응용통계연구 제 19권 2호, 2006년, pp. 361–368

다이아몬드 그래프의 활용 방법

홍종선[1] 고용석[2]

요 약

이차원 범주형 자료를 시각적으로 표현하는 이차원과 삼차원 그래프는 많이 존재한다. 그중에서 Li 등(2003)은 삼차원 그래프를 이차원 평면에 투영시키는 다이아몬드 그래프를 제안하였다. 여기서 세번째 차원은 면적과 높이 그리고 길이가 관찰값에 대응하는 다이아몬드 모양으로 대체하였다. 본 논문에서는 이차원 자료에 대하여는 두 범주형 변수의 독립성을 검정하기 위하여 다이아몬드 그래프를 이용한다. 그리고 삼차원 이상의 자료에 대해서는 자료에 가장 적합한 로그선형모형을 설정하는데 활용할 수 있다.

주요용어: 독립성 검정, 로그선형모형, 적합도 검정.

1. 서론

범주형 자료를 시각적으로 표현하기 위한 연구는 최근까지 많은 방법들이 개발되었으며, 이 방법들 중에서 이차원 범주형 자료를 표현하는 삼차원 (3-D) 그래프들은 빈도수를 나타내는 결과변수와 두 설명변수들 사이의 관계를 표현하는데 역부족인 면이 있다. 즉 3-D 그래프들은 잘못 해석하기 쉽고 잘못 오인할 수도 있으며, 자료를 중복되지 않게 표현하는데 한계가 있다(Cleveland and McGill, 1984; Wilkinson, 1999; Harris, 1999). 이런 단점을 보완하기 위하여 이차원 그래프를 주로 사용하는데, Four-Fold Plot (Friendly 1994b), Association Plot (Cohen, 1980; Friendly, 1991), Mosaic Plot (Hartigan and Kleiner, 1981, 1984; Friendly, 1994a; Wilkinson, 1999), Grouped Bar Graph (Tufte, 1983), Grouped Dot Plot과 Framed Rectangle Chart (Cleveland and McGill, 1984), Trellis Display (Becker, Cleveland and Shyu, 1996), 그리고 Ring Chart(Oh, Hong and Lee, 1999; Hong and Lee, 2000)등이 있다.

최근에 Li, Buechner, Tarwater와 Munoz(2003)는 육각형의 높이, 밑변, 가운데 수평 길이 그리고 면적을 칸비율의 선형식으로 표현하는 '다이아몬드 그래프(Diamond Graph)'를 제안하였다. 이 그래픽 표현 방법은 어느 각도에서도 똑같은 시각적 영향을 미치는 사각 칸에 대한 다각형으로 표시하기 때문에(상하좌우의 대칭적 구조), 두 변수 모두를 동등하게 표현할 수 있다는 점과 다른 변수의 결과에 감춰지는 현상이 발생하지 않는 장점이 있다. 따라서 일반적으로 많이 쓰이고 있는 3-D 그래프와는 달리 원근의 확인이 불필

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수
E-mail: cshong@skku.ac.kr
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 대학원

요하고 인접한 관찰값들을 쉽게 확인할 수 있다. 또한 사각 칸 안에 형성된 영역을 구성하는 요소가 모두 선형 관계식으로 이루어져있고 육각형이란 가장 안정된 형태의 도형을 사용하여 시각적인 효과를 상승시켰다. 다이아몬드 그래프의 원리를 요약하면 다음과 같다: 관찰칸값 f_{ij}들을 모든 관찰칸값 중에서 가장 큰 값으로 나눈 비율 p로 전환시킨다, $p_{ij} = f_{ij}/\max\{f_{ij}\}$. 대각선의 길이가 l인 마름모 모양의 칸의 면적에 p의 비율로 비례하는 면적을 가진 육각형 모양의 다이아몬드를 설정한다. 즉 다이아몬드의 높이$= V(p) = p$, 가운데 수평 길이$= H(p) = 0.5 + 0.5p$, 윗변과 밑변의 길이$= h(p) = 0.5 - 0.5p$로 설정하면, 다이아몬드 모양의 면적은 칸의 면적에 $V(p)[H(p) + h(p)] = p$의 비율로 비례한다.(예를 들어 $l = 1$ 인 경우, 마름모 모양의 칸 면적은 1/2이며 다이아몬드 모양의 면적은 $p/2$이다.)

다이아몬드 그래프는 관찰칸값에 대응하는 비율로 표현되므로 그래프만으로는 두 설명변수의 독립적인 관계에 대하여 판단할 수 없다. 본 논문에서는 관찰값의 비율뿐만 아니라 독립모형하에서의 기대값의 비율도 다이아몬드 그래프로 표현하여, 이차원 범주형 자료에서 두 변수가 독립인 귀무가설에 얼마나 적합한지를 파악하는데 활용할 수 있는 방법을 제안한다. 독립성 검정을 위해서는 관찰칸값과 기대칸값에 대응하는 비율을 새로 설정해야 하는데 이에 대하여는 2.1절에서 논의하고자 한다. 또한 다이아몬드 그래프를 이용하여 삼차원 이상의 범주형자료에 가장 적합한 로그선형모형을 탐색적으로 발견할 수 있는 방법을 2.2절에서 설명한다.

2. 다이아몬드 그래프의 활용

2.1. 이차원 자료에서 독립성 검정

Li 등(2003)은 이차원 분할표 자료를 다이아몬드 그래프로 표현하기 위해서는 관찰칸값 f_{ij}에 대응하는 비율을 $p_{ij} = f_{ij}/\max\{f_{ij}\}$로 설정하였으나, 본 논문에서는 두 변수가 독립이라는 귀무가설에서의 기대칸값 m_{ij}에 대응하는 비율도 함께 고려해야 하기 때문에 관찰칸값과 기대칸값에 대응하는 비율을 각각 다음과 같이 정의한다.

$$p_{ij} = f_{ij}/\max\{f_{ij}, m_{ij}\} \tag{2.1}$$

$$\hat{p}_{ij} = m_{ij}/\max\{f_{ij}, m_{ij}\}, \tag{2.2}$$

여기서 $m_{ij} = f_{i+}f_{+j}/N$이며, 행과 열 각각의 주변합 $f_{i+} = \sum_j f_{ij}$와 $f_{+j} = \sum_i f_{ij}$ 그리고 N은 총빈도수이다.

다음과 같은 실증 예제를 통하여 관찰칸값과 기대칸값에 대응하는 비율 p_{ij}, $\hat{p}_{ij}$로 표현되는 다이아몬드 그래프를 이용하여 이차원 범주형 자료의 독립성 검정에 대하여 토론해보자. 표 2.1은 버클리대학교의 6개 학과의 성별과 입학허가에 관한 2×2 분할표자료(Bickel, Hammel과 O'Connell, 1975)이며 두 변수의 독립성을 검정하는 피어슨 χ^2 통계량값은 92.2053(p-값=0.0001)으로 성별에 따른 입학허가 여부는 연관이 있음을 알 수 있다. 또한, 표 2.2는 심장혈관질병을 앓고 있지 않는 3182명의 사람을 성격과 운동습관에 따라 분류한 자료(홍종선·최현집, 1999, pp. 22)이며, 이 자료에 대한 피어슨 χ^2 통계량값은 0.1559(p-값=0.693)으로 성격과 운동습관과는 독립적이라고 결론내릴 수 있다. 표 2.1과 표

2.2의 자료에서 관찰칸값에 대응하는 비율을 실선으로 그리고 기대칸값에 대응하는 비율을 점선으로 표현하고, 두 개의 그림을 중복시킨 다이아몬드 그래프를 R을 사용하여 그림 2.1에 각각 나타내었다. 관찰값과 기대값에 대응하는 다이아몬드를 중복시킨 그림에서 실선과 점선이 일치하면 두 변수간 연관이 없는 것으로, 그리고 실선과 점선이 차이를 보이고 있으면 두 변수간에 연관성이 있는 것으로 판단할 수 있다.

표 2.1: 성별과 입학허가에 관한 자료

관찰값(기대값)	남성	여성
입학 허가	1198(1043.46)	1493(1647.54)
입학 불허	557(711.54)	1278(1123.46)

표 2.2: 운동습관과 성격에 관한 자료

관찰값(기대값)	A형	B형
정기적인 운동	438(477.89)	477(482.11)
기타	1101(1106.11)	1121(1115.89)

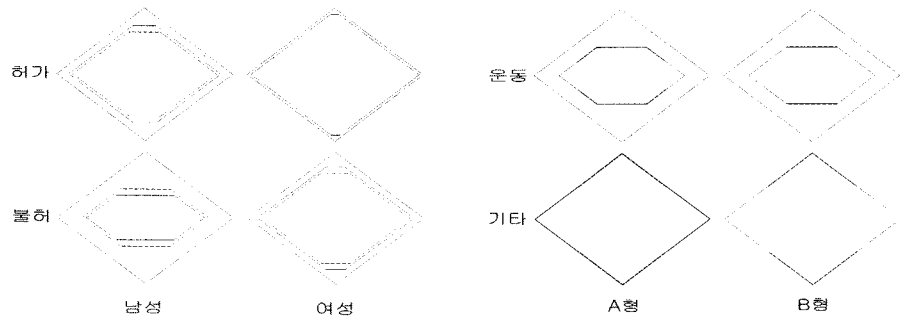

그림 2.1: [표 2.1]과 [표 2.2] 자료의 다이아몬드 그래프

표 2.1과 표 2.2의 자료를 다이아몬드 그래프로 구현한 그림 2.1을 살펴보자. 우선 그림 2.1의 왼쪽 그래프에서는 (2,1)칸의 그림에서 관찰칸값과 기대칸값에 대응하는 실선과 점선의 다이아몬드 모양의 차이가 많이 존재하고, (1,1)칸과 (2,2)칸에서도 적지 않은 차이를 발견할 수 있다. 이와는 반대로 그림 2.1의 오른쪽 그래프에서는 모든 칸에서 점선과 실선 사이의 차이를 발견할 수 없다. 따라서 그림 2.1의 왼쪽 그래프를 통해서는 성별에 따른 입학허가 여부는 연관이 있음을 재확인할 수 있으며, 오른쪽 다이아몬드 그래프에서는 성격과 운동습관과는 연관이 없다고 판단할 수 있다. 이러한 판단은 피어슨 χ^2 통계량을 사용한 독립성 검정결과와 동일하다.

따라서 본 논문에서 제안한 다이아몬드 그래프는 식 (2.1)과 (2.2)에서 정의한 비율 p_{ij}와 $\hat{p}_{ij}$에 비례하는 다이아몬드를 각각 실선과 점선으로 구현하고, 이 다이아몬드들을 중복시켜 그래프를 작성한다. 중복된 다이아몬드 모양의 차이가 많이 발생하지 않으면 두 변수가 독립적이라고 판단내릴 수 있다. 물론 이와 같은 시각적 판단은 상대적이고 주관적이지만,

많은 경험을 가진다면 탐색적 자료분석이 가능하겠다. 또한 본 논문에서 제안한 다이아몬드 그래프를 객관적인 적합도 검정방법들과 병행하여 사용한다면, 자료를 분석하고 이해하는데 많은 도움이 된다.

2.2. 삼차원 자료에서 모형 선택

표 2.3의 삼차원 분할표 자료(홍종선·최현집, 1999, pp. 111)는 스웨덴 교통부에서 속도제한이 교통사고 사망률에 미치는 영향을 분석하기 위해, 주도로보다 차로가 적은 간선도로에서 속도제한(90km/h)을 하였을 경우와 하지 않은 경우의 사망자 수를 1961년과 1962년에 걸쳐 조사한 자료이다.

표 2.3: 교통사고 자료

		연도(변수 1)			
		1961		1962	
도로형태(변수 3)		주도로	간선도로	주도로	간선도로
속도제한	제한	8	42	11	37
(변수 2)	자유	57	106	45	69

홍종선과 최현집(1999)은 포화모형을 제외한 여덟가지 모형중 [1][23] 모형이 표 2.3 자료에 가장 적합한 모형이라고 하였으며, 이 모형을 포함한 계층적 구조(hierarchical structure)에서 [1][2][3] 모형과 [12][23] 모형을 비교해 보면, 가능도비 검정통계량의 차이는 $G^2([1][2][3]) - G^2([1][23]) = 13.8511 - 3.1320 = 10.7191$($p$-값=0.0011)이며 $G^2([1][23]) - G^2([12][23]) = 3.1320 - 1.3351 = 1.7969$($p$-값=0.1801)이다. 이 결과는 자료에 적합한 [1][23] 모형은 적합하지 못한 [1][2][3] 모형보다 훨씬 좋은 모형이며, 적합한 [12][23] 모형과는 차이가 없다고 해석된다. 이와 같은 검정 결과를 본 논문에서 제안한 다이아몬드 그래프와 비교분석하기 위하여, 표 2.3의 자료를 가장 잘 설명하는 [1][23] 모형뿐만 아니라 [1][2][3]모형과 [12][23] 모형하에서 구한 기대값과 관찰값에 대한 다이아몬드 그래프를 각각 그림 2.2부터 그림 2.4에 나타내었다.

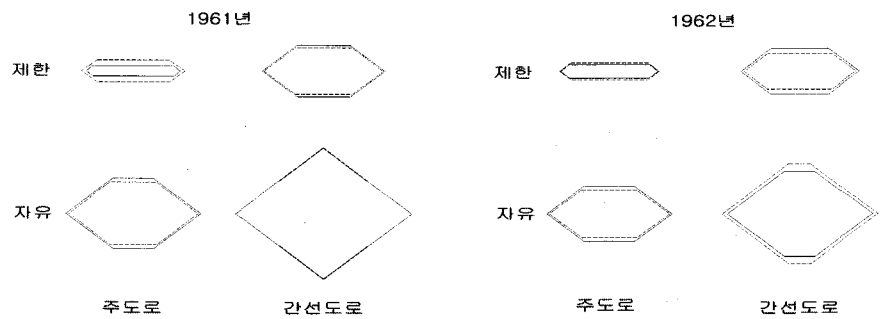

그림 2.2: [1][2][3]모형의 다이아몬드 그래프

그림 2.2를 살펴보면 거의 모든 칸에서 실선과 점선의 차이가 심하며, 그림 2.3과 그림 2.4에서는 실선과 점선의 다이아몬드가 대부분 중복됨을 발견할 수 있다. 또한 그림 2.3과 그

림 2.4에서의 실선과 점선의 차이에 대한 변화는 매우 적다고 판단된다. 그러므로 표 2.3의 자료에 적합한 로그선형모형은 가능도비 검정통계량을 사용한 통계적 분석과 동일하게, [1][23] 모형이라고 탐색적으로 판단할 수 있다.

이와 같이 삼차원 이상의 범주형 자료에서도 다이아몬드 그래프를 이용하면, 다차원 범주형 자료에 적합한 로그선형모형을 탐색적으로 발견할 수 있다.

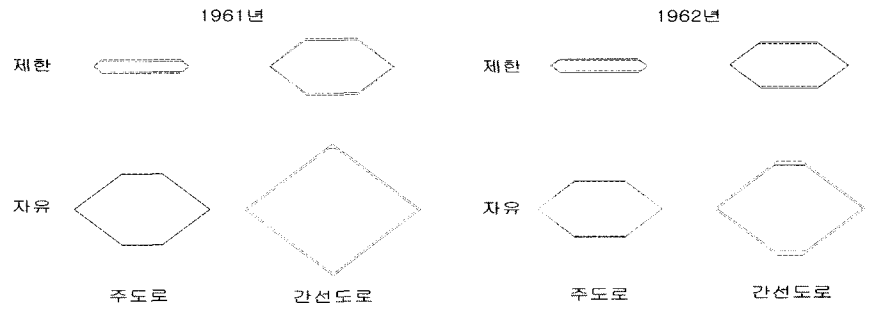

그림 2.3: [1][23]모형의 다이아몬드 그래프

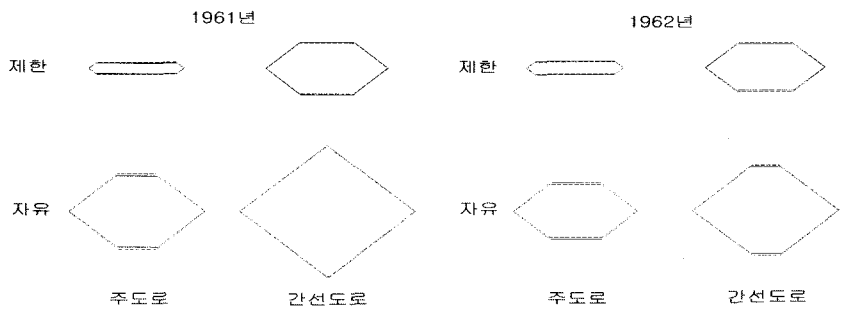

그림 2.4: [12][23]모형의 다이아몬드 그래프

2.3. 육각기둥 모양의 다이아몬드 그래프

이차원 이상의 분할표 자료를 다이아몬드 그래프로 작성할 때, 육각형 모양의 다이아몬드를 보다 시각적으로 표현하기 위하여 육각기둥 모양의 다이아몬드를 고려할 수 있는데, 육각기둥의 윗면(또는 아래 면)의 육각형 모양에서 높이를 $V(p)$, 가운데 수평 길이를 $H(p)$, 윗변(또는 밑변)의 길이를 $h(p)$, 그리고 육각기둥의 높이를 $L(p)$로 다음과 같이 설정하면,

$$\begin{aligned} V(p) &= \sqrt{p} \\ H(p) &= 0.5 + 0.5\sqrt{p} \\ h(p) &= 0.5 - 0.5\sqrt{p} \\ L(p) &= \sqrt{p} \end{aligned}$$

육각기둥의 부피는 다음과 같이 비율 그 자체가 된다.

$$[H(p) + h(p)]\ V(p)\ L(p) = p.$$

위의 방법을 이용하여 표 2.3 자료의 관찰값과 [1][2][3]모형하의 기대값을 중복 표현한 다이아몬드 그래프는 그림 2.5와 같다.

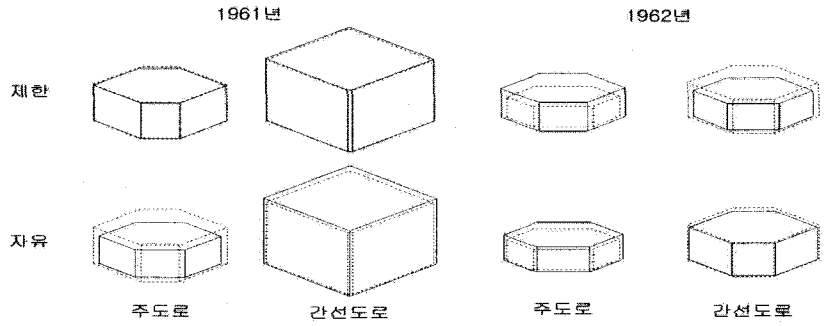

그림 2.5: 육각기둥 모양의 다이아몬드 그래프

육각기둥 모양의 다이아몬드 그래프는, 1절에서 언급한 바와 같이, 일반적인 3-D 그래프에서 발생하는 원근 확인의 어려움과, 육각기둥을 형성하는 선들의 구성요소가 칸비율의 선형 관계식이 아니므로 칸비율과 일차적인 관계를 갖지 않는다는 문제점을 갖고 있다. 그러나 시각적으로 다양한 방법을 제공하고 모형의 적합성 여부를 판단할 수 있다는 장점이 있다.

3. 결론 및 향후 연구

Li 등(2003)이 제안한 다이아몬드 그래프 방법은 이차원 분할표 자료의 관찰칸값에 대응하는 비율을 육각형 모양으로 구현하였는데, 본 연구에서는 관찰칸값에 대응하는 육각형 모양의 다이아몬드와 설정된 귀무가설 모형하의 기대칸값에 대응하는 육각형 모양의 다이아몬드를 중복시켜 표현하는 다이아몬드 그래프를 제안하였다. 제안된 그래프를 이용하여 관찰값과 기대값에 대응하는 다이아몬드의 차이를 살펴봄으로 설정된 귀무가설 모형이 자료를 잘 설명하는지를 판단할 수 있는데, 이차원 범주형 자료에서는 두 범주형 변수의 독립성에 대하여 판단 가능하며, 삼차원 이상의 자료에서는 완전독립모형 외에 부분독립모형, 조건부독립모형, 한 변수의 독립모형등의 로그선형모형들이 자료에 적합한지를 탐색적으로 검정이 가능하여 최적의 모형을 발견할 수 있다. 따라서 적합도 검정을 수행하고 그 결과를 설명할 때, 본 논문에서 제안한 다이아몬드 그래프를 활용하면 자료 해석을 더욱 쉽게 할 수 있다.

참고문헌

홍종선, 최현집(1999). 범주형 자료분석, 자유아카데미.

Becker, R. A., Cleveland, W. S., and Shyu, M. J. (1996). The Visual Design and Control of Trellis Display, *Journal of Computational and Statistical Graphics*, **5**, 123-155.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley, *Science*, **187**, 398-404.

Cleveland, W. S. and McGill, R. (1984). Graphical Perception : Theory, Experimentation, and Application to the Development of Graphical Methods, *Journal of the American Statistical Association*, **79**, 387, 531–554.

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.

Friendly, M. (1991). *The SAS System for Statistical Graphics*, Cary, NC: SAS Institute Inc.

Friendly, M. (1994a). Mosaic Displays for Multi-way Contingency Tables, *Journal of the American Statistical Association*, **89**, 190-200.

Friendly, M. (1994b). SAS/IML Graphics for Fourfold Displays, *Observations*, **3**, 4, 47-56.

Harris, R. L. (1999). *Information Graphics: A Comprehensive Illustrated Reference*, New York: Oxford University Press.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for Contingency Tables, In W. F. Eddy (Ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, New York: Springer-Verlag.

Hartigan, J. A. and Kleiner, B. (1984). A Mosaic of Television Ratings, *The American Statistician*, **38**, 32-35.

Hong, C. S. and Lee, J. C. (2000). Ring Chart II for Multidimensional Categorical Data Analysing using Conditional Ring Charts, *Korean Journal of Applied Statistics*, **13**, 1, 163-178.

Li, X., Buechner, J. M., Tarwater, P. M., and Munoz, A. (2003). A Diamond-Shaped Equiponderant Graphical Display of the Effects of Two Predicts on Continuous Outcomes, *The American Statistician*, **57**, 193-199

Oh, M. G., Hong, C. S., and Lee, J. C. (1999). Ring Chart for Categorical Data, *Korean Journal of Applied Statistics*, **12**, 1, 225-240.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

Wilkinson, L. (1999). *The Grammar of Graphics*, New York: Springer-Verlag.

[2005년 11월 접수, 2006년 2월 채택]

Applications of Diamond Graph

C. S. Hong[1] Y. S. Ko[2]

ABSTRACT

There are lots of two and three dimensional graph representing two dimensional categorical data. Among them, Li, et al. (2003) proposed Diamond Graph that projects three dimensional graph into two dimension whereby the third dimension is replaced with a diamond shape whose area and middle and vertical and horizontal lengths represent the outcome. In this paper, we use the Diamond graph to test the independence of two predictor variables for two dimensional data. And this graph could be applied for finding the best fitted log-linear model to three dimensional data.

Keywords: Goodness-of-fit test, Log-linear model, Independence test.

1) Professor, Department of Statistics, SungKyunKwan University, Seoul, Korea.
E-mail: cshong@skku.ac.kr
2) Research Fellow, Department of Statistics, SungKyunKwan University, Seoul, Korea.