On the quantification of statistical significance of the extent of association projected on the margins of 2x2 tables, when only the aggregate data is available: A pseudo p-value approach – applied to leukaemia relapse data

[a]S. A. Cheema, E. J. Beh and I. L. Hudson

[a]*School of Mathematical and Physical sciences, University of Newcastle, Callaghan campus, NSW, 2308, Australia*
E-mail: salman.cheema@uon.edu.au

Abstract: Aggregate data arises in situations where survey research or other means of collecting individual-level data are either infeasible or inefficient. The recent increasing use of aggregate data in the statistical and allied fields – including epidemiology, education and social sciences – has arisen due to number of reasons. These include the questionable reliability of estimates when sensitive information is required, the imposition of strict confidentiality policies on data by government and other organisational bodies and in some contexts it is impossible to collect the information that is needed. In this paper we present a novel approach to quantify the statistical significance of the extent of association that exists between two dichotomous variables when only the aggregate data is available. This is achieved by examining a newly developed index, called the *aggregate association index* (or the AAI), developed by Beh (2008 and 2010) which enumerates the overall extent of association about individuals that may exist at the aggregate level when individual level data is not available.

The applicability of the technique is demonstrated by using leukaemia relapse data of Cave et al. (1998). This data is presented in the form of a contingency table that cross-classifies the follow up status of leukaemia relapse by whether cancer traces were found (or not) on the basis of polymerase child reaction (PCR) – a modern method used to detect cancerous cells in the body assumed superior than conventional for that period, microscopic identification.

Assuming that the joint cell frequencies of this table are not available, and that the only available information is contained in the aggregate data, we first quantify the extent of association that exists between both variables by calculating the AAI. This index shows that the likelihood of association is high. As the AAI has been developed by exploiting Pearson's chi-squared statistics, the AAI inherently suffers from the well-known large sample size effect that can overshadow the true nature of the association shown in the aggregate data of a given table.

However, in this paper we show that the impact of sample size can be isolated by generating a pseudo population of 2x2 tables under the given sample size. Therefore, the focus of this paper is to present an approach to help answer the question "*is this high AAI value statistically significant or not?*" by using aggregate data only. The answer to this question lies we believe, in the calculation of the p-value of the nominated index. We shall present a new method of numerically quantifying the p-value of the AAI thereby gaining new insights into the statistical significance of the association between two dichotomous variables when only aggregate level information is available. The pseudo p-value approach suggested in this paper enhances the applicability of the AAI and thus can be considered a valuable addition to the literature of aggregate data analysis.

Keywords: *Aggregate data, Aggregate Association Index, pseudo p values, Ecological inference, sample size*

1. INTRODUCTION

Aggregate data analysis often involves the quantification of the association of categorical variables at the individual level when only grouped or aggregate level data is available. This issue is of relevance to a variety of disciplines that span the social, physical, political and health sciences; see for example, Steel et al. (2004) and Salway and Wakefield (2004) and Hudson et al. (2010). While the analysis of aggregate data now spans at least 100 years, there has been an explosion of research in this area over the last two decades. As noted by Achin and Shively (1995), King (1997) and Salway and Wakefield (2004), the number of scholars working on the development of methodologies to analyse this specific data structure is now larger than at any time in the history of the statistical and allied disciplines. Recently, researchers from a variety of disciplines, such as political science, economics, geography, public health and statistics have had no choice but to make inferences using only aggregate information; see King et al. (2004) and Hudson et al. (2005 and 2010).

To date the most dominant techniques designed to analyse aggregate data fall within the area of study commonly referred to as *ecological inference (EI)*; see King (1997). The techniques that lie within EI aim to reconstruct the individual level data (or some function of it) from the aggregate data. Therefore, ecological inference may be viewed as a special case of statistical inference wherein the main difficulties surrounding EI involve the loss of information about the individual data by the aggregation process. Debate over various proposed solutions to EI have appeared in the literature. The general tone of the debate is more critical than glowing; see for example, Rivers (1998), Cho and Gaine (2004) and Hudson et al. (2010). It has been identified that the success of EI depends heavily upon the lost information and the process of recapturing such lost information; see for example Quinn (2004) and Greiner and Quinn (2009). Furthermore, Schuessler (1999) and more recently Hudson et al. (2010 pg, 198) questioned the extent to which EI techniques are sensitive to assumptions, and they noted that

> *"the choice of model is particularly important in EI. Do the assumptions fit the data?"*

Therefore, with the motivation of developing more robust techniques towards lost information, Beh, (2008, 2010) introduced the novel aggregate association index (AAI). The basic principal of the AAI differs from the suite of EI techniques. Instead of reconstructing the individual level data directly (or some function of it), which is the primary aim of EI, the AAI quantifies the overall extent of association between two dichotomous, categorical variables using only the aggregate data. A major advantage of the AAI over traditional EI techniques is that the AAI does not heavily rely on any assumptions about the individual level data. Cheema, Beh and Hudson (2013), Beh et al. (2013, 2015) and Beh, Tran and Hudson (2014), have explored many different aspects of the AAI.

This paper is devoted to answering the question *"how large should the AAI be to be sure that it is a significant measure?"* To quantify the statistical significance of the value of the AAI, we calculate the probability of observing that AAI (or one that is more extreme) thereby quantifying a pseudo (Monte Carlo like) p-value measure. This is obtained by generating a pseudo population of AAI's under some given conditions. The applicability of our technique is demonstrated by analysing the leukaemia relapse data of Cave et al. (1998) which presents a cross-classification of polymerase child reaction (PCR) status and follow up status of relapse (or otherwise) of cancerous cell in a sample of 178 children.

This paper is sub-divided into four major parts. The focus of section 2 is to introduce the AAI and the notation used throughout the paper. In section 3, we present the methodological rationale of our pseudo p-value approach to quantify the statistical significance of the extent of association mirrored in the AAI when only aggregate data is available. Section 4 introduces the data of Cave et al. (1998) and this data for a for demonstration of our novel technique.

2. THE AGGREGATE ASSOCIATION INDEX (AAI)

2.1 Notation

Suppose two variables, A and B, are classified as row and column variables respectively, where the first row category is assigned the generic label A_1 and the second row category is given the generic label A_2. Similarly, the first and second column categories are B_1 and B_2, respectively.

Table 1. A general 2x2 contingency table

A|B	B₁	B₂	Total
A₁	n_{11}	n_{12}	$n_{1.}$
A₂	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Table 1 presents the general form of a 2x2 contingency table, where n is the total number of objects or individuals in the study. Let n_{ij} denote the number in the sample that is classified into the (i, j)th cell with the proportion of the sample allocated into this cell denoted as $p_{ij} = n_{ij}/n$, for $i = 1, 2$ and $j = 1, 2$ such that $\sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1$. Thus, n_{11} denotes the number of objects/individuals that have been classified in the first row and the first column category and p_{11} is the proportion of those objects/individuals in the sample that have been classified into this cell. The frequency of the i'th row margin and j'th column margin is denoted by $n_{i.} = \sum_{j=1}^{2} n_{ij}$ and $n_{.j} = \sum_{i=1}^{2} n_{ij}$ respectively, such that, $\sum_{i=1}^{2} n_{i.} = \sum_{j=1}^{2} n_{.j} = n$. Define $p_{i.} = n_{i.}/n$ and $p_{.j} = n_{.j}/n$ to be the i'th row and j'th column marginal proportion respectively, such that $\sum_{i=1}^{2} p_{i.} = \sum_{j=1}^{2} p_{.j} = 1$. Suppose we now consider $P_1 = n_{11}/n_{1.}$, which is the conditional probability of an individual/unit being classified into "Column 1" given that it has been classified into "Row 1".

When the individual level information (or joint cell values) of a 2x2 contingency table are unknown, Duncan and Davis (1953) provided lower and upper bounds for the (1, 1)'th cell frequency as,

$$A_1 = \max(0, n_{.1} - n_{.2}) \leq n_{11} \leq \min(n_{.1}, n_{1.}) = B_1. \tag{1}$$

It is easy to verify that the bounds of equation (1) can be written in terms of P_1 such that

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

Beh (2008) showed that when only the marginal information is available, and a test of the association between the two dichotomous variables is performed at the α level of significance, the bounds of P_1 can be narrowed to

$$L_\alpha = p_{.1} - p_{2.}\sqrt{\frac{\chi_\alpha^2}{n}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)} < P_1 < p_{.1} + p_{2.}\sqrt{\frac{\chi_\alpha^2}{n}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)} = U_\alpha, \tag{3}$$

where, χ_α^2 is the $100(1 - \alpha)'$th percentile of the chi-squared distribution with 1 degree of freedom. Furthermore, Pearson's chi-squared statistics for a 2x2 contingency table (Table 1) is simplified in terms of individual cell frequencies and marginal totals such as

$$X^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \tag{4}$$

2.2 The Aggregate Association Index

Beh (2008, 2010) transformed Pearson's chi-squared statistic, as given by equation (4) so that it can be expressed as a function of P_1 and the marginal data. Equation (4) can alternatively, but equivalently, expressed as

$$X^2(P_1|p_{1.}, p_{.1}) = n\left(\frac{P_1 - p_{.1}}{p_{2.}}\right)^2\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right). \tag{5}$$

Note that the chi-squared statistic given by equation (5) depends, not only on the aggregate data and the sample size, but it also on the magnitude of the conditional proportion P_1. In turn, P_1 is bounded by equation (2), which is further narrowed to be within the bounds of equation (3) when a test of association is conducted using an α level of significance. Thus, using the expression of the chi-squared statistic given by equation (5), we are able to quantify the values of the statistic over the permissible range of the P_1. Like traditional inferential theory, by using the ranges in (2) and (3) and the statistic given by equation (5), we can now quantify the extent of the deviation of P_1 from its value under the hypothesis of independence, $p_{.1}$. This development enables the analyst to quantify all possible values of the chi-squared statistic for any and all P_1 values that fall within this interval. Therefore, given only aggregate data, the quadratic function given

by equation (5) can be visualised to yield a unique AAI curve over the permissible range of the P_1, which is depicted in Figure 1.
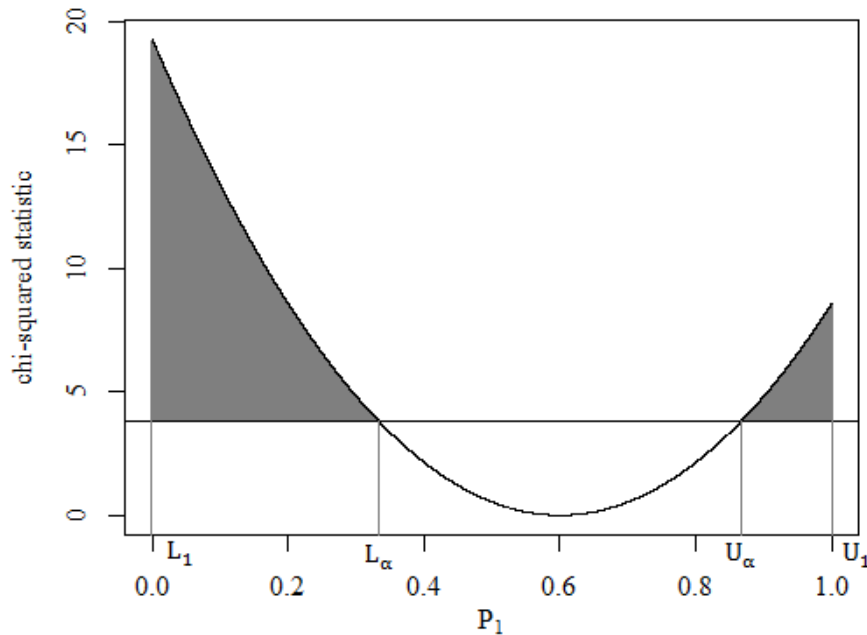

Figure 1. A general graphical display of the AAI

For the usual Pearson's chi-squared test of independence between two dichotomous variables, we reject the null hypothesis of independence at the α level of significance when the observed value of the statistic, $X^2(P_1|p_{1.},p_{.1})$, exceeds the critical value of $\chi_\alpha{}^2$ with one degree of freedom. Thus the region that reflects the strength of a statistically significant association between the variables is the area above the critical line defined by $\chi_\alpha{}^2$ but which lies under the AAI curve – see the shaded region in Figure 1. Quantifying this strength can be undertaken by calculating the proportion of the total area under the AAI curve, defined by equation (5), shown as shaded. Mathematically, the area of this shaded region, quantifying the extent of association between two dichotomous variables using the AAI, can be calculated as,

$$A_\alpha(P_1) = 100\left(1 - \frac{[(L_\alpha - L_1) + (U_1 - U_\alpha)]\chi_\alpha{}^2 + \int_{L_\alpha}^{U_\alpha} X^2(P_1|p_{1.},p_{.1})dP_1}{\int_{L_1}^{U_1} X^2(P_1|p_{1.},p_{.1})dP_1}\right). \tag{6}$$

By evaluating the definite integrals in equation (6), Beh (2010) provided the following simplified form of the AAI for a direct calculation of its magnitude:

$$A_\alpha(P_1) = 100\left[1 - \frac{\chi_\alpha{}^2[(L_\alpha - L_1) + (U_1 - U_\alpha)]}{kn[(U_1 - p_{.1})^3 - (L_1 - p_{.1})^3]} - \frac{(U_\alpha - p_{.1})^3 - (L_\alpha - p_{.1})^3}{(U_1 - p_{.1})^3 - (L_1 - p_{.1})^3}\right]. \tag{7}$$

In equation (7), $k = \frac{1}{3p_{2.}^2}\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right)$. The AAI is bounded by $0 \le A_\alpha(P_1) \le 100$. For a given level of significance, α, the AAI quantifies how likely a particular set of fixed marginal frequencies will enable the researcher to conclude that there exists a statistically significant association between two dichotomous variables. An AAI of $A_\alpha \approx 0$ indicates that the existence of a statistically significant association (at the α level of significance) is not at all likely. However, an AAI of $A_\alpha \approx 100$ suggests that it is highly likely that there is a statistically significant association.

3.   METHODOLOGY

A well-known characteristic of the AAI, as based on the Pearson's chi-squared statistic, is that its magnitude increases as the sample size increases. This is due to Pearson's chi-squared statistic being linearly related to the sample size; see equation (5). Everitt (1977, pg. 56) formally discusses the relationship between the chi-squared statistic and sample size, n. Therefore, the true nature of the association between two variables can be masked by the magnitude or effect of sample size; see Beh, et al. (2013, 2015). However, it may be that, for a given n, a large observed AAI is actually relatively small when compared with other values that the AAI can take for that given sample size. Our aim is therefore to quantify the significance of the AAI

magnitude while isolating the impact of sample size. This will allow us to obtain a clearer indication of the true extent of the association by controlling for the impact of the sample size and will better highlight the underlying structure of the association. To achieve this objective we propose a pseudo p-value approach by adopting the following strategy:

- The first step involves treating the sample size of Table 1 as fixed.

- For the fixed (and observed) sample size, n, we generate all possible row and column marginal totals by imposing the restriction that no marginal frequency is less than 1. By doing this we can generate a population of 2x2 tables for the specified sample size. Therefore, the maximum number of the contingency tables that can be studied for a given sample of size n is $(n-1)^2$.

- For each member of the population of $(n-1)^2$ tables, we calculate the AAI so as to obtain a population of all possible AAI quantities for the observed sample size. Furthermore, when the impact of the sample size is constant for all AAI values, any difference in the magnitude of the AAI will be due to the distribution of the row and column marginal frequencies.

- As the observed 2x2 contingency table is a realization of one possible table from the population of the all possible tables, we can therefore identify the proportion of 2x2 tables that have the same, or higher, AAI than the observed. Such a proportion, defined as the p-value of the AAI, can be quantified by

$$P(AAI \geq AAI_{obs}|n) \tag{8}$$

where AAI_{obs} is the observed AAI of the 2x2 table being analysed.

4.   APPLICATION

4.1   The Data

In the United States of America, 3000 to 4000 people are diagnosed with acute lymphoblastic leukaemia every year and two thirds of them are children. The standard criterion for assessing whether a child with leukaemia is in remission is when a doctor cannot detect any cancerous cells in the child's bone marrow through a microscope; see Cave et al. (1998) and Simonoff (2003, pg. 199). However a more sophisticated technique, polymerase child reaction (PCR) can detect as few as 5 cancer cells in every 100,000 cells, which is a significantly smaller amount than detectable by microscope. To verify the effectiveness of the new PCR technique, Cave et al. (1998) conducted a study examining 178 children who appeared to be in remission using the standard criterion. They found that 75 out of 178 patients were detected with traces of cancer and in a three year follow up, 30 out of 75 suffered a relapse. Of those children who did not appear with traces of cancer (103 children), 8 suffered a relapse. The 2x2 contingency table that summarises the findings of the leukaemia relapse study undertaken by Cave et al. (1998) is given by Table 2 and cross-classifies PCR status by follow up status for the sample of 178 patients.

Table 2. Cave et al. (1998) Leukaemia relapse data

PCR status	Follow up status		
	Relapse	No relapse	**Total**
Cancer traces	30	45	75
Cancer free	8	95	103
Total	**38**	**140**	**178**

4.2   Demonstration of the pseudo p-value Approach

In this section we investigate our method of calculating the probability of observing a specific value of the AAI, given by equation (8), for the observed sample size of the 2x2 contingency table (Table 2).

First we analyse the leukaemia data of Table 2 by assuming the case where the joint cell frequencies are known. The observed proportion, p_{11}, of children who were identified by PCR with traces of cancer and also suffered a relapse after three years is $30/178 = 0.169$. For this cell, the expected proportion, $p_{1.}p_{.1}$, if the hypothesis that PCR was not helpful in detecting traces of cancer is 0.090. The value of the Pearson's chi-squared test statistic, given by equation (4), is $X^2 = 26.85$. Furthermore, with 1 degree of freedom, the p-value of this statistic is less than 0.0001, showing that the there is a significant association between PCR status and a child's leukaemia status three years after treatment. Such conclusions were also made by Cave et al. (1998) and Simonoff (2003) in their analysis of the data.

Let us now consider the scenario where the cell frequencies of the Table 2 are unknown and we have only the marginal totals as available information. Note that with the absence of individual level data the joint cell proportion p_{11} and conditional proportion P_1 are inestimable, as are the usual measures of association. However, by using the equation (2) we can determine the lower and upper bounds of P_1 for the leukaemia relapse data; using only the marginal totals the bounds of P_1 are $(L_1 = 0, U_1 = 0.507)$. Similarly, to perform a test of independence at the 0.05 level of significance the bounds are calculated by equation (3); which gives bounds $(L_{0.05} = 0.143, U_{0.05} = 0.284)$. Therefore, independence between the dichotomous variables of Table 2 can be concluded if $0.143 \leq P_1 \leq 0.284$. On the other hand, the row and column variables are statistically significantly associated at the 5% level of significance if $0 \leq P_1 \leq 0.143$ or $0.284 \leq P_1 \leq 0.507$. The extent of association between PCR status and follow up status is then quantified by the AAI using equation (7) at the 5% level of significance. Doing so gives $A_{0.05} = 82.36$. Due to the high value of the AAI, it appears likely that a strong association exists between PCR status and follow up status at the 0.05 level of significance (when only the aggregate data is being analysed). By applying our approach, we argue that the high extent of association between both variables is not due to the sample size, but due to the underlying association structure of the two variables. Since the total number of children included in the study is n = 178, there will be a population of $177^2 = 31{,}329$ contingency tables of size 2x2 that can be generated for this sample size. Hence, we obtain a population of 31,329 AAI values. Recall that, when testing the association between the dichotomous variables of the Table 2 at the 5% level of significance, the observed AAI is $A_{0.05} = 82.360$ and this is only one of the possible 31,329 AAI values within this population. Figure 2 gives a histogram of all possible AAI values for n = 178.
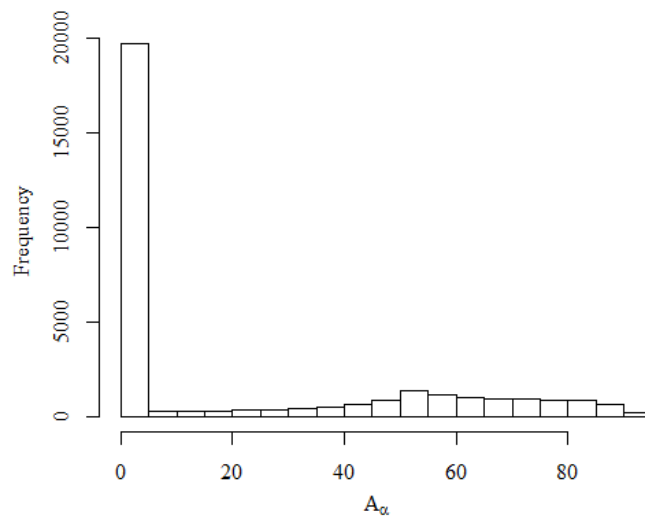

Figure 2. Histogram of the magnitudes of the all possible AAIs for Cave et al. (1998) leukaemia relapse data given the observed sample size, n = 178

The probability of observing an AAI (or one that is more extreme) than $A_{0.05} = 82.360$ is 0.0407 as calculated by equation (8). That is, there are about 4 chances out of 100 that an AAI as big (or bigger) than that observed will be obtained for the given table of size 178 with marginal configuration as in Table 2. This suggests that the aggregate data of Table 2 is not only informative in terms of exploring the underlying association structure between the studied two dichotomous variables, but that the observed AAI of Table 2, is also extremely large in comparison to all possible AAI's that may be obtained for its sample size.

5. DISCUSSION

This paper presents a new approach to quantify the significance of the extent of association in aggregates of 2x2 contingency tables calculated by using the AAI (2008 and 2010). Due to the underlying properties of the AAI, the sample size effect can overshadow the true nature of association displayed in the aggregate data of a table. However, we show that the impact of sample size can be isolated by generating a pseudo population of 2x2 tables under the given sample size. This provides an opportunity for analysts to quantify the likelihood of the value of the AAI for a given marginal configuration for a specific sample size. The pseudo p-value approach suggested in this paper enhances the applicability of the AAI and thus can be considered a valuable addition to the literature of aggregate data analysis. More research is also needed to further explore the mathematical properties when a specific hypothesis needs to be tested at a given

significance level. Future work will also extend our technique for stratified 2x2 tables in addition to tables of higher dimensions.

REFERENCES

Achen, C. H. and W. P. Shively (1995), *Cross-level Inference*. Chicago: University of Chicago Press.

Beh, E. J. (2008), Correspondence analysis of aggregate data: The 2x2 table. *Journal of Statistical Planning and Inference*, 13, 2941 – 2952.

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

Beh. E. J., Tran, D. and Hudson, I. L. (2013), A reformulation of the aggregate association index using the odds ratio. *Computational Statistics & Data Analysis*, 68, 52 – 65.

Beh, E. J., Cheema, S.A., Tran, D. and Hudson, I.L. (2013), Adjusting the aggregate association index for large samples. Proceedings of the *Statistical Italian Societies Conference*: Advances on Latent Variables – Methods, Models and Applications (SIS 2013), Brescia, Italy, ISBN 978 88 343 2556, 8 pages..

Beh, E. J., Cheema, S. A., Tran, D. and Hudson, I. L. (2015), Adjustment to the aggregate association index to minimize the impact of large samples, *Advances in Latent Variables* (eds Carpita, M., Brentari, E. and Qannari, E. M.), pp. 241 – 251, Springer.

Cave, H. C., Bosch, V. W., Suciu, S., et al (1998), Clinical significance of minimal residual disease in children acute lymphoblastic leukemia. *New England Journal of Medicine*, 339, 591 – 598.

Cheema, S.A., Beh, E.J. and Hudson, I.L. (2013), On issues concerning the assessment of information contained in aggregate data using the F-statistics. *MODSIM 2013: 20th International Congress on Modelling and Simulation* (eds Piantadosi, J., Anderssen R.S. and Boland J.), pp. 1966 – 1972. ISBN: 978-0-9872143-3-1.

Cho, W. K.T. and Gains, B. J. (2004), The limits of ecological inference: the case of split ticket voting. *American Journal of Political Sciences*, 48, 152 – 171.

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

Everitt, B. S. (1977), *The Analysis of Contingency Tables*. New York: Wiley.

Greiner, D. and Quinn, K. (2009), R×C ecological inference: bounds, correlations, flexibility and transparency of assumptions, *Journal of the Royal Statistical Society, Series A*, 172, 67 – 81.

Hudson, I.L. Moore, L. Beh, E.J. and Steel, D.G. (2005), Gender counts of historical NZ 1893-1919: A rigorous statistical ecological inference approach. *International Statistical Institute (ISI) Conference 2005*, Sydney. April 5 – 12, pp. 1 – 4.

Hudson, I.L. Moore, L. Beh, E.J. and 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: Princeton University Press.

King, G., Rosen, O. and Tanner, M. (2004). Information in ecological inference: an introduction. *Ecological Inference: New Methodological Strategies,* (King, G., Rosen, O. and Tanner, M. eds), pp. 1 – 12. Cambridge: Cambridge University Press

Quinn, K. M. (2004), Ecological inference in the presence of temporal dependence. *Ecological Inference: New Methodological Strategies,* (King, G., Rosen, O. and Tanner, M. eds), pp. 207 – 232. Cambridge: Cambridge University Press.

Rivers, D. (1998), Review of "A Solution to the Ecological Inference Problem". *American Political Science Review*, 92, 442-443.

Salway, R. and Wakefield, J.A. (2004), A common framework for ecological inference in epidemiology, political science, and sociology. *Ecological Inference: New Methodological Strategies*, (King, G., Rosen, O. and Tanner, M. eds), pp. 303 – 332. Cambridge: Cambridge University Press.

Schuessler, A.A. (1999), Ecological inference. *Proceedings of National Academy of Sciences* (USA), 96, 10549 – 10944.

Simonoff, J.S. (2003), *Analyzing Categorical Data*. New York: Springer.

Steel, D. G., Beh, E. J. and Chambers, R. L. (2004), The information in aggregate data. *Ecological Inference: New Methodological Strategies*, (King, G., Rosen, O. and Tanner, M. eds), pp. 51 – 68. Cambridge: Cambridge University Press.