MATHÉMATIQUES ET SCIENCES HUMAINES

ETIENNE MULLET

La décomposition de l'interaction en analyse multivariée de l'incertitude

Mathématiques et sciences humaines, tome 76 (1981), p. 59-63

<http://www.numdam.org/item?id=MSH_1981__76__59_0>

Math. Sci. hum. (19e année, n°76, 1981, p.59-63)

LA DECOMPOSITION DE L'INTERACTION EN ANALYSE MULTIVARIEE DE L'INCERTITUDE

Etienne MULLET (1)

Le présent document n'est pas un article mais une question que je vais illustrer par un exemple numérique. Cette question est relative à la décomposition de l'interaction en Analyse Multivariée de l'Incertitude, c'est-à-dire à la décomposition (dans le cas orthogonal) de la quantité d'information supplémentaire A (x,t,y) en quantités plus élémentaires correspondant à chaque configuration de deux éventualités de x et de t.

Soit le tableau de contingence suivant dont chacun des éléments exprime une fréquence (exemple réel) :

	1	2	3	4	5	6	7	8	Total
A	.208	.0	.285	.002	.160	.153	.104	.089	1.000
B	.063	.030	.592	.115	.082	.066	.036	.017	1.000
C	.0	.118	.519	.260	.049	.050	.003	.0	1.000
m	.090	.049	.465	.126	.097	.090	.048	.035	1.000

Tab. 1

Les deux entrées de ce tableau correspondent, la première, en ligne, à la variable dépendante (8 éventualités), la seconde, en colonne, à la

(1) Cette étude a utilisé les moyens de travail du Service de recherches de l'INETOP (CNAM) et du Laboratoire de Psychologie Différentielle (EPHE, 3e section et Université René Descartes) ainsi que les moyens mis à la disposition du Laboratoire par le CNRS (ERA n° 79).

variable indépendante (3 éventualités). Ces deux variables sont nominales.

Le calcul de la liaison entre ces deux variables par la méthode de l'information (BACHER, 1957) conduit à la valeur .321 de l'indice T, quantité d'information transmise.

Soit un autre tableau de contingence, exprimant la liaison entre la même variable dépendante et une autre variable indépendante dont les trois éventualités sont X, Y, Z.

	1	2	3	4	5	6	7	8	Total
X	.250	.123	.429	.0	.0	.007	.123	.068	1.000
Y	.021	.024	.672	.010	.064	.149	.021	.038	1.000
Z	.0	.0	.295	.366	.226	.113	.0	.0	1.000
m	.090	.049	.465	.126	.097	.090	.048	.035	1.000

Tab. 2

La quantité d'information transmise est, dans ce cas, de .557.

Le troisième tableau exprime la liaison entre la même variable dépendante et les deux variables indépendantes prises conjointement.

	1	2	3	4	5	6	7	8	Total
AX	.563	.0	.0	.0	.0	.0	.255	.182	1.000
AY	.063	.0	.563	.0	.078	.182	.057	.083	1.000
AZ	.0	.0	.318	.005	.401	.276	.0	.0	1.000
BX	.188	.089	.594	.0	.0	.005	.104	.021	1.000
BY	.0	.0	.729	.005	.068	.161	.005	.031	1.000
BZ	.0	.0	.453	.339	.177	.031	.0	.0	1.000
CX	.0	.281	.693	.0	.0	.016	.010	.0	1.000
CY	.0	.073	.750	.026	.047	.104	.0	.0	1.000
CZ	.0	.0	.115	.755	.099	.031	.0	.0	1.000
m	.090	.049	.465	.126	.097	.090	.048	.035	1.000

Tab. 3

La quantité est cette fois de .980. Cette quantité est supérieure à la somme des deux précédentes : .878. Dans le cas orthogonal qui est le nôtre la quantité supplémentaire d'information transmise (.102) est directement assimilable à l'interaction :

$$T\ (x\ ,\ t\ ;\ y) = T\ (x\ ;\ y\ ) + T\ (t\ ;\ y) + A\ (x\ ,\ t\ ,\ y)$$

où T (x , t ; y) est la quantité transmise par les deux variables indépendantes prises conjointement,

T (x , y) et T (t ; y) les quantités transmises par chaque variable prise séparément,

A (x , t , y) une quantité supplémentaire assimilable à l'interaction.

Outre le calcul de cette quantité A (x , t , y), ce qui m'importe, compte tenu de ma problématique, est de repérer quelles configurations d'éventualités des deux variables indépendantes (de AX à CZ) sont plus précisément responsables de cette quantité A (x , t , y). Pour ce faire j'ai adopté une méthode empirique qui bien que donnant relativement satisfaction mériterait peut-être de recevoir un fondement mathématique plus solide voire d'être grandement améliorée.

Je constitue un tableau intermédiaire, censé représenter les valeurs attendues des fréquences dans le cas de stricte additivité des effets des deux variables. Ce tableau est constitué par application de la formule :

$$p_{i,j} = p_i \cdot p_j \ / \ p_k$$

où $p_{i,j}$ représente chacune des valeurs du tableau intermédiaire,

p_i chacune des valeurs du tableau correspondant au croisement de la première variable indépendante avec le critère (de .208 à .000)

p_j chacune des valeurs du tableau correspondant au croisement de la seconde variable avec le critère (de .250 à .000)

p_k chacune des fréquences moyennes observées des éventualités de la variable dépendante (de .090 à .035).

Dans cet exemple le tableau est le suivant :

	1	2	3	4	5	6	7	8	Total
AX	.577	.0	.262	.0	.0	.012	.267	.170	1.288
AY	.048	.0	.411	.000	.106	.254	.045	.096	.961
AZ	.0	.0	.181	.005	.373	.192	.0	.0	.751
BX	.173	.074	.546	.0	.0	.005	.094	.033	.925
BY	.014	.015	.855	.010	.054	.110	.016	.019	1.092
BZ	.0	.0	.376	.334	.191	.083	.0	.0	.983
CX	.0	.296	.478	.0	.0	.004	.009	.0	.787
CY	.0	.058	.750	.022	.032	.084	.002	.0	.947
CZ	.0	.0	.329	.760	.114	.063	.0	.0	1.266
m	.090	.049	.465	.126	.097	.090	.048	.035	1.000

Tab. 4

Dans ce tableau le total de chaque ligne n'est pas nécessairement égal à 1. Σp_{ij} est généralement différent de $\Sigma p_i \Sigma p_j / \Sigma p_k$. Ce tableau est transformé de telle manière que chaque total de ligne soit égal à 1.000.

1	2	3	4	5	6	7	8	Total
.448	0	.203	0	0	.009	.207	.133	1.000
.050	0	.427	0	.110	.265	.048	.100	1.000
0	0	.241	.006	.497	.256	0	0	1.000
.187	.080	.590	0	0	.005	.102	.036	1.000
.013	.014	.783	.009	.049	.100	.015	.017	1.000
0	0	.383	.340	.193	.084	0	0	1.000
0	.376	.607	0	0	.005	.012	0	1.000
0	.061	.792	.023	.034	.088	.002	0	1.000
0	0	.260	.600	.090	.050	0	0	1.000

Tab. 5

Je calcule alors l'entropie de chaque distribution par ligne, de AX à CZ, dans ce dernier tableau et je la soustrais de l'entropie de chaque ligne correspondante du tableau 3. J'obtiens chaque fois une quantité E (x_M , t_L , y), tantôt positive, tantôt négative, qui représente pour moi la contribution de chaque ligne du tableau 3, de chaque configurations d'éventualités, à la quantité A (x , t , y). Ces quantités sont dans l'ordre, les suivantes : .488, .140, -.062, .034, -.001, 174, .039, -.141, .326.

On observe en général que $\overline{E}$ (x_M , t_L , y) est une bonne approximation de A (x , t , y). (Cette méthode permet outre les lignes de repérer les cases responsables de l'interaction).

C'est à propos de cette méthode empirique et de cet indice E (x_M , t_L , y) que j'aimerais connaître l'avis des mathématiciens, lecteurs et auteurs de cette revue.

REFERENCES

BACHER (F.).- Etude de la liaison statistique entre deux variables par la méthode de l'information. B.I.N.O.P., 1957, 13, 13-25.