



UNIVERSITÀ DI PARMA

ARCHIVIO DELLA RICERCA

University of Parma Research Repository

Measuring variability and association for categorical data

This is the peer reviewed version of the following article:

Original

Measuring variability and association for categorical data / Allaj, E.. - In: FUZZY SETS AND SYSTEMS. - ISSN 0165-0114. - 421:(2021), pp. 29-43. [10.1016/j.fss.2020.11.018]

Availability:

This version is available at: 11381/2911574 since: 2024-11-20T14:27:41Z

Publisher:

Elsevier B.V.

Published

DOI:10.1016/j.fss.2020.11.018

Terms of use:

Anyone can freely access the full text of works made available as "Open Access". Works made available

Publisher copyright

note finali coverpage

(Article begins on next page)

28 March 2025

Measuring variability and association for categorical data

Erindi Allaj

Johannes Kepler University, Austria*

*Address for correspondence: Erindi Allaj, Department of Applied Statistics, Johannes Kepler University Linz, Linz, Austria E-mail: erindi.allaj@jku.at

Abstract

The quantification of the variability of the categorical data is an important topic not only in statistics, but also in many other disciplines. We suggest different variability measures to describe the variability of categorical data. In our approach, any set of categorical data for a determinate categorical variable is treated as a fuzzy set. Therefore, measuring the variability of categorical data is the same as measuring its fuzziness. Different measures of association between two categorical variables are also proposed. The measures can be easily applied to categoric random variables.

Keywords: Variability measures; Association measures; Fuzzy sets; Relative frequencies; Categorical random variables

1 Introduction

Measuring the variability of a given variable in a dataset is an important issue in many scientific fields. Of the same importance is the problem of determining the degree of association between two determinate categorical variables. While there is a general consensus among scholars on the measures for determining the variability and association between continuous variables, there is no general consensus on the appropriate variability and association measures for categorical data.

The main difficulty arises because of the particular nature of categorical data. [Agresti, 2012] provides an excellent introduction to categorical data analysis and [Kader and Perry, 2007] to the concept of variability of these data. Indeed, the majority of authors argue that arithmetic operations are not possible with categorical data. These authors rely on the work of [Stevens, 1946] who asserts that arithmetic operations on these data are a nonsense. On the contrary, other authors inspired from [Velleman and Wilkinson, 1993] believe that doing mathematical operations with categorical data is senseful. In the paper, we will embrace the point of view of the first group of authors.

Even though they are often known under the name diversity or similarity measures, several measures are proposed from the first group of authors to capture the variability in the categorical data. [Gini, 1912], [Shannon, 1948], [Gambaryan *et. al.*, 1964], [Goodall, 1966], [Hill, 1973], [Stanfill and Waltz, 1986] and [Eskin *et. al.*, 2002] suggest different measures capable to deal with the variability of categorical data. The measures are obtained as functions of the frequencies or relative frequencies of the various categories characterizing a given categorical variable. [Boriah *et. al.*, 2008] and [Alamuri *et. al.*, 2014] provide a good overview of different variability measures used for categorical data.

[Burbea and Rao, 1982] propose the ϕ -entropies measures assuming ϕ is a continuous concave function of the probability of a given category of a multinomial random variable. Shannon

entropy [Shannon, 1948] is a special case. [Salicru *et. al*, 1993] generalize these measures by proposing the (h, ϕ) -entropies measures where ϕ is assumed to be concave and h differentiable and increasing or ϕ convex and h differentiable and decreasing. [Pardo, 2018] in addition to offering an introduction to these and other variability measures also gives an excellent discussion of their mathematical and statistical properties. [Rao, 1982a, Rao, 1982b] asserts that any non-negative real valued concave function defined on the space of probability distributions taking the value of zero if and only if the distribution is degenerate is an ideal candidate for a measure of diversity or variability. A classical example regards the Gini-Simpson index [Gini, 1912] defined as one minus the sum of the squares of the probabilities of the single categories. One can safely apply these measures to categorical data by simply taking the maximum likelihood estimator of the probability of the single category (relative frequency) instead of the single probabilities.

[Allaj, 2018] also suggests a measure of variability obtained as one minus the Euclidean norm of the vector of the relative frequencies of the different categories composing a categorical variable. The variability measure it is shown to be a function of [Gini, 1912] index. The measure also falls in the (h, ϕ) -entropies measures if one assumes that ϕ is convex and h differentiable and decreasing.

Measuring the association between categorical variables is still a major issue in the statistics field, and other fields. The classical tool used to measure the degree of association between two categorical variables is the Pearson's chi-squared test of independence [Pearson, 1916]. Many other measures like *phi*, contingency coefficient and Cramer's V are also based on the chi-square test. [Fávero *et. al*, 2019] offer an illustration of these measures. Other measures include, for example, Goodman-Kruskal's lambda and gamma [Goodman *et. al*, 1979], Spearman's rho and Kendall's tau [Kendall, 1938]. An association measure was also recently proposed by [Baak *et. al*, 2020].

Taking inspiration from the work of [Allaj, 2018] and fuzzy set theory [Zadeh, 1965], this paper proposes some new variability and association measures for categorical data.

In our setting, the variability is measured for any fuzzy set in the universe of discourse as given by the set of all categories or classes. The reader might refer to [Zimmermann, 1996] for an introduction to fuzzy set theory. We define thus the variability measures on the fuzzy power set of the universe of discourse. Therefore measuring the variability of a fuzzy set or a set of categorical data is the same as measuring the fuzziness of this set. The membership function gives the degree of the membership of a given category to a fuzzy set and measures the relative frequency of this category. To the best of our knowledge, we are the first to establish a link between the variability of categorical data and fuzziness of fuzzy sets.

As a further step, following the work of [Allaj, 2018], we study some properties of the variability measure proposed in this paper and suggest that any variability measure applied

to categorical data should satisfy these properties. Having defined the properties a variability measure should satisfy, we propose then three new variability measures. The first measure is a function of the L^p norm, $p > 1$, where the norm is applied to the vector of the relative frequencies of a fuzzy set, the second measure is a function of the L^p distance, $p > 1$, between the vectors of the relative frequencies of a fuzzy set and the fuzzy set having the maximum variability, and finally the third measure is a function of the Kullback-Leibler divergence [Kullback and Leibler, 1951] between the vectors of the relative frequencies of a fuzzy set and the maximum variability fuzzy set. The maximum variability fuzzy set is the set which has equal relative frequencies for each possible category. The measure proposed in [Allaj, 2018] is a particular case of the first measure for $p = 2$.

The measures of association between two categorical variables are defined for any fuzzy relation on the Cartesian product of the universe of discourse sets of the single variables. The elements of the universe of discourse sets are given by the categories of each variable. The general form of the association measures is given by one minus the average between the variability measures applied to the single categorical variables. It follows that one can obtain an association measure by simply changing the variability measures used to measure the variability of the single categorical variables. Using the particular form of the association measures, we prove then some properties of these measures which we believe any association measure should satisfy. Contrary to many other measures of association proposed for categorical data which are concerned with establishing whether a given observed frequency distribution differs from a theoretical distribution, our association measures are functions only of the frequency distribution.

Finally, we show through an example that our proposed measures easily apply to general multinomial random variables.

The paper is structured as follows. Section 2 introduces the variability measures in the two category case. Section 3 generalizes the result in the general case of k categories. Section 4 proposes the association measures. In Section 5, we show that the variability and the association measures can also be easily computed for categorical random variables. The final section concludes.

2 Measuring variability in the two category case

Let X denote the set of all categories or classes, the universe of discourse, i.e. the set of positive integers including 0 as given by $X = \{0, 1, 2, \dots, k - 1\}$, where $k \geq 2$ is an integer. A fuzzy set in X is defined as follows.

Definition 2.1 *A fuzzy set \tilde{A} in the universe of discourse X is a set of ordered pairs given as*

follows

$$\tilde{A} = \{(x, f_{\tilde{A}}(x)) : x \in X\} \quad (1)$$

where the mapping $f_{\tilde{A}} : X \rightarrow N$ associates to each value of X the degree of membership in \tilde{A} with N giving the membership space. The space N is the unit interval $[0, 1]$.

The membership function $f_{\tilde{A}}(x)$ measures the relative frequency of the category x in \tilde{A} . A value of zero of this function indicates zero membership of the category x to the fuzzy set \tilde{A} . This means that no elements fall in this category. When, on the other side, $f_{\tilde{A}}(x)$ assumes the value of one, the degree of the membership of x to \tilde{A} is full. Of course, in this case all the elements fall in the category x . Another characteristic of the membership function $f_{\tilde{A}}(x)$ is that the cardinality, the sum of the relative frequencies of each distinct category is equal to one, i.e. $\sum_{x \in X} f_{\tilde{A}}(x) = 1$. These properties of the membership function are summarized in the following definition.

Definition 2.2 *The membership function $f_{\tilde{A}}(x)$ of a fuzzy set \tilde{A} satisfies the properties.*

1. $0 \leq f_{\tilde{A}}(x) \leq 1$.
2. $f_{\tilde{A}}(x) = 0$ when no elements fall in the category x .
3. $f_{\tilde{A}}(x) = 1$ when all the elements fall in the category x .
4. $\sum_{x \in X} f_{\tilde{A}}(x) = 1$.

Suppose now that you collect data on two elements on a given categorical variable having two levels. Then, $\tilde{A}_1 = \{(0, 1), (1, 0)\}$ and $\tilde{A}_2 = \{(0, 0), (1, 1)\}$ are the Crisp sets corresponding to the extreme cases where all the elements lie within a single category. On the other hand, the fuzzy set $\tilde{A}_3 = \{(0, 1/2), (1, 1/2)\}$ represents the situation where elements are equally distributed across the two categories.

The outcome $(1, 1)$, meaning that the first element falls in the first category and the second one in the second category, has the highest variability (see [Allaj, 2018]) if one uses the variability measure

$$v_k = 1 - \|\mathbf{f}_k\| = 1 - \sqrt{f_0^2 + f_1^2 + \dots + f_{k-1}^2} \quad (2)$$

or

$$v_{k,s} = \frac{v_k}{1 - \frac{1}{\sqrt{k}}} \quad (3)$$

where $k \geq 2$ gives the number of categories and $f_i = \frac{n_i}{n}$, with n_i giving the number of elements falling in category $i = 0, 1, \dots, k-1$ and n the total number of elements, measures the proportion of elements falling in category i .

On the opposite, the outcomes $(0, 1)$ and $(1, 0)$ have the lowest variability. [Allaj, 2018] gives also an interesting geometric interpretation of the variability of the different possible outcomes corresponding to the case when we have only two categories. Assuming that the number of elements is equal to n , we have that each possible outcome has a relative frequencies vector that can be represented geometrically as a vector in the positive quadrant of the (f_0, f_1) space. In Figure 1, we show five possible outcomes with relative frequencies vectors given by $\vec{A} = (1/2, 1/2)$, $\vec{B} = (1/3, 2/3)$, $\vec{C} = (2/3, 1/3)$, $\vec{D} = (0, 1)$, and $\vec{E} = (1, 0)$.

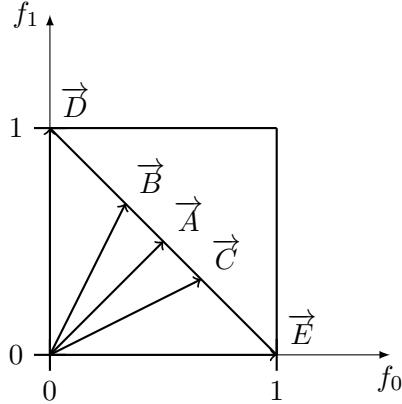


Figure 1: Variability illustration - Two category case

The measures v_k and $v_{k,s}$ assume the maximum value when the vector of the relative frequencies is equal to $\vec{A} = (1/2, 1/2)$. Clearly, the relative frequencies of each category corresponding to each possible outcome can be represented through a point lying on the line segment shown in Figure 1. Also, any relative frequencies vector can be expressed as a linear combination of the two vectors (or vertices) $(0, 1)$ and $(1, 0)$, the outcomes with the lowest variability. Finally, there is an inverse relationship between the Euclidean norm and the variability.

The reader will notice that the geometrical interpretation of the variability presented here is quite similar to the geometrical interpretation of the fuzzy sets offered by [Kosko, 1992]. Indeed, every fuzzy set \tilde{A} can be visualized as a point on the line segment shown in Figure 1 where the x -axis gives the degree of membership of category 0 to \tilde{A} and the y -axis the degree of membership of category 1 to \tilde{A} . Therefore, following [Kosko, 1992], the fuzzy set \tilde{A}_3 has the maximum fuzziness and the fuzzy sets \tilde{A}_1 and \tilde{A}_2 the minimum fuzziness. The last two fuzzy sets as mentioned before are viewed as Crisp sets. We will assume henceforth that the fuzzy set with the maximum fuzziness corresponds to the outcome with the maximum variability. Thus, the degree of fuzziness of a fuzzy set gives also the variability of the corresponding outcome. Note that using again the results in [Allaj, 2018], the number of all possible fuzzy and Crisp sets in the two-category case when n elements are available is equal to $n + 1$.

Hereafter, for a given fuzzy set \tilde{A} we let $v_2(\tilde{A})$ and $v_{2,s}(\tilde{A})$ denote

$$\begin{aligned} v_2(\tilde{A}) &= 1 - \sqrt{f_{\tilde{A}}^2(0) + f_{\tilde{A}}^2(1)} \\ v_{2,s}(\tilde{A}) &= \frac{v_2(\tilde{A})}{1 - \frac{1}{\sqrt{2}}} \end{aligned} \quad (4)$$

We shall now discuss some properties of the second fuzziness or the variability measure. It is not difficult to show that $v_{2,s}(\tilde{A})$ satisfies the following properties.

Definition 2.3 *The measure $v_{2,s}(\tilde{A})$ satisfies the following properties.*

1. Let \tilde{A} be a fuzzy set. Then $v_{2,s}(\tilde{A}) = 0$ if and only if \tilde{A} is a Crisp set, i.e. $\tilde{A} = \{(0, 0), (1, 1)\}$ or $\tilde{A} = \{(0, 1), (1, 0)\}$.
2. Let \tilde{A} be a fuzzy set. Then $v_{2,s}(\tilde{A})$ assumes its absolute maximum equal to one at $f_{\tilde{A}}(0) = 1/2$ and $f_{\tilde{A}}(1) = 1/2$.
3. Let \tilde{A} and \tilde{B} be two fuzzy sets. Then, $v_{2,s}(\tilde{A}) \geq v_{2,s}(\tilde{B})$ if $f_{\tilde{A}}(0) \geq f_{\tilde{B}}(0)$ or $f_{\tilde{A}}(1) \geq f_{\tilde{B}}(1)$ for $f_{\tilde{A}}(0) \leq \frac{1}{2}$ or $f_{\tilde{A}}(1) \leq \frac{1}{2}$. Similarly, $v_{2,s}(\tilde{A}) \geq v_{2,s}(\tilde{B})$ if $f_{\tilde{A}}(0) \leq f_{\tilde{B}}(0)$ or $f_{\tilde{A}}(1) \leq f_{\tilde{B}}(1)$ for $f_{\tilde{A}}(0) \geq \frac{1}{2}$ or $f_{\tilde{A}}(1) \geq \frac{1}{2}$.
4. If \tilde{A} is a fuzzy set, then its complement \tilde{A}^c satisfies the equality $v_{2,s}(\tilde{A}) = v_{2,s}(\tilde{A}^c)$.

The first and the second property are proved in [Allaj, 2018]. They assert, respectively, that the variability measure is bounded below by 0 and above by 1. The last property follows easily by the symmetry of the Euclidean norm. This means that whenever $f_{\tilde{A}}(0)$ is equal to $f_{\tilde{B}}(1)$ and $f_{\tilde{A}}(1)$ equal to $f_{\tilde{B}}(0)$, the variability of the fuzzy sets \tilde{A} and \tilde{B} are the same. For example, outcomes with vector of relative frequencies equal to \vec{B} and \vec{C} in Figure 1 have the same variability. To see this, let $\tilde{A} = \{(0, f_{\tilde{A}}(0)), (1, f_{\tilde{A}}(1))\}$. We have then that $\tilde{A}^c = \{(0, 1 - f_{\tilde{A}}(0)), (1, 1 - f_{\tilde{A}}(1))\}$ and

$$\begin{aligned} v_{2,s}(\tilde{A}) &= \frac{1 - \sqrt{f_{\tilde{A}}^2(0) + f_{\tilde{A}}^2(1)}}{1 - \frac{1}{\sqrt{2}}} \\ v_{2,s}(\tilde{A}^c) &= \frac{1 - \sqrt{(1 - f_{\tilde{A}}(0))^2 + (1 - f_{\tilde{A}}(1))^2}}{1 - \frac{1}{\sqrt{2}}} \end{aligned} \quad (5)$$

The result then follows. The third property says that when a given fuzzy set is Crisper ("closer" to the fuzzy sets $\tilde{A} = \{(0, 0), (1, 1)\}$ or $\tilde{A} = \{(0, 1), (1, 0)\}$) than any other fuzzy set, it has a lower variability. In a similar fashion, a given fuzzy set has a higher variability when it is closer to the fuzzy set $\tilde{A} = \{(0, 1/2), (1, 1/2)\}$. These can be noticed by writing $v_{2,s}(\tilde{A})$ as

$$v_{2,s}(\tilde{A}) = \frac{1 - \sqrt{2f_{\tilde{A}}(0)(f_{\tilde{A}}(0) - 1) + 1}}{1 - \frac{1}{\sqrt{2}}} \quad (6)$$

or

$$v_{2,s}(\tilde{A}) = \frac{1 - \sqrt{2f_{\tilde{A}}(1)(f_{\tilde{A}}(1) - 1) + 1}}{1 - \frac{1}{\sqrt{2}}} \quad (7)$$

and taking then the first derivative of $v_{2,s}(\tilde{A})$ with respect to $f_{\tilde{A}}(0)$

$$v'_{2,s}(\tilde{A}) = \frac{-[2f_{\tilde{A}}(0)(f_{\tilde{A}}(0) - 1) + 1]^{-\frac{1}{2}}(2f_{\tilde{A}}(0) - 1)}{1 - \frac{1}{\sqrt{2}}} \quad (8)$$

or $f_{\tilde{A}}(1)$

$$v'_{2,s}(\tilde{A}) = \frac{-[2f_{\tilde{A}}(1)(f_{\tilde{A}}(1) - 1) + 1]^{-\frac{1}{2}}(2f_{\tilde{A}}(1) - 1)}{1 - \frac{1}{\sqrt{2}}} \quad (9)$$

The derivatives do not exist only at the extreme points where $f_{\tilde{A}}(0) = 0$ ($f_{\tilde{A}}(1) = 1$) or $f_{\tilde{A}}(0) = 1$ ($f_{\tilde{A}}(1) = 0$). But, these are the relative frequencies corresponding to the fuzzy sets having the minimum variability. It follows that property 3 is also valid for these points.

Properties 1-4 of the variability measure $v_{2,s}(\tilde{A})$ are very close to the properties a measure of fuzziness should satisfy [De Luca and Termini, 1972]. See also [Xuecheng, 1992] for a more recent citation. Therefore, from now on, any measure satisfying properties 1-4 is a suitable candidate for measuring the variability of categorical data in the two category case.

Having determined the properties a variability measure should satisfy, it will also be interesting to introduce other measures capable to measure the variability of categorical data. Using results in fuzzy set theory, we can define a new distance measure as follows

$$D_2(\tilde{A}, \tilde{A}^c) = 1 - \frac{\left[\sum_{i=0}^1 |f_{\tilde{A}}(x_i) - f_{\tilde{A}^c}(x_i)|^2 \right]^{\frac{1}{2}}}{\sqrt{2}} \quad (10)$$

where $x_0 = 0$ and $x_1 = 1$.

The variability measure $D_2(\tilde{A}, \tilde{A}^c)$ is similar to the measure of fuzziness introduced by [Yager, 1979]. The only difference is in the denominator. In our setting, the interpretation of this measure is very intuitive. Indeed, suppose for example that $\tilde{B} = \{(0, 1/3), (1, 2/3)\}$. Therefore, $\tilde{B}^c = \{(0, 2/3), (1, 1/3)\}$. We display the vectors of the relative frequencies of these two fuzzy sets graphically in Figure 2, where \vec{B} , \vec{D} and \vec{E} are as before, while $\vec{M} = (1/2, 1/2)$ and $\vec{B}^c = (2/3, 1/3)$.

It is easy to see that the Euclidean distance between vectors \vec{B} and \vec{B}^c or the Euclidean norm of the vector difference $\vec{B} - \vec{B}^c$ is the same as the Euclidean norm of the hypotenuse of the right triangle with vertices $(1/3, 1/3)$, $(1/3, 2/3)$ and $(2/3, 1/3)$. Using a geometrical argument, this norm can also be derived as the sum of the Euclidean norm of the hypotenuses of the right triangle with vertices $(1/2, 1/3)$, $(1/2, 1/2)$ and $(2/3, 1/3)$ and of the right triangle with vertices $(1/3, 1/2)$, $(1/3, 2/3)$ and $(1/2, 1/2)$. Now, by symmetry, it is also not difficult to conclude that

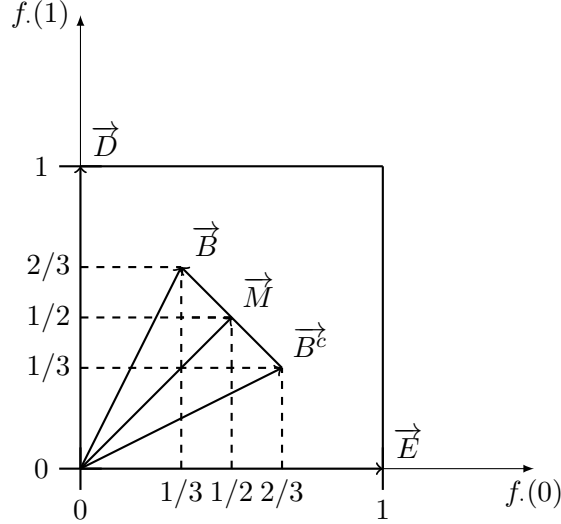


Figure 2: Variability illustration - The $D_2(\tilde{A}, \tilde{A}^c)$ measure case

the Euclidean norms of the hypotenuses of the last two right triangles have the same value. Consequently, $D_2(\tilde{A}, \tilde{A}^c)$ can be re-expressed as

$$D_2(\tilde{A}, \tilde{A}^c) = 1 - \frac{2 \left[\sum_{i=0}^1 |f_{\tilde{A}}(x_i) - f_{\tilde{M}}(x_i)|^2 \right]^{\frac{1}{2}}}{\sqrt{2}} \quad (11)$$

where clearly the vector $\vec{M} = (1/2, 1/2)$ gives the relative frequencies or the membership values of the fuzzy set with the maximum variability.

Using this new expression for the variability measure $D_2(\tilde{A}, \tilde{A}^c)$, we can assert that $D_2(\tilde{A}, \tilde{A}^c)$ depends on the distance between the vector of the relative frequencies of \tilde{A} and the vector of relative frequencies of \tilde{M} , the fuzzy set with the highest variability. As this distance increases, the variability of $D_2(\tilde{A}, \tilde{A}^c)$ approaches zero and as this distance decreases, the variability of $D_2(\tilde{A}, \tilde{A}^c)$ tends to one.

The properties 1-4 valid for measure $v_{2,s}(\tilde{A})$ are also satisfied by $D_2(\tilde{A}, \tilde{A}^c)$. One can prove these geometrically using Figure 2. Algebraically, it is sufficient writing $D_2(\tilde{A}, \tilde{A}^c)$ as

$$D_2(\tilde{A}, \tilde{A}^c) = 1 - \frac{2 \left[|f_{\tilde{A}}(0) - \frac{1}{2}|^2 + |\frac{1}{2} - f_{\tilde{A}}(0)|^2 \right]^{\frac{1}{2}}}{\sqrt{2}} \quad (12)$$

By equating $D_2(\tilde{A}, \tilde{A}^c)$ to zero and squaring both sides of the equation, we get that

$$|f_{\tilde{A}}(0) - \frac{1}{2}|^2 + |\frac{1}{2} - f_{\tilde{A}}(0)|^2 = \left(\frac{\sqrt{2}}{2} \right)^2 = \frac{1}{2} \quad (13)$$

Noticing that $|f_{\tilde{A}}(0) - \frac{1}{2}| = |\frac{1}{2} - f_{\tilde{A}}(0)|$, we can further re-write the above equation as

$$|f_{\tilde{A}}(0) - \frac{1}{2}| = \frac{1}{2} \quad (14)$$

which is true for $f_{\tilde{A}}(0) = 0$ or $f_{\tilde{A}}(0) = 1$. Substituting the latter values into Equation (12), we can also find that $D_2(\tilde{A}, \tilde{A}^c) = 0$. Therefore, we have proved the 'only if' and 'if' part of property 1.

Take now the first derivative of $D_2(\tilde{A}, \tilde{A}^c)$ with respect to $f_{\tilde{A}}(0)$ to get

$$D'_2(\tilde{A}, \tilde{A}^c) = \frac{-\left(|f_{\tilde{A}}(0) - \frac{1}{2}|^2 + \left|\frac{1}{2} - f_{\tilde{A}}(0)\right|^2\right)(4f_{\tilde{A}}(0) - 2)}{\sqrt{2}} \quad (15)$$

It is now clear that $D'_2(\tilde{A}, \tilde{A}^c) \geq 0$ whenever $f_{\tilde{A}}(0) \leq \frac{1}{2}$ and negative whenever $f_{\tilde{A}}(0) \geq \frac{1}{2}$. In the same spirit we can write Equation (12) in terms of $f_{\tilde{A}}(1)$ rather than in terms of $f_{\tilde{A}}(0)$. Thus, we have just shown property 2 and 3 of Definition 3.1. Note that property 2 holds given that the norm function is strictly convex. The last property of Definition 3.1 follows from Equation (10) or (12).

The last measure we would like to discuss is the Shannon entropy [Shannon, 1948]. [De Luca and Termini, 1972] suggest to use the entropy measure in order to measure the fuzziness of a given fuzzy set. Their entropy measure is defined by summing the entropy of a fuzzy set \tilde{A} to the entropy of the complement of this fuzzy set, i.e.

$$d(\tilde{A}) = H(\tilde{A}) + H(\tilde{A}^c) \quad (16)$$

where $H(\tilde{A})$ in the two category case is given by

$$H(\tilde{A}) = -K \sum_{i=0}^1 f_{\tilde{A}}(x_i) \ln f_{\tilde{A}}(x_i) \quad (17)$$

where K is a positive constant. Using the fact that $f_{\tilde{A}^c}(x_i) = 1 - f_{\tilde{A}}(x_i)$ and that $f_{\tilde{A}}(1) = 1 - f_{\tilde{A}}(0)$, we can write $d(\tilde{A})$ as

$$\begin{aligned} d(\tilde{A}) &= -K \sum_{i=0}^1 f_{\tilde{A}}(x_i) \ln f_{\tilde{A}}(x_i) - K \sum_{i=0}^1 f_{\tilde{A}^c}(x_i) \ln f_{\tilde{A}^c}(x_i) \\ &= -K \sum_{i=0}^1 f_{\tilde{A}}(x_i) \ln f_{\tilde{A}}(x_i) - K \sum_{i=0}^1 (1 - f_{\tilde{A}}(x_i)) \ln(1 - f_{\tilde{A}}(x_i)) \\ &= -K f_{\tilde{A}}(0) \ln f_{\tilde{A}}(0) - K f_{\tilde{A}}(1) \ln f_{\tilde{A}}(1) - K f_{\tilde{A}}(1) \ln f_{\tilde{A}}(1) - K f_{\tilde{A}}(0) \ln f_{\tilde{A}}(0) \\ &= -2K \sum_{i=0}^1 f_{\tilde{A}}(x_i) \ln f_{\tilde{A}}(x_i) \end{aligned} \quad (18)$$

We can conclude that in our setting $d(\tilde{A})$ reduces to the classical Shannon entropy measure. As a result, we define the entropy measure to be equal to

$$d_2(\tilde{A}, \tilde{M}) = 1 - \frac{\sum_{i=0}^1 f_{\tilde{A}}(x_i) \ln \frac{f_{\tilde{A}}(x_i)}{f_{\tilde{M}}(x_i)}}{\ln 2} \quad (19)$$

with the convention that $0 \ln 0 = 0$.

The measure $d_2(\tilde{A}, \tilde{M})$ compares the set \tilde{A} to the maximum variability set \tilde{M} . What matters thus in the computation of this measure is the dissimilarity between the fuzzy set \tilde{A} and the fuzzy set having the maximum variability. As can be observed, the measure is similar to the Kullback-Leibler divergence [Kullback and Leibler, 1951]. We can thus interpret the fuzzy set \tilde{M} as the state of the world and the fuzzy set having relative frequencies situated at the midpoint of the line in Figure 1.

Obviously, this measure tends to one when the fuzzy set \tilde{A} gets close to \tilde{M} and approaches zero when \tilde{A} tends to one of the Crisp sets. As with the previous two measures, $d_2(\tilde{A}, \tilde{M})$ satisfies the first property outlined in Definition 3.1. To derive the second and the third property it is sufficient to write $d_2(\tilde{A}, \tilde{M})$ in terms of $f_{\tilde{A}}(0)$ (or $f_{\tilde{A}}(1)$) and compute its first derivative with respect to $f_{\tilde{A}}(0)$ (or $f_{\tilde{A}}(1)$). In fact, we have that the first derivative of $-\sum_{i=0}^1 f_{\tilde{A}}(x_i) \ln \frac{f_{\tilde{A}}(x_i)}{f_{\tilde{M}}(x_i)}$ with respect to $f_{\tilde{A}}(0)$ is given by

$$-\ln \frac{f_{\tilde{A}}(0)}{f_{\tilde{M}}(0)} + \ln \frac{1 - f_{\tilde{A}}(0)}{f_{\tilde{M}}(1)} \quad (20)$$

The last property can be proved by writing $\sum_{i=0}^1 f_{\tilde{A}^c}(x_i) \ln \frac{f_{\tilde{A}^c}(x_i)}{f_{\tilde{M}}(x_i)}$ as

$$\begin{aligned} & (1 - f_{\tilde{A}}(0)) \frac{(1 - f_{\tilde{A}}(0))}{f_{\tilde{M}}(0)} + (1 - f_{\tilde{A}}(1)) \frac{(1 - f_{\tilde{A}}(1))}{f_{\tilde{M}}(1)} \\ &= f_{\tilde{A}}(1) \frac{f_{\tilde{A}}(1)}{f_{\tilde{M}}(0)} + f_{\tilde{A}}(0) \frac{f_{\tilde{A}}(0)}{f_{\tilde{M}}(1)} = \sum_{i=0}^1 f_{\tilde{A}}(x_i) \ln \frac{f_{\tilde{A}}(x_i)}{f_{\tilde{M}}(x_i)} \end{aligned} \quad (21)$$

where we have used the fact that $f_{\tilde{A}}(0) + f_{\tilde{A}}(1) = 1$ and $f_{\tilde{M}}(0) = f_{\tilde{M}}(1)$.

3 Generalization of the previous results

Let $X = \{0, 1, \dots, k - 1\}$ be our universe of discourse with $k \geq 2$. The fuzzy power set (see [Kosko, 1986]) of X denoted by $F(2^X)$ is defined as $\{\tilde{A} | \tilde{A} \text{ is a fuzzy subset of } X\}$. We also let $\vec{f}_{\tilde{A}}$ denote the k -dimensional vector $(f_{\tilde{A}}(0), f_{\tilde{A}}(1), \dots, f_{\tilde{A}}(k - 1))$ associated to the fuzzy set \tilde{A} . A measure of variability (fuzziness) for categorical data having k categories is a mapping M from $F(2^X)$ to $[0, 1]$ satisfying the following four properties.

Definition 3.1 *The measure $M : F(2^X) \rightarrow [0, 1]$ satisfies the following properties.*

1. *Let \tilde{A} be a given fuzzy set in X . Then, $M(\tilde{A}) = 0$ if and only if \tilde{A} is a Crisp set, i.e. $\vec{f}_{\tilde{A}}$ is equal to one of the natural basis vector in the \mathbb{R}^k space.*
2. *Let \tilde{A} be a given fuzzy set in X . Then, $M(\tilde{A})$ assumes its unique maximum equal to one at $\vec{f}_{\tilde{A}} = (1/k, 1/k, \dots, 1/k)$.*

3. Let \tilde{A} and \tilde{B} be two fuzzy sets such that each element of the vectors $\vec{f}_{\tilde{A}}$ and $\vec{f}_{\tilde{B}}$ is different from zero. Then $M(\tilde{A}) \leq M(\tilde{B})$ if $\sum_{x \in C(X)} f_{\tilde{A}}(x) \geq \sum_{x \in C(X)} f_{\tilde{B}}(x) \geq \frac{k-1}{k}$ and $M(\tilde{A}) \leq M(\tilde{B})$ if $\sum_{x \in C(X)} f_{\tilde{A}}(x) \leq \sum_{x \in C(X)} f_{\tilde{B}}(x) \leq \frac{k-1}{k}$, where $C(X)$ is a set contained in the set of all possible combinations of order $k-1$ of X .
4. Suppose \tilde{A} is a fuzzy set with a membership vector equal to $\vec{f}_{\tilde{A}} = (f_{\tilde{A}}(0), f_{\tilde{A}}(1), \dots, f_{\tilde{A}}(k-1))$. Then, any other fuzzy set with membership vector obtained by permuting the elements of $\vec{f}_{\tilde{A}}$ has identical variability as the original fuzzy set \tilde{A} .

Remark 3.2 The reader will have observed that the standard definition of the complement of a fuzzy set is not valid in the general case of data having k categories. Indeed, taking the complement of a fuzzy set we would have that the membership function would fail to satisfy property 4 of Definition 2.2. Note also that property 4 in Definition 2.3 in the case of the two category case is the same as property 4 in Definition 3.1 and that the number of possible permutations (with repetition) including the original set is equal to $\frac{k!}{n_1!n_2!\dots n_k!}$ where $n_i!$ gives the number of times the same numerical value of f_i occurs.

Remark 3.3 As mentioned in the previous section, properties in Definition 3.1 are close to those proposed by [De Luca and Termini, 1972] for the measures of fuzziness. The non-negativity of M and the fact that this measure is equal to zero if and only if the relative frequencies vector is degenerate is a property of many other similarity or entropy measures proposed in literature. For example, focusing on the multinomial case, the measures defined in [Gini, 1912], [Shannon, 1948], [Rényi, 1961] and [Behara and Chawla, 1974] satisfy this property. The maximum for these measures is also achieved when the uniform distribution is used and these measures satisfy property 4 of Definition 3.1. The "if" conditions in property 3 of Definition 3.1 can also be written as $f_{\tilde{A}}(x) \leq f_{\tilde{B}}(x) \leq \frac{1}{k}$ or $f_{\tilde{A}}(x) \geq f_{\tilde{B}}(x) \geq \frac{1}{k}$ where $x \notin C(X)$. This indicates that M is a non-decreasing function of $f(x)$ on the interval $(0, \frac{1}{k}]$ or non-increasing on the interval $[\frac{1}{k}, 1)$ for a value of $x \notin C(X)$. The property extends naturally property 3 of Definition 2.3 to the general case of k categories. The four properties proposed in Definition 3.1 do not define only a specific functional form for the measure M like the Shannon-Khinchin axioms [Shannon, 1948, Khinchin, 1957] and do not require M to be concave [Rao, 1982a, Rao, 1982b] or to satisfy differentiability and concavity-convexity assumptions as in [Burbea and Rao, 1982] and [Salicru et. al, 1993].

Figure 3 illustrates the feasible region for the membership function f in the three category case. As can be seen, the (simplex) region of interest is an equilateral triangle. In general, the region would be a polytope of dimension k [Ziegler, 2012].

We measure the variability (fuzziness) of a fuzzy set \tilde{A} by using one of the following measures.

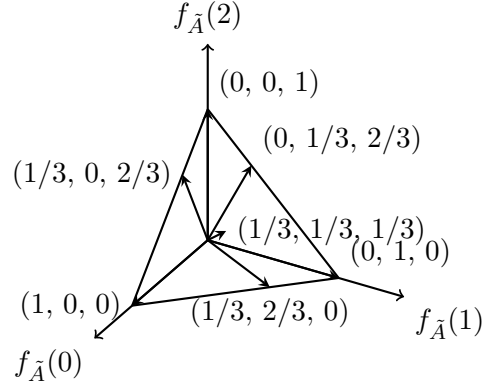


Figure 3: Variability illustration - Three category case

Definition 3.4 The measure of variability $M : F(2^X) \rightarrow [0, 1]$ is assumed to have one of the following forms.

1. $v_{k,s}^p(\tilde{A}) = \frac{1 - \|\vec{f}_{\tilde{A}}\|_p}{1 - \frac{1}{\sqrt[k]{k^{p-1}}}}$, where $\|\vec{f}_{\tilde{A}}\|_p = (\sum_{x \in X} |f_{\tilde{A}}(x)|^p)^{1/p}$ with $p > 1$.
2. $D_{p,k}(\tilde{A}, \tilde{M}) = 1 - \frac{k}{[(k-1) + (k-1)^p]^{1/p}} \|\vec{f}_{\tilde{A}} - \vec{f}_{\tilde{M}}\|_p$, where $\|\vec{f}_{\tilde{A}} - \vec{f}_{\tilde{M}}\|_p = (\sum_{x \in X} |f_{\tilde{A}}(x) - f_{\tilde{M}}(x)|^p)^{1/p}$ with $p > 1$.
3. $d_k(\tilde{A}, \tilde{M}) = 1 - \frac{\sum_{x \in X} f_{\tilde{A}}(x) \ln \frac{f_{\tilde{A}}(x)}{f_{\tilde{M}}(x)}}{\ln k}$.

where the fuzzy set \tilde{M} is the set with the maximum variability.

Remark 3.5 The measures defined in Definition 3.4 are a generalization of the measures seen in the previous section. In particular, taking $p = 2$, $v_{k,s}^p$ is equal to the measure proposed by [Allaj, 2018]. This measure also falls in the (h, ψ) -entropies measures proposed by [Salicru et al, 1993] if one takes $\psi = \sum_{x \in X} \phi(f_{\tilde{A}}(x))$ and $h = \frac{1 - \sqrt{\psi}}{1 - \frac{1}{\sqrt{k}}}$ where $\phi = |f_{\tilde{A}}(x)|^2$. Note also that $v_{k,s}^p$ is different from the entropy measure suggested by [Behara and Chawla, 1974] and the Gini-Simpson index [Gini, 1912] when $p = 2$. The former can be written as $\frac{1 - \|\vec{f}_{\tilde{A}}\|_p}{1 - \frac{1}{\sqrt[k]{k^{p-1}}}}$, $p \neq 1$, $p \geq 0$, and the latter as $1 - \sum_{x \in X} |f_{\tilde{A}}(x)|^2$.

Having reached this point, we have to show that each of the proposed variability measure satisfy the properties described in Definition 3.1. Let's focus on the measure $v_{k,s}^p(\tilde{A})$ and suppose that $v_{k,s}^p(\tilde{A}) = 0$. Then $\sum_{x \in X} |f_{\tilde{A}}(x)|^p = 1$. The inequality

$$\sum_{x \in X} |f_{\tilde{A}}(x)|^p = \sum_{i=0}^{k-1} |f_{\tilde{A}}(x_i)|^p \leq \left(\sum_{i=0}^{k-1} |f_{\tilde{A}}(x_i)| \right)^p \quad (22)$$

shows that $\sum_{x \in X} |f_{\tilde{A}}(x)|^p = 1$ only when the mixed terms on the right-hand side of this inequality are equal to zero.

This is of course achieved when $\vec{f}_{\tilde{A}}$ is equal to one of the natural basis vector in the \mathbb{R}^k space which coincides with one of the vertices of the standard simplex in the \mathbb{R}^k space. On the other hand, when $\vec{f}_{\tilde{A}}$ is equal to one of the natural basis vector in the \mathbb{R}^k space, $v_{k,s}^p(\tilde{A}) = 0$. It follows that $v_{k,s}^p(\tilde{A})$ fulfills the first property.

Taking now the first partial derivatives (see Equation (23)) of $\|\vec{f}_{\tilde{A}}\|_p$ and setting them to zero, we get that $(1/k, 1/k, \dots, 1/k)$ is a stationary point of $\|\vec{f}_{\tilde{A}}\|_p$. The first partial derivatives exist for all vectors $\vec{f}_{\tilde{A}}$ having non-zero elements. Since $\|\vec{f}_{\tilde{A}}\|_p$ is a norm, it is also convex. Results in optimization theory show then that the stationary point is a local minimizer. But given that $\|\vec{f}_{\tilde{A}}\|_p$ is defined on the simplex region $S = \{\vec{f}_{\tilde{A}} : \tilde{A} \in F(2^X), 0 \leq f_{\tilde{A}}(x) \leq 1 \text{ and } \sum_{x \in X} f_{\tilde{A}}(x) = 1, \forall x \in X\}$, this critical point is also a global minimum [Sun and Yuan, 2006]. The fact that $p > 1$ implies that $\|\cdot\|_p$ is strictly convex, and as a result this point is also the unique global minimum. It is then not difficult to conclude that the second property of Definition 3.1 is satisfied.

To check if $v_{k,s}^p(\tilde{A})$ satisfies property 3, we need first to compute the first partial derivatives of $\|\vec{f}_{\tilde{A}}\|_p = (\sum_{x \in X} |f_{\tilde{A}}(x)|^p)^{1/p} = (\sum_{i=0}^{k-1} |f_{\tilde{A}}(x_i)|^p)^{1/p} = (|f_{\tilde{A}}(0)|^p + |f_{\tilde{A}}(1)|^p + \dots + |f_{\tilde{A}}(k-2)|^p + |1 - f_{\tilde{A}}(0) - f_{\tilde{A}}(1) - \dots - f_{\tilde{A}}(k-2)|^p)^{1/p}$. Simple computations show then that these partial derivatives are equal to

$$(|f_{\tilde{A}}(0)|^p + |f_{\tilde{A}}(1)|^p + \dots + |f_{\tilde{A}}(k-2)|^p + |f_{\tilde{A}}(x_{k-1})|^p)^{\frac{1-p}{p}} [f_{\tilde{A}}(x_i) |f_{\tilde{A}}(x_i)|^{p-2} - f_{\tilde{A}}(x_{k-1}) |f_{\tilde{A}}(x_{k-1})|^{p-2}] \quad (23)$$

for every $i = 0, 1, \dots, k-2$. The partial derivatives are all non-negative when

$$f_{\tilde{A}}(x_i) \geq f_{\tilde{A}}(x_{k-1}), \quad i = 0, 1, \dots, k-2 \quad (24)$$

Summing the above inequalities over the index i , we get that

$$\sum_{i=0}^{k-2} f_{\tilde{A}}(x_i) \geq \frac{k-1}{k} \quad (25)$$

We can thus conclude that for any fuzzy set \tilde{A} and \tilde{B} , $(\sum_{x \in X} |f_{\tilde{A}}(x)|^p)^{1/p} \geq (\sum_{x \in X} |f_{\tilde{B}}(x)|^p)^{1/p}$ holds if $\sum_{i=0}^{k-2} f_{\tilde{A}}(x_i) \geq \sum_{i=0}^{k-2} f_{\tilde{B}}(x_i) \geq \frac{k-1}{k}$. The opposite still holds by requiring first partial derivatives to be non-positive. Equation (25) remains true even in the case when we decide to express a different relative frequency from $f_{\tilde{A}}(x_{k-1})$ as a function of the other relative frequencies. It follows then that $v_{k,s}^p(\tilde{A})$ satisfies property 3 of Definition 3.1.

Showing that $v_{k,s}^p(\tilde{A})$ satisfies the last property of Definition 3.1 it is a straightforward exercise.

The first two properties of Definition 3.1 can be easily shown to hold for the $D_{p,k}(\tilde{A}, \tilde{M})$ measure. In particular, for the first property we can write the condition

$$D_{p,k}(\tilde{A}, \tilde{M}) = 1 - \frac{k}{[(k-1) + (k-1)^p]^{1/p}} \|\vec{f}_{\tilde{A}} - \vec{f}_{\tilde{M}}\|_p = 0 \quad (26)$$

as

$$1 = \frac{k}{[(k-1) + (k-1)^p]^{1/p}} \|\vec{f}_{\tilde{A}} - \vec{f}_{\tilde{M}}\|_p \quad (27)$$

Given that the function $\|\vec{f}_{\tilde{A}} - \vec{f}_{\tilde{M}}\|_p$ is strictly convex, for $p > 1$, in $\vec{f}_{\tilde{A}}$, $\tilde{A} \in F(2^X)$, it assumes its maximum at the extreme points of S . The other side of the property 1 is easily proved.

One can take the first partial derivatives of $\|\vec{f}_{\tilde{A}} - \vec{f}_{\tilde{M}}\|_p$ at the points where the elements of $\vec{f}_{\tilde{A}}$ are different from zero to show that these satisfy the same system of equations defined in Equation (24) and hence property 3 of Definition 3.1 is automatically satisfied by $D_{p,k}(\tilde{A}, \tilde{M})$. Property 4 follows from the properties of the distance function.

Regarding the last measure, we can write the condition $d_k(\tilde{A}, \tilde{M}) = 0$ as

$$\frac{\sum_{x \in X} f_{\tilde{A}}(x) \ln \frac{f_{\tilde{A}}(x)}{f_{\tilde{M}}(x)}}{\ln k} = 1 \quad (28)$$

The function in the numerator is strictly convex on the unit simplex and therefore assumes its maximum at the vertices of this simplex. At these points, the above condition is satisfied. Any other point in the unit simplex would fail to satisfy Equation (28). Clearly, when $\vec{f}_{\tilde{A}}$ is equal to one of the natural basis vector, $d_k(\tilde{A}, \tilde{M}) = 0$. We have thus shown that property 1 of Definition 3.1 holds for $d_k(\tilde{A}, \tilde{M})$.

The function $d_k(\tilde{A}, \tilde{M})$ is strictly concave on the unit simplex and assumes its unique maximum at $\vec{f}_{\tilde{A}} = (1/k, 1/k, \dots, 1/k)$. Property 3 of Definition 3.1 also follows by taking the first partial derivatives of $d_k(\tilde{A}, \tilde{M})$ with respect to the relative frequencies having all the elements different from zero. To see this, note that these first partial derivatives are of the form

$$d'_k(\tilde{A}, \tilde{M}) = - \frac{\ln \frac{f_{\tilde{A}}(x_i)}{f_{\tilde{M}}(x_i)} - \ln \frac{f_{\tilde{A}}(x_{k-1})}{f_{\tilde{M}}(x_{k-1})}}{\ln k} \quad (29)$$

where $i = 0, 1, \dots, k-2$ and $f_{\tilde{A}}(x_{k-1})$ is expressed in terms of the other relative frequencies. We can then repeat the same analysis done for $v_{k,s}^p(\tilde{A})$ and $D_{p,k}(\tilde{A}, \tilde{M})$ to get the final result.

Permuting the elements of $\vec{f}_{\tilde{A}}$ does not have any influence on the variability measure.

We collect these results in the following theorem.

Theorem 3.6 *The measures proposed in Definition 3.4 satisfy the properties described in Definition 3.1.*

4 Measures of association

In this section, we introduce some association measures which we will use to measure the strength of the relationship between two categorical variables. We start the illustration of this measure with the following definition.

Definition 4.1 *Let X and Y be two universe of discourse sets of form $\{0, 1, 2, \dots, k_X - 1\}$ and $\{0, 1, 2, \dots, k_Y - 1\}$ with $k_X \geq 2$ and $k_Y \geq 2$. Then, R denote the relation*

$$R = \{(x, y), f_R(x, y) | (x, y) \in X \times Y\} \quad (30)$$

in $X \times Y$.

Clearly, R coincides with the definition of the fuzzy relation in [Zadeh, 1965, Zadeh, 1971]. In our case the membership function $f_R(x, y)$ associated to each (x, y) gives also the joint frequency of the category x and y under the relation R .

Example 4.2 *Suppose we have two sets of categorical data with the following crosstabulation*

	Y		
X	0	1	Tot.
0	1/4	1/4	1/2
1	1/4	1/4	1/2
Tot.	1/2	1/2	1

where X and Y are two universal sets both equal to $\{0, 1\}$. The membership function $f_R(x, y)$ can be expressed in this particular example as

$$f_R(x, y) = \frac{1}{k_X k_Y} = \frac{1}{4}, \quad \forall (x, y) \in X \times Y \quad (31)$$

where R is given by the following matrix

$$R = \begin{matrix} X/Y & 0 & 1 \\ 0 & \begin{bmatrix} 1/4 & 1/4 \end{bmatrix} \\ 1 & \begin{bmatrix} 1/4 & 1/4 \end{bmatrix} \end{matrix}$$

with the related fuzzy graph [Mordeson and Nair, 2012] shown in Figure 4.

The properties of the membership function $f_R(x, y)$ are listed in Definition 4.3.

Definition 4.3 *R is a relation in $X \times Y$ if*

1. $0 \leq f_R(x, y) \leq 1$.

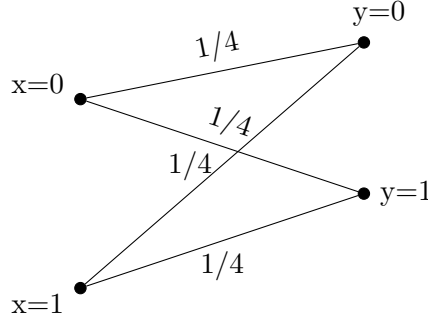


Figure 4: Illustration of the association

2. $f_{\tilde{A}_R^X}(x) = \sum_{y \in Y} f_R(x, y)$, where \tilde{A}_R^X is a fuzzy set in the universe of discourse X given as $\tilde{A}_R^X = \{(x, f_{\tilde{A}_R^X}(x)) : x \in X\}$.
3. $f_{\tilde{A}_R^Y}(y) = \sum_{x \in X} f_R(x, y)$, where \tilde{A}_R^Y is a fuzzy set in the universe of discourse Y given as $\tilde{A}_R^Y = \{(y, f_{\tilde{A}_R^Y}(y)) : y \in Y\}$.
4. $\sum_{x \in X} \sum_{y \in Y} f_R(x, y) = 1$.

A direct consequence of Definition 4.3 is that $\sum_{x \in X} f_{\tilde{A}_R^X}(x) = 1$ and $\sum_{y \in Y} f_{\tilde{A}_R^Y}(y) = 1$ which implies $f_{\tilde{A}_R^X}(x)$ and $f_{\tilde{A}_R^Y}(y)$ define a membership function for the fuzzy sets \tilde{A}_R^X and \tilde{A}_R^Y .

Suppose now that the relation R in $X \times Y$ of Example 4.2 can be expressed as

$$R = \begin{matrix} X/Y & 0 & 1 \\ 0 & \begin{bmatrix} 0 & 1 \end{bmatrix} \\ 1 & \begin{bmatrix} 0 & 0 \end{bmatrix} \end{matrix}$$

This of course corresponds to the situation where all the elements fall in the category 0 of X and category 1 of Y . This means that there is a perfect association between X and Y . Note how in this case $f_{\tilde{A}_R^X}(0) = 1$, $f_{\tilde{A}_R^X}(1) = 0$, $f_{\tilde{A}_R^Y}(0) = 0$ and $f_{\tilde{A}_R^Y}(1) = 1$ implying that $\tilde{A}_R^X = \{(0, 1), (1, 0)\}$ and $\tilde{A}_R^Y = \{(0, 0), (1, 1)\}$. Using one of the variability measures described in the previous section, the fuzzy sets \tilde{A}_R^X and \tilde{A}_R^Y can be easily seen to have the lowest variability or fuzziness. This is true whenever the sets \tilde{A}_R^X and \tilde{A}_R^Y are equal to one of the Crisp sets. We can thus establish that when a perfect association between X and Y exists, the variability of each categorical variable achieves its minimum value equal to zero.

The opposite is true when our data obey the relation R outlined in Example 4.2. In fact, the sets $\tilde{A}_R^X = \{(0, 1/2), (1, 1/2)\}$ and $\tilde{A}_R^Y = \{(0, 1/2), (1, 1/2)\}$ have now the highest variability which is also reflected in the fact that the elements are equally distributed across the joint categories of X and Y .

Motivated by these last considerations, we define the association measure between two categorical variables X and Y governed by a given relation R as in the following definition.

Definition 4.4 Let X, Y and R be defined as in Definitions 4.1 and 4.3. Then,

$$\rho_R(X, Y) = 1 - \frac{1}{2}[M(\tilde{A}_R^X) + M(\tilde{A}_R^Y)] \quad (32)$$

measures the association between X and Y governed by a given relation R .

The measures $M(\tilde{A}_R^X)$ and $M(\tilde{A}_R^Y)$ of the fuzzy sets \tilde{A}_R^X and \tilde{A}_R^Y are given by one of variability measures described in the previous section.

An immediate consequence of Definition 4.4 is that $\rho_R(X, Y)$ assumes its maximum value equal to one when $M(\tilde{A}_R^X) = 0$ and $M(\tilde{A}_R^Y) = 0$, and its minimum value equal to 0 when $M(\tilde{A}_R^X) = 1$ and $M(\tilde{A}_R^Y) = 1$, which was what we expected to get from an association measure between two categorical variables.

The properties of $\rho_R(X, Y)$ shown below follow easily from the properties of the variability measure M and ρ .

Definition 4.5 The measure $\rho_R(X, Y)$ satisfies the following properties.

1. Let R be a relation in $X \times Y$. Then, $0 \leq \rho_R(X, Y) \leq 1$.
2. Let R be a relation in $X \times Y$. Then, $\rho_R(X, Y) = 1$ (highest degree of association) if and only if $M(\tilde{A}_R^X) = 0$ and $M(\tilde{A}_R^Y) = 0$.
3. Let R be a relation in $X \times Y$. Then, $\rho_R(X, Y) = 0$ (lowest degree of association) if and only if $M(\tilde{A}_R^X) = 1$ and $M(\tilde{A}_R^Y) = 1$.
4. Suppose R is a relation in $X \times Y$. We have that $\rho_R(X, X) = M(\tilde{A}_R^X)$ and $\rho_R(Y, Y) = M(\tilde{A}_R^Y)$.
5. Suppose R is a relation in $X \times Y$. Then, $\rho_R(X, Y) = \rho_R(Y, X)$.
6. Let R be a relation in $X \times Y$ and let R' be another relation with vector of relative frequencies $\overrightarrow{f_{\tilde{A}_{R'}^X}}$ and $\overrightarrow{f_{\tilde{A}_{R'}^Y}}$ obtained by permuting the elements of $\overrightarrow{f_{\tilde{A}_R^X}} = (f_{\tilde{A}_R^X}(0), f_{\tilde{A}_R^X}(1), \dots, f_{\tilde{A}_R^X}(k_X - 1))$ or/and $\overrightarrow{f_{\tilde{A}_R^Y}} = (f_{\tilde{A}_R^Y}(0), f_{\tilde{A}_R^Y}(1), \dots, f_{\tilde{A}_R^Y}(k_Y - 1))$. Then, $\rho_{R'}(X, Y) = \rho_R(X, Y)$.
7. Suppose R is a relation in $X \times Y$ and let also $\tilde{A}_R^X, \tilde{A}_{R'}^X$ and $\tilde{A}_R^Y, \tilde{A}_{R'}^Y$ be fuzzy sets in X and Y , respectively. Suppose also that each element of the vectors $\overrightarrow{f_{\tilde{A}_R^X}}, \overrightarrow{f_{\tilde{A}_{R'}^X}}, \overrightarrow{f_{\tilde{A}_R^Y}}$ and $\overrightarrow{f_{\tilde{A}_{R'}^Y}}$ is different from zero. Then, $\rho_R(X, Y) \leq \rho_{R'}(X, Y)$ if $\sum_{i=1}^{k_X-1} f_{\tilde{A}_R^X}(x_i) \geq \sum_{i=1}^{k_X-1} f_{\tilde{A}_{R'}^X}(x_i) \geq \frac{k_X-1}{k_X}$ and $\sum_{i=1}^{k_Y-1} f_{\tilde{A}_R^Y}(y_i) \geq \sum_{i=1}^{k_Y-1} f_{\tilde{A}_{R'}^Y}(y_i) \geq \frac{k_Y-1}{k_Y}$. Similarly, $\rho_R(Y, X) \leq \rho_{R'}(Y, X)$ if $\sum_{i=1}^{k_X-1} f_{\tilde{A}_R^X}(x_i) \leq \sum_{i=1}^{k_X-1} f_{\tilde{A}_{R'}^X}(x_i) \leq \frac{k_X-1}{k_X}$ and $\sum_{i=1}^{k_Y-1} f_{\tilde{A}_R^Y}(y_i) \leq \sum_{i=1}^{k_Y-1} f_{\tilde{A}_{R'}^Y}(y_i) \leq \frac{k_Y-1}{k_Y}$.

In general, any association measure satisfying properties 1-7 is a potential candidate for measuring the association between two categorical variables.

Remark 4.6 Many association measures proposed in literature are based on the chi-squared test. Examples include the phi coefficient and the Cramer's V measure (see [Fávero et. al, 2019] for an overview of these and other measures). Recent studies also propose measures based on the chi-squared test. [Zhang, 2019] proposes an association measure based on the weighted Minkowski distance between the joint probabilities and the product between the marginal probabilities of two categorical variables. [Baak et. al, 2020] obtain a correlation coefficient for categorical variables by inverting the chi-squared test. To our knowledge, our measures of association have not been proposed yet in literature.

Example 4.7 Suppose we are interested in the relationship between gender and college major choice and that we have selected a sample of 500 individuals with data summarized in the following crosstabulation.

	Y			
X	0	1	2	Tot.
0	150	80	20	250
1	70	70	110	250
Tot.	220	150	130	500

where $X = \{Male, Female\}$ is coded as 0, 1 and $Y = \{Engineering, Business, Humanities\}$ is coded as 0, 1, 2.

The relation R is then given as

$$R = \begin{matrix} X/Y & 0 & 1 & 2 \\ 0 & \left[\begin{matrix} 150/500 & 80/500 & 20/500 \end{matrix} \right] \\ 1 & \left[\begin{matrix} 70/500 & 70/500 & 110/500 \end{matrix} \right] \end{matrix}$$

and the association measure $\rho_R(X, Y)$ is equal approximately to 0.018, 0.082 and 0.012 using the variability measures $v_{k,s}^p(\tilde{A})$, $D_{p,k}(\tilde{A}, \tilde{M})$ and $d_k(\tilde{A}, \tilde{M})$ for $p = 2$. The association is low given the high variability in the variables X and Y , equal to one in the case of X for all the variability measures and equal approximately to 0.964, 0.836 and 0.976 in the case of Y for the variability measures $v_{k,s}^p(\tilde{A})$, $D_{p,k}(\tilde{A}, \tilde{M})$ and $d_k(\tilde{A}, \tilde{M})$. We note that $v_{k,s}^p(\tilde{A})$ and $d_k(\tilde{A}, \tilde{M})$ give almostly the same result, while $D_{p,k}(\tilde{A}, \tilde{M})$ gives a slightly different result.

The situation will be slightly different if one has this relation between gender and college major choice.

$$R = \begin{matrix} X/Y & 0 & 1 & 2 \\ 0 & \left[\begin{matrix} 120/500 & 50/500 & 10/500 \end{matrix} \right] \\ 1 & \left[\begin{matrix} 10/500 & 40/500 & 270/500 \end{matrix} \right] \end{matrix}$$

Using the variability measure $v_{k,s}^p(\tilde{A})$ with $p = 2$ we get that the association measure is approximately equal to 0.124 showing a higher degree of association compared to the previous situation.

5 Variability and association measures for categorical random variables

The results obtained in the previous section remain unchanged if instead of considering relative frequencies we consider probabilities. We illustrate this with the following example.

Example 5.1 Let $X = \{Red, Blue, Black\}$ and $Y = \{Male, Female\}$ be two variables measuring respectively the color and gender. Assume also that $R = \{(x, y), p_R(x, y) | (x, y) \in X \times Y\}$ measuring the relation between color and gender has the following form

$$R = \begin{array}{c|cc} X/Y & 0 & 1 \\ \hline 0 & 0.05 & 0.3 \\ 1 & 0.2 & 0.15 \\ 2 & 0.15 & 0.15 \end{array}$$

where we have coded the categories of X by 0, 1 and 2, and the categories of Y by 0 and 1. In this example we are assuming that $p_R(x, y)$ gives the joint probability of occurrence of category x of X and y of Y . It follows that $p_{\tilde{A}_R^X}(x) = \sum_{y \in Y} p_R(x, y)$ and $p_{\tilde{A}_R^Y}(y) = \sum_{x \in X} p_R(x, y)$ give the marginal probabilities of X and Y .

As stated in the previous section, we measure the degree of association between X and Y by using $\rho_R(X, Y)$. This measure is equal approximately to 0.026, 0.125 and 0.016 if one uses the variability measure $v_{k,s}^2(\tilde{A})$, $D_{2,k}(\tilde{A}, \tilde{M})$ and $d_k(\tilde{A}, \tilde{M})$ to compute $\rho_R(X, Y)$. The values of $\rho_R(X, Y)$ show that there is a weak association between the two variables X and Y . This is the result of a high variability in the single variables X and Y . The three measures of variability described in Section 3 are equal respectively to 0.996, 0.950 and 0.998 for the variable X , and 0.952, 0.800, 0.971 for the variable Y .

Note that when the joint probabilities are all equal to $\frac{1}{k_X k_Y} = \frac{1}{6}$, the measure $\rho_R(X, Y)$ takes the value of zero. In this particular case, these probabilities can be computed assuming that the marginal probabilities are equally distributed between each category of the same variable and that X and Y are independent variables.

In general, given two multinomial random variables $X = (X_0, X_1, \dots, X_{k_X-1})$ with parameter $\mathbf{p} = (p_0, p_1, \dots, p_{k_X-1})$ and $Y = (Y_0, Y_1, \dots, Y_{k_Y-1})$ with parameter $\mathbf{q} = (q_0, q_1, \dots, q_{k_Y-1})$ we can measure the variability of the single variables by using one of the variability measures in Definition 3.4 and we can measure the degree of the association between these two random variables using our measure $\rho_R(X, Y)$.

6 Conclusions

We have proposed a link between the variability of categorical data and the degree of fuzziness of fuzzy sets. We have also proposed different properties a variability and an association measure should satisfy. The measures of variability and association for categorical data are of easy application. Overall, we believe our suggestions will improve the understanding of the concepts of variability and association applied to categorical variables.

References

- [Allaj, 2018] Allaj, E. (2018) Two simple measures of variability for categorical data. *Journal of Applied Statistics* 45 (8) pp. 1497-1516.
- [Agresti, 2012] Agresti, A. (2012). *Categorical Data Analysis*. Wiley-Interscience.
- [Alamuri *et. al*, 2014] Alamuri, M., Surampudi, B.R., Negi, A. (2014) A survey of distance/similarity measures for categorical data. In *Proceedings of the International Joint Conference on Neural Networks*, BeiJing, China, 611 July pp. 1907-1914.
- [Baak *et. al*, 2020] Baak, M., Koopman, R., Snoek, H., Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics & Data Analysis*, 152, 107043.
- [Behara and Chawla, 1974] Behara, M., Chawla, J. S. (1974). Generalized gamma-entropy. *Selecta statistica canadiana*, 2, 15-38.
- [Boriah *et. al*, 2008] Boriah, S., Chandola, V., Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the eighth SIAM International Conference on Data Mining*. 243-254.
- [Burbea and Rao, 1982] Burbea, J., Rao, C. R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of Multivariate Analysis*, 12(4), 575-596.
- [De Luca and Termini, 1972] De Luca, A., Termini, S. (1972). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and control*, 20(4), 301-312.
- [Fávero *et. al*, 2019] Fávero, L. P., Belfiore, P. (2019). *Data Science for Business and Decision Making*, Academic Press: Cambridge, MA, USA.

- [Eskin *et. al*, 2002] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In Applications of data mining in computer security (pp. 77-101). Springer, Boston, MA.
- [Gambaryan *et. al*, 1964] Gambaryan. P. (1964). A mathematical model of taxonomy. Izvest. Akad. Nauk Armen. SSR, 17(12):47-53.
- [Gini, 1912] Gini, C. (1912). Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. C. Cuppini, Bologna.
- [Goodman *et. al*, 1979] Goodman L.A., Kruskal W.H. (1979). Measures of association for cross classifications. In: Springer Series in Statistics, 234. Springer-Verlag, New York, NY.
- [Goodall, 1966] Goodall, D.W. (1966) A new similarity index based on probability. Biometrics, 22(4): 882-907.
- [Hill, 1973] Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. Ecology 54:427-432.
- [Kader and Perry, 2007] Kader, G.D., Perry, M. (2007). Variability for Categorical Variables. Journal of Statistics Education, 15(2).
- [Kendall, 1938] Kendall, M.G. (1938). A new measure of rank correlation. Biometrika 30, 81.
- [Khinchin, 1957] Khinchin, A. Y. (1957). Mathematical foundations of information theory. New York: Dover.
- [Kosko, 1986] Kosko, B. (1986). Fuzzy cognitive maps. International journal of man-machine studies, 24(1), 65-75.
- [Kosko, 1992] Kosko, B. (1992). Neural networks and fuzzy systems: A dynamical systems approach to machine intelligence (No. QA76. 76. E95 K86).
- [Kullback and Leibler, 1951] Kullback, S., Leibler, A.R. (1951), Annals of Mathematical Statistics 22, 49.
- [Mordeson and Nair, 2012] Mordeson, J. N., Nair, P. S. (2012). Fuzzy graphs and fuzzy hypergraphs (Vol. 46). Physica.
- [Pardo, 2018] Pardo, L. (2018). Statistical inference based on divergence measures. CRC press.
- [Pearson, 1916] Pearson, K. (1916) On the general theory of multiple contingency with special reference to partial contingency, Biometrika, vol. 11, no. 3, pp. 145-158.

- [Rao, 1982a] Rao, C. R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1), 24-43.
- [Rao, 1982b] Rao, C. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhy: The Indian Journal of Statistics, Series A*, 1-22.
- [Rényi, 1961] Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.
- [Salicru *et. al*, 1993] Salicru, M., Menendez, M. L., Morales, D., & Pardo, L. (1993). Asymptotic distribution of (h, ϕ) -entropies. *Communications in Statistics-Theory and Methods*, 22(7), 2015-2031.
- [Shannon, 1948] Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 379-423.
- [Stanfill and Waltz, 1986] Stanfill, C., Waltz, D. (1986). Toward Memory-based Reasoning. *Communications of the ACM*, 29 (12), 1213-1228.
- [Stevens, 1946] Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.
- [Sun and Yuan, 2006] Sun, W., Yuan, Y. X. (2006). *Optimization theory and methods: nonlinear programming (Vol. 1)*. Springer Science & Business Media.
- [Velleman and Wilkinson, 1993] Velleman, P.F., Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician*, 47(1), 65-72.
- [Xuecheng, 1992] Xuecheng, L. (1992). Entropy, distance measure and similarity measure of fuzzy sets and their relations. *Fuzzy sets and systems*, 52(3), 305-318.
- [Yager, 1979] Yager, R. R. (1979). On the measure of fuzziness and negation part I: membership in the unit interval.
- [Zadeh, 1965] Zadeh, L. (1965). Fuzzy sets. *Information and Control* 8, 338-353.
- [Zadeh, 1971] Zadeh, L. A. (1971). Similarity relations and fuzzy orderings. *Information sciences*, 3(2), 177-200.
- [Zhang, 2019] Zhang, Q. (2019). A Class of Association Measures for Categorical Variables Based on Weighted Minkowski Distance. *Entropy*, 21(10), 990.

[Ziegler, 2012] Ziegler, G. M. (2012). Lectures on polytopes (Vol. 152). Springer Science & Business Media.

[Zimmermann, 1996] Zimmermann, H.J. (1996) Fuzzy set theory and its applications (3rd ed). Kluwer Academic Publishers 69(92): 205-230.