Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 4, pp. 717~729

Contour Method and Collapsibility Criteria for $2\times3\times K$ Contingency Tables

C. S. Hong[1] · B. U. Son[2] · J. Y. Park[3]

Abstract

The contour method which was originally designed for $2\times2\times2$ contingency table is studied for $2\times2\times K$ and $2\times3\times K$ tables. Whereas a contour plot for a $2\times2\times K$ table is represented on unit squared two dimensional plane, a contour plot of a $2\times3\times K$ table can be expressed with a regular hexahedron on three dimensional space. Based on contour plots for categorical data fitted to all possible three dimensional log-linear models, one might identify whether $2\times2\times K$ or $2\times3\times K$ tables are collapsible over the third variable.

Keywords : Categorical data, Collapsibility, Contour plot, Log-linear models, Odds ratio.

1. 서론

이차원 평면에 범주형 자료를 표현하는 다양한 시각적인 방법은 자료구조를 이해하기 위해 유용한 정보를 제공하는 탐색적 자료 분석(EDA)을 수행한다. 잘 알려진 EDA에 의한 시각적인 방법은 히스토그램(histogram), 바차트(bar chart), 파이차트(pie chart), 스타차트(star chart) 등이 있다. 이것들은 하나의 범주형 변수에 대해 다양한 범주들의 확률 또는 빈도를 보여준다. Fienberg(1975, 1980)는 이차원 범주형 자료 중 2×2 분할표를 표현하는 'four-fold circuit display'를 제안하였고, 이것은 각 칸의 빈도에 비례하는 반지름을 갖는 4개의 원으로 구성되어 있다. 이차원 분할표를 표현하는 대표적인 방법으로는 블록차트(block chart)가 있으며, Hartigan과

1) First Author : Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea.
E-mail : cshong@skku.ac.kr
2) Assistant teacher, Department of Information Statistics, Korea National Open University, Seoul, 110-791, Korea.
3) Assistant teacher, Department of Information Statistics, Korea National Open University, Seoul, 110-791, Korea.

Kleiner(1981, 1984)는 타일(tile)이라 불리는 정사면체의 크기를 관찰값의 확률에 비례하도록 표현한 'mosaic plot'을 제안하였다. 독립 모형 하에서 Cohen(1980)과 Friendly(1992)는 각 칸의 편차를 표현한 'association plot'을 제안하였다. Friendly(1992, 1994)는 각 타일에 피어슨 카이제곱의 각 칸에 해당하는 편차의 크기를 고려하여 색상과 빗금으로 표현한 보다 향상된 'mosaic plot'을 제안하였다. Tukey(1977)는 이차원 분할표에 대한 적합도를 표현한 'two-way plot'을 제안하였다.

$I\times J\times K$ 삼차원 자료에 대해 세 번째 변수의 각 범주와 분리된 $I\times J$ 분할표는 'mosaic plot'과 'four-fold circular display'로 분석된다. 사차원 또는 그 이상의 고차원 분할표는 향상된 'mosaic plot'을 통해 표현할 수 있다(Hartigan과 Kleiner 1984). 오민권, 홍종선과 이종철(1999)은 다차원 분할표 자료를 분석하고 이를 시각적인 방법으로 설명한 'ring chart' 방법을 제안하였다. 이는 모든 칸의 확률을 링모양의 그림으로 보여주어 범주형 자료의 전체 구조를 설명할 수 있게 해주며 그 외에 'standardized ring chart', 'double ring chart', 'residual ring chart'를 제안하였다. 이 방법들은 주어진 로그선형모형의 적합도를 평가하는 것뿐만 아니라 일차, 이차교호작용을 포함하는 변수들 사이의 관계를 결정하는데 유용한 정보를 얻을 수 있다(홍종선과 이종철 2000).

Fienberg와 Gilbert(1970)는 2×2 분할표에 대해 사면체 내의 궤적(loci)을 환산하여 변수간의 연관성 측도를 기하학적으로 표현하는 방법을 제안하였다. Darroch, Lauritzen와 Speed(1980)는 다차원 분할표에 대한 독립모형과 조건부 독립모형을 표현할 수 있는 그래픽 모형(graphical model)을 개발하였다. 이러한 그래픽 모형들은 변수들 사이에 연관성의 측도를 나타내는 연관그래프(association graph)로 표현하였다.

$2\times2\times2$ 분할표에 대해 Doi, Nakamura와 Yamamoto(2001)는 연관성 측도들 중의 하나인 세 개의 오즈비(odds ratio)를 사용한 contour 방법을 제안하였다 : 두 개의 오즈비는 각각 $k=1,2$에 대한 두 개의 2×2 분할표에서 계산되고 나머지 하나는 세 번째 변수(예를 들면 교락(confounder)변수)에 대하여 차원축소된 2×2 분할표에서 얻어진다. Yamamoto와 Doi(2001), Yamamoto(2002)는 이러한 오즈비들의 기하학적 구조연구를 가능하게 하는 contour 방법을 확장하였다. 이 Contour 방법은 층화된 분할표의 일반적인 오즈비와 차원축소된 분할표의 오즈비가 같은 지를 확인하는 오즈비 차원축소 가능성(collapsibility)을 평가하는 데에 유용하다.

본 연구에서는 일반적인 $2\times2\times K$와 $2\times3\times K$ 분할표를 표현할 수 있는 Contour plot을 개발하였다. $2\times2\times K$ 분할표에 대해서 Contour plot은 $K+1$개의 오즈비의 값들에 대응하는 곡선들과 점들을 포함한다: K개의 오즈비는 각각 $k=1,2,\ldots,K$에 대한 2×2 분할표에서 얻어지고 나머지 하나는 세 번째 변수에 대하여 차원축소된 2×2 분할표에서 계산된다. $2\times2\times K$ 분할표에 대한 Contour plot은 이차원 평면상에 그려지는 반면에, $2\times3\times K$ 분할표에 대한 Contour plot은 삼차원 공간에서 정육면체로 표현되어질 수 있다.

본 논문의 2절에서는 $2\times3\times K$ 분할표를 시각적인 방법으로 구현할 수 있는 Contour plot을 제안하였다. 이 방법은 Doi, Nakamura와 Yamamoto(2001)가 제안한 $2\times2\times2$ 분할표에 대한 Contour plot을 확장하였는데, $2\times2\times2$ 분할표($2\times2\times K$ 포함)의 Contour plot은 이차원 평면에서 표현되지만 $2\times3\times K$ 분할표의 Contour plot은 삼차원 공간에서 표현된다. 3절에서는 삼차원 로그선형모형들의 종류를 소개하고,

$2\times3\times K$ 범주형 자료에서의 오즈비들의 특성으로 모형들을 몇 개의 집단으로 분류하여 설명하였다. 오즈비들의 특성으로 분류한 로그선형모형에 적합한 $2\times3\times K$ 범주형 자료를 생성하여 삼차원 Contour plot을 4절에서 구현하였다. 5절에서는 오즈비들의 특성으로 분류한 로그선형모형과 이에 적합한 자료를 구현한 Contour plot과의 관계를 연결하여 차원축소에 대하여 설명하였다. 즉 Contour plot에서 나타나는 Contour를 살펴보면 자료에 적합한 로그선형모형이 어떠한 성격을 가졌으며 차원축소에 대하여는 어떻게 설명되는지를 연구하였다. 마지막 6절에서는 4차원 이상의 그림을 구현할 수 없는 $2\times J\times K$ $(J\geq 4)$ 분할표 자료에 대하여 오즈비를 정의하고 오즈비의 특성을 4절과 5절에서 분류한 로그선형모형들에 대하여 토론하였다.

2. $2\times3\times K$ 분할표에 대한 Contour 방법

Doi, Nakamura와 Yamamoto(2001)는 $2\times2\times2$ 분할표에 대한 Contour 방법을 제안하였다. 이 방법은 $2\times2\times K$ 분할표로 확장될 수 있는데 이 확장은 $2\times3\times K$ 분할표에 대한 Contour 방법을 설명한 후에 쉽게 이해할 수 있다.

우선 삼차원 $2\times3\times K$ 분할표에 대한 기본적인 개념과 수실을 정의하기로 한다. 다음의 표 2.1과 같이 분할표에서 칸확률 p_{ijk}를 가지는 $2\times3\times K$ 분할표와 세 번째 변수에 대하여 차원축소된 2×3 분할표를 고려하자.

표 2.1: $2\times3\times K$분할표와 차원축소된 2×3 분할표

p_{111}	p_{121}	p_{131}	p_{112}	p_{122}	p_{132}	.....	p_{11K}	p_{12K}	p_{13K}
p_{211}	p_{221}	p_{231}	p_{212}	p_{222}	p_{232}		p_{21K}	p_{22K}	p_{23K}

p_{11+}	p_{12+}	p_{13+}
p_{21+}	p_{22+}	p_{23+}

표 2.1에서 k번째 2×3 분할표에 대한 p_k, q_k, r_k와 차원축소된 2×3 분할표에 대한 p_c, q_c, r_c를 다음과 같이 정의하자.

$$\begin{aligned} p_k &= \frac{p_{11k}}{p_{+1k}},\ q_k = \frac{p_{12k}}{p_{+2k}},\ r_k = \frac{p_{13k}}{p_{+3k}} \\ p_c &= \frac{p_{11+}}{p_{+1+}},\ q_c = \frac{p_{12+}}{p_{+2+}},\ r_c = \frac{p_{13+}}{p_{+3+}} \end{aligned} \tag{2.1}$$

k번째 2×3 분할표의 오즈비와 차원축소된 분할표의 오즈비는 다음과 같이 표현된다.

$$\theta_{1k} = \frac{p_k/(1-p_k)}{q_k/(1-q_k)}, \quad \theta_{2k} = \frac{p_k/(1-p_k)}{r_k/(1-r_k)}, \quad k=1, 2, \ldots, K$$
$$\theta_{1c} = \frac{p_c/(1-p_c)}{q_c/(1-q_c)}, \quad \theta_{2c} = \frac{p_c/(1-p_c)}{r_c/(1-r_c)} \tag{2.2}$$

여기에서 2열과 3열에 대한 오즈비 $\frac{q_k/(1-q_k)}{r_k/(1-r_k)}$, $\frac{q_c/(1-q_c)}{r_c/(1-r_c)}$는 각각 θ_{2k}/θ_{1k}, θ_{2c}/θ_{1c}으로 표현된다는 것을 추가적으로 언급한다. 따라서 2열과 3열에 대한 오즈비는 1열과 2열 그리고 1열과 3열의 오즈비의 함수이므로 따로 정의할 필요가 없다.

Doi, Nakamura와 Yamamoto(2001)는 $2\times2\times2$ 분할표에 대한 Contour 방법을 설명하기 위하여 함수 $f(p, q)$와 contour $C(\theta)$는 다음과 같이 정의하였다.

$$f(p, q)=\frac{p/(1-p)}{q/(1-q)}, \ (p, q)\in(0,1)^2$$

$$C(\theta)=\{(p, q)\in(0,1)^2 : f(p, q)=\theta\}, \ \theta>0$$

contour $C(\theta)$는 오즈비 값이 θ인 모든 좌표 (p, q)의 집합이며 그림 2.1은 $\theta=0.1, 0.5, 1, 2, 10$에 대한 $C(\theta)$의 형태를 나타낸다.

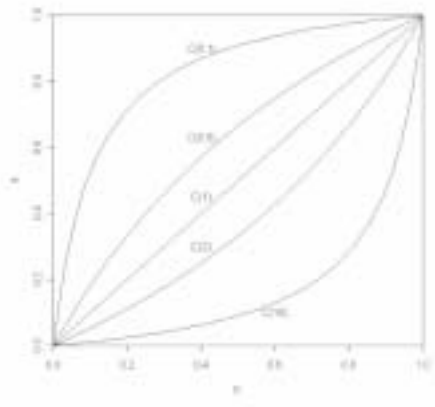

그림 2.1: $C(\theta)$의 형태

$2\times3\times K$ 분할표에 대한 Contour 방법을 확장하기 위하여 contour $C(\theta_1, \theta_2)$를 다음과 같이 정의한다.

$$C(\theta_1, \theta_2)=\left\{(p, q, r)\in(0,1)^3 : \frac{p/(1-p)}{q/(1-q)}=\theta_1, \frac{p/(1-p)}{r/(1-r)}=\theta_2\right\}, \ \theta_1>0, \ \theta_2>0 \tag{2.4}$$

그림 2.2는 정육면체 상의 θ_1, θ_2의 몇몇 값들에 대한 $C(\theta_1, \theta_2)$의 형태를 보여준다.

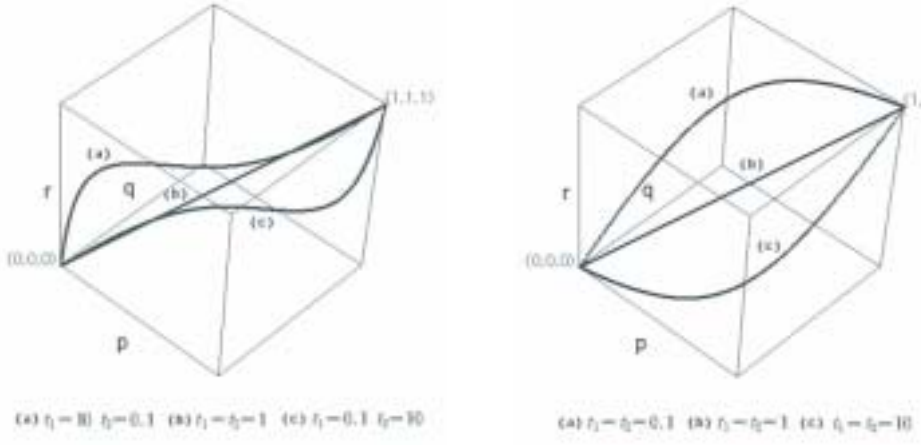

그림 2.2 $C(\theta_1, \theta_2)$의 형태

3. 삼차원 로그선형모형과 오즈비

삼차원 범주형 자료에 대한 로그선형모형들을 Christensen(1990)이 사용한 표기 방법을 따라 분류하면 다음과 같다: 포화모형(saturated model)은 [123], 부분연관모형(partial association model)은 [12][13][23], 조건부독립모형(conditionally independent model)은 [12][13], [12][23], [13][23], 한 변수의 독립모형(model with one factor independent of the other two)은 [12][3], [13][2], [1][23], 그리고 완전독립모형(completely independent model)은 [1][2][3]으로 표기한다.

첫 번째 집단으로 분류하는 모형으로는 첫 번째와 두 번째 변수의 일차 교호작용항이 포함된 [12][3], [12][13], [12][23] 모형이며 이 집단에 속하는 모형의 경우 (2.1) 식에서 정의한 p_{k} q_{k} r_{k}를 구하면 다음과 같다.

$$\left.\begin{aligned} p_k &= \frac{p_{11k}}{p_{+1k}} = \frac{p_{11+}p_{++k}}{p_{+1+}p_{++k}} = \frac{p_{11+}}{p_{+1+}} = p_c \\ q_k &= \frac{p_{12k}}{p_{+2k}} = \frac{p_{12+}p_{++k}}{p_{+2+}p_{++k}} = \frac{p_{12+}}{p_{+2+}} = q_c \\ r_k &= \frac{p_{13k}}{p_{+3k}} = \frac{p_{13+}p_{++k}}{p_{+3+}p_{++k}} = \frac{p_{13+}}{p_{+3+}} = r_c \end{aligned}\right\} \quad [12][3]\text{모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{11+}p_{+1k}/p_{+1+}}{p_{+1+}p_{+1k}/p_{+1+}} = \frac{p_{11+}}{p_{+1+}} = p_c \\ q_k &= \frac{p_{12+}p_{+2k}/p_{+2+}}{p_{+2+}p_{+2k}/p_{+2+}} = \frac{p_{12+}}{p_{+2+}} = q_c \\ r_k &= \frac{p_{13+}p_{+3k}/p_{+3+}}{p_{+3+}p_{+3k}/p_{+3+}} = \frac{p_{13+}}{p_{+3+}} = r_c \end{aligned}\right\} \quad [12][23]\text{모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{11+}p_{1+k}/p_{1++}}{(p_{11+}p_{1+k}/p_{1++})+(p_{21+}p_{2+k}/p_{2++})} \\ q_k &= \frac{p_{12+}p_{1+k}/p_{1++}}{(p_{12+}p_{1+k}/p_{1++})+(p_{22+}p_{2+k}/p_{2++})} \\ r_k &= \frac{p_{13+}p_{1+k}/p_{1++}}{(p_{13+}p_{1+k}/p_{1++})+(p_{23+}p_{2+k}/p_{2++})} \end{aligned}\right\} \quad [12][13]\text{모형}$$

[12][3], [12][23] 모형에서 p_k, q_k, r_k는 k에 의존하지 않으므로 각각 p_c, q_c, r_c와 동일하다. 따라서 모든 k에 대하여 $\theta_{1k}=\theta_{1c}$, $\theta_{2k}=\theta_{2c}$이다. 또한 [12][13] 모형에서 p_k q_k r_k는 k에 독립적이지 않지만 $\theta_{1k}=p_{11+}p_{22+}/p_{12+}p_{21+}$, $\theta_{2k}=p_{11+}p_{23+}/p_{13+}p_{21+}$가 되어 k와 독립이며 따라서 각각 θ_{1c}, θ_{2c}의 값과 동일함을 유도할 수 있다. 그러므로 [12][3], [12][13], [12][23] 모형에서는 모든 k에 대하여 $\theta_{1k}=\theta_{1c}$, $\theta_{2k}=\theta_{2c}$임을 얻을 수 있다.

두 번째 모형 집단에 속하는 [1][2][3], [1][23], [13][2] 모형에 대하여 p_k, q_k, r_k를 구하면 다음과 같다.

$$\left.\begin{aligned} p_k &= \frac{p_{1++}p_{+1+}p_{++k}}{p_{+++}p_{+1+}p_{++k}} = p_{1++} \\ q_k &= \frac{p_{1++}p_{+2+}p_{++k}}{p_{+++}p_{+2+}p_{++k}} = p_{1++} \\ r_k &= \frac{p_{1++}p_{+3+}p_{++k}}{p_{+++}p_{+3+}p_{++k}} = p_{1++} \end{aligned}\right\} \quad [1][2][3]\text{모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{1++}p_{+1k}}{p_{+++}p_{+1k}} = p_{1++} \\ q_k &= \frac{p_{1++}p_{+2k}}{p_{+++}p_{+2k}} = p_{1++} \\ r_k &= \frac{p_{1++}p_{+3k}}{p_{+++}p_{+3k}} = p_{1++} \end{aligned}\right\} \quad [1][23]\text{모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{1+k}p_{+1+}}{p_{++k}p_{+1+}} = \frac{p_{1+k}}{p_{++k}} \\ q_k &= \frac{p_{1+k}p_{+2+}}{p_{++k}p_{+2+}} = \frac{p_{1+k}}{p_{++k}} \\ r_k &= \frac{p_{1+k}p_{+3+}}{p_{++k}p_{+3+}} = \frac{p_{1+k}}{p_{++k}} \end{aligned}\right\} \quad [13][2]\text{모형}$$

[1][2][3], [1][23], [13][2] 모형에 대한 p_k, q_k, r_k는 모두 동일하기 때문에 모든 오즈비의 값은 1이다. 즉 모든 k에 대하여 $\theta_{1k}=\theta_{1c}=\theta_{2k}=\theta_{2c}=1$이다.

앞에서 언급한 두 종류의 모형 집단에 속하지 않은 모형 중 [13][23] 모형의 p_k, q_k, r_k는 다음과 같이 동일하나,

$$\left.\begin{aligned} p_k &= \frac{p_{1+k}p_{+1k}/p_{++k}}{p_{++k}p_{+1k}/p_{++k}} = \frac{p_{1+k}}{p_{++k}} \\ q_k &= \frac{p_{1+k}p_{+2k}/p_{++k}}{p_{++k}p_{+2k}/p_{++k}} = \frac{p_{1+k}}{p_{++k}} \\ r_k &= \frac{p_{1+k}p_{+3k}/p_{++k}}{p_{++k}p_{+3k}/p_{++k}} = \frac{p_{1+k}}{p_{++k}} \end{aligned}\right\} \quad [13][23]\text{모형}$$

이 모형에 대한 p_c, q_c, r_c는 간략히 정리할 수 없다. 따라서 모든 k에 대하여 $\theta_{1k}=\theta_{2k}=1$임을 유도할 수 있으나, 차원축소된 자료에서 θ_c의 값은 θ_k의 값과 동일하지 않다는 것을 발견하였다.

마지막으로 부분연관모형인 [12][13][23] 모형에서는 p_{ijk}에 대한 해를 직접적인 방법으로 구할 수 없기 때문에 p_k, q_k, r_k를 정의할 수 없다.

$2\times3\times K$ 분할표 자료에 대하여 위에서 연구한 오즈비에 관하여 정리하면 다음과 같다. 첫 번째 모형집단인 [12][3], [12][13], [12][23] 모형에서는 다음과 같은 관계식을 얻는다. 모든 k에 대해서,

$$\theta_{1k}=\theta_{1c},\ \theta_{2k}=\theta_{2c}. \tag{3.1}$$

두 번째 모형집단인 [1][2][3], [1][23], [13][2] 모형에서는, 모든 k에 대해서

$$\theta_{1k}=\theta_{1c}=\theta_{2k}=\theta_{2c}=1 \tag{3.2}$$

이며, 그 밖의 모형 중 [13][23] 모형에서는, 모든 k에 대해서

$$\theta_{1k}=\theta_{2k}=1 \tag{3.3}$$

의 관계를 유도하였다. 그리고 [12][13][23] 모형에서는 이차교호작용항이 존재하지 않기 때문에 오직 다음과 같은 관계식만을 유도할 수 있다.

$$\theta_{11} = \cdots = \theta_{1K}, \quad \theta_{21} = \cdots = \theta_{2K} \cdot \tag{3.4}$$

4. Contour 방법

3절에서 논의한 삼차원 로그선형모형들의 오즈비의 특성을 Contour plot을 통하여 살펴보자. 우선 다양한 로그선형모형에 적합한 $2\times3\times3$ 범주형 자료를 생성하는데, 대표적으로 [12][3], [13][2], [13][23] 그리고 [12][13][23] 모형을 따르는 $2\times3\times3$ 자료를 생성한다. [12][3], [13][2], [13][23] 모형은 직접해(direct solution)가 존재하기 때문에 칸확률 $\{p_{ijk}\}$는 충분주변형태(sufficient configuration, sufficient marginal probability)로 정리한다. 즉 각 모형의 칸확률은 표 4.1과 같이 구한다.

표 4.1 칸확률의 직접해

모형	칸확률
[12][3]	$p_{ij+}p_{++k}$
[13][2]	$p_{1+k}p_{+j+}$
[13][23]	$p_{i+k}p_{+jk} / p_{++k}$

주어진 충분주변확률(sufficient marginal probability)로 칸확률을 설정한 후 이 확률과 표본크기(N=1,000)의 다항분포를 이용하여 칸 빈도수 x_{ijk}를 생성한다(Rudas (1984, 1986)를 참조). 예를 들어 [12][3] 모형에 대하여는 주어진 충분주변확률 $\{p_{ij+}\}$, $\{p_{++k}\}$을 이용하여 칸확률 $\{p_{ijk}\}$를 구하고, 설정된 칸확률과 표본크기의 다항분포를 이용하여 칸 빈도수를 생성한다. 그러나 부분연관모형인 [12][13][23] 모형은 직접해가 존재하지 않으므로 정동빈, 홍종선과 윤상호(2003)의 연구 방법을 이용하여 적절한 구간의 균일분포를 따르는 난수를 생성하고 부분연관모형의 여러 모수의 추정값을 얻은 후, 이에 대응하는 칸 빈도수를 생성한다. 표 4.2부터 표 4.5까지는 [12][3], [13][2], [13][23] 그리고 [12][13][23] 모형에 적합한 자료를 생성하여 나타내었다. 그림 4.1부터 그림4.4는 생성된 자료를 각각 Contour plot으로 구현하였다.

표 4.2 [12][3] 모형의 자료

[12][3]	k=1			k=2			k=3		
	j=1	j=2	j=3	j=1	j=2	j=3	j=1	j=2	j=3
i=1	131	53	36	180	58	33	146	42	31
i=2	30	40	34	39	44	33	30	35	29

우도비통계량값 = 5.12 p-값 = 0.8833

표 4.3 [13][2] 모형의 자료

[13][2]	k=1			k=2			k=3		
	j=1	j=2	j=3	j=1	j=2	j=3	j=1	j=2	j=3
i=1	146	195	130	42	68	41	22	41	33
i=2	30	39	25	35	38	19	26	39	31

우도비통계량값 = 11.22 p-값 = 0.3408

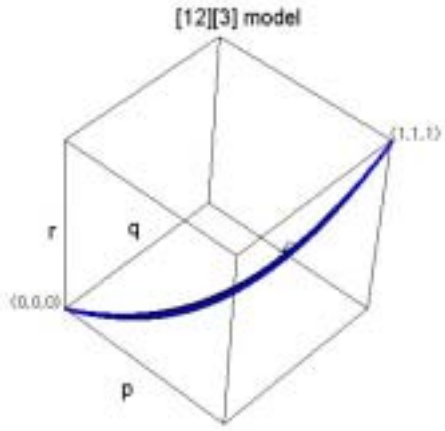

그림 4.1 [12][3] 모형의 Contour plot

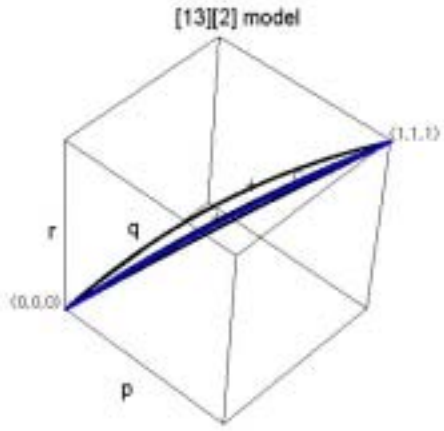

그림 4.2 [13][2] 모형의 Contour plot

표 4.4 [13][23] 모형의 자료

[13][23]	k=1			k=2			k=3		
	j=1	j=2	j=3	j=1	j=2	j=3	j=1	j=2	j=3
i=1	275	136	92	69	50	32	23	24	55
i=2	50	17	15	32	35	13	20	14	48

우도비통계량값 = 5.44 p-값 = 0.4889

표 4.5 [12][13][23] 모형의 자료

[12][13][23]	k=1			k=2			k=3		
	j=1	j=2	j=3	j=1	j=2	j=3	j=1	j=2	j=3
i=1	72	37	40	73	67	40	21	62	68
i=2	39	25	42	31	20	33	30	112	187

우도비통계량값 = 2.95 p-값 = 0.5655

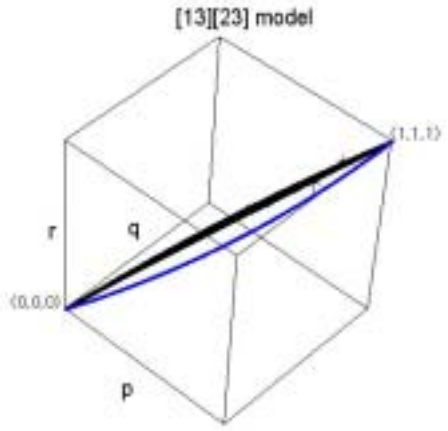

그림 4.3 [13][23] 모형의 Contour plot

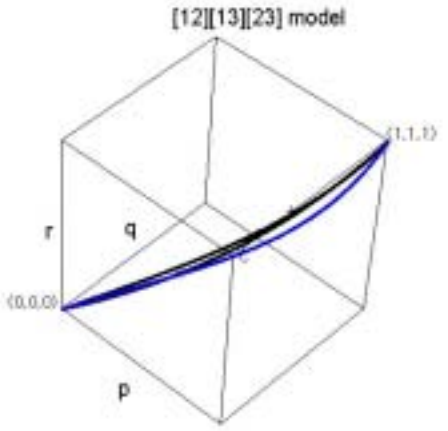

그림 4.4 [12][13][23] 모형의 Contour plot

그림 4.1부터 그림4.4를 살펴보면, 3절에서 논의한 삼차원 로그선형모형들을 분류한 집단의 특징인 (3.1) 식부터 (3.4) 식까지의 관계가 표현되고 있음을 파악할 수 있다. 다음 절에서는 Contour plot에서 표현되는 오즈비의 특성을 나타낸 (3.1) 식부터 (3.4) 식까지가 차원축소(collapsibility)와 어떤 관계를 갖고 있는지를 살펴보기로 하자.

5. 차원축소와 Contour plot

Bishop, Fienberg와 Holland(1975, pp. 39, 47), Agresti(1984, pp. 146), 그리고 Christensen(1990, pp. 114)은 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형은 [12][3], [12][13], [12][23] 모형으로 정의하였으며 그 이외 다른 모형은 차원축소 불가능한 모형이라고 하였다. Whittemore(1978), Ducharme과 Lepage(1986), 그리고 Geng(1992)은 차원축소 가능한 모형인 [12][3], [12][13], [12][23] 모형과 차원축소 불가능 모형 중 [1][2][3], [1][23], [13][2] 모형을 합한 모형들을 strict collapsibility 모형으로 정의하였으며, 이 모형 중에서 차원축소 가능 모형인 [12][3], [12][13], [12][23] 모형을 strong collapsibility라고 정의하였다.

이 이론을 바탕으로 (3.1) 식의 성질을 만족하는 [12][3], [12][13], [12][23] 모형은 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형이며 strong and strict collapsibility 모형으로 정의하는데, 이 모형에 적합한 $2\times3\times K$ 분할표 자료를 Contour plot으로 구현하면, 정육면체의 좌표 (0,0,0)과 (1,1,1)을 연결하는 대각선과는 멀리 떨어진 곳에 모든 contour가 동일하게 위치한다는 것을 파악할 수 있다. 따라서 $2\times3\times K$ 분할표 자료에 대하여 Contour plot을 구현하여 오즈비에 대응하는 contour를 살펴본 후, 모든 k에 대하여 $\theta_{1k}=\theta_{1c}$, $\theta_{2k}=\theta_{2c}$이며, 그 값들이 1이 아니라는 (3.1) 식의 이론과 일치하는 경향을 발견하였다면, 이 자료에 적합한 로그선형 모형들은 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형이며 strong and strict collapsibility 모형이라고 해석할 수 있으며, 특히 [12][3], [12][13], [12][23] 모형 중의 하나임을 탐색적으로 식

별할 수 있다. $2\times2\times K$ 분할표 자료에 대하여는 이차원 Contour plot에서 (0,0)과 (1,1)을 연결하는 대각선과 멀리 떨어진 오목 또는 볼록 형태의 contour로 표현된다.

다음으로 (3.2) 식을 만족하는 [1][2][3], [1][23], [13][2] 모형은 세 번째 변수가 차원축소 불가능한 모형이라고 정의할 수 있으나 다른 한편으로는 strict collapsibility이지만 strong collapsibility는 아닌 모형으로 정의할 수 있는데, 이 모형에 적합한 $2\times3\times K$ 분할표 자료를 Contour plot으로 구현하면, 정육면체의 좌표 (0,0,0)과 (1,1,1)을 연결하는 대각선과 가까운 곳에서 contour가 집중된다는 것을 인식할 수 있다. 따라서 $2\times3\times K$ 분할표 자료에 대하여 Contour plot을 구현하여 오즈비에 대응하는 contour를 살펴본 후, 모든 k에 대하여 오즈비의 값들이 동일하며 그 값들이 1에 근접하다는 (3.2) 식의 이론과 일치하는 경향을 발견하였다면, 이 자료에 적합한 로그선형 모형들은 세 번째 변수가 차원축소 불가능한 모형 또는 strict collapsibility이지만 strong collapsibility는 아닌 모형으로 해석할 수 있으며, 특히 [1][2][3], [1][23], [13][2] 모형 중의 하나임을 탐색할 수 있다. $2\times2\times K$ 분할표 자료에 대하여는 이차원 Contour plot에서 (0,0)과 (1,1)을 연결하는 직선 형태의 contour로 표현된다.

[13][23] 모형과 [12][13][23] 모형에 대하여는 차원축소된 오즈비 θ_{1c}, θ_{2c}와 그에 대응하는 contour에 대하여 뚜렷한 이론을 설정할 수 없기 때문에 이 모형에 대한 Contour plot의 특색은 언급하기에는 한계가 있다.

6. 향후 연구과제와 토론

본 연구에서는 $2\times2\times K$와 $2\times3\times K$ 분할표 자료에 대하여 Contour plot을 구현하였으며 이를 통하여 자료에 적합한 로그선형모형의 특징을 살펴보았다. 일반적인 $2\times J\times K$ $(J\geq4)$ 분할표 자료를 고려하여 보자. 4차원 이상의 그림을 구현하는 것은 쉬운 일이 아니므로 향후 연구 과제로 남겨두기로 하고, 다만 (2.2) 식에서 정의한 오즈비는 다음과 같이 일반화 시킬 수 있다. 모든 $j=2,\cdots,J$, $k=1,\cdots,K$에 대하여,

$$\theta_{(j-1)k}=p_{11k}\,p_{2jk}\,/\,p_{1jk}\,p_{21k}$$
$$\theta_{(j-1)c}=p_{11+}\,p_{2j+}\,/\,p_{1j+}\,p_{21+}$$

5절에서 연구한 오즈비를 $2\times J\times K$ $(J\geq4)$ 분할표 자료에 확장하면 다음과 같이 정리할 수 있다. [12][3], [12][13], [12][23] 모형에서는 주어진 $j=2,\cdots,J$에서 다음을 만족한다. 모든 k에 대해서

$$\theta_{(j-1)k}=\theta_{(j-1)c}$$

[1][2][3], [1][23], [13][2] 모형에서는 모든 $j=2,\cdots,J$, $k=1,\cdots,K$에서

$$\theta_{(j-1)k}=\theta_{(j-1)c}=1$$

이며, [13][23] 모형에서는 모든 $j=2,\cdots,J,\quad k=1,\cdots,K$에서

$$\theta_{(j-1)k}=1$$

의 관계를 유도하였다. 그리고 [12][13][23] 모형에서는 이차교호작용항이 존재하지 않기 때문에 주어진 $j=2,\cdots,J$에서 오직 다음과 같은 관계식만을 유도할 수 있다.

$$\theta_{(j-1)1}=\cdots=\theta_{(j-1)K}$$

$2\times J\times K\quad (J\geq 4)$ 분할표 자료를 4차원 이상의 Contour plot으로 구현하지 못하더라도, 본 절에서 설명한 오즈비의 성격으로 자료에 적합한 로그선형모형의 특징을 파악할 수 있겠다.

참고문헌

1. Agresti, A. (1984). *Analysis of Ordinary Categorical Data*, New York: John Wiley and Sons.
2. Bishop, Yvonne M. M., Fienberg, Steve E., and Holland, Paul W. (1975). *Discrete Multivariate Analysis*, Cambridge, Massachusetts: MIT Press.
3. Christensen, Ronaldo. (1990). *Log-Linear Models*, New York: Springer-Verlag.
4. Cohen, A. (1980). On the Graphical Display of the Significant Components in a Two-way Contingency Table, *Communications in Statistics - Theory and Methods*, A9, 1025-1041.
5. Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov Fields and Log-linear Interaction Models for Contingency Tables, *Annals of Statistics*, 8, 522-539.
6. Doi, M., Nakamura, T., and Yamamoto, E. (2001). Conservative Tendency of the Crude Odds Ratio, *Journal of Japan Statistical Society*, 1(1), 1-19.
7. Ducharme, G. R., and Lepage, Y. (1986). Testing Collapsibility in Contingency Tables, *Journal of the Royal Statistical Society*, B, 48(2), 197-205.
8. Fienberg, S. E., and Gilbert, J. P. (1970). The geometry of 2×2 Contingency Tables, Journal of the American Statistical Association, 65, 694-701.
9. Fienberg, S. E. (1975). Perspective Canada as a Social Report, *Social Indicators Research*, 2, 154-174.
10. Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd ed, The MIT Press.

11. Friendly, M. (1992). Mosaic Displays for Log-linear Models, *Proceedings of the Statistical Graphics Section, the American Statistical Association*, 61-68.
12. Friendly, M. (1994). Mosaic Displays for Multi-way Contingency Tables, *Journal of the American Statistical Association*, 89, 190-200.
13. Geng, Z. (1992). Collapsibility of Relative Risk in Contingency Tables with a Response Variable, *Journal of the Royal Statistical Society*, B, 54(2), 585-593.
14. Hartigan, J. A., and Kleiner, B. (1981). Mosaic for Contingency Tables, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, ED. W. F. Eddy, New York: Springer-Verlag, 268-273.
15. Hartigan, J. A., and Kleiner, B. (1984). A Mosaic of the Television Ratings, *The American Statistician*, 38, 32-35.
16. Hong, C. S., and Lee, J. C. (2000). Ring Chart Ⅱ for Multidimensional Categorical Data, *Korean Journal of Applied Statistics*, 13(1), 163-177.
17. Jeong, D. B., Hong, C. S., and Yoon, S. H. (2003). Empirical Comparisons of Disparity Measures for Partial Association Models in Three Dimensional Contingency Tables, *The Korean Communications in Statistics*, 10(1), 135-144.
18. Oh, M. G., Hong, C. S., and Lee, J. C. (1999). Ring Chart for Categorical Data, *Korean Journal of Applied Statistics*, 12(1), 225-239.
19. Rudas, T. (1984). Testing goodness-of fit of log-linear models based on small sample: a Monte Carlo study, *Colloquia Mathematica Societas Ja'nos Bolyai*, 45, Goodness-of-fit, North-Holland.
20. Rudas, T. (1986). A Monte Carlo Comparison of the Small Sample Behaviour of the Pearson, the Likelihood Ratio and the Cressie-Read Statistic, *Journal of the Statistical Computation and Simulation*, 24, 107-120.
21. Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley Publishing Company.
22. Whittemore, A. S. (1978). Collapsibility of Multidimensional Contingency Tables, *Journal of the Royal Statistical Society*, B, 40(3), 328-340.
23. Yamamoto, E., and Doi, M. (2001). Noncollapsibility of Common Odds Ratios without/with Confounding, *Bulletin of The 53rd Session of the International Statistical Institute*, Book 3, 39-40.
24. Yamamoto, E. (2002). Collapsibility Conditions for Common Odds Ratio in two 2×2 Tables, *The 4th Conference of the Asian Regional Section of the International Association for Statistical Computing*, Session 623, 66-69.

[2004년 6월 접수, 2004년 9월 채택]