Partitioning the Cressie-Read divergence statistic for three-way contingency tables: a study on environmental sustainability data.

Decomposizione delle statistiche divergenti di Cressie-Read per tabelle di contingenza a tre-vie. Test su dati di sostenibilità ambientale.

Rosaria Lombardo and Eric J. Beh

Abstract When studying the association between the variables of a three-way contingency table, Lancaster [10] proposed different partitions of Pearson's three-way chi-squared statistic. This statistic is a special case of the three-way generalisation of the Cressie-Read divergence statistic [6]. To test the association among environmental sustainability variables, this paper presents an additive orthogonal partition of the generalised Cressie-Read divergence statistic under the assumption of complete independence between the variables.

Abstract *Per l'analisi dell'associazione tra le variabili di una tabella di contingenza a tre-vie, Lancaster [10] propose differenti partizioni del chi-quadrato a tre-vie di Pearson. Questa statistica è anche vista come un caso particolare della statistica divergente di Cressie-Read [6], generalizzata per tabelle a tre-vie. Per verificare la significativà statistica dell'associazione tra alcune variabili della sostenibilità ambientale, questo lavoro presenta una partizione ortogonale additiva della statistica divergente di Cressie-Read generalizzata, sotto l'ipotesi di indipendenza completa tra le variabili.*

Key words: Orthogonal Partition, Cressie-Read divergence statistic, Testing association

1 Three-way Cressie-Read Divergence Statistic

To determine whether there exists a statistically significant association between the row, column and tube variables of a three-way contingency table, one may calculate

Rosaria Lombardo
University of Campania L. Vanvitelli, Gran Priorato di Malta, Capua (CE), e-mail: rosaria.lombardo@unicampania.it

Eric J. Beh
University of Newcastle, Australia e-mail: eric.beh@newcastle.edu.au

any number of measures. The most common that are used for such a purpose include the three-way generalisation of Pearson's chi-squared statistic [10, 5] and the log-likelihood ratio statistic. These statistics can be shown to be special cases of the three-way Cressie-Read divergence statistic [6, 14, 15, 16] as are generalisations of the Freeman-Tukey statistic [7], the modified chi-squared statistic [12, 13] and the modified log-likelihood ratio statistic [9]. All of these statistics, and other special cases of the divergence statistic, are chi-squared random variables with $(IJK-1)-(I-1)-(J-1)-(K-1)$ degrees of freedom.

Here, we propose a three-way extension of the Cressie-Read divergence statistic and then examine a method of orthogonally partitioning this statistic. Doing so provides a means of generalising to three categorical variables the benefits of the Cressie-Read divergence statistic. The partition can also be used to examine and test the association between three categorical variables under the assumption of complete independence and in the presence of sparse data. Further generalisations of the partition to the multi-way case can certainly be considered but we shall examine this extension at a later date.

Let $\underline{\mathbf{N}}$ be an $I \times J \times K$ three-way contingency table belonging to the space $\Re^{I\times J\times K}$, where the (i, j, k)th cell entry has a frequency of n_{ijk} for $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$. Define n the grand total of $\underline{\mathbf{N}}$ and let the matrix of relative frequencies be $\underline{\mathbf{P}}$ so that its (i, j, k)th cell entry is $p_{ijk} = n_{ijk}/n$ where $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk} = 1$. Define the ith row marginal proportion by $p_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}$. Similarly, let $p_{\bullet j\bullet} = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ijk}$ be the jth column marginal proportion, and $p_{\bullet\bullet k} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ijk}$ the kth tube marginal proportion.

The Cressie-Read divergence statistic [6] has been extensively studied for two-way contingency tables and has been extended for studying the association in a three-way contingency table; see, for example, [15]. Such a divergence statistic is defined here as

$$\mathrm{CR}(\delta) = \frac{2n}{\delta(\delta+1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}\left\{\left(\frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^{\delta} - 1\right\}, \tag{1}$$

where $\delta \in (-\infty, \infty)$. The general nature of (1) ensures that specific values of δ lead to specific measures of association, all of which are chi-squared random variables.

The most common special cases of (1) were considered by [14]. In this paper, we focus our attention on some of those special cases. Specifically, we focus our attention on Pearson's chi-squared statistic (when $\delta = 1$), on the Cressie-Read statistic (when $\delta = 2/3$) and on the Freeman-Tukey statistic ($\delta = -1/2$) which are, respectively,

$$\mathrm{CR}(\delta = 1) = X^2 = n\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(p_{ijk} - p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}\right)^2}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}} \tag{2}$$

$$\mathrm{CR}(\delta = 2/3) = CR = \frac{9n}{5}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}\left[\left(\frac{p_{ijk}}{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^{2/3} - 1\right] \tag{3}$$

$$\mathrm{CR}\left(\delta=-\frac{1}{2}\right)=T^2=4n\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\sqrt{p_{ijk}}-\sqrt{p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^2. \tag{4}$$

Other measures, which are notable members of the family of Cressie-Read divergence statistics, can be generalised to three-way data and include the modified chi-squared statistic $N^2=CR(\delta=-2)$, the log-likelihood ratio statistic $G^2=CR(\delta=0)$ and its modified version $M^2=CR(\delta=-1)$.

In the context of goodness-of-fit testing for two categorical variables, Cressie and Read [6] examine the appropriate values of δ that one should use. For a two-way contingency table, such that $\delta\neq-1,0$, they also recommend that $\delta\in(0,3/2]$ when $n>10$ and $\min(np_{i\bullet}p_{\bullet j})>1$ for all $i=1,2,\ldots,I$ and $j=1,2,\ldots,J$. Cressie and Read [6] advised that an appropriate choice of δ is $2/3$ leading to their statistic; $\mathrm{CR}(\delta=2/3)=CR$. The selection criteria for δ for a three-way contingency table can be made in an analogous manner.

The term $p_{ijk}/(p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k})$ in (1) is defined as the *Pearson ratio* of the (i,j,k)th cell of the contingency table and is just the ratio of the observed cell frequency to what is expected under complete independence; see [2, p. 123] and [3, 8] for a definition of this ratio in the context of correspondence analysis. When this ratio is equal to 1 for all cells, (1), and hence its special cases (including X^2, G^2, T^2, N^2 and M^2), is zero providing evidence that the three variables of $\underline{\mathbf{N}}$ are completely independent. One advantage of considering the Pearson ratio's is that they ensure that the log-transformation of the cell's proportion, p_{ijk}, is "triple-centred" by the log-transformed row, column and tube proportions, i.e. $\ln\left(p_{ijk}/(p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k})\right)=\ln\left(p_{ijk}\right)-\ln(p_{i\bullet\bullet})-\ln(p_{\bullet j\bullet})-\ln(p_{\bullet\bullet k})$.

1.1 Partitioning the Cressie-Read divergence statistics

Here, we present an additive orthogonal, and ANOVA-like, partition of $\mathrm{CR}(\delta)$, defined by (1), from which the Pearson's chi-squared statistic and its companion measures of association can be derived in a straightforward manner. As an ANOVA-like partition, we consider the classical definition of inner products and orthogonality conditions for partitioning a measure of association belonging to the space $\Re^{I\times J\times K}$; for more details see [5, 11] and [2, Chapter 11]. Therefore, the general partition of (1) can be written as

$$\begin{aligned}\mathrm{CR}(\delta)=&\frac{2n}{\delta(\delta+1)}\sum_{i=1}^{I}\sum_{j=1}^{J}p_{ij\bullet}\left\{\left(\frac{p_{ij\bullet}}{p_{i\bullet\bullet}p_{\bullet j\bullet}}\right)^{\delta}-1\right\}\\&+\frac{2n}{\delta(\delta+1)}\sum_{i=1}^{I}\sum_{k=1}^{K}p_{i\bullet k}\left\{\left(\frac{p_{i\bullet k}}{p_{i\bullet\bullet}p_{\bullet\bullet k}}\right)^{\delta}-1\right\}\\&+\frac{2n}{\delta(\delta+1)}\sum_{j=1}^{J}\sum_{k=1}^{K}p_{\bullet jk}\left\{\left(\frac{p_{\bullet jk}}{p_{\bullet j\bullet}p_{\bullet\bullet k}}\right)^{\delta}-1\right\}\end{aligned}$$

$$+ \frac{2n}{\delta(\delta+1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} p_{ijk} \left\{ \left(\frac{p_{ijk}}{\alpha p_{\bullet j \bullet} p_{\bullet\bullet k}} \right)^{\delta} - 1 \right\}$$
$$= \mathrm{CR}_{IJ}(\delta) + \mathrm{CR}_{IK}(\delta) + \mathrm{CR}_{JK}(\delta) + \mathrm{CR}_{IJK}(\delta). \quad (5)$$

Here we can see that there are four terms in the partition. The first three terms, $\mathrm{CR}_{IJ}(\delta)$, $\mathrm{CR}_{IK}(\delta)$ and $\mathrm{CR}_{JK}(\delta)$, are the bivariate Cressie-Read divergence statistics that assess the row-column, row-tube and column-tube association, respectively. Therefore, these measures are asymptotically chi-squared random variables with $(I-1)(J-1)$, $(I-1)(K-1)$ and $(J-1)(K-1)$ degrees of freedom, respectively. The last term, CR_{IJK}, is the measure of three-way, or *trivariate*, association between all three variables and is asymptotically a chi-squared random variable with $(I-1)(J-1)(K-1)$ degrees of freedom.

2 Application

In recent years, sustainable development goals (SDG) have become increasingly important. Decision makers everywhere need data and statistics that are accurate, timely, sufficiently disaggregated, relevant, accessible and easy to use. Here among the plethora of authoritative SDG data sources (available at https://unstats-undesa.opendata.arcgis.com/), we study the association between the *Renewable energy share* in the *total final energy consumption indicator* (RES), the *indicator of adjusted emission growth rate for black carbon indicator* (BCA) and the *Geographical area* (Africa, America, Asia, Australia, Caribbean islands, Europe) of 186 countries in 2018 (GEO). These three variables (RES, BCA and GEO) can be cross-classified to produce the $4 \times 4 \times 6$ contingency of Table 1. We now describe the nature of the three variables and examine the partition of $\mathrm{CR}(\delta)$ – see (5) – for the table where $\delta = 1, 1/2$ and $2/3$.

The four categories of the row variable RES are (0, 34.5], (34.5, 53.7], (53.7, 80.7] and (80.7, 100]. The column variable *BCA* also has four ordered categories; (0, 8.13], (8.13, 23.1], (23.1, 49.7] and (49.7, 96.4]. These variables are formed by dividing each continuous variable into quantiles. While both are ordinal variables we shall be treating them as nominal. The categories of the tube variable GEO are the six areas mentioned above.

We study the association of these three variables through the partition of (1); see (5). Since Table 1 has many zero cell frequencies, we compare the results of the partition of Pearson's chi-squared statistic with the results obtained from the partition of the three-way versions of the Freeman-Tukey statistic (T^2) and the Cressie-Read statistic (CR). We consider these last two statistics since there is a strong presence of overdispersion in data; see [4].

Pearson's chi-squared of Table 1 is $\mathrm{CR}(1) = 290.035$. With 84 degrees of freedom, this statistic has a p-value that is less than 0.001 and so a statistically significant association exists between the three variables. Table 2 summarises the partition

of this statistic which can be obtained by considering the three-way Cressie-Read divergence statistic with $\delta = 1$. All four terms of the partition are statistically significant with a p-value less than 0.001 except for the trivariate association term whose p-value is 0.002.

We take into account the presence of sparse data by investigating the partition of the Freeman-Tukey statistic ($\delta = 1/2$) and of the Cressie-Read statistic ($\delta = 2/3$). These two statistics are $\mathrm{CR}(1/2) = 250.491$ and $\mathrm{CR}(2/3) = 259.305$, respectively, and like $\mathrm{CR}(1) = X^2$ each has a p-value that is less than 0.001 thereby confirming that there exists a statistically significant association between the variables. Table 2 provides a summary of the partition of T^2 and CR and shows, like X^2, that each of the bivariate terms have a p-value that is less than 0.001. However, the trivariate association term is no longer statistically significant; see Table 2. This is likely due to the sparsity of many of the cell frequencies. Thus we can make the following

Table 1 Cross-classification of RES, BCA and GEO

	BCA			
RES	$(0, 8.13]$	$(8.13, 23.1]$	$(23.1, 49.7]$	$(49.7, 96.4]$
		Africa		
(0, 34.5]	1	0	4	8
(34.5, 53.7]	0	2	4	19
(53.7, 80.7]	3	2	1	4
(80.7, 100]	1	1	0	1
		America		
(0, 34.5]	1	0	6	2
(34.5, 53.7]	0	2	0	1
(53.7, 80.7]	0	3	1	0
(80.7, 100]	1	3	2	1
		Asia		
(0, 34.5]	5	1	2	0
(34.5, 53.7]	6	2	3	1
(53.7, 80.7]	5	3	3	3
(80.7, 100]	4	2	0	0
		Australia		
(0, 34.5]	1	0	0	0
(34.5, 53.7]	0	1	2	0
(53.7, 80.7]	1	1	5	0
(80.7, 100]	1	0	0	0
		Carribean		
(0, 34.5]	3	2	0	0
(34.5, 53.7]	1	1	0	1
(53.7, 80.7]	0	0	0	0
(80.7, 100]	2	3	0	0
		Europe		
(0, 34.5]	1	0	0	0
(34.5, 53.7]	0	0	1	0
(53.7, 80.7]	0	7	5	0
(80.7, 100]	3	11	7	3

Table 2 Partition of $\mathrm{CR}(\delta)$

Association	Term	%	*df*	*p*-value
	$\mathrm{CR}(1) = X^2$			
$I \times J$	34.970	12%	9	<0.001
$I \times K$	82.816	29%	15	<0.001
$J \times K$	95.677	33%	15	<0.001
$I \times J \times K$	76.573	26%	45	0.002
X^2	290.035	100%	84	<0.001
	$\mathrm{CR}(1/2) = T^2$			
$I \times J$	35.112	14%	9	<0.001
$I \times K$	83.340	33%	15	<0.001
$J \times K$	92.4207	37%	15	<0.001
$I \times J \times K$	39.619	16%	45	0.699
T^2	250.491	100%	84	<0.001
	$\mathrm{CR}(2/3) = CR$			
$I \times J$	34.988	13%	9	<0.001
$I \times K$	82.653	32%	15	<0.001
$J \times K$	93.105	36%	15	<0.001
$I \times J \times K$	48.559	19%	45	0.332
CR	259.305	100%	84	<0.001

conclusions about the nature of the association between the variables of Table 1. There is a statistically significant association between

- the total final energy consumption and the indicator of adjusted emission growth rate for black carbon,
- total final energy consumption and the geographical area,
- the indicator of adjusted emission growth rate for black carbon and the geographical area,

while no such association exists between all three variables.

References

1. Agresti, A.: Categorical Data Analysis (2nd ed). Wiley, New York (2002)
2. Beh, E. J., Lombardo, R.: Correspondence Analysis: Theory, Practice and New Strategies. Wiley, Chichester (2014)
3. Beh, E. J.: Simple correspondence analysis: A bibliographic review. International Statistical Review **72**, 257–284 (2004)
4. Beh, E. J., Lombardo, R., Alberti, G.: Correspondence analysis and the Freeman-Tukey statistic: A study of archaeological data. Computational Statistics and Data Analysis **128**, 73–86 (2018)
5. Carlier, A., Kroonenberg, P. M.: Biplots and decompositions in two-way and three-way correspondence analysis. Psychometrika **61**, 355–373 (1996)

6. Cressie, N. A. C., Read, T. R. C.: Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, Series B **46**, 440–464 (1984)
7. Freeman, M. F., Tukey, J. W.: Multinomial goodness-of-fit tests. The Annals of Mathematical Statistics **21**, 607–611 (1950)
8. Greenacre, M.: Power transformations in correspondence analysis. Computational Statistics and Data Analysis **53**, 3107–3116 (2009)
9. Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
10. Lancaster, H. O.: Complex contingency tables treated by the partition of chi-square. Journal of of the Royal Statistical Society B **13**, 242–249 (1951)
11. Lombardo, R., Takane, Y., Beh, E. J.: Familywise decompositions of Pearson's chi-square statistic in the analysis of contingency tables. Advances in Data Analysis and Classification **14(3)**, 629–649 (2019)
12. Neyman, J.: Contribution to the theory of certain test criteria. Bulletin de L'Institut International de Statistique **24**, 44–86 (1940)
13. Neyman, J.: Contributions to the theory of the χ^2 test. In: Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability , 239–273 (1949)
14. Pardo, M. C.: An empirical investigation of Cressie and Read tests for the hypothesis of independence in three-way contingency tables. Kybernetika **32**, 175–183 (1996)
15. Pardo, L., Pardo, M. C.: Minimum power-divergence estimator in three-way contingency tables. Journal of Statistical Computation and Simulation **73**, 819–831 (2003)
16. Pardo, J. A.: An approach to multiway contingency tables based on ϕ-divergence test statistics. Journal of Multivariate Analysis **101**, 2305–2319 (2010)