sciendo

*DOI: 10.2478/bile-2018-0014*

*Biometrical Letters*
*Vol. 55 (2018), No. 2, 233-243*

# Entropy as a measure of dependency for categorized data

**Ewa Skotarczak, Anita Dobek, Krzysztof Moliński**

Department of Mathematical and Statistical Methods, Poznań University of Life Sciences,
Wojska Polskiego 28, Poznań, Poland, efalsa@up.poznan.pl

## SUMMARY

Data arranged in a two-way contingency table can be obtained as a result of many experiments in the life sciences. In some cases the categorized trait is in fact conditioned by an unobservable continuous variable, called liability. It may be interesting to know the relationship between the Pearson correlation coefficient of these two continuous variables and the entropy function measuring the corresponding relation for categorized data. After many simulation trials, a linear regression was estimated between the Pearson correlation coefficient and the normalized mutual information (both on a logarithmic scale). It was observed that the regression coefficients obtained do not depend either on the number of observations classified on a categorical scale or on the continuous random distribution used for the latent variable, but they are influenced by the number of columns in the contingency table. In this paper a known measure of dependency for such data, based on the entropy concept, is applied.

**Keywords:** contingency table, correlation, entropy, liability

## 1. Introduction

The problem of analysis of dependencies arises for both continuous and discrete variables. The case of continuous variables is widely treated in the literature, and hence the estimation and testing procedures are very well recognized. By contrast, the case of discrete variables seems to be more complicated, mainly because for discrete data the assumptions of linear models are difficult or even impossible to fulfill. Generally there are several strategies used in such cases. The first is based on nonparametric statistics. In the second, categorical data are transformed to continuous variables to enable the application of well-known statistical analyses (Snell, 1964). Other options, based on the general linear model and the threshold model, are also noted and widely used in the literature (Gianola and Foulley, 1983; McCullagh and Nelder, 1989; Bilow et al., 2017). Using these approaches we do not need to transform the discontinuous observations, but there are provided some link functions which can be modeled by a linear model. In this paper we will focus on the threshold model, which is appropriate for the analysis of data organized in a contingency table and has been used by many authors to analyze issues in the life sciences (Harville and Mee, 1984; Sorensen et al., 1995; Moliński et al., 2003). The main concept of the threshold model is based on the assumption that the discrete observations of a trait are determined by an unobservable continuous (often normally distributed) variable called liability (Falconer, 1989). This approach is justified for many threshold traits related to reproduction, health and fitness. For instance, it is well known that fertility is determined by many factors both genetic and environmental, which is typical for a continuous trait. However, observations of fertility made for individuals have a binary nature: fertile or not. There are many other threshold traits where the observable values are classified into several mutually exclusive and independent categories, but this classification is determined by an underlying continuous variable, for example hatchability (Dobek et al., 2003), ease of calving (Harville and Mee, 1984) or lodging of grain (Bakinowska and Kala, 2007). The relationship between the liability and visible categorical observations may be briefly presented as follows. When we observe only two categories (success or failure, e.g. fertile or not fertile) we expect success when the value of the liability reaches a sufficient value (threshold) on the unobservable scale, and failure otherwise. Similarly, for more categories, we observe one of several states of the categorical trait as a consequence of the fact that the underlying liability exceeds the corresponding unobservable threshold.

Let as assume that we are interested in the analysis of two threshold traits, both observed on each experimental unit. In medical treatments it is usual that patients are classified into two categories for one trait—for example, as diabetic or not—and into several categories for the other trait, which may be, for example, blood pressure (low, normal, high). The problem of estimation of the correlation between the two observed traits can be very important for the proper diagnostic process. Such data can be arranged into a two-row contingency table, and the question is whether it is possible to estimate the correlation between the underlying liabilities for observed threshold traits on the basis of the categorical data. In this paper, beginning with the assumption that the discrete variable is in fact conditioned by a continuous variable, we attempt to find a relationship between the Pearson correlation coefficient of two continuous variables and the entropy function measuring the corresponding relation when the attributes are observed on a discrete scale. The choice of an entropy-based approach is appropriate for a growing number of applications in various fields, especially in genetics (Kang et al, 2008; Ruiz-Marin et al, 2010). The relation between correlation and relative entropy measures of multivariate dependence was analyzed by Joe (1989). The aim of our work is to develop the approach proposed by Joe (1989) and to verify the possibility of estimating the Pearson correlation coefficient for the threshold traits with observations organized into a two-row contingency table. This was done by means of simulation studies.

To begin with, we shall present an existing measure of dependencies between categorical variables, being a function of entropy. Then, based on simulation studies, we shall present a dependency observed for the Pearson correlation coefficient and the measure of entropy used for data arranged in a two-row contingency table.

## 2. Definition of entropy

For a categorical random variable **A** taking values $\{a_1, a_2,\ldots,a_k\}$ with probabilities $p(a_i)$ ($i$ = 1,2,…,$k$), the well-known entropy introduced by Shannon (1948) is defined as

$$H(\mathbf{A}) = -\sum_{i=1}^{k} p(a_i)\ln\left(p(a_i)\right),$$

where ln() denotes natural logarithm. The value $H(\mathbf{A})$ is often called the Shannon diversity index, and it takes smaller values when there are significant differences between the $p(a_i)$ (Shannon, 1948).

The notion of entropy can be generalized to two or more categorical variables. In the case of two variables it is defined as

$$H(\mathbf{A},\mathbf{B}) = -\sum_{i=1}^{k}\sum_{j=1}^{l} p(a_i, b_j) \ln p(a_i, b_j),$$

where $a_i$ and $b_j$ are the values of **A** and **B**, and again $p$ is the corresponding probability (Jakulin, 2005). The value $H(\mathbf{A}, \mathbf{B})$ is a basis for several measures of dependency between two categorical variables (Moore et al., 2006).

## 3. Measure of dependency

The measure of dependency of categorical variables described below is a function of the value $I(\mathbf{A}, \mathbf{B})$, which represents the so-called mutual information, namely (Yan et al., 2008)

$$\begin{aligned} I(\mathbf{A},\mathbf{B}) &= H(\mathbf{A}) + H(\mathbf{B}) - H(\mathbf{A},\mathbf{B}) \\ &= H(\mathbf{A}) - H(\mathbf{A}/\mathbf{B}) \\ &= H(\mathbf{B}) - H(\mathbf{B}/\mathbf{A}) \\ &= H(\mathbf{A},\mathbf{B}) - H(\mathbf{A}/\mathbf{B}) - H(\mathbf{B}/\mathbf{A}), \end{aligned}$$

where $H(\mathbf{A} / \mathbf{B})$ is the conditional entropy, i.e.

$$H(\mathbf{A}/\mathbf{B}) = -\sum_{j=1}^{l} p(b_j) \sum_{i=1}^{k} p(a_i / b_j) \ln p(a_i / b_j),$$

which describes the uncertainty about variable **A** in the presence of **B** (Jakulin, 2005). The uncertainty about variable **B** in the presence of **A**, $H(\mathbf{B} / \mathbf{A})$, is defined in a symmetric way to $H(\mathbf{A} / \mathbf{B})$.

It should be noted that there are different measures of dependency (Jakulin, 2005; Moliński, et al., 2012), but in this paper we focus on a measure proposed by Joe (1989). According to Joe (1989), the quotient

$$J(\mathbf{A},\mathbf{B}) = \frac{I(\mathbf{A},\mathbf{B})}{\min\{H(\mathbf{A}),H(\mathbf{B})\}},$$

has some desired properties that allow it to be interpreted as the square of a correlation coefficient. Namely, since the entropy of a discrete probability distribution is positive, and mutual information *I*(**A**, **B**) is non-negative, *J*(**A**, **B**) is also non-negative. Furthermore, as was shown by Joe (1989), *J*(**A**, **B**) is bounded between 0 and 1 and the value 1 is reached iff there exists a functional dependency between **A** and **B**. The significance of *I(***A**,**B***)* can be tested based on a $\chi^2$ test (Kang et al., 2008).

## 4. Relation between correlation of continuous variables and mutual information

As has already been mentioned, in the life sciences there are many traits which can be observed on a categorical scale but are determined by many factors including genetic and environmental determinants, for example fertility (fertile or not), resistance to diseases (healthy or not), and resistance of pathogenic bacteria to various antibiotics. Due to the multifactorial determination of traits of this type, it is natural to suppose that their categorical phenotype is really expressed on a continuous, unobservable variable called liability.

In the present study we looked for a relation between the information obtained from the phenotypic data arranged in a contingency table, and the correlation between continuous, unobservable variables determining the classification of the data. To find these relations, simulation studies were performed. The data were simulated according to the following scenario. Two vectors **X** and **Y**, both of length *n*, were generated from a two-dimensional normal distribution with a given correlation coefficient between **X** and **Y** (denoted *r*). Several variants of expectations and variances for **X** and **Y** were tested. Three cases for *n* (100, 200 and 300) and, for every *n*, nine values of *r* (from 0.1 to 0.9 with step 0.1) were taken. The components of the generated **X** and **Y** were next distributed into categorical classes, so that a contingency table could be created. Because we focused on two-row tables, one of the vectors, say **X**, was divided into two classes, and the second vector into two, three, four or five categories. The thresholds between classes were determined taking into account the range of values of **X** and **Y**. For example, for a division of **X** into two categories, the threshold was taken as the minimum value of **X** increased by $^1/_2$ of the range of **X**; for a division of **Y** into three categories, the threshold between the first and the second class was equal to the minimum value of **Y** increased by $^1/_3$ of the range of **Y**, and the threshold between the second and the third category was the minimum **Y** increased by $^2/_3$ of the range of **Y**; etc. Then, for every contingency table the value of *J* was calculated. This process was repeated 100 times for each examined *r* and *n*, i.e. for every considered *n*, 900 2x2 contingency tables, 900 2x3 contingency tables, etc. were built. Finally, on the basis of 900 points, the linear regression between –ln(*r*) and –ln(*J*) was estimated.

The calculations were performed using the R package (R Core Team, 2013).

Figure 1 shows a visualization of the regression obtained for one data variant; however, a similar tendency was observed for the other cases. Each point on the figure denotes one contingency table for which the value of *J* was calculated. As mentioned earlier, for nine assumed values of correlation 100 contingency tables were simulated, hence at every level of assumed *r* 100 points are plotted. Several regression scenarios were checked, and finally the best fitting properties were obtained for linear regression between –ln(*r*) and –ln(*J*). It is clear from Figure 1 that the range of –ln(*J*) increases as the value of *r* becomes smaller. Values of –ln(*J*) greater than seven may suggest low correlation between the continuous variables determining the classification of the data, while values of –ln(*J*) less than two were obtained when the assumed correlation between **X** and **Y** was the highest.

Table 1 gives the regression coefficients ($b_1$) and intercepts ($b_0$) obtained for two sets of simulated data: one is the case in which the same expectations (equal to 10) and the same variances (equal to 1) were used to generate the variables **X** and **Y**, and other is the case in which the random distributions of **X** and **Y** had different parameters. It is clear that the results for both datasets are quite similar.

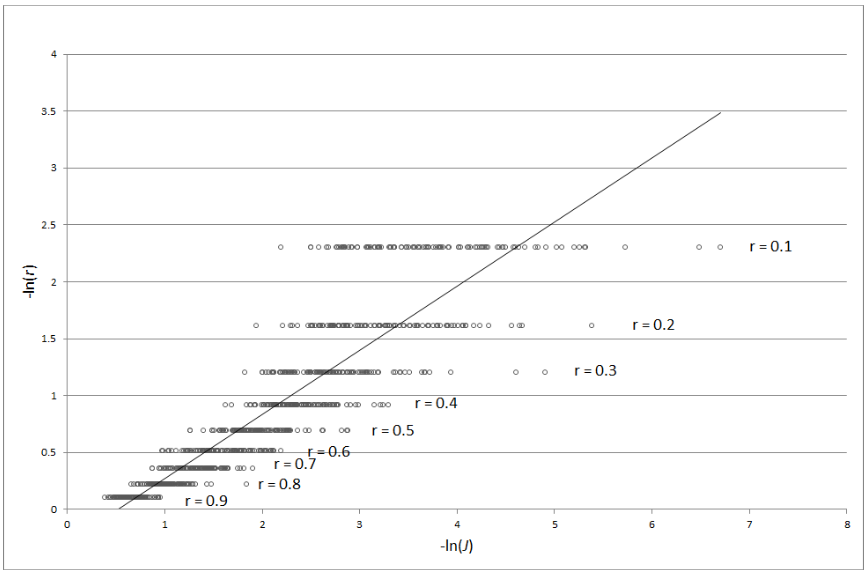

**Figure 1.** Linear regression for data generated with the same expectations (equal to 10) and the same variances (equal to 1) for **X** and **Y**.

**Table 1.** Linear regression coefficients

| | | Data set with E**X**=10, E**Y**=50, V**X**=4, V**Y**=9 | | | | Data set with E**X**=E**Y**=10, V**X**=V**Y**=1 | | | |
|---|---|---|---|---|---|---|---|---|---|
| Table | *n* | $b_1$ | $b_0$ | SD$_{b1}$ | SD$_{b0}$ | $b_1$ | $b_0$ | SD$_{b1}$ | SD$_{b0}$ |
| 2x2 | 100 | 0.3336 | -0.0915 | 0.007 | 0.023 | 0.3113 | -0.0104 | 0.007 | 0.023 |
| | 200 | 0.3390 | -0.1077 | 0.006 | 0.019 | 0.3439 | -0.1142 | 0.006 | 0.020 |
| | 300 | 0.3354 | -0.0920 | 0.005 | 0.018 | 0.3455 | -0.1134 | 0.006 | 0.019 |
| 2x3 | 100 | 0.4342 | -0.2023 | 0.008 | 0.023 | 0.4307 | -0.1929 | 0.008 | 0.023 |
| | 200 | 0.4466 | -0.2831 | 0.006 | 0.018 | 0.4625 | -0.3300 | 0.006 | 0.017 |
| | 300 | 0.4334 | -0.2878 | 0.006 | 0.017 | 0.4470 | -0.3094 | 0.006 | 0.018 |
| 2x4 | 100 | 0.5149 | -0.2703 | 0.009 | 0.022 | 0.5200 | -0.2722 | 0.009 | 0.023 |
| | 200 | 0.5173 | -0.3552 | 0.006 | 0.017 | 0.5054 | -0.3218 | 0.006 | 0.017 |
| | 300 | 0.4903 | -0.3229 | 0.005 | 0.014 | 0.5019 | -0.3487 | 0.005 | 0.015 |
| 2x5 | 100 | 0.5783 | -0.3169 | 0.010 | 0.024 | 0.5642 | -0.2932 | 0.010 | 0.022 |
| | 200 | 0.5396 | -0.3387 | 0.007 | 0.017 | 0.5565 | -0.3658 | 0.007 | 0.017 |
| | 300 | 0.5379 | -0.3745 | 0.005 | 0.013 | 0.5270 | -0.3515 | 0.005 | 0.014 |

E**X** – expectation of **X**, E**Y** – expectation of **Y**, V**X** – variance of **X**,V**Y** – variance of **Y**, *SD* – standard deviation

Analyzing the results, it was possible to observe certain relations between the estimated regression coefficients and the sizes of the data tables. While the values of $b_1$ and $b_0$ seem not to be influenced by the value of the parameter *n*, they are sensitive to changes in the size of the contingency table. An increase in the number of classes in the tables causes an increase in the absolute values of the regression coefficients. To show this tendency better, the means of $b_0$ and $b_1$ were calculated for different sizes of contingency tables. These are given in Table 2.

**Table 2.** Averaged regression coefficients for different sizes of contingency table

| Size of table | Mean $b_1$ | Mean $b_0$ |
|---|---|---|
| 2x2 | 0.3348 | -0.0882 |
| 2x3 | 0.4424 | -0.2676 |
| 2x4 | 0.5083 | -0.3152 |
| 2x5 | 0.5506 | -0.3401 |

The largest increase in the regression coefficients can be observed for 2x3 tables compared with 2x2 tables (Table 2). Further increases in the number of columns did not cause very significant increases in the coefficients. It is known that the discretization of data into *n* categories with very large *n* may lead to infinite values of the logarithms calculated in $J(\mathbf{A}, \mathbf{B})$. However, from a practical perspective, the classification of the data into more than five categories would appear not to be reasonable.

The obtained regression coefficients were applied to the estimation of the correlation for several simulated contingency tables for which the value of *J* was calculated. Then, the correlation coefficient *r* was estimated in two ways: following Joe (1989) as $\sqrt{J}$, and from the regression equation as $r = e^{-b_0}J^{b_1}$, where $b_1$ and $b_0$ were taken from Table 2. Some results of this procedure are presented in Table 3.

As is clear from Table 3, the use of regression made it possible to obtain more accurate correlation coefficients than Joe's approach. The best results were obtained for a true correlation equal to 0.5, and this seems to be a natural consequence of regression properties. For $n$ = 100 and 200 this conclusion remained true, hence these results are not presented.

**Table 3.** True and estimated correlation coefficients for four sizes of contingency tables with $n$=300

| | Correlation coefficient | | |
|---|---|---|---|
| Table | assumed | estimated as $\sqrt{J}$ | estimated via regression |
| | 0.1 | 0.0540 | 0.1547 |
| 2x2 | 0.5 | 0.2916 | 0.4785 |
| | 0.9 | 0.6375 | 0.8080 |
| | 0.1 | 0.0767 | 0.1348 |
| 2x3 | 0.5 | 0.3099 | 0.4634 |
| | 0.9 | 0.5658 | 0.7896 |
| | 0.1 | 0.0391 | 0.0508 |
| 2x4 | 0.5 | 0.4076 | 0.5503 |
| | 0.9 | 0.6161 | 0.8377 |
| | 0.1 | 0.1612 | 0.1883 |
| 2x5 | 0.5 | 0.3792 | 0.4830 |
| | 0.9 | 0.7404 | 1.0092 |

The proposed analysis was also performed for data generated with a Poisson distribution. The estimated regression coefficients were very similar to those presented above; therefore the results appear not to be influenced by the distribution of liability.

## 5. Conclusions

Data arranged in a two-way contingency table may arise as the result of many experiments in which two threshold traits are observed and at least one of them is collected on a binary scale. Knowledge about a potential correlation between underlying continuous random variables can improve the statistical analysis. The theory proposed by Joe (1989) suggested that it is possible to use measures of information and find their relations with the quantitative structure of the data. In this paper we proposed to estimate the Pearson correlation coefficient of unobservable continuous distributed liabilities of threshold variables on the basis of their categorical observations. Using simulation studies it was shown that the regression approach yields more accurate results than Joe’s approach. The obtained regression coefficients did not change significantly with modification of the underlying random distribution or with an increase in the number of observations classified in the given categories. However, they increased when the number of columns in the contingency table was greater. Moreover, the range of calculated values of normalized mutual information (on a logarithmic scale) is smaller for higher correlations. Based on the results, it was possible to propose values of linear regression coefficients which can be used, depending on the size of the contingency table, to estimate correlation. Notwithstanding, the results presented here are based on simulated data, and they can be treated only as a starting point for further analytical and practical research.

## REFERENCES

Bilow M., Crespo F., Pan Z., Eskin E., Eyheramendy S. (2017): Simultaneous modeling of disease status and clinical phenotypes to increase power in GWAS. Genetics 205: 1041-1047.

Bakinowska E., Kala R. (2007): An application of logistic models for comparison of varieties of seed pea with respect to lodging. Biometrical Letters 44(2): 143-154.

Dobek A., Steppa R., Moliński K., Ślósarz P. (2013): Use of entropy in the analysis of nominal traits in sheep. Journal of Applied Genetics 54: 97-102.

Dobek A., Szydłowski M., Szwaczkowski T., Skotarczak E., Moliński K. (2003): Bayesian estimates of genetic variance of fertility and hatchability under a threshold animal model. Journal of Animal and Feed Sciences 12: 307-314.

Falconer D.S. (1989): Introduction to Quantitative Genetics. Ed.3. Longmans Green/John Wiley & Sons, Harlow, Essex, UK/New York.

Gianola D., Foulley J.L. (1983): Sire evaluation for ordered categorical data with a threshold model. Genetics Selection Evolution 15: 201-224.

Harville D.A, Mee R.W. (1984): A mixed model procedure for analyzing ordered categorical data. Biometrics 40: 393–408.

Jakulin A. (2005): Machine learning based on attribute interactions. PhD dissertation. University of Ljubljana.

Joe H. (1989): Relative entropy measures of multivariate dependence. Journal of the American Statistical Association 84(405): 157-164.

Kang G., Yue W., Zhang J., Cui Y., Zuo Y., Zhang D. (2008): An entropy-based approach for testing genetic epistasis underlying complex diseases. Journal of Theoretical Biology 250: 362-374.

McCullagh P., Nelder J.A. (1989): Generalized linear models. Chapman and Hall/CRC.

Moliński K., Dobek A., Tomaszyk K. (2012): The use of information and information gain in the analysis of attribute dependencies. Biometrical Letters 49(2): 149-158.

Moliński K., Szydłowski M., Szwaczkowski T., Dobek A., Skotarczak E. (2003): An algorithm for genetic variance estimation of reproductive traits under a threshold model. Archives Animal Breeding 46: 85-91.

Moore J.H., Gilbert J.C., Tsai C.T., Chiang F.T., Holden T., Barney N., White B.C. (2006): A flexible computational framework for detecting, characterizing and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. Journal of Theoretical Biology 241: 252-261.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ruiz-Marin M., Matilla Garcia M., Cordoba J.A.G., Susillo-Gonzalez J.L., Romo-Astorga A., Gonzalez-Perez A., Ruiz A., Gayan J. (2010): An entropy test for single-locus genetic association analysis. BMC Genetics 11(19).

Shannon C.E. (1948): A mathematical theory of communication. The Bell System Technical Journal (27): 379-423, 623-656.

Snell E.J. (1964): A scaling procedure for ordered categorical data. Biometrics (20): 592-607.

Yan Z., Wang Z., Xie H. (2008): The application of mutual information-based feature selection and fuzzy LS-SVM-based classifier in motion classification. Computer Methods and Programs in Biomedicine (90): 275-284.