The Aggregate Association Index

Eric J. Beh
School of Mathematical & Physical Sciences, University of Newcastle, Australia

Abstract

Recently Beh (2008, *JSPI*) presented an index that helps to identify how likely two dichotomous categorical variables may be associated given only the aggregate (or marginal) information. Such an index was referred to as the aggregate association index. This paper will further consider some of the issues concerned with that index. These include variations of the original index as well as adaptations for quantifying the possibility that there exists a statistically significant positive or negative association between the two dichotomous variables.

Keywords: 2×2 Contingency Table; Correlation; Ecological Inference.

1 The 2×2 Contingency Table

Consider a single two-way contingency table where both variables are dichotomous in nature. Suppose that n individuals/units are classified into this table such that the number classified into the (1, 1)th cell is denoted by n_{11}. Let the i'th row marginal frequency be denoted by $n_{i\bullet}$, for i = 1, 2, and the j'th column marginal frequency by $n_{\bullet j}$, for j = 1, 2. Also, denote the i'th row and j'th column marginal proportion by $p_{i\bullet} = n_{i\bullet}/n$ and $p_{\bullet j} = n_{\bullet j}/n$ respectively. Table 1 provides a description of this notation.

Table 1
Notation for a single 2×2 contingency table

	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Suppose that the cell values in Table 1 are unknown so that only the information in the marginal frequencies is known, and fixed. This is commonly the situation in many studies where it is prohibitive (because of reasons of confidentiality) or impossible (because such information was never obtained) to know the value of the cells.

The problem at hand is to obtain some information concerning the nature of the association between the two dichotomous variables when only the marginal information is provided. When a single 2×2 table is of interest (as is the case in this paper), Fisher (1935) considered this issue and judged there to be very little or no information in the margins for inferring individual (or cellular) level data. More recent discussions, including those by Plackett (1977), Aitkin and Hinde (1984), Barnard (1984) and Beh (2008) concluded that the marginal information was not completely useless for making such inferences. For a set of G 2×2 tables (for example, such tables collected at G

geographical/institutional regions) the problem of understanding the nature of the association between the two dichotomous variables given only the marginal information falls within the realm of ecological inference. Political scientists and statisticians have proposed a variety of solutions to the problem including those of King (1997), Steel, *et al.* (2005) and others, however the issue of the applicability of these results to G 2×2 tables will not be considered here.

Consider the case where, for now, the cell values of Table 1 are known. Define the proportions

$$P_1 = \frac{n_{11}}{n_{1\bullet}} \qquad \text{and} \qquad P_2 = \frac{n_{21}}{n_{2\bullet}}.$$

Here, P_1 is the conditional probability of an individual/unit being classified into "Column 1" given that they are classified in "Row 1". Similarly P_2 is the conditional probability of an individual/unit being classified into "Column 1" given that they are classified in "Row 2". For reasons of simplicity, we shall focus only on P_1 in this paper. Although similar conclusions can be made concerning P_2.

When the joint frequencies of Table 1 are not known P_1 will also be unknown, although since it is a proportion it will lie within the interval [0, 1]. Duncan and Davis (1953) showed that this interval can be narrowed such that

$$L_1 = \max\left(0, \frac{n_{\bullet 1} - n_{2\bullet}}{n_{1\bullet}}\right) \le P_1 \le \min\left(\frac{n_{\bullet 1}}{n_{1\bullet}}, 1\right) = U_1. \tag{1}$$

If one wishes to obtain a $100(1-\alpha)\%$ confidence interval for P_1 given only the marginal information in the 2×2 contingency table, Beh (2008) showed that such an interval is

$$L_\alpha^* = p_{\bullet 1} - p_{2\bullet}\sqrt{\frac{\chi_\alpha^2}{n}\left(\frac{p_{\bullet 1}p_{\bullet 2}}{p_{1\bullet}p_{2\bullet}}\right)} < P_1 < p_{\bullet 1} + p_{2\bullet}\sqrt{\frac{\chi_\alpha^2}{n}\left(\frac{p_{\bullet 1}p_{\bullet 2}}{p_{1\bullet}p_{2\bullet}}\right)} = U_\alpha^* \tag{2}$$

where χ_α^2 is the $1-\alpha$ percentile of the chi-squared distribution with 1 degree of freedom. However, for extremely small sample sizes and/or a small α value, the bounds can lie outside of the permissible range of [0, 1]. For example, the 95% confidence interval of P_1 for the set of marginal frequencies $(n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}, n_{\bullet 2}) = (2, 3, 2, 3)$ is (-0.126, 0.926). This problem can be alleviated by decreasing the level of confidence, say to 85% which gives (0.0137, 0.786). Alternatively increasing the sample size can help resolve the problem. For example, if a sample was selected which leads to a doubling of each of the marginal frequencies that appeared in the original table, the 95% confidence interval is (0.028, 0.772). A more appropriate resolution to the problem is to enforce the bounds to lie within [0, 1] in the same manner as undertaken by Duncan and Davis (1953). That is, we shall consider modifying the bounds of P_1 given by (2) such that

$$L_\alpha = \max(0, L_\alpha^*) < P_1 < \min(1, U_\alpha^*) = U_\alpha.$$

Given the marginal information in the table, and a level of significance α, one may conclude that there may exist a statistically significant association between the two dichotomous variables if $L_1 \le P_1 \le L_\alpha$ or $U_\alpha \le P_1 \le U_1$. If this is the case, then for a given α, the marginal information provides some evidence to suggest that a statistically significant association between the two dichotomous variables may exist. Such an interval is derived assuming that the two dichotomous variables are independent.

The extent to whether (2), or $[L_\alpha, U_\alpha]$, provides an accurate reflection of the coverage of P_1 has not yet been a topic of discussion, although further work to investigate this aspect of the analysis can be made. However, Agresti and Coull (1998) review several procedures that are appropriate for the interval estimation of a binomial proportion and may lead to useful insights into this issue.

2. The Aggregate Association Index

2.1 The Original Index

In practice P_1 can not be estimated precisely and so the relative width of these intervals may only be used as an indication of the extent to which the two dichotomous variables may be associated. It must also be noted that the more statistically significant association structures will arise when P_1 lies at, or near, the boundaries of (1) – see, for example, Fig. 1. To overcome these issues, Beh (2008) proposed that, when only aggregate information is known, the extent to which the variables may be associated can be measured using the aggregate association index (AAI)

$$A_\alpha = 100\left(1 - \frac{\left[(L_\alpha - L_1) + (U_1 - U_\alpha)\right]\chi^2_\alpha + \int_{L_\alpha}^{U_\alpha} X^2(P_1 \mid p_{1\bullet}, p_{\bullet 1})dP_1}{\int_{L_1}^{U_1} X^2(P_1 \mid p_{1\bullet}, p_{\bullet 1})dP_1}\right) \tag{3}$$

where

$$X^2(P_1 \mid p_{1\bullet}, p_{\bullet 1}) = n\left(\frac{P_1 - p_{\bullet 1}}{p_{2\bullet}}\right)^2\left(\frac{p_{1\bullet}p_{2\bullet}}{p_{\bullet 1}p_{\bullet 2}}\right). \tag{4}$$

Equation (4) is the sample chi-squared statistic of the 2×2 contingency table as a function of P_1, when only the marginal information is known. For the AAI, (3), Beh (2008) treated P_1 as a continuous random variable and Fig. 1 provides a graphical representation of its meaning.

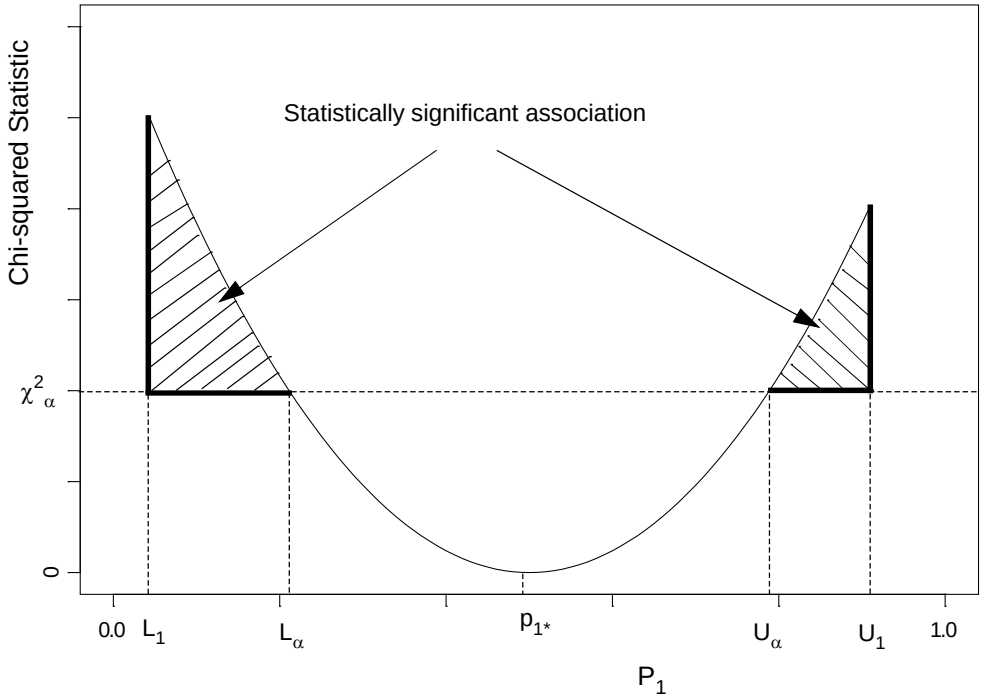

Fig. 1. Graphical perspective of the AAI (3). The shaded area indicates the magnitude of A_α

The index (3) is bounded by [0, 100]. It quantifies, for a given α, how likely a particular set of fixed marginal frequencies will enable the user to conclude that there exists a statistically significant association between the two dichotomous variables. A value of A_α close to zero indicates that there is virtually no information in the margins to suggest that an association might exist between the two variables. On the other hand, an index value close to 100 reflects that it is highly likely that such an association may exist. The choice of index thresholds to define when the marginal information infers that a statistically significant association may exist can be appropriately chosen. For the purposes of exploring the indices in this paper, an index at, or above, 75 will be considered to reflect that there is strong evidence to suggest that the variables may be statistically significantly associated. An index above 50 will highlight that it is more likely that a significant association may exist than not. We will consider that an association is very unlikely, given only the marginal information, if the index is below 25.

The justification for the index may be made by observing that (4) is a quadratic function in terms of P_1 and is maximised at the endpoints of the function and minimised at independence ($P_1 = p_{\bullet 1}$). This is also consistent with comments made in the ecological inference, and related, literature. For example, refer to Beh (2008) and Wakefield (2004).

We can simplify the aggregate association index of (3) by removing the integrals in the expression. After simplification,

$$\int X^2(P_1)dP_1 = kn(P_1 - p_{\bullet 1})^3 + c$$

where c is the constant of integration and $k = \frac{1}{3p_{2\bullet}^2}\left(\frac{p_{1\bullet}p_{2\bullet}}{p_{\bullet 1}p_{\bullet 2}}\right)$.

Thus, the definite integral on the denominator of (3) is $\int_{L_1}^{U_1} X^2(P_1)dP_1 = kn\left[(U_1 - p_{\bullet 1})^3 - (L_1 - p_{\bullet 1})^3\right]$. Similarly, $\int_{L_\alpha}^{U_\alpha} X^2(P_1)dP_1 = kn\left[(U_\alpha - p_{\bullet 1})^3 - (L_\alpha - p_{\bullet 1})^3\right]$ is the definite integral on the numerator of (3). Thus, A_α, as defined by (3), can be written in terms of the bounds of P_1 by

$$A_\alpha = 100\left(1 - \frac{\chi_\alpha^2\left[(L_\alpha - L_1) + (U_1 - U_\alpha)\right]}{kn\left[(U_1 - p_{\bullet 1})^3 - (L_1 - p_{\bullet 1})^3\right]} - \frac{\left[(U_\alpha - p_{\bullet 1})^3 - (L_\alpha - p_{\bullet 1})^3\right]}{\left[(U_1 - p_{\bullet 1})^3 - (L_1 - p_{\bullet 1})^3\right]}\right). \quad (5)$$

When $U_1 = U_\alpha$ and $L_1 = L_\alpha$ (so that the width of the confidence interval of P_1 is at its maximum), $A_\alpha = 0$. Also, in the rare case where $U_1 = L_1 \approx p_{\bullet 1}$ (such as when a very large sample of individuals/items exists) the aggregated data will always provide some information about the association structure of the variables since, in this case, $A_\alpha \approx 100$.

2.2 The Discrete Version of the Index

The AAI of (3) originally considered by Beh (2008), and its alternative derivation (5), assumes that P_1 is a continuous quantity. However, given a set of specific marginal frequencies there are a discrete number of values that n_{11}, and hence P_1, can take. Therefore the AAI can be considered in this context. For Fig. 1, rather than determining the area under the curve defined by the function (4) but above the critical value χ_α^2 using integration, one can instead consider determining the area of this region using a more simple approach involving rectangular regions, or bins – see Fig. 2. The

shaded region of this figure represents the proportion of the total region of interest that describes when an association exists between the two dichotomous variables given the presence of the marginal frequencies only.

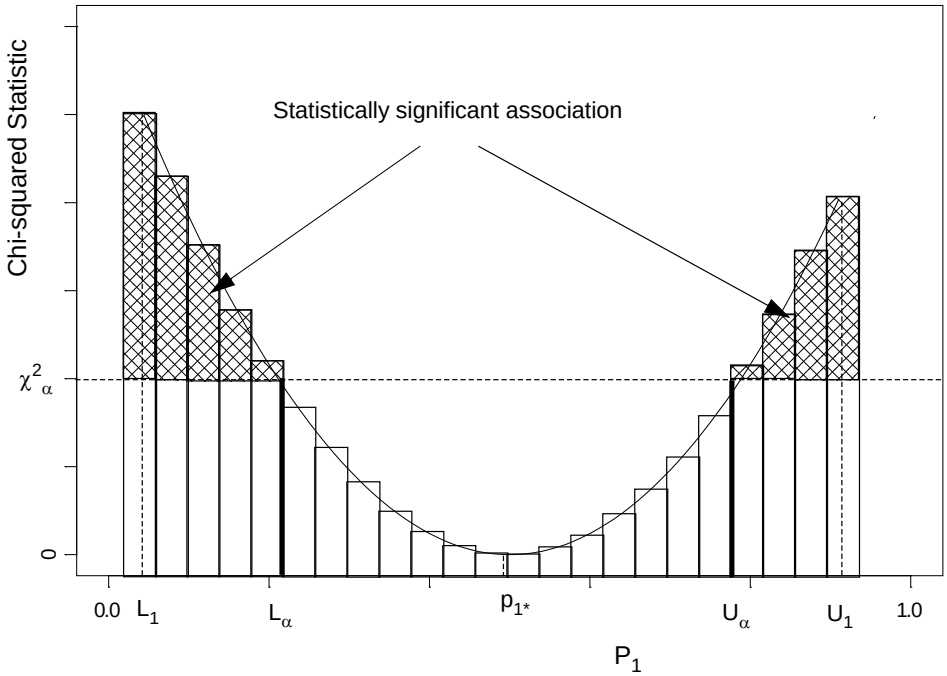

Fig. 2. Graphical perspective of the AAI (6), where $1+n_{1\bullet}(U_1-L_1)=23$.

The number of bins within the interval $[L_1, U_1]$ is $1+n_{1\bullet}(U_1-L_1)$. Therefore Fig. 2 considers an example were an analysis of an artificial contingency table results in $1+n_{1\bullet}(U_1-L_1)=23$ bins. Fig. 2 also highlights where L_α and U_α lie in relation to those bounds of Duncan and Davis (1953); equation (1). The discrete number of possible values P_1 can take given the marginal frequencies is equivalent to the number of possible values that n_{11} can have and is $1+n_{1\bullet}(U_1-L_1)$. Therefore, assuming that each of these bins is identical in width, the proportion each bin width represents in terms of the total width of the Duncan and Davis (1953) bounds (i.e. the bin width) is $\dfrac{U_1-L_1}{1+n_{1\bullet}(U_1-L_1)}$.

Suppose we now consider the height of each bin. Let the height of the i'th bin above the critical value, χ^2_α, be denoted by

$$D_\alpha(i)=\begin{cases} X_i^2(P_1)-\chi^2_\alpha, & \text{if } X_i^2(P_1)>\chi^2_\alpha \\ 0, & \text{if } X_i^2(P_1)<\chi^2_\alpha \end{cases}.$$

Therefore the total area of the $1+n_{1\bullet}(U_1-L_1)$ bins above the critical value (indicated by the shaded region of Fig. 2) is

$$\frac{U_1-L_1}{1+n_{1\bullet}(U_1-L_1)}\sum_{i=1}^{1+n_{1\bullet}(U_1-L_1)}D_\alpha(i).$$

Similarly, the total area of these bins is $\frac{U_1 - L_1}{1 + n_{1\bullet}(U_1 - L_1)} \sum_{i=1}^{1+n_{1\bullet}(U_1-L_1)} X_i^2(P_1)$. Thus, the discrete version of the aggregate association index, (3), is

$$A_{\alpha D} = \frac{\sum_{i=1}^{1+n_{1\bullet}(U_1-L_1)} D_\alpha(i)}{\sum_{i=1}^{1+n_{1\bullet}(U_1-L_1)} X_i^2(P_1)}, \qquad (6)$$

where $0 \le A_{\alpha D} \le 100$ for all α. When the number of possible discrete values of P_1 is small there can be quite a large difference between (6) and (5).

2.3 The Empirical Version of the Index

One may also compare the AAI with an empirical version of the index. Since the bounds of (1) specify the valid interval in which P_1 can lie, the interval $[n_{1\bullet}L_1, n_{1\bullet}U_1]$ indicates the range of values for which n_{11} is valid. Therefore, for each of the $1 + n_{1\bullet}(U_1 - L_1)$ values that n_{11} can take, the p-value associated with its Pearson chi-squared statistic can be calculated. Determining the proportion of those p-values that are less than the level of significance, α, will provide an indication of how likely a set of marginal frequencies will lead to a statistically significant association between the two dichotomous variables. Such a proportion is termed here the empirical version of the AAI and is denoted by $A_{\alpha E}$.

Since this index only measures the proportion of those possible 2×2 contingency tables where a statistically significant association exists it considers only the extent to which the variables may be associated, not the extent to which they are associated. In terms of Fig. 1 and Fig. 2, this index reflects the proportion of possible P_1 values whose chi-squared statistic exceeds the critical value, but will not reflect the area of the shaded regions. Thus such an index will not take into account that the more significant association structures will occur at, and near, the boundaries of P_1 (see Fig 1. and Fig. 2), only that an association exists. Therefore, generally, $A_{\alpha E} < A_\alpha$ and $A_{\alpha E} < A_{\alpha D}$, especially when the sample size is not deemed to be considered too small. In the case of small sample sizes (say, n < 200), or where a cell value is less than 10, the p-value from Fisher's exact test may be considered instead of the p-value from the Pearson chi-squared statistic. However we shall focus our attention on the p-value obtained from a chi-squared test of independence.

3. Examples

3.1 Surface Plot of the Indices

Consider a 2×2 contingency table with a sample of size n = 100. To observe the behaviour of the indices A_α, $A_{\alpha D}$ and $A_{\alpha E}$, surface plots are constructed for $n_{1\bullet}$ and $n_{\bullet 1}$ varying from 1 to 99 at increments of 1. Fig. 3, Fig. 4 and Fig. 5 are these plots for A_α, $A_{\alpha D}$ and $A_{\alpha E}$ respectively and were calculated with $\alpha = 0.05$.

Fig 3. shows that the aggregate association index of (5) is locally maximised along a saddlepoint defined when $n_{1\bullet} = n_{\bullet 1}$ and $n_{1\bullet} = n_{\bullet 2}$. Global maximums of the index exist when either $n_{1\bullet} = nL_1$ or $n_{1\bullet} = nU_1$. It also shows that the index reaches a minimum when $n_{1\bullet}$ lies close to the limits of $[nL_1, nU_1]$ and 20 < $n_{\bullet 1}$ < 80. Such results show that it is extremely difficult to determine

the likely values of the cells of the 2×2 table. However, since we are more concerned with the nature of the association here, the marginal frequencies prove to be useful.

Fig. 4. shows that the index $A_{0.05D}$ behaves in a very similar manner to $A_{0.05}$. In fact, Fig. 4 appears virtually indistinguishable when compared with Fig. 3. As we shall see in Section 3.2 this is because the sample size is relatively large. Thus, if we were to consider a plot similar to Fig 2. it would consist of many bins and provide a value of the index very similar to that of the continuous version.

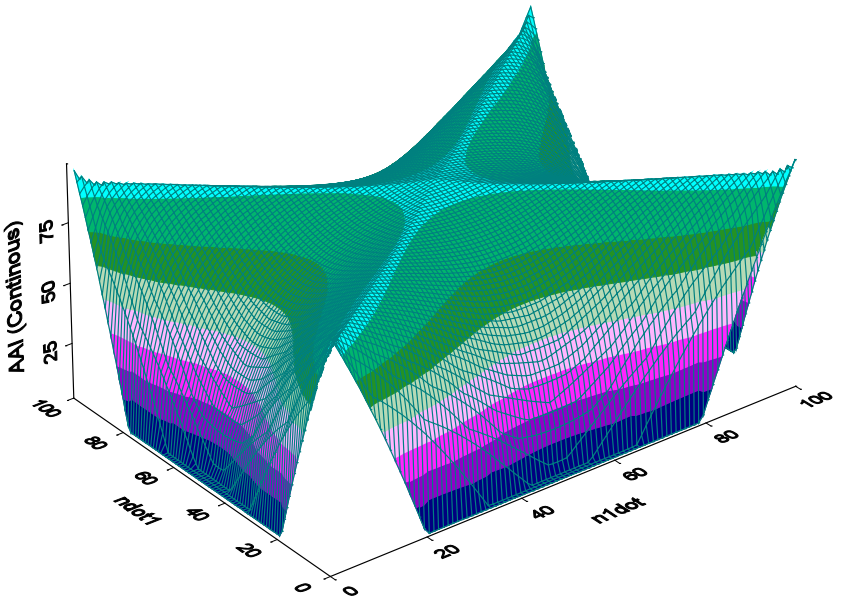

Fig. 3. Surface plot of $A_{0.05}$ for n = 100

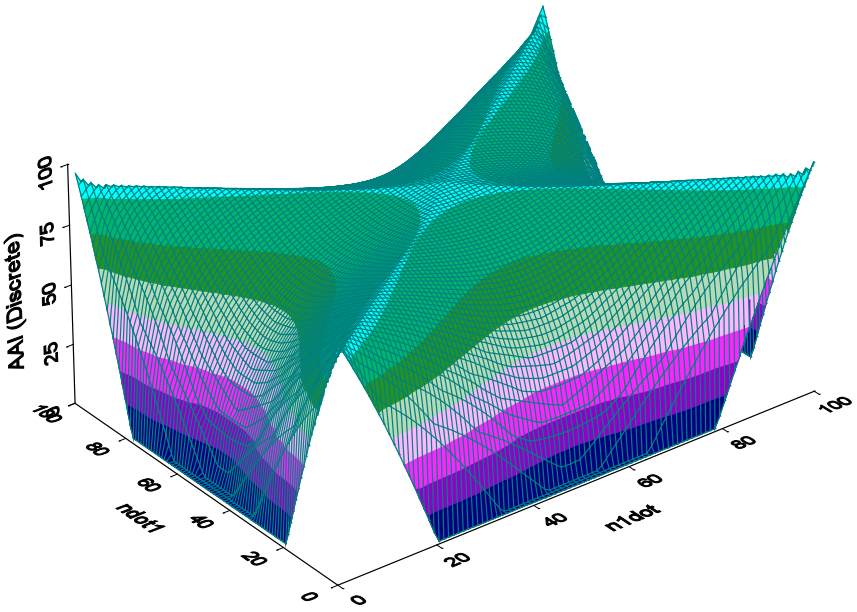

Fig. 4. Surface plot of $A_{0.05D}$ for n = 100

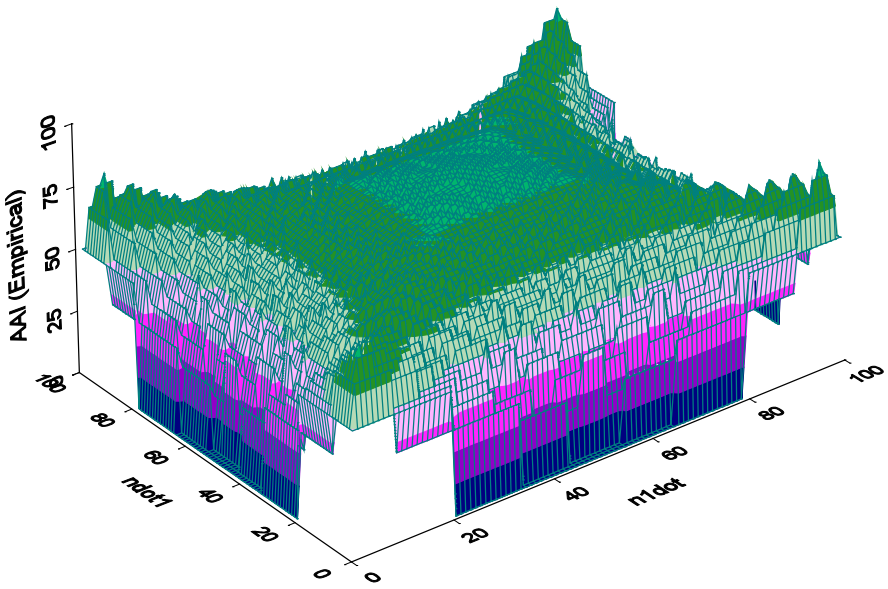

Fig. 5. Surface plot of $A_{0.05E}$ for n = 100

Fig 5. graphically shows the index $A_{0.05E}$ for these computed contingency tables. While the index is generally below 0.75 there appears to be some information in the margins for concluding that an association exists between the two dichotomous variables. Table 3 provides a summary of the $A_{0.05E}$ values for a selection of valid $n_{1\bullet}$ and $n_{\bullet 1}$ marginal frequencies. It shows that, generally there will be at least 50% of contingency tables that will exhibit a significant statistical association at the 0.05 level of significance. For those situations where such an association is not significant, they arise in the same situations as those defined above for small $A_{0.05}$ and $A_{0.05D}$ values. Table 2 provides a tabular summary of $A_{0.05D}$ for a selection of valid of $n_{1\bullet}$ and $n_{\bullet 1}$ values.

Table 2
Value of $A_{0.05D}$ for various $n_{1\bullet}$ and $n_{\bullet 1}$ and n = 100.

	$n_{\bullet 1}$										
$n_{1\bullet}$	1	10	20	30	40	50	60	70	80	90	99
1	96.15	57.04	4.62	0.00	0.00	0.00	0.00	0.00	4.62	57.04	96.15
10	57.04	90.60	79.34	65.22	50.59	43.76	50.59	65.22	79.34	90.60	57.04
20	4.62	79.34	89.14	81.13	72.54	68.43	72.54	81.13	89.14	79.34	4.62
30	0.00	65.22	81.13	87.97	82.39	79.69	82.39	87.97	81.13	65.22	0.00
40	0.00	50.59	72.54	82.39	87.99	86.11	87.99	82.39	72.54	50.59	0.00
50	0.00	43.76	68.43	79.69	86.11	90.33	86.11	79.69	68.43	43.76	0.00
60	0.00	50.59	72.54	82.39	87.99	86.11	87.99	82.39	72.54	50.59	0.00
70	0.00	65.22	81.13	87.97	82.39	79.69	82.39	87.97	81.13	65.22	0.00
80	4.62	79.34	89.14	81.13	72.54	68.43	72.54	81.13	89.14	79.34	4.62
90	57.04	90.60	79.34	65.22	50.59	43.76	50.59	65.22	79.34	90.60	57.04
99	96.15	57.04	4.62	0.00	0.00	0.00	0.00	0.00	4.62	57.04	96.15

Table 3
Value of $A_{0.05E}$ for various $n_{1\bullet}$ and $n_{\bullet 1}$ and n = 100.

	$n_{\bullet 1}$										
$n_{1\bullet}$	1	10	20	30	40	50	60	70	80	90	99
1	50.00	50.00	50.00	0.00	0.00	0.00	0.00	0.00	50.00	50.00	50.00
10	50.00	72.73	54.55	54.55	54.55	54.55	54.55	54.55	54.55	72.73	50.00
20	50.00	54.55	66.67	66.67	66.67	66.67	66.67	66.67	66.67	54.55	50.00
30	0.00	54.55	66.67	70.97	70.97	70.97	70.97	70.97	66.67	54.55	0.00
40	0.00	54.55	66.67	70.97	78.05	78.05	78.05	70.97	66.67	54.55	0.00
50	0.00	54.55	66.67	70.97	78.05	82.35	78.05	70.97	66.67	54.55	0.00
60	0.00	54.55	66.67	70.97	78.05	78.05	78.05	70.97	66.67	54.55	0.00
70	0.00	54.55	66.67	70.97	70.97	70.97	70.97	70.97	66.67	54.55	0.00
80	50.00	54.55	66.67	66.67	66.67	66.67	66.67	66.67	66.67	54.55	50.00
90	50.00	72.73	54.55	54.55	54.55	54.55	54.55	54.55	54.55	72.73	50.00
99	50.00	50.00	50.00	0.00	0.00	0.00	0.00	0.00	50.00	50.00	50.00

3.2 Example – Fisher's (1935) Twin Data

Consider the 2×2 contingency table of Table 4. This table was considered by Fisher (1935) and used by Beh (2008) to illustrate a simple application of (3). Fisher's data studies 30 criminal twins and classifies them according to whether they are a monozygotic twin or dizygotic twin. The table also classifies whether their same sex twin has been convicted of a criminal offence. We shall, for now, overlook the problem surrounding the applicability of using the Pearson chi-squared statistic in cases where the cell frequencies are not greater than five. In such cases Yates continuity correction can be used. However, as we shall see, we will investigate the implications of the indices proposed here when the small sample size of Table 4 increases by a constant positive factor.

The Pearson chi-squared statistic for Table 4 is 13.032, and with a p-value of 0.0003, shows that there is a statistically significant association between the type of criminal twin and whether their same sex sibling has been convicted of a crime. For this data $P_1 = 10/13 = 0.7692$ and shows that about 77% of those monozygotic criminal twins in the sample have a same sex sibling who has also been convicted of a crime.

Table 4.
Fisher's (1935) same sex criminal twin data set

	Convicted	Not Convicted	Total
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

Suppose now that only the marginal information of Table 4 is known. The issue is to determine how likely it is that the two variables are associated with each other using only this information.

For the marginal frequencies of Table 4, the Pearson chi-squared statistic can be written as a function of P_1 such that

$$X^2(P_1) = \frac{221}{216}\left(\frac{30P_1 - 12}{17}\right)^2$$

This function is graphically depicted in Fig. 6. Based on the interval (1), P_1 lies within [0, 0.9231] and the shaded region of the figure indicates where there exists a statistically significant association between the two dichotomous variables of Table 4 at the 0.05 level of significance. Fig. 6 also shows that the Pearson chi-squared statistic is minimised at zero when $P_1 = p_{1\bullet} = 0.4$ which coincides with independence between the two variables. Note also that the statistic has two global maximums existing at the limits of the bounds of P_1 where the a absolute maximum is 26.1537 at $P_1 = 0.9231$.

To determine the extent to which the variables of Table 4 are associated (given only the marginal information), we shall calculate the area of the shaded region of Fig. 6, and consider the discrete version of the region, as well as the empirical version of the AAI. Beh (2008) determined that the AAI using (3), or alternatively (5), is $A_{0.05} = 61.83$. Similarly it was shown that the empirical version of the index is $A_{0.05E} = 61.54$ and compares very well with the original version of the index. By considering (6) the discrete AAI is $A_{0.05D} = 66.09$. Therefore, it is likely that a 2×2 contingency table with the marginal information structure of Table 4 will reflect a statistically significant association between the two dichotomous variables. Such results indicate that even for a relatively small sample size of 30, such an association is likely to exist since these indices are greater than 50.

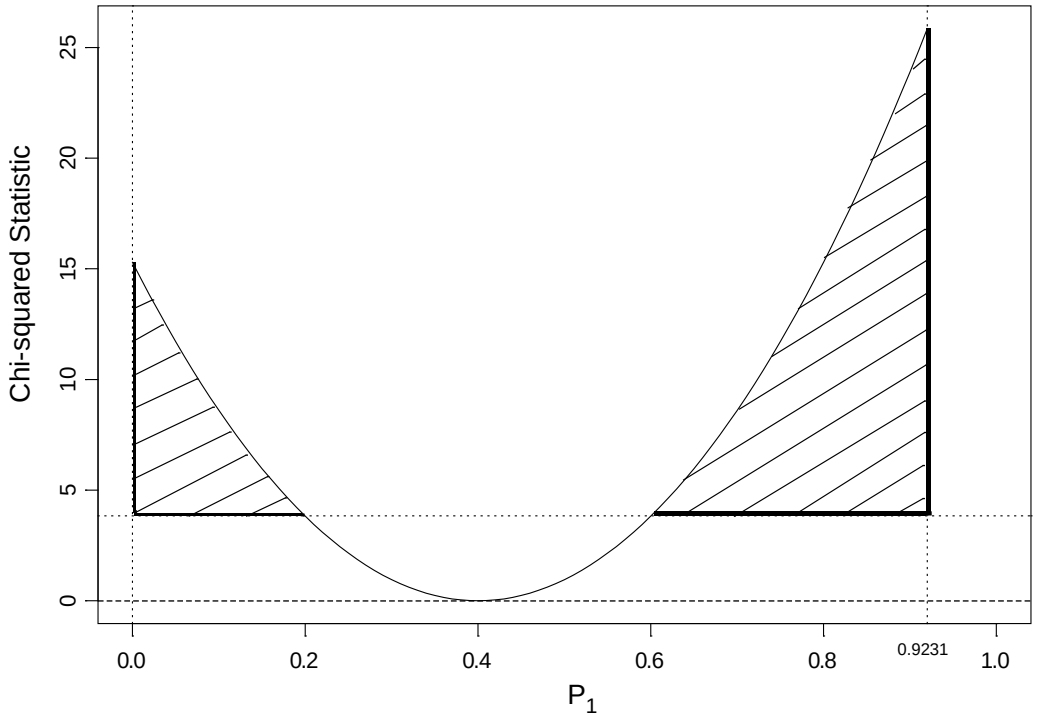

Fig. 6. Plot of $X^2(P_1)$ versus P_1 for Table 2

To investigate the behaviour of the three indices as the sample size increases, consider the case where we multiply each of the marginal frequencies by a positive integer C, and let C vary between 1

and 100 in increments of 1. Therefore we will consider what impact C has on the margins such that $(Cn_{1\bullet}, Cn_{2\bullet}, Cn_{\bullet 1}, Cn_{\bullet 2}) = C(13, 17, 12, 18)$.

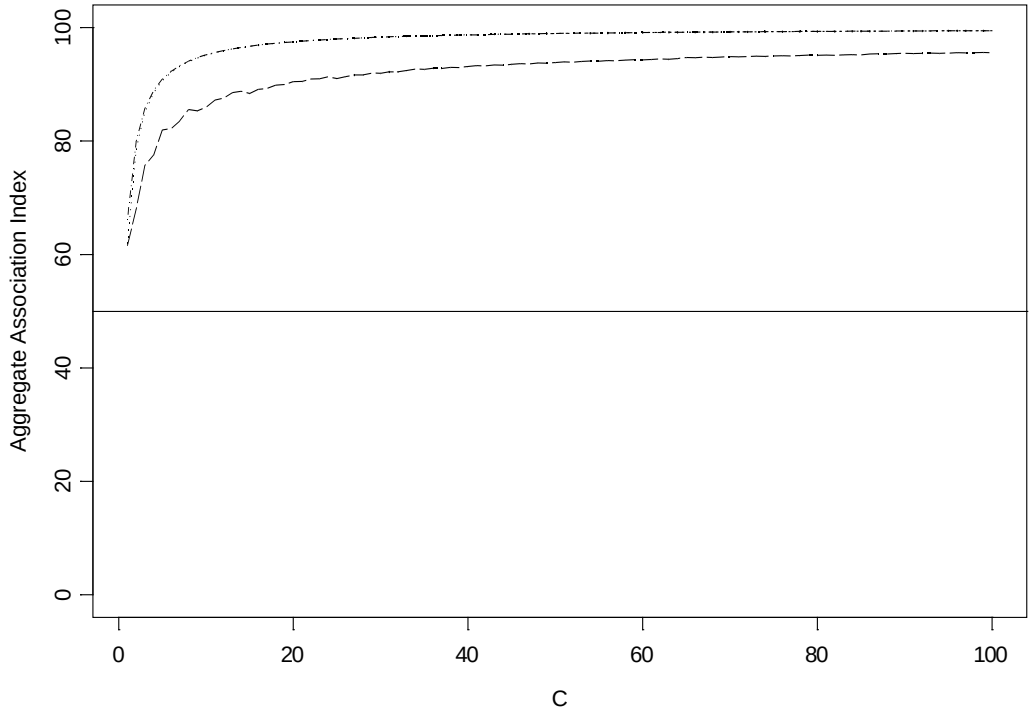

Fig. 7. Comparison of continuous, discrete and empirical versions of the AAI as C increases from 1 to 100 (Short dash line = Eqn (5); Dotted line = Eqn (6), Long dash line = empirical version)

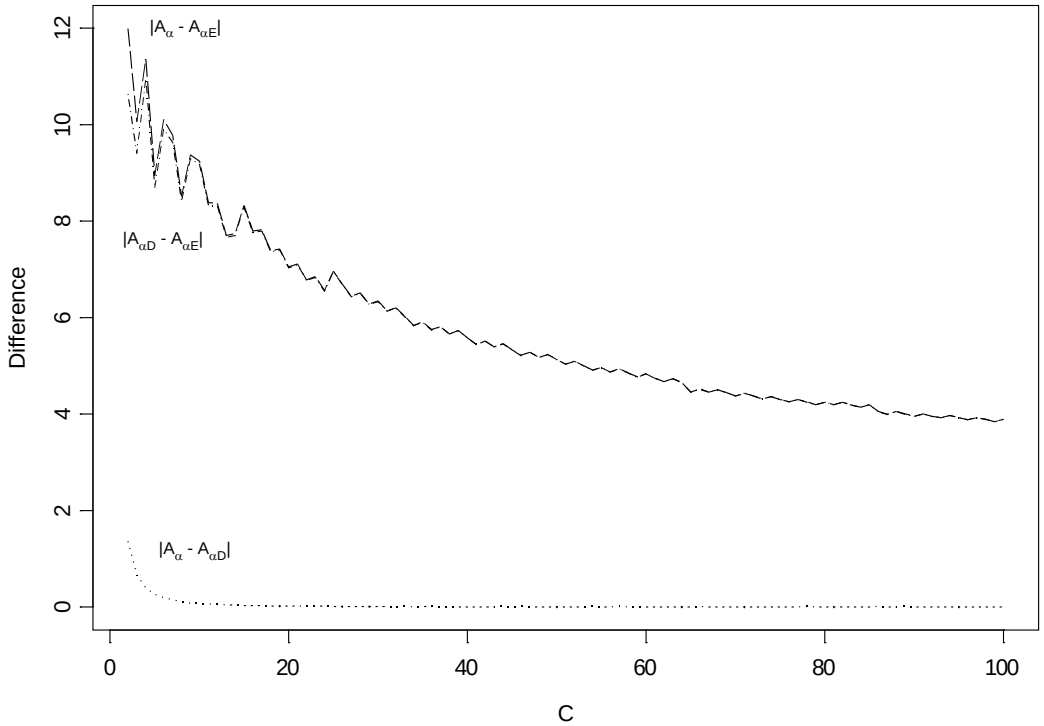

Fig. 8. Absolute Difference between the AAI measures

Fig. 7 shows that since the three indices are well above 75 for nearly all C considered it is highly likely that there will be a statistically significant association between the two dichotomous variables if the cell values were known. Even for relatively small sample sizes (where C is small) the difference between $A_{0.05}$ and $A_{0.05D}$ is practically zero. However, as expected, $A_{0.05E}$ is consistently smaller than its two counterparts, but shows that over 80% of contingency tables generated with this marginal information structure will lead to an association between the two dichotomous variables at the 5% level of significance. A comparison of the difference between each pair of indices is shown in Fig. 8 and reflects that for all multiples of the original sample size of n = 30 there is virtually a zero difference between A_{α} and $A_{\alpha D}$.

4 The Direction of the Association

4.1 Aggregate Positive and Negative Index

The aggregate association index allows one to quantify the extent to which the two dichotomous variables may be associated based only on the information provided by the marginal frequencies. However it also allows us to identify the direction of this association. By keeping the marginal frequencies fixed in a 2×2 table, some n_{11} (and hence P_1) will lead to a significantly positive association between the variables, while other P_1 values will lead to a significant negative association. For a single 2×2 contingency table Beh (2008) showed that the Pearson product moment correlation, ρ, can be expressed in terms of P_1 and the marginal information such that

$$\rho(P_1 \mid p_{\bullet 1}, p_{\bullet 1}) = \left(\frac{P_1 - p_{\bullet 1}}{p_{2\bullet}}\right)\sqrt{\frac{p_{1\bullet}p_{2\bullet}}{p_{\bullet 1}p_{\bullet 2}}}\,. \tag{7}$$

Since P_1 is bounded by (1), this correlation is bounded by the interval

$$-\min\left(\sqrt{\frac{p_{1\bullet}p_{\bullet 1}}{p_{2\bullet}p_{\bullet 2}}}, \sqrt{\frac{p_{2\bullet}p_{\bullet 2}}{p_{1\bullet}p_{\bullet 1}}}\right) \le \rho \le \min\left(\sqrt{\frac{p_{2\bullet}p_{\bullet 1}}{p_{1\bullet}p_{\bullet 2}}}, \sqrt{\frac{p_{1\bullet}p_{\bullet 2}}{p_{2\bullet}p_{\bullet 1}}}\right) \tag{8}$$

and was also considered by Duncan and Davis (1953). Since equation (7) is a linear function in terms of P_1 it may be expressed graphically as a straight line – see Fig. 9 - where the domain of the function (representing the valid P_1 values) is bounded by (1) and the range of the function is bounded by (8).

One may consider instead the aggregate association index (3). By comparing (7) with (4):

$$X^2(P_1 \mid p_{1\bullet}, p_{\bullet 1}) = n\rho^2(P_1 \mid p_{1\bullet}, p_{\bullet 1}).$$

Therefore one may determine, for a given α and known marginal frequencies, the extent to which there is a significant positive association by observing the shaded area on the right side of Fig. 1. By denoting this portion of A_{α} by A_{α}^{+}, we can consider

$$A_{\alpha}^{+} = \frac{\int_{U_{\alpha}}^{U_1}\left[X^2(P_1) - \chi_{\alpha}^2\right]dP_1}{\int_{L_1}^{U_1} X^2(P_1)dP_1} = \frac{kn\left[(U_1 - p_{\bullet 1})^3 - (U_{\alpha} - p_{\bullet 1})^3\right] - (U_1 - U_{\alpha})\chi_{\alpha}^2}{kn\left[(U_1 - p_{\bullet 1})^3 - (L_1 - p_{\bullet 1})^3\right]}$$

as an index that measures the extent to which the marginal information reflect a significant positive association. Here, A_α^+ is referred to as the *aggregate positive association index*. Similarly, we can consider

$$A_\alpha^- = \frac{\int_{L_1}^{L_\alpha}\left[X^2(P_1) - \chi_\alpha^2\right]dP_1}{\int_{L_1}^{U_1} X^2(P_1)dP_1} = \frac{kn\left[(L_\alpha - p_{\bullet 1})^3 - (L_1 - p_{\bullet 1})^3\right] - (L_\alpha - L_1)\chi_\alpha^2}{kn\left[(U_1 - p_{\bullet 1})^3 - (L_1 - p_{\bullet 1})^3\right]}$$

as an index that quantifies the extent to which the marginal information reflect a significant negative association between the two dichotomous variables. Here A_α^- is referred to as the *aggregate negative association index*. Note that $A_\alpha = A_\alpha^+ + A_\alpha^-$. Therefore, this allows us to partition the aggregate association index, A_α, to reflect significant positive association and a significant negative association index.

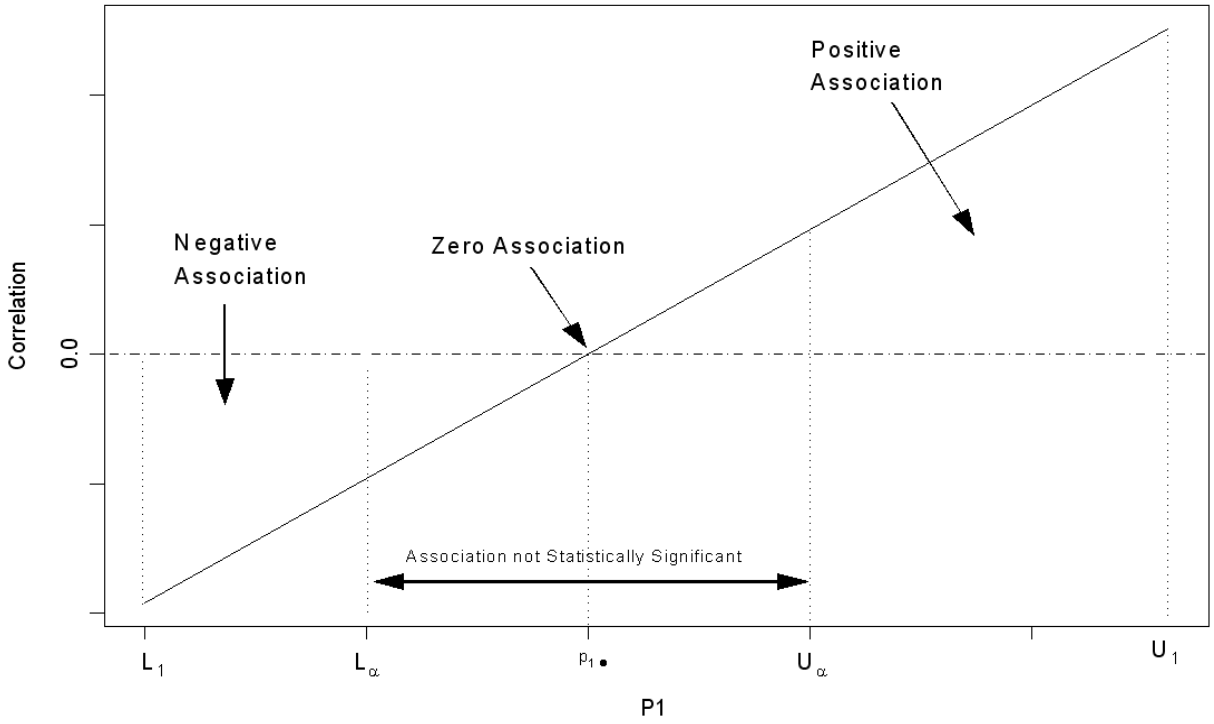

Fig. 9. Graphical representation of the correlation of two dichotomous variables when only the marginal frequencies are known.

Discrete versions of A_α^+ and A_α^- can also be obtained, although they will not be considered in this discussion.

4.1 Surface Plot of $A_{0.05}^+$

Consider again the contingency tables generated in Section 3.1 where n = 100. The aggregate positive association index, $A_{0.05}^+$, is plotted against $n_{1\bullet}$ and $n_{\bullet 1}$ in Fig. 10. Its shows that if $n_{1\bullet} < 50$ and $n_{\bullet 1} < 50$ then it is highly likely that the association between the two dichotomous variables will be positive. The symmetry of Fig. 10 about $n_{\bullet 1} = n_{2\bullet}$ also indicates that is the case if both $n_{1\bullet} > 50$ and $n_{\bullet 1} > 50$. In both cases, the strongest evidence that there exists a significant positive association

exists when $p_{1\bullet} \approx p_{\bullet 1} < 0.4$, or $p_{1\bullet} \approx p_{\bullet 1} > 0.6$. There is some evidence to suggest that a weak positive association will occur when $n_{1\bullet}$ and $n_{\bullet 1}$ lie on the limits of their permissible range. However in these cases Fig. 4 suggests that the marginal frequencies are not very informative in providing an indication of the nature of the association between the variables.

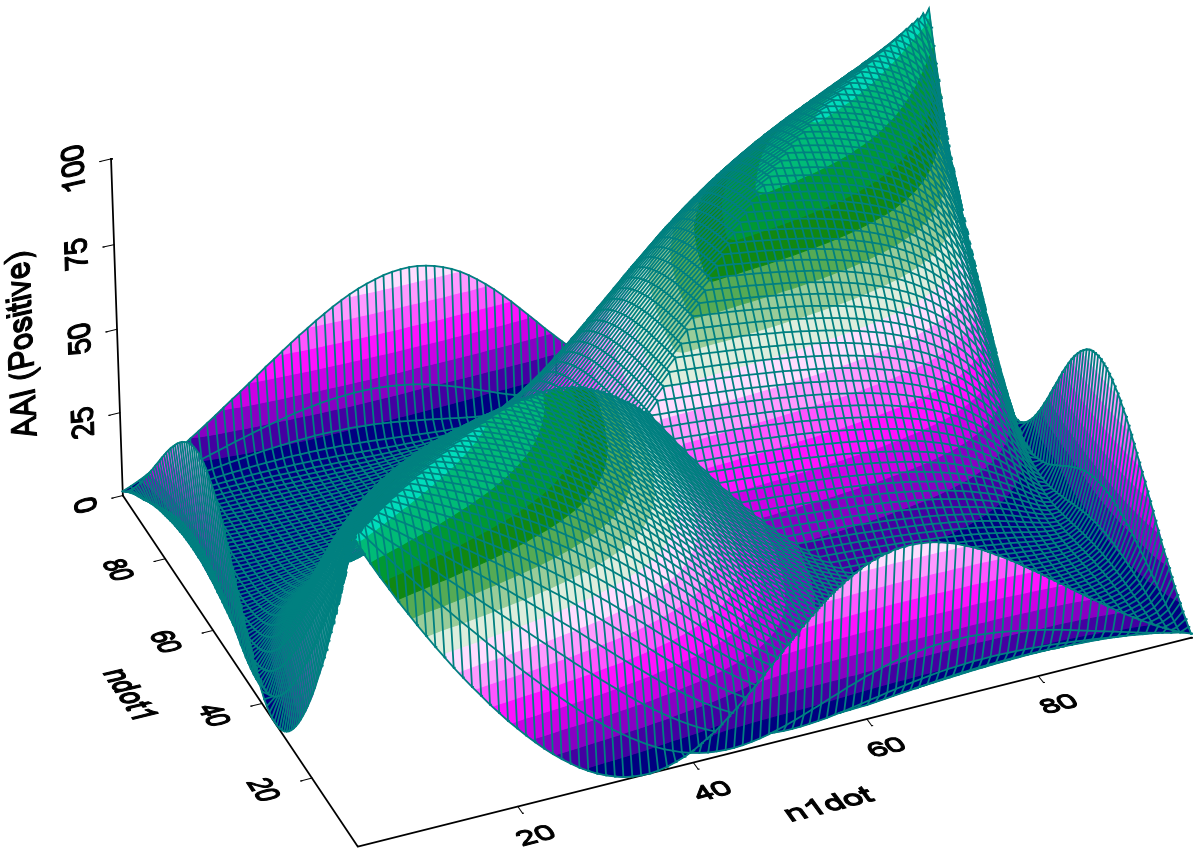

Fig. 10: Surface plot of $A^{+}_{0.05}$ for n = 100

4.2 Fisher's (1935) Data Revisited

Consider again Table 4. Recall that for this 2×2 contingency table $A_{0.05} = 61.83$ of which $A^{+}_{0.05} = 46.43$ and $A^{-}_{0.05} = 15.40$. Therefore based solely on the marginal information of Table 4 we can determine that the dichotomous variables are far more likely to be significantly positively associated than significantly negatively associated. This is not surprising keeping in mind the comments made in section 4.1, since $p_{1\bullet} = 13/30 = 0.43$ and $p_{\bullet 1} = 12/30 = 0.40$. If we assumed that there existed a significant association between the variables (since $A_{\alpha} = 61.83$) then the probability of this significant association being positive is 46.43/61.83 = 0.751

To observe the behaviour of the aggregate positive association index, $A^{+}_{0.05}$, and the aggregate negative association index, $A^{-}_{0.05}$, as the sample size increases by a factor of C (see Section 3.2) consider Fig 11. This figure shows that, just like $A_{0.05}$, the aggregate positive association index $A^{+}_{0.05}$ stabilises to approximately 68.8 as C increases. Thus, there is evidence to suggest that as the sample size increases there is a very good chance that the two dichotomous variables of Table 1 are significantly positively associated.

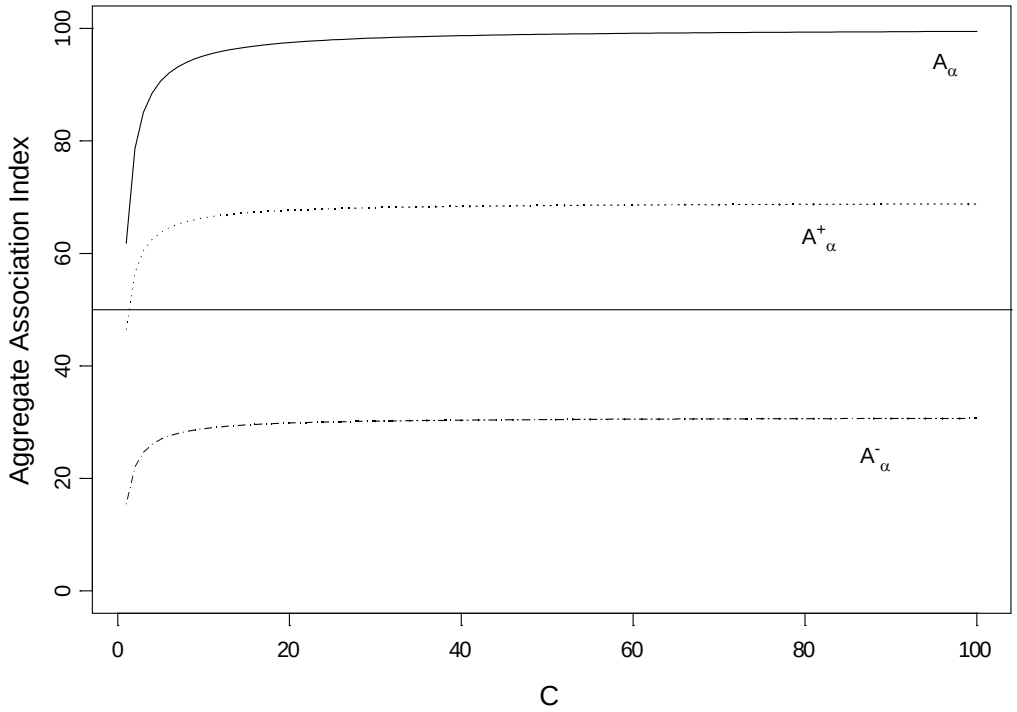

Fig. 11. $A^{+}_{0.05}$ and $A^{-}_{0.05}$ for Table 4 (as C increases from 1 to 100).

5. Discussion

This paper has elaborated further on the aggregate association index proposed by Beh (2008). Such an index provides an indication of the possibility that there exists a statistically significant association between two dichotomous variables given the presence of only the marginal frequencies. Where such an association exists, the direction (positive or negative) of the association can also be examined. However, such an index does not provide a means of inferring the value of the unknown cells. Therefore, the purpose of the indices is not to infer the individual level correlation of the variables, but instead to provide a measure reflecting how likely the two variables *may* be associated.

The issue of determining how much information the margins of a 2×2 table provide for inferring the cell values is a long standing problem R. A. Fisher grappled with in 1935. By considering the indices described in this paper the likely association structure can be determined by considering either the continuous, or discrete, version of the aggregate association index. In practice, the discrete version, $A_{\alpha D}$, is a more ideal measure of the likely association than A_α since $A_{\alpha D}$ takes into account that there are a discrete number of possible values that P_1 can have. However, as demonstrated in the examples, if the sample size is considered large, then there is a negligible difference between them.

An obvious next step to the development of the procedure outlined in this paper is to investigate the applicability of these indices for G (>1) 2×2 contingency tables. One option is to consider the use of the indices where each table is considered separately. Another strategy is to incorporate their use in ecological inference (eg King, 1997; Steel, Beh & Chambers, 2004; Wakefield, 2004). There have been proposals made in the ecological inference literature to incorporate additional (covariate) information to better estimate parameters that reflect individual level data – see, for example, King (1997), Chambers and Steel (2001), and Wakefield (2004). Considering covariate information, in conjunction with the marginal information, when calculating the AAI has the potential to lead to a better understanding of the association structure between two dichotomous variables. However further consideration of these issues is beyond the scope of this paper.

References

Agresti, A., Coull, B. A., 1998. Approximate is better than "Exact" for interval estimation of binomial proportions. *The American Statistician*. 52, 119 – 126.

Aitkin, M., Hinde, J. P., 1984. Comments to "Tests of significance for 2×2 contingency tables". *Journal of the Royal Statistical Association, Series A*. 47, 453–454.

Barnard, G. A., 1984. Comments to "Tests of significance for 2×2 contingency tables". *Journal of the Royal Statistical Society, Series A*. 47, 449–450.

Beh, E. J., 2008. Correspondence analysis of aggregate data: The 2×2 table. *Journal of Statistical Planning and Inference.138, 2941-2952.*

Chambers, R. L., Steel, D. G., 2001. Simple methods for ecological inference in 2×2 tables. *Journal of the Royal Statistical Society, Series A*. 164, 175–192.

Duncan, O.D., Davis, B., 1953. An alternative to ecological correlation. *American Sociological Review* 18, 665–666.

Fisher, R. A., 1935. The logic of inductive inference (with discussions). *Journal of the Royal Statistical Society, Series A*. 98, 39–82.

King, G., 1997. A Solution to the Ecological Inference Problem. Princeton University Press: Princeton, USA.

Plackett, R.L., 1977. The marginal totals of a 2×2 table. *Biometrika*. 64, 37–42.

Steel, D. G., Beh, E. J., Chambers, R. L., 2004. The information in aggregate data, in: King, G., Rosen, O., Tanner, M. (Eds), Ecological Inference: New Methodological Strategies, Cambridge University Press: New York, pp 51-68.

Wakefield, J. (2004). Ecological inference for 2×2 tables (with discussion), *Journal of the Royal Statistical Society, Series A*. 167, 385-424.