ALFREDO RIZZI (*)

Indices of agreement relative to categorical variables

1. INTRODUCTION

Let $X_{n,k}$, be a data matrix of n statistical units and k categorical variables that can only take the value 0 or 1. In general, the k variables may be correlated as is the case for instance when the k variables code the opinions given about the various services offered by a company. An example of what may be asked in a questionnaire which yeld the matrix $X_{n,k}$ is as follows:

- Do you think party A is innovative?
  - Yes, very much
  - Quite
  - A little
  - No, not at all
  - I don't know
- Do you think party A addresses issues of interest to youth?
  - Yes, very much
  - Quite
  - A little
  - No, not at all
  - I don't know

(*) Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università «La Sapienza», Roma.

– Do you think party A seeks to precede new issues of interest to civil life?

- Yes, very much
- Quite
- A little
- No, not at all
- I don't know

A dichotomic codification may associate the number code 1 to those answers that are substantially positive such as "Yes, very much" and "Quite" while assigning the number code of 0 to those answers which are negative such as "A little", "No, not at all", and "I don't know". This binary codification makes it possible to eliminate the subjective interpretation of the terms which result in their not being read in a homogeneous manner by the interviewed persons. It also make possible to eliminate a subjective interpretation of the questions (from the side of the interviewed persons) as well as a subjective interpretation of the answers (from the side of the interviewers). This is important as well as having more possibilities when answering gives the respondents a feeling of greater freedom of expression.

The responses to the questions, moreover, are correlated in that a person who believes a party is innovative will probably feel that the same party addresses issues of interest to youth and society in general.

The index of "perceived quality" measures the rating with which the opinion has been expressed and therefore the quality as perceived by the respondent. The "opinion" is the index that could regard any other aspect of the inquiry, or any other thing.

An index of quality can be calculated, by adding up all of the positive responses and dividing the number by the total number of people interviewed multiplied by the number of variables:

$$Q = \frac{R}{nk}$$

where R is equal to the sum of all positive responses. This index varies between 0 and 1. In this manner, however, the correlation that exists between the variables is not taken into account in the sense that positive results originated from other positive results (as is the case for negative results). For this reason, then, the value of the index should be reappraised. In this paper we will study some of the indices that eliminate the presence of correlation between the variables.

The indices proposed here refer to research situations in which categorical variables are present and therefore, especially for nominal scales, derived by a process of specific measuring. The term "measuring" in the social sciences is used, as is well known, with a wider meaning than that used by the natural and physical sciences, etc. Here, "measuring" means the rules of assignation of the codes (number or words) as a property of objects or events in such a way as to make it possible to attribute some characteristics of the members to the properties themselves. Within the orbit of social phenomena, understood in the same way, the attribution of a measure is often "indirect" (or assigned according to standardized rules for measuring), as is the case, for example, for attributes such as prestige, quality of a service, etc. These concepts are highly abstract and can therefore, be expressed differently by different people.

A dichotomic codification is the first and most basic classification of a variable; it is always possible and despite implying the simple algebraic structure of equal or different, it allows for sophisticated statistical elaborations which are based on the data derived from codes having a very reduced subjectivity level.

2. Methodology

The matrix $X_{n,k}$ is relative to a set of n statistical units and k variables that can only take the value 0 or 1.

The matrix $X_{n,k}$ will be, for istance, as follows:

$$X_{n,k} \equiv \begin{pmatrix} 1 & 0 & 1 & 0 & \dots & 1 \\ 0 & 1 & 0 & 1 & \dots & 1 \\ \dots & & & & & \\ 0 & 1 & 1 & 0 & \dots & 1 \end{pmatrix} \qquad (1)$$

$R_{k,k}$ is the correlation matrix of $X_{n,k}$ that is obtained, as is known, from the transformation:

$$R_{k,k} = D_{k,k}\, \Sigma_{k,k}\, D_{k,k}$$

where:

– $D_{k,k}$ is the diagonal matrix, where the main diagonal has as elements the inverses of the standard deviation of the k variables;

– $\Sigma_{k,k}$ is the matrix of variance and covariance of $X_{n,k}$ and is equal to:

$$\frac{1}{n} S'_{n,k} \cdot S_{n,k}$$

where $S_{n,k}$ is the matrix of the difference from the mean of $X_{n,k}$ and the transposed matrix of $S_{n,k}$ has been indicated with an apex.
With the aim of eliminating the correlation of the variables, any orthogonal transformation which transforms the matrix $X_{n,k}$ with the correlation matrix $R_{k,k}$ in $Y_{n,k}$ with a matrix of correlation $I_{k,k}$, where $I_{k,k}$ is diagonal with all ones in the principal diagonal and zero in the other cells, can be carried out. In this way one gets variables that are not correlated with each other. Among the infinite number of orthogonal transformations that transform $X_{n,k}$ in $Y_{n,k}$ one can consider:

$$Y_{n,k} = Z_{n,k}\, Q_{k,k}\, \Lambda_{k,k}^{-\frac{1}{2}}$$

where:

$Z_{n,k}$ is the transformation of $X_{n,k}$ standardized per column (zero mean and unitary variance);

$Q_{k,k}$ is the matrix of the normalizing eigeinvectors (such that the sum of the squared components is equal to one);

$\Lambda_{k,k}^{-\frac{1}{2}}$ is the diagonal matrix, on the main diagonal of which, are the inverses of the square roots of the eigein values of $R_{k,k}$.

This implies that the eigeinvalues of $R_{k,k}$ are positive, as occurs when $R_{k,k}$ is definitely positive and therefore no coefficient of linear correlation is ± 1.

Furthermore, even the transformation of the principal components allows for the transformation of the $X_{n,k}$ matrix into a $Y_{n,k}$ matrix using non correlated components ordered according to decreasing variance. The principal k components given:

$$Y_i = Z_{n,k} \cdot a_i \quad (i = 1, 2, \ldots, k) \tag{3}$$

where $Z_{n,k}$ is the matrix of the standardized difference and a_i is the eigein vector which corresponds to the i-mo eigein value λ_i of $R_{k,k}$ in decreasing order.

The presence of correlation is not inconvenient. It is, however, when the questions are fomulated in such a way as to regard aspects of the subject under study that overlap. The methodology for arriving at pure data although somewhat artificial, makes it possible to arrive at variables for which it is not always possible to give an exact interpretation.

Let 1 be the vector of ones having all components equal to:

$$1 \equiv (1, 1, \dots, 1)$$

that is, for all favorable answers to questions supplied by a respondent and inversely with 0 vector having all the components equal to 0 (complete disagreement or a don't know answer to all questions posed). With the aim of defining an index of perceived quality, we consider the matrix in which only one statistical units is characterized by the vector 1 and $n-1$ by the vector 0, that is:

$$X_{n,k} \equiv \begin{pmatrix} 1, 1, \dots, 1 \\ 0, 0, \dots, 0 \\ \dots \\ 0, 0, \dots, 0 \end{pmatrix} \tag{4}$$

For such a matrix the vector of the mean is equal to:

$$\overline{X} \equiv \left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n} \right) \tag{5}$$

and that of variance, as shown, with a series of simple transformations:

$$\sigma^2 \equiv \left(\frac{n-1}{n^2}, \frac{n-1}{n^2}, \dots, \frac{n-1}{n^2} \right) \tag{6}$$

The correlation matrix is the $U_{k,k}$ or the matrix formed by all ones.
The matrix (4) is indicated as "the maximum distance matrix".
The eigein values will be:

$$\lambda_1 = k;\ \lambda_2 = \lambda_3 = \dots = \lambda_k = 0$$

In fact, these are obtained from the expression:

$$\begin{vmatrix} 1-\lambda & 1 \dots 1 & 1 \\ 1 & 1-\lambda \dots 1 & 1 \\ \dots & & \\ 1 & 1 \dots 1 & 1-\lambda \end{vmatrix}$$

The determinant at first member is a particular case of:

$$|\Delta| = \begin{vmatrix} a_1 & x \dots x & x \\ x & a_2 \dots x & x \\ \dots & & \\ x & x \dots x & a_k \end{vmatrix}$$

which is the same as:

$$|\Delta| = (a_1 - x)(a_2 - x) \dots (a_k - x)\left(\frac{x}{a_1 - x} + \frac{x}{a_2 - x} + \dots + \frac{x}{a_{k-1} - x} - \frac{a_k}{a_{k-1}}\right) =$$

$$= x(a_1 - x) \dots (a_k - x)\left(\frac{1}{x} + \frac{1}{a_1 - x} + \dots + \frac{1}{a_k - x}\right).$$

In fact, if one subtract the k-th column from the first $(k-1)$ column, putting in evidence the factors: $(a_1 - x)$, $(a_2 - x)$, ..., $(a_k - x)$ respectively in the first, second, ..., k-th line, and adding to the last line the sum of all the others, one obtains a determinant in which all of the elements below the principal diagonal are zero: with further algebraic simplifications one obtains the previous formula.

By putting in Δ the value $x = 1$ and $a_1 = a_2 = \dots = a_k = 1 - \lambda$, one gets:

$$|\Delta| = (-\lambda)^k\left(1 + \frac{1}{-\lambda} + \dots + \frac{1}{-\lambda}\right) = (-1)^k \lambda^k \left(\frac{\lambda - k}{\lambda}\right) = (-1)^k\, \lambda^{k-1}(\lambda - k)$$

from which only one eigein value is equal to k and $(k-1)$ are zero.

The same result could be arrived at by considering the characteristic equation of the matrix:

$$\lambda^k + c_1 \lambda^{k-1} + \ldots + c_k = 0$$

and by remembering (Rizzi 1985) that the coefficient c_r $(r = 1, 2, \ldots, k)$ of the characteristic polynomial is the same as the sum of all the *principal minor* of r of Δ multiplied by $(-1)^r$. Given that with:

$$c_k = |U| = 0$$
$$c_2 = c_3 = \ldots = c_{k-1} = 0$$
$$c_1 = (-1) + rU = -k$$

one gets:

$$|U - \lambda I| = \lambda^k - k\lambda^{k-1}$$

The first eigeinvector (corresponding to $\lambda = k$) is obtained by solving the homogenous system of k equations in the unknown k; a_i $(i = 1, 2, \ldots, k)$ $(R - kI)\ a = 0$
that is

$$\begin{cases} (1-k)a_1 + a_2 + \ldots + a_k = 0 \\ a_1 + (1-k)a_2 + \ldots + a_k = 0 \\ \ldots \\ a_1 + a_2 + \ldots + (1-k)a_k = 0 \end{cases}$$

which gives the solution:

$$a_1 = a_2 = \ldots = a_k.$$

The eigeinvector of normalized positive components, or that for which the sum of the squares of the components is equal to one, is

$$a_1 = \frac{1}{\sqrt{k}};\ a_2 = \frac{1}{\sqrt{k}};\ \ldots;\ a_k = \frac{1}{\sqrt{k}}$$

The first principal components of the matrix (4) of maximum distance will have as elements:

$$P_1 \equiv \left(\sqrt{k(n-1)};\ -\sqrt{\frac{k}{n-1}}, \ldots;\ -\sqrt{\frac{k}{n-1}} \right)$$

which is obtained from the equation:

$$P_1 = Z_{n,k} \cdot a_1 = \frac{1}{\sigma} \begin{Vmatrix} 1-\frac{1}{n} & 1-\frac{1}{n} & \dots & 1-\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ & \dots & & \\ -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \end{Vmatrix} \begin{Vmatrix} \frac{1}{\sqrt{k}} \\ \frac{1}{\sqrt{k}} \\ \frac{1}{\sqrt{k}} \end{Vmatrix}$$

where $Z_{n,k}$ is the matrix of the standardized differences per column and

$$\sigma^2 = \frac{n-1}{n^2}$$

The variance of the first principal components is equal to k.

The other $k - 1$ principal components have null elements, in that the corresponding eigeinvalue is zero and only one component expresses all of the variability of the data matrix.

In essence, the transformation of the principal components makes it possible to pass from the matrix of maximum distance:

$$X_{n,k} \equiv \begin{pmatrix} 1, 1, \dots, 1 \\ 0, 0, \dots, 0 \\ \dots \\ 0, 0, \dots, 0 \end{pmatrix}$$

to the matrix:

$$P_{n,k} \equiv \begin{pmatrix} \sqrt{k(n-1)}, 0, 0, \dots, 0 \\ -\sqrt{\frac{k}{n-1}}, 0, 0, \dots, 0 \\ \dots \\ -\sqrt{\frac{k}{n-1}}, 0, 0, \dots, 0 \end{pmatrix}$$

where k are the variables and n the statistical units.

The term D_1 is used to indicate the sum of the differences, in absolute value, of the principal components of the $(n-1)$ statistical units from those of a statistical unit having unitary components. One calculates, using the data in matrix (4), the sum of the absolute values of the distances of the principal components from the units having unitary components. Similarly, D_2 is the square root of the sum of the squared differences. D_1 is the distance of the "city" blocks and D_2 is the Euclidean distance.

With the terms D_{1_j} and D_{2j}, we calculate the distance between unit 1 and respectively, by the distance from the city blocks and Euclidean distance for the variable *j*.

For the sum of the distances of the $(n-1)$ statistical units from the first unit, that of maximum consensus, we get:

$$\max D_1 = \sum_{i=2}^{n} \left| D_{1i} \right| = (n-1)\left[\sqrt{\frac{k}{n-1}} + \sqrt{k(n-1)}\right] = n\sqrt{k(n-1)}$$

for the "city blocks" and for Euclidean distance it is as follows:

$$\max D_2 = \sum_{i=2}^{n} \sqrt{D_{2i}^2} = \sqrt{\left(\frac{\sqrt{k}}{\sqrt{n-1}} + \sqrt{k}\sqrt{n-1}\right)^2 (n-1)} = n\sqrt{k}$$

3. Indices of perceived quality

An index of "perceived quality" can be defined as follows:

$$Q_1 = 1 - \frac{\sum_{i=2}^{n} D_{1i}}{\max D_1} = 1 - \frac{D_1}{\max D_1} = 1 - \frac{D_1}{n\sqrt{k(n-1)}}$$

where D_{1_i} is equal to the sum of the absolute values from the distance between unit *i* and the unit which has all the unitary components, calculated in the matrix of principal components. Obviously:

$$0 \leq Q_1 \leq 1$$

Similarly, it can be defined as:

$$Q_2 = 1 - \frac{\sum_{i=2}^{n} D_{2i}}{\max D_2} = 1 - \frac{D_2}{\max D_2} = 1 - \frac{D_2}{n\sqrt{k}}$$

If we consider a total numerosity equal to a cn with c being constant, the maximum of D_1 becomes:

$$\max D_1 = cn\sqrt{k(cn-1)}$$

If the numerator in Q_1 becomes $D_1' = cD_1$, as in the case where each statistical units has a "weight" equal to c, the Q_1 index assumes the value:

$$Q_1 = 1 - \frac{cD_1}{cn\sqrt{k(cn-1)}} = 1 - \frac{D_1}{n\sqrt{k(cn-1)}}$$

Thus, in order to compare two indices relative to the numerosities n_1 and n_2 with $n_2 > n_1$, we must consider:

$$Q_1 = 1 - \frac{D_1}{n_1\sqrt{k(n_1-1)}}$$

$$Q_1^1 = \frac{1 - D'_1 \frac{n_1}{n_2}}{n_2\sqrt{k(n_{2-1})}} \sqrt{\frac{n_2-1}{n_1-1}\,\frac{n_2}{n_1}} = 1 - \frac{D'_1}{n_2\sqrt{k(n_1-1)}}$$

Q_1^1 is obtained directly from Q_1 by multiplying the second addendum by the relation of the maximum that refers to the units n_1 and n_2 respectively.

It should be noted that Q_1 can assume a negative value if the ratio $\frac{n_2}{n_1}$ is particularly high.

For the index Q_2 one has, by the transformation from an n numerosity to a numerosity equal to cn, the maximum which becomes $cn\sqrt{k}$. Therefore, the index Q_2 makes it possible to compare situations with different numerosities.

It should be noted that when k_1 and k_2 respectively are considered in two different situations, in an instance of equality between n, the maximums of both Q_1 and Q_2 will vary in the relation $\sqrt{k_1} / \sqrt{k_2}$. The relation between the relative value of D_1 (or of D_2) should vary only slightly given that when increasing the number of variables, the importance of the latter components diminishes. Nothing precise however, can be said.

In practice, particularly in studies that regard the perception of quality of services to a broad public, it is difficult that the absolute maximum of disagreement will be reached with respect to an ideal situation. Thus, both the indices Q_1 and Q_2 tend to assume rather high values. In this case it is possible to follow two different paths.

In the first, the index of disagreement is considered

$${}_1Q = \frac{D_1}{\max D_1} \qquad \text{or} \qquad {}_2Q_1 = \frac{D_2}{\max D_2}$$

and the index is used to measure the temporal variations of disagreement by considering the relation:

$$\frac{{}_1^tQ_i}{{}_1^{t-1}Q_i} \qquad \text{or} \qquad \frac{{}_2^tQ_i}{{}_1^{t-i}Q_i}$$

where ${}_1^tQ_i$ is the index of quality at a certain time t ($i = 1, 2$).

In the second, a percentage of the plausable absolute maximum is fixed (empirically) and the index is then calculated:

$$Q_1 = 1 - \frac{D_1}{p_1 \max D_1} \qquad\qquad Q_2 = 1 - \frac{D_2}{p_2 \max D_2}$$

where p_1 and p_2 range between 0 and 1 and are established subjectively, based on the knowledge of the statistician to other temporal and spatial situations or similar research etc. Naturally, p_1 and p_2 must remain constant in order to understand the development of the phenomenon which is of course, the most important aspect of the process.

4. Applications

In order to calculate the indices that are proposed here, it is sufficient to utilize a program for principal component analysis, such as SAS or SPSS, that can be operated in tandem with the "file" of the respondents answers and has a first record made up of all codes relative to "positive" dichotomic answers (all ones for example). On the output of the principal components it is easy to calculate D_1 (or D_2) as the sum of the difference in absolute value of the single components of $n - 1$ statistical unit from those of the first record artificially composed and formatted by all ones. The maximum of D_1 (or D_2) is easily ottainable.

This approach was tried with a sample group of 1,000 telephone users mangaged by Italy Telecom. The study was carried out by a marketing research company on the behalf of Telecom in march, 1994. Twelve indicators were considered and were as follows:

Satisfaction with the service, Overall quality, Telecom efforts with new lines, Introduction of new services, Bureaucratic procedures, Attention to customer needs, Bill paying, Evaluation of effectiveness of service, Availability of Information, Facility in speaking with a Telecom representative, Evaluation of communication with Telecom, Evaluation of discussion about Telecom.

Each of the questions posed could be answered in five different ways (two were relative to a good opinion on the quality of the service, two were negative and one was don't know-missing). These answers were then codified in either "1" (for the first two answers) or "0" for the other three. The matrix of maximum distance (4) was composed of 1,001 rows (the first was all "1" and the other 1,000 were of "0" and "1" according to the opinion supplied by the respondents regarding the quality of the service).

By applying the procedure of principal components, a matrix of 1,001 rows with 12 independent columns, was obtained. The sum of the absolute values of the difference between the various elements for each of the 1,000 units from that in the first row (that in which a $X_{1000.112}$ alla correspondent to "1") was:

$$D_1 = 28878.7$$

For the maximum of D_1 you got:

$$\max D_1 = n\sqrt{k(n-1)} = 109489,7$$

Thus the index:

$$Q_1 = 1 - \frac{D_1}{\max D_1}$$

is equal to 73,6%.

This value of the index allows one to evaluate the variations in the quality of services as perceived for example, by students in different universities, or different academic years.

The value of the index calculated on the basis of only one sample does not permit making generalizations on the opinions expressed on a service supplied by the university for example, because reference is being made to an absolute maximum of "distance", and it is unrealistic that the absolute maximum of "disagreement" will ever be attained with respect to and ideal situation (par. 3).

In general, the index Q_1 is more sensitive than Q_2 in measuring the substantial variation in the variables under consideration.

Finally, in its application to empirical situations, sampling groups of consistent size will have to be considered so that the variability of the index not be elevated, as happens with sample groups of a thousand or more units. Naturally, this empirical affermation should be validated by analyses relative to the theoretical distribution of the index as confirmed in real situations.

REFERENCES

CHATFIELD, C. and COLLINS, A.J. (1980) *Introduction to multivariate analysis*, Chapman and Hall, London.

FRAIRE, M. (1991) *Complementi ed applicazioni di analisi dei dati*, Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università La Sapienza, Roma.

RIZZI, A. (1985) *Il linguaggio delle matrici*, La Nuova Italia Scientifica.

RIZZI, A. (1990) *Analisi dei dati*, La Nuova Italia Scientifica.

SAS INSTITUTE INC. *SAS/STAT Guide for personale computer*, version 6 edition.

Indices of agreement relative to categorical variables

SUMMARY

In this paper the A. proposes a new index of agreement relative to categorical variables. Substantially we first transform the matrix units variables in the matrix of the principal components. Then we calculate the sum of the differences, in absolute value, of the primary components of $(n-1)$ statistical units from those of a statistical unit having unitary components. This sum is divided by the maximum that is calculate in the paper.

Indici di consenso relativi a variabili dicotome

RIASSUNTO

Nel lavoro viene proposto un nuovo "indice della qualtià percepita" che misura il grado con cui un giudizio su, ad esempio, un particolare servizio è espresso con *k* items binari da *n* unità statistiche. In sostanza si sottopone la matrice ad una trasformazione che elimina la correlazione tra le variabili (componenti principali) e si rapporta la somma delle differenze, in valore assoluto, delle componenti principali di tutte le *n* unità statistiche da quelle di una unità statistiche aventi componenti unitarie, cioè di massimo consenso.

Molte applicazioni indicano che gli indici statistici proposti sono particolarmente adatti nelle situazioni concrete di ricerca.

KEY WORDS

Categorical variables; consensus.