XLI Riunione Scientifica (Milano, 5-7 giugno 2002)
ISBN 88-7178-589-4
Editore: CLEUP - Padova

Measures of association in the Fréchet class[1]

Misure di Associazione nella Classe di Fréchet

Francesca Greselin, Michele Zenga[*]

Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali
Università degli Studi di Milano Bicocca, P.zza dell'Ateneo Nuovo 1, Mi, I-20126, Italy
francesca.greselin@unimib.it, michele.zenga@unimib.it

Riassunto:

Data una tabella a doppia entrata T, è possibile introdurre ordinamenti parziali e totali di associazione nella classe di Fréchet cui T appartiene. La posizione relativa che la tabella assume in tali ordinamenti può essere una significativa misura di associazione, in quanto permette di discriminare, tra tutte le tabelle della classe, quante di esse rispecchiano una situazione di maggiore indipendenza tra i caratteri e quante, al contrario, riflettano un più alto grado di dipendenza. Al variare dell'ordinamento scelto, si definisce una famiglia di indici, che gode di interessanti proprietà: essi sono coerenti con gli ordinamenti e autonormalizzati per definizione, sono invarianti a permutazioni di righe e colonne, assumono i valori estremi sulle situazioni estreme di dipendenza, inoltre si comportano come asintoticamente invarianti su distribuzioni simili.

Keywords: Association ordering; Fréchet class; Association measure**.**

1. Introduction

When populations are cross-classified with respect to two or more qualitative characters, the question often arise of the degree of association existing between the several polytomies. The main unsolved problem with traditional measures of association is that it is difficult to meaningfully compare their values. Given a particular r x c table T, our approach to association starts from the consideration of its entire Fréchet class $\mathcal{T}(n_{i\bullet}\ ;\ n_{\bullet j})$: chosen a particular association ordering for all the bivariate distributions, as defined in Zenga and Greselin 2001, the relative position assumed by T in $\mathcal{T}(n_{i\bullet}\ ;\ n_{\bullet j})$ can be a meaningful and sound index of association. Moreover, this family of indexes has several desirable and interesting properties that we will show in detail.

2. Terminology

A double polytomy between characters A and B over a population of N individuals may be represented by a double table, where classification A divides the population into the c classes $a_1, a_2,\ldots a_c$ and classification B divides the population into the r classes $b_1, b_2,\ldots b_r$:

[1] The present paper is financially supported by the project COFIN 2001 prot.2001137492.

[*] The present work reflects the common thinking of the authors, even if, more specifically, M.Zenga wrote the Introduction and the Concluding remarks, while F.Greselin wrote the remaining paragraphs.

B\A	a_1	…	a_j	…	a_c	
b_1	n_{11}	n_{12}				$n_{1\bullet}$
…						…
b_i	n_{i1}		n_{ij}			$n_{i\bullet}$
…						…
b_r						$n_{r\bullet}$
	$n_{\bullet 1}$	…	$n_{\bullet j}$	…	$n_{\bullet c}$	N

Characters A and B are supposed qualitative and without ordering among the categories. In the hypothesis of independence between characters A and B, the joint frequencies are given by: $\hat{n}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{N}$ $i = 1,\dots,r$ and $j = 1,\dots,c$. Conversely, complete association refers to the case in which each modality of A uniquely identifies one of B: in any column we have only one joint frequency not null, and hence c ≥ r.

Measures of association are single summary numbers that describe the type and the extent of relationships between two classified variables. They differ in defining intermediate stages of association. Except in the 2 x 2 case, a single measure of association cannot reflect the large variety of ways in which a table can depart from independence. It is this fundamental mathematical fact, connected to the degrees of freedom, that leads to the variety of measures and to the inherent difficulty in choosing a single measure on any given situation.

3. Our proposal

Among the three distinct possibilities that classically arise in the study of association between characters A and B, with regard to the typology of statistical data (Leti (1983), Zanella (1988)) we choose to consider as fixed both marginal frequencies. Let us recall:

Definition 1: The Fréchet Class (see, for its first definition, Fréchet (1951))

The set of all double tables, with the same given marginal frequencies {$n_{i\bullet}$,i=1,…,r; $n_{\bullet j}$,j=1,…,s}, constitutes the Fréchet class, denoted by $\mathcal{T}$ ($n_{i\bullet}$; $n_{\bullet j}$).

Let T be the generic table pertaining to $\mathcal{T}$ ($n_{i\bullet}$; $n_{\bullet j}$).

Our approach moves his steps defining association orderings in the Fréchet class:

Definition 2: Total ordering on the Fréchet class

Let f be any function f: $\mathcal{T}$ ($n_{i\bullet}$; n_j) → $\Re$ that associates a real number to each table in $\mathcal{T}$ ($n_{i\bullet}$; n_j). f induces a total ordering relation on $\mathcal{T}$ ($n_{i\bullet}$; n_j), denoted by $\le_f$: ∀ M, P ∈ $\mathcal{T}$ ($n_{i\bullet}$; $n_{\bullet j}$) M precedes P according to $\le_f$, writing: $M \le_f P$, if and only if: $f(M) \le f(P)$.

Definition 3: The family of indexes $I_f(\mathrm{T})$

Given a bivariate distribution T, the relative position T assumes among $\mathcal{T}$ ($n_{i\bullet}$; $n_{\bullet j}$), according to the sorting induced by f, is a measure of association $I_f(\mathrm{T})$, for T:

$$I_f(\mathrm{T}) = \frac{\#\left\{\mathrm{M} \,\middle|\, \mathrm{M} \in \mathbf{T}(n_{i\bullet}\,; n_{\bullet j})_{\prec}\,;\ \mathrm{M} \le_f \mathrm{T}\right\}}{\#\left\{\mathrm{M} \,\middle|\, \mathrm{M} \in \mathbf{T}(n_{i\bullet}\,; n_{\bullet j})\right\}}$$

Observe that I_f is autonormalized, being a *relative frequency*[2,3].

The family of indices I_f (T) possesses the following desirable properties:

- they assume values in [0,1];
- they are invariant to permutation of rows or columns;

[2] The function f: $\mathcal{T}$ ($n_{i\bullet}$; $n_{\bullet j}$) → $\Re$, for example, can be X^2, or $\mathrm{M}_1(|\rho|)$ or $\mathrm{M}_2(|\rho|)$. Those expressions are non-negative strictly increasing functions on each (relative or) absolute contingency.

[3] Similarly, we can define a second family of indices $I_{\prec}$(T), based on the relative position the table T assumes in a partial ordering $\prec$, defined over $\mathbf{T}$($n_{i\bullet}$; $n_{\bullet j}$), as done in Zenga and Greselin 2001.

- they attain their minimum value, only if T is the nearest table to the table of independence;
- they assume their maximum value 1 only in the cases of maximum dependence, compatibly with the fixed margins.

Moreover, they are relative frequencies, so their meaning is straightforward. A value of $I_{X^2}(\mathrm{T})$=.45 means that, among all bivariate distributions in $\mathbf{T}(n_{i\bullet} ; n_{\bullet j})$ sorted by non-decreasing values of X^2, T lies in a position that divides the class in two parts, 45% of tables in $\mathbf{T}(n_{i\bullet} ; n_{\bullet j})$ has a value of $X^2 \leq X^2_T$ and the remaining tables have $X^2 > X^2_T$.

The ***case of minimum dependence*** deserves a particular care. The observed frequencies n_{ij} are integers, while the frequencies of independence $\hat{n}_{ij}$ are rational numbers. In the very rare case in which the table of independence $\hat{\mathrm{T}}$ belongs to the Fréchet class $\mathbf{T}(n_{i\bullet} ; n_{\bullet j})$, the numerators of $I_f(\mathrm{T})$ attain their minimum value, i.e.1, if T $=\hat{\mathrm{T}}$. Generally, in a given Fréchet class there can be more than one table 'near' to the independence table.

Definition 4: Minimum constrained dependence

We say that T is a table nearest to the independence table $\hat{\mathrm{T}}$ according to the function f if and only if: $\min\limits_{\mathrm{P}\in T(n_{i\bullet};n_{\bullet j})} f(\mathrm{P}) = f(\mathrm{T})$.

If T is a table nearest to $\hat{\mathrm{T}}$ according to f, $I_f(\mathrm{T})$ approach the value 0 from positive values while the size of the class $\mathbf{T}(n_{i\bullet} ; n_{\bullet j})$ increases (i.e. r, c or N increase).

The second extreme situation is in presence of ***complete*** or ***absolute association.*** These cases are not always compatible with a set of given margins. So we need an analogous definition for the *maximum constrained dependence,* and for the farthest table from independence. In this case the functions f attain their maximum values, so that $I_f(\mathrm{T}) = 1$.

The $I_f(\mathrm{T})$ also possess ***invariant properties***, as traditional indexes do. They are invariant to permutation of rows and columns and we would expect them to be invariant also with respect to similar distribution.

Definition 5: α-similar distribution. Let T be a given bivariate table T=$\{n_{ij}\}$ and let α be a natural number. The α-similar distribution with respect to T is the bivariate distribution, denoted by αT, with size αN, defined by the joint frequencies $n_{ij}^{(\alpha)} = \alpha \cdot n_{ij}$.

In the case of low size populations ($N \approx 10,..,100$) this property doesn't hold for $I_f(\mathrm{T})$ as the multiplication of a table by an integer factor α modifies in an unpredictable direction $I_f(\mathrm{T})$. Considering the granularity of the distributions in the Fréchet class $\mathbf{T}(n_{i\bullet} ; n_{\bullet j})$, due to the requirement that frequencies are integer numbers, we expect that this anomaly disappears as N increases. We verified that $I_f(\mathrm{T})$ attain a substantial stabilization, presenting only fluctuations on their second decimal, whenever the factor α or N are as high as αN reaches some hundreds.

4. Computational results

As an example, let us show how $I_f(\mathrm{T})$ behaves starting with the table:

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

Successively multiplying its internal frequencies by a positive integer factor α, we reached interesting results.

A program[4] has been developed to generate the entire Fréchet class for each bivariate distribution and calculate the indexes based on the relative position, choosing $f = \mathrm{M}_1(|\rho|)$ and $f = \mathrm{M}_2(|\rho|)$. The corresponding values for the Mortara's normalized index and the Cramer index are M'= 0.6 and C = 0.68313.

Computation time in s:c	α	α **N**	$I_{M_1}(|\rho|)$	$I_{M_2}(|\rho|)$
< 0:01	1	10	0.75000	0.75000
< 0:01	3	30	0.85000	0.90000
< 0:01	5	50	0.87500	0.91667
0:05	10	100	0.91202	0.93548
0:06	20	200	0.91413	0.95004
0:33	50	500	0.91560	0.95559
0:87	80	800	0.91599	0.95687
1:10	90	900	0.91606	0.95758
1:38	100	1000	0.91612	0.95790

We can observe that the two indices $I_{\mathrm{M}_1}(|\rho|)$ and $I_{\mathrm{M}_2}(|\rho|)$ attain a substantial stabilization as αN reaches some hundreds.

5. Concluding remarks

The introduction of partial and total orderings between tables in a given Fréchet class naturally suggests considering the relative position a table assumes in a chosen association ordering as a meaningful index of association. Discussing the properties of those indexes, we have shown that, as measures of association in a Fréchet class, they are normalized: they attain their extreme values in correspondence with extreme situations, constrained to the given margins, they are invariant to permutations of rows and columns, they behave as they were asymptotically invariant to similar distributions.

In social sciences, researchers generally work with a high number of observations, and their data usually refer to populations whose size overcomes the threshold that assures the stability of I_f(T). In these contexts, the property of immediate interpretation, provided by our indices, joint with the possibility of meaningfully comparing their values, could be an important improvement in measuring association.

In presence of bivariate distributions referring to small populations (N < 200), we can stabilize the index in the following way: generating the multiple table αT, with α such that αN > 200, the index $I_f(\alpha\mathrm{T})$ is a good measure of association for T.

Conversely, in presence of large populations, we use the property of invariance in the reverse mode: the approximated table αT' with $\alpha\mathrm{N} \approx 200$ allows us to obtain $I_f(\alpha\mathrm{T}')$ in a considerably reduced time and $I_f(\alpha\mathrm{T}')$ is a good approximation of the exact index.

References

Fréchet, M. (1951) *Sur les tableaux de corrélation dont les marges sont donnés*, Annales de l 'Université de Lyon, II Fac., 4, Sciences, 1951.

Greselin F. (2001), *Counting and enumerating rectangular arrays with given margins*, presented to be published by Discrete Mathematics, Elsevier Science.

Goodman Leo A., Kruskal William H.(1979), *Measures of association for cross classification*, Springer series in Statistics, Springer-Verlag, New York.

Leti, G. (1983), *Lezioni di Statistica*, Il Mulino, Bologna.

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

Zenga M., Greselin F. (2001), *Partial and total orderings on tables in the Fréchet class*, to be submitted for publishing.

[4] See Greselin (2001) for more details about the software program developed and adapted for this research.