사회과학연구에서의 3차원 범주형 자료분석

박 옥 희*

명목변수나 서열변수와 같이 여러 항목의 범주(category)를 갖고 있는 변수들이 상호 교차되어 측정된 빈도수가 분할표의 형태로 표현되는 자료를 범주형 자료(categorical data)라 하는데 대부분의 사회과학적 연구에서는 2차원 분할표로써 두 범주형 변수간의 관계를 나타내고 널리 알려진 Pearson χ^2검정 또는 우도비 G^2검정을 사용하여 상호연관성을 분석하고 있다.

그러나 변수가 3개 이상인 고차원 범주형 자료를 분석할 경우에는 2차원 분할표와 독립성 검정은 커다란 제한점을 지니게 된다. 본 논문에서는 따라서 3차원 범주형 자료를 적절히 분석할 수 있는 방법으로 대수선형모형(loglinear model)을 소개하고, 사회학적 자료의 사례분석을 통하여 독립성 검정법으로는 파악할 수 없는 세 가지 범주형 변수들간의 교호관계를 규명해 보았다.

Ⅰ. 서 론

많은 사회과학자료의 형태는 여러 항목의 범주(category)들을 갖고 있는 변수(variable)들로 구성되어 있으며 각 변수의 범주조합에 해당하는 칸(cell)에는 빈도수(frequency)가 나타나 있다. 특히 여러 범주를 갖고 있는 변수들이 상호 교차분류되어 분할표(contingency table)의 형태로 표현되는 자료들이 많이 있는데 이러한 형태의 자료를 범주형 자료(categorical data)라고 한다.

범주형 자료를 요약정리하여 전반적인 상황을 파악, 설명하고 또 이러한 자료를 바탕으로 의사결정에서 정책수립까지 수행하는 과정은 사회과학분야의 연구자에게 있어 매우 중요한 과정이다. 이러한 범주형 자료형태를 갖고 있는 대부분의 조사자료들은 여러 가지 범주형 변수들로 구성되어 있으나 대부분의 연구보고서에서는 보통 이들 자료를 통하여 1차원 또는 많아야 2차원 분할표만을 만들어 분석, 발표하는 것이 예사이다.

2차원 분할표를 바탕으로 두 변수간의 독립성에 관하여 통계분석을 할 때에는 널리 알려진 Pearson χ^2검정이나 우도비 G^2검정을 사용

* 한국보건사회연구원 책임연구원

한다. 이 때 두 검정방법에 대한 귀무가설(null hypothesis)은 "두 변수는 독립적이다"이며, 이러한 귀무가설을 기각 또는 채택하는 통계적 의사결정을 내리게 된다.

사회과학분야의 연구에서 변수가 3개 이상인 고차원 범주형 자료를 분석할 경우에도 관심있는 두 변수의 쌍(pair)들에 대한 몇 가지 2차원 주변분할표(marginal contingency table)로 표현하여 χ^2 또는 우도비 G^2 검정분석을 하는 것이 대부분이다. 그러나 이러한 분석방법은 두 변수간의 관련성에 대해서는 자세히 알 수 있는 장점이 있으나 다음과 같은 단점이 있다.

(1) 한 쌍의 범주형 변수들간의 주변관련성과 다른 변수가 새로 등장하였을 때의 관련성을 복합적으로 설명할 수 없다.

(2) 여러 쌍에 대한 변수들의 관련성을 동시에 측정할 수 없다.

(3) 변수들의 3차나 그 이상의 차수의 교호작용(interaction)의 가능성이 무시된다.

이러한 단점들을 보완하는 방법으로서 본 논문에서는 3차원 상황에서 교차분류된 범주형 자료의 통계적 분석을 위하여 대수선형모형(loglinear model)을 이용한 자료분석을 사례를 들어 설명하고자 한다.

II. 3차원표 분석

1. 대수선형모형

범주형 변수가 3개 이상인 다차원 상황의 자료에서 상호 교차분류된 분할표를 통계적 방법으로 분석하고자 할 때 대수선형모형(loglinear model)을 이용할 수 있다. 우선 간단한 2차원 분할표에서 대수선형모형을 고려하여 보자.

O_{ij}를 행(row)변수의 i번째 항(i=1, 2, ……, I)과 열(column)변수의 j번째 열(j=1, 2, ……, J)에 해당하는 관찰값(observation)이라 할 때 귀무가설(Ho : 두 변수는 서로 독립적이다) 상황하에서 (i, j)칸에 해당하는 기대값(expected value)은

$$E_{ij}=\sum_{i=1}^{I}O_{ij}\sum_{j=1}^{J}O_{ij}/n,\quad n=\sum_{i=1}^{I}\sum_{j=1}^{J}O_{ij} \quad \cdots\cdots\cdots (1)$$

이며, Pearson χ^2검정 통계량은 다음과 같다.

$$\chi^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

독립성의 귀무가설 하에 (1)식을 다음과 같이 쓸 수 있다.

$$E_{ij}=\frac{1}{n}O_{i+}+O_{+j}$$

여기서, $O_{i+}=\sum_{j}O_{ij}$, $O_{+j}=\sum_{i}O_{ij}$ 이다. 그리고 위 식의 양변에 자연대수를 취하면 다음과 같다.

$$\log E_{ij}=\log O_{i+}+\log O_{+j}-\log n$$

위 식은 분산분석기호와 유사하다는 것을 파악할 수 있으며, 또한 E_{+j}의 모수를 M_{+j}로 한다면 다음과 같은 가산형으로 표현할 수 있다.

$$\log M_{ij}=U+U_{1(i)}+U_{2(j)} \quad \cdots\cdots\cdots\cdots (2)$$

여기서, $U=\frac{1}{IJ}\sum_{i}\sum_{j}\log M_{ij}$

$$U+U_{1(i)}=\frac{1}{J}\sum_{j=1}^{J}\log M_{ij}$$

$$U+U_{2(j)}=\frac{1}{I}\sum_{i=1}^{I}\log M_{ij}$$

으로 정의된다. 또한 모형의 제약식으로는

$$\sum_{i}U_{1(i)}=\sum_{j}U_{2(j)}=0 \quad \cdots\cdots\cdots\cdots\cdots (3)$$

이다. 위 (2)번 모형은 두 범주형 변수가 독립이라는 가정하에 설정된 것이므로 우리는 두 변수간의 교호작용항을 첨가시킨 다음과 같은 모형을 고려할 수 있다.

$$\log M_{ij}=U+U_{1(i)}+U_{2(j)}+U_{12(ij)}$$

다음으로 3개의 범주형 변수가 상호교차된 3차원인 I×J×K 분할표를 고려하자. 여기서, (i, j, k)칸에 관찰된 도수를 O_{ijk}이라 하고 대응되는 기대모수값을 M_{ijk}라 하자. 3차원 분할표에 대한 일반적인 대수모형식은 다음과 같이 정의된다.

$$\log M_{ijk}=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)}+U_{13(ik)}+U_{23(jk)}+U_{123(ijk)} \quad (4)$$

위 모형의 제약식은 (3)식을 확장하여 생각하면 된다.

위 모형 (4)를 완전모형(full model)이라고 할 때 이 모형식으로부터 여러 종류의 부분모형(partial model)을 고려할 수 있다. 여기에서 몇 가지 중요한 부분모형을 고려하여 보자. 모든 ijk에 대하여

$$U_{12(ij)}+U_{13(ik)}+U_{23(jk)}+U_{123(ijk)}=0$$

인 경우에는 세 변수들이 완전한 독립을 나타내주는 모형이다. 위 모형식에서

$$U_{12(ij)}=U_{123(ijk)}=0$$

라고 가정하면, 변수 3의 값이 고정된 경우 변수 1과 변수 2가 상호독립적임을 의미한다.

또한 모형 (4)에서

$$U_{12(ij)}=U_{23(jk)}=U_{123(ijk)}=0$$

이라 하면, 변수 1과 변수 2의 결합된 상태가 변수 3과 독립임을 표현한다. 마지막으로 모형 (4)에서

$$U_{123(ijk)}=0$$

이라는 조건을 부여한다면 2차 교호작용이 없는 모형이라 할 수 있다.

설정된 여러 형태의 모형을 통하여 각 칸의 기대값의 모수(M_{ijk})를 추정할 수 있고 추정된 기대값(E_{ijk})을 바탕으로 이미 설정한 대수선형모형의 적합성을 검정할 수 있다. 적합성을 검정하기 위한 가설은 다음과 같다.

H_0 : 어느 특정한 대수선형모형이 자료에 적합하다.

H_1 : 어느 특정한 대수선형모형이 자료에 적합하지 않다.

귀무가설에서 설정된 모형을 통하여 기대값(E_{ijk})을 구하고 Pearson χ^2나 우도비 G^2 통계량 중 하나를 사용하여 모형의 적합여부를 검정할 수 있다.

Pearson χ^2 통계량 :

$$\chi^2=\sum_{ijk}\frac{(O_{ijk}-E_{ijk})^2}{E_{ijk}}$$

우도비 G^2 통계량 :

$$G^2=2\sum_{ijk}O_{ijk}\log\left(\frac{O_{ijk}}{E_{ijk}}\right)$$

위의 두 통계량은 근사적으로 χ^2분포를 따르며 자유도는 다음과 같다.

자유도=칸의 갯수
−적합된 모형에서의 모수의 수

위 가설의 검정방법으로는 통계량의 값이 크면 귀무가설을 기각한다. 즉 귀무가설에서 설정한 대수선형모형이 자료에 적합하지 않다는 의미이며 따라서 자료를 적절히 설명할 수 있는 다른 모형을 설정하여 재분석하여야 할 것이다. 덧붙여서, 4차원 이상의 범주형 자료분석은 3차원 자료의 분석을 확장하여 실시한다.

2. 사례분석

가. 사례분석 1

최근 수년간 한국사회학지에 실린 논문들의 자료분석 형태를 살펴보면, 많은 경우에서 2차원 분할표를 사용하고 있다. 이러한 논문들 중 박숙자(1989)는 한국 노동시장에서의 남녀 성차별에 관한 연구를 하기 위하여 기업체의 규모와 직원선발방법 또는 선발기준을 성별에 따

라 비교하는 분할표를 작성하여 비교분석하였는데 이 표들을 정리하면 다음과 같다.

Table 1. Method of Selection by Size of Company

남녀직원의 기업체 규모별 선발방법

Method of Selection	Male		Female	
	Large	Small	Large	Small
Examination	2	1	2	0
Recommendation/interview	8	7	13	9
Exam, recommendation, interview	36	14	31	18
	χ^2=1.84(n.s.)		χ^2=1.85(n.s.)	

Source : Park, Sook-Ja, "Structural Aspects of Sex Discrimination in the Korean Labor Market -the Screening Process of Recruitment", Korean Journal of Sociology, Vol. 23, Summer, 1989, p. 57.

Table 2. Standard of Selection by Size of Company

남녀직원의 기업체 규모별 선발기준

Standard of Selection	Male		Female	
	Large	Small	Large	Small
Record/education	16	7	17	5
Birth place	1	0	1	6
Appearance	1	1	6	7
Professional sense	16	9	10	5
Specialty	2	3	7	2
Personality	8	1	2	0
Home background	0	0	1	0
Recommendation	0	0	0	2
	χ^2=4.55(n.s.)		χ^2=10.21(n.s.)	

Source : Park, Sook-Ja, "Structural Aspects of Sex Discrimination in the Korean Labor Market -the Screening Process of Recruitment", Korean Journal of Sociology, Vol. 23, Summer, 1989, p. 57.

위의 표들을 살펴보면, 3차원 범주형 자료임에도 불구하고 성별에 따라 각각 2차원 범주형 자료로 나누어 분석하였으며 4개의 χ^2검정 통계량 값들은 모두 유의하지 않다고 결론이 났다. 그러나 서론에서 언급했듯이 이러한 분석방법은 성별과 선발방법(또는 선발기준)간의 관계성, 또는 성별과 기업체 규모간의 관계성을 전혀 알 수 없다는 단점이 있으며, 특히 이 논문의 주요 논점인 성차별 측면에 대해서는 언급을 할 수 없다.

위 논문의 필자는 Table 1에서 성별에 따라 분포상의 차이가 있다고 해석하였으며, 더 나아가 남자직원보다 여자직원을 채용할 때 고용주의 자의성이 더 많이 개입될 수 있는 가능성을 암시한다고까지 주장하였다. 그러나 과연 이러한 해석이 옳은 것인가를 알아보기 위하여 위 자료를 대수선형모형으로 분석해 보자.

성별과 선발방법, 그리고 기업체 규모라는 세가지 범주형 변수로 이루어진 2×3×2 분할표를 세 변수가 모두 독립인 대수선형모형

$$\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)}$$

으로 Table 1의 자료를 적합시켜 보았다. 그 결과 우도비 G^2 통계량 값은 5.74이며, 자유도 7의 χ^2분포와 비교하면 p값이 0.57로 매우 높았다. 따라서 우리는 세 범주형 변수가 모두 독립적이라고 할 수 있으며 성별과 선발방법, 그리고 성별과 기업체 규모가 각각 독립적이고 또한 성별간의 차이도 존재하지 않는다는 결론을 유도할 수 있다. 따라서 위 논자의 주장은 비약된 오류적 판단이라 할 수 있다.

다음으로 선발기준에 대하여 살펴보자. 위 논자는 Table 2를 통하여 남녀 모두에 있어 직업의식과 성적 및 학벌이 중요한 선발기준이지만 선발기준의 하나인 인상 및 용모는 남자직원의 경우 빈도수가 총 2명(3.1%)인 반면, 여

자직원의 경우에는 총 13명(20%)이므로 선발기준의 성차별적 측면을 암시한다고 하였다.

이같은 결론이 옳은 지를 규명하기 위하여 Table 2의 자료를 대수선형모형으로 적합시켜 보자. 성별과 선별기준, 그리고 기업체 규모의 세가지 범주형 변수로 구성된 2×8×2 분할표를 바탕으로 우선 세 변수가 독립인 대수선형모형으로 적합시켜 본 결과 우도비 G^2검정 통계량 값은 35.60으로 매우 컸으며, 자유도 22인 χ^2분포와 비교해 보면 p값이 0.03이었다. 따라서 세 변수가 독립적이라는 가설은 기각된다.

다음으로 위 논자의 암시적 결론에 기반하여 성별과 선발기준과의 연관성이 존재하는 대수선형모형

$$\log M_{ijk}=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)}$$

을 자료에 적합시켜 보았다. 위 모형에 대하여 우도비 G^2검정 통계량 값은 15.08이었고 자유도 15인 χ^2분포와 비교할 때 p값은 0.45였다. 따라서 성별과 선발방법에 관한 Table 1의 자료분석과는 대조적으로 성별과 선별기준과는 매우 밀접한 관계가 존재하며 위 논자의 결론은 타당하다고 인정할 수 있다. 그 외의 다른 쌍들의 연관성에 대하여 살펴보기 위하여 여러 종류의 대수선형모형을 Table 2의 자료에 적합시켜 본 결과 기업체의 규모와 나머지 두 변수들간의 상호연관성은 존재하지 않음이 나타났다.

나. 사례분석 2

한상진은 그의 논문(1987)에서 근자에 이르러 우리 사회에서 중산층이란 용어가 널리 사용되고 있음에도 불구하고 이의 실체를 밝히려는 노력이 극히 부진한 상태에 있음을 지적하고 그동안 다소 막연하게 사용된 중산층의 개념을 보다 명확히 하기 위하여 중산층의 기준을 정하고 그 규모를 추정하며 이렇게 해서 형성된 범주가 사회학적으로 의미있는 범주인가를 분석하고자 시도하였다.

Table 3. Relationship Between Attitude Toward the Cause of a Labor-Management Dispute and the Middle Stratum When Controlling Classes

계급을 통제했을 때 노사분규의 원인에 대한 태도와 중산층 여부

Attitude Due to	Working class			Old middle class			New middle class		
	Non-middle stratum	Middle stratum	Total	Non-M.S.	M.S.	Total	Non-M.S.	M.S.	Total
Government	30.6 (701)	44.8 (128)	32.2 (829)	37.6 (92)	35.9 (146)	36.5 (238)	37.8 (260)	42.1 (469)	40.5 (729)
Employer	37.4 (855)	31.8 (91)	36.7 (946)	39.6 (97)	29.0 (118)	33.0 (215)	31.1 (214)	25.9 (288)	27.9 (502)
Worker	32.0 (733)	23.4 (67)	31.1 (800)	22.9 (56)	35.1 (143)	30.5 (199)	31.1 (214)	32.1 (357)	31.7 (571)
Total	100.0 (2,289)	100.0 (286)	100.0 (2,575)	100.1 (245)	100.0 (407)	100.0 (652)	100.0 (688)	100.1 (1,114)	100.0 (1,882)
	$\chi^2=23.934$ df=2 $p<0.000$ v=0.096			$\chi^2=12.882$ df=2 $p<0.002$ v=0.141			$\chi^2=6.283$ df=2 $p<0.043$ v=0.059		

Source : Han, Sang-Jin, "An Implementation for Conceptualization of the Middle Stratum in Korea", Korean Journal of Sociology, Vol. 21, Summer, 1987, p. 136.

이 논문의 가장 핵심적 문제는 중산층으로 분류된 집단이 단순한 통계적 범주가 아니라 사회학적으로 유의미한 실천적 성격을 공유하는 범주인가, 즉 이 집단이 정치적, 이데올로기적으로 유사한 실천적 지향을 갖는가를 보여주는 것으로서 이를 위하여 위 논자는 1981년과 1986년에 실시된 경제기획원의 사회통계 조사자료 중 중산층의 기준으로서 직업, 교육, 주택, 주관적 귀속의식, 가구별 연간소득 등의 변수를 채택, 이용하였고 또 다른 자료로서 노동자와 화이트칼라, 자영업주를 대상으로 하여 논자가 1985년에 실시하였던 노동실태와 의식에 관한 전국 표본조사를 분석하였다.

논자는 우선 중산층이라 불리울 수 있는 집단을 경제적 기준과 사회적 기준을 분류기준으로 하여 비중산층과 구분하고 그 규모와 변화양상을 분석하고 있다. 그리고 이러한 중산층이 과연 비중산층과 다른 나름대로의 이데올로기적 성격을 보유하고 있는지를 Table 3과 Table 4에서 보고자 하였다.

Table 4. Relationship Between Attitude Toward the Cause of a Labor-Management Dispute and Classes When Controlling the Middle Stratum

중산층 여부를 통제했을 때 노사분규의 원인에 대한 태도와 계급과의 관계

Attitude Due to	Non middle stratum				Middle stratum			
	Working class	Old middle class	New middle class	Total	Working class	Old middle class	New middle class	Total
Government	30.6 (701)	37.6 (92)	37.8 (260)	32.7 (1,053)	44.8 (128)	35.9 (146)	42.1 (469)	41.1 (743)
Employer	37.4 (855)	39.6 (97)	31.1 (214)	36.2 (1,166)	31.8 (91)	29.0 (118)	25.9 (288)	27.5 (497)
Worker	32.0 (733)	22.9 (56)	31.1 (214)	31.1 (1,003)	23.4 (67)	35.1 (143)	32.1 (357)	31.4 (567)
Total	100.0 (2,289)	100.1 (245)	100.0 (688)	100.0 (3,222)	100.0 (286)	100.0 (407)	100.1 (1,114)	100.0 (1,827)
	$\chi^2=22.763$				$\chi^2=15.026$			
	df=4				df=4			
	$p < 0.000$				$p < 0.005$			
	v=0.059				v=0.064			

Source : Han, S. J., "An Implementation for Conceptualization of the Middle Stratum in Korea", Korean Journal of Sociology, Vol. 21, Summer, 1987, p. 137.

위의 표들은 노사분규의 원인에 대한 태도(변수 1)와 중산층 여부(변수 2), 계급(변수 3)간의 3차원 범주형 자료이다. Table 3은 계급을 통제하고 태도와 중산층 여부의 2차원 분할표로 나누어 세번의 χ^2검정을 수행하였고, Table 4는 중산층 여부를 통제하고 태도와 계급간의 2개의 2차원 분할표를 χ^2검정하였다.

논자는 Table 3에서 다음과 같은 경향을 알 수 있다고 하였다. 첫째, 중산층화된 노동자계급은 다른 집단보다 현저히 정부비관적 성향이

높고 둘째, 구중간계급일지라도 비중산층은 중간계급의 일반적 성향과는 달리 사용자에 대한 비관의식이 매우 높으며 셋째, 중산층화된 구중간계급은 다른 어느 층보다도 노사분규를 노동자 잘못으로 돌리는 경향이 강하고 넷째, 중산층화될수록 정부비관의식이 높아진다는 의미에서 중산층의 특징이 드러나는 것처럼 보이며 다섯째, 비중산층화될수록 사용자에 대한 비관의식이 높아진다는 점에서 비중산층의 어떤 특징이 드러나는 것처럼 보이고 여섯째, 신중간계급의 경우에는 중산층 여부가 태도에 뚜렷한 영향을 미친다고 말하기 어려울 것 같다. 즉 이 경우엔 중산층 여부에 관계없이 계급적 성격이 태도에 큰 영향을 미친다고 할 수 있다.

논자는 Table 3과 Table 4에서 보듯이 여러 개의 2차원 분할표로 나누어진 통계적 검증을 통하여 중산층화된 집단과 비중산층화된 집단간에 이데올로기적 성향면에서 대체로 어느 정도 차이를 보이고 있다고 보고 있다.

그러나 이 자료를 대수선형모형으로 고려한 결과 논자가 유도한 결론은 사실을 정확히 설명하고 있지 않다고 말할 수 있다. 우선 Table 3에 적절한 대수선형모형을 알아보자. 논자가 제시한 가설에 대한 모형은 계급이 통제된 상황에서 태도와 중산층 여부가 독립인 모형으로서 다음과 같다.

$$\log M_{ijk}=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)}+U_{23(jk)}$$
$$(\text{or } U_{13(ik)}=U_{123(ijk)}=0)$$

여기서, 우도비 G^2 통계량 값은 42.35이며, 자유도 6의 χ^2분포와 비교하면 p값은 0.0000이다. 따라서 위 가설은 기각되며 계급이 통제된 상황에서 두 변수는 독립적이 아니라고 설명된다.

그리고 Table 4에서 설정된 귀무가설은 중산층 여부를 통제했을 때 태도와 계급간은 독립적이다이며, 설정된 대수선형모형은 다음과 같다.

$$\log M_{ijk}=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{23(jk)}+U_{13(ik)}$$
$$(\text{or } U_{12(ij)}=U_{123(ijk)}=0)$$

여기서, 우도비 G^2 통계량 값은 38.78이며, 자유도는 8, p값은 0.0000이다. 그러므로 위 가설은 기각되어 중산층 여부가 통제된 상황에서 태도와 계급간은 독립적이 아니라고 설명된다. 이와 같은 결과는 동 논문에서 주장된 것과 다르지 않은 것처럼 보인다. 그러나 위 자료를 보다 더 적절히 설명하고 있는 대수선형모형이 존재하는지의 여부를 알아보기 위하여 가능한 모든 대수선형모형에 대하여 자료를 적합시킨 표는 다음과 같다.

Table 5. Analysis for Loglinear Models 1

대수선형모형 분석결과표 1

Models	df	G^2	p-value
$U+U_1+U_2+U_3$	12	1,601.05	0.0000
$+U_{12}$	8	1,554.82	0.000
$+U_{23}$	10	88.58	0.000
$+U_{13}$	10	1,551.24	0.000
$+U_{12}+U_{23}$	6	42.35	0.000
$+U_{13}+U_{23}$	8	38.78	0.000
$+U_{12}+U_{13}$	6	1,505.01	0.000
$+U_{12}+U_{23}+U_{13}$	4	23.01	0.0001

위 표를 자세히 살펴보면 p값이 모두 매우 작아 모든 귀무가설 하의 모형을 기각하므로 어떠한 모형도 자료를 설명하지 못하는 것을 발견할 수 있다. 자료를 구성하는 세 변수는 모두 상호 밀접한 관계가 있음을 알 수 있으며, 1차 교호작용뿐 아니라 2차 교호작용도 존재한다고 주장할 수 있다. 따라서 논자가 Table 3과 Table 4에서 통계적 검증과 함께 설

명한 것은 정확하다고 보기 어렵다.

또한 Table 3에서 태도와 중산층 여부가 관계가 있고 Table 4에서 태도와 계급간에 관계가 있으므로 중산층 여부와 계급간도 독립적 관계가 아니라고 할 수 있으며, 따라서 자료를 이루고 있는 세 변수 모두가 상호관계를 갖고 있음을 유도할 수 있다. 그러므로 논자는 세 변수간의 관계를 동시에 고려하지 않음으로써 사실을 정확히 설명할 수 없는 결론을 유도하였음을 지적할 수 있다.

이와 동일하게 동 논문중 Table 6의 자료를 분석하여 내린 결론 역시 마찬가지로 오류를 범하고 있음을 알 수 있다. Table 6은 계급(변수 2)을 통제했을 때 노사갈등과 사회질서에 대한 태도유형(변수 1)과 중산층 여부(변수 3)와의 관계를 보고 있다. 노사갈등과 사회질서에 관한 태도유형은 완전혼란, 절제된 비관, 관리가능, 완전개방 등의 네 가지 범주로 분류하였다.

Table 6. Relationship Between Attitude Toward a Labor-Management Conflict/Social Order and the Middle Stratum When Controlling Classes

계급을 통제했을 때 노사갈등과 사회질서에 관한 태도유형과 중산층 여부

Attitude	Working class			Old middle class			New middle class		
	Non-middle stratum	Middle stratum	Total	Non-M.S.	M.S.	Total	Non-M.S.	M.S.	Total
Complete disorder	6.8 (126)	9.3 (23)	7.1 (149)	12.8 (25)	17.3 (62)	15.7 (87)	15.0 (94)	19.1 (197)	17.6 (291)
Restrained pessimism	17.1 (319)	14.9 (35)	16.9 (354)	6.7 (13)	5.3 (19)	5.8 (32)	5.8 (36)	4.6 (47)	5.0 (83)
Manageable	34.3 (639)	38.7 (91)	34.7 (730)	39.5 (138)	38.6 (138)	38.8 (215)	46.0 (288)	51.8 (534)	49.6 (822)
Complete opening	41.9 (782)	36.6 (86)	41.3 (868)	41.0 (80)	38.8 (139)	39.6 (219)	33.2 (208)	24.5 (252)	27.8 (460)
Total	100.1 (1,866)	100.0 (235)	100.0 (2,101)	100.0 (195)	100.0 (358)	99.9 (553)	100.0 (626)	100.0 (1,030)	100.0 (1,656)
	$\chi^2=5.942$			$\chi^2=2.209$			$\chi^2=18.271$		
	df=3			df=3			df=3		
	$p < 0.114$			$p < 0.530$			$p < 0.000$		
	v=0.053			v=0.063			v=0.105		

Source : Han, S. J., "An Implementation for Conceptualization of the Middle Stratum in Korea", Korean Journal of Sociology, Vol. 21, Summer, 1987, p. 140.

논자는 Table 6을 통하여 첫째, 중산층화될수록 어느 계급에서나 완전혼란의 보수적 태도가 증가하며, 동시에 구중간계급을 제외하면 관리가능의 태도도 증가한다. 둘째, 비중산층화될수록 어느 계급에서나 완전개방의 태도가 증가하는 추세를 보이며, 특히 구중간계급에서는 그 경향이 노동자에 근접해 간다. 셋째, 관리가능의 점진적 개혁의 태도는 다른 어디에서보다

도 신중산층(중산층화된 신중간계급)에서 압도적으로 많다, 넷째, 중산층 또는 비중산층 내부에서 신중간계급과 구중간계급의 차이가 상당히 뚜렷하다는 사실 등을 발견할 수 있다고 하였다.

여기에서 Table 3과 Table 4의 자료를 대수선형모형으로 분석한 것과 동일하게 Table 6의 자료를 적절하게 설명하고 있는 대수선형모형을 찾아보기 위하여 모든 대수선형모형에 대한 결과를 Table 7에 나타내었다.

Table 7. Analysis for Loglinear Models 2

대수선형모형 분석결과표 2

Models	df	G^2	p-value
$U+U_1+U_2+U_3$	17	1,640.53	0.0000
$+U_{12}$	11	1,304.41	0.0000
$+U_{23}$	15	376.43	0.0000
$+U_{13}$	14	1,485.90	0.0000
$+U_{12}+U_{23}$	9	40.31	0.0000
$+U_{13}+U_{23}$	12	221.80	0.0000
$+U_{12}+U_{13}$	8	1,149.78	0.0000
$+U_{12}+U_{23}+U_{13}$	6	22.57	0.0010

위 표를 살펴보면 Table 5와 동일하게 모든 모형에 대한 p값이 매우 작아 귀무가설 하의 대수선형모형들을 모두 기각함을 알 수 있다. 따라서 어떠한 모형도 이 자료를 적절하게 설명하고 있지 않다. 그러므로 우리는 Table 6의 자료에 대하여 다음과 같은 결론을 내릴 수 있다. 즉, 세 변수 모두 상호 매우 밀접한 관계를 갖고 있으며 1차 교호작용과 2차 교호작용이 모두 존재한다. 따라서 위 자료를 분석하여 논자가 내린 결론은 매우 국지적이라고 말할 수 있다.

이상의 두 가지 사례분석을 통하여 3차원 범주형 자료를 2차원 분할표로 나누어 각기 분석하는 경우 미약한 결론을 유도할 수 밖에 없으며 그러한 결론은 오류를 범하기 쉽다는 것을 알 수 있다. 따라서 고차원 범주형 자료는 대수선형모형을 이용하여 분석하는 것이 바람직하다.

III. 결 론

대부분의 사회과학분야의 자료형태가 범주형 자료임에도 불구하고 많은 범주형 변수들로 구성된 자료가 기껏해야 1차원 또는 2차원 분할표로 표현되고 이러한 표에 의한 통계분석을 통하여 사회현상을 설명하고 있다. 그러나 이와 같은 분석방법은 각 변수의 특성 또는 두 변수간의 상호관련성에 대해서는 파악할 수 있으나 서론에서 언급한 많은 점들을 간과하기 쉽다.

따라서 본 논문에서는 다차원 분할표를 분석할 수 있는 통계적 대수선형모형에 대하여 설명하고 사례분석을 통하여 연구해 보았다. 사례분석은 최근 한국사회학지에 실린 남녀 성차별에 관한 연구논문과 한국의 중산층에 관한 논문에서 고차원 자료임에도 불구하고 2차원 자료로 나누어 분석함으로써 발생된 오류에 대하여 살펴보았다.

앞으로 많은 사회과학분야의 자료형태인 범주형 자료의 분석에는 2차원 분할표 중심의 표현에서 한 걸음 더 나아가 일목요연하게 자료를 나타낼 수 있는 고차원 분할표의 표현방식을 채택함이 바람직하며 이 때의 자료분석으로는 대수선형모형 분석기법을 활용하여 정확하게 자료를 분석함으로써 과오없는 정책수립과 정확한 예측을 도모해야 할 것이다.

참 고 문 헌

남궁 평, 홍종선, 범주형 자료의 통계분석, 서울, 자유아카데미, 1991.

박숙자, "한국 노동시장에서의 남녀고용차별", 한국사회학, 제23집 여름호, 1989, pp. 48~74.

한상진, "한국 중산층의 개념화를 위한 시도 : 중산층의 규모와 이데올로기적 성격을 중심으로", 한국사회학, 제21집 여름호, 1987, pp. 121~148.

<Summary>

Analysis of Three Dimensional Categorical Data in Social Science Research

Ok-Hee Park*

Most data in social science contain variables that have several categories, and each cell, which is the combination of categories of each variable, represents frequency. Especially, there are many data whose types are contingency tables, cross-classified by some variables in several categories. These types are called categorical data.

Even though these categorical data have many categorical variables, most articles or reports in social science have used at most two-dimensional contingency tables when alayzing this kind of data. The well-known Pearson χ^2 test statistic and likelihood ratio G^2 test statistic are used when independence is analyzed based on two-dimensional contingency tables.

While several two-dimensional marginal contingency tables are analyzed out of high-dimensional categorical data which have more than three categorical variables, this kind of analysis is only appropriate for discovering the relationship between two variables. It is therefore, maintained in this article that it is good to make use of loglinear models for the analysis of high-dimensional categorical data.

Let us assume that there are three categorical variables and each variable has several categories. The data is represented by $I \times J \times K$ contingency table. Each cell, for example (i, j, k) cell, has O_{ijk} frequencies and E_{ijk} expected value.

Now let M_{ijk} be the parameter of E_{ijk}. Then we have the following additive model similar to the ANOVA model :

$$\log M_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)} + U_{23(jk)} + U_{123(ijk)}$$

where

$$U = \frac{1}{IJK} \sum_{ijk} \log M_{ijk}$$

$$U_{1(i)} = \frac{1}{JK} \sum_{jk} \log M_{ijk}$$

$$U_{2(j)} = \frac{1}{IK} \sum_{ik} \log M_{ijk}$$

$$U_{3(k)} = \frac{1}{IJ} \sum_{ij} \log M_{ijk}$$

$$U_{12(ij)} = \frac{1}{K} \sum_{k} \log M_{ijk}$$

This additive model is called a loglinear model of full model on three-dimensional categorical

* Senior Researcher, Korea Institute for Health and Social Affairs.

data. We also consider many partial models out of the full model. We can estimate the expected value of E_{ijk} via a loglinear model which is set up based on the data. Then the suitability of the loglinear model whose test hypothesis is as follows might be tested.

H_0 : The loglinear model fits the data well.

H_1 : The loglinear model does not fit the data.

We can also use Pearson χ^2 or likelihood ratio G^2 test statistics to test the suitability. These test statistics follow approximate χ^2 distribution with a degree of freedom such as :

degree of freedom = number of cells − number of parameters in the suitable model.

If the value of the test statistic is too big, the null hypothesis is rejected, and then we search for another loglinear model and reanalyze the data. The analysis of more than four-dimensional data can also be extended.

In this article, some case of analysis of loglinear models have been done using the data in two recent articles in the Korean Journal of Sociology, in which the data had been incorrectly analyzed because loglinear models were not used.