50 FF

ISSN 0759-1063

Contains RC 33 Newsletter

BMS

BULLETIN DE METHODOLOGIE SOCIOLOGIQUE

N. 44 SEPTEMBER 1994

Bulletin de Méthodologie Sociologique, September 1994, N. 44, pp. 93 - 102

ON-GOING RESEARCH
RECHERCHE EN COURS

LIAISON ET INFORMATION DANS LES TABLEAUX DE CONTINGENCE: COMMENTAIRE DE L'ARTICLE DE FREDERIC MICHEL

par

Philippe Cibois
(Université René-Descartes - Paris V,
12 rue Cujas, 75005 Paris)

Abstract. Links and Information in Contingency Tables - Commentary on Frédéric Michel's Article. Frédéric Michel's article ("Link Strength and Volume of Information in Contingency Tables - Two Problems of Measure in Data Analysis") published in this issue, proposes solutions to a certain number of problems. It would be helpful for research on these problems to situate the author's solutions in the framework of preceding research on the question. **Data Analysis, Contingency Tables, Measures of Information, Link Strength.**

Résumé. L'article de Frédéric Michel ("Intensité de liaison et masse d'information des tableaux de contingence: deux problèmes de mesure en analyse des données") publié dans le présent numéro, propose des solutions à un certain nombre de questions. Il serait utile pour éclairer le problème, de replacer la solution de l'auteur dans la logique de ce qui a été fait antérieurement. **Analyse des données, Tableaux de contingence, Mesure d'information, Indice de liaison.**

LES COEFFICIENTS DE CONTINGENCE

La logique du premier problème proposé est celle-ci: quand on a un tableau de contingence quelconque, on observe une liaison entre les lignes et les colonnes. Pour pouvoir la mesurer on a besoin d'une référence quant à son minimum et à son maximum afin de pouvoir juger si elle est importante ou non.

En ce qui concerne le minimum, la solution est assez simple puisque l'absence d'écarts à l'indépendance correspond à l'absence de liaison. Par contre pour le maximum les avis divergent et nous allons examiner les diverses solutions que l'on trouve dans la littérature pour en voir la logique et y voir la place que l'indice GT proposé par l'auteur peut y tenir.

Pour mesurer la liaison on se sert en général de l'indice Khi-deux en mettant en rapport le Khi-deux observé et un Khi-deux qui correspondrait au maximum de la liaison. Dans ce domaine la solution la plus radicale est celle proposée par Harald Cramér: "If q is the smallest of the numbers r and s, it follows (...) that:

$$0 <= f^2 / q - 1 = Khi^2 / n (q - 1) <= 1$$

The upper limit 1 is attained when and only when each row (when r >= s) or each column (when r <= s) contains one single element different from zero. Thus,

$$Khi^2 / n (q - 1)$$

may be regarded as a measure of the degree of association indicated by the sample." (Cramér 1946: 443). Il ajoute ensuite que cet indice suit évidemment une loi du Khi-deux et que d'autres mesures d'association sont proposées dans le Yule-Kendall (1940).

Cette solution est radicale car comme l'indique Cramér, le cas d'égalité à 1 peut se rencontrer, non seulement dans un tableau carré comme semble le croire Frédéric Michel mais aussi dans un tableau rectangulaire si, sur la plus petite dimension du tableau, tout l'effectif se trouve sur une case comme dans le cas suivant:

	COL1	COL2	COL3	Total
LGN1	3	0	0	3
LGN2	0	2	0	2
LGN3	0	0	4	4
LGN4	0	0	1	1
Total	3	2	5	10

En traduisant les notations de Cramér et en notant désormais par l le nombre de lignes, par c le nombre de colonnes, par k le minimum de l et de c, par Phi-deux le f^2 et par N l'effectif total on a dans ce cas:

l = 4, c = 3, k = 3, N = 10, Phi-deux = 2, Khi-deux = 20

Le Phi-deux maximum est de k - 1 soit 2 et le Khi-deux maximum est de N (k - 1) soit 20. Le Phi-deux observé est égal au Phi- deux maximum et le coefficient de Cramér est égal à 1.

On peut dire que cette solution est radicale car il est bien rare que dans un tableau observé, tout l'effectif d'une ligne (ou d'une colonne) se concentre sur une seule case: on sent bien que l'on a avec le coefficient de Cramér (désigné souvent par la lettre V) une limite toute théorique et somme toute assez peu réaliste. De ce fait le Phi-deux observé ne représente qu'une petite part du Phi-deux maximum et ce coefficient semble assez pessimiste. Par exemple sur le premier exemple proposé par Frédéric Michel (sur le vote syndical, tableau à 4 lignes et 7 colonnes) où le Phi-deux observé est de 0,2619 (Khi-deux de 221,1), le Phi-deux maximum est de k - 1 = 3, le V de Cramér est de: 0,2619 / 3 = 0,087 ce qui peut s'interpréter en disant que le Phi-deux observé (ou le Khi-deux observé) représente 8,7% du Phi-deux maximum (resp. du Khi-deux maximum), ce qui semble bien faible.

On trouve donc dans la littérature plusieurs autres coefficients qui vont proposer des limites pour le Phi-deux maximum qui soient plus réalistes que celle proposées par Cramér. On a d'abord un coefficient attribué à Pearson sous divers nom: Carré moyen de contingence pour Yule-Kendall ou simplement Coefficient de contingence pour McNemar ou Siegel (Conover 1971: 181) ou Coefficient de Pearson (Rouanet 1987: 163). Il est défini de la manière suivante: C^2 = Phi-deux / 1 + Phi-deux. Son maximum théorique n'est pas 1 mais k - 1 / k qui est évidemment proche de 1 quand k est assez grand.

Si l'on reprend le même exemple, C^2 = 0,2619/1 + 0,2619 = 0,208 ce qui peut s'interpréter en disant que le Phi-deux observé représente 20,8% d'un maximum réaliste, même si ce n'est pas le maximum théorique puisqu'ici le maximum de C^2 est 0,75 (pour k=4).

On trouve enfin le coefficient de Tchuprov (Conover 1971: 182, Rouanet 1987: 163) défini par: T^2 = Phi-deux $/\sqrt{(l - 1)(c - 1)}$ c'est à dire par un Phi-deux maximum défini comme étant égal à la racine carrée du degré de liberté du tableau. On voit facilement la logique de ce choix: si le tableau est carré, la racine carrée du degré de liberté est égale à la plus petite dimension diminuée de l'unité (k - 1), c'est à dire au maximum théorique du Phi-deux. T^2 est identique à V dans ce cas et s'en éloigne d'autant plus que le nombre de lignes est différent de celui des colonnes comme le montre le tableau suivant qui donne les valeurs du dénominateur de T^2 pour quelques valeurs du nombre de lignes et du nombre de colonnes:

	nombre de colonnes \	2	3	4	5	6	7
nombre	2	1	.71	.58	.50	.45	.41
de	3		1	.82	.71	.63	.58
lignes	4			1	.87	.77	.71
	5				1	.89	.82
	6					1	.91
	7						1

Pour l'exemple considéré, T^2 est égal à 0,2619 / 18 = 0,062 ce qui signifie que le Phi-deux observé représente 6,2% du Phi-deux maximum réaliste au vu du format du tableau.

Récapitulons pour l'exemple considéré (format 4 x 7, Phi-deux observé de 0,2619):

Coef. de Cramér	V = 0,087 soit en pourcentage	8,7% du max.
Pearson	C^2 = 0,208	20,8%
Tchuprov	T^2 = 0,062	6,2%

Pour un format donné on remarque que le rapport de T^2 à son maximum théorique redonne le V de Cramér. Pour l'exemple considéré:

$$T^2/T^2max = 0,062 / 0,71 = 0,087 = V.$$

On remarquera que l'on compare V à C^2 et à T^2 car dans les trois cas on a le même Phi-deux en numérateur et seuls les dénominateurs changent. Pour simplifier les notations et faciliter les comparaisons on peut évidemment définir le coefficient de Cramér comme V^2 = Phi-deux / k - 1 (par ex. Rouanet 1987: 163) mais on va délibérément contre la définition de l'auteur et je m'en tiendrai à celle-ci.

EXAMEN DU COEFFICIENT GT DE FREDERIC MICHEL

Si l'on revient maintenant à l'examen du coefficient GT proposé, on s'aperçoit qu'il correspond strictement au coefficient T de Tchuprov. Du point de vue des exigences de l'auteur, il est peu satisfaisant puisqu'en général il n'est pas borné par 1. Par ailleurs, il n'est pas très optimiste si l'on fait les comparaisons correctes entre T^2, C^2 et V (ou entre leurs racines carrées).

Pour améliorer le résultat, Frédéric Michel reprend la logique de l'analyse des données et propose, si les données ne sont pas de nature ordonnée, de les réorganiser sous forme "scalogrammatique", c'est à dire en classant par poids marginal décroissant et de chercher ensuite la liaison maximale correspondant à une diagonale. On aurait ainsi une référence plus réaliste pour le Phi-deux (ou le Khi-deux) maximum.

Cette procédure est injustifiable pour le sociologue dans la mesure où l'ordre proposé doit être cohérent avec les données. Cette cohérence n'a aucune raison de s'en tenir à la seule nature ordonnée ou non, des modalités des lignes ou des colonnes. Ce n'est pas parce que les modalités sont "nominales" que tout ordre est indifférent et tout voisinage licite: il faut respecter les données telles qu'on les connait comme sociologue et ne pas les traiter avec une technique qui autoriserait toute licence à partir du moment où les donnée ne sont pas réputées "ordonnées".

Par exemple il est tout à fait déraisonnable dans le tableau 3 de l'article de Frédéric Michel, de faire cohabiter la colonne d qui correspond au vote pour la CGT avec la colonne g qui correspond au refus de vote, tout simplement du fait qu'elles sont en correspondance avec les lignes caractérisées par la confiance dans les syndicats, cela sous prétexte que ce sont les deux plus forts effectifs de colonne. Indépendamment de toute connaissance sur le sujet (ce qui n'est pas réaliste chez un sociologue), les données du tableau nous disent exactement le contraire pour les non votants dont on observe dans le tableau étudié qu'ils choisissent plutôt la réponse de non-confiance dans l'action des syndicats, ce qui va à l'encontre du regroupement proposé.

Si l'on décide de faire confiance aux données, il faut être cohérent avec cette attitude et laisser les données nous dire dans quel ordre doivent être classées lignes et colonnes afin que cet ordre nous suggère quelle pourrait être la liaison maximale. A cette fin j'avais proposé (Cibois 1984: 48, 97) de prendre l'ordre des lignes et des colonnes donné par le premier facteur d'une analyse des correspondances du tableau, dont il a été montré qu'il mettait en avant cette structure d'ordre si elle existait (Benzécri 1973: T.I 278).

On peut dans certains cas discuter pour savoir si le meilleur ordre sera donné par une analyse des correspondances ou par une analyse en composantes principales sur les écarts à l'indépendance car l'AFC met en relief les liaisons entre modalités à faible effectif tandis que l'ACP met en relief les liaisons entre modalités à plus fort effectif (effet de "distinction" en AFC discuté dans Cibois 1992).

Si l'on applique une telle technique de recherche de la liaison maximale aux données syndicales, on construit un Phi-deux maximum de 1.9506 qui mis en rapport avec le Phi-deux observé, donne un résultat de 0,134 soit un Pourcentage du Khi-deux Maximum ou PKM de 13,4%. Ce Phi-deux maximum est non seulement plus réaliste que le maximum théorique de 3 mais il est aussi cohérent avec l'information apportée par les données elles-mêmes. On notera que l'ordre du premier facteur d'une AFC respecte l'ordre antérieur des lignes et qu'il classe les votes en regroupant d'abord les syndicats les plus revendicatifs (CGT, CFDT), puis les syndicats modérés (CFTC, FO, Autonomes), puis les modalités de refus (vote pour des non- syndiqués ou non-votant).

Ajoutons donc ce nouveau coefficient à la liste précédente et comparons la liste ordonnée des résultats:

Coefficient de	Tchuprov en pourcentage	6,2% du maximum
	Cramér	8,7%
	PKM	13,4%
	Pearson	20,8%

Il est possible de "faire mieux" en conservant la logique du PKM mais en substituant au Phi-deux un autre indicateur.

L'ECART A L'INDEPENDANCE

On peut mettre en cause le choix du Khi-deux ou du Phi-deux comme numérateur de l'indice. En effet la contribution au Khi-deux d'une case peut être considérée comme l'écart à l'indépendance multiplié par un coefficient sans dimension (rapport écart sur théorique) qui pondère cet écart selon qu'il était attendu ou non. Si l'écart est inférieur au théorique, la case apporte peu d'information, si l'écart est supérieur au théorique, elle en apporte beaucoup. Le Khi-deux global de ce fait est un indicateur de l'information apportée par les écarts du tableau. C'est bien sûr quelque chose d'intéressant de savoir si des écarts apportent peu de surprise où s'ils sont totalement inattendus, mais cela ne doit pas nous faire oublier les écarts eux-mêmes, la structure du croisement qui fait qu'il y a des attractions, des répulsions et des situations d'indépendance entre lignes et colonnes du tableau.

Le terme "information" a ici un double sens car il désigne à la fois le contenu de l'information et sa valeur informationnelle au sens de la théorie de l'information. Dans un tableau croisé, le premier

sens est relatif aux écarts à l'indépendance, le deuxième à la contribution au Khi-deux de ces écarts. De même qu'un fait divers peut avoir une plus ou moins grande valeur pour un journaliste (il s'intéresse à sa valeur informationnelle), ce même événement est une donnée qui doit être étudiée dans sa structure, ses relations à son environnement, ses causes.

De la même façon, on peut se servir de l'information apportée par les écarts pour mesurer la force d'une liaison. Une solution simple consiste à cumuler les écarts positifs à l'indépendance, et à les mettre en rapport avec la situation de liaison maximale telle qu'elle a été définie plus haut. On peut définir ainsi un PEM (déjà présenté dans Cibois 1993) qui est le Pourcentage de l'Ecart Maximum.

En reprenant l'exemple précédent on fait la somme de tous les écarts à l'indépendance positifs et on arrive au total de 176,3. Ensuite on se place dans le cas où la liaison serait maximale en réordonnant lignes et colonnes selon l'ordre du premier facteur d'une analyse des correspondances du tableau. On charge au maximum une diagonale et on refait la somme des écarts positifs dans ce cas. La somme des écarts positifs est alors de 464,5. Le rapport 176,3 / 464,5 est de 0,379 ce qui signifie que les écarts observés représentent 37,9% du maximum observable. Ce pourcentage, plus fort que les précédents semble à l'expérience réaliste car on s'aperçoit que des valeurs plus fortes du PEM correspondent souvent à des cas de redondance où la question en ligne et la question en colonne correspondent à une même dimension. Inversement l'expérience montre qu'un PEM est souvent intéressant alors qu'il tourne autour de 10%

L'INFORMATION CONTENUE DANS UN TABLEAU

A la fin de son article Frédéric Michel se pose la question de l'information contenue dans le tableau relatif à l'indépendance et veut nous mettre en garde contre la tentation de le considérer comme d'information nulle puisqu'il n'apporte pas de Khi-deux. Pour cela il compare la situation de l'analyse des correspondances avec celle de l'analyse en composantes principales où, la première approximation toute positive des données d'origine contribue fortement à l'inertie comme le manifeste l'exemple numérique donné par l'auteur.

Il convient de distinguer une nouvelle fois les différents sens du mot information. En analyse des correspondances, la trace correspond au Phi-deux et l'information est entendue au sens de

valeur de surprise ou non apportée par un écart à l'indépendance. S'il n'y a pas d'écart, il n'y a pas d'information. En analyse en composantes principales sur des données brutes, l'indicateur d'information est la somme des carrés, ce qui signifie que tout effectif est informatif et il l'est d'autant plus qu'il est important (puisque le carré est pris en compte).

Quel est le statut, dans chacun des deux cas, de la première approximation en nombres positifs du tableau? Dans le cas de l'analyse des correspondances, si ce tableau n'apporte pas de Phi-deux il apporte cependant des renseignements, une information au sens banal du terme puisqu'il nous décrit une structure de croisement hypothétique s'il n'y avait pas de relation entre lignes et colonnes. Cette structure est entièrement dépendante des marges du tableau et ne reflète donc que l'information contenue dans la distribution marginale. En quelque sorte le tableau d'indépendance ne fait que nous donner l'information contenue isolément dans chacune des distributions des deux questions que l'on croise. Distributions que l'on connait antérieurement au croisement et qui ont déjà pu être examinées et commentées.

Ce niveau zéro de la relation entre lignes et colonnes est déjà à lui-même une information comme peut l'être le niveau de la mer: dans un port à marée, le niveau moyen est bien sûr une information mais ce qui intéresse chacun c'est de savoir si la mer est haute ou si elle est basse.

De la même façon, quand on croise deux questions, on suppose que l'on a examiné au préalable la distribution de chacune des questions et que l'on s'est penché sur les fréquences relatives de tel ou tel item. Quand le tableau d'indépendance se présente à nos yeux, il nous sert de repère pour juger des attractions ou répulsions entre lignes et colonnes mais en lui-même il ne nous dit rien de plus que nous ne savions déjà.

En composantes principales la situation est toute différente puisque tout effectif contribue à la somme des carrés, y compris les effectifs qui ne font que correspondre à la situation d'indépendance. Comme les écarts à l'indépendance sont toujours faibles par rapport à la situation d'indépendance, il est bien normal que le premier facteur (analogue au tableau d'indépendance et souvent proche de lui) apporte une contribution gigantesque par rapport aux facteurs suivants. C'est comme si nous mesurions la marée par rapport au centre de la terre et non par rapport au niveau moyen. Dire que la marée est au niveau 0 ou au niveau 6.366.198 ne doit pas nous faire dire qu'elle est faible dans un cas et forte dans l'autre: ce n'est que le point de repère qui change. Une amplitude de marée de

quelques mètres passera inaperçue dans un repère et non dans l'autre alors qu'il s'agit du même phénomène.

Entre Phi-deux et somme des carrés, il y a une différence de repère d'origine, il y a aussi une différence de pondération et on peut discuter du problème de savoir quel est l'indice le plus approprié pour mesurer des écarts à l'indépendance. On peut en effet s'affranchir du problème d'origine en décidant que le tableau dont on doit faire l'analyse factorielle est le tableau des écarts à l'indépendance quelle que soit le type d'analyse. Comme il s'agit d'une tableau centré et non réduit, l'analyse en composantes principales aura l'avantage par rapport à l'analyse des correspondances de donner des facteurs qui mettront en relation les attractions entre modalités à fort effectif (alors que les correspondances mettent en avant les "minorités actives"). On trouvera une comparaison des deux méthodes dans Cibois 1992 (indépendamment de la comparaison sur des données qui donnent le même résultat comme dans Cibois 1983: 120).

En résumé, il n'y a qu'une information pertinente, c'est la structure des écarts à l'indépendance, étant bien entendu que la situation d'indépendance est le niveau zéro du croisement. Il reste qu'antérieurement au croisement, la distribution de chaque question est une information tout à fait pertinente qui doit être examinée au préalable. Quant aux indicateurs numériques de cette information, il ne faut pas les fétichiser: le phénomène social que l'on étudie n'est ni un Khi-deux, ni une somme de carrés, c'est une attraction ou une répulsion entre un comportement et une croyance; ou encore des cohérences ou des incohérences entre croyances. C'est le conducteur d'un véhicule qui juge s'il va trop vite ou non, non son compteur de vitesse. Ceci vaut en particulier pour les taux d'explication en analyse factorielle d'un tableau de Burt où le fétichisme du compteur fait oublier au chercheur que c'est lui qui juge de la cohérence de l'interprétation.

En conclusion on ne peut qu'être d'accord avec l'auteur dans sa volonté de ne pas se laisser emprisonner par les techniques existantes et dans son désir d'en créer de nouvelles plus adaptées. Son intuition et ses résultats recoupent sur ce point des travaux antérieurs. J'ajouterai simplement que le respect profond des données doit être notre guide, autant pour leur structuration que pour leur interprétation, qui ne sont d'ailleurs qu'une seule et même démarche. L'analyse des données permet ainsi de coupler la démarche statistique et la démarche interprétative en proposant des méthodes qui tirent leur cohérence des données elles-mêmes.

BIBLIOGRAPHIE

Benzécri, J.P. *et al.* (1973). *l'Analyse des données*, Paris: Dunod (2 tomes).

Cibois, Ph. (1983). *l'Analyse factorielle*, Paris: Presses universitaires de France.

Cibois, Ph. (1984). *l'Analyse des données en sociologie*, Paris: Presses universitaires de France.

Cibois, Ph. (1992). Soixante styles de loisirs et de pratiques culturelles, Paris: rapport Cersof.

Cibois, Ph. (1993). le PEM, Pourcentage de l'Ecart Maximum: un indice de liaison entre modalités d'un tableau de contingence, *Bulletin de Méthodologie Sociologique*, 40.

Conover, W.J. (1971). *Practical Nonparametric Statistics*, New York: John Wiley & Sons.

Cramér, H. (1946). *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

Rouanet, H., Le Roux, B., Bert, M.-Cl. (1987). *Statistiques en sciences humaines: procédures naturelles*, Paris: Dunod.

Yule, G.U., Kendall, M.G. (1940). *An Introduction to the Theory of Statistics*, Londres.

==========