

REVUE DE STATISTIQUE APPLIQUÉE

F. FACY

Y. LECHEVALLIER

Traitement des non réponses et des données manquantes pour des variables qualitatives après classification automatique

Revue de statistique appliquée, tome 26, n° 4 (1978), p. 39-53

http://www.numdam.org/item?id=RSA_1978__26_4_39_0

© Société française de statistique, 1978, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

TRAITEMENT DES NON RÉPONSES ET DES DONNÉES MANQUANTES POUR DES VARIABLES QUALITATIVES APRÈS CLASSIFICATION AUTOMATIQUE

F. FACY

INSERM – 44, Chemin de ronde – 78110 LE VÉSINET

Y. LECHEVALLIER

IRIA-LABORIA – Domaine de Voluceau – 78150 LE CHESNAY

RESUME

La pratique des enquêtes par questionnaire révèle une fréquence importante des non réponses. La méthode proposée ici traite la non-réponse dans le cas des variables qualitatives. La population étudiée est supposée divisée en k groupes homogènes. Le but poursuivi est de décider pour chaque groupe homogène et pour chaque variable entre les deux éventualités suivantes :

- Soit la non-réponse doit être conservée comme une modalité à part entière ;
- soit elle doit être estimée par une autre modalité de réponse à la question.

Cette décision est fondée sur des calculs de chi-deux de liaison entre chaque variable comportant des non-réponses et la variable présence-absence associée à chaque groupe.

Un exemple d'utilisation est ensuite fourni, extrait d'une enquête de Santé Publique, abordant le problème des tentatives de suicide chez les adolescents.

INTRODUCTION

Par expérience des enquêtes conduites à travers des questionnaires, on a pu remarquer une relative fréquence des non-réponses, dues à différentes causes :

- soit par faute de l'enquêteur ou absence du sujet ;
- soit par la nature, trop vague par exemple, de la question ;
- soit par la volonté délibérée du sujet de ne pas répondre à la question.

Pour recouvrir ces diverses situations, le terme de non-réponse est entendu ici dans un sens très étendu. Il englobe l'erreur, l'oubli et la volonté du sujet interrogé pour chaque question.

La multiplicité des non-réponses, observée dans certaines enquêtes et la variété de leurs motifs indiquent qu'il faut leur donner une signification différente

suyvant les individus et suivant les questions. Pour atteindre cet objectif, on est amené à faire une approche globale du problème pour proposer une solution qui suivant les cas, conserve la non réponse, en lui donnant un sens propre, ou la remplace par d'autres modalités de réponse présentes dans le questionnaire.

Pour illustrer le problème, un exemple est choisi dans une enquête de Santé Publique, concernant des adolescents ayant accompli une tentative de suicide cf. [Da 76a], [Da 76b]. [Fa 78]. L'enquête est constituée par un ensemble de questions caractérisant, pour chaque sujet, sa situation socio-familiale, son comportement, ses antécédents cliniques, son acte. Chaque question forme ainsi, pour l'échantillon des sujets interrogés, une variable qualitative, à plusieurs modalités de réponse, la non-réponse étant l'une d'entre elles. Cet exemple sert d'application de la méthode proposée, par la suite.

De nombreuses méthodes ont été développées dans le but de réduire le moins possible l'ensemble des observations, ce qui permet une certaine robustesse dans les résultats.

On peut distinguer 3 voies :

- la plus ancienne revient à estimer les paramètres d'un modèle, (souvent la loi normale) et à remplacer la non-réponse par une donnée simulée cf. [Wi 32], [Af 66], [Ho 68] ;
- d'autres études ont été faites dans le cas multivarié sans hypothèse préalable sur les lois des variables. Leur but principal est de pouvoir utiliser simultanément ou par la suite des méthodes factorielles sur l'ensemble des objets analysés. On peut citer Christofferson [Ch 70] pour l'analyse en composantes principales, et pour le traitement des variables qualitatives, en vue d'une analyse des correspondances, J.P. Benzecri [Be 73] et Nora [No 75]. On peut utiliser indirectement aussi les travaux de B. Escofier [Es 76] ;
- enfin, si on observe une hétérogénéité de l'ensemble analysé, un traitement par une classification automatique s'impose, d'où la nécessité d'adapter des méthodes de classification à des données manquantes cf. [Ok 75], [Ha 75], [Di 76].

Ces méthodes s'appliquent à des données quantitatives, car schématiquement, elles remplacent la donnée manquante par une valeur numérique assimilée sans problème à une observation possible sur cette variable (par exemple centre de gravité). Par contre, avec des données qualitatives, il faut absolument remplacer la non-réponse par une autre modalité, d'où l'impossibilité d'utiliser les méthodes précédentes. Ainsi dans le traitement d'un tableau disjonctif complet, la suppression de la non-réponse entraîne la redistribution des individus à non-réponse sur les autres modalités cf. [Le 77].

Dans cette optique, nous proposons ici un traitement qui permet d'appliquer les méthodes de classification à des données dans le cas de variables qualitatives, avec le souci important de ne pas supprimer à tout prix la modalité de "non-réponse" cf. [Fa 78].

1. PRINCIPE DE LA METHODE

Hypothèse préliminaire

Afin de considérer les non-réponses par rapport à l'ensemble des informations du fichier, on fait l'hypothèse suivante :

- la population étudiée se décompose en groupes homogènes au regard des variables analysées et la modalité de non-réponse peut se différencier suivant ces groupes. On suppose que la non-réponse n'a pas une signification unique pour l'ensemble des sujets qui est hétérogène a priori.

1.1. Position du problème

Le problème est d'analyser la non-réponse dans le cadre des variables qualitatives. Cette analyse doit nous permettre pour chaque groupe homogène de décider entre les deux éventualités suivantes :

- soit la non-réponse doit être conservée comme une modalité à part entière ;
- soit elle doit être estimée par une autre modalité de réponse à cette question.

Pour analyser la non réponse d'une variable dans une population par rapport aux autres modalités, on cherche quelle est la part d'explication de chaque modalité de réponse dans l'organisation de la population en K groupes homogènes ; la mesure de l'explication est faite, ici, par le χ^2 . Ces K groupes homogènes peuvent être fournis soit par l'utilisateur soit par une méthode de classification automatique, par exemple la Méthode des Nuées Dynamiques [Di 72], [Le 74], [Di 75]. L'objectif est d'approcher la population dans une perspective très orientée, celle d'expliquer successivement K variables dichotomiques. Ces K variables sont les fonctions indicatrices des K groupes déterminés au départ. Dans le cas où les résultats sont obtenus à partir d'une classification automatique qui regroupe les sujets en types :

- la variable à expliquer δ_i est l'appartenance à un groupe E_i de la typologie :

$$\delta_i(x) = 1 \text{ si } x \in E_i, \delta_i(x) = 0 \text{ si } x \notin E_i$$

- les variables explicatives sont les différentes modalités j_q d'une variable j parmi lesquelles se trouve la non-réponse j_0 . (La même étude est répétée pour toutes les variables comportant une modalité de non-réponse).

1.2. Méthodologie

Pour chaque groupe E_i de la population on considère une partition en s classes de l'ensemble des modalités de la variable j. Le but de l'étude est de rechercher le découpage des modalités j_q en un certain nombre de classes qui caractérise le mieux le groupe E_i . Pour interpréter la non-réponse d'une variable dans un groupe homogène de sujets, il faut déterminer sa meilleure position relative, par rapport aux autres modalités. Celle-ci est donnée lorsque la correspondance entre les 2 variables δ_i et j, dont certaines modalités peuvent être regroupées, assure un χ^2 maximum.

définitions

On considère le tableau de données $T_0 = (E \times J)$ qui est le tableau disjonctif complet des deux ensembles :

$$E = \{\text{sujets}\} \text{ et } J = \{\text{variables}\}$$

	J	variable 1 ----	variable j j ₀ ... j _q ... j _{r-1}	variable v
E				
			t _{ℓj_q}	
		ℓ		

$$t_{\ell j_q} = 1 \text{ si le sujet } \ell \text{ est caractérisé par la modalité } j_q \text{ de } j, 0 \text{ sinon.}$$

Après classification P des sujets sur l'ensemble des variables, il existe pour chaque variable j le tableau $T_1^j = (P \times j)$

	j	j ₀	j _q	j _{r-1}
P				
groupe E ₁				
⋮				
⋮				
groupe E _i			$\sum_{\ell \in E_i} t_{\ell j_q}$	
⋮				
⋮				
groupe E _k				

$\sum_{\ell \in E_i} t_{\ell j_q}$ est le nombre de sujets du groupe E_i ayant la modalité j_q pour la variable j.

Propriétés

Le tableau $(\delta_i, j) = T_2^j$ se fait à partir de T_0 . T_2^j est la concaténation de T_1^j , si on ne considère qu'un groupe E_i et son complément \bar{E}_i .

δ_i	j	$j_0 \dots\dots\dots j_q \dots\dots\dots j_{r-1}$
E_i		n_{ij_q}
\bar{E}_i		$n_{i\bar{j}_q}$

n_{ij_q} est le nombre de sujets de la classe E_i ayant la modalité j_q pour j :

$$n_{ij_q} = \sum_{\varrho \in E_i} t_{\varrho j_q}$$

$$n_{i\bar{j}_q} = \sum_{\varrho \notin E_i} t_{\varrho j_q}$$

Le tableau T_3^j se fait à partir de T_2^j par la répartition des r modalités en s classes :

δ_i	S	$S_1 \dots\dots\dots S_h \dots\dots\dots S_s$
E_i		$\sum_{j_q \in S_h} n_{ij_q}$
\bar{E}_i		$\sum_{j_q \in S_h} n_{i\bar{j}_q}$

La variable à expliquer δ_i est dichotomique, présentant deux états : l'appartenance à E_i et l'appartenance au reste.

Un χ^2 est calculé sur chaque tableau T_3^j correspondant à une distribution des j_q en s classes.

D'après les propriétés du χ^2 , en retenant le découpage qui donne le χ^2 maximum, on a, au regard des modalités, un *regroupement qui assure la meilleure homogénéité de ces classes vis-à-vis de δ_i* .

Ainsi : un pronostic est possible à partir de la connaissance des j_q d'un individu de la population pour caractériser δ_i et réciproquement à partir de δ_i , prévoir les j_q : ce dernier point est très utile ici.

Dans le tableau (δ_i, S') dont le χ^2 est maximum pour une certaine distribution des j_q , la répartition des modalités est telle que la non-réponse j_0 se trouve seule ou avec d'autres modalités de façon "très homogène".

L'interprétation de la non-réponse en découle : suivant la présence de modalités dans la même classe que j_0 , on peut dire que j_0 et celles-ci sont proches pour les sujets de E_i par opposition aux modalités restantes.

Cette proximité éventuelle de j_0 avec d'autres modalités donne le sens de la non-réponse d'une variable j pour un groupe donné de sujets. Par la suite, le remplacement de j_0 peut être effectué en fonction des modalités présentes dans la même classe. En pratique, il peut se faire à la proportionnelle, suivant la fréquence de ces dernières. Ceci ne change pas globalement la probabilité pour un sujet du groupe E_i d'avoir un type de réponse, chaque type correspondant à une classe.

A la fin, on peut obtenir pour chaque groupe de sujets et pour chaque variable, une signification particulière de la non-réponse et éventuellement une proposition de remplacement en fonction des autres modalités.

Remarque :

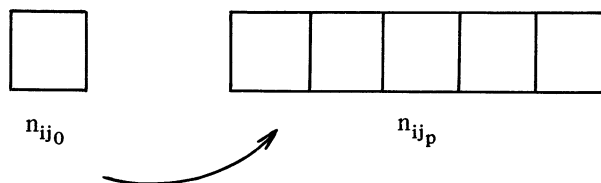
La part d'explication de chaque modalité j_p de j pour le groupe E_i dans la classe s_1 telle que $j_0 \in s_1$ est donnée par le rapport :

$$\frac{n_{ijp}}{n_{is_1}} \quad \text{avec} \quad \sum_{j_p \in s_1} n_{ijp} = n_{is_1}$$

et pour la non-réponse :

$$\frac{n_{ij_0}}{n_{is_1}}$$

La configuration des n_{ij_0} est donnée par l'ensemble des autres n_{ijp} . Si un tirage au sort peut être effectué, pour remplacer les n_{ij_0} par des différents n_{ijp} dans un rapport : $\frac{n_{ijp}}{n_{is_1} - n_{ij_0}}$ (cf. [Le 77]).



la part que représente n_{ij_0} pour le groupe E_i aura la même configuration que le reste des n_{ij_p} .

Ainsi :

$$\sum_{j_p \neq j_0} \frac{n_{ij_p}}{n_{is_1} - n_{ij_0}} \times n_{ij_0} + \sum_{j_p \neq j_0} n_{ij_p} = n_{is_1}$$

La probabilité d'interprétation de j_0 par j_p est :

$$\frac{n_{ij_p}}{n_{is_1} - n_{ij_0}}$$

2. CONDITIONS DE VALIDITE DE LA METHODE

Dans le déroulement des calculs, on est amené à :

- considérer un tableau initial à 2 lignes et r colonnes :

	j			
		j ₀	j _q	j _{r-1}
δ _i				
E _i		n _{ij₀}	n _{ij_q}	
Ē _i		n̄ _{ij₀}	n̄ _{ij₀}	

- construire la suite des tableaux à 2 lignes et s colonnes correspondant chacun à une répartition des j_q modalités de j en s classes. s est fixé par l'utilisateur en fonction de la nature des variables initiales et du nombre de leurs modalités. Comme on cherche un remplacement éventuel de la non-réponse, le nombre de classes est obligatoirement inférieur au nombre initial de modalités ;

d'où

$s < r$

En pratique pour une variable initialement en OUI – NON et non-réponse, le nombre de classes peut être 2.

Pour une variable comportant plus de 3 modalités, 3 classes ou moins peuvent être retenues.

Tableau (2,s)

δ_i \ classes	S_1	$\dots S_k \dots$	S_s
E_i	n_{i1}	n_{ik}	n_{is}
\bar{E}_i	$n_{i\bar{1}}$	$n_{i\bar{k}}$	$n_{i\bar{s}}$

Avec

$$n_{ik} = \sum_{j_q \in S_k} n_{ij_q} \quad \text{et} \quad n_{i\bar{k}} = \sum_{j_q \in S_k} n_{ij_q}$$

S_k étant la $k^{\text{ième}}$ classe

– effectuer sur chaque tableau (2,s) le test du χ^2 .

Tout d'abord, pour un tableau en modalités, la condition habituelle d'utilisation du χ^2 est que chaque effectif observé soit supérieur ou égal à 5 :

$$\forall_i, \forall_k : n_{ik} \geq 5$$

D'autre part, dans la recherche du tableau, parmi ceux de dimensions (2, s), ayant le χ^2 maximum, l'utilisateur peut se fixer un seuil de signification du test : par exemple, il peut admettre qu'il y a liaison entre la variable δ_i et la variable j à la seule condition que α soit inférieur ou égal à 5 %. Dans le cas où α est supérieur à 5 %, la méthode n'est pas applicable et un autre traitement peut être envisagé.

Au total, les conditions à vérifier sont trois :

$$\begin{aligned} \forall_i ; \forall_q : n_{ij_q} &\geq 5 \\ s &< r \\ \alpha &\leq 5 \% \end{aligned}$$

3. SIGNIFICATION DU CRITERE OPTIMISE

Le test du χ^2 est un test d'indépendance entre deux variables qualitatives I et J. Il est donné pratiquement, après détermination des effectifs calculés :

$$c_{ij} = \frac{n_{j.} \times n_{i.}}{n}, \text{ par la formule}$$

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - c_{ij})^2}{c_{ij}}$$

Pour la signification de ce test, on cherche le risque α associé par lecture de la table de χ^2 , pour le nombre de degrés de liberté : d.d.l. = $(k - 1) (s - 1)$

Dans le cas présent où $k=2$ et où s sera souvent fixé égal à 2, on a le tableau :

Variable j	classe S_1	classe S_2	Total de lignes
δ_i			
E_i	n_{i1}	n_{i2}	b_i
\bar{E}_i	$n_{\bar{i}1}$	$n_{\bar{i}2}$	$b_{\bar{i}}$
Total des colonnes	n_1	n_2	n

Et par rapport au tableau initial T_0 des individus avec $t_{hj_q} = 0$ ou 1 pour la ligne h et la colonne j_q , on a :

$$n_{i1} = \sum_{h \in E_i} \sum_{j_q \in S_1} t_{hj_q} \quad n_{i2} = \sum_{h \in E_i} \sum_{j_q \in S_2} t_{hj_q}$$

$$n_{\bar{i}1} = \sum_{h \in \bar{E}_i} \sum_{j_q \in S_1} t_{hj_q} \quad n_{\bar{i}2} = \sum_{h \in \bar{E}_i} \sum_{j_q \in S_2} t_{hj_q}$$

Si on pose $f_1 = \frac{n_{i1}}{n_1}$ et $f_2 = \frac{n_{\bar{i}2}}{n_2}$, la méthode revient à maximiser les expressions suivantes, pour lesquelles on peut se référer à [Fa 78] pour le détail des calculs. très classiques par ailleurs.

$$S = \frac{n}{b_i b_{\bar{i}}} [n n_1 f_1^2 + n n_2 f_2^2 - b_i^2] \quad (1)$$

$$S = \frac{n^2}{b_i b_{\bar{i}}} [n_1 f_1^2 + n_2 f_2^2 - \frac{b_i^2}{n}] \quad (2)$$

$$S = \frac{n}{b_i b_{\bar{i}}} n_1 n_2 (f_1 - f_2)^2 \quad (3)$$

Comme b_i , $b_{\bar{i}}$ et n sont indépendants de la subdivision effectuée, maximiser S revient à maximiser $E = \boxed{n_1 f_1^2 + n_2 f_2^2}$.

La somme des variances inter-classes est égale à :

$$n_1 (f_1 - a)^2 + n_2 (f_2 - a)^2 = n_1 f_1^2 + n_2 f_2^2 - n a^2, \text{ si } a = \frac{b_{\bar{i}}}{n}$$

Elle est maximale quand E est maximum

Et la somme des variances intra-classes est égale à :

$$n_1 f_1 (1 - f_1) + n_2 f_2 (1 - f_2) = na^2 - (n_1 f_1^2 + n_2 f_2^2)$$

Elle est minimale quand E est maximum.

Dans le cas où $s = k$: le détail des calculs se trouve en annexe.

L'expression à maximiser est : $E = n_1 f_1^2 + \dots + n_k f_k^2$

Remarque 1

On pourrait aussi considérer le cas où le nombre de classes s n'est pas fixé. Quand s est fixé, on compare les valeurs prises par $\chi^2 (s - 1)$ pour les différentes subdivisions des modalités de la variable j alors que dans l'autre cas, on compare les $\chi^2 (s - 1)$ avec s variable, en passant par l'intermédiaire de la probabilité correspondante, fournie par les tables de distribution de la loi du χ^2 . C'est une possibilité intéressante de la méthode du χ^2 que d'effectuer des subdivisions en nombres différents d'états.

Remarque 2

Ainsi dans la méthode de classification appliquée pour la détermination des K groupes, on a défini l'homogénéité des groupes par la somme des variances intra-classes.

L'algorithme de classification fait décroître ce critère et notre méthode de regroupement des modalités augmente cette décroissance. On peut remarquer que ces deux méthodes font décroître le même critère, on peut ainsi les utiliser alternativement. Par exemple, avec les nuées dynamiques on obtient un optimum local ; le regroupement de certaines modalités permet de dégager un autre optimum en simplifiant l'ensemble des modalités du questionnaire, ce qui apporte de plus un intérêt, lorsque celui-ci est lourd au départ.

4. DEROULEMENT DES CALCULS

Commentaire de l'organigramme suivant.

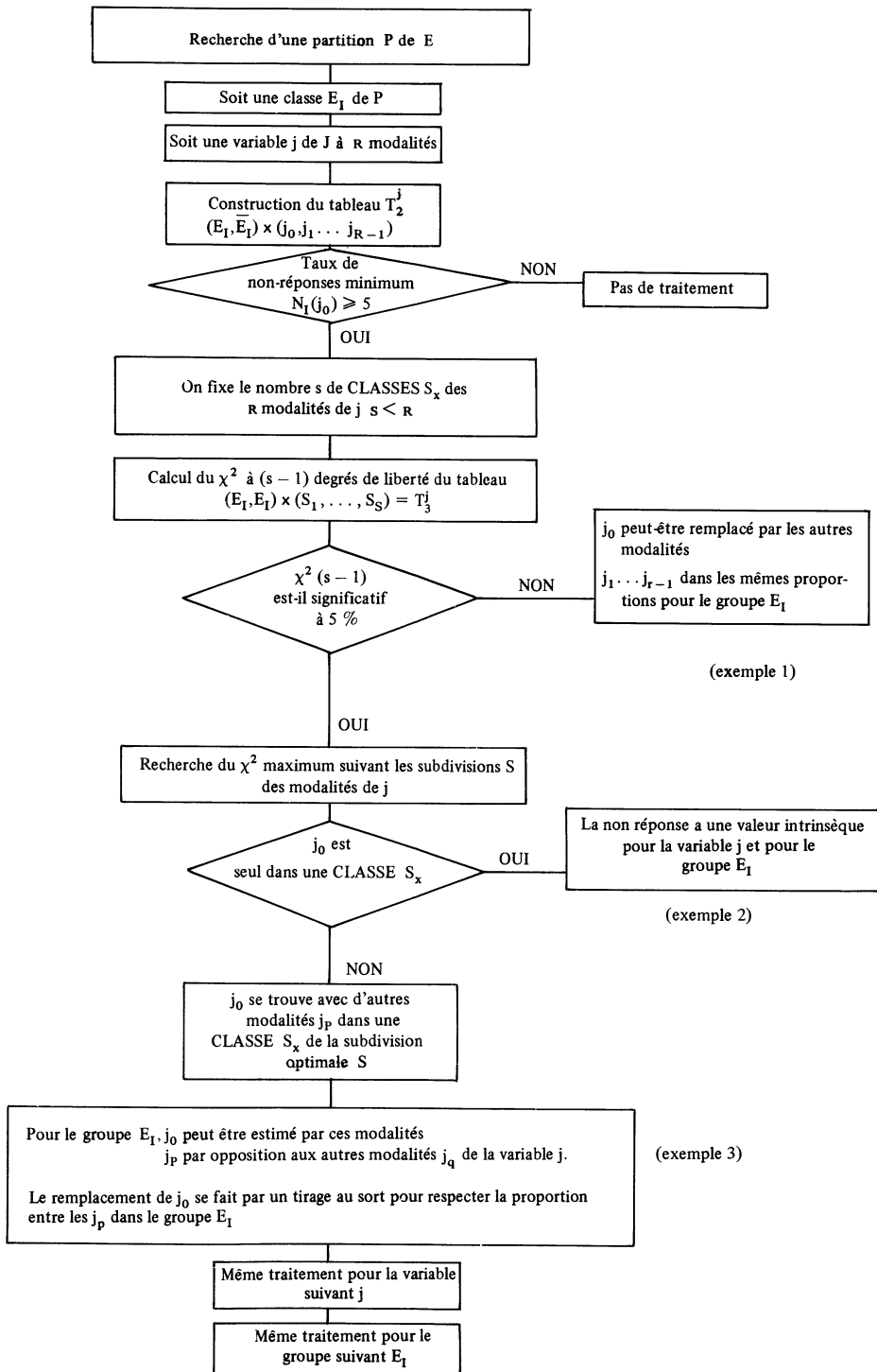
3 étapes peuvent être schématisées :

4.1 – En considérant tous les sujets, on effectue une classification, du type “nuées dynamiques” par exemple. On obtient ainsi une partition des sujets en K groupes E_i , les non-réponses sont considérées comme des modalités à part entière pour chaque variable.

On part d'une juxtaposition de tableaux $T_0^j, j = 1, v$ et d'une juxtaposition de tableaux $T_1^j, j = 1, v$.

Pour plus de commodité, on a en principe réservé la modalité j_0 pour les non-réponses, quand elles sont présentes dans l'ensemble E pour la variable j .

ORGANIGRAMME



4.2 – En étudiant chaque groupe E_i séparément, on recherche pour chaque variable j le découpage optimal de l'ensemble de ses modalités j_q au sens de la distance du χ^2 . Pour un nombre de classes fixé à 2, le test effectué consiste à maximiser le χ^2 de contingence du tableau T_3^j avec $s = 2$.

Pour appliquer cette méthode, la seule condition à vérifier est que $n_{i0} \geq 5$ donc en particulier pour le nombre de non-réponses, il faut qu'il soit supérieur à 5 pour une variable et pour un groupe.

4.3 – L'analyse des résultats de la méthode exposée conduit à l'interprétation de j_0

Premier cas

Le test du χ^2 correspondant au tableau initial n'est pas significatif au seuil fixé⁺. La variable ne représente pas E_i mais s'il est indispensable de supprimer la modalité j_0 pour cette variable qualitative, on peut proposer une représentation de j_0 proportionnelle à la répartition des autres modalités dans chaque groupe, réalisée par un tirage au hasard des sujets. Ceci est à rapprocher d'une pratique assez répandue, consistant à remplacer les données manquantes par le "centre de gravité" du nuage de points.

Exemple 1 : A partir du fichier initial des suicidants, un recodage s'est révélé nécessaire, après une première classification. La fréquence des non-réponses étant trop faible pour certaines variables pour obtenir un χ^2 significatif, les non-réponses ont été réparties sur les autres modalités. Par contre, pour d'autres variables, la fréquence était telle que la modalité de non-réponse a été conservée.

Deuxième cas

Le test du χ^2 est significatif au seuil fixé a priori (la variable j est liée à la classe E_i).

Le découpage dichotomique de j peut se présenter sous différentes formes :
A) j_0 est seul contre toutes les autres modalités. On ne peut donc que garder j_0 qui a une valeur intrinsèque pour le groupe E_i .

Exemple 2 : D'après une classification sur les variables socio-familiales des jeunes suicidants, un groupe de sujets apparaît comme étant plutôt constitué par des filles de 15 à 16 ans, habitant à la campagne, dont le père est ouvrier ou agriculteur, et ayant de nombreux frères et soeurs ; elles vivent au foyer familial et ont subi une réaction d'agressivité de leur entourage après leur tentative. Les familles n'ont pas répondu aux questions concernant l'éducation et les difficultés scolaires éventuelles des adolescentes. Or, d'après les tests du χ^2 effectués sur ces variables pour ce groupe, aucune autre modalité ne s'est révélée proche de la non-réponse. Elle doit être conservée et montre un refus de répondre de la part de la famille dans ce cas.

+ Le seuil est souvent pris dans la pratique égal à 5 %.

Exemple 2 bis : Dans une classification des sujets d'après les variables psychologiques et relationnelles, un groupe se distingue, caractérisé essentiellement par des non-réponses. Aucune proximité avec d'autres modalités n'apparaît. La non-réponse ne peut être remplacée pour ces variables.

B) j_0 est accompagné d'autres modalités j_p contre j_q . Donc les sujets du groupe E_i qui ont j_0 pour la variable j ont un comportement proche de ceux qui ont les modalités j_p , s'opposant à j_q . On peut donc estimer, pour E_i , j_0 par les j_p , estimation qui est proportionnelle à la répartition des sujets du groupe E_i sur les modalités j_p , ce qui donne une probabilité d'interprétation et d'affectation de j_0 par les autres j_p .

Exemple 3 : Une partition des suicidants sur la situation socio-familiale montre un petit groupe où il y a certaines non-réponses et quelques caractéristiques significatives précisées, qui sont retrouvées pour un autre groupe. Les tests du χ^2 montrent une proximité des non-réponses avec les modalités significatives de l'autre groupe. Ainsi l'absence de réponse pour la situation familiale indique en fait la séparation du milieu familial pour ce groupe de sujets.

Il en a été de même pour un deuxième groupe de sujets dont les non-réponses ont pu être estimées "proches" des réponses d'un autre groupe.

Contrôle de la méthode

Après le remplacement des modalités j_0 des variables suivant les différents groupes, on peut refaire une classification de tous les individus avec leurs nouvelles modalités et voir si la nouvelle partition obtenue diffère beaucoup de la première.

5. CONCLUSION

La fréquence importante des non-réponses dans ce fichier de Santé Publique, atteignant 10 % des effectifs pour certaines variables, justifiait un traitement particulier, dont les résultats, de plus, se révèlent intéressants.

En effet l'analyse des données montre que les non-réponses sont plutôt spécifiques des types de suicidants les plus graves. Si on avait écarté les sujets dont certains renseignements manquaient et qui étaient assez nombreux, on aurait probablement sous-évalué l'importance des tendances suicidaires.

On voit donc par cet exemple l'intérêt du non regroupement systématique des non-réponses et la cohérence entre la méthode proposée de partition de modalités et la méthode de classification des individus.

Elle peut être généralisée au traitement de modalités autres que la non-réponse dans un but de regrouper des modalités voisines.

ANNEXE

cas où $s = k$: calcul du χ^2

$\delta_i \backslash S$	$S_1 \dots \dots \dots S_k$	Total
E_i	$n_1(1 - f_1) \dots \dots \dots n_k(1 - f_k)$	b_i
\bar{E}_i	$n_1 f_1 \dots \dots \dots n_k f_k$	b_i^-
Total	$n_1 \dots \dots \dots n_k$	n

Le χ^2 total a pour expression :

$$S = \sum_{n=1}^k \left[\frac{\left(n_m(1 - f_m) - \frac{n_m b_i}{n} \right)^2}{\frac{n_m b_i}{n}} + \frac{\left(n_m f_m - \frac{n_m b_i^-}{n} \right)^2}{\frac{n_m b_i^-}{n}} \right]$$

$$S = \frac{n}{b_i b_i^-} \sum_{m=1}^k \left[n_m b_i^- \left(1 - f_m - \frac{b_i}{n} \right)^2 + n_m b_i \left(f_m - \frac{b_i^-}{n} \right)^2 \right]$$

soit, en tenant compte de ce que $b_i + b_i^- = n$:

$$S = \frac{n}{b_i b_i^-} \sum_{m=1}^k n_m \left(b_i^- + n f_m^2 + \frac{b_i b_i^-}{n} - 2 f_m b_i^- - 2 \frac{b_i b_i^-}{n} \right)$$

$$S = \frac{n^2}{b_i b_i^-} \left(b_i^- + \sum_{m=1}^k n_m f_m^2 - \frac{b_i b_i^-}{n} - 2 \frac{b_i^2}{n} \right)$$

soit puisque $\frac{b_i^- - b_i b_i^-}{n} = \frac{b_i^-}{n}$:

$$S = \frac{n^2}{b_i b_i^-} \left(\sum_{m=1}^k n_m f_m^2 - \frac{b_i^2}{n} \right)$$

BIBLIOGRAPHIE

- [Af 66] AFIFI A. et ELASHOFF R.M. — “Missing observations in Multivariate Statistics I review of the litterature”. *J.A.S.A.*, 1966, 61, pp. 595-604.
- [Be 73] BENZECRI J.P. — “L’analyse des données”, Tome 2, DUNOD, 1973.
- [Ch 70] CHRISTOFFERSON A. — “The One-Component Model with Incomplete Data” Dissertation for the Degree of Doctor of Philosophy, Faculty of Social Sciences of the University of Uppsala, 1970.
- [Da 76a] DAVIDSON F., CHOQUET M., et FACY F. — “La notion de risque dans le domaine du suicide chez l’adolescent”. *Revue épidémiologique et Santé Publique*, 1976, 24, pp. 283-300.
- [Da 76b] DAVIDSON F., et CHOQUET M. — “Etude épidémiologique du suicide de l’adolescent : comparaison entre suicidants primaires et suicidants récidivistes”. *Revue épidémiologique et santé publique*, 1976, 24, pp 11-26
- [Di 72] DIDAY E. — “Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes”. Thèse d’Etat, Paris VI, 1972.
- [Di 75] DIDAY E. — “Classification automatique séquentielle sur grands tableaux”. RAIRO-B1, pp. 29-61.
- [Di 76] DIDAY E., OK Y., JACOTTET, TOMASSONNE. — “Lissage typologique”. Rapport IRIA-LABORIA n° 164.
- [Es 76] ESCOFIER B., et LEROUX B. — “Influence d’un élément sur les facteurs en analyse des correspondances”. *Les cahiers de l’analyse des données*, n° 3, 1976, pp. 297-318.
- [Fa 78] FACY-MARCHAL F. — “Classification sur tableaux juxtaposés. Application à l’étude psychosociologique des adolescents suicidants”. Thèse 3^e cycle, Paris VI. 1978.
- [Ha 75] HARTIGAN J.A. — “Clustering Algorithms”. N.Y. Wiley 75.
- [Ho 68] HOCKING R.R. et SMITH W.B. — “Estimation of Parameters in the Multivariate Normal Distribution with Missing observations”. *J.A.S.A.*, 1968, 63, pp. 159-173.
- [Le 74] LECHEVALLIER Y. — “Optimisation de quelques critères en classification automatique et application à l’étude des modifications des protéines sériques en pathologie clinique”. Thèse 3^e cycle, Paris VI, 1974.
- [Le 77] LEBART L., MORINEAU A., TABARD N. — “Techniques de la Description Statistique, Méthodes et logiciels pour l’analyse des grands tableaux”. Dunod 1977.
- [No 75] NORA-CHOUTEAU C. — “Une méthode de reconstitution et d’analyse de données incomplètes”. Thèse d’Etat, Paris, VI, 1975.
- [Ok 75] OK-SAKUN Y.— “Analyse factorielle typologique et lissage typologique”. Thèse 3^e cycle, Paris VI, 1975.
- [Wi 32] WILKS. — “Moments and distributions of estimates of population parameters from fragmentary samples”. *Annals of Mathematical Statistics*, 1932, Vol. 3.