The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 305-317

The Changes in χ^2 Statistic when a Row is Deleted from a Contingency Table

Heesook Lee[1], Honggie Kim[2]

Abstract

We suggest methods to measure the changes in χ^2 statistic when a row is deleted from a two-way contingency table. The influence function is extended and the deletion method is applied. Two examples are presented and we compare the results obtained from the influence function method and the deletion method.

Keywords: Influence function, Deletion method, Chi-square statistics, Contingency table

1. 서론

분할표의 대표적인 형태인 이차원 분할표의 분석에 있어 연구자가 최초로 관심을 갖게 되는 것은 분할표를 이루는 두 변수 간의 독립성(independence) 혹은 행을 이루는 집단들 간의 동질성(homogeneity)에 관한 가설이다. 이러한 가설 검정에 가장 널리 쓰이는 통계량은 Pearson이 제안한 카이제곱 통계량이다. 독립성 혹은 동질성의 가설인 귀무가설이 기각되면 그 다음의 분석 절차는 어느 칸(셀)이 귀무가설의 기각에 영향을 미쳤는가를 밝히는 과정이 될 것이다. 이러한 목적으로 많이 쓰이는 통계량값는 칸 카이제곱 성분(cell Chi-square components)이다. 칸 카이제곱 성분은 각 칸의 관측도수와 귀무가설 아래에서 얻은 기대도수의 차이를 제곱하여 그것을 다시 기대도수로 나눈 값이다.

카이제곱 통계량과 칸 성분에 대한 연구들 중에 대표적인 것에는 Irwin(1949), Kimball(1954), Kastenbaum(1960), Kass(1980) 등이 있다.

영향함수(influence function)는 Hampel(1974)에 의해 처음으로 소개 되었다. 그 이후, Campbell(1978)은 판별분석(discriminant analysis)에서 이상값(outlier) 탐지에 영향함수를 사용하였고 Radhakrishnan & Kshirsagar(1981)은 다변량 분석에서 여러 가지 모수에 대한 이론적인 영향함수들을 유도했다. 또한 Cook & Weisberg(1982)는 회귀분석에서 회귀진단 방법으로 영향함수를 이용하였고, Critchley(1985)는 주성분 분석에서 영향력 있는 관측치를 찾아내기 위해 이 방법

1) Post Doc., Department of Engineering, Information and Communication University, Daejon, KOREA.
E-mail: hslee@icu.ac.kr

2) Professor, Department of Statistics, Chungnam University, Daejon, KOREA.
E-mail: hgkim@stat.cnu.ac.kr

을 적용하였다. Kim(1992)은 이차원 분할표의 대응분석(correspondence analysis)에서 얻어진 고유치들(eigenvalues)에 대한 영향을 측정하였고, Kim(1994)은 다차원 대응분석으로 이를 확장하였다.

이차원 분할표의 카이제곱 통계량에 대한 영향함수 연구들을 살펴보면, Kim & Lee(1996)는 카이제곱 통계량에 대한 각 관찰도수의 영향함수를 유도했고, Kim(1998)은 칸 성분과 같은 역할을 하는 칸 영향함수를 유도했다.

본 논문에서는 이차원 분할표에서 한 행이 제거되었을 때 카이제곱 통계량의 변화를 연구 해보고자 한다. 2절에서는 한 행이 제거되었을 때 카이제곱 통계량의 변화를 Kim & Lee(1996)와 Kim(1998)의 결과를 이용하여 영향함수를 확장하며 3절에서는 소거법(deletion method)을 이용하여 직접 수식으로 유도한다. 4절에서는 범주의 개수가 비교적 적은 자료와 많은 자료인 2개의 자료를 이용하여 적용 예를 보이고 마지막으로 결론에서는 본 연구과제의 결과를 정리하고 실제 자료 분석에서 어떻게 적용될 수 있는지를 설명한다. 본 연구의 결과는 분할표에서 행과 열의 제거 혹은 통합에 응용될 수도 있을 것이며. 또한 예비조사를 통해 얻어진 분할표에 적용해 봄으로써 본 조사의 표본추출 시 표본수의 최적 분배에도 이용될 수 있을 것으로 보인다. 예를 들어, 한 대학에서 학년(1학년, 2학년, 3학년, 4학년)에 따라 지방선거에 대한 참여여부가 동일한 지를 조사하기 위해 설문조사를 실시한다고 하자. 예비조사에서 얻어진 자료에 본 논문의 결과를 이용하면 귀무가설인 동일성의 기각에 영향을 적게 미치는 학년을 밝힐 수 있으며 이 학년은 본조사에서 다른 학년에 비해 상대적으로 적은 표본 수를 할당할 수도 있을 것이다.

2. 영향함수의 확장

행(raw) 범주의 개수가 I개이고, 열(column) 범주의 개수가 J개인 $I \times J$ 분할표를 고려하자. 여기서 $n_{i+}\,(i=1,\cdots,I)$은 i번째 행의 도수 합계, $n_{+j}\,(j=1,\cdots,J)$은 j번째 열의 도수합계, n은 총 도수이다. 확률행렬 $P=p_{ij}\,, i=1,\cdots,I,\quad j=1,\cdots,J$의 추정값 $\hat{p}_{ij}$는 n_{ij} 각각을 총 도수 n으로 나누어 구한다. 즉, $\hat{p}_{ij}=\frac{n_{ij}}{n}$이고, 주변 확률 r_i와 c_j의 추정값는 각각 $\hat{r}_i=\frac{n_{i+}}{n}, i=1,\cdots,I$와 $\hat{c}_j=\frac{n_{+j}}{n}, j=1,\cdots,J$ 으로 구한다. 본 논문에서는 편의상 추정값를 나타내는 ' ^ '을 생략한다.

두 벡터 $\boldsymbol{r}=r_i$와 $\boldsymbol{c}=c_j$을 고려하자. $\boldsymbol{r}$과 $\boldsymbol{c}$를 대각원소로 갖는 대각행렬을 각각 D_r과 D_c라 하면 카이제곱 통계량 X^2은 다음과 같이 행렬의 형태로 주어진다.(Greenacre, 1984)

$$X^2=n\ \mathbf{tr}(\mathrm{D}_r^{-1}(\mathrm{P}-\mathrm{rc}^t)\mathrm{D}_c^{-1}(\mathrm{P}-\mathrm{rc}^t)^t)$$

Kim & Lee(1996)는 i번째 행, j번째 열인 (i,j) 칸에 속하는 하나의 도수 y_{ij}의 영향을 측정하기 위해 Hampel(1974)의 영향함수(Influence function)를 사용하여 카이제곱 통계량 $X^2=T(F)$의

영향함수를 다음과 같이 유도하였다.

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

또한 Kim(1998)은 칸 성분과 같은 역할을 하는 칸 영향함수를 다음과 같이 유도했다.

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

즉, 각 칸의 영향은 하나의 관측치가 칸에 추가될 때의 영향에 칸 도수를 곱한 것과 같아졌다.

Kim & Lee(1996)와 Kim(1998)의 확장으로 카이제곱 통계량에 대한 행 영향은 그 행의 모든 칸 영향들의 합으로 나타낼 수 있다. 즉, i번째 행 도수 벡터 R_i을 $R_i = (n_{i1}, n_{i2}, \cdots, n_{iJ})$라 놓으면 i행에 대한 영향은 그 행에 있는 J개 칸들의 영향을 합한 것과 같아진다. 이를 수식으로 표현하면 다음과 같다.

$$IF(X^2, R_i) = \sum_{j=1}^{J} IF(X^2, n_{ij})$$

3. 소거법(Deletion Method)의 적용

한 행이 제거되었을 때 카이제곱 통계량과 제거되지 않았을 때의 차이를 직접 수식으로 유도해 본다.

i번째 행을 제거한 후 카이제곱 통계량을 행렬로 표시하면 다음과 같이 표현할 수 있다.

$$X^2(i) = n(i)\ \mathbf{tr}\,(\mathrm{D_r(i)}^{-1}(\mathrm{P(i)} - \mathrm{r(i)c(i)}^t)\mathrm{D_c(i)}^{-1}(\mathrm{P(i)} - \mathrm{r(i)c(i)}^t)^t) \tag{3.1}$$

여기서 $n(i)$는 i번째 행을 제거한 후의 총 도수 즉, $n - n_{i+}$을 의미한다. i번째 행을 제거한 확률행렬 $P(i) = \frac{n}{n(i)} I(i)P$ 으로 나타낼 수 있다. 여기서 $I(i)$는 i번째 대각원소가 0인 단위행렬을 의미한다. 식(3.1)에서 $D_r(i)^{-1}$과 $D_c(i)^{-1}$ 은 일반화 역행렬(generalized inverse matrix)로써 대각행렬인 $D_r(i)$와 $D_c(i)$의 0이 아닌 각 대각원소의 역수를 대각원소로 갖는 대각행렬들이다. $D_r(i)$와 $D_c(i)$의 역행렬은 이 행렬들의 i번째 대각원소가 0이므로 존재하지 않는다.

$\mathbf{1}$을 모든 원소가 1인 벡터라 하면 $r(i), c(i), D_r(i), D_c(i)$는 $P(i)$로부터 각각 다음과 같이 표현할 수 있다.

$$\begin{aligned} r(i) &= P(i)\ \mathbf{1} = \frac{n}{n(i)} I(i)P\mathbf{1} = \frac{n}{n(i)} I(i)r \\ c(i) &= \mathbf{1}^t P(i) = \mathbf{1}^t \frac{n}{n(i)} I(i)P \\ &= \mathbf{1}^t [\frac{n}{n(i)} P - \frac{n}{n(i)}(I - I(i))P] \\ &= \frac{n}{n(i)} c - \frac{1}{n(i)} n_i \end{aligned} \tag{3.2}$$

$$D_r(i) = \frac{n}{n(i)} I(i) D_r$$

$$D_c(i) = \frac{n}{n(i)} D_c - \frac{1}{n(i)} D_{n_i}$$

여기서, $\boldsymbol{r}$, $\boldsymbol{c}$, D_r, D_c는 2절에서와 같고, n_i는 $n_i = (n_{i1}, n_{i2}, \cdots, n_{ij})^t$이고 D_{n_i}는 i번째 행의 칸 도수들을 대각원소로 갖는 대각행렬을 나타낸다. 즉, $D_{n_i} = \mathbf{diag}(n_{i1}, n_{i2}, \cdots, n_{ij})$ 이다.

$D_r(i)^{-1}$을 먼저 구해보면 $D_r(i)^{-1}$은 $\frac{n}{n(i)} I(i) D_r$의 0이 아닌 각 대각원소의 역수를 취한 행렬이므로 $D_r(i)^{-1} = \frac{n(i)}{n} I(i) D_r^{-1}$ 이 된다. 또한, $D_c(i)^{-1}$은

$$\begin{aligned} D_c(i)^{-1} &= [\frac{n}{n(i)} D_c - \frac{1}{n(i)} D_{n_i}]^{-1} \\ &= n(i) D_{n_i}^{-1} [n(i) D_{n_i}^{-1} - \frac{n(i)}{n} D_c^{-1}]^{-1} \frac{n(i)}{n} D_c^{-1} \\ &= \frac{n(i)}{n} D_{n_i}^{-1} (D_{n_i}^{-1} - \frac{1}{n} D_c^{-1})^{-1} D_c^{-1} \end{aligned} \tag{3.3}$$

이 된다. 식(3.3)의 $D_c(i)^{-1}$의 1행 1열의 원소 $\frac{n(i)}{n_{+1} - n_{i1}}$을 Taylor 급수전개 해 보면

$$\begin{aligned} \frac{n(i)}{n_{+1} - n_{i1}} &= \frac{n(i)}{n_{+1}} (\frac{n_{+1}}{n_{+1} - n_{i1}}) \\ &= \frac{n(i)}{n_{+1}} \left(\frac{1}{1 - \frac{n_{i1}}{n_{+1}}} \right) \\ &\approx \frac{n(i)}{n_{+1}} [1 + \frac{n_{i1}}{n_{+1}}] \qquad , n_{+1} \gg n_{i1} \\ &= n(i) [\frac{1}{n_{+1}} + \frac{n_{i1}}{(n_{+1})^2}] \qquad , n_{+1} \gg n_{i1} \\ &\approx \frac{n(i)}{n_{+1}} = \frac{n(i)}{n} \cdot \frac{n}{n_{+1}} \end{aligned}$$

이 된다. 이 방법을 $D_c(i)^{-1}$의 모든 대각원소에 적용 해 보면

$$D_c(i)^{-1} \approx \frac{n(i)}{n} D_c^{-1}$$

로 근사 시킬 수 있다.

i행을 제거한 후 카이제곱 통계량을 행렬로 표시한 식(3.1)에 $r(i) = \frac{n}{n(i)} I(i) r$, $c(i) = \frac{n}{n(i)} c - \frac{1}{n(i)} n_i$, $D_r(i)^{-1} = \frac{n}{n(i)} I(i) D_r^{-1}$, $D_c(i)^{-1} \approx \frac{n}{n(i)} D_c^{-1}$ 을 대입하면, $D_c(i)$만 근사시켰을 때 i행을 제거한 후 카이제곱 통계량 $X_1^2(i)$은 다음과 같이 표현된다.

$$\begin{aligned} X_1^2(i) &\approx n(i)\ \mathrm{tr}(D_r(i)^{-1}(P(i)-r(i)c(i)^t)D_c(i)^{-1}(P(i)-r(i)c(i)^t)^t) \\ &= n(i)\ \mathrm{tr}(A_1+A_2+A_3+A_4) \\ &= n(i)[\ \mathrm{tr}(A_1)+\ \mathrm{tr}(A_2)+\ \mathrm{tr}(A_3)+\ \mathrm{tr}(A_4)] \end{aligned} \tag{3.4}$$

여기서, A_1, A_2, A_3, A_4는

$A_1 = I(i)D_r^{-1}(P-rc^t)D_c^{-1}(P-rc^t)^t I(i)$, $\quad A_2 = I(i)D_r^{-1}HD_c^{-1}(P-rc^t)^t I(i)$,

$A_3 = I(i)D_r^{-1}(P-rc^t)D_c^{-1}H^t I(i)$, $\quad A_4 = I(i)D_r^{-1}HD_c^{-1}H^t I(i)$

이 되며, H는 $H=\frac{1}{n(i)}rn_i^t-\frac{n_{i+}}{n(i)}rc^t$ 이다. 또한, A_1, A_2, A_3, A_4의 트레이스를 구해보면 다음과 같이 나타낼 수 있다.

$$\mathrm{tr}(A_1)=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}-\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}$$

$$\mathrm{tr}(A_2)=\frac{1}{n(i)}\sum_{k=1,k\neq i}^{I}\sum_{j=1}^{J}c_j^{-1}(n_{ij}-n_{i+}c_j)(p_{kj}-r_kc_j)$$

$$\mathrm{tr}(A_3)=\ \mathrm{tr}(A_2)$$

$$\mathrm{tr}(A_4)=(\frac{1}{n(i)})^2\sum_{j=1}^{J}c_j^{-1}(n_{ij}-n_{i+}c_j)(1-r_i)$$

식(3.4)를 정리하면 i행을 제거한 후 카이제곱 통계량 $X_1^2(i)$은

$$\begin{aligned} X_1^2(i) \approx\ & n(i)[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}-\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}+2\frac{1}{n(i)}\sum_{k\neq i}^{I}\sum_{j=1}^{J}c_j^{-1}(n_{ij}-n_{i+}c_j)(p_{kj}-r_kc_j) \\ & +(\frac{1}{n(i)})^2\sum_{j=1}^{J}c_j^{-1}(n_{ij}-n_{i+}c_j)(1-r_i)] \\ =\ & \frac{n(i)}{n}X^2-n(i)\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}+2\sum_{k\neq i}^{I}\sum_{j=1}^{J}c_j^{-1}(n_{ij}-n_{i+}c_j)(p_{kj}-r_kc_j) \\ & +\frac{1}{n(i)}\sum_{j=1}^{J}c_j^{-1}(n_{ij}-n_{i+}c_j)(1-r_i) \end{aligned} \tag{3.5}$$

이 된다. 식(3.5)에서 볼 수 있듯이 i행을 제거한 후의 카이제곱 통계량 $X_1^2(i)$은 근사적으로 제거하기 전의 카이제곱 통계량 X^2로 표현되어진다. 따라서, i행을 제거하기 전의 카이제곱 통계량 X^2과 제거한 후의 카이제곱 통계량 $X_1^2(i)$의 차이는

$$\begin{aligned} X^2-X_1^2(i) \approx\ & \frac{n_{i+}}{n}X^2 \\ & +n(i)\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}-2\sum_{k\neq i}^{I}\sum_{j=1}^{J}c_j^{-1}(n_{ij}-n_{i+}c_j)(p_{kj}-r_kc_j) \\ & -\frac{1}{n(i)}\sum_{j=1}^{J}c_j^{-1}(n_{ij}-n_{i+}c_j)(1-r_i) \end{aligned}$$

이 된다.

식(3.2)에서 i행의 칸 도수 벡터 n_i는 i행을 제거한 후 총 도수 $n(i)$에 비해 상당히 작은 수로 생각 할 수 있으므로 $c(i)=\frac{n}{n(i)}c-\frac{1}{n(i)}n_i$, $n_i^t=(n_{i1},n_{i2},\cdots,n_{ij})$을 $c(i)\approx\frac{n}{n(i)}c$ 로 근사 시킬 수 있다.

i행을 제거한 후 카이제곱 통계량 $X^2(i)$을 행렬로 표시한 식(3.1)에 $D_c(i)$와 $c(i)$ 모두 근사시킨 $r(i)=\frac{n}{n(i)}I(i)r$, $c(i)\approx\frac{n}{n(i)}c$, $D_r(i)^{-1}=\frac{n}{n(i)}I(i)D_r^{-1}$, $D_c(i)^{-1}\approx\frac{n}{n(i)}D_c^{-1}$ 을 대입하면, $D_c(i)$와 $c(i)$ 모두 근사시킨 i행을 제거한 후 카이제곱 통계량 $X_2^2(i)$은

$$\begin{aligned} X_2^2(i) &\approx n(i)\ \mathrm{tr}(D_r(i)^{-1}(P(i)-r(i)c(i)^t)D_c(i)^{-1}(P(i)-r(i)c(i)^t)^t) \\ &= n(i)\ \mathrm{tr}(A_1-A_2-A_3+A_4) \\ &= n(i)[\ \mathrm{tr}(A_1)-\ \mathrm{tr}(A_2)-\ \mathrm{tr}(A_3)+\ \mathrm{tr}(A_4)] \end{aligned} \tag{3.6}$$

이 된다. 여기서, A_1,A_2,A_3,A_4는 $A_1=I(i)D_r^{-1}(P-rc^t)D_c^{-1}(P-rc^t)^tI(i)$, $A_2=I(i)D_r^{-1}HD_c^{-1}(P-rc^t)^tI(i)$, $A_3=I(i)D_r^{-1}(P-rc^t)D_c^{-1}H^tI(i)$, $A_4=I(i)D_r^{-1}HD_c^{-1}H^tI(i)$ 이 되며, H는 $H=\frac{n_{i+}}{n(i)}rc^t$ 이다.

또한, A_1,A_2,A_3,A_4의 트레이스를 구해보면 다음과 같이 나타낼 수 있다.

$$\mathrm{tr}(A_1)=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}-\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}$$

$$\mathrm{tr}(A_2)=\frac{n_{i+}}{n(i)}[\sum_{i=1}^{I}\sum_{j=1}^{J}(p_{ij}-r_ic_j)-\sum_{j=1}^{J}(p_{ij}-r_ic_j)]$$

$$\mathrm{tr}(A_3)=\ \mathrm{tr}(A_2)$$

$$\mathrm{tr}(A_4)=(\frac{n_{i+}}{n(i)})^2(1-r_i)$$

식(3.6)을 정리하면

$$\begin{aligned} X_2^2(i)\approx\ & n(i)[\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}-\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j} \\ & -2\frac{n_{i+}}{n(i)}(\sum_{i=1}^{I}\sum_{j=1}^{J}(p_{ij}-r_ic_j)-\sum_{j=1}^{J}(p_{ij}-r_ic_j)) \\ & +(\frac{n_{i+}}{n(i)})^2(1-r_i)] \\ =\ & \frac{n(i)}{n}X^2-n(i)\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j}-2n_{i+}\sum_{i=1}^{I}\sum_{j=1}^{J}(p_{ij}-r_ic_j) \\ & +2n_{i+}\sum_{j=1}^{J}(p_{ij}-r_ic_j)+\frac{n_{i+}^2}{n(i)}(1-r_i) \end{aligned}$$

이 된다. 따라서, i행을 제거한 후 카이제곱 통계량과 제거하기 전의 카이제곱 통계량의 차이 $X^2-X_2^2(i)$는 근사적으로 다음과 같이 된다.

$$X^2 - X_2^2(i) \approx \frac{n_{i+}}{n} X^2 + n(i) \sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} - 2n_{i+} \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - r_i c_j)$$
$$- 2n_{i+} \sum_{j=1}^{J} (p_{ij} - r_i c_j) - \frac{n_{i+}^2}{n(i)} (1 - r_i)$$

4. 적용 예

본 절에서는 2절과 3절에서 언급한 세 가지 방법을 적용하기 위해 실제 자료를 이용한다. 상대적으로 범주의 개수가 적은 자료와 많은 자료, 두 가지를 가지고 비교 해 본다.

[표 4-1] 이스라엘 성인 1554명의 주요근심거리와 출신지

	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

[표 4-1]는 Guttman(1971)으로부터 얻은 8×5 분할표이다. 이 분할표는 이스라엘에서 1554명의 성인을 대상으로 주요걱정거리와 출신지를 조사한 결과이다. 이 자료는 Greenacre(1974)와 Kim & Lee(1996), Kim(1998)에서도 실제 자료로 사용되었다.

[표 4-1]의 열 범주(Country of origin)에 대한 설명은 다음과 같다.

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

[표 4-2]은 [표 4-1]의 자료를 가지고 한 행을 제거한 후 카이제곱 통계량의 변화량을 보여준다. $D_c(i)$만 근사한 경우와 $D_c(i)$와 $c(i)$ 모두 근사한 경우 중 $D_c(i)$만 근사한 경우가 실제 변화량과 비슷한 것을 알 수 있다. 특히, 8행을 제거했을 때는 실제 변화량이 77.73이고 $D_c(i)$만 근사한 경우는 83.13으로 거의 비슷함을 알 수 있다. 또한 실제 변화량이 큰 것이 영향함수 값도 큰 것을

알 수 있다. 이 자료에서는 세 가지 방법 모두 8번째 행이 카이제곱 통계량에 가장 크게 작용함을 알 수 있다. 다시 말해서 8번째 행은 독립성 가설인 귀무가설의 기각에 가장 큰 영향을 미친다고 볼 수 있다. 이러한 결과는 자료 분석에서 범주의 통합(pooling)에 응용될 수 있다. 이 자료의 경우 범주의 개수를 줄이기 위해 8번째 행은 다른 행과 통합되어서는 안 될 것으로 보여 진다.

[표 4-3]는 [표 4-2]의 실제 변화량과 세 가지 방법으로 구한 변화량과의 상관계수 행렬을 나타낸다. 실제 변화량과 $D_c(i)$만 근사한 경우와의 상관계수가 0.9553으로 상당히 큰 양의 상관을 보여주고 있다. 영향함수 값과의 상관계수도 0.9111로 나타났다. $D_c(i)$와 $c(i)$ 모두 근사한 경우에는 0.6519로 위의 두 가지 보다 약한 상관을 보이고 있다.

[부록 1]은 비교적 범주가 많은 자료로써 Lebart, Morineau and M. Warwick(1984)으로부터 얻은 자료이다. 행(row)은 직업의 종류로 26가지 범주로 나뉘어져 있고 열(column)은 직업의 좋은 점으로 17가지의 범주로 되어 있다. [부록 2]는 [부록 1]의 분할표에서 행과 열에 대한 설명이다.

[표 4-2] 표 4-1의 한 행이 제거되었을 때 카이제곱 통계량의 변화량

		실제 변화량		$D_c(i)$만 근사		$D_c(i)$와 $c(i)$ 근사		영향함수
제거된 행 번호	X^2	$X^2(i)$	$X^2-X^2(i)$	$X_1^2(i)$	$X^2-X_1^2(i)$	$X_2^2(i)$	$X^2-X_2^2(i)$	IF
1	120.44	119.48	0.96	104.18	16.26	130.02	-9.58	-14.61
2		119.86	0.58	102.44	18.00	135.66	-15.22	-17.08
3		101.84	18.60	78.98	41.47	169.65	-49.21	-13.78
4		98.78	21.66	87.22	33.22	112.96	7.48	4.90
5		119.39	1.05	118.02	2.42	118.27	2.17	-0.43
6		102.79	17.65	84.07	36.37	139.69	-19.25	-9.14
7		115.42	5.02	110.42	10.02	113.78	6.66	-0.51
8		42.71	77.73	37.31	83.13	69.18	51.26	50.66

[표 4-4]에서 볼 수 있듯이 범주가 많은 [부록 1]자료에서는 $D_c(i)$만 근사한 경우보다 $D_c(i)$와 $c(i)$ 모두 근사한 경우가 실제 변화량과 비슷함을 알 수 있다. 그렇지만 그 차이는 근소하다.
영향함수 값을 보면 실제 변화량이 큰 행에서 영향함수 값도 큰 것을 볼 수 있다. 이는 영향함수 값으로도 카이제곱 통계량에 대한 행의 영향을 충분히 설명 할 수 있음을 의미한다.

1, 13, 18, 19행 등은 실제 변화량에서 큰 값을 갖는 행들이다. 즉, 독립성 기각에 영향을 크게 미친 행들이라고 볼 수 있다. 이 행들의 세 가지 경우 즉, $D_c(i)$만 근사한 경우, $D_c(i)$와 $c(i)$ 모두 근사한 경우, 영향함수 값을 살펴보면 모두 값이 큰 것을 알 수 있다. 따라서 이 세 가지 모두 실제 변화량을 대신해서 사용할 수 있는 충분한 가치가 있다.

[표 4-3] 표 4-1의 실제변화량과 세 가지 방법의 상관계수 행렬

	실제변화량	$D_c(i)$만 근사	$D_c(i)$와 $c(i)$근사	영향함수
실제 변화량	1.0	.	.	.
$D_c(i)$만 근사	0.9553	1.0	.	.
$D_c(i)$와 $c(i)$근사	0.6519	0.4290	1.0	.
영향함수	0.9111	0.7498	0.8810	1.0

[표 4-4] 부록 1의 한 행이 제거되었을 때 카이제곱 통계량의 변화량

		실제 변화량		$D_c(i)$만 근사		$D_c(i)$와 $c(i)$ 근사		영향함수
제거된 행 번호	X^2	$X^2(i)$	$X^2-X^2(i)$	$X_1^2(i)$	$X^2-X_1^2(i)$	$X_2^2(i)$	$X^2-X_2^2(i)$	IF
1	1777.19	1302.96	474.23	1223.36	553.83	1281.12	496.07	245.79
2		1747.30	29.89	1695.66	81.53	1699.70	77.48	-26.44
3		1749.47	27.72	1718.12	59.07	1719.34	57.85	0.61
4		1712.61	64.57	1670.69	106.50	1674.86	102.33	12.94
5		1744.20	32.98	1717.50	59.68	1719.00	58.19	1.19
6		1752.50	24.69	1740.70	36.48	1741.09	36.10	10.24
7		1716.25	60.94	1664.29	112.90	1669.38	107.81	3.29
8		1719.18	58.01	1670.51	106.68	1676.04	101.15	-5.96
9		1750.24	26.95	1726.88	50.31	1727.99	49.19	-0.42
10		1672.49	104.70	1607.68	169.51	1619.33	157.85	14.17
11		1716.77	60.41	1611.84	165.35	1629.47	147.72	-59.30
12		1724.53	52.66	1664.43	112.76	1670.16	107.02	-10.49
13		1694.19	83.00	1573.93	203.26	1596.34	180.85	-47.84
14		1754.37	22.82	1690.22	86.97	1696.09	81.09	-48.40
15		1591.28	185.91	1454.91	322.28	1490.08	287.10	9.38
16		1742.86	34.32	1707.83	69.35	1709.76	67.43	-1.13
17		1733.41	43.77	1703.12	74.06	1704.90	72.28	11.23
18		1709.18	68.01	1620.09	157.09	1632.86	144.33	-31.38
19		1542.95	234.24	1419.90	357.29	1458.59	318.60	41.99
20		1745.10	32.09	1686.17	91.02	1690.57	86.62	-24.84
21		1734.56	42.62	1671.18	106.01	1676.55	100.63	-19.76
22		1731.26	45.92	1701.89	75.30	1703.73	73.46	12.38
23		1728.78	48.41	1632.61	144.58	1645.65	131.54	-56.23
24		1732.04	45.14	1688.60	88.58	1692.11	85.07	-4.56
25		1757.28	19.91	1700.57	76.62	1704.83	72.36	-40.03
26		1725.81	51.38	1689.83	87.36	1692.11	85.08	13.58

[표 4-5] 부록 1의 실제변화량과 세 가지 방법의 상관계수 행렬

	실제변화량	$D_c(i)$만 근사	$D_c(i)$와 $c(i)$근사	영향함수
실제 변화량	1.0	.	.	.
$D_c(i)$만 근사	0.9664	1.0	.	.
$D_c(i)$와 $c(i)$근사	0.9694	0.9998	1.0	.
영향함수	0.8624	0.7050	0.7141	1.0

[표 4-5]은 상관계수 행렬을 보여준다. [부록 1]의 자료에서는 세 가지 방법 모두 실제 변화량과의 상관계수가 0.85이상으로 나타났다. 특히, $D_c(i)$만 근사한 경우와 $D_c(i)$와 $c(i)$ 근사한 경우 모두 0.95이상으로 상당히 큰 양의 상관을 보여주고 있다.
범주의 개수가 많은 [부록 1]자료에서는 세 가지 방법의 어떤 값을 사용해도 한 행이 제거되었을 때 카이제곱 통계량에 대한 변화를 충분히 설명할 수 있다.

5. 결론

본 논문에서는 분할표에서 한 행이 제거되었을 때 카이제곱 통계량의 변화량을 알아보기 위해, 소거법을 이용하여 한 행이 제거되었을 때 카이제곱 통계량의 값을 직접 구한 다음 식을 간단히 하기 위해서 첫 번째는 $D_c(i)$만 근사한 경우와 두 번째는 $D_c(i)$와 $c(i)$을 근사한 경우로 나누어 변화량을 구해 보았다. 두 경우 모두 근사적으로 제거되지 않았을 때의 카이제곱 통계량 X^2으로 표현할 수 있었다. 특히, $D_c(i)$와 $c(i)$을 근사한 경우에는 i행이 제거되었을 때 카이제곱 통계량 $X^2(i)$이 제거되지 않았을 때 카이제곱 통계량 X^2과 i행의 칸 도수 합 n_{i+}의 함수에 대한 합으로 표현되었다. 즉, $X_2^2(i) \approx \frac{n(i)}{n} X^2 + f(n_{i+})$ 으로 나타낼 수 있었다. 또한 세 번째 방법으로, 영향 함수를 이용하여 변화량을 설명하였다. 행에 대한 영향은 그 행의 칸 영향들을 합한 것으로 설명되어 졌다.

실제 자료를 가지고 이들 세 가지 방법으로 카이제곱 통계량의 변화량을 구해 본 결과 상대적으로 범주가 적은 자료와 많은 자료 모두, 실제 변화량과 상당히 비슷함을 보였다. 특히 범주가 적은 자료에서는 $D_c(i)$만 근사한 경우가 $D_c(i)$와 $c(i)$을 근사한 경우보다 상당히 근사가 잘 되었다. 그러나 범주가 많은 자료에서는 두 경우 모두 근사가 잘 되었다. 따라서 범주가 많은 자료에서는 식의 표현이 좀 더 간단한 $D_c(i)$와 $c(i)$을 근사한 경우를 사용하여 카이제곱 통계량의 변화량을 구하는 것이 더 좋을 것으로 보인다. 또한 영향함수 값도 한 행이 제거되었을 때 카이제곱 통계량의 변화를 설명할 수 있는 좋은 측도로 사용됨을 실제 자료에서 볼 수 있었다.

본 논문의 결과를 적용하면 이차원 분할표에서 행 범주의 개수를 줄일 수 있는 근거를 제시할 수 있다. 즉, 앞의 세 가지 방법에 의해, 한 행을 제거했을 때 카이제곱 통계량의 변화를 조사하여

그 차이가 작은 행들은 서로 통합하여 범주의 수를 줄일 수 있다. 반대로 행을 제거했을 때 변화량의 값이 큰 행은 다른 행과 합쳐서는 안 될 것이다. 또한 설문조사시 표본 추출에 쓰이는 비용의 절감효과도 기대할 수 있다. 예를 들어, 한 대학에서 학년(1학년, 2학년, 3학년, 4학년)과 지방선거에 대한 참여여부가 독립인지를 조사하기 위해 설문조사를 실시한다고 하자. 예비조사에서 얻어진 자료에 본 논문의 결과를 이용하면 귀무가설인 독립성의 기각에 영향을 적게 미치는 학년을 밝힐 수 있으며 이 학년은 본조사에서 다른 학년에 비해 상대적으로 적은 표본의 수로 동일한 결과를 얻을 수 있다. 따라서 본 논문의 결과는 표본추출을 시행하는 연구에서 예비조사의 시행 결과로 독립성이 기각되면 본조사 또는 차후조사에서 표본의 수를 줄일 수 있으므로 비용의 절감효과도 기대된다.

참 고 문 헌

[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. (1982). Residuals and Influence in Regression, Chapman and Hall, New York.

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

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

[5] Guttman, L. (1971). Measurement as structural theory, *Psychometrika,* Vol. 36, 329-347.

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

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

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

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

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

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

[12] Kim, H. (1998). A study on Cell Influences to Chi-square statistic in Contingency Tables, *The Korean Communications in Statistics*, Vol. 5, 35-42.

[13] Kim, H. and Lee, H. (1996). Influence Functions on χ^2 statistic in Contingency Tables, *The Korean Communications in Statistics*, Vol. 3, 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.

[15] Lebart L., Morineau A. and Warwick. K. M. (1984). *Multivariate Descriptive Statistical Analysis - Correspondence analysis and related techniques for large matrices*, John Wiley & Sons, New York.

[16] Radhakrishnan, R. & Kshirsagar, A. M. (1981). Influence functions for certain parameters in multivariate analysis, *Communications in Statistics* A 10, 515-529.

[2003년 1월 접수, 2003년 6월 채택]

[부록 1] 3278명에 대한 직업의 종류와 직업의 좋은 점

열 / 행	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	4	189	0	3	2	2	9	3	12	2	1	4	11	15	12	8	1
2	1	13	3	10	17	12	4	1	8	3	5	1	9	5	11	0	0
3	1	9	1	0	4	13	0	2	2	0	2	1	4	3	6	1	0
4	5	5	2	9	18	5	3	2	6	5	5	0	2	3	22	0	0
5	2	7	1	4	15	5	2	1	6	1	2	2	3	0	5	0	1
6	2	5	0	4	1	0	3	0	2	1	1	1	1	0	3	0	2
7	2	3	1	8	16	17	1	8	7	2	4	3	6	1	24	0	1
8	3	18	0	6	16	5	4	4	13	4	2	3	6	2	26	0	2
9	3	7	3	6	6	0	0	2	6	3	3	0	2	1	8	0	0
10	0	18	1	12	31	7	0	8	19	11	3	2	10	4	26	0	6
11	7	63	2	9	31	9	4	6	9	10	3	4	14	8	35	2	2
12	2	43	16	7	6	4	7	1	8	2	0	1	6	1	7	0	3
13	8	95	23	15	15	2	13	7	9	5	2	3	13	4	18	1	3
14	5	32	9	9	17	4	5	4	7	4	3	0	8	3	18	0	3
15	8	26	10	24	24	80	10	17	11	3	8	2	6	9	16	3	4
16	1	7	2	11	3	14	2	6	3	1	1	2	1	3	5	0	2
17	4	10	10	8	2	1	6	4	2	3	1	0	3	2	1	0	1
18	3	31	16	15	11	19	5	19	10	2	3	7	24	1	5	0	5
19	2	33	27	31	9	18	27	24	3	4	43	8	18	3	11	1	3
20	2	19	2	12	12	21	0	1	4	5	5	1	3	3	13	0	1
21	8	12	4	8	13	21	2	10	4	2	5	6	3	1	10	0	3
22	0	8	0	4	5	2	7	1	5	7	2	1	2	2	11	1	1
23	8	35	14	13	16	10	6	25	6	4	10	9	11	4	14	0	1
24	2	13	2	14	5	8	0	10	0	8	3	2	5	4	11	0	2
25	3	26	9	3	12	5	8	8	4	4	2	3	10	3	8	0	2
26	0	1	1	3	4	4	3	4	1	1	2	1	5	1	3	0	3

[부록 2] 부록 1의 행과 열에 대한 설명

행번호	설명(직업의 종류)	열번호	설명(직업의 좋은 점)
1	Farming-fishing	1	Variety of Work
2	Farm-food industry	2	Freedom
3	Energy-mines	3	Human Contact
4	Steel	4	Schedules
5	Chemical-glass-oil	5	Salaries
6	Wood-paper	6	Job Security
7	Auto-aviation-shipping	7	Family Life
8	Textile-leather-shoes	8	Interesting
9	Pharmaceutical-industries	9	Near Home
10	Manufacturing	10	Good Atmosphere
11	Construction	11	Social Advantages
12	Food-grocery	12	I Am my Own Boss
13	Small business	13	I Like It
14	Miscellaneous business	14	Other
15	Administrative services	15	None
16	Telecommunications	16	Work Outdoors
17	Social services	17	No answer
18	Health services		
19	Teaching-research		
20	Transportation		
21	Insurance-banking		
22	Domestic worker		
23	Other services		
24	Printing-publishing		
25	Private services		
26	No answer		