The Korean Communications
in Statistics Vol. 7, No. 3, 2000
pp. 687-698

Identification of Multiple Outlying Cells in Multi-way Tables

Jong Cheol Lee[1], Chong Sun Hong[2]

Abstract

An identification method is proposed in order to detect more than one outlying cells in multi-way contingency tables. The iterative proportional fitting method is applied to get expected values of several suspected outlying cells. Since the proposed method uses minimal sufficient statistics under quasi log-linear models, expected counts of outlying cells could be estimated under any hierarchical log-linear models. This method is an extension of the backwards-stepping method of Simonoff (1988) and requires less iteration to identify outlying cells.

Keywords : backwards-stepping; Bonfferoni bound; deleted residual; influence; masking effect; minimal sufficient statistics; outlying cell; quasi-independent; swamping effect.

1. 서 론

분할표에 특정한 모형을 적합시키는 경우 적합결여의 한 원인으로 이상칸(outlying cell)을 고려할 수 있다. 이상칸을 식별하기 위한 다양한 기준들이 제시되었는데 한 종류는 다양한 형태의 잔차들이고 다른 종류는 적합도 검증통계량의 차이의 함수로 표현된다. Barnet과 Lewis(1994)는 이러한 다양한 기준들에 대하여 설명하고 이중 몇 가지 기준을 이용하여 모형에 불일치한 칸을 이상칸으로 식별하였다. Fienberg(1969), Kotz와 Hawkins(1984)는 로그 오즈(log odds)로 표현되는 테트라드(tetrads)를 기준으로, 그리고 Mostellar와 Parunak(1985)는 탐색적 접근방법으로 이상칸을 식별할 수 있다고 하였다.

Haberman(1973)은 식 (1.1)과 같은 수정된 잔차(adjusted residual)을 제안하였다. 이는 관찰값 x_{ij}와 칸 기대값의 추정량 $\hat{m}_{ij}$과의 차이를 측정하였다. 그는 정규확률그림에 수정된 잔차를 표시하여 극단적인 칸을 이상칸으로 식별하였다.

$$\tilde{r}_{ij} = \frac{x_{ij} - \hat{m}_{ij}}{\sqrt{var(\hat{m}_{ij})}} = \frac{x_{ij} - (x_{i+}\, x_{+j})/N}{\sqrt{x_{i+}\, x_{+j}(N-x_{i+})(N-x_{+j})/N^3}}, \tag{1.1}$$

1) Lecturer, Department of Statistics, Sungkyunkwan University, Seoul, Korea (110-745).
2) Professor, Department of Statistics, Sungkyunkwan University, Seoul, Korea (110-745)
E-mail : cshong@skku.ac.kr.

여기서 $N=\sum_i\sum_j x_{ij}$ $x_{i+}=\sum_j x_{ij}$, $x_{+j}=\sum_i x_{ij}$는 각각 총합, i번째 행 주변합, 그리고 j번째 열 주변합을 나타낸다.

Brown(1974)은 $I\times J$인 분할표에서 하나의 칸 (i,j)를 제거한 후 (i,j)칸의 기대값을 식 (1.2)와 같이 추정하였다.

$$m_{ij}^{*}=\frac{(x_{i+}-x_{ij})(x_{+j}-x_{ij})}{N-x_{i+}-x_{+j}+x_{ij}}. \tag{1.2}$$

추정량 m_{ij}^{*}를 (i,j)칸의 관찰값으로 간주하고 준독립성(quasi-independence) 모형하에서의 기대값으로 고려한다(Goodman, 1968). 적합도 검증통계량에 영향을 주는 칸은 식 (1.3)과 같은 삭제된 잔차(deleted residual)를 이용하여 이상칸으로 식별되었다.

$$r_{ij}^{*}=\frac{x_{ij}-m_{ij}^{*}}{\sqrt{m_{ij}^{*}}}. \tag{1.3}$$

다차원 분할표에서 직접계산이 가능한 최우추정량이 존재하는 분해 모형(decomposable model)인 경우, Upton과 Guillen(1995)은 식 (1.3)의 삭제된 잔차를 확장하여 하나의 이상칸을 식별하기 위한 일반적인 공식을 제안하였다. 그러나 설정된 모형의 최소충분통계량(minimal sufficient statistics)의 수가 증가하거나 둘 이상의 이상칸이 존재하는 경우에는 제안된 공식을 이용하는 것이 어렵다.

이들 기준에 의해 하나 이상의 이상칸을 식별하기 위한 몇 가지 식별방법들이 제안되었는데 그 중 하나는 Fuchs와 Kenett(1980)이 제안한 전진단계방법(forwards-stepping method)으로 이상칸으로 의심나는 가장 극단적인 칸으로부터 덜 극단적인 칸의 순서로 이상칸을 식별하는 방법이다. 다른 하나는 Simonoff(1988)가 제안한 후진단계방법(backwards-stepping method)으로 덜 극단적인 칸으로부터 가장 극단적인 칸의 순서로 이상칸을 식별하는 방법이다. Simonoff는 하나이상의 이상칸이 존재하는 경우 전진단계방법을 이용하여 이상칸을 식별하면 가장효과(masking effect ; 이상칸을 이상칸으로 식별하지 못하는 효과)를 유발할 수 있고 삭제된 잔차에 의한 후진단계방법을 이용하면 가장효과와 편승효과(swamping effect ; 이상칸이 아닌 칸을 이상칸으로 식별하는 효과)가 제한된다고 하였다. 그럼에도 불구하고 이들 방법은 범주형 변수의 범주 수가 증가할수록 이상칸을 식별하는데 많은 시간이 소요될 뿐만 아니라 다차원 분할표에 대해서는 이상칸을 식별할 수 없다.

본 논문에서는 다차원 분할표에서 다중 이상칸을 식별하기 위한 방법을 제안하고자 한다. 우선 삭제된 잔차를 이용하여 초기에 이상칸으로 의심나는 칸들을 제거한 후, 이 칸들의 기대값을 추정하기 위해 반복비율적합(Iterative Proportional Fitting : IPF)에 의하여 추정한다. 이 추정방법은 반복의 각 단계에서 로그선형모형(log-linear model)하에서 최소충분통계량을 이용하므로 모든 계층적 로그선형모형(hierarchical log-linear model)에 대해서도 이상칸으로 의심되어 제거된 칸들의 기대값을 추정할 수 있다. 이 방법으로는 다차원 분할표에서 하나 이상 제거된 칸들의 기대값을 추정할 수 있기 때문에 삭제된 잔차를 통해서 다차원 분할표에서도 다중 이상칸의 식별이 가능하

다. 제안된 추정방법과 이상칸을 식별하는 방법은 후진단계방법을 응용하였기 때문에 가장효과와 편승효과에 대해 제한적이다. 그러나 제안된 방법은 후진단계방법에 비하여 훨씬 적은 계산에 의해 다중 이상칸을 식별할 수 있음을 모의실험을 통하여 연구하고자 한다.

2. 이상칸으로 의심나는 칸값의 추정방법

2차원 분할표에서 식 (1.3)과 같은 삭제된 잔차를 구하기 위해, Brown(1974)은 하나의 칸을 결측칸으로 간주하고 해당 칸의 기대값을 추정하는 방법을 제안하였다. 다차원 분할표인 경우에는 Upton과 Guillen(1995)이 직접해 모형인 경우 하나의 칸이 제거된 분할표에 대해 완전한 칸값(perfect cell value)이라는 추정량을 이용하여 이상칸을 식별하였다. 이 절에서는 Brown의 추정량을 확장하고 IPF에 의해 다차원 분할표의 로그선형모형에 대해 하나 이상의 칸이 제거된 경우 해당 칸의 기대값을 추정하는 방법을 제안한다. 이 추정방법은 최현집과 신상준(2000)이 제안한 다차원 분할표에서의 결측값 추정방법을 응용한 것이다.

일반적인 상황에서 하나 이상의 칸이 제거된 분할표에서의 칸 기대값의 추정방법을 설명하기 위해, $I\times J\times K$인 3차원 분할표인 경우에 대해 살펴보기로 한다. 3차원 분할표의 칸값을 $\{x_{ijk} \mid i=1,\ldots,I;\ j=1,\ldots,J;\ k=1,\ldots,K\}$라 하고 T를 전체 칸의 집합, S를 이상칸으로 의심되는 칸으로 이루어진 T의 부분집합이라고 하자. 변수 A, B, C를 갖는 3차원 분할표에 대해 직접해가 존재하지 않는 부분연관모형 $[AB][AC][BC]$의 최소충분통계량이 $\{x_{ij+}, x_{i+j}, x_{+jk}\}$이므로 이들 통계량을 이용하여 다음과 같은 과정으로 S에 속하는 칸 기대값을 추정한다.

과정 1. S에 속하는 칸에 대해 초기값을 설정한다. 예를 들어 $q=1$번째 반복에서 $\hat{m}^{*q}_{ijk}=1$이라 하자.

과정 2. 준로그선형모형의 최소충분통계량$\{x^*_{ij+}, x^*_{i+j}, x^*_{+jk}\}$을 이용하여 다음의 칸 기대값이 수렴할 때까지 반복한다. 모든 $r=1,2,\ldots$ 에 대하여

$$\begin{aligned} a)\ \hat{m}^{*q(3r)}_{ijk} &= \hat{m}^{*q(3r-1)}_{ijk}\ x^{*q}_{+jk} \,/\, \hat{m}^{*q(3r-1)}_{+jk} \\ b)\ \hat{m}^{*q(3r-1)}_{ijk} &= \hat{m}^{*q(3r-2)}_{ijk}\ x^{*q}_{i+k} \,/\, \hat{m}^{*q(3r-2)}_{i+k} \\ c)\ \hat{m}^{*q(3r-2)}_{ijk} &= \hat{m}^{*q(3r-3)}_{ijk}\ x^{*q}_{ij+} \,/\, \hat{m}^{*q(3r-3)}_{ij+}\ , \end{aligned} \tag{2.1}$$

여기서 $x^{*q}_{ij+}=(x_{ij+}-x_{ijk}+\hat{m}^{*q(3r-2)}_{ijk})$, $x^{*q}_{i+k}=(x_{i+k}-x_{ijk}+\hat{m}^{*q(3r-1)}_{ijk})$,
$x^{*q}_{+jk}=(x_{+jk}-x_{ijk}+\hat{m}^{*q(3r)}_{ijk})$ 이며 $\hat{m}^{*q(0)}_{ijk}$는 초기값인 $\hat{m}^{*q}_{ijk}=1$을 나타낸다.

과정 3. <과정 2>을 통해 얻어진 q번째 단계에서의 잠정적인 추정값을 $(q+1)$단계에서의 (i,j,k)칸의 초기값으로 고려한다. q번째 추정값과 $(q+1)$번째 추정량의 차이가 매우 작으면 $\hat{m}^{*q}_{ijk}$를 칸 기대값의 추정량으로 설정하고 그렇지 않으면 <과정 2>를 반복 시행한다.

초기값으로 임의의 양수를 설정하면 위 반복적인 방법은 초기값에 의해 큰 영향을 받지 않는다.

또한 양의 초기값을 설정함으로써 주변합은 항상 양수가 되고 $\hat{m}_{ijk}^{*q}$의 가능도(likelihood)가 단조감소함수이므로 위 과정에 의해 얻어진 칸 기대값의 추정량 $\hat{m}_{ijk}^{*q}$는 수렴한다. 이 수렴성을 살펴보기 위하여 임의의 $r=1,2,\ldots$ 에서 $\hat{m}_{ijk}^{*q}$의 로그가능도비함수(log-likelihood ratio function)는 식 (2.2)와 같다.

$$D^{(3r)}=\sum_i\sum_j\sum_k\left(x_{ijk}\ln x_{ijk}-x_{ijk}\ln \hat{m}_{ijk}^{*q(3r)}\right). \tag{2.2}$$

위 식에서 $\hat{m}_{ijk}^{*q(3r)}$에 식 (2.1)을 대입하면

$$\begin{aligned} D^{(3r)} &= \sum_i\sum_j\sum_k\left(x_{ijk}\ln x_{ijk}-x_{ijk}\ln\left(\hat{m}_{ijk}^{*q(3r-1)}\ x_{+jk}^{*q}\ /\ \hat{m}_{+jk}^{*q(3r-1)}\right)\right)\\ &= \sum_i\sum_j\sum_k\left(x_{ijk}\ln x_{ijk}-x_{ijk}\ln \hat{m}_{ijk}^{*q(3r-1)}-x_{ijk}\ln x_{+jk}^{*q}+x_{ijk}\ln \hat{m}_{+jk}^{*q(3r-1)}\right)\\ &= D^{(3r-1)}-\sum_i\sum_j\sum_k\left(x_{ijk}\ln x_{+jk}^{*q}-x_{ijk}\ln \hat{m}_{+jk}^{*q(3r-1)}\right)\ , \end{aligned}$$

여기서 $\hat{m}_{+jk}^{*q(3r-1)}=x_{+jk}^{*q}$일 때 $\sum_j\sum_k\left(x_{+jk}\ln \hat{m}_{+jk}^{*q(3r-1)}\right)$이 최대이므로 다음을 만족한다.

$$\sum_i\sum_j\sum_k\left(x_{ijk}\ln x_{+jk}^{*q}-x_{ijk}\ln \hat{m}_{+jk}^{*q(3r-1)}\right)\geq 0.$$

따라서 모든 $r=1,2,\ldots$에 대하여

$$D^{(3r)}\leq D^{(3r-1)}$$

이 성립하므로 $D^{(3r)}$은 단조감소함수이고 추정량 m_{ijk}^{*q}는 항상 수렴한다는 사실을 얻는다.

2차원 분할표인 경우 준독립성 모형을 적합시켰을 때 얻어질 추정량은 Brown(1974)이 제시한 추정량 m_{ij}^{*}과 동일하다. 그가 제안한 추정량은 하나의 칸만이 제거된 경우에 추정이 가능하지만, 본 논문에서의 추정방법은 하나 이상의 칸이 제거된 경우 즉, 집합 S에 대한 각 칸들의 기대값을 동시에 추정할 수 있다. 다차원 분할표에 대해 하나의 칸이 제거된 경우 칸값을 추정하기 위해, Upton과 Guillen(1995)은 완전한 칸값이라는 추정량을 제시하였고 직접해 모형인 경우 추정량을 구할 수 있다. 그러나 여기에서 제안된 추정방법은 다차원 분할표인 경우 임의의 로그선형모형에 대해서도 칸 추정량을 구할 수 있을 뿐만 아니라 하나 이상의 칸이 제거된 경우에도 칸 기대값의 추정이 가능함이 특징이다.

3. 다중 이상칸을 식별하기 위한 방법

$I\times J$인 2차원 분할표에 대해 덜 극단적인 칸으로부터 가장 극단적인 칸의 순서로 이상칸을 식별하는 후진단계방법을 Simonoff(1988)가 제안하였다. 그는 Haberman(1973)이 제안한 수정된 잔

차가 편승효과가 유발되므로 이를 피하기 위하여 삭제된 잔차를 이용하여 이상칸을 식별하였다. 최소충분통계량의 수가 증가할수록 후진단계방법은 많은 계산을 해야 하므로 이상칸을 식별하는데 시간이 많이 소요되고 이차원 분할표에 대해서만 적용이 가능하다. 본 논문에서는 후진단계방법을 응용하여 한 번에 두 개 이상의 다중 이상칸의 식별이 가능할뿐만 아니라 후진단계방법에 비하여 적은 계산으로 이상칸을 식별하고 임의의 다차원 로그선형모형에 대해서도 적용할 수 있는 식별방법을 제안하고자 한다.

본 논문에서는 삭제된 잔차를 이용하여 극단적인 값을 갖는 칸들을 이상칸의 대상인 집합 S_q와 다음 단계에서 고려 대상인 집합 S_{q+1}을 설정한 후 이 두 집합들을 대상으로 이상칸 여부를 식별한다. 설정한 칸들이 이상칸인지를 검증하기 위해 다음과 같은 가설을 세운다.

$$\begin{aligned} H_0 &: S_{q+1}\text{에 포함된 칸들이 이상칸이다.} \\ H_1 &: S_q\text{에 포함된 칸들이 이상칸이다.} \end{aligned} \tag{3.1}$$

가설 (3.1)을 검증하기 위해 일반화 가능도비 검증통계량(generalized likelihood ratio test statistic), G^2, 의 차이를 나타내는 다음과 같은 검증통계량 ΔG_q^2 를 이용한다.

$$\Delta G_q^2 = G^2_{S_{q+1}} - G^2_{S_q}. \tag{3.2}$$

이 통계량은 Cook의 D-통계량과 동일한 형태를 이루고 있으며 해당 칸들의 영향력(influence)을 의미한다(Christensen, 1990). 따라서 가설 (3.1)도 다음과 같이 영향관측칸(influence cell)을 결정하는 가설로 변환될 수 있다.

H_0 : S_{q+1}에는 포함되어있지 않으며 S_q에는 포함되어있는 칸이 영향관측칸이 아니다.
H_1 : S_{q+1}에는 포함되어있지 않으며 S_q에는 포함되어있는 칸이 영향관측칸이다.

ΔG_q^2은 두 집합에 대한 원소의 개수 차이를 자유도로 갖는 카이제곱분포를 따르고 이 값이 작으면 두 집합간에 적합도 검증통계량의 변화가 작으므로 S_{q+1}에 포함되는 칸들을 모형에 대해 이상칸으로 고려할 수 있다. 반대로 큰 값을 나타내면 S_q에 포함되는 칸들을 이상칸으로 고려할 수 있음을 의미한다. 이는 S_q에는 포함되나 S_{q+1}에 포함되지 않은 원소(칸)가 이상칸이 아니다는 결론을 유도할 수 있다. 따라서 이는 만일 하나의 칸을 대상으로 이상칸 여부를 식별하는 경우 또 다른 이상칸에 의해 가장효과를 유발될 수 있으므로 제안된 방법에서는 가장효과를 감소시키기 위해 초기에 이상칸의 대상을 가능한 많이 선정하여 후진단계방법과 유사한 방법으로 덜 극단적인 칸으로부터 이상칸 여부를 식별한다. 그러므로 초기에 설정한 이상칸의 집합 S_1에 포함된 칸들의 수는 S_2에 포함된 칸들의 수보다 많으며 이는 S_3보다 많음을 알 수 있다. 즉,

$$S_1\text{의 원소의 수} > S_2\text{의 원소의 수} > S_3\text{의 원소의 수} > \cdots.$$

다중 이상칸 식별방법(MOCI 방법)을 다음과 같이 제안한다.

▫ *MOCI* 방법 *(Multiple Outlying Cells Identification method)*

과정 1. 주어진 분할표에 대해 적절한 모형을 적합시킨다. 만일 적합결여가 발생한다면 다음 단계를 수행한다. 이때 $q=1$로 설정한다.

과정 2. q번째 단계에서 모든 칸에 대해 삭제된 잔차 r^*를 계산하여 큰 잔차를 갖는 칸들을 초기에 의심나는 이상칸 집합 S_q로 설정한다. 여기서 삭제된 잔차를 본페로니 바운드(Bonferroni bound ; $\Phi^{-1}(1-\alpha/k)$, 여기서 α는 유의수준, k는 분할표에서 비교하고자 하는 총 칸의 수)와 비교하여 집합 S_q에 포함될 칸을 선택한다.

과정 3. S_q에 속하는 칸들에 대하여 2절에서 제안한 방법으로 칸값을 추정한다. 추정된 값을 m^*이라 한다. 이 추정량을 칸값으로 고려하여 S_q에 속하는 칸에 대해서만 삭제된 잔차를 재계산한 후, 본페로니 바운드와 비교하여 유의하지 않은 칸을 이상칸이 아닌 것으로 간주하고 이 칸들을 집합 S_q 로부터 제외하여 집합 S_{q+1}을 설정한다.

과정 4. 식 (3.2)에서 제시한 검증통계량 ΔG_q^2을 이용하여 집합 S_{q+1}이 이상칸 집합인지를 검증한다. 만일 검증통계량 ΔG_q^2의 값이 작다면 과정을 중단하고 집합 S_{q+1}에 속하는 칸들을 이상칸으로 식별하고, ΔG_q^2의 값이 크다면 S_{q+1}에 포함되는 칸의 삭제된 잔차를 다시 구하기 위해 <과정 3>을 다시 수행하여 한다.

Simonoff(1988)가 작성한 <표 3-1>과 같이 (1,2), (1,3) 그리고 (2,1)칸을 이상칸으로 설정한 5×5 분할표에 대해 *MOCI* 방법으로 이상칸을 식별하여 후진단계방법과 비교하였다.

<표 3-1> 가상자료

	열(B)				
행(A)	1	2	3	4	5
1	18	41	41	20	21
2	39	20	20	22	22
3	24	20	20	16	18
4	20	20	19	19	19
5	23	19	20	17	20

준독립성 모형하에서 각 칸을 제거한 후 칸의 기대값을 이용하여 삭제된 잔차와 본페로니 바운

드를 비교하여 초기에 이상칸으로 의심나는 칸들을 집합 S_1으로 선별하기 위하여 각 칸이 제거되었을 때의 칸 추정량과 삭제된 잔차가 <표 3-2>에 나타내었다. 여기서 칸값은 Brown(1974)의 칸 추정량 m_{ij}^*을 이용한 결과와 일치하고 ( )는 삭제된 잔차를 나타낸다.

<표 3-2>을 살펴보면 (1,1), (1,2), (1,3) 그리고 (2,1) 칸의 기대값이 관찰값과 유의한 차이를 나타내는 것을 볼 수 있다(유의수준 0.05에서 본페로니 바운드의 값 $\Phi^{-1}(1-0.05/25)=2.87$보다 삭제된 잔차의 절대값이 크다). 그러므로 이들 4개의 칸을 초기에 고려할 이상칸으로 의심나는 칸들로 고려할 수 있다. 즉,

$$S_1=\{(1,1),(1,2),(1,3),(2,1)\}.$$

<표 3-2> 준독립성하에서의 기대값과 삭제된 잔차

행	열				
	1	2	3	4	5
1	41.9228 (-3.6948)	23.3728 (3.6461)	23.3728 (3.6461)	26.1050 (-1.1949)	28.0473 (-1.3307)
2	20.4 (4.1181)	30.7463 (-1.9380)	30.7463 (-1.9380)	20.0331 (0.4395)	22.0672 (-0.0143)
3	20.5556 (0.7597)	21.6667 (-0.3581)	21.6667 (-0.3581)	16.7435 (-0.1817)	17.3545 (0.1549)
4	22.4314 (-0.5134)	21.3296 (-0.2879)	21.8833 (-0.6164)	15.1554 (0.98756)	16.6263 (0.5821)
5	21.4413 (0.3366)	22.5698 (-0.7514)	22.0056 (-0.4275)	16.5288 (0.1159)	16.6755 (0.8141)

다음 과정에서 S_1에 속한 4개의 칸을 결측칸으로 간주하여 <과정 2>과 <과정 3>에서 언급한 바와 같이 2절에서 제안한 반복적인 추정방법으로 S_1에 속한 칸들에 대해서만 칸값을 추정하여 삭제된 잔차를 계산하여 유의한 값을 갖는지를 검증한다. 이를 정리한 결과는 <표 3-3>과 같다.

<표 3-3> *MOCI*-방법을 이용한 이상칸 식별과정

칸	칸값	<단계 1> 본페로니 바운드 = 2.24		<단계 2> 본페로니 바운드 = 2.13	
		칸 추정량	삭제된 잔차	칸 추정량	삭제된 잔차
(1,1)	18	24.5981	-1.3304		
(1,2)	41	21.1699	4.3099	19.1567	4.9907
(1,3)	41	21.1699	4.3099	19.1567	4.9907
(2,1)	39	24.7930	2.8532	23.3095	3.2499
		$G^2_{S_1}$=1.34		$G^2_{S_2}$=2.59	
ΔG^2_q		$\Delta G^2_1 = G^2_{S_2} - G^2_{S_1} = 2.59 - 1.34 = 1.25$			

<표 3-3>의 <단계 1>에 나타난 칸 추정량을 살펴보면 (1,1) 칸값이 <표 3-2>의 추정량에 비해 향상된 것을 볼 수 있다. 특히 본페로니 바운드의 값인 $\Phi^{-1}(1-0.05/4)=2.24$와 삭제된 잔차를 비교하면 (1,1)칸의 삭제된 잔차의 절대값이 작은 것을 볼 수 있다. 그러므로 (1,1)칸이 이상칸이 아니라는 것을 보여준다. 이러한 결과에 의해 다음에 고려할 이상칸 집합은 다음과 같다.

$$S_2=\{(1,2),(1,3),(2,1)\}.$$

S_2에 속하는 칸에 대한 제거된 칸의 기대값과 삭제된 잔차가 <표 3-3>의 <단계 2>에 나타나있다. 이들 칸에 대한 삭제된 잔차가 <단계 1>과 비교할 때 변화가 적음을 알 수 있다. S_2에 포함된 칸들의 삭제된 잔차들은 모두 본페로니 바운드 $\Phi^{-1}(1-0.05/3)=2.13$보다 크다. 또한 식 (3.2)에서 제안한 검증통계량의 값이 $\Delta G_1^2=G_{S_2}^2-G_{S_1}^2=1.25$(p-값=0.263)으로 작으므로 <표 3-1>의 자료에 대해 S_2에 속하는 (1,2), (1,3), (2,1) 칸들이 이상칸이라 결론을 내릴 수 있다.

이상의 과정에서 볼 수 있듯이 하나 이상의 칸을 이상칸의 대상으로 고려하여 검증함으로써 가장효과를 덜 유발하게 되고 또한 모형에 대한 칸들의 영향력을 측정하는 Cook의 D-통계량이 ΔG_q^2로 표현되므로 검증통계량을 통해 영향력도 살펴볼 수 있다. 더구나 MOCI 방법으로는 단 두 단계만을 통해 이상칸을 식별하므로 전체 칸 수의 20-30%를 초기에 이상칸으로 고려하는 후진단계방법보다 적은 시간이 소요된다.

4. 모 의 실 험

이 절에서는 5×5 분할표에 로그선형모형을 적합시켜 얻어진 분할표를 이용하여 제안된 방법의 특성을 살펴보고 후진단계방법과 비교하고자 한다. 모의실험을 통해 다음의 3가지 속성에 대해 이상칸을 식별하는 과정을 조사하여 후진단계방법과 MOCI 방법을 비교하였다.

(1) 검증의 수준과 유의수준 $\alpha=0.05$의 비교
(2) 이상칸을 식별하는 정도
(3) 식별되지 않는 이상칸의 정도

Simonoff가 모의실험을 한 것과 마찬가지로 5×5인 분할표를 International Mathematical and Statistical Libraries(IMSL)의 GGMTN을 응용하여 생성하였다: 총 1,600개의 분할표를 통해 MOCI 방법과 후진단계방법을 비교하였다.

다양한 이상칸의 위치 형태에 대해 분할표를 생성하여, 각 형태에 따라 검증력(β_1), 설정한 이상칸만을 식별할 확률(β_2), 설정한 이상칸중 최소한 하나의 이상칸을 식별할 확률(β_3), 설정한 이상칸중 식별된 평균 칸수(N_c), 그리고 이상칸으로 잘못 식별한 평균 칸수(N_i)를 조사하였다. 귀무가설에서 고려한 칸 확률은 균일한 확률 $p_{ij}=1/25$로 설정하고 대립가설에 설정한 확률은

$$p_{ij}^{*}=\begin{cases} p_{ij}(1+\Delta_{ij}/\sqrt{N}), & (i,j)\in T_A, \\ p_{ij}\left(1-\sum_{(i,j)\in T_A}p_{ij}^{*}\right)\Big/\sum_{(i,j)\in T_O}p_{ij}, & (i,j)\in T_O, \end{cases} \tag{4.1}$$

여기서 Δ_{ij}는 (i,j)칸을 이상칸으로 설정하기 위해 칸값을 조정할 크기, T_A는 설정된 이상칸들의 집합, T_0는 이상칸을 제외한 칸들의 집합을 나타내고 총 표본의 크기는 $N=500$으로 설정하였다.

<표 4-1>과 같이 3개의 칸을 이상칸으로 설정한 4가지 경우에 대해 두 방법을 통해 얻어진 값들을 비교하여 보자. (1)과 (2)의 경우는 후진단계방법의 $N_c=0.908,\ 1.288$이고 MOCI 방법의 $N_c=0.426, 0.787$로 후진단계방법의 식별정도가 높게 나타났다. 반면에 (3)과 (4)의 경우는 MOCI 방법의 $N_c=2.403, 2.952$가 후진단계방법의 $N_c=2,260, 2.695$보다 크므로 더 높은 식별정도를 보여준다. MOCI 방법은 최소충분통계량에 의해 칸값을 추정하므로 동일한 행에 하나 이상의 이상칸이 존재하는 (1)과 (2)의 경우에서처럼 행 주변합 또는 열 주변합에 많은 영향을 주는 경우 삭제된 잔차의 크기가 오히려 작은 값을 갖게 되어 가장효과를 유발하게 된다. 동일한 행에 이상칸이 존재하더라도 (3)의 경우와 같이 행 주변합에 덜 영향을 주는 경우와 행과 열이 일치하지 않은 (4)의 경우에서는 MOCI 방법의 β_2 또는 N_c가 더 큰 값을 가지므로 후진단계방법보다 가장효과가 작은 것을 볼 수 있다.

이상칸을 식별하는 데 필요한 계산의 양을 살펴보면 후진단계방법은 첫 번째 단계에서 총 25개 칸의 삭제된 잔차를 계산한 후 가장 큰 값을 갖는 칸을 선택한 후 나머지 24개 칸에 대해서 동일한 과정을 수행하여 하나의 칸을 선택한 후 이상칸 여부를 식별하므로 매 단계에서 고려된 칸 전체에 대해 삭제된 잔차를 재계산해야 한다. 따라서 <표 4-1>에 나타난 것과 같이 평균적으로 115(=25+24+23+22+21)번의 계산을 필요로 한다. 그러나 MOCI 방법을 적용할 경우 초기에 고려된 칸에 대해서 다음 단계에서 삭제된 잔차를 계산하기 때문에 평균적으로 33(=25+5+3)번의 삭제된 잔차를 계산하게 된다. 만일 범주의 수준수가 증가하면 이상칸의 식별을 위해 필요한 계산의 양은 MOCI 방법에 비해 후진단계방법이 훨씬 많은 것을 알 수 있다.

<표 4-1> 5×5 분할표에 대한 모의실험 결과

(1) $\Delta_{11}=\Delta_{12}=\Delta_{13}=30$ 인 경우

	β_1	β_2	β_3	N_c	N_i	N_r
후진단계방법	0.806	0.206	0.389	0.908	0.795	115
MOCI	0.994	0.000	0.318	0.426	1.773	33

(2) $\Delta_{11}=20, \Delta_{12}=30, \Delta_{13}=40$인 경우

	β_1	β_2	β_3	N_c	N_i	N_r
후진단계방법	0.851	0.212	0.627	1.288	0.470	115
MOCI	0.998	0.002	0.577	0.787	1.660	33

(3) $\Delta_{11}=-20, \Delta_{12}=20, \Delta_{13}=40$인 경우

	β_1	β_2	β_3	N_c	N_i	N_r
후진단계방법	1	0.281	0.999	2.260	0.579	115
MOCI	1	0.309	1	2.403	0.555	33

(4) $\Delta_{11}=30, \Delta_{22}=30, \Delta_{33}=30$인 경우

	β_1	β_2	β_3	N_c	N_i	N_r
후진단계방법	1	0.655	1	2.695	0.069	115
MOCI	1	0.768	1	2.956	0.263	33

※ β_1 : 검증력, β_2 : 3개의 이상칸을 정확히 식별할 확률, β_3 : 최소한 하나의 이상칸을 정확히 식별할 확률

N_c : 정확히 식별한 평균 이상칸수, N_i : 이상칸으로 잘못 식별한 평균 칸수 N_r : 삭제된 잔차의 평균 추정수

5. 다차원 분할표에서의 이상칸 식별

다차원 분할표의 로그선형모형에 대해 이상칸을 식별하는 정도를 살펴보기 위해 특별히 3×3×3인 3차원 분할표를 고려한다. 직접해가 존재하는 $[AB][C]$ 모형에 대해 이상칸의 식별정도를 살펴보자. 설정한 모형에 적합한 칸 기대값 m_{ijk} 은 다음과 같은 로그선형모형으로부터 구할 수 있다.

$$\log m_{ijk}=u+u_{1(i)}+u_{2(j)}+u_{3(k)}+u_{12(ij)}.$$

칸 기대값 m_{ijk}으로부터 칸 확률 $p_{ijk}=m_{ijk}/m_{+++}$을 구할 수 있으므로 4절에서 설명한 방법으로 각 칸에 대한 확률표본을 얻을 수 있을 것이다. 이때 대립가설에 대한 확률은 2차원에서 고려한 식 (4.1)과 같은 방법을 적용하여 모형을 설정하기로 한다.

2차원 분할표에 대해 이상칸의 식별정도를 조사한 것과 동일하게 검증력(β_1), 설정한 이상칸만을 식별할 확률(β_2), 설정한 이상칸중 최소한 하나의 이상칸을 식별할 확률(β_3), 설정한 이상칸중 식별된 평균 칸수(N_c), 그리고 이상칸으로 잘못 식별한 평균 칸수(N_i)를 조사하였다.

<표 5-1> $[AB][C]$모형에 대해 이상칸 식별정도

이상칸의 설정	β_1	β_2	β_3	N_c	N_i	N_r
$\Delta_{111}=30$	1.000	0.059	1.000	1.000	2.583	34
$\Delta_{111}=30,\ \Delta_{332}=30$	1.000	0.04	1.000	1.636	2.482	34
$\Delta_{111}=30,\ \Delta_{332}=30,\ \Delta_{223}=30$	1.000	0.054	1.000	2.485	2.345	36

직접해가 존재하지 않는 부분연관모형 $[AB][AC][BC]$에 대해 이상칸의 식별정도를 살펴보면 <표 5-2>과 같다.

<표 5-2> $[AB][AC][BC]$모형에서의 이상칸 식별정도

이상칸의 설정	β_1	β_2	β_3	N_c	N_i	N_r
$\Delta_{111}=30$	0.978	0.101	0.853	0.853	2.439	34
$\Delta_{111}=30,\ \Delta_{332}=30$	0.986	0.057	0.917	1.41	2.89	35

<표 5-1>과 <표 5-2>를 살펴보면 올바르게 식별된 이상칸의 수 N_c가 설정한 이상칸의 수보다 작은 것을 알 수 있다. 이들 모형에 대해 처음에 하나 이상의 칸을 이상칸의 대상으로 선택하는 과정에서 편승효과가 나타나는 것을 보여준다. N_i의 값이 크게 나타난 결과는 두 모형 $[AB][AC][BC]$와 $[AB][C]$가 편승효과를 갖는다는 것을 나타낸다. 특히 β_2의 값이 매우 작은 것을 보여주는 데 이는 설정한 이상칸만을 모두 식별하는 것이 어려움을 보여준다. 그러나 제안된 MOCI 방법을 통하여 다차원 분할표에서도 이상칸들을 식별이 가능하고 적은 계산으로 이상칸을 식별할 수 있다는 사실을 기억해야 한다.

6. 결 론

분할표 자료에 대해서 적합한 모형을 설정하는 과정에서 이상칸으로 인해 적합결여가 발생할 수 있다. 이들 칸은 잔차를 이용한 기준과 적합도 검증통계량을 이용하여 이상칸을 식별할 수 있음이 알려져 있고, Fuchs와 Kenett(1980)의 전진단계방법과 Simonoff(1988)에 의해 제안된 후진단계방법에 의해 하나 이상의 이상칸 식별이 가능하다. 전진단계방법과 후진단계방법은 한 번에 하나의 칸에 대해 이상칸 여부를 순차적으로 검증하므로 하나 이상의 칸이 이상칸인 경우에는 많은 계산이 필요할 뿐만 아니라 2차원 분할표에 대해서만 적용이 가능하다.

본 논문에서 제안한 MOCI 방법을 이용하면 이상칸으로 의심나는 칸을 초기에 하나 이상 선택하여 이들 칸들에 대한 검증을 통해 이상칸을 식별하므로 후진단계방법보다 적은 계산으로 이상칸을 식별할 수 있다. 더구나 기존의 이상칸의 식별방법들은 2차원 분할표에 대해서만 가능한데 비해 MOCI 방법은 다차원 분할표인 경우에도 반복적인 추정방법을 통해 다양한 로그선형모형, 선형 대 선형 연관모형(linear-by-linear model), 행과 열효과 모형(row and column effect model) 등에 대해서도 추정이 가능하므로 이상칸의 식별이 가능하다. MOCI 방법은 최소충분통계량에 의해 칸값의 추정량을 얻기 때문에 한 행 또는 한 열에만 여러 개의 이상칸이 존재하는 경우에는 다중이상칸을 식별하는데 어려움이 발생된다. 이와 같은 경우는 분할표를 잘 설명해주는 대안적인 모형을 적합시켜 식별할 것을 제안한다.

참 고 문 헌

[1] 최현집, 신상준 (2000). 반복비율적합에 의한 다차원 분할표의 결측칸값 추정, 「응용통계연구」, 제13권 제1호, 197-205.

[2] 홍종선, 최현집 (1999). 「로그선형모형을 이용한 범주형 자료분석」, 자유아카데미.

[3] Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd, John Wiley & Sons.

[4] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press.

[5] Brown, M. L. (1974). Identification of the sources of significance in two-way contingency tables, *Applied Statistics*, Vol. 23, 405-413.

[6] Christensen, R. (1990). *Log-linear models*, New York : John Wiley & Sons.

[7] Fienberg, S.E. (1969). Preliminary Graphical analysis and quasi-independence for two-way contingency table, *Applied Statistics*, Vol. 18, 153-168.

[8] Fuchs, C. and Kenett, R. (1980). A test for outlying cells in the multinomial distribution and two-way contingency tables, *Journal of the American Statistical Association*, Vol. 75, 395-398.

[9] Goodman, L. A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with or without missing cells, *Journal of the American Statistical Association*, Vol. 63, 1091-1131.

[10] Harberman, S. J. (1973). The analysis of residuals in cross-classified tables, *Biometrics*, Vol. 29, 205-220.

[11] Kotze, T. J. W. and Hawkins, D. M. (1984). The identification of outliers in two-way contingency tables using 2x2 subtables, *Applied Statistics*, Vol. 33, 215-223.

[12] Mosteller, F. and Parunak, A. (1985). Identifying extreme cells in a sizable contingency table : probabilistic and exploratory approaches, *In Exploring Data Tables, Trends and Shapes*, Wiley, pp. 189-224.

[13] Simonoff, J. S. (1988). Detecting outlying cells in two-way contingency tables via backwards-stepping, *Technometrics*, Vol. 30, 339-345.

[14] Upton, G. J. G. and Guillen, M. (1995). Perfect cells, direct models and contingency table outliers, *Communications in Statistics, Part A - Theory and Methods*, Vol. 24, 1843-1862.