碩士學位論文

順序形 $2\times J$ 分割表에서의 趨勢檢定에 대한 Score 選擇

濟州大學校 大學院

電算統計學科

金 奉 模

2000年 12月

順序形 $2\times J$ 分割表에서의 趨勢檢定에 대한 Score 選擇

指導敎授 金 鐵 洙

金 奉 模

이 論文을 電算統計學 碩士學位 論文으로 提出함

2000年 12月

金奉模의 電算統計學 碩士學位 論文을 確認함

審査委員長 _____

委 員 _____

委 員 _____

濟州大學校 大學院

2000年 12月

<초 록>

順序形 $2\times J$ 分割表에서의 趨勢檢定에 대한 Score 選擇

김 봉 모

제주대학교 대학원 전산통계학과

지도교수 김 철 수

범주형 자료분석이란 자료의 값이 범주형으로 측정되는 변수들 사이의 관계를 분석하는 통계적 분석 방법으로 설문조사 자료분석이나 사회조사 자료 분석 등에 많이 이용되는 통계적 분석방법이다.

범주형자료를 분석하는 방법으로는 χ^2-검정을 실시하거나 주어진 모형의 가정에서 우도비검정의 방법을 이용하여 검정통계량을 구하는 방법이 있다.

두 방법에 의한 검정통계량은 근사적으로 χ^2 분포를 따른다고 알려져 있다. 그러나, 순서형 자료인 경우에는 명목형 자료와 달리 관측값의 표본수에 따라 다른 기대치를 갖기 때문에 그에 따른 검정통계량인 M^2 검정을 많이 사용하여 왔다.

본 논문에서는 $2\times K$ 분할표에서 행자료가 순서를 갖는 경우에 그에 적합한 임의의 score를 부여하여 M^2 통계량을 산출, 최적의 p-value를 만들어내는 데 목적을 두고 있으며 본 논문에서는 난수를 사용하여 정밀한 척도를 만드는 방법을 제시하여 보았다.

χ^2나 G^2검정인 경우에는 주어진 자료를 명목형 변수로 취급, 순서형 변수를 전혀 고려하지 않은 결과를 만들어내기 때문에, M^2검정을 사용하여 여러 가지 방법으로 score를 변경시키면서 p-value를 비교하여 보았다.

그 결과 제Ⅱ장과 제Ⅲ장에 제시된 기존의 방법들은 범주간의 표본수에 대한 자료간의 거리를 고려한 결과를 보여주었으며, 이에 본 논문은 각 범주간의 효과적인 거리를 산출하기 위하여 난수를 사용하는 방법을 택하게 되었다.

따라서, 제Ⅳ장에서는 Uniform 난수를 사용한 최적의 p-value를 만들어 낼 수 있도록 일반화된 방법을 찾아낼 수 있게 되었다.

따라서, 보다 더 정밀하게 귀무가설을 기각할 수 있는 보수적인 임계치의 기각역을 구할 수 있게 되었다.

<차 례>

<표 차 례>

<그 림 차 례>

Ⅰ. 서 론

범주형 자료분석(categorical data analysis)이란 자료의 값이 범주형으로 측정되는 변수들 사이의 관계를 분석하는 통계적 분석방법으로 설문조사 자료분석이나 사회조사 자료 분석 등에 많이 이용되는 통계적 분석방법이다.

이와 같은 범주형 자료분석에서는 대부분의 경우 두 변수 또는 세 변수간에 서로 독립인가의 검정을 실시하게 되는데 독립성 검정은 관측된 자료의 독립성 가정하에서 주어진 모형에 의한 관측값(observed value)과 기대값(expected value)을 이용한 χ^2-검정을 실시하거나 주어진 모형의 가정에서 우도비검정(likelihood ratio test)방법을 이용하여 검정통계량을 구하는 방법이 있다.

두 방법에 의한 검정통계량은 근사적으로 χ^2분포를 따른다고 알려져 있다. 그러나, 순서형 자료인 경우에는 명목형 자료와 달리 관측값의 빈도수에 따라 다른 기대치를 갖기 때문에 그에 따른 검정통계량인 M^2검정을 많이 사용하여 왔다.

의학통계분석 같은 경우는 $2\times K$분할표의 행이 증가하는 순서형일 때 반응변수가 이항반응변수를 갖는 경우의 분석이 많이 있기 때문에 이에 대한 자료들에 대해서는 χ^2검정을 사용하기보다는 그에 맞는 적합한 분석방법을 권장하기도 하고 (e.g. Moses, Emerson, and Hosseini, 1984) 적합한 분석의 방법들을 제안하기도 하였다.(Agresti, 1984)

분석에 있어서 이항변량의 경우에는 로지스틱회귀분석을 사용하기도 하나, 행 score가 존재하지 않는 경우에는 순서(rank)를 부여하여 검정하는 Wilcoxen rank-sum 검정을 사용하기도 하였다.

본 논문에서는 $2\times K$분할표에서 행자료가 순서를 갖는 경우에 그에 적합한 임의의 score를 부여하여 M^2통계량을 산출, 최적의 p-값을 만들어내는데 목적을 두고 있다.

M^2검정을 사용할 경우 score는 대부분의 경우 조사자의 직관적인 판단을 사용하여 score를 만들어 검정을 시도하였으나, 본 논문에서는 난수를 사용하여 정밀한 척도를 만드는 방법을 제시하여 보았다.

Ⅱ. 범주형자료 분석방법

범주형 자료를 분석하는 방법은 새로운 의학 치료법에 대한 가치를 평가하는 문제에서부터 다양한 쟁점에 관한 사람들의 의견에 영향을 주는 인자들을 평가하는 문제에 이르기까지 여러 분야에 걸쳐 널리 사용되고 있다.

범주형 변수는 측정된 척도가 여러 범주들의 집합으로 구성되어 있는 변수를 의미한다. 예를 들어, 정치 성향은 "진보적", "중립" 또는 "보수적"으로 구분되어 측정되고 아침 식사의 종류는 "밥", "빵", 또는 "먹지 않음"등의 범주를 사용하여 측정될 수 있다. 알츠하이머병에 대한 검사 결과는 "병의 징후가 있음", "병의 징후가 없음"으로 측정될 수 있다. 이러한 경우 각 문항에 대해서 한 범주값만 관측되어야 한다.

범주형 척도들은 사회과학 분야에서 어떤 논점에 대한 사람들의 태도와 의견을 측정하기 위해 널리 사용된다. 또한 범주형 척도들은 의학 및 보건학 분야에서도 많이 이용된다. 예를 들면, 수술 후 환자의 생존 여부(예, 아니오), 상처의 증상 정도(괜찮음, 약간, 보통, 심각함) 그리고 병의 단계(초기, 중기, 말기)등을 측정하는데 사용된다.

범주형 변수들은 사회과학, 의학 및 보건학 분야뿐만 아니라 그 외 다른 분야에서도 널리 사용된다. 범주형 변수들이 많이 사용되는 또 다른 분야는 행동과학(예, 정신 질병의 진단을 위한 범주: 정신분열증, 우울증, 노이로제), 공중보건학(예, AIDS로 인해 콘돔의 사용이 증가하는지에 대한 범주: "예", "아니오"), 동물학(예, 악어의 먹이 선택에 대한 범주: 물고기, 무척추동물, 파충류), 교육학(예, 시험 문제에 대한 학생의 응답에 대한 범주: "맞음", "틀림") 그리고 마케팅 분야(예, 생산품의 세 가지 상표 중에서 소비자의 선호도에 대한 범주: "상표 A", "상표 B", "상표 C")등이다. 또한 기술적 분야인 엔지니어링분야나 품질관리 분야에서도, 어떤 품목이 표준에 적합한 지의 여부를 분류할 때도 범주형 변수들이 사용된다.

1. 범주형 자료의 구조

연구하고자 하는 모집단의 각 개체가 t개의 범주에 오직 하나씩 분류될 수 있는 경우에 그 범주는 상호배반(mutually exclusive and exhaustive)이라 한다. 이러한 모집단으로부터 임의추출된 표본이 t범주에 속할 확률은 $\{p_i\}$가 된다. 즉,

$$\{p_i\} = (p_1, p_2, \cdots, p_t)$$

여기서, $\sum_{i=1}^{t} p_i = 1$ 이다. 이제 N개의 무작위 표본을 취했다하면, t개 범주에 속하는 표본도수(sample counts) $\{x_i\}$는

$$\{x_i\} = (x_1, x_2, \cdots, x_t)$$

여기서 $\sum_{i=1}^{t} x_i = N$이며, 해당되는 각 범주의 기대도수(expected counts) $\{m_i\}$는

$$\{m_i\} = (m_1, m_2, \cdots, m_t)$$

가 되고, 여기서

$$E(x_i) = m_i, \quad i = 1, \cdots, t$$

$$m_i = Np_i$$

이다.

두 개의 범주형 변수에 의한 자료구조는 행은 한 범주형 변수, 그리고 열은 나머지 다른 한 범주형 변수에 의해 결정되는 이차원 분할표로 나타나게 된다. 만일

<표 1> 다차원분할표

구 분		C						Total
		1	2	…	j	…	J	
R	1	p_{11}	p_{12}	…	p_{1j}	…	p_{1J}	p_{1+}
	2	p_{21}	p_{22}	…	p_{2j}	…	p_{2J}	p_{2+}
	…	…	…	…	…	…	…	…
	i	p_{i1}	p_{i2}	…	p_{ij}	…	p_{iJ}	p_{i+}
	…			…	…	…	…	…
	I	p_{I1}	p_{I2}	…	p_{Ij}	…	p_{IJ}	p_{I+}
Total		p_{+1}	p_{+2}	…	p_{+j}	…	p_{+J}	$p_{++}=1$

I개의 범주를 갖는 행 범주형 변수를 R, J개의 범주를 갖는 열 범주형 변수를 C라 하면, 이들 두 범주형 변수에 의해 생성되는 이차원분할표는 다음과 같이 나타낼 수 있다.

즉, 범주형 변수 R과 C에 의해 결정되는 분할표는 $I \times J$개의 칸을 가지며

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = p_{++} = 1$$

이 된다. 또한 위 분할표에서

$$p_{i+} = \sum_{j=1}^{J} p_{ij}, \qquad p_{+j} = \sum_{i=1}^{I} p_{ij}$$

는 각각 행과 열의 주변합으로 p_{i+}는 범주형 변수 R이 i번째 행에 속하는 확률을, p_{+j}는 범주형 변수 C가 j번째 열에 속하는 확률을 각각 나타낸다.

이제 N개의 무작위 표본을 취했다고 하면, 표본도수에 의한 이차원분할표의 칸 값은 $\{x_{ij}; i=1,\cdots,I; j=1,\cdots,J\}$가 될 것이며 해당 칸의 기대도수 $\{m_{ij}\}$는

$$E(x_{ij}) = Np_{ij}, \qquad i=1,\cdots,I, \quad j=1,\cdots,J$$

가 될 것이다. 이러한 범주형 변수에 의한 자료구조는 셋 이상의 범주형 변수에 의해 생성되는 다차원분할표로 직접 확장이 가능하다.

2. 반응변수와 설명변수

많은 통계분석에서는 반응변수(response variable)와 설명변수(explanatory variable)를 구분하고 있다. 예를 들어, 회귀분석 모형은 연수입 같은 연속반응변수의 분포가 교육을 받은 연수나 근로경험의 연수 같은 설명변수들의 수준에 따라서 어떻게 변하는지를 묘사한다. 반응변수들은 종속변수(dependent variable) 또는 Y변수라고 부르고, 설명변수는 독립변수(independent variable) 또는 X변수라고 부른다.

정치 성향 같은 경우에는 반응변수로 사용되어지기도 하고 때로는 설명변수로 사용될 수도 있다. 즉, 동일한 변수라 할지라도 연구 목적에 따라 반응변수가 될 수 있고 설명변수가 될 수 있다. 범주형 반응변수들에 대한 통계모형은 반응변수들이

설명변수들에 의하여 어떻게 영향을 받는지를 분석한다. 예를 들어 정치 성향이 그 사람의 수입, 교육 수준, 종교, 나이, 성별 그리고 인종 같은 인자들에 의해 어떻게 영향을 받는지를 연구한다. 설명변수들은 범주형이거나 연속형일 수 있다.

3. 연속변수와 이산형 변수

질병현상을 더욱 정확하게 측정하기 위해 계속적으로 측정의 단위를 세분화시킬 수 있는 측정의 척도를 연속형(continuous) 변수라 하는데 뒤에서 설명할 구간척도와 비척도의 변수가 여기에 속한다. 예를 들어 두 지점 사이의 거리는 지금 현재의 측정이 얼마나 정확한 가에 상관없이 측정을 좀 더 세분화하면 할수록 보다 더 세분화된 단위로 측정할 수 있다.

하지만 측정을 더 이상 세분하는 것이 항상 가능하지는 않은 데, 이런 경우를 이산형(discrete) 변수라 한다. 예를 들어 남자의 수를 측정하는데 있어 0과 1, 그리고, 1과 2 사이에는 아무리 세분화 하려 해도 가능한 값이 없으므로 측정을 연속적으로 세분하는 것이 불가능하다. 명목척도와 순위척도의 변수가 이산형 변수에 속하는데, 이를 범주형(categorical) 자료라고도 부른다. 범주형 자료중에서 이분성으로 갈라지는 변수를 이분성 변수 혹은 양분성(dichotomous) 변수라 하며(예: 성별-남자 vs 여자) 두 가지 이상의 범주를 가지는 변수를 다분성(polychotomous) 혹은 다항성(multinomial) 변수라 한다.

4. 명목형과 순서형 변수의 구분

범주형 변수들에 대하여 두 가지 형태의 측정 척도가 있다. 범주형 척도들이 자연스럽게 순서를 갖는 경우가 많은데, 예를 들면 유산의 합법성에 대한 태도를 조사했을 때 그 반응이 (모든 경우에 반대, 특정한 경우에만 찬성, 모든 경우에 찬성)으로 나타나고, 회사의 재고 수준을 평가했을 때 결과가 (너무 낮다, 적절하다, 너무 높다)로 나타나게 된다. 또한 의학 치료에 대한 반응이 (우수, 양호, 보통, 나쁨)으로 나타나고, 환자가 정신질환을 앓고 있는지에 대한 진단 결과가 (확실함, 그럴 것

같음, 아닌 것 같은, 아님)등으로 나타나는 경우이다. 이와 같이 범주들간에 순서척도가 있는 범주형 변수들을 순서형 변수(ordinal variable)이라고 부른다.

순서가 없는 범주형 변수를 명목형 변수(nominal variable)라고 부른다. 예를 들어 종교를 (카톨릭, 유대교, 신교, 기타)로 구분하고, 교통수단을 (오토바이, 자전거, 버스, 지하철, 도보)로 구분하며 좋아하는 음악을 (고전, 컨트리, 포크, 재즈, 록)으로 구분하고, 거주지를 (아파트, 콘도, 주택, 기타)등으로 구분했을 때 이들 범주들 간에는 아무런 순서가 존재하지 않음을 알 수 있다.

명목형 변수들에 있어서, 범주를 나열하는 순서는 의미가 없다. 또한 명목형 변수에 대한 통계적 분석 방법들은 범주가 어떻게 나열되는지에 관계없이 같은 결과를 제공해야 한다. 그러나, 순서형 변수에 대한 분석방법들은 범주의 순서를 이용하게 되므로, 범주를 낮은 데서 높은 순서대로 혹은 반대로 높은 데서 낮은 데로 나열하여 분석하면 문제가 없지만, 순서를 임의로 바꾸게 되면 분석 결과가 다르게 된다.

순서형 변수의 분석을 위해 제안된 방법은 순서정보를 갖고 있지 않는 명목형 변수의 분석에는 사용할 수 없다. 그러나 명목형 변수에 대한 분석방법은 범주의 척도만을 요구하므로 명목형 또는 순서형 변수의 분석에 모두 사용할 수 있다. 하지만 이 경우에도 순서에 대한 정보를 이용하지 않기 때문에 검정력에서 많은 손실을 볼 수 있다. 그러므로 실제 척도에 적절한 방법을 적용하는 것이 바람직하다.

범주형 변수는 흔히 몸무게, 나이, 수입, 체포된 회수 등과 같은 양적 변수(quantitative variable) 와 구별하기 위해 질적변수(qualitative variable)로 부르기도 한다. 그러나 때때로 순서형 자료들을 범주들에 일정한 score를 할당하여 양적인 자료의 형태로 다루는 것이 유익할 때도 있다.

5. 척도의 종류(Scales of the Variables)

의학분야를 포함한 모든 과학분야 연구에서 사용되는 변수는 다음의 네 가지 척도 중 하나로 분류된다.

1) 명목척도(Norminal scale)

이 척도로 분류되는 변수는 측정된 현상을 상호 배타적으로 서로 겹치지 않는 상

태에서 범주(Category)로 구분하는 의미만을 가진다. 즉, 혈액형, 성별, 인종, 실험군(대조군, 치료군), 치료결과(호전, 재발, 사망)등과 같이 특정 상태를 지칭해 주는 기능 외에는 아무런 의미를 가지지 못하는 것으로 숫자가 가지는 특성인 순서의 개념이나 구간의 개념 그리고 가감승제의 수학적 연산 기능도 물론 가지지 못하는 변수들이다. 수학적으로는 가장 원시적인 형태의 변수이나 의학 연구에서는 흔히 사용되는 변수이다.

2) 순위척도(Ordinal scale)

순위척도로 분류되는 변수들은 특정 상태를 지칭해 주는 명목척도로서의 기능뿐만 아니라 각 범주간의 대소관계, 즉 서열성에 관한 정보도 내포하고 있는 변수를 말한다. 교육 정도, 사회 경제적 수준, 병리조직학적 소견(－, ±, ＋, ＋＋, ＋＋＋), 치료의 정도(반응, 중간 반응, 무반응)등 실로 많은 예가 순위 척도로 분류될 수 있는 것들인데, 역시 가감승제 같은 수학적 조작은 불가능하다.

3) 구간척도(Interval scale)

특정 상태의 지칭이나 대소관계 개념 외에도 측정치간의 구간에 의미를 부여할 수 있을 때 이러한 변수를 구간척도라 한다. 순위척도로 분류되는 병리 소견인 경우에는 '－' 와 '＋' 사이의 구간이 '＋' 와 '＋＋＋' 사이의 구간과 같다고 인정할 수 없지만, 온도의 경우와 같이 20℃와 30℃의 구간(차이)인 10℃와, 50℃와 60℃의 차이인 10℃는 본질적으로 같다고 할 수 있다. 수학적으로는 가감의 조작은 가능하지만, 승제의 조작은 불가능하다.

즉, 100℃/ 50℃ ≠ 212°F/122°F에서 보는바와 같이 이 척도로 분류되는 변수는 구간의 의미 이외에 비(ratio)의 개념은 가지고 있지 못함을 알 수 있는 데, 이는 온도라는 변수는 자연적으로 존재하는 무(natural zero point)의 상태를 수량화한 것이 아니라 특정 조건(1기압)하에서 특정 상태(물이 어는 상태)를 인위적으로 영점으로 정하였기 때문이다.

4) 비 척도(Ratio scale)

구간척도에 비해 절대 영점을 가지기 때문에 수학적으로 가장 완전한 형태의 변수들로서 가감승제등의 모든 수학적 조작이 가능하다. 즉, 연령이나 혈압등과 같은 변

수는 태어나기 직전의 상태나 사망자의 혈압처럼 원천적인 영점을 가지고 있고 또 40세는 20세에 비해 20년의 구간이 엄연히 존재하면서 동시에 2배에 해당한다는 뜻이다. 앞서 언급한 구간척도의 변수를 두 측정치간의 차이로 환산하면 이는 다시 비 척도가 된다. 지금까지 설명한 대로 모든 변수들은 네 가지 척도중의 하나로 분류될 수 있다.

6. 변수의 척도와 통계학적 분석법

모든 변수는 각기 고유의 척도를 가지지만 분석 과정에서 변수의 형태를 변환시키면 그 척도도 달라지며, 이 때, 적용해야 할 통계학적 분석법도 따라서 새로 변형된 변수의 척도에 맞게 달라져야 한다. 예를 들어 연령은 12세, 13세, … 등의 연속적인 고유의 값을 가지는 것으로 비 척도(ratio scale)에 해당되는 값이다. 그러나, 자료를 분석하는 과정에서는 연구의 목적에 따라 연령을 ~9세, 10~19세, … 등의 범주로 변수를 변환하게 되는데 이런 경우에는 원래의 연속적인 비 척도가 순위 척도(ordinal scale)로 바뀌게 된다. 즉 해당 변수의 원래의 척도에 상관 없이 분석과정에서 가지는 변형된 변수의 척도에 따라 통계학적 분석법이 결정된다는 뜻이다.

Ⅲ. 범주형자료에 대한 검정방법

범주형 자료는 반응범주에서 관찰되는 빈도수로 구성된다. 두 개의 범주형 변수 X와 Y에서 X는 I개의 수준을 , Y는 J개의 수준을 갖는다. X의 범주는 I개의 행을 나타내고 Y의 범주는 J개의 열을 나타내는 IJ개의 조합들로 이루어진다. 표의 모든 칸(cell)들은 IJ개의 가능한 결과를 표현한다. 표의 칸에 빈도수를 나타낸 표를 분할표(contingency table)라 한다. 두 변수들을 교차 분류하는 분할표를 이원분할표(two-way table), 세 개 변수를 교차분류하는 경우 삼원표등으로 부른다. I개의 행과 J개의 열을 갖는 이원표를 $I \times J$ 표라 한다.

1. 2×2분할표의 자료분석(Two-by-Two Table Analysis)

의학 및 보건학 분야의 연구에서는 수집된 자료가 2×2 table의 형태로 요약되는 경우가 흔치 않다. 역학적 연구의 결과 변수인 질병발생은 '환자군 및 대조군' 혹은 '사망의 발생 및 미발생'으로 요약되어 이분성 변수가 되는 것이 일반적이지만, 반면에 독립변수인 '요인에의 폭로여부(exposure to risk factors)'는 대개의 경우 연속변수(continuous variable)이거나 아니면 다항으로 분류되는 순위변수(multinomial ordinal variable)가 되기 때문이다.

그럼에도 불구하고 역학적 자료분석론에서 가장 기본이 되는 분석법이 '2×2 table'의 형태로 요약된 자료에 대한 이분성 자료분석법이다. 연관성 분석은 대부분의 다변량 통계 분석에서 핵심적 위치에 있으며, 이원분할표에서 비율의 차이 및 비율의 비를 분석하는 방법으로는 오즈비(odds ratio)가 있다.

1) 오즈비(Odds ratio)

일반적인 'odds'의 개념은 발생하지 않은 사상의 확률과 발생한 사상의 확률의 비(ratio)를 의미한다. 즉, 어떤 사상이 확률 P를 가지고 발생했다면 그 때의 비(ratio), $\frac{p}{q}$ ($q = 1 - p$)

를 odds라고 부른다. 따라서, odds ratio란 두 개의 확률비(odds)의 비(ratio)를 뜻한다. 이 기본 개념을 바탕으로 질병과 관련 인자간의 관계를 나타내는 환자-대조군 연구(case-control study)의 자료분석에서 'odds'라는 용어는 논하는 상황(case-control study의 유형)에 따라 다르게 정의될 수 있다. 즉, 환자-대조군 연구(case-control study)는 두 가지 유형이 있는데, 첫 째는 임상의들이 진단에 도움이 되기 위해 병력(case histories)을 조사할 때 일반적으로 사용되는 것으로 효과에서 원인에 이르기까지를 종속변수로 포함시켜서 확인하는 것이고, 둘 째 유형은 어떤 특정 인자를 선택해서 그 인자가 질병에 영향을 미치는가를 알아 보기 위해 대조군(controls)을 선정하여 비교하는 연구를 말한다.

질병의 발생 여부와 폭로여부에 따라 표본 분석에 대해 2×2분할표(contingency table)를 작성하면 다음과 같다.

<표 2> 2×2분할표 (contingency table)

	exposure	unexposure	
Disease	a	b	$a+b$
No Disease	c	d	$c+d$

상관의 측도에 대한 관심은 1880년도에 미국의 과학 써클에서 일기 시작하였다. 19세기말 Jozsef Körösy는 천연두 왁진 주사(small pox vaccination)의 효과에 대해 다방면으로 조사했는데, 이 연구에서 그는 수집된 많은 양의 자료를 해석하고 요약하기 위해서 2×2분할표에 대한 여러 가지 상관의 측도를 소개하기에 이르렀다. 이때, Körösy(1884)에 의해 제안된 여러 가지 상관의 측도들 중의 하나가 odds ratio이다. 그러나, 비록 그는 자신의 연구 자료를 해석하기 위해 odds ratio를 소개했지만 그 측도의 활용상의 해설은 언급하지 않았다. 그래서, 2×2분할표에 있는 관측된 값을 의미있게 비교하기 위한 측도로 사용하기엔 적절하지 못하여, 19세기말 이후로 오랫동안 사용되지 않았다. 그러다가 1950년대에 들어서서 Goodman 과 Kruskall(1954-1963)에 의해 소개되었다.

그들은 Pearson(1904)에 의해 제안된 χ^2검정법은 단지 독립성 여부만을 측정할 수 있기 때문에 $I\times J$분할표에서 관측치를 의미 있게 비교하는 것은 곤란하므로, $I\times J$분할표에서 독립성(independence) 가설이 기각되는 경우에도 적용될 수 있는 분할표(cross-classification)에 대한 새로운 상관의 측도를 찾는 것이 필수적이라고 느꼈다.

원래, odds ratio가 상관의 측도로 역학 연구의 자료분석에 활용된 것은 Cornfied(1951)에 의해서이다. Confied(1951)는 사망률(motality)처럼 원인 인자와 결과사상(outcome event)간

의 상관측도를 측정하기 위해 상대위험도(relative risk)의 근사치인 odds ratio를 제시했다.

또한, Woolf(1955)에 의해 combined odds ratio 추정량이 제안된 이후로는 여러 학자들에 의해 odds ratio의 활용에 대한 연구가 매우 활발히 진행되었다. Mantel과 Haenszel(1959)은 비교호작용 가정이 아닐 때도 적용할 수 있는 추정량을 제안했으며 Gart(1962)는 최우추정량(Method of Maximum Likelihood Estimator)을 적용하여 combined odds ratio 추정량인 최우추정량(Maximum Likelihood Estimator)을 제안했다. 또한, 최우추정법의 조건 중 일치성에 위반되었을 때 이용되는 조건최우추정량(conditional MLE)을 해결하는데서 오는 수리적 어려움은 Birch(1964)와 Goodman(1969)으로 하여금 보다 쉽게 계산할 수 있는 근사치를 제시하도록 했으며, Gart(1970)도 조건최우추정량의 근사치를 제안했다.

2×2 분할표에서 첫 번째 행의 성공확률을 $\pi_1 = \frac{a}{a+b}$ 으로, 두 번째 행의 성공확률은 $\pi_2 = \frac{c}{c+d}$ 로 나타내면 첫 번째 행에서 성공의 오즈(odds)는 다음과 같이 정의된다.

$$\text{odds}_1 = \frac{\pi_1}{1-\pi_1} = \frac{\frac{a}{a+b}}{1-\frac{a}{a+b}} = \frac{a}{b}$$

$$\text{odds}_2 = \frac{\pi_2}{1-\pi_2} = \frac{\frac{c}{c+d}}{1-\frac{c}{c+d}} = \frac{c}{d}$$

이 때, 두 행에서 계산된 오즈의 비율을 오즈비(odds ratio)라고 하며 다음과 같이 정의된다.

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{ad}{bc}$$

오즈비는 오즈값의 비로 음이 아닌 실수이며 두 변수 X와 Y가 독립일 때는 $\pi_1 = \pi_2$ 및 $\text{odds}_1 = \text{odds}_2$가 성립하므로 오즈비 $\theta = \frac{\text{odds}_1}{\text{odds}_2} = 1$ 이 된다. 독립성을 나타내는 $\theta = 1$ 은 두 집단을 비교할 때 기준값이 된다. 그 이유는 1을 기준으로 오즈비가 이보다 큰가 또는 작은가에 따라 어떤 유형의 연관성을 반영하기 때문이다. $1 < \theta < \infty$ 일 때는 첫 행의 오즈가 둘 째 행보다 더 크다. 예를 들면 $\theta = 4$ 인 경우 첫 행의 오즈는 둘 째 행의 4배이므로 첫 행의 성공가능성이 더 높다는 것을 나타낸다. 즉, $\pi_1 > \pi_2$이다. 이와 반대로

$0<\theta<1$ 일 때는 첫 행에서 성공가능성이 더 낮게 되며 $\pi_1<\pi_2$ 이다.

오즈비가 1.0에서 멀리 떨어져 있을수록 연관성이 더 강하게 된다. 오즈비가 4일 때는 오즈비 2일 때보다 독립성에서 더 벗어난 것을 나타내고, 마찬가지로 오즈비 0.25는 0.50보다 독립성과 더 동떨어져 있음을 의미한다.

2) 상대위험도(ralative risk)와 Odds ratio

잠복기가 길거나 만성 질환처럼 발병률이 낮은 질병(발병율 0.03%)을 연구하는 데는 코호트연구(Cohort study)를 이용하면 모든 노력이 개개인의 추적조사(follow-up)에 소비되므로 아울러 시간과 경비가 많이 소모된다. 그러므로 이러한 질병을 연구하는 데는 환자-대조군 연구(case-control study)가 적당하다.

19세기 중엽, Quetelet(1849)가 그의 연구를 분석하기 위해 제안하여 사용한 두 이항사상(bionomial event)의 비(ratio)인 $\frac{a/(a+b)}{c/(c+d)}$ 를 Cornfield가 그의 연구에서 상대위험도라는 의미로 다시 제시하여 사용해 왔지만 이 상대위험도는 환자-대조군 연구에서는 산출이 불가능하므로 이를 보완하기 위해 새로운 상관정도의 측도를 제시했다. 이것이 odds ratio이다. Odds ratio는 상대위험도의 근사치(approximation)로 상대위험도에서 유도될 수 있다.

우선 상대위험도의 개념을 설명하면

$$RR = \frac{\text{발병군/위험 노출군}}{\text{발병군/비위험 노출군}}$$

으로 이를 <표 2>에 의해 공식으로 나타내면

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a(c+d)}{c(a+b)} = \frac{ac+ad}{ac+bc} \text{ 이다.}$$

<표 2>에서 폭로율이 낮은 경우 ac는 무시해도 될 만큼 적어져서 0에 가까우므로 지워버리면

$$RR = \frac{ac+ad}{ac+bc} \cong \frac{ad}{bc}$$

로 odds ratio 공식과 같게 된다. 그러므로 odds ratio는 상대위험도의 근사치가 된다. 또한 상대위험도(relative risk)는 코호트연구(Cohort study)에서만 추정될 수 있지만, odds ratio는 코호트 연구(Cohort study)와 환자-대조군 연구(case-control study) 모두에서 추정된다.

2. 카이제곱검정법

각 칸확률이 어떤 고정된 확률값 $\{\pi_{ij}\}$와 같다는 귀무가설(H_0)을 어떻게 검정하는지 살펴보자. 칸도수가 $\{n_{ij}\}$이고 전체 표본크기가 n인 표본에서 $\{\mu_{ij}=n\pi_{ij}\}$를 기대도수라고 한다. 이것은 귀무가설이 참일 때의 기대값 $\{E(n_{ij})\}$를 나타낸다. 이 기호는 이차원 분할표에 대해 정의되었지만 다차원분할표나 하나의 범주형변수를 갖는 도수표에 대하여도 똑같이 적용할 수 있다. 이를 설명하기 위해서 동전을 n번 던지는 실험을 생각해 보자. 동전의 앞면이 나올 확률을 π라고 하고 뒷면이 나올 확률을 $1-\pi$라고 하자. 이 동전이 공정(fair)하다는 귀무가설은 $\pi=1-\pi$도수와 일치한다. 만일 H_0이 참이라면 반은 앞면이 나오고 반은 뒷면이 나오리라 기대할 수 있다.

관측된 자료가 H_0과 모순이 되는지 알아보기 위하여 표본칸도수와 기대칸도수를 비교해보자. 이차원분할표에서 H_0이 참이라면 n_{ij}는 각 칸의 μ_{ij}에 가까운 값을 가져야 하므로 $\{n_{ij}-\mu_{ij}\}$의 차이가 클수록 H_0을 기각할 수 있는 근거가 커지게 된다. 이러한 비교를 위해 사용하는 검정통계량은 대표본카이제곱분포를 따른다.

1) Pearson 통계량과 카이제곱분포

H_0을 검정하기 위한 Pearson 카이제곱통계량은 다음과 같다.

$$\chi^2=\sum\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}}$$

이 통계량은 Pearson의 곱적률상관계수(product-moment correlation)로 알려져 있는데, 영국인 통계학자 Karl Pearson이 1900년에 제안한 것이다. 이 통계량은 모든 n_{ij}가 μ_{ij}와 같을 때 최소값 0을 갖는다. 표본크기가 고정되어 있을 때, n_{ij}와 μ_{ij} 사이의 차이가 커지면 χ^2가 커져서 결과적으로 H_0을 반증하는 증거가 더 강하게 된다.

χ^2값이 클수록 H_0과 더 모순되기 때문에 이 검정법의 P-값은 귀무가설하에서 χ^2의 관측값보다 더 큰 값들이 관측될 확률이 된다. χ^2 통계량은 표본이 커지면

근사적으로 카이제곱분포를 따른다. 표본이 어는 정도 커져야 하는지를 객관적으로 정하는 것은 어렵지만 대개 $\{\mu \geq 5\}$이면 충분히 크다고 할 수 있다. P-값은 관측된 χ^2값보다 더 큰 값을 가질 확률로 카이제곱분포에서 오른쪽 꼬리 부분의 확률이다.

카이제곱분포는 df로 표시되는 자유도에 의해 분포모양이 정해진다. 카이제곱분포의 평균은 df와 같고 표준편차는 $\sqrt{2df}$와 같다. 자유도가 커질수록 이 분포는 보다 큰 값 쪽으로 이동하고 더 넓게 퍼지게 된다. 이 분포는 오직 음이 아닌 실수에 대해서만 정의되고 오른쪽으로 긴 꼬리를 갖고 있으나 df가 커질수록 종 모양(정규분포의 모양)이 된다.

2) 2×2분할표에서의 χ^2검정

범주형자료 분석의 score test 방법으로 가장 널리 애용되고 있는 것으로서 흔히 χ^2-검정법(chi-square test, 카이제곱 검정법)이 있다. 아래에서 보는 바와 같이 [위험 요인 -> 질병 발생] 의 가설을 증명하는 연구에서 위험요인 E에 폭로된 경우를 E(+), 폭로되지 않은 경우를 E(-), 질병이 발생된 혹은 질병에 걸린 경우를 D(+), 그리고 질병에 걸리지 않은 경우를 D(-)라 할 때, 전체 연구 대상 n명에 대해 조사한 결과는 다음의 표와 같이 2×2table 형태로 요약될 수 있다.

여기서 귀무가설은 [H_0: 위험요인에의 폭로 정도와 질병과는 서로 무관하다]로 설정하게 되는데, 이와같은 귀무가설 아래에서는 전체 관찰수 n과 주변도수들의 분포가 고정된 상태이므로 일정한 질병확률 $\left(\frac{n_{1+}}{n} \text{ 및 } \frac{n_{2+}}{n}\right)$과 특정 요인에의 폭로 확률 $\left(\frac{n_{+1}}{n} \text{ 및 } \frac{n_{+2}}{n}\right)$에 따라 a, b, c, d의 네 가지로 나누어 분포하게 된다.

<표 3> 2×2분할표 (contingency table)

	exposure	unexposure	Total
Disease	a	b	n_{1+}
No Disease	c	d	n_{2+}
Total	n_{+1}	n_{+2}	n

여기서 실제로 관찰된 값을 관측값(observed frequencies, o_{ij})이라 하고, 귀무가설하에서 각 칸(cell)에 들어갈 수 있는 이론적인 값을 기대값(expected frequencies, e_{ij})이라 하는데, 예를 들어 a칸에 들어갈 수 있는 기댓값은 귀무가설[H_0: $\pi_1 = \pi_2 = \pi$]하에서는 n_{+1}명이 $\pi = \frac{n_{1+}}{n}$ 의 확률을 가지면서 D(+)군에 속하게 되는 경우이므로 결국 $n_{+1} \times \frac{n_{1+}}{n}$ 이 된다.

이를 일반화시키면, 2×2table에서 특정 칸에 들어갈 기댓값은 귀무가설하에서 주변도수의 값이 고정되어 있는 경우에 다음의 공식과 같이 된다.

$$e_{ij} = \frac{\text{row total} \times \text{coulumn total}}{\text{grand total}}$$

이 때 각각의 관측값들(o_{ij})와 기대값들(e_{ij})의 차들의 합은 아래와 같은 검정통계량이 되는데 이를 피어슨의 카이제곱(Pearson's chi-square)검정통계량이라 하며, 이 값은 χ^2-분포를 따른다. 이 검정통계량은 2×2table뿐만 아니라 가로-세로의 항이 훨씬 많은 $r \times c$table의 분석에까지도 모두 이용할 수 있는 매우 훌륭한 통계분석법으로 $r \times c$table의 경우에는 $(r-1)(c-1)$의 자유도를 가지는 χ^2-분포를 따른다.

$$S^2 = \frac{\sum (o_{ij} - e_{ij})^2}{e_{ij}} \approx \chi^2(\text{df})$$

χ^2-검정법은 계산이 간편하고 그 적용범위가 넓어서 매우 유용하게 사용되고 있는 방법이다. 그러나, 전체 관찰대상의 수가 클수록, 줄(row)의 수와 칸(column)의 수가 많으면 많을수록, 즉, 자유도의 값이 클수록, 위험요인과 질병과의 관련성이 강하면 강할수록, 그리고 특정칸의 기대값이 매우 작을 경우에 χ^2-값은 커진다. 역설적으로 말하면 어떤 통계학적 검정에서 유의성을 보장받기 위해서는 표본의 수를 무작정 늘리기만 하면 된다.

3) 독립성검정

독립성검정(independence test)은 어떤 모집단으로부터 랜덤추출한 표본이 두 가지 특성에 의해 범주로 분류될 경우 두 가지 특성간에 상호 관련성이 있는지를 검정하는 방법이다. 예를 들어, 폐암과 흡연 사이에 상호 관련성이 있는지의 여부, 연령층

<표 4> 독립성검정을 위한 $r\times c$분할표

속성 A \ 속성 B	1	2	…	c	행의 합계
1	o_{11}	o_{12}	…	o_{1c}	n_{1+}
2	o_{21}	o_{22}	…	o_{2c}	n_{2+}
…	…	…	…	…	…
r	o_{r1}	o_{r2}	…	o_{rc}	n_{r+}
열의 합계	n_{+1}	n_{+2}	…	n_{+c}	n

과 정치적 성향 사이에 관계가 있는지, 체중과 혈압 사이에 관계가 있는지 등을 분석하고자 할 때 적용된다. 표본의 크기가 n이면 두 가지 속성 A, B에 의해 각각 r개와 c개의 범주로 분류될 때 랜덤추출된 표본은 어느 한 칸(cell)에 속하게 되며 아래와 같은 $r\times c$분할표($r\times c$ contingency table)를 얻게 된다.

여기서, $o_{ij}=i$번째 행의 j번째 열에 속하는 관측도수

$$n_{i+}=\sum_{j=1}^{c} o_{ij}$$

$$n_{+j}=\sum_{i=1}^{r} o_{ij}$$

< 검정의 절차 >

(1) 가설

H_0: 두 가지 특성 A와 B는 독립이다.

H_1: 두 가지 특성 A와 B는 종속이다.

(2)검정통계량

n이 충분히 클 때,

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

는 귀무가설하에서 자유도 $(r-1)(c-1)$인 χ^2분포를 근사적으로 따른다.

여기서, $e_{ij}=\frac{n_{i+}\times n_{+j}}{n}$ 이다.

(3) 기각영역

$\chi^2 \geq \chi^2((r-1)(c-1),\alpha)$

단, $\chi^2((r-1)(c-1),\alpha)$는 자유도 $(r-1)(c-1)$인 χ^2분할표에서 상위 확률값이 α인 점을 나타낸다.

3. 우도비검정법

H_0을 검정하기 위한 또 다른 통계량은 유의성검정에 대한 우도비(likelihood-ratio) 방법을 이용한 통계량이다. 이 검정법은 먼저 H_0이 참일 때 우도함수를 최대화시키는 모수값을 결정한다. 다음으로 H_0에 관계없이 일반적인 상황에서 우도함수를 최대화시키는 모수값을 결정한다. 각각 최대화된 우도함수값들의 비를 기초로 다음과 같은 검정통계량을 정의한다.

$$\Lambda=\frac{\text{모수가 } H_0 \text{을 만족할 때의 최대우도값}}{\text{모수가 제한되어 있지 않을 때의 최대우도값}}$$

이 비는 1을 초과할 수 없다. 만일 모수가 제한되어 있지 않을 때의 최대우도함수값이 훨씬 크다면 Λ값은 1보다 매우 작게 되어 H_0를 반증하는 강한 증거를 보여주게 된다. 최대우도검정법의 검정통계량은 $-2\log(\Lambda)$이다. 이 값은 음수가 아니며 Λ가 작은 값을 가지면 $-2\log(\Lambda)$는 큰 값을 갖게 된다. 로그변환을 하는 이유는 이 통계량의 분포가 근사적으로 카이제곱분포로 수렴하기 때문이다. 이차원 분할표에서 이 검정통계량은 다음 식과 같이 간단하게 정리된다.

$$G^2=2\sum n_{ij}\log\left(\frac{n_{ij}}{\mu_{ij}}\right)$$

이 G^2 통계량을 우도비카이제곱통계량(likelihood-ratio chi-squared statistic)이라고 부른다. Pearson 통계량처럼 G^2는 모든 i,j에 대하여 $n_{ij}=\mu_{ij}$일 때 최소값 0

<표 5> *Presence or absence of congenital sex organ malformation categorized by alchol consumption of the mother*

Malformation	Alchol consumption(average#drinks/day)				
	0	< 1	1~2	3~5	>= 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1
Total	17,114	14,502	793	127	38

출처 : B. I. Graubard and E. L. Korn, *Biometrics* 43 : 473 (1987)

을 가지며 G^2값이 클수록 H_0를 기각할 수 있는 강력한 증거를 보여준다.

비록, Pearson χ^2와 우도비 G^2가 서로 다른 검정통계량이지만 많은 성질을 공유하고 있으며 대개는 동일한 결론을 갖는다. H_0이 참이고 표본칸도수가 클 때 두 통계량은 같은 카이제곱분포를 따르고 수치적으로 매우 유사한 값을 갖는다.

이차원분할표에서 두 반응변수의 독립성검정을 위한 귀무가설은 모든 i, j에 대하여 다음과 같이 표현할 수 있다.

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$

즉, 독립성이 만족되면 주변확률로부터 결합확률을 구할 수 있다. H_0을 검정하기 위한 기대도수는 $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$이다. 이 때, μ_{ij}는 독립성가정 하에서 n_{ij}의 기대값이다. 일반적으로 $\{\pi_{i+}\}$와 $\{\pi_{+j}\}$는 이 기대값과 마찬가지로 미지의 모수들이다.

기대도수는 미지의 확률에 표본비율을 대입시켜서 다음과 같이 추정할 수 있다.

$$\widehat{\mu_{ij}} = n p_{i+} p_{+j} = n \frac{n_{i+}}{n} \frac{n_{+j}}{n} = \frac{n_{i+} n_{+j}}{n}$$

여기서 $\{\widehat{\mu_{ij}}\}$는 추정된 기대도수를 나타낸다. $\{\widehat{\mu_{ij}}\}$는 관측도수 $\{n_{ij}\}$와 동일한 주변행합과 주변열합을 갖지만 독립성을 만족한다.

$I \times J$분할표의 독립성검정에 대하여 Pearson과 우도비통계량은 다음과 같다.

$$\chi^2 = \sum \frac{(n_{ij} - \widehat{\mu_{ij}})^2}{\widehat{\mu_{ij}}}, \qquad G^2 = 2\sum n_{ij} \log\left(\frac{n_{ij}}{\widehat{\mu_{ij}}}\right)$$

$2 \times J$분할표인 <표 4>에 대한 χ^2와 G^2에 대한 독립성검정의 결과는 다음과 같다.

χ^2를 계산하기 위한 $\widehat{\mu_{ij}}$ 테이블에서 값을 계산하면 다음과 같다.

$$\chi^2 = \frac{(17066-17065.13888)^2}{17065.13888} + \cdots + \frac{(1-0.10849)^2}{0.10849} = 12.08205$$

자유도 $df = (5-1)\times(2-1) = 4$이므로, p-value $= 0.01675$

$$G^2 = 2\left\{17066\times \ln\left(\frac{17066}{17065.13888}\right) + \cdots + \left\{1\times \ln\left(\frac{1}{0.10849}\right)\right\}\right. = 6.2020$$

자유도 $df = (5-1)\times(2-1) = 4$이므로, p-value $= 0.18456$

이를 부록의 SAS list와 비교하면 <표 6>과 같다.

4. M^2 검정법

다른 유의성검정과 같이 카이제곱독립성검정도 심각한 제한점이 있는데 그것은 이 검정이 단순히 연관성의 정도만을 나타낸다는 것이다. 이것을 통하여 자료에 관한 모든 의문점에 대해 답을 찾아내는 것은 적합하지 않다. 따라서 이 검정의 결과에만 의존하지 말고 연관성의 성질을 연구해야 한다. 이를 위하여 카이제곱통계량을 분할하고, 잔차를 연구하고, 연관성의 정도를 나타내는 오즈비와 같은 모수를 추정하는 것이다.

χ^2과 G^2검정은 자료의 유형에 따라 제한을 받는데, 예를 들면 이들의 대표본근사를 이용하기 위해서는 표본크기에 대한 조건이 만족되어야 한다. χ^2과 G^2의 표본추출분포는 전체 칸의 수인 IJ에 비하여 표본크기 n이 상대적으로 커지면 카이제곱분포로 수렴한다. 이 수렴속도는 G^2보다 χ^2가 더 빠르다. G^2의 카이제곱분포로의 근사는 $\frac{n}{IJ}<5$ 일 때는 형편없다. I나 J가 크면 어떤 칸의 기대도수가 1만

<표 6> <표 4>에 대한 $\widehat{u_{ij}}$

	Alchol consumption(average#drinks/day)				
Malformation	0	< 1	1~2	3~5	>= 6
Absent	17065.13888	14460.59624	790.73596	126.63741	37.89151
Present	48.86112	41.40376	2.26404	0.36259	0.10849

큼이나 작더라도 χ^2통계량의 근사는 타당해진다.

χ^2나 G^2를 계산하는 데 있어서 $\left\{\widehat{\mu_{ij}} = \frac{n_{i+}n_{+j}}{n}\right\}$는 열과 행들의 주변합에 의존하지만 열과 행들이 나열된 순서와는 상관이 없다. 따라서, χ^2나 G^2은 열과 행을 임의대로 다시 나열하더라도 통계량값들은 변하지 않는다. 이는 이 검정법들이 두 변수의 범주를 명목적인 것으로 다룬다는 의미이다. 따라서, 만약 순서형인 두 변수에 대한 독립성검정을 하기 위해 이 검정통계량을 사용하게 되면 정보를 어느 정도 무시하게 되는 것이다. 일반적으로 한 변수 혹은 두 변수 모두가 순서형일 때, 검정력이 더 뛰어난 방법들이 존재한다.

1) Pearson 상관

가장 많이 사용되는 상관연구는 19세기 후반에 Karl Pearson에 의해 개발되었는데, 그는 두 변인간의 관계를 효과적으로 양화 시키는 방법을 보여주었다. 그래서 그에게 경의를 표하는 의미에서 Pearson 적률상관계수(Pearson product moment correlation)라 명명하였다.

Pearson 상관은 문자 r로 나타내고 개념상 이 상관은 다음과 같이 계산된다.

$$r = \frac{X와\ 함께\ Y가\ 함께\ 변하는\ 정도}{X와\ Y가\ 따로\ 변하는\ 정도} = \frac{X와\ Y의\ 공변량}{X와\ Y의\ 각\ 변량}$$

완전한 직선 관계일 때, X변인의 모든 변화는 Y변인의 모든 변화에 대응된다. 예를 들면, 그림에서 X값이 증가할 때마다 Y에는 완벽하게 예측 가능한 감소가 있다. 결과는 X와 Y가 항상 함께 변하는 완전한 직선관계이다. 이 경우에 공변량은(X와 Y 함께) X와 Y의 독립된 변량과 동일하고 그 공식은 1의 상관을 낳게 된다. 반대로 직선관계가 없을 때 X변인의 변화는 Y의 예측 가능한 변화에 대응되지 않고 이 경우에는 공변량이 없으므로 상관의 결과는 0이다.

2) 상관의 이해와 해석

상관은 단지 두 변인간의 관계를 나타낸다. 두 변인이 왜 관련되었는지는 설명하지 않는다. 특히, 상관은 두 변인간 인과관계의 증거로 해석되지 않으며 될 수도 없으며, 상관값은 자료에 나타난 score 범위에 의해 크게 영향을 받을 수 있다. 관계

가 얼마나 양호한지 판정할 때, 상관의 숫자값에 중심을 두도록 유혹을 받는다. 예를 들어, 0.5의 상관은 0과 1.00 사이의 중간값이므로 적당한 관계도를 나타낼 것 같다. 그러나 상관의 비율로 해석해서는 안 된다. 1.00의 상관이 X와 Y간의 100% 완전 예측 가능한 관계임을 뜻하지만 0.5의 상관은 50% 정확도의 예측을 할 수 있다는 뜻은 아니다. 한 변인이 다른 것을 얼마나 정확하게 예측하는지를 설명하려면 상관을 자승해야 한다. 그러므로 $r=0.5$의 상관은 $r^2=0.5^2=0.25$ 또는 25% 정확도를 제공한다.

상관을 해석하는 데 가장 흔히 일어나는 실수중의 하나는 상관이 반드시 두 변인간의 인과관계를 뜻한다는 가정을 한다는 것이다. 흡연은 심장질환과 관계가 있다. 당근 섭취는 좋은 시력과 관계가 있다. 이런 관계들은 담배가 심장질환의 원인이고 당근이 좋은 시력의 원인이 됨을 뜻하는가? 대답은 '아니오'이다. 인과관계가 있을진 모르지만 상관의 단순한 존재는 그것을 증명하지 않는다.

상관의 흔한 용도중의 하나는 예측을 하기 위한 것이다. 두 변인이 상관되어 있을 때 하나의 변인값을 이용해서 다른 것을 예측할 수 있다. 예를 들면 대학입학 사무관은 어떤 지원자가 잘 해낼지 단지 추측만을 할 수 있다. 그들은 다른 변인들(SAT 점수, 고등학교 성적 등)을 이용해서 어느 학생들이 성공적으로 할지를 예측한다.

이런 예측은 상관들을 근거로 한 것이다. 입학 사무관은 상관을 이용해서 단지 추측으로 얻은 것보다 더욱 정확한 예측을 할 것을 기대한다. 일반적으로 자승된 상관(r^2)은 추측 대신에 예측의 상관을 이용하여 얻은 정확성에 그 이점이 있다.

3) 순서형 자료의 독립성검정

검정통계량 χ^2와 G^2을 사용하는 독립성검정은 두 변수를 명목형 변수로 간주한다. 그러나 행 또는 열이 순서형일 때는 순서적인 개념을 활용하는 검정통계량을 사용하는 것이 더 적절하다.

(1) 독립성가설과 선형추세의 대립가설

행변수 X와 열변수 Y가 순서형일 때 흔히 추세 연관성(trend association)이 존재한다. 즉, X의 수준이 증가할 때 Y의 반응수준이 높아지거나 감소하는 경향이

있다. 이런 경우에 하나의 모수를 사용하여 이와 같은 순서형 추세연관성을 설명할 수 있다. 가장 일반적인 분석방법은 범주 수준에 score를 할당하여 선형추세(linear trend)나 상관관계를 측정하는 것이다.

$u_1 \le u_2 \le \cdots \le u_I$를 행score라고 하고 $v_1 \le v_2 \le \cdots \le v_J$를 열score라고 하자. 이 score는 범주의 수준과 같은 순서를 갖는 단조점수(monotone score)이다 이 score는 범주 사이의 거리를 나타내며 범주들 사이의 거리가 크면 범주들이 서로 더 멀리 떨어진 것으로 간주한다.

발생도수 n_{ij}를 가중치로 사용하여 구한 score의 교차곱 $u_i v_j$의 합은 $\sum_{i,j} u_i v_j n_{ij}$ 이며 이 가중교차합은 변수 X와 Y의 공분산과 연관되어 있다. 이 score를 사용하여 X와 Y사이의 Pearson 교차적률(cross moment) 상관계수를 구하면

$$r = \frac{\sum_{i,j} u_i v_j n_{ij} - \frac{(\sum_i u_i n_{i+})(\sum_j u_j n_{+j})}{n}}{\sqrt{\left[\sum_i u_i^2 n_{i+} - \frac{(\sum_i u_i n_{i+})^2}{n}\right]\left[\sum_j v_j^2 n_{+j} - \frac{(\sum_j v_j n_{+j})^2}{n}\right]}}$$

이며 이 값은 가중교차합을 표준화시킨 것이다. 이 방법은 각각의 개체에 대하여 행과 열에 score를 입력하여 계산할 수 있다. 상관계수 r의 값은 -1과 1사이이며, 두 변수들 사이의 독립성은 이 상관계수의 참값이 0이라는 것을 의미한다. 상관계수의 절대값이 클수록 자료가 독립성으로부터 선형적으로 멀어지는 것을 나타낸다.

상관계수의 참값이 0이라는 독립성가설과 상관계수의 참값이 0이 아니라는 양측대립가설을 검정하기 위한 검정통계량은

$$M^2 = (n-1)r^2$$

이다. 이 통계량은 표본상관계수 r의 크기가 증가하고 표본크기 n이 증가하면 함께 증가한다. 또한 표본크기가 크면 이 통계량은 근사적으로 자유도 $df=1$인 카이제곱의 분포를 따른다. 이 통계량이 큰 값을 가지면 χ^2와 G^2와 같이 독립성을 반증하며 이 때의 P-값은 관측값의 오른쪽 꼬리부분의 확률이 된다. 또한 $M=\sqrt{n-1}\,r$ 은 근사적으로 표준정규분포를 따른다. 이 통계량은 두 변수 사이에 상관관계의 방향성까지 함께 나타내는 단측대립가설에 적용할 수 있다.

M^2검정은 변수들을 대칭적으로 간주한다. 즉, $I\times J$표에서 열과 행을 바꾸고 score도 바꿔서 구한 $J\times I$표로부터도 동일한 M^2의 값을 얻게 된다.

(2) 순서형 검정통계량의 추가적인 검정력

독립성검정을 할 때 χ^2와 G^2은 가능한 가장 일반적인 형태의 대립가설을 고려하지만 칸확률은 어떤 특정한 형태의 종속성을 나타낼 수 있다. $(I-1)(J-1)$개의 자유도를 갖는 이 통계량들은 귀무가설보다 $(I-1)(J-1)$개나 많은 모수를 갖는 대립가설을 반영하고 있으며 이 추가적인 모수들에 대한 모든 유형의 패턴을 알아내기 위해서 개발된 것이다. 따라서 이러한 일반성을 얻는 반면에 어떤 특별한 유형의 연관성을 알아내는 것에는 민감하지 않다.

행과 열변수들이 순서형일 때는 하나의 추가 모수만을 사용하여 연관성을 설명할 수 있다. 예를 들어 검정통계량 M^2는 선형추세의 상관관계의 척도에 기초를 두고 있다. 검정통계량이 하나의 모수와 관련되어 있으면 $df=1$이 된다.

자유도 df는 카이제곱분포의 평균과 같으므로 $df=1$을 갖는 M^2는 $df=(I-1)(J-1)$인 χ^2나 G^2와 비슷한 값을 갖는 경우에 상대적으로 오른쪽 꼬리부분으로 훨씬 더 먼 곳에 위치하게 된다. 따라서 더 작은 P-값을 갖게 된다. 실제로 선형추세가 있을 때 M^2는 χ^2나 G^2와 비슷한 크기의 값을 갖는 경향이 있으므로 결과적으로 더 작은 P-값을 제공하여 더 높은 검정력을 갖게 된다. 일반적으로 어떤 특정한 유형의 종속성을 검정할 때, χ^2와 G^2 통계량은 그러한 종속성을 검정하기 위하여 고안된 통계량에 비하여 상대적으로 낮은 검정력을 갖는다.

작은 자유도를 갖는 카이제곱통계량의 또 다른 장점은 카이제곱 근사의 정확도와 관련이 있다. 소표본에서부터 보통크기의 표본에 대하여 실제 표본추출분포는 자유도가 작아지면 카이제곱분포에 더 가까워지는 경향이 있다. 따라서 많은 칸도수가 작을 때에는 χ^2나 G^2의 카이제곱근사가 보다 더 나빠지는 경향이 있다.

<표 7> 직관적 선택 score에 대한 p-value

알콜소비량	0	< 1	1~2	3~5	>= 6	r	M^2	p-value
직관적선택	0	0.5	1.5	4.0	7.0	0.01420	6.56993	0.01037

<표 8> p-value의 비교

구분	SAS result			비교치		
	DF	Value	p-value	DF	Value	p-value
χ^2	4	12.082	0.017	4	12.08205	0.01675
G^2	4	6.202	0.185	4	6.20200	0.18456
M^2	1	6.570	0.010	1	6.56993	0.01037

Ⅳ. 추세검정을 위한 Score 선택법

1. 순위변수의 추세검정(Test for trend for Ordinal Variable)

1) 순위변수의 분석

임상적 혹은 실험적 연구에서도 흔히 볼 수 있는 자료의 하나는 종속변수는 이분성 명목척도(dichotomous nominal variable)인데 독립변수(위험요인)는 2개 이상의 범주를 가지면서 순서의 개념을 가지고 있는 순위변수(ordinal variable)로 구성된 $2\times J$테이블이다.

추세에 대한 검정의 방법으로는 폭로의 위험에 대한 위험률의 증가에 대한 귀무가설의 판정으로 Cochran-Armitage-Mantel(CAM)이 많이 사용되어져 오고 있으며, 특히나 순서형 범주를 갖는 $2\times K$ 분할표의 경우에는 행score를 많이 쓰고 있다.

일반적으로 순서형 분할표인 경우에는 범주내에 score를 갖고 있지 않은 경우와 모든 범주가 score를 갖고 있는 경우, 행score의 마지막 범주에서 score를 갖고 있지 않은 경우(open-ended category)로 나눌 수 있다.

첫 번째의 경우는 순서형 범주의 행score로서 등간격의 score인 정수값(midrank)을 사용할 수 있으며, 두 번째인 경우에는 행범주의 score(midpoint)를 사용할 수 있다.(Graubard and Korn, 1987)

2) 양-반응관계 분석

<표 9> *Alternative scoring systems for solumn categories with exact one-sided P-values*

	Alchol consumption (average# drinks/day)				
	0	<1	1~2	3~5	>=6
Midpoints	0	0.5	1.5	4.0	7.0
			P-value = 0.01037		
Midranks	8557.5	24,365.5	32,013.0	32,473.0	32,555.5
			P-value = 0.55330		
Equally spaced	1.0	2.0	3.0	4.0	5.0
			P-value = 0.17639		

독립변수가 순위변수인 범주형 자료분석에서는 [H_0: 독립변수에 폭로되는 양이 변동함에 따라 그 결과로 야기되는 종속변수의 양도 변함]이라는 가설을 검정하게 된다.

여기서 독립변수는 세 개 이상의 범주로 분류되는 순위변수(ordinal variable)이며, 그 결과는 독립변수의 각 수준에 따른 질병위험의 분율(proportion) 혹은 비율(rate)로 요약된다. 이러한 순위자료의 분석론을 소위[순위변수의 양-반응 관계(dose-response relationship) 분석]이라 하는데, 어떤 두 사상간에 이러한 [양-반응 관계]가 관찰되면 이들 두 변수는 서로 인과적 관계에 있을 가능성이 매우 높기 때문에 질병의 원인을 추구하는 의학 연구에서는 매우 중요하게 취급되고 있다.

연속변수를 범주형 자료로 변환시키면 대개의 경우는 그 척도가 순위변수로 변한다. 한편 연속적 변수를 극단적으로 범주화하여 이분성 자료(예: 2×2분할표)로 만들면 통계적인 안정성은 좋아질지 몰라도 독립변수와 종속변수간의 관계를 폭넓게 보지 못하는 단점이 있다. 따라서 비율이나 분율과 같은 범주형 자료들은 $2\times J$분할표 형태로 분석하는 것이 바람직하다.

2. score 선택법

대부분의 자료에서 score의 선택은 결과에 영향을 거의 미치지 않는다. 여러 유형의 단조점수들은 유사한 결과를 보여준다. 그러나 자료가 매우 불균형한 경우, 즉, 어떤 범주들이 다른 범주보다 더 많은 관측값을 가지고 있을 때 결과는 달라진다. <표 9>의 자료를 통해 이 경우를 살펴보자. 동일한 구간의 행score들 (1, 2, 3, 4, 5)에 대하여, 검정통계량 $M^2=1.82776$ 은 매우 약한 연관성을 나타낸다. (P-값

<표 10> 동일 구간 score에 대한 p-value

알콜소비량	0	< 1	1~2	3~5	>= 6	r	M^2	p-value
score	1	2	3	4	5	0.00749	1.82776	0.17639
	0	1	2	3	4	0.00749	1.82776	0.17639
	2	4	6	8	10	0.00749	1.82776	0.17639
	10	20	30	40	50	0.00749	1.82776	0.17639

=0.017639)

r과 M^2의 값은 동일한 구간을 유지하는 score의 변환에 대해서는 변하지 않는다.

예를 들어 <표 10>에 제시한 것처럼 score (1, 2, 3, 4, 5) 대신에 (0, 1, 2, 3, 4), (2, 4, 6, 8, 10) 또는 (10, 20, 30, 40, 50)과 같은 score들을 사용하더라도 동일한 통계량값을 얻게 된다.

score를 선택하지 않고 검정할 수 있는 방법은 자료로부터 자동적으로 score를 계산하는 방법이다. 즉, 각 관측개체에 순위를 매긴 후에 그 순위를 범주 score로 사용하는 방법이다. 한 범주에 속한 모든 개체들에 대하여 동일한 score를 사용하는데 이 score는 전체 개체에 매겨진 1부터 n까지의 순위 중에서 그 범주에 속하는 개체들의 순위의 평균값이 된다. 이 score를 중간순위(midranks)라고 부른다. 이 중간순위의 개념을 <표 9>의 알코올 소비량의 수준에 적용해 보자. 알코올 소비량의 수준 0에서 17,114명의 개체들은 1부터 17,114의 순위를 갖게된다. 따라서, 17,114명에게 중간순위 (1+17,114)/2=8557.5를 각각 할당한다. 또한 알코올 소비량의 수준이 1보다 큰 14,502명의 개체들은 17,115부터 17,114+14,502=31,616의 순위를 갖게 되므로 중간순위가 (17,115+31,616)/2 = 24,366.5가 된다. 이와 마찬가지로 마지막 세 범주의 중간순위들은 각각 32,013.0 , 32,473.0 , 32,555.5가 된다. 이 score들을 이용하여 구한 값이 0.35이고 P-값이 0.55이므로 매우 약한 상관관계를 보인다고 결론을 내릴 수 있다.

이 방법은 van Elteren(1960)이 "Wilcoxon rank sum test"를 인용하여 다음과 같이 식을 정의하였다.

$$a_j = \frac{2\left(\sum_{k=1}^{j} n_{+k}\right) - n_{+j} + 1}{2(n+1)}$$

이 때, a_j는 0과 1 사이에 존재하게 되나 그 결과는 중간순위를 사용했을 때와

<표 11> midrank score에 대한 p-value

알콜소비량	0	< 1	1~2	3~5	>= 6	r	M^2	p-value
midranks	8557.5	24365.5	32013	32473	32555.5	0.00329	0.35143	0.55331
a_j	0.2627	0.7480	0.9827	0.9969	0.9994	0.00329	0.35143	0.55331

동일한 결과를 갖게 된다. 위 식을 modified ridit scores라 하여 SAS 프로그램의 옵션으로 사용하여 만들 때 만들 수 있는 식이다.

상대적으로 적은 관측값들을 가진 서로 인접한 범주들은 서로 유사한 중간순위들을 갖게 된다. 예를 들어, <표 11>의 midranks(8,557.5, 24,365.5, 32,013.0, 32,473.0, 32,555.5)나 a_j(modified ridit scores) (0.2627, 0.7480, 0.9827, 0.9969, 0.9994)는 마지막 세 범주에서 유사한 값을 갖는데, 이 범주들이 처음 두 범주들보다 상당히 적은 관측값을 갖고 있기 때문이다. 이 방법에 의하면 알코올 소비량이 1~2(범주 3)인 경우보다 수준이 6이상인(범주 5)경우와 훨씬 비슷하다고 간주하는 것인데, 결과적으로 이 가정은 매우 부적절하다. 일반적으로 조사자의 판단에 따라 범주간의 거리를 반영하는 score를 직관적으로 선택하는 것이 오히려 바람직하다고 할 수 있다.

3. 개선된 score선택법에 의한 score검정

1) 난수(random numbers)

일반적으로 "랜덤으로 선택된 수"라고 하는데 완전한 랜덤의 한 숫자를 말하지 않고 어떤 확률분포를 따르는 서로 독립된 난수의 수열이라는 의미로 많이 사용된다. 그리고, 여러개의 난수들이 있을 때 각 난수들이 갖는 확률은 서로 같아야 한다. 예를 들면, 십진법수 236에서 첫 번째 자리수인 2는 0~9 사이의 10개의 숫자중에서 1/10의 확률로서 그를 취하게 된 것이고 3과 6도 마찬가지로 각각 1/10의 확률로서 얻어진 것이다. 난수는 컴퓨터 시뮬레이션, 의사결정론, 컴퓨터 프로그래밍등에 많이 쓰이며, 구하는 방법은 합동법(congrueutial method)을 많이 쓰고 있다.

최근 수학은 컴퓨터의 발달과 더불어 자연현상, 사회현상 등의 여러 가지 문제를 합리적으로 해결하는데 많은 공헌을 하고 있다. 컴퓨터는 가감승제를 특히, 고속으로 처리하고 지금까지 실행불가능이라고 생각되어 온 많은 문제가 이것으로 간단히 처리되고 있지만 자연현상, 사회현상 등에 있어서 결정론적 혹은 확률론적으로 표현된 문제 가운데에는 이와 같은 컴퓨터를 쓰더라도 해답이 얻어지지 않는 것이 상당히 많다. 이러한 문제들을 해결하려면 많은 계산 시간이 요구되거나 대량의 기억용량이 필요하여 문제 해결에 있어서의 경제성뿐만 아니라 현재 개발된 컴퓨터로 개발된 경우도 있고 처음부터 문제를 수식화 즉, 모형화 시킬 수 없는 경우도 있다.

이러한 요구에 부응하여 최초에는 "결정론적인 문제를 난수를 써서 풀이하는 방법"으로부터 출발하여 확률적 오차를 갖는 근사 해법으로 Monte Carlo 법이 등장하였다. 이것은 어떤 문제에 있어서 방정식을 세우기가 곤란하거나 방정식을 세우더라도 풀기가 곤란한 경우에 현상의 모델을 작성해서 현상의 모습을 조사하려는 방법으로 이것은 통계학에 있어서 모집단의 각 요소를 하나 하나 점검하는 것은 거의 무한의 시간과 노력을 요하므로 표본을 점검하면 간단하다는 표본론과 생각하는 방법이 같다.

자연 현상의 시뮬레이션(simulation), 표본의 추출, Monte Carlo법 등에서 필요불가결한 것이 난수이다. 난수란 어떤 특정한 확률분포를 따르는 확률변수의 표현치라고 보여진다. 난수를 Uniform난수와 특수난수(예, 정규난수, 이항난수)로 크게 나눌 수 있다.

1부터 10000까지의 정수값을 갖는 난수(이산형 균일분포)를 생성하기 위해서는 서로 구분이 안 되는 구슬 10000개에 1부터 10000까지의 숫자를 적어 잘 섞은 후 복원추출(random sampling with replacement)하여 난수를 생성한다. 그러나, 이 실험은 시간도 많이 걸릴 뿐 아니라 잘 섞는 과정 또한 쉬운 일이 아니므로 많은 값을 필요로 하는 통계적 실험에는 적합하지 못하다.

'백만난수표(Rand Corporation, 1955)'와 같이 잘 만들어진 난수표를 이용하여 난수를 생성할 수 있겠지만 표를 이용하는 것은 시간이 많이 걸리고 생성개수가 표에 나타난 난수의 개수보다 큰 경우에는 사용이 애매해지는 단점이 있다. 뿐만 아니라 통계적 실험을 위하여 전산프로그램에 포함시키기에는 표의 전체값들을 저장하여야 하기 때문에 매우 비효율적인 방법이 된다.

이러한 문제를 해결하기 위하여 전산 알고리즘을 이용한 난수생성법을 사용하게 되며 일반적으로 순환공식(recursive formula, 앞의 값의 함수 형태로 다음 값을 구하는 결정적 모형식)에 의하여 생성이 이루어진다. 이것은 순수한 의미의 난수라고 할 수 없다. 왜냐하면 난수란 표준균일분포를 하는 모집단으로부터의 확률표본의 결과값이 되기 때문에 표준균일분포에 따르고 자료들 사이에 독립성이 만족되어야 하기 때문이다. 그러므로 순환공식을 이용하는 전산 알고리즘에 의하여 생성된 난수는 엄밀히 말하면 의사난수(pseudo-random number)인 것이다. 생성된 의사난수의 형태 및 성질이 실제 난수의 형태와 성질과 아주 가깝다면 난수의 역할을 한다고 보아도 무방할 것이다. 이러한 실제 난수와 차이가 거의 없는 의사난수 생성 알

고리즘을 구하는 것이 확률변수값 생성의 첫 번째 단계가 된다.

실제 난수와 차이가 거의 없는 의사난수를 생성하는 전산 알고리즘을 난수생성자(random number generator)라고 부른다.

좋은 난수생성자는 생성된 난수가 실제 난수의 성질과 형태를 가질 수 있도록 다음의 네가지 조건을 만족하여야 한다.

① 생성된 난수들의 분포가 표준균일분포에 따라야 한다.

② 생성된 자료들 사이에 독립성이 있어야 한다.

③ 동일한 난수 수열의 재생성이 가능하여야 한다.

④ 생성에 효율적이어야 한다.

순환공식에 의한 난수생성 알고리즘의 효시는 1951년 Von Newmann과 Metropolis에 의하여 제안된 중앙제곱법(mid-square method)이다. 중앙제곱법은 하나의 초기치(seed-value)를 필요로 하며, 표준균일분포에 따르는 난수생성을 목적으로 한다. 그러나, 이 방법은 생성 시간이 많이 걸리고, 영에 가까운 값이 생성 중간에 나타나면 난수의 형태가 나빠지며 가장 큰 문제는 발생되는 난수가 0으로 수렴하는 성질이 있어 오늘날에는 사용이 되지 않고 있다. 중앙제곱법의 생성 알고리즘은 다음과 같다.

<중앙제곱법을 이용한 난수 생성 알고리즘>

단계 1] 초기화 단계 : 초기치로 임의의 4자리의 자연수 x_0을 선정한다.

단계 2] 반복 생성 단계 : n을 0, 1, 2, … 으로 변화시키며 다음의 수식으로 난수를 생성한다.

$$x_{n+1} = (x_n^2\text{의 중앙 4자리의 수})$$

다음의 나눗셈에 의하여 0부터 9999까지의 정수값을 0부터 1까지의 난수로 변환시킨다.

$$u_{n+1} = x_{n+1}/10000$$

요즘에는 레머(Lehmer, 1951)에 의하여 제안된 선형합동법(linear congruential method)이 가장 많이 사용되고 있다. 선형합동법은 다음과 같은 단순한 계산으로 이루어져 생성에 효율적이다.

$$x_n = a \cdot x_{n-1} + c \pmod{m}, \quad n = 1, 2, 3, \cdots$$

여기서 a, c, m은 양의 정수이며 $a, c, x_0 < m$을 만족해야 하고, mod는 법수 연산을 의미한다. a는 곱하는 값이므로 승수(multiplier)라고 부르고, c는 더하는 값이므로 가수(increment), m은 법수(modulus)라고 부른다. 선형합동법을 이용하여 0과 1 사이의 난수를 생성하기 위한 기본 알고리즘은 다음과 같다.

<선형합동법을 이용한 난수 생성 기본 알고리즘>

단계 1] 초기화 단계 : a, c, m 값을 결정한다.

생성될 수열의 초기값(seed value) x_0를 정해준다.

단계 2] 반복 생성 단계 : n을 1, 2, … 으로 변화시키며 다음의 수식으로 난수를 생성한다.

$$x_n = a \cdot x_{n-1} + c \pmod{m}$$

다음의 나눗셈에 의하여 0부터 ($m-1$)까지의 정수값을 0부터 1 사이의 난수로 면환시킨다.

$$u_n = \frac{x_n}{m}$$

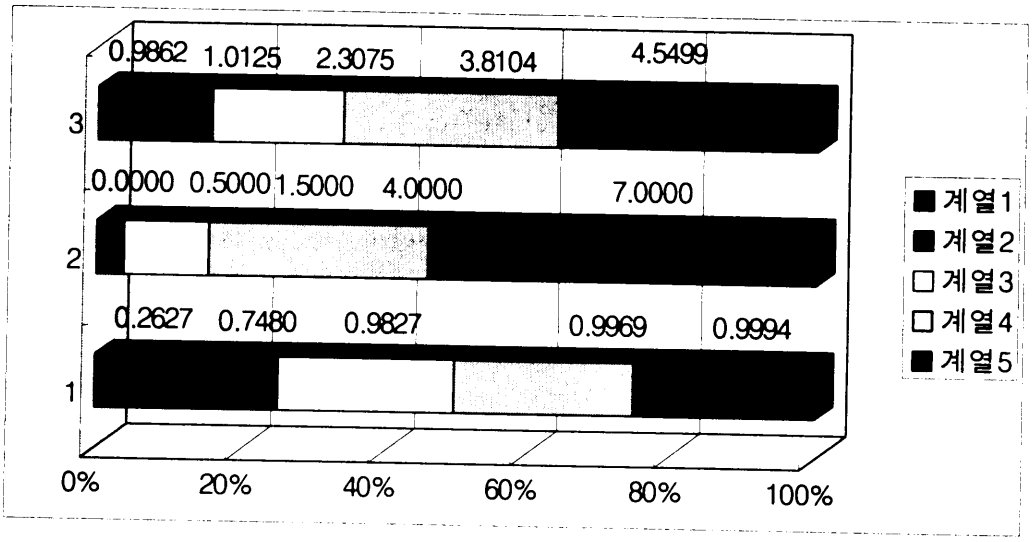
단계 3] 통계적 실험 단계 : 생성된 난수 수열을 통계적 실험에 사용한다.

2) 개선된 방법

순위변수를 갖는 $2 \times K$분할표의 개선된 score test의 방법은 앞에서 기술한 바와 같이 χ^2검정이나 우도비검정의 경우에는 주어진 분할표를 명목형변수로 취급하기 때문에 순서의 특징을 나타내지 못한다. 그러나, M^2인 경우에는 행score를 어떻게 부여하느냐에 따라 척도가 달라지는 특징을 가지고 있다.

따라서, 여러 가지 방법으로 score를 부여하여 보았으나, 표본수를 고려한 행score를 직관적으로 부여할 때 최적의 결과를 만들어낼 수 있었으며, 기존의 방법들이 행 score간의 간격의 비에 따라 최적의 p-value를 만들어내는 데 착안하여, 본 논문은 범주간의 최적의 거리를 찾아낼 수 있도록 하는 데 초점을 맞추었다.

시뮬레이션의 방법은 먼저 등간격(Equally spaced)의 범주척도를 부여한 다음, 등간격의 score를 기준으로 난수를 발생시켜 나가는 데 있다. 난수 발생의 방법은 범주척도의 각 구간에 대하여 동일확률의 구간을 만들어 낼 수 있도록 MATLAB에



<그림 1> score 척도간 간격의 비교

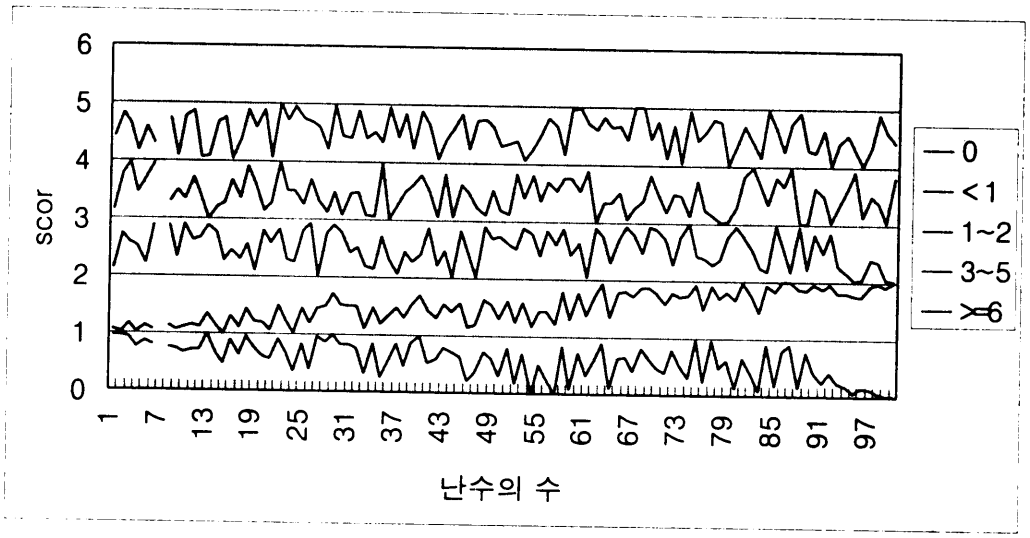
내장되어 있는 Uniform 난수를 사용하였으며, $2\times J$ table인 경우에는 부록에 있는 MATLAB Program list에 나타낸 경우와 같이 J개의 난수를 발생시켜 난수가 만들어질 때마다 동시에 J개의 범주척도에 대하여 상관계수 r를 계산하고 동시에 M^2와 p-value를 계산한다.

본 논문을 처리하는 데는 통계 Package인 SAS와 MATLAB을 사용하였다.Score Test를 위한 MATLAB 파일인 M-file은 Pearson 상관계수부분을 Script file로 처리하여 다른 프로그램에서 수시로 호출하여 사용할 수 있도록 만들었으며, 다른 하나는 M^2를 계산하는 부분이다.

결과는 엑셀을 이용하여 분석을 수행하였으며, 수행의 결과 범주척도에 의한 score의 추세선은 직관적인 score와 유사한 형태를 보여 주었으며, 각종 검정의 결과는 SAS를 가지고 수행하였을 때와 동일하게 출력함을 확인할 수 있었다.

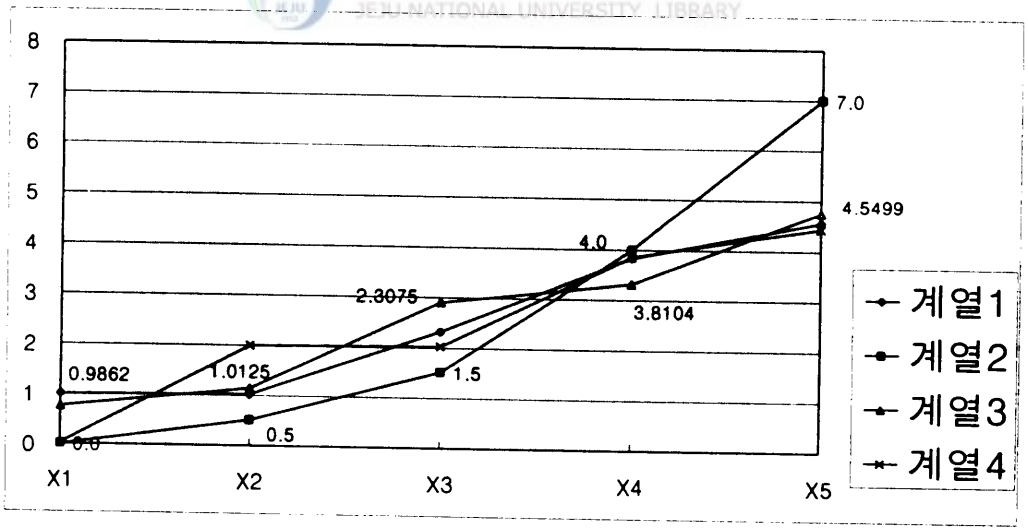
프로그래밍은 부록에 주어진 형태로 프로그램이 이루어졌으며, 가능하면 최적의 난수발생의 개수를 임의로 정할 수 있도록 만들었으며, 최소 100쌍의 난수를 만들어 내면 최적의 score를 구할 수 있었다.

자료는 <표 5> *Presence or absence of congenital sex organ malformation categorized by alchol consumption of the mother*를 사용하여 앞의 자료들과 p-value를 비교한 다음,



<그림 2> 난수를 이용한 score간 거리의 비교

<표 13> *Data from Devore and Peck(1993) on TV viewing and physical fitness*를 사용하여 그 유용성을 확인하였다.



<그림 3> <표 4>의 score에 대한 추세 경향

<표 12> <표 4>의 엑셀 분석 결과

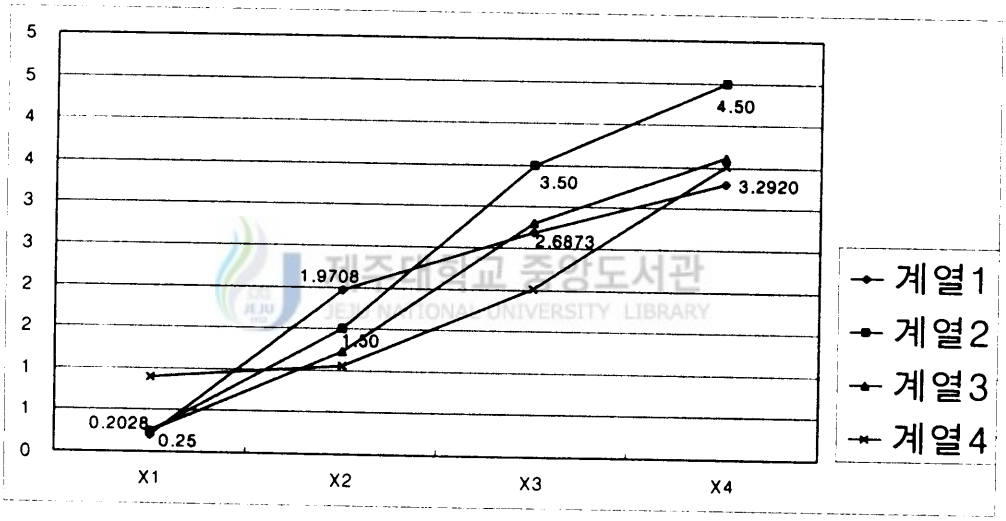
No	0	< 1	1~2	3~5	>= 6	r	M^2	p-value
1	0.9862	1.0125	2.3075	3.8104	4.5499	0.0168	9.1463	0.0025
2	0.9485	1.0745	2.1236	3.1467	4.4146	0.0165	8.9134	0.0028
3	0.9537	1.0221	2.7384	3.7775	4.8297	0.0164	8.7993	0.0030
4	0.9452	1.1721	2.5950	3.9773	4.6264	0.0151	7.4497	0.0063
5	0.7634	1.0191	2.5134	3.4581	4.1807	0.0146	6.9842	0.0082
6	0.8613	1.1363	2.2243	3.7106	4.5976	0.0146	6.9680	0.0083
7	0.7957	1.0702	2.8906	3.9567	4.3030	0.0144	6.7528	0.0094
8	0.0000	0.5000	1.5000	4.0000	7.0000	0.0142	6.5700	0.0100
9	0.7563	1.1313	2.8977	3.3053	4.7369	0.0139	6.2786	0.0122
10	0.7314	1.0569	2.3352	3.5027	4.0821	0.0137	6.1390	0.0132
11	0.6765	1.0986	2.9139	3.3532	4.7811	0.0135	5.9704	0.0145
12	0.7184	1.1447	2.6214	3.7182	4.8642	0.0135	5.8947	0.0152
13	0.7193	1.0999	2.6697	3.3944	4.0684	0.0132	5.6857	0.0171
14	0.9971	1.3540	2.8953	3.0044	4.0907	0.0131	5.5697	0.0183
15	0.6492	1.1345	2.7660	3.2111	4.6803	0.0127	5.2906	0.0214
16	0.4881	1.0007	2.2841	3.2917	4.7534	0.0124	5.0281	0.0249
17	0.8853	1.3009	2.4619	3.6823	4.0424	0.0123	4.9306	0.0264
18	0.6271	1.1021	2.3001	3.3681	4.3988	0.0123	4.9103	0.0267
19	0.9522	1.4428	2.5584	3.9218	4.8989	0.0122	4.8356	0.0279
20	0.7603	1.2150	2.1061	3.6035	4.5769	0.0121	4.7825	0.0288
21	0.6155	1.1910	2.8132	3.1421	4.8842	0.0120	4.6980	0.0302
22	0.5647	1.0720	2.5839	3.2859	4.0608	0.0120	4.6619	0.0308
23	0.9100	1.4974	2.8230	3.9851	4.9943	0.0116	4.3502	0.0370
24	0.6695	1.2275	2.3124	3.5104	4.7092	0.0114	4.2607	0.0390
25	0.3462	1.0195	2.2537	3.4938	4.9480	0.0110	3.9771	0.0461
26	0.8327	1.4593	2.8095	3.2558	4.7418	0.0108	3.8279	0.0504
27	0.3935	1.1856	2.9449	3.7023	4.6944	0.0105	3.5825	0.0584
28	0.9617	1.4624	2.0140	3.3214	4.6150	0.0104	3.5444	0.0597
29	0.8704	1.4943	2.7544	3.1404	4.2253	0.0101	3.3457	0.0674
30	0.9912	1.7144	2.9146	3.4798	4.9763	0.0098	3.1073	0.0779
31	0.8215	1.5177	2.7793	3.0848	4.4596	0.0097	3.0650	0.0800
32	0.8352	1.4945	2.4245	3.4609	4.4204	0.0095	2.9682	0.0849
33	0.7616	1.5102	2.5272	3.4715	4.8988	0.0094	2.8965	0.0888
34	0.3173	1.1031	2.1879	3.0961	4.4033	0.0092	2.7649	0.0964
35	0.8459	1.4742	2.1484	3.0714	4.5139	0.0092	2.7448	0.0976
36	0.2685	1.1961	2.7276	3.9876	4.3415	0.0092	2.7338	0.0982
37	0.5226	1.3255	2.3271	3.0138	4.9536	0.0091	2.6922	0.1008
38	0.8479	1.4690	2.0746	3.2751	4.4426	0.0091	2.6904	0.1010
39	0.4467	1.3139	2.4610	3.4962	4.8462	0.0090	2.6579	0.1030
40	0.8722	1.5301	2.2806	3.6157	4.2435	0.0090	2.6221	0.1054
41	0.9701	1.6853	2.4159	3.7800	4.8920	0.0090	2.6165	0.1058
42	0.5122	1.4357	2.8681	3.5159	4.6606	0.0089	2.5600	0.1096
43	0.5912	1.3065	2.2260	3.0673	4.0562	0.0088	2.5198	0.1124
44	0.7820	1.5441	2.4762	3.7942	4.3873	0.0088	2.5034	0.1136
45	0.7153	1.4051	2.0168	3.0558	4.5766	0.0086	2.4307	0.1190
46	0.6232	1.5544	2.8360	3.6220	4.8518	0.0086	2.4276	0.1192
47	0.2269	1.1538	2.4319	3.4891	4.2408	0.0086	2.4014	0.1212
48	0.3607	1.2061	2.0121	3.2472	4.7330	0.0085	2.3301	0.1269
49	0.7233	1.6357	2.9132	3.1120	4.7601	0.0084	2.2716	0.1318
50	0.6177	1.5363	2.6859	3.5401	4.6246	0.0083	2.2378	0.1347

<표 13> *Data from Devore and Peck(1993) on TV viewing and physical fitness*를 사용하여 그 유용성을 확인하였다.

<표 13> *Data from Devore and Peck*(1993) *on TV viewing and physical fitness*

Physical fitness		Hours of TV viewed 0	1 ~ 2	3 ~ 4	5 +	Total
Not fit	1	147	629	222	34	1032
Fit	2	35	101	28	4	168
Total		182	730	250	38	1200

출처 : Shiva Gautam, *Biometrics* 53 : 1168 (1987)



<그림 4> <표 13>의 score에 대한 추세 경향

<표 14> <표 13>의 엑셀 분석 결과

No	0	1 ~ 2	3 ~ 4	5 +	r	M^2	p-value
1	0.2028	1.9708	2.6873	3.2920	-0.0714	6.1054	0.0135
2	0.2844	1.8439	2.6739	3.3418	-0.0712	6.0750	0.0137
3	0.2714	1.8656	2.9827	3.3143	-0.0710	6.0387	0.0140
4	0.1389	1.7680	2.7349	3.5534	-0.0708	6.0114	0.0142
5	0.1988	1.4983	2.1911	3.0534	-0.0704	5.9372	0.0148
6	0.6813	1.8392	2.6932	3.2731	-0.0702	5.9045	0.0151
7	0.2987	1.6408	2.5668	3.5414	-0.0696	5.8124	0.0159
8	0.1897	1.5751	2.7334	3.4740	-0.0695	5.7928	0.0161
9	0.4289	1.6299	2.1988	3.2713	-0.0693	5.7568	0.0164
10	0.5226	1.7159	2.7764	3.4093	-0.0692	5.7396	0.0166
11	0.4660	1.7266	2.4608	3.6166	-0.0691	5.7219	0.0168
12	0.3412	1.7176	2.6555	3.8386	-0.0691	5.7218	0.0168
13	0.5341	1.6927	2.3919	3.4516	-0.0690	5.7165	0.0168
14	0.4449	1.7275	2.4253	3.6400	-0.0690	5.7046	0.0169
15	0.9318	1.9601	2.8159	3.5625	-0.0689	5.6943	0.0170
16	0.7373	1.9084	2.4850	3.6210	-0.0689	5.6864	0.0171
17	0.8801	1.8928	2.4893	3.4635	-0.0688	5.6805	0.0172
18	0.3795	1.6288	2.6501	3.6262	-0.0688	5.6736	0.0172
19	0.8757	1.8049	2.7036	3.1750	-0.0687	5.6637	0.0173
20	0.0196	1.2126	2.0492	3.2009	-0.0685	5.6268	0.0177
21	0.3784	1.6831	2.8447	3.8295	-0.0685	5.6228	0.0177
22	0.7095	1.6072	2.4001	3.0890	-0.0685	5.6222	0.0177
23	0.6038	1.7889	2.1660	3.3358	-0.0682	5.5828	0.0181
24	0.5936	1.6124	2.7313	3.0287	-0.0682	5.5795	0.0182
25	0.3093	1.4544	2.6991	3.1472	-0.0681	5.5671	0.0183
26	0.2311	1.4235	2.3400	3.5485	-0.0680	5.5468	0.0185
27	0.1987	1.9901	2.3461	3.8580	-0.0680	5.5434	0.0186
28	0.1365	1.2319	2.1146	3.2460	-0.0678	5.5068	0.0189
29	0.4186	1.4120	2.4574	3.1133	-0.0678	5.5063	0.0189
30	0.4447	1.6405	2.8699	3.7400	-0.0677	5.5003	0.0190
31	0.0153	1.2140	2.4225	3.3567	-0.0677	5.4941	0.0191
32	0.0579	1.4154	2.1370	3.7327	-0.0676	5.4859	0.0192
33	0.6068	1.5155	2.3142	3.2618	-0.0674	5.4462	0.0196
34	0.4103	1.6029	2.8704	3.8735	-0.0672	5.4083	0.0200
35	0.7382	1.6808	2.8729	3.0839	-0.0671	5.3994	0.0201
36	0.1934	1.4514	2.3759	3.9090	-0.0671	5.3963	0.0202
37	0.6213	1.5548	2.5657	3.3527	-0.0670	5.3880	0.0203
38	0.4057	1.5678	2.6458	3.9159	-0.0669	5.3739	0.0204
39	0.0118	1.2393	2.6649	3.5874	-0.0668	5.3536	0.0207
40	0.2722	1.4387	2.1556	3.6802	-0.0668	5.3502	0.0207
41	0.2026	1.4399	2.9016	3.7911	-0.0668	5.3471	0.0208
42	0.3046	1.3705	2.6252	3.4091	-0.0668	5.3444	0.0208
43	0.1763	1.4611	2.2379	3.9455	-0.0667	5.3295	0.0210
44	0.7027	1.6756	2.8376	3.6206	-0.0665	5.2974	0.0214
45	0.5466	1.6992	2.3716	3.9517	-0.0664	5.2882	0.0215
46	0.1730	1.2731	2.1859	3.6109	-0.0664	5.2850	0.0215
47	0.4565	1.5798	2.1197	3.7009	-0.0660	5.2305	0.0222
48	0.8132	1.8744	2.4302	3.9614	-0.0659	5.2055	0.0225
49	0.0185	1.7604	2.0381	3.9623	-0.0658	5.1986	0.0226
50	0.2500	1.5000	3.5000	4.5000	-0.0641	4.9270	0.0260

V. 결 론

본 논문에서는 순서형 $2\times J$ 분할표에서의 독립성검정을 다루었다.

기존의 범주형 자료 분석에서 순서형변수에 대한 반응변수를 다루는 경우에는 분할표내에서의 범주에 대한 score를 활용한 검정의 방법을 찾기 위해 여러 가지 시도를 하게 되었다.

그 방법으로는 순서형변수를 갖는 경우에는 rank를 사용하거나 등간격(Equally spaced)의 score를 사용하기도 하였으며, 가장 일반적인 경우 범주척도가 행score를 갖고 있는 경우에는 행score를 기준으로 새로운 척도(midrank, midpoint)를 산출하여 검정을 하기도 하였다.

χ^2나 G^2검정인 경우에는 주어진 자료를 명목형 변수로 취급, 순서형 변수를 전혀 고려하지 않은 결과를 만들어 낸다, M^2검정인 경우에는 순서형 변수를 고려한 행score를 사용하여 보다 정밀한 p-value를 찾을 수 있다. 이 때, 여러 가지 방법으로 score를 변경시키면서 p-value를 비교하여 보았다.

그 결과 기존의 방법들은 범주간의 표본수에 대한 자료간의 거리를 고려한 결과를 보여주었으며, 이에 본 논문은 각 범주간의 효과적인 거리의 비를 산출할 수 있도록 난수를 사용하는 방법을 택하게 되었다.

따라서, 제4장의 시뮬레이션에서 보는 것처럼 Uniform 난수를 사용한 일반화된 방법을 찾아내어 최적의 p-value를 만들어 낼 수 있었으며, 귀무가설을 기각할 수 있는 보수적인 임계치의 기각역을 구할 수 있게 되었다.

參考文獻

구자성 (1988) "종속적인 자료에 대한 기수비(odds ratio)의 추정에 관한 연구" 연세대대학원 석사

김성태 (1986) "난수와 그 응용" 동아대학교 교육대학원 수학교육전공 석사

김정문 (1995) "2차 정방형 분할표 대칭 검정 통계량의 검정력 연구" 연세대학교 대학원 응용통계학과 석사

김종호 · 김주환 편저 (1996) 「통계적 가설 검정」 서울:자유아카데미

박경희 (1986) "Odds Ratio의 이론적 특성 및 그 활용범위에 대한 연구" 서울대학교 보건대학원 보건학석사

박태성 · 이승연 (1998) 「범주형자료분석개론」 서울:자유아카데미.

손건태 (1996) 「전산통계개론」 서울:자유아카데미

유근영 (1996) 「의학-보건학을 위한 범주형자료 분석론」 서울:서울대학교출판부.

윤미숙 (1991) "범주형자료분석의 모형적합에 관한 연구" 중앙대학교 대학원 통계학과 수리통계학석사

정광모 · 최용석 (1999) 「SAS를 활용한 범주형자료분석」 서울:자유아카데미.

최완식 (1994) 「난수발생기 "RANDU"와 "DRAND" 에 관한 분석 및 그것을 위한 일반 도구 컴퓨터 프로그램 개발」 대한공업교육학회지 제19권 제1호 80-90

통계용어사전 (1987) 한국통계학회편, 자유아카데미

허명회 (1992) 「비교연구를 위한 통계적방법론」 서울:자유아카데비.

Agresti (1990). Categorical Data Analysis, New York: Willy.

Armitage, P. (1955). Tests for linear trends in proportions. Biometrics 11, 375-389.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. Biometrics 10, 417-451.

Cox, D. R. and Hinkley, D. V. (1974). Theoretical Statistics. New York: Chapman and Hall.

D. J. Best and J. C. W. Rayner and L. G. Stephens (1998). Small-sample comparison of McCulla호 and Nair analysis for nominal-ordinal categorical data. Computational Statistics & Data Analysis 28 (1998) 217-223.

Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered $2\times K$ contingency tables. Biometrics 43, 471-476.

Kendall, M. G. (1948) Rank correlation methods. London: Griffin.

Kimeldorf, G., Sampson, R. A., and Whitaker, L. R. (1992). Min and max scoring for two-sampled ordinal data. Journel of the American Statistical Association 87, 241-247.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion analysis of variance. J. Amer. Statist. Ass., 47, 583-621.

Maclure, M. and Greenland, S. (1992). Tests for trend and dose response: Misinterpretations and alternatives. American Journel of Epidemiology 135, 96-104

Markus Neuhäuser and Ludwig A. Hothorn (1999). An exact Cochran-Armitage test for trend when dose-response shapes are a priori unknown. Computational Statistics & Data Analysis 30 (1999) 403-412.

Maura E. Stokes, Charles S. Davis and Gary G. Koch (1995). Categorical Data Analysis Using the SAS System

Shiva Gautam. (1997). Test for linear trend in $2\times K$ ordered tables with open-ended categories. Biometrics 53, 1163-1169.

Silvapulle, M. J. and Sivapulle, P. (1995). A score test against one-sided alternatives Journel of the American Statistical Association 90, 342-349

Qing Liu (1998). An Order-Directed score test for trend in ordered $2\times K$ tables. Biometrics, September 1998

Yoo K. Y. (1990). A review of test for trend: score tests and likelihood ratio tests using linear logistic model and log-linearmodel with computer programs. Korean J Epidemiol 1990; 12(2): 115-30.

<Abstract>

The Choice of Score for The Trend Test in the Ordinal $2\times J$ Table

Kim, Bong-mo

Computational and Statistical Major

Graduate School, Cheju National University

Cheju, Korea

Supervised by Professor Kim, Chul-Soo

Categorical data analysis is a statistical method for analysing the relation between variables in which we take the values of data as categorical data. It is frequently used to analyse data about a questionary and social inquiry data.

There are two important method for analysis of categorical data; chi-squared test and likelihood ratio test which uses a given model of hypothesis test.

It is known that the test statistics from two tests above follow chi-squared distribution. But unlike nominal data, ordinal one has much used M-squared test because it has expectations according to the numbers of observational value.

The object of this thesis is to suggest suitable scores, producing optimal p-value if row variables in $2\times K$ table has a kind of trend.

Chapter 2 and Chapter 3 survey standard descriptive and traditional methods of lists such as chi-squared test, likelihood test and M-squared test.

Chapter 4 introduces a new method for the choice of scores in $2\times J$ contingency table which produces an optimal p-value by using random numbers instead of traditional scores.

As χ^2 and G tests brought about the results considering given data nominal variable, not ordinal one, I compared p-values varying scores in many ways by using test in Chapter 4.

As a result, Chapter 5 showed that suggested method could the distance

between data about the numbers of sample categories. So I get the way of using random numbers for the purpose of producing optimal distance. This method took the rejection region of conservative critical value which would accurately reject null hypothesis.

附錄

SAS Program Example (1)

```
data infants;
input malform alcohol count @@;
cards;
1 0   17066  1 0.5 14464  1 1.5   788  1 4.0   126  1 7.0    37
2 0      48  2 0.5    38  2 1.5     5  2 4.0     1  2 7.0     1
;

proc freq; weight count;
   tables malform*alcohol / chisq measure cmh1;
run;
```

SAS Program Example (1) List

The SAS System 1

TABLE OF MALFORM BY ALCOHOL

MALFORM ALCOHOL

Frequency / Percent / Row Pct / Col Pct	0	0.5	1.5	4	7	Total
1	17066	14464	788	126	37	32481
	52.39	44.40	2.42	0.39	0.11	99.71
	52.54	44.53	2.43	0.39	0.11	
	99.72	99.74	99.37	99.21	97.37	
2	48	38	5	1	1	93
	0.15	0.12	0.02	0.00	0.00	0.29
	51.61	40.86	5.38	1.08	1.08	
	0.28	0.26	0.63	0.79	2.63	
Total	17114	14502	793	127	38	32574
	52.54	44.52	2.43	0.39	0.12	100.00

STATISTICS FOR TABLE OF MALFORM BY ALCOHOL

Statistic	DF	Value	Prob
Chi-Square	4	12.082	0.017
Likelihood Ratio Chi-Square	4	6.202	0.185
Mantel-Haenszel Chi-Square	1	6.570	0.010
Phi Coefficient		0.019	
Contingency Coefficient		0.019	
Cramer's V		0.019	

Sample Size = 32574
WARNING: 30% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

SUMMARY STATISTICS FOR MALFORM BY ALCOHOL

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.570	0.010

Total Sample Size = 32574

SAS Program Example (2)

```
data viewing;
input fitness view count @@;
cards;
1 0.25  147  1  1.5  629  1  3.5  222  1  4.5  34
2 0.25   35  2  1.5  101  2  3.5   28  2  4.5   4
;

proc freq;
   weight count;
   tables fitness*view / chisq cmh1;
run;
```

SAS Program Example (2) List

The SAS System 1

TABLE OF FITNESS BY VIEW

FITNESS / VIEW	0.25	1.5	3.5	4.5	Total
1	147	629	222	34	1032
	12.25	52.42	18.50	2.83	86.00
	14.24	60.95	21.51	3.29	
	80.77	86.16	88.80	89.47	
2	35	101	28	4	168
	2.92	8.42	2.33	0.33	14.00
	20.83	60.12	16.67	2.38	
	19.23	13.84	11.20	10.53	
Total	182	730	250	38	1200
	15.17	60.83	20.83	3.17	100.00

STATISTICS FOR TABLE OF FITNESS BY VIEW

Statistic	DF	Value	Prob
Chi-Square	3	6.161	0.104
Likelihood Ratio Chi-Square	3	5.930	0.115
Mantel-Haenszel Chi-Square	1	4.927	0.026
Phi Coefficient		0.072	
Contingency Coefficient		0.071	
Cramer's V		0.072	

Sample Size = 1200

SUMMARY STATISTICS FOR FITNESS BY VIEW

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	4.927	0.026

Total Sample Size = 1200

MATLAB PROGRAM

```
% file name : score.m
%=====================================================
% 2*J table에서의 피어슨 교차적률 상관계수 구하기
% Pearson cross moment correlation coefficient  Base file
%=====================================================

clear all

A = [147 629 222 34; 35 101 28 4]                    % Viewing TV Table

counti = size(A,1);
countj = size(A,2);                                  % Table의 열의 수 즉,
score의 수

countrand = 100                          % Table score의 수  즉, 동일 score의 수

randscore = rand(countrand,countj);


for(bi=1:1:counti)
   u(bi) = bi;                                       % 행 score
end


for(bj=1:1:countrand)

   for(bk=1:1:countj)
      score(bj,bk) = randscore(bj,bk) + (bk-1);      % 열 score , Ordinal score
   end

end
```

```
for(bl=1:1:countrand)

    for(bm=1:1:countj)
        v(bm) = score(bl,bm);
    end

    pearcorr;                                   % script file, pearson cross moment

    Msquare = (ntotal-1)*result(7)^2;           % Msquare 값

    table(bl,1) = bl;                           % random 의 수
    table(bl,2) = result(7);                    % r : 피어슨 상관계수
    table(bl,3) = Msquare;                      % Msquare
    table(bl,4) = (sqrt(ntotal-1)) * result(7); % M value
    table(bl,5) = 1 - chi2cdf(Msquare,1);       % p-value

    for(bm=6:1:countj+5)
        table(bl,bm) = score(bl,bm-5);          % table score
    end

    bl
    tablex(bl,:)=result;

end

wk1write('c:\My Documents\sheet',table)
wk1write('c:\My Documents\sheetx',tablex)

fx = table(:,1);
fy = table(:,3);

plot(fx,fy)
```

```
% file name : pearcorr.m
%=====================================================
% 피어슨 교차적률 상관계수
% Pearson cross moment correlation coefficient  Script file
%=====================================================

counti = size(A,1);
countj = size(A,2);

for(si=1:1:counti)
    niplus(si) = sum(A(si,:));
end

for(sj=1:1:countj)
    nplusj(sj) = sum(A(:,sj));
end

ntotal=sum(A(:));

for(si=1:1:counti)

    for(sj=1:1:countj)

        uvn(si,sj) = u(si) * v(sj) * A(si,sj);

    end

end


uniplus = 0; usqniplus = 0;
vnplusj = 0; vsqnplusj = 0;

for(si=1:1:counti)
```

```
    uniplus    = uniplus    + u(si)    * niplus(si);
    usqniplus  = usqniplus  + u(si)^2  * niplus(si);

end


for(sj=1:1:countj)

    vnplusj    = vnplusj    + v(sj)    * nplusj(sj);
    vsqnplusj  = vsqnplusj  + v(sj)^2  * nplusj(sj);

end

result(:,:)=0;

% 분자
result(1) = sum(uvn(:));
result(2) = (uniplus * vnplusj) / ntotal;

% 분모
result(3) = usqniplus;
result(4) = uniplus^2 / ntotal;
result(5) = vsqnplusj;
result(6) = vnplusj^2 / ntotal;

% correlation coefficient : r
result(7) = (result(1) - result(2)) / sqrt((result(3) - result(4)) * (result(5) -
result(6)));
```