응용통계연구 제 12권 1호, 1999년, pp. 225-239

다차원 범주형 자료에 대한 링차트 *

오민권 [1] 홍종선[2] 이종철[3]

요 약

범주형 자료에 대하여 탐색적 자료분석을 할 수 있는 기존의 여러 그림들은 변수의 수가 많아지면 시각적인 식별이 어렵다는 단점이 있다. 본 논문에서는 삼차원이상의 다차원 범주형 자료를 이차원 평면상에 표현할 수 있는 링차트(ring chart)를 제안한다. 각 칸의 확률값을 표현하는 링차트는 범주형 자료의 구조 전체를 시각적으로 파악할 수 있으며, 관측값을 표준화한 링차트는 변수들간의 연관성 여부를 시각적으로 판단하는데 유용한 정보를 제공한다. 삼차원이상의 자료에서는 이중 링차트(조건부 링차트)를 개발하여 일차 및 이차교호작용 검정까지도 가능하다. 또한, 관측값과 잔차를 동시에 표현한 잔차 링차트는 설정된 모형의 적합성 여부를 시각적으로 평가할 수 있는 장점이 있다.

1. 서론

범주형 자료를 시각적으로 표현하며 탐색적 자료 분석(exploratory data analysis)을 수행할 수 있는 여러 그림들은 자료의 구조를 파악하는데 유용한 정보를 제공한다. 이러한 탐색적 자료 분석의 여러 그림들 중에서 바차트(bar chart), 히스토그램(histogram), 파이차트(pie chart), 스타차트(star chart) 등은 한 범주형 변수에 대한 여러 범주의 확률이나 관측값을 이차원 평면상에 표현하는 대표적인 그림이다. Fienberg(1975)는 2×2 분할표를 각 칸 도수의 크기에 비례하는 반지름을 갖는 1/4크기의 원(quarter circles)으로 나타내는 "four-fold circular display"를 제안하였다. 일반적인 $I\times J$ 이차원 분할표를 시각적으로 표현하는 대표적인 그림으로는 블럭차트(block chart)가 있다. Hartigan과 Kleiner(1981, 1984)는 관측값의 확률을 사각형 또는 타일(tile)의 면적으로 표현한 모자익 그림(mosaic plot)을 제안하였고, Friendly(1992, 1994)는 각 칸에 해당하는 편차의 크기를 고려하여 각 타일에 색깔을 주고, 편차의 부호와 밀도를 빗금으로 표현하는 개선된 모자익 그림을 제안하였다. 또한 Tukey(1977)는 이차원 분할표의 적합을 그림으로 표현하는 "two-way plot"을 제안하였다. $I\times J\times K$인 삼차원 분할표 자료의 경우에는 한 변수의 각 범주에 대하여 재구성한 $I\times J$ 이차원 분할표 자료를 "four-fold circular display"($2\times2\times K$ 인 경우로 확장가능)과 모자익 그림으로 나타낼 수 있으며, 모자익 그림을 확장하여 삼차원뿐만 아니라 사차원 이상의 분할표를 표현할 수 있다(Hartigan and Kleiner (1984)).

* 이 논문은 성균관대학교의 1998년도 성균학술연구비에 의하여 연구되었으며, 1997년도 한국학술진흥재단 자유공모과제 연구비에 의하여 지원되었음.

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 강사
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 교수
3) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 강사

분할표를 시각적으로 표현하는 측면과는 달리 범주형 변수들간의 연관정도를 나타내는 그림으로는 앞서 소개된 모자익 그림과 "four-fold circular display"등이 있다. 또한, Fienberg(1968), Fienberg와 Gilbert(1970)는 이차원으로 구성된 2×2 분할표 자료에 대하여 연관성 측도를 사면체내의 궤적(loci)에 의해 기하학적으로 표현하는 방법을 제안하였다. 이들 그림들과는 달리 Darroch, Lauritzen과 Speed(1980)는 다차원 분할표 자료에 적합하는 독립모형과 조건부 독립모형에 의해 일련의 그림으로 표시되는 그림모형(graphical models)을 제안하였는데, 그림모형은 삼차원 이상의 분할표에서 변수들간의 연관을 선으로 연결하는 연관그림(association graph)으로 나타내어질 수 있다. 연관그림은 초기 모형을 설정하거나 최적의 모형을 찾는데 유용한 정보를 제공한다.

앞에서 살펴본 탐색적 자료 분석의 여러 그림들이나 연관정도를 표현한 그림들은 범주의 수나 변수의 수가 많아지는 경우 시각적인 식별이 어렵고 변수들간의 관계를 구체적으로 파악할 수 없다는 단점이 있다. 본 논문에서는 범주형 자료의 차원과 범주의 수에 상관없이 항상 이차원 평면상에 표현되는 "링차트(ring chart)" 를 제안하고자 한다. 본 논문의 2절에서는 2×2 분할표를 링차트로 나타내는 방법에 대해서 자세히 설명하고 있다. 그리고 관측값을 표준화한 분할표를 표현하는 "표준화된 링차트(standardized ring chart)"에서는 두 변수들간의 연관정도를 교차적비와의 관계를 이용하여 시각적으로 설명할 수 있음을 발견하였다. 3절에서는 링차트를 일반적인 $I \times J$ 이차원 분할표로 확장하여 설명하였다. 4절에서는 삼차원인 $2 \times 2 \times 2$ 분할표에 대한 링차트를 설명하고 일차와 이차 교호작용항과의 관계를 논하였다. 그리고 어떤 한 변수의 두 범주에 대한 "이중 링차트(double ring chart)" 또는 "조건부 링차트(conditional ring chart)"를 제안하여 교호작용항의 검정에 대하여 토론하였다. 5절에서는 링차트를 일반적인 $I \times J \times K$ 삼차원 분할표에 대해서도 확장할 수 있음을 설명하고, 예제를 통해 자료의 구조뿐만 아니라 변수들간의 연관정도를 시각적으로 파악하여 최적의 모형을 유추할 수 있음을 논의하였다. 6절에서는 관측값과 잔차를 동시에 표현할 수 있는 "잔차 링차트(residual ring chart)"를 제안하였다. 이 잔차 링차트를 이용하여 설정된 모형의 적합성 여부를 시각적으로 평가할 수 있음을 논의하였다. 2절과 3절에서는 가상의 예제로 설명하였으나 4절부터 6절까지는 실제로 연구된 예제를 갖고 논의하였다. 여기서 제안된 링차트를 표현하는 프로그램은 JAVA로 구현되었으며 MS Internet Explorer(4.0 버젼이상)을 사용하여 http://stat.skku.ac.kr/~omg0906/Ring.html을 접속하면 이용할수 있다.

2. 2×2 분할표의 링차트

2×2 이차원 분할표에서 각 칸의 확률 $p_{ij}(i = 1, 2,\ j = 1, 2)$에 대해서 첫 번째 변수와 두 번째 변수의 주변합의 확률을 각각 p_{i+}와 p_{+j}라 하면 다음과 같이 정의된다.

$$p_{i+} = \sum_j p_{ij}\ ,\ p_{+j} = \sum_i p_{ij}\ .$$

분할표의 각 칸의 확률 p_{ij}와 주변합의 확률 p_{i+}, p_{+j}에 대응되는 각도 θ_{ij}와 θ_{i+}, θ_{+j}는 각 확률에 2π를 곱하여 구할 수 있다. 예를 들어 $\theta_{ij} = p_{ij} \times 2\pi$ 이다. 2×2 분할표에서 행

변수에 대한 주변합의 확률 p_{i+}에 대응되는 각도 θ_{i+}를 이용하여 우선 파이차트로 나타내고 파이차트의 외곽에 확률 p_{ij}에 대응하는 각도 θ_{ij}들을 $\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$ 순서로 색깔 있는 링(ring)을 만들어 나타낸 그림을 고려하자. 예를 들어 표 2.1과 같은 분할표 자료를 이와 같은 방법으로 표현한 결과가 그림 2.1과 같으며 이를 "링차트"라 하자.

표 2.1: 2×2 분할표

		변수 B	
		B_1	B_2
변수 A	A_1	0.28	0.12
	A_2	0.42	0.18

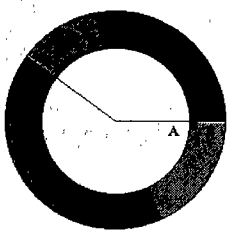

그림 2.1: 링차트

그림 2.1은 표 2.1의 분할표에 대하여 변수 A의 주변합의 확률 0.4와 0.6에 대응되는 각도를 가운데 흰 부분에서 파이차트로 나타내고 각 칸의 확률 0.28, 0.12, 0.42, 0.18에 대응되는 각도를 외곽 링에 색깔을 주어 표현한 링차트이다. 그림 2.1의 링차트에서 각각 변수 A의 주변합에 대한 확률들의 크기를 시각적으로 비교할 수 있을 뿐만 아니라 외곽의 색깔이 있는 부분에서는 변수 A의 주변합의 확률에 대한 변수 B의 각 범주의 확률의 크기를 비교할 수 있다. 따라서 링차트는 이차원 분할표의 구조를 이차원 평면상에서 각 칸의 확률 p_{ij}를 주변합 p_{i+}와 같이 시각적으로 한꺼번에 파악할 수 있는 장점이 있다.

그림 2.1의 링차트에서 각각의 주변합의 확률 p_{i+}에 대해 칸확률 p_{ij}의 크기가 어느 정도인지를 파악하기 위하여 주변합의 확률을 같게 해주는 표준화(standardized)를 고려하자. 여기서 표준화란 특정 변수에 대해 주변합의 비율을 같게 한다는 의미이며, Fienberg(1980)가 표준화된 2×2 분할표를 "four-fold circlar display"에 적용할 때 처음 사용하였다(du Toit, Steyn과 Stumpf(1986) 참조). 예를 들어 2×2 분할표를 표준화하는 경우 주변합의 확률은 범주의 수가 2개이므로 0.5이다. 그리고 변수 A에 대해 표준화된 분할표의 각 칸의 확률을 q_{Aij}라 하면 다음과 같이 정의할 수 있다.

$$q_{Aij} = \frac{p_{ij}}{p_{i+}} \times \frac{1}{I} . \tag{2.1}$$

표 2.2는 표 2.1의 분할표를 변수 A에 대하여 표준화된 분할표이고, 그림 2.2는 표 2.2의 분할표를 표현한 표준화된 링차트이다. 그림 2.2에서는 a와 b가 원점 o와 연결하는 분할선

들이 일직선을 나타내고 수평선과 교차함을 주의해서 보아야하며 그 이론적 배경은 변수 B를 먼저 고려하여 작성한 링차트를 설명한 뒤에 설명하고자 한다.

표 2.2: 변수 A를 표준화한 2×2 분할표

		변수B		
		B_1	B_2	주변합
변수A	A_1	0.35	0.15	0.5
	A_2	0.35	0.15	0.5

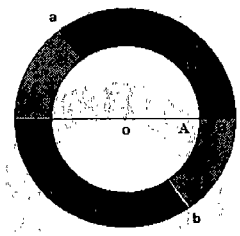

그림 2.2: 변수 A에 대해 표준화된 링차트

변수 B를 먼저 고려한 링차트는 2×2 분할표에서 열 변수에 대한 주변합의 확률 p_{+j} 에 대하여 우선 파이차트로 나타내고 파이차트의 외곽에 확률 p_{ij}에 대응하는 각도 θ_{ij}들을 외곽 링에 색깔을 주며 $\theta_{11}, \theta_{21}, \theta_{12}, \theta_{22}$ 순서로 나타낸 그림이다. 그림 2.3의 왼쪽 링차트는 표 2.1의 분할표에 대하여 변수 B의 주변합의 확률 0.7과 0.3에 대응되는 각도를 가운데 파이차트로 나타내고 각 칸의 확률 0.28, 0.42, 0.12, 0.18에 대응되는 각도를 외곽에 표현하였다. 그림 2.3의 왼쪽 링차트에서는, 변수 B의 주변합의 확률들과 외곽의 색깔이 있는 링에 표현된 확률 p_{ij}들의 크기를 상대적으로 비교할 수 있다. 이제 그림 2.3의 왼쪽 링차트에서 변수 B의 주변 확률 p_{+j}에 대해 표준화를 고려하자. 변수 B에 대해 표준화된 분할표의 각 칸의 확률 q_{Bij}는 (2.1)식과 같은 형태로 정의할 수 있다.

$$q_{Bij} = \frac{p_{ij}}{p_{+j}} \times \frac{1}{J}.$$

표 2.3은 변수 B에 대하여 표준화된 분할표이며 그림 2.3의 오른쪽 링차트는 변수 B에 대하여 표준화된 링차트이다.

그림 2.2와 그림 2.3의 표준화된 링차트를 살펴보면, a와 b, 원점 o를 연결하는 분할선이 직선으로 나타나고, 수평선과 교차함의 이유를 설명하기 위하여 우선 표 2.1과 같은 이차원 분할표에 대해서 다음과 같은 대수선형모형을 정의하여 보자.

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)}, \tag{2.2}$$

표 2.3: 변수 B를 표준화한 2×2 분할표

		변수B	
		B_1	B_2
변수A	A_1	0.2	0.2
	A_2	0.3	0.3
주변합		0.5	0.5

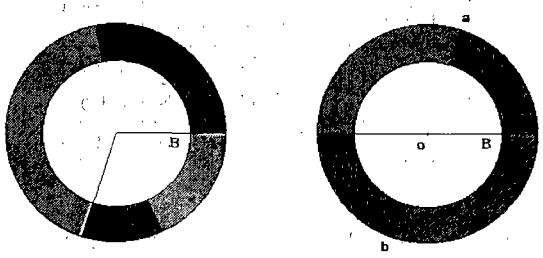

그림 2.3: 변수 B에 대한 링차트와 표준화된 링차트

여기서 μ는 총 평균이고, $\mu_{1(i)}$와 $\mu_{2(j)}$는 변수 A와 변수 B의 주효과이고 $\mu_{12(ij)}$는 일차 교호작용항으로 다음과 같이 정의함을 알 수 있다(홍종선(1995) 참조).

$$\mu_{12(11)} = \frac{1}{4} \log \left(\frac{p_{11}p_{22}}{p_{12}p_{21}} \right). \tag{2.3}$$

(2.3)식의 오른쪽항에 있는 교차적비(cross-product ratio, odds ratio)를 표준화된 표 2.2와 표 2.3의 분할표에 적용하여 살펴보면 다음과 같은 등식이 성립함을 발견할 수 있다.

$$\begin{aligned} \frac{p_{11}p_{22}}{p_{12}p_{21}} &= \frac{q_{A11}q_{A22}}{q_{A12}q_{A21}} \\ &= \frac{q_{B11}q_{B22}}{q_{B12}q_{B21}}. \end{aligned} \tag{2.4}$$

만일 표 2.1, 표 2.2, 표 2.3의 분할표에서 교차적비가 1이라면 변수 A와 변수 B는 독립이며, 일차 교호작용항 $\mu_{12(ij)}$은 0을 나타낸다. 그리고 독립인 경우의 표준화된 링차트는 상단부의 반원을 구분하는 분할선 $\overline{ao}$와 하단부의 반원을 구분하는 분할선 $\overline{bo}$를 연결하면 일직선으로 나타날 것이다. 즉 표준화된 링차트에서 확률 q_{Aij}와 q_{Bij}들을 구분하는 선들이 일직선으로 수평선과 서로 'X'자로 교차한다면, (2.4)식의 관계로 인하여 두 변수들이 독립적인 관계를 갖고 있다고 판단할 수 있는 사실을 발견하였다. 그림 2.3에 나타난 링차트는 열변수를 먼저 고려할 수 있음을 보여주었으며 특히 표준화된 링차트는 어느 변수를 먼저 고려하여도 동일한 정보를 제공함을 알수 있다.

3. $I \times J$ 분할표의 링차트

링차트와 표준화된 링차트는 일반적인 $I \times J$ 이차원 분할표에 대해서 쉽게 확장할 수 있다.

표 3.1: 4×3 분할표

		변수B		
		B_1	B_2	B_3
변수A	A_1	0.08	0.12	0.20
	A_2	0.04	0.06	0.10
	A_3	0.02	0.03	0.05
	A_4	0.06	0.09	0.15

예를 들어 표 3.1과 같은 4×3인 이차원 분할표를 고려하자. 그림 3.1의 왼쪽 그림은 표 3.1의 분할표에서 변수 A의 주변합을 우선 고려하여 나타낸 링차트이고, 오른쪽 그림은 변수 A에 대해서 표준화된 링차트이다. 오른쪽 그림에서 p_{A11}, p_{A21}, p_{A31}, p_{A41}은 동일한 확률값을 갖고 있고 따라서 분할선인 $\overline{a_1o}$, $\overline{a_2o}$, $\overline{a_3o}$, $\overline{a_4o}$ 가 동일한 각도인 90도 각도로 벌어져있음을 알수 있다. $\overline{b_1o}$, $\overline{b_2o}$, $\overline{b_3o}$, $\overline{b_4o}$의 분할선도 직각인 관계라는 사실은 변수 A와 변수 B가 독립이라고 결론내릴수 있다. 그러므로 범주의 수준수가 I개일때 i번째와 $(i+1)$번째의 분할선(예를들어, $\overline{a_1o}$와 $\overline{a_2o}$)이 이루는 각도가 $360/I$도로 표현되면 변수는 독립적이라고 가정한다. 그리고 그림 3.1의 오른쪽 그림인 표준화된 링차트에서 각 표준화된 칸 확률 q_{Aij} $(i = 1, \ldots, I,\ j = 1, \ldots, J)$가 다른 모든 $q_{Ai'j}$ $(i' \neq i)$와 동일한 크기를 갖고 있다면 변수 A와 B는 독립적인 관계를 갖고 있음을 안다.

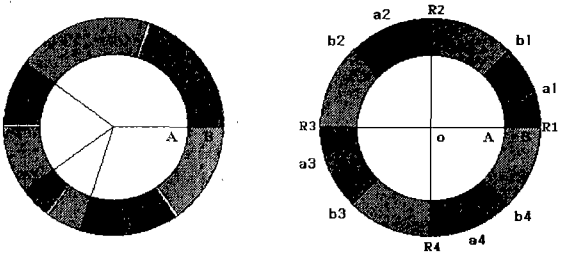

그림 3.1: 이차원 자료의 링차트

일반적인 $I \times J$ 분할표에서 두 변수의 독립성을 주장하기 어려울 때에는 여러 종류의 연관성 측도(association measure)를 링차트 주변에 나타내주어 자료분석에 도움을 줄 수 있다(저자들이 구현한 프로그램 참조). 순위변수(ordinal variable)로 구성된 분할표 자료는 명목변수(nominal variable)일때와 동일한 방법으로 링차트를 표현할수 있다. 다만 명목변수일때는 그 변수의 범주를 임의대로 바꾸어 표현하여도 무관하나 순위변수일때

는 그 범주의 순서를 유지하며 표현하여야 한다. 표 3.1에서 두 변수가 모두 순위변수라고 가정하고, $A_1 \prec A_2 \prec A_3 \prec A_4$ 그리고 $B_1 \prec B_2 \prec B_3 \prec B_4$의 순서성이 있을때 두 변수간의 양의 연관성은 변수 A의 각 범주내에서 변수 B의 조건부 확률들이 통계적 서열(stochastic ordering)을 만족할 때 발생한다. 즉 그림 3.1의 오른쪽 표준화된 링차트에서 $\angle R_1oa_1 \geq \angle R_2oa_2 \geq \angle R_3oa_3 \geq \angle R_4oa_4$ 그리고 $\angle R_1ob_1 \geq \angle R_2ob_2 \geq \angle R_3ob_3 \geq \angle R_4ob_4$가 동시에 만족되면(적어도 하나 이상의 완전 부등관계 필요), 두 변수사이에는 양의 연관성이 존재한다고 판단할 수 있다. 또한 모든 부등호가 반대로 ($\leq$)으로 되어있으면 음의 연관성이 있다고 판단할 수 있으며, 모든 부등호대신 등호(=)로 이루어진다면 두 변수는 서로 독립이라고 위에서와 동일한 논지로 결론 지을 수 있다.

4. $2 \times 2 \times 2$ 인 삼차원 분할표

표 4.1는 스웨덴 교통부에서 속도제한이 교통사고 사망률에 미치는 효과를 분석하기 위해 조사한 자료이다(Andersen(1991), p158). 표 4.1와 같은 $2 \times 2 \times 2$ 분할표인 경우 어느 한 변수의 두 범주에 대해 새로운 두 개의 2×2 분할표를 고려하여 2절에서 언급한 링차트를 그대로 적용하여 두 개의 링차트를 동시에 살펴봄으로 나머지 두 변수의 독립여부를 시각적으로 판단할 수 있다.

표 4.1: $2 \times 2 \times 2$ 분할표

		변수B	변수C	
			주도로	이차도로
변수A	1961	제한	8	45
		자유	57	106
	1962	제한	11	37
		자유	45	69

변수 A, B, C에 대한 $2 \times 2 \times 2$분할표에서 변수 C의 첫 번째 범주에 대해서 재구성한 2×2 분할표에 대한 교차적비를 $\alpha^{(1)}$, 두 번째 범주에 대해서 재구성한 2×2 분할표에 대한 교차적비를 $\alpha^{(2)}$라 할 때, 칸확률 p_{ijk} 와 변수 A에 대하여 표준화된 확률 q_{Aijk}로 표현되는 교차적비는 다음과 같이 정의된다(변수 B에 대하여 표준화된 확률 q_{Bijk}로 표현된 정의는 생략함).

$$\begin{aligned} \alpha^{(1)} &= \frac{p_{111}p_{221}}{p_{121}p_{211}} \\ &= \frac{q_{A111}q_{A221}}{q_{A121}q_{A211}}, \end{aligned}$$

$$\begin{aligned}\alpha^{(2)} &= \frac{p_{112}p_{222}}{p_{122}p_{212}} \\ &= \frac{q_{A112}q_{A222}}{q_{A122}q_{A212}}.\end{aligned}$$

대수선형모형의 일차교호작용항 $\mu_{12(11)}$와 이차교호작용항 $\mu_{123(111)}$은 다음과 같이 교차적비 $\alpha^{(1)}$와 $\alpha^{(2)}$의 함수로 정의할 수 있다(홍종선(1995) 참조).

$$\begin{aligned}\mu_{12(11)} &= \frac{1}{8}\log(\alpha^{(1)}\alpha^{(2)}), \\ \mu_{123(111)} &= \frac{1}{8}\log(\frac{\alpha^{(1)}}{\alpha^{(2)}}).\end{aligned}$$

만일 $\alpha^{(1)}$와 $\alpha^{(2)}$의 값이 역수 관계를 갖는다면, 대수선형모형에 일차교호작용항 $\mu_{12(11)}$이 포함되지 않으며 변수 A와 변수 B가 독립이라고 한다. 또한, $\alpha^{(1)}$와 $\alpha^{(2)}$의 값이 같다면 대수선형모형에 이차 교호작용항 $\mu_{123(111)}$이 포함되지 않을 것이다. 이와 같은 관계를 효과적으로 나타내기 위해 $\alpha^{(1)}$와 $\alpha^{(2)}$에 해당하는 두 개의 분할표(변수 A의 주변합을 기준으로 그리고 변수 A에 대하여 표준화시킨 분할표)에 대해서 크기가 서로 다른 두 개의 링차트를 각각을 겹쳐 그림 4.1과 같이 작성하였다. 우리는 이 차트를 "이중 링차트(double ring chart)" 또는 "조건부 링차트(conditional ring chart)"라 한다.

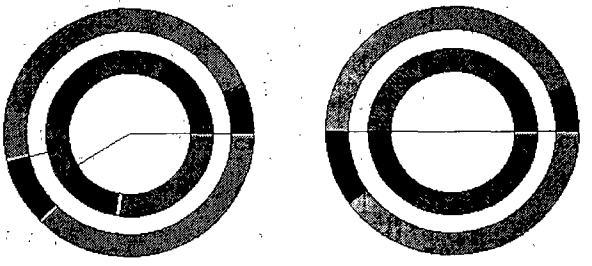

그림 4.1: 이중 링차트 (조건부 링차트)

그림 4.1의 왼쪽 이중 링차트에서 외곽의 링은 표 4.1의 분할표에서 변수 C의 첫 번째 범주(주도로)에 대해 재구성한 2×2 분할표를, 변수 A의 주변합을 기준으로, 나타낸 것이다. 안에 있는 작은 링은 두 번째 범주(이차도로)에 대해 새롭게 구성한 2×2분할표를, 변수 A의 주변합을 기준으로, 나타낸 링차트이다. 또한, 오른쪽 그림은 두 개의 2×2분할표에서 변수 A의 주변합을 표준화하여 그린 이중 링차트이다. 그림 4.1의 오른쪽 두 개의 링차트에서 변수 C에 의해 재구성된 분할표에서 변수 A와 변수 B를 구분하는 두 개의 분할선이 각각 일직선으로 나타나고 있으므로, 변수 A와 변수 B가 독립적임을 인지할 수 있을 뿐만 아니라 변수 A, B, C에 대한 이차교호작용은 존재하지 않음을 시각적으로 파악할 수

있다. 그러므로 이중 링차트에서는 선정된 두 개의 변수에 대한 일차교호작용과 세 변수에 대한 이차교호작용의 존재함까지도 시각적으로 판단할 수 있다.

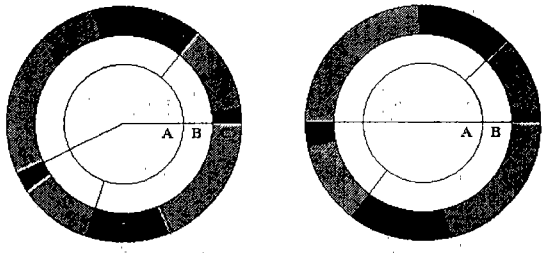

그림 4.2: 삼차원 자료의 링차트 (1)

삼차원인 $2 \times 2 \times 2$ 분할표를 그림 2.1과 같이 하나의 링으로 구성된 링차트로 표현하여보자. 그림 4.2는 삼차원 분할표에서 8개의 칸확률 p_{ijk}를 변수 A와 변수 B 그리고 변수 C에 대하여 차례로 나열하여 $p_{111}, p_{112}, p_{121}, p_{122}, p_{211}, p_{212}, p_{221}, p_{222}$ 의 순으로 외곽링에 적절한 색을 주어 나타내고, 제일 안쪽에는 변수 A와 변수 B로만 구성된 분할표의 자료를 그림 2.1과 같은 링차트로 구성하여 색깔없이 표현하였다. 즉 링 안쪽 그림은 변수 A의 주변확률 p_{i++}와 그 주변확률에서 변수 B에 대한 확률 p_{ij+}들을 설명하는 선들로 나타내어 그림 4.2의 왼쪽과 같이 표현하였다. 그림 4.2의 왼쪽 링차트는 표 4.1의 분할표를 변수 A의 주변합을 기준으로 변수 B와 변수 C의 순서로 각 확률들을 나타낸 링차트이다. 오른쪽의 링차트는 변수 A의 주변합에 대하여만 표준화하여 변수 B와 변수 C의 순서로 표현한 표준화된 링차트이다. 그림 4.2의 오른쪽 링차트에서 링 안쪽 그림은 2절에서 언급한 바와 같이 변수 A와 변수 B로만 구성된 2×2분할표의 표준화된 링차트와 같은 방법으로 해석하면 된다. 즉 그림 2.3과 같은 표준화된 링차트에서 링의 색깔을 제거한 모양이다. 그림 4.2의 오른쪽 링차트에서 우리는 변수 A와 변수 B는 독립에 가깝다는 사실을 시각적으로 파악할 수 있다. 또한 그림 4.2와 같은 링차트에서 변수의 종류와 순서를 바꾸어 가면서 그림 4.3과 그림 4.4와 같이 링차트를 작성해보면 다른 두 변수들간의 독립성 여부를 파악할 수 있기 때문에 모형을 설정하는데 있어서 어떤 모수가 모형에 포함되어야 하는지를 알 수 있을 것이다. 그림 4.3의 경우 변수 B와 변수 C는 독립이 아님을 파악할수 있고, 그림 4.4를 살펴보면 변수 C와 변수 A가 독립임을 알 수 있다. 그러므로 그림 4.2부터 그림 4.4를 통해서 변수 B와 변수 C 만이 연관을 갖고 있다는 것을 볼 수 있다. 참고로 표 4.1의 자료를 가장 잘 적합하는 대수선형모형은 [A][BC]로 알려져 있다($G^2 = 3.13$, p-값 $= 0.3717$).

5. $I \times J \times K$ 분할표와 고차원 링차트

링차트는 일반적인 $I \times J \times K$ 삼차원 자료뿐만 아니라 그 이상의 고차원 자료도 쉽게 확장할 수 있으나, 여기에서는 $2 \times 2 \times 3$ 인 삼차원 자료에 대하여 다루어보자. 표 5.1은 50-60대 노인들을 대상으로 성별, 결혼상태 그리고 주거형태에 대해 조사한 자료이다(Andersen(1991),

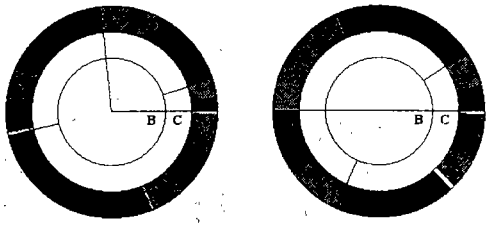

그림 4.3: 삼차원 자료의 링차트 (2)

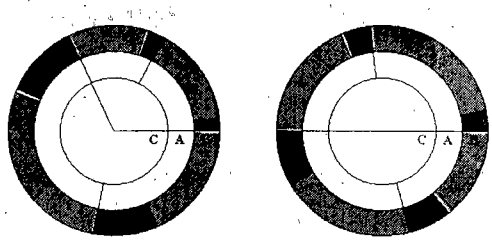

그림 4.4: 삼차원 자료의 링차트 (3)

p176). 표 5.1에 나타난 분할표 자료를 그림 4.2과 같이 동일한 방법으로 링차트를 작성하여 그림 5.1에 표현하였다.

표 5.1: 2 × 2 × 3 분할표

성별(A)	결혼상태(B)	주거형태(C)		
		아파트	주택	농장
남자	미혼	30	32	5
	결혼	46	229	14
여자	미혼	68	41	5
	결혼	76	193	44

그림 5.1의 왼쪽 그림은 표 5.1의 자료에 대해서 변수 A, 변수 B, 변수 C의 순서로 표현한 링차트이다. 오른쪽 그림은 변수 A에 대해 표준화된 링차트이며 이들 두 개의 링차트를 통해 각 변수의 주변합의 확률의 크기를 비교할 수 있으며 특히, 오른쪽 표준화된 링차트에서는 변수 A와 변수 B는 독립임을 시각적으로 알 수 있으므로 이 자료를 가장 잘 적합시키는 최적의 모형 [AC][BC](G^2=6.52, p-값=0.0888)을 유추하는데 유용한 정보를 얻을 수 있다.

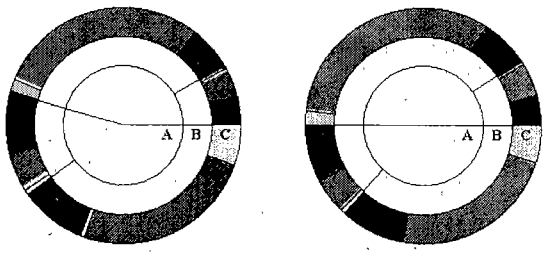

그림 5.1: 링차트

6. 잔차 링차트

앞에서 제안되고 설명한 링차트를 응용하여, 회귀분석의 잔차분석과 같이 설정된 대수선형모형을 적용하여 구한 각 칸의 잔차를 링차트 상에 표현할 수 있다. 링차트의 제일 외곽 링에서 색을 제거하고, 각 칸에 해당되는 양의 잔차나 음의 잔차를 다른 색(여기서는 적색과 청색)으로 그 크기만큼을 표현하여 그림 6.1와 같이 나타내었는데 우리는 이를 "잔차 링차트(residual ring chart)라고 하자. 다음의 $3 \times 2 \times 2 \times 2$ 사차원 분할표 자료인 표 6.1은 Ries와 Smith(1963)에 의해 수집되었고 Cox와 Lauh(1967), Ku와 Kullback (1968), Goodman(1971a) 그리고 홍종선(1995) 등 많은 학자들에 의해 분석된 세제상표 선호도 자료이다. 이 자료에 대한 최적의 모형식은 [AD][BC][BD]으로 잘 알려져 있으며, 이 모형에 대한 검정통계량 G^2은 11.88이고 p-값은 0.6154이다. 반면에 독립모형 [A][B][C][D]에 대한 검정통계량 G^2은 42.92이고 p-값은 0.0008이며, 두 모형에 대한 잔차 링차트를 그림 6.1에 작성하였다.

표 6.1: 세제선호조사자료에 대한 분할표

변수 A	변수 B	변수 C			
		예		아니오	
		변수 D			
		고	저	고	저
부드러움	X	19	57	29	63
	M	29	49	27	53
중간	X	23	47	33	66
	M	47	55	23	50
거침	X	24	37	42	68
	M	43	52	30	42

그림 6.1에 있는 두 개의 잔차 링차트는 앞 절에서 설명한 링차트에서 색깔이 있었던 외곽 링의 색을 제거하고 각 칸에 양(+)의 값을 갖는 잔차는 적색으로, 음(-)의 값을 갖는 잔

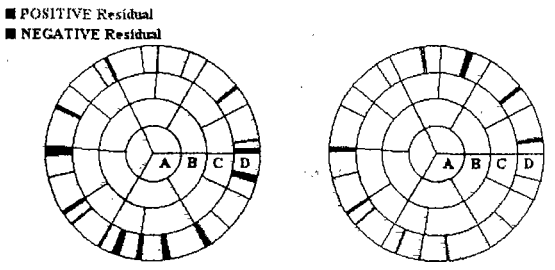

그림 6.1: 모형 [A][B][C][D]와 [AD][BC][BD]에 대한 잔차 링차트

차는 청색으로 잔차의 크기만큼을 표현한 그림이다. 잔차 링차트를 이용하여 외곽 링에 색이 칠해져 있는 부분이 많고 적음에 따라 임의로 선정된 모형이 자료에 어느 정도 적합한지를 알 수 있으며, 여러 모형들을 적용시켜서 살펴보면 최적의 모형을 발견할 수 있다. 그리고 각 모형의 잔차에 대한 링차트를 통하여 어떤 변수의 어느 범주에서 잔차가 크게 나타나는지를 파악하여 구조적인 측면에서 자료를 분석할 수 있다.

7. 결론

본 논문에서는 모든 다차원 범주형 자료를 이차원 평면상에 시각적으로 표현해주는 링차트를 제안하였는데 이는 변수의 수가 많아도 식별이 가능하다. 그리고 2×2 분할표 자료에 대한 링차트를 이용하여 각 변수의 주변확률과 각 칸확률값 모두의 크기를 시각적으로 비교할 수 있으며, 또한 표준화된 링차트는 두 변수간의 연관정도를 파악하는데 유용한 정보를 제공한다. 삼차원 분할표 자료에 대해서는 한 변수의 각 범주에 따라 나머지 두 변수로 구성된 새로운 분할표를 재구성하여 이중 링차트를 작성할 수 있는데 이를 통하여는 대수선형모형의 일차교호작용과 이차교호작용을 설명할 수 있다. 그리고 세 변수들 중에서 미리 선정된 두 변수의 설명이 첨가된 링차트를 작성하여 두 변수의 연관성 여부를 쉽게 파악할 수 있다. 다차원 분할표 자료에 대해서는, 임의의 대수선형모형에 적용하여 각 칸의 잔차를 나타내어 잔차의 크기와 부호를 링차트에 표현한 잔차 링차트는 모형의 적합성 여부를 시각적으로 평가할 수 있다는 장점이 있다. 이러한 링차트를 활용하면, 자료의 구조를 전반적으로 쉽게 파악할 수 있으며, 변수들의 연관성 정도도 검토할 수 있다. 나아가 링차트는 자료에 적합하는 최적의 대수선형모형을 발견하는 방법으로 활용할 수 있다. 다차원 범주형 자료에서 최적의 모형을 찾는 방법은 많이 있는데 그 중에서도 최현집과 홍종선(1995)이 제안한 다면체 그림을 이용하여 최적의 모형을 찾는 방법과 링차트에서 최적의 모형을 선택하는 방법을 연결하고자 한다. 임의의 계층구조하의 모형을 적용하여 링차트와 표준화된 링차트 그리고 잔차분석을 통한 잔차 링차트를 표현하는 시스템을 개발하였다. Explorer(version 4.0 이상)을 이용하여 http://stat.skku.ac.kr/~omg0906/Ring.html을 접속하면 논문에서 제안한 링차트와 표준화된 링차트 그리고 잔차 링차트를 구현할 수 있다.

감사의 글

가치있는 조언을 해주신 익명의 심사위원께 감사드립니다.

참고문헌

[1] 홍종선 (1995). 〈 대수선형모형 〉. 자유아카데미.

[2] 최현집, 홍종선(1995). Graphical Descriptions for Hierarchical Log Linear Models. 〈 한국통계학회 논문집 〉, 제2권 2호, 310-319.

[3] Andersen, E. B. (1991). *The Statistical Analysis of Categorical Data,* Springer-Verlag, New York Inc.

[4] Cox, D. R. and Lauh, E. (1967). A note on the graphical analysis of multidimesional contigency tables, *Technometrics,* Vol. 9, 481-488.

[5] Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables, *Ann. of Statist,* Vol. 8, 522-539.

[6] du Toit, S. H. C., Steyn, A. G. W. and Stumpf, R. H. (1986). *Graphical Exploratory Data Analysis,* Springer-Verlag, New York Inc.

[7] Fienberg, S. E. (1968). *The Estimation of Exponential Probabilities in Two-way Contigency Tables,* Ph. D. Thesis, Department of Statistics, Harvard University.

[8] Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a 2×2 contingency tables, *Journal of the American Statistical Association,* Vol. 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,* 4th 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,* Vol. 89, 190-200.

[13] Goodman, L. A. (1971). The analysis of multidimensional contigency tables : stepwise procedures and direct estimation methods for building models for multiple classification, *Technometrics,* Vol. 13, 33-61.

[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,* Vol. 38, 32-35.

[16] Ku, H. H. and Kullback, S. (1968). Loglinear models in contingency table analysis, *The American Statistician,* Vol. 28, 115-122.

[17] Ries, P. N. and Smith, H. (1963). The use of chi-square for preference testing in multidimensional problems, *Chemical Engineering Progress,* Vol. 59, 39-43.

[18] Tukey, J. W. (1977). *Exploratory Data Analysis,* Addison-Wesley Publishing Company.

[1998년 8월 접수, 1998년 11월 최종수정]

Ring Chart for Categorical Data

Min Gweon OH [1] Chong Sun HONG [2] Jong Cheol LEE [3]

ABSTRACT

The current various graphics which can do exploratory data analysis provide some difficulties in representing and evaluating categorical data visually when the number of variables increases. In this paper we propose a "ring chart" which can represent multidimensional categorical data on a two dimensional plane. The "ring chart", which represents probabilities of all cells, can explore the whole structure of the categorical data, and the "standardized ring chart", which represents standardized frequencies, provides useful information in determining the relationship among variables. For three dimensional data, it is possible to test the first and second-order interactions by developing a "double ring chart." Also, a "residual ring chart", which overlaps residuals on observed values, has some advantages in evaluating the goodness of fit of a given log-linear model.

1) Lecturer, Department of Statistic, SungKyunKwan University, Seoul, 110-745, Korea.
2) Professor, Department of Statistic, SungKyunKwan University, Seoul, 110-745, Korea.
3) Lecturer, Department of Statistic, SungKyunKwan University, Seoul, 110-745, Korea.