Ewings, P.O. (1980), "Chi-Squared Test with Survey Data", J. Roy. Statist. Soc. A, 143, 302-330.

[22] Holt, D. and Smith, T.M.F. (1979), "Regression Analysis of Data from Complex Surveys", Roy. Statist. Soc. Conf. Oxford.

[23] Holt, D., Smith, T.M.F. and Winter, P.O. (1980), "Regression Analysis of Data from Complex Surveys", Jour. Roy. Statist. Soc. A, 143, 474-483.

[24] Imvrey, P., Sobel, E. and Francis, M. (1980), "Modeling Contingency Tables from Complex Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 213-217.

[25] Jonrup, H. and Rennermalm, B. (1976), "Regression Analysis in Samples from Finite Population", Scand. Jour. Statist., 3, 33-37.

[26] Kaplan, B., Francis, I., and Sedransk, J. (1979), "A Comparison of Methods and Programs for Computing Variances of Estimators from Complex Sample Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 97-100.

[27] Kish, Leslie and Frankel, M.R. (1970), "Balanced Repeated Replication for Standard Errors", J. Amer. Statist. Assoc., 65, 1071-1094.

[28] Kish, L. and Frankel, M.R. (1974), "Inference from Complex Samples", (avec commentaires), J. Roy. Statist. Soc. B, 36, 1-37.

[29] Koch, G.G., Freeman, D.J., Jr., and Freeman, J.L. (1975), "Strategies in The Multivariate Analysis of Data from Complex Surveys", Inter. Statist. Rev. 43, 59-78.

[30] Koch, G.G., Stokes, M.E. and Brock, D. (1980), "Applications of Weighted Least Squares Methods for Fitting Variational Models to Health Survey Data", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 218-223.

[31] Konijn, H.S. (1962), "Regression Analysis for Sample Surveys", J. Amer. Statist. Assoc. 57, 590-606.

[32] Krewski, D., and Rao, J.N.K. (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods", Ann. Statist., 9 (5/ 1010-1019).

[33] Lepkowski, J.M. and Landis, J.R. (1980), "Design Effects for Linear Contrasts of Proportions and Logits", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 224-229.

[34] McCarthy, P.J. (1969), "Pseudo-Replication: Half-Samples", Inter. Statist. Rev. 37, 239-264.

[35] Miller, R.G. (1974), "The Jackknife--A Review", Biometrika 61, 1-15.

[36] Nathan, G. (1969), "Tests of Independence in Contingency Tables from Stratified Samples", New Developments in Survey Sampling (N.L. Johnson and H. Smith, eds.), New York: Wiley, 578-600.

[37] Nathan, G. (1972), "On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples", J. Amer. Statist. Assoc., 67, 917-920.

[38] Nathan, G. (1973), "Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples", National Center for Health Statistics, Vital and Health Statistics Series 2, No. 53, Washington, D.C.

[39] Nathan, G. (1975), "Tests of Independence in Contingency Tables from Stratified Proportional Samples", Sankhya C, 37, 77-87. [erratum: Sankhya C, 40, (1978), 190].

[40] Nathan, G. and Holt, D. (1980), "The Effect of Survey Design on Regression Analysis", J. Roy. Statist. Soc. B, 42, 377-386.

[41] Pfeffermann, D., and Nathan, G. (1981), "Regression Analysis of Data from Complex Samples", J. Amer. Statist. Assoc. 76, 681-689.

[42] Porter, R.M. (1973), "On the Use of Survey Sample Weights in the Linear Model", Annals of Economic and Social Measurement, 2, 141-158.

[43] Rao, J.N.K. (1975), "Analytic Studies of Sample Survey Data", Survey Methodology, Vol. 1, Supplementary Issue.

[44] Rao, J.N.K. and Scott, A.J. (1981), "The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables", J. Amer. Statist. Assoc. 76, 221-230.

[45] Richards, V. and Freeman, D.H. Jr. (1980), "A Comparison of Replicated and Pseudo-Replicated Covariance Matrix Estimators for the Analysis of Contingency Tables", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 209-211.

[46] Särndal, C.E. (1978), "Design-Based and Model-Based Inference in Survey Sampling", Scand. J. Statist., 5, 27-52.

[47] Sedransk, S. and Meyer, J. (1978), "Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling", J. Roy. Statist. Soc. B., 40, 239-252.

[48] Scott, A. and Holt, D. (1981), "The effect of Two-Stage Sampling on Ordinary Least Squares Methods", (non publié).

[49] Shah, B.V. (1978), "SUDAAN: Survey Data Analysis Software", Proc. Statist. Comp. Sect., Amer. Statist. Assoc., 146-151.

[50] Shah, B.V., Holt, M.M. and Folsom, R.E. (1977), "Inference About Regression Models from Sample Survey Data", Bull. Inter. Statist. Inst. 47, Bk. 3, 43-57.

[51] Shuster, J.J. and Downing, D.J. (1976), "Two-Way Contingency Tables for Complex Sampling Schemes", Biometrika 63, 271-278.

[52] Smith, T.M.F. (1976), "The Foundations of Survey Sampling: A Review (Avec Commentaires), J. Roy. Statist. Soc. A., 139, 183-195.

[53] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 66, 411-414.

[54] Thomsen, I. (1978), "Design and Estimation Problems When Estimating a Regression Coefficient From Survey Data", Metrika 25, 27-35.

[55] Tomberlin, T.J. (1979), "The Analysis of Contigency Tables of Data from Complex Samples", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 152-157.

[56] Woodruff, Ralph S. (1952), "Confidence Intervals for Medians and Other Position Measures", J. Amer. Statist. Assoc. 47, 635-646.

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.

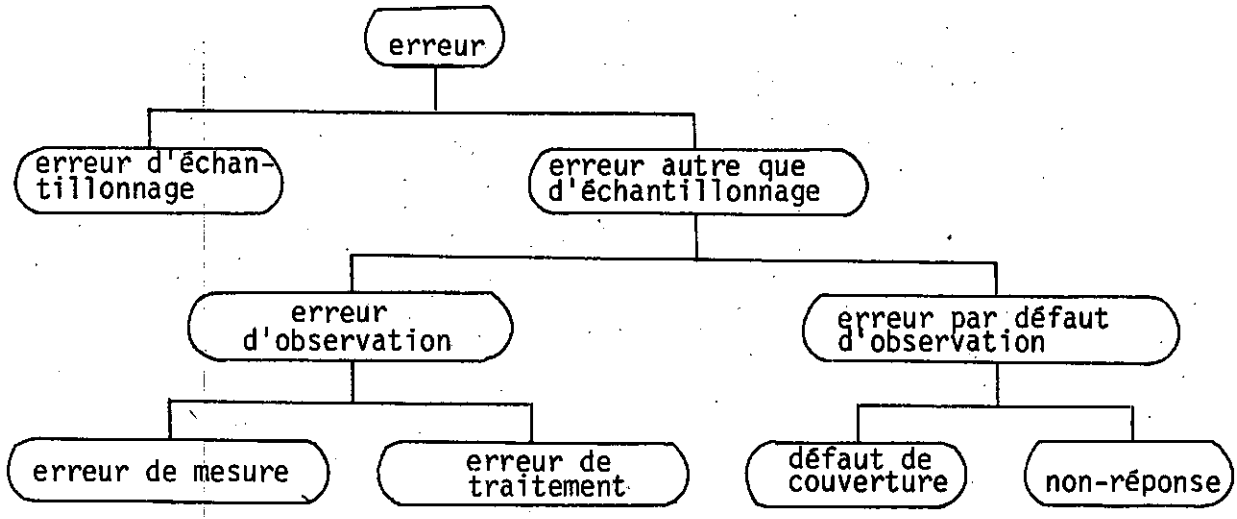
2.1 Terminologie

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

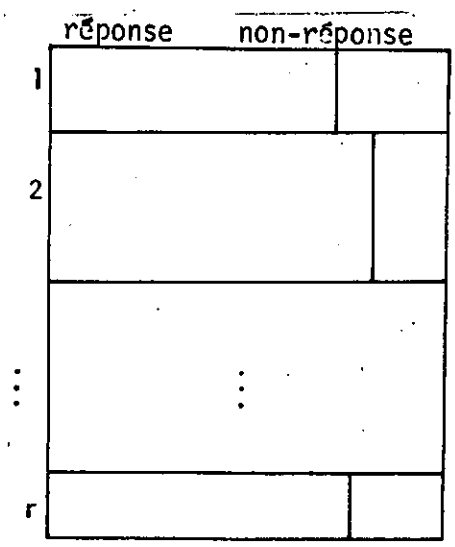
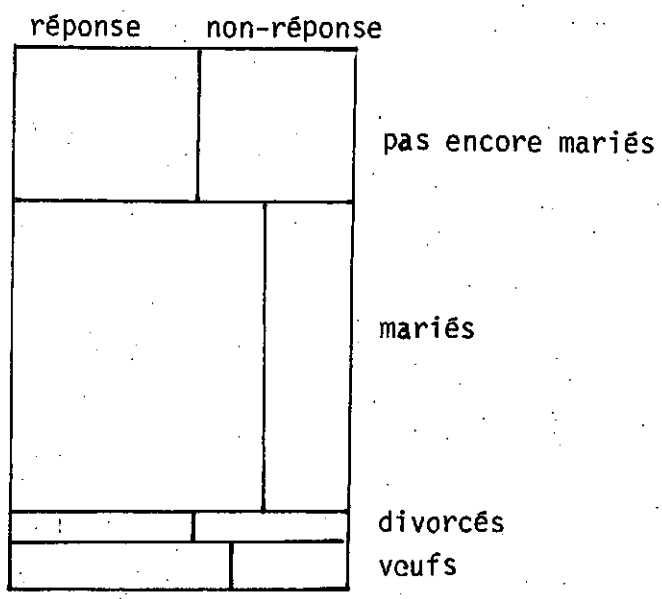


Figure 3. DIAGRAMME A CASES INDIQUANT LA SITUATION MATRIMONIALE DES ENQUETES LORS DE L'ENQUETE DE 1977/78 SUR LA DEMANDE DE LOGEMENTS A AMSTERDAM



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:

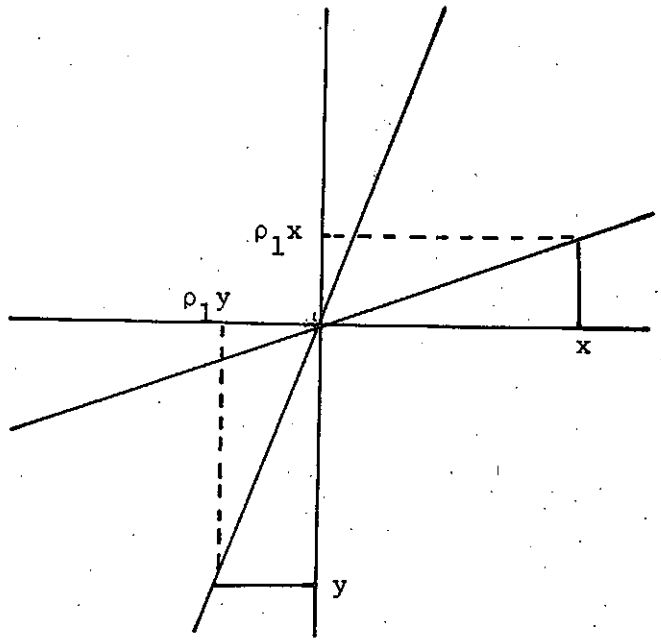
$$y = \rho_1 x \tag{1}$$

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

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

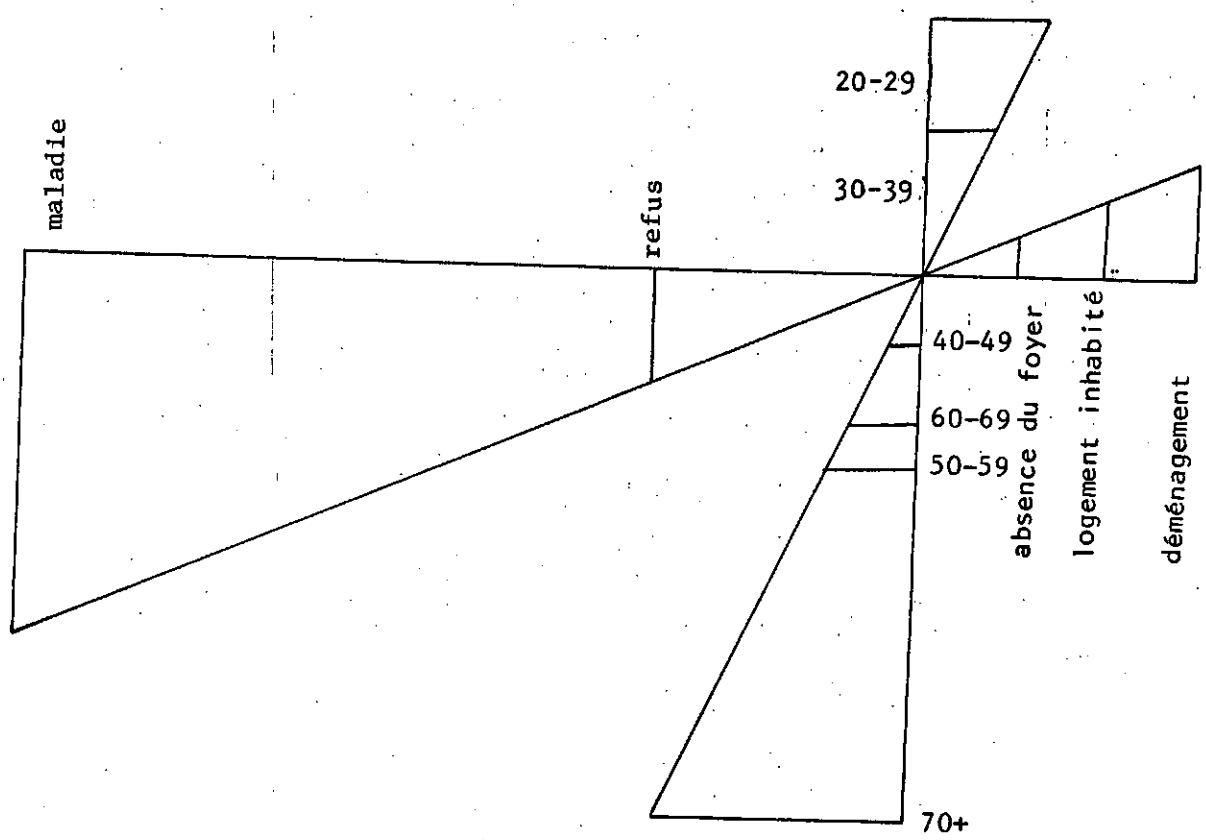
où X^2 est la valeur de chi-carré pour le tableau et N le total général,

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

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 (τ_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 $\underline{u}_1^*$, $\underline{u}_2^*$, .., $\underline{u}_m^*$ ($\underline{m} = \underline{r}_1 + \underline{r}_2 + \ldots, + \underline{r}_n$), 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 $\underline{\bar{y}}_h^*$ 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 $w_1, w_2, ..., w_H$.

Si l'on connaît les dimensions $N_1, N_2, ..., N_H$ 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 ($n = n_1 + n_2 + ... + n_H$).

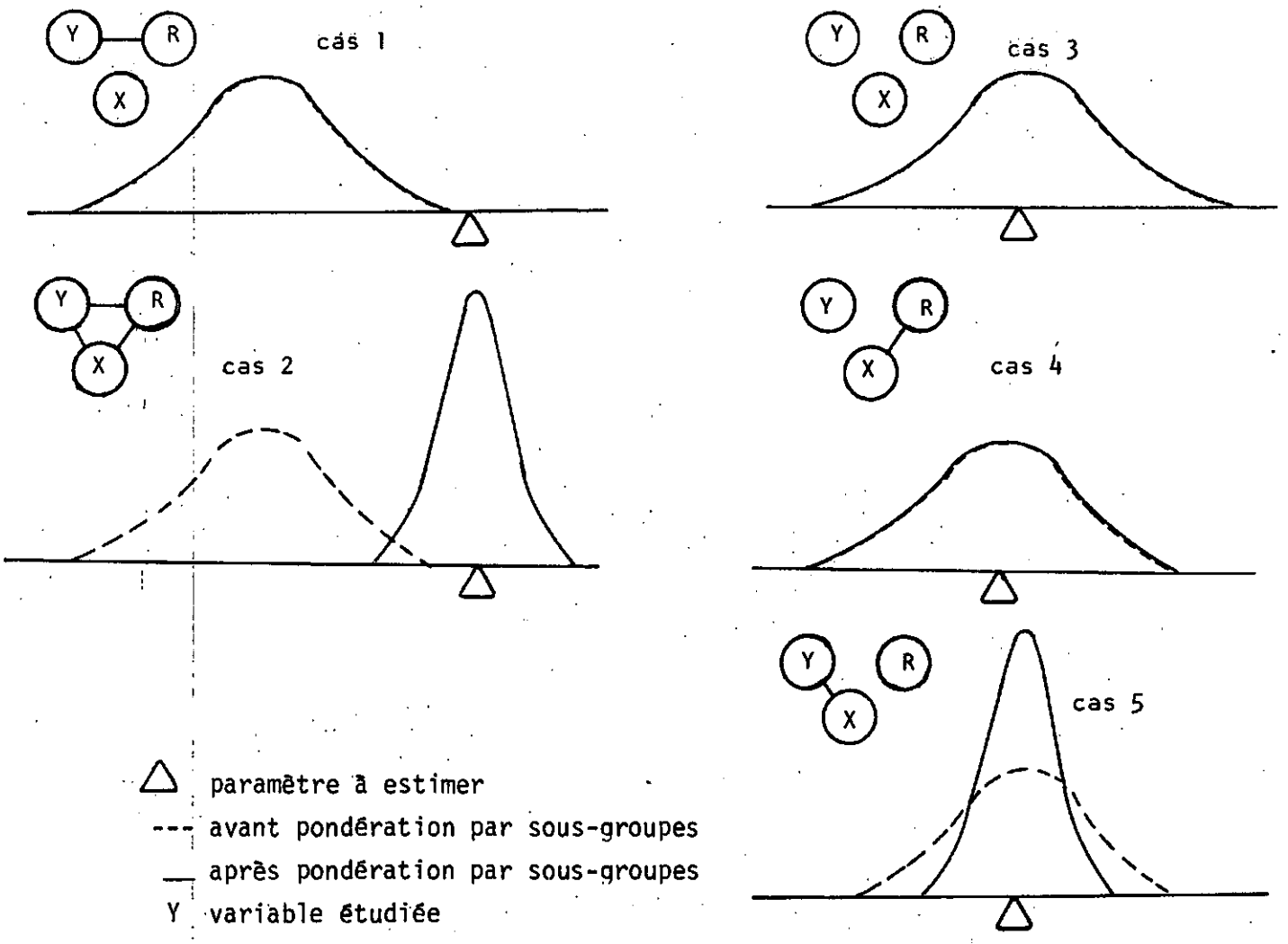
Dans une situation intermédiaire où l'on utilise deux variables auxiliaires X_1 et X_2 et où l'on ne connaît que les totaux marginaux des deux variables, on peut appliquer une autre méthode pour estimer les poids (voir par exemple Chapman, 1976). Supposons que X_1 donne G

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



△ paramètre à estimer
--- avant pondération par sous-groupes
— après pondération par sous-groupes
Y variable étudiée
R variable "réponses"
X variable auxiliaire

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).

TECHNIQUES D'ENQUÊTE 1981, VOL. 7 NO. 2

LES VARIANCES D'ESTIMATEURS ASYMPTOTIQUEMENT NORMAUX BASES SUR DES ENQUÊTES COMPLEXES

David A. Binder[1]

Nous discutons du problème de la spécification et de l'estimation de la variance de paramètres estimés basés sur des plans d'échantillonnage complexes provenant de populations finies. Les résultats présentés dans cet article sont particulièrement utiles lorsque les estimateurs des paramètres ne sont pas définis explicitement comme étant une fonction des autres statistiques de l'échantillon. Nous montrons comment des résultats peuvent s'appliqer à la régression linéaire, à la régression logistique et aux modèles linéaires logarithmiques de tables de contingence.

I. INTRODUCTION

Nous avons observé ces dernières années une tendance croissante à utiliser des données d'enquête pour estimer les paramètres de modèles classiques, tels que les paramètres de régression, les fonctions discriminantes et les paramètres de logit et de probit. Toutefois, dans nombre d'enquêtes, l'objet premier est l'estimation des valeurs moyennes, des valeurs totales, des tendances et autres caractéristiques d'une population ou d'une

1 D.A. Binder, Division des méthodes d'enquête-institution et agriculture, Statistique Canada.

sous-population. Pour cette raison et pour d'autres considérations d'ordre pratique, le plan d'enquête comprend rarement un échantillon aléatoire simple, mais plutôt un échantillon stratifié, souvent à plusieurs degrés, pouvant présenter des probabilités inégales à certains degrés.

Aussi, on s'est beaucoup interrogé sur le bien-fondé de l'utilisation de coefficients de pondération de l'échantillonnage dans les inférences faites au sujet de ces paramètres de modèles (voir notamment Särndal, 1978). Pour pouvoir répondre à une telle question, il faut d'abord déterminer si un modèle de superpopulation convient à toutes les unités de la population. Dans l'affirmative, les inférences portant sur les paramètres de la superpopulation constituent souvent la première occupation. Il s'agit alors d'inférences basées sur des modèles, c'est-à-dire que, pour un échantillon donné, elles ne dépendent pas des coefficients de pondération de l'échantillonnage.

La question qui nous vient à l'esprit est la suivante, si le modèle de superpopulation ne convient pas, à quoi correspondent les paramètres que nous estimons? Il nous faut admettre que dans de nombreuses études, particulièrement dans le domaine des sciences sociales, le modèle utilisé (par exemple la régression linéaire) n'est qu'une approximation utile de l'univers réel et que les paramètres de ce modèle (par exemple les corrélations et les corrélations partielles) ont plus souvent pour objet de faciliter la compréhension des interdépendances approximatives des variables que de donner lieu à une interprétation scientifique particulière. Par conséquent, les paramètres que nous estimons ne se rapportent pas nécessairement à un modèle de superpopulation réel; ils ont plutôt un caractère descriptif.

Dans cet article, nous partons du principe que nous cherchons à faire des inférences au sujet de ces paramètres "descriptifs" de la population. À titre d'exemple, supposons que $\underset{\sim}{X}$ et $\underset{\sim}{Y}$ désignent respectivement des matrices $N \times p$ et $N \times 1$, chaque ligne de $\underset{\sim}{X}$ et $\underset{\sim}{Y}$ correspondant à un individu différent de la population. Nous nous intéressons au paramètre descriptif, $\underset{\sim}{B}$, un vecteur pxl qui satisfait l'équation:

Les paramètres examinés dans cet article ne sont pas définis par une équation explicite comme l'équation (1.3), mais ils sont plutôt définis implicitement sous la forme $\underset{\sim}{U}(\underset{\sim}{Z},\underset{\sim}{\theta}) = 0$. Un exemple simple d'une telle distinction pourrait être le paramètre du rapport :

$$R = \frac{\Sigma Y_k}{\Sigma X_k},$$

qu'on pourrait aussi définir implicitement sous la forme

$$\Sigma Y_k - R\Sigma X_k = 0.$$

Pour certains types de modèles, comme les modèles linéaires logarithmiques indirects ou les modèles de régression logistique, les paramètres peuvent êtres définis seulement en fonction de relation implicites.Bien qu'elle n'apparaise encore sons une forme générale dans la literature l'extension des résultats de Tepping (1968) est relativement simple appliquer dans ce cas. Il existe toutefois quelques exemples précis de son application,voir notamment Fuller (1975) et Freeman et Koch (1976).

Le cadre général et les principaux résultats de notre étude sont exposés dans la deuxième partie du présent article et des exemples d'un certain nombre de modèles sont présentés dans la troisième partie.

2. CADRE GÉNÉRAL

2.1 Cadre

Les unités de la population sont désignées 1, ..., N. Un vecteur de données q-dimensionnel $\underset{\sim}{X}_i$ est associé à la i-ième unité. Nous avons un espace des paramètres $\Theta \subseteq R^p$. Le paramètre $\underset{\sim}{\theta}_o = (\theta_{1o}, \ldots, \theta_{po})$ est défini par les équations p:

$$\underset{\sim}{X}^T \underset{\sim}{X} \underset{\sim}{B} = \underset{\sim}{X}^T \underset{\sim}{Y} \tag{1.1}$$

Cette conception des paramètres descriptifs est la même que celle qui a été considérée par Frankel (1971) et Kish et Frankel (1974)

Pour l'estimation de tels paramètres, on tient compte habituellement des coefficients de pondération de l'échantillonnage. Si nous indiquons par Π_i la probabilité que la i-ième unité de l'échantillon soit prélevée et si $\underset{\sim}{\Pi} = \text{diag}\ (\pi_1, \ldots, \pi_n)$, alors l'estimateur pondéré du paramètre $\underset{\sim}{B}$ satisfait:

$$\underset{\sim}{x}^T \underset{\sim}{\Pi}^{-1} \underset{\sim}{x}\ \underset{\sim}{B} = \underset{\sim}{x}^T \underset{\sim}{\Pi}^{-1} \underset{\sim}{y} \tag{1.2}$$

où $\underset{\sim}{x}$ et $\underset{\sim}{y}$ représentent respectivement des matrices nxp et nxl dont les lignes correspondent aux lignes échantillonnées de $\underset{\sim}{X}$ et de $\underset{\sim}{Y}$.

Supposons maintenant qu'un estimateur d'un paramètre d'une population peut être défini de la façon suivante:

$$\hat{\theta} = g(z_1, \ldots, z_k), \tag{1.3}$$

où $E(z_i) = Z_i$. Dans cette équation, $\hat{\theta}$ est un estimateur de $g(Z_i, \ldots Z_k)$. Selon Tepping (1968) et Woodruff (1971), un développement de la série de Taylor pour $\hat{\theta}$ donne:

$$V[\hat{\theta}] \doteq V[\sum_{i=1}^{k} (\frac{\partial g}{\partial Z_i})\ (z_i - Z_i)]\ . \tag{1.4}$$

Des exemples de ces formules appliquées à l'estimation des coefficients de régression (1.1) ont été fournis par Tepping (1968). Toutefois, pour le calcul des variances des coefficients de régression, les expressions qui découlent de l'équation (1.4) sont un peu plus complexes que celles qui ont été dérivées par Fuller (1975).

$$U_i(\underset{\sim}{x},\underset{\sim}{\theta}_o) = \sum_{k=1}^{N} u_i(\underset{\sim}{x}_k,\underset{\sim}{\theta}_o) - v_i(\underset{\sim}{\theta}_o) = 0 \ , \tag{2.1}$$

pour i=1, ..., p. Nous supposons que les équations (2.1) définissent θ_o uniquement dans Θ. Nous supposons également que $\partial u_i(\underset{\sim}{X},\underset{\sim}{\theta})/\partial\underset{\sim}{\theta}$ et $\partial v_i(\underset{\sim}{\theta})/\partial\underset{\sim}{\theta}$ existent dans le voisinage de $\underset{\sim}{\theta}_o$. Un exemple simple de (2.1) est lorsque θ_o représente un total de la population et que nous avons $U(\underset{\sim}{X},\theta_o)= \sum_{k=1}^{N} X_k-\underset{\sim}{\theta}_o$. Dans ce cas, $u(X_k,\theta_o) = X_k$ et $v(\theta_o) = \theta_o$.

Nous tirons un échantillon d'unités selon une distribution de probabilité définie en sur l'ensemble de tous les sous-ensembles non vides de {1, ..., N}. Nous désignons par $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ les valeurs choisies de $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_N$. Nous supposons que pour tout $\underset{\sim}{\theta} \in \Theta$, nous pouvons construire un estimateur convergent asymptotiquement normal de $U_i(X,\underset{\sim}{\theta})$. Nous désignons cet estimateur par $\hat{U}_i(\underset{\sim}{x},\underset{\sim}{\theta})$. Par exemple, pour nombre de plans d'échantillonnage sans remise,

$$\hat{U}_i(\underset{\sim}{x},\underset{\sim}{\theta}) = \sum_{k=1}^{n} (\underset{\sim}{x}_k,\underset{\sim}{\theta})/\pi_k - v_i(\underset{\sim}{\theta}) \tag{2.2}$$

est un estimateur convergent asymptotiquement normal, où π_k représente la probabilité d'inclusion de la k-ième unité.
Posons $\sigma_{ij}(\underset{\sim}{X},\underset{\sim}{\theta}) = \mathrm{Cov}[\hat{U}_i(\underset{\sim}{x},\underset{\sim}{\theta}), \hat{U}_j(\underset{\sim}{x},\underset{\sim}{\theta})]$. Par exemple, pour l'estimateur (2.2), nous constatons que

$$\sigma_{ij}(\underset{\sim}{X},\underset{\sim}{\theta}) = \sum_{k=1}^{N} \sum_{\ell=1}^{N} u_i(\underset{\sim}{X}_k,\underset{\sim}{\theta})\ u_j(\underset{\sim}{X}_\ell,\underset{\sim}{\theta})\ (\pi_{k\ell}-\Pi_k\pi_\ell)/\pi_k\pi_\ell , \tag{2.3}$$

où $\pi_{k\ell}$ représente la probabilité que la k-ième et la ℓ-ième unité soient dans l'échantillon.

Soit $\underset{\sim}{\Sigma}(\underset{\sim}{X},\underset{\sim}{\theta})$ la matrice pxp avec les élements $\sigma_{ij}(\underset{\sim}{X},\underset{\sim}{\theta})$, et $\hat{\underset{\sim}{\Sigma}}(\underset{\sim}{x},\underset{\sim}{\theta})$ un estimateur convergent de $\underset{\sim}{\Sigma}$. Maintenant, pour tout $\underset{\sim}{\theta}$ donné,

$$U_i(\underset{\sim}{X},\underset{\sim}{\theta}) + v_i(\underset{\sim}{\theta}) = \sum_{k=1}^{N} u_i(\underset{\sim}{X}_k,\underset{\sim}{\theta}),$$

de sorte que les estimateurs $\hat{U}_i(\underset{\sim}{X},\underset{\sim}{\theta})$ et $\hat{\underset{\sim}{\Sigma}}(x,\underset{\sim}{\theta})$ peuvent être spécifiés pour tout plan dans lequel nous pouvons dériver des estimateurs convergents asymptotiquement normaux des valeurs totales de la population, de même que des estimateurs convergents des variances des estimateurs des valeurs totales.

L'estimateur de Horvitz-Thompson pour (2.3) est

$$\sum_{k=1}^{n} \sum_{\ell=1}^{n} u_i(\underset{\sim}{x}_k,\underset{\sim}{\theta})\ u_j(\underset{\sim}{x}_\ell,\underset{\sim}{\theta})(\pi_{k\ell} - \pi_k\pi_\ell)/\pi_k\pi_\ell\pi_{k\ell}. \tag{2.4}$$

Dans le cas d'un échantillon de taille déterminée, l'estimateur de Yates-Grundy pour (2.3) est:

$$\sum_{k<\ell}\sum \left[\frac{u_i(\underset{\sim}{x}_k,\underset{\sim}{\theta})}{\pi_k} - \frac{u_i(\underset{\sim}{x}_\ell,\underset{\sim}{\theta})}{\pi_\ell}\right]\left[\frac{u_j(\underset{\sim}{x}_k,\underset{\sim}{\theta})}{\pi_k} - \frac{u_j(\underset{\sim}{x}_\ell,\underset{\sim}{\theta})}{\pi_\ell}\right](\pi_k\pi_\ell - \pi_{k\ell}). \tag{2.5}$$

Soit $\underset{\sim}{U}(\underset{\sim}{X},\underset{\sim}{\theta})$ et $\hat{\underset{\sim}{U}}(\underset{\sim}{x},\underset{\sim}{\theta})$ les vecteurs p-dimensionnels dont les composantes sont respectivement $U_i(\underset{\sim}{X},\underset{\sim}{\theta})$ et $\hat{U}_i(\underset{\sim}{x},\underset{\sim}{\theta})$, nous établissons que

$$\underset{\sim}{J}(\underset{\sim}{X},\underset{\sim}{\theta}) = \partial\underset{\sim}{U}(\underset{\sim}{X},\underset{\sim}{\theta})/\partial\underset{\sim}{\theta} \tag{2.6}$$

$$\hat{\underset{\sim}{J}}(\underset{\sim}{x},\underset{\sim}{\theta}) = U(\underset{\sim}{x},\theta)/\partial\underset{\sim}{\theta} \tag{2.7}$$

où $\underset{\sim}{J}$ et $\hat{\underset{\sim}{J}}$ sont des matrices de dérivées partielles pxp. Supposons que les matrices sont des fonctions continues de $\underset{\sim}{\theta}$ et que les dérivées partielles relatives à existent dans le voisinage de $\underset{\sim}{\theta}_o$. Supposons égalemet que

$\hat{J}(x,\theta)$ est un estimateur convergent de $J(X,\theta)$.

Notre estimateur de θ est donné par $\hat{\theta}$, la solution à :

$$\hat{U}_i(x,\hat{\theta}) = 0, \text{ puisque } i=1, \ldots, p. \tag{2.8}$$

Nous suppons que la taille de l'échantillon est suffisamment grande pour que la solution de l'équation (2.8) soit unique dans Θ. Nous verrons dans la prochaine section que la matrice des covariances de $\hat{\theta}$ peut être estimée de façon convergente par:

$$[\hat{J}^{-1}(x,\hat{\theta})]\ \hat{\Sigma}(x,\hat{\theta})\ \ \hat{J}^{-1}(x,\hat{\theta})]^T .$$

2.2 Théorie asymptotique

Conformément aux raisonnements asymptotiques de Madow (1948) et Hajek (1960), nous étudions une séquence de populations avec l'indice t, de tailles $N^{(t)}$ et à partir de données $X^{(t)}$. Nous supposons que $N^{(t)} \to \infty$ quand $t \to \infty$. Pour la population t, nous tirons un échantillon de taille $n^{(t)}$ et faisons porter nos observations sur les données $x^{(t)}$. Nous posons que $\nu^{(t)} = E(n^{(t)})$ et nous supposons que $\lim_{t\to\infty} \nu^{(t)} = \infty$ et $\lim_{t\to\infty} (N^{(t)} - \nu^{(t)}) = \infty$.

Pour tout θ dans le voisinage de $\theta_0^{(t)}$, nous supposons que

$$[\nu^{(t)}]^{\frac{1}{2}}\ [\hat{U}(x^{(t)}, \theta) - U(X^{(t)},\theta)]/N^{(t)}$$

suit asymptotiquement une loi $N[0,S(\theta)]$, où

$$S(\theta) = \lim[\nu^{(t)}\ \Sigma(X^{(t)},\theta)/\{N^{(t)}\}^2]$$

existe. Nous supposons également l'existence de

$$K(\theta) = \lim J(X^{(t)},\theta)/N^{(t)} \quad \text{et aussi}$$

$$\text{plim}\ \hat{J}(x^{(t)},\theta)/N^{(t)} = K(\theta).$$

Supposons en outre que

$$\lim [\text{rang}\ \{J(X^{(t)},\theta)\}] = \text{plim}\ [\text{rang}\{J\ (x^{(t)}\theta)\}] = p.$$

Nous définissons $\hat{\theta}^{(t)}$ pour satisfaire

$$\hat{U}(x^{(t)},\ \hat{\theta}^{(t)}) = 0\ .$$

Grâce à un développement de la série de Taylor, nous obtenouns

$$U(x^{(t)},\ \theta_o^{(t)} \doteq -\ \hat{J}(x^{(t)},\ \hat{\theta}^{(t)})\ (\theta^{(t)} - \theta_o^{(t)}). \tag{2.9}$$

Comme le premier membre de l'équation (2.9) est asymptotiquement normal, il s'ensuit que

$$(n^{(t)})^{\frac{1}{2}}\ (\theta^{(t)} - \theta_o^{(t)})$$

Soit asymptotiquement une loi $N[0,\ G(\theta_o)]$, où

$$S\ (\theta) = K(\theta_o)\ G(\theta_o)[K(\theta_o)]^T\ .$$

Par conséquent,

$$G(\theta_o) = [K^{-1}(\theta_o)]\ S(\theta_o)\ [K^{-1}(\theta_o)]^T \tag{2.10}$$

et

$$n^{(t)}[\hat{J}^{-1}\ (x,\hat{\theta})]\ \hat{\Sigma}\ (x,\hat{\theta})\ [\hat{J}^{-1}(x,\hat{\theta})]^T. \tag{2.11}$$

est un estimateur convergent de $G(\theta_o)$

En conclusion, lorsque la forme fonctionnelle de $\hat{U}(x,\theta)$ et de $\hat{\Sigma}(x,\theta)$ est spécifiée, il ne nous reste plus qu'à dériver la matrice $\hat{J}(x,\theta_o)$ et son estimateur $\hat{J}(x,\hat{\theta})$ pour pouvoir utiliser ces résulats.

3. EXEMPLES

3.1 Introduction

Dans la présente partie, nous analysons en détail les applications des énoncés généraux formulés dans la deuxième partie au sujet de l'estimation des variances de certains estimateurs des paramètres de la population. Nous examinerons tout particulièrement les rapports, les coefficients de régression et les modèles linéaires logarithmiques se rapportant à des données qualitatives. D'autres modèles comme les modèles de probit pourraient être analysés de façon analogue.

Nous nous fondons généralement sur les notations suivantes. Si $\underset{\sim}{W}_1, \ldots, \underset{\sim}{W}_N$ représentent des valeurs de la population et que $\underset{\sim}{W} = \Sigma \underset{\sim}{W}_k$, lorsque nous tirons un échantillon $\underset{\sim}{w}_1, \ldots, \underset{\sim}{w}_n$, nous avons donc un estimateur sans biais de $\underset{\sim}{W}$ indiqué par $\hat{\underset{\sim}{W}}$. Nous supposons que $\underset{\sim}{V}(\hat{\underset{\sim}{W}})$ représente la matrice des covariances de $\hat{\underset{\sim}{W}}$ et que $\hat{\underset{\sim}{V}}(\hat{\underset{\sim}{W}})$ est un estimateur convergent de $\underset{\sim}{V}(\hat{\underset{\sim}{W}})$. La forme particulière que prendra cet estimateur dépend du plan de d'échantillonage, par exemple, stratifié à plusieurs degrés, ppt avec remise, etc..

3.2 Rapports

Supposons que nous nous intéressions au rapport $R = \Sigma X_{k2}/\Sigma X_{k1}$. Nous définissons

$$U(\underset{\sim}{X},R) = \Sigma X_{k2} - R\Sigma X_{k1}.$$

Par conséquent, dans le cas d'un échantillonnage sans remise, nous observons que

$$\hat{U}(\underset{\sim}{x},R) = \hat{X}_2 - R\hat{X}_1.$$

Si $\hat{\underset{\sim}{U}}(\underset{\sim}{x},\hat{R}) = 0$, nous obtenons

$$\hat{R} = \hat{X}_2/\hat{X}_1. \tag{3.1}$$

Nous posons que $W_k = X_{k2} - R\, X_{k1}$.

Comme $J\,(\underset{\sim}{X},R) = -\Sigma X_{k1}$, nous pouvons affirmer que $(V(\hat{R})$ peut s'approximer par $V(\hat{W})/(\Sigma X_{k1})^2$, estimé par $\hat{V}(\hat{W})/\hat{X}_1^2$. Dans le cas d'un échantillonage stratifé, le résultat obtenu serait le même que dans l'étude de Woodruff (1971).

3.3 Coefficients de régression et R

Supposons que notre matrice de données $\underset{\sim}{X}$ est répartie en $[\underset{\sim}{Z}|\underset{\sim}{Y}]$, la première colonne de $\underset{\sim}{Z}$ étant le vecteur des 1, $\underset{\sim}{Y}$ étant un vecteur par Nx1. Les paramètres d'intérêt θ, $\underset{\sim}{B}$, et R^2 sont définis par:

$$U_1 = \theta - \underset{\sim}{Y}^T \underset{\sim}{1} = 0, \tag{3.2a}$$

$$\underset{\sim}{U}_2 = \underset{\sim}{Z}^T \underset{\sim}{Z}\, \underset{\sim}{B} - \underset{\sim}{Z}^T \underset{\sim}{Y} = 0, \tag{3.2b}$$

$$U_3 = (\underset{\sim}{Y}^T\underset{\sim}{Y}-N^{-1}\theta^2)(R^2-1) + \underset{\sim}{Y}^T \underset{\sim}{Y} - \underset{\sim}{Y}^T \underset{\sim}{Z}\, \underset{\sim}{B} = 0. \tag{3.2c}$$

Dans cet exemple, $\underset{\sim}{B}$ représente le vecteur des coefficients de régression, R^2 le coefficient de détermination multiple et θ, le total de tous les Y. Prenons d'abord le cas où N est connu. Nous supposons que $SSY = \underset{\sim}{Y}^T\underset{\sim}{Y} - N^{-1}\theta^2$. Définissons également, $\underset{\sim}{S}_{ZZ}$ l'estimateur de $\underset{\sim}{Z}^T\underset{\sim}{Z}$, S_{YY}, l'estimateur de $\underset{\sim}{Y}^T\underset{\sim}{Y}$ et $\underset{\sim}{S}_{ZY}$, l'estimateur de $\underset{\sim}{Z}^T\underset{\sim}{Y}$. Par conséquent, nous avons:

$$\hat{\theta} = \hat{Y}\,, \tag{3.3a}$$

$$\hat{\underset{\sim}{B}} = \underset{\sim}{S}_{ZZ}^{-1}\, \underset{\sim}{S}_{ZY}\,, \tag{3.3b}$$

$$\hat{R}^2 = 1 - \frac{S_{YY} - \hat{\underset{\sim}{B}}^T \underset{\sim}{S}_{ZY}}{S_{YY} - N^{-1}\,\hat{Y}^2}\,, \tag{3.3c}$$

et

$$\underset{\sim}{J} = \partial \underset{\sim}{U}(\underset{\sim}{Z},\underset{\sim}{Y},\underset{\sim}{B},R^2,\theta)/\partial(\underset{\sim}{B},\hat{R},\theta) = \begin{bmatrix} \underset{\sim}{0}^T & 0 & 1 \\ \underset{\sim}{Z}^T\underset{\sim}{Z} & \underset{\sim}{0} & \underset{\sim}{0} \\ -\underset{\sim}{Y}^T\underset{\sim}{Z} & SSY & 2\bar{Y}(1-R^2) \end{bmatrix},$$

où $\bar{Y} = \theta/N$.

Il résulte donc que

$$\underset{\sim}{J}^{-1} = \begin{bmatrix} \underset{\sim}{0} & (\underset{\sim}{Z}^T\underset{\sim}{Z})^{-1} & \underset{\sim}{0} \\ -2\bar{Y}(1-R^2)/SSY & B^T/SSY & 1/SSY \\ 1 & \underset{\sim}{0}^T & 0 \end{bmatrix}.$$

Si nous supposons maintenant que $\underset{\sim}{W}_k^T(B) = (Z_{k1}\, e_k, \ldots, Z_{kp}\, e_k)$, où $e_k = Y_k - \sum_j Z_{kj}\, B_j$, nous obtenons:

$$\underset{\sim}{V}[\hat{\underset{\sim}{B}}] \doteq (\underset{\sim}{Z}^T\underset{\sim}{Z})^{-1}\, \underset{\sim}{V}[\hat{\underset{\sim}{W}}(\underset{\sim}{B})](\underset{\sim}{Z}^T\underset{\sim}{Z})^{-1}\,. \tag{3.4}$$

Il s'agit là d'une conséquence directe de l'équation (2.10). Notons que l'ensemble des vecteurs de $\underset{\sim}{W}_k(\underset{\sim}{B})$ correspond à $\underset{\sim}{U}_2$ dans l'équation (3.2b). Fuller (1975) obtient le même résultat avec un échantillon stratifié ou un échantillon stratifié à deux degrés.

Pour estimer (3.4), nous utilisons:

$$\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{B}}] = \underset{\sim}{S}_{ZZ}^{-1}\, \hat{\underset{\sim}{V}}[\hat{\underset{\sim}{W}}(\hat{\underset{\sim}{B}})]\, \underset{\sim}{S}_{ZZ}^{-1}\,.$$

En supposant que $\underset{\sim}{B}$ est l'estimateur du maximum de vraisemblance de $\underset{\sim}{\beta}$, alors $\underset{\sim}{B}$ satisfait

$$\underset{\sim}{U} = \underset{\sim}{Z}^T \underset{\sim}{P}(\underset{\sim}{B}) - \underset{\sim}{Z}^T \underset{\sim}{Y} = 0 , \qquad (3.8)$$

où $\underset{\sim}{P}(\underset{\sim}{B})^T = [p_1(\underset{\sim}{B}), \ldots, p_N(\underset{\sim}{B})]$.

Pour une population finie donnée, nous définissons $\underset{\sim}{B}$ comme notre paramètre d'intérêt.

Nous choisissons $\underset{\sim}{C}(\underset{\sim}{B})$ comme notre estimateur de $\underset{\sim}{Z}^T\underset{\sim}{P}(\underset{\sim}{B})$ et $\underset{\sim}{S}_{ZY}$ comme notre estimateur de $\underset{\sim}{Z}^T\underset{\sim}{Y}$. Par conséquent, $\hat{\underset{\sim}{B}}$ satisfait $\underset{\sim}{C}(\hat{\underset{\sim}{B}}) = \underset{\sim}{S}_{ZY}$. Ces équations doivent habituellement être résolues de façon itérative. Posons également

$$\underset{\sim}{J} = \frac{\partial \underset{\sim}{U}}{\partial \underset{\sim}{B}} .$$

La (i, j)-ième composante de $\underset{\sim}{J}$ est $\sum_k Z_{ki} Z_{kj} p_k(\underset{\sim}{B}) [1-p_k(\underset{\sim}{B})]$. Nous notons l'estimateur de $\underset{\sim}{J}$ par $\hat{\underset{\sim}{J}}$.
Pour estimer la variance de $\hat{\underset{\sim}{B}}$, nous posons:

$$\underset{\sim}{W}_k^T = (Z_{k1}\hat{e}_k, \ldots, Z_{kr}\hat{e}_k)$$

où $\hat{e}_k = p_k(\hat{\underset{\sim}{B}}) - Y_k$. L'estimateur de $\underset{\sim}{V}[\hat{\underset{\sim}{B}}]$ est donné par:

$$\hat{\underset{\sim}{J}}^{-1} \hat{\underset{\sim}{V}}(\hat{\underset{\sim}{W}}) \hat{\underset{\sim}{J}}^{-1}.$$

3.5 Modèles linéaires logarithmiques pour des données qualitatives

Supposons que chaque membre de la population appartienne à une seule catégorie d'un nombre q de catégories distinctes. Un vecteur a_i de rxl est associé à la catégorie i de sorte que la proportion d'individus faisant partie de la i-ième catégorie est approximativement:

$$p_i(\underset{\sim}{\beta}) = \frac{\exp(\underset{\sim}{a}_i^T \underset{\sim}{\beta})}{\sum_j \exp(\underset{\sim}{a}_j^T \underset{\sim}{\beta})} .$$

Nous pouvons également estimer la variance de $\hat{R}^2$. Si $\underset{\sim}{W}_k^T(B,R^2) = [Y_k, Z_{k1}\, e_k, \ldots, Z_{kp}\, e_k, Y_k\,(\sum_j Z_{kj}\, B_j - R^2 Y_k)]$ et $\underset{\sim}{c}^T = [-2\hat{Y}(1-\hat{R}^2)/N, \hat{\underset{\sim}{B}}^T, 1]/(S_{YY}-1^{-1}\,\hat{Y}^2)$, nous obtenons:

$$\hat{V}[\hat{R}^2] \doteq \underset{\sim}{c}^T\, \hat{\underset{\sim}{V}}[\hat{\underset{\sim}{W}}(\hat{\underset{\sim}{B}},\hat{R}^2)]\, \underset{\sim}{c}. \tag{3.5}$$

Dans le cas où N est inconnu, par exemple lorsque les unités primaires d'échantillonnage sont des régions géographiques, nous obtenons l'équation supplémentaire suivante:

$$U_4 = N - \Sigma 1, \tag{3.6}$$

Après avoir ajouté les unités de la ligne et de la colonne appropriées à $\underset{\sim}{J}$ et après inversion, nous pouvons estimer $V[\hat{R}^2]$ de la façon suivante

Nous posons: $\underset{\sim}{W}_k^T(\underset{\sim}{B},R^2) = [Y_k, Z_{k1}\, e_k, \ldots, Z_{kp}\, e_k, Y_k(\sum_j Z_{kj}\, B_j - R^2 Y_k), 1]$

et que $\underset{\sim}{c}^T = [-2\hat{Y}(1-\hat{R}^2)/\hat{N}, \hat{B}^T, 1, \hat{Y}^2(1-\hat{R}^2)/\hat{N}^2]/(S_{YY} - \hat{N}^{-1}\,\hat{Y}^2)$.

Nous pouvons alors déterminer $\hat{V}[\hat{R}^2]$ à l'aide de l'équation (3.5) pour ces nouvelles valeurs de $\underset{\sim}{W}_k(\underset{\sim}{B},R^2)$ et $\underset{\sim}{c}$.

3.4 Régression logistique

Comme dans la section précédente, nous posons comme hypothèse que la matrice de données $\underset{\sim}{X}$ peut être répartie en $[\underset{\sim}{Z}|\underset{\sim}{Y}]$, mais ici $\underset{\sim}{Y}$ est un vecteur de 0 et des 1. Dans le cadre de l'analyse statistique classique, le modèle de régression logistique de $\underset{\sim}{Y}$ conditionnel à $\underset{\sim}{Z}$ implique que les valeurs de $Y_1, \ldots, Y_N$ sont indépendantes et que $Pr(Y_k = 1) = p_k\,(\underset{\sim}{\beta})$, où:
où

$$p_k(\beta) = \frac{\exp(\underset{\sim}{\beta}^T\, \underset{\sim}{z}_k)}{1 + \exp(\underset{\sim}{\beta}^T\, \underset{\sim}{z}_k)}\ . \tag{3.7}$$

Supposons que $\underset{\sim}{p}(\underset{\sim}{\beta})^T = [p_1(\underset{\sim}{\beta}), \ldots, p_q(\underset{\sim}{\beta})]$ et que $\underset{\sim}{N}^T = (N_1, \ldots, N_q)$, où N_i représente le nombre d'individus dans la i-ième catégorie. Si la population est générée à partir d'une distribution multinomiale de probabilité $\underset{\sim}{p}(\underset{\sim}{\beta})$, l'estimateur à maximum du vraisemblance de $\underset{\sim}{\beta}$, donné par $\underset{\sim}{B}$, satisfait

$$\underset{\sim}{U} = \underset{\sim}{A}^T \underset{\sim}{N} - [\underset{\sim}{A}^T \underset{\sim}{p}(\underset{\sim}{B})] \underset{\sim}{1}^T \underset{\sim}{N} = 0,$$

où $\underset{\sim}{A}$ est une matrice qxr dont la i-ième ligne correspond à $\underset{\sim}{a}_i^T$. Nous considérons $\underset{\sim}{B}$ comme notre paramètre d'intérêt de toute population finie donnée.

Soit $\hat{\underset{\sim}{N}}$ un estimateur convergent asymptotiquement normal de $\underset{\sim}{N}$, $\underset{\sim}{V}[\hat{\underset{\sim}{N}}]$ la matrice des variances-covariances et $\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{N}}]$ la matrice estimée. Notre estimateur, $\hat{\underset{\sim}{B}}$, satisfait:

$$\underset{\sim}{A}^T \hat{\underset{\sim}{N}} - [\underset{\sim}{A}^T \underset{\sim}{p}(\hat{B})] \underset{\sim}{1}^T \hat{\underset{\sim}{N}} = 0. \qquad (3.9)$$

Cette méthode d'estimation a été proposée par Freeman et Koch (1976). Elle est peut-être moins efficace que la méthode de régression asymptotique fonctionnelle de Imrey, Koch et Stokes (1981, 1982), mais il n'est pas nécessaire de calculer toutes les composantes de $\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{N}}]$ pour appliquer la formule (3.9).

Nous supposons que $\underset{\sim}{D}(\underset{\sim}{B})$ correspond à $\mathrm{diag}[\underset{\sim}{p}(\underset{\sim}{B})]$ et $\underset{\sim}{H}(\underset{\sim}{B}) = \underset{\sim}{D}(\underset{\sim}{B}) - \underset{\sim}{p}(\underset{\sim}{B})\, \underset{\sim}{p}(\underset{\sim}{B})^T$. Alors

$$\underset{\sim}{J} = \frac{\partial \underset{\sim}{U}}{\partial \underset{\sim}{B}} = - (\underset{\sim}{1}^T \underset{\sim}{N})\, \underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B})\, \underset{\sim}{A}\,.$$

Par conséquent, la matrice de variance asymptotique pour B est donnée par:

$$\underset{\sim}{V}[\hat{\underset{\sim}{B}}] = (\underset{\sim}{N}^T \underset{\sim}{1})^{-2} (\underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B})\, \underset{\sim}{A})^{-1}$$

$$\underset{\sim}{A}^T(\underset{\sim}{I} - \underset{\sim}{p}(\underset{\sim}{B})\underset{\sim}{1}^T)\, V[\hat{N}]\, (\underset{\sim}{I} - \underset{\sim}{1}\, \underset{\sim}{p}(\underset{\sim}{B})^T)\, \underset{\sim}{A}(\underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B})\, \underset{\sim}{A})^{-1}\,. \qquad (3.10)$$

Cette expression peut quelquefois être simplifiée de la manière suivante. Si on peut supposer que $\underset{\sim}{N}/\underset{\sim}{N}^T\underset{\sim}{1} \doteq \underset{\sim}{p}(\underset{\sim}{B})$, alors pour $\hat{\underset{\sim}{\pi}} = \underset{\sim}{N}/n^T\underset{\sim}{1}$ nous avons :

$$\underset{\sim}{V}[\hat{\underset{\sim}{\pi}}] \doteq (\underset{\sim}{N}^T\underset{\sim}{1})^{-2} (\underset{\sim}{I}-\underset{\sim}{p}(\underset{\sim}{B})\underset{\sim}{1}^T) \underset{\sim}{V}[\hat{\underset{\sim}{N}}](\underset{\sim}{I}-\underset{\sim}{1}\ \underset{\sim}{p}(\underset{\sim}{B})^T),$$

de sorte que

$$\underset{\sim}{V}[\underset{\sim}{B}] \doteq (\underset{\sim}{A}^T \underset{\sim}{H}(\underset{\sim}{B})\ \underset{\sim}{A})^{-1}\ \underset{\sim}{A}^T\ \underset{\sim}{V}[\hat{\underset{\sim}{\pi}}]\ \underset{\sim}{A}(\underset{\sim}{A}^T\ \underset{\sim}{H}\ (\underset{\sim}{B})\ \underset{\sim}{A})^{-1}. \qquad (3.11)$$

En outre, la matrice des covariances de $\underset{\sim}{p}(\hat{\underset{\sim}{B}})$, soient les probabilités estimées par case, nous est donnée par:

$$\underset{\sim}{V}[\underset{\sim}{p}(\hat{\underset{\sim}{B}})] = \underset{\sim}{H}(\underset{\sim}{B})\ \underset{\sim}{A}\ \underset{\sim}{V}[\hat{\underset{\sim}{B}}]\ \underset{\sim}{A}^T\ \underset{\sim}{H}(\underset{\sim}{B}).$$

Les estimateurs de $\underset{\sim}{V}[\hat{\underset{\sim}{B}}]$ et de $\underset{\sim}{V}[\underset{\sim}{p}(\underset{\sim}{B})]$ sont des expressions semblables dans lesquelles $\underset{\sim}{N}$ et $\underset{\sim}{B}$ sont remplacés respectivement par $\hat{\underset{\sim}{N}}$ et $\hat{\underset{\sim}{B}}$. On pourrait alors déterminer directement la valeur de $\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{N}}]$. Dans le cas où q est relativement plus quand que r, il serait plus efficace de procéder de la façon suivante. Supposons que

Y_{ki} = 1 si la k-ième unité appartient à la i-ième catégorie
= 0 dans tout autre cas,

par k=1, ..., N; i-1, ..., q. Soient $\underset{\sim}{Y}^T = (Y_{k1}, \ldots, Y_{kq})$, et $\underset{\sim}{W}_k = \underset{\sim}{A}^T\ [\underset{\sim}{I}-\underset{\sim}{p}(\hat{\underset{\sim}{B}})\ \underset{\sim}{1}^T]\ \underset{\sim}{Y}_k$.

Alors

$$\hat{\underset{\sim}{V}}[\hat{\underset{\sim}{B}}] = (\hat{\underset{\sim}{N}}^T\underset{\sim}{1})^2\ (\underset{\sim}{A}^T\ \underset{\sim}{H}(\hat{\underset{\sim}{B}})\ \underset{\sim}{A})^{-1}\ \hat{\underset{\sim}{V}}(\hat{\underset{\sim}{W}})\ (\underset{\sim}{A}^T\ \underset{\sim}{H}(\underset{\sim}{B})\ \underset{\sim}{A})^{-1}.$$

Soulignons que les méthodes décrites dans la présente section peuvent facilement être appliquées aux modèles de distribution multinomiale du produit, dans lesquels il existe un modèle linéaire logarithmique pour $\{N_{ij}\}$, mais où les marges $\{\sum_j N_{ij}\}$ sont connues.

4. DISCUSSION

Les méthodes décrites dans le présent article s'appliquent à certains modèles précis, à titre d'exemple, voir Fuller (1975) et Freeman et Koch (1976). Les résultats généraux n'ont toutefois pas été présentés de façon explicite. Il est possible d'utiliser de nombreux progammes statistiques types pour estimer les paramètres des modèles décrits, mais les variances et les tests d'hypothèses compris dans ces programmes ne seront pas valides.

Les résultats présentés dans cet article sont fondés sur l'hypothèse que les estimateurs sont asymptotiquement normaux. Il importe d'effectuer des études empiriques sur la validité de ces approximations.

D'autres méthodes d'estimation de bon nombre de paramètres décrits dans cet article sont présentées par Imrey, Koch et Stokes (1981, 1982). Leur méthode de régression asymptotique fonctionnelle s'inscrit également dans le cadre général de la présente analyse, quant à l'estimation et la dérivation des variances.

BIBLIOGRAPHIE

[1] Frankel, M.R. (1971), Inference from Survey Samples , University of Michigan, Ann Arbor.

[2] Freeman, D.H. Jr., and Koch, G.G. (1976), "An Asymptotic Covariance Structure for Testing Hypotheses on Raked Contingency Tables from Complex Sample Surveys", Proc. Amer. Statist. Ass. (Social Statistics Section), Part 1, 330-335.

[3] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sankya, Series C, 37, 117-132.

[4] Hajek, J. (1960), "Limiting Distributions in Simple Random Sampling from a Finite Population", Publ. Math. Inst. Hung. Acad. Sci., 5, 361-374.

[5] Imrey, P.B., Koch, G.G., Stokes, M.E. (1981-1982), "Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part I: Historical and Methodological Overview. Part II: Data Analysis", International Statistical Review. À paraître.

[6] Kish, L., and Frankel, M.R. (1974), "Inference from Complex Samples", J. Roy. Statistic. Soc. B, 36, 1-22.

[7] Madow, W.G. (1948), "On the Limiting Distribution of Estimates Based on Samples from Finite Universes", Ann. Math. Statist., 19, 535-545.

[8] Särndal, C.E. (1978), "Design-based and Model-based Inference in Survey Sampling", Scand. J. Statist. 5, 27-52.

[9] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Amer. Statist. Assoc. (Social Statistics Section , 11-18.

[10] Woodruff, R.S. (1977), "A Simple Method for Approximating the Variance of a Complicated Estimate", J. Amer. Statist. Assoc. 66, 411-414.

LA STATISTIQUE DE LA SANTÉ AU CANADA: RÉTROSPECTIVE ET JALONS POUR L'AVENIR[1]

Lorne Rowebottom[2]

L'auteur décrit brièvement les facteurs qui déterminent la production de statistiques dans le domaine de la santé au Canada. Il insiste surtout sur les différentes sources de données et sur la longue tradition de collaboration entre les plusieurs organismes impliqués dans la production d'information dans le domaine de la santé.

Monsieur le président, je tiens tout d'abord à vous dire combien je suis heureux de faire partie de cette table ronde, car cela me permet de rendre hommage à Mme Dorothy Rice et à ses collègues du National Center for Health Statistics qui, depuis vingt-cinq ans déjà, mène des enquêtes dans le domaine de la santé. Les membres de Statistique Canada, en particulier ceux de la Division de la santé, ont toujours admiré le travail accompli par le personnel du Centre et ils se joignent à moi pour féliciter Mme Rice.

Notre président m'a demandé d'examiner des questions qui traduisent non seulement la situation actuelle dans le domaine de la santé au Canada, mais qui peuvent également être représentatives des préoccupations des statisticiens à l'étranger. J'espère que le tableau très général que j'ai choisi de brosser sur les tendances qui se manifestent dans le secteur de la santé au Canada et sur les facteurs qui peuvent en influencer l'évolution répond à ses voeux et saura, par conséquent, vous intéresser.

---

[1] Allocution prononcée à l'assemblée annuelle de l'American Statistical Association tenue à Détroit, en août 1981

[2] Lorne Rowebottom, Statisticien en chef adjoint, Direction de la statistique des institutions et de l'agriculture, Statistique Canada.

Pour produire des statistiques sur la santé, nous utilisons, aujourd'hui comme par le passé, deux grandes sources de renseignements. La première regroupe les établissements de santé, en particulier les hôpitaux généraux et psychiatriques. A partir des données contenues dans leurs dossiers, nous pouvons établir un éventail de statistiques sur les caractéristiques mêmes de ces établissements ainsi que sur leurs patients et les maladies soignées. La statistique hospitalière du Canada compte parmi les plus détaillées et les plus complètes au monde.

Les registres de l'état civil (registres des naissances, des mariages et des décès) constituent notre seconde source de renseignements. Nous en tirons d'importantes statistiques sur les causes de décès.

Des statistiques très variées sont produites à partir de ces deux sources de données précieuses, et certaines statistiques importantes sont tirées d'autres sources, notamment des registres des maladies à déclaration obligatoire et, dans le cas du nombre de cas de cancer, des registres provinciaux des tumeurs. J'ai apporté quelques exemplaires du Répertoire de données de la santé. Les personnes qui ne peuvent en obtenir un aujourd'hui pourront m'écrire à Statistique Canada pour que je leur en fasse parvenir un par la poste.

Les questions que je désire aborder cet avant-midi sont les suivantes:

- Premièrement, nos statistiques ne portent malheureusement que sur les personnes qui s'adressent à des établissements de santé pour s'y faire soigner.

- Deuxièmement, parce qu'elles sont tirées de résultats d'enquêtes ou de documents administratifs, ces statistiques, même si elles sont très utiles, comportent néanmoins des lacunes.

- Troisièmement, l'existence d'un volume aussi considérable de données témoigne bien de la collaboration étroite et constante qui existe depuis de nombreuses années entre Statistique Canada, Bien-être et

Santé Canada, ministères provinciaux de la santé, les établissements de santé, les associations hospitalières et enfin les services d'archives responsables des registres de l'état civil.

Je reviendrai plus loin à ces trois grandes questions que interviennent dans le domaine de la statistique de la santé, c'est-à-dire les renseignements qu'elle nous fournit et ceux qu'elle ne nous fournit pas, les points forts et les lacunes des sources de renseignements consultées, ainsi que le niveau de collaboration existant entre les organismes intéressés.

Comment expliquer les caractéristiques de notre statistique de la santé? Si l'on examine son évolution depuis une soixantaine d'années, j'estime qu'il peut être instructif d'y réfléchir. Il est impossible d'évaluer, en toute objectivité, de quelle façon a été établi l'ordre des priorités dans le passé. D'ailleurs, même aujourd'hui, le choix des priorités comporte une certaine part de subjectivité. Bref, dans le passé, je crois que, d'une part, nous avons cherché à répondre aux nouveaux besoins formulés par les utilisateurs ou exprimés dans les rapports de commissions royales d'enquête et que, d'autre part, nous avons pressenti ces nouveaux besoins et cherché à les satisfaire en exploitant les sources de renseignements existantes parce qu'elles offraient des possibilités intéressantes. Ces sources de renseignements représentaient pour nous le filon d'or que le prospecteur cherche ou découvre par hasard. Soulignons aussi que nous avons dû suivre le courant amorcé par les nombreux progrès technologiques et que nous avons su en tirer profit. En outre, le climat de collaboration dans lequel nous travaillons a influencé aussi la nature de nos travaux. Enfin, les ressources à notre disposition, soit les ressources financières et humaines, et le matériel de traitement des données,nous ont permis d'entreprendre certains travaux mais pas d'autres.

En résumé, l'importance de nos réalisations a été fonction des quelques facteurs suivants:

- les besoins qui étaient portés à notre attention et ceux que nous avions pressentis,

- les sources de renseignements existantes,
- les changements technologiques en matière de traitement et d'analyse des données,
- la collaboration entre les organismes intéressés et enfin
- les budgets consentis.

Il va de soi que ces facteurs décident encore aujourd'hui et décideront demain des travaux que nous pouvons mener à bonne fin.

Ces facteurs changent et se recoupent de sorte qu'il se peut qu'au cours de périodes données, certains aient plus d'influence que d'autres sur le cours et la teneur de nos travaux.

Quoi qu'il en soit, au fil des ans, la statistique de la santé a beaucoup évolué au Canada et il semble évident que dans l'avenir, cette évolution ira en s'accélérant. Ce sont les besoins des utilisateurs qui doivent orienter les travaux statistiques et c'est effectivement ce qui se produit même si nous n'obtenons pas toujours les résultats escomptés. J'ai déjà souligné que les statistiques sur les décès au Canada étaient surtout compilées à partir des données sur la morbidité hospitalière. Toutes les personnes intéressés s'accordaient pour dire que cet état de choses n'était pas satisfaisant. Pour corriger cette situation, l'administration fédérale a établi, il y a quelques années, une enquête permanente sur la santé nationale qui a fait l'objet de travaux de planification détaillés et de tests rigoureux tant en ce qui a trait au cadre théorique qu'en ce qui concerne les méthodes utilisées. Des données ont pu être recueillies pendant dix mois seulement, les compressions budgétaires imposées dans toute l'administration fédérale ayant entraîné l'abandon de l'enquête. Les premiers résultats viennent d'être publiés et la base de données constituée promet d'être une source de renseignements précieuse. Elle constituera en outre un instrument très utile aux responsables de la prise de décisions en matière de santé. Bien entendu, il faudra utiliser ces données avec beaucoup de prudence, étant donné qu'elles portent sur une très courte période. A l'heure actuelle, nous ne sommes pas encore en mesure de prévoir quand l'Enquête Santé Canada sera rétablie. Toutefois, j'ai bon espoir que les travaux reprendront dans un

proche avenir, étant donné que cette enquête permettra de connaître les effets de notre mode de vie sur notre santé et qu'elle aidera la population et les gouvernements à réagir.

Nous venons de voir qu'au Canada, il a été souvent difficile de satisfaire les besoins des utilisateurs par le biais d'enquêtes auprès des ménages. Toutefois, il existe un domaine où nous avons su prévoir ces besoins et avons réussi à y répondre efficacement. Je pense à nos études épidémiologiques dont l'objet est de déterminer les risques que présentent certains facteurs démographiques, sociaux, professionnels et écologiques pour la santé des citoyens. Grâce à la prévoyance et à la ténacité du personnel responsable de la statistique de l'état civil, qui travaille en collaboration avec quelques personnes-ressources de Statistique Canada et d'ailleurs, nous possédons maintenant un fichier informatique sur la mortalité où tous les décès survenus au Canada depuis trente ans sont stockés selon cause du décès. De plus, nous disposons d'un système général de couplages des données qui sert à relier aux données du fichier sur la mortalité les groupes de population soumis aux facteurs énumérés plus haut. Il est également possible d'effectuer le couplage avec le fichier sur l'incidence du cancer qui, bien qu'encore incomplet, est tout de même assez important puisqu'il contient des données recueillies au cours d'une période de dix ans.

Mme Martha Smith, chef de la Section de la santé environnementale et professionnelle, se rendra en Écosse avant la fin du mois pour y faire un exposé sur la question, dans lequel elle présentera une bibliographie composée surtout de documents canadiens. La transcription de cet exposé est offerte sur demande. J'aimerais souligner que Mme Smith et M. John Silins, chef de la Section de la statistique de l'état civil et des registres des maladies, sont dans l'assistance aujourd'hui.

Je me permets de signaler d'autres sources de données qui pourraient contribuer à l'orientation future de l'appareil statistique de la santé au Canada. En effet, certaines bases de données sont très volumineuses et, vrai semblablement, très précieuses, mais, en général, elles ne sont pas beaucoup utilisées pour l'établissement de statistiques nationales.

Je pense en particulier aux dossiers administratifs concernant les régimes d'assurance-maladie provinciaux. On y enregistre annuellement au delà de trente millions de cas où les médecins canadiens doivent donner des soins médicaux primaires. Nous avons déjà démontré les possibilités que ces bases de données offrent sur le plan statistique et nous formulons actuellement des propositions concernant leur exploitation au cours des prochaines années. L'affectation des crédits nécessaires peut cependant présenter certaines difficultés.

Les nouveaux besoins devraient déterminer l'orientation des travaux statistiques. Dans les faits, ce sont les progrès réalisés dans le domaine de l'information qui décident de cette orientation. L'ordinateur exerce une influence prépondérante sur la statistique de la santé et, grâce à lui, des changements profonds vont s'opérer au chapitre de la production et de l'utilisation des statistiques.

L'apparition de l'ordinateur a permis aux gestionnaires, aux spécialistes de la santé, aux administrateurs hospitaliers, aux gouvernements des États, aux administrations provinciales, aux universités et à diverses associations d'utiliser des données sous une forme statistique. Dans la Fonction publique fédérale, l'ordinateur est venu modifier le traitement des données ainsi que leurs applications. En outre, les responsables de nombreux organismes oeuvrant dans le domaine de la santé utilisent maintenant les données informatisées, non seulement dans le cadre de recherches, mais aussi afin de mieux comprendrece secteur d'activité et de prendre des décisions en toute connaissance de cause. Autrefois, les données disponibles étaient peu, voire pas du tout consultées.

J'exagérerai à peine en disant qu'il n'y a pas si longtemps, pratiquement seuls les organismes statistiques nationaux pouvaient s'offrir des ordinateursde grande puissance. Les temps ont beaucoup changé, car aujourd'hui, les grands et petits organismes peuvent disposer sans trop de frais d'installations informatiques très complexes et faciles à exploiter. Les changements qu'entraîne une utilisation répandue de l'ordinateur seront profonds et je crois que nous n'en saisissons pas encore toute la portée.

- Nous pouvons cependant entrevoir l'apparition de nombreux producteurs de statistiques. Dans bien des cas, il s'agira d'organismes qui se considéreront comme des opérateurs de SIG. Il n'en demeure pas moins que la statistique occupera une place prépondérante dans leurs activités.

- Ces mêmes organismes seront également d'importants utilisateurs de statistiques, en particulier en ce qui a trait à leur structure interne et à leur secteur de compétence.

- Par conséquent, ils se feront une idée beaucoup plus juste du milieu dans lequel ils évoluent.

- Ces organismes acquerront donc beaucoup d'autonomie et seront moins portés à se tourner vers l'extérieur pour obtenir des statistiques.

- Il est évident que lorsqu'ils seront en mesure de produire des statistiques à partir des données contenues dans leurs propres documents administratifs, ils seront sans doute moins disposés à répondre à nos questionnaires. Par conséquent, il nous faudra aller chercher les informations à même leurs documents. Il pourrait en résulter une diminution de la production statistique à cause d'un manque de comparabilité des données entre les organismes ou les différents paliers administratifs.

J'ai du mal à prévoir l'incidence que ces changements auront sur l'esprit de collaboration qui doit animer les nombreux organismes qui participent à l'élaboration et à la mise à jour d'un ensemble exhaustif et nécessairement complexe de statistiques sur la santé. Permettez-moi seulement de dire qu'au Canada, en dépit de toutes les pressions exercées, cet esprit ne s'est pas érodé, bien au contraire. J'envisage donc l'avenir avec beaucoup d'optimisme, mais j'ajouterai que les organismes statistiques nationaux devront repenser leur rôle en fonction des besoins dont je viens de vous entretenir. Il me semble évident que notre vocation sera non plus de produire des statistiques, mais d'assurer la coordination des travaux statistiques de tous les organismesparticipants.

En conclusion, j'aimerais aborder ce qui doit, selon moi, devenir prioritaire, soit l'analyse statistique de nos bases de données en vue d'en extrairetoute l'information et d'en définir le sens et la portée pour en tirer vraiment des applications pratiques.

Depuis trop longtemps à Statistique Canada, nous nous sommes bornés à publier des myriades de chiffres sans en tirer d'indicateurs. Nous avons donc laissé aux autres le soin d'extraire l'or du minerai. Bien que nous comptions des réussites à ce chapitre, cette lacune a coûté cher à Statistique Canada et à tous ceux qui oeuvrent dans le domaine de la santé. En effet, nous avons privé les utilisateurs d'une ressource extrêmement précieuse, à savoir les moyens de leur permettre d'améliorer les conditions de vie.

MODELES D'ESTIMATION DES ERREURS D'ECHANTILLONNAGE[1]

P.D. Ghangurde[2]

Ce document présente les résultats d'une étude empirique sur l'adjustement de modèles linéaires logarithmiques en fonction des estimations de caractéristiques de l'enquête sur la population active et de leurs coefficients de variation (CV). Ces caractéristiques ont été regroupées en fonction des effets du plan d'échantillonnage, et des modèles ont été ajustés en fonction des estimations de leur valeur totale et de leurs CV calculés sur une période de douze mois. De tels modèles peuvent être utilisés lorsqu'on doit estimer les CV de nouvelles caractéristiques ou pour fournir des estimations plus précises de la fiabilité des estimations fondées sur des données antérieures. Le problème de l'évaluation de la validité de l'ajustement des modèles est également examiné dans ce document.

1. INTRODUCTION

Ce document présente les résultats d'une étude d'évaluation des modèles de calcul des coefficients de variation (CV) d'estimations de caractéristiques de l'enquête sur la population active du Canada (EPA). L'EPA est une enquête mensuelle menée auprès des ménages, qui est basée sur un plan d'échantillonnage aréolaire stratifié à plusieurs degrés et dont l'échantillon comprend environ 55 000 ménages.

On estime chaque mois les CV d'un ensemble de caractéristiques en appliquant la méthode d'estimation de la variance de Keyfitz selon une approximation de la série de Taylor [4], [5].

---

1 Exposé présenté à l'assemblée annuelle de l'American Statistical Association, en août 1981, à Détroit.

2 P.D. Ghangurde, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.

Cependant, pour des raisons de temps et d'argent, il est impossible de calculer la variance approximative de toutes les estimations obtenues à partir d'une enquête à grande échelle comme l'EPA. Les estimations des CV basées sur des modèles peuvent être utilisées lorsqu'on veut obtenir des estimations provisoires de la fiabilité de nouvelles caractéristiques fondées sur des données antérieures ou lorsqu'il faut étendre la période d'observation (une année par exemple). Les modèles peuvent également être utilisés pour obtenir des estimations précises de la fiabilité, notamment des indicateurs alphabétiques des intervalles des CV.

Dans la partie 2, on explique les modèles linéaires et non linéaires appliqués au calcul des estimations totales et proportionnelles. Les parties 3 et 4 portent sur les facteurs de groupement des caractéristiques, l'ajustement des modèles et l'évaluation de la validité des ajustements.

2. MODÈLES

L'EPA est une enquête mensuelle menée auprès des ménages dans laquelle le logement est l'unité d'échantillonnage du dernier degré. Chaque province du Canada est divisée en régions économiques qui sont formées de groupes de comtés ayant une structure économique semblable. Les régions économiques sont subdivisées en strates géographiques et on tire des échantillons aréolaires sans remise à plusieurs degrés, soit à deux degrés dans les strates autoreprésentatives dans les grands centres urbains et à trois ou quatre degrés dans les strates non autoreprésentatives dans les régions rurales. Aux premiers degrés, la sélection de l'échantillon se fait selon une probabilité proportionnelle à la taille de la population et, au dernier degré, où les logements sont choisis dans les grappes, la sélection est systématique.

Pour produire les estimations relatives aux strates (basées sur le plan d'échantillonnage), on applique aux données un coefficient de pondération qui équivaut à l'inverse des probabilités de sélection. Les estimations finales sont obtenues à la suite d'un ajustement du coefficient de pondération de base pour tenir compte de la non-réponse et par l'estimation

par quotient à l'intérieur des groupes âge/sexe qui sont des strates établies à posteriori. Les projections démographiques du recensement par groupe âge/sexe et par province sont les variables auxiliaires utilisées dans l'estimation par quotient. Pour plus de détails sur le plan d'échantillonnage et l'estimation, voir la publication [5].

Les estimations de la variance de diverses caractéristiques au niveau provincial sont obtenues par une approximation de la série de Taylor, en supposant que les unités primaires d'échantillonnage (UPÉ) à l'intérieur des strates non autoreprésentatives sont choisies de façon indépendante. Dans les strates autoreprésentatives, les grappes échantillonnées sont réparties en deux groupes qui deviennent des pseudo-UPÉ et sont supposées avoir été choisies de façon indépendante. L'estimation de la variance du total estimé d'une caractéristique à l'échelle nationale consiste à faire la somme des estimations correspondantes de la variance à l'échelon provincial [5]. La variance d'une estimation $\hat{X}$ du total X d'une caractéristique dans une province donnée prend donc la forme:

$$V(\hat{X}) = F\ (W-1)\ X\ (1 - \frac{X}{P}),$$

où P = la population de la province

W = l'inverse du taux de sondage

F = l'effet du plan d'échantillonnage pour la caractéristique

n = la taille de l'échantillon (nombre de personnes).

L'expression (1) pour calculer $V(\hat{X})$ relie la variance de l'estimation complexe par quotient qui est fondée sur un plan d'échantillonnage stratifié à plusieurs degrés à la variance de l'estimation basée sur un échantillon aléatoire simple de même taille tiré de la population finie de taille P. La variance d'échantillonnage d'une estimation de la valeur totale d'une caractéristique fondée sur un échantillon aléatoire simple de taille $n(=\frac{P}{W})$ correspond à la variance binomiale courante corrigée en fonction de la population finie. Le terme F, l'effet du plan d'échantillonnage, est un facteur d'accroissement de la variance qui réagit à des facteurs comme la méthode d'échantillonnage à chaque degré, le niveau de stratification et de stratification à posteriori, la taille des unités aux divers degrés et la

constitution de grappes à partir des totaux obtenus pour la caractéristique dans la province. On doit souligner ici que la stratification et la stratification à posteriori ont habituellement pour effet de réduire la variance d'une estimation, alors que la constitution de grappes l'augmente.

De façon générale, les effets du plan d'échantillonnage tendent à être plus importants que les effets dus à l'échantillonnage par grappes des unités de l'EPA. Les effets du plan tendent à être moins élevés dans le cas des catégories de la population active comme celles des "personnes occupées" et des "chômeurs" par groupe âge/sexe, en raison de la stratification à posteriori par âge et sexe qui réduit leur variance. Lorsque la population active est classée selon le secteur d'activité, les effets du plan tendent à être plus élevés à cause de la concentration dans des régions précises. On sait que les effets du plan d'échantillonnage sont fonction des mesures d'homogénéité et de la taille moyenne des grappes. Des modèles fondés sur de tels rapports ont été conçus pour nombre d'enquêtes. Dans une étude sur les composantes de la variance intervenant dans l'EPA, on a analysé les effets du plan d'échantillonnage et les mesures d'homogénéité d'un certain nombre de caractéristiques [2].

Le coefficient de variation nous donne une mesure de précision des estimations sans égard à leur niveau et à l'échelle. Le CV $(\hat{X})$ est donné par

$$CV\ (\hat{X}) = F(W - 1)\left(\frac{1}{X} - \frac{1}{P}\right) . \qquad (2)$$

En utilisant des logarithmes pour déterminer la base e des deux membres de l'expression (2), nous obtenons une équation reliant CV, X et P qui prend la forme:

$$\log CV\ (\hat{X}) = \frac{1}{2} \log F(W - 1) - \frac{1}{2} \log X + \frac{1}{2} \log \left(1 - \frac{X}{P}\right) . \qquad (3)$$

À cause du troisième terme du deuxième membre, l'équation (3) n'est pas linéaire sous la forme log X, même si F(W-1) est supposé constant.

Toutefois, dans le cas des petites valeurs de X, le troisième terme a un effet négligeable. Un modèle fondé sur (3) est exprimé par

$$\log CV(\hat{X}) = A + B \log X + E, \tag{4}$$

où A et B sont les paramètres du modèle et E le terme de l'erreur. L'estimation du paramètre B sera différente de $-\frac{1}{2}$ selon que B log X est plus ou moins une approximation de $\frac{1}{2} \log [X/(1 - \frac{X}{P})]$ dans l'étendue de X. Dans une évaluation des ajustements du modèle (4) et du modèle de rechange (5) donné par

$$\log CV(\hat{X}) = A + B \log \frac{X}{(1 - \frac{X}{P})} E, \tag{5}$$

on a constaté que la validité de l'ajustement, exprimée par R^2, soit le rapport de la somme des carrés de la régression sur la somme des carrés totale, était sensiblement la même pour les deux modèles. Le modèle (4) est linéaire sous la forme log X et log CV et peut être appliqué plus facilement que le modèle (5).

Un modèle non linéaire correspondant à (4) prend la forme:

$$CV(\hat{X}) = A' X^{B'} + E, \tag{6}$$

où A' et B' sont les paramètres du modèle et ε est le terme de l'erreur. Les deux modèles (4) et (6) ont été ajustés en fonction des estimations mensuelles et du (CV) de 90 caractéristiques dans chacune des 10 provinces.

3. GROUPEMENT DES CARACTÉRISTIQUES

On a calculé la moyenne des effets du plan mensuels des estimations de l'EPA pour la période de janvier à décembre 1980 et pour chacune des 90 caractéristiques, sauf la population totale de chaque province et du Canada, puis on les a portées sur un graphique afin de déterminer l'étendue

des deux groupes. Dans chaque province, le premier groupe comprenait les caractéristiques affichant des effets plus grands que D.

Le tableau 1 indique les valeurs limites D des groupes I et II dans chaque province et pour le Canada, de même que le nombre de caractéristiques incluses dans le groupe II. Pour grouper les caractéristiques, on les a classées par ordre croissant de la moyenne des effets du plan. La valeur limite D a été choisie de façon que l'hypothèse de l'égalité des effets du plan soit satisfaite le plus possible dans le groupe I. Le deuxième groupe comprend toutes les autres caractéristiques pour lesquelles l'égalité des effets du plan est supposée plus rudimentaire. La plupart des caractéristiques relatives à la situation vis-à-vis de l'activité, classées par groupe âge/sexe se retrouvent dans le groupe I. Les personnes occupées par secteur d'activité et la durée de la période de chômage entrent habituellement dans le groupe II. Les effets moyens varient beaucoup entre les provinces et pour le Canada. On examine actuellement l'application d'un groupement des caractéristiques plus détaillé basé sur des modèles des effets du plan.

Il convient de souligner qu'environ 80 % des caractéristiques dans chaque province et pour le Canada ont été classées dans le groupe I. Pour produire une estimation prudente du CV de nouvelles caractéristiques, on peut utiliser des modèles basés sur le groupe II. Dans le cas d'une caractéristique dont les estimations mensuelles du CV sont calculées régulièrement, on peut appliquer les modèles construits pour le groupe dans lequel cette caractéristique est classée, si on veut obtenir une estimation approximative du CV qui sera plus précise que celle basée sur les données mensuelles.

Dans la partie suivante, nous examinons les hypothèses appliquées à l'ajustement des modèles (4) et (6) et nous évaluons ces ajustements.

4. ÉVALUATION DES MODÈLES

Pour ajuster le modèle linéaire logarithmique (4), il faut essentiellement considérer ce modèle comme un modèle de régression linéaire simple sous la forme $y = \log CV(\hat{X})$ et $x = \log X$, et calculer les estimations des paramètres A et B selon la base de régression linéaire. Les hypothèses usuelles d'indépendance des erreurs et de variance constante s'appliquent et, selon ces hypothèses, R^2 fournit une mesure de l'ajustement du modèle. Les valeurs des paramètres estimés et les coefficients de détermination, R^2, dans les groupes I et II, relativement aux 10 provinces et au Canada, figurent au tableau 2. L'ajustement effectif de ces modèles a été fait à l'aide du programme SAS.

Toutes les valeurs R^2 sont significatives et assez élevées, ce qui démontre que les ajustements sont très précis. Le graphique des erreurs ne révèle pas de tendance qui permette de conclure que l'hypothèse de la variance constante n'est pas satisfaite. En vertu de ces hypothèses et compte tenu de la normalité des erreurs, $CV(\hat{X})$ présente une distribution log-normale avec un CV constant pour n'importe quelle valeur de X.

Le modèle non linéaire (6) a été ajusté selon la méthode Gauss-Newton à l'aide du programme SAS. On a supposé que les valeurs initiales des paramètres A' et B' étaient 1.00 et -0.50 respectivement. Le nombre d'itérations requises pour atteindre la convergence a été de 8 au maximum pour chaque province et pour le Canada, la norme de convergence étant que l'écart relatif entre plusieurs sommes des carrés résiduelles doit être inférieur à 10^{-8}. Le tableau 3 contient les valeurs des paramètres estimés et la somme des carrés résiduelle pour le Canada, dans le groupe II. Les erreurs affichent une distribution approximativement normale puisque les points sur le graphique révèlent une distribution normale de probabilités.

Comme la comparaison des ajustements du modèle non linéaire appliqué aux provinces, au Canada, pour les deux groupes, présente un certain intérêt, il est nécessaire de définir une norme de validité de l'ajustement. Dans le modèle non linéaire, la somme des carrés totale n'est pas égale au total de

la régression et de la somme des carrés résiduelle. La norme R'^2 peut être exprimée

$$R'^2 = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y}_i)^2},$$

où les $\hat{Y}_i$ sont les CV estimés basés sur le modèle, les $\hat{Y}_i$ sont les CV observés et $\bar{Y}$ leur moyenne. La sommation s'étend à N, qui est le nombre de caractéristiques dans le groupe, multiplié par 12, le nombre de mois. Dans le modèle linéaire, $R^2 = R'^2$. Cependant, dans le modèle non linéaire, R^2 est différent de R'^2, car la somme des carrés totale n'est pas égale à la somme des carrés de la régression plus la somme des carrés résiduelle parce que le produit n'est pas zéro.

Les erreurs $(Y_i - \hat{Y}_i)$ sont petites lorsque l'ajustement est précis et produit une valeur de R'^2 proche de 1; les erreurs $(Y_i - \hat{Y}_i)$ sont importantes lorsque l'ajustement est faible et produit une petite valeur de R'^2. Lorsque tous les points se trouvent sur la courbe ajustée, c'est-à-dire que $Y_i = \hat{Y}_i$ pour tous les i, alors $R'^2 = 1$. Toutefois, règle générale, il ne semble pas y avoir de borne inférieure à R'^2. Les valeurs de R'^2 présentées dans le tableau 4 semblent être plus élevées dans le groupe I que dans le groupe II qui comprend de 13 à 21 caractéristiques sur 90.

Bien que le modèle linéaire logarithmique (4) ait été ajusté en fonction des logarithmes des estimations et de leur CV et que l'ajustement semble précis, les modèles ajustés pour les provinces et le Canada sont appliqués à l'estimation des CV des estimations. Afin de pouvoir comparer l'ajustement du modèle transformé aux estimations initiales et à leur CV, les données et le nouveau modèle (4) ont été portés sur un graphique, et ce pour les deux groupes relatifs aux 10 provinces et au Canada. À partir de ces diagrammes, on peut conclure que ce nouveau modèle (4) peut ajuster les estimations et leur CV de façon plus précise que le modèle non linéaire (6), surtout dans

le cas des petites valeurs. Les graphiques 1 et 2 illustrent ces modèles appliqués aux valeurs du groupe II à l'échelle du Canada.

5. CONCLUSION

Les caractéristiques étudiées sont le nombre total de personnes classées selon leur situation vis-à-vis de l'activité par groupe âge/sexe, secteur d'activité et état matrimonial, de même que le nombre total de personnes selon la durée de la période de chômage. Les modèles peuvent être utilisés aussi bien pour la production de proportions que pour celle des totaux, mais ils ne peuvent pas être appliqués aux estimations relatives aux régions infraprovinciales, notamment les centres urbains ou les groupes de régions économiques, parce que les effets du plan d'échantillonnage dans ces régions sont moins stables et peuvent être beaucoup plus élevés à cause de l'effet de la correction du quotient effectuée à l'aide des projections démographiques provinciales [1].

On a posé comme hypothèse d'application de modèles pour une nouvelle caractéristique que l'effet du plan d'échantillonnage est proche de la moyenne établie pour le groupe. Il faut alors détailler encore plus le groupement des caractéristiques en fonction de modèles qui permettent de lier les effets du plan aux mesures d'homogénéité de ces caractéristiques. Pour l'ajustement des modèles, on a supposé que les erreurs ne sont pas corrélées et que la variable indépendante est fixe. Comme on a utilisé 12 estimations mensuelles de chaque caractéristique, il est possible que les erreurs relatives aux estimations d'une caractéristique particulière soient corrélées. On examine actuellement la possibilité d'étendre cette étude de façon à englober les modèles qui tiennent compte des erreurs produites avec la variable indépendante et des erreurs corrélées.

Le problème qui se pose quant à l'évaluation de l'ajustement de modèles non linéaires, qu'ils soient effectivement ajustés aux données ou qu'il s'agisse de modèles linéaires transformés, est l'absence d'une norme de comparaison des ajustements de divers modèles. Il est probable que la norme proposée à la partie 4 convienne pour la comparaison des ajustements de modèles à

diverses séries de données, mais il est probable également que cela ne convienne pas à des modèles différents.

TABLEAU 1: VALEURS LIMITES DE L'EFFET DU PLAN D'ÉCHANTILLONNAGE ET NOMBRE DE CARACTÉRISTIQUES DANS LES GROUPES I ET II*

Province	Valeur limite (D)	Nombre de caractéristiques	
		Groupe I	Groupe II
Terre-Neuve	2.3	75	15
Île-du-Prince-Édouard	1.9	73	17
Nouvelle-Écosse	1.9	74	16
Nouveau-Brunswick	2.2	77	13
Québec	1.9	73	17
Ontario	1.7	69	21
Manitoba	2.0	76	14
Saskatchewan	2.8	76	14
Alberta	2.1	71	19
Colombie-Britannique	2.3	73	17
Canada	1.9	77	13

* Une caractéristique est classée dans le groupe I si l'effet du plan d'échantillonnage pour cette caractéristique (dont la moyenne est calculée sur une période de 12 mois, soit de janvier à décembre 1980) est inférieur ou égal à la valeur limite D. Si l'effet moyen est supérieur à D, la caractéristique est alors incluse dans le groupe II.

TABLEAU 2: COEFFICIENTS DE RÉGRESSION ET R^2 SELON LE MODÈLE LINÉAIRE LOGARITHMIQUE

Province	Groupe	Coefficient de régression A	B	R^2
Terre-Neuve	I	3.3119	-0.5723	0.9534
	II	3.7757	-0.6101	0.9377
Île-du-Prince-Édouard	I	2.7962	-0.5617	0.9485
	II	3.1796	-0.5885	0.8887
Nouvelle-Écosse	I	3.4612	-0.5837	0.9702
	II	3.6412	0.5257	0.8717
Nouveau-Brunswick	I	3.2782	-0.5545	0.9606
	II	3.7544	-0.6017	0.9357
Québec	I	4.3298	-0.5942	0.9686
	II	4.3093	-0.5216	0.9127
Ontario	I	4.3825	-0.6053	0.9736
	II	4.1796	-0.5009	0.9633
Manitoba	I	3.5155	-0.5926	0.9619
	II	3.8769	-0.5640	0.9166
Saskatchewan	I	3.3796	-0.5700	0.9544
	II	3.5478	-0.4423	0.8994
Alberta	I	3.6960	-0.5968	0.9678
	II	3.7526	-0.5090	0.9513
C.-B.	I	3.9847	-0.5750	0.9621
	II	3.9814	-0.4708	0.8410
Canada	I	4.3458	-0.5936	0.9703
	II	4.2357	-0.5191	0.9699

TABLEAU 3: MOINDRES CARRÉS NON LINÉAIRES: MÉTHODE DE GAUSS-NEWTON

CANADA (GROUPE II)

Itération	A'	B'	Somme des carrés résiduelle
0	1.00000000	-0.50000000	3401.93232121
1	15.22076853	-0.23647629	461.76322678
2	26.47981387	-0.36743343	322.67707190
3	51.94184546	-0.51147529	248.68405130
4	57.29455529	-0.47434886	99.32440727
5	58.32558100	-0.48419609	96.57832290
6	58.28627964	-0.48409502	96.57810754
7	58.28746710	-0.48409960	96.57810746

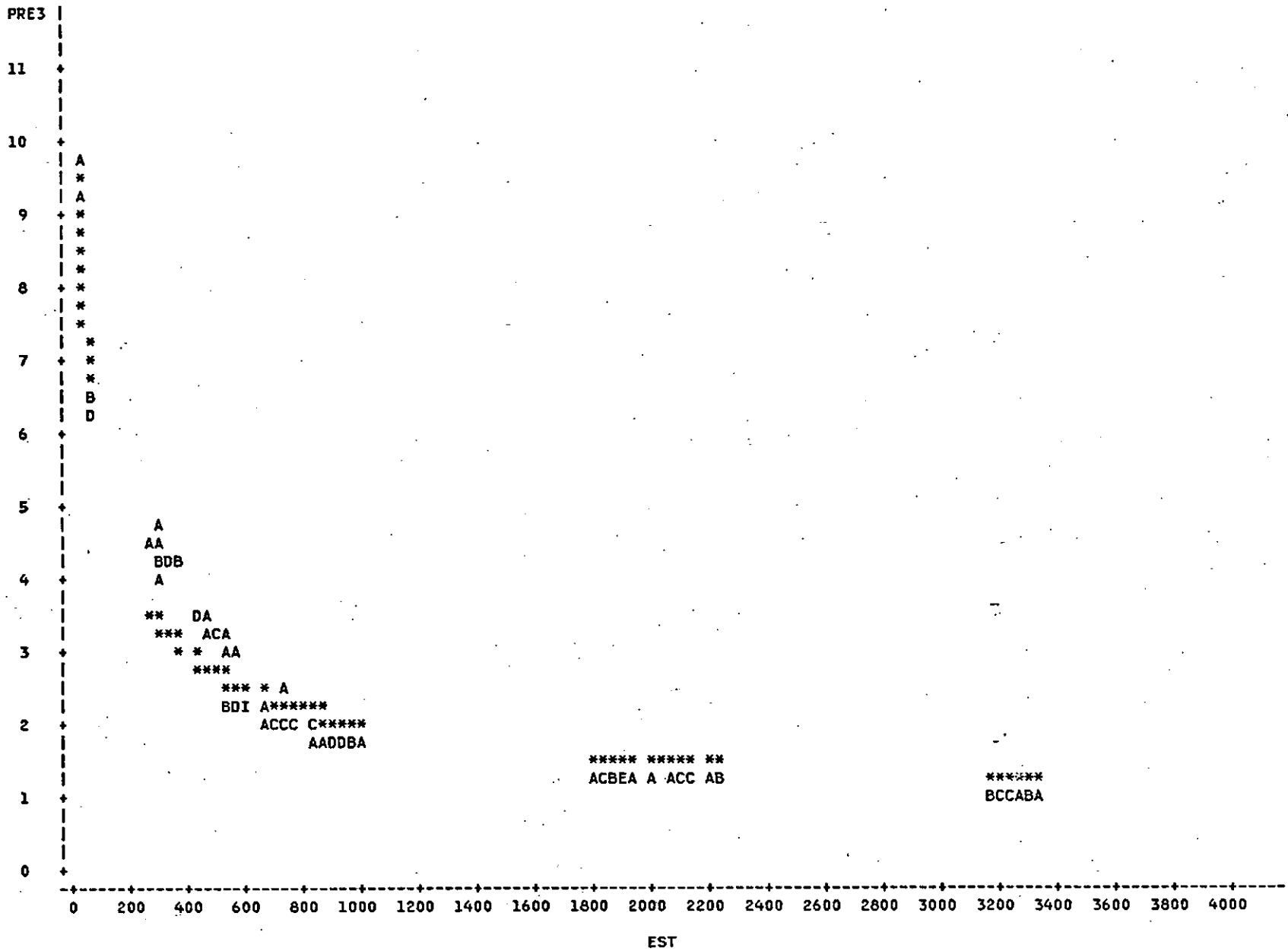
TABLEAU 4: R'^2 DANS LES GROUPES I ET II

Province	Groupe	N*	$R'^2 = 1 - \frac{\text{Somme des carrés résiduelle}}{\text{Somme des carrés totale}}$
Terre-Neuve	I	866	0.9362
	II	190	0.8835
Île-du-Prince-Édouard	I	827	0.8925
	II	294	0.7285
Nouvelle-Écosse	I	872	0.9790
	II	192	0.7813
Nouveau-Brunswick	I	908	0.9990
	II	156	0.8639
Québec	I	859	0.9800
	II	204	0.7804
Ontario	I	823	0.9632
	II	252	0.9208
Manitoba	I	895	0.9691
	II	168	0.8137
Saskatchewan	I	896	0.9436
	II	168	0.8196
Alberta	I	845	0.9701
	II	228	0.8852
C.-B.	I	868	0.9319
	II	204	0.7786
Canada	I	923	0.9665
	II	156	0.9286

* N dans le groupe I peut être inférieur à 12 (nombre de caractéristiques) à cause de l'exclusion des caractéristiques sans estimation.

#3 SIMPLE POWER MODEL
PROV=CANADA GROUP=2

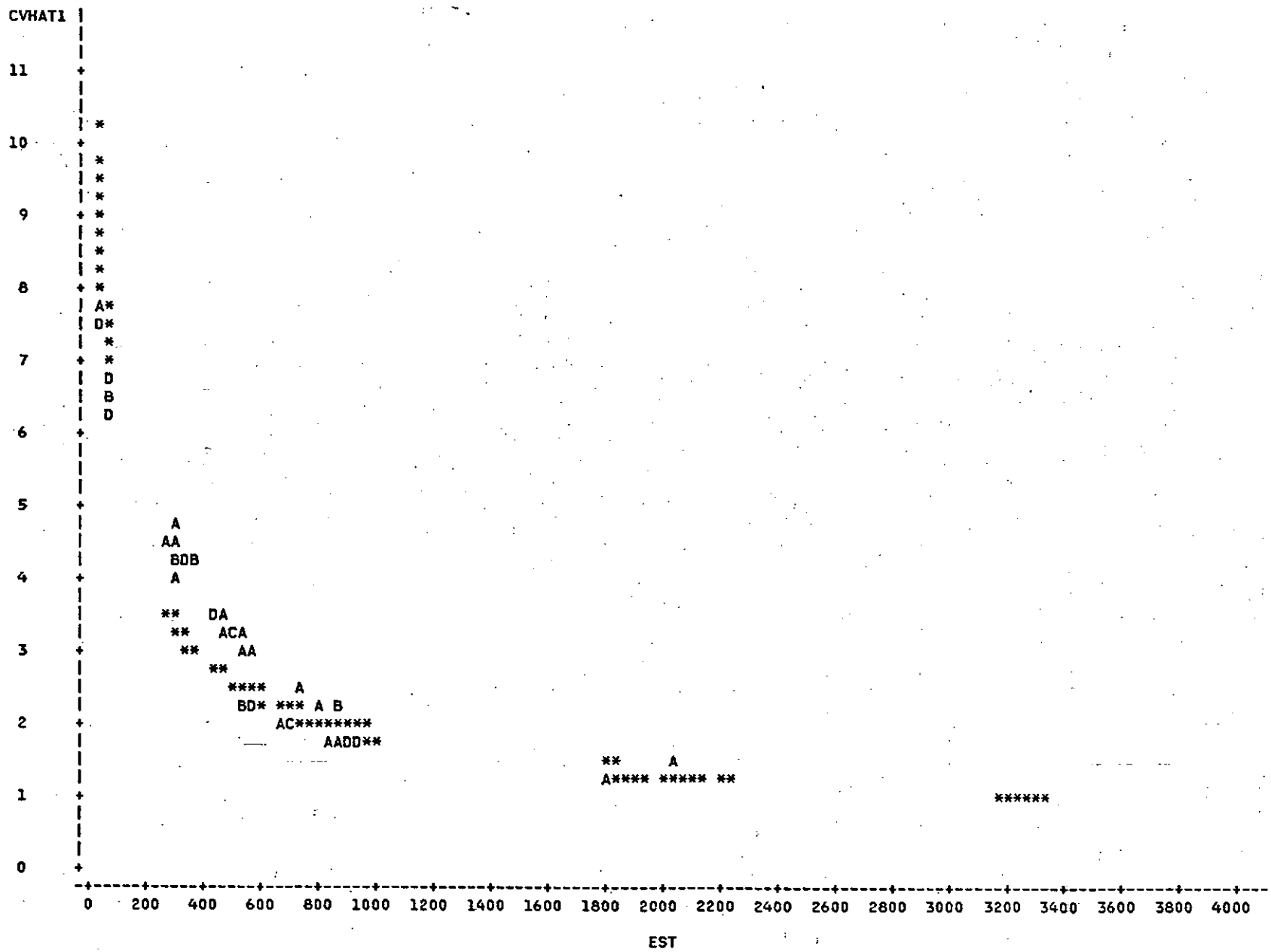
PLOT OF PRE3*EST SYMBOL USED IS *
PLOT OF CV*EST LEGEND: A = 1 OBS, B = 2 OBS, ETC.



NOTE: 101 OBS HIDDEN

#1 LOG-LINEAR MODEL
PROV=CANADA GROUP=2

PLOT OF CVHAT1*EST SYMBOL USED IS *
PLOT OF CV*EST LEGEND: A = 1 OBS, B = 2 OBS, ETC.



NOTE: 98 OBS HIDDEN

REMERCIEMENTS

L'auteur aimerait remercier l'arbitre pour ses commentaires et ses observations.

BIBLIOGRAPHIE

[1] Ghangurde, P.D. et Gray, G.B. (1981), "Estimation for Small Areas in Household Surveys", Communications in Statistics Theory and Methods, A 10(22), 2327-38.

[2] Gray, G.B. et Platek, R. (1976), "Analysis of Design Effects and Variance Components in Multistage Surveys", Techniques d'enquête, vol. 2, n° 1, 1-30.

[3] Kalton, G. (1977), "Practical Methods for Estimating Survey Sampling Errors", exposé présenté à une assemblée de l'International Association of Survey Statisticians.

[4] Keyfitz, N. (1957), "Estimates of Sampling Variance Where Two Units are Selected from Each Stratum", Journal of the American Statistical Association, 52, 503-510.

[5] Platek, R. et Singh, M.P. (1976), Méthodologie de l'enquête sur la population active, n° 71-526 au catalogue, hors-série.

[6] Sprent, P. (1969) "Models in Regresion and Related Topics", Methuen and Co.

Le comité de rédaction désire remercier M. R.E. Drover, Président, Comité de publication et le personnel de la Division des services administratifs pour leur appui constant dans le publication de cette revue.

De même que N. Brien, M. Fluet, P. Foy et G. Kriger pour leur aide lors de la révision des textes et la préparation de la version finale.

Nous remercions également Mme. D. Edirisinghe, qui a dactylographié soigneusement la revue, et qui s'est acquittée de nombreuses autres tâches associées à sa publication.

Le comité de rédaction désire enfin remercier les personnes suivantes, qui ont bien voulu faire la critique des articles présentés au cours de l'année dernière.

D.A. Binder
R.G. Carter
G.H. Choudhry
D.P. Dixon
J.D. Drew
S. Earwaker
M. Fluet
P. Foy
G.B. Gray
M.A. Hidiroglou
G. Kriger
S. Kumar
M.L. Lawes
I. Macredie
M.J. March
A. Satin
K.P. Srinath
L. Swain
P.F. Timmons

TECHNIQUES D'ENQUÊTE

juin 1981 Vol. 7 No. 1

Préparé par les méthodologistes de Statistique Canada

TABLE DES MATIÈRES

Coordination des enquêtes - philosophie et pratique
F. MAYDA et P. TIMMONS 1

Imputation dans les enquêtes: affrontes la réalité
J.G. SANDE 21

Le ramaniement des enquêtes permanentes dans un milieu en mutation
M.P. SINGH et J.D. DREW 44

La méthodologie de l'énquête sur le transport rontier de marchantises pour le compte d'autrin
R. LUSSIER 74

Calcul des "probabilités de travail" et des probabilitiés conjointes de sélection de la méthode d'échantillonnage à probabilités inégales de Fellegi
G.H. CHOUDHRY 93