An Order on Cross-Tabulations and Degrees of Association

Philippe Cibois
(Printemps, University of Versailles - St. Quentin)

BMS - Bulletin de Méthodologie Sociologique 2013, n°119, p. 24–43

Résumé.
La structure d'ordre dans un tableau croisé : degré d'association. Il est souvent affirmé qu'une structure d'ordre existe dans un tableau croisé quand les marges du tableau disposent d'une telle structure. On peut s'affranchir de ce point de vue et définir précisément une structure d'ordre du tableau lui-même. Comme l'avait déjà remarqué Louis Guttman dans le cas de la *scalogram analysis,* il faut souvent des déplacements alternatifs des lignes et des colonnes pour parvenir à repérer une échelle. Dans ce cas, c'est bien l'ordre du tableau structuré qui induit un ordre sur les marges et non l'inverse. Cependant Goodman et Kruskal au moment où ils présentent l'indice *gamma* qui permet de définir l'intensité d'une liaison dans le cas ordonné, n'utilisent que l'ordre des marges, et ils ont été suivis depuis. Il convient de revenir à l'intuition de Guttman et d'utiliser des résultats obtenus ultérieurement pour montrer qu'au moins une approximation d'une structure d'ordre est pratiquement toujours présente sur un tableau. Le tableau croisé issu de questions ordonnées n'est qu'un cas parmi d'autres et inversement un tableau disposant d'une liaison ordonnée forte induit un ordre interprétable sur les modalités des questions. En partant d'exemples réels on montrera que l'on dispose de critères pour définir un ordre sur un tableau, de méthodes formalisées pour rendre apparente la structure associée, de différents indices pour mesurer l'intensité de la liaison, de tests pour en évaluer le degré de signification.

Abstract.
It is often argued that an order exists in a cross-tabulation when the table's margins have such a structure. We can free ourselves from this point of view and clearly define an order on the table itself. As Louis Guttman noted previously in the case of *scalogram analysis*, one must often move rows and columns about to be able to create a scale. In this case, it is the order of the table's structure which induces an order on the margins and not the reverse. However, Goodman and Kruskal, when they proposed the *gamma* index that defines the strength of an association in the ordered case, only use the margins' order, and they have since then been followed by most researchers. One should return to the original intuition of Guttman and show that at least an approximate order is almost always present in a table. The ordered cross-tabulation generated by ordered questions is only one case among many others and conversely a table with a strong order structure induces an order on question modalities. With real examples, we show that the criteria are available to define an order on a table, that there are formalized methods to reveal the associated structure, that there are also different indices to measure the degree of association, and finally that there are tests to assess the level of significance.

Mots clefs
Tableau croisé, structure d'ordre, indice gamma, indice PEM

Keywords
Cross-tabulations, Order Structures, Gamma Index, PEM Index

Measure the degree of association between rows and columns in a cross-tabulation is an issue that has been discussed for more than a century. If we retain only the methods still used, we find Karl Pearson's contingency coefficient (1904), Tschuprow's coefficient (1925) and Cramér's coefficient (1946). This issue was addressed in a series of four articles in the *Journal of the American Statistical Association* by Goodman and Kruskal (1954; 1959; 1963; 1972) who then assembled the articles together it in a book (1979). On the other hand, there is a question that has engendered little

research: if there is an order structure on the rows and columns, then does the resulting table have a particular structure that can be identified? This question is addressed in this article's first section, before the question of the degree of association. The result we obtain is a unification of all types of cross-tabulations, ordered or not, in a single type of table where cross-tabulations differ only by the intensity, whether significant or not, of the order we found.[1]

Order Structure of an Ordered Margins Table

Louis Guttman in the fourth volume of *The American Soldier* (1950) laid the foundations of *scalogram analysis* with which he ordered a table crossing answers and individuals, seeking by alternative shifts between rows and columns to achieve a homogeneous form which he deduces an order he called a scale. It is the ordered table that induces an order on the margins and not the inverse. This technique was further developed by Bertin (1967). Before returning to this idea, we show through examples how the problem arises.

Let us suppose we have a 2 x 2 table A_iB_j where margins A_i and B_j have a defined order structure as follows $A_1 > A_2$ and $B_1 > B_2$. Consider the following example:

Table2X2	B_1	B_2	Totals
A_1			140
A_2			60
Totals	80	120	200

The margins are ordered, but the table itself does not necessarily have an order structure, as in the following case:

Table2X2	B_1	B_2	Totals
A_1	56	84	140
A_2	24	36	60
Totals	80	120	200

Indeed, for the first cell A_1B_1, we see that the product margins divided by the total equals 56, the expected frequency. We are therefore in the case of independence between rows and columns, yet ordered.

To change this situation, you can either add or subtract an individual in the cell A_1B_1. If added, it is assumed that there is an attraction between A_1 and B_1, and therefore, since this is a fixed margins framework, an opposition between A_1 and B_2 and between A_2 and B_1 and finally another attraction between A_2 and B_2. The elementary displacement is the following:

	B_1	B_2
A_1	+1	-1
A_2	-1	+1

The table resulting from this change is as follows:

[1] This research should be considered a tribute to George T. Guilbaud (1912-2008) whose lectures in his seminary in the 1970s are at the origin of the methods presented here.

	B_1	B_2	Totals
A_1	57	83	140
A_2	23	37	60
Totals	80	120	200

This table has an order structure that is defined by the structure of the signs of deviations from independence. Before considering a definition of the order structure, we can say that a 2 x 2 table is ordered when one of the diagonals has positive deviations and the other negative ones.

We can repeat elementary changes several times, but a maximum is reached when the cell A_1B_1 is 80 because of the margin constraint (and the A_2B_1 cell is thus equal to zero):

	B_1	B_2	Totals
A_1	80	60	140
A_2	0	60	60
Totals	80	120	200

One has thus made 24 elementary changes or shifts that corresponded to a progressive decrease in the A_2B_1 cell from 24 to 0.

The elementary shifts with reverse signs leads to a table that has the reverse order for A and B:

	B_1	B_2
A_1	-1	+1
A_2	+1	-1

And the maximum association in opposite direction of the association between A and B is:

	B_1	B_2	Totals
A_1	20	120	140
A_2	60	0	60
Totals	80	120	200

In this case, 36 elementary shifts were necessary. There is a total of 60 shifts to which must be added the case of independence; that is to say, 61 possible situations (which corresponds to the smallest margins + one unit). As one cell defines the entire table for one single degree of freedom, we can summarize all possible cases as follows, taking as reference A_2B_2:

60 – 59 – 58 – 57 – 56	37 – **36** – 35	4 – 3 – 2 – 1 – **0**
Maximum association	independence	Max. inverse association

If A_2B_2 is between 37 and 60, the table is ordered in the direction of A and B, if A_2B_2 is between 35 and 0, the table is ordered in the opposite direction of A and B. All of these tables (except for independence) have an order structure defined by the two elementary displacements and their possible repetitions.

As the situation of independence plays a central role, we can, by subtracting the value of independence from the previous scale, consider the scale of deviations from independence, always for the A_2B_2 cell:

+24 +23 +22 +21 +20	+1 **0** -1	-32 -33 -34 -35 **-36**
Max. association	independence	Max. . inverse association

Consider for example the table of deviations for $A_2B_2 = 50$:

	B_1	B_2
A_1	+14	-14
A_2	-14	+14

And the table for $A_2B_2 = 30$:

	B_1	B_2
A_1	-6	+6
A_2	+6	-6

This scale also provides us with an index of intensity of the link or association by indicating a deviation from the maximum. For the cell $A_2B_2 = 50$, the difference is 14, compared to a maximum of 24, and is therefore 14/24 x 100 = 58.3 percent of the maximum, an index which will be called percentage of maximum deviation from independence or, in French, the PEM for *Pourcentage de l'Écart Maximum* (Cibois, 1993). For the cell $A_2B_2 = 30$, the maximum deviation is in the negative direction, -36, the difference is -6 which is -6 / -36 x 100 = 16.7 percent of the maximum, which is by convention given a negative sign to indicate that it is a negative deviation.

As we have shown (Cibois 1993), it is possible to extend this procedure of looking for a PEM to each of the table's cells. All that is needed is to isolate the cell for which we want to know the intensity of association and reorganize all the other lines in a single line and all the other columns in a single column, all of which comes back to the 2 x 2 table.

In conclusion, as soon as a 2 x 2 table in not a situation of independence, it always has an order structure identifiable by the existing margin order. Since in general, the situation of independence rarely occurs with observed data, one can say that the order structure is practically the general case.

A Real Example - London 1911

We will now work on a real table from Kendal and Stuart (1961: 558), showing the results of a survey made in London in 1911 (London noted 4x6[2]). The table shows the distribution of 1,725 school children who were classified (1) in rows according to their standard of clothing (Very well clad, Well clad, Poor but passable, Very badly clad), and (2) in columns according to their intelligence (Very able, Distinctly capable, Fairly intelligent, Slow but intelligent, Dull, Mentally deficient or slow and dull), respectively:

London46	VABL	DCAP	FINT	SLBI	DULL	DEFI	Totals
VWEL	39	194	209	113	48	33	636
WELL	15	138	255	202	100	41	751
POOR	4	33	61	70	58	39	265
VBAD	1	10	10	22	13	17	73
Totals	59	375	535	407	219	130	1725

[2] *London* in original format with 4 rows and 6 columns

The following table is 3 rows and 3 columns (noted London 3x3) obtained by combining the columns in pairs and lines 3 and 4:

London33	Intel+	Intel=	Intel-	Totals
Clad+	233	322	81	636
Clad=	153	457	141	751
Clad-	48	163	127	338
Totals	434	942	349	1725

London 3x3 can be decomposed into the sum of two tables corresponding to independence and the deviations from independence:

Independence	Intel+	Intel=	Intel-	Totals
Clad+	160,0	347,3	128,7	636
Clad=	188,9	410,1	151,9	751
Clad-	85,0	184,6	68,4	338
Totals	434	942	349	1725

Deviations	Intel+	Intel=	Intel-
Clad+	**73,0**	-25,3	-47,7
Clad=	-35,9	**46,9**	-10,9
Clad-	-37,0	-21,6	**58,6**

Around the first diagonal where the differences are all positive (in bold), all differences are negative. However, the notion of a diagonal must be specified if the number of rows and columns are not equal, and even in the event of equality when the margin structure has distorting effects.

Number of Rows and Columns Are Different

Let us form a new table (London 2x3) where the columns are grouped as above and 2-4 lines are grouped. We have the following decomposition:

London23	Intel+	Intel=	Intel-	Totals
CladSup	233	322	81	636
CladInf	201	620	268	1089
Totals	434	942	349	1725

Independence	Intel+	Intel=	Intel-	Totals
CladSup	160	347,3	128,7	636
CladInf	274	594,7	220,3	1089
Total	434	942	349	1725

Deviations	Intel+	Intel=	Intel-
CladSup	**73**	-25,3	-47,7
CladInf Inf	-73	**25,3**	**47,7**

We see that in column 2, the positive deviation is in the second row.

Constrained Margins

A new table, London 3 x 3 (London33B), is made, keeping the same column grouping but by making less balanced lines in the margins. It includes lines 1 and 2 (now CladA) and left the two remaining lines identical (POOR becomes CladB and VBAD becomes CladC). We have the following decomposition:

London33B	Intel+	Intel=	Intel-	Totals
CladA	386	779	222	1387
CladB	37	131	97	265
CladC	11	32	30	73
Totals	434	942	349	1725

Independence	Intel+	Intel=	Intel-	Totals
CladA	349,0	757,4	280,6	1387
CladB	66,7	144,7	53,6	265
CladC	18,4	39,9	14,8	73
Totals	434	942	349	1725

Deviations	Intel+	Intel=	Intel-
CladA	**37,0**	**21,6**	-58,6
CladB	-29,7	-13,7	**43,4**
CladC	-7,4	-7,9	**15,2**

We see this time that the effect of the diagonal is still present, but it has been deformed (positive differences in bold). The high weight of the margin CladA pulled the diagonal of positive deviations to the right and upward.[2] So, for reasons of size or for reasons of margin constraints, only the extreme diagonal cells (for a diagonal following the margin order) have always positive deviations. To go from one end to the other, the path of positive deviations may deviate more or less from the diagonal; positive deviations are always contiguous (laterally) or adjacent (diagonally). It is the existence of this "ridge" – where there are the positive deviations isolating all the negative differences – which will be the definition of a table with an order.

Definition: a table has an order when the diagonal, which connects (by lateral contiguity or diagonal adjacency) the extreme cells defined by the margin order, has positive deviations from independence. Negative deviations are on both sides of the diagonal.

Finally, let us decompose the original table of London 1911 whose deviations from independence are:

London46	VABL	DCAP	FINT	SLBI	DULL	DEFI
VWEL	**17,2**	**55,7**	**11,7**	-37,1	-32,7	-14,9
WELL	-10,7	-25,3	**22,1**	**24,8**	**4,7**	-15,6
POOR	-5,1	-24,6	-21,2	**7,5**	**24,4**	**19,0**
VBAD	-1,5	-5,9	-12,6	**4,8**	**3,7**	**11,5**

A "ridge" runs clearly from both ends of the diagonal, and it is more or less wide. All positive deviations situated on this line are contiguous and/or adjacent; all the negative deviations are located on either side of the ridge.

Reciprocal Situation

Let us now look at the problem initiated by Guttman and ask the reverse question: if we find an order structure in a table, what does this imply for its rows and columns? Take for example the following table: it is a table from a survey of political and union opinions of French workers in 1970 (Adam, 1970) from which we extract a table of confidence in unions depending on the union chosen during voting on the job.

The table rows are ordered respectively ("To defends your interests in labor disputes, you're Very confident in them, Somewhat confident, Not confident, Not confident at all"), but the columns are responses to the question "in case of union elections in your firm, would you prefer to vote for a list led by FO ("Force Ouvrière"), CFDT ("Confédération Française Démocratique du Travail"), non-unionized workers, CGT ("Confédération Générale du Travail"), an autonomous or independent union, CFTC ("Confédération Française des Travailleurs Chrétiens"), you not vote at all?" Here is the observed table and the table of deviations from independence:

Confidence in unions	FO	CFDT	Non-union	CGT	Auto	CFTC	Non-Vote	Totals
Very confident	14	24	12	137	11	4	6	208
Somewhat confident	38	43	22	137	40	12	45	337
Not very confident	15	7	19	25	25	4	34	129
Not confident at all	11	13	38	18	25	3	62	170
Total	78	87	91	317	101	23	147	844

Deviations	FO	CFDT	Non-union	CGT	Auto	CFTC	Non-Vote
Very confident	-5,2	**2,6**	-10,4	**58,9**	-13,9	-1,7	-30,2
Somewhat confident	**6,9**	**8,3**	-14,3	**10,4**	-0,3	**2,8**	-13,7
Not very confident	**3,1**	-6,3	**5,1**	-23,5	**9,6**	**0,5**	**11,5**
Not confident at all	-4,7	-4,5	**19,7**	-45,9	**4,7**	-1,6	**32,4**

Graphically highlighting the positive deviations from independence, one can note a similarity of profiles between CGT and CFDT (positive differences for high degrees of confidence), between FO and CFTC (positive differences for intermediate degrees), and between Autonomous, non-unionized and non-voters (positive differences for the lowest degrees of confidence). We can reorder the table so as to find the order structure previously defined where positive deviations partition the table around the first diagonal, the negative differences being on either side:

Deviations	CGT	CFDT	CFTC	FO	Auto	Non-union	Non-Vote
Very confident	**58,9**	**2,6**	-1,7	-5,2	-13,9	-10,4	-30,2
Somewhat confident	**10,4**	**8,3**	**2,8**	**6,9**	-0,3	-14,3	-13,7
Not very confident	-23,5	-6,3	**0,5**	**3,1**	**9,6**	**5,1**	**11,5**
Not confident at all	-45,9	-4,5	-1,6	-4,7	**4,7**	**19,7**	**32,4**

It remains to define the order of the columns: as in France, the two unions CGT and CFDT are the protest unions while CFTC and FO positions are less radical and independent unions are most often unions created by employers and used to oppose union protests, we can reinterpret the question based on the responses obtained. The order on the unions shows the degree of opposition to the established order (Cibois, 1984: 20-21).

Searching for an Order Structure

The previous problem was particularly simple since there was already an order structure on the rows, and it was enough to make a few permutations on the columns to reset the order structure of the table. To address the generalized problem, we will use the technique of correspondence analysis since Benzécri (1976: 279-80) shows that if there is an order structure on the rows and columns, the first factor of a correspondence factor analysis manifests that order. We can verify it with the table above in the following bi-plot (Figure 1) with the first factor as the horizontal axis and second factor as the vertical axis:

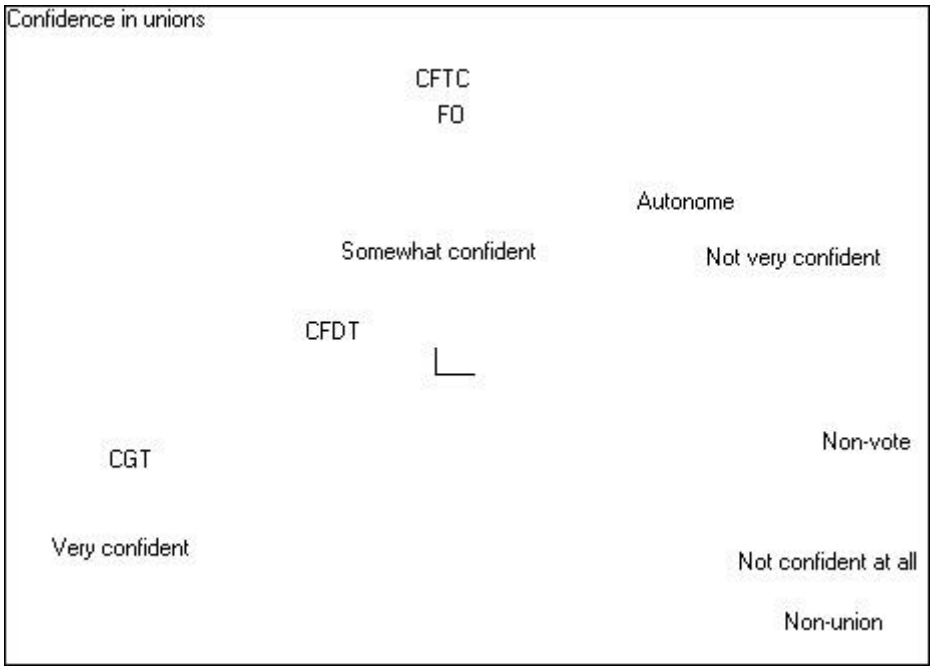


Figure 1. French workers' opinions on unions with the first factor as the horizontal axis and the second factor as the vertical axis

We can complete this graphic with the representation of the intensity of ties by computing all PEM positive cells and then connecting the dots with a line whose thickness corresponds to the strength of the PEM (Figure 2).

With this example, we can specify the procedure for calculating the PEM for a cell (such a PEM is called “local”). We seek the degree of attraction between the row "very confident" and the CGT union. We reduced the table to a 2 x 2 table in which one can operate as before:

Table 2 x 2	CGT	Other columns	Totals
Very confident	137	71	208
Other lines	180	456	636
Totals	317	527	844

Observed deviation from independence is 37 – (208 x 318 / 844) = 58.9. Deviation from independence in the maximum case is 208 - (208 x 318 / 844) = 129.9. Local PEM is 58.9 / 129.9 x 100 = 45.3 percent. We proceed in this way for each cell of the table:

PEM	CGT	CFDT	CFTC	FO	Auto	Non-Union	Non-Vote
Very confident	45,3	3,9	-29,4	-27,2	-55,8	-46,5	-83,4
Somewhat confident	5,5	15,8	20,4	14,6	-0,8	-39,5	-23,3
Not very confident	-48,4	-47,4	2,5	4,7	11,2	6,6	10,8
Not confident at all	-71,8	-25,8	-35,2	-30	5,8	27,1	27,6

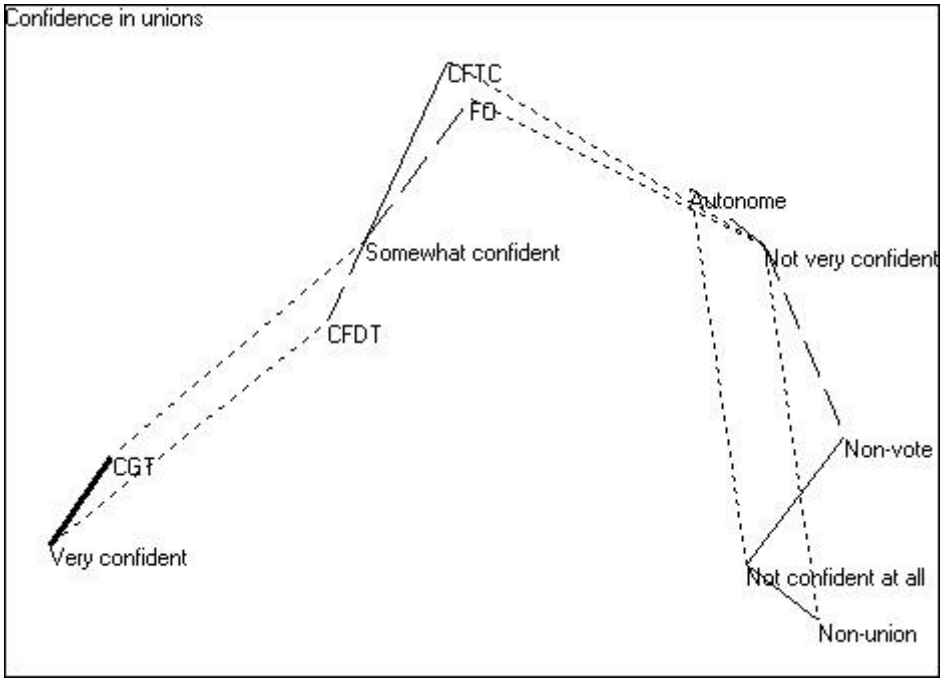


Figure 2. Visualization of the strength of ties for local PEMs in the factorial plan

The order structure of the table indeed follows the correspondence analysis first factor (horizontal axis).

We now have a procedure for finding an order structure for any cross-tabulation. Now let us consider the degree of association.

Degree of Association

We are looking for an indicator giving us the degree of association between the order of rows and the order of columns. We start with the work of Goodman and Kruskal (1954) who took on the problem completely and proposed indices of association that were no longer based on the chi-square because "The fact that an excellent test of independence may be based on χ^2 does not at all mean that χ^2, or some simple function of it, is an appropriate *measure* of degree of association" (1954: 740). Then we criticize this index and propose a generalization of the PEM.

Goodman and Kruskal's Gamma

To present this indicator, we will reuse the London data first as the 2 x 2 table as follows:

London22	IntellSup	IntelInf	Totals
CladSup	850	537	1387
CladInf	119	219	338
Totals	969	756	1725

On such a 2 x 2 table, Yule (1900) had defined a coefficient of association using cross products (850 x 219 = 186,150 and 119 x 537 = 63,903); if they are equal, there is independence and association coefficient Q is the ratio of the sum of their difference:

$$Q = (186{,}150 - 63{,}903) / (186{,}150 + 63{,}903) = 122{,}247/250{,}053 = 0.489$$

Goodman and Kruskal return to this idea of using the cross products: they call *concordant pairs* the cross product of the first diagonal 850 x 219 (and the symmetric 219 x 850) which, when moving from one cell to another, the rank order rises for both rows and columns. Symmetrically, they call *discordant pairs* when the rank order increases for the lines, but decreases for columns (or *vice versa*). This is the case in the second diagonal where from 119 to 537, we are going from cell Clad-Inf – Intelligence-Sup to cell Clad-Sup – Intelligence-Inf: we go onward in the order of clothing, but downward in order of intelligence. It is therefore a case of discordant pairs. Formally, Goodman and Kruskal (1954: 749) define the cases (in proportion) as follows:

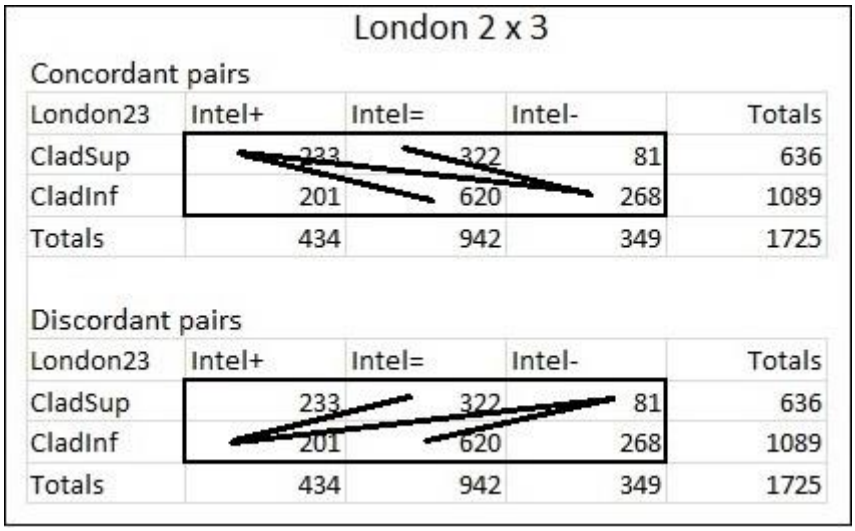
Π_s= Pr $\{a_1 < a_2$ and $b_1 < b_2$; or $a_1 > a_2$ and $b_1 > b_2\}$, same order
Π_d =Pr $\{ a_1 < a_2$ and $b_1 > b_2$; or $a_1 > a_2$ and $b_1 < b_2\}$, discordant order
Π_t = Pr $\{ a_1 = a_2$ or $b_1 = b_2\}$, ties are equal.

The case of equality here corresponds to the pairs 119-850, 119-219, etc., and pairs corresponding to the identity 119-119, etc. They are not taken into account in the calculation of Gamma.

Goodman and Kruskal's Gamma is defined as $\gamma = (\Pi s - \Pi d) / (\Pi s + \Pi d)$, which in the case of a 2 x 2 table corresponds to Yule's Q. But they generalize: to understand what happens, let's return to data of London 2 x 3.

London23	Intel+	Intel=	Intel-	Totals
CladSup	233	322	81	636
CladInf	201	620	268	1089
Totals	434	942	349	1725

If we take the pair of cells in opposition in the first diagonal, we see that if we start with 268, compared to 233, it goes in the order of rows and columns. But it is also the case for 268 to 322 and 620 to 233. Let us visualize these concordant and discordant pairs:



We calculate Gamma from product values of concordant pairs and discordant pairs, and we have:

233 x 620 =	144460
233 x 268 =	62444
322 x 268 =	86296
Concordant pairs =	293200

201 x 322 =	64722
201 x 81 =	16281
620 x 81 =	50220
Discordant pairs =	131223

Gamma : C-D / C + D =	0,382

The rationale for these calculations by Goodman and Kruskal is as follows: Suppose that two individuals are taken independently and at random from the population. Each falls into some (A_a, B_b) cell. (…) If there is high association one expects that the order of the *a*'s would generally be the same as that of the *b*'s.

Taking the products of the concordant pairs is equivalent of counting the pairs of individuals in situations of order, and making the products of the discordant pairs is equivalent of counting the pairs of individuals that are not in a position of order. The more the situation resembles that of the total order, the greater the association. Several coefficients use counts of the number of pairs: Kendall's Tau, Stuart's Tau-C, Somers' asymmetric D. When the rows and columns do not have an order structure, these techniques cannot be used and we observe that users often return to indicators derived from chi-square, despite of the criticism of Goodman and Kruskal.

The difficulty with this procedure is that the search for concordant and discordant pairs does not take account of the observed structure of the table and is based only on the order of rows and and columns, whereas an order structure may exist. We overcome this difficulty by ordering the

rows and columns with the first factor of a correspondence analysis, which always gives an order that we can use for developing an index of association derived from the PEM.

The Global PEM

The proposed general association coefficient is a measure of association between rows and columns and assumed that:

• If an order structure is known for the rows and columns, it can be observed empirically and conversely.
• If an order structure has not been identified, it may however exist, even if the order is not very pronounced.
• The coefficient can be used to determine the degree of the association for a table cell and for the entire table.
• As recommended by Goodman and Kruskal above, it will not use the chi-square.
• Its value will be zero in case of independence.
• It will vary between -1 and 1 from dependence in one direction to dependence in the other (the sign is conventional). Values close to the maximum must correspond to situations that occur empirically.
• The index values must be comparable from one table to another, even if they are different on the size of the populations or concerning the numbers of rows or columns.
• The principle of the coefficient should be simple to understand, even if it is the result of lengthy operations that cannot be done by hand in the elementary cases.

As the situation of independence is well defined and is still indicating no association, as a principle to measure the association, we consider (in the logic of local PEMs) the ratio of the sum of the positive deviations from independence observed, to the sum of the positive deviations in the case where the link would be at its maximum.

Let’s consider the table ordered by the first factor of the correspondence analysis of the survey on confidence in unions. Returning to the table of deviation from independence, we see that the sum of the positive deviations from independence is equal to 176.26.
We must now define the maximum. We have an ordered table of which we retain only the margins: by the fact that the table is ordered, the diagonal of the positive differences either starts from the attraction between CGT and "very confident", or from the cell Non-voting - "no confidence at all". The choice of starting point is irrelevant and leads in both cases to the same result. Let starts from the cell at the top left. All CGT, which are 317, cannot be "very confident in the unions” because the corresponding margin is only 208, but conversely all "very confident" can be put in the CGT cell. There remain 317 - 208 = 109 CGT that we will put in the adjoining cell (laterally) the nearest "somewhat confident". The table will be as follows:

Confidence in unions	CGT	CFDT	CFTC	FO	Auto	Non-Union	Non-Vote	Totals
Very confident	208							208
Somewhat confident	109							337
Not very confident								129
Not confident at all								170
Totals	317	87	23	78	101	91	147	844

All numbers in row 1 and column 1 are now distributed. In line 2, there is 337 (margin) - 109 (CGT) =~~r~~ 228 “reasonably confident”. They can be divided into CFDT (87), CFTC (23), FO (78); there are 40 that will be put in "autonomous". The entire second line is placed, and we go back to the column where there are still 101-40 = 61 autonomous to be placed which we will place in the adjoining "not very confident" cell. All autonomous are placed, but there are still the 129 - 61 = 68 "not very confident", which will be placed in non-unionized workers, whose 23 and all remaining non-voters will be “no confidence at all”. This gives the final table (which could be obtained with the same algorithm starting from the "Non-voting" - "Not at all confident" cell. The solution is unique and the algorithm is used in the program in the Annex.

Confidence in unions	CGT	CFDT	CFTC	FO	Auto	Non-Union	Non-Vote	Totals
Very confident	208							208
Somewhat confident	109	87	23	78	40			337
Not very confident					61	68		129
Not confident at all						23	147	170
Totals	317	87	23	78	101	91	147	844

We can easily verify that the sum of the positive differences from independence in the case of this maximum table is 464.53. The global PEM is the ratio of the two sums (positive differences observed, differences in the case of maximum), in percentage: 176.26 / 464.53 x 100 = 37.9 percent.

On real data from tables, as it is always possible to order the data according to the first factor of the correspondences analysis, we can say that there is always a structure of order and it is always possible to calculate a global PEM. This result may seem hazardous because in some cases, this order can be entirely due to a random structure of data that are not actually ordered. We are going to confront this situation in the following case where we know *a priori* that there is no order on the rows and columns.

Compatibility of Astrological Signs for Married People

We study a case where the order structure is absent and we submit it to the procedures for searching for an order structure. Below is a table that was constructed to show the meaninglessness of astrology (data presented in Cibois, 1997). For a population of 68,000 married couples, we construct a table of 12 rows and 12 columns, the rows corresponding to the astrological signs of the

men and the columns for those of the women. We note at the intersection of a row and a column, the number of couples for given signs.

	Women												
Men	Aqu	Pis	Ari	Tau	Gem	Can	Leo	Vir	Lib	Sco	Sag	Cap	Totals
H-Aquarius	536	478	518	535	532	500	451	478	478	413	430	502	5851
H-Pisces	482	592	536	541	525	506	484	463	503	475	443	482	6032
H-Aries	555	560	596	584	525	508	543	452	525	461	451	521	6281
H-Taurus	511	508	582	607	552	523	527	462	490	448	438	460	6108
H-Gemini	488	497	557	520	577	496	469	461	433	433	421	458	5810
H-Cancer	487	508	512	530	478	504	446	436	462	397	420	456	5636
H-Leo	456	502	522	482	478	461	466	431	455	440	402	472	5567
H-Virgo	445	463	489	500	426	464	413	457	409	381	395	434	5276
H-Libra	490	494	482	493	481	450	482	406	494	392	449	440	5553
H-Scorpio	441	437	459	483	464	433	426	382	434	392	432	401	5184
H-Sagittarius	455	445	475	436	456	423	411	395	443	377	419	435	5170
HCapricorn	498	496	445	554	456	461	443	398	469	411	398	494	5523
Totals	5844	5980	6173	6265	5950	5729	5561	5221	5595	5020	5098	5555	67991

If we do a correspondence factor analysis of this table, one could be disturbed by the factor graph (Figure 3) that highlights similarities between signs which are outlined below by ovals. Indeed, 10 out of 12 signs are nearby (the only clear exception being the sign of Taurus, Aquarius is less clear)

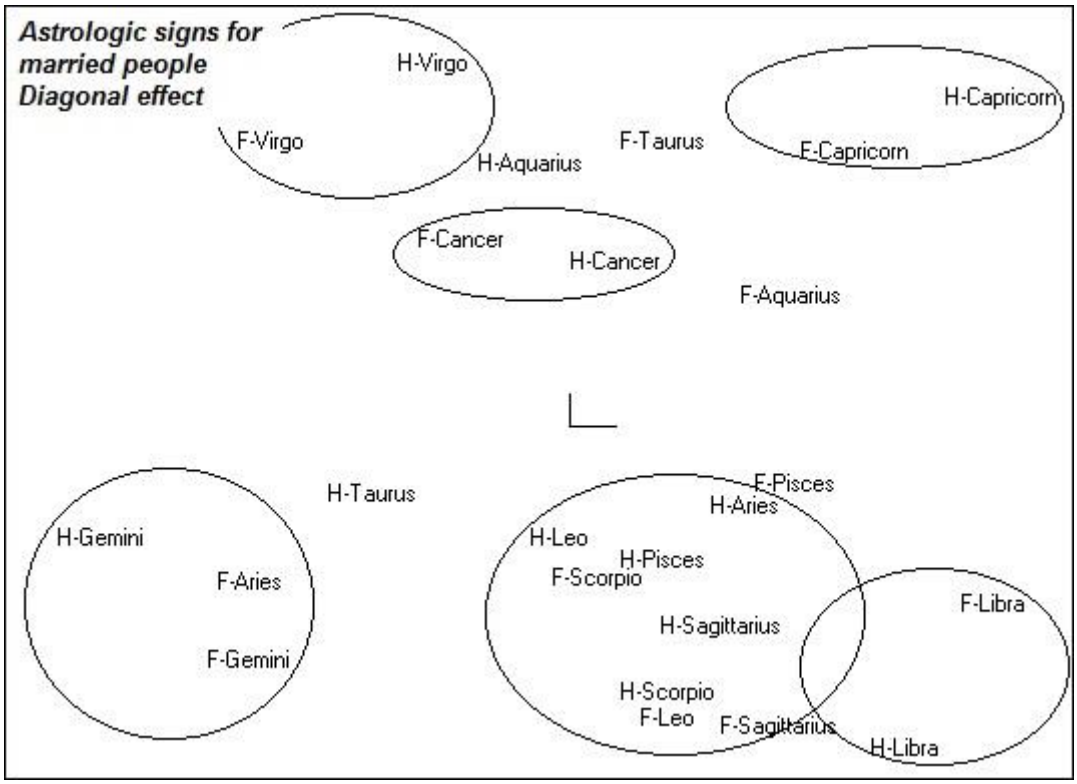


Figure 3. Astrologic signs for married people – Diagonal effect

This can be explained if we examine the deviations from independence: here we retained only the positive deviations greater than 9. We see that all deviations from the diagonal are positive, which explains previous proximities:

	Women											
Men	Aqu	Pis	Ari	Tau	Gem	Can	Leo	Vir	Lib	Sco	Sag	Cap
H-Aquarius	**33**				20			29				24
H-Pisces		**61**								30		
H-Aries	15		**26**				29					
H-Taurus			27	**44**	17		27					
H-Gemini			30		**69**			15				
H-Cancer		12		11		**29**						
H-Leo		12	17				**11**			29		17
H-Virgo			10	14		19		**52**				
H-Libra	13						28		**37**		33	
H-Scorpio					10					**9**	43	
H-Sagittarius	11								18		**31**	13
HCapricorn	23	10		45					15			**43**

However, if we look at all the individual PEMs, we see that these diagonal deviations are in the same order of magnitude as the others and they are also fewer:

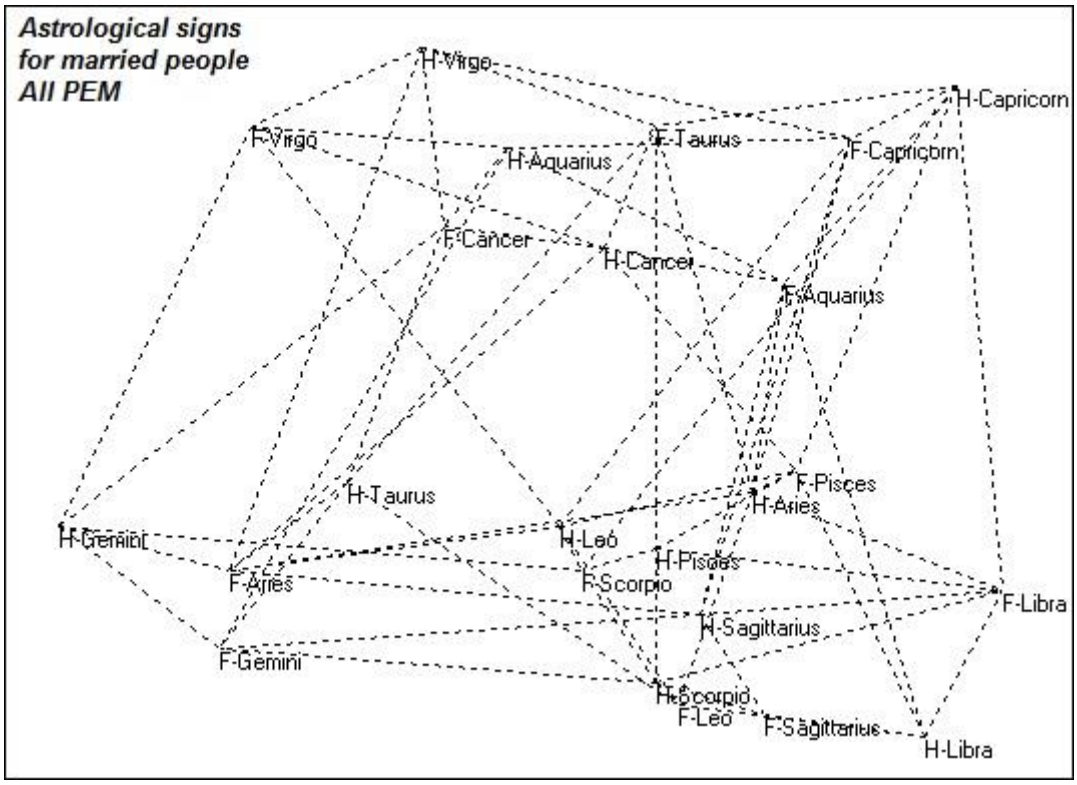


Figure 4. Astrological signs for married people – All PEMs

If the first factor (horizontal) of a correspondence analysis really offers an order, the order corresponds to a virtual absence of ties because the global PEM is equal to 2.0 percent. Another

clue that the table is very close to independence is provided by the first eigen value of the correspondence analysis which is very low and equal to 0.0006. As for the diagonal effect, it is explained by the fact that people who believe in astrology believe that people of the same sign attract each other. It is therefore a small but self-fulfilling and noticeable effect.

Before studying the problems of significance of the results, let us compare the different indices used for different examples discussed here:

	PEM	First eigen-value	Gamma	V Cramér	% Cramér	Chi-square	p=
London23	26,6%	0,049	0,382	0,220	4,85%	83,63	0,000
London33	20,4%	0,066	0,368	0,198	3,90%	134,69	0,000
London33b	21,7%	0,050	0,415	0,158	2,51%	86,62	0,000
London46	23,3%	0,079	0,332	0,184	3,38%	174,82	0,000
Unions-conf.	37,9%	0,238	0,525	0,295	8,73%	221,05	0,000
Astrologic Signs	2,0%	0,0006	0,017	0,014	0,02%	139,17	0,124

• The Gamma index has been calculated on the last two tables, assuming they had an order structure obtained by the first factor of a correspondences analysis. In this case, Gamma yields results similar to those of PEM which are equally interpretable.
• Although criticized for its use of the chi-square, Cramér's V reacts like the other indices, but on another range; like the others, it is very weak when it is close to independence (zodiac signs).
• We call the Cramér percentage the index defined by Cramér himself and not as it has been interpreted by other authors who then took the square root of it. Indeed, Cramér said (1946) that "φ2 / q-1 [q is the smallest dimension of the table] may be used as a measure, on a standardized scale, of the degree of dependence between the variables" (1946: 282). The proportion of this maximum can be read as a percentage. We note that this index is very pessimistic.
• The significance of the PEM depends on the significance of the table from which it came. When the PEM is calculated on a non-significant table, we cannot exclude the situation of independence and therefore the nullity of the PEM. This is the case here for the last table.
• Concerning the ranges of PEM use, experience shows that interesting PEM s range between 10 percent and 50 percent. The stronger ties are often indicating a redundancy between indicators. When the tie is less than 10 percent, it may be the result of chance and the chi-square test can show this.

Conclusion

The global PEM can be used as an indicator of the intensity of a link between rows and columns in all cross tables. If there is an existing order on the rows and columns, it will be found by the first factor of the factorial analysis of correspondences. If this is not the case, it may be necessary to challenge the order defined a priori or understand why there is a difference. If we trust the order defined previously, we can then use it to calculate the maximum table. If the order determined by the first factor of the factorial analysis of correspondences is not interpretable, we are then in a situation related to a structure of random deviations and the PEM will probably be small and the table not significant in the sense of the chi-square.

The PEM has the advantage of not being calculated with an index derived from the chi-square (in opposition to Cramér's V). It does not assume there is an order on the table margins, but identifies such an order, if there is one (in opposition to indices calculated from matched pairs). The minimum corresponds to independence and the maximum is well defined and is realistic in the

sense that a value close to 100 percent can actually be observed (if we cross two indicators of the same dimension). It does not depend on the size or the number of rows and columns. For detailed analysis of a table, it can be used for each cell in its local version. And it's easy to understand.

The PEM is available in the Trideux[3] and Modalisa[4] softwares. Programming does not pose special problems: one will find in the Appendix the program in R made by Nicolas Robette[5].

Appendix

Calculation of the PEM

```
# ================================
# R function for calculating the PEM
# (Percentage of the maximum deviation,
# proposed by Philippe Cibois)
# ================================
# X must be an object table or matrix
# The function returns the local PEM ($peml) and the global PEM ($pemg)

pem <- function(x) {
  tota <- colSums(x)
  totb <- rowSums(x)
  total <- sum(x)
  theo <- matrix(nrow=nrow(x),ncol=ncol(x))
  for(i in 1:nrow(x)) { for(j in 1:ncol(x)) theo[i,j] <- tota[j]*totb[i]/total }
  ecart <- x-theo
  max <- matrix(nrow=nrow(x),ncol=ncol(x))
  emax <- matrix(nrow=nrow(x),ncol=ncol(x))
  pem <- matrix(nrow=nrow(x),ncol=ncol(x))
  for(i in 1:nrow(x)) { for(j in 1:ncol(x)) {
    if(ecart[i,j]>=0) max[i,j] <- min(tota[j],totb[i])
    if(ecart[i,j]<0&tota[j]<=(total-totb[i])) max[i,j] <- 0
    if(ecart[i,j]<0&tota[j]>(total-totb[i])) max[i,j] <- tota[j]+totb[i]-total
    emax[i,j] <- max[i,j] - theo[i,j]
    pem[i,j] <- ifelse(ecart[i,j]>=0,ecart[i,j]/emax[i,j]*100,0-ecart[i,j]/emax[i,j]*100)
    }}
  dimnames(pem) <- dimnames(x)
  cor <- corresp(x,nf=1)
  z <- x[order(cor$rscore),order(cor$cscore)]
  tota <- colSums(z)
  totb <- rowSums(z)
  maxc <- matrix(0,nrow=nrow(z),ncol=ncol(z))
  i <- 1; j <- 1
  repeat {
    m <- min(tota[j],totb[i])
    maxc[i,j] <- m
    tota[j] <- tota[j] - m
    totb[i] <- totb[i] - m
    if(sum(tota)+sum(totb)==0) break
```

[3] http://cibois.pagesperso-orange.fr/Trideux.html
[4] http://www.modalisa.com/
[5] http://nicolas.robette.free.fr/outils.html I thank Nicolas Robette for this procedure as well as the comments made about my text

```
    if(tota[j]==0) j <- j+1
    if(totb[i]==0) i <- i+1
   }
  pemg <- (sum(ecart)+sum(abs(ecart)))/(sum(maxc-
theo[order(cor$rscore),order(cor$cscore)])+sum(abs(maxc-
theo[order(cor$rscore),order(cor$cscore)])))
  rm(tota,totb,total,theo,ecart,max,emax,cor,z,m,maxc,i,j)
  PEM <- list(peml=round(pem,1),pemg=round(100*pemg,1))
  return(PEM)
}
```

References

Adam G, Bon F, Capdevielle J and Mouriaux R(1970) *L'ouvrier français en 1970*. Paris: Presses de la FNSP.

Benzécri JP (1976) *L'analyse des données. Tome I La taxinomie*. Paris: Dunod.

Bertin J (1967) *Sémiologie graphique*. Paris and La Haye: Mouton.

Cibois P (1984) *L'analyse des données en sociologie*. Paris: Presses universitaires de France.

Cibois P (1993) Le PEM, pourcentage de l'écart maximum - Un indice de liaison entre modalités d'un tableau de contingence. *Bulletin de Méthodologie Sociologique*, 40: 43-63.

Cibois P (1997) Les pièges de l'analyse des correspondances. *Histoire & Mesure*, 12(3/4): 299-320.

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

Goodman LA and Kruskal WH (1954) Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, 49: 732-64.

Goodman LA and Kruskal WH (1959) Measures of Association for Cross Classifications. II: Further Discussion and References. *Journal of the American Statistical Association*, 54, 123-63.

Goodman LA and Kruskal WH (1963) Measures of Association for Cross Classifications. III: Approximate Sampling Theory. *Journal of the American Statistical Association*, 58, 310-64.

Goodman LA and Kruskal WH (1972) Measures of Association for Cross Classifications. IV: Simplification of Asymptotic Variances. *Journal of the American Statistical Association*, 67, 415-21.

Goodman LA and Kruskal WH (1979) *Measures of Association for Cross Classifications*. New York: Springer-Verlag.

Guttman L (1950) The Basis for Scalogram Analysis. In Stouffer SA., Guttman L, Suchman EA, Lazarsfeld PF, Star SA and Clausen JA (eds) *Measurement and Prediction*. Princeton NJ: Princeton University Press, 60-90.

Kendal M and Stuart A (1961) *The Advanced Theory of Statistics, Volume 2*. London: C. Griffin and Company.

Pearson K (1904) On the Theory of Contingency and Its Relation to Association and Normal Correlation. *Biometric Series No. 1*, London: Drapers Company Research Memoirs (reprinted in 1948 in *Karl Pearson's Early Statistical Papers*. London: Cambridge University Press).

Tschuprow AA (1925) *Grundbegriffe und Grundprobleme der Korrelationstheorie.* Leipzig and Berlin: Teubner.

Yule GU (1900) On the Association of Attributes in Statistics - With Illustrations from the Material of the Childhood Society, &c. *Philosophical Transactions of the Royal Society of London*, series A, vol. 194.