Adjusting the Aggregate Association Index for Large Samples

Eric, J. Beh[1], Salman, A. Cheema[1], Duy Tran[1], Irene, L. Hudson[1]

Abstract Recently, the aggregate association index (or AAI) was proposed to quantify the strength of the association between two dichotomous variables given only the marginal, or aggregate, data from a 2x2 contingency table. One feature of this index is that it is susceptible to changes in the sample size; as the sample size increases, so too does the AAI even when the relative distribution of the aggregate data remains unchanged. This paper proposes two adjustments to the AAI that help to overcome this problem. We consider a simple example using Fisher's twin criminal data to demonstrate the application of the AAI and its adjustments.

Keywords: 2x2 contingency tables, aggregate association index, large sample size, marginal information

1 Introduction

The analysis of aggregate information for the study of two dichotomous variables has a long history. Fisher (1935, page 48) was one of the first to consider such a study and invited us to "blot out the contents of the table, leaving only the marginal frequencies". In his discussion, Fisher (1935) concluded that the marginal information provided only "ancillary information" for inferring the missing cell values. Others to have discussed this issue include, but are not limited to Plackett (1977), Berkson (1978), Haber (1989) and Yates (1984).

Related to the issue are those techniques concerned with the ecological inference of aggregate data. They involve the estimation of the cells (or some simple transformation of them) for stratified 2x2 contingency tables given only aggregate data. One may refer to, for example, Goodman (1953), Freedman et al (1991),

[1] School of Mathematical and Physical Sciences, Newcastle University

King (1997), Wakefield (2004) and Steel, Beh and Chambers (2004) for detailed strategies for making such inferences. Hudson, Moore, Beh and Steel (2010) demonstrated the effectiveness of a variety of strategies for performing ecological inference by considering early New Zealand gendered election data.

The fundamental problem of all of these techniques is that, since the cells of the contingency table are unknown, numerous assumptions (all of which are untestable) need to be made about their structure. As an alternative approach to analysing aggregate data, involves the aggregate association index, or AAI, proposed by Beh (2008, 2010). Underlying the theory of the AAI is Pearson's chi-squared statistic. Therefore, rather than estimating the cells (or some function of them) of multiple 2x2 contingency tables, the purpose of the AAI is to quantify the likelihood that a statistically significant association exists between the two dichotomous variables. Unlike the numerous ecological inference techniques that are available, the AAI is applicable to the analysis of a single 2x2 table.

One feature of the AAI is that, as one considers an increase in the sample size, the AAI increases. This is because Pearson's chi-squared statistic is susceptible to changes in the sample size of the contingency table. This, therefore, can lead to the true nature of the association between the variables being masked by the magnitude of the sample size. Therefore, this paper explores two adjustments to the AAI that reduce the impact of the sample size on the index, when the relative marginal frequencies remain constant. A simple empirical study of the AAI and its adjustments will be considered through the analysis of Fisher's (1935) twin criminal data which motivated his discussion of the analysis of aggregate data.

2 The Aggregate Association Index

2.1 Notation

Table 1 gives the general form of a 2x2 contingency table, where two dichotomous variables are cross classified. Suppose the original sample size of the table is n_0 and n_{ij} denotes the (i,j)th cell frequency. Therefore, let $p_{ij} = n_{ij}/n_0$ be the proportion of classifications made into this cell for $i = 1,2$ and $j = 1,2$. The ith row jth column marginal frequencies are denoted by $n_{i.} = \sum_{j=1}^{2} n_{ij}$ and $n_{.j} = \sum_{i=1}^{2} n_{ij}$ respectively so that $\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} = \sum_{i=1}^{2} n_{i.} = \sum_{j=1}^{2} n_{.j} = n_0$. Thus, let $p_{i.} = n_{i.}/n_0$ and $p_{.j} = n_{.j}/n_0$ be the *i*th row marginal and *j*th column marginal proportions respectively.

Table 1: *A general 2x2 contingency table*

	Column 1	*Column 2*	*Total*
Row 1	n_{11}	n_{12}	$n_{1.}$
Row 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n_0

When cell frequencies of Table 1 are unknown, so that only the aggregate data is available, Duncan and Davis (1953) considered the upper and lower bounds of n_{11}

$$A_1 = \max(0, n_{.1} - n_{2.}) \leq n_{11} \leq \min(n_{.1}, n_{1.}) = B_1$$

Rather than considering n_{11}, much of the attention given to the ecological inference techniques focuses on the conditional proportion $P_1 = n_{11}/n_{1.}$. Therefore, P_1 is bounded by

$$L_1 = \max\left(0, \frac{p_{.1} - p_{2.}}{p_{1.}}\right) \leq P_1 \leq \min\left(\frac{p_{.1}}{p_{1.}}, 1\right) = U_1. \quad (1)$$

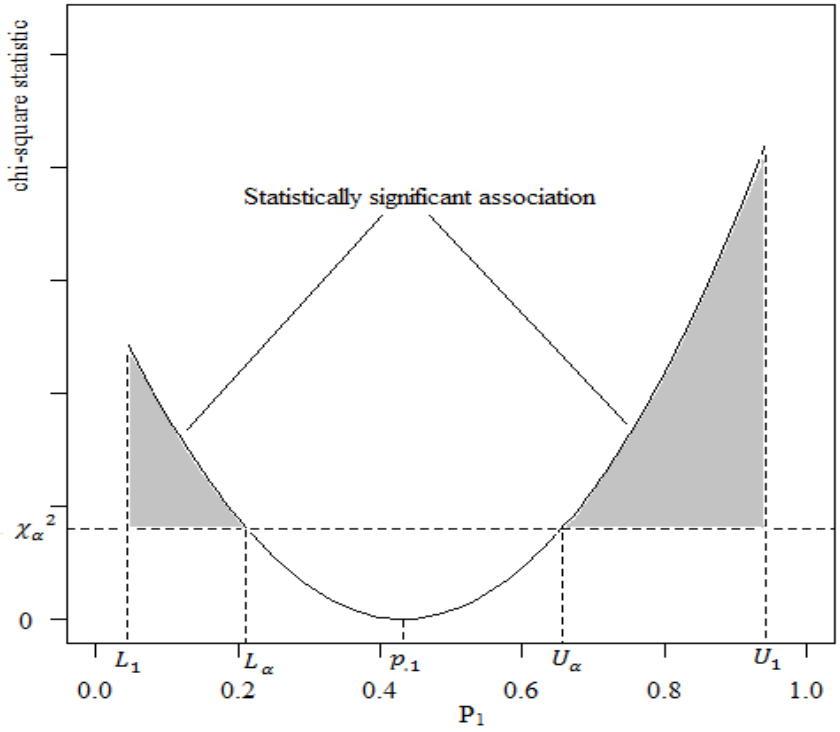
Beh (2010) showed that when only marginal information is available, and a test of the association is made at the α level of significance, the bounds of P_1 are

$$L_\alpha(n_0) = max\left(0, p_{.1} - p_{2.}\sqrt{\frac{\chi^2_\alpha}{n_0}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)}\right) < P_1 < min\left(1, p_{.1} + p_{2.}\sqrt{\frac{\chi^2_\alpha}{n_0}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)}\right) = U_\alpha(n_0) \quad (2)$$

where χ^2_α is 1 – α percentile of the chi-squared distribution with 1 degree of freedom. Since this paper considers the case where each of the cell frequencies of Table 1 is unknown, the proportion of interest, P_1, is therefore also unknown. Despite this, Beh (2008, 2010) demonstrated that Pearson's chi-squared statistic of Table 1 can be expressed as a quadratic function of this proportion such that

$$X^2(P_1|p_{1.}, p_{.1}) = n_0\left(\frac{P_1 - p_{.1}}{p_{2.}}\right)^2\left(\frac{p_{.1}p_{2.}}{p_{.1}p_{.2}}\right) \quad (3)$$

Figure 1: *A general graphical display of the AAI*



2.2 The Index

Figure 1 provides a graphical representation of the relationship between Pearson's statistic, (3), and the bounds of (1) and (2); note that U_α and L_α in the figure refer to the extremes of (2). The null hypothesis of independence between the dichotomous variables is rejected when the observed Pearson chi-squared value (at some value of P_1) exceeds the critical value of χ^2_α. Therefore the region under the curved defined by (3) but lying above the critical value χ^2_α, indicates where a statistically significant association exists between the variables. The relative size of this region, when compared with the total area under the curve, is quantified by

$$A_\alpha = 100\left[1 - \frac{\chi^2_\alpha\{(L_\alpha(n_0)-L_1)+(U_1-U_\alpha(n_0))\}}{kn_0\{(U_1-p_{\cdot 1})^3-(L_1-p_{\cdot 1})^3\}} - \frac{\{(U_\alpha(n_0)-p_{\cdot 1})^3-(L_\alpha(n_0)-p_{\cdot 1})^3\}}{\{(U_1-p_{\cdot 1})^3-(L_1-p_{\cdot 1})^3\}}\right] \quad (4)$$

where $k = \frac{1}{3p_{2\cdot}^2}\left(\frac{p_{1\cdot}p_{2\cdot}}{p_{\cdot 1}p_{\cdot 2}}\right)$ and $0 \le A_\alpha \le 100$; see Beh (2010). Equation (4) is referred to as the aggregate association index, or more simply the AAI of Table 1, and 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 variables. If $A_\alpha \approx 100$ then, given only the aggregate data, it is highly likely that an association exists. However, $A_\alpha \approx 0$ reflects that it is highly unlikely that such an association exists.

3 Adjusted Aggregate Association Index

Equation (3) shows that the magnitude of Pearson's chi-squared statistic is highly dependent on the sample size, n_0. For example, if the original sample size of Table 1 is increased by a factor of $C > 1$ so that $n = Cn_0$, then Pearson's statistic increases by a factor of C. This has been long understood and prompted Pearson to consider his phi-squared statistic. Everitt (1977, pp. 56), and many others, also discussed this feature of the statistic. As we shall now discuss, and propose a remedy to, the magnitude of the AAI is very much affected by the magnitude by which the sample size is increased, C.

Consider equation (4). It may be alternatively expressed as

$$A_\alpha = 100\left[1 - f_{(n_0)}\left(\frac{U_1-L_1}{U_\alpha(n_0)-L_\alpha(n_0)}\right)\times \left\{\frac{\chi^2_\alpha\{(L_\alpha(n_0)-L_1)+(U_1-U_\alpha(n_0))\}}{kn_0\{(U_1-p_{\cdot 1})^3-(L_1-p_{\cdot 1})^3\}} - \frac{\{(U_\alpha(n_0)-p_{\cdot 1})^3-(L_\alpha(n_0)-p_{\cdot 1})^3\}}{(U_1-p_{\cdot 1})^3-(L_1-p_{\cdot 1})^3}\right\}\right] \quad (5)$$

where

$$f_{(n_0)} = \frac{U_\alpha(n_0)-L_\alpha(n_0)}{U_1-L_1} \quad (6)$$

Suppose the level of significance, α, at which a test of independence is made remains fixed, as does the relative marginal proportions for the row and column categories. Multiplying the sample size by a $C > 1$ does not change the relative marginal information. However, it does impact on the sample size and the row and column totals of the contingency table. Increasing the original sample size of Table 1, n_0, by multiplying it by $C > 1$ will result a new sample size $n = Cn_0$ and increase Pearson's chi-squared statistic.

Therefore, this narrows the interval (2) and decreases the magnitude of $f_{(Cn_0)}$. This therefore leads to an increase in the AAI, even though the relative marginal information remains unchanged. Specifically, as $C \to \infty$, $f_{(Cn_0)} \to 0^+$, and $A_\alpha \to 100^-$. Similarly, as $C \to 0^+$, $f_{(Cn_0)} \to 1^+$, and $A_\alpha \to 0^+$. Therefore, to help minimise the impact that increasing the sample size has on the AAI, we shall adjust its calculation by considering various specifications of $f_{(n_0)}$, subject to $0 \leq f_{(n_0)} \leq 1$, that may be considered as an alternative to (6). This adjusted AAI is denoted by

$$A'_\alpha = 100\left[1 - f'_{(n_0)}\left(\frac{U_1 - L_1}{U_\alpha(n) - L_\alpha(n)}\right)\left\{\frac{\chi^2_\alpha\{(L_\alpha(n) - L_1) + (U_1 - U_\alpha(n))\}}{kn_0\{(U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\}} - \frac{\{(U_\alpha(n) - p_{.1})^3 - (L_\alpha(n) - p_{.1})^3\}}{(U_1 - p_{.1})^3 - (L_1 - p_{.1})^3}\right\}\right] \quad (7)$$

where $f'_{(n_0)}$ is the adjustment of (6) and may be subjectively, or objectively, determined so that $0 \leq f'_{(n_0)} \leq 1$. Here we shall consider two simple adjustments.

Adjustment 1

The first adjustment is to consider a subjective choice of $f_{(n_0)}$ that remains constant for all C. A conservative value, and one that is used in the following section, is $f'_{(n_0)} = 0.5$.

Adjustment 2

One may note that, from (2), $U_\alpha(n_0) - L_\alpha(n_0) = 2p_{2.}\sqrt{\frac{\chi^2_\alpha}{n_0}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)}$. Therefore, a second adjustment to (6) is to consider

$$f'_{(n_0)} = \frac{2p_{2.}}{U_1 - L_1}\sqrt{\frac{\chi^2_\alpha}{n_0}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)}$$

Since this adjustment is equivalent to (6), the adjusted AAI, (7), is

$$A'_\alpha = 100\left[1 - \sqrt{\frac{n}{n_0}}\left\{\frac{\chi^2_\alpha\{(L_\alpha - L_1) + (U_1 - U_\alpha)\}}{kn\{(U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\}} - \frac{\{(U_\alpha - p_{.1})^3 - (L_\alpha - p_{.1})^3\}}{\{(U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\}}\right\}\right] \quad (8)$$

Hence, under this adjustment, the relationship between the original AAI, (4), and its adjusted index, (8), is

$$A'_\alpha = A_\alpha\sqrt{\frac{n}{n_0}} - 100\left(\sqrt{\frac{n}{n_0}} - 1\right)$$

Thus, if the original sample size is increased by a factor of C, where $C > 1$, so that $n = Cn_0$, then

$$A'_\alpha = A_\alpha\sqrt{C} - 100\left(\sqrt{C} - 1\right) \quad (9)$$

Hence, $A'_\alpha < A_\alpha$ for any reasonably large C.

4 Empirical Study

Consider Table 1 that was originally studied by Fisher (1935). It cross-classifies 30 criminal twins according to whether they are a monzygotic twin or a dizygotic twin. The table also classifies whether their same sex twin has been convicted of a criminal offence. Beh (2010) analyses Table 1 using the original AAI, A_α, and here we consider the adjusted index, A'_α, (7). Pearson's chi-squared statistic for Table 1 is 13.032, and with a p-value of 0.0003, shows that there is a statistically significant association between the two dichotomous variables. 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. Fisher (1935, page 48) considered the case where the reader was invited to "blot out the contents of the table, leaving only the marginal frequencies". Doing so, when testing the association at the 5% level of significance, the AAI, calculated from (4), is 61.83. Therefore, based only on the analysis of the aggregate data of Table 2, it is likely that there exists a statistically significant association between the variables.

Table 1: *Fisher's (1935) criminal twin data*

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

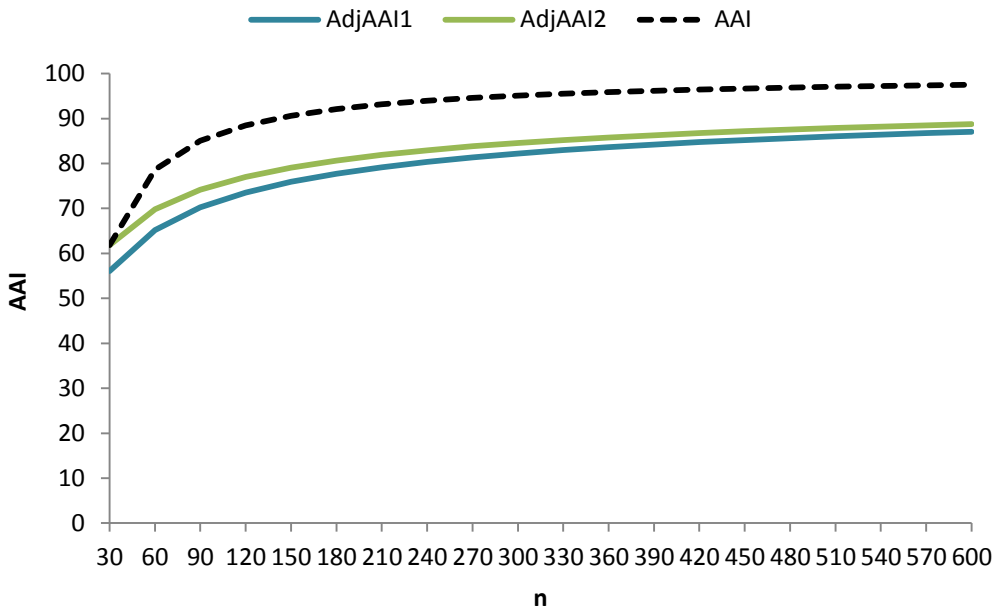
Suppose we now consider the case where the larger samples were selected, but retaining the same distribution of the marginal proportions. For example, doubling the sample size leads to twice the Pearson chi-squared statistic of the original data. Hence, the AAI increases. Figure 2 graphically shows the impact of the AAI as n increases for C ranging from 1 to 20; that is, considering a sample size ranging from 30 to 600. When $n = 600$, $A_{0.05} = 97.49$, indicating that it is now extremely likely that an association exists between the variables of Table 1 (given only the aggregate data).

Our aim is therefore to stabilise the magnitude of the AAI as the sample size increases. This will allow us to obtain a clearer indication of the nature of the association by reducing the impact of the magnitude of n, and can be done by considering the two adjusted AAI's. With the original sample size, the first adjustment – $f'_{(n)} = 0.5 - A'_{0.05} = 56.06$. For the second adjustment, using equation (8), as expected $A'_{0.05} = 61.83$; identical to the original AAI. As the sample size increase, these adjusted versions of A_α do increase, but more slowly than A_α. At $n = 600$ ($C = 20$), the first adjustment yields $A'_{0.05} = 87.06$, while the second adjustment leads to $A'_{0.05} = 88.77$. Note that, this second adjusted AAI can be obtained by considering equation (9):

$$A'_\alpha = 97.49\sqrt{20} - 100\left(\sqrt{20} - 1\right) = 88.77$$

Figure 2 shows that the rate of change of both $A'_{0.05}$ indices more stable, and less than, with the original magnitude of the AAI as C increases from 1 to 20.

Figure 2: *Comparison of $A'_{0.05}$ using the first adjustment (blue line), and the second adjustment (green line) with the original AAI (dashed line) as n increases.*



5 Discussion

In this article we have presented two adjustments that aim to minimize the impact that the sample size of a 2x2 contingency table has on the AAI when analysing the association between the variables using only the aggregate information. Both adjustments do not inflate the magnitude of the index as severely as the sample size does on the original index. Of the two proposed adjustments, the simplicity of the first is very appealing. Based on the empirical study presented, it performs just as well as equation (9). However, equation (9) is consistent with A_α at the original sample size and, as Figure 2 of our empirical study shows, appears more consistent as C increases. This study provides only an introduction to possible adjustments of the AAI that stabilise the impact of the sample size. More comprehensive research still needs to be undertaken to reveal the features of these, and other, adjustments. For example, one area of that requires further investigation is the identification $f'_{(n_0)}$ that minimises the rate of change of A'_α, and hence providing a more stable index, as the sample size increases.

References

1. Beh, E.J.: Correspondence analysis of aggregate data: The 2x2 table, Journal of Statistical Planning and Inference, 138, 2941 – 2952 (2008)
2. Beh, E.J.: The aggregate association index. Computational Statistics and Data Analysis, 54, 1570 – 1580 (2010)
3. Berkson, J.: In dispraise of the exact test: Do the marginal totals of the 2x2 table contain relevant information respecting the table proportion. Journal of Statistical Planning and Inference, 2, 27 – 42 (1978)
4. Duncan, O.D. and Davis, B.: An alternative to ecological correlation. American Sociological Review, 18, 665 – 666 (1953)
5. Everitt, B.S.: The Analysis of Contingency Tables. Wiley: New York (1977)
6. Fisher, R.A.: The logic of inductive inference (with discussion). Journal of Royal Statistical Association, Series A, 98, 39 – 82 (1935)
7. Freedman, D.A., Klein, S.P., Sacks, J, Smyth, C.A., Everett, C.G.: Ecological regression and voting rights. Evaluation review. 15, 673-711 (1991)
8. Goodman, L.: Ecological regressions and the behavior of individuals. American Sociological Review, 18, 663 – 666 (1953)
9. Haber, M.: Do the marginal total of a 2x2 contingency table contain information regarding the table proportion. Communication in Statistics: Theory and Methods, 18, 147 – 156 (1989)
10. Hudson, I.L., Moore, L., Beh, E.J., Steel, D.G.: Ecological inference techniques: an empirical evaluation using data describing gender and voter turnout at New Zealand elections 1893-1919. Journal of the Royal Statistical Society: Series A, vol. 173, 185-213 (2010)
11. King, G.: A Solution to Ecological Inference Problem. Princeton University Press: Princeton, U.S.A. (1997)
12. Plackett, R.L.: The marginal totals of a 2x2 table. Biometrika, 64, 37 – 42 (1977)
13. Steel, D.G., Beh, E.J., Chambers, R.L.: The information in aggregate data. In: King, G., Rosen, O., Tanner, M. (eds) Ecological Inference: New Methodological Strategies, pp. 51-68. Cambridge University Press (2004)
14. Wakefield, J.: Ecological inference for 2x2 tables. Journal of Royal Statistical Society, Series A, 167, 385 – 445 (2004)
15. Yates, F.: Tests of significance for 2x2 contingency tables (with discussion). Journal of Royal Statistical Society, Series A, 147, 426 – 463 (1984)