AN ENTROPY BASED CORRELATION COEFFICIENT FOR CATEGORIZED DATA

by

Ilkka Virtanen

Reprint from

Tarja Meristö - Kari Heinistö (eds.)

ESSAYS IN MANAGEMENT STUDIES
VISIT TO BOSTON 1983

Virtanen, Ilkka

AN ENTROPY BASED CORRELATION COEFFICIENT FOR CATEGORIZED DATA

1. INTRODUCTION

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

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

due to the good properties of H as a measure of dispersion: H is nonnegative, 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).

Now we consider, however, data which are presented as a two-way contingency table. For a more detailed discussion on the subject see Astola and Virtanen (1983). 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,...,r, j = 1,...,c.

The entropies of the marginal distributions of X and Y are

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

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

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

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

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

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

$$(5)\quad \max\left\{H_X, H_Y\right\} \leq H_{XY} \leq H_X + H_Y$$

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

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

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

2. ENTROPY CORRELATION COEFFICIENT

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

(7) $$I_{XY} = -(H_{XY} - H_X - H_Y) = H_X + H_Y - H_{XY} \, .$$

I_{XY} has also been called the expected mutual information (Theil 1969, 470) or mean information (Kullback 1959, 5).

The 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

(8) $0 \leq I_{XY} \leq \frac{1}{2}(H_x + H_Y)$

(9) $0 \leq I_{XY} \leq \min \{\log r, \log c\}$

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

(11) $I_{XY} = \frac{1}{2}(H_X + H_Y)$ if and only if X and Y are completely dependent, i.e.

$p_{i_1 j} p_{i_2 j} = 0$ if $i_1 \neq i_2$, $j = 1,\dots,c$ and

$p_{i j_1} p_{i j_2} = 0$ if $j_1 \neq j_2$, $i = 1,\dots,r$.

The statements (8) and (9) 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

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

The division by $\frac{1}{2}(H_X + H_Y)$ in (12) 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

(13) $$0 \leq \rho_H \leq 1$$

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

(15) $$\rho_H = 1, \text{ iff } \quad p_{i_1 j}p_{i_2 j} = 0, \quad \forall i_1 \neq i_2, \quad j = 1,\dots,c. \qquad p_{ij_1}p_{ij_2} = 0, \quad \forall j_1 \neq j_2, \quad i = 1,\dots,r.$$

We see that ρ_H has been scaled between 0 and 1, 0 indicating full independence and 1 complete dependence. 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 does not depend on the forms of marginal distributions. And at last, the population size has no effect on ρ_H. 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).

3. GENERALIZATIONS TO THREE-WAY TABLES

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

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

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

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

I_{XYZ} has the following properties

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

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

(20) $$I_{XYZ} = \frac{1}{3}(H_X+H_Y+H_Z)$$

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 $p_{ijk} > 0$

(21) $$I_{XYZ} = -\frac{1}{3}(H_X+H_Y+H_Z)$$

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:

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

Using the properties (18)-(21) 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 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.
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 1981.

Astola, J. and Virtanen I.
A measure of overall statistical dependence based on the entropy concept. Manuscript, 31 p., submitted to Communications in statistics, 1983.

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

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

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

Shannon, C.E.
A mathematical theory of communication. Bell system technical journal, vol. 27, 1948, pp. 379-423.

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