Association for the Chi-square Test

David J. Olive*

Southern Illinois University

February 8, 2012

Abstract

A problem with measures of association for the chi-square test is that the measures depend on the number of observations N and on the dimension of the $r \times c$ contingency table. Hence $C = 0.5$ for one contingency table and $C = 0.2$ for another contingency table does not necessarily mean that the association is higher in the first table than the second. There are two measures of association that tend to be small when the chi-square test statistic X^2 is not significant provided $N > 10(r-1)(c-1)$.

KEY WORDS: categorical data, contingency coefficient C, Cramer's V

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu.

1. THE MAXIMUM VALUE OF X^2

The chi-square test is used to test whether there is an association between two categorical variables: the row variable with r categories and the column variable with c categories. The chi–square test statistic =

$$X^2 = \sum_i^r \sum_j^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed count of the ijth cell and the expected cell count under independence is $E_{ij} = (i\text{th row total})(j\text{th column total})/N$ where $N = \sum_i \sum_j O_{ij}$ is the total number of observations.

Note that $X^2 = 0$ if all of the rc observed counts equal the expected counts: $O_{ij} = E_{ij}$. Let $q = \min(r, c)$. Cramér (1946, p. 443) showed that the maximum value of X^2 is $X^2_M = N(q-1)$, and that the maximum occurs when all of the cell counts are zeros except q nonempty cells such that there is at most one nonempty cell in each row and each column. Thus X^2 is smallest under independence or no association, and X^2 is largest if the categorical variables are "functions of each other" in that if $q = r$, then the ith level of the row variable was observed only with the $j(i)$th level of the column variable and vice verca. If $q = c$, then the jth level of the column variable was observed only with the $i(j)$th level of the column variable and vice verca.

For example, in the following table, let a_i be the count of the nonempty cell in the ith row. Then categories $r1$ and $c2$, $r2$ and $c5$, and $r3$ and $c3$ occur together.

Row/Column	c1	c2	c3	c4	c5	row total
r1	0	a_1	0	0	0	a_1
r2	0	0	0	0	a_2	a_2
r3	0	0	a_3	0	0	a_3
column total	0	a_1	a_3	0	a_2	N

To see that such a configuration of q nonempty cells has $X^2 = N(q-1)$, define $(0-0)^2/0 = 0$ in the sum for X^2. Since the variables are categorical, the categories of each variable can be arranged so that the nonempty cells counts are $O_{11} = a_1$, ..., $O_{qq} = a_q$ and the $r-q$ rows or $c-q$ columns where all of the counts are zeros can be omitted, resulting in the following "computational table" with

$$X^2 = \sum_i^q \sum_j^q \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

and $N = \sum_i^q a_i$. Note that $N^2 = (\sum_i^q a_i)(\sum_j^q a_j) = \sum_i \sum_j a_i a_j = \sum_i a_i^2 + \sum \sum_{i \neq j} a_i a_j$.

Row/Column	c1	c2	c3	$\cdots$	cq	row total
r1	a_1	0	0	$\cdots$	0	a_1
r2	0	a_2	0	$\cdots$	0	a_2
r3	0	0	a_3	$\cdots$	0	a_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
rq	0	0	0	$\cdots$	a_q	a_q
column total	a_1	a_2	a_3	$\cdots$	a_q	N

Hence

$$X^2 = \sum_i \frac{(a_i - \frac{a_i^2}{N})^2}{\frac{a_i^2}{N}} + \sum_i \sum_{\neq j} \frac{(0 - \frac{a_i a_j}{N})^2}{\frac{a_i a_j}{N}} = \sum_i \frac{(\frac{Na_i - a_i^2}{N})^2}{\frac{a_i^2}{N}} + \sum_i \sum_{\neq j} \frac{a_i a_j}{N} =$$

$$\sum_i \frac{[a_i(N-a_i)]^2}{a_i^2\ N} + \frac{1}{N}[\sum_i \sum_{\neq\ j} a_i a_j + \sum_i a_i^2] - \frac{1}{N}\sum_i a_i^2 =$$

$$\sum_i \frac{(N-a_i)^2}{N} + \frac{1}{N}N^2 - \frac{1}{N}\sum_i a_i^2 = N + \sum_i \left[\frac{(N-a_i)^2 - a_i^2}{N}\right] =$$

$$N + \sum_i \left[\frac{N^2 - 2Na_i + a_i^2 - a_i^2}{N}\right] = N + \sum_i \left[\frac{N^2 - 2Na_i}{N}\right] = N + \sum_i [N - 2a_i]$$

$$= N + qN - 2\sum_i a_i = N + qN - 2N = N(q-1).$$

2. MEASURES OF ASSOCIATION

Gibbons (1985), Goodman and Kruskal (1954) and Wikipedia (2012) review measures of association for the chi-square test. The Cramér (1946, p. 443) contingency coefficient

$$V^2 = \frac{X^2}{N(q-1)}$$

and Cramer's $V = \sqrt{V^2}$. The coefficient of mean square contingency or contingency coefficient

$$C = \sqrt{\frac{X^2}{N + X^2}}.$$

The Sakoda (1977) adjusted contingency coefficient

$$C^* = \sqrt{\frac{q}{q-1}}C = \sqrt{\frac{q}{q-1}}\sqrt{\frac{X^2}{N + X^2}}.$$

Let A be V^2, V, C^* or C^{*2}. Then association measure A satisfies $0 \leq A \leq 1$ with $A = 0$ if all of the $O_{ij} = E_{ij}$ and $A = 1$ if $X^2 = N(q-1)$. Hence $C^* > C$ and A is near 0 if X^2 is small and A is near 1 if X^2 is near its maximum so that the association is large.

Measures of association need to be used with care. For multiple linear regression, the coefficient of multiple determination R^2 is a measure of linear association. If the population coefficient is $\delta \neq 0$, then for large enough sample size n, the Anova F statistic will be significant and R^2 close to δ. Cramér (1946, pp. 414-415) suggests that R^2 should be considerably larger than p/n if the p predictors are useful, and for iid normal errors and 0 slopes, notes that $E(R^2) = (p-1)/(n-1)$. If $n_1 >> p_1$ and $n_2 >> p_2$ then $R_1^2 = 0.7$ suggests stronger linear association than $R_2^2 = 0.6$, but if $n_1 = k_1 p_1$ and $n_2 = k_2 p_2$ where k_1 and k_2 are small, then no such comparison can be made. In fact, $R_i^2 = 1$ if $k_1 = k_2 = 1$.

These types of problems are compounded for association measures for contingency tables. Goodman and Kruskal (1954, p. 740) note that such association measures depend on r and c. Hence $A = 0.6$ for one contingency table and $A = 0.2$ for another contingency table does not necessarily mean that the association is higher in the first table than the second. Conover (1971, p. 177) notes that V^2 depends on r and c for its interpretation since X^2 tends to be larger the larger $(r-1)(c-1)$ is, and dividing by $q-1$ "only partially offsets this tendency." Smith and Albaum (2004, p. 631) suggests that C can be larger if $r \neq c$ than if $q = r = c$.

To further examine these problems, first note that the 98th percentile of the χ_d^2 distribution is approximately $d + 3\sqrt{d} \approx d + 2.121\sqrt{2d}$. Let $d = (r-1)(c-1)$. Association measures A seem comparable for tables of the same dimension and N. As $N \to \infty$, $A \to 0$ if X^2 is close to $d + 3\sqrt{d}$. If $X^2 = d$, and $N = kd$, then $V^2 = 1/[k(q-1)]$ and

$$C^* = \sqrt{\frac{q}{q-1}}\sqrt{\frac{1}{k+1}}.$$

Hence $N > 10d$ suggests C^{*2} and V^2 will be small (< 0.2) when X^2 is not significant.

REFERENCES

Conover, J.W. (1971), *Practical Nonparametric Statistics*, Wiley, New York, NY.

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

Gibbons, J.A. (1985), "Shrinkage Formulas for Two Nominal Level Measures of Association," *Educational and Psychological Measurement*, 45, 551-566.

Goodman, L.A., and Kruskal, W.H. (1954), "Measures for Association for Cross-Classification," *Journal of the American Statistical Association*, 49: 732-764.

Sakoda, J.M. (1977), "Measures of Association for Multivariate Contingency Tables," *Proceedings of the Social Statistics Section of the American Statistical Association* (Part III), 777-780.

Smith, S.C., and Albaum, G.S. (2004), *Fundamentals of Marketing Research*, Sage Publications, Thousand Oaks, CA.

Wikipedia (2012), "Contingency Table," online at (http://en.wikipedia.org/wiki/Contingency_table)