Statistica & Applicazioni
Vol. IV, n. 2, 2006
pp. 3-25

Dependence measures based on partial and total orderings[1]

Francesca Greselin [§]

Summary: *The aim of this paper is to propose a new operational measure for evaluating the degree of dependence existing between two nominal categorical variables. Given an r×c table T, representing bivariate statistical data, our approach to measure the strength of this relation is based on the consideration of the class $\mathcal{F}$ of all contingency tables with the same given margins as T. Once a partial or total ordering of dependence in $\mathcal{F}$ (as defined in Greselin and Zenga [2004b]) has been given, the relative position assumed by T in $\mathcal{F}$ can be a meaningful measure of dependence. Some desirable properties of these indexes are presented: by construction they are normalized, coherent with each level of ordering and attain extreme values in extreme dependence situations. They are invariant to permutation of rows and columns in the table and to transposition (as qualitative variables classification requires), and, finally they show a sort of stability behaviour with respect to similar populations. Furthermore, their straightforward interpretability is compared with the classical interpretation of some well-known normalized indexes. Interesting remarks arise when the comparison is carried out on the discussion of their values, particularly on the extreme dependence situations.*

Keywords: *partial ordering of dependence, association measure, dependence measure.*

1. Introduction

Dependence relations between variables is one of the most widely studied subjects in Statistics. Many studies in the literature are devoted to explore the nature and the extent of the relationship between two variables. Relatively few works deal with nominal categorical variables, as this paper

[1] Preliminary findings of this work have been presented at SIS (Società Italiana di Statistica) Annual Meeting, Milan, 2002.

[§] Quantitative Methods for Economics and Business Sciences – University of Milano-Bicocca – P.za dell'Ateneo Nuovo 1, 20126 MILANO (e-mail: francesca.greselin@unimib.it).

does (for a specific review, see Greselin and Zenga (2004a)).

The aim of this work is to employ the hierarchy of partial and total orderings of dependence recently introduced by Greselin and Zenga (2004b) to define a new class of measures of dependence. Our proposal is to express the strength of dependence of a given table T by the relative position that T assumes among a finite set of tables, endowed by an ordering of dependence. The reference set is the class of all bivariate distributions having the same pair of margins.

The remainder of the paper is organized as follows. Section 2 states the terminology and recalls some concepts related to partial and total dependence orderings. In Section 3 the definition of the new measures is recalled and some desirable properties of these indexes are stated. Section 4 deals with the computational cost of the index. In Section 5, by computational results, a sort of asymptotic behaviour of the indexes with respect to similar population is shown. Section 6 shows some graphical comparisons between two classical indexes and the new measures of dependence we deal with. The cases of very high strength of dependence (respectively very low) deserves interesting comments. Section 7 gives some concluding remarks.

2. Brief review and terminology

Let N statistical units of a given population be classified according to the qualitative variables A and B, both with nominal scale, whose unordered categories are denoted respectively by $a_1,\ldots,a_j,\ldots,a_c$ and $b_1,\ldots,b_i,\ldots,b_r$. As usual, the joint frequency $n(b_i,a_j)$ of the pair of categories b_i and a_j is denoted by n_{ij}, while marginal frequencies are denoted by $n_{i\bullet} = n(b_i)$, $i = 1,2,\ldots,r$ for variable B and $n_{\bullet j} = n(a_j)$, $j = 1,2,\ldots,c$ for variable A. Bivariate statistical data are generally represented in a table with r rows and c columns, as Figure 1 shows:

$B\backslash A$	a_1		a_j		a_c	*Total*
b_1	n_{11}	…	n_{1j}	…	n_{1c}	$n_{1\bullet}$
…	…		…		…	…
b_i	n_{i1}	…	n_{ij}	…	n_{ic}	$n_{i\bullet}$
…	…		…		…	…
b_r	n_{r1}	…	n_{rj}	…	n_{rc}	$n_{r\bullet}$
Total	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet c}$	N

Figure 1. *A bivariate table and its notation*

In the hypothesis of independence (lack of association) the joint

frequencies are given by:

$$\hat{n}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{N} \quad i = 1,...,r \quad \text{and} \quad j = 1,...,c \tag{1}$$

and therefore independence is a symmetric relation.

The literature on measures of dependence for cross-classification is really wide, and the reader can refer to Goodman and Kruskal's works (1954, 1959, 1963, 1972 reprinted as a volume in 1979) which provide a broad survey, careful discussion as well as brief historical and bibliographical remarks.

Association measures are single summary numbers that describe the type and the intensity of the relationships between two classified variables. When properly used, they provide a useful description of the dependence structure displayed by a two-dimensional table.

In this paper variables *A* and *B* are supposed to play a symmetrical role, so that a classical approach to evaluate deviation from independence can be based on the contingencies:

$$c_{ij} = n_{ij} - \hat{n}_{ij} \quad (i = 1,2,..,r;\ j = 1,2,...,c) \tag{2}$$

or on the relative contingencies:

$$\rho_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\hat{n}_{ij}}. \tag{3}$$

A careful observation of the table of contingencies offers an analytical view over the kind and the intensity of the relation between each pair of the categories of the two variables *A* and *B*. In order to obtain a synthetic measure of its strength, a suitable normalized weighted mean of contingencies yields two well-known association indexes, the first proposed by Pearson (1904) and successively normalized by Cramer (1946), the second due to Mortara (1922).

These widely used indexes possess many desirable and well-known properties:

- they assume values in [0,1];
- they are null if and only if there is independence (lack of association);
- they assume their maximum value, i.e. 1, if and only if there is complete or absolute association[2];

[2] For the maximum dependence situations we will adopt Kendall and Stuart (1979) terminology, briefly recalled here: "Considering a population classified according to the presence or absence of two attributes *A* and *B*, we say that association is *complete* if all *A*'s are *B*'s. *Absolute* association arises when all *A*'s are *B*'s and all

- they are invariant under transposition and permutation of rows or columns[3];
- they are invariant with respect to 'similar' populations.

The term '*similar*' is given by the following definition:

DEFINITION 1
*Let T be a given bivariate population T ={n_{ij}} and let α a positive integer. The α-**similar** population with respect to T is the bivariate distribution αT, defined by the joint frequencies $n_{ij}^{(\alpha)} = \alpha \cdot n_{ij}$, multiples according to α of the initial frequencies n_{ij}.*

Denoting with an apical (α) each index, when it refers to the α-similar population obtained from the original by replicating it α times, it is well-known that:

- $M_1(|\rho|) = M_1^{(\alpha)}(|\rho|)$, where $M_1(|\rho|)$ is the arithmetic mean of the absolute value of the relative contingencies, weighted by the joint frequencies $\hat{n}_{ij}$;
- $M_2(|\rho|) = M_2^{(\alpha)}(|\rho|)$, being $M_2(|\rho|)$ the quadratic mean of the absolute value of the relative contingencies, weighted by the joint frequencies $\hat{n}_{ij}$;
- $M' = M'^{(\alpha)}$ where M' stands for the Mortara's normalized index, defined by $M' = M_1(|\rho|)/2\hat{E}$ and $\hat{E}$ is the minimum Gini's etherogeneity index of the two marginal distributions of *A* and *B*;
- $C = C^{(\alpha)}$ where C is Cramer's norming position for $M_2(|\rho|)$, i.e.: $C = M_2(|\rho|)/(k\text{-}1)^{1/2}$, with $k = \min(r,c)$.

Concluding these remarks, it is worth noting that, except in the 2×2 table case, a single function (and hence a single measure of association) cannot reflect the large diversity of ways in which a table can depart from independence.

This leads to the variety of measures and to the difficulty inherent in choosing a single measure in any given situation. Their weakness and the major difficulty in their use is their lack of a clear interpretation. Therefore, partial orderings appear to be an adequate method in approaching dependence and the introduction of a class of measures based on partial

---

B's are *A*'s." These definitions corresponds to *unilateral* and *bilateral dependence*, in the Italian literature.

[3] This property corresponds to the discretionary position of the variables *A* and *B* – and of their categories – in the rows or columns of the table.

orderings tries to overcome this issue.

3. Our proposal

In Greselin and Zenga (2004a) a *cyclic frequency transfer* is defined: it transforms a bivariate distribution T in a new one, say T', which shows a lower degree of dependence between the variables A and B, and T' has the same pair of margins as T. A cyclic frequency transfer acts over a paired set of cells in T, decrementing the absolute value of their contingencies and maintaining the margins. In the cited work, hence, the authors say that in T' there is lower *directional dependence* between A and B than in T. This notion corresponds to the definition of a partial ordering of dependence in the reference set $\mathcal{F}$, the class of all $r \times c$ contingency tables with non negative integer entries n_{ij}, whose row sums $n_{i\cdot}$ and column totals $n_{\cdot j}$ are given.

In partial ordering relations, only a subset of pairs belongs to the relation, therefore to enable the ranking of all pairs of tables of $\mathcal{F}$ a total ordering is needed:

DEFINITION 2 ***Total ordering of dependence***

Let $f: \mathcal{F} \to P$ be a real valued function, mapping $r \times c$ tables from $\mathcal{F}$ onto the real line P. Let $T, T' \in \mathcal{F}$, then T precedes T' according to $\leq_f$ if and only if:

$$f(T) \leq f(T') \tag{4}$$

and it will be denoted by $T \leq_f T'$.

The function $f: \mathcal{F} \to \mathrm{P}$ that induces the total ordering on $\mathcal{F}$, in order to be adequate in measuring dependence, has to be coherent with the *Directional dependence ordering* (and also with the *Intensity of dependence ordering* in Greselin and Zenga (2004b), based only on the absolute values of the contingencies).

A sufficient condition for $f = f(\rho_{ij};\ \hat{n}_{ij})$ is to be a bounded, symmetric, non-negative and non-decreasing functional form of its arguments, the absolute (or relative) contingencies ρ_{ij}, having the independence frequencies $\hat{n}_{ij}$ as parameters. For example, f can be $\mathrm{M}_1(|\rho|)$ or $\mathrm{M}_2(|\rho|)$: their domain is a set of bivariate distributions and their co-domain is the set of real numbers[4]. Each of these functions induces a total ordering relation, denoted by $\leq_{\mathrm{M}_1(|\rho|)}$ and $\leq_{\mathrm{M}_2(|\rho|)}$.

[4] For a comparison between the indexes $\mathrm{M}_1(|\rho|)$ and $\mathrm{M}_2(|\rho|)$, and for a discussion of their joint use in measuring dependence, see Greselin and Zenga (2004b) .

Let T indicates the generic table pertaining to $\mathcal{F}$. As sketched above, the second step of our approach is then to consider, as a measure of dependence for T, the relative position that T assumes in its ordered class $\mathcal{F}$.

DEFINITION 3 ***The class of indexes $I_f(T)$***

Given a bivariate distribution T, the relative position that T assumes in $\mathcal{F}$, is a measure of dependence $I_f(T)$, for T:[5]

$$\mathrm{I}_f(T) = \frac{\#\left\{S \mid S \in \mathcal{F};\ S \leq_f T\right\}}{\#\ \mathcal{F}} \tag{5}$$

The index I_f will also appear in the foregoing with explicit mention of the chosen function: we will specifically deal with $\mathrm{I}_{\mathrm{M}_1(|\rho|)}$ and $\mathrm{I}_{\mathrm{M}_2(|\rho|)}$.

3.1 *Properties of $I_f(T)$*

First of all, the index $\mathrm{I}_f(T)$ is a relative frequency, so its meaning is straightforward. For example, if T is a table in $\mathcal{F}$ and $\mathrm{I}_{\mathrm{M}_2(|\rho|)}(T)$ =.23, it means that, among all bivariate distributions in $\mathcal{F}$, ordered by non-decreasing values of $\mathrm{M}_2(|\rho|)$, T lies in a position that divides the class in two parts, 23 percent of tables in $\mathcal{F}$ has a lower or equal value of $\mathrm{M}_2(|\rho|)$, and the remaining tables have an higher value.

Moreover, the index I_f has the following desirable properties:

- it takes values in [0,1], i.e. it is autonormalized, by construction.
- The minimum value of $\mathrm{I}_f(T)$ is never 0, but it is a very good approximation of 0, given by the value $1/(\#\mathcal{F})$, if and only if T is a minimum table for the dependence ordering in $\mathcal{F}$. For almost all tables T of real statistical data, $\#\mathcal{F}$ has a high value: as it is shown in (Greselin, 2003) $\#\mathcal{F}$ grows exponentially with the population size N, the number r of rows and the number c of columns in the table[6].
- Conversely, the maximum value $\mathrm{I}_f(T) = 1$ is reached if and only if T is a maximum table for the ordering function f in $\mathcal{F}$. This property is coherent with our choice of comparing a table T in its class, choosing one of the three distinct possibilities (both sets of margins fixed, one

---

[5] The symbol $\#\mathcal{F}$ stands for "number of the elements of the set $\mathcal{F}$", i.e. the cardinality of $\mathcal{F}$.

[6] See also Examples 3, 4 and 5 in the following Section, giving the values $\#\mathcal{F}$ in tables 3, 4 and 5 respectively.

margin fixed, total population size fixed) as in Zanella (1988)[7].

- it is invariant under transposing and under permutation of rows and columns, that is, for Q and S tables in a given class $\mathcal{F}$:

  if $Q \leq_f S$ then $Q^T \leq_f S^T$ and $P_1 Q P_2 \leq_f P_1 S P_2$

  for all $r \times r$ permutation matrices P_1 and all $c \times c$ permutation matrices P_2.

Agreeing with Goodman and Kruskal's remark (1954): "one difficulty with the use of traditional normalized measures is that it is difficult to compare meaningfully their values for two cross-classification tables", our proposal overcomes this issue: being a relative position, $I_f(T)$ allows comparisons between tables with different dimensions r and c, and different population sizes N.

4. The computational cost of $I_f(T)$

Now a remark about the computational cost to compute the index $I_f(T)$ is needed. To evaluate $I_f(T)$, the whole class $\mathcal{F}$ of tables with the same margins as T has to be generated. Then a measure of dependence, given by $M_1(|\rho|)$ or $M_2(|\rho|)$, or, more generally, by the function $f = f(\rho_{ij}; \hat{n}_{ij})$ has to be evaluated on each table S in $\mathcal{F}$, observing if it produces a value $f(S) < f(T)$ so that the relative position of T in $\mathcal{F}$ can be evaluated in the meanwhile. These steps can be executed in a few seconds or minutes, by a software program written in C and running on a common pc Pentium V, whenever $N < 1000$ and the number of the rows and columns of the table T does not exceed 10.

In the other cases the computational time required to generate the class $\mathcal{F}$ grows, becoming quickly intractable. The following Section, showing an observed behaviour of the index, will be very useful to face this issue.

5. The behaviour of $I_f(T)$ with respect to similar distributions

It could be expected that $I_f(T)$ possesses the invariance property with respect to similar populations, as Mortara's and Cramer's normalized indexes and other traditional indexes do.

---

[7] It is worth noting that, whenever complete or absolute association is compatible with the given margins, all the functions f attain their maximum values on T and hence $I_f(T) = 1$;

Actually, it has been verified that for low size populations (N≈10,..,100), this property does not hold and the multiplication of a table by an integer factor α can modify in an unpredictable direction $I_f(T)$. In the following, Example 1 has been chosen to show how $I_f(T)$ may change when T lies in a small class $\mathcal{F}$.

Indeed, considering the "granularity" of the distributions in $\mathcal{F}$, due to the requirement that joint frequencies are integers, this anomaly is expected to disappear by increasing αN. This is what we have verified: as α or N increase so that αN reaches some hundreds, the index $I_f(T)$ attains a substantial stabilization, presenting fluctuations only on its second decimal. Examples 2, 3 and 4 have been chosen, among a set of more results, to show the stabilization process of the index.

The following Tables 1-5 show the results of computing the index $I_f(T)$ for similar distributions. Starting from a given table T and iteratively computing its double, triple, etc, multiplying its joint frequencies by a positive integer α, we have reached interesting results. A software program has been developed: beginning with α=1 and incrementing its value, the algorithm generates the class $\alpha\mathcal{F}$ and calculates the index $I_f(\alpha T)$ for each value of α.

Example 1: Let us consider table T in Figure 2:

1	2	3
3	3	6
1	0	1
5	5	10

Figure 2. *The bivariate table T*

The aim of this first example is to examine in detail the class $\mathcal{F}$ to which table T belongs, here denoted also by the given margins: $\mathcal{F}\{(3,6,1);(5,5)\}$, and composed by the following eight tables (printed without margins for shortness):

0 3	0 3	1 2	1 2	2 1	2 1	3 0	3 0
4 2	5 1	3 3	4 2	2 4	3 3	1 5	2 4
1 0	0 1	1 0	0 1	1 0	0 1	1 0	0 1
T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8

Figure 3. *The class $\mathcal{F}${(3,6,1);(5,5)}*

First of all, we recognize table T in the third enumerated table: T_3. For each of the bivariate distributions that pertain to $\mathcal{F}\{(3,6,1);(5,5)\}$, the quantities $M_1(|\rho|)$, $M_2(|\rho|)$, Mortara's index M′ and C were evaluated,

obtaining the distributions of those indexes over $\mathcal{F}$.

Different total orderings can be considered for enumerating the tables in $\mathcal{F}\{(3,6,1);(5,5)\}$: by non decreasing values of $M_1(|\rho|)$, or of $M_2(|\rho|)$ etc.

As a second step, we consider the 2-similar distribution $2T_3$ obtained by doubling all observed frequencies:

2	4	6
6	6	12
2	0	2
10	10	20

Figure 4. *The bivariate table $2T_3$*

For table $2T_3$ the corresponding class $2\mathcal{F}= \mathcal{F}\{(6,12,2);(10,10)\}$ is then generated, all indexes cited above are evaluated on each table in $2\mathcal{F}$ and the different orderings on $2\mathcal{F}$ are considered. The class $\mathcal{F}\{(6,12,2);(10,10)\}$ is composed by 21 tables. Among them there are the eight tables that are 'doubles' of those in $\mathcal{F}\{(3,6,1);(5,5)\}$. As the eight initial tables were all comparable using a chosen (partial or total) ordering with T_3, in the same way and with the same kind of inequalities their 'doubles' are comparable with $2T_3$. The remaining 13 'new' matrices (that are not multiples of any of the preceding tables) do not split uniformly at the left and at the right of $2T_3$ with respect to the selected ordering. Hence the dependence index based on the relative position in the ordering changes from $I_{M_1(|\rho|)}= 0.25$ on T_3 to $I_{M_1(|\rho|)}= 0.333333$ on $2T_3$. Conversely, because of the invariance property, $M_1(|\rho|) = M' = 0.2$ and $M_2(|\rho|) = C = 0.365$ for all tables αT_3.

The analysis of table T_3 goes on with the generation of the following multiples of T_3 and of the corresponding classes $\alpha\mathcal{F}$, in order to study the behaviour of each of the indexes of interest. Passing from $2T_3$ to $3T_3$, the number of the elements in $\mathcal{F}\{(9,18,3);(15,15)\}$ increases up to 40; only 9 tables precede $3T_3$ in the ordering induced by $M_1(|\rho|)$, so that $I_{M_1(|\rho|)}= 0.25$ on $3T_3$. Observing $4T_3$, the cardinality of its class $4\mathcal{F}$ attains 65, only 18 tables precede $4T_3$ in the ordering induced by $M_1(|\rho|)$ and $I_{M_1(|\rho|)}= 0.292308$ on $4T_3$. Table 1 shows, from left to right, all results on αT_3: the multiplying factor α in the first column; then, the size of the α-similar population αN, the cardinality of the class $\alpha\mathcal{F}$, and the indexes $I_{M_1(|\rho|)}$ and $I_{M_2(|\rho|)}$ evaluated on αT_3.

α	αN	$\#(\alpha\mathcal{F})$	$I_{M_1(\lvert\rho\rvert)}$	$I_{M_2(\lvert\rho\rvert)}$
1	10	8	0.250000	0.250000
2	20	21	0.333333	0.333333
3	30	40	0.250000	0.400000
4	40	65	0.292308	0.384615
5	50	96	0.250000	0.395833
6	60	133	0.278195	0.413534
7	70	176	0.250000	0.420455
8	80	225	0.271111	0.413333
9	90	280	0.250000	0.421429
10	100	341	0.266862	0.413490
20	200	1,281	0.258392	0.433255
30	300	2,821	0.255583	0.441333
40	400	4,961	0.254183	0.443257
50	500	7,701	0.253344	0.446306
60	600	11,041	0.252785	0.447514
70	700	14,981	0.252386	0.448635
80	800	19,521	0.252087	0.448696
90	900	24,661	0.251855	0.450185
100	1,000	30,401	0.251669	0.450215
200	2,000	120,801	0.250834	0.452074
300	3,000	271,201	0.250556	0.452701
400	4,000	481,601	0.250417	0.453074
500	5,000	752,001	0.250333	0.453296
600	6,000	1,082,401	0.250278	0.453378
700	7,000	1,472,801	0.250238	0.453494
800	8,000	1,923,201	0.250208	0.453570
900	9,000	2,433,601	0.250185	0.453594
1,000	10,000	3,004,001	0.250167	0.453668
1,500	15,000	6,756,001	0.250111	0.453792
2,000	20,000	12,008,001	0.250083	0.453860
2,500	25,000	18,760,001	0.250067	0.453897
3,000	30,000	27,012,001	0.250056	0.453924
3,500	35,000	36,764,001	0.250048	0.453943
4,000	40,000	48,016,001	0.250042	0.453956
4,500	45,000	60,768,001	0.250037	0.453967

Table 1. *Behaviour of the indexes* $I_{M_1(\lvert\rho\rvert)}$ *and* $I_{M_2(\lvert\rho\rvert)}$ *on tables* αT_3

The behaviour of each index can be easily resumed: $I_{M_1(\lvert\rho\rvert)}$, after oscillating, steadily decreases, while $I_{M_2(\lvert\rho\rvert)}$ increases from the multiple α=10 upwards. Both show a stabilization behaviour for high values of αN.

We can investigate if this is a general behaviour of the indexes, developing the working scheme selected for T_3, and repeating it in detail for each table

in the same class $\mathcal{F}\{(3,6,1);(5,5)\}$.

Hence, carrying out the same analysis on table T_4:

1	2	3
4	2	6
0	1	1
5	5	10

Figure 5. *The bivariate table T_4*

as the tables pertaining to the double, triple and quadruple classes $2\mathcal{F}$, $3\mathcal{F}$, $4\mathcal{F}$ (and comparable with $2T_4$, $3T_4$ and $4T_4$) do not split uniformly at the left and at the right of the corresponding multiple of T_4 w.r.t. the ordering, the value of the index $I_{M_1(|\rho|)}$ varies:

$I_{M_1(|\rho|)}(T_4) = 0.5$ $\qquad$ $I_{M_1(|\rho|)}(2T_4) = 0.619$

$I_{M_1(|\rho|)}(3T_4) = 0.55$ $\qquad$ and $\qquad$ $I_{M_1(|\rho|)}(4T_4) = 0.6$

while $M' = 0.4$ is the value of Mortara's normalized index; and $C = 0.447$ is the Cramer's index for all similar tables αT_4.

Analogous considerations can be drawn with respect to the table T_1, pertaining to $\mathcal{F}\{(3,6,1);(5,5)\}$:

0	3	3
4	2	6
1	0	1
5	5	10

Figure 6. *The bivariate table T_1*

All results on replications αT_1 are resumed in Table 2.

For all tables αT_1 the values of Mortara's normalized index and the Cramer's index are $M' = 0.6$ and $C = 0.683$, respectively. Also in this case, after oscillating for the first multiples α, the two indexes $I_{M_1(|\rho|)}$ and $I_{M_2(|\rho|)}$ attain a substantial stabilization around the values 0.92 and 0.96, respectively, as the product αN reaches a hundred. The difference among the two families of indexes is remarkable: while Mortara's index indicates a fair dependence, measured by a 60 percent of the maximum dependence, the index based on the total ordering points out that only an 8 percent of tables in $\mathcal{F}$ reflect a higher degree of dependence. This remarkable difference will be further discussed in Section 5.

α	αN	$\#(\alpha\mathcal{F})$	$I_{M_1(\|\rho\|)}$	$I_{M_2(\|\rho\|)}$
1	10	8	0.75000	0.75000
2	20	21	0.90476	0.90476
3	30	40	0.85000	0.90000
4	40	65	0.90769	0.90769
5	50	96	0.87500	0.91667
6	60	133	0.90977	0.93985
7	70	176	0.88636	0.93182
8	80	225	0.91111	0.93778
9	90	280	0.89286	0.94286
10	100	341	0.91202	0.93548
100	1000	30,401	0.91612	0.95790
200	2000	120,801	0.91639	0.95871
300	3000	271,201	0.91648	0.95900
400	4000	481,601	0.91653	0.95923
500	5000	752,001	0.91656	0.95931
600	6000	1,082,401	0.91657	0.95937
700	7000	1,472,801	0.91659	0.95943
800	8000	1,923,201	0.91660	0.95947
900	9000	2,433,601	0.91661	0.95948
1,000	10,000	3,004,001	0.91661	0.95952
1,400	14,000	5,885,601	0.91663	0.95957
1,800	18,000	9,727,201	0.91664	0.95960

Table 2. *Behaviour of the indexes* $I_{M_1(|\rho|)}$ *and* $I_{M_2(|\rho|)}$ *on tables* αT_1

Table T_2, also pertaining to $\mathcal{F}\{(3,6,1);(5,5)\}$, shows a different behaviour for the indexes, due to its particular position in the dependence orderings:

0	3	3
5	1	6
0	1	1
5	5	10

Figure 7. *The bivariate table* T_2

For this table, as the tables belonging to the double, triple and quadruple classes $\mathcal{F}$ (and comparable with $2T_2$, $3T_2$ and $4T_2$) rank all to the left of $2T_2$, $3T_2$ and $4T_2$, respectively. This happens because in $\mathcal{F}$ the maximum values of $M_1(|\rho|)$, $M_2(|\rho|)$, of Mortara's M′ and C are assumed on table T_2, and this property hold also for its multiples αT_2. Hence the two indexes $I_{M_1(|\rho|)}$ and $I_{M_2(|\rho|)}$ attain steadily their maximum, depicting the evidence that the initial table T_2 and its multiples are the 'farthest' from independence with respect to

the entire class $\mathcal{F}$. Conversely, the normalized indexes C and M′ do not attain their maximum value, because of the constraint on margins: M′= 0.8 and C = 0.817 for all tables αT_2.

Finally, observing that the remaining four tables in $\mathcal{F}\{(3,6,1);(5,5)\}$ are obtained by T_1, T_2, T_3 and T_4 interchanging the columns (so they lead to the same results), the analysis of the class is completed.

For the sake of brevity, and with the purpose of exploring one entire class $\mathcal{F}$ in a reasonable amount of computing time, this first example refers to tables with low statistical meaning, while in the following cases we apply the replications to more realistic bivariate distributions.

Example 2: Now we analyze a 2×3 table *S*, referring to a population with *N*= 837 and a low dependence, measured by M′ = 0.222 and C = 0.193:

80	80	45	205
387	161	84	632
467	241	129	837

Figure 8. *The bivariate table S*

Proceeding as before with multiples, now focusing on studying the stabilization of $I_{M_1(|\rho|)}$ and $I_{M_2(|\rho|)}$, the following Table 3 gives in the fifth and in the seventh column the relative variation in comparing the indexes evaluated on a table and on its double:

α	αN	$\#(\alpha\mathcal{F})$	$I_{M_1(\|\rho\|)}$	$I_{M_1}(2\alpha S)/I_{M_1}(\alpha S)$	$I_{M_2(\|\rho\|)}$	$I_{M_2}(2\alpha S)/I_{M_2}(\alpha S)$
1	837	18,395	0.188366	-	0.129492	-
2	1,674	73,038	0.188806	1.002336	0.130302	1.006255
4	3,348	291,071	0.189930	1.005953	0.130762	1.003530
8	6,696	1,162,125	0.190037	1.000563	0.131001	1.001828
16	13,392	4,644,185	0.190100	1.000332	0.131125	1.000947
32	26,784	18,568,113	0.190132	1.000168	0.131184	1.000450
64	53,568	74,255,201	0.190115	0.999911	0.131214	1.000229
128	107,136	296,986,305	0.190122	1.000037	0.131230	1.000122

Table 3. *Behaviour of the indexes* $I_{M_1(|\rho|)}$ I_{M_1} *and* $I_{M_2(|\rho|)}$ *on tables* αS

The starting size *N* = 837 of the population assures that the relative variation of the two indexes $I_{M_1(|\rho|)}$and $I_{M_2(|\rho|)}$ has an order of magnitude of

10^{-3}, till from $\alpha=1$.

Example 3: Now we analyze a 2×4 table *V*, shown in Figure 9, referring to a population with $N = 84$ and reflecting a strong dependence between variables, for $M_1 = M' = 0.738$ and $M_2 = C = 0.772$:

0	5	17	20	42
22	14	5	1	42
22	19	22	21	84

Figure 9. *The bivariate table V*

The results on αV are presented in Table 4:

α	αN	$\#(\alpha\mathcal{F})$	$I_{M_1}(\lvert\rho\rvert)$	$I_{M_1}(2\alpha S)/I_{M_1}(\alpha S)$	$I_{M_2}(\lvert\rho\rvert)$	$I_{M_2}(2\alpha S)/I_{M_2}(\alpha S)$
1	84	7,040	0.936080	-	0.934091	-
2	168	52,507	0.944579	1.009079	0.946826	1.013634
4	336	405,405	0.946241	1.001760	0.952813	1.006323
8	672	3,185,817	0.947102	1.000910	0.955755	1.003088
16	1,344	25,259,185	0.947539	1.000461	0.957167	1.001477
32	2,688	201,168,737	0.947760	1.000233	0.957890	1.000755
64	5,376	1,605,740,225	0.947871	1.000117	0.958247	1.000373
128	10,752	12,831,501,697	0.947926	1.000058	0.958425	1.000186

Table 4. *Behaviour of the indexes* $I_{M_1}(\lvert\rho\rvert)$ *and* $I_{M_2}(\lvert\rho\rvert)$ *on tables* αV

Also in this case a sort of stability of the two indexes $I_{M_1}(\lvert\rho\rvert)$ and $I_{M_2}(\lvert\rho\rvert)$ is expressed by their low relative variations.

Example 4: Finally, we analyze a 3×3 table *W*, shown in Figure 10, referring to a population with N = 27 and with an intermediate degree of dependence between characters, expressed by $M' = 0.321$ and $C = 0.316$:

3	3	3	9
8	1	2	11
5	0	2	7
16	4	7	27

Figure 10. *The bivariate table W*

As before, starting from the given table, generating its class and successively doubling the distributions and the corresponding classes, we

obtained the following results:

α	αN	$\#(\alpha\mathcal{F})$	$I_{M_1}(\lvert\rho\rvert)$	$I_{M_1}(2\alpha S)/I_{M_1}(\alpha S)$	$I_{M_2}(\lvert\rho\rvert)$	$I_{M_2}(2\alpha S)/I_{M_2}(\alpha S)$
1	27	500	0.214000	-	0.172000	-
2	54	5,035	0.286395	1.338294	0.250447	1.456087
4	108	62,349	0.329917	1.151965	0.300598	1.200246
8	216	871,097	0.357033	1.082190	0.330207	1.098500
16	432	12,997,105	0.370535	1.037817	0.346283	1.048685
32	864	200,705,505	0.376607	1.016387	0.354601	1.024021
64	1,728	3,154,383,809	0.379930	1.008824	0.358831	1.011929

Table 5. *Behaviour of the indexes* $I_{M_1}(\lvert\rho\rvert)$ *and* $I_{M_2}(\lvert\rho\rvert)$ *on tables* αW

In the first table $N = 27$, so the process of stabilization is slower than above. The relative variation of the indexes reaches an order of magnitude of 10^{-2} only arriving at $\alpha=64$.

More results can be given, also increasing the dimensions of the tables, even if the computing time increases when dealing with the generation of the multiples $\alpha\mathcal{F}$. In any case, all results that we have achieved confirm this stabilization of the indexes when replicated on similar populations as αN reaches some (few) hundreds.

This behaviour of $I_f(T)$ with respect to α–similar distributions can be very useful for evaluating this index in two opposite situations.

In presence of bivariate distributions referring to small population sizes ($N < 400$), we can consider the multiple table αT (for α: $\alpha N > 400$), and select the value of $I_f(\alpha T)$ as a good measure of association for T. This evaluation of the index gives a stabilized value, as it does not depend on the 'granularity' of the initial class $\mathcal{F}$ of table T.

Conversely, as in social sciences researchers generally work with a large number of observations and their cross-classification tables usually refer to populations whose size passes the threshold that assures the stability of $I_f(T)$, the observed behaviour of the indexes can be used in the reverse mode. The approximated table $(\alpha T)'$ with $\alpha < 1$ can be evaluated and the total population can be reduced to the threshold $N \approx 400$. In spite of generating the whole class $\mathcal{F}$ of T, whose cardinality can be so high to force the software program to run for a big amount of time, by means of the given approximation it has to generate only the smaller class $\mathcal{F}'$ corresponding to $(\alpha T)'$. Hence, the software program can calculate the relative position of

$(\alpha T)'$ within an acceptable response time. The index $I_f((\alpha T)')$ is a good approximation ($\sim 10^{-3}$) of the exact index $I_f(T)$.

6. Comparing I_f to Mortara's and Cramer's indexes

The purpose of this section is to offer a graphical representation of I_f with respect to Mortara's and Cramer's normalized indexes. Chosen a class $\mathcal{F}$, for each enumerated table in $\mathcal{F}$, the indexes M′ and C and the measures of dependence based on the relative position $I_{M_1(|\rho|)}$ and $I_{M_2(|\rho|)}$ have been evaluated. Obviously, the total ordering induced by $M_1(|\rho|)$ coincides with that induced by M′, hence $I_{M_1(|\rho|)} \equiv I_{M'}$; analogously for $M_2(|\rho|)$ and C, hence $I_{M_2(|\rho|)} \equiv I_C$.

Let us begin focusing our analysis on $\mathcal{F}\{(42,42)(22,19,22,21)\}$, i.e. the class of 2×4 tables presented in Example 3 of Section 4. This class is composed by 7,040 tables: each representing a different bivariate distribution w.r.t. a population of 84 statistical units. The following Cartesian diagram represents the relation between M′ and $I_{M'}$: a point (x,y) in the graph corresponds to the pair of values $(M'(T), I_{M'}(T))$ evaluated on the same table T in $\mathcal{F}$, considering all tables T in $\mathcal{F}$.

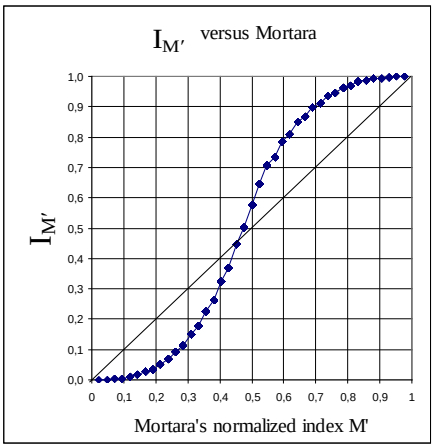

Figure 11. *The index $I_{M'}$ versus Mortara's index in $\mathcal{F}\{(42,42);(22,19,22,21)\}$*

The first remark is that there are very few tables in $\mathcal{F}$ with extreme values of Mortara's index: only 10 percent of tables in $\mathcal{F}$ have $M' \leq 0.285714$ and only another 10 percent of them have $M' \geq 0.714286$. These considerations can be appreciated also by the diagram representing the distribution of M′ in

$\mathcal{F}$ (the range [0,1] of values has been divided in 20 classes, each of width 0.05).

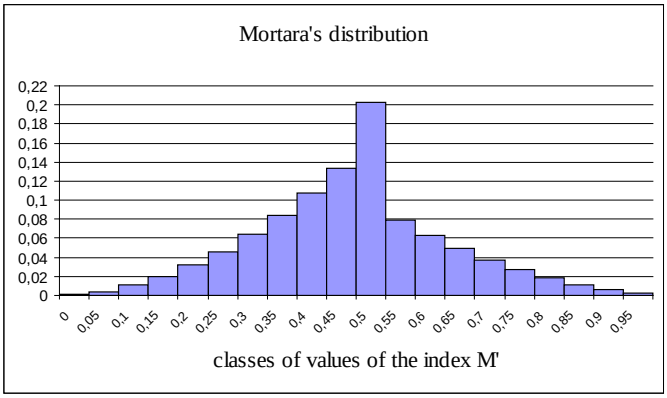

Figure 12. *The distribution of Mortara's index in $\mathcal{F}${(42,42);(22,19,22,21)}*

The following quantiles describe more analytically the distribution of Mortara's index in $\mathcal{F}$: the first decile $D_1(M') = 0.285714$; the first quartile $Q_1(M') = 0.380952$, the median $Me\ (M') = 0.47619$; the third quartile $Q_3(M') = 0.595238$ and, finally, the ninth decile $D_9(M') = 0.714285$.

Hence, a large proportion of tables corresponds to intermediate values of this classical normalized index: namely, 80 percent of tables are described by a Mortara's index in a narrow range of central values: 0.714286-0.285714 $\approx$ 0.43.

The comparison between I_C and C, on the same class $\mathcal{F}${(42,42);(22,19,22,21)}, deserves now our attention. In the following diagram, the value of the index C is plotted in the *X*-axis, while the corresponding value of I_C is plotted in the *Y*-axis:

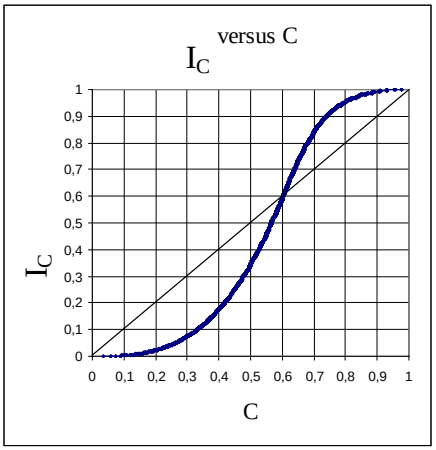

Figure 13. *The index I_C versus Cramer index in $\mathcal{F}\{(42,42);(22,19,22,21)\}$*

Once again, we observe that the lower values and the higher values of C correspond to few tables. The graphical representation of the distribution of C in $\mathcal{F}$, is given in Figure 14.

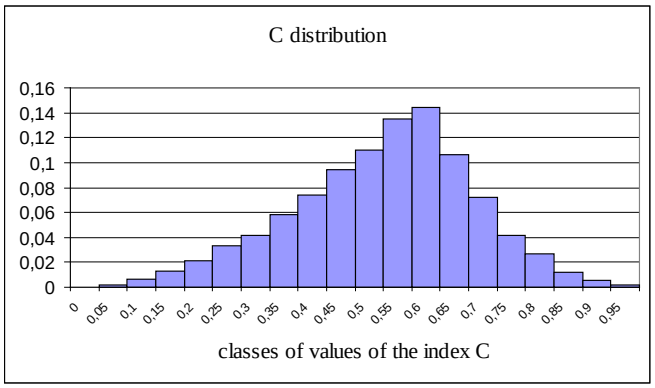

Figure 14. *The distribution of Cramer's index in $\mathcal{F}\{(42,42);(22,19,22,21)\}$*

Only 2 percent of all tables in $\mathcal{F}$ has a value of $C < 0.2$ and the tables with $C > 0.8$ are 3.7 percent of all tables in $\mathcal{F}$. The quantiles of C in $\mathcal{F}$ are the following: (with the same notation as above) $D_1(C) = 0.330752$; $Q_1(C) = 0.449991$; $Me(C) = 0.565737$; $Q_3(C) = 0.658759$ and $D_9(C) = 0.739896$.

Now, let us consider $\mathcal{F}\{(9,11,7);(16,4,7)\}$, the class of 3×3 tables in Example 4 of the above section, with $\#\mathcal{F}=500$ and referring to a population with $N = 27$. Figure 15 represents the relation between Mortara's index and the measure of dependence $I_{M'}$.

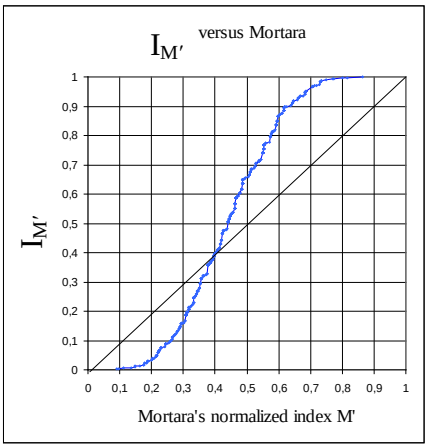

Figure 15. *The index $I_{M'}$ versus Mortara's index in $\mathcal{F}\{(9,11,7);(16,4,7)\}$*

Once again we observe that extreme values of Mortara's index are really rare in the class: only 10 percent of tables in $\mathcal{F}$ have $M' \leq 0.259804$ and only another 10 percent of them have $M' \geq 0.637255$. Conversely, 80 percent of tables are described by values of the classical indexes that have a narrow range of $0.637255-0.259804 \approx 0.38$ for M'. Furthermore, the minimum value of M' in $\mathcal{F}$ is 0.090686; the maximum value is 0.862745; the median of M' is $Me(M') = 0.441176$ in the class $\mathcal{F}$. These considerations can be appreciated also by inspecting the distribution of M':

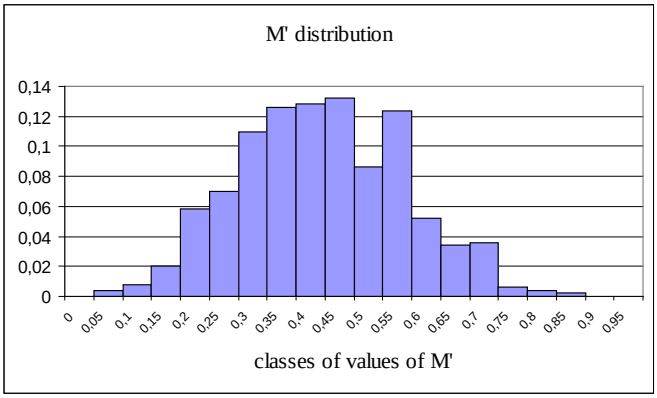

Figure 16. *The distribution of Mortara's index in $\mathcal{F}\{(9,11,7);(16,4,7)\}$*

Among all tables in $\mathcal{F}$, only 3.2 percent have $M' < 0.2$, and only 2.8 percent have $M' > 0.8$.

Figure 17 depicts clearly the relation between C and I_C:

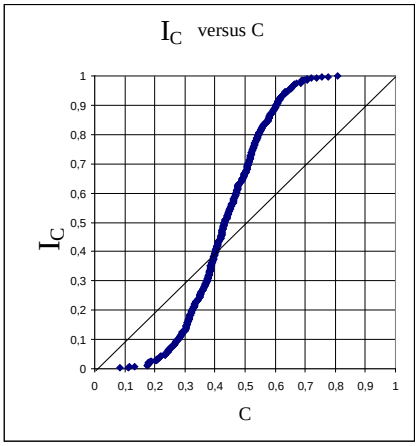

Figure 17. *The index I_C versus Cramer's index in $\mathcal{F}\{(9,11,7);(16,4,7)\}$*

Only 10 percent of tables in $\mathcal{F}$ have $C \leq 0.275428$ and only another 10 percent of them have $C > 0.603791$. Conversely, a high proportion of tables corresponds to intermediate values of this classical normalized index. The median of C is $Me(C) = 0.430513$ in the class $\mathcal{F}$. The minimum value of C in $\mathcal{F}$ is 0.083101 and the maximum value is 0.807947. Figure 18 shows the C distribution on $\mathcal{F}$.

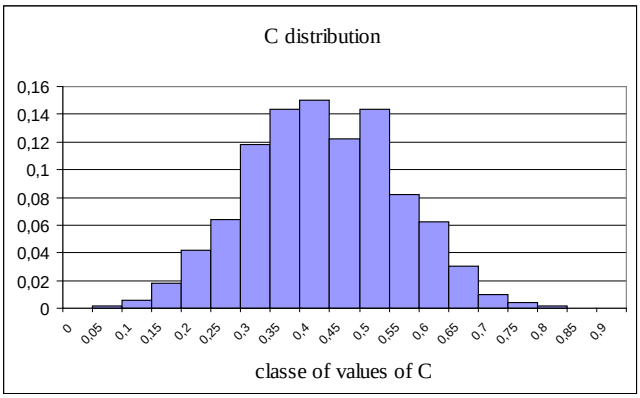

Figure 18. *The distribution of Cramer's index in $\mathcal{F}\{(9,11,7);(16,4,7)\}$*

From these examples (more cases were analysed in detail), we can argue that the indexes M′ and C overmeasure the degree of dependence in their low values: the tables in the class $\mathcal{F}$ with low values of M′ and C are really rare and their relative position expressed by $I_{M'}$ and I_C detects their extreme position in the ordering. Similarly, the indexes M′ and C tend to

underevaluate the strength of dependence in $\mathcal{F}$ providing values far from 1 on a large set of tables with high dependence.

Comparing the four diagrams representing the indexes based on the total ordering versus the classical measures M′ and C, the line bisecting the first quadrant crosses the curves for different values of M′ and C. In the four cases that we are considering, the crossing points are approximately 0.404 for M′ and 0.403 for C in $\mathcal{F}\{(9,11,7);(16,4,7)\}$, and 0.446 for M′ and 0.603 for the C plotting in $\mathcal{F}\{(42,42);(22,19,22,21)\}$. As we verified in many other cases, tables ranked in the same position in the ordering of dependence can have different values of the indexes M′ and C.

Furthermore, while $M_1(|\rho|)$, being the order of magnitude of the relative deviation from independence, have an immediate meaning, the same can not be said for Mortara's M′ and for C, in the context of the class $\mathcal{F}$. Their lack of a clear interpretation, particularly for intermediate values (where they are dense and they can poorly distinguish between a high proportion of tables) and for nearly extreme values (other than the unity and the null value), suggests better to use $M_1(|\rho|)$ jointly with the indexes proposed in this work. For example, in case of $M_1(|\rho|) = 0.699588$ and $I_{M'} = 0.9$ for an assigned table *T*, we can argue that:

- the joint frequencies in the table differ from the independence frequencies, in average, of 70 percent of their value, but
- the 90 percent of tables in the same class $\mathcal{F}$ have a degree of dependence - measured by $M_1(|\rho|)$ or by M′- lower than *T*.

7. Concludine remarks

The relative position a table assumes in a chosen dependence ordering is a meaningful index of dependence, denoted by $I_f(T)$, whose interpretation is immediate and straightforward.

The properties of $I_f(T)$ are:

- it is normalized;
- it attains the extreme values in correspondence with the extreme situations of dependence, constrained by the given margins;
- it inherits all invariance properties the ordering induced by *f* has (as bivariate distributions of qualitative variables require);
- it allows comparisons between cross-classification tables with different dimensions (*r* and *c*) and population size *N*;
- it behaves as it possesses a sort of invariance property with respect to

similar populations: as αN reaches some hundreds $I_f(T)$ attains a substantial stabilization, presenting fluctuations only on its second decimal.

This last property can be very useful for evaluating the index in the two opposite situations of bivariate distribution referring to small population sizes and, in the reverse mode, facing computational complexity for cross classification tables referred to populations whose size passes the threshold that assures the stability of $I_f(T)$.

The analysis of the relation between M′ and C w.r.t. the corresponding indexes $I_{M'}$ and I_C shows that the former measures are remarkably concentrated on their intermediate values. They assign a narrow interval of central values to a wide set of tables, poorly discriminating between their degree of dependence. On the other side, only a low proportion of tables corresponds to the extreme values of those indexes. Hence, in case of very high strength of dependence (respectively very low) the indexes C and M′ can not reveal this situation by their values, while $I_{M'}$ and I_C depict it precisely.

Acknowledgements

The author wish to thank M.Zenga for his continuous suggestions on a preliminary draft of the paper and the Italian MIUR for financial support, within the project: 'Descriptive and Inferential Aspects for Categorical Data Analysis', led by professor A. Forcina, University of Perugia (COFIN prot. 2002133957_004). Sincere thanks to an anonymous referee for his constructive remarks and comments.

References

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

Greselin F. (2003). Counting and enumerating frequency tables with given margins. *Statistica & Applicazioni,* **I**, 87-104.

Greselin F. and Zenga M. (2004a). A partial ordering of dependence for contingency tables. *Statistica & Applicazioni,* **II**, 53-71.

Greselin F. and Zenga M. (2004b). Partial and total orderings of dependence on tables with given margins. *Quaderni di Statistica,* **6**, 129-155.

Goodman L.A. and Kruskal W.H. (1954). Measures of Association for Cross Classification. *Journal of the American Statistical Association,* **49**,723-764.

---------- (1959). Measures of Association for Cross Classification, II: Further Discussion and References. *Journal of the American Statistical Association,* **54**, 123-163.

---------- (1963). Measures of Association for Cross Classification, III: Approximate Sampling Theory. *Journal of the American Statistical Association,* **58**, 310-364.

---------- (1972). Measures of Association for Cross Classification, IV: Simplification of Asymptotic Variances. *Journal of the American Statistical Association,* **67**, 415-421.

---------- (1979). *Measures of association for cross classification*. Springer series in Statistics, Springer-Verlag, New York.

Kendall M. and Stuart A. (1979). *The Advanced Theory of Statistics, vol 2: Inference and Relationship*. 4th edn. New York: Macmillan.

Mortara G. (1922). *Lezioni di statistica metodologica.* Città di Castello, Società tipografica "Leonardo da Vinci".

Pearson K. (1904). Mathematical Contributions to the Theory of Evolution. XIII. On the Theory of Contingency and Its Relation to Association and Normal Correlation. *Drapers' Company Res. Mem., Biometric Ser.,* **1**.

Zanella A. (1988). *Lezioni di Statistica*, parte II: Strutture di dati in due o più dimensioni, sez. II: La Connessione. Vita e Pensiero, Milano.