Statistique et analyse des données

J. J Daudin

Coefficient de Tschuprow partiel et indépendance conditionnelle

Statistique et analyse des données, tome 4, n° 3 (1979), p. 55-58.

<http://www.numdam.org/item?id=SAD_1979__4_3_55_0>

Statistique et Analyse des Données
3 - 1979 pp. 55-58

COEFFICIENT DE TSCHUPROW PARTIEL ET INDEPENDANCE CONDITIONNELLE

J.J DAUDIN

Institut National Agronomique Paris Grignon
Service de Mathématiques
16, rue Claude Bernard - 75005 PARIS

Résumé

Nous étudions les relations logiques entre indépendance conditionnelle et la nullité du coefficient de liaison partielle proposé par Saporta.

1 - RAPPEL DE LA DEFINITION DU COEFFICIENT DE TSCHUPROW PARTIEL

Soit X,Y,Z 3 variables aléatoires discrètes, à valeurs respectivement sur les ensembles [1,2,...I] , [1,2..J] et[1,2,..K]

Soit $P_{ijk} = p\ (X=i \cap Y=j \cap z=k)$

$$p_{ij.} = \sum_k p_{ijk}$$

et des définitions semblables pour $p_{i.k}$, $p_{.jk}$, $p_{.j.}$ et $p_{..k}$

$$\text{Soit } \Phi^2_{XY} = \sum_{ij} \frac{p^2_{ij.}}{p_{i..}p_{.j.}} - 1$$

$$\Phi^2_{XZ} = \sum_{ij} \frac{p^2_{i.k}}{p_{i..}p_{..k}} - 1$$

$$\Phi^2_{YZ} = \sum_{jk} \frac{p^2_{.jk}}{p_{.j.}p_{..k}} - 1$$

soit enfin

$$T_{XY} = \frac{\Phi^2_{XY}}{\sqrt{(I-1)\ (k-1)}}$$

$$T_{XZ} = \frac{\Phi^2_{XZ}}{\sqrt{(I-1)\ (k-1)}}$$

$$T_{YZ} = \frac{\Phi^2_{YZ}}{\sqrt{(J-1)\ (k-1)}}$$

T_{XY} est le carré du coefficient de Tschuprow qui est une mesure de la liaison entre X et Y.(voir Kendall et Stuart tome 2 p.557)
Saporta (1976) a montré que l'on peut définir Φ^2_{XY} comme produit scalaire de 2 opérateurs. Comme les normes de ces 2 opérateurs sont respectivement $\sqrt{I-1}$ et $\sqrt{J-1}$, T_{XY} peut être défini comme le produit scalaire entre 2 opérateurs divisés par le produit de leurs normes.
Cette propriété de T_{XY} qui en fait un indice de liaison entre variables discrètes analogue au coefficient de corrélation entre variables continues a conduit Saporta a définir le coefficient $T_{XY,Z}$ par analogie avec le coefficient de correlation partiel.

$$T_{XY,Z}=\frac{T_{XY}-T_{XZ}\,T_{YZ}}{\sqrt{1-T^2_{XZ}}\,\sqrt{1-T^2_{YZ}}}$$

Saporta a proposé $T_{XY,Z}$ comme indice de liaison partielle entre X et Y à Z fixé et l'a appelé coefficient de Tschuprow partiel. Il est naturel alors de vérifier si la nullité de $T_{XY.Z}$ caractérise l'indépendance conditionnelle, c'est à dire si on obtient l'équivalence :

$T_{XY.Z}=0 \Leftrightarrow$ X,Y indépendantes à Z fixé.

En fait on va montrer que les 2 implications sont fausses.

2 - $T_{XY.Z}=0$ et NON INDEPENDANCE CONDITIONNELLE

On considère la table suivante de probabilités (p_{ijk}) avec I=J=K=2

k	j	i=1	i=2
1	1	1/4	1/12
1	2	0	1/6
2	1	1/2	1/6
2	2	1/4	1/2

Table 1

Les 3 marges de la table 1 sont :

Y \ X	i=1	i=2
j=1	1/4	1/4
j=2	1/4	1/4

$T_{XY}=0$

Z \ X	i=1	i=2
k=1	1/4	1/4
k=2	1/4	1/4

$T_{XZ}=0$

Z \ Y	j=1	j=2
k=1	1/3	1/6
k=2	1/6	1/3

$T_{YZ}\neq 0$

Alors $T_{XY.Z}=0$ tandis que

$$p_{111}= 1/4 \neq \frac{p_{1.1}p_{.11}}{p_{..1}} = \frac{1/4.1/3}{1/2} = 1/6$$

Ce qui implique que X et Y ne sont pas independantes conditionnellement à Z.

Soit la table de probabilité suivante et ses marges :

k	j	i=1	i=2
1	1	1/2	0
1	2	0	1/2
2	1	o	1/2
2	2	1/2	0

	i=1	i=2
j=1	1/2	1/2
j=2	1/2	1/2

$T_{XY}=0$

	i=1	i=2
k=1	1/2	1/2
k=2	1/2	1/2

	j=1	j=2
k=1	1/2	1/2
k=2	1/2	1/2

$T_{YZ}=0$

On obtient $T_{XY.Z}=0$ et pourtant la liaison entre X et Y à Z fixé est maximale puisque la connaissance de la valeur de X implique celle de Y.
Ces contres exemples indiquent que $T_{XY.Z}=0$ n'implique pas l'indépendance conditionnelle et que on peut avoir à la fois $T_{XY.Z}=0$ et une dépendance entre X et Y à Z fixé.

3- INDEPENDANCE CONDITIONNELLE ET $T_{XY.Z} \neq 0$

On suppose que X et Y sont indépendantes conditionnellement à Z. En fait on n'obtient alors $T_{XY.Z}=0$ que dans le cas particulier où K=2.
Dans le cas où $K \neq 2$, $T_{XY.Z}$ peut être très différent de 0 comme on le voit sur l'exemple suivant.
On considère la table de probabilité (X10) suivante : Table 2

Table 2

	i=1	i=2	i=1	i=2	i=1	i=2
j=1	2	1	0	0	2	1
j=2	0	0	1	3	0	0
	k=1		k=2		k=3	

Les marges de la table 2 sont les suivantes :

X \ Z	k=1	k=2	k=3
i=1	2	1	2
i=2	1	3	1

Y \ Z	k=1	k=2	k=3
j=1	3	0	3
j=2	0	4	0

X \ Y	j=1	j=2
i=1	4	1
i=2	1	3

On obtient T_{XY} =.322
T_{XZ} =.118
T_{YZ} =.707
$T_{XY.Z}$=.339

La nullité de $T_{XY.Z}$ pour K=2 provient de l'égalité connue
$\Phi^2_{XY} = \Phi^2_{XZ} \cdot \Phi^2_{YZ}$
Remarquons pour conclure que la non vérification de l'implication

$T_{XY.Z}=0 \Rightarrow$ indépendance conditionnelle

n'est pas tragique:

elle n'est pas non plus réalisée par le coefficient de corrélation pour des variables non gaussiennes.
Par contre la non vérification de l'implication

indépendance conditionnelle$\Rightarrow T_{XY.Z}=0$

nous semble plus grave et conduit à mettre en question l'utilisation de ce coefficient.

Je remercie G.Saporta et G.Drouet d'Aubigny pour leurs utiles remarques.

BIBLIOGRAPHIE

M.G. KENDALL, A.STUART. The advanced theory of statistics volume 2 Griffin 3eme edition

G.SAPORTA (1976) "quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives" Bulletin de l'Association des Statisticiens Universitaires - ler trimestre 1976, 38-46.