VAASAN KORKEAKOULUN JULKAISUJA

TUTKIMUKSIA No 91

Tilastotiede 12

Jaakko Astola and Ilkka Virtanen

A MEASURE OF OVERALL STATISTICAL DEPENDENCE BASED ON THE ENTROPY CONCEPT

Vaasa 1983

1. INTRODUCTION

1.1 On the concept and use of entropy

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 tool for planning purposes in the wide and fast developing area of system modelling.

The concept of entropy originated in physics from the basic principle of the second law of thermodynamics. One of the many possible statements of this law is expressed in entropy form: the entropy of a physical system always increases. This statement simply asserts that the system cannot receive more in energy than the amount of external work supplied, and conversely, the system cannot transfer more energy to its environment, in the form of work, than it has energy available. For the use of entropy in physics see e.g. Van Wylen and Sonntag (1976, 193-265), see also the discussion in Wilson (1970b, 255-256).

The form of the concept of entropy that has found the most applications in various branches of science originated in information theory. Shannon (1948) discovered that there was a unique, unambiquous criterion for the amount of uncertainty represented by a discrete probability distribution, which agreed with the intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one and satisfied all other conditions which made it reasonable. He defined this measure of uncertainty, called the entropy of the probability distribution $(p_1,p_2,\ldots,p_n)$, as

$$(1.1) \qquad S(p_1,p_2,\ldots,p_n) = -k \sum_{i=1}^{n} p_i \log p_i \ .$$

2. BIVARIATE ENTROPY CORRELATION

2.1 Notation

In the present section we consider data which are presented as two-way contingency tables. These tables are most appropriate when the data are qualitative or categorical, but may also be used for discrete but ordered or for continuous but grouped data. The two variables, X and Y, to be considered are assumed to have r and c classes, respectively. In connection with the frequency tables we use the notation presented in Table 2.1. We assume throughout the paper that the cell frequencies N_{ij} are the theoretical or true frequencies, i.e. that the whole population has been under classification.

X \ Y	$F_1 \cdots F_j \cdots F_c$	Σ
E_1	$N_{11} \cdots N_{1j} \cdots N_{1c}$	$N_{1.}$
.	. . .	.
.	. . .	.
.	. . .	.
E_i	$N_{i1} \cdots N_{ij} \cdots N_{ic}$	$N_{i.}$
.	. . .	.
.	. . .	.
.	. . .	.
E_r	$N_{r1} \cdots N_{rj} \cdots N_{rc}$	$N_{r.}$
Σ	$N_{.1} \cdots N_{.j} \cdots N_{.c}$	N

Table 2.1. The frequency table of two categorized variables X and Y.

From the two-way table we can as marginal frequencies obtain the class frequencies for the variables X and Y:

$$N_{i.} = \sum_{j=1}^{c} N_{ij}, \qquad i = 1,2,\ldots,r, \tag{2.1}$$

$$N_{.j} = \sum_{i=1}^{r} N_{ij}, \qquad j = 1,2,\ldots,c. \tag{2.2}$$

The size of the population, N, is then

$$N = \sum_{i=1}^{r} N_{i.} = \sum_{j=1}^{c} N_{.j} = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij}. \tag{2.3}$$

By dividing the frequencies in Table 2.1 by the population size N, we get the relative frequencies or probabilities as given in Table 2.2. The cell probabilities p_{ij} define

X \ Y	$F_1 \cdots F_j \cdots F_c$	Σ
E_1	$p_{11} \cdots p_{1j} \cdots p_{1c}$	$p_{1.}$
.	. . .	.
.	. . .	.
.	. . .	.
E_i	$p_{i1} \cdots p_{ij} \cdots p_{ic}$	$p_{i.}$
.	. . .	.
.	. . .	.
.	. . .	.
E_r	$p_{r1} \cdots p_{rj} \cdots p_{rc}$	$p_{r.}$
Σ	$p_{.1} \cdots p_{.j} \cdots p_{.c}$	1

Table 2.2. The joint probability distribution and the marginal distributions of the random variables X and Y.

the joint probability (or relative frequency) distribution of X and Y. The one-dimensional distributions of X and Y are obtained as the marginal probabilities $p_{i.}$ and $p_{.j}$. The following relations are evident

$$(2.4) \qquad p_{ij} = N_{ij}/N, \quad i = 1,2,\ldots,r, \quad j = 1,2,\ldots,c,$$

$$(2.5) \qquad p_{i.} = N_{i.}/N, \quad i = 1,2,\ldots,r,$$

$$(2.6) \qquad p_{.j} = N_{.j}/N, \quad j = 1,2,\ldots,c,$$

$$(2.7) \qquad p_{i.} = \sum_{j=1}^{c} p_{ij}, \quad i = 1,2,\ldots,r,$$

$$(2.8) \qquad p_{.j} = \sum_{i=1}^{r} p_{ij}, \quad j = 1,2,\ldots,c,$$

$$(2.9) \qquad \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = \sum_{i=1}^{r} p_{i.} = \sum_{j=1}^{c} p_{.j} = 1 .$$

In what follows, the symbol log x is used to denote $\log_2 x$, i.e. the logarithm of x to base 2.

2.2 Coentropy and marginal entropies

As was pointed out in Section 1 already, the main role of entropy in statistics is its use as a measure of dispersion in connection with one-dimensional categorical variables. Our aim is now to extend the concept of entropy for two dimensional distributions in order to get an appropriate quantitative measure for the degree of dependence (or association) appearing in a two-way frequency table. The basic quantity in formulating this measure is the entropy of the joint distribution which can be shown to reveal both the dispersion

and the dependence existing in the distribution. We call this two-dimensional entropy *coentropy* because the pair entropy-coentropy can be shown to possess in connection with qualitative variables analogical relations and interpretation as the pair variance (or its square root standard deviation) - covariance has in connection with measurable data. The definition of coentropy in a two-way frequency table is based on the ideas presented by Theil (1969). Coentropy has also been called joint entropy (Theil 1969, 469-472) and overall entropy (Preuss 1980, 1566).

Definition 2.1. (Coentropy and marginal entropies). Let the bivariate distribution and the marginal distributions of X and Y be as presented in Table 2.2. The entropies of the marginal distributions of X and Y are defined as

$$(2.10) \qquad H_X = -\sum_{i=1}^{r} p_{i.} \log p_{i.}$$

$$(2.11) \qquad H_Y = -\sum_{j=1}^{c} p_{.j} \log p_{.j}$$

and the coentropy of the joint distribution of X and Y as

$$(2.12) \qquad H_{XY} = -\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \log p_{ij} .$$

Next we shall present some general properties of entropy. The proofs can be found e.g. in Aczél and Daróczy (1975), Astola and Virtanen (1981, 7-9).

Theorem 2.1. (Shannons Lemma). Consider two variables Z and Z' with distributions $(p_1,\ldots,p_M)$ and $(p'_1,\ldots,p'_M)$ respectively. Suppose that $p'_i > 0$ for $i = 1,\ldots,M$. Then

An important application of entropy in information theory is its use as a measure of the expected information of a message and as a tool for matching information streams with channel capacities.

Entropy is also being used with increasing frequency in the analysis of business and economic data. This was initiated by Theil (1969) and followed up by a number of authors. Empirical applications have been presented in economics, as well as in each of the major functional areas of business, viz. accounting, finance, management, marketing and production. A good survey and critique of the early business and economic applications is found e.g. in Horowitz and Horowitz (1976).

The concept of entropy has also been widely used in geography, especially in building models for urban and regional systems and for transportation. As a pioneer in this area may be named A.G. Wilson, see e.g. Wilson (1970a), who has also considered entropy as a general tool of system modelling in the context of operational research (Wilson 1970b).

1.2 Entropy in statistics

The use of entropy in statistics has its origin in information theory. Shannon's measure for uncertainty, for example, has been introduced as a measure of dispersion for qualitative data. For the connections between statistics and information theory, see Kullback (1959).

For a qualitative variable X, the values (symbols of the equivalence classes) of the variable may be quite arbitrary. The whole information of the distribution is in the class frequencies or probabilities. In order to get, for example, a location or dispersion index for the distribution, we have to use these probabilities. As a measure of the degree of dispersion of a distribution X: $(p_1, p_2, \ldots, p_n)$ the entropy of this distribution is used

$$(1.2) \qquad H = -\sum_{i=1}^{n} p_i \log_2 p_i \quad .$$

When we compare H with Shannon's original S given by (1.1), we see that the coefficient k in the expression of S has been fixed by choosing the base of the logarithm as 2.
It is easy to show that entropy H is a welldefined measure for the dispersion of the distribution: H is non-negative, $H = 0$ if and only if some $p_i = 1$, and H gets its maximum value ($= \log_2 n$) for the uniform distribution, i.e. when $p_1 = p_2 = \ldots = p_n = 1/n$.

It is possible to calculate entropy also for a two-dimensional distribution of two qualitative variables, i.e. for a bivariate distribution given as a frequency table. In this case entropy reveals both the dispersion of the distribution and the dependence between the two variables (Theil 1969, 469-472).
The analysis of entropy as a measure of dependence has remained, however, quite slight.

Our aim is now to carry out a more detailed analysis of the concept of entropy defined for two-way frequency tables. We also give entropy an interpretation as the mean uncertainty appearing in the table and demonstrate its definitional analogy with the covariance of two quantitative variables. Further, we construct an entropy-based measure for the degree of dependence and scale this measure in order to get a measure of dependence that fulfills both the theoretical and intuitively rational requirements for a well-defined correlation coefficient. Finally, a measure of dependence for three-way tables is introduced and analyzed.

$$(2.13) \qquad -\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log p_i'$$

and the equality holds only if $p_i = p_i'$ for $i = 1,\ldots,M$.

Theorem 2.2. Let Z be a random variable with the distribution $(p_1,\ldots,p_M)$. Then $0 \leq H_Z \leq \log M$

and
$$\begin{cases} H_Z = 0 \text{ if and only if some } p_i = 1 \\ H_Z = \log M \text{ if and only if } p_1 = p_2 = \ldots = p_M = \frac{1}{M}. \end{cases}$$

Entropy will thus be maximized when the population is uniformly distributed into all of the classes of the variable Z, i.e. when the dispersion of the distribution is at largest. Using probabilistic interpretation we may also say that the uncertainty connected with the distribution is then at its maximum: for a randomly chosen individual all the classes are equiprobable a' priori.

In the following theorem we list some properties of coentropy. From these properties 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.

Theorem 2.3. The following properties hold for the coentropy H_{XY} of the bivariate distribution of X and Y and for the entropies H_X and H_Y of its marginal distributions:

$$(2.14) \qquad \max\{H_X, H_Y\} \leq H_{XY} \leq H_X + H_Y$$

such that

$$(2.15) \qquad H_{XY} = H_X + H_Y$$

if and only if X and Y are independent. As numerical bounds for H_{XY} we have

$$(2.16) \qquad 0 \leq H_{XY} \leq \log(rc).$$

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.

Let us first consider the entropy H_Z of an one-dimensional distribution $Z:(p_1,p_2,\ldots,p_M)$. We can write

$$(2.17) \qquad H_Z = -\sum_{i=1}^{M} p_i \log p_i = \sum_{i=1}^{M} p_i \log(1/p_i).$$

Introducing a random variable $H = H(Z)$, which has the value $\eta_i = \log(1/p_i)$ when the value of the variable Z belongs to the i'th class, $i = 1,2,\ldots,M$, we can write

$$(2.18) \qquad H_Z = \sum_{i=1}^{M} p_i \log(1/p_i) = \sum_{i=1}^{M} p_i \eta_i = E\{H\},$$

i.e. entropy 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 (2.18) with the definition of the standard deviation of a quantitative variable Z, i.e. with

(2.19) $$D(Z) = \sqrt{E\{Z - E\{Z\}\}^2} ,$$

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 (2.12) we get analogously to (2.18)

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

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log(1/p_{ij})$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\eta_{ij} ,$$

where the quantities $\eta_{ij} = \log(1/p_{ij})$, $i = 1,2,\ldots,r$, $j = 1,2,\ldots,c$, may now be interpreted as the values of a two-dimensional random variable $H(X,Y)$, as the amount of uncertainty in the cells of the table. We have again

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

i.e. coentropy H_{XY} may be interpreted as the mean uncertainty appearing in the frequency table. The analogy with the covariance of a two-dimensional quantitative variable (X,Y), viz.

(2.22) $$Cov(X,Y) = E\{(X-E\{X\})(Y-E\{Y\})\},$$

is again evident: covariance gives the mean inaccuracy (about the mean) included in the distribution.

2.3 Mean dependence information

As we have seen, the coentropy measures both the dispersion of the joint distribution and the degree of independence of the two variables in the margins. In order to get an appropriate measure for the degree of dependence and for it only, we must eliminate the effects of marginal entropies from the coentropy and move over to the opposite quantity. Because we are working with logarithms, the natural way to carry out these modifications is subtraction. We get a measure of the degree of dependence called the *mean dependence information* and denoted by I_{XY}. Theil (1969) calls I_{XY} the expected mutual information, the term mean information has been used e.g. by Kullback (1959) in somewhat more general information theoretical circumstances.

Definition 2.2. (Mean dependence information). The mean dependence information of the bivariate distribution is defined as

(2.23) $$I_{XY} = -(H_{XY} - H_X - H_Y) = H_X + H_Y - H_{XY} .$$

The role of I_{XY} as the mean dependence information can be justified, however, as follows. We write

(2.24) $$I_{XY} = H_X + H_Y - H_{XY}$$

$$= -\sum_{i=1}^{r} p_{i.}\log p_{i.} - \sum_{i=1}^{c} p_{.j}\log p_{.j} + \sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij}\log p_{ij}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log(p_{ij}/p_{i.}p_{.j})$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\, l_{ij} ,$$

where $\iota_{ij} = \log(p_{ij}/p_{i.}p_{.j})$ is the amount of information about dependence in the cell (E_i,F_j): if it holds for a certain cell (E_i,F_j) the condition $p_{ij} = p_{i.}p_{.j}$ (which is the rule for all the 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 (2.24) we see that I_{XY} is expressed as the mean or expected value of this information. Analogously to (2.18) and (2.21) we can write

$$(2.25) \qquad I_{XY} = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij}\,\iota_{ij} = E\{I(X,Y)\},$$

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

From (2.23) we can see that the definition of the mean dependence information is analogous to the definition of the product moment correlation coefficient $\rho(X,Y)$ defined for quantitative variables:

$$(2.26) \qquad \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{D(X)D(Y)} \ .$$

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 (variances). In (2.23) 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 theorem considers the possible values of I_{XY} and shows that I_{XY} can be used as a measure of the degree of dependence.

Theorem 2.4. The following statements hold for the mean dependence information I_{XY}

$$(2.27) \qquad 0 \leq I_{XY} \leq \frac{1}{2}(H_X+H_Y)$$

$$(2.28) \qquad 0 \leq I_{XY} \leq \min\{\log r, \log c\}$$

$$(2.29) \qquad I_{XY} = 0 \quad \text{if and only if X and Y are independent}$$

$$(2.30) \qquad I_{XY} = \frac{1}{2}(H_X+H_Y) \text{ if and only if X and Y are completely dependent, i.e. } p_{i_1j}p_{i_2j} = 0 \text{ if } i_1 \neq i_2,\ j = 1,\ldots,c \text{ and } p_{ij_1}p_{ij_2} = 0 \text{ if } j_1 \neq j_2,\ i = 1,\ldots,r.$$

2.4 Entropy correlation coefficient

In the previous subsection we considered the quantity I_{XY}, the mean dependence information, as a measure of the degree of dependence of two qualitative variables and demonstrated its definitional analogy with the product moment correlation coefficient of quantitative variables. As a measure of dependence I_{XY} has, however, 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.

In order to get the final index to vary between 0 and 1 such that 0 shall indicate full independence and 1 complete dependence we must divide the value of I_{XY} by its maximal value. It is worth to note here that by complete dependence we 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 (in the contingency table there exists at most one positive p_{ij} in each row and in each column). This degree of dependence is sometimes called absolute dependence (Kendall and Stuart 1979, 570). In order to get the other two requirements satisfied as well, we use the maximal value $\frac{1}{2}(H_X+H_Y)$ in the scaling , cf. condition (2.27).

After having divided I_{XY} by the term $\frac{1}{2}(H_X+H_Y)$ we have obtained an index for the degree of dependence which is theoretically justified and completely matches the general idea about the nature and degree of dependence at the extreme cases of full independence and complete dependence: scores 0 and 1, respectively. But does this index posses a consistent behaviour between these extreme values as well? Although it is very difficult to say which numerical value of a measure of dependence in each particular case best corresponds to our intuitive idea of the degree of dependence, it has become via several numerical examples apperent that the above index appears to have quite small values, especially in the cases of low or moderate dependence. This has lead to use a square root transformation for magnifying variations near zero and for quaranteeing the final measure an intuitively rational behaviour in the whole interval [0, 1], as will be demonstrated later. We call this new derived index *entropy correlation coefficient* and denote it by the symbol ρ_H.

Definition 2.3. (Entropy correlation coefficient). The entropy correlation coefficient between two variables X and Y, the joint distribution of which is given by Table 2.2, is defined as

$$\rho_H = \sqrt{\frac{I_{XY}}{\frac{1}{2}(H_X+H_Y)}} = \sqrt{2\left(1 - \frac{H_{XY}}{H_X+H_Y}\right)} \quad . \tag{2.31}$$

Theorem 2.5 presents the theoretical foundations for entropy correlation coefficient as a well-behaving measure of the degree of dependence. They are direct consequences from the corresponding properties of the mean dependence information I_{XY} (Theorem 2.4).

Theorem 2.5. For the entropy correlation coefficient ρ_H it holds:

$$0 \le \rho_H \le 1 \tag{2.32}$$

$$\rho_H = 0 \text{ iff } p_{ij} = p_{i.}p_{.j}, \ \forall\, i = 1,2,\ldots,r, \ j = 1,2,\ldots,c \tag{2.33}$$

$$\rho_H = 1 \text{ iff } \begin{cases} p_{i_1 j}p_{i_2 j} = 0, & \forall\, i_1 \neq i_2, \ j = 1,2,\ldots,c \\ p_{i j_1}p_{i j_2} = 0, & \forall\, j_1 \neq j_2, \ i = 1,2,\ldots,r \ . \end{cases} \tag{2.34}$$

Theorem 2.5 thus shows that ρ_H has been scaled between 0 and 1 (property (2.32)), 0 indicating full independence (property (2.33)) and 1 complete dependence (property (2.34)). From the properties (2.33) and (2.34) 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 (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 N has no effect on the values of 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 logarithm 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 the degree of dependence that fulfills all the properties verified for ρ_H above, cf. for example the discussion in Kendall and Stuart (1979, 586-590).

2.5 Discussion

Our next aim is to construct a procedure by which it will be possible to demonstrate and test the behaviour of an association index on the whole range of the degree of dependence, from the case of full independence to the case of complete dependence.

We utilize a contingency table with a fixed population size and with fixed marginal frequency distributions (the notation is as presented in Table 2.1). We start (table 0) with the case of full independence: the cell frequencies are determined by the marginal frequencies as

$$(2.35) \qquad N_{ij} = \frac{N_{i.} N_{.j}}{N} \quad , \quad i = 1,\ldots,r, \quad j = 1,\ldots,c \; .$$

In the first phase (table 1) we modify, keeping the margins fixed, the starting table towards the case of complete dependence just as little as possible: for some i both N_{ii} and $N_{i+1,i+1}$ are increased by one, and both $N_{i,i+1}$ and $N_{i+1,i}$ are decreased by one. In the following phases we proceed with similar elementary modifications until we reach the final table (table n, say) which represents the case of complete dependence: we have

$$(2.36) \qquad N_{i_1 j} N_{i_2 j} = 0, \quad \forall\, i_1 \neq i_2, \quad j = 1,\ldots,c$$

and

$$(2.37) \qquad N_{i j_1} N_{i j_2} = 0, \quad \forall\, i = 1,\ldots,r, \quad j_1 \neq j_2 \; .$$

Without loss of generality, the positive frequencies of the final table may be assumed to be located on the main diagonal, because we are working with categorized data only. This adoption was indicated by the definitions of the intermediate steps already.

Table 2.3 presents the procedure of elementary modifications for increasing the degree of dependence from full independence to complete dependence in one particular case. We have a 2 x 2-table with uniform margins. The population size N = 100 leads to 25 steps (26 tables) in the procedure.

25	25	50
25	25	50
50	50	100

Table 0

26	24	50
24	26	50
50	50	100

Table 1

27	23	50
23	27	50
50	50	100

Table 2

. . .

48	2	50
2	48	50
50	50	100

Table 23

49	1	50
1	49	50
50	50	100

Table 24

50	0	50
0	50	50
50	50	100

Table 25

Table 2.3. The procedure of elementary modifications for varying the degree of dependence.

The procedure of elementary modifications offers now a method to test the consistent behaviour of an index of dependence. The first and last table represent two quite unambiguous situations: the cases of full independence and complete dependence must have the index scores 0 and 1, respectively. And because the steps from the first table to the final table were made as slight as possible, we can assume a linear growth in the degree of dependence along with the modificated tables. Therefore, we also require a linear relationship between the measure of dependence and the number of elementary steps made in our procedure.

In Fig. 2.1 the behaviour of ρ_H, the entropy correlation coefficient, and C, Pearson's coefficient of contingency, is presented as a function of the degree of dependence, the latter being measured by the number of modifications made into the original table of full independence. The graphs in Fig. 2.1 correspond to the data of Table 2.3. We can see, that the graph of ρ_H quite well fits the angle bisector which is used to stand for the ideal relationship. Only a slight underestimation may be noted in the behaviour of ρ_H, especially in the middle of the relevant domain. The behaviour of C, on the other hand, is less satisfactory. The graph of C departs strongly from the bisector at the right end (in the cases of moderate and high dependence). The underestimation of the "true" degree of dependence is highest in the case of complete dependence: C can never reach the value 1.

Although the deductions made above are based on one particular case only, they can be shown to hold generally. Several numerical computations with different population sizes, with different number of equivalence classes in the margins and with different forms of the table have shown that the qualitative behaviour of ρ_H remains as presented in Fig. 2.1.

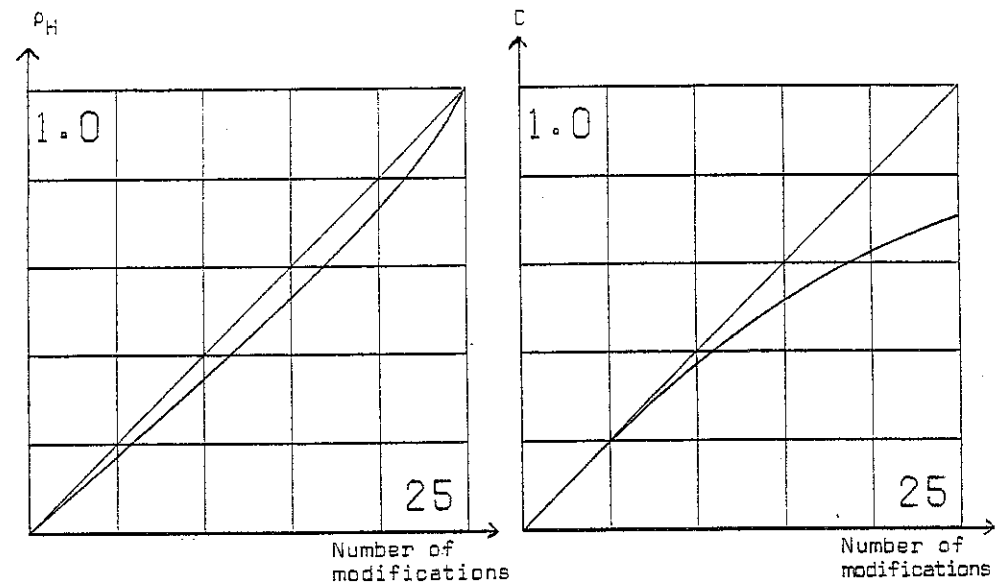


Figure 2.1. The behaviour of ρ_H and C as a function of the degree of dependence.

As a conclusion from the discussion concerning the entropy correlation coefficient ρ_H in the bivariate case we can state that in ρ_H we have a measure of dependence which exceptionally well fulfills both the theoretical requirements and intuitive expectations we have set for a correlation coefficient. And remembering its definitional analogies with the product moment correlation coefficient, the foundations for its use become even more firm.

3. GENERALIZATIONS TO THREE-WAY TABLES

3.1 Preliminary information-based concepts

The entropy-based concepts derived for the bivariate case can be quite straightforwardly generalized to contingency tables with any number of dimensions. In the following, however, only three-way tables are considered. This is mainly done to keep the notation simple. In higher dimensions we would also encounter interpretational difficulties.

Consider three categorized variables X, Y and Z having the joint distribution (p_{ijk}), $i = 1,\ldots,r$, $j = 1,\ldots,c$, $k = 1,\ldots,l$, i.e. the table has r rows, c columns and l layers with cell probabilities (or relative frequencies) p_{ijk}. The coentropy of the joint distribution is now defined as

$$H_{XYZ} = -\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} p_{ijk} \log p_{ijk} . \tag{3.1}$$

We can also calculate the coentropies of the two-dimensional marginal distributions as

$$H_{XY} = -\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij.} \log p_{ij.} \tag{3.2}$$

$$H_{YZ} = -\sum_{j=1}^{c}\sum_{k=1}^{l} p_{.jk} \log p_{.jk} \tag{3.3}$$

$$H_{XZ} = -\sum_{i=1}^{r}\sum_{k=1}^{l} p_{i.k} \log p_{i.k} \tag{3.4}$$

and the entropies of the one-dimensional marginal distributions as

$$H_X = -\sum_{i=1}^{r} p_{i..} \log p_{i..} \tag{3.5}$$

$$H_Y = -\sum_{j=1}^{c} p_{.j.} \log p_{.j.} \tag{3.6}$$

$$H_Z = -\sum_{k=1}^{l} p_{..k} \log p_{..k} . \tag{3.7}$$

In order to measure the overall dependence between the three variables in the form of one single index we must eliminate the effects of lower order dependences and dispersion from the covariation (measured by the coentropy) of the variables. Analogically to the bivariate case we define the quantity *mean (total) dependence information* as a measure of overall dependence prevailing in the distribution.

Definition 3.1. (Mean total dependence information). The mean total dependence information I_{XYZ} of a three-dimensional distribution is defined as

$$\begin{aligned} I_{XYZ} &= H_{XYZ} - (H_{XY} - H_X - H_Y) - (H_{XZ} - H_X - H_Z) \\ &\quad - (H_{YZ} - H_Y - H_Z) - H_X - H_Y - H_Z \\ &= H_{XYZ} - H_{XY} - H_{XZ} - H_{YZ} + H_X + H_Y + H_Z . \end{aligned} \tag{3.8}$$

The role of I_{XYZ} as the expected value of the amount of information about the overall dependence can be justified analogically to the bivariate case (equations (2.24) and (2.25)).

In the following theorem we present some properties of I_{XYZ} that describe I_{XYZ} from the point of view of a dependence index.

Theorem 3.1. The mean total dependence information I_{XYZ} has the following properties

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

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

(3.11) $I_{XYZ} = \frac{1}{3}(H_X + H_Y + H_Z)$ 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 $p_{ijk} > 0$

(3.12) $I_{XYZ} = -\frac{1}{3}(H_X + H_Y + H_Z)$ 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 is minimum of the numbers of nonzero $p_{i..}$, nonzero $p_{.j.}$ or nonzero $p_{..k}$.

Proof. For the left hand side of double inequality (3.9) we write

(3.13)
$$\begin{aligned} I_{XYZ} &= (H_X + H_Y - H_{XY}) + (H_Y + H_Z - H_{YZ}) \\ &\quad + (H_{XYZ} - H_{XZ}) - H_Y \\ &= I_{XY} + I_{YZ} + (H_{XYZ} - H_{XZ}) - H_Y \\ &\geq -H_Y , \end{aligned}$$

where the inequality holds on the basis of (2.27) and (2.14). Similarly we have $I_{XYZ} \geq -H_X$ and $I_{XYZ} \geq -H_Z$. These there inequalities together confirm the left hand side of (3.9). For the right hand side of (3.9) we write first

(3.14)
$$\begin{aligned} I_{XYZ} &= (H_X + H_Y - H_{XY}) - (H_{XZ} + H_{YZ} - H_{XYZ}) + H_Z \\ &\leq (H_X + H_Y - H_{XY}) - (H_X + H_{YZ} - H_{XYZ}) + H_Z \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} p_{ijk} \log \frac{p_{ij.}}{p_{i..}\, p_{.j.}} \\ &\quad - \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} p_{ijk} \log \frac{p_{ijk}}{p_{i..}\, p_{.jk}} + H_Z \\ &= -\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} p_{ijk} \log p_{ijk} \\ &\quad + \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} p_{ijk} \log \frac{p_{ij.}\, p_{.jk}}{p_{.j.}} + H_Z \\ &\leq H_Z , \end{aligned}$$

where the last inequality results from Shannon's lemma. Similarly we have $I_{XYZ} \leq H_X$ and $I_{XYZ} \leq H_Y$, and so also the right hand side of (3.9) is shown to be true.

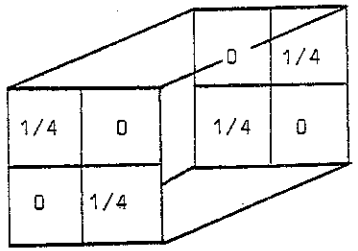
For proving (3.10) we write

(3.15)
$$\begin{aligned} I_{XYZ} &= -\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} p_{ijk} \log \frac{p_{ij.}\, p_{i.k}\, p_{.jk}}{p_{ijk}\, p_{i..}\, p_{.j.}\, p_{..k}} \\ &= -\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} p_{ijk} \log \frac{p_{i..}p_{.j.}p_{i..}p_{..k}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}p_{i..}p_{.j.}p_{..k}} = 0, \end{aligned}$$

where the second equality holds because of the mutual independence.

From (3.11) we see that I_{XYZ} reaches its maximum value in the case of complete dependence, i.e. in the case where the positive frequencies are located on the main diagonal of the cube (the physical site on the main diagonal may require a renumbering of the classes but this is always possible for true categorical data).

As (3.12) shows we may also encounter negative association when three categorical variables are considered together (in the bivariate case the sign of the association has no meaning for true categorical variables). A negative association exists, when two of the variables, X and Y say, are conditionally (with respect to Z) dependent, but the form of dependence varies with different values of Z. In Table 3.2. we have a distribution where X and Y conditionally have a complete dependence in both of the layers of Z but the forms of dependence are opposite to each others: I_{XYZ} reaches its minimum value -log 2.



I_{XYZ} = -log 2

Table 3.2. An example about complete negative dependence between three variables.

3.2 Total entropy correlation coefficient

As a measure of dependence, the mean total dependence information I_{XYZ} has the same disadvantages which were pointed out for I_{XY} in the bivariate case: it is not satisfactorily scaled, its maximum (now also minimum) value depends on the size and type of the contingency table and on the marginal distributions. Analogically to the bivariate case we use scaling and an algebraic transformation to obtain the final rationally behaving measure of overall statistical dependence, called the *total entropy correlation coefficient*.

Definition 3.2. (Total entropy correlation coefficient). The total entropy correlation coefficient ρ_H of a three-dimensional distribution is defined as

$$\rho_H = \sqrt[3]{\frac{I_{XYZ}}{\frac{1}{3}(H_X + H_Y + H_Z)}} \qquad (3.16)$$

Using Theorem 3.1 we see that ρ_H varies between -1 and +1. The minimum -1 (indicating maximal negative association) is reached by a distribution where there is a diagonal conditional distribution in each layer but these distributions situate in different positions in different layers (cf. the distribution in Table 3.2.). The maximum value +1 is reached by a diagonal distribution, i.e. in the case of complete (or absolute) positive dependence.

For mutually independent variables we have $\rho_H = 0$, as required. In the cases of certain conditional independencies or degeneracies we also have $\rho_H = 0$ (cf. the distributions in Table 3.1). The different cases for $\rho_H = 0$ can be distinquished for example by considering the two-dimensional conditional distributions.

The three critical values of ρ_H considered above match extremely well the general idea about the nature and degree of dependence in the cases of complete negative or positive dependence and full independence, respectively. Scaling of I_{XYZ} by the coefficient $3/(H_X + H_Y + H_Z)$ makes the index theoretically well justified. The cubic root transformation is needed to quarantee ρ_H an intuitively rational behaviour between the extreme values, too.

In order to demonstrate and test the behaviour of ρ_H on the whole range of the degree of dependence, we utilize an analogical procedure as described in the bivariate case. We take again a contingency table with a fixed population size and with fixed marginal frequency distributions. Then we start with the case of complete negative dependence, condition (3.12) being satisfied. In the following phases we modify the table towards the case of mutual independence, keeping the steps as small as possible. After having reached the mutual independence we proceed further with elementary modifications towards the case of complete positive dependence. Due to the elementary modifications we can again assume a linear growth in the degree of dependence along with the modificated tables. We should, therefore, also have a linear functionship between ρ_H and the number of elementary steps made in the procedure.

Table 3.3 presents the procedure of elementary modifications in a three-dimensional 2 x 2 x 2-table with uniform margins and with N = 400.

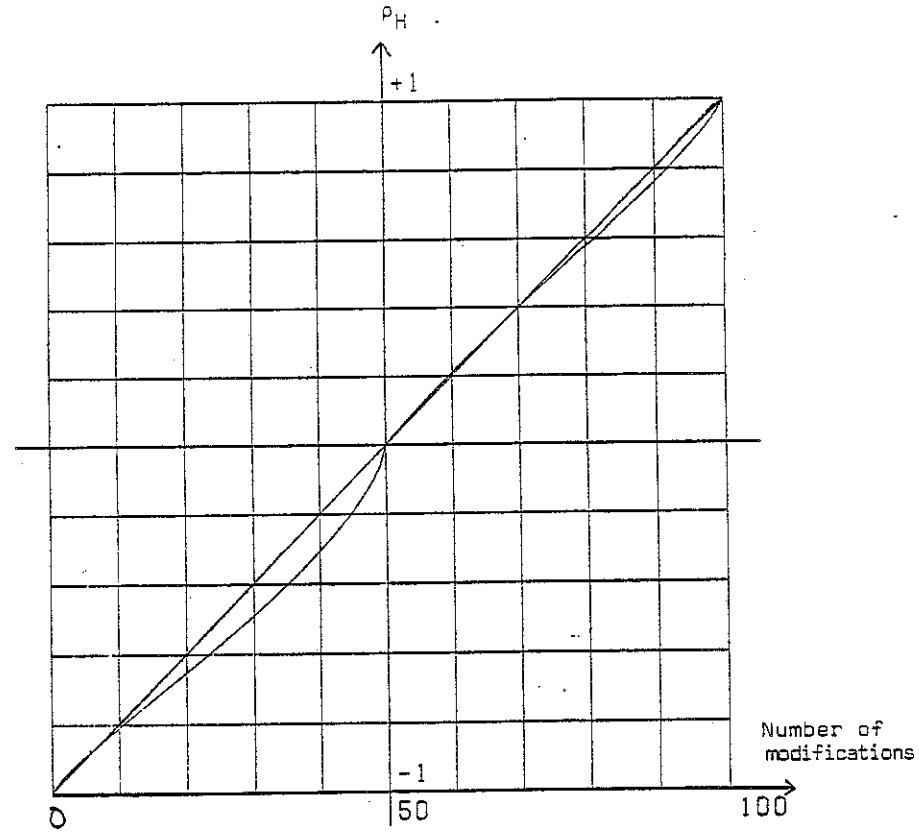


Figure 3.1. The behaviour of ρ_H as a function of the degree of dependence.

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. There exist, for example, several different kinds of situations where $\rho_H = 0$. More information about dependence can be obtained when different types of partial or multiple correlation coefficients are introduced. These correlation coefficients can also be based on entropy and coentropy concepts as has been recently shown by Preuss (1980).

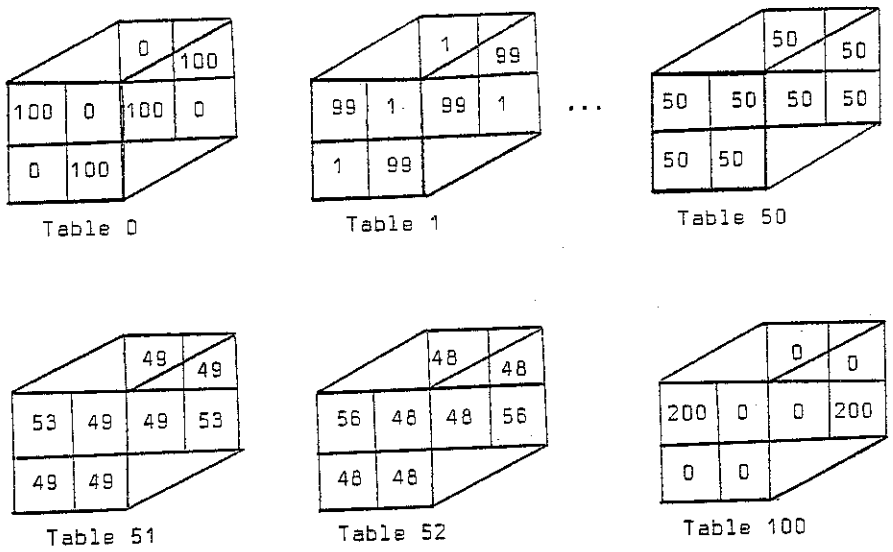


Table 3.3. The procedure of elementary modifications in the trivariate case.

Figure 3.1 presents, using the data of Table 3.3, the behaviour of ρ_H as a function of the degree of dependence. The latter is again measured by the number of modifications made into the starting table. We see that the graph of ρ_H also in the trivariate case quite well fits the angle bisector, the hypotetized ideal relationship. The fit is nearly perfect on the area of positive dependence, whereas on the negative side a slight underestimation can be seen to occur. If compared, however, e.g. with Pearson's coefficient of contingency C, the fit is exceedingly good: on the positive side C would behave as described in Fig. 2.1, and the negative sign of the inverse association would not be revealed at all (on the negative side only the absolute value of the degree of dependence would be obtained).

4. CONCLUSION

As a conclusion, the entropy based indices (total) mean dependence information and (total) entropy correlation coefficient appear to be theoretically justified and intuitively well-behaving measures of dependence in the connection of categorized data. In three (and higher) dimensions especially, the indices posses the ability to reveal also inverse association between the variables.

Throughout the paper we have assumed that the data is related to the whole population, i.e. the problem of statistical inference has not been relevant. If, however, the data is to be considered as a sample, the sample distribution of the entities should be derived for estimation and testing purposes. It may be conjectured that I_{XY} (I_{XYZ}), the purely theoretical one of the indices, will play a central role in these considerations.

REFERENCES

Aczél J., Daróczy Z., (1975): On measures of information and their characterizations. New York, Academic Press.

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.