

ANNALES DE L'I. H. P., SECTION B

DOMINIQUE COLLOMBIER

Un algorithme d'aide à l'analyse des tables de contingence et des plans d'expériences incomplets

Annales de l'I. H. P., section B, tome 14, n° 2 (1978), p. 215-231

http://www.numdam.org/item?id=AIHPB_1978__14_2_215_0

© Gauthier-Villars, 1978, tous droits réservés.

L'accès aux archives de la revue « Annales de l'I. H. P., section B » (<http://www.elsevier.com/locate/anihpb>) 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/>

Un algorithme d'aide à l'analyse des tables de contingence et des plans d'expériences incomplets

par

Dominique COLLOMBIER

Laboratoire de Statistique,
Université Paul-Sabatier, Toulouse, ERA-C. N. R. S. 591

RÉSUMÉ. — Dans cet article nous décrivons deux types de formulation des hypothèses d'analyse d'une table de contingence tronquée ou d'un plan d'expérience incomplet au moyen de sous-espaces vectoriels de l'espace des fonctions numériques définies sur l'ensemble des cellules de la table ou du plan. Puis nous proposons un algorithme de construction d'une base et de calcul de la dimension de ces sous-espaces.

SUMMARY. — In this article we describe two ways of formulating hypotheses in the analysis of a truncated contingency table or an incomplete experimental design by means of subspaces in the vector space of the numerical functions defined over the cells of the table or the design. Then we propose an algorithm for constructing a basis and for computing the dimension of these subspaces.

INTRODUCTION

La formulation d'une hypothèse dans l'analyse d'un plan d'expérience (par un modèle linéaire) ou d'une table de contingence (par un modèle log-linéaire) suppose fixée une décomposition en somme de l'espace vec-

toriel des fonctions numériques définies sur l'ensemble des cellules de ce plan ou de cette table.

Or, plusieurs décompositions sont pratiquement envisageables, et du choix de l'une d'entre elles dépend l'aisance avec laquelle certains problèmes sont résolus. Ainsi en est-il de la détermination d'une base et du calcul de la dimension du sous-espace associé à une hypothèse dans l'analyse d'un plan incomplet ou d'une table tronquée.

A ce jour, les solutions apportées sont partielles ou de peu d'intérêt pratique [4] [9]. Il nous semble que cela est dû pour une bonne part à la façon dont sont formulées les hypothèses d'analyse. Une formulation, étudiée dans le paragraphe 1, connue mais peu employée explicitement, permet de donner au problème envisagé une solution complète et pratique sous forme d'un algorithme décrit dans le paragraphe 2 de cet article.

1. FORMULATION DES HYPOTHÈSES

Notons E l'ensemble des cellules e d'un plan d'expérience ou d'une table de contingence. C'est un sous-ensemble du produit cartésien $E_1 = \prod \{ E_i \mid i \in I \}$, d'une famille finie I d'ensembles finis E_i . Chaque E_i représente un facteur ou l'un des caractères qualitatifs observés. Tout élément e_i de E_i représente une modalité d'un facteur ou d'un caractère. Si $E = E_1$ on dit que le plan est complet ou que la table est complète; sinon on dit que le plan est incomplet ou que la table est tronquée (ou incomplète).

Pour analyser un plan ou une table on fait l'hypothèse qu'en chacune des cellules figure la valeur observée d'une variable aléatoire d'espérance mathématique $\mu(e)$ (fréquemment appelée en Analyse de Variance : effet du traitement e).

Soit un sous-ensemble J de la famille I des facteurs ou des caractères et \bar{J} son complémentaire dans I . Toute cellule $e \in E \subset E_1$ peut être considérée comme un couple d'éléments des produits cartésiens des ensembles E_i pour i parcourant J d'une part et \bar{J} d'autre part :

$$e = (e_J, e_{\bar{J}}) \text{ avec } \begin{cases} e_J \in E_J = \prod \{ E_i \mid i \in J \} \\ e_{\bar{J}} \in E_{\bar{J}} = \prod \{ E_i \mid i \in \bar{J} \} \end{cases}$$

Nous appellerons désormais :

1) Tranche (ou section) de E , tout sous-ensemble de E défini, pour e_J fixé, par :

$$E_J(e_J) = \{ (e_J, e_{\bar{J}}) \mid e_{\bar{J}} \in E_{\bar{J}} : (e_J, e_{\bar{J}}) \in E \}, \quad J \neq \phi \text{ et } J \neq I;$$

Par convention nous appellerons tranches le singleton $\{e\} \in E$ et l'ensemble E ; nous les noterons respectivement $E_\phi(e)$ et $E_I(e_\phi)$.

2) Indicatrice de $E_J(e_j)$, la fonction numérique γ^{e_j} définie sur E égale à 1 sur $E_J(e_j)$ et nulle ailleurs. Pour simplifier nous noterons γ^ϕ l'indicatrice de E .

3) J-contraste, toute fonction numérique définie sur E constante égale à $v_j(e_j)$ sur $E_J(e_j)$ et telle que la fonction v_j de $E'_j = \{e_j \mid e_j \in E_j : E_j(e_j) \neq \phi\}$ dans \mathbb{R} soit de moyenne nulle sur toute tranche $E'_{J \setminus \{i\}}(e_{i\phi})$ de E'_j , $e_i \in E_i$ et $i \in J \neq \phi$.

Premier type de formulation [1] [2] [14]

En Analyse de Variance et dans la plupart des cas pour l'analyse des tables de contingence, on pose :

$$R^E = \Sigma \{ \Omega_J \mid J \in \mathcal{P}(I) \},$$

avec :

Ω_ϕ sous-espace de \mathbb{R}^E des fonctions constantes sur E ,

Ω_J sous-espace de \mathbb{R}^E des J-contrastes sur E pour $J \neq \phi$.

Si le plan (ou la table) est complet(e), cette somme est directe et ses composantes sont orthogonales pour le produit scalaire canonique de \mathbb{R}^E .

Lorsqu'on analyse un plan (modèle linéaire) on suppose que le vecteur des effets principaux, $\mu = [\mu(e) \mid e \in E]$, appartient à un sous-espace de \mathbb{R}^E du type $\Omega_{\mathcal{H}} = \Sigma \{ \Omega_J \mid J \in \mathcal{H} \}$ où \mathcal{H} est un ensemble non vide de parties de I . Dans les modèles log-linéaires d'analyse des tables c'est le vecteur $\lambda = [\lambda(e) = \text{Log} [\mu(e)] \mid e \in E]$ qui est supposé appartenir à $\Omega_{\mathcal{H}}$.

On appelle ensemble hiérarchique de parties de I une famille non vide \mathcal{H} de parties de I telle que :

$$(J \in \mathcal{H} \text{ et } J' \subset J) \Rightarrow (J' \in \mathcal{H})$$

On appelle sous-ensemble générateur par inclusion de \mathcal{H} le sous-ensemble \mathcal{H}^g de \mathcal{H} tel que

$$(J \in \mathcal{H}^g \text{ et } J \subset J') \Rightarrow (J' \notin \mathcal{H}).$$

Dans le cas d'ensembles hiérarchiques, on peut énoncer les résultats suivants.

PROPOSITION 1. — Si \mathcal{H} est un ensemble hiérarchique de parties de I , $\Omega_{\mathcal{H}}$ est engendré par l'ensemble d'indicatrices

$$\cup \{ \{ \gamma^{e_j} \mid e_j \in E_j : E_j(e_j) \neq \phi \} \mid J \in \mathcal{H}^g \}.$$

Preuve. — E'_j est le sous-ensemble de E_j des éléments e_j tels que $E_j(e_j) \neq \phi$. Entre l'espace $\mathbb{R}^{E'_j}$ des fonctions numériques définies sur E'_j et le sous-espace \mathcal{V}_j de \mathbb{R}^E engendré par la famille d'indicatrices $\{\gamma^{e_j} | e_j \in E'_j\}$ il existe un isomorphisme naturel \mathcal{I} . Tout vecteur de $\mathbb{R}^{E'_j}$ de coordonnées $v_j(e_j)$ dans la base canonique a pour image par \mathcal{I} la fonction de \mathbb{R}^E constante égale à $v_j(e_j)$ sur la tranche $E_j(e_j)$ pour e_j parcourant E'_j . Or

$$\mathbb{R}^{E'_j} = \Sigma \{ \Omega'_K | K \in \mathcal{P}(J) \}$$

avec

$$\begin{cases} \Omega'_\phi & \text{sous-espace de } \mathbb{R}^{E'_j} \text{ des fonctions constantes sur } E'_j, \\ \Omega'_K & \text{sous-espace de } \mathbb{R}^{E'_j} \text{ des } K\text{-contrastes sur } E'_j \text{ pour } K \neq \phi \end{cases}$$

et l'image par l'isomorphisme \mathcal{I} de tout sous-espace Ω'_K de $\mathbb{R}^{E'_j}$ est le sous-espace Ω_K de \mathbb{R}^E . Donc $\mathcal{V}_j = \Sigma \{ \Omega_K | K \in \mathcal{P}(J) \} = \Omega_{\mathcal{P}(J)}$.

Mais puisque \mathcal{H} est hiérarchique $\Omega_{\mathcal{H}} = \Sigma \{ \Omega_{\mathcal{P}(J)} | J \in \mathcal{H}^g \}$.

Il s'ensuit que $\Omega_{\mathcal{H}} = \Sigma \{ \mathcal{V}_j | J \in \mathcal{H}^g \}$ sous-espace de \mathbb{R}^E engendré par la famille d'indicatrices $\cup \{ \{ \gamma^{e_j} | e_j \in E'_j \} | J \in \mathcal{H}^g \}$ \square

COROLLAIRE 1. — Soit \mathcal{V}_E le sous-espace de \mathbb{R}^{E_1} des fonctions numériques définies sur E_1 nulles en tout point de $E_1 \setminus E$. Considérons un ensemble hiérarchique \mathcal{H} de parties de I . A l'isomorphisme canonique $\mathcal{V}_E \approx \mathbb{R}^E$ près, le sous-espace $\Omega_{\mathcal{H}}$ de \mathbb{R}^E est image par la projection orthogonale (pour le produit scalaire usuel) de \mathbb{R}^{E_1} sur \mathcal{V}_E du sous-espace $\Omega_{\mathcal{H}}$ de \mathbb{R}^{E_1} .

Ce corollaire est d'un grand intérêt pratique. On sait en effet aisément construire une base d'un sous-espace $\Omega_{\mathcal{H}}$ de \mathbb{R}^{E_1} par produit tensoriel [10] [11] [13]. Par restriction à E on en déduit un système générateur du sous-espace $\Omega_{\mathcal{H}}$ de \mathbb{R}^E .

Illustrons-le par un exemple. Soient $I = \{ 1, 2 \}$, $E_1 = \{ a, b \} \times \{ a, b \}$ et $E = E_1 \setminus \{ (b, b) \}$. \mathbb{R}^{E_1} a pour base :

	a	b									
a	1	1		1	1		1	-1		-1	1
b	1	1		-1	-1		1	-1		-1	-1
$(\text{sur } E_1 :)\gamma^\phi$	$\{ 1 \}$ -contraste			$\{ 2 \}$ -contraste			$\{ 1, 2 \}$ -contraste				

Les images de ces vecteurs par la projection orthogonale de \mathbb{R}^{E_1} sur \mathcal{V}_E sont respectivement :

	1	1		1	1		-1	1
	1	1		-1	-1		1	-1
$(\text{sur } E :)\gamma^\phi$	$\{ 1 \}$ -contraste : $v_{\{1\}}$			$\{ 2 \}$ -contraste : $v_{\{2\}}$			f	

f n'est pas un $\{1, 2\}$ -contraste sur E . Si $\mathcal{H} = \{\phi, \{1\}, \{2\}, \{1, 2\}\}$, hiérarchique, le sous-espace $\Omega_{\mathcal{H}}$ de \mathbb{R}^E est engendré par γ^ϕ , $v_{\{1\}}$ et $v_{\{2\}}$; il coïncide avec l'espace engendré par γ^ϕ , $v_{\{1\}}$, $v_{\{2\}}$ et f puisque

$$f = \gamma^\phi - v_{\{1\}} - v_{\{2\}}.$$

Par contre, si $\mathcal{H} = \{\phi, \{1\}, \{1, 2\}\}$, non hiérarchique, $\Omega_{\mathcal{H}}$ sous-espace de \mathbb{R}^E engendré par γ^ϕ et $v_{\{1\}}$ ne coïncide pas avec l'espace engendré par γ^ϕ , $v_{\{1\}}$ et f en vertu de l'unicité de la décomposition précédente de f .

Second type de formulation

Lorsque pour des raisons pratiques on se voit obligé de passer par un programme de régression linéaire multiple pour analyser un plan d'expérience (par exemple pour des plans incomplets ou/et déséquilibrés), on prend souvent pour variables explicatives γ^ϕ et les indicatrices des tranches $E_J(e_J)$ non vides pour $e_J \in \Pi \{E_i \setminus \{e_i^0\} \mid i \in J\}$, $J \in \mathcal{H}$ ($J \neq \phi$), avec

$$e^0 = (e_J^0, e_J^0) = (e_i^0 \mid i \in I)$$

élément fixé de E_i .

Cette pratique suppose implicitement : $\mathbb{R}^E = \Sigma \{ \Theta_J \mid J \in \mathcal{P}(I) \}$, avec $\left\{ \begin{array}{l} \Theta_\phi \text{ sous-espace de } \mathbb{R}^E \text{ des fonctions constantes sur } E \text{ (} \Theta_\phi = \Omega_\phi \text{)} \\ \Theta_J, J \neq \phi, \text{ sous-espace de } \mathbb{R}^E \text{ des fonctions nulles en tout point } (e_J, e_J) \in E \text{ tel que, pour au moins un } i \text{ la } i\text{ème composante } e_i \text{ de } e_J = (e_i \mid i \in J) \text{ soit égale à } e_i^0, \text{ autrement dit sous-espace engendré par les indicatrices } \gamma^{e_J}, e_J \in \Pi \{ E_i \setminus \{e_i^0\} \mid i \in J \} : E_J(e_J) \neq \phi \end{array} \right.$

et :

$$\mu \in \Theta_{\mathcal{H}} = \Sigma \{ \Theta_J \mid J \in \mathcal{H} \}$$

Si le plan est complet ces sommes sont directes mais les composantes ne sont pas orthogonales.

PROPOSITION 2. — *Si \mathcal{H} est un ensemble hiérarchique de parties de I , alors $\Theta_{\mathcal{H}} = \Omega_{\mathcal{H}}$.*

Compte tenu de la proposition 1 il suffit de prouver que $\Theta_{\mathcal{H}}$ est engendré par la famille d'indicatrices $\cup \{ \{ \gamma^{e_J} \mid e_J \in E_J : E_J(e_J) \} \mid J \in \mathcal{H}^g \}$ pour démontrer cette proposition. Cela s'obtient par une démonstration analogue à celle de la proposition 1.

Remarque. — C'est cette proposition qui justifie la pratique décrite plus haut. Pour tester l'hypothèse de nullité de l'effet d'un facteur ou d'un effet d'interaction au moyen d'un programme de régression linéaire multiple, on choisit deux ensembles hiérarchiques de parties de I : \mathcal{H}_1 et \mathcal{H}_2 , tels que $\mathcal{H}_2 \setminus \mathcal{H}_1 = \{ J \}$, facteur sur lequel porte l'hypothèse d'absence

d'effet ou groupe de facteurs sur lequel porte l'hypothèse d'absence d'interactions. On effectue deux analyses successives, la première sous l'hypothèse $\mu \in \Theta_{\mathcal{H}_1}$, la seconde sous l'hypothèse $\mu \in \Theta_{\mathcal{H}_2}$. Puis on calcule la valeur observée de la statistique de test à partir des résultats de ces deux analyses.

Ce parti pris dans le choix des variables explicatives (à savoir : γ^ϕ et les indicatrices des tranches $E_J(e_J) \neq \phi$, $e_J \in \Pi \{ E_i \setminus \{ e_i^0 \} \mid i \in J \}$, $J \in \mathcal{H} : J \neq \phi$) a pour but d'éliminer toute dépendance vectorielle entre celles-ci. Cependant ce but n'est pas atteint dans certains cas dits de confusion d'effets ; certaines des indicatrices retenues sont alors dépendantes. L'algorithme décrit dans le paragraphe suivant a précisément pour objet d'extraire une base d'une famille génératrice de ce genre.

Ce type de décomposition de \mathbb{R}^E est également utilisé par quelques auteurs pour l'analyse des tables de contingence complètes.

Certains cherchent par là uniquement à simplifier et à réduire le plus possible le système des contraintes de non-interaction dans les modèles qui sont dits hiérarchiques [15]. Soulignons que cette notion n'a rien à voir avec celle de « plan » hiérarchique : on désigne simplement ici par « modèle » hiérarchique, un modèle où les seules hypothèses qui interviennent se formulent λ (ou μ) $\in \Omega_{\mathcal{H}}$ ou λ (ou μ) $\in \Theta_{\mathcal{H}}$ avec \mathcal{H} ensemble hiérarchique de parties de I. La simplification intervient parce que si $E = E_1$ l'orthogonal de $\Theta_{\mathcal{H}}$ a pour base l'ensemble des J-contrastes, pour J parcourant \mathcal{H} , qui

1) sont nuls en tout point de E autre que les $2^{|J|}$ sommets de chacun des hypercubes s'appuyant sur les sommets diagonalement opposés $e^0 = (e_i^0 \mid i \in I)$ et $e = (e_i \mid i \in I)$ avec $e_i \neq e_i^0$ pour $i \in J$ et $e_i = e_i^0$ pour $i \notin J$,

2) prennent des valeurs opposées sur les 2 moitiés enchevêtrées de ces hypercubes : 1 sur l'une et - 1 sur l'autre [5, ch. 2].

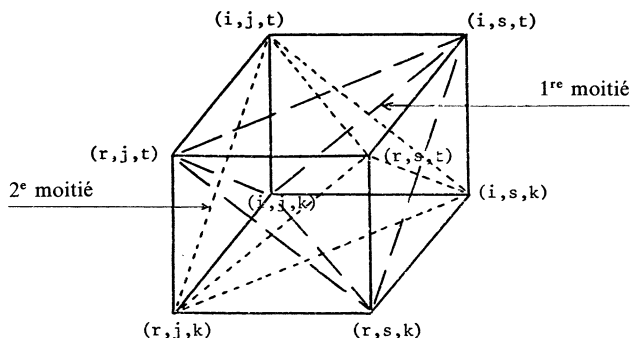


Fig. 1.

Ainsi, par exemple, dans une table complète à $r \times s \times t$ cellules repérées par le triplet (i, j, k) avec $i \in [1, r]$, $j \in [1, s]$ et $k \in [1, t]$, si $e^0 = (r, s, t)$, l'hypothèse de non-interaction d'ordre 2 prend la forme proposée par S. N. Roy et M. A. Kastenbaum [16] à savoir

$$(\mu_{ijk} \cdot \mu_{rsk}) / (\mu_{rjk} \cdot \mu_{isk}) = (\mu_{ijt} \mu_{rst}) / (\mu_{rit} \cdot \mu_{ist})$$

pour $i \in [1, r - 1]$, $j \in [1, s - 1]$ et $k \in [1, t - 1]$.

D'autres cherchent par cette deuxième formulation à rendre les résultats des analyses plus interprétables lorsque l'une des modalités e_i^0 de chacun des facteurs ou des caractères observés joue le rôle de témoin. Ainsi en est-il des utilisateurs des modèles logistiques [7].

Enfin, c'est ce type de formulation qui est utilisé implicitement dans le produit-programme G. L. I. M. qui assure l'analyse de données au moyen du modèle dit linéaire généralisé de J. A. Nelder et R. W. M. Wedderburn [12].

2. RECHERCHE D'UNE BASE DE $\Omega_{\mathcal{H}}$ OU DE $\Theta_{\mathcal{H}}$

Pour tout plan complet ou toute table complète, il est relativement aisé de construire par produit tensoriel une base de tout espace $\Omega_{\mathcal{H}}$ ou $\Theta_{\mathcal{H}}$ et d'en calculer la dimension [5] [11] [13]. Sauf exception, il n'en est pas de même dans le cas des plans incomplets ou des tables tronquées [4]. S. J. Haberman [9, pages 233 à 244] propose un algorithme de construction d'une base de $\Omega_{\mathcal{H}}$ lorsque \mathcal{H} est hiérarchique par extraction de vecteurs d'une famille génératrice de type fixé. Mais cet algorithme est limité au traitement des modèles hiérarchiques. De plus, pour éviter les dépassements de capacité en mémoire d'ordinateur, il requiert :

1) la détermination d'un ou plusieurs nombres premiers qui sont représentables en machine, dont le produit dépasse une borne qu'il faut calculer pour chaque exemple et qui est en général très grande.

2) plusieurs essais avec, pour le i ème essai, des calculs en arithmétique modulo le i ème nombre premier retenu.

A chaque essai on extrait un sous-ensemble de vecteurs indépendants de la famille génératrice ; les sous-ensembles de cardinal maximum sont des bases.

On conçoit aisément que cet algorithme soit d'un intérêt pratique limité. Nous proposons ci-dessous un nouvel algorithme de construction d'une

base de $\Theta_{\mathcal{H}}$, \mathcal{H} quelconque, qui ne requiert que des calculs en arithmétique ordinaire. Il est d'un emploi aisé comme le prouve la procédure Fortran qui le met en œuvre [6] et les exemples traités ci-dessous.

Nouvel algorithme

Cet algorithme permet d'extraire une base d'une famille génératrice de $\Theta_{\mathcal{H}}$, que \mathcal{H} soit hiérarchique ou non. Si \mathcal{H} est hiérarchique il assure donc également la construction d'une base $\Omega_{\mathcal{H}}$ en vertu de la proposition 2.

Deux types de familles génératrices sont utilisés :

Type 1 (valable quel que soit \mathcal{H}) :

$$\{\gamma^\phi\} \cup \{ \{ \gamma^{e_j} \mid e_j \in \Pi \{ E_i \setminus \{ e_i^0 \} \mid i \in J \} \mid e_j \in E_J : E_J(e_j) \neq \phi \} \mid J \in \mathcal{H} : J \neq \phi \}$$

Type 2 (valable pour \mathcal{H} hiérarchique) :

$$\cup \{ \{ \gamma^{e_j} \mid e_j \in E_J : E_J(e_j) \neq \phi \} \mid J \in \mathcal{H}^g \}$$

Lorsque \mathcal{H} est hiérarchique, on a donc le choix entre ces deux types de familles. Le choix de la première de ces familles résulte directement de la définition de $\Theta_{\mathcal{H}}$, celui de la deuxième de la proposition 1. Ces familles ne sont constituées que d'indicatrices. Si nous notons A la matrice dans la base canonique de \mathbb{R}^E des vecteurs de l'une ou l'autre de ces familles génératrices (vecteurs lignes), A ne contient donc que des éléments entiers égaux à 0 ou 1.

Pour en extraire une base nous procédons en deux étapes.

PREMIÈRE ÉTAPE. — Notons r le nombre d'éléments de la famille étudiée et s le cardinal de E . Par permutation des r lignes et des s colonnes de A nous transformons cette matrice en la matrice semi-triangulaire $A^{(p)}$.

$$A^{(p)} = \begin{matrix} & \begin{matrix} p & s \end{matrix} \\ \begin{matrix} r \\ \hline \end{matrix} & \begin{array}{|c|c|} \hline \text{T} & * \\ \hline * & * \\ \hline \end{array} \end{matrix} \quad \text{avec} \quad \begin{cases} p \leq r \\ p \leq s \end{cases}$$

où les $*$ sont des blocs quelconques et T est un bloc $p \times p$ triangulaire supérieur d'éléments diagonaux égaux à 1 (donc $\det \{ T \} = 1$).

Cette étape n'a pour but que de rendre l'algorithme plus performant, à moins que l'on ne soit dans les conditions de la proposition 3 ci-dessous.

DEUXIÈME ÉTAPE. — Partant de $A^{(p)}$, on construit une séquence finie de matrices selon le schéma itératif suivant :

- (1) Si $p = r$ ou $p = s$ on s'arrête, sinon on pose $j = p + 1$ et $k = p + 1$.
- (2) Considérons les blocs

$B_{11}^{(p)}$ formé par les éléments des p premières lignes et colonnes de $A^{(p)}$,
 $B_{21}^{(p)}$ par ceux de la k ième ligne et des p premières colonnes,
 $B_{12}^{(p)}$ par ceux de la j ième colonne et des p premières lignes,
 $B_{22}^{(p)}$ par le seul élément $a_{kj}^{(p)}$

		p		j	
	p	$B_{11}^{(p)}$		$B_{12}^{(p)}$	
	k	$B_{21}^{(p)}$		$B_{22}^{(p)}$	

on résout le système d'équations ('B désignant la transposée de B) :

$${}^t B_{11}^{(p)} X = \det \{ B_{11}^{(p)} \} {}^t B_{21}^{(p)} \quad (a)$$

où X est une matrice colonne à p lignes à éléments entiers puisque le i ème est le déterminant d'une matrice à éléments entiers formée en substituant $B_{21}^{(p)}$ à la i ème ligne de $B_{11}^{(p)}$.

- (3) On pose $\Delta = \det \{ B_{11}^{(p)} \} B_{22}^{(p)} - {}^t B_{12}^{(p)} X$

Si Δ n'est pas nul, le k ième vecteur ligne de $A^{(p)}$ est indépendant des p premiers vecteurs lignes. On constitue alors, si nécessaire par permutation de lignes ou de colonnes, la matrice

$$A^{(p+1)} = \begin{matrix} & & p + 1 \\ & & \begin{matrix} B_{11}^{(p)} & B_{12}^{(p)} \\ B_{21}^{(p)} & B_{22}^{(p)} \end{matrix} \\ p + 1 & \begin{matrix} B_{11}^{(p)} & B_{12}^{(p)} \\ B_{21}^{(p)} & B_{22}^{(p)} \end{matrix} & \end{matrix}$$

qui est telle que $\det \{ B_{11}^{(p+1)} \} = \Delta$. On incrémente p de 1 et on reprend en (1).

- Si Δ est nul :
 - si $k < r$ on incrémente k de 1 et on reprend en (2);
 - si $k = r$ et $j < s$ on pose $k = p + 1$, on incrémente j de 1 et on reprend en (2);
 - si $k = r$ et $j = s$ on s'arrête.

N. B. — La résolution du système (a) et l'évaluation de Δ donc du déter-

minant de $B_{11}^{(p+1)}$ sont assurées simultanément par un procédé d'élimination de Gauss où tous les calculs sont effectués *en nombres entiers*.

Au terme de l'exécution de cet algorithme p est la dimension de $\Theta_{\mathcal{H}}$ et les p premiers vecteurs lignes de $A^{(p)}$ définissent la base extraite.

Cet algorithme n'exclut pas tout dépassement de capacité; mais cet événement est beaucoup moins fréquent que lors de l'emploi de l'algorithme de S. J. Haberman avec calculs effectués en arithmétique ordinaire.

S'il advient un tel dépassement on fait en sorte qu'il soit automatiquement détecté.

A titre d'exemple reprenons le cas traité en [9, p. 237] :

1) E_1 est le produit cartésien de trois ensembles $E_1 = E_2 = E_3 = \{1, 2, 3\}$

2) $E = E_1 \setminus \Pi \{E_i \setminus \{2\} \mid i \in I\}$

3) $\mathcal{H}^g = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$: hypothèse de non interaction d'ordre 2.

Sur cet exemple l'algorithme de S. J. Haberman, avec calculs en arithmétique ordinaire, provoque un dépassement de capacité sur tout ordinateur n'acceptant que des entiers représentables en vingt chiffres binaires au plus. Pour éviter cela il faut déterminer une famille de nombres premiers représentables dont le produit soit supérieur à $(14)^3 (10)^6 (19)^{1/2}$, par exemple les 2 nombres 254.039 et 254.047. A priori l'extraction d'une base devrait donc nécessiter l'usage de l'algorithme de S. J. Haberman à deux reprises avec calculs effectués modulo 254.039 puis modulo 254.047. Cependant la famille extraite au premier essai coïncide avec la famille génératrice. Cette dernière est donc une base et un seul essai suffit dans ce cas très particulier.

Sur le même exemple, notre algorithme donne les résultats suivants :

1) si on utilise une famille génératrice du type 1 avec :

- . $e^0 = (1, 1, 1)$: alors 15 des 19 vecteurs de base sont extraits dès la première étape et les résultats intermédiaires ne dépassent pas 1 en valeur absolue dans la deuxième étape,
- . $e^0 = (3, 3, 3)$: alors 6 des vecteurs de base sont extraits lors de la première étape et les résultats intermédiaires ne dépassent pas 4 en valeur absolue dans la deuxième étape,

2) si on utilise une famille génératrice de type 2, la base est construite dès la première étape.

En fait, il existe ici un élément $e^0 = (e_i^0 \mid i \in I) \in E_b$, précisément $e^0 = (2, 2, 2)$, tel que $E_1 \setminus E = \Pi \{E_i \setminus \{e_i^0\} \mid i \in I\}$ et l'on sait que, dans ce cas particulier, la famille génératrice de type 1 correspondante est une base [5, ch. 2].

PROPOSITION 3. — Si \mathcal{H} est un ensemble hiérarchique de parties de I tel que $|\mathcal{H}^s| = 2$, si l'on considère la famille génératrice de $\Theta_{\mathcal{H}}$ (ou $\Omega_{\mathcal{H}}$) de type 2 et si, au terme de la première étape de l'algorithme précédent, la matrice $A^{(p)}$ est semi-triangulaire maximale, alors l'exécution de la deuxième étape est inutile.

Remarque. — Dans ce cas l'hypothèse ainsi formulée est du type additivité généralisée des effets pour les plans d'expérience et quasi-indépendance généralisée pour les tables de contingence (« quasi » signifiant que la table est éventuellement tronquée).

Preuve. — La famille génératrice considérée est constituée d'indicatrices de tranches qui recouvrent E . Notons \mathcal{E} l'ensemble de ces tranches; (E, \mathcal{E}) est un hypergraphe dont 'A est la matrice d'incidence [3]. Pour tout $J \in \mathcal{H}^s$ les tranches $E_J(e_j)$, $e_j \in E_j$, sont disjointes; tout sous-ensemble d'entre elles est donc un couplage de l'hypergraphe.

Soit $\mathcal{H}^s = \{J_0, J_1\}$. Considérons une composante connexe de (E, \mathcal{E}) ; il y a dépendance vectorielle entre les indicatrices des tranches de \mathcal{E} qui sont contenues dans cette composante. En effet, tout sommet appartient à deux et seulement deux tranches, l'une du type $E_{J_0}(e_{j_0})$ l'autre du type $E_{J_1}(e_{j_1})$ et la combinaison linéaire

$$\Sigma \{ \gamma^{e_{j_0}} | e_{j_0} \in E_{J_0} : E_{J_0}(e_{j_0}) \neq \phi \} - \Sigma \{ \gamma^{e_{j_1}} | e_{j_1} \in E_{J_1} : E_{J_1}(e_{j_1}) \neq \phi \}$$

est donc nulle.

Nous allons montrer qu'il suffit d'éliminer une quelconque et une seule tranche par composante connexe pour qu'il y ait indépendance vectorielle des indicatrices. De plus, nous montrerons que la matrice d'incidence de l'hypergraphe partiel (E, \mathcal{E}_1) de (E, \mathcal{E}) où \mathcal{E}_1 est le sous-ensemble des tranches non éliminées est de la forme (à une permutation près des lignes et des colonnes) :

T
*

où T est un bloc carré triangulaire inférieur d'éléments diagonaux égaux à 1.

Éliminons une tranche par composante connexe, par exemple une tranche du type $E_{J_0}(e_{j_0})$; \mathcal{E}_1 est le sous-ensemble des tranches restantes. Nous allons construire une séquence finie de sous-ensembles de tranches

$$\mathcal{E}_1 \supset \mathcal{E}_2 \supset \mathcal{E}_3 \supset \dots \supset \mathcal{E}_p \supset \dots$$

en extrayant alternativement des arêtes de

$$\{ E_{J_0}(e_{j_0}) | e_{j_0} \in E_{J_0} : E_{J_0}(e_{j_0}) \in \mathcal{E}_1 \}$$

et de

$$\{ E_{J_1}(e_{\bar{J}_1}) \mid e_{J_1} \in E_{J_1} : E_{\bar{J}_1}(e_{\bar{J}_1}) \neq \phi \}.$$

Pour tout p entier supérieur ou égal à 1 nous définissons sur E une fonction à valeurs dans $\{0, 1, 2\}$ telle que $i_p(e) = |\{ \mathcal{A} \mid \mathcal{A} \in \mathcal{C}_p : e \in \mathcal{A} \}|$.

Nous construirons alors cette séquence par la récurrence suivante.

Soit E'_p le sous-ensemble des éléments de E tels que $i_p(e) = 1$.

Tant que $E'_p \neq \phi$ on pose $\mathcal{C}_{p+1} = \mathcal{C}_p \setminus \mathcal{C}_p$ où \mathcal{C}_p est le couplage de (E, \mathcal{E}_1) constitué par les tranches $E_{\bar{J}_m}(e_{J_m})$, $m \equiv p$ modulo 2, d'intersection non vide avec E'_p .

Notons R_p le sous-ensemble des cellules de E recouvertes par les arêtes du couplage \mathcal{C}_p , $R_p = \cup \{ \mathcal{A} \mid \mathcal{A} \in \mathcal{C}_p \}$.

Considérons un sommet e_q de E tel que $i_1(e_q) = 2$. Nécessairement il existe :

- 1) un sommet e_1 qui appartient à la composante connexe de (E, \mathcal{E}) contenant e_q et à E'_1 ,
- 2) une chaîne reliant e_1 à $e_q : (e_1, \mathcal{A}_1, e_2, \mathcal{A}_2, \dots, \mathcal{A}_{q-1}, e_q)$ avec \mathcal{A}_p arête du couplage \mathcal{C}_p contenant e_p et e_{p+1} .

En conséquence :

- 1) tout élément de E appartient à un sous-ensemble E'_p ;
- 2) tout élément e de E'_p est contenu dans une arête du couplage \mathcal{C}_p , donc $E'_p \subset R_p$ et puisque $i_p(e) = 1, i_{p+1}(e) = 0$, alors $e \notin E'_{p+1}$ et $E'_p \cap E'_{p+1} = \phi$;
- 3) tout élément de E'_{p+1} pour $p > 0$ est contenu dans une arête du couplage \mathcal{C}_p , donc $E'_{p+1} \subset R_p$.

On en déduit que le couple (E'_p, E'_{p+1}) est une partition de R_p et que la famille des E'_p est une partition de E . Soit E'_p un ensemble constitué d'un élément de E'_p par arête du couplage \mathcal{C}_p . La matrice d'incidence de l'hypergraphe (E, \mathcal{E}_1) peut alors se mettre sous la forme :

T_1	0	0	
0	T_2		0	
		⋮	
0	0	T_i	
	⋮		
*				

où les 0 sont des blocs d'éléments nuls et T_i est un bloc carré associé à la i ème composante connexe de (E, \mathcal{E}) qui est triangulaire inférieur, d'éléments diagonaux égaux à 1 et de structure type suivante :

	\mathcal{C}_1	\mathcal{C}_2	\mathcal{C}_3		\mathcal{C}_p	\mathcal{C}_{p+1}	
E_1''	I	0	0		0	0	
E_2''	*	I	0		0	0	
E_3''	0	*	I		0	0	
				⋮			
E_p''	0	0	0		I	0	
E_{p+1}''	0	0	0		*	I	
							⋮

où tout I désigne une matrice unité d'ordre convenable.

Toute combinaison linéaire de colonnes de cette matrice ne peut donc être nulle sans que tous ses coefficients le soient. Les indicatrices des tranches qui appartiennent à \mathcal{E}_1 sont donc indépendantes et la matrice d'incidence de l'hypergraphe a bien la forme souhaitée à une permutation près des lignes et des colonnes \square

3. EXEMPLES

Dans ce paragraphe nous présentons quelques exemples types de plans incomplets ou de tables tronquées. Pour chacun d'entre eux nous envisageons une hypothèse caractérisée par un ensemble \mathcal{H} de parties de I, ou éventuellement par le sous-ensemble \mathcal{H}^g engendrant \mathcal{H} par inclusion lorsque \mathcal{H} est hiérarchique, et nous indiquons les données et résultats des calculs effectués au moyen d'une procédure Fortran disponible [6] qui met en œuvre l'algorithme que nous venons de présenter. Plus précisément nous mentionnons :

- 1) le type de la famille génératrice utilisée et, s'il s'agit du second type, l'élément e^0 choisi,
- 2) classés automatiquement les vecteurs de base extraits de cette famille lors de la première étape et, s'il y a lieu, de la seconde,

3) la codimension du sous-espace $\Theta_{\mathcal{H}}$ de \mathbb{R}^E , autrement dit le degré de liberté résiduel pour les analyses statistiques les plus courantes.

Les exemples présentés ci-dessus ont été choisis de façon à ce qu'on puisse aisément vérifier les résultats obtenus.

La famille I des facteurs ou caractères observés est ordonnée arbitrairement et chaque élément i de I est repéré par son rang. Toute partie J de I est repérée par le $|J|$ -uple des rangs des éléments qu'elle contient. Les ensembles E_i des modalités des facteurs ou des caractères, composant le produit cartésien E_I , sont ordonnés arbitrairement et chaque élément e_i est repéré par son rang, $rg(e_i)$, dans E_i .

Tout élément $e = (e_i | i \in I)$ de E_I ou de $E \subset E_I$ est repéré par un $|I|$ -uple d'entiers positifs

$$(z_i | i \in I) \quad \text{avec} \quad z_i = rg_i(e_i)$$

Tout élément $e_j = (e_i | i \in J)$ de E_J est repéré par un $|I|$ -uple d'entiers positifs ou nuls

$$(z_i | i \in I) \quad \text{avec} \quad \begin{cases} z_i = 0 & \text{pour } i \in \bar{J} \\ z_i = rg_i(e_i) & \text{pour } i \in J. \end{cases}$$

E est défini par énumération des éléments e de E_I qu'il contient.

L'indicatrice γ^{e_j} de la tranche $E_J(e_j)$ est repérée par le même $|I|$ -uple que e_j enfin γ^ϕ est représentée par un $|I|$ -uple de zéros.

EXEMPLE 1. — $|I| = 2, |E_1| = |E_2| = 3$

$$E = \begin{array}{|c|c|c|} \hline 1, 1 & 1, 2 & \\ \hline 2, 1 & 2, 2 & \\ \hline & & 3, 3 \\ \hline \end{array}$$

Hypothèse représentée par μ ou $\lambda \in \Theta_{\mathcal{H}}$ (ou $\Omega_{\mathcal{H}}$) avec $\mathcal{H} = \{ \phi, \{1\}, \{2\} \}$

Famille génératrice de type 1 avec $e^0 = (1, 1)$

Base de $\Theta_{\mathcal{H}}$: vecteurs extraits lors de la 1^{re} étape : $(0, 0), (2, 0), (0, 2), (3, 0)$

Codim $[\Theta_{\mathcal{H}} = \Omega_{\mathcal{H}}] = 1$

Il s'agit ici d'un premier exemple élémentaire de confusion d'effets.

EXEMPLE 2. — $|I| = 3, |E_1| = |E_2| = |E_3| = 3$

$$E = \begin{array}{|c|c|c|} \hline 1, 1, 1 & 1, 2, 1 & \\ \hline 2, 1, 1 & 2, 2, 1 & \\ \hline & & 3, 3, 1 \\ \hline \end{array} \quad \begin{array}{|c|c|c|} \hline 1, 1, 2 & 1, 2, 2 & \\ \hline 2, 1, 2 & 2, 2, 2 & \\ \hline & & 3, 3, 2 \\ \hline \end{array} \quad \begin{array}{|c|c|c|} \hline & & 1, 3, 3 \\ \hline & & 2, 3, 3 \\ \hline 3, 1, 3 & 3, 2, 3 & 3, 3, 3 \\ \hline \end{array}$$

Hypothèse représentée par μ ou $\lambda \in \Theta_{\mathcal{H}}$ (ou $\Omega_{\mathcal{H}}$) avec

$$\mathcal{H}^g = \{ \{ 1, 2 \}, \{ 1, 3 \}, \{ 2, 3 \} \}$$

Famille génératrice de type 2

Base de $\Theta_{\mathcal{H}}$: vecteurs extraits lors de la 1^{re} étape : (1, 1, 0), (2, 1, 0), (1, 2, 0), (2, 2, 0), (3, 3, 0), (1, 0, 2), (2, 0, 2), (0, 2, 2), (3, 0, 2), (3, 0, 3), (3, 1, 0), (3, 2, 0), (1, 3, 0), (2, 3, 0)

Codim [$\Theta_{\mathcal{H}} = \Omega_{\mathcal{H}}$] = 1

Il s'agit ici d'un deuxième exemple de confusion d'effets. Ce plan (ou cette table) n'est pas connexe (ou est « séparable »). En effet, l'hypergraphe (E, \mathcal{E}) associé à deux composantes connexes.

EXEMPLE 3. — |I| = 3, |E₁| = 2, |E₂| = |E₃| = 3

E =

1, 1, 1		

	1, 2, 2	1, 3, 2
2, 1, 2		

	2, 2, 3	2, 3, 3

Hypothèse représentée par μ ou $\lambda \in \Theta_{\mathcal{H}}$ (ou $\Omega_{\mathcal{H}}$) avec $\mathcal{H}^g = \{ \{ 1 \}, \{ 2 \}, \{ 3 \} \}$

Famille génératrice de type 2

Base de $\Theta_{\mathcal{H}}$: vecteurs extraits lors de la première étape : (1, 0, 0), (2, 0, 0), (0, 2, 0), (0, 3, 0), (0, 0, 3)

Codim [$\Theta_{\mathcal{H}} = \Omega_{\mathcal{H}}$] = 0

Il s'agit ici d'un troisième exemple de confusion d'effets qui ne provient pas, cette fois, d'un défaut de connexité de (E, \mathcal{E}).

EXEMPLE 4. — |I| = 4, |E₁| = |E₂| = |E₃| = |E₄| = 4

E : carré gréco-latin $e = (i, j, k, l)$ représenté par $i \begin{matrix} j \\ \boxed{k, l} \end{matrix}$

	1	2	3	4
1	1, 1	2, 2	3, 3	4, 4
2	2, 3	1, 4	4, 1	3, 2
3	3, 4	4, 3	1, 2	2, 1
4	4, 2	3, 1	2, 4	1, 3

Hypothèse représentée par μ ou $\lambda \in \Theta_{\mathcal{H}}$ (ou $\Omega_{\mathcal{H}}$) avec

$$\mathcal{H} = \{ \phi, \{ 1 \}, \{ 2 \}, \{ 3 \}, \{ 4 \} \}$$

Famille génératrice de type 1 avec $e^0 = (4, 4, 4, 4)$

Base de $\Theta_{\mathcal{H}}$: vecteurs extraits lors de la première étape : $(0, 0, 0, 0)$,
 $(0, 0, 0, 2)$, $(0, 2, 0, 0)$, $(0, 0, 3, 0)$, $(3, 0, 0, 0)$, $(0, 0, 0, 3)$, $(2, 0, 0, 0)$,
 $(0, 3, 0, 0)$;
 vecteurs extraits lors de la seconde étape : $(1, 0, 0, 0)$,
 $(0, 0, 1, 0)$, $(0, 0, 2, 0)$, $(0, 0, 0, 1)$, $(0, 1, 0, 0)$

Codim $[\Theta_{\mathcal{H}} = \Omega_{\mathcal{H}}] = 3$

CONCLUSION

L'automatisation du traitement des plans d'expériences incomplets et des tables tronquées se heurte souvent à deux écueils en ce qui concerne l'estimation des paramètres et les tests d'hypothèses :

1) la réduction des équations de vraisemblance à un système minimal, ce qui nécessite la construction d'une base du sous-espace $\Omega_{\mathcal{H}}$ (ou $\Theta_{\mathcal{H}}$) de \mathbb{R}^E qui formalise le modèle retenu ou l'hypothèse à éprouver.

2) le calcul des degrés de liberté des statistiques de test usuelles.

Nous pensons que l'algorithme proposé dans cet article permet de franchir ces écueils sans entraîner des calculs trop onéreux.

De plus, comme la base retenue du sous-espace $\Theta_{\mathcal{H}}$ (ou $\Omega_{\mathcal{H}}$ lorsque \mathcal{H} est hiérarchique) de \mathbb{R}^E est constituée d'indicatrices, le système des équations de vraisemblance conserve une forme simple. Cela permet notamment au niveau des calculs :

1) d'éviter le stockage en mémoire de la matrice des vecteurs de base, stockage souvent prohibitif,

2) d'utiliser des procédures classiques pour la résolution de ce système, par exemple pour l'analyse des tables de contingence les « Iterative Scaling Procedures » [8] [4].

La programmation et l'exécution des calculs s'en trouvent donc simplifiées.

BIBLIOGRAPHIE

- [1] J. R. BARRA, *Notions Fondamentales de Statistique Mathématique*, Dunod, 1971, p. 258.
- [2] J. P. BENZECRI, *Lois de probabilité sur un ensemble produit : les diverses notions d'indépendance et le critère d'entropie maximale*. Document ronéotypé de l'I. S. U. P., C20, 1967, p. 50.
- [3] C. BERGE, *Graphes et Hypergraphes*. Dunod, 1972, p. 502.
- [4] Y. M. M. BISHOP, S. FIENBERG, P. W. HOLLAND, *Discrete Multivariate Analysis: Theory and Practice*, MIT-Press, 1975, p. 557.
- [5] D. COLLOMBIER, *Sur la résolution de problèmes d'affectation avec contraintes*. Thèse de 3^e cycle, Toulouse, 1972.

- [6] D. COLLOMBIER, Procédures Fortran pour l'Analyse de Tables de contingence. *Publications du Laboratoire de Statistique*, Université Paul Sabatier, Toulouse, t. **01-77**, 1977.
- [7] D. R. COX, *Analyse des données binaires*. Dunod, 1972, p. 122.
- [8] D. V. GOKHALE, An iterative procedure for analyzing log-linear models, *Biometrics*, t. **27**, 1971, p. 681-687.
- [9] S. J. HABERMAN, *The Analysis of Frequency data*. The University of Chicago Press, 1974, p. 419.
- [10] S. J. HABERMAN, Direct products and linear models for complete factorial tables. *Annals of Statistics*, t. **3**, 1975, p. 314-333.
- [11] J. R. HOWELL, Factorial Analysis of Variance. *Collected Algorithms from C. A. C. M.*, algorithm 359, 1969.
- [12] J. A. NELDER, R. W. M. WEDDERBURN, Generalized linear models. *Journal of the Royal Statistical Society, Serie A*, t. **135**, 1972, p. 370-384.
- [13] I. OLIVER, Factorial analysis of variance. *Collected algorithms from C. A. C. M.*, algorithm 330, 1968.
- [14] T. PHAM DINH, *Contribution à l'analyse de la variance et aux plans d'expérience*. Thèse de 3^e cycle, Grenoble, 1970.
- [15] R. L. PLACKETT, *The analysis of categorical data*. Griffin, 1974, p. 159.
- [16] S. N. ROY, M. A. KASTENBAUM, On the hypothesis of non-interaction in multiway contingency tables. *Annals of Mathematical Statistics*, t. **27**, 1956, p. 749-757.

(Manuscrit reçu le 1^{er} février 1978)