THE UNIVERSITY OF NEWCASTLE AUSTRALIA

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

	Column 1	Column 2	Total
Row 1	p_{11}	p_{12}	$p_{1\bullet}$
Row 2	p_{21}	p_{22}	$p_{2\bullet}$
Total	$p_{\bullet 1}$	$p_{\bullet 2}$	1

"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

$$P_1 = \frac{p_{11}}{p_{1\bullet}}$$

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

Bounds of P_1

Duncan & Davis (1953) Bounds

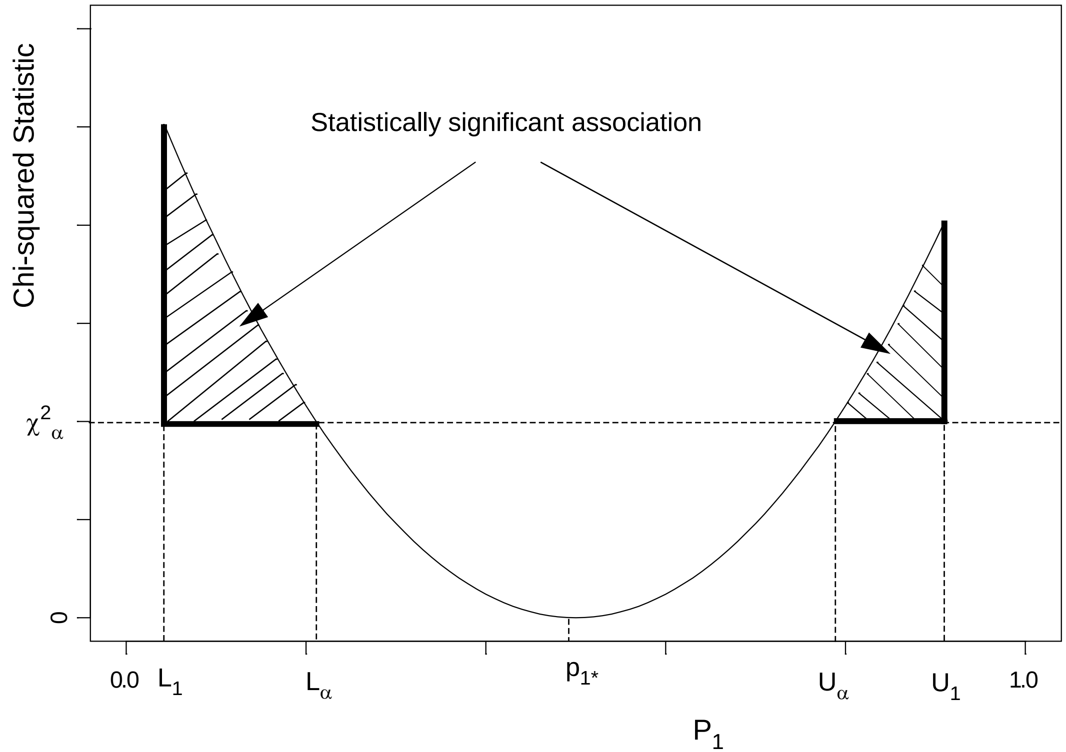
$$L_1 = \max\left(0, \frac{n_{\bullet 1} - n_{2\bullet}}{n_{1\bullet}}\right) \le P_1 \le \min\left(\frac{n_{\bullet 1}}{n_{1\bullet}}, 1\right) = U_1$$

100(1 – α)% Confidence Bounds

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

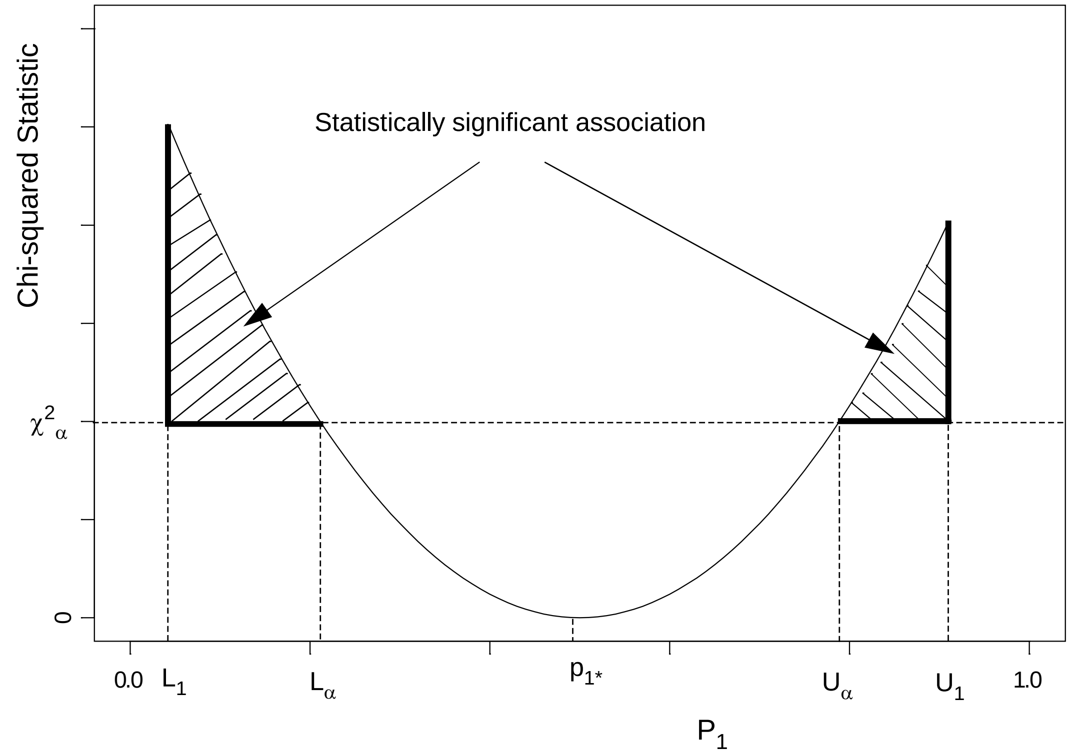
$$L_\alpha = \max(0, L_\alpha^*) < P_1 < \min(1, U_\alpha^*) = U_\alpha$$

Aggregate Association Index (AAI)



If the area under $X^2(P_1)$ but above χ^2_α is large than there may be evidence to suggest that there is a significant association (at the α level of significance) between the two dichotomous variables.

Aggregate Association Index (AAI)



$$A_\alpha = 100\left(1 - \frac{\left[\left(L_\alpha - L_1\right) + \left(U_1 - U_\alpha\right)\right]\chi^2_\alpha + \int_{L_\alpha}^{U_\alpha} X^2\left(P_1 \mid p_{1\bullet}, p_{\bullet 1}\right) dP_1}{\int_{L_1}^{U_1} X^2\left(P_1 \mid p_{1\bullet}, p_{\bullet 1}\right) dP_1}\right)$$

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.

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

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

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

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

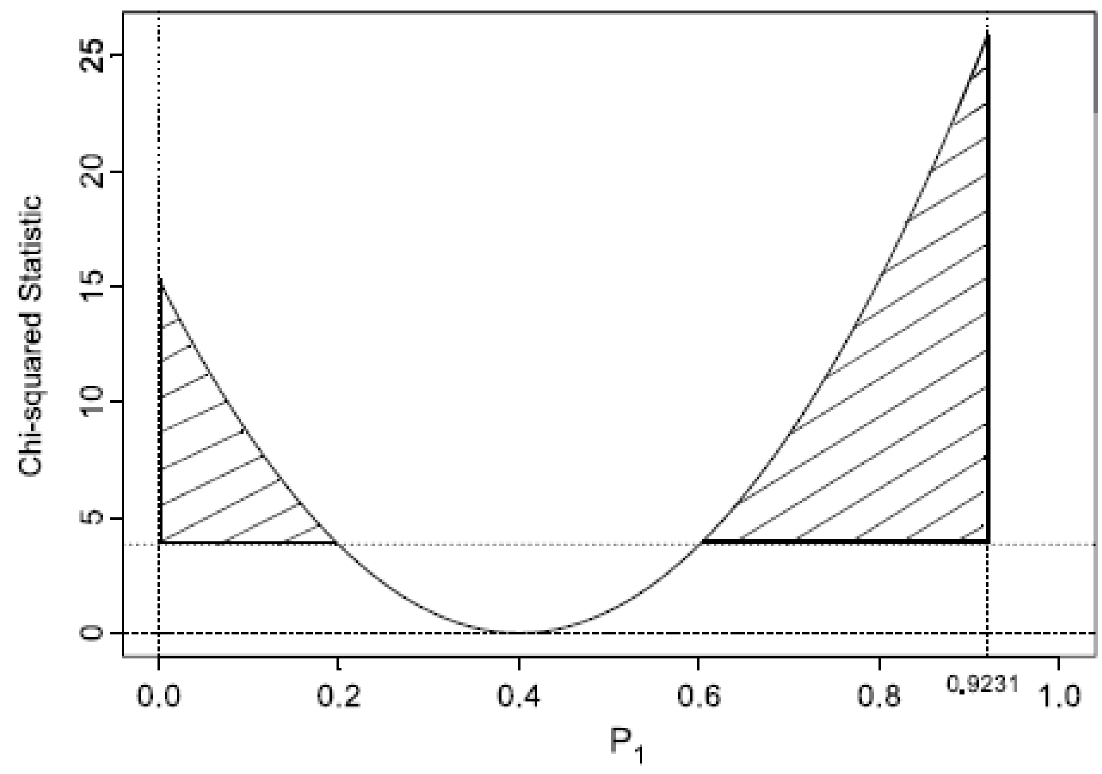
Question: What information do the margins provide in understanding the extent to which the variables are associated.

We shall calculate the aggregate association index

Example – Fisher's Twin Data

$A_{0.05} = 61.83$

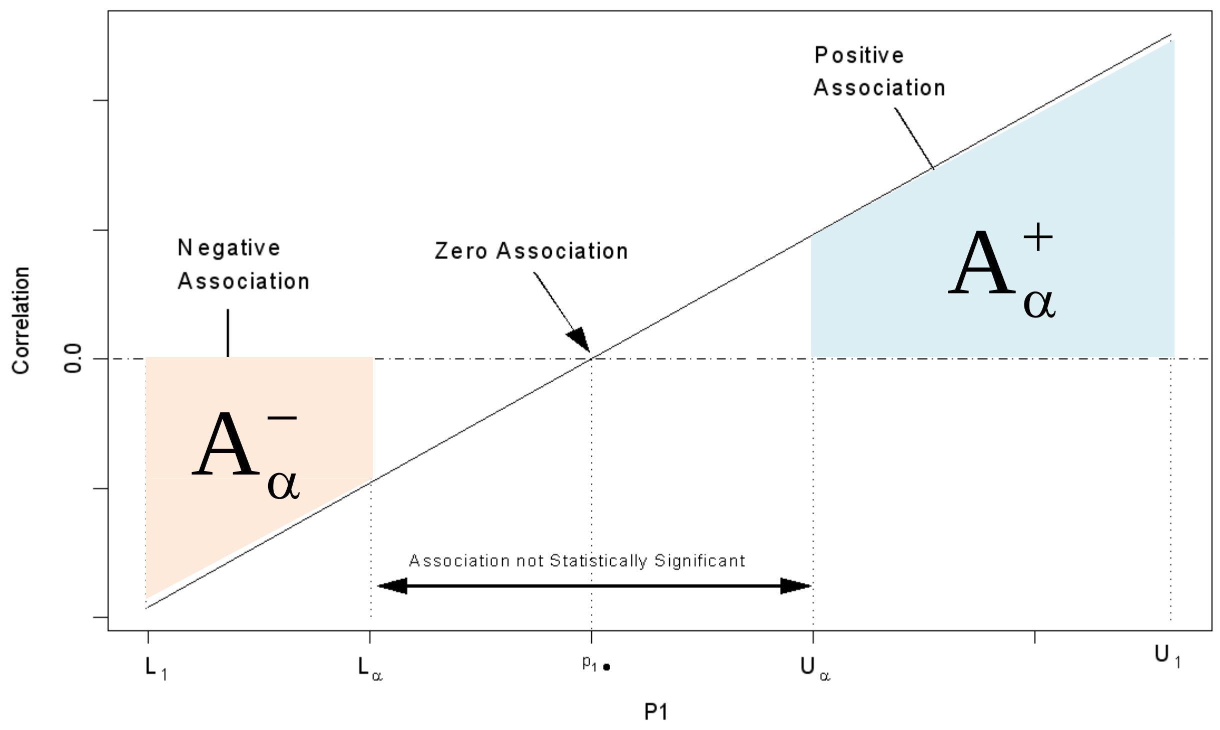
If we consider the 5% level of significance, the margins provide **strong evidence that there *may* exist a significant association** between twin type & conviction status



$$X^2(P_1) = \frac{221}{216}\left(\frac{30P_1 - 12}{17}\right)^2$$

where $0 \le P_1 \le 0.9231$

Direction of the Association



$$A_\alpha = A_\alpha^+ + A_\alpha^-$$

Fisher's Twin Data (…revisited)

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

$$A_{0.05} = 61.83$$

$$A^{+}_{0.05} = 46.43 \quad A^{-}_{0.05} = 15.40$$

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 UNIVERSITY OF NEWCASTLE AUSTRALIA

- The index provides an indication of the extent to which two dichotomous variables are statistically significantly association 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