LAPPEENRANNAN TEKNILLINEN KORKEAKOULU

YLEISTENTIETEIDEN LAITOS

LAPPEENRANTA UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF PHYSICS AND MATHEMATICS

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

by

Jaakko Astola and Ilkka Virtanen

LAPPEENRANTA
FINLAND



## 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 [8], pp. 193-265, see also the discussion in Wilson [11], pp. 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 [6] 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

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

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 [7] 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 [1].

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. [10]), who has also considered entropy as a general tool of system modelling in the context of operational research (see [11]).

## 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 [3].

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, (see e.g. Vasama and Vartia [9] pp. 99-104)



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

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 (see subsection 2.2 in this paper), that entropy H is a well-defined 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 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. This two-dimensional entropy has been shown to reveal both the dispersion of the distribution and the dependence between the two variables (Theil [7], pp. 469-472, Malaska and Reponen [5], pp. 179-185). The analysis of entropy as a measure of dependence has remained, however, quite slight.

Our aim is now to carry out a 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 analyze and interprete this measure. And finally, we scale this measure in order to get a measure of dependence that fulfills the requirements set for a correlation coefficient.

## 2. ENTROPY CORRELATION

### 2.1 Notation

In the present paper we consider data which are presented as two-way frequency or 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 classified into r and c categories, respectively. In the formation of the frequency tables we use the notation presented in Table 2.1. We assume throughout the paper

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

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

that the cell frequencies $N_{ij}$ are theoretical or true frequencies, i.e. that the whole population has been under classification.

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



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

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

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 proportional frequencies or probabilities given in Table 2.2. The cell probabilities

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

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

$p_{ij}$ define the joint probability distribution of X and Y which may now be considered as random variables. 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. There is no theoretical or practical reason to use particularly logarithms to base 2, it is, however, a tradition in information theory.

## 2.2 Coentropy of a two-dimensional distribution

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 ([7], pp. 469-472) and Malaska-Reponen ([5], pp. 179-180). These authors call, however, their two-dimensional entropies simply entropy (Theil also joint entropy).

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 and coentropy. We begin with the bounds of the one-dimensional entropy.

Theorem 2.1. Let Z be a random variable with the distribution $(p_1, p_2, \ldots, p_M)$. Then $H_Z \geq 0$, with equality if and only if some $p_i = 1$.

Proof. Theorem 2.1 follows immediately from the properties of log x and the fact that $0 \leq p_i \leq 1$ for $i = 1,\ldots,M$ and $p_1 + \ldots + p_M = 1$.

Theorem 2.1 thus states that entropy gets its minimum value (= 0) when all the individuals in the population belong to exactly one class with respect to the variable Z in question, i.e. when there is no dispersion in the values of Z. Or, using probabilistic interpretation, the entropy of a distribution equals zero when there is no uncertainly connected with the distribution, the value (class) of the variable is known with probability one a'priori.

Theorem 2.2 shows the dependence of the maximum value of the entropy on the number of classes of the distribution. Entropy reaches its maximum value, the logarithm of the number of classes, for uniform distribution. In the proof of the theorem we use two lemmas which we, therefore, present first.

Lemma 2.1. For all $x > 0$ we have

$$\log x \leq (x-1) \log e \,. \tag{2.13}$$

The equality holds only for $x = 1$.

Proof. Consider the function

$$f(x) = (x-1) \log e - \log x. \tag{2.14}$$

Now $f(1) = 0$ and by inspecting its derivative

$$f'(x) = \log e - \frac{\log e}{x} \tag{2.15}$$

we see that $f(x) > 0$ if $x \neq 1$.

Lemma 2.2. 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

$$-\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log p'_i \tag{2.16}$$

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



Proof. Suppose first that $p_i > 0$ for $i = 1,\ldots,M$. By lemma 2.1

$$\begin{aligned} (2.17) \qquad \sum_{i=1}^{M} p_i \log p'_i - \sum_{i=1}^{M} p_i \log p_i &= \sum_{i=1}^{M} p_i \log \frac{p'_i}{p_i} \\ &\leq \sum_{i=1}^{M} p_i \left(\frac{p'_i}{p_i} - 1\right) \log e = \log e \sum_{i=1}^{M} (p'_i - p_i) = 0 \end{aligned}$$

and the equality holds only if $p_i = p'_i$ for all $i = 1,\ldots,M$. Suppose then that $p_1 > 0,\ldots,p_l > 0$ and $p_{l+1} = \ldots = p_M = 0$. Let us write $p' = \sum_{i=1}^{l} p'_i$ and $p''_i = p'_i/p'$ $i = 1,\ldots,l$. Then also $\sum_{i=1}^{l} p''_i = 1$. By (2.17) we have

$$\begin{aligned} (2.18) \qquad -\sum_{i=1}^{l} p_i \log p_i &\leq -\sum_{i=1}^{l} p_i \log p''_i = -\sum_{i=1}^{l} p_i \log p'_i + \log p' \\ &< -\sum_{i=1}^{l} p_i \log p'_i + \sum_{i=l+1}^{M} 0 \cdot \log p'_i = -\sum_{i=1}^{M} p_i \log p'_i \,, \end{aligned}$$

and lemma 2.1 is seen to hold also in this case.

Theorem 2.2. Let Z be a variable with the distribution $(p_1,\ldots,p_M)$. Then $H_Z \leq \log M$ and the equality holds only if $p_1 = p_2 = \ldots = p_M = 1/M$.

Proof. We use lemma 2.2 with $p'_i = 1/M$ for $i = 1,\ldots,M$. It implies that

$$H_Z = -\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log \frac{1}{M} = \log M \tag{2.19}$$

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

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.

From the following theorem we can see that the upper bound for the coentropy of a bivariate distribution is formed as the sum of the entropies of the marginal distributions. The theorem also shows that both the dispersion of the joint distribution (which is revealed as the entropies of the marginal distributions) and the degree of the independence of the two variables have a contribution to the value of the coentropy.

Theorem 2.3. For the coentropy $H_{XY}$ of the bivariate distribution of X and Y and the entropies $H_X$ and $H_Y$ of its marginal distributions we have

$$H_{XY} \leq H_X + H_Y \,. \tag{2.20}$$

The equality in (2.20) holds only if X and Y are independent.

Proof. By definition 2.1

$$H_X = -\sum_{i=1}^{r} p_{i.} \log p_{i.} = -\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \log p_{i.} \tag{2.21}$$

and

$$H_Y = -\sum_{j=1}^{c} p_{.j} \log p_{.j} = -\sum_{j=1}^{c} \sum_{i=1}^{r} p_{ij} \log p_{.j} \,. \tag{2.22}$$

Thus

$$H_X + H_Y = -\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} (\log p_{i.} + \log p_{.j}) \tag{2.23}$$

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

On the other hand

$$H_{XY} = -\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \log p_{ij} \,. \tag{2.24}$$



Suppose first that all $p_{i.} p_{.j} > 0$. An application of lemma 2.2 with $p_{i.} p_{.j} = p'_{ij}$ then gives that

$$H_{XY} \leq H_X + H_Y \tag{2.25}$$

and that the equality holds only if $p_{i.} p_{.j} = p_{ij}$ $i = 1,\ldots,r$, $j = 1,\ldots,c$, i.e. only if X and Y are independent. Suppose then that some $p_{i.}$ and $p_{.j}$ are zero. Since $p_{i.} = 0$ or $p_{.j} = 0$ implies that the corresponding row or column consists entirely of zeros these can be deleted and then lemma 2.2 applied to the reduced table.

Theorem 2.4, on the other hand, shows that the entropies of the marginal distributions can never exceed the coentropy of the joint distribution.

Theorem 2.4. Let $H_{XY}$ be the coentropy of the distributions of X and Y and $H_X$ and $H_Y$ the entropies of the distributions of X and Y respectively. Then

$$H_{XY} \geq H_X \text{ and } H_{XY} \geq H_Y \,. \tag{2.26}$$

Proof. By symmetry it is enough to show that $H_{XY} \geq H_X$.
Now, for $i = 1,\ldots,r$

$$-\sum_{j=1}^{c} p_{ij} \log p_{ij} = -\sum_{j=1}^{c} p_{ij} (\log p_{ij} - \log p_{i.}) - p_{i.} \log p_{i.} \tag{2.27}$$

$$= \underbrace{-p_{i.} \sum_{j=1}^{c} \frac{p_{ij}}{p_{i.}} \log \frac{p_{ij}}{p_{i.}}}_{\geq 0} - p_{i.} \log p_{i.} \geq -p_{i.} \log p_{i.}$$

Thus

$$(2.28)\qquad -\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log p_{ij} \geq -\sum_{i=1}^{r} p_{i.}\log p_{i.}$$

as required.

By combining the results of the theorems 2.1 to 2.4 we obtain the following corollary which gives the absolute numerical bounds for the two-dimensional coentropy.

Corollary 2.1. For the coentropy $H_{XY}$ it holds

$$(2.29)\qquad 0 \leq H_{XY} \leq \log(rc).$$

Next we shall present for the entropy and coentropy an interpretation as the mean uncertainty appearing in the distribution. This interpretation clearly shows the analogy of the definitions of entropy and coentropy with those of variance (or its square root standard deviation) and covariance used in connection with quantitative and measurable variables.

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

$$(2.30)\qquad H_Z = -\sum_{i=1}^{M} p_i\log p_i = \sum_{i=1}^{M} p_i\log(1/p_i).$$

Let us now consider 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 then write



$$(2.31)\qquad H_Z = \sum_{i=1}^{M} p_i\log(1/p_i) = \sum_{i=1}^{M} p_i\eta_i = E\{H\},$$

i.e. the entropy H is expressed as the mean value of the random variable $H$. The quantity $\eta_i = \log(1/p_i)$ may be interpreted as the uncertainty of the i'th class: the uncertainty of the class equals zero, if $p_i$ equals one, the uncertainty increases monotonically when $p_i$ decreases, and approaches infinity when $p_i$ approaches zero. Entropy H thus expresses the mean uncertainty appearing in the distribution. If we compare (2.31) with the definition of the standard deviation of a quantitative variable Z, i.e. with

$$(2.32)\qquad D(Z) = \sqrt{E\{Z - E\{Z\}\}^2},$$

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 the coentropy (2.12) defined in a two-way frequency table we get analogously to (2.31)

$$(2.33)\qquad H_{XY} = -\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log p_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\log(1/p_{ij}) = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij}\eta_{ij},$$

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

We have again

$$(2.34) \qquad H_{XY} = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \eta_{ij} = E\{H(X,Y)\},$$

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

$$(2.35) \qquad Cov(X,Y) = E\{(X-E\{X\})(Y-E\{Y\})\},$$

is again evident.

## 2.3. Mean dependence information

As theorem 2.3 shows, the coentropy measures both the dispersion of the joint distribution (which is due to the dispersion of the marginal distributions) and the degree of the independence of the two variables in the marginals. 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 the logarithms, the natural way to carry out these modifications is subtraction. We get a measure of the degree of dependence, $I_{XY}$, called the *mean dependence information* (Theil [7], p. 470, calls $I_{XY}$ the expected mutual information, the term mean information has been used e.g. by Kullback [3], p. 5, in somewhat more general information theoretical circumstances).

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

$$(2.36) \qquad I_{XY} = -(H_{XY} - H_X - H_Y)$$

$$= H_X + H_Y - H_{XY} .$$



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

$$(2.37) \qquad I_{XY} = H_X + H_Y - H_{XY}$$

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

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

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

where $\iota_{ij} = \log(p_{ij}/p_{i.}p_{.j})$ is the dependence information of the cell $(E_i,F_j)$. If it holds for a certain cell $(E_i,F_j)$ $p_{ij} = p_{i.}p_{.j}$ (which is the rule for all the cells in the case of totally independent variables), the cell gives no information about the dependence of the variables, otherwise $\iota_{ij} \neq 0$ and there exists some dependence information in the cell. From (2.37) we see that $I_{XY}$ is formed as the mean or expected value of this information. Analogously to (2.31) and (2.34) we can write

$$(2.38) \qquad I_{XY} = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \iota_{ij} = E\{I(X,Y)\},$$

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

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 2.5. The following statements hold for the mean dependence information $I_{XY}$

(2.39) $$0 \leq I_{XY} \leq \frac{1}{2}(H_X+H_Y)$$

(2.40) $$0 \leq I_{XY} \leq \min\{\log r, \log c\}$$

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

(2.42) $I_{XY} = \frac{1}{2}(H_X+H_Y)$ 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 $j_1 \neq j_2$, $i = 1,\ldots,r$.

Proof. From (2.26) we find that

(2.43) $$H_{XY} \geq \max\{H_X, H_Y\}$$

which in turn implies (2.39). Combining (2.43) and (2.29) in corollary 2.1 we obtain (2.40). The statement (2.41) follows immediately from theorem 2.3. From the inequality (2.43) and the definition of $I_{XY}$ we see that $I_{XY} = \frac{1}{2}(H_X+H_Y)$ if and only if $H_{XY} = H_X = H_Y$. From (2.27) in the proof of theorem 2.4 we see that this happens solely in the case in which there is at most one nonzero cell in each row and column of the frequency table.

From (2.36) we can see that the definition of the mean dependence information is quite analogous to the definition of the product moment correlation coefficient $\rho(X,Y)$ defined for quantitative variables:

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

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 (2.36) 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.

## 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 the maximum value of $I_{XY}$ determined by the table depends on the marginal distributions (we require reaching of the maximum value independently of the marginal distributions in the case of complete dependence). We define, therefore, a new derived measure for dependence that fulfills all the requirements presented above. We call this measure of the degree of dependence *entropy correlation coefficient*. The definition of the entropy correlation coefficient is based on the idea presented by Malaska and Reponen ([5], p. 182). They have introduced a quantity called the index of information which turns out to be one half of our entropy correlation coefficient. Malaska and Reponen have not, however, much analyzed their index, e.g. they do not know its maximum value.

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

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

Theorem 2.6 presents the properties of the entropy correlation coefficient as a well-behaving measure of the degree of dependence. The theorem is a direct consequence from the properties of the mean dependence information $I_{XY}$ (Theorem 2.5 ).

Theorem 2.6. For the entropy correlation coefficient $\rho_H$ it holds:

$$
(2.46)\quad \begin{cases} \text{(i)} & 0 \le \rho_H \le 1 \\ \text{(ii)} & \rho_H = 0 \ \text{ iff } \ p_{ij} = p_{i.}p_{.j}, \ \forall\, i = 1,2,\ldots,r, \ j = 1,2,\ldots,c \\ \text{(iii)} & \rho_H = 1 \ \text{ iff } \begin{cases} p_{i_1 j}\, p_{i_2 j} = 0, & \forall\, i_1 \neq i_2, \ j = 1,2,\ldots,c \\ p_{i j_1}\, p_{i j_2} = 0, & \forall\, j_1 \neq j_2, \ i = 1,2,\ldots,r. \end{cases} \end{cases}
$$

Theorem 2.6 thus shows that $\rho_H$ has been scaled between 0 and 1 (property (i)), 0 indicating full independence (property (ii)) and 1 complete dependence (property (iii)). 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 (there exists at most one positive $p_{ij}$ in each row and in each column). This degree of dependence is sometimes called absolute dependence (see e.g. Kendall and Stuart [2], p. 570). From the properties (ii) and (iii) we can also see that the values of $\rho_H$ are independent of the size and type of the table: $\rho_H$ can reach all the values between 0 and 1 both in square and rectangular tables. Further, the values of $\rho_H$ are independent of the forms of marginal distributions (of the number classes in these, of the shape, e.g. location and dispersion indices of these etc.): there are no special requirements for the marginal probabilities $p_{i.}$ and $p_{.j}$ to reach the end values 0 and 1, for instance. And at last, the population size N has no effect on the values of $\rho_H$. From the point of view of mathematics, it is interesting to note that $\rho_H$ does not depend



on the base of the logarithm to be used. For, from the latter form of equation (2.45) we see that the module of the base of the logarithm appearing in the entropy quantities is cancelled out in the division. As a summary of the properties of $\rho_H$ 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 $\rho_H$ above (cf. for example the discussion in Kendall and Stuart [2], pp. 586-590).

## 2.5 An example

The following table (from Luoma and Taanonen [4], p. 9) shows the distribution of 87 students of Vaasa School of Economics who were classified (X) according to their drinking habits and (Y) according to their smoking habits. The standards in the first are $E_1$ = drinks never or seldom, $E_2$ = drinks sometimes, $E_3$ = drinks often and the standards in the latter are $F_1$ = does not smoke, $F_2$ = smokes sometimes, $F_3$ = smokes heavily.

| X \ Y | $F_1$ | $F_2$ | $F_3$ | $\Sigma$ |
|---|---|---|---|---|
| $E_1$ | 34 | 9 | 4 | 47 |
| $E_2$ | 7 | 5 | 6 | 18 |
| $E_3$ | 9 | 3 | 10 | 22 |
| $\Sigma$ | 50 | 17 | 20 | 87 |

Table 2.3. The population of 87 students of Vaasa School of Economics cross-tabulated according to their drinking and smoking habits.

From a frequency table the entropies are most easily calculated by the following formulae

$$
(2.47)\qquad H_X = \log N - \frac{1}{N} \sum_i N_{i.} \log N_{i.}
$$

(2.48) $$H_Y = \log N - \frac{1}{N} \sum_j N_{.j} \log N_{.j}$$

(2.49) $$H_{XY} = \log N - \frac{1}{N} \sum_i \sum_j N_{ij} \log N_{ij} .$$

From Table 2.3 we have

(2.50) $$H_X = 1.452$$

(2.51) $$H_Y = 1.407$$

(2.52) $$H_{XY} = 2.732$$

From these the mean dependence information is obtained

(2.53) $$I_{XY} = H_X + H_Y - H_{XY} = 0.126$$

We have seen above that the mean dependence information varies between 0 and $\frac{1}{2}(H_X + H_Y)$, the first being attained when X and Y are independent and the second being attained when X and Y are completely dependent, i.e. when the class of X completely determines the class of Y and vice versa. Entropy correlation coefficient $\rho_M$ was defined as the ratio of mean dependence information and its maximal value. From Table 2.3 we have

(2.54) $$\rho_H = \frac{2I_{XY}}{H_X + H_Y} = 0.088 .$$

The value of entropy correlation coefficient appears to be small compared for instance to Pearsons coefficient of contingency which for these data is approximately 0.38. On the other hand, it is difficult to say which numerical value of a measure of dependence best corresponds to our intuitive idea of the degree of dependence. The consistent behaviour of the measure is more



important. Also, if there is a generally accepted standard, it is easy to modify the measure to meet the standard by a suitable algebraic operation. In connection with the measures of dependence it has been customary to take a square root, an operation which tremendously magnifies variations near zero.

REFERENCES


[1] Horowitz, A.R., Horowitz, I., The Real and Illusory Virtues of Entropy-Based Measures for Business and Economic Analysis, Decision Sci., Vol. 7 (1976), pp. 121-136.

[2] Kendall, M. and Stuart, A., The Advanced Theory of Statistics, Vol. 2: Inference and Relationship, 4th Edition, Griffin, London 1979.

[3] Kullback, S., Information Theory and Statistics, Wiley, New York 1959.

[4] Luoma, M., Taanonen, M., Analysis of Contingency Tables with Log-Linear Models, Research Report 69, Publications of Vaasa School of Economics, (in finnish), Vaasa 1980.

[5] Malaska, P. and Reponen T., Problem Areas in Management in the 1980's, in Conference of the Researches of the Turku School of Economics 1979, Publications of the Turku School of Economics, series A-7:1979 (in finnish), Turku 1979.

[6] Shannon, C.E., A Mathematical Theory of Communication, Bell System Techn. J., Vol. 27 (1948), pp. 379-423, 623-656.

[7] Theil, H., On the Use of Information Theory Concepts in the Analysis of Financial Statements, Mgmt Sci. 15 (1969), No. 9, pp. 459-480.

[8] Van Wylen, G.J. and Sonntag, R.E., Fundamentals of Classical Thermodynamics, Wiley, New York 1976.

[9] Vasama, P.-M., Vartia, Y., Introduction to Statistics, Vol. I, (in finnish), Helsinki 1970.

[10] Wilson, A.G., Entropy in Urban and Regional Modelling, Pion., London 1970.

[11] Wilson, A.G., The Use of the Concept of Entropy in System Modelling, Operational Res. Quart., Vol. 21 (1970), No. 2, pp. 247-265.