REVUE DE STATISTIQUE APPLIQUÉE

V. CHOULAKIAN

Analyse factorielle des correspondances de tableaux multiples

Revue de statistique appliquée, tome 36, n° 4 (1988), p. 33-41

<http://www.numdam.org/item?id=RSA_1988__36_4_33_0>

Rev. Statistique Appliquée, 1988, **XXXVI** (4), 33-42

ANALYSE FACTORIELLE DES CORRESPONDANCES DE TABLEAUX MULTIPLES

V. CHOULAKIAN*

Université de Moncton, Dept de mathématique, Moncton, N.-B. Canada, E1A 3E9

RÉSUMÉ

L'analyse factorielle des correspondances (AFC) s'applique aux tableaux rectangulaires. La méthode proposée pour étudier des tableaux multiples est de choisir le tableau binaire "optimal", c'est-à-dire ayant la plus grande variance et d'en faire l'AFC. Pour cela, on utilise la décomposition linéaire additive des interactions dans un tableau multiple de Lancaster (1951, 1969) et un lemme démontré par Choulakian (1985). On présente des exemples.

ABSTRACT

A method to analyse a multidimensional contingency table to flatten it into a binary table, then to carry out simple correspondence analysis was proposed by Benzecri (1976, p. 27). Lancaster's (1951, 1969) partitioning of K. Pearson's coefficient of contingency into additive terms is advocated; a simple criterion proposed by Choulakian (1985) will aid to select the appropriate flattened table. The procedure is illustrated on three-way and four-way contingency tables.

Mots clés: *Analyse factorielle des correspondances, Tableaux multiples, Décomposition additive des interactions.*

1. Introduction

Soit $k = (k_{ijs})$ pour $i = 1,\ldots,I, j = 1,\ldots,J$ et $s = 1,\ldots,S$ un tableau de correspondance ternaire. On peut présenter ce tableau ternaire comme un tableau rectangulaire et ce de trois manières différentes : $K_{(IJ)S}, K_{(IS)J}$ et $K_{(JS)I}$; de plus, il y a trois tableaux rectangulaires de marges notés K_{IJ}, K_{IS} et K_{JS}. A tous ces tableaux sont associés des lois de probabilités (fréquences) qu'on note en utilisant la lettre F au lieu de K.

La littérature statistique propose principalement trois méthodes d'analyse factorielle des correspondances (AFC) d'un tableau ternaire :

a) Analyse du tableau de Burt construit à partir des tableaux de marges, cf. Benzecri (1976).

---

* L'auteur remercie les referees pour leurs remarques constructives.

b) Si $S = T$ représente le facteur temps, analyse du tableau K_{IJ} en mettant les tableaux $K_{(IS)J}$ et $K_{I(SJ)}$ en supplémentaire à K_{IJ} ; voir, par exemple, Foucart (1984, p. 114), Cazes (1982) parmi d'autres.

c) Analyse séparée de différents tableaux binaires parmi les 6 tableaux rectangulaires cités ci-dessus; voir, par exemple, Feghali et Benzecri (1983).

Le présent article a pour but de proposer un critère pour nous guider dans le choix d'une méthode appropriée. Ce critère sera basé sur la décomposition linéaire additive des interactions dans un tableau multiple de Lancaster (1951, 1969) et un lemme démontré par Choulakian (1985). On discute le cas de tableaux ternaires et quaternaires; la généralisation au cas supérieur suivra sans difficulté.

2. Critère pour choisir le "meilleur" tableau binaire

Essentiellement, il y a deux théories pour décrire des interactions dans un tableau multiple : multiplicative et additive. La première a été mise au point par Bartlett (1935), Roy et Kastenbaum (1956), Ku et Kullback (1968) et d'autres. La méthode additive est due à Lancaster (1951, 1969, pp. 256-260). Une comparaison des deux méthodes est faite par Darroch (1974) et Lancaster (1971). Bener (1982) a traité la théorie additive dans un cadre géométrique.

La théorie additive des interactions d'un tableau ternaire (f_{ijs}) s'énonce comme suit :

$$\phi^2(I,J,S) = \phi^2(I,J) + \phi^2(J,S) + \phi^2(I,S) + \text{int}(I,J,S) \tag{1}$$

où

$$\phi^2(I,J) = \Sigma_{i,j}(f_{ij} - f_i f_j)^2/(f_i f_j) \tag{2}$$

représente la variance du tableau binaire K_{IJ}, et

$$\phi^2(I,J,S) = \Sigma_{i,j,s}(f_{ijs} - f_i f_j f_s)^2/(f_i f_j f_s) \tag{3}$$

représente la variance totale du tableau ternaire; $\text{int}(I,J,S)$ représente l'interaction ternaire contenue dans le tableau ternaire; f_i, f_j, f_k correspondent aux termes généraux des marges d'ordre 1 du tableau (f_{ijk}).

Un résultat similaire existe pour des tableaux de contingence quaternaire, soit (f_{ijsd}), alors :

$$\begin{aligned}\phi^2(I,J,S,D) = {} & \phi^2(I,J) + \phi^2(I,S) + \phi^2(I,D) + \phi^2(J,S) + \phi^2(J,D) + \phi^2(S,D) \\ & + \text{int}(I,J,S) + \text{int}(I,J,D) + \text{int}(J,S,D) + \text{int}(I,S,D) + \text{int}(I,J,S,D)\end{aligned} \tag{4}$$

où la variance totale est décomposée additivement en termes de six interactions binaires, quatre interactions ternaires et une interaction quaternaire.

Lemme :

a) Dans un tableau ternaire si $\phi^2(J,S)$ est négligeable[1] , alors

$$\phi^2(I,JS) \approx \phi^2(I,J) + \phi^2(I,S) + \text{int}(I,J,S) \tag{5}$$

[1] De façon précise, nous dirons que $\phi^2(J,S)$(resp.$\phi^2(J,S,D)$) est négligeable si toutes les quantités $r_{js} = (f_{js} - f_j f_s)/(f_j f_s)$ (resp. $r_{jsd} = (f_{jsd} - f_j f_s f_d)/(f_j f_s f_d)$) sont négligeables par rapport à 1.

et (5) est une égalité si et seulement si $\phi^2(J,S) = 0$.

b) Dans un tableau quaternaire, si $\phi^2(J,S,D)$ est négligeable[1], alors

$$\phi^2(I,JSD) \approx \phi^2(I,J) + \phi^2(I,S) + \phi^2(I,D) + \text{int}(I,J,S) + \text{int}(I,J,D) + \text{int}(I,S,D) + \text{int}(I,J,S,D) \quad (6)$$

et (6) est une égalité si et seulement si $\phi^2(J,S,D) = 0$ preuve : voir Choulakian (1985).

Les relations (5) et (1) impliquent que

$$\phi^2(I,J,S) \approx \phi^2(J,S) + \phi^2(I,JS) \quad (7)$$

Dans le cadre de la théorie de l'information une relation similaire à (7) a été donnée par Choulakian (1983).

De même les relations (6) et (4) impliquent que

$$\phi^2(I,J,S,D) \approx \phi^2(J,S,D) + \phi^2(I,JSD) \quad (8)$$

Les relations (1, 5, 7) ou (4, 6, 8) peuvent nous aider à choisir le tableau binaire "optimal", c'est-à-dire de variance maximale.

Dans le lemme ci-dessus, l'hypothèse peut être confirmée soit a) en calculant le % de la variance totale expliquée par $\phi^2(J,S)$ ou $\phi^2(J,S,D)$; ou b) en appliquant les tests classiques du khi-deux.

Voici une énumération des différents cas que l'on peut avoir pour un tableau ternaire (le cas d'un tableau quaternaire ou supérieur peut être déduit assez facilement) :

a) si seulement $\phi^2(J,S)$ est négligeable, alors on effectue l'AFC du tableau $f_{I(JS)}$ en mettant les tableaux f_{IJ} et f_{IS} en supplémentaires;

b) si seulement $\phi^2(I,J)$ est important, alors on effectue l'AFC du tableau f_{IJ} ;

c) si $\phi^2(J,S)$ et $\phi^2(I,S)$ sont négligeables, alors on effectue l'AFC du tableau $f_{I(JS)}$ ou $f_{J(IS)}$, en mettant le tableau f_{IJ} en supplémentaire;

d) si seulement int(I,J,K) est négligeable, on effectue l'AFC du tableau de Burt construit à partir de f_{IJ}, f_{IS}, et f_{JS} ;

e) si $\phi^2(J,S)$ et int(I,J,K) sont négligeables, on effectue, compte tenu de (1), l'AFC du sous-tableau de Burt composé de f_{IJ} et f_{IS} ;

f) si toutes les interactions binaires et ternaires sont non-négligeables, il n'est pas possible d'effectuer une seule AFC; on peut appliquer dans ce cas-là l'AFC généralisée, ayant la forme suivante, voir Choulakian (1988) :

$$f_{ijs} \cong f_i f_j f_s \left[1 + \sum_{\alpha=1}^{A} \left(\lambda_\alpha^{IJ} \psi_i^\alpha \psi_j^\alpha + \lambda_\alpha^{IS} \psi_i^\alpha \psi_s^\alpha + \lambda_\alpha^{JS} \psi_j^\alpha \psi_s^\alpha + \lambda_\alpha^{IJS} \psi_i^\alpha \psi_j^\alpha \psi_s^\alpha\right)\right]$$

(1) Cf. bas de la page précédente.

sous les contraintes

$$\Sigma_i f_i \psi^i_\alpha \psi^i_{\alpha'} = \Sigma_j f_j \psi^j_\alpha \psi^j_{\alpha'} = \Sigma_s f_s \psi^s_\alpha \psi^s_{\alpha'} = 1 \quad \text{si } \alpha' = \alpha \text{ avec } \alpha,\ \alpha' = 0;\ldots, A$$
$$= 0 \quad \text{si } \alpha' \neq \alpha$$

où

$$\psi^i_0 = \psi^j_0 = \psi^s_0 = 1 \quad \text{et} \quad A = \text{nombre de facteurs}$$

Les classifications faites ci-dessous s'appliquent aux tableaux ternaires qui n'ont pas de structure particulière (tableaux de confusion, etc.).

3. Exemples

Plusieurs tableaux multiples de différentes dimensions ont été analysés par la méthode proposée; ici, on n'en présente que deux.

a) Le tableau I est reproduit de Grizzle, Starmer et Koch (1969) ou Koch et Imrey (1985, p. 218). Les données représentent des patients ayant un ulcère duodénal et classifiés par rapport aux variables suivantes : hôpital ($S = 4$), type d'opération effectuée ($J = 4$) et gravité de l'opération ($I = 3$). Les 4 types d'opérations sont représentés par les lettres A (drainage de l'estomac), B (enlèvement de 25 % de l'estomac), C (enlèvement de 50 % de l'estomac) et D (enlèvement de 75 % de l'estomac). Les 3 modalités de la gravité de l'opération sont : nulle, faible et modérée.

TABLEAU I
Les données représentent des patients ayant un ulcère duodénal et classifiés par rapport aux variables suivantes : hôpital, type d'opération effectuée et gravité de l'opération

	Gravité de l'opération											
	Hôpital I			Hôpital II			Hôpital III			Hôpital IV		
Opération	nulle	faible	mod.	nulle	faible	mod.	nulle	faible	mod.	nulle	faible	mod.
A	23	7	2	18	6	1	8	6	3	12	9	1
B	23	10	5	18	6	2	12	4	4	15	3	2
C	20	13	5	13	13	2	11	6	2	14	8	3
D	24	10	6	9	15	2	7	7	4	13	6	4

Source : Grizzle, Starmer, and Koch (1969).

Les interactions additives sont : $\phi^2(I,J,S) = 0.0779$, $\phi^2(I,J) = 0.0253$, $\phi^2(J,S) = 0.0023$, $\phi^2(I,S) = 0.0195$ et $\text{int}(I,J,S) = 0.0308$. En utilisant le critère mentionné ci-dessus, on en déduit que $\phi^2(J,S)$ est négligeable, donc on effectue l'AFC du tableau $f_{I(JS)}$ en mettant les tableaux binaires f_{IJ} et f_{IS} en supplémentaires. $\phi^2(I,JS) = 0.0768$, alors qu'en appliquant (5), on obtient la valeur approximative 0.0756; la différence entre les deux valeurs est minime. On peut dire que le tableau $f_{I(JS)}$ explique approximativement 97.05 % de la dispersion totale. Il y a 2 axes principaux d'inertie : $\lambda_1 = 0.052$ et $\lambda_2 = 0.024$ sont les moments d'inertie d'ordre 1 et 2 respectivement. La figure 1 reproduit le plan principal. Sur la 1re bissectrice, les

modalités de l'opération et de la gravité sont ordonnées, donc il y a une association positive entre elles. On voit aussi les points "hôpital × opération" (par exemple 2A veut dire l'hôpital 2 × l'opération A).

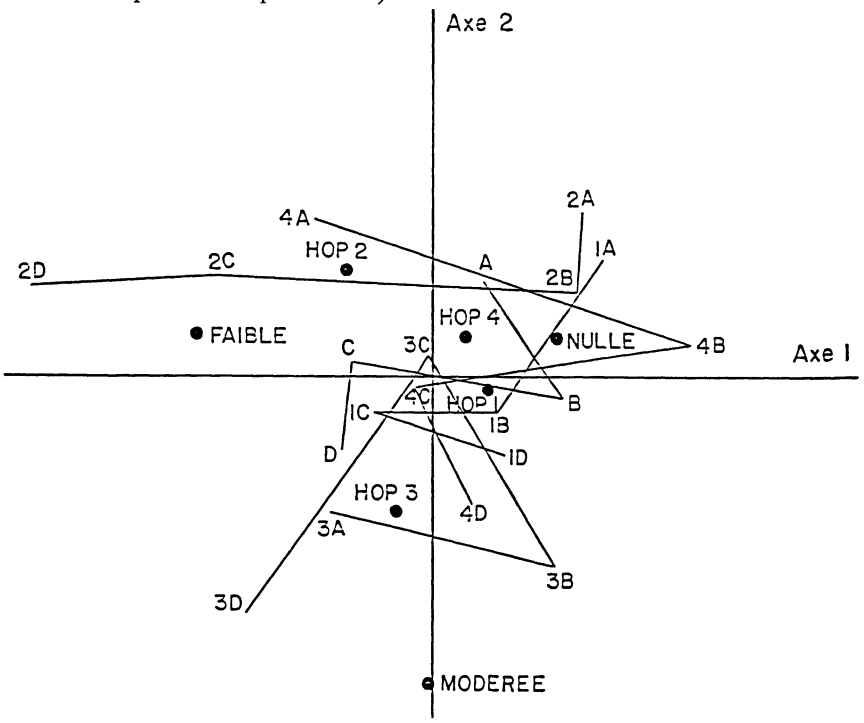

FIGURE 1

Plan (1-2) de l'AFC du tableau I. On a indiqué les principales variables et les variables supplémentaires

Le même tableau a été analysé aussi par Agresti (1983) par des méthodes log-linéaires utilisant une approche inférentielle.

b) Le tableau II est aussi reproduit de Koch et Imrey (1985, p. 140). C'est un tableau quaternaire où des femmes qui ont eu un enfant récemment sont classifiées par rapport aux variables suivantes : hôpital ($S = 4$), douleur sentie avant l'accouchement ($D = 2$), traitements ($I = 4$) et nombre d'heures sans douleur après l'accouchement ($J = 9$). Les traitements sont notés par : placebo, B = médicament B seulement, A = médicament A seulement et AB = combinaison des médicaments B et A.

Le tableau III représente les interactions additive du tableau II. On en déduit que l'on a intérêt à effectuer l'AFC du tableau $f_{J(DSI)}$ en mettant les tableaux f_{SJ}, f_{DJ} et f_{IJ} en supplémentaires. $\phi^2\ (J, SDI) = 0.6309$, tandis que sa valeur approximative calculée à partir de (6) est 0.6383. Le tableau $f_{J(SDI)}$ représente approximativement 97.51 %.

TABLEAU II

Les données représentent des femmes qui ont eu un enfant récemment, classifiées par rapport aux variables suivantes : hôpital, douleur sentie avant l'accouchement, traitements et nombre d'heures sans douleur après l'accouchement

hôpital	douleur initiale	traitement	nombre d'heures sans douleur après l'accouchement								
			0	1	2	3	4	5	6	7	8
1	peu	Placebo	1	0	3	0	2	2	4	4	2
1	peu	B	0	0	0	1	0	3	7	6	2
1	peu	A	2	1	0	2	1	2	4	5	1
1	peu	AB	0	0	0	0	1	3	5	4	6
1	beaucoup	Placebo	6	1	2	2	2	3	7	3	0
1	beaucoup	B	3	1	0	4	2	3	11	4	0
1	beaucoup	A	6	3	1	2	4	4	6	1	0
1	beaucoup	AB	0	0	0	1	1	7	9	6	2
2	peu	Placebo	2	0	2	1	3	1	2	5	4
2	peu	B	0	2	0	1	0	1	4	6	6
2	peu	A	0	0	0	1	1	1	8	1	7
2	peu	AB	0	0	0	1	3	0	4	7	5
2	beaucoup	Placebo	7	2	3	2	3	2	2	2	2
2	beaucoup	B	0	0	0	1	1	5	8	7	4
2	beaucoup	A	3	1	0	0	3	2	9	7	1
2	beaucoup	AB	0	1	0	0	1	2	8	9	5
3	peu	Placebo	5	0	0	1	3	1	4	4	5
3	peu	B	3	0	1	1	0	0	3	7	11
3	peu	A	1	0	0	1	3	5	3	3	6
3	peu	AB	0	0	0	1	1	4	2	4	13
3	beaucoup	Placebo	6	0	2	2	2	6	1	2	1
3	beaucoup	B	5	0	2	3	1	0	2	6	7
3	beaucoup	A	4	2	1	5	1	1	3	2	3
3	beaucoup	AB	3	2	1	0	0	2	5	9	4
4	peu	Placebo	1	0	1	1	4	1	1	0	10
4	peu	B	0	0	0	1	1	1	1	5	11
4	peu	A	0	0	0	1	0	2	2	1	13
4	peu	AB	1	0	0	0	0	2	2	2	14
4	beaucoup	Placebo	4	0	1	3	2	1	1	2	2
4	beaucoup	B	0	0	0	0	2	7	2	2	9
4	beaucoup	A	0	0	0	0	2	7	2	2	9
4	beaucoup	AB	1	0	3	0	1	2	3	4	8

Source : Koch and Imrey (1985).

TABLEAU III
Interactions additives dans le tableau II

Interactions binaires	$\varphi^2 \times 10^{+4}$
Traitement × nombre d'heures	1275.8
Traitement × douleur initiale	3.8
Douleur initiale × nombre d'heures	859.9
Hôpital × nombre d'heures	1622.6
Hôpital × traitement	15.2
Hôpital × douleur initiale	62.7
Interactions ternaires	**int × 10^{+4}**
Douleur initiale × traitement × nombre d'heures	365.9
Hôpital × traitement × nombre d'heures	1090.8
Hôpital × douleur initiale × nombre d'heures	326.1
Hôpital × douleur initiale × traitement	5.1
Interaction quaternaire	842.2
Variance totale	6470.1

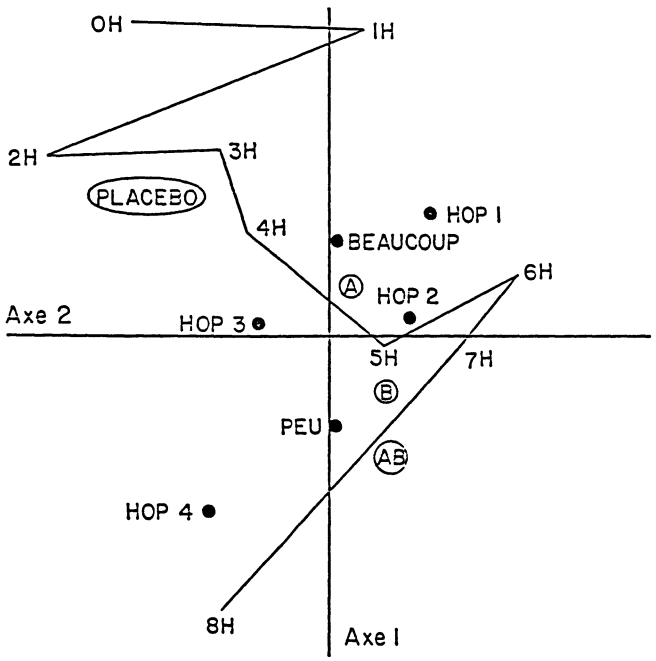

FIGURE 2
Plan (1-2) de l'AFC du tableau II. On a indiqué quelques variables principales et supplémentaires

de l'inertie totale. Il y a 8 axes principaux d'inertie : les quatre premiers moments d'inertie sont (entre parenthèses se trouvent le % de la variance expliquée de l'axe correspondant) :

$$\lambda_1 = 0.235\ (37.25\ \%), \quad \lambda_2 = 0.1409\ (22.33\ \%), \quad \lambda_3 = 0.0757\ (12\ \%),$$
$$\text{et } \lambda_4 = 0.057\ (9.09\ \%)$$

Les deux premières dimensions sont importantes. La figure 2 représente le plan principal. Sur le 1er axe, on remarque que les modalités de la variable nombre d'heures sans douleur après l'accouchement (iH, pour $i = 0, 1, \ldots 9$) sont approximativement ordonnées, ainsi que les modalités des différents traitements : placebo, A, B et AB, et les deux modalités de la variable douleur initiale : peu et beaucoup. Sur le 2e axe 5H, 6H et 7H s'opposent en particulier au reste.

4.Conclusion

La méthode proposée pour analyser des tableaux multiples par une AFC simple ne s'applique que si au moins l'une des interactions binaires dans un tableau ternaire est négligeable. Le même critère peut être appliqué aussi à l'analyse de tableaux multiples par des modèles généralisés log-linéaire, voir Goodman (1986) et Choulakian (1988).

Références

[1] A. AGRESTI (1983). — A survey of strategies for modeling cross-classifications having ordinal variables, *Journal of American Statistical Society*, 1983, 184-198.

[2] M.S. BARTLETT (1935). — Contingency table interactions, *Journal of Royal Statistical Society, Supp. 2*, 148-259.

[3] A. BENER (1982). — Décomposition des interactions dans une correspondance multiple, *Les Cahiers de l'Analyse des Données*, 7, 25-32.

[4] J.P. BENZECRI (1976). — L'Analyse des données : vol. 2, *L'analyse des Correspondances* (2e édition), Paris; Dunod.

[5] P. CAZES (1982). — Note sur les éléments supplémentaires en analyse des correspondances; I : Pratique et utilisation, II : Tableaux multiples, *Les Cahiers de l'Analyse des Données*, 7, 9-23, 133-154.

[6] V. CHOULAKIAN (1983). — Sur l'information associée aux différents tableaux issus d'un tableau ternaire, *Les Cahiers de l'Analyse des Données*, 8, 7-9.

[7] V. CHOULAKIAN (1985). — Relation entre les termes d'interaction supérieurs et les traces des correspondances binaires associées à une correspondance multiple, *Les Cahiers de l'Analyse des Données*, 10, 85-90.

[8] V. CHOULAKIAN (1988). — Exploratory analysis of contingency tables by log-linear formulation and generalizations of correspondence analysis, à paraître dans Psychometrika.

[9] J.N. DARROCH (1974). — Multiplicative and additive interaction in contingency tables, *Biometrika*, 61, 207-214.

[10] CH. FEGHALI et J.P. BENZECRI (1983). — L'articulation des voyelles : analyse sur micro ordinateur et interprétation des résultats en comparaison avec ceux issus d'une analyse ternaire, *Les Cahiers de l'Analyse des Données*, 8, 159-180.

[11] T. FOUCART (1984). — Analyse factorielle de tableaux multiples, Paris : Masson.

[12] L. GOODMAN (1986). — Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables (with discussions), *International Statistical Review*, 54, 243-309.

[13] GRIZZLE STARMER et KOCH (1969). — Analysis of categorical data by linear models, *Biometrics*, 25, 489-514.

[14] G.G. KOCH et P.B. IMREY (1985). — Analysis of categorical data, *Les Presses de l'Université de Montréal* : Montréal.

[15] H.H. KU et S. KULLBACK (1968). — Interaction in multidimensional tables : an information theoretic approach, *Journal of Research of the National Bureau of Standard-Mathematical Sciences*, 72B, 159-199.

[16] H.O. LANCASTER (1951). — Complex contingency tables treated by the partition of chi-square, *Journal of Royal Statistical Society B*, 13, 242-249.

[17] H.O. LANCASTER (1969). — The chi-squared distribution, Wiley, N.Y.

[18] H.O. LANCASTER (1971). — The multiplicative definition of Interaction, *Australian Journal of Statistics*, 13, 36-44.

[19] S.N. ROY et M.A. KASTENBAUM (1956). — On the hypothesis of no "interaction" in a multiway contingency table, *Annals of Mathematical Statistics*, 27, 749-757.