The Korean Communications
in Statistics Vol. 5, No. 1, 1998
pp. 35-42

A Study on Cell Influences to Chi-square Statistic in Contingency Tables[1]

Honggie Kim[2]

Abstract

Once a contingency table is constructed, the first interest will be the hypotheses of either homogeneity or independence depending on the sampling scheme. The most widely used test statistic in practice is the classical Pearson's χ^2 statistic. When the null hypothesis is rejected, another natural interest becomes which cell contributed to the rejection of the null hypothesis more than others. For this purpose, so called cell χ^2 components are investigated. In this paper, the influence function of a cell to the χ^2 statistic is derived, which can be used for the same purpose. This function measures the effect of each cell to the χ^2 statistic. A numerical example is given to demonstrate the role of the new function.

1. Introduction

Contingency tables are well known summarized forms of catagorical data arising in many fields of researches. In spite of those excellent advanced statistical techniques such as log-linear model, correspondence analysis, the most widely used statistic is still the classical Pearson's χ^2 statistic, since it is the easiest.

In an analysis of high dimensional contingency tables, log-linear model analysis may be more appropriate, but most researchers are still slicing their tables and apply two-way analysis. It's simply because they don't have enough understanding of the theoretically complex techniques. Confronting all these situations, the importance of the role of the classical χ^2 statistic can not be underestimated.

In an analysis of a two-way contingency table, the first interest will be a hypothesis of independence between two categorical variables which the rows and colums of the contingency

[1] This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1996.

[2] Associate Professor, Department of Statistics, Chungnam National University, Daejon, 305-764, Korea.

table consist of, or a hypothesis of homogeneity among rows of the contingency table, depending on the sampling scheme.

The most popular statistic among reseachers for testing either of these hypotheses as a null hypothesis is the Pearson's χ^2 statistic, the theory of which is well introduced in most elementary statistical texts.

When the null hypothesis is rejected according to the χ^2 test, the next natural interest becomes which cells contributed to the rejection more than others. For this purpose, cell χ^2 components are used.

A cell χ^2 component is the square of the differences between the observed and expected cell frequencies divided by the expected cell frequency, where the expected cell frequencies are obtained under the null hypothesis. Irwin (1949), Kimball (1954), Kastenbaum (1960), and Kass (1980) are among those researcher who were interested in χ^2 statistic and its components.

The idea of influence function is first introduced by Hampel (1974). Cook and Weisberg (1980) used this technique in detection of outliers in regression. Critchley (1985) studied influence in principal component analysis, and Campbell (1978) obtained some interesting results on influence in discriminant analysis. Kim (1992) derived influence functions in correspondence analysis, which has been extended to multiple correspondence analysis in Kim (1994).

By applying Hampel's idea and treating the χ^2 statistic as the trace of a multiplication of matrices, which are obtained from the observed contingency table, Kim and Lee (1996) derived the influence of an observation to the χ^2 statistic as a function.

The result of Kim and Lee (1996) will be extended to derive the influence function of a cell, instead of an observation, to the χ^2 statistic, which can be used for the same purpose as the cell χ^2 components.

2. Extension

Let $N=\{n_{ij}\}$ be an $(I\times J)$ contingency table with n_{i+} $(i=1,\cdots,I)$ being the i^{th} row total, n_{+j} $(j=1,\cdots,J)$ being the j^{th} column total and n being the total frequencies in N. Under the null hypothesis of independence or homogeneity, the expected cell frequency is given by

$$e_{ij}=\frac{n_{i+}\cdot n_{+j}}{n} \quad , \quad i=1,\cdots,I \quad , \quad j=1,\cdots,J \quad ,$$

and the Pearson's χ^2 statistic is then

$$X^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

The estimated probability matrix F is obtained by dividing the entries of N by n.

Let

$$r_i = \frac{n_{i+}}{n} \quad , \ i = 1, \cdots, I \qquad \text{and} \qquad c_j = \frac{n_{+j}}{n} \quad , \ j = 1, \cdots, J$$

be the estimated marginal probabilities. Consider the two vectors $r = \{r_i\}$ and $c = \{c_j\}$. Letting D_r and D_c be diagonal matrices with r and c as their diagonals, the χ^2 statistic is given by (Greenacre, 1984)

$$X^2 = n\, trace\{D_r^{-1}(P - rc^t)D_c^{-1}(P - rc^t)^t\} \ . \qquad (1)$$

We will regard the probability matrix F as known to derive an influence function. When F is in fact an estimated probability matrix, the influence function is called an empirical influence function, which is an estimated influence function. Of course, the latter will be the one we can use in practice.

Define an $(I \times J)$ random matrix Y so that its $(i, j)^{th}$ element is 1 and others are 0 when a randomly chosen subject is classified into i^{th} row category and j^{th} column category. Let Y have a distribution F, which is multinomial $M(1, P)$. Now, we can see that the probability matrix F is a functional evaluated on F. That is,

$$P = EY = \int Y dF \ . \qquad (2)$$

Also X^2 given by (1) is a functional evaluated on F, since X^2 is a function of F.

Given i and j, let y_{ij} be an $(I \times J)$ matrix where the $(i, j)^{th}$ element is 1 and the others are 0. That is, y_{ij} is a realization of the random matrix Y. To measure the influence of an observation y_{ij} on $X^2 = T(F)$, we use the influence function (IF) of Hampel (1974) which is defined as :

$$IF(X^2, y_{ij}) = \lim_{\varepsilon \to 0} [\, T(F_\varepsilon) - T(F)\,]/\varepsilon$$

where $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\delta_{y_{ij}}$, is a perturbation of F by $\delta_{y_{ij}}$, a measure with point mass one at y_{ij}.

This influence function is derived in Kim and Lee (1996) as

$$IF(X^2, y_{ij}) = 2n\frac{P_{ij} - r_i c_j}{r_i c_j} - \frac{n}{r_i}\sum_{j=1}^{J}\frac{(P_{ij} - r_i c_j)^2}{r_i c_j} - \frac{n}{c_j}\sum_{i=1}^{I}\frac{(P_{ij} - r_i c_j)^2}{r_i c_j} \ .$$

Now, using the idea of perturbation by multiple observations introduced in Kim (1992), we perturb F by $n_{ij}\delta_{y_{ij}}$ to assess influence of a cell to the χ^2statistic. That is, the perturbed F

now becomes

$$F_\varepsilon = (1-\varepsilon n_{ij})F + \varepsilon n_{ij}\delta_{y_{ij}}$$

Perhurbing F produces a perturbation of P, hence a perturbation of X^2, denoted X^2_ε. The influence of the (i , j) cell on X^2 can be measured by

$$IF(X^2, n_{ij}) = \lim_{\varepsilon\to 0}[X^2_\varepsilon - X^2]/\varepsilon\ . \qquad (3)$$

As in Kim and Lee (1996), we replace P, r, c, D_r^{-1}, and D_c^{-1} in (1) with the corresponding perturbations, P_ε , r_ε , c_ε , $(D_r^{-1})_\varepsilon$ and $(D_c^{-1})_\varepsilon$, with the subscript ε meaning a perturbation.

Similar algebraic manipulation as in Kim and Lee (1996) gives

$$\begin{aligned} P_\varepsilon &= \int YdF_\varepsilon \\ &= \int Yd\,[(1-\varepsilon n_{ij})F + \varepsilon n_{ij}\delta_{y_{ij}}] \\ &= (1-\varepsilon n_{ij})\int YdF + \varepsilon n_{ij}\int Yd\delta_{y_{ij}} \\ &= (1-\varepsilon n_{ij})P + \varepsilon n_{ij}y_{ij}\ , \end{aligned}$$

$$\begin{aligned} r_\varepsilon &= P_\varepsilon 1 \\ &= (1-\varepsilon n_{ij})r + \varepsilon n_{ij}y_i\ , \end{aligned}$$

$$\begin{aligned} c_\varepsilon &= P_\varepsilon^t\ 1 \\ &= (1-\varepsilon n_{ij})c + \varepsilon n_{ij}y_j\ , \end{aligned}$$

where y_i and y_j are $(I\times 1)$ and $(J\times 1)$ unit vectors with i^{th} and j^{th} elements 1, and

$$P_\varepsilon - r_\varepsilon c_\varepsilon^t = (1-\varepsilon n_{ij})(P - rc^t) + \varepsilon n_{ij}(y_{ij} + rc^t - ry_j^t - y_i c^t) + O(\varepsilon^2)\ .$$

Letting

$$M = D_r^{-1}(P - rc^t)D_c^{-1}(P - rc^t)^t$$

gives the χ^2 statistic

$$X^2 = n\ trace(M).$$

The influence function given in (3) becomes

$$IF(X^2, n_{ij}) = \lim_{\varepsilon\to 0}[n\ trace(M_\varepsilon) - n\ trace(M)]/\varepsilon\ , \qquad (4)$$

where M_ε is the perturbed M.

Expanding M_ε gives

$$M_\varepsilon = M + \varepsilon n_{ij}(A_1 + A_2 + A_3 + A_4) + O(\varepsilon^2) \ ,$$

where

$$\begin{aligned}
A_1 &= -\frac{1}{r_i^2}\, diag(y_i)(P - rc^t)D_c^{-1}(P - rc^t)^t \\
A_2 &= D_r^{-1}(y_{ij} + rc^t - ry_j^t - y_i c^t)D_c^{-1}(P - rc^t)^t \\
A_3 &= -\frac{1}{c_j^2}\, D_r^{-1}(P - rc^t)diag(y_j)(P - rc^t)^t \\
A_4 &= D_r^{-1}(P - rc^t)D_c^{-1}(y_{ij}^t + cr^t - y_j r^t - cy_i^t) \ .
\end{aligned}$$

By Kim and Lee (1996), the influence given in (4) has the form

$$IF(X^2,\ n_{ij}) = n_{ij} IF(X^2, y_{ij})$$

or more specifically

$$IF(X^2, n_{ij}) = 2nn_{ij}\frac{P_{ij} - r_i c_j}{r_i c_j} - n\frac{n_{ij}}{r_i}\sum_{j=1}^{J}\frac{(P_{ij} - r_i c_j)^2}{r_i c_j} - n\frac{n_{ij}}{c_j}\sum_{i=1}^{I}\frac{(P_{ij} - r_i c_j)^2}{r_i c_j} \qquad (5)$$

This function measures the instantaneous rate of change in χ^2 statistic when a cell frequency is doubled.

3. Numerical Example

Table 1 contains a 8×5 contingency table taken from Guttman (1971). It represents 1554 Israeli adults cross-classified according to their types of principal worries(rows), and countries of origin(columns). The data are also used by Greenacre (1984) to illustrate correspondence analysis, and by Kim and Lee (1996) to demonstrate influence functions of single observation in contingency table.

The χ^2 statistic computed from this contingency table is 120.44, which leads us rejection of the null hypothesis of independence.

Table 2 includes three entries, the difference between observed and expected cell frequencies, the cell χ^2 component, and the evaluation of the influence function given by (5). The third entry is in fact

$$IF(X^2, n_{ij})\ \frac{1}{n+1} \ ,$$

which represents the estimated changes in χ^2 statistic when a cell frequency is doubled.

As we can see from the table, those cells with large χ^2 components such as cell (4, 1), cell (6, 4), cell (8, 1), and cell (8, 2), also show large values of influence function in absolute sense. Note that the signs of the evaluations of the influence function conincide with those of the differences between observed and expected frequencies. When the difference is positive, that is, observed frequency is larger than expected, the influence function value is also positive, and vice versa. Of course, there are some reversed signs when the cell χ^2 component is very small. Our influence function tells which cells are causing the rejection of the null hypothesis through its magnitude and why (observed frequency is less than or greater than expected) through its sign, while the cell χ^2 component loses this second property because of the squaring process. Note that the difference between observed and expected frequencies should also be interpreted relatively to the expected frequencies.

Table 1. Principal worries of Israeli adults. Description of categories of variable B, country of origin, is given at the foot of the table.

	Country of origin (B)				
Principal worry (A)	1	2 .	3	4	5
Enlisted relative (1)	61	104	8	22	5
Sabotage (2)	70	117	9	24	7
Military situation (3)	97	218	12	28	14
Political situation (4)	32	118·	6	28	7
Economic situation (5)	4	11	1	2	1
Other (6)	81	128	14	52	12
More than one worry (7)	20	42	2	6	0
Personal economics (8)	104	48	14	16	9

1: From Asia or Africa
2: From Europe or America
3: From Israel and their father from Asia or Africa
4: From Israel and their father from Europe or America
5: From Israel and their father from Israel

Table 2. Differencies between observed and expected frequencies, cell χ^2 components, and evaluated influence functions.

O-E cell χ^2 IF	1	2	3	4	5
1	0.64 0.0068 -5.6364	2.84 0.0798 0.4160	-0.49 0.0288 -1.7344	-0.91 0.0360 -4.2922	-2.08 0.6103 -3.3655
2	1.49 0.0324 -4.7670	2.19 0.0416 -1.4157	-0.64 0.0426 -2.0826	-2.00 0.1540 -6.4224	-1.03 0.1331 -2.3898
3	-14.36 1.8529 -39.0425	31.36 5.2703 55.0886	-3.67 0.8603 -7.1988	-14.27 4.8154 -23.0020	0.94 0.0677 0.3738
4	-25.64 11.4080 -35.0496	21.39 4.7377 35.1522	-2.11 0.5499 -4.2810	6.12 1.7133 9.8280	0.34 0.0085 -0.7546
5	-1.73 0.5245 -3.0560	1.39 0.2010 2.1010	0.19 0.0462 0.3317	-0.18 0.0143 -0.6474	0.33 0.1595 0.8414
6	-5.62 0.3643 -23.3766	-17.16 2.0290 -42.7392	1.81 0.2690 2.1140	19.13 11.1280 52.0884	1.84 0.3342 2.7720
7	-1.13 0.0600 -5.6380	6.39 1.2283 10.8486	-0.97 0.3184 -1.6328	-1.98 0.5079 -4.0836	-2.48 2.4775 0.0000
8	46.36 37.2780 118.5600	-48.61 24.4560 -67.7520	5.89 4.2738 13.9510	-5.88 1.5791 -16.1040	2.33 0.7423 2.0016

References

[1] Campbell, N.A. (1978). The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, Vol. 27, 251-258.

[2] Cook, R.D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression, *Technometris*, Vol. 22, 495-508.

[3] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, London.

[4] Critchley, F. (1985). Influence in principal components analysis, *Biometrika,* Vol. 72, 627-636.

[5] Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, New York.

[6] Guttman, L. (1974). Measurement as structural theory, *Psychometrika,* Vol. 36, 329-347.

[7] Hampel, F. (1974). The influence curve and its role in robust estimation, *Journal of American Statistical Association*, Vol. 69, 383-393.

[8] Irwin, J.O. (1949). A note on the subdivision of χ^2 into components, *Biometrika,* Vol. 36, 130-134.

[9] Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, Vol. 29, 119-127.

[10] Kastenbaum, M.A. (1960). A note on the additive partitioning of chi-square in contingency tables, *Biometrics*, Vol. 16, 416-422.

[11] Kim, H. (1992). Measures of influence in correspondence analysis, *Journal of Statistical Computation and Simulation*, Vol. 40, 201-217.

[12] Kim, H. (1994). Influence functions in multiple correspondence analysis, *The Korean Journal of Applied Statistics*, Vol. 7, 69-74.

[13] Kim, H. and Lee, H.S. (1996). Influence Functions on χ^2 statistic in Contingency Tables, *Journal of the Korean Communications in Statistics*, Vol. 3, No. 2, 69-76.

[14] Kimball, A.W. (1954). Short-cut formulars for the exact partitioning of χ^2 in contingency tables. *Biometrics*. Vol. 10. 452-458.