PROCEEDINGS
OF THE UNIVERSITY
OF VAASA

DISCUSSION PAPERS 44

Jaakko Astola and Ilkka Virtanen

ENTROPY CORRELATION COEFFICIENT, A MEASURE OF STATISTICAL DEPENDENCE FOR CATEGORIZED DATA

Vaasa 1982

Osoite / Address:

Vaasan korkeakoulu
University of Vaasa

Raastuvank. 31
65100 VAASA 10
Finland

ISBN 951-683-165-6
ISSN 0358-870X

PROCEEDINGS OF THE UNIVERSITY OF VAASA

DISCUSSION PAPERS 44

Jaakko Astola and Ilkka Virtanen

ENTROPY CORRELATION COEFFICIENT, A MEASURE OF STATISTICAL DEPENDENCE FOR CATEGORIZED DATA

Paper presented at The 15th European Meeting of Statisticians (EMS' 82) in Palermo, September 13-17, 1982.

Vaasa 1982

ABSTRACT

Astola, Jaakko and *Virtanen, Ilkka* (1982). Entropy correlation coefficient, a measure of statistical dependence for categorized data. *Proceedings of the University of Vaasa. Discussion Papers* No 44, 12 p.

The main use of Shannon's *entropy* in statistics has been in measuring the dispersion of one-dimensional categorized data. However, entropy can also be defined for a two- or multi-dimensional distribution given as a contingency table. This generalized entropy, called *coentropy*, forms a basis for a measure of overall dependence between the variables in the table. By reducing the lower order entropies from the coentropy and using an appropriate scaling, such a measure of dependence, that fulfills the criteria for a well-defined correlation coefficient, can be constructed. This *entropy correlation coefficient* is introduced and then analyzed in this paper.

Jaakko Astola, Lappeenranta University of Technology, Box 20, SF-53851 Lappeenranta 85, Finland.

Ilkka Virtanen, School of Business Studies, University of Vaasa, Raastuvankatu 31, SF-65100 Vaasa 10, Finland.

1. INTRODUCTION

The concept of *entropy* has been widely used in physics and information theory. Over the years the idea has been borrowed by other disciplines and has been applied in several problem areas within the social sciences, especially in statistics, economics, business, geography and operational research. Entropy has become an important planning tool in the area of system modelling.

The use of entropy in statistics has its origin in information theory. Entropy, Shannon's measure for uncertainty

(Shannon 1948), has been especially used as a measure of dispersion for qualitative data. The dispersion of a distribution $X: (p_1, p_2, \ldots, p_n)$ can be measured by its entropy

$$(1) \qquad H = - \sum_{i=1}^{n} p_i \log p_i$$

due to the good properties of H as a measure of dispersion: H is non-negative, $H = 0$ if and only if some $p_i = 1$, and H gets its maximum value ($= \log n$) for the uniform distribution $p_1 = p_2 = \ldots = p_n = 1/n$, see e.g. Astola and Virtanen (1981, 4-10).

It is possible to calculate entropy also for a two-dimensional distribution given as frequency data. In this case entropy reveals both the dispersion of the distribution and the dependence between the two variables, see Theil (1969, 469-472), Astola and Virtanen (1981). In this paper the main results of the report Astola and Virtanen (1981), concerning the entropy-based measures of statistical dependence, are summarized: we define the concept of *coentropy*, analyze and interprete it, demonstrate its definitional analogy with covariance and especially, construct an entropy-based measure, called *entropy correlation coefficient*, for the degree of dependence between the variables. Finally, some preliminary results concerning the three-way tables are presented.

2. COENTROPY OF A BIVARIATE DISTRIBUTION

In this section we consider data which are presented as a two-way contingency table. The two variables, X and Y, to be considered are assumed to have r and c classes, respectively. Let the joint probability (or relative frequency) distribution of X and Y be (p_{ij}), $i = 1,\ldots,r$, $j = 1,\ldots,c$.

The entropies of the marginal distributions of X and Y are

$$(2) \qquad H_X = - \sum_{i=1}^{r} p_{i.} \log p_{i.}$$

$$(3) \qquad H_Y = - \sum_{j=1}^{c} p_{.j} \log p_{.j} \ ,$$

where $p_{i.}$, $i = 1,\ldots,r$ and $p_{.j}$, $j = 1,\ldots,c$ are the marginal probabilities (relative frequencies) of the distributions of X and Y, respectively.

The entropy of the joint distribution, called now *coentropy*, is defined as

$$(4) \qquad H_{XY} = - \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \log p_{ij} \ .$$

H_{XY} has also been called joint entropy (Theil 1969, 469-472) and overall entropy (Preuss 1980, 1566).

In the following the elementary properties of univariate entropy are assumed to be known. Next we list some general properties of coentropy. The proofs can be found e.g. in Astola and Virtanen (1981). First we have

$$(5) \qquad \max\{H_X, H_Y\} \leq H_{XY} \leq H_X + H_Y$$

such that equality holds in the right hand side of (5) if and only if X and Y are independent. As numerical bounds for H_{XY} we have

$$(6) \qquad 0 \leq H_{XY} \leq \log(rc).$$

From (5) and (6) we see that both the dispersion of the joint distribution (the entropies of the marginal distributions) and the degree of independence of the two variables have a contribution to the value of coentropy.

Next we shall present for the concepts entropy and coentropy an interpretation that shows the analogy of their definitions with those of variance (or its square root standard deviation) and covariance of quantitative and measurable variables, respectively.

We can write the entropy of an one-dimensional distribution $Z: (p_1,\ldots,p_n)$ also in the form

$$(7) \qquad H_Z = \sum_{i=1}^{n} p_i \log(1/p_i) .$$

Introducing a random variable $H = H(Z)$, which has the value $\eta_i = \log(1/p_i)$ when the value of Z belongs to the i'th class, we can write

$$(8) \qquad H_Z = \sum_{i=1}^{n} p_i \eta_i = E\{H\} ,$$

i.e. H_Z is expressed as the mean value of the random variable H. The quantity $\eta_i = \log(1/p_i)$ may be interpreted as the amount of uncertainty in the i'th class: the uncertainty equals zero, if p_i equals one, it increases monotonically when p_i decreases, and approaches infinity when p_i approaches zero. Entropy thus expresses the mean uncertainty appearing in the distribution. If we compare (8) with the definition of the usual standard deviation of a quantitative variable Z, i.e. with $D(Z) = [E\{Z-E\{Z\}\}^2]^{1/2}$, the analogy of these two definitions is evident. The standard deviation expresses the mean inaccuracy appearing in the distribution, the mean inaccuracy being measured as the root mean square deviation about the mean.

For coentropy (4) we get analogously to (8)

$$(9) \qquad H_{XY} = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log(1/p_{ij}) = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \eta_{ij} ,$$

where the quantity $\eta_{ij} = \log(1/p_{ij})$ may now be interpreted as the value of a two-dimensional random variable $H(X,Y)$,

as the amount of uncertainty in cell (i,j). We have again

$$(10) \qquad H_{XY} = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \eta_{ij} = E\{H(X,Y)\} ,$$

i.e. coentropy H_{XY} expresses the mean uncertainty appearing in the frequency table. The analogy with the covariance of a two-dimensional quantitative variable (X,Y), viz. with $\mathrm{Cov}(X,Y) = E\{(X-E\{X\})(Y-E\{Y\})\}$ is again evident: covariance gives the mean inaccuracy (about the mean) appearing in the distribution.

3. ENTROPY CORRELATION COEFFICIENT

As we have seen, coentropy measures both the dispersion of the joint distribution and the degree of independence between the variables. In order to get an appropriate measure for the degree of dependence, we must eliminate the effects of the marginal entropies from coentropy and change its sign. Because we are working with logarithms, the natural way to carry out these modifications is subtraction.

We define the quantity *mean dependence information*, denoted by I_{XY}, as

$$(11) \qquad I_{XY} = -(H_{XY} - H_X - H_Y) = H_X + H_Y - H_{XY} .$$

I_{XY} has also been called the expected mutual information (Theil 1969, 470) or mean information (Kullback 1959, 5). The role of I_{XY} as the mean dependence information can be justified, however, as follows. We can, after some manipulation, write

$$(12) \qquad I_{XY} = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \log(p_{ij}/p_{i.}p_{.j}) = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\, \iota_{ij} ,$$

where $\iota_{ij} = \log(p_{ij}/p_{i.}p_{.j})$ is the amount of information

about dependence in cell (i,j): if it holds for a certain cell (i,j) $p_{ij} = p_{i.}p_{.j}$ (which is the rule for all cells in the case of independent variables), the cell gives no contribution to the amount of dependence of the variables, otherwise $\iota_{ij} \neq 0$ and the cell contains some information about the dependence of the variables. From (12) we see that I_{XY} expresses the mean value of this information. Analogously to (8) and (9) we can write

$$(13) \qquad I_{XY} = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij}\, \iota_{ij} = E\{\mathrm{I}(X,Y)\}\,,$$

where $\mathrm{I} = \mathrm{I}(X,Y)$ is a two-dimensional random variable describing the dependence information of the cells.

From (11) we can also see that the definition of the mean dependence information is analogous to the definition of the product moment correlation coefficient $\rho(X,Y) = \mathrm{Cov}(X,Y)/[D(X)D(Y)]$ defined for quantitative variables: the quantities I_{XY} and $\rho(X,Y)$ are formed with the help of the two-dimensional coentropy (covariance) and the one-dimensional marginal entropies (standard deviations). In (11) we, however, instead of multiplication and division use addition and subtraction. This is, of course, due to the use of logarithms in the definition of the entropy quantities.

The following statements consider the possible values of I_{XY} and show that I_{XY} can be used as a measure of the degree of dependence. For proofs, see e.g. Astola and Virtanen (1981, 16). We have

$$(14) \qquad 0 \leq I_{XY} \leq \tfrac{1}{2}(H_X+H_Y)$$

$$(15) \qquad 0 \leq I_{XY} \leq \min\{\log r, \log c\}$$

(16) $I_{XY} = 0$ if and only if X and Y are independent

(17) $I_{XY} = \frac{1}{2}(H_X+H_Y)$ if and only if X and Y are completely dependent, i.e. $p_{i_1 j}p_{i_2 j} = 0$ if $i_1 \neq i_2$, $j = 1,\ldots,c$ and $p_{ij_1}p_{ij_2} = 0$ if $j_1 \neq j_2$, $i = 1,\ldots,r$.

The statements (14) and (15) show that I_{XY} as a measure of dependence still has some disadvantages. It is not satisfactorily scaled (we prefer scaling between 0 and 1). The maximum value of I_{XY} depends on the size and type of the frequency table (we require independence on the formation of the table). And at last, reaching of this maximum value depends on the marginal distributions (we require reaching of the maximum value independently of the marginal distributions in the case of complete dependence). We need, therefore, another derived measure for dependence that fulfills all the requirements presented above.

The new measure of dependence, called *entropy correlation coefficient* and denoted by ρ_H, is now defined as

$$(18) \qquad \rho_H = \sqrt{\frac{I_{XY}}{\frac{1}{2}(H_X+H_Y)}} = \sqrt{2\left(1 - \frac{H_{XY}}{H_X+H_Y}\right)}\,.$$

The division by $\frac{1}{2}(H_X+H_Y)$ in (18) is needed to meet the requirements set above for the final measure. The square root in the definition is not necessary from the theoretical point of view, but by magnifying variations especially near zero it gives ρ_H a behaviour which surprisingly well matches our intuitive ideas of the degree of dependence.

In the following we present the properties of entropy correlation coefficients as a well-behaving measure of dependence. They are direct consequences from the corresponding properties of the mean dependence information I_{XY}. We have

(19) $$0 \leq \rho_H \leq 1$$

(20) $$\rho_H = 0 \text{ , iff } p_{ij} = p_{i.}p_{.j} \text{ , } \forall\, i = i,\ldots,r \text{ , } j = 1,\ldots,c$$

(21) $$\rho_H = 1 \text{ , iff } \begin{cases} p_{i_1 j}\, p_{i_2 j} = 0 \text{ , } \forall\, i_1 \neq i_2 \text{ , } j = 1,\ldots,c \\ p_{i j_1}\, p_{i j_2} = 0 \text{ , } \forall\, j_1 \neq j_2 \text{ , } i = 1,\ldots,r \text{ .} \end{cases}$$

We see that ρ_H has been scaled between 0 and 1, 0 indicating full independence (property (20)) and 1 complete dependence (property (21)). By complete dependence we here mean the highest degree of dependence: if we for an individual know the class of X we also know the class of Y it belongs to, and vice versa. This degree of dependence is sometimes called absolute dependence (Kendall and Stuart 1979, 570). We can also see that the values of ρ_H are independent of the size and type of the table: ρ_H can reach all the values between 0 and 1 both in square and rectangular tables. Further, ρ_H doesnot depend on the forms of marginal distributions (the number of classes in these, the location and dispersion indices of these etc.): there are no special requirements for the marginal probabilities $p_{i.}$ and $p_{.j}$ for ρ_H to reach the end values 0 and 1. And at last, the population size has no effect on ρ_H. From the point of view of purely mathematics, it is interesting to note that ρ_H does not depend on the base of the logarithms to be used. As a summary of the properties of ρ_H we can state that for qualitative categorical variables it is difficult to find another measure of dependence that fulfills all the properties verified for ρ_H above, cf. for example the discussion in Kendall and Stuart (1979, 586-590).

4. GENERALIZATIONS TO THREE-WAY TABLES

The entropy-based concepts presented in the previous sections can be generalized to contingency tables with any number of dimensions. In the following, however, only three-way tables

are considered. This is mainly to keep the notation simple. Further, the results are to be taken as preliminary due to the ongoing unfinished research by the authors. The more spesific results will be published in a near future.

Consider three variables X, Y and Z having the joint distribution (p_{ijk}), $i = 1,\ldots,r$, $j = 1,\ldots,c$, $k = 1,\ldots,l$. The coentropy of the joint distribution is now defined as

(22) $$H_{XYZ} = -\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} p_{ijk} \log p_{ijk} \text{ .}$$

We can also calculate the coentropies H_{XY}, H_{XZ} and H_{YZ} of the two-dimensional marginal distributions and the entropies H_X, H_Y and H_Z of the one-dimensional marginal distributions analogously to the bivariate and univariate case, respectively.

In order to get an overall measure for total dependence between the three variables, we eliminate the effects of lower order dependences and dispersion. We define the *mean total dependence information* as

(23) $$\begin{aligned} I_{XYZ} &= H_{XYZ} - (H_{XY} - H_X - H_Y) - (H_{XZ} - H_X - H_Z) \\ &\quad - (H_{YZ} - H_Y - H_Z) - H_X - H_Y - H_Z \\ &= H_X + H_Y + H_Z - H_{XY} - H_{XZ} - H_{YZ} + H_{XYZ} \text{ .} \end{aligned}$$

The role of I_{XYZ} as the mean total dependence information can be justified analogously to the bivariate case (equations (12) and (13)). It is also possible to show that I_{XYZ} has the following properties

(24) $$-\tfrac{1}{3}(H_X + H_Y + H_Z) \leq I_{XYZ} \leq \tfrac{1}{3}(H_X + H_Y + H_Z)$$

(25) $I_{XYZ} = 0$, if X, Y and Z are mutually independent

(26) $I_{XYZ} = \frac{1}{3}(H_X + H_Y + H_Z)$ if and only if for each i, j and k there is at most one pair (j,k), (k,i) and (i,j), respectively, such that $p_{ijk} > 0$

(27) $I_{XYZ} = -\frac{1}{3}(H_X + H_Y + H_Z)$ if and only if for each (i,j), (j,k) and (k,i) there is at most one k, i and j, respectively, such that $p_{ijk} = 1/m^2$, where $m = \min\{r,c,l\}$.

The final rationally scaled measure of dependence, called *total entropy correlation coefficient*, is defined analogically to the bivariate case:

(28) $$\rho_H = \sqrt[3]{\frac{I_{XYZ}}{\frac{1}{3}(H_X+H_Y+H_Z)}} .$$

Using the properties (24) - (27) derived for I_{XYZ} we see that ρ_H varies between -1 and 1. The minimum -1 (indicating maximal negative correlation) is reached for a distribution in which there is a diagonal distribution in each layer but these distributions situate in different positions in different layers. The maximum value 1 is reached for a diagonal distribution, i.e. in the case of complete (or absolute) dependence. And at last, independence is scored as 0. All these critical values of ρ_H match extremely well our intuitive idea of the nature and degree of dependence. The cubic root transformation is needed to quarantee ρ_H an intuitively rational behaviour between the extreme values, too.

It is clear that in three or higher dimensions ρ_H can highlight dependence from only one point of view, from the point of view of total correlation. More information about dependence can be obtained when different types of partial correlation coefficients are introduced. These partial correlation coefficients can also be based on entropy and coentropy concepts.

REFERENCES

Astola, J. and Virtanen, I. (1981). Entropy correlation coefficient, a measure of statistical dependence for categorized data. Lappeenranta University of Technology, Department of Physics and Mathematics, Research Report 4/1981, 22 p. Lappeenranta.

Kendall, M. and Stuart, A. (1979). The advanced theory of statistics, Vol. 2: Inference and relationship, 4th edition. London: Charles Griffin & Co Ltd.

Kullback, S. (1959). Information theory and statistics. New York: John Wiley & Sons.

Preuss, L.G. (1980). A class of statistics based on the information concept. Communications in statistics, theory and methods, Vol. A 9, No 15, 1563-1586.

Shannon, C.E. (1948). A mathematical theory of communication. Bell system technical journal, Vol. 27, 379-423.

Theil, H. (1969). On the use of information theory concepts in the analysis of financial statements. Management science, Vol. 15, No 9, 459-480.