한국통계학회논문집 제4권 3호
1997년 12월 pp. 645 - 661

범주형 자료에서 연관성 측도들의 비교 분석

홍 종 선[1], 임 한 승[2]

요 약

연속형 변수들의 상관관계와 범주형 변수들의 연관성 측도들을 비교 연구하였다. 이 연구를 위하여 연속형 변수들이며 +1에서 -1까지 완벽한 상관관계를 갖고 있는 2 변량 정규분포를 이용하여 2×2 분할표와 확장하여 일반적인 $I \times J$ 분할표를 대신하는 3×3 분할표를 생성하였다. 2 차원 분할표에서 정의된 연관성 측도들을 구하여 논의하였는데 2×2 분할표에서는 교차적비 α 통계량과 교차적비의 함수로 표현되는 Yule [1912]의 Q와 Y 통계량 그리고 상관계수 R 통계량과 R 통계량의 함수인 P 통계량을 설명하고 생성된 분할표에서 구한 통계량값을 분석하였으며, 3×3 분할표에서는 Pearson의 독립성 검정통계량 X^2의 함수로 표현되는 P, T, V 통계량과 Goodman과 Kruskal [1954]의 $\lambda_{C|R}$ 통계량과 Light와 Margolin [1971]의 $\tau_{R|C}$ 통계량을 설명하고 그 값들을 Pearson의 상관계수와 비교 분석하였다.

1. 서론

두 개의 연속형 변수의 연관성을 측정하는 통계량은 이미 많은 연구가 있었으며 그 중에서 Pearson의 상관계수 ρ가 일반적으로 많이 사용되고 있다. 반면에 범주형 변수들에 관한 연관성 측도들도 많은 연구가 진행되었는데 특히 Goodman과 Kruskal [1959]는 연관성 측도들의 연구를 역사적으로 서술하였으며, 그의 저서 [1979]에서는 연관성 측도들에 대하여 연구한 4 개의 논문들 [1954, 1959, 1963, 1972]을 재 집성하였다. 그러나 그 통계량들이 나타내는 값들의 의미는 연속형 변수들의 의미보다는 충분하지 않다(Bishop, Fienberg와 Holland [1975] 참조).

본 논문은 상관계수 ρ의 값의 범위가 -1에서 +1까지 변하는 완벽한 연관성을 나타내 주고 있는 2 변량 정규분포로부터 2 차원 분할표를 생성하여, 연속형 변수들에 관한 Pearson의 상관계수의 값과 범주형 변수들의 연관성 측도들의 값과 비교 연구를 하고자 한다. 우선 주어진 상관계수를 갖고 있는 2 변량 정규분포를 이용하여 여러 종류의 2×2 분할표를 생성하여 이항 범주형 변수의 연관성을 측정할 수 있는 5 가지의 통계량의 값을 구하여 분석하고자 한다. 그리고 $I \times J$ 분할표로 확장하고자 하는데 그 중에서 비교적 간단한 3×3 분할표를 생성하여 다항 범주형 변수의 연관성을 측정하는 또 다른 종류의 5 가지 통계량의 값을 구하여 비교 분석하였다.

1) (110-745) 서울특별시 종로구 명륜동 3-53, 성균관대학교 통계학과 부교수.
2) (150-717) 서울특별시 영등포구 여의도동 23-5, 한화경제연구원 주임연구원.

논문의 구성을 살펴보면 다음과 같다. 2 장에서는 2×2 분할표에서 정의할 수 있는 5 종류의 연관성 측도들을 열거하여 설명하였으며, 3 장에서는 2 변량 정규분포로부터 2×2 분할표를 생성하는 과정을 설명하고 이 과정을 통하여 얻은 여러 종류의 분할표로부터 구한 2 장에서 언급한 연관성 측도의 값들을 정규분포의 상관계수와 비교 분석하였다. 4 장에서는 일반적인 $I\times J$ 분할표에서 정의된 연관성 측도들을 제시하였고, 5 장에서는 2 변량 정규분포로부터 $I\times J$ 분할표를 대신하여 3×3 분할표를 생성하는 과정을 간단히 설명하고 이 과정을 통하여 얻은 연관성 측도의 값들을 정규분포의 상관계수와 비교 분석하여 설명하였다.

2. 2×2 분할표의 연관성 측도

두 범주형 변수에 의한 2×2 분할표를 먼저 정의하여 보자. 분할표상의 (i, j)칸의 관찰 도수를 x_{ij} $(i,j=1,2)$라고 하면 이에 대응하는 비율 p_{ij}에 의한 표는 다음과 같다.

<표 1> 칸을 확률로 나타낸 2 차원 분할표

		변수 2		
		1	2	
변수 1	1	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	$1-p_{1+}$
		p_{+1}	$1-p_{+1}$	

위에서 정의한 2×2 분할표의 두 변수간의 연관성을 측정할 수 있는 여러 통계량을 우선 살펴보자.

1) 교차적비 α 통계량

잘 알려진 교차적비(cross product ratio, odds ratio) α는 p_{ij}에 대한 비율로 다음과 같이 정의된다.

$$\alpha=\frac{p_{11}/p_{12}}{p_{21}/p_{22}}=\frac{p_{11}p_{22}}{p_{12}p_{21}} .$$

교차적비의 값은 0에서 ∞까지의 값을 가지며 두 변수가 독립이면 1의 값을 갖는다. 따라서 양단 조증가함수(positive monotonic increasing function)인 $f(\cdot)$를 근거한 (단, $f(1)=1$을 만족하여야 함) 교차적비의 일반적인 함수를 다음과 정의한다.

$$g(\alpha)=\frac{f(\alpha)-1}{f(\alpha)+1} .$$

여기서 $\alpha=1$이면 $g(\alpha)=0$으로 한다. 함수 $g(\alpha)$를 이용한 α의 함수로 표현되는 연관성 측도인

Q 통계량과 Y 통계량은 다음과 같다.

2) Q 통계량

Yule [1900]은 $f(x)=x$를 고려하여 연관성 측도 Q를 제안하였다.

$$Q=\frac{p_{11}p_{22}-p_{12}p_{21}}{p_{11}p_{22}+p_{12}p_{21}}=\frac{\alpha-1}{\alpha+1}\ .$$

3) Y 통계량

또한 Yule [1912]은 $f(x)=\sqrt{x}$를 고려하여 Y 통계량을 제안하였다.

$$Y=\frac{\sqrt{p_{11}p_{22}}-\sqrt{p_{12}p_{21}}}{\sqrt{p_{11}p_{22}}+\sqrt{p_{12}p_{21}}}=\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}\ .$$

통계량 Q와 Y의 특징으로는 임의의 두 분할표에서 $Q_1>Q_2$이면 $Y_1>Y_2$로 일치된 양상을 보인다. 두 통계량 모두 [-1, 1]의 구간을 갖고, 독립이면 0의 값을 갖는다. $p_{12}=p_{21}=0$ 혹은 $p_{11}=p_{22}=0$이면 각각 1 또는 -1의 값을 가지며 완전한 양의 연관과 음의 연관을 나타낸다.

4) 상관계수 R 통계량

2×2 분할표의 첫 행과 열에 점수 0이 부여되고 두 번째 행과 열에는 점수 1이 부여되었다고 하면, 행과 열의 주변분포의 평균은 다음과 같다.

$$\mu_r=1-p_{1+},\quad \mu_c=1-p_{+1}\ .$$

각각의 분산과 공분산도 구할 수 있으므로 상관계수 R을 다음과 같이 구할 수 있다.

$$R=\frac{p_{11}p_{22}-p_{12}p_{21}}{\sqrt{(p_{1+}-{p_{1+}}^2)(p_{+1}-{p_{+1}}^2)}}\ .$$

상관계수 R의 특징으로는 두 범주형 변수가 서로 독립이면 $R=0$이며, $p_{12}=p_{21}=0$이면 $R=1$, $p_{11}=p_{22}=0$이면 $R=-1$의 값을 갖는다. 그리고 $|R|\leq|Y|$의 성격도 갖고 있다.

5) 평균제곱분할계수 P 통계량

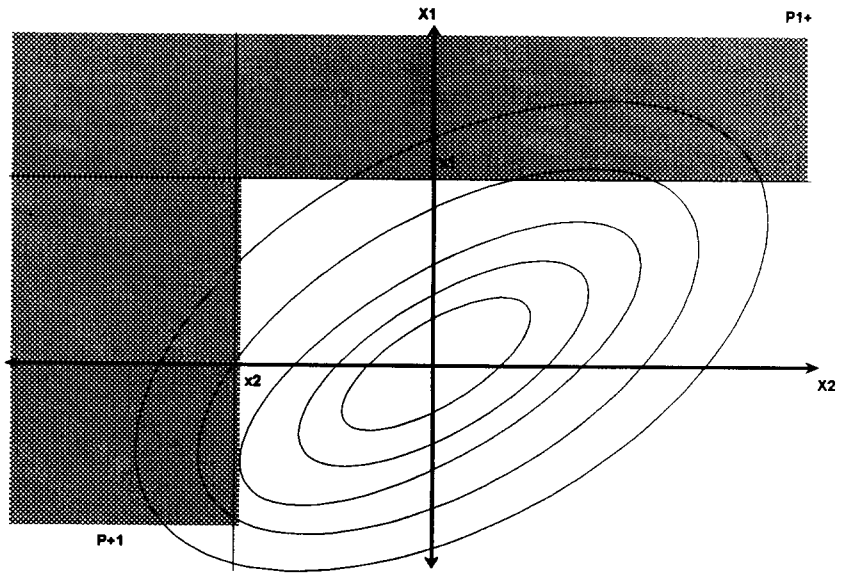
R의 함수로 나타나는 여러 연관성 측도들 중에서 Pearson [1904]은 다음과 같은 평균제곱분할계수(coefficient of mean square contingency)를 제안하였다.

$$P=\sqrt{\frac{R^2}{R^2+1}}\ .$$

P 통계량의 단점으로는 두 변수가 완전한 연관을 갖더라고 상한값이 1이 되지 않는 특징이 있다.

3. 2×2 분할표의 연관성 측도들의 비교

주어진 상관계수를 갖고 있는 2 변량(확률변수 X_1과 X_2) 표준정규분포를 이용하여 2×2 분할표를 생성하기 위하여 <그림 1>과 같은 2 차원 평면을 고려하여 보자. 1 장에서 언급한 <표 1>과 비교를 쉽게 하기 위하여 수평선을 X_2축, 그리고 수직선을 X_1축 이라고 하자. 우리는 이 평면 위에 2 변량 정규분포의 확률밀도함수가 펼쳐져 있다고 가정하고 이 확률밀도함수를 임의의 상수 x_1과 x_2를 기준으로 <그림 1>과 같이 4 개로 분할하여 2×2 분할표의 4 개의 칸 확률을 구하고자 한다.



<그림 1>

다시 주어진 상관계수 ρ의 값에 대하여 <표 1>에서 정의된 p_{1+}와 p_{+1}를 다음과 같이 설정할 수 있다.

$$p_{1+} = P(X_1 \geq x_1),$$

$$p_{+1} = P(X_2 \leq x_2).$$

따라서 정규분포를 따르는 상수 x_1과 x_2은

$$x_1 = \text{probit}(1 - p_{1+}), \quad x_2 = \text{probit}(p_{+1})$$

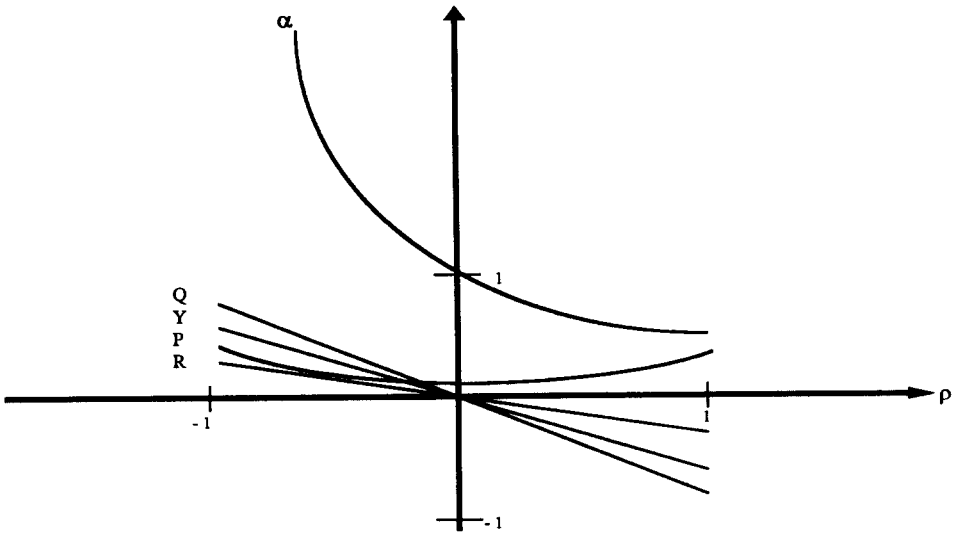
을 이용하여 구한다.

주어진 상관계수의 값을 갖고 있는 표준정규분포로부터 그리고 주변합 p_{1+}, p_{+1}의 값을 0.1에서 0.9까지 0.1의 간격으로 변화시키면서, 상수값 x_1과 x_2을 구한 뒤 각 칸의 확률 p_{11}, p_{12},

p_{21}, p_{22}를 다음과 같이 구할 수 있다.

$$\begin{aligned} p_{21} &= \Phi(x_1, x_2), \\ p_{11} &= \Phi(x_2) - \Phi(x_1, x_2), \\ p_{22} &= \Phi(x_1) - \Phi(x_1, x_2), \\ p_{12} &= 1 - \Phi(x_1) - \Phi(x_2) + \Phi(x_1, x_2) \\ &= 1 - p_{11} - p_{21} - p_{22} \ , \end{aligned}$$

여기서 $\Phi(\cdot)$와 $\Phi(\cdot\,,\cdot)$는 각각 단일변량과 2 변량 표준정규분포의 누적분포함수이다.



<그림 2>

위에서 언급한 방법으로 주어진 조건하에서 구한 칸의 확률들을 이용해서 2×2 분할표의 연관성 측도인 α와 Q, Y, R, P 등을 구할 수 있다. 우선 ρ의 값을 -0.9부터 +0.9까지 0.1씩 변화시키면서 구한 5 개의 연관성 측도들의 전체적인 반응형태를 살펴보면 <그림 2>와 같다.

<그림 2>를 살펴보면 ρ의 값이 -0.9에서 +0.9로 증가함에 따라 교차적비 α의 값은 ∞에서 0으로 감소하는 현상을 파악할 수 있다. 그리고 ρ의 값이 증가함에 따라 Q, Y, R 의 값은 +1에서 -1로 선형적으로 감소하는데 $|Q| \geq |Y| \geq |R|$인 관계를 갖고 있다. $(0, \infty)$의 구간의 값을 갖고 있는 α의 값은 독립성을 나타내는 1을 중심으로 비대칭적으로 나타나지만, 이에 비해 Q, Y, R는 (-1, 1)의 구간을 갖으며 원점에 대해 대칭적이다. 범주형 변수들의 상관계수 R 통계량의 값의 범위는 연속형 변수의 상관계수 ρ의 값의 범위에 비교하여 수축된 현상이 나타남을 알 수 있다. R의 함수로 표현되는 P 통계량의 값은 수직 축에 대해서 대칭인 이차함수형태를 나타낸다. 따라

서 ρ의 부호와는 무관함을 파악할 수 있다.

다음으로 ρ의 값이 고정되어 있을 때를 살펴보자. ρ의 값이 +0.9, 0, 그리고 -0.9인 대표적인 경우를 <그림 3>, <그림 4>, 그리고 <그림 5>에 윤곽그림으로 표현하였다. 주어진 상관계수의 값에 대하여 p_{1+}과 p_{+1}의 값을 0.1부터 0.9까지 0.1씩 증가시킬 때마다 각 칸의 확률을 구하여 5개의 연관성 측도값을 구하였다. <그림 3>에 있는 5 개의 윤곽그림들 중에서 첫 번째 그림은 교차적비 α 통계량값의 분포를 나타내는데 표현된 α의 값을 예를 들어 설명하면 다음과 같다. 변수 1의 주변확률 p_{1+}=0.5 그리고 변수 2의 주변확률 p_{+1}=0.5인 경우에 네 개의 칸확률 p_{11}, p_{12}, p_{21}, p_{22}를 구하여 α 통계량의 값을 구하면 약 0.03임을 알 수 있다. 이런 과정을 통한 <그림 3>부터 <그림 5>까지 5 개의 통계량값들을 살펴보면, α와 Q, Y의 최대값, 최소값은 모두 동일한 추세를 보이고 있다. 반면, R과 P는 $\rho<0$인 경우에는 동일한 분포 형태를 유지하고 모두 양의 값을 가지나, $\rho\geq 0$인 경우에는 R과 P의 값이 $p_{1+} = p_{+1}$(즉, 기울기가 45° 인 직선)을 나타내는 대각선 축에 완전 대칭인 분포를 갖는 것을 알 수 있다.

1) ρ=0.9인 경우

α와 Q, Y는 대략적으로 $p_{1+}+p_{+1}=1$(즉, 기울기가 -45° 인 직선) 근처에서 큰 값을 나타내며, 윤곽그림의 형태가 ρ의 역수값을 갖고 있는 2 변량 정규분포의 형태를 갖는다. α의 윤곽그림에서 특이한 점 하나는 주변확률 p_{1+}과 p_{+1}이 0.1 근처에서 최대값이(α인 경우에 0.06정도) 등장한다는 사실이다. ρ=0.9인 경우 p_{1+}과 p_{+1}의 확률이 0.1 일때는 p_{11}의 확률이 0에 수렴하고 p_{22}의 확률이 1에 접근하는 경우인데, 이런 수렴하는 확률들을 이용하여 α 통계량값을 구하였기 때문에 발생하였다. 우리는 수렴값을 반올림함으로 발생한 이러한 특이값들을 무시해도 무리가 없을 것이다. 이런 현상은 Q와 Y의 윤곽그림에서도 동일하게 발생하는데 그 이유는 Q와 Y 통계량은 α 통계량의 함수이기 때문이다. 그러므로 α와 Q, Y의 통계량값은 $p_{1+}+p_{+1}=1$인 직선 근처에서 최대값을 갖고 이 선에서 멀어질수록 값이 작아진다. 그리고 R과 P의 윤곽 그림은 α와 Q, Y와 비슷한 형태를 갖고 있지만, α와 Q, Y 보다 넓게 퍼져 있다. 그리고 R의 최대값과 최소값이 P의 최소값과 최대값으로 서로 대칭적인 값을 가지며, $p_{1+}+p_{+1}=1$(즉, 기울기가 45° 인 직선) 근처에서 R는 최소값 그리고 P는 최대값을 갖는다. ρ=0.9인 경우에 각 연관성 측도의 실질적인 값의 범위는 다음과 같다.

$\alpha\in(0, 0.03)$, $Q\in(-1.008, -0.936)$, $Y\in(-1, -0.68)$,
$R\in(-0.733, -0.067)$, $P\in(0.1, 0.6)$.

2) ρ=-0.9인 경우

2 변량 정규분포의 상관계수가 ρ=-0.9인 경우에 각 연관성측도들이 갖는 값의 범위는 다음과 같다.

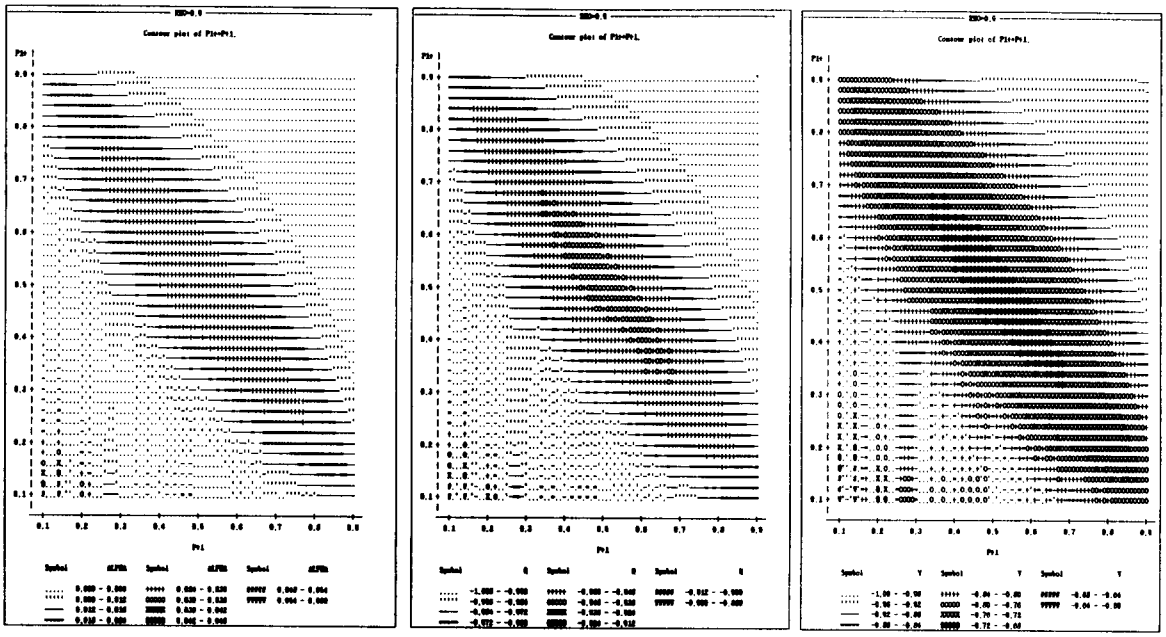
$\alpha \in (17.37,\ 87090998.2)$, $Q \in (0.891,\ 1.001)$, $Y \in (0.6,\ 1.0)$,
$R \in (0.0667,\ 0.7333)$, $P \in (0.1,\ 0.6)$.

전반적인 윤곽그림은 ρ=0.9일 경우와 대칭적으로 표현되며, α와 Q, Y 는 대략적으로 $p_{1+} = p_{+1}$ 인 부근에서 최소값을 가지며, 그 선에서 멀리 떨어질수록 값들이 커진다. R과 P 의 윤곽그림은 최대값과 최소값이 반대방향으로 나타나는데 여기에서는 $p_{1+} = p_{+1}$일 때 모두 최대값을 갖으며, 그 선에서 멀리 떨어질수록 값이 감소한다. 또한 R과 P 의 윤곽 그림은 α와 Q, Y 의 윤곽그림보다 보다 넓게 퍼져 있음을 알 수 있다. α, Q와 Y 의 윤곽그림에서도 p_{1+}의 주변확률이 0.1 그리고 p_{+1}의 주변확률이 0.9 근처에서 특이값이 등장하는데 이 경우에는 p_{11}의 확률이 1에 수렴하고 p_{22}의 확률이 0에 접근하는 경우이다. 이런 수렴값들의 반올림을 이용하여 통계량값을 구하였기 때문에 계산상에서 발생한 특이값들을 무시한다.

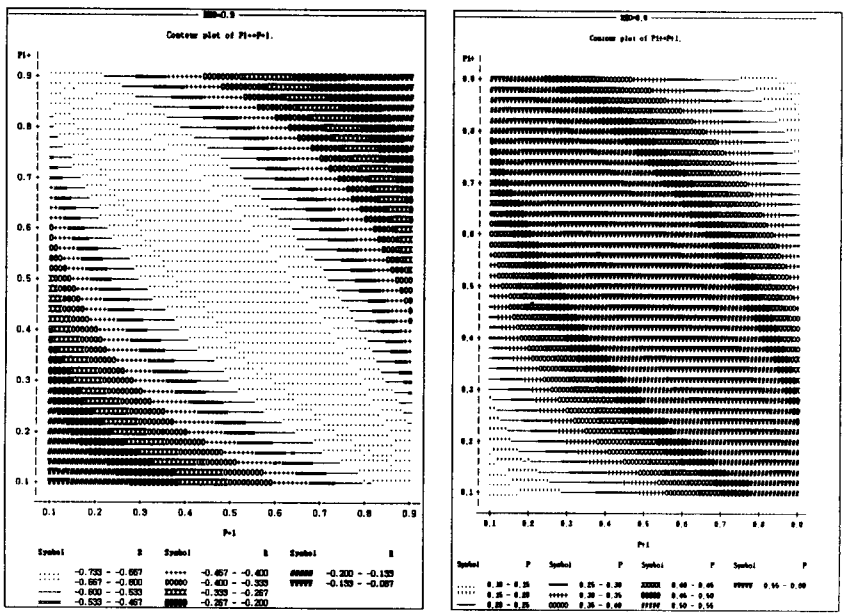
3) ρ=0인 경우

α의 값은 모든 p_{1+}값에 대해 p_{+1}=0.1에서 최대값을 갖고, p_{+1}=0.9에서 최소값을 갖는다. 그리고 Q, Y 의 값은 모든 p_{1+}값에 대해 p_{+1}=0.1에서 최대값을, p_{+1}=0.5에서 최소값을 갖는다. 또한 R의 값은 모든 p_{1+}값에 대해 p_{+1}=0.2에서 최대값을, p_{+1}=0.5에서 최소값을 갖는다. P 의 값은 모든 p_{1+}값에 대해 p_{+1}=0.5에서 최대값을 갖고, p_{+1}=0.7과 0.3에서 최소값을 갖는다. 즉, 모든 연관성 측도들은 최대값과 최소값을 비롯하여 전반적인 분포가 p_{1+}보다는 p_{+1}에 의한 영향을 많이 받는다. 그러나 전체적으로 최대값과 최소값의 범위가 좁아 위에서 언급한 추세가 유위하다고 할 수 없으며 이는 ρ가 0일 때 즉, 연속형 변수들이 독립임을 가정한다면 이에 대하여 생성된 범주형 변수들도 어떠한 주변확률의 상황하에서도 독립적인 관계임을 잘 반영하고 있다.

즉 결과적으로 ρ값의 변화에 따른 연관성 측도들의 반응 방향을 제외하고 값들의 범위를 살펴보면, α 보다는 Q 와 Y 의 값의 범위가 적다. 이는 α 보다는 Q 와 Y 가 ρ의 변화에 따른 2변량 정규분포에서 연관성 정도의 변화를 변동이 적게 반영하고 있다고 할 수 있다. 또한 R과 P 보다도 Q 와 Y 의 값의 범위가 적으므로 Q 와 Y 통계량이 더욱 ρ의 변화에 민감하다고 할 수 없다. 그러므로 위에서 언급한 상황하에서 2×2 분할표로 나타난 범주형 자료에서의 연관성 측도 통계량은 Q 와 Y 통계량이 비교적 안정적이며 그 중에서도 Q 통계량이 가장 안정적인 연관성 측도라 할 수 있다.

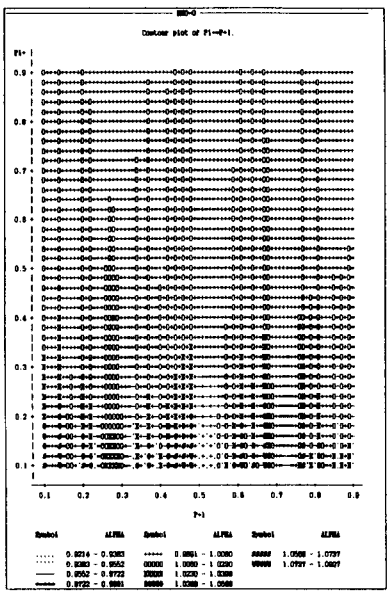


α *Q* *Y*

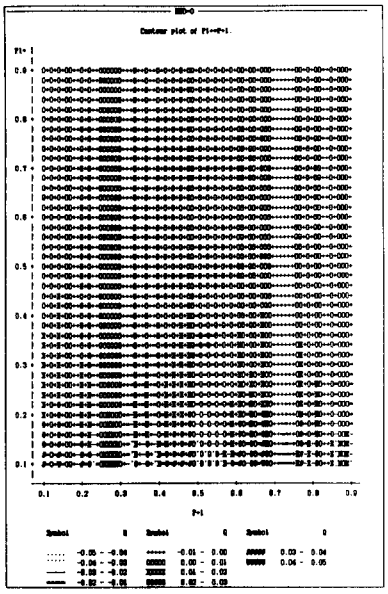


R *P*

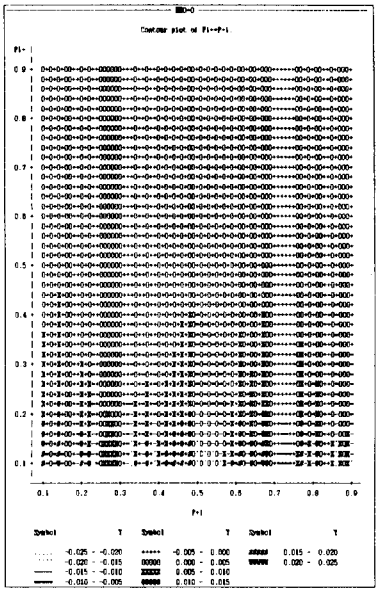
<그림 3> ρ=0.9인 경우



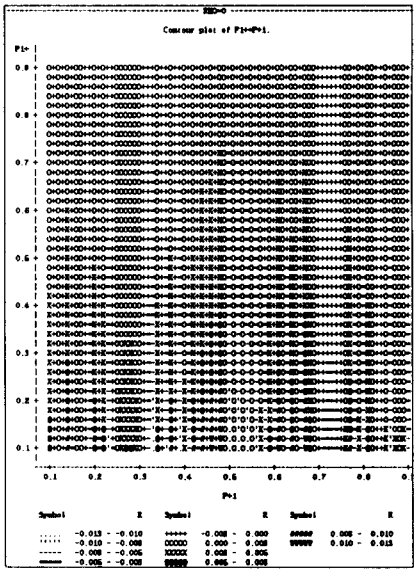
a



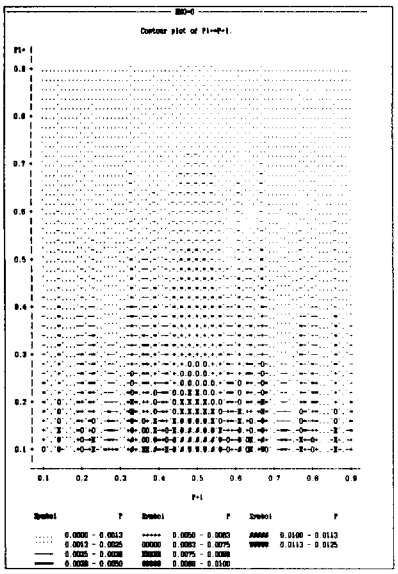
Q



Y

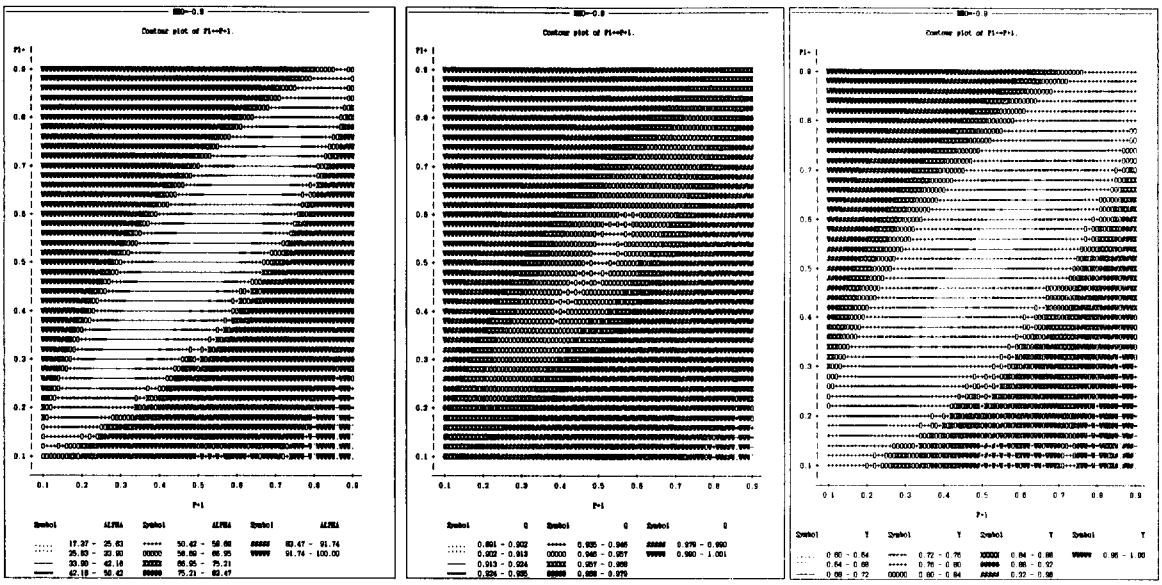


R

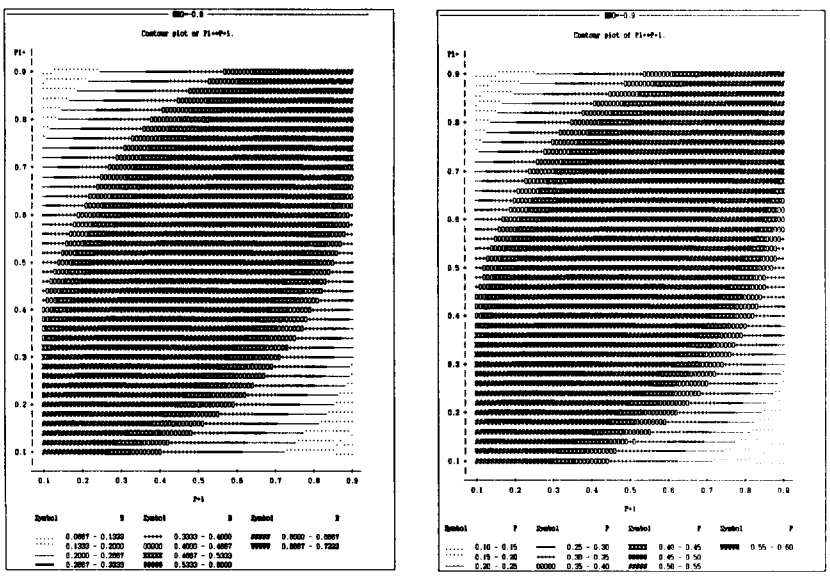


P

<그림 4> $\rho=0$인 경우



α *Q* *Y*



R *P*

<그림 5> ρ=-0.9인 경우

4. $I\times J$ 분할표의 연관성 측도

1) P 통계량

$I\times J$ 분할표에서 Pearson의 독립성 검정통계량으로 잘 알려진

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

은 $I>2$, $J>2$이면 더 이상 구간 [0, 1]사이의 값을 가지지 않는다. 이러한 단점을 보완하기 위해 Pearson은 2절에서 언급한 평균제곱분할계수를 확장한 다음과 같은 통계량을 제안하였다.

$$P=\sqrt{\frac{X^2}{X^2+1}}\ .$$

이 통계량은 $I>2$, $J>2$일 때도 구간 [0, 1]사이의 값을 갖는다. 그러나 상한값이 항상 1이 될 수 없다는 단점을 갖고 있다.

2) T 통계량

P 통계량의 한계를 극복하기 위하여 상한값이 1이 되도록 T 통계량을 제안하였다.

$$T=\left[\frac{X^2}{[(I-1)(J-1)]^{1/2}}\right]^{1/2}.$$

그러나 이 통계량 역시 $I\neq J$ 이면 1이 되지 않는 성질이 있다(자세한 설명은 Bishop, Fienberg와 Holland [1975] 그리고 홍종선 [1995]을 참조).

3) V 통계량

Cramer [1946]는 다음과 같은 V 통계량

$$V=\left[\frac{X^2}{\min[(I-1),(J-1)]}\right]^{1/2}$$

을 제안하였는데 $I>2$, $J>2$에 대하여 $V\geq T$ 가 된다는 것을 알 수 있다.

4) $\lambda_{C|R}$ 통계량

$I\times J$ 분할표에서 Goodman과 Kruskal [1954]은 X^2통계량에 근거한 측도들에 결여되어있는 확률에 의한 명확한 해석을 갖는 오차비례감소측도(proportional reduction in error measure; PRE)라는 통계량을 개발하였다. 그리고 Goodman과 Kruskal [1954]은 이 통계량 보다 직접적인 해석이 가능한 여러 측도를 제안하고 있는데 그 중에서 가장 널리 알려져 있는 것으로는 행변수의 범주로부터 열변수의 범주를 예측하거나 혹은 열범주로부터 행범주를 예측하기 위한 분할표들을 위한 방법이다. 만약 행변수 R로부터 열변수 C의 범주를 예측하고자 한다면 다음과 같은 측

도를 이끌어 낼 수 있다.

$$\lambda_{C|R}=\frac{\sum_{i=1}^{I} p_{im}-p_{+m}}{1-p_{+m}},$$

여기서 $p_{im}=\max_j\{p_{ij}\}$, $p_{+m}=\max_j\{p_{+j}\}$이다. $\lambda_{C|R}$ 통계량은 구간 [0, 1] 사이의 값을 가지며, 각 행에 0이 아닌 칸이 하나씩만 나타날 때 1의 값을 갖는다.

5) U^2 통계량

$I\times J$ 분할표에 대해 Light와 Margolin [1971]은 결정계수와는 다르지만, 결정계수와 유사한 형태로 나타나는 연관성측도를 고려하였다. 우선 Gini [1912]는 분산분석에서 총변동처럼 $I\times J$ 분할표의 총변동(TSS)을 정의하였고 Light와 Margolin은 총변동이 그룹내 제곱합(WSS), 그룹간 제곱합(BSS)으로 변동분해된다는 것을 보였으며 이를 이용하여 열변수에 비례한 행변수의 변동비율 BSS/TSS으로 다음과 같은 통계량을 제시하였다.

$$\tau_{R|C}=\frac{\sum_{ij}\frac{(p_{ij}-p_{i+}p_{+j})^2}{p_{+j}}}{1-\sum_i p_{i+}^2}.$$

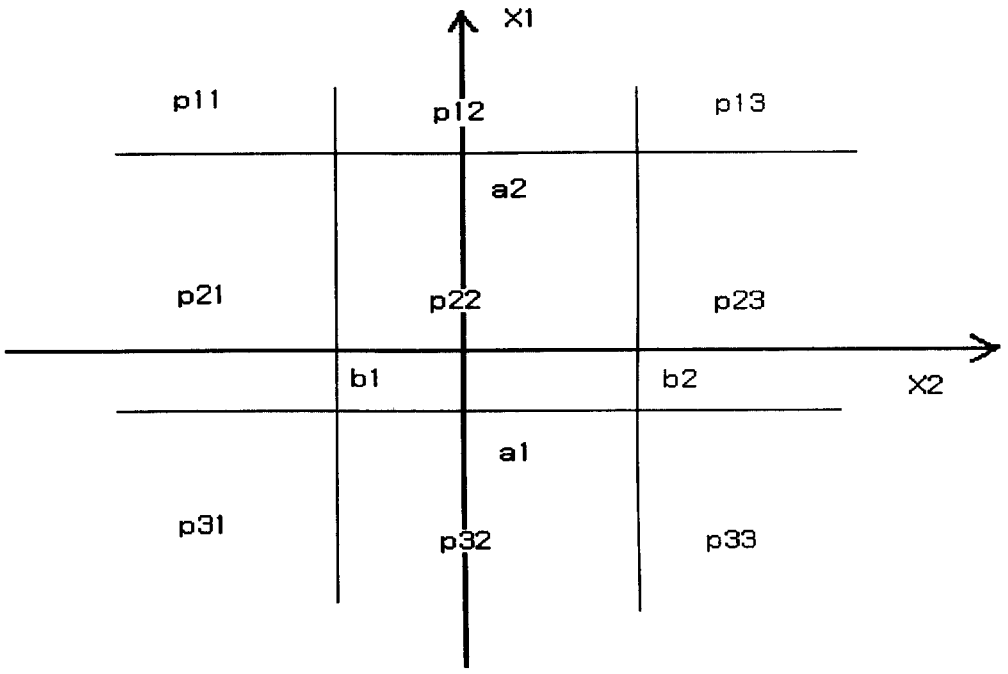
$\tau_{R|C}$ 통계량을 이용해서 연관이 존재하지 않는가에 대한 검정은 검정통계량

$$U^2=(n-1)(I-1)\tau_{R|C}$$

에 의해 검정을 수행하며, 독립이라는 귀무가설하에서 U^2 통계량은 자유도 $(I-1)(J-1)$을 갖는 χ^2 분포를 따르게 된다.

5. 3×3 분할표의 연관성 측도비교

$I\times J$ 분할표를 구하는데는 각각의 확률변수에서 $(I-1)$개와 $(J-1)$개의 임의의 값에 대하여 분할하여 $I\times J$ 개의 분할된 칸확률을 구하여야하는 힘든 과정이 필요로 하기 때문에 여기에서는 비교적 간단한 3×3 분할표를 생성하여 논의하고자 한다. 2×2 분할표와 마찬가지로 3×3 분할표에서도 9 개의 칸 확률을 구하기 위하여 2 변량 표준정규분포를 따르는 확률변수 X_1과 X_2를 <그림 6>과 같이 2 차원 평면에서 고려하자. X_1축과 X_2축에 각각 2 개씩의 분할선을 찾아서 각각 a_1, a_2, b_1, b_2라 놓고 이 점들로 분할된 면적에 해당되는 확률을 3 장에서와 같이 구하면 3×3 분할표를 만들 수 있다. 2×2 분할표와 마찬가지로 3×3 분할표에서도 각 칸의 확률 p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33}등을 구할 수 있는데 예를 들어 $p_{31}=\Phi(a_1,b_1)$이며 $p_{11}=\Phi(b_1)-\Phi(a_2,b_1)$ 이다.



<그림 6>

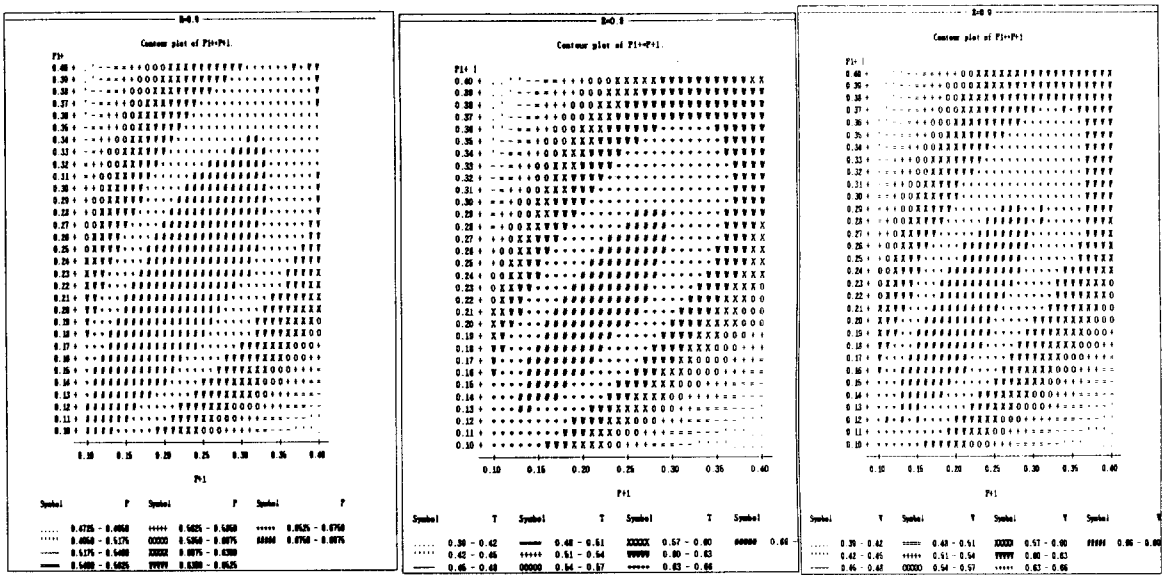
이때 a_1, a_2, b_1, b_2를 설정함에 따라 각 칸의 확률이 달라지게 되는데 이를 위하여 주변확률을 변화시키고자 한다. 주변확률을 설정하기 위하여는 여러 가지의 방법이 존재할 수 있지만 여기서는 p_{1+}, p_{3+}, p_{+1}, p_{+3}를 변화시키면서 칸 확률을 구하였다. 특히 p_{1+}와 p_{3+}의 비율이 2 : 1, 1 : 1, 그리고 1 : 2 과 같은 3 가지를 고려할 수 있으며 p_{+1}과 p_{+3}의 비율도 동일하게 고려한다면 모두 다음과 같은 9 가지 경우를 생각할 수 있다.

<Case 1-1> $2p_{1+}=p_{3+}$, $2p_{+1}=p_{+3}$
<Case 1-2> $2p_{1+}=p_{3+}$, $p_{+1}=p_{+3}$
<Case 1-3> $2p_{1+}=p_{3+}$, $p_{+1}=2p_{+3}$
<Case 2-1> $p_{1+}=p_{3+}$, $2p_{+1}=p_{+3}$
<Case 2-2> $p_{1+}=p_{3+}$, $p_{+1}=p_{+3}$
<Case 2-3> $p_{1+}=p_{3+}$, $p_{+1}=2p_{+3}$
<Case 3-1> $p_{1+}=2p_{3+}$, $2p_{+1}=p_{+3}$
<Case 3-2> $p_{1+}=2p_{3+}$, $p_{+1}=p_{+3}$
<Case 3-3> $p_{1+}=2p_{3+}$, $p_{+1}=2p_{+3}$

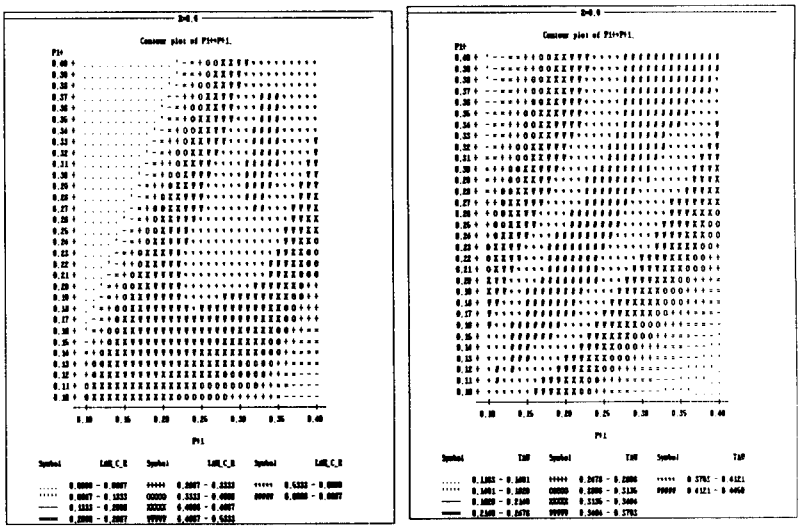
본 논문에서는 p_{1+}와 p_{+1}의 경우의 수가 제일 많이 발생하는 <Case 2-2>의 경우를 살펴보도록 하겠다. ρ의 값을 -0.9에서 0 그리고 0.9로 변화시키고 동시에 p_{1+}와 p_{+1}를 0.1에서 0.4까지 0.1씩 변화시키면서 각 경우의 연관성 측도들을 구한 결과를 <그림 7>, <그림 8>에 윤곽그림으로 표현하였다. U^2 통계량은 $\tau_{R|C}$의 선형 함수이면서 총빈도수 n을 알아야 하므로 여기에서는 U^2 통계량 대신에 $\tau_{R|C}$를 사용하고자 한다. ρ가 -0.9일 때 각 연관성 측도들의 범위는 $T \in (0.39, 0.69)$, $P \in (0.4725, 0.6975)$, $V \in (0.39, 0.69)$, $\lambda_{C|R} \in (0, 0.6667)$, $\tau_{R|C} \in (0.1164, 0.4369)$ 등을 갖으며, ρ가 0.9일 때도 이와 거의 동일한 범위의 값들을 갖는다. 즉 측도의 형태들이 제곱의 함수형태를 갖고 있어 ρ에 대한 부호와는 관계없이 연관정도에만 영향을 받는다. 또한 ρ가 0일 경우에는 모든 연관성 측도들이 0 근처의 값을 나타내고 있다.

<그림 7>과 <그림 8>을 자세히 살펴보면 X^2에서 발전된 통계량 T, P, V 등은 서로 비슷한 값의 추세를 보이고 있는데 p_{1+}과 p_{+1}의 값이 동일하면서 그 값이 각각 0.3보다 미만인 선상에서 최고값을 가지며 그 선에서 멀어질수록 작은 값을 갖는다. $\lambda_{C|R}$의 값은 T, P, V 등의 추세와 비슷한 현상을 나타내고 있으나 p_{1+}과 p_{+1}이 0.3에서 0.35 근처에서 최고값을 갖는다. 반면 $\tau_{R|C}$ 통계량값은 $p_{1+} = p_{+1}$에서 대칭으로 나타나고 있으며 p_{1+}과 p_{+1}의 값이 증가할수록 값의 폭이 넓어지는 특징을 지니고 있다.

여기에서 논의된 5 가지의 통계량들을 살펴보면 모든 통계량들은 비슷한 분포 형태를 나타내고 있으며 단지 각 통계량들의 최대값의 위치들이 약간씩 차이가 있다. 그 중에서도 P 통계량이 나타내주는 값들이 큰 수를 나타내 주며 그 값들의 범위가 가장 적으며 여러 통계량 중에서도 P 통계량이 가장 안정적이라는 사실을 지적할 수 있다.

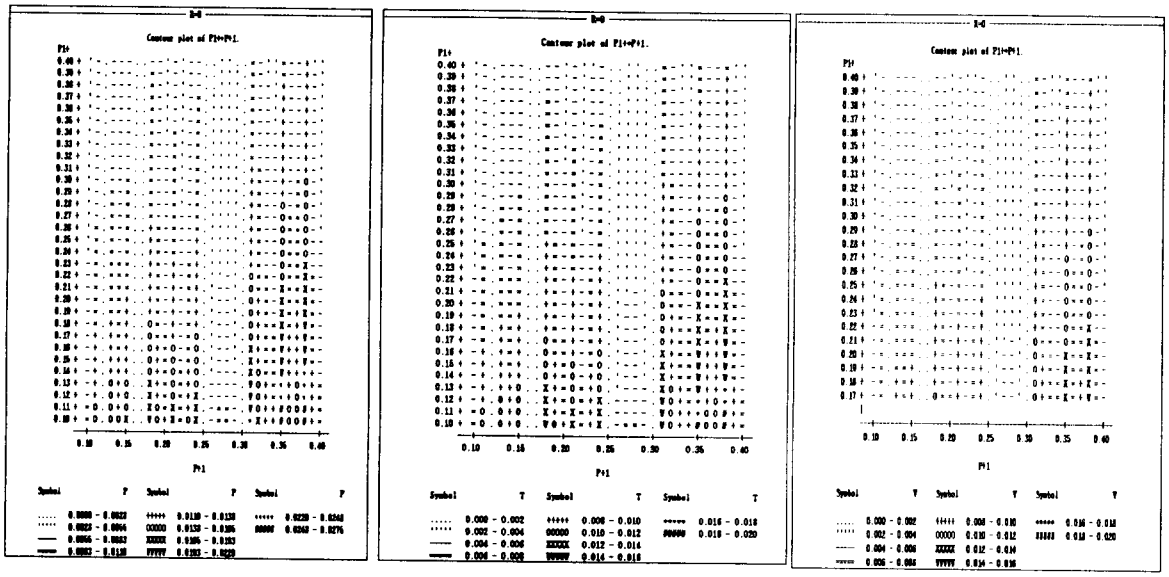


P T V



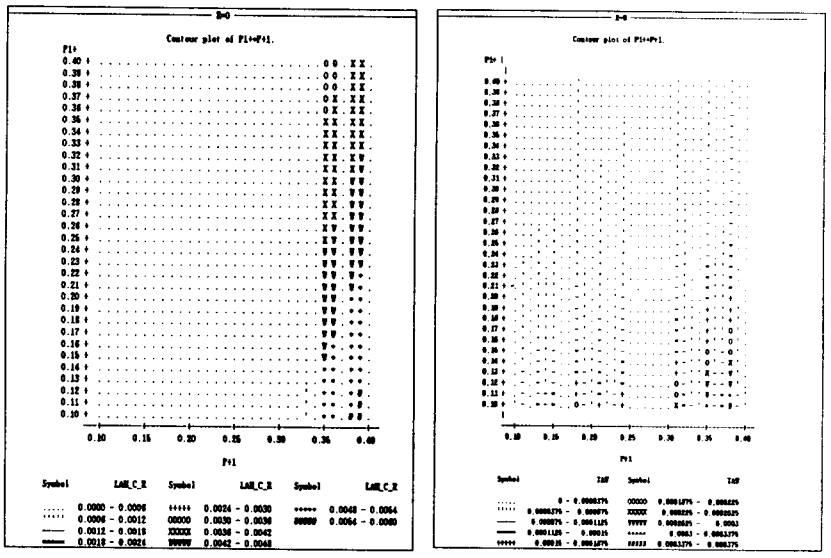
$\lambda_{C|R}$ $\tau_{R|C}$

<그림 7> $\rho=0.9$일 경우



P T

V



λ_{aR} τ_{RC}

<그림 8> $\rho=0$일 경우

6. 결론

연속형 변수들이며 +1에서 -1까지 완벽한 상관관계를 갖고 있는 2 변량 정규분포를 이용하여 2×2 분할표와 3×3 분할표를 생성하였으며 각각의 분할표에서 대표되는 5가지 통계량의 값들을 구하여 분석하였다.

2×2 분할표에서는 α 보다는 Q 와 Y 가 ρ의 변화에 따른 2 변량 정규분포에서 연관성 정도의 변화를 잘 반영하고 있으며, Q 와 Y 보다도 R과 P 가 더욱 ρ의 변화에 민감함을 알 수 있다. 결론적으로 2×2 분할표로 나타난 범주형 자료에서의 연관성 측도 통계량은 Q 통계량이 가장 안정적인 연관성 측도라 할 수 있다.

3×3 분할표에서 논의된 5 가지의 통계량들을 살펴보면 모든 통계량들은 비슷한 분포 형태를 나타내고 있다. 그 중에서도 P 통계량이 나타내주는 값들이 큰 수를 나타내 주며 그 값들의 범위가 가장 적어 가장 안정적이라는 사실을 지적한다. 그러므로 3×3 분할표에서의 연관성 측도 통계량으로는 P 통계량을 추천한다고 할 수 있다. 그러나 이런 현상은 3×3 분할표만을 그리고 p_{1+}와 p_{3+} 그리고 p_{+1}과 p_{+3}의 비율이 동일한 경우에만 살펴보았기 때문에 일반적인 다항범주형 변수로 구성된 $I \times J$ 분할표에서의 연관성 측도의 현상이라고 단정할 수는 없다. 따라서 향후 많은 경우의 2 차원 분할표에 대하여 연구가 필요하다.

참 고 문 헌

[1] 홍 종 선 [1995]. 「대수선형모형」, 자유아카데미.

[2] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. [1975]. *Discrete Multivariate Analysis : Theory and Practice*, Cambridge, Mass., The MIT Press.

[3] Cramer, H. [1946]. *Mathematical Methods of Statistics*, Princeton Univ. Press.

[4] Gini [1912]. *Valiabilit e mutabilit, contributo allo dello distribuzioni; relazione statische,* In StudiEconomico-Giuridici della R. Universit di Carliari.

[5] Goodman, L. A., and Kruskal, W. H. [1954]. *Measures of Association for Cross Classifications*, J. of Amer. Statist. Assoc. 49, 732-764.

[6] _____ [1979]. *Measures of Association for Cross Classifications,* New York: Springer-Verlag(contains articles appearing in J. of Amer. Statist. Assoc. in 1954, 1959, 1963, 1972).

[7] Light, R. J., and Margolin, B. H. [1971]. An anlaysis of variance for categorical data, *J. Amer. Statst. Assoc.,* 66, 534-544.

[8] Pearson, K. [1904]. Mathematical contributions to the theory of evolution Ⅷ: On the theory of contingency and its relation to association and normal correlation. *Draper's Co. Research Memoirs. Biometric Series*, no. 1.

[9] Yule, G. U. [1900], On the association of attributes in ststistics, *Phil. Trans.* Ser. A 194, 257-319.

[10] _____ [1912], On the methods of measuring association between two attributes, *J. Roy. Statist. Soc.* 75, 579-642.