The aggregate association index and its links with common measurements of association in a 2x2 table: An analysis of early New Zealand gendered voting data

D. Tran [a], **E.J. Beh** [a], **I.L. Hudson** [a], **L. M. Moore** [b]

[a] *School of Mathematical and Physical Sciences, University of Newcastle*
[b] *Statistics New Zealand*
Email: c3141292@uon.edu.au

Abstract: The analysis of aggregate, or marginal, data for contingency tables is an increasingly important area of statistics, especially in political science and epidemiology. Aggregation often exists due to confidentiality issues or by source of the data itself. Aggregate data alone makes drawing conclusions about the true association between categorical variables difficult, especially in dealing with the aggregate analysis of single or stratified 2x2 contingency tables. These tables are the most fundamental of data structures when dealing with cross-classifying categorical variables hence it is not surprising that the analysis of this type of data has received an enormous amount of attention in the statistical, and related, literature. However, the information, from which the aggregate data can provide for the inference of association between the variables, is still a long standing issue. In order to analyse the association that exists between the variables of a 2x2 table, or stratified 2x2 tables, based only on the aggregate data, numerous approaches that lie within the area of Ecological Inference (EI) have been proposed. As an application of this new development, we shall analyse a unique record of New Zealand gendered election data from 1893 when it was the first self-governing country in the world allowing women to vote, this trend quickly spread across the globe. Since the NZ data structure consists of stratified 2x2 tables, where the stratifier is electorate, the issue of analysing a single 2x2 table shall not be discussed. For stratified 2x2 tables, a number of ecological inference techniques exist but these rely on strong, yet untestable assumptions, which are not applicable to a single 2x2 table. To remedy this, one may analyse the association between two dichotomous variables, given only the aggregate data, by using the Aggregate Association Index (AAI). To date, the AAI has been expressed as a function of a conditional probability and been used to test if a statistically significant association is likely to exist given only aggregate data. Nevertheless, the interpretation about the strength and direction of association cannot be obtained through the current AAI. As a result, the purpose of this study is to broaden our understanding of the AAI by establishing its functional link with other classical association measurements, such as the standardised residual, Pearson's ratio, contingency and correlation indices. For brevity, only the standardised residual shall be considered here as a foundational baseline for the other association measures. This work will allow us to confirm the characteristics of the AAI's generalizability and enable analysis of aggregate data in terms of common association measurements. In other words, we show that the analysis of aggregate data of the 2x2 tables can be extended from justifying the existence of an association to that of determining the strength and direction of the association, if it exists, given only aggregate data. The important nature of association between gender and voting in the election shall be carefully examined given only aggregate data and compared to the information from a complete data analysis reported in (Hudson, Moore, Beh and Steel, 2010). Moore (2005) confirmed that gender was a significant factor in determining voting in early New Zealand elections from 1893. This paper shows that the AAI can provide the same result for testing the statistical association between the two dichotomous categorical variables, voting and gender, irrespective of the association measurements considered, and given only the aggregate data, or marginal information, of a 2x2 contingency table. It is noteworthy that, it is also possible to establish a relationship between the traditional AAI with other association measurements such as the standardised residual, so as to obtain a better understanding of association in terms of strength and direction. This new development thus extends the application of the AAI from not only justifying the existence of association, but also to interpretation of how strong or weak the association is and which direction (positive or negative) it is likely to be. Future developments of the AAI will involve the formulation of how to combine multiple AAI (Z_g) curves from each electorate into a single index for an election year. This may well allow us to compare the trends between politics and gender among different NZ elections (from 1893 to 1919) and to provide a better and unique methodology for ecological inference.

Keywords: *Aggregate Data, Aggregate Association Index (AAI), 2x2 tables, ecological inference, standardised residual, New Zealand election*

1. INTRODUCTION

In statistics, aggregate data has always been an interesting topic which attracts a lot of attention, especially for aggregate categorical data. In general, data aggregation is any process in which information is gathered and expressed in a summary form for purposes of statistical analysis. For categorical analysis, the data is often summarised in term of contingency tables. A common aggregation purpose is to get more information about association within particular groups based on specific variables such as age, profession, or gender. However, due to confidentiality and availability of the data, aggregate categorical data can become very difficult to analyse. In contingency tables, it means that only the aggregate, or marginal, totals are available. The scope of this paper will only focus on issues of 2x2 contingency tables.

In order to analyse the association that exists between two dichotomous variables of a 2x2 table, or stratified 2x2 tables, based only on the aggregate data, numerous approaches that lie within the area of Ecological Inference (EI) have dealt with this problem to varying degrees. However, current EI techniques still suffer from major shortfalls in the assumptions that are required (Hudson, Moore, Beh and Steel, 2010). An approach that does not require ensuring the integrity of the untestable EI assumptions given only aggregate data is the Aggregate Association Index (AAI). This technique was proposed by Beh (2008, 2010) and it is currently applied for the case of a single, or stratified 2x2 tables.

The core of the current AAI depends largely on a conditional probability P_1, which can be expressed as a linear function of p_{11}- the proportion of the sample size appearing in the $(1, 1)^{th}$ cell of a 2x2 table. However, there are a variety of classic association measurements that can also be expressed as a linear function of p_{11}. Therefore, the main purpose of this paper is to generalise the current AAI and explore its connection with one such classic association measure - the standardised residual. We shall demonstrate the properties of this generalised AAI, and the residual (as a special case), by analysing the early New Zealand voting data of 1893.

2. THE DATA AND NOTATIONS

2.1 New Zealand Voting Data

New Zealand (NZ) is a democratic country in which members of Parliament (MPs) are chosen in free and fair elections. Voting in NZ was introduced after colonisation by British settlers and, as it was back then, is not compulsory; however citizens and permanent residents aged 18 years and over must enrol to vote before Election Day. The history of the NZ election system is an interesting one because, in 1893, it was the first self-governing nation in the world providing women the right to vote in federal elections; however 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 women voting in NZ.

Table 1. Summary of the 11 NZ elections, 1893-1919

Year	No. Of Electorates	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

Table 2. Cross-classification of registered voters by gender for electorate 1, 1893

1st electorate 1893	Vote	No vote	Total
Women	1,443	289	1,732
Men	1,747	842	2,589
Total	3,190	1,131	4,321

Table 1 provides a summary of the number of men and women voters as well as the number of registered voters for each gender for 11 national elections held from 1893 to 1919. 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 2, which provides a summary of the men and women who registered to vote in electorate 1 of the 1893 election.

2.2 Notations

For each election, the electorate data can be summarised as a 2x2 contingency table; consider Table 2 to be one such example. Therefore, for a particular election, denote the total number of registered voters in the g^{th} electorate by n_g and the overall NZ population for a particular election is $N = \sum_g^G n_g$where G is the total number of electorates. Suppose that the number of voters in the i^{th} row and j^{th} column (for i = 1, 2 and j =1, 2) of the 2x2 table is n_{ijg} with an electorate proportion of $p_{ijg} = n_{ijg} / n_g$, the i^{th} and j^{th} marginal proportions can be denoted as $p_{i.g =} n_{i.g} / n_g$ and $p_{.jg =} n_{.jg} / n_g$ respectively.

For the study of the NZ voting data, the row variable consists of the gender categories "Women" (for i = 1) and "Men" (i = 2). Similarly, the column variable reflects whether a registered individual voted or not with categories "Vote" (j = 1) for a registered individual who voted and "No Vote" (j = 2) for a registered individual who did not vote. Table 3 provides a summary of the notation used in this paper.

g^{th} electorate	Vote	No vote	Total
Women	n_{11g}	n_{12g}	$n_{1.g}$
Men	n_{21g}	n_{22g}	$n_{2.g}$
Total	$n_{.1g}$	$n_{.2g}$	n_g

Table 3. A 2x2 table of registered voters in the g^{th} electorate of an election

Typically, analysing a contingency table involves answering the two questions: 1) Is there enough evidence to suggest an there exists a statistically significant association between the categorical variables? 2) If the variables are associated, how can we measure or quantify the association among them? The first question can easily be achieved by examining the Pearson chi-squared statistic calculated from the counts and marginals of a contingency table. In the case of the g^{th} electorate, the Pearson chi-squared statistic can be considered.

$$X_g^2 = n_g \frac{\left(n_{11g}n_{22g} - n_{12g}n_{21g}\right)^2}{n_{1.g}n_{2.g}n_{.1g}n_{.2g}} \quad (1)$$

For the second question, the Pearson product moment correlation can be used to determine the direction and magnitude of linear trend within a table.

$$r_g = \frac{n_{11g}n_{22g} - n_{12g}n_{21g}}{\sqrt{n_{1.g}n_{2.g}n_{.1g}n_{.2g}}} \quad (2)$$

The correlation coefficient ranges from -1 to 1. A value of ±1 implies that a perfect association, while $r_g = 0$ implies no association. The two questions above can be easily answered if the cell values of the contingency table are known. Given that, in this paper, only the marginal totals (or aggregate data) are available, Beh (2008, 2010) derived the Aggregate Association Index and showed that these questions can be answered for a 2x2 table. In the following sections, this index will be described and applied to analyse the NZ voting data.

3. THE AAI

Denote $P_{1g} = n_{11g}/n_{1.g}$ as the conditional probability of an individual being classified into 'Column 1' given that they are classified in 'Row 1' of the g^{th} electorate. Beh (2008, 2010) showed that the Pearson chi-squared statistic, (1), can be expressed as a function of P_{1g} and the aggregate data from the g^{th} electorate by:

$$X_g^2\left(P_{1g}\middle|p_{1.g}, p_{.1g}\right) = n_g \left(\frac{P_{1g} - p_{.1g}}{p_{2.g}}\right)^2 \left(\frac{p_{1.g}p_{2.g}}{p_{.1g}p_{.2g}}\right) \quad (3)$$

It can be seen that the chi-squared statistic is a quadratic function (with positive concavity) in terms of the conditional proportion P_{1g}.When the cell values of Table 2 are unknown, it is not possible to calculate P_{1g}, nor is it possible to determine $X_g^2\left(P_{1g}\middle|p_{1.g}, p_{.1g}\right)$. However, the extremes of the permissible values of the (1, 1)th cell frequency (Duncan and Davis, 1953) are well understood to lie within the interval

$$L_{n_{11g}} = max\left(0, n_{.1g} - n_{2.g}\right) \leq n_{11g} \leq min\left(n_{.1g}, n_{1.g}\right) = U_{n_{11g}} \quad (4)$$

Hence from (4) the value of P_{1g} lies within the interval:

$$L_{P_{1g}} = max\left(0, \frac{n_{.1g} - n_{2.g}}{n_{1.g}}\right) \leq P_{1g} \leq min\left(\frac{n_{.1g}}{n_{1.g}}, 1\right) = U_{P_{1g}} \quad (5)$$

Since L_{1g} and U_{1g} only depend on the marginal information, (3) can be investigated by using only the marginals. By taking into account the above properties of P_{1g}, Beh (2010) showed that when only aggregate data from a 2x2 table is available, and a test of the association is made at the α level of significance, the bounds of P_{1g} are

$$L_{P_{1g}}^{\alpha} = max\left(0, p_{.1g} - p_{2.g}\sqrt{\frac{\chi_\alpha^2}{n_g}\left(\frac{p_{.1g}p_{.2g}}{p_{1.g}p_{2.g}}\right)}\right) \le P_{1g} \le min\left(1, p_{.1g} + p_{2.g}\sqrt{\frac{\chi_\alpha^2}{n_g}\left(\frac{p_{.1g}p_{.2g}}{p_{1.g}p_{2.g}}\right)}\right) = U_{P_{1g}}^{\alpha} \qquad (6)$$

where χ_α^2 is the 1 – α percentile of the chi-squared distribution with 1 degree of freedom. Thus, for the gth electorate, Beh (2010) derived the following Aggregate Association Index (AAI):

$$A_{\alpha g} = 100\left(1 - \frac{\chi_\alpha^2\left[\left(L_{P_{1g}}^{\alpha} - L_{P_{1g}}\right) + \left(U_{P_{1g}} - U_{P_{1g}}^{\alpha}\right)\right]}{kn_g\left[\left(U_{P_{1g}} - p_{.1g}\right)^3 + \left(L_{P_{1g}} - p_{.1g}\right)^3\right]} - \frac{\left(U_{P_{1g}}^{\alpha} - p_{.1g}\right)^3 - \left(L_{P_{1g}}^{\alpha} - p_{.1g}\right)^3}{\left(U_{P_{1g}} - p_{.1g}\right)^3 - \left(L_{P_{1g}} - p_{.1g}\right)^3}\right) \qquad (7)$$

where $k = \frac{1}{3p_{2.}^2}\left(\frac{p_{1.}\,p_{2.}}{p_{.1}\,p_{.2}}\right)$. For a given α level of significance, this index is the ratio of the total region that lies under the curved defined by (3) but above the critical value of χ_α^2. Hence, given only the aggregate data, this index quantifies how likely it is that a statistically significant association will exist between the two dichotomous variables at the α level of significance. Figure 1 provides a graphical representation of the index. The index $A_{\alpha g}$ is bounded by [0,100] where a value of zero indicates that, at the α level of significance, there is no evidence of an association between the variables. A value close to 100 indicates that, at the α level of significance, there is enough evidence to suggest an association between the two variables (based on the available aggregate data).

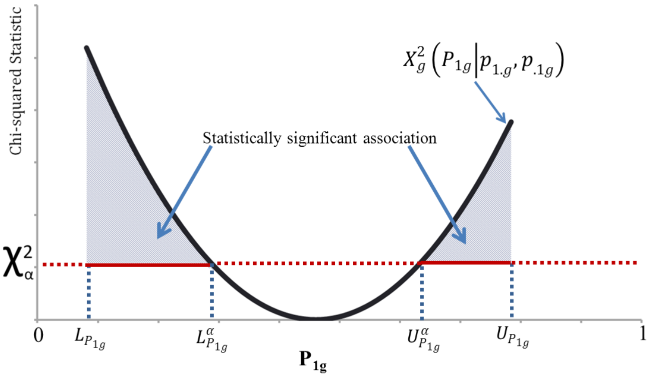

Figure 1. Graphical illustration of AAI concept for the gth electorate

4. THE TRANSFORMATION OF p_{11} IN THE AAI

Central to the original AAI, is the conditional proportion $P_{1g} = p_{11g}/p_{1.g}$, which is just a simple linear transformation of the p_{11g}. However, there are many other simple linear transformations of p_{11g} that yield other popular measures of the association between two dichotomous variables. Therefore, this suggests that the AAI is expressible as a function of association measurement other than P_{1g}. The form for any linear transformation of p_{11g} can be generalised as $l(p_{11g}) = ap_{11g} + b$. The choice of measurement that is used to reflect the association structure between the row and column categories of a 2x2 contingency table can be considered by choosing the appropriate value of *a* and *b*. We shall consider how the relationship between the generalised form of p_{11} and the AAI is shaped in the following sections. In this section, we shall establish a relationship between the linear transformation of p_{11g} and the AAI. By considering the transformation of $l(p_{11g})$, the bounds of (5) generalise to:

$$L_{lg} = amax(0, p_{.1g} - p_{2.g}) + b \le l(p_{11g}) \le amin(p_{.1g}, p_{1.g}) + b = U_{lg} \qquad (8)$$

while (6) becomes

$$L_{lg}^{\alpha} = max\left[b, ap_{1.g}\left(p_{.1g} - p_{2.g}\sqrt{\frac{\chi_\alpha^2}{n_g}\left(\frac{p_{.1g}p_{.2g}}{p_{1.g}p_{2.g}}\right)}\right) + b\right] \le l(p_{11g})$$
$$\le min\left[ap_{.1g} + b, ap_{.1g}\left(p_{.1g} + p_{2.g}\sqrt{\frac{\chi_\alpha^2}{n_g}\left(\frac{p_{.1g}p_{.2g}}{p_{1.g}p_{2.g}}\right)}\right) + b\right] = U_{lg}^{\alpha} \qquad (9)$$

We may also express the chi-squared statistic in terms of $l\left(p_{11g}\right)$ by:

$$X_g^2\left(l\left(p_{11g}\right)|p_{1.g},p_{.1g}\right)=n_g\left[\frac{l\left(p_{11g}\right)-E\left(l\left(p_{11g}\right)\right)}{ap_{1.g}\left(1-p_{1.g}\right)}\right]^2\left(\frac{p_{1.g}p_{2g.}}{p_{.1g}p_{.2g}}\right) \tag{10}$$

where $E\left(l\left(p_{11g}\right)\right)=\mu\left(p_{11g}\right)=aE\left(p_{11g}\right)+b$ is a linear function of the expectation of the (1, 1)th cell value under the hypothesis of independence. Note that, irrespective of the choice of a and b (other than a = 0, which we won't consider here), the chi-squared statistic is a quadratic function in terms of $l\left(p_{11g}\right)$. By the definition of the AAI, the area under the curve defined by (10) but above the critical value χ_α^2 reflects a significant association hence the generalised AAI for any simple linear transformation may be expressed by

$$\begin{aligned}A_{\alpha g}\left(l(p_{11})\right)=100\Bigg(1&-\frac{\chi_\alpha^2\left[\left(L_{lg}^\alpha-L_{lg}\right)+\left(U_{lg}-U_{lg}^\alpha\right)\right]}{k_{lg}n_g\left[\left(U_{lg}-\mu\left(p_{11g}\right)\right)^3+\left(L_{lg}-\mu\left(p_{11g}\right)\right)^3\right]}\\&-\frac{\left(U_{lg}^\alpha-\mu\left(p_{11g}\right)\right)^3-\left(L_{lg}^\alpha-\mu\left(p_{11g}\right)\right)^3}{\left(U_{lg}-\mu\left(p_{11g}\right)\right)^3-\left(L_{P1g}-\mu\left(p_{11g}\right)\right)^3}\Bigg)\end{aligned} \tag{11}$$

where $k_{lg}=\left(3a^2\left(p_{1.g}p_{2.g}p_{.1g}p_{.2g}\right)\right)^{-1}$. From (8), (9), (10), and (11) it can be seen that the AAI can be expressed in term of any linear function of p_{11g} and its value remains constant regardless of which linear function of p_{11g}. In terms of association measurements, it means that the value of AAI is constant regardless of which association measurement is considered, as long as the measurement can be expressed as a linear function of p_{11g}.This new characteristic of the AAI shall be examined more closely in the following section.

5. STANDARDISED RESIDUAL AND THE AAI

Suppose we consider the standardised residual of the (1, 1)th cell value of the 2x2 contingency table

$$Z_g=\frac{p_{11g}-p_{1.g}p_{.1g}}{\sqrt{p_{1.g}p_{.1g}}} \tag{12}$$

See, for example, Haberman (1973) and Agresti (2002). This is a special case of $l(p_{11})$ where $a=\left(p_{1.g}p_{.1g}\right)^{-1/2}$and $b=-\left(p_{1.g}p_{.1g}\right)^{1/2}$. Therefore, from (8), Z_g is bounded by

$$L_{Z_g}=-\sqrt{p_{1.g}p_{.1g}}min\left(1,\frac{p_{2.g}p_{.2g}}{p_{1.g}p_{.1g}}\right)\leq Z_g\leq\sqrt{p_{1.g}p_{.1g}}min\left(\frac{p_{2.g}}{p_{1.g}},\frac{p_{.2g}}{p_{.1g}}\right)=U_{Z_g} \tag{13}$$

while (9) in term of Z_g becomes

$$L_{Z_g}^\alpha=-p_{2.g}\sqrt{\frac{p_{1.g}}{p_{.1g}}}min\left(\frac{p_{.1g}}{p_{2.g}},\sqrt{\frac{\chi_\alpha^2}{n_g}\left(\frac{p_{.1g}p_{.2g}}{p_{1.g}p_{2.g}}\right)}\right)\leq Z_g\leq p_{2.g}\sqrt{\frac{p_{1.g}}{p_{.1g}}}min\left(\frac{p_{.1g}p_{2.g}}{p_{1.g}},\sqrt{\frac{\chi_\alpha^2}{n_g}\left(\frac{p_{.1g}p_{.2g}}{p_{1.g}p_{2.g}}\right)}\right)=U_{Z_g}^\alpha \tag{14}$$

The expected value of Z_g, under the hypothesis of independence is $E\left(Z_g\right)=0$. So the chi-squared statistic in term of $Z_g$and marginal information can be expressed as:

$$X_g^2\left(Z_g|p_{1.g},p_{.1g}\right)=n_gZ_g^2\left(\frac{1}{p_{2.g}p_{.2g}}\right) \tag{15}$$

The quadratic relationship between $X_g^2\left(Z_g|p_{1.g},p_{.1g}\right)$ and Z_g can also be graphically depicted in the same way that one can graphically show the relationship between $X_g^2\left(P_{1g}|p_{1.g},p_{.1g}\right)$ and P_{1g}. For such a relationship, the vertex of $X_g^2\left(Z_g|p_{1.g},p_{.1g}\right)$ is at $Z_g=0$ which corresponds to independence between the two dichotomous variables. Given (13), (14), and (15) the AAI may be expressed in terms of Z_g such that

$$A_{\alpha g}\left(Z_g\right)=100\left(1-\frac{\chi_\alpha^2\left[\left(L_{Z_g}^\alpha-L_{Z_g}\right)+\left(U_{Z_g}+U_{Z_g}^\alpha\right)\right]}{k_{Z_g}\left[U_{Z_g}^3-L_{Z_g}^3\right]}-\frac{\left(U_{Z_g}^\alpha\right)^3-\left(L_{Z_g}^\alpha\right)^3}{U_{Z_g}^3-L_{Z_g}^3}\right);k_{Z_g}=\frac{1}{3p_{2.g}p_{.2g}} \tag{16}$$

6. THE AAI AND THE NZ ELECTION DATA

6.1 2x2 contingency table cells are known

Consider Table 2 where its cells are assumed to be known, a Pearson chi-squared test of independence gives a test statistic of 134.68 and a P-value < 0.0001 at the level of significance α = 0.05. Thus, there is ample evidence to suggest that there exists a significant association between the gender and voting patterns at the α = 0.05. The direction of this association can also be determined by considering the Pearson correlation coefficient: r_1 = +0.18 which suggests that the association within the electorate 1, 1893 is likely to be positive but weak. That is, women who were registered are more likely to vote than registered men in the electorate 1 while registered men are more likely not to vote than registered women. In addition, from (12) the standardised residual $Z_1 = +0.07$ and its Monte Carlo P- values < 0.0001 (100,000 simulations with Poisson randomly selected cell values) also reflect a statistically weak positive association within the Table 2 - note that this interpretation is similar to the interpretation of the Pearson correlation coefficient.

6.2 Aggregate data analysis

Given only the marginal totals of Table 2, the analysis of association is now undertaken by considering only the marginal information. For electorate 1 of the 1893 election, from (7) the AAI (α = 0.05) is $A_{0.05,1}$ = 99.37. Therefore, when testing for the association at the 5% level of significance, 99.37% of contingency tables randomly generated with the marginal frequencies $n_{1.1}$ = 1,732, $n_{2.1}$ = 2,589, $n_{.11}$ = 3,190 and $n_{.21}$ = 1,131 will exhibit a significant association between the two dichotomous categorical variables: gender and voting patterns. That is, at the 5% level of significance and analysing only the aggregate information, the very high AAI indicates that there is very strong evidence to conclude that an association exists between the variables at the level of significance α = 0.05. However, the strength and direction of the association is unknown.

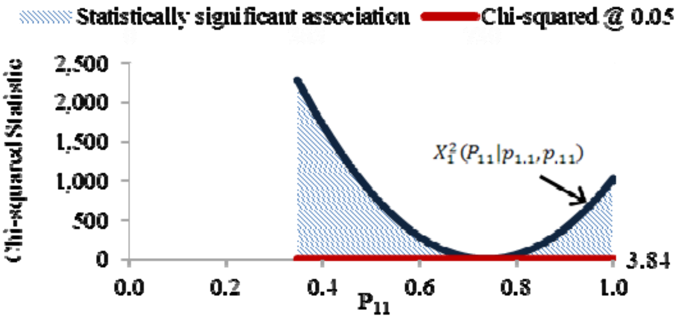

Figure 2. AAI (P_{11}) graphical presentation of electorate 1, 1893

Figure 2 shows that P_{11} of electorate 1 is bounded by the lower limit $L_{P_{11}}$ = 0.35 and the upper limit $U_{P_{11}}$ = 1 and that the Pearson chi-squared statistic is maximised at these bounds. Similarly, the AAI of electorate 1 (1893) can also be calculated in terms of the standardised residual Z_1 by using (16) and from (13) its bounds are $L_{Z_1} = -0.29$ and U_{Z_1} = 0.19 as shown in Figure 3. Note that at the 5% level of significance the value of AAI is 99.37 and constant because the calculation concept of the AAI remains the same when implementing the linear transformation (discussed in section 4). The only difference between Figure 2 and Figure 3 is in the horizontal scale where the original scaling system in terms of $P_{11}$is converted into Z_1. By doing so, the relationship between the AAI and the association measurement is established and the AAI remains unchanged.

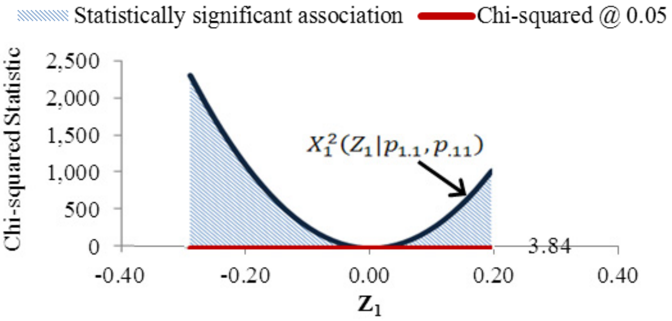

Figure 3. AAI (Z_1) graphical presentation of electorate 1, 1893

The underlying theory described here makes it possible to determine the likely strength of the association between voting patterns and gender given only the aggregate data. Since we know that when $Z_1 = 0$ the two dichotomous variables are independent, the further the limits of Z_1 are away from 0, the more statistically significant the association is likely to be. This outcome is defined by comparing the L_{Z_1} and U_{Z_1} with (14).

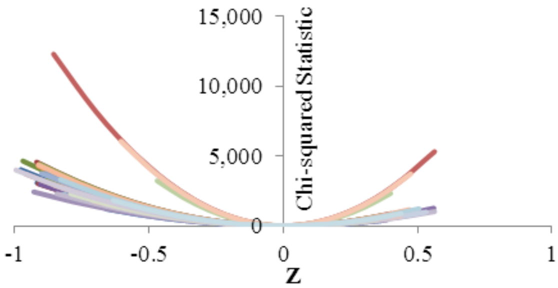

Figure 4. Multiple AAI (Z_g) curves from 57 electorates, 1893

The direction of the association can be based on the sign of Z_1 and the area under the curve for $Z_1 > 0$ (which reflects a positive association) and $Z_1 < 0$ (which reflects negative association). Figure 3 shows that the association between voting patterns and gender within electorate 1 is more likely to be negative (note that this is different from the result in section 6.1 as we consider the aggregate data here). Additionally, one may simultaneously consider the AAI (Z_g) curves for all 57 electorates in the 1893 election. For example, Figure 4 illustrates how the AAI curves for 57 electorates (1893) are presented, and shows that for most of the electorates, the association is likely to be negative.

7. DISCUSSION

Given only the aggregate data, or marginal information, of a 2x2 contingency table, this paper shows that the AAI can provide the same result for testing the statistical significant association between the two dichotomous categorical variables irrespective of the association measure considered.. Moore (2005) confirmed that gender was a significant factor at New Zealand elections from 1893. On the other hand, it is also possible to establish a relationship between the traditional AAI with classic association measurements such as the standardised residual to obtain better understanding of the association in term of strength and direction. In other words, this new development extends the purpose of the AAI from only justifying if an association exists, it is now possible to interpret how strong or weak the association is and which direction (positive or negative) of the association is likely to be.

Future developments of the AAI can be made by investigating how to combine multiple AAI (Z_g) curves from each electorate into a single index for an election year. This may well allow us to compare the trends between politics and gender among different NZ elections (from 1893 to 1919) and provide better perspective when performing ecological inference. Discussions of this aspect of aggregate data analysis can be found by referring to, for example, King (1997) and King et al. (2004). For a study of this issue to the NZ voting data of 1893-1919, refer to Hudson, Moore, Beh and Steel (2005, 2010). In addition, the effect of sample size to the result should be examined and it is useful to extend the connection of the traditional AAI with other well-known association measurements such as adjusted standardised residual, Pearson's ratio, contingency, and correlation. Further generalisations for non-linear transformations of p_{11}, including the odds ratio (Beh, Tran & Hudson, 2013) can also be considered.

REFERENCES

Agresti, A. (2002). "Categorical Data Analysis" (Second edition). John Wiley and Sons.

Beh, E.J. (2008). "Correspondence analysis of aggregate data: The 2x2 table", Journal of Statistical Planning and Inference, 138, 2941-2952.

Beh, E.J. (2010). "The aggregate association index", Computational Statistics and Data Analysis, 54, 1570 – 1580.

Beh, E.J., Tran, D., Hudson, I.L. (2013). A reformulation of the aggregate association index using the odds ratio. Computational Statistics & Data Analysis (in press)

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

Duncan, O.D., Davis, B. (1953). "An alternative to ecological correlation", American Sociological Review, 18, 665-666.

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

Goodman, L. A. and Kruskal, W. H. (1954). "Measures of association for cross classifications", Journal of the American Statistical Association, 49, 732– 764.

Haberman, S. J. (1973). "The Analysis of Residuals in Corss-Classified Tables", Biometrics, Vol.29, No.1, 205-220.

Hudson, I.L. , Moore, L., Beh, E.J., Steel, D.G. (2005). "Gendered counts of historical voting in NZ 1893 – 1919: A rigorous statistical ecological inference approach, in 55th Session of the International Statistical Institute (ISI) (Invited Special Session), Sydney, April 5-12, 1-4.

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.

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

King, G., Rosen, O., Tanner, M.A. (2004). "Ecological Inference: New Methodological Strategies", Princeton University Press: Princeton, USA.

Moore, L. (2005). Was gender a factor in voter participation at New Zealand elections?, in *Class, Gender and the Vote* (eds M. Fairburn, E. Olssen), Otago University Press: Otago, NZ, 129-142.