The Odds Ratio and Aggregate Data: The 2×2 Contingency Table

Eric J. Beh

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA
eric.beh@newcastle.edu.au

Abstract

The odds ratio remains one of the simplest of measures for quantifying the association structure between two dichotomous variables. Its use is especially applicable when the cell values of a 2×2 contingency table are known. However, there are cases where this information is not known. This may be due to reasons of confidentiality or because the data was not collected at the time of the study. Therefore one must resort to considering other means of quantifying the association between the variables. One strategy is to consider the aggregate association index (AAI) proposed by [1]. This paper will explore the characteristics of the AAI when considering the odds ratio of the 2×2 contingency table.

Keywords: 2×2 contingency table, aggregate association index, aggregate data, odds ratio.

1. Introduction

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 ith row and jth column by $p_{i\bullet} = p_{i1} + p_{i2}$ and $p_{\bullet j} = p_{1j} + p_{2j}$ respectively. Table 1 provides a description of the notation used in this paper.

	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

Table 1: Notation for a 2x2 contingency table

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 and magnitude of the association may be determined by considering the Pearson product moment correlation

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

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

This paper will examine the structure of the association between two dichotomous variables based only on the marginal information. We shall do so by considering the aggregate association index proposed by [1, 2] in terms of the odds ratio, a very common measure of association for 2×2 contingency tables. The point of our discussion though is not to make inferences about the magnitude of the odds ratio, but to use its properties and the marginal frequencies (or proportions), to explore the association structure of the variables.

2. Aggregate Association Index

Let $P_1 = n_{11}/n_{1\bullet}$ and $P_2 = n_{21}/n_{2\bullet}$. 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'. The following comments apply to P_1 only but may be amended if one wishes to consider P_2.

When the cells of Table 1 are unknown, the bounds of the (1, 1)th cell frequency are well understood [4] to lie within the interval

$$\max(0, n_{\bullet 1} - n_{2\bullet}) \le n_{11} \le \min(n_{\bullet 1}, n_{1\bullet}).$$

Therefore, the bounds for P_1 are

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

[2] showed that when only marginal information is available the 95% confidence interval for P_1 is

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

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 $L_1 < P_1 < L_\alpha$ or $U_\alpha < P_1 < U_1$ then there is evidence to suggest that the variables are associated. From this interval, [1] proposed the following index

$$A_\alpha = 100\left(1 - \frac{\chi^2_\alpha\left[(L_\alpha - L_1) + (U_1 - U_\alpha)\right] + Int(L_\alpha, U_\alpha)}{Int(L_1, U_1)}\right) \quad (2)$$

where

$$Int(a, b) = \int_a^b X^2\left(P_1 \mid p_{1\bullet}, p_{\bullet 1}\right) dP_1$$

and

$$X^2\left(P_1 \mid p_{1\bullet}, p_{\bullet 1}\right) = n\left(\frac{P_1 - p_{\bullet 1}}{p_{2\bullet}}\right)^2\left(\frac{p_{1\bullet} p_{2\bullet}}{p_{\bullet 1} p_{\bullet 2}}\right) \quad (3)$$

Equation (2) is termed the aggregate association index (AAI). For a given α, this index quantifies how likely there will be a statistically significant association between the two dichotomous variables, given only the marginal information. A value of A_α close to zero suggests there is no association between the two variables. On the other hand, an index value close to 100 suggests that such an association may exist. 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.

3 The Odds Ratio

One of the most common measures of association for a 2×2 contingency table is the odds ratio

$$\theta = \frac{p_{11} p_{22}}{p_{21} p_{12}} = \frac{p_{11}\{p_{11} - (p_{1\bullet} + p_{\bullet 1} - 1)\}}{(p_{1\bullet} - p_{11})(p_{\bullet 1} - p_{11})}. \quad (4)$$

Often the logarithm of the odds ratio (also simply referred to as the log-odds ratio) is considered as a measure of association between two dichotomous variables. When the cell frequencies are known, the $100(1 - \alpha)\%$ confidence interval for log-odds ratio is

$$\ln(\theta) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

It is demonstrated in [9] 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. The underlying premise of the AAI is not to infer the magnitude of a measure of association. Instead it is to determine how likely a particular set of fixed marginal variables will enable to researcher to conclude that there exists a statistically significant association between the two dichotomous variables. In this paper, we tackle the problem by considering the odds ratio.

Since p_{11} is unknown here, one may express this proportion in terms of the marginal proportions and the odds ratio. If one considers (4), p_{11} may be expressed as a quadratic function in terms of the odds ratio. By solving this quadratic expression, we get

$$p_{11} = \frac{B - \sqrt{B^2 - 4 p_{1\bullet} p_{\bullet 1}\theta(\theta - 1)}}{2(\theta - 1)}$$

where

$$B = \theta(p_{1\bullet} + p_{\bullet 1}) + (p_{2\bullet} + p_{\bullet 1})$$

This result has been long studied and was considered by, for example, [8, pg 7] and [6, section 6.6]. Therefore, $P_1(\theta \mid p_{1\bullet}, p_{\bullet 1})$ may be expressed as

$$P_1(\theta) = \frac{B - \sqrt{B^2 - 4p_{1\bullet}p_{\bullet 1}\theta(\theta-1)}}{2p_{1\bullet}(\theta-1)} \qquad (5)$$

when $p_{1\bullet} \neq 0$. By substituting (5) into (3), the chi-squared statistic can be expressed as a function of the odds ratio.

It is very difficult to directly determine the 100(1 – α)% confidence intervals for the odds ratio based only on the marginal information. Such an interval, which we will denote by $\hat{L}_\alpha < \theta < \hat{U}_\alpha$, can be derived by considering those θ that satisfy

$$X^2(\theta \mid p_{1\bullet}, p_{\bullet 1}) = n\left(\frac{P_1(\theta) - p_{\bullet 1}}{p_{2\bullet}}\right)^2 \left(\frac{p_{1\bullet}p_{2\bullet}}{p_{\bullet 1}p_{\bullet 2}}\right) < \chi^2_\alpha$$

where χ^2_α is the 100(1 − α) percentile of a chi-squared distribution with 1 degree of freedom. Calculating $\hat{L}_\alpha$ and $\hat{U}_\alpha$ is computationally difficult. Therefore, for the purposes of our discussion, we shall approximate the bounds based on a graphical inspection of $X^2(\theta \mid p_{1\bullet}, p_{\bullet 1})$ versus θ.

We shall also be exploring the use of the log-odds ratio in the context of the AAI in the following section.

4 Example – Fisher's Twin Data

Consider the 2×2 contingency table of Table 4 analysed by [1, 2, 5]. These data concern 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 surrounding the applicability of using the Pearson chi-squared statistic in cases where the cell frequencies are not greater than five. [6] provides an excellent review of strategies for including Yate's continuity correction [11]. However, studies have revealed that incorporating the correction is not essential (eg [3, 7]) and so we will not consider its inclusion here.

The 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 product moment correlation of r = +0.6591 indicates that this association is 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.

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

Table 2: Criminal twin data original considered by [5]

[2] considered the AAI of Table 2 in terms of P_1 and showed that $A_{0.05}$ = 61.83. Therefore, it is likely that a 2 $\times$ 2 contingency table with the marginal information of Table 2 will reflect a statistically significant association (at the 5% level) between the two dichotomous variables. Figure 1 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 15.29 and 26.15. It can also be seen that the shaded region exceeding the critical value of $\chi^2_{0.05}(df = 1) = 3.84$ 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.

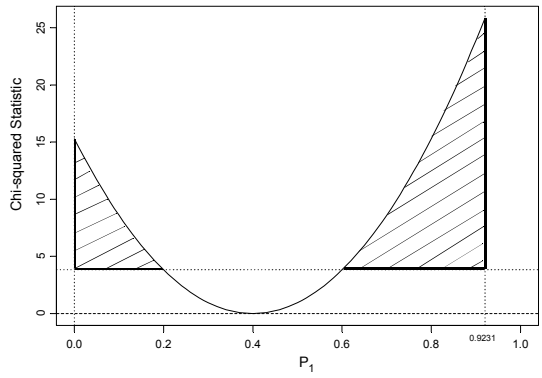

Figure 1: Plot of $X^2(P_1)$ versus P_1 for Table 1

For Table 2, $\theta = 25.00$ and the log-odds ratio of 3.22 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 significant positive association between the two dichotomous variables at the 5% level of significance. This is consistent with the findings made regarding the Pearson product moment correlation. We shall now consider the case where the cell frequencies are unknown.

Despite the simplicity and popularity of the odds ratio, the issue of determining the AAI becomes a little more complicated, but equally revealing. Let us first consider the relationship between the Pearson chi-squared statistic and the odds ratio – see Figure 3. This figure graphically shows that a maximum chi-squared statistic is reached when the odds ratio approaches zero or reaches infinity.

Similarly, the chi-squared statistic achieves its minimum of zero when the odds ratio is 1.

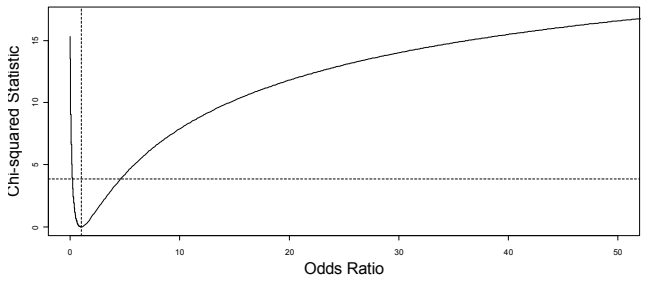

Figure 3: Plot of $X^2(\theta)$ versus θ.

Figure 3 shows the relationship between the chi-squared statistic and the odds ratio using (5). We can see that the chi-squared statistic is exceeded by the critical value of 3.84, at the 5% level of significance, when (approximately) $0.11 < \theta < 7.7$. However, since the shape of the curve is biased towards those odds ratios greater than 1, determining whether there may exist a positive or negative association using the odds ratio can produce misleading conclusions.

To overcome this problem we may also consider the log-odds ratio. Figure 4 shows the relationship between the Pearson chi-squared statistic and the log-odds ratio using (5). It reveals that when using (5) the local maximums of the Pearson chi-squared statistic (15.2941 and 26.1538) are reached as the log-odds ratio approaches negative, and positive, infinity.

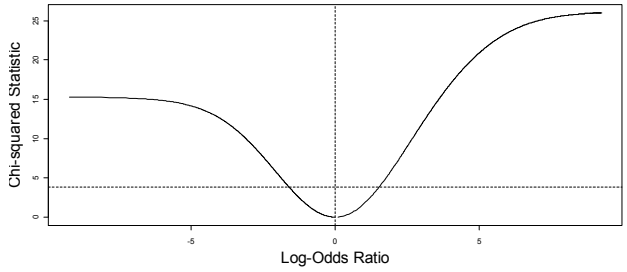

Figure 4: Plot of $X^2(\ln\theta)$ versus $\ln\theta$.

Figure 4 shows that, given only the marginal information of Table 2, there appears to be some evidence that a strong association exists. This is evident by considering the area under the curve that lies above the critical value of 3.84. In fact, by considering a log-odds ratio greater than zero, we can see that the area under the curve, using (5) 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 significant association between the two dichotomous variables, but that the association is more likely to be positive than negative.

5 Discussion

This paper discusses the use of the aggregate association index in terms of the 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 using the most popular of simple measures of association. Of course, we may explore the behaviour of this index in terms of other simple measures of association, including $\beta_{11} = p_{11}/(p_{1\bullet}p_{\bullet 1})$ which is referred to as the (1, 1)th Pearson ratio.

Our focus has been concerned with the chi-squared statistic but the index may be generalised for other measures of association such as the Goodman-Kruskal tau index. Other popular measures for 2×2 contingency tables such as Yule's Q ("coefficient of association") or Yule's Y ("coefficient of colligation") may also be examined in this context. On may also consider extending this index for multiple 2×2 tables or larger sized contingency tables. We shall consider these, and other, issues in future discussions of the index.

References

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

[2] E.J. Beh, The aggregate association index, Computational Statistics & Data Analysis, 54, 1570 – 1580, 2010.

[3] W.J. Conover, Some reasons for not using Yates continuity correction on 2×2 contingency tables (with discussion), Journal of the American Statistical Association, 69, 374 – 382, 1974.

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

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

[6] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions (3rd ed), Wiley: NJ, 2003

[7] J.E. Grizzle, Continuity correction in the χ^2 for 2×2 tables, American Statistician, 21 (October), 28 – 32, 1967.

[8] F. Mosteller, Association and estimation in contingency tables, Journal of the American Statistical Association, 63, 1 – 28, 1968.

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

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

[11] F. Yates, Contingency tables involving small numbers and the χ^2 test, Journal of the Royal Statistical Society Supplement, 1, 217 – 235, 1934.