The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 773-781

A Measures of Association for Two-way Contingency Tables by Maximum Association[1)]

Sang Giun Kim[2)] and Woo Rhee Lee[3)]

Abstract

Pearson χ^2 statistics can be used as the measure of association for two-way contingency tables. However, it's magnitude depends on the sample size N, the normalized index $\Phi^2 = \chi^2/N$ has been widely used in many researches instead of that. In this study, we firstly point out that the actual upper bound of Φ^2 can be affected by the given marginals of the tables, so that Φ^2 should be adjusted the actual maximum to get the proper range $[0,1]$. Also we point out that the maximum could be found easily by using several algorithms like Hu and Mukerjee(2002). With the actual maximum, we propose the relative measure of association which is normalized by it.

Keywords : Measure of association, Φ^2 Statistics, Contingency tables.

1. 서 론

범주형 변수들에 의해 교차 분류된 분할표 분석의 주된 관심은 표를 구성하는 변수들 사이의 연관관계를 식별하는 것으로 판단할 수 있다. 특히, 두 명목형 범주형 변수들에 의해 교차 분류된 이차원 분할표의 경우에 Pearson의 카이제곱(χ^2) 통계량에 의한 독립성 검정은 연관관계 식별을 위해 매우 중요한 역할을 수행하고 있다. 이와 연계하여 두 변수의 독립성 검정과는 별개로 두 변수의 연관성을 측정하는 측도 역시 이차원 분할표 분석에서 매우 중요한 관심사 중의 하나이다. 이러한 이유로 이차원 분할표의 연관을 측정하기 위하여 많은 측도들이 제안되어 있으며, Gooman과 Kruskal(1979)을 포함한 다수의 범주형 자료 분석이론 저술들에서 이들 측도들의 특징과 제한점등을 참고할 수 있다.

두 명목형 변수에 의한 이차원 분할표의 연관성을 측정하기 위한 측도들은 크게 Pearson의 χ^2

1) This work was supported by 2001 Kyonggi University Research Fund for Sang Giun Kim.
2) Professor, Department of Applied Information Statistics, Kyonggi University, Suwon 443-760, Korea
E-mail : sgkim@kyonggi.ac.kr
3) Professor, Department of Applied Information Statistics, Kyonggi University, Suwon 443-760, Korea

통계량에 기반한 측도들과 PRE(proportional reduction in error) 원칙에 의한 측도들로 나눌 수 있다. 이들 제안된 대부분의 측도들은 '0'과 '1' 사이의 값을 가지게 되며, '0' 값은 두 변수가 서로 독립일 때 갖는 측도의 최소값 그리고 '1'은 두 변수가 서로 완전한 연관(perfect association)임을 나타내는 최대값을 의미하게 된다. 이때 '0' 값을 갖는 최소연관은 독립성에 의하여 설명할 수 있는데 반하여 '1' 값을 갖는 완전한 연관은 Reynolds(1977)와 Mirkin(2001)이 지적한 바와 같이 분할표의 모양 즉, 두 변수의 범주의 수가 다른 경우와 주변분포(marginal distribution)의 영향에 의하여 만족할 만한 정의를 이끌어 내는 것에 많은 제한점을 가지고 있다. 그러므로 Pearson의 χ^2 통계량에 기반한 측도들은 두 변수가 서로 독립인 경우에 '0' 값을 갖게 되지만 표의 형태와 주변분포의 영향에 의하여 대부분 완전한 연관을 나타내는 '1' 값을 최대값으로 갖지 못한다.

이차원 분할표를 위한 연관성 측도들이 갖는 이러한 문제점들을 해결하기 위한 방법으로 분할표를 주어진 주변분포에 의하여 표준화하는 방법과 주어진 표본크기와 주변분포에 의한 측도의 최대값, 즉 최대연관(maximum association)을 얻고 분할표로부터 측정한 값을 이 값으로 나누어 주는 방법을 고려할 수 있다. 이들 중에서 주변분포의 영향을 제거하기 위하여 표준화 표(standardized table)를 얻어 연관성을 측정하는 방법은 IPF(iterative proportional fitting) 방법을 직접 적용하는 것으로 쉽게 얻을 수 있기 때문에 가장 널리 사용되고 있으며, 둘 이상의 분할표들의 연관성을 비교하기 위하여 이용하기도 한다. 그러나 표준화 표로부터 얻어지는 연관성 측도 역시 표의 모양에 따라 최대값 '1'을 갖지 못할 수 있다.

표준화에 의한 방법은 원 분할표로부터 연관성을 측정하는 것이 아니라 원하는 주변분포를 갖는 표준화 표를 얻어 연관성을 측정하는 방법으로 이해할 수 있다. 같은 맥락에서 최대연관으로 나누어 주는 방법은 먼저, 주어진 주변분포로부터 최대연관을 갖는 분할표를 찾고, 원 분할표로부터 측정된 측도값을 최대연관을 갖는 분할표에서 측정한 측도값으로 나누어주는 방법이라고 할 수 있다. 다시 말해 완전한 연관은 분할표의 형태와 주변분포에 의해 결정되므로 연관성 측도의 최대값을 찾는 문제는 주어진 조건, 즉 행과 열의 수, 표본크기 그리고 주변분포로부터 최대연관을 갖는 분할표를 찾는 문제와 동등한 것으로 판단할 수 있다.

본 연구에서는 주어진 표본크기와 주변분포로부터 얻을 수 있는 Pearson의 χ^2 통계량의 최대값을 찾고, 이를 이용하여 원 분할표로부터 얻어진 측도값을 나누어준 측도에 관하여 연구하고자 한다. 제2장에서는 연관성 측도로서의 Pearson의 χ^2 통계량과 이에 기반한 표준화된 측도에 관하여 간략히 정리하고, 이들이 갖는 주변분포에 의한 영향력에 의한 문제점을 지적하였다. 또한 이러한 문제점을 해결하기 위하여 최대연관을 갖는 분할표를 찾는 방법들을 소개하고, 이를 통하여 얻어진 최대값에 의한 표준화된 측도를 제안하고자 한다. 제3장에서는 실제 자료를 이용하여 제안된 측도의 적용 예를 소개하고자 하며, 마지막으로 제4장에서는 본 연구의 결과를 정리하고, 제안된 방법의 PRE 원칙에 의한 측도들로의 적용문제에 관하여 토론하였다.

2. 최대연관에 의한 상대적 연관성 측도

2.1 연관성 측도로서의 χ^2 통계량

I개 행 범주와 J개 열 범주를 가진 두 명목형 범주형 변수 R과 C에 의한 이차원 분할표의

(i,j)번째 관찰칸 값을 N_{ij}, i번째 행의 주변합과 j번째 열의 주변합을 각각 $N_{i+}=\sum_{j=1}^{J}N_{ij}$와 $N_{+j}=\sum_{i=1}^{I}N_{ij}$로 나타내기로 한다. 그리고 (i,j)칸의 비율은 $p_{ij}=N_{ij}/N$, i번째 행과 j번째 열의 주변비율은 각각 $p_{i+}=N_{i+}/N$, $p_{+j}=N_{+j}/N$와 같이 나타내도록 한다. 또한 π_{ij}는 (i,j)번째 칸의 모비율 그리고 π_{i+} , π_{+j}는 각각 i번째 행과 j번째 열의 모비율을 나타내기로 한다. 여기서 만일 모든 i와 j에 대하여

$$\pi_{ij}=\pi_{i+}\pi_{+j} \tag{2.1}$$

을 만족한다면 두 변수 R과 C는 서로 독립인 관계를 가지며, 이를 검정하기 위하여 다음과 같은 Pearson의 카이제곱 통계량을 고려할 수 있다.

$$\chi^2=N\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-\pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}}$$

그러나 π_{i+}와 π_{+j}는 일반적으로 알려져 있지 않으므로 이들의 최우추정량인 p_{i+}과 p_{+j}을 이용한 다음과 같은 통계량에 의하여 두 변수 R과 C의 독립성 유무를 판단하게 된다.

$$\begin{aligned}X^2&=N\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-p_{i+}p_{+j})^2}{p_{i+}p_{j}}\\&=N\Phi^2\end{aligned} \tag{2.2}$$

통계량 (2.2)는 자유도 $(I-1)(J-1)$을 갖는 χ^2 분포를 따르며, 만일 (2.2)가 상당히 크다면, 즉 해당 분포에서의 p-값이 상당히 작다면 (2.1)의 관계는 옳지 않고 두 변수는 서로 연관을 가지고 있을 수 있다는 판단이 가능하다. 이러한 사실로부터 통계량 (2.2)는 두 변수 R과 C의 연관을 측정하는 측도로 이용되기도 한다. 그러나 통계량 (2.2)는 표본의 총수, $N=\sum_{i=i}^{I}\sum_{j=1}^{J}N_{ij}$의 영향을 받기 때문에 이의 영향력을 제거한 파이 계수(phi squared coefficient)로 알려진 $\Phi^2=X^2/N$가 두 변수의 연관을 측정하는 측도로 이용되고 있다.

2.2 최소연관과 최대연관에 의한 표준화된 측도

식 (2.2)로부터 두 변수가 서로 독립일 경우에 $\Phi^2=0$인 것을 알 수 있다. 이렇듯 $\Phi^2=0$값을 가질 때를 두 변수는 최소연관(minimum association)을 갖는 다고 하며, 두 변수는 서로 연관이 없는 것으로 판단한다. 이제 두 변수가 가질 수 있는 최대연관(maximum association)을 나타내는

값을 유도하기 위하여 Φ^2를 다음과 같이 표현하기로 한다.

<표 2.1> 완전한 연관을 나타내는 분할표들의 예

(가) $I=J$이고 $\Phi^2=2$

3/10	0	0	3/10
0	3/10	0	3/10
0	0	4/10	4/10
3/10	3/10	4/10	

(나) $I<J$이고 $\Phi^2=2$

2/10	0	0	0	2/10
0	0	3/10	0	3/10
0	2/10	0	3/10	5/10
2/10	2/10	3/10	3/10	

$$\Phi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i+}p_{+j}} - 1 \tag{2.3}$$

Φ^2의 최대값은 윗 식의 우변 합이 최대값을 가질 때 얻어지므로 $I<J$라 가정하고 다음과 같이 합 부분의 각 행에서의 열의 합을 고려하기로 하자.

$$\begin{aligned}\Phi_i &= \sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i+}p_{+j}} \\ &= \sum_{j=1}^{J}\left(\frac{p_{ij}}{p_{i+}}\right)\left(\frac{p_{ij}}{p_{+j}}\right)\end{aligned} \tag{2.4}$$

식 (2.4) 우변의 괄호 안 비율은 행과 열의 조건부 확률로 $\sum_{j=1}^{J}(p_{ij}/p_{i+})=1$이므로 Φ_i는 행의 조건부 분포에 의한 각 열의 조건부 확률 p_{ij}/p_{+j}의 평균임을 나타낸다. 그러므로 $\Phi_i \le \max_j(p_{ij}/p_{+j}) \le 1$과 같은 관계가 성립하며 최대값 1을 얻기 위해서는 특정한 열, j'에서 (i,j')칸의 비율이 주변비율 $p_{+j'}$과 같아야 한다. 다시 말해 $p_{ij'}=p_{+j'}$이어야 하고, 이는 i번째 행에는 j'열만이 칸 비율을 갖고 나머지는 모두 '0' 값을 가져야 하는 것을 의미한다. 이러한 사실로부터, 결국 Φ^2의 최대값은 $I<J$라 가정했으므로 $\min(I,J)-1$인 것을 알 수 있다. 따라서 Φ^2이 최대값을 갖기 위해서는 주어진 분할표의 모든 행 주변비율이 '0'보다 커야하며, 특정 행에서 열 비율은 오직 한 칸에서만 나타나야 하는 것을 의미한다.

<표 2.1>은 최대값 $\min(I,J)-1$을 값을 갖는 분할표의 예들을 보여주고 있다. <표 2.1>의 비율에 의한 분할표 (가)는 $I=J$인 경우로 주어진 한 행에서 반드시 열 비율은 한 칸에만 존재하는 것을 알 수 있다. 그러나 행과 열의 수가 다른 분할표 (나)는 주어진 한 행에서 열 비율이 한 칸 혹은 두 칸에 나타나고 있는 것을 알 수 있다. 이러한 경우에 두 변수 R과 C는 서로 완전한 연

관(complete association)을 갖는다고 하며, 두 변수가 완전한 연관을 가질 때 Φ^2은 최대값 $\min(I,J)-1$을 갖게 된다. 분할표 (나)와 달리 $I>J$인 경우에도 유사한 설명이 가능하다.

이러한 사실에 근거하여 두 변수가 독립인 경우와 완전한 연관을 가질 때 Φ^2이 '0'과 '1' 사이 값을 갖도록 표준화(normalized)한 다음과 같은 Cramer의 C^2 통계량과 같은 표준화된 측도들이 제안되어 있다.

<표 2.2> 주변합의 영향에 의해 최대값 $\min(I,J)-1$을 갖지 못하는 분할표들

(가) $\Phi^2=1.5333$

2/10	0	0	2/10
0	3/10	0	3/10
1/10	0	4/10	5/10
3/10	3/10	4/10	

(나) $\Phi^2=1.3247$

1/10	0	0	0	1/10
5/100	0	3/10	0	35/100
5/100	2/10	0	3/10	55/100
2/10	2/10	3/10	3/10	

<표 2.3> 표준화된 행 주변합에 의해 최대연관을 갖는 표

(가) $\Phi^2=1.65$

3/10	0	1/30	1/3
0	3/10	1/30	1/3
0	0	1/3	1/3
3/10	3/10	4/10	

(나) $\Phi^2=1.7$

1/30	0	0	3/10	1/3
4/30	2/10	0	0	1/3
1/30	0	3/10	0	1/3
2/10	2/10	3/10	3/10	

<표 2.4> 표준화된 행과 열 주변합에 의해 최대연관을 갖는 표

(가) $\Phi^2=2$

1/3	0	0	1/3
0	1/3	0	1/3
0	0	1/3	1/3
1/3	1/3	1/3	

(나) $\Phi^2=1.5$

1/4	0	0	1/12	1/3
0	1/4	0	1/12	1/3
0	0	1/4	1/12	1/3
1/4	1/4	1/4	1/4	

$$C^2=\frac{\Phi^2}{\min(I,J)-1} \tag{2.5}$$

여기서 <표 2.1>에 제시된 분할표들과 같은 행과 열 범주수를 갖는 <표 2.2>의 분할표들을 고려해보기로 한다. 먼저, <표 2.2>의 분할표 (가)는 <표 2.1>의 분할표 (가)와 열 변수의 주변분포는 같으나 행 변수의 주변분포가 다르다는 차이점을 가지며, 이러한 주변분포에 의해 최대연관을 갖는 분할표의 칸 확률이 결정되었다. 그러나 <표 2.1>에서와는 달리 주어진 한 행에서 반드시

열 비율은 한 칸에만 존재하지 않고, 최대값 $\min(I,J)-1=2$를 갖지 못한다. 즉, 행 변수의 주변분포가 Φ^2의 최대값에 영향을 미치고 있으며, 행과 열의 수가 다른 분할표 (나) 역시 최대값 '2'를 갖지 못하는 것을 알 수 있다.

이러한 행 주변분포에 따른 영향을 제거하기 위하여 행 주변분포를 균등분포로 한 후에 표준화된 표에서의 최대연관을 갖도록 결정된 칸 확률에 의한 분할표가 <표 2.3>에 제시되어 있다. <표 2.2>에서와 마찬가지로 주어진 한 행에서 반드시 열 비율은 한 칸에만 존재하지 않고, 최대값 '2'를 갖지 못한다. 즉, 행 변수의 주변분포가 Φ^2의 최대값에 영향을 미치고 있다. 같은 맥락에서 행과 열의 수가 다른 분할표 (나) 역시 최대값을 갖지 못하는 것을 알 수 있다.

행과 열 주변분포를 모두 균등분포로 한 후에 표준화된 표에서 최대연관을 갖도록 결정된 칸 확률에 의한 분할표들인 <표 2.4>의 분할표 (가)는 주어진 한 행에서 열 비율은 한 칸에만 존재하므로 최대값 '2'를 얻게 된다. 반면에 행과 열 범주수가 다른 분할표 (나)는 표준화를 수행하더라도 최대값을 얻지 못하는 것을 알 수 있다. Φ^2의 최대값은 행과 열 주변분포에 의해 영향을 받고 행과 열 범주수의 차이 역시 영향을 미치므로 표준화된 분할표를 통해서 항상 최대값 $\min(I,J)-1$을 갖을 수 없다. 그러므로 식 (2.5)의 Cramer의 C^2 통계량은 행과 열 주변분포 그리고 행과 열 범주수의 차이에 따라 최대값은 $(\min(I,J)-1)$보다 작거나 같을 수 있으므로 두 변수의 연관을 과소평가할 위험을 갖게 된다.

이러한 사실들로부터 Cramer의 C^2 통계량과 유사한 두 변수의 범주수와 주변분포에 의해 결정되는 최대값으로 표준화한 다음과 같은 연관성 측도를 고려할 수 있다.

$$M^2=\frac{\Phi^2}{\Phi^2_{\max}} \tag{2.6}$$

여기서 $\Phi^2_{\max}$은 두 변수의 범주수와 주변분포에 의해 결정되는 최대값을 의미하며, M^2은 최대값에 의해 표준화되었으므로 $0\le M^2\le 1$인 것은 너무도 분명하다. 그러나 M^2을 얻기 위해서는 반드시 주어진 조건들, 즉 두 변수의 범주수와 주변분포에 의해 결정되는 $\Phi^2_{\max}$을 먼저 얻어야만 한다.

최대값 $\Phi^2_{\max}$을 찾는 문제는 Lee(1997)에서 지적한 바와 같이 목적함수 (2.4)와 주어진 주변분포를 조건식으로 한 LP(Linear Programing) 문제라고 볼 수 있다. 이때 주변합이 주어진 분할표의 경우에는 독립성에서 가장 많이 벗어난 최대연관을 갖는 분할표를 찾는 문제와 동등하며, 이를 찾기 위한 발견법적인 방법을 포함한 여러 알고리즘이 Kalantari 등(1993)에 의해 제안되어 있다. 특히 주변분포를 조건으로 한 최대값 $\Phi^2_{\max}$은 해가 존재하는 영역의 꼭지점(vertex)들에서 찾을 수 있기 때문에 Hu와 Mukerjee(2002)가 제안한 모든 가능한 극단표(extreme table)를 찾는 알고리즘에 의해 이러한 정점에 위치한 분할표들을 찾을 수 있다. 따라서 주어진 주변분포를 만족하는 최대연관을 갖는 분할표는 이들 방법들에 의해 어렵지 않게 얻을 수 있으며, 제안된 연관성 측도 M^2은 최대연관을 갖는 분할표를 찾고 이때 얻어진 $\Phi^2_{\max}$에 의해 원 분할표에서 얻어진 Φ^2을 통하여 구할 수 있다.

3. 실제자료 분석 예

Agresti(1990)에서 인용한 DiFranceisco와 Giltelman(1984)에 의해 조사된 <표 3.1>은 구소련 국민의 정치참여도를 여타 국가와 비교하기 위하여, 국가 그리고 교육수준이 정치집회에 정기적으로 참석하는지의 여부에 어떠한 영향을 미치는지 분석하기 위한 자료이다. 자료는 다음과 같은 세 변수에 의한 5×3×2 분할표이다.

<표 3.1> 국가, 교육별 정기적인 정치집회 참석여부

정치집회 참석	구소련		미국		영국		이탈리아		멕시코	
	예	아니오	예	아니오	예	아니오	예	아니오	예	아니오
초등	94	84	227	112	356	144	166	526	447	430
중등	318	120	371	71	256	76	142	103	78	25
고등	473	72	180	8	22	2	47	7	22	2

<표 3.2> Φ^2_{max}을 갖는 국가별 분할표

정치집회 참석	구소련		미국		영국		이탈리아		멕시코	
	예	아니오	예	아니오	예	아니오	예	아니오	예	아니오
초등	0	178	336	3	500	0	56	636	547	330
중등	340	98	442	0	110	222	245	0	0	103
고등	545	0	0	188	24	0	54	0	0	24

<표 3.3> 정치집회 참석여부 자료의 연관성 측도들

국가	원 분할표			표준화된 분할표
	Φ^2	Φ^2_{max}	M^2	Φ^2
구소련	0.0780	0.6384	0.1222	0.0815
미국	0.0724	0.9806	0.0738	0.1044
영국	0.0089	0.5527	0.0161	0.0585
이탈리아	0.1574	0.7741	0.2033	0.2581
멕시코	0.0363	0.1733	0.2095	0.1843

변수 A : 국가(구소련, 미국, 영국, 이탈리아, 멕시코)
변수 B : 교육수준(초등교육, 중등교육, 고등교육)
변수 C : 정기적인 정치집회 참석여부(예, 아니오)

이들 세 변수의 연관관계를 식별하기 위하여 삼차원 로그선형모형을 적합시킨 결과, 각 두 변수들의 상호연관에 의한 부분연관모형 $[AB][AC][BC]$는 우도비 검정통계량 값 $G^2=35.09$을 갖고, 모형의 자유도는 '8'로 '0'에 가까운 p-값을 갖는다. 즉, 세 변수가 서로 완전한 연관을 갖는 포화모형이 세 변수들의 연관구조를 가장 잘 설명하는 모형이라고 판단할 수 있다. 이러한 근거에 의하여 5개 국가별로 교육수준과 정치집회 참석여부의 연관을 비교하기 위하여 얻은 Φ^2값이 <표 3.3>에 정리되어 있으며, 이때 국가별로 $\Phi^2_{\max}$을 갖는 분할표들은 <표 3.2>에 제시되어 있다.

<표 3.3>의 두 번째 열에서 이탈리아의 교육수준과 참석여부의 연관이 가장 높은 것을 알 수 있다. 그리고 구소련이 미국에 비해 다소 높은 연관을 갖으며, 이러한 결과는 DiFranceisco와 Giltelman(1984)의 연구결과와 일치한다. 그러나 세 번째 열에 제시된 최대값의 의해 표준화된 M^2에 의하면 역시 이탈리아의 연관이 가장 높고, 구소련의 연관이 미국에 비하여 높은 연관관계를 갖는 것을 알 수 있다. 한 가지 특기할 점은 멕시코의 연관이 M^2에 의하면 구소련과 미국에 비하여 높게 나타난 점이다.

이러한 차이가 국가별 교육수준의 주변합의 차이에서 기인한 것인지의 여부를 판단하기 위하여 교육수준을 균등분포로 표준화한 분할표로부터 얻은 값들이 <표 3.3>의 오른쪽 '표준화된 분할표' 열에 정리되어 있다. 우선 표준화된 분할표에서 얻은 Φ^2 값들로부터 이탈리아가 가장 높은 연관을 갖고, 미국은 구소련에 비하여 다소 높은 연관을 갖는다는 차이가 나타난다. 그러나 멕시코의 연관은 구소련과 미국에 비하여 높게 나타난 것을 볼 수 있다. 이러한 현상은 멕시코의 교육수준별 주변합이 매우 큰 차이를 가지는 것이 연관에 영향을 주고 있는 것으로 해석할 수 있으며, 이 결과는 주변합에 의해 결정된 최대연관에 의한 M^2에서 얻은 사실과 유사한 결과를 보이고 있다.

4. 결 론

Peason의 χ^2 통계량은 두 명목형 변수에 의한 이차원 분할표의 연관을 측정하는 측도로 이용될 수 있다. 그러나 χ^2 통계량은 표본크기 N에 영향을 받기 때문에 이의 영향력을 제거한 파이계수 $\Phi^2=\chi^2/N$이 널리 이용되고 있다. 그러나 Φ^2이 갖는 값에 의해 연관관계를 해석하는 것은 쉽지 않기 때문에 Cramer의 C^2과 같이 구간 $[0,1]$을 갖도록 표준화된 측도들이 제안되고 있다.

본 연구에서는 먼저, Φ^2의 상한, 즉 최대값 $\min(I,J)-1$이 분할표를 구성하는 두 변수의 범주수와 주변분포에 의하여 영향을 받는다는 것을 지적하고, 최대값 '1'을 갖기 위하여 이론적인 최대값 $\min(I,J)-1$이 주변분포에 의해서 결정된 최대값인 $\Phi^2_{\max}$로 조정되어야 한다는 것 역시 지적하였다. 이를 바탕으로 $\Phi^2_{\max}$에 의해 파이계수를 표준화한 연관성 측도 M^2을 제안하였으며, 이때 최대값 $\Phi^2_{\max}\leq(\min(I,J)-1)$과 같은 관계를 갖는다.

제안된 M^2을 얻기 위해서는 두 변수의 범주수와 주변분포에 의해 결정되는 최대연관 $\Phi^2_{\max}$을 얻어야만 한다. 이는 주변합이 선형조건이고 Φ^2이 목적함수인 선형계획문제에 의하여 해결할 수

있으며, 이와 동등한 결과를 얻어내는 여러 알고리즘을 소개하고 특히, Hu와 Mukerjee(2002)가 제안한 알고리즘에 의해 쉽게 얻을 수 있다는 사실을 지적하였다. 따라서 M^2은 두 변수의 연관관계를 측정하는 측도로서 뿐만 아니라 실제 자료 분석 예에서와 같이 연관관계를 비교하는 측도로 응용될 수 있다.

마지막으로 한 가지 지적할 점은 본 연구에서 제안된 방법은 pearson의 χ^2 통계량에 기반한 표준화된 측도 뿐만 아니라, PRE 원칙에 의한 측도들 역시 주변분포에 따라 최대값인 '1'을 얻지 못하는 상황이 발생할 수 있으므로, 이차원 분할표의 연관관계를 측정하기 위한 여러 다른 측도들에도 직접 응용 될 수 있다는 점이다. 이때 최대값을 갖는 문제는 최대연관을 갖는 분할표를 찾는 문제가 되기 때문에 각 연관성 측도들의 최대값은 선형계획문제의 경우에 목적함수만이 다르게 되며, 역시 해가 존재하는 영역의 정점에 최대값이 놓이게 되므로 본 연구에서 적용한 Hu와 Mukerjee(2002)의 알고리즘에 의해 어렵지 않게 구할 수 있다.

참고문헌

[1] Agresti, A.(1990). *Categorical Data Analysis*, Wiley, New York.

[2] DiFranceisco, W. and Gitelman, Z.(1984). Soviet Political Culture and Covert Participation in Policy Implementation, *American Political Science Review,* Vol. 78, 603-621.

[3] Goodman, L.A. and Kruskal, W.H.(1979). *Measures of Association for Cross Classifications*, Springer-Velag, New York.

[4] Hu, X. and Mukerjee, H.(2002). Constructing All Extremes of Contingency Tables with Given Marginals, *Journal of Computational and Graphical Statistics*, Vol. 11, 910-919.

[5] Kalantari, B., Lari I., Rizzi, A., and Simeon, B.(1993). Sharp Bounds for the Maximum of the Chi-Square index in a Class of Contingency Tables with given Marginals, *Computational Statistics & Data Analysis,* Vol. 16, 19-34.

[6] Lee, A.J.(1997). Some Simple Methods for Generating Correlated Categorical Variates, *Computational Statistics & Data Analysis,* Vol. 26, 133-148.

[7] Mirkin, B.(2001). Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables, *American Statistician*, Vol. 55, 111-120.

[8] Reynolds, H.T.(1977). *The Analysis of Cross-Classifications*, The Free Press, New York.

[2005년 9 월 접수, 2005년 11 월 채택]