Journal of Statistical Theory and Practice (2023) 17:47
https://doi.org/10.1007/s42519-023-00344-5

ORIGINAL ARTICLE

Association Between Nominal Categorical Variables: New Measure Formulation Based on Metric Distances and Value Validity

Tarald O. Kvålseth[1,2]

Accepted: 4 September 2023 / Published online: 27 September 2023

Abstract
When dealing with nominal categorical data, it is often desirable to know the degree of association or dependence between the categorical variables. While there is literally no limit to the number of alternative association measures that have been proposed over the years, they all yield greatly varying, contradictory, and unreliable results due to their lack of an important property: value validity. After discussing the value-validity property, this paper introduces a new measure of association (dependence) based on the mean Euclidean distance between probability distributions, one being a distribution under independence. Both the asymmetric form, when one variable can be considered as the explanatory (independent) variable and one as the response (dependent) variable, and the symmetric form of the measure are introduced. Particular emphasis is given to the important 2 × 2 case when each variable has two categories, but the general case of any number of categories is also covered. Besides having the value-validity property, the new measure has all the prerequisites of a good association measure. Comparisons are made with the well-known Goodman–Kruskal lambda and tau measures. Statistical inference procedure for the new measure is also derived and numerical examples are provided.

Keywords Association measures · Contingency tables · Nominal categorical data · Value validity

✉ Tarald O. Kvålseth
kvals001@umn.edu

[1] Department of Mechanical Engineering, University of Minnesota, 111 Church Street NE, Minneapolis, MN 55455, USA

[2] Department of Industrial and Systems Engineering, University of Minnesota, 207 Church Street NE, Minneapolis, MN 55455, USA

1 Introduction

As expressed by Upton and Cook [30], p. 19]:

> Two variables are associated if they are not independent, i.e.,
> if the value of one variable affects the value, or the distribution
> of the values, of the other.

Measures of association, or of the synonymous term "dependence", reflect the extent of the departure from independence. In the case of two nominal categorical variables X and Y, with the respective number of categories I and J and with joint probabilities p_{ij} and marginal probabilities $p_{i+} = \sum_{j=1}^{J} p_{ij}$ and $p_{+j} = \sum_{i=1}^{I} p_{ij}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, a wide variety of measures of association (dependence) have been proposed over the years. Table 1 provides a concise historical account of such measures.

For an *asymmetric* measure of association $A(Y|X)$, X is considered to be the explanatory or independent variable and Y the response or dependent one. An example of such a situation may involve different medical treatments (e.g., surgery versus radiation) as explanatory variable X that has a potential causal effect on the medical outcome (cancer being controlled versus not controlled) as response variable Y. Another example of Y depending on X would be the potential relationship between an electorate's party identification (X) and voting pattern (Y). However, in situations when one variable cannot reasonably be assumed to depend upon the other, symmetric measures of association $A(X, Y)$ have been introduced as in Table 1.

The large number of alternative association measures proposed to date has at least two implications: first, it implies that the measurement of association is an important subject matter; second, there is no clear consensus as to any generally preferred measure. There is also limited consistency between measures. Different measures may produce widely differing results for the same data sets. As stated by Reynolds [24], p. 55]:

> [M]easures of association … sometimes mislead as much as
> they inform. An index's numerical value should, of course,
> reflect the "true" relationship.

What is needed, as discussed in the present paper, is an additional requirement specifically related to the potential values taken on by an association measure.

The additional property requirement is the *value-validity property*. Introduced by Kvålseth [16] and based on 2 × 2 contingency tables with uniform (even) marginal probabilities, the value-validity requirement is generalized to one involving nonuniform (uneven) marginal probabilities. Since the various proposed association measures do not meet the condition imposed by the value-validity property, an alternative measure with the requisite properties is discussed in this paper. With the exception of the perhaps most popular Goodman–Kruskal measure $\lambda(Y|X)$ in Table 1, other existing measures from Table 1 are not able to distinguish between asymmetric and symmetric measures unless the categories are $I > 2$ or $J > 2$or both. The new measure proposed here has both an asymmetric and a symmetric form for the case of $I = J = 2$, which is important for many real situations, as well as for $I > 2$ and $J > 2$. The new measure,

Table 1 Historical account of asymmetric and symmetric association measures between nominal categorical variable X and Y

Measure formula	References
$P(X,Y)=\sqrt{\frac{\phi^2}{\phi^2+1}},\quad \phi^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-p_{i+}p_{+j})^2}{p_{i+}p_{+j}}=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i+}p_{+j}}-1$	Pearson [21]
$T(X,Y)=\sqrt{\frac{\phi^2}{\sqrt{(I-1)(J-1)}}}$	Tschuprow [29]
$S(X,Y)=\sqrt{\frac{m\phi^2}{(m-1)(1+\phi^2)}},\quad m=\min\{I,J\}$	Sakoda [25]
$V(X,Y)=\sqrt{\frac{\phi^2}{m-1}},\quad m=\min\{I,J\}$	Cramér [8]
$\lambda(Y\|X)=\frac{\sum_{i=1}^{I}p_{im}-p_{+m}}{1-p_{+m}},\quad p_{im}=\max_j\{p_{ij}\},\ p_{+m}=\max_j\{p_{+j}\}$	Goodman and Kruskal [11]
$\tau(Y\|X)=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}^2/p_{i+}-\sum_{j=1}^{J}p_{+j}^2}{1-\sum_{j=1}^{J}p_{+j}^2}$	Goodman and Kruskal [11]
$\lambda(X,Y)=\frac{\sum_{i=1}^{I}p_{im}+\sum_{j=1}^{J}p_{mj}-p_{+m}-p_{m+}}{2-p_{+m}-p_{m+}},\quad p_{mj}=\max_i\{p_{ij}\},\quad p_{m+}=\max_i\{p_{i+}\}$	Goodman and Kruskal [11]
$\tau(X,Y)=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}^2/p_{i+}+\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}^2/p_{+j}-\sum_{J=1}^{J}p_{+j}^2-\sum_{i=1}^{I}p_{i+}^2}{2-\sum_{j=1}^{J}p_{+j}^2-\sum_{i=1}^{I}p_{i+}^2}$	Goodman and Kruskal [12]
$I^*(Y\|X)=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\log(p_{ij}/p_{i+}p_{+j})}{-\sum_{j=1}^{J}p_{+j}\log p_{+j}}$	Attneave [2]
$I^*(X,Y)=\frac{2\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\log(p_{ij}/p_{i+}p_{+j})}{-\sum_{i=1}^{I}p_{i+}\log p_{i+}-\sum_{j=1}^{J}p_{+j}\log p_{+j}}$	Kvålseth [14]
$I^{**}(X,Y)=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\log(p_{ij}/p_{i+}p_{+j})}{\min\{\log I,\log J\}}$	Reshef [23]
$I_\alpha^*(X,Y)=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij}\log(p_{ij}/p_{i+}p_{+j})}{\left[\left(-\sum_{i=1}^{I}p_{i+}\log p_{i+}\right)^\alpha/2+\left(-\sum_{j=1}^{J}p_{+j}\log p_{+j}\right)^\alpha/2\right]^{1/\alpha}}$	Kvålseth [15]
$\lambda^{(2)}(Y\|X)=\frac{\sqrt{\sum_{i=1}^{I}p_{im}^2/p_{i+}}-p_{+m}}{1-p_{+m}}$	Kvålseth [16]
$\lambda^{(2)}(X,Y)=\frac{\sqrt{\sum_{i=1}^{I}p_{im}^2/p_{i+}}+\sqrt{\sum_{j=1}^{J}p_{mj}^2/p_{+j}}-p_{+m}-p_{m+}}{2-p_{+m}-p_{m+}}$	Kvålseth [16]
$S_\alpha(Y\|X)=\frac{\sum_{i=1}^{I}p_{i+}\sum_{j=1}^{J}(p_{ij}/p_{i+})^{\alpha+1}/\sum_{j=1}^{J}(p_{ij}/p_{i+})^{\alpha}-\sum_{j=1}^{J}p_{+j}^{\alpha+1}/\sum_{j=1}^{J}p_{+j}^{\alpha}}{1-\sum_{j=1}^{J}p_{+j}^{\alpha+1}/\sum_{j=1}^{J}p_{+j}^{\alpha}},\alpha\geq 1$	Särndal [26]

Table 1 (continued)

Measure formula	References
$V_\alpha(Y\|X) = \dfrac{\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^{\alpha+1}}{p_{i+}^{\alpha}p_{+j}^{\alpha}}-1}{\sum_{j=1}^{J}p_{+j}^{1-\alpha}-1},\quad \alpha>0$	Särndal [26], Tomizawa et al. [28]
$V_\alpha(X,Y) = \dfrac{2\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^{\alpha+1}}{p_{i+}^{\alpha}p_{+j}^{\alpha}}-1\right)}{\sum_{i=1}^{I}p_{i+}^{1-\alpha}+\sum_{j=1}^{J}p_{+j}^{1-\alpha}-2},\quad \alpha>0$	Tomizawa et al. [28]
$W(X,Y) = \sqrt{\dfrac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(p_{ij}-p_{i+}p_{+j}\right)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}p_{i+}p_{+j}\left(p_{i+}p_{+j}-2p_{ij}\right)+\min\left\{\sum_{i=1}^{I}p_{i+}^2,\sum_{j=1}^{J}p_{+j}^2\right\}}}$	Kvålseth [17]

which is proved to have some clear advantages over $\lambda(Y|X)$ and the symmetric $\lambda(X, Y)$, also turns out to be closely related to the Goodman–Kruskal $\tau(Y|X)$ and $\tau(X, Y)$ in Table 1. Finally, statistical inferences are discussed for the new association measure.

2 Value Validity

2.1 Why?

As with any summary measure or descriptive statistic, the purpose of a measure of association is to summarize by means of a single number the strength of the potential relationship between variables X and Y for some given data set. Consider, for instance, a cancer study involving medical treatment as the two-category explanatory variable X (surgery, radiation) and the two-category response variable Y (patient survives after 5 years, patient dies within first 5 years). For an association-measure value of, say $A(Y|X) = 0.11$, one would like to be able to interpret this result to mean that the outcome of the medical treatment depends only to a very limited extent on the type of treatment given. If another study using a different data set produced the value $A(Y|X) = 0.20$, one would like to be justified in making the order ("larger than") comparison that the second study showed greater dependence between treatment types and their outcome and "substantially" so. Similarly, from a third data set with $A(Y|X) = 0.15$, it would be informative to conclude that 0.20–0.11 > 0.20–0.15 provides a true difference comparison.

However, there is no adequate or rigorous basis for assuming that such interpretations and comparisons with existing association measures are valid or admissible in the sense that they actually do provide true and realistic representations of the attribute being measured, i.e., association (dependence) as opposed to simply being a representation of a measure itself. Such concern is evidenced by the fact that different measures can produce widely different results for the same data sets. It is not hard to find data sets for which two measures such a $\lambda(Y|X)$ and $\tau(Y|X)$ in Table 1 may produce entirely opposite results for even simple order comparisons. Such concern about the considerable variation in values between measures and the unreasonable values taken on by some has been expressed by various authors over the years (e.g., [6], pp. 302–303], [7], pp. 244–245], [9], p. 61], [24], pp. 55–57], [26]).

What is needed is some kind of condition or constraint on an association measure such that its numerical values are indeed "reasonable" with respect to some generally acceptable criterion. That is, a measure has to have the *value-validity property*. Then, if an association measure A has this and other desirable properties, there will be a sound basis for making different types of comparisons between association values and for interpreting the extent of association.

2.2 Condition

The term *validity*, an important concept in measurement theory and not the least for the behavioral and social sciences, generally means that a measure has validity if it measures what it is supposed to measure. This rather vague concept can be further refined by defining different types of validity: *content*, *construct*, *predictive*, *criterion*, and *concurrent* (e.g., [4], Ch. 11], [13], pp. 129–134]). These types of validity relate to the indirect measurement of an attribute using intermediary variables. However, none of these fit the present concern of determining if an association measure takes on “reasonable” numerical values throughout its range. Hence, the term *value validity* is chosen.

In order to address the value-validity property, let $A(P_{I\times J})$ denote the value of an association measure, either symmetric or asymmetric, for a joint distribution $P_{I\times J} = \{p_{ij}\}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ as represented by an $I \times J$ contingency table. Then for any $P_{I\times J}$ and $A(P_{I\times J}) \in [0, 1]$, there necessarily exists a distribution $P_{2\times 2}$ such that $A(P_{I\times J}) = A(P_{2\times 2})$. Thus, without any loss of generality, it is sufficient to use $A(P_{2\times 2})$ when considering any condition on value validity.

Specifically, consider the $P_{2\times 2}$ distributions or the 2×2 contingency Tables 2(a) and 2(b) with marginal probabilities r and $1 - r$. Table 2(a) represents perfect association with $A\left(P^{r1}_{2\times 2}\right) = 1$ for the distribution denoted by $P^{r1}_{2\times 2}$ while Table 2(b) represents $A\left(P^{r0}_{2\times 2}\right) = 0$ for the independence distribution denoted by $P^{r0}_{2\times 2}$. Then, one can define the distribution P^{rw}_{2x2} such that each component is the weighted mean of the corresponding components of P^{r0}_{2x2} and P^{r1}_{2x2} represented as:

$$P^{rw}_{2x2} = wP^{r1}_{2x2} + (1-w)P^{r0}_{2x2}, \quad w \in [0, 1] \tag{1}$$

and as the contingency table in Table 2(c). It is then postulated as a logical requirement that such a weighted mean should similarly be reflected by an association measure A as:

$$\begin{aligned} A\left(P^{rw}_{2x2}\right) &= wA\left(P^{r1}_{2x2}\right) + (1-w)A\left(P^{r0}_{2x2}\right), \quad w \in [0, 1] \\ &= w \text{ for } A\left(P^{r1}_{2x2}\right) = 1 \quad \text{and} \quad A\left(P^{r0}_{2x2}\right) = 0 \end{aligned} \tag{2}$$

Table 2 Binary tables with marginal probabilities r and $1-r$, with $r \in [0,1]$, and showing (a) perfect association, (b) zero association, and (c) weighted mean of (a) and (b)

X	Y			X	Y		
	1	2	Total		1	2	Total
(a) Perfect association P_{2x2}^{r1}				(b) Zero association P_{2x2}^{r0}			
1	r	0	r	1	r^2	$r(1-r)$	r
2	0	$1-r$	$1-r$	2	$r(1-r)$	$(1-r)^2$	$1-r$
Total	r	$1-r$		Total	r	$1-r$	

X	Y		
	1	2	Total
(c) Weighted mean P_{2x2}^{rw} of (a) and (b)			
1	$wr+(1-w)r^2$	$(1-w)r(1-r)$	r
2	$(1-\mathrm{w})\mathrm{r}(1-\mathrm{r})$	$w(1-r)+(1-w)(1-r)^2$	$1-r$
Total	r	$1-r$	

for all values of $r \in [0,1]$.

The expression in (2) may appropriately be called the *mean-value condition* (*criterion*) for the value-validity property of an association measure A. For the particular case of the simple arithmetic mean when $w=1/2$, (2) requires that $A\left(P_{2x2}^{r(1/2)}\right)=1/2$ as the arithmetic mean of $A\left(P_{2x2}^{r0}\right)=0$ and $A\left(P_{2x2}^{r1}\right)=1$ for any marginal distribution $(r, 1-r)$. It is hard to imagine hat anyone would disagree with the proposition that such a requirement is entirely appropriate. Nevertheless, most of the measures defined in Table 1 fail to meet this requirement.

The weighting factor w in (1)–(2) can also be viewed as an association parameter with $w=0$ in the case of zero association and $w=1$ in the case of perfect association. Such an interpretation of w can be justified from metric distances between the distributions in Table 2 considered as points in 4-dimensional Euclidean space. Thus, in terms of the Euclidean distance $d()$ (and for all other members of the Minkowski family of distance metrics), it is seen that

$$\frac{d\left(P_{2x2}^{rw}, P_{2x2}^{r0}\right)}{d\left(P_{2x2}^{r1}, P_{2x2}^{r0}\right)}=w \tag{3}$$

which shows that w is the normalized distance between P_{2x2}^{rw} and the zero-association distribution P_{2x2}^{r0} in Table 2(b). The condition in (2) requires that $\mathrm{A}(P_{2x2}^{rw})$ should equal the association parameter w, which seems entirely reasonable.

While Tables 2(a)–(c) all involve *marginal homogeneity* (i.e., $p_{1+}=p_{+1}$ and $p_{2+}=p_{+2}$), permutations of the rows or columns of these tables should not affect the values of an association measure. Also, note that the representations in Table

2(a)–(c) are the most general ones that permit zero-association and perfect-association distributions to be represented with the same marginal distributions.

2.3 Assessment of Current Measures

Of the various association measures defined in Table 1, only two measures can be seen to meet the mean-value condition in (2): T and V. However, both of these have other limitations related to their dependence on the number of categories I and J of X and Y, respectively, and to their lack of any intuitively meaningful interpretations [11], pp. 732–764; 17]. Neither of the two well-known measures λ and τ are seen to satisfy the condition in (2). For example, with $r = 0.80$ and $w = 0.50$ in Table 2(c), the values of λ and τ are, respectively, 0.20 and 0.25, substantially less than the 0.50 required by the condition in (2). Similarly, $I^*(Y|X) = 0.21$.

Among the various measures proposed to date, the Goodman–Kruskal's λ and τ seem to be among the most popular ones because of their meaningful interpretations. Both of these measures have a so-called proportional reduction in error (PRE) interpretation, i.e., they measure the relative decrease in the probability of incorrectly predicting the Y-category of a random observation when its X-category is given versus not given. Both measures use a different prediction strategy: λ(Y|X) is based on predicting only the modal (largest) Y-category whereas $\tau(Y|X)$ uses the more complex strategy of making the predictions so as to reproduce the conditional and marginal probability distributions. However, neither λ nor τ can be assumed to provide reliable or representative association values since they, as well as nearly all other association measures, fail to comply with the value-validity requirement in (2).

In spite of the large number of alternative association measures available, there is clearly a need for a new measure with the requisite properties. Such a proposal is developed next, starting with the important 2 × 2 case, i.e., when the number of categories $I = J = 2$.

3 Measure Derivation

3.1 2 × 2 Case

Since association (dependence) between nominal categorical variables X and Y implies departure from independence, it seems rather logical to consider a measure of association between X and Y as reflecting the distance between the joint distribution $\{p_{ij}\}$ and the corresponding independence distribution $\left\{p_{i+}p_{+j}\right\}$. In the case of $I = J = 2$ categories, $\{p_{ij}\}$ and $\left\{p_{i+}p_{+j}\right\}$ can be considered as points in 4-dimensional Euclidean space with the Euclidean distance between them being

$$d = \left(\sum\nolimits_{i=1}^{2}\sum\nolimits_{j=1}^{2}\left|p_{ij} - p_{i+}p_{+j}\right|^2\right)^{1/2} = 2|p_{11} - p_{1+}p_{+1}| \qquad (4)$$

where the last term follows from the fact that all of the terms $\left|p_{ij} - p_{i+}p_{+j}\right|$ in (4) are equal. The second expression in (4) can equivalently be expressed as

$$d = 2|p_{11}p_{22} - p_{12}p_{21}|. \tag{5}$$

Note that because of the 4 terms $\left|p_{ij} - p_{i+}p_{+j}\right|$ being equal, any member of Minkowski's class of metric distances would result in (5) except for the form of the multiplicative factor.

The metric distance function in (5) can also be expressed as

$$d = 2|p_{11}p_{2+} + p_{22}p_{1+} - p_{1+}p_{2+}| = 2p_{1+}p_{2+}\left|\frac{p_{11}}{p_{1+}} + \frac{p_{22}}{p_{2+}} - 1\right|. \tag{6}$$

Since the upper bound on the absolute-value term in (6) is clearly 1, $d \leq 2p_{1+}p_{2+}$ and hence

$$\delta(Y|X) = \frac{|p_{11}p_{22} - p_{12}p_{21}|}{p_{1+}p_{2+}} = \left|\frac{p_{11}}{p_{1+}} + \frac{p_{22}}{p_{2+}} - 1\right| \tag{7}$$

becomes the proposed measure of association when X is the explanatory variable and Y is the response. This measure has an intuitively appealing interpretation: it is the metric distance between $\{p_{ij}\}$ and $\left\{p_{i+}p_{+j}\right\}$ relative to the maximum distance when both categories of X contain at most one $p_{ij} \neq 0$.

Some important properties of $\delta(Y|X)$ are as follows:

(P1) $\delta(Y|X)$ is well defined unless all $p_{ij} > 0$ fall in one row.
(P2) $\delta(Y|X)$ has the value-validity property since it meets the condition in (2).
(P3) $\delta(Y|X)$ takes on values between 0 and 1, with $\delta(Y|X) = 0$ if and only if X and Y are independent and $\delta(Y|X) = 1$ if and only if each row of the contingency table contains at most one $p_{ij} \neq 0$.
(P4) $\delta(Y|X)$ is invariant under permutations of rows or columns.

Interestingly, although derived entirely independently, the expression in (7), except for the absolute value symbol, turns out to be the same as a formula proposed by Peirce [22] and Youden [31] for evaluating the performance of a prediction rule (Peirce) and a diagnostic test (Youden). However, the approach and context of their work differed from the present one. See also [3]. Evidently, $\delta(Y|X)$ in (7) can then be viewed as the distance of the Peirce-Youden statistic from zero.

3.2 Comparison with $\lambda(Y|X)$

An important correspondence between $\delta(Y|X)$ in (7) and $\lambda(Y|X)$ defined in Table 1 is that even though their respective numerical values can differ greatly for any given 2 × 2 table, the two measures provide the same order ("larger than") comparisons. That is, when reversing the roles of the variables X and Y in any given 2 × 2 table,

$$\delta(Y|X) \geq \delta(X|Y) \quad \text{implies} \quad \lambda(Y|X) \geq \lambda(X|Y) \tag{8}$$

where

$$\delta(X|Y) = \frac{|p_{11}p_{22} - p_{12}p_{21}|}{p_{+1}p_{+2}} = \left|\frac{p_{11}}{p_{+1}} + \frac{p_{22}}{p_{+2}} - 1\right| \tag{9}$$

and

$$\lambda(X|Y) = \frac{p_{m1} + p_{m2} - p_{m+}}{1 - p_{m+}}, \ p_{mj} = \max\{p_{1j}, p_{2j}\}, \ p_{m+} = \max\{p_{1+}, p_{2+}\}. \tag{10}$$

In order to prove this relationship property, note that for any 2 × 2 table with $\lambda(Y|X) \neq 0$, $p_{1m} + p_{2m} = p_{m1} + p_{m2} = p_m$ so that

$$\lambda(Y|X) = \frac{p_m - p_{+m}}{1 - p_{+m}}, \lambda(X|Y) = \frac{p_m - p_{m+}}{1 - p_{m+}}. \tag{11}$$

Since $(p_m - a)/(1 - a)$ is strictly decreasing in a, it follows from (11) that

$$\lambda(Y|X) \geq \lambda(X|Y) \quad \text{if} \quad p_{m+} \geq p_{+m}, \quad \text{i.e., if} \quad p_{m+}(1 - p_{m+}) \leq p_{+m}(1 - p_{+m}) \tag{12}$$

where the last inequality is the result of $p_{m+} \geq 1/2$ and $p_{+m} \geq 1/2$. The last inequality in (12) is obviously implied by $p_{1+}p_{2+} \leq p_{+1}p_{+2}$, which, together with (7), (9), and (12) leads to (8), completing the proof.

Any relationship between $\delta(Y|X)$ in (7) and $\lambda(Y|X)$ in Table 1 is fairly limited to the order inequalities in (8). Since $\lambda(Y|X)$ fails to meet the value-validity condition in (2) and is well-known to be extremely sensitive to the unevenness (skewness) of the marginal distributions (p_{1+}, p_{2+}) and (p_{+1}, p_{+2}), values of $\delta(Y|X)$ and $\lambda(Y|X)$ can differ greatly for the same 2 × 2 table. As an illustration, consider the following 2 × 2 table:

$$\begin{array}{ll} p_{11} = 0.50 & p_{12} = 0 \\ p_{21} = 0.25 & p_{22} = 0.25 \end{array}$$

for which $\delta(Y|X) = 0.50$ whereas $\lambda(Y|X) = 0$ (and $\delta(X|Y) = 0.67$, $\lambda(X|Y) = 0.50$). This example also illustrates the fact that $\lambda(Y|X) = 0$ without statistical independence (if p_{1m} and p_{2m} fall in the same column) whereas other association measures equal 0 if and only if X and Y are independent. For the $\lambda^{(2)}(Y|X)$ defined in Table 1, which is most comparable to $\lambda(Y|X)$, $\lambda^{(2)}(Y|X) = 0.16$ for this 2 × 2 table. However, this value is substantially less than $\delta(Y|X) = 0.50$ and is explainable by the fact that $\lambda^{(2)}(Y|X)$ does not meet the general condition in (2) except when $r = 1/2$.

The sensitivity of $\lambda(Y|X)$ to the unevenness of the marginal distributions can also be determined analytically be the use of Table 2(c). While $\delta(Y|X) = w$ for this table irrespective of the marginal distribution $(r, 1 - r)$ as required by the value-validity condition in (2), the expression for $\lambda(Y|X)$ becomes

$$\lambda(Y|X) = \begin{cases} 1 - 2(1-w)r & \text{for} \quad r \geq 1/2 \quad \text{and} \quad w \geq 1 - 1/2r \\ 0 & \text{for} \quad r \geq 1/2 \quad \text{and} \quad w < 1 - 1/2r \end{cases} \tag{13}$$

where the restriction $r \geq 1/2$ creates no lack of generality since the value of $\lambda(Y|X)$ for any $r = r'$ equals its value for $r = 1 - r'$. From (13) and $\Delta = w - [1 - 2(1-w)r] = (1-w)(2r-1)$, it is seen that $\lambda(Y|X)$ departs from the value-validity requirement in (2) at a rate that increases with r, i.e., as the unevenness of the marginal distribution $(r, 1-r)$ increases. It is only when $r = 1/2$ that $\lambda(Y|X)$ complies with the condition in (2) and when $\lambda(Y|X) = \delta(Y|X)$. Otherwise, $\lambda(Y|X)$ understates the true extent of the association and can generally be expected to have lower values than those of $\delta(Y|X)$.

3.3 Comparison with Odds Ratio

In order to provide a definition for the so-called odds ratio, consider that "treatment 1" and "treatment 2" are the categories of X and that "success" and "failure" are those of Y. Then, the odds (in favor) of success for treatment 1 becomes $(p_{11}/p_{1+})/(1 - p_{11}/p_{1+})$ and that of treatment 2 becomes $(p_{21}/p_{2+})/(1 - p_{21}/p_{2+})$ so that their ratio, the odds ratio OR, can be expressed as

$$OR = \frac{p_{11}p_{22}}{p_{12}p_{21}}. \tag{14}$$

The $OR = 1$ under independence of X and Y and increasing $|OR - 1|$ indicates increasing degree of association.

However, since OR is unbounded above so that any interpretation of OR values as reflecting the extent of association becomes difficult, alternative measures have been proposed as functions of OR: Yule's [32] $Q = (OR - 1)/(OR + 1)$ and Yule's [33] $Y = \left(\sqrt{OR} - 1\right)/\left(\sqrt{OR} + 1\right)$. Both of these measures, however, fail to meet the value-validity condition in (2). Note also that $|OR - 1|(p_{12}p_{21}/p_{1+}p_{2+}) = \delta(Y|X)$.

As an interesting example comparing OR and $\delta(Y|X)$ consider the data in Table 3 from a meta-analysis by Zheng et al. [34] comparing the outcomes of surgery versus

Table 3 Meta-analysis results of lung-cancer treatment of $N = 11{,}921$ patients after 1, 3, and 5 years post treatment

Treatment	Results after 1 year		Results after 3 years		Results after 5 years	
	Survive	Die	Survive	Die	Survive	Die
Surgery	6569	502	5607	1464	4872	2199
Radiation	4045	805	2745	2105	1998	2852
Total	10,614	1307	8352	3569	6870	5051

Source: Zheng et al. [34]

radiation for treating lung cancer at 1, 3, and 5 years after treatment. The surgery data are based on the mean survival rates for surgeries involving both lobectomy and limited lung resections. From the frequency data $\{n_{ij}\}$ in Table 3 with sample size $N = \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij} = 11{,}921$ and when replacing the p_{ij}'s in (7) and (14) with frequencies to produce the following expressions:

$$\delta(Y|X) = \left| \frac{n_{11}}{n_{1+}} + \frac{n_{22}}{n_{2+}} - 1 \right|, \quad OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{15}$$

it is found that $\delta(Y|X) = 0.10$, 0.23, and 0.28 and $OR = 2.60$, 2.94, and 3.16 after 1, 3, and 5 years post treatment.

These results would seem to indicate that the odds of successful treatment by means of surgery is about three times that of radiation. Such results may seem to favor surgery to a potentially misleading extent, especially when compared with the $\delta(Y|X)$ values that show a rather limited degree of association between the two types of cancer treatment (X) and the outcome (Y). The consecutive values of $\delta(Y|X)$ indicate that the strength of the causal relationship (dependence) between X and Y, although rather low, increases with time since the treatment, with the absolute increase between years 1 and 3 after treatment being more than twice that between years 3 and 5. Such interpretation and comparisons are considered permissible due to the value-validity property of $\delta(Y|X)$. Note also that, as pointed out by the authors of this paper [34], the advantage of surgery becomes further reduced when adjusting for other factors such as patients' age and operability.

3.4 $I \times J$ Case

Just as $\delta(Y|X)$ in (7) for the 2 × 2 case is based on a metric-distance formulation, a similar approach can be used for the case when either I or J or both are greater than 2. Thus, consider the conditional probability distribution $\left\{p_{ij}/p_{i+}\right\} = (p_{i1}/p_{i+}, \ldots, p_{iJ}/p_{i+})$ for $i = 1, \ldots, I$ and the corresponding independence distribution $\left\{p_{+j}\right\} = (p_{+1}, \ldots, p_{+J})$ as points or vectors in J-dimensional Euclidean space. The Euclidean distance between the two points $\left\{p_{ij}/p_{i+}\right\}$ and $\{p_{+j}\}$ is given by

$$d_i = d\left(\left\{p_{ij}/p_{i+}\right\}, \left\{p_{+j}\right\}\right) = \sqrt{\sum\nolimits_{j=1}^{J} \left(p_{ij}/p_{i+} - p_{+j}\right)^2}, \quad i = 1, \ldots, I. \tag{16}$$

Using the same order as in (16), the second-order weighted arithmetic mean of the distances in (16) becomes

$$\overline{d} = \sqrt{\sum\nolimits_{i=1}^{I} p_{i+} d_i^2} = \sqrt{\sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} p_{ij}^2 \Big/ p_{i+} - \sum\nolimits_{j=1}^{J} p_{+j}^2}. \tag{17}$$

From the inequality

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}^2 \Big/ p_{i+} \leq \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}p_{i+} \Big/ p_{i+} = 1$$

the $\overline{d}$ in (17) can be normalized into the following measure:

$$\delta(Y|X) = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}^2/p_{i+} - \sum_{j=1}^{J} p_{+j}^2}{1-\sum_{j=1}^{J} p_{+j}^2}} \in [0, 1] \tag{18}$$

which is proposed as a distance-based measure of association when X is the explanatory variable and Y is the response. Also, by interchanging the roles of X and Y,

$$\delta(X|Y) = \sqrt{\frac{\sum_{j=1}^{J}\sum_{i=1}^{I} p_{ij}^2/p_{+j} - \sum_{i=1}^{I} p_{i+}^2}{1-\sum_{i=1}^{I} p_{i+}^2}} \in [0, 1]. \tag{19}$$

The Euclidean distance is used in (16) since it is the standard distance metric used in various scientific fields. Also, the second-order mean used in (17) is a commonly used one. Although one could potentially consider other members of the power mean, only the second-order member is seen to comply with the value-validity condition in (2).

The $\delta(Y|X)$ in (18) has an intuitively appealing interpretation: it measures how far the probability distribution between Y and X, conditional on X, is on the average from the independence distribution relative to its maximum. It also has all the same properties (P1)-(P4) as those of $\delta(Y|X)$ outlined above for the 2 × 2 case.

Interestingly, the $\delta(Y|X)$ in (18) has the same expression as the square root of the Goodman–Kruskal $\tau(Y|X)$ in Table 1. However, $\tau(Y|X)$ was derived from a very different approach based on PRE prediction as well as on an analysis of variance approach [20].

3.5 Symmetric Case

When one variable cannot reasonably be assumed to depend on the other, a symmetric association measure is needed. Such a measure $A(X, Y)$ may be formulated as some mean of the corresponding asymmetric forms $A(Y|X)$ and $A(X|Y)$ as the roles of X and Y are interchanged. Such means can also be considered as weighted means as, for example, in the case of $\lambda(X, Y)$, $\tau(X, Y)$, $I^*(X, Y)$, and $I_\alpha^*(X, Y)$ in Table 1.

In the case of 2 × 2 contingency tables when $I = 2$ and $J = 2$, a most simple symmetric equivalent of the new asymmetric measures $\delta(Y|X)$ in (7) and $\delta(X|Y)$ in (9) would be their arithmetic mean $[\delta(Y|X) + \delta(X|Y)]/2$. Another alternative would be the geometric mean giving the following symmetric measure:

$$\delta(X, Y) = \frac{|p_{11}p_{22} - p_{12}p_{21}|}{\sqrt{p_{1+}p_{2+}p_{+1}p_{+2}}}. \tag{20}$$

It is interesting to note that this expression is the same as the absolute value $|\rho|$ of the Pearson product-moment correlation coefficient ρ obtained when denoting the two categories of X and Y as 0 and 1 (e.g., [5], pp. 380–382], [18], pp. 43–44], [27], pp. 54–55]). This $\delta(X, Y)$ is also equivalent to Pearson's ϕ in Table 1.

For the case when $I \geq 2$ and $J \geq 2$, a symmetric form $\delta(X, Y)$ can be based on the mean of $\delta(Y|X)$ in (18) and $\delta(X|Y)$ in (19). Their simple arithmetic mean or their geometric mean would be obvious potential choices. Alternatively, since $\delta(Y|X)$ and $\delta(X|Y)$ turn out to equal $\sqrt{\tau(Y|X)}$ and $\sqrt{\tau(X|Y)}$ for the Goodman–Kruskal tau in Table 1, one could consider the second-order weighted mean of $\delta(Y|X)$ and $\delta(X|Y)$ using weights based on their respective denominators (i.e., weights $d_1/(d_1 + d_2)$ and $d_2/(d_1 + d_2)$ from their denominators d_1 and d_2). The resulting weighted mean becomes

$$\delta(X, Y) = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}^2 \Big/ p_{i+} + \sum_{j=1}^{J}\sum_{i=1}^{I} p_{ij}^2 \Big/ p_{+j} - \sum_{j=1}^{J} p_{+j}^2 - \sum_{i=1}^{I} p_{i+}^2}{2 - \sum_{j=1}^{J} p_{+j}^2 - \sum_{i=1}^{I} p_{i+}^2}} \tag{21}$$

which is also seen to equal the square root of $\tau(X, Y)$ in Table 1.

In the case of $I = J = 2$, when Goodman–Kruskal's $\tau(Y|X) = \tau(X|Y)$ for any 2 × 2 table whereas $\delta(Y|X)$ and $\delta(X|Y)$ may differ considerably, it can be verified that the symmetric $\delta(Y, X)$ in (21) is, in fact, equivalent to the $\delta(X, Y)$ in (20). Thus, $\delta(Y|X)$ in (7) or (15) is an appropriate asymmetric measure when $I = J = 2$ and the symmetric $\delta(X, Y)$ in (21) can be used when $I \geq 2$ and/or $J \geq 2$.

The $\delta(X, Y)$ in (21) represents the mean normalized Euclidean distance between the joint probability distribution of X and Y and the corresponding independence distribution. Also, since $\delta^2(X, Y) = \tau(X, Y)$, $\delta^2(X, Y)$ has the same PRE interpretation as $\tau(X, Y)$. Other properties of $\delta(X, Y)$ are the same as those of $\delta(Y|X)$ in (7) and (18), with obvious modifications. Also, being a mean of $\delta(Y|X)$ and $\delta(X|Y)$, values of $\delta(X, Y)$ will always fall between those of $\delta(Y|X)$ and $\delta(X|Y)$ inclusive.

4 Statistical Inferences About δ

4.1 $\delta(Y|X)$ in (7)

Consider now that $\delta(Y|X)$ is based on multinomial sample probabilities $p_{ij} = n_{ij}/N$ with sample size $N = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}$ and let Δ(Y|X) denote the corresponding population measure based on the population probabilities π_{ij} for $i = 1, 2$ and $j = 1, 2$. One may then want to perform statistical inferences about $\Delta(Y|X)$, especially the construction of confidence intervals, but also perhaps testing of hypotheses about $\Delta(Y|X)$. Besides the use of resampling methods (jackknife, bootstrap), such inferences can be done by using the *delta method* [1], Ch.16] as briefly outlined next.

Accordingly, the following convergence-in-distribution holds:

$$\sqrt{N}[\delta(Y|X) - \Delta(Y|X)] \rightarrow^{d} Normal\left(0, \sigma_{\delta}^{2}\right) \tag{22}$$

so that for a large multinominal sample of size N, the estimator $\delta(Y|X)$ is approximately normal with mean $\Delta(Y|X)$ and variance σ_{δ}^{2}/N, or standard error $\sigma_{\delta}/\sqrt{N}$. By taking the partial derivatives of $\Delta(Y|X)$ with respect to each π_{ij} and then substituting those with the corresponding sample estimates p_{ij} for all i and j, the estimated variance in (22) becomes

$$\hat{\sigma}_{\delta}^{2} = \sum\nolimits_{i=1}^{2} \sum\nolimits_{j=1}^{2} p_{ij}\phi_{ij}^{2} - \left(\sum\nolimits_{i=1}^{2} \sum\nolimits_{j=1}^{2} p_{ij}\phi_{ij}\right)^{2} \tag{23}$$

where $\phi_{ij} = \partial\delta(Y|X)/\partial p_{ij}$ for $i = 1, 2$ and $j = 1, 2$. From (7),

$$\phi_{11} = p_{1+}^{-2}p_{12}, \quad \phi_{12} = -p_{1+}^{-2}p_{11}, \quad \phi_{21} = -p_{2+}^{-2}p_{22}, \quad \phi_{22} = p_{2+}^{-2}p_{21}. \tag{24}$$

It is found that $\sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij}\phi_{ij} = 0$ as also follows from a theorem by Fleiss [10]. Therefore, from (23)–(24),

$$\hat{\sigma}_{\delta}^{2} = \frac{p_{11}p_{12}}{p_{1+}^{3}} + \frac{p_{21}p_{22}}{p_{2+}^{3}} = N\left(\frac{n_{11}n_{12}}{n_{1+}^{3}} + \frac{n_{21}n_{22}}{n_{2+}^{3}}\right). \tag{25}$$

Instead of performing statistical inferences directly on $\Delta(Y|X)$, it is preferable to use the following logarithmic transformation and it inverse:

$$L = \log\left(\frac{\delta(Y|X)}{1 - \delta(Y|X)}\right), \delta(Y|X) = \frac{\exp(L)}{1 + \exp(L)} \tag{26}$$

since this transformation (a) provides a more rapid convergence to normality, especially for large or small $\delta(Y|X)$, and (b) ensures that a confidence interval will always fall inside the [0, 1]-interval (e.g., [1], pp. 70, 618], [19], p. 106]). The estimated variance of L becomes

$$\hat{\sigma}_{L}^{2} = [dL/d\delta(Y|X)]^{2}\hat{\sigma}_{\delta}^{2} = [\delta(Y|X)(1 - \delta(Y|X))]^{-2}\hat{\sigma}_{\delta}^{2} \tag{27}$$

with the following confidence interval (CI):

$$100(1- \propto)\%CI \quad \text{for} \quad L_{\Delta}{:}L \pm z_{\alpha/2}\hat{\sigma}_{L}/\sqrt{N} \tag{28}$$

where $z_{\alpha/2}$ is the standard normal quantile (e.g., $z_{\alpha/2} = 1.96$ for $\alpha = 0.05$ and for 95% confidence). The corresponding confidence interval for $\Delta(Y|X)$ is then obtained by using the inverse transformation in (26) to each side of the interval in (28).

As a numerical example, consider again the data in Table 3 for the results after 5 years. From (7) or (15), $\delta(Y|X) = 0.2771$ and, from (25), $\hat{\sigma}_{\delta}^{2} = 0.9567$ so that from (27), $\hat{\sigma}_L^2 = 23.8422$. Therefore, with L = − 0.9589 from (26), a 95% *CI* from (28) becomes $-0.9589 \pm 1.96\sqrt{23.8422/11,921}$, i.e., [− 1.0466, − 0.8712], which, from the inverse transformation in (26), gives the following 95% *CI* for $\Delta(Y|X)$: [0.26, 0.30].

4.2 $\delta(Y|X)$ in (18)

For the case when $I > 2$ and $J > 2$, it is found from the expression in (18) that

$$\phi_{ij} = \frac{\partial \delta(Y|X)}{\partial p_{ij}} = \frac{1}{\delta(Y|X)\left(1 - \sum_{j=1}^{J} p_{+j}^2\right)}\left[p_{ij}/p_{i+} - (1/2)\sum_{k=1}^{J}\left(p_{ik}/p_{i+}\right)^2 - \left(1 - \delta^2(Y|X)\right)p_{+j}\right] \tag{29}$$

or, in terms of frequencies,

$$\phi_{ij} = \frac{N^2}{\delta(Y|X)\left(N^2 - \sum_{j=1}^{J} n_{+j}^2\right)}\left[n_{ij}/n_{i+} - (1/2)\sum_{k=1}^{J}(n_{ik}/n_{i+})^2 - \left(1 - \delta^2(Y|X)\right)n_{+j}/N\right]. \tag{30}$$

The inference procedure is then equivalent to the one outlined above for the case $I = J = 2$ (except for the summations now being from $i = 1$ to $i = I$ and from $j = 1$ to $j = J$). Thus, the computed values for ϕ_{ij} are used in (23) to obtain the variance $\hat{\sigma}_\delta^2$ for $\delta(Y|X)$. Then, the logarithmic transformation in (26) leads to $\hat{\sigma}_L^2$ in (27) and *CI* in (28) from which the *CI* for the population measure $\Delta(Y|X)$ is derived from the exponentiation in (26).

4.3 $\delta(X, Y)$ in (21)

The same inference procedure as that of $\delta(Y|X)$ can be used for the symmetric measure $\delta(X, Y)$ in (21) and its equivalent population measure $\Delta(X, Y)$. For the estimated variance $\hat{\sigma}_\delta^2$ for $\delta(X, Y)$ in (21), it is found that

$$\begin{aligned}\phi_{ij} = \frac{\partial \delta(X,Y)}{\partial p_{ij}} = \frac{1}{\delta(X,Y)\left(2 - \sum_{i=1}^{I} p_{i+}^2 - \sum_{j=1}^{J} p_{+j}^2\right)}&\left[p_{ij}\left(\frac{1}{p_{i+}} + \frac{1}{p_{+j}}\right)\right.\\ &\left. - (1/2)\sum_{k=1}^{J}(p_{ik}/p_{i+})^2 - (1/2)\sum_{h=1}^{I}\left(p_{hj}/p_{+j}\right)^2 - \left(1 - \delta^2(X,Y)\right)\left(p_{i+} + p_{+j}\right)\right]\end{aligned} \tag{31}$$

or, in terms of frequencies,

$$\phi_{ij} = \frac{N^2}{\delta(X, Y)\left(2N^2 - \sum_{i=1}^{I} n_{i+}^2 - \sum_{j=1}^{J} n_{+j}^2\right)} \left[n_{ij}\left(\frac{1}{n_{i+}} + \frac{1}{n_{+j}}\right) - (1/2)\sum_{k=1}^{J} (n_{ik}/n_{i+})^2 \right.$$
$$\left. -(1/2)\sum_{h=1}^{I} \left(n_{hj}/n_{+j}\right)^2 - \left(1 - \delta^2(X, Y)\right)\left(n_{i+} + n_{+j}\right)/N \right]. \quad (32)$$

As a numerical example, consider the data in Table 3 (results after 5 years). It is then found from (2) that $\delta(X, Y) = 0.2754$ and, from (32), $\phi_{11} = -0.9798$, $\phi_{12} = -2.7419$, $\phi_{21} = -2.8316$, and $\phi_{22} = -0.4741$, which substituted into (23) gives $\hat{\sigma}_{\delta}^2 = 0.9441$. From (27), $\hat{\sigma}_{L}^2 = 23.7055$ and hence, from (28), a 95% *CI* for L_{Δ} becomes [− 1.1084, − 0.9336] and, from the inverse transformation in (26), a 95% *CI* for $\Delta(X, Y)$ becomes [0.25, 0.28].

5 Concluding Comments

When seeking a way to measure the degree of association or dependence between two nominal categorical variables X and Y, one is faced with an effectively infinite number of choices of alternative measures. Furthermore, different measures may yield widely different results for the same data sets so that the results may depend as much on the measure chosen as on the true relationship between the variables. One way to deal with such overabundance of measures producing potentially greatly varying and misleading results is to introduce an additional requirement on such measures as done in this paper.

The value-validity property with potentially nonuniform marginal distributions is introduced here as such an additional requirement to ensure that all values of an association measure provide true and realistic representations of this characteristic between categorical variables. Since only two of the existing measures meet the value-validity condition, but both have other limitations, a new measure, δ, has been introduced in this paper: the asymmetric $\delta(Y|X)$ in (7) and (18) when the number of categories $I = J = 2$ and $I, J > 2$, respectively; the symmetric $\delta(X, Y)$ in (20) and (21) when $I = J = 2$ and $I, J > 2$, respectively. The new measure would seem to have all of the properties required of an appropriate association measure.

When comparing δ with the apparently most popular measure, Goodman–Kruskal's λ, δ has some clear advantages. First, δ satisfies the value-validity condition in (2) while λ does not. Second, δ equals 0 only under independence as is the case with all association measures with the exception of λ for which the asymmetric $\lambda(Y|X) = 0$ when the largest row probabilities fall in a single column of the contingency table. Third, as explained above in some detail, λ is very sensitive to the unevenness (skewness) of the marginal distributions, resulting in values that may potentially appear to be unreasonably low. The δ does not suffer from any of those limitations. There is every reason to suggest that δ is to be generally preferred over λ.

Because of its unique value-validity property as well as other necessary properties, the asymmetric $\delta(Y|X)$ and the symmetric $\delta(X, Y)$ are designed to provide true and valid representations of the association attribute. Thus, values of this measure can be used to make reliable and meaningful interpretations, comparisons, and conclusions about the extent or strength of association and about various comparisons. In order for the interpretation of the extent or strength of an association to be consistent, the following guideline is offered as verbal descriptions:

$$\begin{aligned} 0 \le A &\le 0.20\text{: Very low} \\ 0.20 < A &\le 0.40\text{: Low} \\ 0.40 < A &\le 0.60\text{: Moderate} \\ 0.60 < A &\le 0.80\text{: High} \\ 0.80 < A &\le 1.00\text{: Very high.} \end{aligned}$$

As with any measure that summarizes results into a single number, the new δ may, of course, be supplemented by more extensive analyses. However, there may be any number of real situations in which summary information as provided by δ can be interesting and useful. The results from the meta-analysis of lung cancer discussed above is one such example, showing a clear, although rather low, degree of association (dependence) between treatment (surgery, radiation) and medical outcome (survival, death). This causal relationship tended to favor surgery over radiation and increasingly so with time. Specifically, the greatest relative increase in $\delta(Y|X)$ at 130% occurred from year 1 to year 3 after treatment as compared to about 22% from year 3 to 5. Because of the properties of δ, including the value-validity property, such relative comparisons can be expected to provide true representations of the association characteristic and not simply such changes in a measure itself.

Such medical information would certainly be of considerable help when making important decisions by both medical providers and patients. As other real and diverse examples of the potential utility of an association measure such as δ, consider the data relating different types of office work involving VDU screens (explanatory variable X) and eyestrain (response variable Y) reported by Lloyd [19], pp. 131–132]. Those data show that $\delta(Y|X) = 0.85$, which can be interpreted as a "very high" degree of association and as a potentially serious problem. By comparison, the $\lambda(Y|X) = 0$ as an entirely unreasonable and misleading result.

As an example of particular interest to political scientists, Reynolds [24], pp. 1–2] reported data on party identification X (democrat, republican. independent) and vote Y (democrat, republican). When the value of $\delta(Y|X)$ is computed from those data, a "high" degree of association is established as $\delta(Y|X) = 0.62$. As a comparison, the $\tau(Y|X)$ defined in Table 1, which lacks the value-validity property, shows a misleadingly low value of $\tau(Y|X) = 0.38$. In an example given by Tang et al. [27], p. 61] relating gender (X) to different levels of depression (Y), the result $\delta(Y|X) = 0.15$ indicates a very low association, but females tended to have more severe depression than males. However, $\lambda(Y|X)$ and $\tau(Y|X)$ provide the even lower values of 0 and 0.02, respectively.

In conclusion, the new association measure δ, whether in asymmetric or symmetric form, would seem to meet all of the requirements expected of a good association measure. It has an intuitively appealing interpretation in terms of normalized Euclidean distances between joint probability distributions. In addition to sharing the appropriate properties with other measures, it has the important value-validity property. Consequently, there would appear to be no particular reason why δ should not become the measure of choice for the association between nominal categorical variables.

Funding No funding was received for this work and no financial interest is declared.

Declarations

Conflict of interest No conflict of interest is declared.

References

1. Agresti A (2013) Categorical data analysis, 3rd edn. Wiley, Hoboken
2. Attneave F (1959) Applications of information theory to psychology. Holt, Rinehart, and Winston, New York
3. Baker SG, Kramer BS (2007) Peirce, Youden, and receiver operating characteristic curves. Am Stat 61(4):343–346
4. Bandalos DL (2018) Measurement theory and applications for the social sciences. The Guildford Press, New York
5. Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge
6. Blalock HM Jr (1972) Social statistics. McGraw-Hill, New York
7. Bors D (2018) Data analysis for the social sciences. Sage, Thousand Oaks
8. Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton, N.J.
9. Everitt BS (1977) The analysis of contingency tables. Chapman and Hall, London
10. Fleiss JL (1982) A simplification of the classic large-sample standard error of a function of multinomial proportions. Am Stat 36(4):377–378
11. Goodman LA, Kruskal WH (1954) Measures of association for cross classifications. J Am Stat Assoc 49:732–764
12. Goodman LA, Kruskal WH (1959) Measures of association for cross classification II: further discussion and references. J Am Stat Assoc 54:123–163
13. Hand DJ (2004) Measurement theory and practice. Wiley, Chichester
14. Kvålseth TO (1987) Entropy and correlation: some comments. IEEE Trans Syst Man Cybern SMC-17:517–519
15. Kvålseth TO (2017) On normalized mutual information: measure derivations and properties. Entropy 19:1–15
16. Kvålseth TO (2018) Measuring association between nominal categorical variables: an alternative to the Goodman–Kruskal lambda. J Appl Stat 45(6):1118–1132
17. Kvålseth TO (2018) An alternative to Cramér's coefficient of association. Comm Stat Theory Methods 47(23):5662–5674. https://doi.org/10.1080/03610926.2017.1400056
18. Liebetrau AM (1983) Measures of association. Sage, Beverly Hills
19. Lloyd CJ (1999) Statistical analysis of categorical data. Wiley, New York
20. Margolin BH, Light RJ (1974) An analysis of variance for categorical data II: small sample comparisons with Chi square and other competitors. J Am Stat Assoc 69:755–764

21. Pearson K (1904) On the theory of contingency and its relation to association and normal correlation. Drapers Company Research Memoirs, Biometric Series, No. 1, London
22. Peirce CS (1884) The numerical measure of the success of predictions. Science 4:453–454
23. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC (2011) Detecting novel associations in large data sets. Science 334:1518–1524
24. Reynolds HT (1977) The analysis of cross-classification. The Free Press, New York
25. Sakoda JM (1977) Measures of association for multivariate contingency tables. Soc Stat Sect Proc Am Stat Assoc 66:777–780
26. Särndal CE (1974) A comparative study of association measures. Psychometrika 39:165–187
27. Tang W, He H, Tu XM (2012) Applied categorical and count data analysis. CRC Press, Boca Raton
28. Tomizawa S, Miyamoto N, Houya H (2004) Generalization of Cramer's coefficient of association for contingency tables. S Afr Stat J 38(1):1–24
29. Tschuprow AA (1939) Principles of the mathematical theory of correlation, translated by M. Kantorowitsch. W. Hodge & Co., London
30. Upton G, Cook I (2014) Oxford dictionary of statistics. Oxford University Press, Oxford
31. Youden WJ (1950) Index for rating diagnostic tests. Cancer 3:32–35
32. Yule GU (1900) VII. On the association of attributes in statistics: with illustrations from the material of the childhood society&c. Philos Trans R Stat Soc 194(251–269):257–319
33. Yule GU (1912) On the methods of measuring association between two attributes. J R Stat Soc 75:579–652
34. Zheng X, Schipper M, Kidwell K, Lin J, Reddy R, Ren Y, Chang A, Lv F, Orringer M, Kong F-MS (2014) Survival outcome after stereotactic body radiation therapy and surgery for stage I non-small lung cancer: a meta-analysis. Int J Radiat Oncol Biol Phys 90:603–611