NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Beh, Eric J.; Tran, Duy; Hudson, Irene L. "A reformulation of the aggregate association index using the odds ratio" Computational Statistics and Data Analysis Vol. 68, p. 52-65 (2013)

Available from: <http://dx.doi.org/10.1177/1049732310385253>

Accessed from: <http://hdl.handle.net/1959.13/1054048>

A Reformulation of the Aggregate Association Index using the Odds Ratio

Eric J. Beh, Duy Tran, Irene L. Hudson

School of Mathematical and Physical Sciences, University of Newcastle, Australia

ABSTRACT

Since its inception in the 1950's the odds ratio has become one of the most simple and popular measures available for analysing the association between two dichotomous variables. Since the direction and magnitude of the association can be captured in such a simple measure, its impact has been felt throughout much of scientific research, in particular in epidemiology and clinical trials. Despite this, its applicability for analysing aggregate data has rarely been considered. In this paper we shall express a new measure of association (the aggregate association index, or AAI), in terms of the classic odds ratio. The advantage of doing so is that we are able to explore the use of the odds ratio in a context for which it was not originally intended, and that is for the analysis of a 2×2 table where only the aggregate data is known.

Keywords: 2×2 contingency table; Aggregate Association Index (AAI); Chi-squared Statistic; Fisher's criminal twin data; New Zealand Voting Data (1883 – 1919).

1 Introduction

The odds ratio remains one of the most simple, influential and diversely used measures of association available to the analyst. Due to the widespread use of logistic regression, the odds ratio is widely used in many fields of medical and social science research. It is commonly used in survey research, in epidemiology (Rothman, 2002; Rothman, Greenland and Lash, 2008) and to express the results of some clinical trials, such as in case-control studies (Miettinen, 1976). The odds ratio underpins the field of meta-analysis (Cheung, Ho, Lim and Mak, 2012). Meta analysis is a statistical method used to compare and combine effect sizes from a pool of relevant empirical studies. It is now a

standard approach to synthesize research findings in many disciplines, including medical and healthcare research, and climate change research (Hudson, 2010) and increasingly in genome-wide studies (Nakaoka and Inoue, 2009; Kraft *et al.*, 2009; Schurink, 2012) and drug discovery (Hudson, *et al.*, 2012). The odds ratio is often used as an alternative to the relative risk measure (Zhang and Yu, 1998; Montreuil, Bendavid and Brophy, 2005; Schmidt and Kohlmann, 2008; Viera, 2008) in many applications where it is important to measure the strength, and direction, of the association between two dichotomous variables from a 2×2 table. Despite its popularity, using the odds ratio in cases where only the margins of the 2×2 table are available has rarely been considered. One exception to this is Plackett (1977) who showed that the margins do not provide enough information to make inferences about the cell values. There are a host of techniques that lie within the ecological inference literature that one may consider for inferring cell values, or some function of them; none of them, however, consider the odds ratio. For example, King's (1997) groundbreaking parametric and non-parametric approaches may be considered. King (1997) also describes the ecological inference problem at length. Other strategies include Goodman's (1953) ecological regression, Freedman *et al.*'s (1991) neighbourhood model, Chamber and Steel's (2001) semi-parametric approach, Steel, Beh and Chambers (2004) homogeneous model and Wakefield's (2004) Bayesian extension of this model. A comprehensive review of these ecological inference techniques, and their application to early New Zealand gender and voter turnout data was given by Hudson, Moore, Beh and Steel (2010). Wakefield, Haneuse, Dobra and Teeple (2011) further discuss strategies for determining individual level information based only on aggregate data and Imai, Lu and Strauss (2011) provide an R package, *eco*, for performing ecological inference.

While these techniques all have their merits, they are all applicable only to the case where multiple, or stratified, 2×2 tables are simultaneously considered; they cannot be used for analysing the aggregate data of a single 2×2 table. They also involve the estimation of simple transformations of the cell frequencies, rather than the general association structure of the data. Therefore, all of these techniques are also subject to a variety of untestable assumptions which are problematic for evaluating their effectiveness. It is therefore appropriate to consider a strategy that does not rest on assumptions that are not testable while being applicable to a single 2×2 table. Thus, such a strategy should not estimate the (1, 1)'th cell frequency, or some transformation of it, but rather examine the structure of the association between two dichotomous variables based only on the marginal information. This

issue has a long history and dates back as far as Fisher (1935, page 48) who focused on the case where one may “blot out the contents of the table”. We shall consider this same issue but examine it by incorporating the classic odds ratio into a new index of association called the aggregate association index, also simply referred to as the AAI, proposed by Beh (2008, 2010). It shall be shown that considering the odds ratio offers a simple alternative to considering the AAI and whose calculation is as easy as that considered in Beh (2008). This will be achieved in the following 5 sections. Section 2 provides a review of Beh’s (2008, 2010) AAI and Section 3 demonstrates how the odds ratio can be incorporated into this measure. The direction of the association structure, when only the marginal information is available is explored in Section 4 while Section 5 considers the application of the odds ratio – AAI link using two data sets. The first is the classic twin data of Fisher (1935) and was considered by Beh (2008, 2010) in his development of the AAI. The second 2×2 table is based on the study conducted by Hudson, Moore, Beh and Steel (2010) and considers data obtained from the 1893 national election held in New Zealand. This data describes the gendered voting rates and is important because it was in this election that New Zealand became the first self-governing country in the world where women could vote at the national level. Some final comments are made in Section 6.

It must be pointed out that it is not the aim of this paper to formulate strategies for making inferences about the magnitude of the odds ratio given only the marginal information. Instead we shall use the properties of the odds ratio and the AAI, given only the marginal information of a 2×2 contingency table, to explore the association structure of the variables.

2. Aggregate Association Index

Consider a single two-way contingency table where both variables are dichotomous. Suppose that n individuals/units are classified into this table such that the number classified into the (i, j) th cell is denoted by n_{ij} and the proportion of those in this cell $p_{ij} = n_{ij} / n$ for $i = 1, 2$ and $j = 1, 2$. Denote the proportion of the sample classified into the i ’th row and j ’th column $p_{i\bullet} = p_{i1} + p_{i2}$ and $p_{\bullet j} = p_{1j} + p_{2j}$ respectively. Table 1 provides a description of the notation used in this paper.

Table 1:Notation for a 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

Typically, measuring the extent to which the row and column variables are associated is achieved by considering the Pearson chi-squared statistic calculated from the counts and margins of a contingency table. For a 2×2 table of the form described by Table 1, this statistic is

$$X^2 = n \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}.$$

The direction of the association may be determined by considering Pearson's (1900, page 12) estimate of his tetrachoric correlation. Such an estimate, and one of the most popular measures of correlation for 2×2 contingency tables due to its relative simplicity is

$$r = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}}$$

so that $X^2 = nr^2$.

Another very common measure of association for a 2×2 table is the odds ratio (Cornfield, 1951):

$$\theta = \frac{n_{11}n_{22}}{n_{21}n_{12}}. \quad (1)$$

We shall consider the odds ratio in more detail in Section 3.

Suppose, for now, that the cell values of Table 1 are known. Define $P_1 = n_{11} / n_{1\bullet}$; this is the conditional proportion of an individual/unit being classified into "Column 1" given that they are classified in "Row 1". For the analysis of marginal information of the 2×2 table, most of the ecological inference techniques consider this quantity as did Beh (2008, 2010) for his derivation of the aggregate association index.

When the cell values of Table 1 are unknown, it is not possible to calculate P_1 . However, the extremes of the permissible values of the (I, I) th cell frequency are well understood (Duncan and Davis, 1953) to lie within the interval

$$L_n = \max(0, n_{\bullet 1} - n_{2\bullet}) \leq n_{11} \leq \min(n_{\bullet 1}, n_{1\bullet}) = U_n. \quad (2)$$

Therefore, the bounds for P_1 are

$$L_p = \max\left(0, \frac{n_{\bullet 1} - n_{2\bullet}}{n_{1\bullet}}\right) \leq P_1 \leq \min\left(\frac{n_{\bullet 1}}{n_{1\bullet}}, 1\right) = U_p. \quad (3)$$

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 narrowed to

$$L_\alpha = \max\left(0, p_{\bullet 1} - p_{2\bullet} \sqrt{\frac{\chi_\alpha^2}{n} \left(\frac{p_{\bullet 1} p_{2\bullet}}{p_{1\bullet} p_{2\bullet}}\right)}\right) < P_1 < \min\left(1, p_{\bullet 1} + p_{2\bullet} \sqrt{\frac{\chi_\alpha^2}{n} \left(\frac{p_{\bullet 1} p_{2\bullet}}{p_{1\bullet} p_{2\bullet}}\right)}\right) = U_\alpha$$

where χ_α^2 is the $1 - \alpha$ percentile of the chi-squared distribution with 1 degree of freedom. If $L_\alpha < P_1 < U_\alpha$ then there is evidence that the row and column variables are independent at the α level of significance. However, if $L_p < P_1 < L_\alpha$ or $U_\alpha < P_1 < U_p$ then there is evidence to suggest that the variables are associated.

Within the interval (3), the magnitude of Pearson's chi-squared varies. Rather than consider whether a statistically significant association exists (or not) for all possible values of P_1 within this interval, the strength of the association may also be considered. By doing so, Beh (2008) proposed the following index to assess how likely a statistically significant association exists between the two dichotomous variables given only the aggregate data:

$$A_\alpha = 100 \left(1 - \frac{\chi_\alpha^2 [(L_\alpha - L_p) + (U_p - U_\alpha)] + \int_{L_\alpha}^{U_\alpha} X^2(P_1 | p_{1\bullet}, p_{\bullet 1}) dP_1}{\int_{L_p}^{U_p} X^2(P_1 | p_{1\bullet}, p_{\bullet 1}) dP_1} \right) \quad (4)$$

where

$$X^2(P_1 | p_{1\bullet}, p_{\bullet 1}) = n \left(\frac{P_1 - p_{\bullet 1}}{p_{2\bullet}} \right)^2 \left(\frac{p_{1\bullet} p_{2\bullet}}{p_{\bullet 1} p_{\bullet 2}} \right). \quad (5)$$

is (asymptotically) a chi-squared random variable with one degree of freedom. The index, A_α , is termed the aggregate association index (AAI). It quantifies for a given α how likely a particular set of fixed marginal frequencies will enable the analyst 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. Fig. 1 provides a visual description of the AAI; the magnitude of the AAI is reflected by the shaded area as a proportion of the total area under the curve.

Beh (2010) provided an alternative, and equivalent, expression of (4):

$$A_\alpha = 100 \left(1 - \frac{\chi_\alpha^2 [(L_\alpha - L_P) + (U_P - U_\alpha)]}{kn [(U_P - p_{\bullet 1})^3 - (L_P - p_{\bullet 1})^3]} - \frac{(U_\alpha - p_{\bullet 1})^3 - (L_\alpha - p_{\bullet 1})^3}{(U_P - p_{\bullet 1})^3 - (L_P - p_{\bullet 1})^3} \right)$$

where

$$k = \frac{1}{3p_{2\bullet}^2} \left(\frac{p_{1\bullet} p_{2\bullet}}{p_{\bullet 1} p_{\bullet 2}} \right).$$

Fig. 1. A graphical interpretation of the aggregate association index.

The shaded regions reflect the magnitude of the AAI.

Due to the popularity of the odds ratio for describing the association between two dichotomous variables, we shall now explore the AAI in terms this measure of association. Since the conceptual issues concerned with the interpretation of the AAI remain unchanged, so will A_α when expressed in terms of this ratio - only the parameterisation of the chi-squared statistic differs as does the relative level of difficulty of calculating the index. The concept of expressing the strength of the association in a 2×2 contingency table using the area under curve has been a topic of previous discussion. For example, Glas, Lijmer, Prins, Bonsel and Bossuyt (2003) discussed the area under the curve (AUC) of a receiver operating characteristic (ROC) function for use in clinical trials – the quantification of this area is based on a plot of true positive rates in a clinical trial against false positive rates. Their discussion of the AUC was made in terms of the odds ratio, as will our discussion of the AAI in the following section. However, unlike the AAI, the AUC relies on the cell values of the 2×2 table being known.

3. The AAI and the Odds Ratio

3.1. The Odds Ratio

One of the most common measures of association for a 2×2 contingency table is the odds ratio, (1), which may be alternatively written as

$$\theta = \frac{p_{11}p_{22}}{p_{21}p_{12}} = \frac{p_{11}\{p_{11} - (p_{1\cdot} + p_{\cdot 1} - 1)\}}{(p_{1\cdot} - p_{11})(p_{\cdot 1} - p_{11})}. \quad (6)$$

The odds ratio is a measure of effect size, describing the strength of association, or non-independence, between two dichotomous variables. It is commonly used as a descriptive statistic, and plays an important role in logistic regression. Unlike other measures of association, such as the relative risk, the odds ratio treats the two variables being compared symmetrically, and can be estimated using some types of non-random samples. This ratio, originally considered to have been first proposed by Cornfield (1951), was considered by Yule (1900, page 273). However, Yule did not explicitly state it in the above form. Rather, in his discussion of his “coefficient of association” (which is now better known as Yule’s Q), Yule (1900) considered the reciprocal of the odds ratio which he defined as κ . Other early

statistical studies of the odds ratio include, but are not limited to, Edwards (1963) and Mosteller (1968).

The sample odds ratio, (1), is easy to calculate, and for moderate and large samples, performs well as an estimator of the population odds ratio. Inferences of the population odds ratio may be made by considering Woolf's (1955) $100(1 - \alpha)\%$ confidence interval

$$\left(\ln(\theta) - Z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}, \ln(\theta) + Z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \right). \quad (7)$$

where $Z_{\alpha/2}$ is (asymptotically) a standard normal random variable.

There is an extensive amount of literature now available on the odds ratio and the coverage of the interval (7). For example, one may consider Mehta and Walsh (1992), Agresti (1999), Brown (1981) and Lloyd and Moldovan (2007a, b). More recently, Demidenko (2012) proposed a simple iterative algorithm that obtains an interval even more narrow than (7), but still maintaining the same coverage. While Demidenko (2012) focused his attention on the link that the odds ratio has with logistic regression, his contribution is easily amended for the analysis of a 2×2 table. Wilson and Langenberg (1999) also considered the coverage of the odds ratio from logistic regression. Alternative definitions of the odds ratio (7) including those of Woolf (1955), Haldane (1955) and Jewell (1986) may be considered – Walter and Cook (1991) conducted a comparative study of these with (1). Anscombe (1981, Chapter 12) and Molenberghs and Lesaffre (1994, pf 636) provided extensions of (1) for a $2 \times 2 \times 2$ table. There has been considerable attention given to the calculation of a single odds ratio for stratified 2×2 tables. In particular, one may consider those of Woolf (1955) and Mantel and Haenszel (1959) for common approaches and Hauck (1989) for an overview of these, and other, measures. Agresti and Coull (2002) discussed, amongst other things, four different measures of common odds ratio's for $I \times J$ ordered contingency tables and their link to log-linear and association models. Extensive discussions on the odds ratio, including the calculation of its standard error, derivation and coverage of its confidence issue, have been made by many including Hauck (1989). Excellent overviews of many of these issues may be found by referring to Bishop, Fienberg and Holland (1974), Agresti (2002) and Fleiss, Levin and Paik (2003).

3.2. Patnaik's Expectation

Placket (1977) demonstrated that, based only on the marginal frequencies of a 2×2 contingency table, there is not enough information available to infer the magnitude of the odds ratio. Since the underlying premise of the AAI is not to infer the magnitude of a measure of association, but rather to determine the strength to which an association might exist, we can tackle our problem by considering the odds ratio, and its natural logarithm, in the following manner.

Consider firstly the expected value, and variance, of n_{11} suggested by Patnaik (1948, pg 162):

$$E(n_{11}) = \frac{n_{1\bullet} n_{\bullet 1}}{n} + \text{Var}(n_{11}) \ln(\theta) \quad \text{Var}(n_{11}) = \frac{n_{1\bullet} n_{\bullet 1} n_{2\bullet} n_{\bullet 2}}{n^2(n-1)}. \quad (8)$$

Therefore, P_1 may be expressed as a function of the log-odds ratio given the marginal information, such that

$$P_1(\ln(\theta) | p_{1\bullet}, p_{\bullet 1}) = p_{\bullet 1} \left[1 + \left(\frac{n}{n-1} \right) p_{2\bullet} p_{\bullet 2} \ln(\theta) \right]. \quad (9)$$

Note that, for (9), when the row and column variables are independent (ie $\theta = 1$), then $P_1 = p_{\bullet 1}$. This is also consistent with comments made in the ecological inference, and related, literature; see, for example, Beh (2008) and Wakefield (2004). Therefore, by considering (5) and (9) the Pearson chi-squared statistic for a 2×2 contingency table can be expressed in terms of the log-odds ratio, given the marginal information, by

$$X^2(\ln(\theta) | p_{1\bullet}, p_{\bullet 1}) = n \left(\frac{P_1(\ln(\theta)) - p_{\bullet 1}}{p_{2\bullet}} \right)^2 \left(\frac{p_{1\bullet} p_{2\bullet}}{p_{\bullet 1} p_{\bullet 2}} \right). \quad (10)$$

Note that substituting (9) into (10) yields the succinct chi-squared function in terms of the log-odds ratio

$$X^2(\ln(\theta) | p_{1\bullet}, p_{\bullet 1}) = \frac{n_{1\bullet} n_{2\bullet} n_{\bullet 1} n_{\bullet 2}}{n(n-1)^2} (\ln(\theta))^2. \quad (11)$$

From (9), it can be shown that, by taking into consideration the bounds of P_1 given by (3), the log-odds ratio is bounded by

$$\tilde{L}_\theta = -\left(\frac{n-1}{n}\right) \min\left(\frac{1}{p_{1\bullet}p_{\bullet 1}}, \frac{1}{p_{2\bullet}p_{\bullet 2}}\right) < \ln(\theta) < \left(\frac{n-1}{n}\right) \min\left(\frac{1}{p_{1\bullet}p_{\bullet 2}}, \frac{1}{p_{2\bullet}p_{\bullet 1}}\right) = \tilde{U}_\theta \quad (12)$$

when only the marginal information is available. Similarly, when only the marginal information is available, and the association between the two dichotomous variables are being tested at the α level of significance, the log-odds ratio is bounded by

$$\begin{aligned} \tilde{L}_\alpha &= -\left(\frac{n-1}{n}\right) \min\left(\frac{1}{p_{2\bullet}p_{\bullet 2}}, \sqrt{\frac{\chi_\alpha^2}{n}} \cdot \frac{1}{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}\right) < \ln(\theta) \\ &< \left(\frac{n-1}{n}\right) \min\left(\frac{p_{\bullet 2}}{p_{\bullet 1}p_{2\bullet}}, \sqrt{\frac{\chi_\alpha^2}{n}} \cdot \frac{1}{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}\right) = \tilde{U}_\alpha. \end{aligned} \quad (13)$$

When $\tilde{L}_\alpha < \ln(\theta) < \tilde{U}_\alpha$ then there is evidence that the row and column variables are independent at the α level of significance. However, if $\tilde{L}_\theta < \ln(\theta) < \tilde{L}_\alpha$ or $\tilde{U}_\alpha < \ln(\theta) < \tilde{U}_\theta$ then there is evidence to suggest that the variables are associated.

By incorporating the above log-odds ratio results into the area under the curve defined by (11), but above χ_α^2 , the AAI, when considering the log-odds ratio, is

$$\tilde{A}_\alpha = 100 \left(1 - \frac{\chi_\alpha^2 \left[(\tilde{L}_\alpha - \tilde{L}_\theta) + (\tilde{U}_\theta - \tilde{U}_\alpha) \right] + \int_{\tilde{L}_\alpha}^{\tilde{U}_\alpha} X^2(\ln(\theta) | p_{1\bullet}, p_{\bullet 1}) d(\ln(\theta))}{\int_{\tilde{L}_\theta}^{\tilde{U}_\theta} X^2(\ln(\theta) | p_{1\bullet}, p_{\bullet 1}) d(\ln(\theta))} \right). \quad (14)$$

This AAI expression may be simplified by directly evaluating the integrals in this index. By considering (11), when $b > a$,

$$\int_a^b X^2(\ln(\theta) | p_{1\bullet}, p_{\bullet 1}) d(\ln(\theta)) = \frac{1}{3} \left(\frac{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}}{n(n-1)^2} \right) [b^3 - a^3]$$

Therefore, (14) may be alternatively expressed as

$$\tilde{A}_\alpha = 100 \left(1 - \frac{\chi_\alpha^2 \left[(\tilde{L}_\alpha - \tilde{L}_\theta) + (\tilde{U}_\theta - \tilde{U}_\alpha) \right]}{\tilde{k}(\tilde{U}_\theta^3 - \tilde{L}_\theta^3)} - \frac{\tilde{U}_\alpha^3 - \tilde{L}_\alpha^3}{\tilde{U}_\theta^3 - \tilde{L}_\theta^3} \right) \quad (15)$$

where $\tilde{k} = \frac{1}{3} \left(\frac{n_{1\bullet} n_{1\bullet} n_{2\bullet} n_{2\bullet}}{n(n-1)^2} \right)$.

3.3. Solving the Quadratic

These results, based on using the expectation and variance of n_{11} , (8), are appropriate only when the expectation produces a positive result. However, as pointed out by Stevens (1951), Patnaik's (1948) expectation of n_{11} can lead to negative quantities. This is clearly inappropriate in some cases and an alternative strategy can be considered.

By rearranging (6) p_{11} can be expressed as a quadratic function in terms of the odds ratio. Solving this quadratic expression yields

$$p_{11}(\theta | p_{1\bullet}, p_{\bullet 1}) = \frac{\theta(p_{1\bullet} + p_{\bullet 1}) + (p_{2\bullet} - p_{\bullet 1}) - \sqrt{\{\theta(p_{1\bullet} + p_{\bullet 1}) + (p_{2\bullet} - p_{\bullet 1})\}^2 - 4p_{1\bullet}p_{\bullet 1}\theta(\theta - 1)}}{2(\theta - 1)} \quad (16)$$

for $\theta \neq 1$. This result has been long studied and was considered by, for example, Mosteller (1968, page 7). More recently Fleiss, Levin and Paik (2003, section 6.6) also considered an expression much like (16) and noted that it provides accurate results for large samples. While not working directly with the odds ratio, Yule (1900, page 273) proposed a similar equation to (16). Since $P_1 = p_{11}/p_{1\bullet}$, P_1 can be expressed in terms of the marginal information and the odds ratio, $P_1(\theta | p_{1\bullet}, p_{\bullet 1})$, by

$$P_1(\theta | p_{1\bullet}, p_{\bullet 1}) = \frac{\theta(p_{1\bullet} + p_{\bullet 1}) + (p_{2\bullet} - p_{\bullet 1}) - \sqrt{\{\theta(p_{1\bullet} + p_{\bullet 1}) + (p_{2\bullet} - p_{\bullet 1})\}^2 - 4p_{1\bullet}p_{\bullet 1}\theta(\theta - 1)}}{2p_{1\bullet}(\theta - 1)}. \quad (17)$$

Stevens (1951) noted that Patnaik's expressions are applicable, and valid, when $\theta \rightarrow 1$ and $n \rightarrow \infty$. The impact this property has on the evaluation of the AAI when considering the odds ratio, or log-odds ratio, is negligible when the bounds of θ , or $\ln(\theta)$, are relatively narrow.

4. A Positive or Negative Association?

An advantage of considering the AAI is that, given only the marginal information, Beh (2010) showed that it is possible to determine whether the association is more likely to be positive or negative. This can be achieved by partitioning A_α such that $A_\alpha = A_\alpha^+ + A_\alpha^-$. Here, A_α^+ is the aggregate positive association index and A_α^- is the aggregate negative association index. Given a level of significance α under which a test of association is being made, these quantities measure the extent to which the marginal information reflect a significant positive, or negative, association.

When the AAI is expressed in terms of P_1 , Beh (2010) showed that

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

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

Similar results can also be obtained when expressing the AAI in terms of the odds ratio.

When considering (14), or its alternative form of (15), $\check{A}_\alpha = \check{A}_\alpha^+ + \check{A}_\alpha^-$ where

$$\check{A}_\alpha^+ = \frac{\int_{\check{U}_\alpha}^{\check{U}_\theta} [X^2(\ln \theta | p_{1\bullet}, p_{\bullet 1}) - \chi_\alpha^2] d(\ln \theta)}{\int_{\check{L}_\theta}^{\check{U}_\theta} X^2(\ln \theta | p_{1\bullet}, p_{\bullet 1}) d(\ln \theta)} = \frac{\check{k}(\check{U}_\theta^3 - \check{U}_\alpha^3) - (\check{U}_\theta - \check{U}_\alpha)\chi_\alpha^2}{\check{k}(\check{U}_\theta^3 - \check{L}_\theta^3)}, \quad (18)$$

$$\check{A}_\alpha^- = \frac{\int_{\check{L}_\theta}^{\check{L}_\alpha} [X^2(\ln \theta | p_{1\bullet}, p_{\bullet 1}) - \chi_\alpha^2] d(\ln \theta)}{\int_{\check{L}_\theta}^{\check{U}_\theta} X^2(\ln \theta | p_{1\bullet}, p_{\bullet 1}) d(\ln \theta)} = \frac{\check{k}(\check{L}_\alpha^3 - \check{L}_\theta^3) - (\check{L}_\alpha - \check{L}_\theta)\chi_\alpha^2}{\check{k}(\check{U}_\theta^3 - \check{L}_\theta^3)}. \quad (19)$$

Whether the conditional proportion P_1 or the log-odds ratio is considered for the calculation of the AAI, the association structure remains unchanged. The key difference is that the AAI has been transformed into these three simple measures of association. Therefore, while $A_\alpha = \check{A}_\alpha$ so too will it be that $A_\alpha^+ = \check{A}_\alpha^+$ and $A_\alpha^- = \check{A}_\alpha^-$.

In the *Appendix*, R code is provided which calculates the AAI, using (15), for a given contingency table. The code also calculates $\check{A}_{\alpha}^{+}$ and $\check{A}_{\alpha}^{-}$.

5. Examples

5.1. Fisher's Twin Data

5.1.1 Fisher's 2×2 Contingency Table

Consider the 2×2 contingency table of Table 2. This table was first analysed by Fisher (1935) and used by Beh (2008, 2010) to illustrate the simple application of (4) – (5). Fisher's data studies 30 criminal twins and classifies them according to whether they are a monozygotic twin or a 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 concerning the applicability of using the Pearson chi-squared statistic in cases where the cell frequencies are not greater than five. Fleiss, Levin and Paik (2003) provide an excellent review of strategies for including Yate's continuity correction (Yates, 1934). However, other studies have also revealed that incorporating the correction is not essential (see, for example, Pearson, 1947; Plackett, 1964; Grizzle, 1967; Conover, 1974) and so we will not consider its inclusion here.

Table 2

Fisher's (1935) criminal twin data

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

The Pearson chi-squared statistic for Table 2 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. The correlation of the two variables is $r = +0.6591$ (p-value < 0.0002) indicates that this association is statistically and

significantly positive. Therefore a monozygotic twin of a convicted criminal is associated with being convicted of a crime, while a dizygotic twin of a convicted criminal tends not to be a convicted criminal.

The odds ratio of Table 2 is 25.00. Therefore the log-odds ratio is 3.22 and has a 95% confidence interval of (1.26, 5.18). Thus, the 95% confidence interval for the odds ratio is (3.52, 177.48). Both these intervals indicate that there is a statistically significant positive association between the two dichotomous variables at the 5% level of significance. This is consistent with the findings made above regarding the correlation of the variables.

Fig. 2. Plot of $X^2(P_1 | p_{1\cdot} = 12/30, p_{\cdot 1} = 13/30)$ versus P_1 for Table 2

5.1.2 The AAI in terms of P_1

Beh (2010) considered the AAI of Table 2 in terms of P_1 and showed that $A_{0.05} = 61.83$. This index shows that it is likely that a 2×2 contingency table with the marginal information of Table 2 will reflect a statistically significant association (at the 5% level of significance) between the two dichotomous variables. Fig. 2 provides a graphical inspection of the meaning of this index. It shows that the Pearson chi-squared statistic is maximised at the bounds of P_1 ; the local maximum chi-squared values are 15.2941 and 26.1538. It can also

be seen that the shaded region exceeding the critical value of $\chi^2_{0.05}(df=1)=3.84$ (depicted by the dashed horizontal line) but below the chi-squared curve defined by (2) is quite large. This region represents 61.83% of the area under the curve and it is this quantity that is the AAI.

5.1.3 The AAI in terms of the Log-Odds Ratio, $\ln\theta$

Suppose we now consider the AAI in terms of θ and $\ln\theta$. Fig. 3 provides a view of the relationship between the chi-squared statistic and the odds ratio, θ , for Table 2; the solid line is obtained by considering the results of (9) proposed by Patnaik (1948). The curved dashed line is obtained by considering (17) and the horizontal dashed line coincides with the critical value of the chi-squared statistic at the 5% level of significance with 1 degree of freedom; 3.81. The logarithm of the intervals (12) and (13) are also included in Fig. 2.

Fig. 3. Plot of $X^2(\theta | p_{1\bullet} = 12/30, p_{\bullet 1} = 13/30)$ versus θ for Table 2.

Solid line based on (9), dashed (curved) line based on (17)

Given the results derived in Section 3, the log-odds ratio of Table 2 is bounded by $(\tilde{L}_\theta, \tilde{U}_\theta) = (-2.84, 3.718)$. If we consider that the bounds of n_{11} as defined by (3) range between 0 to 12 (and thereby yielding a log-odds ratio that does not exist), such an approximation works well. Suppose, we eliminate the problematic limits of n_{11} and confine our attention to the case where $n_{11} = 1$ and $n_{11} = 11$, then the log odds ratio is -3.09 and 4.48 respectively. Given only the marginal information, and (13), if we were to test the association

structure of the two dichotomous variables at the 5% level of significance, statistical independence occurs when the log-odds ratio is bounded by the interval $(\tilde{L}_{0.05}, \tilde{U}_{0.05}) = (-1.425, 1.425)$. A graphical comparison of the relationship between the chi-squared statistic and the log-odds ratio is given by Fig. 4; the dashed horizontal line in this figure is, again, the critical value of the chi-squared statistic at the 5% level of significance - $\chi^2_{0.05}(df = 1) = 3.84$.

One can see that when considering the relationship between the chi-squared statistic and the log-odds ratio as defined by (9) and (17), they are very similar when compared with its alternative result, especially when the odds ratio is confined to lie within the interval $(\tilde{L}_{0.05}, \tilde{U}_{0.05})$. Any deviation between (9) and (17) exists at the bounds of the log-odds ratio $(\tilde{L}_\theta, \tilde{U}_\theta)$. Thus, the AAI may be accurately, and simply, derived using (14), or alternatively (15). Now,

$$\tilde{k} = \frac{1}{3} \left(\frac{13 \times 17 \times 12 \times 18}{30 \times 29^2} \right) = 0.6307.$$

Therefore, when testing the association between the two variables of Table 2 at the 5% level of significance, the AAI is, from (15),

$$\tilde{A}_{0.05} = 100 \left(1 - \frac{3.84 \times [(-1.425 + 2.84) + (3.718 - 1.425)]}{0.6307 \times (3.718^3 - (-2.84)^3)} - \frac{1.425^3 - (-1.425)^3}{3.718^3 - (-2.84)^3} \right) = 61.83.$$

This is exactly the same as the AAI derived when considering P_1 . R code is provided in the appendix to calculate the AAI, when considered as a function of the log-odds ratio. This code also calculates the positive and negative contributions of the AAI as defined by equations (18) and (19) which we shall consider in section 5.1.4.

Fig. 4, and $\tilde{A}_{0.05} = 61.83$, shows that, given the marginal information only of Table 2, there appears to be some evidence that a strong association exists. This is evident by considering the area under the solid curve that lies above the critical value of 3.84 but for the log-odds ratio confined to its bounds. Despite the continuous nature of the odds ratio in our discussion, it is problematic to quantifying the AAI using (17). However, by considering plots of the form that Fig. 4 takes, we can determine that the AAI is bounded such that

$$AAI_{\alpha} < 100 \left[1 - \frac{1}{2} \left(\frac{\chi_{\alpha}^2(df=1)}{X^2(P_1(\ln \theta) = L_{\alpha})} + \frac{\chi_{\alpha}^2(df=1)}{X^2(P_1(\ln \theta) = U_{\alpha})} \right) \right].$$

For Table 2, this suggests that if one considers its log-odds ratio given only the marginal information, the upper bound for the AAI at the 0.05 level of significance is

$$AAI_{0.05} < 100 \left[1 - \frac{1}{2} \left(\frac{3.84}{15.2941} + \frac{3.84}{26.1538} \right) \right] = 80.1 .$$

In fact, by considering a log-odds ratio greater than zero, we can see that the area under the curve is far greater than the area under the curve when the log-odds ratio is negative. This suggests that, not only is there strong evidence of a statistically significant association between the two dichotomous variables at the α level of significance, but that the association is more likely to be positive than negative.

Fig. 4. Plot of $X^2(\ln \theta | p_{1\bullet} = 12/30, p_{\bullet 1} = 13/30)$ versus $\ln \theta$ for Table 2.

Solid line based on (9), dashed (curved) line based on (17)

5.1.4 The Direction of the Association

Irrespective of which simple measure of association is considered, the AAI is 61.83 and concludes that, given the marginal information, it is likely that there is a statistically significant association at the 5% level of significance. The question now is whether we can determine the direction of the association.

Suppose we first consider the AAI in terms of P_1 . By partitioning the AAI such that $A_\alpha = A_\alpha^+ + A_\alpha^-$ we see that $A_\alpha^+ = 46.43$ and $A_\alpha^- = 15.40$. Beh (2010) determined these quantities which suggests that the association is three times more likely to be positive than negative. We may also determine these quantities when considering the log-odds ratio by deriving the partition $\tilde{A}_\alpha = \tilde{A}_\alpha^+ + \tilde{A}_\alpha^-$. As we described above, since these AAI terms are related by a simple reparameterisation, we find that $A_\alpha^+ = \tilde{A}_\alpha^+ = 46.43$ while $A_\alpha^- = \tilde{A}_\alpha^- = 15.40$. The magnitude of these aggregate positive and negative association indices is reflected by the relatively large area above the critical value, but below the AAI curves, of Fig. 4 taking into consideration the bounds associated with $\ln \theta$.

5.2 *New Zealand Voting Data, 1893 - 1919*

The history of the New Zealand (NZ) election system is an interesting one because, in 1893, it is the first self-governing nation in the world to grant women the right to vote in federal elections; even though they were not eligible to stand as candidates until 1919. The trend was quickly spread across the globe including Australia: South Australia enfranchised women in 1894, Western Australia in 1899, and the Australian Commonwealth government in 1902. One may consult the following URL www.elections.org.nz/study/education-centre/history/votes-for-women.html for an extensive history of the voting status for women in NZ. Further information may be found by consulting Moore (2005).

The data from the NZ federal elections held between 1893 and 1919 provides a wealth of information for the analysis of early voting behaviour. For each year that a national election was held, Table 3 provides a summary of the number of men and women voters as well as the number of registered voters for each gender. This table is derived from Table 1 of Hudson, Moore, Beh and Steel (2010). Fortunately for analysts studying this issue, data at the

electorate level were also kept that records the gender of those that voted and those that did not. An example of this data can be seen by considering Table 4 which provides a summary of the men and women who registered to vote in the 1893 election. A more comprehensive discussion of the application of the AAI (in terms of P_1) to the New Zealand voting data of Table 3 was made by Tran, Beh and Hudson (2012).

Table 3

Summary of the 11 NZ elections, 1893 – 1919.

Year	No. Of Electoral districts	No. Of registered men	No. Of registered women	Men's votes	Women's votes
1893	57	175,915	147,567	126,183	88,484
1894	62	191,881	157,942	74,366	47,862
1896	62	197,002	142,305	149,471	108,783
1899	59	202,044	157,974	159,780	119,550
1902	68	229,845	185,944	180,294	138,565
1905	76	263,597	212,876	221,611	175,046
1908	76	294,073	242,930	238,534	190,114
1911	76	321,033	269,009	271,054	221,878
1914	76	335,697	280,346	286,799	234,726
Apr 1919	76	321,773	304,859	241,524	241,510
Dec 1919	76	355,300	328,320	289,244	261,083

If we consider for the moment that the cell values of Table 2 are known, the Pearson chi-squared statistic is 4978.115 and, with a p-value < 0.0001 , shows that there is a statistically significant association between gender and whether the voter turned out to vote in the 1893 election. It must be noted that the very large sample size has influenced the very large chi-squared statistic, thus we would expect that, for an analysis of the marginal information only, the AAI to be very high. However, it is the direction of the association that is of interest to us here for this data. We can see that 88484/147567, or 60% of women voted compared with 71.7% of men.

Table 4

Cross-classification of registered voters by gender for 1893

	Vote	No Vote	Total
Women	88484	59083	147567
Men	126183	49732	175915
Total	214667	108815	323482

Based on the counts in Table 4, women are $88484/59083 = 1.49$ times more likely to vote than not vote. Similarly, men are $126183/49732 = 2.53$ times more likely to vote than not vote. Therefore, with an odds ratio of $1.49/2.53 = 0.59$ thus a women are 0.59 times more likely to vote than a male (we may more succinctly express this by saying that, for the 1893 New Zealand election, males were 1.7 times more likely to vote than females). Therefore the log-odds ratio is -0.527 and has a 95% confidence interval of (-0.54, -0.51). Thus, the 95% confidence interval for the odds ratio is (0.58, 0.60). Both these intervals indicate that there is a statistically significant negative association between the two dichotomous variables of Table 4 at the 5% level of significance. The 95% confidence interval of the log-odds ratio also suggests that the association between gender and voting status is negative. Therefore, analysing Table 4 with the cell values assumed known, this interval suggests that there are statistically significantly more men voting that females.

Fig. 5. Plot of $X^2(\theta | p_{i\cdot}, p_{\cdot i})$ versus θ for Table 4 using (9)

Consider now the case where only the marginal information of Table 4 is known. Not surprising the AAI is 99.9932 suggesting that there is ample evidence in the marginal frequencies of Table 4 to suggest that the association between the two dichotomous variables is statistically significant at the 0.05 level of significance. Again, this is largely due to the large sample size for the 1893 election. Fig. 5 provides a plot of the odds ratio against the Pearson chi-squared statistic for Table 4; the dashed horizontal line is where $\chi^2_{0.05}(df = 1) = 3.84$. Fig. 6 gives an analogous plot but considers the log-odds ratio.

Fig. 6. Plot of $X^2(\theta | p_{1\bullet}, p_{\bullet 1})$ versus $\ln\theta$ for Table 4 using (9)

We can, however, confine our attention to whether the association is likely to be a positive or negative by considering equations (18) and (19). Fig. 6 suggests that when considering the relationship between gender and voting status of Table 4, and keeping in mind the reformulation of the AAI in terms of the odds ratio, the association is far more likely to be negative than positive. Using the R code given in the *Appendix* we indeed find

that $\check{A}_\alpha^+ = 37.11$ and $\check{A}_\alpha^- = 62.88$. Such figures suggest that, based only on the marginal information of Table 4, the association between gender and voter turnout is far more likely to be negative than positive. Therefore, based only on the marginal information, these quantities suggest that the proportion of voters is more likely to be male than female; this is consistent with our findings when the cell values of Table 4 are assumed known. From an ecological inference perspective, the estimated values of the gendered proportions studied by Hudson, Moore, Beh and Steel (2010) agree that males were more likely than females to vote in the 1893 national election held in New Zealand.

6 Discussion

This paper has discussed the reformulation of the aggregate association index in terms of the odds ratio and log-odds ratio for a single 2×2 contingency table. By considering the index in this manner, we can identify how likely two categorical variables will be associated based only on the marginal frequencies. However, measures of association, other than the odds ratio may also be considered by incorporating Edwards' criteria (Edwards, 1963). This criterion suggests that, for the analysis of 2×2 tables, measures of association should be expressible as functions of the odds ratio. While Edwards (1963) originally considered his study when the marginal totals are not assumed fixed, the criterion may be considered in the context of the AAI. Bishop, Fienberg and Holland (1975, page 378) suggest that such a relationship be structured so that

$$g(\theta) = \frac{f(\theta) - 1}{f(\theta) + 1}$$

where $f(\theta)$ is an invertible and monotonically increasing function in terms of the odds ratio, θ , and $g(\theta)$ is a measure of association. For example, when $f(\theta) = \theta$, $f(\theta) = \sqrt{\theta}$, $f(\theta) = \theta^{\pi/4}$, $f(\theta) = \theta^{3/4}$, $g(\theta)$ is Yule's Q (or "coefficient of association"; Yule, 1900, page 272), Yule's Y (or "coefficient of colligation"; Yule, 1912, page 592), Edwards' J (Edwards, 1957) and Digby's H (Digby, 1983, page 754) respectively. All these measures of association are designed to estimate (with varying degrees of accuracy) Pearson's (1900) tetrachoric correlation. One may also consider global measures of association other than

Pearson's chi-squared statistic, such as the Goodman-Kruskal tau index for asymmetrically structured variables (Goodman and Kruskal, 1954). Such generalisations require a more lengthy discussion of the issues that have been made here and so will be left for future consideration.

Preliminary work has been undertaken to investigate the extension of the AAI for multiple 2×2 tables (Tran, Beh and Hudson, 2012) when considered in terms of P_1 . Thus, extending the current discussion to multiple 2×2 tables by incorporating common odds ratios such as the Mantel-Haenszel and Woolf estimate is an avenue for future consideration. Establishing the links between the AAI and the Cochran-Mantel-Haenszel (CMH) test can also be considered.

One must note that the AAI we have considered here implies that the odds ratio and the log-odds ratio are continuous. Strictly speaking, since there are a discrete number of values that n_{11} may take, there are also a discrete number of values that θ can be. Therefore adjusting the AAI to take into consideration this characteristic can be made. Although, as Beh (2010) shows, there is little difference if one considers a continuous or discrete version of the AAI, especially when the interval length of (1) is wide and/or the sample size is large. Further consideration of this issue will also be left for future discussion.

References

- Agresti, A., 2002. *Categorical Data Analysis* (2nd ed), Wiley: NJ.
- Agresti, A., 1999. On logit confidence intervals for the odds ratio with small sample. *Biometrics* 55, 597 – 602.
- Agresti, A., Coull, B.A., 2002. The analysis of contingency tables under inequality constraints. *Journal of Statistical Planning and Inference* 107, 45 – 73.
- Anscombe, F., 1981. *Computing in Statistical Science through APL*. SpringerVerlag, New York.
- Beh, E.J., 2008. Correspondence analysis of aggregate data: The 2×2 table. *Journal of Statistical Planning and Inference* 138, 2941 – 2952.
- Beh, E.J., 2010. The aggregate association index. *Computational Statistics & Data Analysis* 54, 1570 – 1580.

- Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, MA.
- Brown, C.C., 1981. The validity of approximation methods for interval estimation of the odds ratio. *American Journal of Epidemiology* 113, 474 – 480.
- 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.
- Cheung, M.W.-L., Ho, R.C.M., Lim, Y., Mak, A., 2012. Conducting a meta-analysis: basics and good practices. *International Journal of Rheumatic Diseases* 15, 129 – 135.
- Conover, W.J., 1974. Some reasons for not using Yates continuity correction on 2×2 contingency tables (with discussion). *Journal of the American Statistical Association* 69, 374 – 382.
- Cornfield, J., 1951. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast, and cervix, *Journal of the National Cancer Institute* 11, 1269 – 1275.
- Demidenko, E., 2012. The shortest width confidence interval for odds ratio in logistic regression. *Open Journal of Statistics* 2, 305 – 308.
- Digby, P.G.N., 1983. Approximating the tetrachoric correlation coefficient. *Biometrics* 39, 753 – 757.
- Duncan, O.D., Davis, B., 1953. An alternative to ecological correlation. *American Sociological Review* 18, 665 – 666.
- Edwards, A.W.F., 1957. A note on the practical interpretation of 2×2 tables. *British Journal of Preventative and Social Medicine* 11, 73-78.
- Fisher, R.A., 1935. The logic of inductive inference (with discussions). *Journal of the Royal Statistical Society, Series A* 98, 39 – 82.
- Fleiss, J.L., Levin, B., Paik, M.C., 2003. *Statistical Methods for Rates and Proportions* (3rd ed). Wiley: NJ.
- Freedman, D.A., Klein, S.P., Sacks, J., Smyth, C.A., Everett, C.G., 1991. Ecological regression and voting rights. *Evaluation Review* 15, 673–711.
- Glas, A.S., Lijmer, J.G., Prins, M.H., Bonsel, G.J., Bossuyt, P.M.M., 2003. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology* 56, 1129 – 1135.
- Goodman, L.A., 1953. Ecological regressions and behavior of individuals. *American Sociological Review* 18, 663–664.

- Goodman, L.A., 1996. A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *Journal of the American Statistical Association* 91, 408–428
- Goodman, L.A., Kruskal, W.H., 1954. Measures of association for cross classifications. *Journal of the American Statistical Association* 49, 732–764.
- Grizzle, J.E., 1967. Continuity correction in the χ^2 for 2×2 tables. *The American Statistician* 21 (October), 28 – 32.
- Haldane J.B., 1955. The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics* 20, 309 - 311.
- Hauck, W.W., 1989. Odds ratio inference from stratified samples. *Communications in Statistics – Theory and Methods* 18, 767-800.
- Hudson, I. L., 2010. Meta-analysis and its application in phonological research: a review and new statistical approaches. In “Phonological Research: Methods for Environmental and Climate Change Analysis” (Hudson, I.L., Keatley, M.R. eds). Dordrecht: Springer, pp. 463 – 509.
- Hudson, I.L., Moore, L., Beh, E.J., Steel, D.G., 2010. 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* 173, 185 – 213.
- Hudson, I.L., Shafi, S., Lee, S., Hudson, S., Abell, A. (2012), Druggability in drug discovery: Self organising maps with a mixture discrimination approach. *Australian Statistics Conference*, Adelaide.
- Imai, K., Lu, Y., Strauss, A., 2011. eco: R package for ecological inference in 2×2 tables. *Journal of Statistical Software* 42, 23 pages.
- Jewell, N.P., 1986. On the bias of commonly used measures of association for 2×2 tables. *Biometrics* 42, 351 358.
- King, G., 1997. *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.
- Kraft, P., Wacholder, S., Cornelis, M.C., Hu, F.B., Thomas, G., Hoover, R., Hunter, D.J., Chanock, S., 2009. Beyond odds-ratios – communicating disease risk based on genetic profiles. *Nature Reviews Genetics* 10, 264 – 269.

- Lloyd, C.J., Moldovan, M.V., 2007a. Unconditional efficient one-sided confidence limits for the odds ratio based on unconditional likelihood. *Statistics in Medicine* 26, 5136 – 5146.
- Lloyd, C.J., Moldovan, M.V., 2007b. Exact one-sided confidence bounds for the risk ratio in 2×2 tables with structural zero. *Biometrical Journal* 49, 952 – 963.
- Mantel, N., Haenszel, W., 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22, 719 – 748.
- Mehta, C.R., Walsh, S.J., 1992. Comparison of exact, mid-p, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables. *The American Statistician* 46, 146-150.
- Miettinen, O.S., 1976. Estimability and estimation in case-referent studies. *American Journal of Epidemiology* 103, 226 – 236.
- Molenberghs, G., Lesaffre, E., 1994. Marginal modeling of correlated ordinal data using multivariate Plackett distribution. *Journal of the American Statistical Association* 89, 633 644.
- Montreuil, B., Bendavid, Y., Brophy, J., 2005. What is so odd about odds? *Canadian Journal of Surgery* 48, 400 – 408.
- Moore, L., 2005. Was gender a factor in voter participation at New Zealand elections? In “Class, Gender and the Vote” (Fairburn, M., Olssen, E. eds), Otago University Press: Otago, NZ, pp 129 - 142.
- Mosteller, F., 1968. Association and estimation in contingency tables. *Journal of the American Statistical Association* 63, 1 – 28.
- Nakaoka, H., Inoue, I., 2009. Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner’s curse. *Journal of Human Genetics* 54, 615 – 623.
- Patnaik, P.B., 1948. The power function of the test for the difference between two proportions in a 2×2 table. *Biometrika* 35, 157 – 173.
- Pearson, K., 1900. Mathematical contributions to the theory of evolution VII: On the correlation of characters not quantitatively measureable. *Philosophical Transactions of the Royal Society of London, Series A* 195, 1 – 47
- Pearson, E.S., 1947. The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika* 34, 139 – 167.
- Plackett, R.L., 1964. The continuity correction in 2×2 tables. *Biometrika* 51, 327 – 337.

- Plackett, R.L., 1977. The marginal totals of a 2×2 table. *Biometrika* 64, 37 – 42.
- Rothman, K., 2002. *Epidemiology: An Introduction*. Oxford, England: Oxford University Press.
- Rothman, K.J., Greenland, S., Lash, T.L., 2008. *Modern Epidemiology* (3rd ed). Lippincott Williams & Wilkins.
- Schmidt, C.O., Kohlmann, T., 2008. When to use the odds ratio or the relative risk? *International Journal of Public Health* 53, 165 – 167.
- Schurink, A., Ducro, B.J., Bastiaansen, J.W., Frankena, K., van Arendonk, J.A., 2012. Genome-wide association study of insect bite hypersensitivity in Dutch Shetland pony mares. *Animal Genetics* 10, 264 – 269.
- Steel, D.G., Beh, E.J., Chambers, R.L., 2004. The information in aggregate data. In “Ecological Inference: New Methodological Strategies” (King, G., Rosen, O., Tanner, M. eds). Cambridge: Cambridge, University Press, pp. 51 – 68.
- Stevens, W.L., 1951. Mean and variance of an entry in a contingency table. *Biometrika* 38, 468 – 470.
- Tran, D., Beh, E.J., Hudson, I.L., 2012. The aggregate association index and its application in the 1893 New Zealand election. *Proceedings of the Fifth ASEARC Conference*, (Hudson, I.L., Thandrayen, J., Stone, G., Gulati, C. eds), University of Wollongong, pp. 22 – 25.
- Viera, A.J., 2008. Odds ratios and risk ratios: What’s the difference and why does it matter? *Southern Medical Journal* 101, 730 – 734.
- Wakefield, J., 2004. Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A* 167, 385 – 424.
- Wakefield, J., Haneuse, S., Dobra, A., Teeple, E., 2011. Bayes computation for ecological inference. *Statistics in Medicine* 30, 1381 – 1396.
- Walter, S.D., Cook, R.J., 1991. A comparison of several point estimators of the odds ratio in a single 2×2 contingency tables. *Biometrics* 47, 795 811.
- Wilson, P.D., Langenberg, P., 1999. Usual and shortest confidence intervals on odds ratios from logistic regression. *The American Statistician* 53, 332 – 335.
- Woolf, B., 1955. On estimating the relation between blood group and disease. *Annals of Human Genetics* 19, 251 – 253.

- Yates, F., 1934. Contingency tables involving small numbers and the χ^2 test. Journal of the Royal Statistical Society Supplement 1, 217 – 235.
- Yule. G.U., 1900. On the association of attributes in statistics. Philosophical Transactions of the Royal Society, Series A 194, 257 – 319.
- Yule, G.U., 1912. On the methods of measuring the association between two attributes. Journal of the Royal Statistical Society 75, 579 – 642.
- Zhang, J., Yu, K.F., 1998. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. Journal of the American Medical Association 280, 1690 – 1691.

Appendix – R Code

Here we shall give R code that calculates the AAI when expressed in terms of the log-odds ratio. It does so using equation (15) and incorporates the bounds of (12) and (13). The analyst only needs to specify the contingency table, N, being analysed and the level of significance alpha; by default alpha is set at 0.05. The code includes as its output the AAI (defined by equation (15), $\tilde{A}_\alpha^+$ and $\tilde{A}_\alpha^-$).

```
AAIlogtheta <- function(N, alpha = 0.05){
  n <- sum(N)
  p<- N/n

  rows <- apply(p, 1, sum)
  cols <- apply(p, 2, sum)

  # Critical value of the chi-squared statistic with one degree
  # of freedom and the alpha level of significance

  critchisq <- qchisq(1 - alpha, 1)

  # The bounds of ln(theta); given by equation (12)
  #
  Ltheta <- -((n - 1)/n) * min(1/(rows[1] * cols[1]),
    1/(rows[2] * cols[2]))

  Utheta <- ((n - 1)/n)*min(1/(rows[1] * cols[2]), 1/(rows[2]
    * cols[1]))

  # The bounds of ln(theta), when testing the association at the
  # alpha level of significance; given by equation (13)

  Lalpha <- -((n - 1)/n)*min(1/(rows[2] * cols[2]),
    sqrt(critchisq/(n * rows[1] * rows[2] * cols[1] *
    cols[2]))))

  Ualpha <- ((n - 1)/n) * min(cols[2]/(rows[1] * rows[2]),
    sqrt(critchisq/(n * rows[1] * rows[2] * cols[1] *
    cols[2]))))
}
```

```

k <- (1/3) * ((n^3 * rows[1] * rows[2] * cols[1] *
             cols[2])/((n - 1)^2))

# The positive and negative quantities of the AAI; given by
# equations (18) and (19) respectively

AAIpos <- 100 * (k * (Utheta^3 - Ualpha^3) - (Utheta - Ualpha)
               * critchisq)/(k * (Utheta^3 - Ltheta^3))

AAIneg <- 100 * (k * (Lalpha^3 - Ltheta^3) - (Lalpha - Ltheta)
               * critchisq)/(k * (Utheta^3 - Ltheta^3))

# The AAI, our quantity of interesting lying between 0 and 100

AAI <- AAIpos + AAIneg

return(list(AAIpos = AAIpos, AAIneg = AAIneg, AAI = AAI))
}

```

For example, suppose we consider Fisher's (1935) twin data of Table 2. Then, expressing it as the R object `fisher.dat` such that

```

> fisher.dat<- matrix(c(10,2,3,15), nrow = 2)
> dimnames(fisher.dat) <- list(paste(c("Monozygotic","Dizygotic")),
+                               paste(c("Convicted","Not Convicted")))
> fisher.dat

```

	Convicted	Not Convicted
Monozygotic	10	3
Dizygotic	2	15

```

>

```

Therefore, we may calculate the AAI and its positive and negative terms by

```

> AAIlogtheta(fisher.dat)
$AAIpos
Monozygotic
  46.43128

$AAIneg
Monozygotic
  15.39580

$AAI
Monozygotic
  61.82707

>

```

Acknowledgement

We would like to thank Dr Linda Moore (Statistics New Zealand) for making available the 1893 New Zealand voting data that is the focus of the application in Section 5.2