ISSN 2007-9737

Similarity Correlation of Frequency Distributions of Categorical Data in Analysis of Cognitive Decline Severity in Asthmatics

Ildar Z. Batyrshin[1], Imre J. Rudas[2], Nailya Kubysheva[3], Svetlana Akhtyamova[4]

[1] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

[2] Obuda University, Budapest,
Hungary

[3] Kazan Federal University, Kazan,
Russia

[4] Kazan National Research Technological University, Kazan,
Russia

batyr1@gmail.com, rudas@uni-obuda.hu, aibolit70@mail.ru, ahtjamovasve@yandex.ru

Abstract. The paper presents the method of measuring the similarity and difference in the frequency distributions of one categorical variable for different levels of another variable. This method calculates the similarity and correlation between the rows of the contingency table. In this work, we use it for the analysis of associations of cognitive indicators. The proposed categorical data association analysis method can be used as an addition to the classical chi-square test.

Keywords. Correlation, frequency distribution, categorical data, cognitive impairment.

1 Introduction

Recently it was proposed a general approach to the representation and construction of correlation and association coefficients [1-3]. They are considered as (correlation or association) functions of two arguments defined on a set with involutive operation, taking values in $[-1,1]$ and satisfying several properties. The involutive operation is mapping elements of the domain into "opposite" elements, such that the correlation between mutually opposite elements equals -1. It was shown [2-3] that most of the traditional correlation and association coefficients considered in statistics [4-5] and taking values in the interval $[-1,1]$ are correlation functions.

Correlation functions can be generated by similarity and dissimilarity functions satisfying suitable properties [1-3,6]. Moreover, a one-to-one correspondence exists between correlation and bipolar similarity functions [2]. For this reason, correlation functions are also referred to as invertible similarity correlations.

These results pave the way to construct correlation functions on almost any set if one can define an involutive operation and suitable similarity or dissimilarity function on this set. This approach was used in [7] to introduce a bipolar dissimilarity function and corresponding correlation function on the set of probability and relative frequency distributions. This proposal used the involutive negation of probability distributions [8]. The constructed correlation function surprisingly coincided with the Pearson product-moment correlation coefficient [4,5]. The authors of [7] used the introduced correlation function for calculating the correlation between frequency distributions in contingency tables.

Usually, these frequency distributions are defined on a set of categories of categorical variables presented in contingency tables [9-12].

In this paper, we apply the dissimilarity and correlation functions proposed in [7] to analyze the cognitive decline severity in asthmatics.

Currently, the manifestation of cognitive impairment in various diseases, including bronchial asthma, is being actively investigated [13]. Cognitive dysfunctions affect the control and development of asthma, which determines the relevance of more detailed studies of the role of cognitive impairment in the course of the disease. Various factors, including age, disease duration, education, lifestyle, etc., determine the degree of cognitive decline in asthmatics.

As a rule, existing tests for assessing cognitive impairment are analyzed by traditional statistical analysis methods allowing us to determine the significance of associations between the studied indicators [9,10]. In this paper, we analyze new data about relationships between the sociodemographic data of patients and the severity of cognitive decline in asthmatics published in [14].

The paper has the following structure. Section 2 discusses the basics of the theory of invertible similarity correlations.

The dissimilarity and correlation functions introduced in [7] for the analysis of relationships between frequency distributions and categorical data are considered in Section 3. Section 4 presents the data from [14]. Section 5 describes the results of the analysis of these data using the proposed method. The last Section discusses obtained results, a conclusion, and future work.

2 Basics of Correlation Functions

Consider the basic definitions and results related with correlation functions [1-3]. Let Ω be a set with more than one element. A function $N: \Omega \to \Omega$ is referred to as a *reflection* or a *negation* on Ω if, for all x in Ω, it satisfies the *involutivity* property:

$$N(N(x)) = x, \tag{1}$$

and N is not an identity function, i.e., $N(x) \neq x$ for some x in Ω. An element x in Ω satisfying the property:

$$N(x) = x$$

is called a *fixed point* of the negation N. The set of all fixed points of the negation N in Ω is denoted as $FP(\Omega)$. This set can be empty.

Denote V a non-empty subset of $\Omega \setminus FP(\Omega)$ closed under N. The set V does not contain fixed points of N, and if x in V, then $N(x)$ is also in V.

For any element x in V, its negation $N(x)$ can be considered as an element *opposite* to x. From the involutivity property (1), it follows that the element x is also opposite to $N(x)$. Hence for every x in V the elements x and $N(x)$ are mutually opposite, and $N(x)$ is also in V.

A correlation function (association measure) on V is a function $A: V \times V \to [-1,1]$ satisfying for all x, y in V the following properties [1]:

Symmetry:

A1. $A(x,y) = A(y,x)$,

Reflexivity:

A2. $A(x,x) = 1$,

Inverse relationship:

A3. $A(x,N(y)) = -A(x,y)$.

Usually, correlation and association coefficients used in statistics are calculated between real variables, dichotomous variables, rankings, etc., without considering some involutive operation on the corresponding set of variables [4, 5]. For such coefficients taking values in the interval $[-1,1]$, only the properties A1 and A2 are considered. But it was shown [1-3] that involutive operation can be introduced on the domains of traditional correlation coefficients like Pearson, Spearman, Kendall correlation, etc., and they will satisfy the inverse relationship property A3. For this reason, correlation functions satisfying property A3 also referred to as *invertible correlation functions* [2].

From properties A1-A3, it follows that the correlation function satisfies for all x in V the property:

Opposite elements:

$$A\big(x, N(x)\big) = -1.$$

Hence the correlation between opposite elements equals -1.

Correlation functions can be constructed from similarity and dissimilarity functions [1-3].

A function $D: V \times V \to [0,1]$ is a *dissimilarity function* on V if for all x, y in V, it is *symmetric*:

$$D(x,y) = D(y,x)$$

and *irreflexive*:

$$D(x,x) = 0.$$

A function $S: V \times V \to [0,1]$ is a *similarity function* on V if for all x, y in V, it is *symmetric*:

$$S(x,y) = S(y,x),$$

and *reflexive*:

$$S(x,x) = 1.$$

Similarity S and dissimilarity D functions are *dual* if for all x, y in V it is fulfilled:

$$S(x,y) = 1 - D(x,y), \quad (2)$$

$$D(x,y) = 1 - S(x,y). \quad (3)$$

These functions are called *bipolar* if, for all x, y in V, they satisfy the following properties [2]:

$$S(x,y) + S\big(x, N(y)\big) = 1,$$

$$D(x,y) + D\big(x, N(y)\big) = 1.$$

There exists a one-to-one correspondence between invertible correlation functions and bipolar similarity (dissimilarity) functions [2]:

$$A(x,y) = 2S(x,y) - 1, \quad (4)$$

$$A(x,y) = 1 - 2D(x,y). \quad (5)$$

These three functions compose complementary triplet (S, D, A) and are also related as follows:

$$A(x,y) = \mathrm{S(x,y)} - D(x,y). \quad (6)$$

As it follows from (4), the invertible correlation function is nothing else but a rescaled bipolar similarity function. For this reason, the correlation function is also referred to as a *similarity correlation function*. From (4), we see that the similarity values from interval $[0,1]$ are linearly transformed into correlation values in the interval $[-1,1]$. For example, we have:

$$A(x,y) = 1 \quad \text{if} \quad S(x,y) = 1,$$

$$A(x,y) = 0 \quad \text{if} \quad S(x,y) = 0.5,$$

$$A(x,y) = -1 \quad \text{if} \quad S(x,y) = 0.$$

As we can see, the correlation is positive if $S(x,y) > 0.5$ (similarity value is high) and negative if $S(x,y) < 0.5$ (similarity value is low). Dually, we see from (5) that the correlation is positive if $D(x,y) < 0.5$, and negative, if $D(x,y) > 0.5$.

In the following Section, we show how the correlation between frequency distributions is constructed using the involutive negation of probability distribution and suitable dissimilarity function between distributions.

3 Correlation of Frequency and Probability Distributions

The correlation of frequency distributions was introduced in [7]. We give a short description of the steps described in the previous Section and used for constructing the corresponding correlation function. Suppose, $F = \{f_1, \dots, f_n\}$ is a frequency distribution of n categories, where f_i is a non-negative integer value of the frequency of appearance of the i-th category in some experiments. Transform F in relative frequency distribution $P = \{p_1, \dots, p_n\}$, where:

$$p_i = \frac{f_i}{\sum_{i=1}^{n} f_i},\ i = 1, \dots, n.$$

We can consider P as a probability distribution, where p_i is a probability that the result of a random experiment will fall in i-th category. We have:

$$0 \le p_i \le 1, \qquad \sum_{i=1}^{n} p_i = 1.$$

Let Ω be a set of possible probability distributions with n elements. General methods of construction of negations of probability distributions are considered in [15]. The involutive negation of probability distributions is defined as follows [8]:

$$N(p_i) = \frac{MP - p_i}{nMP - 1},$$

where $MP = \max(P) + \min(P)$, and $\max(P) = \max\{p_1, \dots, p_n\}$, $\min(P) = \min\{p_1, \dots, p_n\}$.

The uniform probability distribution

$$P_U = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

is a unique fixed point of the negation N.

The bipolar dissimilarity function D on the set of probability distributions $V = \Omega \backslash \{P_U\}$ is defined as follows [7]:

$$D(P,Q) = \frac{1}{4}\sum_{i=1}^{n}\left[\frac{np_i - 1}{\sqrt{\sum_{i=1}^{n}(np_i-1)^2}} - \frac{nq_i - 1}{\sqrt{\sum_{i=1}^{n}(nq_i-1)^2}}\right]^2. \quad (7)$$

It defines by (5) the invertible correlation function [7]:

$$A(P,Q) = \frac{\sum_{i=1}^{n}(np_i-1)(nq_i-1)}{\sqrt{\sum_{i=1}^{n}(np_i-1)^2}\sqrt{\sum_{i=1}^{n}(nq_i-1)^2}}. \quad (8)$$

This correlation function coincides with Pearson product-moment correlation coefficient.

In the following Section, we use this correlation function to analyze cognitive decline severity in asthmatics.

4 Data of Cognitive Declines in Asthmatics

When analyzing the literature devoted to assessing the effect of cognitive impairment on various characteristics in patients with asthma, we were attracted by the article of Haq Satti et al. [14]. These authors studied associations between sociodemographic factors and cognitive decline severity in asthmatics. Table 1 presents the results obtained in this work.

This study showed [14] that the Duration of Illness and the use of Poly-Pharmacy were closely associated with the presence and severity of cognitive decline (p=0.005 and p=0.019, respectively).

In this paper, we analyzed the presented data using the similarity and correlation of frequency distributions considered in the previous Section.

5 Results

Since the number of severe cognitive declines in Table 1 is small, we combined the last two columns into one column.

Table 1. Characteristics of the asthmatic patients and their cognitive decline severity. Adapted from [14]

Factors	No Cog. Decline	Mild Cog. Decline	Moder. Cog. Decline	Severe Cog. Decline
Total	N (%)	N (%)	N (%)	N (%)
	68 (50.4%)	45 (33.3%)	16 (11.8%)	6 (4.4%)
		Age		
25-40	30 (44.1%)	17 (37.8%)	6 (37.5%)	2 (33.3%)
>40	38 (55.9%)	28 (62.2%)	10 (62.5%)	4 (66.7%)
		Education		
10 or less	53 (77.9%)	32 (71.1%)	12 (75%)	4 (66.7%)
>10	15 (22.1%)	13 (28.9%)	4 (25%)	2 (33.3%)
		Duration of Illness		
<5 years	63 (92.6%)	32 (71.1%)	13 (81.2%)	3 (50%)
>5 years	05 (7.4%)	13 (28.9%)	3 (18.8%)	3 (50%)
		Tobacco Smoking		
Non Smoker	34 (50%)	16 (35.5%)	5 (31.2%)	2 (33.3%)
Smoker	34 (50%)	29 (64.5%)	11 (68.2%)	4 (66.7%)
		Poly-Pharmacy		
No	36 (52.9%)	12 (26.6%)	4 (25%)	3 (50%)
Yes	32 (47.1%)	33 (73.4%)	12 (75%)	3 (50%)

Note: Cog. – cognitive; Moder. – moderate

In addition, we transformed the frequency of patients in each cell of the table into relative frequency such that their sum in every string equals to 1 (see Table 2).

As a result, we obtain for each of the five factors two relative frequency (probability) distributions of the categorical variable Cognitive Decline Severity containing three levels (categories): No Cognitive Decline, Mild Cognitive Decline, and Moderate or Severe Cognitive Decline.

Denoting the first of each pair of distributions by P and the second by Q we calculate relationships between them as follows: dissimilarity $D(P,Q)$ by (7), similarity $S(P,Q)$ by (2) and correlation $A(P,Q)$ by (8). The results are presented below:

Age:

25-40: P = (0.545, 0.31, 0.145),
>40: Q = (0.475, 0.35, 0.175),
$D(P,Q)$: 0.0100,
$S(P,Q)$: 0.9900;
$A(P,Q)$: 0.9800;

Education:

10 or less: P = (0.52, 0.32, 0.16),
>10: Q = (0.44, 0.38, 0.18),
$D(P,Q)$: 0.0372,
$S(P,Q)$: 0.9628,
$A(P,Q)$: 0.9256;

Duration of Illness:

<5 years: P = (0.57, 0.29, 0.14),
>5 years : Q = (0.21, 0.54, 0.25),
$\boldsymbol{D(P,Q)}$**: 0.6464,**
$\boldsymbol{S(P,Q)}$**: 0.3536,**
$\boldsymbol{A(P,Q)}$**: -0.2928**;

Tobacco Smoking:

Non Smoker: P = (0.60, 0.28, 0.12),
Smoker: Q = (0.44, 0.37, 0.19),
$D(P,Q)$: 0.0513,
$S(P,Q)$: 0.9487,
$A(P,Q)$: 0.8973;

Poly-Pharmacy:

No: P = (0.65, 0.22, 0.13),
Yes: Q = (0.40, 0.41, 0.19),
$D(P,Q)$: 0.2030,
$S(P,Q)$: 0.7970,
$A(P,Q)$: 0.5941.

One can see that relative frequency distributions P and Q of the categorical variable Cognitive Decline Severity for both levels of factors Age, Education, and Tobacco Smoking are very similar. The similarity between them is more than 0.94, and the correlation is more than 0.89. A change in levels of factors does not cause a significant change in distributions. For this reason, one can conclude that Cognitive Decline Severity is not associated with these factors or that these associations are insignificant.

Table 2. Characteristics of the asthmatic patients and their cognitive decline severity (modified Table 1)

Factors	No Cog. Decline	Mild. Cog. Decline	Moder. + Severe Cognitive Decline
Total	N=68	N=45	N=22
Age			
25-40 (n=55)	30 (0.545)	17 (0.31)	8 (0.145)
>40 (n=80)	38 (0.475)	28 (0.35)	14 (0.175)
Education			
10 or less (n=101)	53 (0.52)	32 (0.32)	16 (0.16)
>10 (n=34)	15 (0.44)	13 (0.38)	6 (0.18)
Duration of Illness			
<5 years (n=111)	63 (0.57)	32 (0.29)	16 (0.14)
>5 years (n=24)	5 (0.21)	13 (0.54)	6 (0.25)
Tobacco Smoking			
Non Smoker (n=57)	34 (0.60)	16 (0.28)	7 (0.12)
Smoker (n=78)	34 (0.44)	29 (0.37)	15 (0.19)
Poly-Pharmacy			
No (n=55)	36 (0.65)	12 (0.22)	7 (0.13)
Yes (n=80)	32 (0.4)	33 (0.41)	15 (0.19)

Note: Cog. – cognitive; Moder. - moderate

On the contrary, the frequency distributions P and Q of the two Duration of Illness factor levels have considerable differences.

The similarity is less than 0.5, the dissimilarity is greater than 0.5, and the correlation is negative (see also (6) for the relationship between these three functions).

These results allow us to conclude that Cognitive Decline Severity is associated with the factor Duration of Illness because the change in the levels of this factor causes a considerable

change in corresponding distributions. This result is consistent with the results of the work [14].

Although the difference between distributions of Poly-Pharmacy is more considerable than for the first three factors, the similarity between distributions is high, and correlation has a high positive value.

For this reason, we can conclude that the association between Cognitive Decline Severity and Poly-Pharmacy is not very high.

6 Discussion and Conclusion

The method presented in this paper allows us to measure the similarity and difference in the frequency distributions of one categorical variable for different levels of another variable. Our method is based on calculating the similarities and correlations between the rows of the contingency table.

The proposed categorical data association analysis method can be used as an additional relationship assessment to the classical chi-square analysis method.

A comparative analysis of the results obtained in our work and [14], in which the Pearson chi-square test was used, showed the same associations for four factors. At the same time, our calculations revealed a significant similarity and positive correlation (r=0.6) between the degree of cognitive decline for different levels of Poly-Pharmacy, indicating a not large association between the considered variables.

The differences obtained require more detailed further research on the relationship between our and classical methods used to analyze the association of categorical data.

Frequency distributions appear in social-behavioral sciences, biology, medicine, marketing, business, etc. [9-12, 16-18]. We plan to apply the proposed method to data analysis in some of these areas. Another possible application of the considered methods is an analysis of relationships between subjective probability distributions and subjective weight distributions in models of probability reasoning and multicriteria decision-making [19].

Acknowledgments

This work was supported by the Strategic Academic Leadership Program of Kazan Federal University, the program for developing the Scientific-Educational Mathematical Center of Volga Federal District, and the project IPN SIP 20220857.

References

1. **Batyrshin, I. Z. (2015).** On definition and construction of association measures. Journal of Intelligent & Fuzzy Systems, Vol. 29, No. 6, pp. 2319–2326. DOI: 10.3233/IFS-151930.
2. **Batyrshin, I. Z. (2019).** Constructing correlation coefficients from similarity and dissimilarity functions. Acta Polytechnica Hungarica. Vol. 16, No. 10, pp. 191–204.
3. **Batyrshin, I. Z. (2019).** Data science: Similarity, dissimilarity and correlation functions. In: Artificial Intelligence, Springer, Cham, pp. 13-28. DOI: 10.10 07/978-3-030-33274-7_2.
4. **Chen, P. Y., Popovich, P. M. (2002).** Correlation: Parametric and nonparametric measures. Sage, Thousand Oaks, CA.
5. **Gibbons, J. D., Chakraborti, S. (2003).** Nonparametric statistical inference. 4th ed. Dekker, New York.
6. **Batyrshin, I. (2019).** Towards a general theory of similarity and association measures: similarity, dissimilarity, and correlation functions. Journal of Intelligent and Fuzzy Systems, Vol. 36, No. 4, pp. 2977–3004. DOI: 10.3233/JIFS-181503.
7. **Rudas, I. J., Batyrshin I.Z. (2023).** Explainable correlation of categorical data and bar charts. Recent Developments and the New Directions of Research, Foundations, and Applications. Springer Cham. vol. 1.
8. **Batyrshin, I. Z. (2021).** Contracting and involutive negations of probability distributions. Mathematics, Vol. 9, No. 19, p. 2389. DOI:10.3390/math9192389.

9. **Agresti, A. (2002).** Categorical data analysis. 2nd ed. John Wiley & Sons, Hoboken, New Jersey.
10. **Tang, W., He, H., Tu, X. M. (2012).** Applied categorical and count data analysis. CRC Press.
11. **Simonoff, J. S. (2003).** Analyzing categorical data. Springer, New York. Vol. 496.
12. **Azen, R., Walker, C. M. (2021).** Categorical data analysis for the behavioral and social sciences. 2nd ed. Routledge, New York. Pp. 296, DOI: 10.4324/97 80203843611.
13. **Irani, F., Barbone, J. M., Beausoleil, J., Gerald, L. (2017).** Is asthma associated with cognitive impairments? A metaanalytic review. Journal of clinical and experimental neuropsychology, Vol. 39, No. 10, pp. 965–978. DOI: 10.1080/13803395.20 17.1288802.
14. **Haq Satti, R. R. U., Rasheed, S. A., Gul, R., Athar, M.H. (2022).** Frequency of cognitive decline in asthma patients and associated socio-demographic factors. PAFMJ, Vol. 72, pp. S114– S117.
15. **Batyrshin, I. Z., Kubysheva, N. I., Bayrasheva, V. R., Kosheleva, O., Kreinovich, V. (2021).** Negations of probability distributions: A survey. Computación y Sistemas, Vol. 25, No. 4, pp. 775–781. DOI: 10.13053/cys-25-4-4094.
16. **Hancock, J.T., Khoshgoftaar, T. M. (2020).** Survey on categorical data for neural networks. Journal of Big Data, Vol. 7, No. 1, pp. 1–41. DOI: 10.1186/s40537-020-00305-w.
17. **Albright, S. C., Winston, W. L. {2019).** Business analytics: Data analysis & decision making. 7th ed, Cengage Learning.
18. **Camm, J., Cochran, J., Fry, M., Ohlmann, J., Anderson, D. (2019).** Business Analytics: descriptive, predictive, prescriptive. 3rd ed. Cengage Learning.
19. **Batyrshin, I.Z. (2022).** Fuzzy Distribution Sets. Computación y Sistemas, Vol. 26, No. 3, pp. 1411–1416. DOI: 10.13053/CyS-26-3-4360.

Article received on 08/07/2022; accepted on 25/10/2022.
Corresponding author is Ildar Z. Batyrshin.