*Journal of the Korean*
*Data & Information Science Society*
*1999, Vol. 10, No. 2, pp. 457 ~ 464*

# 2 x 2 분할표에서 동적 그래픽스로 구현된 겹쳐진 모자익 그림을 이용한 범주형 자료의 연관성 측정

윤 여 창[1] · 오 민 권[2]

**요약**

Hartigan과 Kleiner(1981)는 분할표 자료에서 주변합의 비율에 대한 각 칸의 관측도수의 비율을 사각형의 면적으로 표현한 모자익 그림을 제안하였는데, 본 연구에서는 2 x 2 분할표에서 관측도수와 기대도수에 대한 두 개의 모자익 그림을 겹쳐서 나타낸 겹쳐진 모자익 그림을 이용한 범주형 자료의 연관성을 측정하고자 한다. 동적 그래픽스기법으로 개선시킨 겹쳐진 모자익 그림을 이용하면 범주형 변수의 연관성을 시각적으로 쉽게 파악할 수 있는데, 이러한 그림은 자료분석이나 통계 패키지에서 제공되고 있지 않다. 겹쳐진 모자익 그림은 변수들의 종속성 여부, 관측도수와 기대도수의 차이등을 제시된 통계량과 함께 시각적으로 파악할 수 있기 때문에 모형 설정시 매우 유용한 정보를 얻을 수 있다.

주제어: 겹쳐진 모자익 그림, 동적그래픽스, 연관성 측정.

## 1. 서 론

실험이나 설문지에 의해 수집되는 자료들 중에서 여러 범주형 변수에 대한 빈도수가 상호 교차 분류된 분할표(contingency table) 형태로 얻어진 자료를 범주형(categorical) 자료라고 한다. 범주형 자료에서 변수들에 대한 연관성(association) 측정은 연관성 측도들을 이용하는 방법과 변수들의 연관관계를 나타내는 그림을 이용하여 연관성 여부를 시각적으로 파악하는 방법이 있다.

범주형 자료에서 변수들이 갖는 연관성을 측정할 수 있는 통계량으로는 Pearson의 $X^2$, $G^2$, Yate의 수정된 $X^2$, Cramer의 V, Gamma, Kendall의 tau-b, Somer의 D C—R, Somer의 D R—C, Pearson의 상관계수, Spearman의 상관계수, 교차적비(odds ratio, cross-product ratio)등이 있다(Agresti(1984,1990,1996), Christensen(1990), Fienberg(1980), Plackett(1981) 참조).

[1](565 - 701) 전북 완주군 삼례읍 후정리 490 우석대학교 전산통계학과 조교수
[2](565 - 701) 전북 완주군 삼례읍 후정리 490 우석대학교 전산통계학과 시간강사

분할표 자료에서 변수들의 연관관계를 시각적으로 파악하기 위해서는 적절한 그림으로 나타내어 자료구조뿐만 아니라 변수들의 연관성을 시각적으로 탐색할 수 있는 방법을 이용할 수 있는데, 이와같은 연구는 다음과 같다.

Fienberg와 Gilbert(1970)는 2 x 2 분할표에서 연관항을 사면체내의 궤적으로 표현하는 방법을 제안하였으며, 2차원에서 고려되는 연관성 측도를 궤적에 의해 기하학적으로 표현하는 방법을 제안하였다. Darroch, Lauritzen 그리고 Speed(1980)는 독립과 조건부 독립에 의해 일련의 그림으로 표시되는 그림 모형(graphical model)을 정의하였다. 그림 모형은 3차원 이상인 분할표에서 변수들간의 연관관계를 선으로 연결하는 연관그림(association graph)으로 나타낼 수 있다.

Hartigan과 Kleiner(1984)는 분할표에서 각 칸의 도수를 정사면체의 면적으로 표시한 모자익(mosaic) 그림을 제안하였다. Friendly(1992,1994)는 모자익 그림을 대수선형모형에 적합시키는 도구로 확장시켰으며, 다차원인 경우에는 모자익 그림의 식별이 어렵고 두 변수들간의 관계를 구체적으로 파악할 수 없다는 단점을 보완하기 위해서 각 타일(tile)의 연관관계와 Pearson $X^2$ 에서 각 칸에 해당하는 편차의 크기를 고려하여 색상과 함께 빗금을 친 모자익 그림을 제안하였다.

그러나 이와 같은 그림들을 이용하여 변수들의 연관관계를 측정하는 경우에는 변수들의 연관정도를 시각적으로 파악할 수는 있지만 독립이 아닌 경우에 관측값들이 독립인 것으로부터 얼마나 멀리 떨어져 있고, 어느 정도이면 독립이 되며, 어떤 특정 셀의 관측값이 연관성 여부에 크게 영향을 미치고 있는가에 대한 문제들에 대해서는 적절한 해결방안을 제공하지는 못한다.

따라서 본 연구에서는 관측도수와 기대도수에 대한 겹쳐진 모자익 그림을 이용하여 이러한 문제들을 통계량들과 함께 시각적으로 해결하는 방법을 제시하고자 한다.

## 2. 2 x 2 분할표에 대한 모자익 그림

### 2.1 관측값에 대한 모자익 그림

두 개의 범주형 변수 A와 B에 대한 2차원 분할표에서 행과 열의 범주 개수가 각각 두 개인 2 x 2 분할표를 고려하면 표 1과 같다.

표 1에서 $n_{ij}$(i=1, 2 와 j=1, 2 )는 각 칸의 관측도수를 나타내며, 변수 B의 주변합을 $n_{\cdot j}$, 변수 A의 주변합을 $n_{i\cdot}$, 총 관측도수를 $n_{\cdot\cdot}$이라 하면 각각 다음과 같이 정의된다.

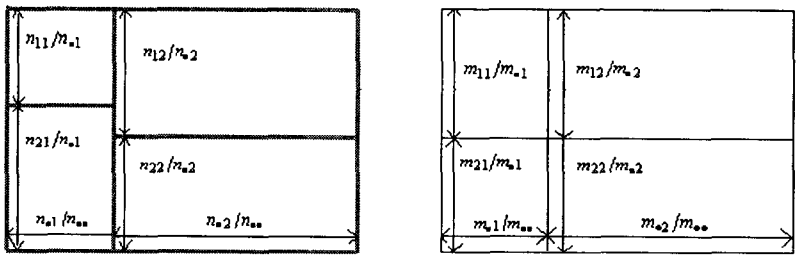
$$n_{\cdot j} = \sum n_{ij}. \tag{1}$$

$$n_{i\cdot} = \sum n_{ij}. \tag{2}$$

$$n_{\cdot\cdot} = \sum \sum n_{ij}. \tag{3}$$

표 1: 2 x 2 분할표

| | | 변수 B | | 계 |
|---|---|---|---|---|
| | | $B_1$ | $B_2$ | |
| 변수 A | $A_1$ | $n_{11}$ | $n_{12}$ | $n_{1.}$ |
| | $A_2$ | $n_{21}$ | $n_{22}$ | $n_{2.}$ |
| 계 | | $n_{.1}$ | $n_{.2}$ | $n_{..}$ |




(a) 관측값에 대한 모자익 그림. (b) 기대도수에 대한 모자익 그림.

그림 1: 2 x 2 분할표에 대한 모자익 그림

각 칸의 관측도수를 사면체의 면적으로 표시한 네 개의 사각형으로 이루어진 그림을 모자익 그림이라 하며, 네 개의 사각형을 그리는데 필요한 가로와 세로의 길이를 계산하는 식은 그림 1 (a)의 대응되는 변에 각각 나타나있다.

만일 그림 1 (a)와 같이 관측도수에 대한 모자익 그림에서 네 개의 사각형이 서로 어긋나게 교차되어 나타나면, 두 변수들간에 연관성이 존재함을 의미한다. 그러나 어긋나는 정도의 유무에 따라서 두 변수들간의 연관성 여부를 판단하기에는 쉽지 않다.

### 2.2 독립성 모형하에서의 모자익 그림

분할표의 독립성 모형($p_{ij} = p_{i.}p_{.j}$)에 대한 모자익 그림은 분할표의 각 칸의 기대도수를 이용하여 표현되는데 각 칸의 기대도수의 추정량은 다음 식 (4)와 같다.

$$\hat{m_{ij}} = \frac{m_{i.}m_{.j}}{m_{..}}. \tag{4}$$

관측값에 대한 모자익 그림과 같은 방법으로 각 칸의 기대도수를 사면체의 면적으로 표시하면 그림 1 (b)와 같은 기대도수에 대한 모자익 그림을 얻을 수 있다. 기대도수에 대한 모자익 그림에서 각 타일들은 정확하게 서로 교차되어 나타난다.

표 2: Wakely(1954)의 자료를 이용한 분할표

| | | 생존여부 | | 계 |
|---|---|---|---|---|
| | | 죽는다 | 살아난다 | |
| 식목깊이 | 얕다 | 41 | 59 | 100 |
| | 깊다 | 11 | 89 | 100 |
| 계 | | 52 | 148 | 200 |

## 3. 겹쳐진 모자익 그림을 이용한 연관성 측정

앞에서 설명한 관측도수와 기대도수에 의한 각각의 모자익 그림들은 관측값들의 독립여부를 파악하기가 어렵다. 따라서 변수의 연관성을 시각적으로 보다 쉽게 측정하기 위하여 두 가지 그림에 대한 겹쳐진 모자익 그림(overlapped mosaic plot)을 고려할 수 있다.

동적그래픽스로 구현된 겹쳐진 모자익 그림에서 임의의 특정 타일의 크기(관측 도수)를 증가시키거나 감소시켜가면서, 변화되는 네 개의 연관성 측도와 관측값을 통해, 두 변수가 독립이려면 관측값이 어느 정도여야 되는지를 시각적으로 파악할 수 있다 따라서 겹쳐진 모자익 그림은 범주형 자료의 구조를 시각적으로 표현해주는 단순한 기능뿐만 아니라, 종속인 경우에는 어느 정도 종속인지도 파악할 수 있으므로 다차원 분할표 자료를 분석하는데 유용한 정보를 제공한다고 할 수 있다.

겹쳐진 모자익 그림의 효과를 살펴보기 위하여 표 2와 같은 자료를 고려해 본다. 이 자료는 Wakely(1954)에서 잎이 긴 소나무의 묘목들을 겨울에 0.5인치 얕게 혹은 깊게 심은 후에, 다음해 가을까지 얼마 만큼 묘목이 살아 남아 있는지를 조사한 자료이다. 표 2 자료에 대한 교차적비는 5.622로서, 식목깊이와 생존여부가 매우 높은 연관관계를 가지고 있다고 볼 수 있다.

표 3은 식목깊이와 생존여부라는 두 변수에 대해, 서로 독립이라는 가정하에서의 기대도수를 구하여 구성한 분할표이다.

Wakely(1954)의 자료에 대한 연관성 측정을 하기 위하여 표 2의 분할표 자료에 대한 모자익 그림과 표 3의 기대도수들의 분할표에 대한 모자익 그림을 각각 그리면 그림 2와 같다.

그림 2 (a)에서는 식목깊이와 생존여부에 대한 네 개의 사각형이 서로 엇갈리게 나타나 있는 정도가 매우 크다. 따라서 식목깊이와 생존여부는 아주 밀접한 관계를 가지고 있다는 것을 알 수 있다. 또한, 각 사각형의 면적을 비교함으로써 두 변수에 대한 범주들의 비율에 대한 차이가 매우 크다는 사실을 시각적으로 파악할 수 있다. 그림 2 (b)와 비교할 때 두 경우의 모자익 그림들은 상당한 차이를 보이고 있으므로 식목깊이와 생존유무는 아주 밀접

표 3: 표2의 자료에 대한 기대도수의 분할표

| | | 생존여부 | | 계 |
|---|---|---|---|---|
| | | 죽는다 | 살아난다 | |
| 식목깊이 | 얕다 | 26 | 74 | 100 |
| | 깊다 | 26 | 74 | 100 |
| 계 | | 52 | 148 | 200 |

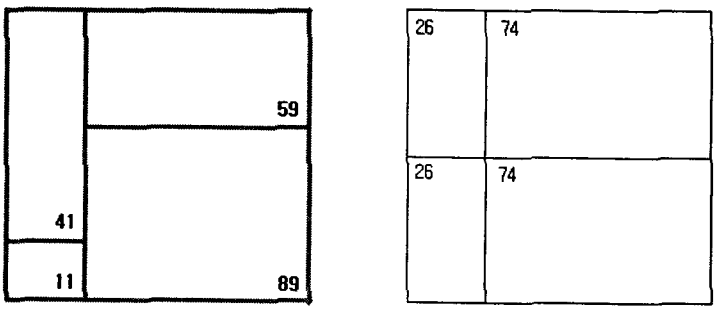



그림 2: 모자익 그림

한 관계를 가지고 있다는 것을 알 수 있다.

그림 3은 관측도수에 대한 모자익 그림인 그림 2 (a)와, 독립이라는 가정하에서의 기대도수에 대한 모자익 그림인 그림 2 (b)를 겹쳐서 나타낸 모자익 그림이다. 겹쳐진 모자익 그림과 함께 그림 3의 오른쪽에는 네 가지 연관성 측도들과 통계량들을 함께 제시하여 그들의 변화 정도를 파악할 수 있게 하였다.

그림 3의 겹쳐진 모자익 그림에서 소나무 묘목이 죽은 경우에는 식목깊이에 대한 두 범주의 관측도수와 기대도수의 차이가 크게 나타나 있음을 알 수 있다. 따라서 두 변수가 종속적인 관계임을 알 수 있다. 이와같이 겹쳐진 모자익 그림을 이용하면 두 변수의 연관정도 뿐만 아니라 관측값의 종속관계 정도와 각 관측값이 어느 정도이면 독립이 되는지를 시각적으로 파악할 수 있다.

그림 4는 그림 3에서 관측도수와 기대도수의 차이가 가장 크게 나타나고 있는 2행 1열의 관측도수 11을 20으로 증가시킨 경우의 겹쳐진 모자익 그림과 연관성 통계량들을 나타내고 있다. 이 경우 1행 2열과 2행 2열의 관측도수는 각각 59와 89로 변화되지 않았으나 1행 1열과 2행 1열의 관측도수가 각각 32와 20으로 변화되었기 때문에 네 개의 기대도수가 변화되고, 이에 따라 변화된 겹쳐진 모자익 그림이 나타나 있다. 변화된 관측값에 대한 교차적비가 2.326으로 최초 관측값에 대한 교차적비 5.622에 비해 훨씬 1에 가까워 졌으므로, 두

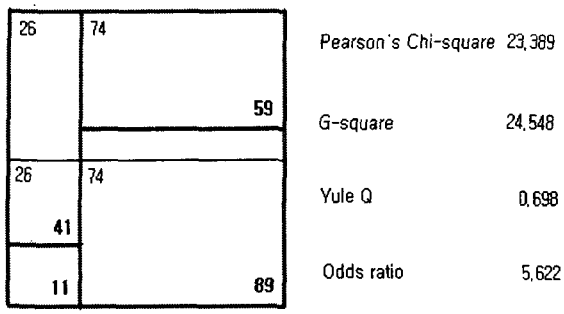



그림 3: 겹쳐진 모자익 그림

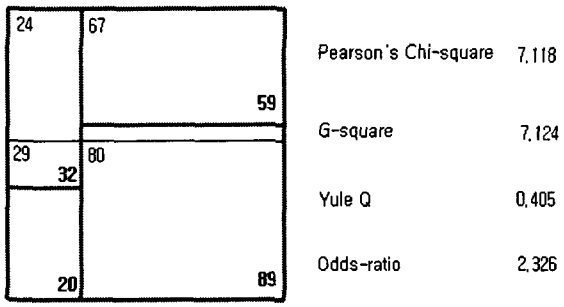



그림 4: 동적 그래픽스로 구현된 겹쳐진 모자익 그림의 예제

변수간의 연관정도가 멀어지고 독립쪽으로 가까워 졌다는 사실을 쉽게 알 수 있다.

## 4. 결론

본 연구에서는 범주형 자료에서 변수들의 연관관계를 주어진 통계량 뿐만 아니라 시각적으로도 쉽게 파악하기 위하여, 관측도수와 기대도수에 대한 두개의 겹쳐진 모자익 그림을 동적 그래픽스 기법을 이용하여 살펴 보았다. 동적 그래픽스로 구현된 겹쳐진 모자익 그림을 이용하면, 범주형 자료에서 변수들의 연관관계가 독립이 아닌 경우에, 각 관측값들의 종속관계를 시각적으로 쉽게 살펴볼 수 있고, 각 관측값들이 어느 정도이면 독립이 되는가를 판단할 수 있으며, 어떤 특정 셀의 관측값이 연관성 여부에 영향을 크게 미치고 있는가를 확인할 수 있다.

따라서 동적그래픽스로 구현된 겹쳐진 모자익 그림은 두 변수들의 자료구조뿐만 아니라 연관정도를 통계량과 더불어 시각적으로도 파악할 수 있으므로 범주형 자료를 분석하는데 매우 유용한 보조 도구라고 할 수 있다.

본 연구에서는 시각적인 분석 효과를 높이기 위하여 2 x 2 분할표에 대한 모자익 그림만을 고려하였다. 그러나 겹쳐진 모자익 그림을 다차원 분할표로 확장하면서 범주형 자료에 대한 초기 모형을 설정하는데 유용한 정보를 제공할 수 있도록 하는 연구가 필요하다.

## 참고문헌


1. Agresti, A. (1984). *Analysis of Ordinal Categorical Data,* John Wiley and Sons.
2. Agresti, A. (1990). *Categorical Data Analysis,* John Wiley and Sons.
3. Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, John Wiley and Sons.
4. Christensen, R. (1990). *Log-Linear Models,* Springer-Verlag.
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*, vol. 8, 522-539.
6. Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data,* MIT Press.
7. Fienberg, S. E., and Gilbert, J. P. (1970). The geometry of a contingency tables, *Journal of the American Statistical Association*, Vol. 65, 694-701.
8. Friendly, M. (1992). Mosaic displays for log-linear models, Proceedings of the Statistical Graphics Section, *Journal of the American Statistical Association*, 61-68.
9. Friendly, M. (1994). Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association*, Vol. 89, 190-200.
10. 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.
11. Hartigan, J. A., and Kleiner, B. (1984). A mosaic of the television ratings, *The American Statistician*, vol. 38, 32-35.
12. Plackett, R. L. (1981). *The Analysis of Categorical Data,* Charles Griffin and Company Ltd.
13. Wakely, P. C. (1954). *Planting the southern pines,* U. S. Dept. Agr. Foresr. Serv. Agr. Monogr. 18, 1-233.

# Measurement of Association of Categorical Data Using The Overlapped Mosaic Plot : Dynamic Graphics Approach for 2 x 2 Contingency Table

**Yeo-Chang Yoon** [3] · **Min-Gweon Oh** [4]

**요약**

In this paper, we propose an overlapped mosaic plot which proposed by Hartigan and Kleiner(1981) represents the counts in 2 x 2 contingency table directly by tiles whose area is proportional to the cell frequency. Overlapped mosaic plot provides some measurements of association including dynamic graphics for mosaic plots. Dynamic graphics for mosaic plots give some useful informations when one gets some measurements of association and selects a model, and current statistical software does not provide this feature.

We can see the deviations between observation and estimate of independence from overlapped mosaic plot. This dynamic graphics give some useful informations how far this data are apart from independence.

*Key words and Phrases*: Dynamic Graphics, Measurement of association, Overlapped mosaic plot

[3]Assistant Professor, Dept. of Computer Science and Statistics, Woosuk University, Chonbuk, 565-701, Korea
[4]Lecturer, Dept. of Computer Science and Statistics, Woosuk University, Chonbuk, 565-701, Korea