The Aggregate Association Index

Eric J. Beh

School of Mathematical and Physical Sciences
University of Newcastle, Australia

COMPSTAT 2010, Paris, France – August 24
The 2x2 Contingency Table

Cross-classify a sample of size n according to two dichotomous variables

<table>
<thead>
<tr>
<th></th>
<th>Column 1</th>
<th>Column 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>( p_{11} )</td>
<td>( p_{12} )</td>
<td>( p_{1.} )</td>
</tr>
<tr>
<td>Row 2</td>
<td>( p_{21} )</td>
<td>( p_{22} )</td>
<td>( p_{2.} )</td>
</tr>
<tr>
<td>Total</td>
<td>( p_{.1} )</td>
<td>( p_{.2} )</td>
<td>1</td>
</tr>
</tbody>
</table>

“Let us blot out the contents of the table, leaving only the marginal frequencies . . . [they] by themselves supply no information on . . . the proportionality of the frequencies in the body of the table . . . ”

– Fisher (1935)

Define

\[ p_1 = \frac{p_{11}}{p_{1.}} \]

\[ x^2(P_1 \mid p_{.1}, p_{.1}) = n \left( \frac{P_1 - p_{.1}}{p_{2.}} \right)^2 \left( \frac{p_{1.}p_{2.}}{p_{1.}p_{2.}} \right) \]
Bounds of $P_1$

Duncan & Davis (1953) Bounds

$$L_1 = \max \left( 0, \frac{n_\cdot_1 - n_\cdot_2}{n_\cdot_1} \right) \leq P_1 \leq \min \left( \frac{n_\cdot_1}{n_\cdot_1}, 1 \right) = U_1$$

100(1 – $\alpha$)% Confidence Bounds

$$L^*_\alpha = p_\cdot_1 - p_\cdot_2 \sqrt{\frac{\chi^2_{\alpha}}{n} \left( \frac{p_\cdot_1 p_\cdot_2}{p_\cdot_1 p_\cdot_2} \right)} < P_1 < p_\cdot_1 + p_\cdot_2 \sqrt{\frac{\chi^2_{\alpha}}{n} \left( \frac{p_\cdot_1 p_\cdot_2}{p_\cdot_1 p_\cdot_2} \right)} = U^*_\alpha$$

$$L_\alpha = \max \left( 0, L^*_\alpha \right) < P_1 < \min \left( 1, U^*_\alpha \right) = U_\alpha$$
Aggregate Association Index (AAI)

If the area under $X^2(P_1)$ but above $\chi^2_\alpha$ is large than there may be evidence to suggest that there is a significant association (at the $\alpha$ level of significance) between the two dichotomous variables.
Aggregate Association Index (AAI)

\[ A_\alpha = 100 \left( 1 - \frac{\left[ (L_\alpha - L_1) + (U_1 - U_\alpha) \right] \chi^2_\alpha + \int_{L_\alpha}^{U_\alpha} X^2 \left( P_1 \mid p_{i1}, p_{.1} \right) dP_1}{\int_{L_1}^{U_1} X^2 \left( P_1 \mid p_{i1}, p_{.1} \right) dP_1} \right) \]
Example – Fisher’s Twin Data

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.

<table>
<thead>
<tr>
<th></th>
<th>Convicted</th>
<th>Not convicted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monozygotic</td>
<td>10</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Dizygotic</td>
<td>2</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>18</td>
<td>30</td>
</tr>
</tbody>
</table>

Pearson chi-squared statistic is 13.032.

- p-value = 0.0003 → there is evidence of a strong association between the two variables.
- The product moment correlation = 0.6591 → positive association
Example – Fisher’s Twin Data

<table>
<thead>
<tr>
<th></th>
<th>Convicted</th>
<th>Not convicted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monozygotic</td>
<td>10</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Dizygotic</td>
<td>2</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12</strong></td>
<td><strong>18</strong></td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>

But, as Fisher (1935) did, suppose we “blot out” the cells of the table.

**Question**: What information do the margins provide in understanding the extent to which the variables are associated.

**We shall calculate the aggregate association index**
Example – Fisher’s Twin Data

\[ A_{0.05} = 61.83 \]

If we consider the 5% level of significance, the margins provide strong evidence that there may exist a significant association between twin type & conviction status.

\[ X^2(P_1) = \frac{221}{216} \left( \frac{30P_1 - 12}{17} \right)^2 \quad \text{where } 0 \leq P_1 \leq 0.9231 \]
Direction of the Association

\[ A_{\alpha} = A_{\alpha}^+ + A_{\alpha}^- \]
Therefore based solely on the marginal information we can determine that the variables are three times more likely to be positively associated than negatively associated.
Discussion

- The index provides an indication of the extent to which two dichotomous variables are statistically significantly associated given only the marginal information.

- Index is not meant to infer the individual level correlation of the variables, but to provide a measure reflecting how likely the two variables may be associated.

Further Issues:

- Investigate the applicability of index for $G (>1)$ 2x2 tables, including incorporating covariate information (ecological inference).

- Has links with the correspondence analysis of aggregate data.

- Link with Fisher’s exact test.