Duy Tran, University of Newcastle – The Aggregate Association Index and its application in the New Zealand election 1893-1919

Session 3A, 1.30 pm Thursday

The Aggregate Association Index and its Application in the 1893 New Zealand Election

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

UNIVERSITY OF NEWCASTLE, NSW, AUSTRALIA
C3141292@uon.edu.au

Abstract

Politics and gender have always been a hot topic for at least four decades, especially for politicians, social scientists, feminists and historians. Numerous statistical methods have been developed to try to establish and understand if any difference occurs between men's and women's voting patterns. However, due to confidentiality issues or policies in many real-life situations, only aggregate information is often available. As a result, a significant amount of valuable information about politics and gender may have been lost over the years.

The objective of this paper is to consider the development and application of a new method called Aggregate Association Index (AAI), which can be applied in such situations of gendered aggregate political data. Moreover, these two aspects of the AAI shall be examined by considering the New Zealand election data in 1893 and this may well help us establish the future development for the index.

Keywords: aggregate data, aggregate association index, New Zealand historical voting data

1. Introduction

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).

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 1: Summary of the 11 NZ elections, 1893-1919

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 1 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 2 which provides a summary of the men and women who registered to vote in electorate 1 of the 1893 election.

1st electorate 1983	Vote	No vote	Total
Women	1,443	289	1,732
Men	1,747	842	2,589
Total	3,190	1,131	4,321

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

The aim of this paper is to introduce a new index - the Aggregate Association Index, or just AAI (Beh, 2008) - to analyse a 2x2 contingency table when only the marginal totals are available. In particular, we will discuss how the index, and its application, can help uncover voting patterns in the NZ electorate voting data. Since the cell frequencies are known, the appropriateness of the AAI for analysing aggregate voting data can also be assessed.

2. The 2x2 Contingency Table

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 $p_{i.g}$ and $p_{.jg}$.

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 following three questions:

1. Is there enough evidence to suggest an association between the categorical variables?
2. If the variables are associated, how can we measure or quantify the association among them?
3. If the variables are associated, how can we visualise their association?

The first question can easily be achieved by examining the Pearson chi-squared statistic calculated from the counts and margins of a contingency table. In the case of the g^{th} electorate, the Pearson chi-squared statistic can be considered:

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

For the second question, the Pearson product moment correlation can be used to determine the direction and magnitude of the association.

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

The correlation coefficient ranges from -1 to 1. A value of ±1 implies that a perfect agreement or disagreement, while 0 implies no relationship.

For the third question, Correspondence Analysis (CA) can be carried out to produce a graphical presentation of the existing association between the variables.

The three questions above can be easily answered if the cell values of the contingency table are known. However, due to reasons concerned with addressing confidentiality issues or because the data was not collected at the time of the study, only aggregate data may be available and this makes it difficult to answer the three questions. Beh (2008) derived the Aggregate Association Index and showed that these three questions can be answered when only marginal information from the 2x2 tables is available. In the next section, this index will be described and applied to analyse the NZ voting data.

3. The Aggregate Association Index (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'. Beh (2008, 2010) showed that the Pearson chi-squared statistic, (1), can be expressed as a function of P_{1g} and the marginal proportions for the g^{th} electorate by:

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

where $p_{1.g} = n_{1.g}/n_g$, $p_{2.g} = n_{2.g}/n_g$, $p_{.1g} = n_{.1g}/n_g$, and $p_{.2g} = n_{.2g}/n_g$. It can be seen that the chi-squared statistic is a (concave down) quadratic function in terms of the conditional proportion P_{1g}. When only margins are known, the value of P_{1g} is well understood to lie within the interval (Duncan and Davis, 1953):

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

Since L_{1g} and U_{1g} only depend on the marginal information, the chi-squared statistic $\chi_g^2\left(P_{1g}|p_{1.g}, p_{.1g}\right)$ can also be investigated by using only the margins.

By taking into account the above properties of P_{1g}, Beh (2008) proposed the **Aggregate Association Index**. For the g^{th} electorate, this index is defined as :

$$A_{\alpha,g} = 100\left(1 - \frac{\left[(L_{\alpha g}-L_{1g})+(U_{1g}-U_{\alpha g})\right]\chi_\alpha^2 + Int(L_{\alpha g}, U_{\alpha g})}{Int(L_{1g}, U_{1g})}\right) \quad (5)$$

where $Int(a,b) = \int_a^b \chi_g^2\left(P_{1g}|p_{1.g}, p_{.1g}\right) dP_{1g}$.

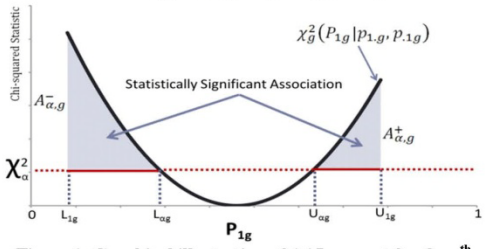

Figure 1: Graphical illustration of AAI concept for the g^{th} electorate

For a given α level of significance, this index is the ratio of the total region that lies under the curved defined by $\chi_g^2\left(P_{1g}|p_{1.g}, p_{.1g}\right)$ but above the critical value. Hence, given only the marginal information, 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).

Beh (2010) also shows that the AAI index can also be partitioned such that $A_{\alpha,g} = A^{-}_{\alpha,g} + A^{+}_{\alpha,g}$ where:

$$A^{-}_{\alpha,g} = \frac{\int_{L_{1g}}^{L_{\alpha g}} [\chi^2_g(P_{1g}|p_{1.g}, p_{.1g}) - \chi^2_\alpha] dP_{1g}}{\int_{L_{1g}}^{U_{1g}} \chi^2_g(P_{1g}|p_{1.g}, p_{.1g}) dP_{1g}} \qquad (6)$$

is referred to as aggregate negative association index and

$$A^{+}_{\alpha,g} = \frac{\int_{U_{\alpha g}}^{U_{1g}} [\chi^2_g(P_{1g}|p_{1.g}, p_{.1g}) - \chi^2_\alpha] dP_{1g}}{\int_{L_{1g}}^{U_{1g}} \chi^2_g(P_{1g}|p_{1.g}, p_{.1g}) dP_{1g}} \qquad (7)$$

is referred to as aggregate positive association index. Furthermore, it is also possible to define the Pearson product moment correlation, (2), in term of P_{1g} and the marginal information to determine the direction of the association – see Figure 2:

$$r_g(P_{1g}|p_{1.g}, p_{.1g}) = \left(\frac{P_{1g} - p_{.1g}}{p_{2.g}}\right)\sqrt{\frac{p_{1.g}p_{2.g}}{p_{.1g}p_{.2g}}} \qquad (8)$$

Visualising the association between the variables can also be achieved by considering a graphical display from performing a correspondence analysis (Greenacre, 1984) on the 2x2 table – such a plot, referred to as a correspondence plot, will consist of only a single dimension. Beh (2008) showed that, for such a plot, the coordinates associated with the row and column categories can also be expressed in terms of P_{1g} and the table's margins. An illustration of such a plot will be discussed further in the following section.

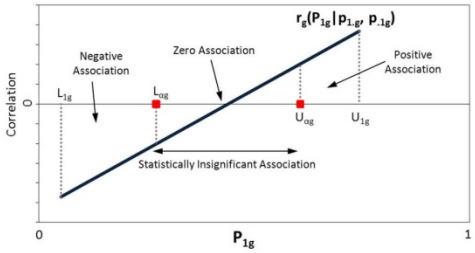

Figure 2: Pearson correlation derived from the AAI

4. The AAI and NZ election data

Consider again Table 2 where the cells are assumed to be known. A chi-squared test of independence gives a test statistic of 134.68 and a p-value <0.0001. Thus, there is ample evidence to suggest that there exists a significant association between the gender and voting patterns. In addition, the direction of this association may also be determined by considering (2); $r_1 = +0.18$ which suggests that the association is likely to be weak and positive. That is, women who were registered are more likely to vote than registered men, while registered men are more likely not to vote than registered women.

A visual representation of this association can be made by considering a correspondence analysis of Table 2. Figure 3 provides a graphical representation of the association between the rows and columns of Table 2.

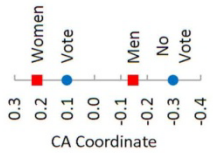

Figure 3: Correspondence plot for electorate 1, 1893

It can be seen from Figure 3 that registered female voters are more likely to vote than registered male voters, whereas registered male voters are more likely not to vote than their female counterparts. Note that such an interpretation reflects the direction of the correlation coefficient.

Consider now the case where the joint cell frequencies of Table 2 are not known so that the analysis of association is undertaken by considering only the marginal information. For the 1st electorate of the 1893 election, the AAI $(\alpha = 0.05)$ 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 α level of significance.

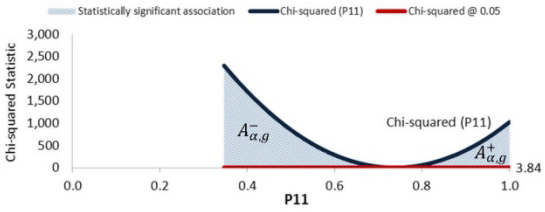

Figure 4: AAI graphical presentation of electorate 1, 1893

Figure 4 shows that P_{11} for electorate 1 is bounded by the lower limit $L_{11} = 0.35$ and the upper limit $U_{11} = 1$ and that the Pearson chi-squared statistic is maximised at these bounds.

Partitioning $A_{0.05,1}$ produces an aggregate negative association index $A^{-}_{0.05,1} = 76.58$ and an aggregate positive association $A^{+}_{0.05,1} = 22.79$. Hence, the overall

direction of the above association is more likely to reflect a negative association than a positive one. This is reflected in Figure 5 where the triangular region on the left hand side of $P_{11} = p_{.11}$ is larger than the triangular region on its right hand side. Recall that based on the cell frequencies of Table 2, the correlation of the variables was weak, but positive.

The apparent contradiction in the direction of the association can be attributed to the fact that the AAI reflects the contingency tables that may be derived given the specified marginal information. Given this combination of marginal frequencies it is more likely to derive a table that has a negative association than a positive one. On the other hand, (2) reflects one possible contingency table that can be derived given the marginal totals of Table 2.

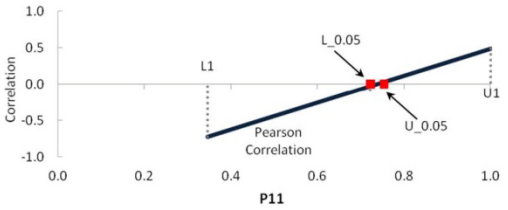

Figure 5: Pearson correlation for electorate 1, 1893

Figure 6 is a generalisation of the correspondence plot given by Figure 3 when only the marginal information is available. Therefore, the P_{11} axis is included to reflect the possible cell values that are possible and the diagonal lines on this plot reflect the variation in the position of the row and column coordinates as P_{11} varies within the Duncan and Davis (1953) of (4) – the circle and square dots of this plot reflect the position of the coordinates when the joint cell frequencies of Table 2 are known. Figure 6 shows that irrespective of the value of P_{11} (and hence n_{11}) registered women tend to vote more than registered men while registered men are again more likely not to vote than registered women.

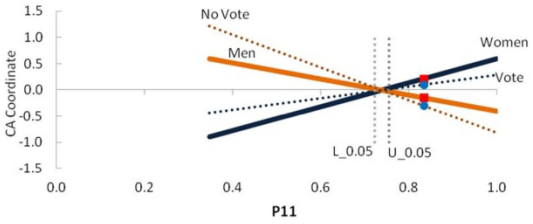

Figure 6: Extended CA plot for electorate 1, 1893

Moreover, one may simultaneously consider the AAI curves for all 57 electorates in the 1893 election. For example, Figure 7 illustrates how the AAI curves for these 57 electorates compare. It shows at least four clusters of homogenous 2x2 tables with four individual electorates (electorate 32, 38, 45, and 54) whose AAI curves are very different. Further investigations of these clusters can be made, but will not be undertaken further in this paper.

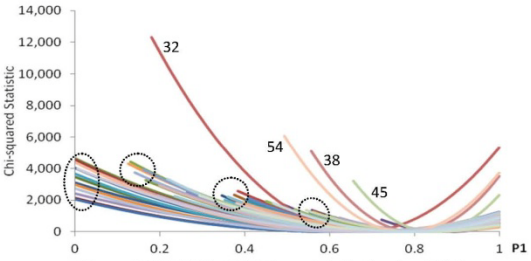

Figure 7: Multiple AAIs from 57 electorates, 1893

5. Discussion and Future Development

Given only the marginal information of a 2x2 table, this paper shows that AAI can provide the same result for testing a significant association between two dichotomous categorical variables: gender and voting patterns as with the typical chi-squared test for independence of a complete 2x2 table. It is also possible to quantify the association directions of the AAI and visualise how the variables are associated by extending the standard CA plot.

Future development of the AAI can be made including investigating how to combine the AAI for each electorate into a single index for an election year. One may also broaden this concept when considering the Pearson correlation line (such as Figure 5) and the correspondence plot (Figure 6) for each electorate. 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).

References

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

[2] E.J. Beh, "The aggregate association index", *Computational Statistics and Data Analysis*, 54, 1570 - 1580, 2010.

[3] I.L. Hudson, L. Moore, E.J. Beh, D.G. Steel, "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, 2010.

[4] I.L. Hudson, L. Moore, E.J. Beh, D.G. Steel, "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, pp 1-4, 2005.

[5] O.D. Duncan, B. Davis, "An alternative to ecological correlation", *American Sociological Review*, 18, 665-666, 1953.

[6] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press: London, 1984.

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

[8] G. King, O. Rosen, M.A. Tanner, *Ecological Inference: New Methodological Strategies*, Princeton University Press: Princeton, USA, 2004.

[9] L. Moore, 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, pp 129-142, 2005.