REVUE DE STATISTIQUE APPLIQUÉE

M. C. WEISS

Décomposition hiérarchique du Khi-deux associé à un tableau de contingence à plusieurs entrées

Revue de statistique appliquée, tome 26, n° 1 (1978), p. 23-33

<http://www.numdam.org/item?id=RSA_1978__26_1_23_0>

DÉCOMPOSITION HIÉRARCHIQUE DU KHI-DEUX ASSOCIÉ A UN TABLEAU DE CONTINGENCE A PLUSIEURS ENTRÉES

M.C. WEISS
Maître Assistant à l'Université René Descartes

INTRODUCTION

Le but de ce texte est de donner une présentation unifiée de résultats sur la décomposition du Khi-deux associé à un tableau de contingence à plusieurs entrées dans le cas où on dispose d'une classification hiérarchique sur chacune des entrées.

Le résultat principal est repris de la thèse de troisième cycle de CAPERAA (1968) mais les justifications, au lieu d'utiliser des formulations matricielles, sont fondées sur la décomposition des mesures considérées comme vecteurs d'un espace linéaire telle qu'elle a été mise en œuvre par H. ROUANET et D. LEPINE (1976).

Après avoir rappelé le concept de classification hiérarchique et en particulier dichotomique, on introduit la notion de base centrée selon une mesure de probabilité et on montre qu'on peut construire une telle base associée à un produit de classifications dichotomiques complètes. En se plaçant dans le cadre d'une mesure-produit on donne les formules générales pour chaque terme de la décomposition, d'une part du Khi-deux d'adéquation, d'autre part du Khi-deux d'indépendance. Un exemple concret correspondant à un tableau à trois entrées est traité numériquement.

1. CLASSIFICATION HIERARCHIQUE ; CLASSIFICATION DICHOTOMIQUE.

Notation allégée pour les partitions

Si $I = \{a, b, c, d, e, f, g\}$ la partition $\{\{a, b\}, \{c\}, \{d, e, f, g\}\}$ sera écrite $[a, b ; c ; d, e, f, g]$.

a) Classification hiérarchique

On appelle *dichotomie* une partition en deux classes d'un ensemble ; la partition grossière sera appelée *pseudo-dichotomie* (on l'assimile à une dichotomie qui serait constituée par l'ensemble lui-même et l'ensemble vide).

Dans ce qui suit on considérera un ensemble I fini de cardinal supérieur ou égal à deux, et des partitions de parties I' de cet ensemble que l'on appellera partitions *dans* I, on précisera partielles lorsque I' sera strictement inclus dans I, et partition *sur* I lorsque $I = I'$.

Mots-clés : **Tableau de contingence. Hiérarchies de parties. Décomposition Algébrique. Khi-deux.**

Une famille de partitions dans I sera appelée *classification hiérarchique* notée $\mathcal{H}$, ou *classification*, si, ordonnée par inclusion des parties :

1) il y a une seule partition initiale (sur I) ;

2) toute classe I′ à plus d'un élément engendre une partition (sur I′) de la classification ;

3) toutes les classes à un élément de I appartiennent à des partitions de la classification.

Exemple avec I = {a, b, c, d, e, f} A [a, b ; c ; d, e, f] il faut adjoindre [d, e ; f] pour avoir une classification.
[a ; b] [d ; e]

Si $\mathcal{H}$ est une classification hiérarchique de I on peut la compléter par la pseudo-dichotomie. On l'appellera classification complète notée $*\mathcal{H} = \mathcal{H} \cup [I ; \phi]$. Les partitions, rangées dans l'ordre indiqué, sont les sommets d'un arbre.

b) Classification dichotomique

On appelle classification dichotomique (resp. complète) notée $\mathcal{C}$ (resp. $*\mathcal{C}$) toute classification hiérarchique (resp. complète) formée uniquement de dichotomies. Une telle famille ordonnée forme un arbre à $|I| - 1$ (resp. $|I|$) sommets dont chaque embranchement est au maximum d'ordre deux.

A toute classification hiérarchique $\mathcal{H}$ on peut associer au moins une classification dichotomique faisant apparaître les mêmes classes. On appelle décomposition dichotomique $\mathcal{D}$ une sous-famille de cette classification dichotomique associée aux classes d'une partition de $\mathcal{H}$. En général il y a plusieurs classifications dichotomiques, et donc plusieurs décompositions des partitions.

Par exemple à [a, b ; c ; d ; e] on peut associer :

[a, b ; c, d, e] ou [a, b ; c, d, e]
[c ; d, e] [c, d ; e]
[d ; e] [c ; d] etc. (il y a six décompositions de cette partition).

On peut compléter une telle décomposition en classification dichotomique $\mathcal{C}$. Exemple : Si I = {a, b, c, d, e, f} il suffira d'adjoindre [a, b, c, d, e ; f] et [a ; b] à l'une des décompositions ci-dessus.

c) Produit de classifications

Soit $I = I^1 \times I^2 \times \ldots \times I^{|K|}$ noté $I = \prod_K I^k$, $k \in K = \{1, 2, \ldots, |K|\}$ (supposé aussi fini avec $|K| \geqslant 2$). On notera $i_1, i_2, \ldots, i_{|K|}$ un élément de I. Pour chaque k, soit $\beta_{L_k} = (B_{l_k}), \ell_k \in L_k$ une partition dans ℓ^k dont les B_{l_k} sont les $|L_k|$ parties. ($\beta_{L_k} \in P^k$ ensemble des partitions dans ℓ^k). $\prod_K B_{l_k}$ est (pour les ℓ_k fixés) une partie de ℓ.

Si P^K est l'ensemble des partitions dans $I = \prod_K I^k$ on appellera produit de classifications dans $I^k (k \in K)$ le résultat de l'opération qui associe aux β_{L_k} $(k \in K)$ la partition notée $\bigotimes_K \beta_{L_k} \in P^K$ formée des parties $\prod_{k \in K} B_{l_k}$, $l_k \in L_k$ et qui a donc $\prod_K |L_k|$ classes.

Si $\mathcal{H}^k$ est une classification de I^k, on définira $\mathcal{H}^K$ comme la famille des produits des partitions de chaque ensemble composant et on notera $\mathcal{H}^K = \underset{K}{\otimes} \mathcal{H}^k$, $k \in K$.

De même on définira ${}^*\mathcal{H}^K = \underset{K}{\otimes} {}^*\mathcal{H}^k$ à partir des classifications complètes dans chaque I^k. Si toujours $|L_k| = 2 \ \forall k \in K$ on obtient $\mathcal{C}^K = \underset{K}{\otimes} \mathcal{C}^k$ d'où ${}^*\mathcal{C}^K = \otimes\, {}^*\mathcal{C}^k$.

Exemple : $I^1 = \{a, b, c, d, e\}$ $\quad I^2 = \{f, g, h, i\}$

$\mathcal{H}^1 = [a, b ; c ; d ; e]$ et $\mathcal{H}^2 = [f, g ; h, i]$

$[a ; b]$ $\quad [f ; g]\ [h ; i]$

$\mathcal{H}^2$ \ $\mathcal{H}^1$	$[a, b ; c ; d ; e]$	$[a ; b]$
$[f, g ; h, i]$	$[\{a, b\} \times \{f, g\} ; \{c\} \times \{f, g\} ; \{d\} \times \{f, g\} ; \{e\} \times \{f, g\} ; \{a, b\} \times \{h, i\} ; \{c\} \times \{h, i\} ; \{d\} \times \{h, i\} ; \{e\} \times \{h, i\}]$	$[\{a\} \times \{f, g\} ; \{b\} \times \{f, g\} ; \{a\} \times \{h, i\} ; \{b\} \times \{h, i\}]$
$[f ; g]$	$[af, bf ; cf ; df ; ef ; ag, bg ; cg ; dg ; eg]$	$[af ; bf ; ag ; bg]$
$[h ; i]$	$[ah, bh ; ch ; dh ; eh ; ai, bi ; ci ; di ; ei]$	$[ah ; bh ; ai ; bi]$

Les partitions $[a, b ; c ; d ; e] \otimes [f ; g]$ et $[a ; b] \otimes [f, g ; h, i]$ ne sont pas comparables en finesse sur $\{a, b\} \times \{f, g\}$ et ne peuvent être engendrées l'une par l'autre. (Un produit $\mathcal{H}^K$ de classifications hiérarchiques n'est pas une classification).

2. DECOMPOSITION DE MESURES ET CLASSIFICATIONS DICHOTOMIQUES

Dans ce paragraphe on se donnera sur l'ensemble I une mesure de probabilité strictement positive $\underline{p}_I$ ($p_i > 0, \forall i \in I$).

a) Préalables et notations

La mesure $\underline{p}_I$ sera considérée comme un vecteur de l'espace des mesures sur I (on notera cet espace $\mathbf{R}_I$) muni de la base canonique $(\underline{\delta}^i_I)$, $i \in I$ (mesures de probabilité ponctuelles) : $\underline{p}_I = \sum_I p_i \, \underline{\delta}^i_I$.

Comme tous les coefficients p_i sont strictement positifs on peut prendre $\underline{p}_I$ comme mesure de référence et définir le produit scalaire de deux mesures sur I, $\underline{x}_I$ et $\underline{y}_I$, par $\langle \underline{x}_I \mid \underline{y}_I \rangle = \sum_I \frac{x_i \, y_i}{p_i}$. Avec ce produit scalaire la base canonique est orthogonale mais n'est pas normée $\| \underline{\delta}^i_I \| = \frac{1}{\sqrt{p_i}}$

On appelle *base centrée selon* $\underline{p}_I$ toute base orthogonale de $\mathbf{R}_I$ notée $(\underline{\alpha}_I^\ell)$, $\ell \in L = \{0, 1, \ldots, |\ell| - 1\}$ dont le vecteur $\underline{\alpha}_I^0$ coïncide avec $\underline{p}_I$. Il résulte de cette définition que pour $\ell = 0, <\underline{\alpha}_I^0 \mid \underline{\alpha}_I^0> = \| \alpha_I^0 \|^2 = 1$, et pour $\ell \neq 0 <\underline{\alpha}_I^0 \mid \underline{\alpha}_I^\ell> = 0$.

Or $<\underline{\alpha}_I^0 \mid \underline{\alpha}_I^\ell> = \sum_I \alpha_i^\ell$, donc $\ell \neq 0 \Rightarrow \sum_I \alpha_i^\ell = 0$.

Les vecteur de $\mathbf{R}_I$ orthogonaux à $\underline{\alpha}_I^0$ sont donc les mesures dont la somme des coefficients est nulle. On les appellera des contrastes et leur ensemble, comparaison globale sur I, les mesures proportionnelles à $\underline{\alpha}_I^0$ sont appelées pseudo-contrastes sur I.

Si $\underline{x}_I = \sum_I x_i \underline{\delta}_I^i$, et si les z_ℓ sont les coordonnées de $\underline{x}_I$ dans une base centrée $(\underline{\alpha}_I^\ell), \ell \in L$ (soit $\underline{x}_I = \sum_L z_\ell \underline{\alpha}_I^\ell$), on a $z_\ell = \dfrac{\langle \underline{x}_I \mid \underline{\alpha}_I^\ell \rangle}{\langle \underline{\alpha}_I^\ell \mid \underline{\alpha}_I^\ell \rangle}$.

Soit maintenant $\underline{f}_I$ une mesure de fréquence sur I et u_ℓ ses coordonnées dans la même base. $\langle \underline{\alpha}_I^0 \mid \underline{\alpha}_I^0 \rangle = 1$:

d'où $$u_0 = \langle \underline{f}_I \mid \underline{\alpha}_I^0 \rangle = \sum_I \frac{f_i p_i}{p_i} = \sum_I f_i = 1.$$

Donc $$\underline{f}_I = \underline{p}_I + \sum_{\ell=1}^{|I|-1} u_\ell \underline{\alpha}_I^\ell$$

Dans la suite on indexera souvent par a, b, c des parties disjointes de I. Ainsi on notera $x_a = \sum_{i \in a} x_i = \sum_a x_i$. Quand une sommation a lieu sur I dans son entier, on pourra omettre l'indice. Ex., si a et b indexent les deux classes d'une dichotomie sur I, on écrira $\sum_a x_i + \sum_b x_i = x_a + x_b = x = x_I$ s'il est besoin de le préciser.

De même avec plusieurs indices : ainsi si $i_1 \in I^1$ et si a_1 indexe une partie de I^1, si $i_2 \in I^2$ et que a_2 et b_2 indexent les deux classes d'une dichotomie sur I^2 :

$$\sum_{i_1 \in a_1} \left(\sum_{i_2 \in a_2} x_{i_1 i_2} + \sum_{i_2 \in b_2} x_{i_1 i_2} \right) = \sum_{i_1 \in a_1} \sum_{i_2 \in I^2} x_{i_1 i_2} = \sum_{a_1, I^2} x_{i_1 i_2} = x_{a_1 I^2} = x_{a_1}$$

b) Base centrée associée à une classification dichotomique complète sur I.

A la pseudo-dichotomie $[I ; \phi]$ on associera $\underline{\alpha}_I^0 = \underline{p}_I$. Il reste $|I| - 1$ dichotomies de la classification que nous numéroterons selon $\ell \in \{1, 2, \ldots, |I| - 1\}$.

Associons à la dichotomie $[a^\ell ; b^\ell]$ la mesure ayant pour coefficient : p_i / p_a^ℓ si $i \in a^\ell$; $- p_i / p_b^\ell$ si $i \in b^\ell$, 0 sinon.

Cette mesure associée à la dichotomie sera appelée *contraste normalisé* car la somme de ses coefficients positifs est égale à 1 (et celle de ses coefficients négatifs à -1). Elle est évidemment orthogonale à $\underline{\alpha}_I^0$.

Le carré de sa norme est $\| \underline{\alpha}_I^1 \|^2 = \frac{1}{p_{a1}} + \frac{1}{p_{b1}} = \frac{p_{a1} + p_{b1}}{p_{a1}\, p_{b1}}$.

Soient deux dichotomies $[a^1 ; b^1]$ et $[a^2 ; b^2]$ et $\underline{\alpha}_I^1$ et $\underline{\alpha}_I^2$ les contrastes associés. Ils sont orthogonaux, soit parce que leurs supports sont disjoints, soit parce que le support de l'un est inclus dans l'une des parties de l'autre.

Nous avons donc bien construit une base associée à la classification. Elle est définie au signe près car on a utilisé l'ordre d'écriture des parties a^ℓ et b^ℓ des dichotomies.

Remarque : si on a une base centrée telle que pour $\ell \neq 0$ les coefficients sont proportionnels à p_i/p_{a^ℓ}, $i \in a^\ell$, à $-p_i/p_{b^\ell}$, $i \in b^\ell$, 0 sinon (avec $p_{a^\ell} = \sum_{a^\ell} p_i$ et $p_{b^\ell} = \sum_{b^\ell} p_i$), on peut associer à chaque vecteur de la base la dichotomie $[a^\ell ; b^\ell]$. La famille de ces dichotomies formera une classification dichotomique de I.

c) Base centrée associée à un produit de classifications dichotomiques complètes

Soit $I = \prod_k I^k$, $k \in K = \{1, 2, \ldots, |K|\}$, chaque I^k étant muni d'une mesure de probabilité strictement positive $\underline{p}_{Ik}$ et $\mathbf{R}_{Ik}$ du produit scalaire associé.

On associera à I la mesure produit $\underline{p}_I = \bigotimes_K \underline{p}_{Ik}$ que l'on peut également écrire $\underline{p}_I = \sum_i (\prod_K p_{i_k})\, \underline{\delta}_I^i$ avec $i = (i_1 . i_2, \ldots, i_K)$, et on munira $\mathbf{R}_I$ du produit scalaire correspondant.

On vérifiera que lorsque I est muni de la mesure produit $\underline{p}_I$ le produit scalaire de deux mesures produits $\underline{\mu}_I$ et $\underline{\nu}_I$ est égal au produit des produits scalaires composants sur chaque I^k.

Considérons maintenant un produit de classifications dichotomiques complètes sur les I^k, $* \mathcal{C}^k = \bigotimes_K * \mathcal{C}^k$ on peut lui associer la base obtenue par le produit tensoriel des bases centrées associées à chaque $* \mathcal{C}^k$; $\underline{\alpha}_I^\ell = \bigotimes_K \underline{\alpha}_{Ik}^{\ell k}$, $\ell \in \prod L_k$ et $L_k = \{0, 1, \ldots, |I^k| - 1\}$. Montrons que cette base est centrée (selon la mesure produit $\underline{p}_I$). En effet premièrement elle comporte $|I| = \prod |I^k|$ vecteurs, deuxièmement $\underline{\alpha}_I^{(0,0,..0)} = \bigotimes_K \underline{\alpha}_{Ik}^0 = \bigotimes_K \underline{p}_{Ik} = \underline{p}_I$, troisièmement les vecteurs sont bien orthogonaux. Selon le théorème précédent le produit scalaire sur $\mathbf{R}_I$ est nul dès que l'un des produits scalaires composants s'annule, ce qui arrive lorsque deux indices sont différents pour un I^k, soit dès que $i \neq i'$, $i,i' \in I$.

Contrastes globaux et contrastes d'interaction

Soit $\underline{\gamma}_{Ik}$ un contraste sur I^k, $k \in K = \{1, 2, \ldots, |K|\}$ muni de la mesure de probabilité $\underline{p}_{Ik}$: on définira les notions suivantes relatives à des contrastes sur $I = \prod_K I^k$ muni de la mesure-produit $\bigotimes_K \underline{p}_{Ik}$:

- un contraste de la forme $\underline{\gamma}_{Ik} \otimes \prod_{k \in K - \{k\}} \underline{p}_{Ik}$ sera dit appartenir à la comparaison globale sur I^k ;

– un contraste de la forme $\underline{\gamma}_{Ik_1} \otimes \underline{\gamma}_{Ik_2} \otimes \prod_{k \in K - \{k_1, k_2\}} \underline{p}_{Ik}$ sera dit appartenir à la comparaison d'interaction (simple) entre I^{k_1} et I^{k_2} $(k_1 \neq k_2)$ notée $I^{k_1} . I^{k_2}$;

– un contraste de la forme $\underline{\gamma}_{Ik_1} \otimes \underline{\gamma}_{Ik_2} \otimes \underline{\gamma}_{Ik_3} \otimes \prod_{k \in K - \{k_1, k_2, k_3\}} \underline{p}_{Ik}$ sera dit appartenir à la comparaison d'interaction (double) entre I^{k_1}, I^{k_2} et I^{k_3} $(k_1 \neq k_2 \neq k_3)$ notée $I^{k_1} . I^{k_2} . I^{k_3}$, et ainsi de suite jusqu'à la comparaison d'interaction d'ordre $|K| - 1$ notée $I^1 . I^2 \ldots I^{|K|}$.

On peut maintenant décomposer $\mathbf{R}_I$ en ces sous-espaces orthogonaux.

Les projections d'un vecteur quelconque sur I^k, $I^k . I^{k'}, \ldots, I^1 . I^2 \ldots I^K$ sont orthogonales, donc le carré de sa norme est la somme des carrés des normes des projections.

3. VARIABLE MULTINOMIALE, VARIABLE Q^2 ET SA DECOMPOSITION

a) Un ensemble I

Soit $\underline{p}_I$ la mesure de probabilité sur I (strictement positive), on notera N_i les variables-effectifs, $i \in I$, qui suivent une loi multinomiale de paramètres n et $\underline{p}_I$. On considèrera également les variables-fréquences $F_i = N_i/n$; on notera $\underline{F}_I$ la mesure constituée par celles-ci.

Soit maintenant $(\underline{\alpha}_I^\ell)$, $\ell \in L$, une base centrée orthonormée selon $\underline{p}_I$ (cf. 2a). On a $\underline{F}_I = \underline{p}_I + \sum_{\ell=1}^{|I|-1} U_\ell \, \underline{\alpha}_I^\ell$ et $\underline{F}_I - \underline{p}_I = \sum_{\ell=1}^{|I|-1} U_\ell \, \underline{\alpha}_I^\ell$. $\underline{F}_I$ étant un ensemble de variables, les U_ℓ sont aussi des variables dont on peut connaître la distribution. On vérifie les propriétés : $E(U_\ell) = 0$, var $U_\ell = 1/n$. cov $(U_\ell, U_{\ell'}) = 0$ $(\ell \neq \ell')$.

On a donc $|I| - \ell$ variables non-corrélées de variance égale à $1/n$.

Dans les conditions asymptotiques où les $F_i - p_i$ suivent une loi de Gauss, il en est de même des U_ℓ et la famille $(\sqrt{n}\, U_\ell)$ est constituée de variables normales réduites indépendantes. On en déduit que $\sum_{\ell=1}^{|I|-1} (\sqrt{n}\, U_\ell)^2 = n \sum_{\ell=1}^{|I|-1} U_1^2 \simeq \chi^2$ à $|I| - 1$ d.ℓ.

Or $\| \underline{F}_I - \underline{p}_I \|^2 = \| \sum_{\ell=1}^{|I|-1} U_\ell \, \underline{\alpha}_I^\ell \|^2 = \sum_{\ell=1}^{|I|-1} U_\ell^2 \, \| \underline{\alpha}_I^\ell \|^2 = \sum_{\ell=1}^{|I|-1} U_\ell^2$, et en multipliant par n : $n \sum_I \frac{(F_i - p_i)^2}{p_i} = \sum_I \frac{(N_i - np_i)^2}{np_i} = n \sum_{\ell=1}^{|I|-1} U_\ell^2 \simeq \chi^2 \, |I| - 1$.

Posons $Q^2 = n \| \underline{F}_I - \underline{p}_I \|^2$, $Q^2 \cong \chi^2_{|I|-1}$ pour n grand.

Supposons maintenant que $(\underline{\alpha}_I^\ell)$, $\ell \in L$, soit la base associée à une classification dichotomique complète. A chaque dichotomie on peut associer la variable nU^2 correspondante distribuée asymptotiquement comme un χ^2 à 1 d.ℓ. Q^2 (lire khi-deux) est la somme des variables associées à toutes ces dichotomies.

Soit $\beta_M = (B_m)$, $m \in M$ une partition sur $I' \subset I$ ($|M| \geqslant 2$) dont la décomposition dichotomique (cf. 1b) appartient à la classification ci-dessus. On peut associer à β_M le sous-espace engendré par les vecteurs de base associés aux dichotomies de cette décomposition. Il ne dépend en fait que des parties B_m de la partition initiale. Soit U_M la projection de $\underline{F}_I$ sur ce sous-espace. nU_M^2 est un élément de la décomposition de Q^2 selon une hiérarchie comprenant β_M.

Si on pose $N_m = \sum_{B_m} N_i$ et $p_m = \sum_{B_m} p_i$, on vérifie que

$$nU_M^2 = n^{-1}\left(\sum \frac{N_m^2}{p_m} - \frac{N_{I'}^2}{p_{I'}}\right);$$

cette variable associée à β_M suit asymptotiquement un χ^2 à $|M| - 1$ d.l.

b) Un produit cartésien d'ensembles muni d'une mesure produit

Nous prendrons les mêmes conventions qu'en 2c. Si $I = \prod_K I^k$, $\underline{p}_I = \bigotimes_K \underline{p}_{I^k}$ mesure de probabilité strictement positive, on désignera par $N_i = N_{i_1,i_2\ldots i_{|K|}}$ les variables-effectifs suivant la loi multinomiale de paramètres n et $\underline{p}_I$.

Suivant les conventions indiquées en 2a, on écrira les variables-effectifs marginaux N_{i_k}, $i_k \in I^k$. La variable Q^2 définie plus haut sur I sera désormais écrite Q_T^2 (comme Q^2 total).

Munissons chaque I^k d'une base centrée orthonormée $(\underline{\alpha}_{I^k}^{\ell_k})$, $\ell_k \in L^{k_1}$, et I du produit tensoriel de ces bases. Q_T^2/n est le carré de la norme de $\underline{F}_I - \underline{p}_I$ qui est un contraste orthogonal à $\underline{\alpha}_I^\circ$ et appartient donc à un sous-espace vectoriel à $|I| - 1$ dimensions de $\mathbf{R}_I$. Ce dernier est la somme directe des sous-espaces orthogonaux suivants : C^k comparaison globale sur I^k et de toutes les comparaisons d'interactions $C^{K'} = I^{k_1} \ldots I^{k_r}$ avec $K' \subset K$ et $K' = \{k_1, k_2, \ldots, k_r\}$, $|K'| \geqslant 2$; $C^{K'}$ est engendré par $\left(\bigotimes_{k\in K-K'} \underline{\alpha}_{I^k}^\circ \bigotimes_{k\in K'} \underline{\alpha}_{I^k}^{\ell_k}\right)$, $\ell_k \neq 0$ et a par conséquent $\prod_{k\in K'} (|I^k| - 1)$ dimensions. On notera $Q_{K'}^2$ le carré de la norme de la projection de $\sqrt{n}\,(\underline{F}_I - \underline{p}_I)$ sur $C^{K'}$. Toutes ces projections étant orthogonales on a :

$$Q_T^2 = \sum_K Q_k^2 + \sum_{k_1<k_2\in K} Q_{k_1k_2}^2 + \ldots + Q_K^2.$$

Q_k^2 sera appelé Q^2 associé à la comparaison globale sur I^k. Pour $|K'| > 1$. $Q_{K'}^2$ sera appelé Q^2 d'interaction d'ordre $|K'| - 1$ entre les $I^{k'}$, $(k' \in K')$.

On a $Q_k^2 = \sum_{I^k} \dfrac{(N_{i_k} - np_{i_k})^2}{np_{i_k}}$. Q_k^2 est distribué asymptotiquement comme un χ^2 à $|I^k| - 1$ d.l., et peut se décomposer selon les éléments d'une classification dichotomique associée à I^k.

Si $K' \subset K$, $Q_{K'}^2$ est asymptotiquement distribué comme un χ^2 à $\prod_{k\in K'} (|I^k| - 1)$ d.l. et ne dépend que de $\underline{F}_{I^{K'}}$ (ou $\underline{N}_{I^{K'}}$) et $\underline{p}_{I^{K'}}$ (avec $I^{K'} = \prod_{k\in K'} I^k$). En

effet comme la base comprend les vecteurs $\underline{\alpha}^{\circ}_{I^k}$ pour $k \notin K'$ la projection sur $C^{K'}$ se fait par addition sur tous ces I^k. Ce $Q^2_{K'}$ peut se décomposer selon toute base de contrastes de $C^{K'}$ par projection de $\sqrt{n}(\underline{F}_I - \underline{p}_I)$ sur chaque vecteur de base, un terme de la décomposition suivant asymptotiquement un χ^2 à 1 d.l.

Supposons maintenant que $(\underline{\alpha}^{l}_{I^k})$, $l \in L_k$ soit la base associée à une classification dichotomique complète sur I^k et considérons une partition sur $I' \subset I^k$ pouvant être engendrée par les dichotomies ci-dessus, soit $\beta^k_{M_k} = \{B^k_m\}$, $m \in M_k$ ($|M_k| \geqslant 2$); on lui associera les $|M_k| - 1$ vecteurs correspondants de la base, et $\underline{\alpha}^{M_k}_{I^k}$ le sous-espace engendré.

Faisons de même pour tous les I^k. Au produit $\beta = \prod_{k \in K'} \beta^k_{M_k}$, on associera le sous-espace de $C^{K'}$ engendré par $\bigotimes_{k \in K - K'} \underline{\alpha}^{\circ}_{I^k} \bigotimes_{k \in K'} \underline{\alpha}^{M_k}_{I^k}$. Soit U_β la projection de $\underline{F}_I - \underline{p}_I$, nU^2_β sera la variable associée à β, élément de la décomposition de $Q^2_{K'}$, suivant asymptotiquement un χ^2 à $\prod_{k \in K'} (|M_k| - 1)$ d.l. On obtient

$$Q^2_\beta = nU^2_\beta = n^{-1} \left(\sum_{j_1 j_2 \ldots j_{|K'|}} (-1)^{n_{j_1 j_2 \ldots j_{|K'|}}} \frac{N^2_{j_1 j_2 \ldots j_{|K'|}}}{p_{j_1} p_{j_2} \ldots p_{j_{|K'|}}} \right)$$

avec $K' = \{k_1, k_2, \ldots, k_{|K'|}\}$ désignés par $\{1, 2, \ldots, |K'|\}$

$j_k \in J_k = \{B^k_1, B^k_2, \ldots, B^k_{|M_k|}, I'^k\}$, $n_{j_1 j_2 \ldots j_{|K'|}}$ = nombre d'indices tels que $j_k = I'^k$. Les indices inférieurs des N et p indiquent les sommations effectuées selon les conventions indiquées en 2a.

4. DECOMPOSITION DE Q^2_T DANS LE CAS D'UN PRODUIT CARTESIEN D'ENSEMBLES MUNI D'UNE MESURE-PRODUIT, LES PARAMETRES ETANT INCONNUS

On reprendra les mêmes notations que dans le paragraphe précédent.

Les probabilités marginales, inconnues, seront estimées par les fréquences correspondantes : $\forall k \in K\ \hat{p}_{i_k} = F_{i_k} = N_{i_k}/n$. On supposera dans la suite $N_{i_k} \neq 0$, $\forall i_k \in I^k$. L'estimation de Q^2_T est alors la statistique :

$$\hat{Q}^2_T = \sum_I n^{|K|-1} \frac{(N_i - n^{-|K|+1} \prod_K N_{i_k})^2}{\prod_K N_{i_k}}, \quad i = (i_1 i_2 \ldots i_{|K|}) \in \prod_K I^k.$$

Prenons $\underline{\hat{p}}_{I^k}$ comme mesure de référence sur I^k (et $\bigotimes_K \underline{\hat{p}}_{I^k}$ pour I). Q^2_T est le carré de la norme de $\sqrt{n}\ (\underline{F}_I - \hat{p}_I)$. Soit $^*\mathcal{C}^k$ une classification dichotomique complète de I^k et $(\underline{\alpha}^{l_k}_{I^k})$ $l_k \in L_k$ la base centrée orthonormée associée, elle est formée des estimateurs de la base associée à $^*\mathcal{C}^k$ avec $\underline{\hat{p}}_{I^k}$ comme mesure de référence. De même la base associée à $^*\mathcal{C}^k = \bigotimes_K {}^*\mathcal{C}^k$ avec $\underline{\hat{p}}_I$, $(\underline{\alpha}^{l}_I) = \bigotimes_K (\underline{\alpha}^{l_k}_{I^k_o})$, $l_k \in L_k$, $l \in L = \prod_K L_k$. est la famille des estimateurs de la base centrée orthonormée avec

$\underline{p}_I = \underset{\underline{p}_{Ik}}{\otimes}$ comme mesure de référence. On peut projeter $\sqrt{n}(\underline{F}_I - \underline{\hat{p}}_I)$ sur les différents vecteurs de $(\underline{\hat{\alpha}}_I^{\ell})$ et obtenir une décomposition de $\hat{Q}_T^2$ comme précédemment (cf. 3b).

a) Les termes associés aux comparaisons globales de chaque I^k sont nuls

Soit $\beta^k_{M_k} = \{B^k_m\}$, $m \in M_k$ une partition sur $I'^k \subset I^k$ (dont la décomposition dichotomique appartient à $\mathcal{C}^k$), et $\hat{U}_{M_k}$ la projection de $\underline{F}_I - \underline{\hat{p}}_I$ sur $\underline{\hat{\alpha}}_{I^k}^{M_k} \cdot \underset{k \in K - \{k\}}{\otimes} \underline{\hat{\alpha}}^{\circ}_{I^k}$. On a (cf. 3b) :

$$\hat{Q}^2_{M_k} = n\,\hat{U}^2_{M_k} = n^{-1}\left(\sum_{M_k} \frac{N_m^2}{\hat{p}_m} - \frac{N_{I'^k}^2}{\hat{p}_{I'^k}}\right) = \sum_{M_k} \frac{N_m^2}{N_m} - \frac{N_{I'^k}^2}{N_{I'^k}} = 0$$

Comme il y a $|I| - 1$ dichotomies associées à I^k, $\forall k \in K$ il y a $\sum_K (|I^k| - 1)$ vecteurs pour lesquels la projection est identiquement nulle.

b) Termes associés aux comparaisons d'interactions

Soit $K' \subset K$ ($|K'| > 1$), $\beta^k_{M_k} = \{B^k_m\}$, $m \in M_k$, $\forall k \in K'$ des partitions sur $I'^k \subset I^k$ (dont les décompositions dichotomiques appartiennent à chaque $\mathcal{C}^k$) et $\beta = \underset{K}{\otimes} \beta^k_{M_k}$, le terme de $\hat{Q}^2_{K'}$ qui lui est associé est :

$$\hat{Q}^2_\beta = n^{(|K'|-1)} \sum_{J_1 J_2 \ldots J_{|K'|}} (-1)^{n_{j_1 j_2 \ldots j_{|K'|}}} \frac{N^2_{j_1 j_2 \ldots j_{|K'|}}}{\prod_{K'} N_{j_k}}$$

avec $K' = \{k_1, k_2, \ldots, k_{|K'|}\}$ désignés par $\{1, 2, \ldots |K'|\}$, $J_k = \{B^k_1, B^k_2, \ldots, B^k_{|m_k|}, I'^k\}$, $n_{j_1 j_2 \ldots j_{|K'|}}$ = nombre d'indices tels que $j_k = I'^k$. $\hat{Q}^2_\beta$ suit asymptotiquement un χ^2 à $\prod_{K'} (|M_k| - 1)$ d.l..

Dans les conditions où la distribution de $(N_i - n\,p_i)/\sqrt{n \prod_K p^k_{i_k} (1 - \prod_K p^k_{i_k})}$, $i = i_1 i_2 \ldots i_{|K|}$, peut être approchée par une loi de Gauss centrée réduite à $\prod_K |I^k| - 1$ dimensions $\hat{Q}^2_T$ est distribué asymptotiquement comme un χ^2 à $|I^1|\,|I^2| \ldots |I^{|K|}| - 1 - \Sigma\,(|I^k| - 1)$ degrés de liberté. Les degrés de liberté perdus correspondent aux projections sur les comparaisons globales (et au nombre de paramètres estimés).

c) Applications à des données de mobilité sociale de Glass

Glass (1954) a étudié la répartition de 3 450 individus selon trois caractères I, J, K :

- la catégorie socio-professionnelle du père en cinq catégories : I = {a, b, c, d, e} de la plus favorisée à la moins favorisée : Cadre supérieur, cadre, cadre moyen, ouvrier qualifié, ouvrier spécialisé ;

– le niveau d'instruction du fils en quatre catégories : J = {f, g, h, i} du niveau supérieur au niveau inférieur : secondaire +, secondaire, primaire +, primaire ;

– la catégorie socio-professionnelle du fils (mêmes que pour le père) K = {s, t, u, v, w}

L'hypothèse d'indépendance totale signifierait une grande mobilité sociale ; $\hat{Q}_T$ est une distance à cette hypothèse.

Nous donnerons successivement : le tableau des données, $\hat{Q}^2_{I.J.K}$, $\hat{Q}^2_{I.J}$, $\hat{Q}^2_{I.K}$ et leurs degrés de liberté en faisant les calculs directement sur les partitions les plus fines pour chaque ensemble, et nous les reconstituerons en utilisant pour I et J les hiérarchies données en exemple en 1c. et pour K la partition la fine. Ceci illustrera la méthode de décomposition exposée.

K	s				t				u				v				w			
I \ J	f	g	h	i	f	g	h	i	f	g	h	i	f	g	h	i	f	g	h	i
a	62	32	12	18	23	9	13	14	7	5	3	10	10	10	8	21	1	3	4	11
b	21	7	8	10	12	13	12	27	6	11	26	23	12	8	28	74	0	2	8	34
c	15	3	6	7	9	2	24	21	21	10	22	55	12	11	53	144	2	2	15	77
d	11	5	22	9	14	1	52	45	8	3	63	108	26	21	175	480	5	12	47	380
e	5	0	3	1	4	1	11	17	0	0	23	46	9	12	61	234	1	6	33	367

Notations :

Partitions dans I : A = [a, b ; c ; d ; e], B = [a ; b] , I = [a ; b ; c ; d ; e]

dans J : C = [f, g ; h, i] , D = [f ; g] , E = [h ; i], J = [f ; g ; h ; i],

sur K : K = [s ; t ; u ; v ; w].

Ceci permettra de voir si pour une "distance à l'indépendance" il y a avantage à regrouper les CSP du père cadre supérieur et cadre, et pour les études du fils à ne considérer que deux niveaux, primaire ou primaire +, secondaire ou secondaire +, ce qui a pour effet d'augmenter les effectifs des catégories.

Tableau des $\hat{Q}^2_\beta$ avec le d.l. entre parenthèses après la spécification de β :

Partition la plus fine :

IJK (48) 1896,93	IJ (12) 681,02	JK (12) 924,93	IK (16) 1043,98
ACK (12) 1100,17	AC(3) 499,21	CK(4) 671,31	AK(12) 813,53
BCK(4) 679,92	BC(1) 127,80		
ADK(12) 44,09	AD(3) 2,73		
BDK(4) 26,76	BD(1) 4,12	DK(4) 41,93	BK(4) 230,46
AEK(12) 29,33	AE(3) 46,91		
BEK(4) 15,70	BE(1) 0,25	EK(4) 211,74	

On peut voir que les $\hat{Q}^2$ d'interaction sont reconstitués à la deuxième décimale près, ainsi que $\hat{Q}^2_T$ qui vaut 4545,92 si on le calcule directement par la formule

$\hat{Q}^2_T = n^2 \sum_{I,J,K} (N^2_{ijk}/N_iN_jN_k) - n$ (formule valable pour trois caractères).

Tous les $\hat{Q}^2$ d'interaction simple correspondent à des effectifs théoriques estimés supérieurs à 14. Ils sont tous significatifs à un seuil $\alpha < 0{,}001$ sauf $\hat{Q}^2_{AD}$, $\hat{Q}^2_{BE}$ non significatifs, et $\hat{Q}^2_{BD}$ significatif à 0,05. La non-significativité de $\hat{Q}^2_{AD}$ (resp. $\hat{Q}^2_{BE}$) veut dire que la séparation de f et g pour C (resp. a et b pour A, et h et i pour C) n'augmente pas la significativité de $\hat{Q}^2_{A\otimes[f;g;h,i]}$ (resp. $\hat{Q}^2_{I\otimes[f,g;h;i]}$).

Pour les $\hat{Q}^2$ d'interaction double seuls les effectifs théoriques minimaux de ACF et AEK sont supérieurs à 5. Ils sont cependant tous supérieurs à 1.

Les seuils observés lus dans la table du χ^2 sont tous inférieurs à 0,01. Si on admet que les distributions de $\hat{Q}^2$ sous l'hypothèse d'indépendance ne s'éloignent pas trop de celles des χ^2 vers les grandes valeurs on peut dire que chaque $\hat{Q}^2$ triple contribue à la significativité de $\hat{Q}^2_{IJ.K}$ et donc de $\hat{Q}^2_T$.

n étant facteur dans les $\hat{Q}^2$, leur valeur croît avec l'effectif global, pour un même tableau de fréquences. On peut essayer d'utiliser cette décomposition pour utiliser les regroupements qui rendent compte d'une forte proportion de la valeur globale. En retenant ceux qui sont supérieurs à 100 on considère le tableau regroupé : $I \otimes C \otimes K$ car A et B interviennent partout (on ne prend pas E introduit par $\hat{Q}^2_{EK}$ seulement)

$\hat{Q}^2_{I.C.K} = 1773{,}08$ (93 % de $\hat{Q}^2_{I.J.K}$)

$\hat{Q}^2_{I.C} = 627{,}01$ (92 % de $\hat{Q}^2_{I.J}$)

$\hat{Q}^2_{C.K} = 671{,}31$ (73 % de $\hat{Q}^2_{J.K}$)

$\hat{Q}^2_T$ (tableau regroupé) $= 4\,115{,}38$ (91 % de $\hat{Q}^2_m$)

Dans ce tableau les CSP du père sont toutes distinguées mais les études du fils sont regroupées en deux niveaux.

N.B. Le programme de calcul qui a été utilisé pour le calcul des différents $\hat{Q}^2_\beta$, directement à partir du tableau des effectifs à trois entrées, a été rédigé en Fortran par J.P. LE MOAN et est disponible à l'U.E.R. de Mathématiques, Logique Formelle et Informatique de l'Université René Descartes (Paris V). Un programme général est envisagé par H. ROUANET et M.O. LEBEAUX.

BIBLIOGRAPHIE

P. CAPERAA. – Sur un test applicable au contenu partiel d'un tableau de contingence, Thèse de 3ème cycle. Paris 1968 (ISUP).

D. GLASS. – *Social mobility in Britain*, London, Routledge et Kegan Paul, 1954.

H. ROUANET, D. LEPINE. – Structure linéaire et analyse des comparaisons, *Mathématiques et Sciences Humaines*, 56, 1976.