INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 4, Issue - 1, Jan – 2018
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 5.60 Publication Date: 31/01/2018

AGGREGATE ASSOCIATION INDEX FOR ANALYSING THE ASSOCIATION BETWEEN TWO DICHOTOMOUS VARIABLES WITH MISSING INFORMATION

Titus K. Kibua and Cornelius M. Ndunda
Statistics and Actuarial Science Department, Kenyatta University, Nairobi, Kenya
Email - kibua.titus@ku.ac.ke

Abstract: *Huge amount of time and resources has been provided over the past decades to the expansion of the techniques that can be used to analyze the association between aggregated categorical data. Improvement of the aggregate association index* (AAI) *has been one of the most recent additions to this. When analyzing the association between two dichotomous variables with missing information the* $AAI's$ *magnitude is seriously affected by the sample size. An increase in the sample size simultaneously increases the* AAI *even when the marginal proportions remain constant. Changes to the* AAI *are proposed that reduce the effect brought about by increasing the sample size. Empirical study results indicate that the adjusted* AAI *is steadier than the existing ones in reply to any increase in the sample size and could be recommended for use in the event of missing information.*

Key Words: *Dichotomous, Aggregate Association Index, Contingency Table.*

1. INTRODUCTION:

In statistical data analysis, establishing the association between variables has been one of the most important tasks to researchers. The association between variables means that one variable's value relates somehow to the values of the others. Many times in real life situations, knowing how one variable relates to another gives the statistician a valuable information that can be used. For example, in health sector, if two variables' age and height are related to each other and each group's average age – to height information is known and a doctor meets a child who happens to be much shorter than expected for his or her age, then based on the information the doctor can proceed to inform the parents to figure out whether there is a problem. Another example of a relationship that a statistician may have interest is the connection between political party bolstered and opinion concerning socio – economic issues.

In collecting data, statisticians normally encounter some data missing. This occurs when no data value is available. This is a typical incidence and can significantly affect the conclusions made using the available data. There are many reasons for this. For example, in automatic equipment sensor errors, respondent's refusal to answer questions that affects confidentiality, researchers' errors in data handling and even strict privacy restrictions forced by government and corporate organizations that are hesitant to discharge information pertaining to individual level. However, researchers who require the best results from their research need to be prepared to deal with missing data in the most appropriate and desirable way possible.

New techniques have been developed to analyze the association between variables given such a data especially when given the aggregate data. Aggregate data analysis for two dichotomous variables has a long history. Techniques concerning the ecological inference of aggregate data like frequencies of cell estimation for a stratified 2×2 contingency tables given only aggregate data have been well thought out. It is a worrying matter for aggregate data analysis that the clarification of the parameters from EI models may be altogether unique to similar restrictions for the investigation of individual level information. The key issue of these strategies is that, since the cell frequencies of the contingency table are unfamiliar, various presumptions should be made about their structure. Late commitments on the subject matter have moved from demonstrating the cell frequencies provided just the aggregate data to the investigation of the association structure between variables. The aggregated association index (AAI) development has provided the analyst with the opportunity to measure the degree of association between the two dichotomous variables rather than modeling the marginal data.

The problem is that, as one ponders a sample size increase; there is an increase in AAI. This is on the grounds that Pearson's chi-squared statistic is defenseless to variations in the sample size. This can change the original idea of the association among the variables being masked by the greatness of the sample size. Cheema et.al (2013) proposed

$$A_\alpha^l = A_\alpha\sqrt{Cn_0} - 100\left(\sqrt{Cn_0} - 1\right) \qquad (1)$$

as an adjustment to the aggregate association but still was not stable since the sample size increased gradually. A_α is the proposed index by Beh (2008) and C is a factor that is used to increase the sample size by multiplying it with the original sample size and can take values from one and above.

In our study we make adjustments to the proposed (AAI) and propose an index that will aim at minimizing the impact of increase in sample size while investigating the relationship between the variables given just the aggregate information.

2. AGGREGATE ASSOCIATION INDEX (AAI):

We define $P_1 = \frac{n_{11}}{n_{1.}}$ as the provisional probability of the classification of an individual/unit in to column one as long as the category is row one. When in a 2×2 contingency table cells are unknown, the bounds of cell $(1, 1)$ lies within the interval $\max(0,\, n_{.1} - n_{2.}) \le n_{11} \le \min(n_{.1},\, n_{1.})$.

P_1 is a proportion and lies within the intervals of [0,1]. When deriving the bounds for the Pearson product moment correlation of a 2×2 contingency table, Duncan and Davis (1953) showed that by involving the marginal frequencies these bounds could be pointed out like;

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

When there is accessibility of marginal information and an association test is made only at the α level of significance the limits of P_1 are;

$$L_\alpha(n_O) = \max\left(0, p_{.1} - p_{2.}\sqrt{\frac{\chi_\alpha^2}{n_O}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)}\right) < P_1 < \min\left(1, p_{.1} + p_{2.}\sqrt{\frac{\chi_\alpha^2}{n_O}\left(\frac{p_{.1}p_{.2}}{p_{1.}p_{2.}}\right)}\right) = U_\alpha(n_O) \qquad (3)$$

Where χ_α^2 is the $1-\alpha$ percentile of the chi-squared distribution with one degree of freedom. Pearson's chi-squared statistic could be linked to a quadratic function as;

$$\chi^2(p_1/p_{1.}, p_{.1}) = n\left(\frac{p_1 - p_{.1}}{p_{2.}}\right)^2\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right)$$

The asymptotic $100(1-\alpha)\%$ under the null hypothesis of independence amongst the dichotomous variables is:

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

Therefore, we may reason out that, as long as a level of significance and the total data is provided there is also a remarkable relationship amongst the dichotomous factors if;

$$U_\alpha \le P_1 \le U_1$$

or

$$L_1 \le P_1 \le L_\alpha$$

From this interval Beh (2008) proposed the Aggregate Association Index (AAI),

$$A_\alpha = 100\left(1 - \left\{\frac{\chi_\alpha^2[(L_\alpha - L_1) + (U_1 - U_\alpha)] + Int(L_\alpha, U_\alpha)}{Int(L_1, U_1)}\right\}\right) \qquad (4)$$

which measured the scope of association that may be between two dichotomous variables at α level of significance assuming that only the total data for a single 2×2 contingency table is available.

Where

$$Int(L_\alpha, U_\alpha) = \int_{L_\alpha}^{U_\alpha} \chi^2\left(\frac{P_1}{P_{1.,}\, P_{.1}}\right) dP_1$$

and

$$Int(L_1, U_1) = \int_{L_1}^{U_1} \chi^2\left(\frac{P_1}{P_{1.,}\, P_{.1}}\right) dP_1$$

Hence

$$\chi^2\left(\frac{P_1}{P_{1.,}\, P_{.1}}\right) = n_o\left(\frac{P_1 - P_{.1}}{P_{2.}}\right)^2\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)$$

L_α , U_α, L_1, U_1 are boundaries within which P_1 lies, where $P_1=\frac{n_{11}}{n_{1.}}$ is the conditional probability of categorizing an individual/unit to column one as long as it is characterized to row one in a contingency table. $P_{1.}, P_{2.}$, $P_{.1}$ and $P_{.2}$ are marginal cell proportions and n_o is the sample size. This aggregated association index is not stable because it depends on the bounds of P_1 and χ^2_α which are septic to change when the sample size varies. Hence there is need for an adjustment so that it can perform better.

3. PROPOSED AGGREGATE ASSOCIATION INDEX:

Consider the aggregate association index (4).

$$A_\alpha = 100\left(1-\left\{\frac{\chi^2_\alpha[(L_\alpha-L_1)+(U_1-U_\alpha)]+Int(L_\alpha,U_\alpha)}{Int(L_1,U_1)}\right\}\right)$$

$$=100\left(1-\frac{\chi^2_\alpha[(L_\alpha-L_1)+(U_1-U_\alpha)]}{Int(L_1,U_1)}-\frac{Int(L_\alpha,U_\alpha)}{Int(L_1,U_1)}\right) \qquad (5)$$

where

$$Int(L_\alpha,U_\alpha)=\int_{L_\alpha}^{U_\alpha}\chi^2\left(\frac{P_1}{P_{1.,}P_{.1}}\right)dP_1 \quad , \quad Int(L_1,U_1)=\int_{L_1}^{U_1}\chi^2\left(\frac{P_1}{P_{1.,}P_{.1}}\right)dP_1$$

and

$$\chi^2\left(\frac{P_1}{P_{1.,}P_{.1}}\right)=n_o\left(\frac{P_1-P_{.1}}{P_{2.}}\right)^2\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)$$

Integrating with respect to P_1 and assuming that P_1 is continuous we obtain

$$Int(L_\alpha,U_\alpha)=\int_{L_\alpha}^{U_\alpha}\chi^2\left(\frac{P_1}{P_{1.,}P_{.1}}\right)dP_1=\int_{L_\alpha}^{U_\alpha}n_o\left(\frac{P_1-P_{.1}}{P_{2.}}\right)^2\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)dP_1$$

$$=\frac{1}{P^2_{2.}}\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)n_o\int_{L\alpha}^{U_\alpha}(P_1-P_{.1})^2\,dP_1 \qquad (6)$$

This can also be expressed as

$$\frac{1}{P^2_{2.}}\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)n_o\int_{L\alpha}^{U_\alpha}(P_1^2-2P_1P_{.1}+P^2{}_{.1})dP_1=\frac{1}{P^2_{2.}}\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)n_o\left[\frac{P_1^3}{3}-P_1^2P_{.1}+P_{.1}^2P_1\right]_{L_\alpha}^{U_\alpha}$$

$$=\frac{1}{P^2_{2.}}\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)n_o\left\{\frac{(U^3_\alpha-3U^2_\alpha P_{.1}+3P^2_{.1}U_\alpha)-(L^3_\alpha-3L^2_\alpha P_{.1}+3L_\alpha P^2_{.1})}{3}\right\} \qquad (7)$$

Subtracting a constant $P_{.1}{}^3$ from (7) and adding the same constant we have,

$$=\frac{1}{3P^2_{2.}}\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)n_o\left\{(U^3_\alpha-3U^2_\alpha P_{.1}+3U_\alpha P^2_{.1}-P^3_{.1})-(L^3_\alpha-3L^2_\alpha P_{.1}+3L_\alpha P^2_{.1}-P^3_{.1})\right\} \qquad (8)$$

Clearly this forms a polynomial of degree three with variables U_α and L_α. Further, (8) can be expressed as

$$\frac{1}{3P^2_{2.}}\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)n_o\left\{(U_\alpha-P_{.1})^3-(L_\alpha-P_{.1})^3\right\}$$

Similarly

$$Int(L_1,U_1)=\int_{L_1}^{U_1}\chi^2\left(\frac{P_1}{P_{1.,}P_{.1}}\right)dP_1$$

$$=\frac{1}{3P^2_{2.}}\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)n_o\left\{(U_1-P_{.1})^3-(L_1-P_{.1})^3\right\}$$

Substituting the integral parts in (5), we have

$$A_\alpha = 100\left[1-\frac{\chi_\alpha^2\{(L_\alpha(n_o)-L_1)+(U_1-U_\alpha(n_o))\}}{Kn_o\{(U_1-p_{.1})^3-(L_1-p_{.1})^3\}}-\frac{\{(U_\alpha(n_o)-p_{.1})^3-(L_\alpha(n_o)-p_{.1})^3\}}{\{(U_1-p_{.1})^3-(L_1-p_{.1})^3\}}\right] \qquad (9)$$

where $K=\frac{1}{3P_{2.}^2}\left(\frac{P_{1.}P_{2.}}{P_{.1}P_{.2}}\right)$ and $0 \le A_\alpha < 100$

L_α and U_α are stated as functions of the sample size. For a specified α, (9) will measure how probable a specific set of fixed marginal frequencies can empower the client to decide that there is a measurably significant relationship between the variables.

Assuming $A_\alpha \approx 100$ at that point, provided just the aggregate data, it is profoundly possible that a significant relationship exists here. However, if $A_\alpha \approx 0$ it is very unlikely that such a relationship really exists. Equation (9) demonstrates that the extent of the index is extremely reliant on the Pearson's chi-squared statistic that is also dependent on the sample size. For instance, if the sample size is increased by a factor of $C > 1$ so that $n = Cn_o$, the Pearson's chi-squared statistic increases by a factor of C. As the sample size increases, $U_\alpha(n_o)$ and $L_\alpha(n_o)$ approaches $P_{.1}$. In this manner, AAI approaches 100 as the sample size increases.

We now propose a simple approach to assure that the AAI is less influenced by the sample size increase, when the marginal proportions are constant.

We express (9) as

$$A_\alpha = 100\left[1-f(n_o)\left(\frac{U_1-L_1}{U_\alpha(n_o)-L_\alpha(n_o)}\right)\times\left\{\begin{array}{l}\frac{\chi_\alpha^2\{(L_\alpha(n_o)-L_1)+(U_1-U_\alpha(n_o))\}}{Kn_o\{(U_1-p_{.1})^3-(L_1-p_{.1})^3\}}-\\ \left\{\frac{(U_\alpha(n_o)-p_{.1})^3-(L_\alpha(n_o)-p_{.1})^3}{(U_1-p_{.1})^3-(L_1-p_{.1})^3}\right\}\end{array}\right\}\right] \qquad (10)$$

where $f(n_o)=\frac{U_\alpha(n_o)-L_\alpha(n_o)}{U_1-L_1}$

Increasing the original sample size by $C > 1$ will certainly result to another sample size $n = Cn_o$ and Pearson's chi-squared statistic will increase. This accordingly, prompts an increase in the AAI, despite the fact that the relative marginal information stays unaltered. In particular, as $C \to \infty$ then $A_\alpha \to 100$.

Subsequently, changing the AAI computation by considering two details of $f^{/}(n_o)$, is the help brought about by limiting the effect that increasing the sample size has on it.

Let

$$A_\alpha^{/} = 100\left[1-f^{/}(n_o)\left(\frac{U_1-L_1}{U_\alpha(n)-L_\alpha(n)}\right)\left\{\frac{\chi_\alpha^2\{(L_\alpha(n)-L_1)+(U_1-U_\alpha(n))\}}{Kn\{(U_1-p_{.1})^3-(L_1-p_{.1})^3\}}-\frac{\{(U_\alpha(n)-p_{.1})^3-(L_\alpha(n)-p_{.1})^3\}}{(U_1-p_{.1})^3-(L_1-p_{.1})^3}\right\}\right]$$

(11)

Where $f^{/}(n_o)$ is to be subjectively or objectively determined so that $0 \le f^{/}(n_o) \le 1$.

Adjustment 1

If a subjective $f^{/}(n_o)$ remains almost constant when the sample size increases then we consider an average of the boundaries U_1 and L_1 of P_1 according to equation (1). This average gives a value in the range specified in equation (2). Let $f^{/}(n_o)=\frac{U_1-L_1}{2}$ and put it in (11) to obtain

$$A_\alpha^{//} = 100\left[1-\left(\frac{U_1-L_1}{2}\right)\left(\frac{U_1-L_1}{U_\alpha(n)-L_\alpha(n)}\right)\left\{\frac{\chi_\alpha^2\{(L_\alpha(n)-L_1)+(U_1-U_\alpha(n))\}}{Kn\{(U_1-p_{.1})^3-(L_1-p_{.1})^3\}}-\frac{\{(U_\alpha(n)-p_{.1})^3-(L_\alpha(n)-p_{.1})^3\}}{(U_1-p_{.1})^3-(L_1-p_{.1})^3}\right\}\right] \qquad (12)$$

Adjustment 2

Since our index is highly depended on χ_α^2 which is affected by changes in the sample size we propose $f^{/}(n_o)$ which is also a function of the bounds of P_1 so that it can help regulate the index. Let $f^{/}(n_o)=\dfrac{U_\alpha(n)+L_\alpha(n)}{U_1-L_1}$. Then our index becomes

$$A_\alpha^{///}=100\left[1-f^{/}(n_o)\left(\frac{U_1-L_1}{U_\alpha(n)-L_\alpha(n)}\right)\left\{\frac{\chi_\alpha^2\{(L_\alpha(n)-L_1)+(U_1-U_\alpha(n))\}}{Kn\{(U_1-p_{.1})^3-(L_1-p_{.1})^3\}}-\frac{\{(U_\alpha(n)-p_{.1})^3-(L_\alpha(n)-p_{.1})^3\}}{(U_1-p_{.1})^3-(L_1-p_{.1})^3}\right\}\right] \quad (13)$$

4. EMPIRICAL STUDY RESULTS AND CONCLUSION:

We consider real data as in Fisher (1935) where cross-checks of whether 30 criminal twins are monozygotic or dizygotic twins and also whether their same sex twin has previously been sentenced on a criminal offence

Table 1: Fisher's (1935) criminal twin data

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

The Pearson's chi-squared statistic for this data is 13.03167, and p-value is 0.001221 showing that, there was a statistically significant relationship between the two dichotomous factors.

The product moment correlation of this data is $r=0.659082$ showing that there is positive association. Hence a monozygotic twin of indicted criminal is linked with being sentenced of an offence, while a dizygotic twin of an indicted criminal tends not to be indicted of an offense. The proportion $p_1=0.7692$ shows that about 77% of those monozygotic criminal twins have a same sex kin in the sample who has also been indicted of an offense. The odds ratio is $\theta=25$. This shows that criminals who were monozygotic were twenty five times as likely to be convicted as criminals as to those who were dizygotic.

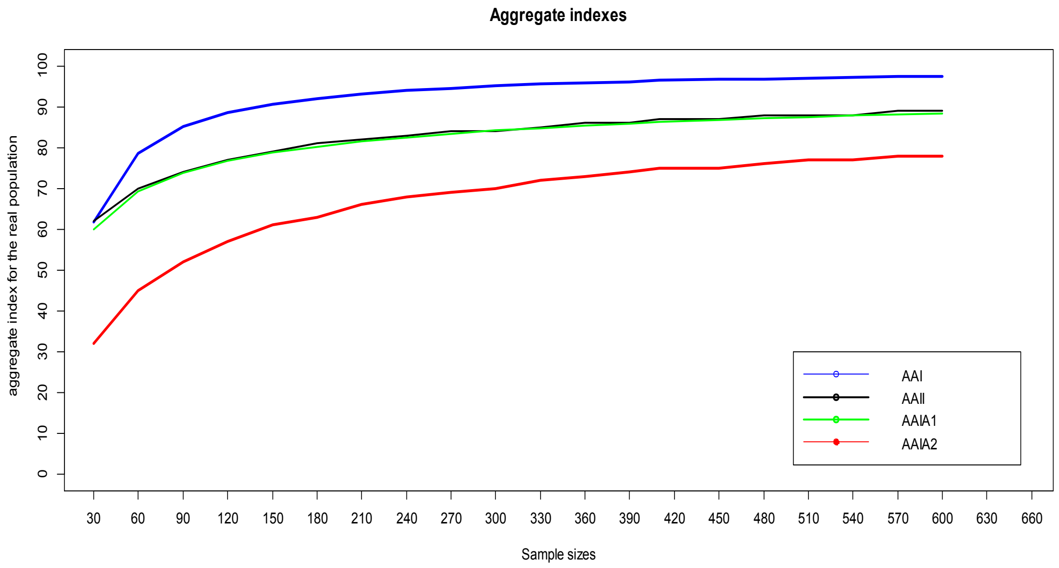
The confidence interval is $[3.52146, 177.4831]$. The phi-coefficient of the data is $\vartheta=0.65158$ portraying that there is a significant relationship between the variables.

Suppose now in table 1, the cell frequencies were missing and only the marginal frequencies were available, this means that we would not be able to calculate the product moment correlation, Pearson's chi-squared statistic, and even the odds ratio. Thus establishing how the dichotomous variables associate would be a challenge and now the aggregated association index becomes the solution to establish this association.

At 5% level of significance, the AAI, reckoned from (9), is 61.827. It is probably that there is a statistically significant relationship between the variables. However, this index changes as the sample size is increased because it depends on the chi-square which is highly depended on the sample size.

Taking a sample size of 30 and increasing it upto 600 the proposed indices behaves as shown in figure 1.

Figure 1: Aggregate association indices as sample size increases.



AAI	Aggregate association index – Beh (2008).
AAII	Aggregate association index adjusted – Chemma et.al(2013).
AAIA1	First adjusted aggregated association index
AAIA2	Second adjusted aggregated association index

When $n = 570$, $A_{0.05} = 97$ showing that it is to a great degree probable that a relationship exists between the variables provided just the aggregate data. As seen all the indices increase fast initially but there after stabilizes as there is an increase in the sample size. This is because the aggregated association indices are dependent on $U_{\alpha}(n)$ and $L_{\alpha}(n)$ which changes with sample size. The $AAI's$ first increase at a higher rate as the sample size increases but thereafter stabilizes when $U_{\alpha}(n)$ and $L_{\alpha}(n)$ are approximately the same at large values of n.

Our aim is to stabilize the extent of the AAI as the sample size increases. For the sample size $n = 30$, the first adjustment $A^{//}_{0.05} = 59.924$ and the second adjustment $A^{//}_{0.05} = 31.576$. The original index is $AAI = 61.82$ and adjustment proposed by Cheema at.al (2013) is $AAI = 61.82$. As the sample size increases, the adjusted versions of A_{α} increase more gradually than before. The second adjustment is more stable as compared to the others because the $f^{/}(n_0)$ used is fairly constant even when the sample size increases making it perform better compared to the other indices. From Figure 1 we can conclude that the rate of change for both adjusted indices is more stable than the existing ones as the sample size increases and hence when some information is missing they could be recommended for better results.

REFERENCES:

1. Acock, A.(2000). Working with missing values. *Journal of marriage and family*, 67(4),1012-1028.
2. Beh, E.J. (2008). Correspondence analysis of aggregate data: The 2×2 table. *Journal of Statistical Planning and Inference,*138, 2941-2952.
3. Beh, E.J. (2010). The aggregate association index. *Computational Statistics and Data Analysis,*54, 1570-1580.
4. Beh, E.J., Cheema, S.A., Tran, D., Hudson, I.L.: Adjusting the aggregate association index for large samples. In: Proceedings of advances on latent variables/methods models and applications, Brescia, Italy (2013)
5. Berkson, J.(1978). In dispraise of the exact test: Do the marginal totals of the 2×2 table contain relevant information respecting the table proportion? *Journal of Statistical Planning and Inference,* 2, 27 – 42.
6. Conover, W.J. (1974). Some reasons for not using the Yate’s continuity correction on 2×2 contingency tables (with discussions). *Journal of the American Statistical Association*, 69,374-382.
7. Duncan, O.D. and Davis, B.(1953). An alternative to ecological correlation. *American Sociological Review,* 18, 665-666.
8. Fisher, R.A. (1935). The logic of inductive inference (with discussion). *Journal of Royal Statistical Association,* Series A, 98, 39-82.
9. Freedman, D.A., Klein, S.P., Sacks,J., Smyth, C.A.,& Everett, C.G. (1991). Ecological regression and voting rights. *Evaluation review*. 15, 673-711.
10. Goodman, L. (1953). Ecological regressions and the behavior of individuals. *American Sociological Review*. 18, 663-666.
11. Haber, M. (1989). Do the marginal total of a 2×2 contingency table contain information regarding the table proportion? *Communication in Statistics*: Theory and Methods, 18, 147 – 156.
12. 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, vol. 173, 185-213.
13. Plackett, R.L. (1977). The marginal totals of a 2×2 table. *Biometrics*, 64, 37 – 42.
14. Wakefield, J. (2004). Ecological inference for 2×2 tables. *Journal of Royal Statistical Society*, Series A, 167, 385 – 445.
15. Yates, F.(1984). Tests of significance for 2×2 contingency tables (with discussion). *Journal of Royal Statistical Society*, Series A, 147,426-463.