NOTE **Communicated by James V. Stone**

On Suspicious Coincidences and Pointwise Mutual Information

Christopher K. I. Williams
ckiw@inf.ed.ac.uk
School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, U.K.

Barlow (1985) hypothesized that the co-occurrence of two events A and B is "suspicious" if $P(A, B) \gg P(A)P(B)$. We first review classical measures of association for 2×2 contingency tables, including Yule's Y (Yule, 1912), which depends only on the odds ratio λ and is independent of the marginal probabilities of the table. We then discuss the mutual information (MI) and pointwise mutual information (PMI), which depend on the ratio $P(A, B)/P(A)P(B)$, as measures of association. We show that once the effect of the marginals is removed, MI and PMI behave similarly to Y as functions of λ. The pointwise mutual information is used extensively in some research communities for flagging suspicious coincidences. We discuss the pros and cons of using it in this way, bearing in mind the sensitivity of the PMI to the marginals, with increased scores for sparser events.

1 Introduction

Barlow (1985) hypothesized that "the cortex behaves like a gifted detective, noting suspicious coincidences in its afferent input, and thereby gaining knowledge of the non-random, causally related, features in its environment." More specifically, he wrote (p. 40):

> The coincident occurrence of two events A and B is "suspicious" if they occur jointly more than would be expected from the probabilities of their individual occurrence, i.e. the coincidence $A\&B$ is suspicious if $P(A\&B) \gg P(A) \times P(B)$.[1] Any detective knows that, for a coincidence to be suspicious, the events themselves must be rare ones, and that if they are rare enough, even a single occurrence is significant.

Edelman, Hiles, Yang, and Intrator (2002) refer to the *principle of suspicious coincidences* as where "two candidate fragments A and B should be combined into a composite object if the probability of their joint appearance $P(A, B)$ is much higher than $P(A)P(B)$."

[1] In fact in Barlow (1985), the inequality is written $\ll$ rather than $\gg$, but it is clear the latter was intended. The same paper was also published as Barlow (1987); there, the inequality is the correct way round.

Neural Computation **34, 2037–2046** (2022)
https://doi.org/10.1162/neco_a_01533

The fundamental problem here is to detect if there is a significant association between events A and B. This can arise in many different contexts—for example:

- An animal detecting that eating a certain plant is associated with subsequent illness
- Detecting that a certain drug is associated with a particular adverse drug reaction
- Detecting the association between a visual stimulus that contains an image of the subject's grandmother or not and the response of a putative "grandmother cell"
- Detecting that particular successive words in text are associated more frequently than by chance—called a *collocation*, an example being the bigram "carbon dioxide"
- A geneticist determining that two genes are in linkage disequilibrium (Lewontin, 1964)
- Detecting that the pattern of two edges in a visual scene making a corner junction occurs more frequently than by chance

Below we review various measures of association from the literature, notably Yule's Y (Yule, 1912), which depends solely on the odds ratio and is invariant to the marginal distributions of the two variables. We then discuss measures of association based on the mutual information and pointwise mutual information, which make use of the ratio $P(A, B)/P(A)P(B)$, as proposed by Barlow and others across diverse literatures. Finally, we consider the pros and cons of using pointwise mutual information (PMI) to flag suspicious coincidences and discuss its estimation from data when (some of) the counts in the table are low.

2 2×2 Contingency Tables

Consider two random variables X and Y that take on values of 0 or 1. The 2×2 contingency table has the form

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}, \tag{2.1}$$

where, for example, $p_{01} = p(X = 0, Y = 1)$. We will also say the event x occurs if $X = 1$, and similarly for y. We denote the marginals with "dot" notation, so that, for example, $p_{1\cdot} = p(X = 1) = p_{10} + p_{11}$.

P is defined by 3 degrees of freedom (as the entries sum to 1). Two of these are taken up by the marginals, leaving 1 additional degree of freedom. Given a table P, we can manipulate the marginals by multiplying the rows and columns with positive numbers and renormalizing. Such a transformation is shown in, for example, Hasenclever and Scholz (2016, eq. 1):

$$g_{\mu,\nu}(P) = \frac{1}{Z(\mu,\nu)} \begin{pmatrix} \mu\nu p_{00} & \mu p_{01} \\ \nu p_{10} & p_{11} \end{pmatrix}, \tag{2.2}$$

where $Z(\mu,\nu) = \mu\nu p_{00} + \mu p_{01} + \nu p_{10} + p_{11}$. The odds ratio

$$\lambda = \frac{p_{00}p_{11}}{p_{01}p_{10}} \tag{2.3}$$

can be seen to be invariant to the action of this margin manipulation transformation and thus defines the third degree of freedom. An odds ratio of 1 implies that there is no association and that P is equal to the product of the marginals.

The "canonical" table with marginals of $1/2$ but with the same odds ratio as P is given by

$$P_{can} = \begin{pmatrix} \frac{\sqrt{\lambda}}{2(1+\sqrt{\lambda})} & \frac{1}{2(1+\sqrt{\lambda})} \\ \frac{1}{2(1+\sqrt{\lambda})} & \frac{\sqrt{\lambda}}{2(1+\sqrt{\lambda})} \end{pmatrix}, \tag{2.4}$$

as shown by Yule (1912). Like a copula for continuous variables, this allows a separation of the marginals from the dependence structure between X and Y.

The table P can also be expressed in terms of a deviation from the product of the marginals (see Hasenclever & Scholz, 2016, p. 24) as

$$P = \begin{pmatrix} p_{0\cdot}p_{\cdot 0} + D & p_{0\cdot}p_{\cdot 1} - D \\ p_{1\cdot}p_{\cdot 0} - D & p_{1\cdot}p_{\cdot 1} + D \end{pmatrix}, \tag{2.5}$$

where $D = p_{00}p_{11} - p_{01}p_{10} = p_{11} - p_{1\cdot}p_{\cdot 1}$, and so on. In genetics, D is known as the coefficient of linkage disequilibrium for two genes.

2.1 Estimation from Data. Equation 2.1 is given in terms of probabilities such as p_{01}. However, observational data do not directly provide such probabilities but counts associated with the corresponding cells. The maximum likelihood estimator (MLE) for p_{ij} is, of course, n_{ij}/n, where n_{ij} is the count associated with cell ij, and n is the total number of counts. The MLE has well-known issues when (some of) the counts are small. Bayesian approaches to address this are discussed in section 5.

3 Classical Measures of Association

For two gaussian continuous random variables, there is a natural measure of their association, the correlation coefficient. This is independent of the

individual (marginal) variances of each variable, and lies in the interval $[-1, 1]$.

For the 2×2 table, many measures of association have been devised. One is Yule's Y (Yule, 1912), where

$$Y = \frac{\sqrt{\lambda} - 1}{\sqrt{\lambda} + 1}. \tag{3.1}$$

Like the correlation coefficient, Y also lies in the range of $[-1, 1]$, with a value of 0 reflecting that there is no association. Its dependence only on λ means that it is invariant to the marginals in the table. $Y(1/\lambda) = -Y(\lambda)$, so Y is an odd function of $\log(\lambda)$. Edwards (1963) argued that measures of association must be functions of the odds ratio.

There are a number of desirable properties for a measure of association η between binary variables. For example, Hasenclever and Scholz (2016, p. 22) list these:

1. η is zero on independent tables.
2. η is a strictly increasing function of the odds ratio when restricted to tables with fixed margins.
3. η respects the symmetry group D_4, namely, η is symmetric in the markers (i.e., invariant to matrix transposition), and η changes sign when the states of a marker are transposed (row or column transposition).
4. The range of the function is restricted to $(-1, 1)$.

As well as Yule's Y,[2] several other measures of association have been proposed; indeed Tan, Kumar, and Srivastva (2004) list 21. Other measures of association include Lewontin's D' (1964), which standardizes D from equation 2.5 by dividing it by the maximum value it can take on, which depends on the marginals of the table, and the binary correlation coefficient r, which standardizes D by $\sqrt{p_{0\cdot}\, p_{\cdot 0}\, p_{1\cdot}\, p_{\cdot 1}}$. For the canonical table, it turns out that $D' = r = Y$.

4 Information-Theoretic Measures of Association

Barlow's definition of a suspicious coincidence suggests consideration of the quantity

$$i(x, y) = \log \frac{p(x, y)}{p(x)p(y)}. \tag{4.1}$$

[2] Yule (1900) had earlier proposed $Q = (\lambda - 1)/(\lambda + 1)$ as a measure of association, but his discussion on p. 592 of Yule (1912) gives a number of reasons for preferring Y to Q.

Indeed $i(x, y)$ has been proposed in different literatures. For example, Church and Hanks (1990) studied it for word associations in linguistics. $i(x, y)$ is termed the *pointwise mutual information* (PMI) in, for example, the statistical natural language processing textbook of Manning and Schütze (1999). In pharmacovigilance, Bate et al. (1998) call $i(x, y)$ the *information component* (IC), as it is one component of the mutual information calculation in a 2×2 table, and it is also studied in DuMouchel (1999). And in the data mining literature, Silverstein, Brin, and Motwani (1998) define the *interest* to be the ratio $p(x, y)/(p(x)p(y))$ (i.e., without the log).

Note that while Y, D', and r consider the difference, $D = p_{11} - p_{1\cdot}p_{\cdot 1} = p(x, y) - p(x)p(y)$, $i(x, y)$ considers the log ratio of these terms. Thus, $i(x, y)$ considers the ratio of the observed and expected probabilities for the event (x, y), where the expected model is that of independence.

The mutual information (MI) is defined as

$$I(X; Y) = \sum_{i,\, j \in \{0,1\}} p(X = i, Y = j) \log \frac{P(X = i, Y = j)}{P(X = i)P(Y = j)}. \tag{4.2}$$

We have that $I(X; Y) \geq 0$, with $I(X; Y) = 0$ when X and Y are independent.

Both PMI and MI as defined above depend on the marginal probabilities in the table. To see this, use $p(x, y) \leq p(x)$ or $p(x, y) \leq p(y)$, so $i(x, y) \leq \min(-\log p(x), -\log p(y))$, that is, favoring "sparsity" (low probability). The MI is maximal for a diagonal (or antidiagonal) table with marginals of $1/2$, the opposite trend to PMI.

There have been various proposals to normalize the PMI and MI to make them fit in the range $[-1, 1]$ and $[0, 1]$, respectively. For example, Bouma (2009) defined the normalized PMI (NPMI) as $i_n(x, y) = i(x, y)/h(x, y)$ for $p(x, y) > 0$, where $h(x, y) = -\log p(x, y)$. NPMI ranges from $+1$ when events x and y only occur together, through 0, when they are independent, to -1 when x and y occur separately but not together. Similarly there are a number of proposals for normalizing the mutual information, Bouma (2009) suggests $I_n(X; Y) = I(X; Y)/H(X, Y)$, where $H(X, Y)$ is the joint entropy of X and Y. $I_n(X; Y)$ (termed the normalized MI or NMI) takes on a value of $+1$ if X and Y are perfectly associated and 0 if they are independent. Alternative normalizations of the MI by $H(X)$ or $H(Y)$ have also been proposed; Press, Teukolsky, Vetterling, and Flannery (2007, sec. 14.7.4) term these the uncertainty coefficients. NMI is not strictly a measure of association as defined above, as it does not take on negative values, but following the construction in Hasenclever and Scholz (2016), one can, for example, define the *signed* NMI as $\text{sign}(D)I_n(X; Y)$.

Given that the canonical table removes the effect of the marginals, it is natural to consider PMI and MI as a function of λ. Using the canonical table from equation 2.4, we obtain

$$i_\lambda(x, y) = \log \frac{2\sqrt{\lambda}}{1 + \sqrt{\lambda}}, \tag{4.3}$$

which takes on a value of 0 for $\lambda = 1$ (independence) and tends to a value of log(2) as λ tends to infinity. For $\lambda < 1$ the value of $i_\lambda(x, y)$ becomes negative and diverges to $-\infty$ as $\lambda \to 0$. However, study of the canonical table indicates would make more sense in this case to consider one of the "antidiagonal" cells in P_{can}, which will have a probability greater than 1/2 as the "event." In general we can treat all four cells of the contingency table as the joint event, compute the PMI for each, and return the maximum. For the canonical table with $\lambda < 1$, this means that we transform $\lambda \to 1/\lambda$ and compute i_λ as per equation 4.3.

For the MI of the canonical table, we obtain (after some manipulation)

$$I_\lambda(X; Y) = \frac{\sqrt{\lambda}}{1 + \sqrt{\lambda}} \log \sqrt{\lambda} - \log(1 + \sqrt{\lambda}) + \log 2. \tag{4.4}$$

Analysis of $I_\lambda(X; Y)$ shows that it is invariant if we transform λ to $1/\lambda$, so a plot of $I_\lambda(X; Y)$ against $\log(\lambda)$ is symmetric around 0 and tends to the value log(2) as λ tends to 0 or infinity.

Plots of Y, i_λ, and I_λ for $\lambda \geq 1$ in Figure 1 show similar behavior, monotonically increasing to a maximum value as $\lambda \to \infty$. If we choose logs to base 2, then the maximum value is 1 in all three cases. As Y is already well established (since 1912!), it does not seem necessary to promote i_λ or I_λ as alternatives, when considering the canonical table.

5 Detecting Associations with Pointwise Mutual Information

As we have seen, the raw PMI score is not invariant to the distribution of the marginals. This can be seen in Table 1, which concerns the association between vaccination and death from smallpox; the original proportions in panel a are based on the Sheffield data in Table I of Yule (1912). In panel b, the marginals of the table with regard to vaccination have been adjusted to 50/50 (as may have happened if these data had been collected in a randomized, controlled trial), and in panel c, we have the canonical table where both marginals are 50/50.[3] Notice that the PMI is highest for the original (unbalanced) table and decreases as the marginals are balanced. Conversely, the MI is lowest in the the original (unbalanced) table and increases as the marginals are balanced. Of course, Yule's Y is constant throughout, by construction.

[3] Yule (1912) comments that on the canonical table, "These are, of course, not the actual proportions, but the proportions that would have resulted if an omnipotent demon of unpleasant character (no relation of Maxwell's friend) could have visited Sheffield . . . ,

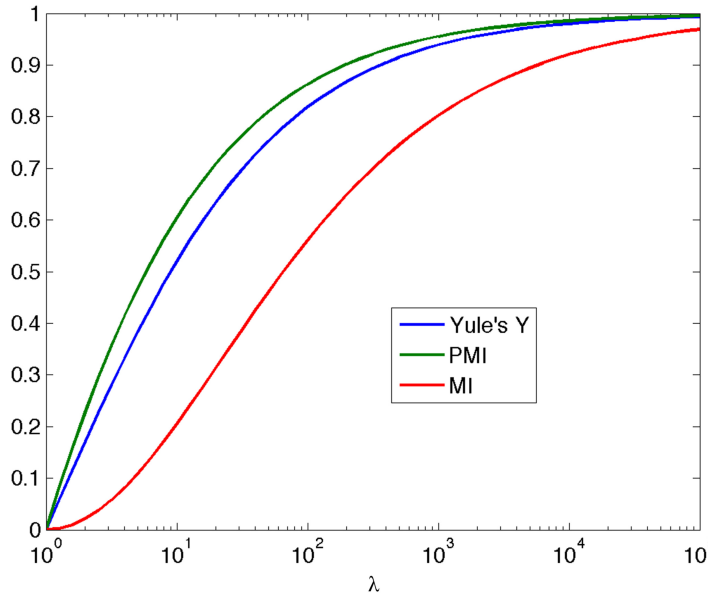

Figure 1: Plots of Y, i_λ (PMI), and I_λ (MI) against λ (log scale) for $\lambda \geq 1$.

Table 1: 2×2 Contingency Tables for the Association between Vaccination and Death from Smallpox.

	(a) Original Table			(b) Vaccination Rate 50%			(c) Canonical Table		
	PMI = 2.300, MI = 0.108			PMI = 0.866, MI = 0.205			PMI = 0.705, MI = 0.310		
	Recover	Die	Marginals	Recover	Die	Marginals	Recover	Die	Marginals
Vaccinated	0.840	0.043	0.883	0.476	0.024	0.500	0.408	0.092	0.500
Unvaccinated	0.059	0.058	0.117	0.252	0.248	0.500	0.092	0.408	0.500
Marginals	0.899	0.101		0.728	0.272		0.500	0.500	

Note: Panel a is the original table based on the data in Yule (1912), panel b adjusts the marginals for vaccinated/unvaccinated to be 50/50, and panel c is the canonical table where the marginals are both 50/50. In all three tables, Yule's $Y = 0.630$.

As another example, consider fixing λ but adjusting the marginal probabilities of events x and y. For example, for $\lambda = 16$, PMI takes on the values of 0.678, 1.642, 2.293, 3.642, and 3.958 (using logs to base 2) as $p(x) = p(y)$ varies from 0.5, 0.2, 0.1, 0.01, and 0.001. This is particularly problematic as low counts will give rise to uncertainty in the estimation of the required

and raised the fatality rate and the proportion of unvaccinated . . . to 50 per cent without otherwise altering the facts."

probabilities (especially of the joint event). In the context of word associations, Manning and Schütze (1999, sec. 5.4) argue that PMI "does not capture the intuitive notion of an interesting collocation very well" and mention work that multiplies it by $p(x, y)$ as one strategy to compensate for the bias in favor of rare events.

Barlow (1985) suggested that sparsity is important for the detection of suspicious coincidences, that is, that "the events themselves must be rare ones." It is true that a low $p(y)$ gives more "headroom" for the ratio $p(y|x)/p(y)$ to be large. The PMI score is used extensively in pharmacovigilance, where the aim is to detect associations between drugs taken and adverse drug reactions (ADRs). In this context, the ratio $p(x, y)/p(x)p(y) = p(y|x)/p(y)$ is termed the *relative reporting ratio* (RRR) and compares the relative probability of an adverse drug reaction y given treatment with drug x, compared to the base rate $p(y)$. Another commonly used measure is the *proportional reporting ratio* (PRR), defined as $p(y|x)/p(y|\neg x)$. A US Food and Drug Administration (FDA) white paper (Duggirala et al., 2018) describes the use of both RRR and PRR for detecting ADRs in routine surveillance activities.

Above, we have described the maximum likelihood estimation for the probabilities in the 2×2 table, based on counts. However, there are well-known issues with the MLE when (some of) the counts are small. This naturally suggests a Bayesian approach, and there is a considerable literature on the Bayesian analysis of contingency tables, as reviewed, for example, in Agresti (2013). There are different sampling models depending on how the data are assumed to be generated, as described in Agresti (2013, sec. 2.1.5). If all four counts are unrestricted, a natural assumption is that each n_{ij} is drawn from a Poisson distribution with mean μ_{ij}, which can be given a gamma prior. Alternatively, if n is fixed, the sampling model is a multinomial, and the conjugate prior is a Dirichlet distribution. If one set of marginals is fixed, then the data are drawn from two binomial distributions, each of which can be given a beta prior. If both marginal totals are fixed, this corresponds to Fisher's famous "lady tasting tea" experiment, and the sampling distribution of any cell in the table follows a hypergeometric distribution. Section 3.6 of Agresti (2013) covers Bayesian inference for two-way contingency tables, and Agresti and Min (2005) discuss Bayesian confidence intervals for association parameters, such as the odds ratio.

DuMouchel (1999) applied an empirical Bayes approach to consider sampling variability for PMI (a.k.a. RRR) in the context of adverse drug reactions. He assumed that each n_{11} is a draw from a Poisson distribution with unknown mean μ_{11} and that the object of interest is $\rho_{11} = \mu_{11}/E_{11}$, where E_{11} is the expected count (assumed known) under the assumption that the variables are independent. Using a mixture of gamma distributions prior for ρ_{11}, DuMouchel obtained the posterior mean $E[\log(\rho_{11})|n_{11}]$ rather than just considering the sample estimate n_{11}/E_{11}. The mixture prior was used to express the belief that when testing many associations, most will have a PMI

of near zero, but there will be some with significantly larger values. This method is known as the multi-item gamma Poisson shrinker (MGPS). The value of this approach is that Bayesian shrinkage corrects for the high variability in the RRR sample estimate n_{11}/E_{11} that results from small counts.

6 Summary

Motivated by Barlow's hypothesis about suspicious coincidences, we have reviewed the properties of 2×2 contingency tables for association analysis, with a focus on the odds ratio λ and Yule's Y. We have considered the mutual information and pointwise mutual information as measures of association, along with normalized versions thereof. We have shown that, considered as functions of λ in the canonical table, MI and PMI behave similar to Y for $\lambda \geq 1$, increasing monotonically with λ (and can be made similar for $0 > \lambda > 1$).

As well as Y, the PMI measure $i(x, y) = \log p(x, y)/(p(x)p(y)$ can also be used to identify suspicious coincidences, and it is used in practice—for example, in pharmacovigilance. We have discussed the pros and cons of using it in this way, bearing in mind the sensitivity of the PMI to the marginals, with increased scores for sparser events. When some of the counts in the table are low, Bayesian approaches can be useful for estimating PMI from raw counts.

Acknowledgments

I thank Peter Dayan and Iain Murray for helpful comments on an early draft of this note and the anonymous reviewers whose comments helped to improve the note.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Agresti, A., & Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency tables. *Biometrics*, *61*, 515–523. 10.1111/j.1541-0420.2005.031228.x, PubMed: 16011699

Barlow, H. B. (1985). Cerebral cortex as model builder. In D. Rose & V. G. Dobson (Eds.), *Models of the visual cortex* (pp. 37–46). New York: Wiley.

Barlow, H. B. (1987). Cerebral cortex as model builder. In L. M. Vaina (Ed.), *Matters of Intelligence: Conceptual structures in cognitive neuroscience* (pp. 395–406). Dordrecht: D. Reidel.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., & De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal detection. *European Journal of Clinical Pharmacology*, *54*, 315–321. 10.1007/s002280050466, PubMed: 9696956

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference 2009*.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, *16*(1), 22–29.

Duggirala, H. J., Tonning, J. M., Smith, E., Bright, R., Baker, J. D., Ball, R., . . . Kass-Hout, T. (2018). *Data mining at FDA* (White Paper). https://www.fda.gov/science-research/data-mining/data-mining-fda-white-paper

DuMouchel, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *American Statistician*, *53*(3), 177–190.

Edelman, S., Hiles, B., Yang, H., & Intrator, N. (2002). Probabilistic principles in unsupervised learning of visual structure: Human data and a model. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*, *14*. Cambridge, MA: MIT Press.

Edwards, A. W. F. (1963). The measure of association in a 2×2 table. *Journal of the Royal Statistical Society, Series A (General)*, *126*(1), 109–114. 10.2307/2982448

Hasenclever, D., & Scholz, M. (2016). Comparing measures of association in 2×2 probability tables. *Open Statistics and Probability Journal*, *7*, 20–35. 10.2174/1876527001607010020

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; Heterotic models. *Genetics*, *49*(1), 49–67. 10.1093/genetics/49.1.49, PubMed: 17248194

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing* (3rd ed.). Cambridge: Cambridge University Press.

Silverstein, C., Brin, S., & Motwani, R. (1998). Beyond market baskets, Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, *2*(1), 39–68. 10.1023/A:1009713703947

Tan, P.-N., Kumar, V., & Srivastva, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, *29*, 293–313. 10.1016/S0306-4379(03)00072-3

Yule, G. U. (1900). On the association of attributes in statistics. *Phil. Trans. Roy. Soc., A*, *194*, 257.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *75*(6), 579–652. 10.2307/2340126

Received March 15, 2022; accepted June 2, 2022.