Generalized Correlation Coefficient Based on Log Likelihood Ratio Test Statistic

Hsiang-Chuan Liu[a]

Department of Informatics and Biomedical Engineering, Asia University, Taichung 41354, Taiwan, ROC
Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taichung 40306, Taiwan

Abstract. In this paper, I point out that both Joe's and Ding's strength statistics can only be used for testing the pair-wise independence, and I propose a novel G-square based strength statistic, called Liu's generalized correlation coefficient, it can be used to detect and compare the strength of not only the pair-wise independence but also the mutual independence of any multivariate variables. Furthermore, I proved that only Liu's generalized correlation coefficient is strictly increasing on its number of variables, it is more sensitive and useful than Cramer's V coefficient, in other words, Liu generalized correlation coefficient is not only the G-square based strength statistic, but also an improved statistic for detecting and comparing the strengths of deferent associations of any two or more sets of multivariate variables, moreover, this new strength statistic can also be tested by G^2.

1 Introduction

For any quantity study, two kinds of statistics must be considered, one is the significant statistic, the other is the strength statistic, in other words, any significant test statistic should always be coupled with an appropriate statistic of strength, because if the association strength is weak, but the sample size is large, then the result of the null hypothesis is always rejected by a significant test, that is meaningless [1]. For comparing the pair-wise independent among some sets of multivariate variable, if the four rules are satisfied; we always use Pearson Chi square statistic [2], χ_P^2, or G^2 to test the significance; (a) Cochran's rule [3] addresses that there should be no expected frequency value under 1 and no more than 0.2 of expected frequency value should be under than 5, (b) Wikens' rule [4] suggests that the total sample size should be at less five times the number of cells of the contingency table, (c) McHugh's rule [5] suggests that the number of cells of contingency table is not larger than 20, (d) Liu's rule [6] suggests that all values of the correlation coefficients of variables are positive. If these rules are not be satisfied, the interpretation of the significant test results will be difficult, we can exploit the improving methods, such as combining some cells with smaller expected frequency of the contingency table, Yates' correct method or Fisher's exact test. And then, we can use Cramer's V coefficient [1] to test its strength.

For comparing the mutual independent of some sets of multivariate variable, if the above-mentioned four rules are satisfied; Sokal & Rohlf [7] suggested to use the log likelihood ratio statistic, G^2, rather than χ_P^2, since for a large dimensional contingency table, G^2 can be neatly decomposed into smaller components, it can be not done exactly with χ_P^2, and G^2 is the usual statistic for log-linear analyses. The most traditional researches always use G^2 or χ_P^2 to test the pair-wise independence of multivariate variables, since it only needs smaller sample size than that of mutual-independence, but only mutual-independence is the complete independence of multivariate variables. In recent years, big data researches and developments are more emphasized, how to compare the relations among some deferent sets of multivariate variables are the more important issues, and the strength statistic becomes more important than its significant statistic, since their mutual independence samples size are large, the null hypotheses of significant testing are always be rejected. We can use G^2 or χ_P^2, to test the mutual independence of them whether is significant or not, and we know that Cramer's V is the χ_P^2 based strength statistic, but how to find the G^2 based strength statistic? That is a very important issue.

In this paper, I will point out that two well-known mutual information based strength statistics can only be used for testing pair-wise independence, one is Joe's generalized correlation coefficient [7], and the other is Ding's generalized correlation coefficient [8], but both of them can still be not used for mutual independence of the three or more multivariate variables, since the value of mutual information of three variables may be negative [10], and its generalized cases can be not defined, and I will propose a novel G^2 based strength statistic, called Liu's generalized correlation coefficient, this new strength statistic can be used for detecting and comparing not only pair-wise independence but also mutual independence of any multivariate variables.

Furthermore, I will prove that the proposed statistic not only has some good properties as which of Pearson's determinant coefficient, but also its strength is strictly increasing on variable number n, therefore, the new

[a] Corresponding author: lhc@asia.edu.tw

strength statistic is more sensitive and useful than Cramer's V_n.

2 Significance statistics for detecting the independence

2.1. For detecting the pair-wise independence of multivariate variables

For detecting the pair-wise independence we can use χ_P^2 and G^2, below;

Definition 1: Pearson Chi square statistic, χ_P^2, [2] is defined below;

$$\chi_P^2=\sum_{i=1}^{r}\sum_{=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}=N\sum_{i=1}^{r}\sum_{=1}^{c}p_{ij}\left(\frac{O_{ij}/N-p_{ij}}{p_{ij}}\right)^2\sim\chi^2_{(r-1)(c-1)} \quad (1)$$

Where O_{ij} is the observed value of row i column j, E_{ij} is the expected value of row column j, N is the sample size, p_{ij} is the probability of row i and column j, $i=1,2,...r,\quad j=1,2,...,c$,

Definition 2: Log likelihood ratio Chi square statistic, G^2 [7], is defined below;

$$G^2=-2\sum_{i=1}^{r}\sum_{j=1}^{c}p_{ij}\log\left(\frac{p_{ij}}{p_ip_j}\right)\sim\chi^2_{(r-1)(c-1)} \quad (2)$$

Where p_{ij} is the joint probability of row i column j, p_i & p_j are the marginal probability of row i and column j, respectively.

2.2. For detecting the mutual independence of multivariate variables

χ_P^2 & G^2 can be extended to the generalized case for detecting the mutual independence below,

Definition 3. Generalized χ_P^2 and, G^2 are defined below, respectively,

(i) $$\chi^2_{P(n)}=\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}\cdots\sum_{i_n=q}^{r_n}\frac{\left(O_{i_1i_2\ldots i_n}-E_{i_1i_2\ldots i_n}\right)^2}{E_{i_1i_2\ldots i_n}}\sim\chi^2_{\prod_{k=1}^{n}(i_k-1)} \quad (3)$$

(ii) $$G^2_n=-2\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}\cdots\sum_{i_n=q}^{r_n}p_{i_1i_2\ldots i_n}\log\left(\frac{p_{i_1i_2\ldots i_n}}{\prod_{k=1}^{n}p_{i_k}}\right)\sim\chi^2_{\prod_{k=1}^{n}(i_k-1)} \quad (4)$$

3 The Chi square based strength statistics

Definition 4: The Chi square based strength statistics, Cramer's V [1, 6], is defined below;

(i) For pair-wise independent:

$$V=\sqrt{\chi_P^2/N\min\left(r-1,c-1\right)}, \quad (5)$$

(ii) For mutual independent:

$$V_n=\sqrt{\chi^2_{P(n)}/N\min\left(r_1-1,r_2-1,...,r_n-1\right)} \quad (6)$$

where $0\le V,V_n\le 1$

Note that: if $\chi^2_{P(n)}=\chi^2_{P(n+1)}$ and

$$X_{n+1}\ne X_i,i=1,2,...,n,\min\left(r_1-1,r_2-1,...,r_n-1\right)=\min\left(r_1-1,r_2-1,...,r_{n+1}-1\right), \quad (7)$$

then $V_n=V_{n+1}$, in other words, Cramer's V is not strictly increasing on its number of variables.

4. Two mutual information based strength statistics

4.1. Joe's generalized correlation coefficient

Definition 5: Joe's generalized correlation coefficient of $\left(X_1,X_2\right)$ [8] is defined below;

$$GCC_J\left(X_1,X_2\right)=\sqrt{1-\exp\left[-2\times I\left(X_1:X_2\right)\right]} \quad (8)$$

where

$$0\le GCC_{J-2}\left(X_1,X_2\right)\le 1,\ I\left(X_1:X_2\right)=H\left(X_1\right)+H\left(X_2\right)-H\left(X_1,X_2\right) \quad (9)$$

Note that Cover and Thomas [10] pointed out that the value of $I\left(X_1:X_2:X_3\right)$ may be negative, therefore, the generalized cases can be not defined.

4.2. Ding's generalized correlation coefficient

Definition 6: Ding's generalized correlation coefficient of $\left(X_1,X_2,...,X_n\right)$ [9] is defined below;

$$GCC_D\left(X_1,X_2,...,X_n\right)=\frac{\sum_{i=1}^{n}H\left(X_i\right)-H\left(X_1,X_2,...,X_n\right)}{\left[\prod_{i=1}^{n}H\left(X_i\right)\right]^{\frac{1}{n}}} \quad (10)$$

Note that if $n>2,H\left(X_i\right)=H\left(X\right),i=1,2,...,n$, then

$$GCC_D\left(X_1,X_2,...,X_n\right)=n-1>1,$$

therefore, $GCC_D\left(X_1,X_2,...,X_n\right)$ is not well defined, and it still can be not used for detecting the mutual independence of multivariate variables.

5. The G^2 based strength test statistic for testing independence

In this paper, a novel G^2 based strength test statistic is proposed as follows,

Definition 7: The G^2 based strength test statistic; Liu's generalized correlation coefficient of $\left(X_1,X_2,...,X_n\right)$ is defined below;

$$\alpha_L(n)=GCC_L(X_1,X_2,...,X_n)=\frac{n}{n-1}\left[1-\frac{H(X_1,X_2,...,X_n)}{\sum_{i=1}^{n}H(X_i)}\right] \tag{11}$$

Where n is the number of variables, N is the sample size, $H(X_1,X_2,...,X_n)$ is the joint entropy of random variables $X_1,X_2,...,X_n$, $H(X_i)$ is the marginal entropy of random variables $X_i, i=1,2,...,n$.

Theorem 1: The Important properties of Liu' $\alpha_L(n)$ coefficient are listed below;

(i) $$G_n^2=-2\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}...\sum_{i_n=q}^{r_n}p_{i_1i_2...i_n}\log\left(\frac{p_{i_1i_2...i_n}}{\prod_{k=1}^{n}p_{i_k}}\right)$$
$$=2\left[\sum_{k=1}^{n}H(X_k)-H(X_1,X_2,...,X_n)\right] \tag{12}$$

(ii) $$0\le\alpha_L(n)=\frac{n}{n-1}\left[1-\frac{H(X_1,X_2,...,X_n)}{\sum_{i=1}^{n}H(X_i)}\right]\le 1 \tag{13}$$

(a) If $X_1,X_2,...,X_n$ are mutual independent, then $\alpha_L(n)=0$.

(b) If they are all equal, then $\alpha_L(n)=1$.

(iii)
$$H(X_{n+1})>0, X_{n+1}\ne X_i, i=1,2,...,n,\ \alpha_L(n)\in(0,1)$$
$$\Rightarrow\alpha_L(n)<\alpha_L(n+1),\forall n\in N, n\ge 2 \tag{14}$$

(v) $$G_n^2=2\left(1-\frac{1}{n}\right)\left[\sum_{k=1}^{n}H(X_k)\right]\alpha_L(n)\sim\chi^2_{\prod_{k=1}^{n}(i_k-1)} \tag{15}$$

Proof:(i)

$$G_n^2=-2\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}...\sum_{i_n=q}^{r_n}p_{i_1i_2...i_n}\log\left(\frac{p_{i_1i_2...i_n}}{\prod_{k=1}^{n}p_{i_k}}\right)$$
$$=-2\left[\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}...\sum_{i_n=q}^{r_n}p_{i_1i_2...i_n}\left(\log p_{i_1i_2...i_n}\right)-\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}...\sum_{i_n=q}^{r_n}p_{i_1i_2...i_n}\left(\log\prod_{k=1}^{n}p_{i_k}\right)\right]$$
$$=-2\left[H(X_1,X_2,...,X_n)-\sum_{k=1}^{n}H(X_k)\right]$$
$$=2\left[\sum_{k=1}^{n}H(X_k)-H(X_1,X_2,...,X_n)\right] \tag{16}$$

(ii) From the Entropy theory [10], we know that;

$$0\le H(X_1,X_2,...,X_n)\le\sum_{i=1}^{n}H(X_i) \tag{17}$$

Thus $$0\le 1-\frac{H(X_1,X_2,...,X_n)}{\sum_{i=1}^{n}H(X_i)}\le\frac{n-1}{n} \tag{18}$$

We can obtain

$$0\le\alpha_L(n)=\frac{n}{n-1}\left[1-\frac{H(X_1,X_2,...,X_n)}{\sum_{i=1}^{n}H(X_i)}\right]\le 1 \tag{19}$$

(a) If $X_1,X_2,...,X_n$ are mutual independent, then $\alpha_L(n)=0$.

(b) If they are all equal, then $\alpha_L(n)=1$.

(iii): According to Entropy theory [10], we know that

$$\sum_{I=1}^{n}H(X_i)\ge H(X_i)_{i=1}^{n}\ge H(X_{n+1}\mid X_i)_{i=1}^{n}\ge 0 \tag{20}$$

$$H(X_{n+1}\mid X_i)_{i=1}^{n}+H(X_i)_{i=1}^{n}=H(X_i)_{i=1}^{n+1} \tag{21}$$

Since $H(X_{n+1})>0$, then

$$H(X_{n+1}\mid X_i)_{i=1}^{n}>0,\ \sum_{i=1}^{n+1}H(X_i)>\sum_{i=1}^{n}H(X_i), \tag{22}$$

We can obtain

$$\frac{\sum_{I=1}^{n}H(X_i)}{\sum_{I=1}^{n+1}H(X_i)}<\frac{\sum_{I=1}^{n}H(X_i)}{\sum_{I=1}^{n}H(X_i)+H(X_{n+1})}\le\frac{\sum_{I=1}^{n}H(X_i)}{\sum_{I=1}^{n}H(X_i)+H(X_{n+1}\mid X_i)_{i=1}^{n}}$$
$$\le\frac{H(X_i)_{i=1}^{n}}{H(X_i)_{i=1}^{n}+H(X_{n+1}\mid X_i)_{i=1}^{n}}=\frac{H(X_i)_{i=1}^{n}}{H(X_i)_{i=1}^{n+1}} \tag{23}$$

and

$$\frac{\sum_{I=1}^{n}H(X_i)}{\sum_{I=1}^{n+1}H(X_i)}<\frac{H(X_i)_{i=1}^{n}}{H(X_i)_{i=1}^{n+1}},\ \frac{H(X_i)_{i=1}^{n+1}}{\sum_{I=1}^{n+1}H(X_i)}<\frac{H(X_i)_{i=1}^{n}}{\sum_{I=1}^{n}H(X_i)} \tag{24}$$

Thus

$$\left(\frac{n}{n+1}\right)\frac{\sum_{I=1}^{n}H(X_i)}{\sum_{I=1}^{n+1}H(X_i)}<\left(\frac{n}{n+1}\right)\frac{H(X_i)_{i=1}^{n}}{H(X_i)_{i=1}^{n+1}}, \tag{25}$$

$$\frac{(n+1)H(X_i)_{i=1}^{n+1}}{(n+1)\sum_{I=1}^{n+1}H(X_i)}<\frac{nH(X_i)_{i=1}^{n}}{n\sum_{I=1}^{n}H(X_i)} \tag{26}$$

We have

$$\frac{nH(X_i)_{i=1}^{n}-\sum_{I=1}^{n}H(X_i)}{n\sum_{I=1}^{n}H(X_i)-\sum_{I=1}^{n}H(X_i)}>\frac{(n+1)H(X_i)_{i=1}^{n+1}-\sum_{I=1}^{n+1}H(X_i)}{(n+1)\sum_{I=1}^{n+1}H(X_i)-\sum_{I=1}^{n+1}H(X_i)}\ge 0 \tag{27}$$

Then $$\frac{nH(X_i)_{i=1}^{n}-\sum_{I=1}^{n}H(X_i)}{(n-1)\sum_{I=1}^{n}H(X_i)}>\frac{(n+1)H(X_i)_{i=1}^{n+1}-\sum_{I=1}^{n-1}H(X_i)}{(n)\sum_{I=1}^{n+1}H(X_i)} \tag{28}$$

and

$$1-\frac{nH(X_i)_{i=1}^{n}-\sum_{I=1}^{n}H(X_i)}{(n-1)\sum_{I=1}^{n}H(X_i)}<1-\frac{(n+1)H(X_i)_{i=1}^{n+1}-\sum_{I=1}^{n-1}H(X_i)}{(n)\sum_{I=1}^{n+1}H(X_i)} \tag{29}$$

Therefore

$$0 < \alpha_L(n) = \frac{n}{n-1}\left[1 - \frac{H(X_i)_{i=1}^{n}}{\sum_{I=1}^{n} H(X_i)}\right].$$

$$< \alpha_L(n+1) = \frac{n+1}{n}\left[1 - \frac{H(X_i)_{i=1}^{n+1}}{\sum_{I=1}^{n+1} H(X_i)}\right], \forall n \in N, n \geq 2 \quad (30)$$

(v) It is trivial.

Note that the proposed strength statistic is strictly increasing on its number of variables, but Cramer's V is not. Therefore, Liu's generalized correlation coefficient is more sensitive and useful than Cramer's V, and it can be tested by G^2.

6. Conclusion

In this paper, I pointed out that both Joe's and Ding's generalized correlation coefficients can be not used for detecting the mutual independence of three or more multivariate variables, further, I proposed a novel strength statistic based on the significant statistic G^2, called Liu's generalized correlation coefficient, it can be used to detect and compare the strength of not only the pair-wise but also the mutual independence of any multivariate variables, moreover, I proved that only the new strength statistic is a strictly increasing function on its number of variables, it is more sensitive and useful than Cramer's V, moreover, this new strength statistic can also be tested by G^2.

This work is partially supported by the National Science Council grant (NSC 100-2511-S-468-001).

References

1. H. Cramer, *Mathematical Methods of Statistics* (Princeton: Princeton University Press, 1946)
2. K. Pearson, Philos. Mag **50** (1900)
3. W.G. Cochran, Biometrics **10** (1954)
4. T.D. Wickens, *Multiway contingency tables analysis for the social sciences* (Erlbaum, NJ, 1989)
5. M.L. McHugh, Biochem. Medica **23**, 2 (2013)
6. H.-C. Liu, Proceeding of 2015 Global Chinese Conference on Educational Information and Assessment (2015)
7. R.R. Sokal, F.J. Rohlf, *Biometry: The principles and practices of statistics in biological research* (WH Freeman and company, San Francisco, 1994)
8. H. Joe, J. Am. Stat. Assoc **84** (1989)
9. J .Ding, W. S. Wang, Y. L. Zhao, J. Sichuan. Univ: Eng. Sci. Ed **34,** 3 (2002)
10. T. Cover, and J. Thomas, *Elements of Information Theory* (John Wiley and Sons, 1991)