응용통계연구 제 19권 3호, 2006년, pp. 545–554

# 다차원 임의 분할표 생성*

최현집[1)]

요 약

로그선형모형에 기반을 둔 다차원 임의 분할표를 생성하는 방법을 제안하였다. 이를 위해 Lee(1997)가 제안한 선형결합에 의한 결합분포 생성 방법을 적용하였으며, Pearson 통계량을 연관성 측도로 사용하는 것을 제안하였다. 세 변수가 서로 완전한 연관을 갖는 삼차원 결합분포를 생성할 수 있으므로 본 연구에서 제안한 방법은 사차원 이상 다차원 임의 분할표를 생성하는 문제로 확장될 수 있다.

주요용어: 난수생성, Pearson 통계량, 다차원 분할표, 로그선형모형

## 1. 서론

여러 범주형 변수들의 범주가 교차 분류된 분할표 분석의 주된 관심은 표를 구성하는 변수들 사이의 연관관계를 식별하는 것이라고 할 수 있다. 이를 위해 모형에 기반을 둔 다수의 분석방법들이 제안되어 있으며, 이들 방법들은 대부분 대표본 이론에 근거하고 있다. 이러한 이유로 모의실험을 통해 제안된 방법들의 소표본 특성을 파악하기 위한 연구 역시 활발히 수행되고 있다. 분할표 분석을 위하여 제안된 모형들은 모형내에 포함된 모수들에 의해 변수들의 연관구조를 식별하며, 이들 모수들은 변수들의 주변분포에 의존하여 결정된다. 그러므로 주어진 주변분포와 변수들간의 연관정도에 의하여 결합분포를 얻는 방법은 모의실험을 위해 매우 중요한 문제이다.

Emrich와 Piedmonte(1991)는 이진 범주형 변수들간의 연관구조가 지정된 상관행렬에 의한 다변량 정규분포로부터 결합분포를 얻는 방법을 제안하였다. 이때 상관행렬을 구성하는 연관성 측도는 Peason 상관계수만이 아닌 두 범주형 변수들의 연관을 측정하기 위해 제안된 여러 측도들을 이용할 수 있다고 지적하고 있다. 그러나 대다수의 연관성 측도들은 주변분포에 의존하며 상관계수가 갖는 구간 $[-1, 1]$을 만족하지 못하는 문제가 발생한다. 특히, Lee(1993)는 이러한 문제를 지적하고 주변분포에 영향을 받지 않는 선형계획법과 Archimedian copulas를 이용한 방법을 제안하고 있다. 이외에 이진 범주형 변수들에 의한 결합분포 생성에 관한 연구들은 Park 등(1996)과 Kang과 Jung(2001) 등을 참고할 수 있다.

다범주를 갖는 변수들에 의한 다차원 결합분포를 생성하기 위하여 Gange(1995)는 반복비율적합(iterative proportional fitting)을 이용한 방법을 제안하였다. 이 방법은 먼저 사

* 이 연구는 2004년 경기대학교 연구비 지원에 의하여 연구되었음.

1) (443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 경제학부 응용정보통계전공, 부교수
E-mail: hjchoi@kyonggi.ac.kr

전 정보를 이용하여 저차 주변분포를 결정한 후에, 이들 주변분포를 이용한 반복비율적합을 통해 결합분포를 생성하게 된다. 여기서 저차 주변분포들은 로그선형모형(log-linear models)의 경우에 분할표의 칸 추정값을 얻기 위한 충분주변합(sufficient marginals) 혹은 형상(configurations)과 동등하며, 결국 이 방법은 칸 확률을 얻기 위한 제약조건인 주변분포를 이용하여 반복비율적합 방법을 통해 고차 결합분포를 얻는 방법으로 이해할 수 있다. 이와는 달리 Lee(1997)는 Theil의 불확실성 계수 $U$ 혹은 Goodman과 Kruskal의 $\tau$와 같은 연관성 측도를 이용하여, 서로 독립인 결합분포와 이로부터 가장 멀리 떨어진 최대연관을 갖는 결합분포의 선형조합에 의해 주어진 측도값을 만족하는 결합분포를 결정하는 방법을 제안하고 있다. 최대연관을 갖는 결합분포는 일차 주변분포가 제약조건인 선형계획법에 의해 얻으며, 이렇게 얻어진 이차원 주변분포들로부터 삼차원 결합분포는 이차원 주변분포가 제약조건으로 추가된 선형계획법에 의해 얻게 된다. 그러나 세 변수가 서로 완전한 연관을 갖는 경우에는 칸 확률 자체를 제약조건에 추가시켜야 하므로 이 방법은 삼차원 이상 다차원 결합분포를 생성하기 위해 확장하기 어렵다.

본 연구에서는 로그선형모형에 기반한 다차원 임의 분할표를 생성하는 방법을 제안하고자 한다. 제2절에서는 이차원 결합분포를 얻기 위한 방법을 정리할 것이며, 특히 선형계획법이 아닌 최대연관을 갖는 결합분포를 얻는 방법과 Pearson의 $\chi^2$ 통계량을 연관성 측도로 이용한 선형결합 방법을 제안할 것이다. 제3절에서는 2절의 결과를 확장하여 부분연관에 의해 결정된 결합분포와 이로부터 가장 멀리 떨어진 세 변수가 완전히 연관된 결합분포의 선형결합을 통해 삼차원 결합분포를 생성하는 방법을 제안하게 될 것이다. 이를 통해 로그선형모형의 형상들을 얻고, 모형에 의해 결정되는 생성하고자 하는 연관구조를 가진 다차원 결합분포는 반복비율적합 방법에 의해 얻고자 한다. 이렇게 결정된 결합분포로부터 다차원 임의 분할표는 inversion 알고리즘 등에 의하여 쉽게 얻을 수 있다. 마지막으로 제4절에서는 본 연구의 결과를 정리하였다.

## 2. 이차원 임의 분할표의 생성

$I_1, I_2$개 범주를 갖는 두 범주형 변수 $X_1$과 $X_2$를 고려하기로 하자. 이들 두 변수에 의한 이차원 임의 분할표를 생성하기 위해서 사전에 알려져 있거나 혹은 연구자에 의해 결정된 주변분포들을 다음과 같이 나타내기로 한다.

$$\begin{aligned} p_{i+} &= \Pr[X_1 = x_i],\ i = 1, 2, \ldots I_1, \\ p_{+j} &= \Pr[X_2 = x_j],\ j = 1, 2, \ldots I_2. \end{aligned}$$

만일 두 변수가 서로 독립이라면 이차원 결합분포는 이들 두 주변분포에 의하여

$$p_{ij} = p_{i+}p_{+j}, i = 1, \ldots, I_1, j = 1, \ldots, I_2 \tag{2.1}$$

와 같이 결정할 수 있으며, 이로부터 다음과 같은 inversion 알고리즘을 통하여 이차원 임의 분할표를 얻을 수 있다.

1: $z_0 = 0$,
$z_i = z_{i-1} + p_{i-1}$ , $i = 1, 2, \ldots, I$,
$z_{I+1} = 1$.
2: 균일분포 $U[0,1]$로부터 난수 $U$를 생성
3: $z_i \leq U < z_{i+1}$ 이면 $x_i$값 증가

여기서 $I = I_1 \times I_2$이며, $p_i$, $i = 1, 2, \ldots, I$는 $p_{ij}$의 첨자에 따라 오른차순으로 정렬한 확률을 나타낸다. 그리고 $x_i$는 해당 첨자에 대응하는 임의 분할표의 칸 도수를 나타낸다. 이외에 주어진 결합분포로 부터 난수를 얻기 위한 여러 알고리즘은 Davis(1993)에 잘 정리되어 있다. 그러나 이와는 달리 두 변수가 서로 연관관계를 가지고 있는 임의 분할표를 얻기 위해서는 주어진 주변분포들과 사전에 정해진 연관관계를 만족하는 이차원 결합분포를 먼저 결정하여야 한다.

두 변수의 연관관계를 결정하기 위하여 Gange(1995)는 결합확률을 칸 확률로 갖는 $I_1 \times I_2$표의 모든 가능한 $2 \times 2$ 부분표의 교차적비를 지정하는 방법을 제안하고 있다. 또한 Lee(1997)는 Theil의 불확실성 측도와 Goodman과 Kruskal의 측도를 제안하였으나, 이들 두 통계량은 그의 논문에서 지적하고 있는 바와 같이 행과 열수가 서로 다른 경우에 서로 다른 값을 가지므로 행과 열에 대칭이 되도록 조정되어야 한다.

Peason의 $\chi^2$ 통계량은 두 범주형 변수의 독립성을 평가하기 위한 통계량이기도 하지만 Mirkin(2001) 등에서 지적한 바와 같이 두 변수의 연관관계를 측정하기 위한 측도로 널리 이용되고 있다. $\chi^2$ 통계량은 행과 열의 수에 영향을 받지 않는 다는 것은 이미 알려진 사실이다. 그러나 $\chi^2$ 통계량은 표본의 총수에 의해 크기가 영향을 받으므로 이의 영향을 제거한 다음과 같은 $\Phi^2$ 통계량이 연관성 측도로 이용되고 있다.

$$\Phi^2 = \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}}.$$

이제 주어진 주변분포들과 고정된 $\Phi^2$ 값을 갖는 이차원 결합분포를 얻기 위하여, 식 (2.1)에 의해 결정된 서로 독립인 결합분포를 $P_{IND}$ 그리고 두 변수가 최대연관을 갖는 결합분포를 $P_{MAX}$와 같이 나타내기로 한다. 이들로부터 주어진 두 변수의 연관정도를 나타내는 $\Phi_0^2$ 값을 갖는 결합분포를 얻기 위해 Lee(1997)가 제안한 다음과 같은 결합분포를 고려하기로 한다.

$$P(\lambda) = \lambda P_{IND} + (1 - \lambda) P_{MAX}, \tag{2.2}$$

여기서 $0 \leq \lambda \leq 1$이며, 두 결합분포의 선형결합인 $P(\lambda)$는 $\lambda$의 연속함수이므로 주어진 $\Phi_0^2$를 갖는 $P(\lambda_0)$를 위한 $\lambda_0$는 항상 존재한다. 이때 $P_{IND}$와 $P_{MAX}$의 일차 주변분포는 주어진 주변분포와 같으므로 $P(\lambda_0)$ 역시 주어진 주변분포와 같은 주변분포를 갖게 된다. 따라서 식 (2.2)에 의해 주어진 연관 $\Phi_0^2$을 갖는 결합분포를 얻기 위해서는 두 변수가 최대연관을 갖는, 즉 $\Phi^2$을 최대로하는 $P_{MAX}$를 결정하여야 한다.

두 변수가 서로 독립인 $P_{IND}$에서 가장 멀리 떨어진 $P_{MAX}$를 찾기 위해 Lee(1997)에서와 같이 선형계획법을 이용하는 방법을 고려할 수 있다. 그러나 주어진 주변분포에 의해

**Algorithm 1** : ExtremeForTwo

**Require:** 일차 주변분포 $p_{i+}$ , $p_{+j}$, $\forall\, i, j$

**Ensure:** 결합분포 $P = \{p_{11}, p_{12}, \cdots, p_{I_1 I_2}\}$

$p_{ij} = 0$, $\forall i, j$
$p'_{i+} = p_{i+}$ , $\forall i$
$p'_{+j} = p_{+j}$ , $\forall j$
**for** $i = 1, \ldots, I_1$ **do**
  **for** $j = 1, \ldots, I_2$ **do**
    $q_{ij} = \min(p'_{i+}, p'_{+j})$, $\forall i, j$
    **if** ($p'_{i+} > 0$ and $p'_{+j} > 0$) **then**
      $p_{ij} = q_{ij}$
      $p'_{i+} = p'_{i+} - q_{ij}$
      $p'_{+j} = p'_{+j} - q_{ij}$
    **end if**
  **end for**
**end for**

그림 2.1: 최대연관을 갖는 이차원 결합분포 생성 알고리즘

표 2.1: 두 주변분포에 의한 최대연관을 갖는 결합분포

| | | |
|---|---|---|
| 0.00 | 0.00 | 0.10 |
| 0.00 | 0.20 | 0.00 |
| 0.25 | 0.05 | 0.00 |
| 0.00 | 0.00 | 0.40 |

표 2.2: 주어진 연관을 갖는 결합분포들

(가) $\lambda = 0.5757$, $\Phi_0^2 = 0.3$

| | | |
|---|---|---|
| 0.0144 | 0.0144 | 0.0712 |
| 0.0288 | 0.1136 | 0.0576 |
| 0.1492 | 0.0644 | 0.0864 |
| 0.05758 | 0.05758 | 0.2849 |

(나) $\lambda = 0.3072$, $\Phi_0^2 = 0.8$

| | | |
|---|---|---|
| 0.0077 | 0.0077 | 0.0846 |
| 0.0154 | 0.1539 | 0.0307 |
| 0.1962 | 0.0577 | 0.0461 |
| 0.0307 | 0.0307 | 0.3386 |

(다) $\lambda = 0.1168$, $\Phi_0^2 = 1.3$

| | | |
|---|---|---|
| 0.0029 | 0.0029 | 0.0942 |
| 0.0058 | 0.1825 | 0.0117 |
| 0.2296 | 0.0529 | 0.0175 |
| 0.0117 | 0.0117 | 0.3766 |

$\Phi^2$을 가장 크게하는 결합분포는 그림 2.1과 같은 greedy 알고리즘을 통해 쉽게 얻을 수 있다. 그림 2.1의 알고리즘 이외에 두 변수가 최대연관을 갖는 극단표(extreme tables)는 Kalantari 등(1993)의 발견법적인 방법들과 Hu와 Murkerjee(2002)의 나무구조를 이용한 알고리즘 등에 의해 구할 수 있다.

예제 2.1: 행 주변분포 $\{0.1, 0.2, 0.3, 0.4\}$와 열 주변분포 $\{0.25, 0.25, 0.5\}$로 부터 적절한 연관정도를 갖는 이차원 결합분포를 생성해보기로 한다. 먼저 그림 2.1의 알고리즘을 적용하면 표 2.1과 같은 최대연관을 갖는 $P_{MAX}$를 얻을 수 있고, 이 때 $\Phi^2 = 1.6667$이다. 이로부터

$\Phi_0^2$값이 0.7, 1.0, 1.3인 식 (2.2)의 $\lambda$를 이분법(bisection methods)을 이용하여 각각 0.3519, 0.2254, 0.1168와 같이 구할 수 있다. 결국, 이들 $\lambda$와 $P_{MAX}$ 그리고 식 (2.2)에 의하여 표 2.2와 같은 결합분포들을 얻을 수 있다.

## 3. 다차원 임의 분할표로의 확장

여러 범주형 변수에 의한 다차원 분할표 분석의 주된 관심은 분할표를 구성하는 변수들 간의 연관구조를 식별하는데 있다. 이를 위해 다수의 모형이 제안되어 있으며, 특히 로그선형모형은 계층구조(hierarchicla structure) 하에서 모형에 포함된 모수들에 의해 조건부 독립성을 이용하여 변수들의 연관구조를 해석할 수 있기 때문에 가장 널리 이용되고 있다. 또한 계층구조 하에서 모형에 의한 기대칸 확률의 추정값은 충분주변합 혹은 형상의 함수인 직접해(direct estimator) 또는 반복비율적합 방법에 의해 구하게 된다. 다시 말해 칸 확률에 의해 정의된 로그선형모형은 주변분포가 곧 형상이 되며, 만일 이들 형상이 주어진다면 이들의 함수인 직접해를 위한 식 또는 반복비율적합 방법에 의하여 구하고자 하는 다차원 결합분포를 얻을 수 있다.

각각 $I_1$, $I_2$, $I_3$개 범주를 갖는 세 범주형 변수 $X_1$, $X_2$, $X_3$를 고려하기로 하자. 세 변수가 완전한 독립이라면 식 (2.1)에서와 같이 다음과 같은 삼차원 결합분포를 얻을 수 있다.

$$p_{ijk} = p_{i++}p_{+j+}p_{++k},\ i = 1,\ldots,I_1,\ j = 1,\ldots,I_2,\ k = 1,\ldots,I_3, \tag{3.1}$$

여기서 $p_{i++}$, $p_{+j+}$, $p_{++k}$ 는 변수 $X_1$, $X_2$, $X_3$의 주변분포를 나타낸다. 그러므로 식 (3.1)에서 결정된 결합분포 $p_{ijk}$를 이용하여 inversion 알고리즘 등에 의해 삼차원 임의 분할표를 얻을 수 있다. 그러나 이차원에서와 마찬가지로 삼차원 역시 세 변수들간의 연관구조가 결정되고 이들 변수들이 적절한 연관을 갖는 분할표를 생성하는 문제에 더 관심을 갖게 된다.

만일 세 변수들 중에서 두 변수는 서로 연관이 있고 나머지 다른 한 변수와는 독립이라면, 이러한 연관구조를 가진 삼차원 결합분포는 다음과 같은 관계를 통해 얻을 수 있다.

$$\begin{aligned} p_{ijk} &= p_{ij+}p_{++k}, \\ p_{ijk} &= p_{i+k}p_{+j+}, \\ p_{ijk} &= p_{+jk}p_{i++}, \end{aligned} \tag{3.2}$$

여기서 $p_{ij+}$는 변수 $X_1$과 $X_2$의 이차 주변분포를 의미하며, $p_{i+k}$는 변수 $X_1$과 $X_3$ 그리고 $p_{+jk}$는 변수 $X_2$와 $X_3$의 이차 주변분포를 의미한다. 이때 이들 세 이차원 주변분포는 제2절에서 제안한 방법에 의하여 적절한 연관정도를 가지도록 결정될 수 있음을 상기하자. 이제 이러한 두 변수와 한 변수의 독립을 나타내는 연관구조 이외에 다음과 같은 한 변수가 주어진 경우에 다른 두 변수의 조건부 독립을 나타내는 연관구조를 고려하기로 한다.

$$\begin{aligned} p_{ijk} &= \frac{p_{ij+}p_{i+k}}{p_{i++}}, \\ p_{ijk} &= \frac{p_{ij+}p_{+jk}}{p_{+j+}}, \\ p_{ijk} &= \frac{p_{i+k}p_{+jk}}{p_{++k}}. \end{aligned} \tag{3.3}$$

이 경우 역시 차원축소(collapsibility) 성질에 의해 두 이차원 주변분포에 의해 삼차원 결합분포가 결정되는 점에 주목하자. 따라서 한 변수가 주어진 경우에 다른 두 변수가 서로 독립인 조건부 독립 구조를 가진 삼차원 결합분포 역시 제2절에서 제안한 방법에 의하여 해당 이차원 주변분포를 생성하여 얻을 수 있다는 사실을 알 수 있다. 이에 더하여 세 변수가 모든 가능한 쌍에 대한 연관이 존재하는 부분연관(partial association)을 갖는 경우를 고려해보기로 한다.

부분연관이 존재하는 경우에는 식 (3.1), (3.2), (3.3)과 같은 식이 존재하지 않는다. 따라서 세 변수간 부분연관이 존재하는 경우에는 Gange(1995)에서와 같이 반복비율적합 방법과 같은 반복 계산법에 의해 삼차 결합분포를 얻어야 한다. 반복은 주어진 연관정도를 가진 세 이차 주변분포 $p_{ij+}$, $p_{ij+}$, $p_{i+k}$와 적절한 초기값 $p_{ijk}^{(0)}$를 이용하여 다음과 같은 세 단계를 거친다.

1: $p_{ijk}^{(3r)} = p_{ijk}^{(3r-1)} \frac{p_{ij+}}{p_{ij+}^{(3r-1)}}$

2: $p_{ijk}^{(3r-1)} = p_{ijk}^{(3r-2)} \frac{p_{i+k}}{p_{i+k}^{(3r-2)}}$

3: $p_{ijk}^{(3r-2)} = p_{ijk}^{(3r-3)} \frac{p_{+jk}}{p_{+jk}^{(3r-3)}}$

이들 세 단계는 임의의 허용오차 $\delta > 0$에 이를 때까지 반복된다. 즉, $|p_{ijk}^{(3r)} - p_{ijk}^{(3r-3)}| < \delta$, 여기서 $r$은 반복주기를 나타낸다. 결국 세 이차 주변분포에 의한 부분연관을 만족하는 삼차 결합분포는 $p_{ijk} = p_{ijk}^{(3r)}$과 같이 구할 수 있다. 한가지 지적할 점은 식 (3.1), (3.2) 그리고 (3.3)과 같이 직접해를 갖는 경우에도 반복비율적합을 이용하여 결합분포를 얻을 수 있다는 점이다.

예제 3.1: 예 2.1에서 사용하였던 $\{0.1, 0.2, 0.3, 0.4\}$, $\{0.25, 0.25, 0.5\}$와 $\{0.4, 0.6\}$를 각각 세 변수 $X_1$, $X_2$, $X_3$의 일차 주변분포라 하자. 이들 세 주변분포를 이용하여 부분연관관계를 갖는 삼차원 결합분포를 생성하고자 한다. 먼저 두 변수 $X_1$, $X_2$를 위한 이차 주변분포는 예제 1의 결과를 이용하기로 하며, 즉 $\Phi_{12}^2 = 0.3$, 변수 $X_1$과 $X_3$ 그리고 $X_2$와 $X_3$의 연관은 각각 $\Phi_{13}^2 = 0.2$, $\Phi_{23}^2 = 0.1$과 같이 부여하기로 한다. 표 3.1은 이러한 세 연관정도에 의한 이차 결합분포들을 보여주고 있다. 표 3.1의 세 이차 주변분포를 이용하여 부분연관을 만족하는 삼차원 결합분포는 반복비율적합을 적용하여 표 3.2와 같이 얻을 수 있다.

이상과 같은 삼차원 임의 분할표를 얻기 위한 결합분포를 구하는 방법을 사차원 이상으로 확장하기 위해서는 세 변수가 서로 완전히 연관된 결합분포를 구할 수 있어야 한다. 이

표 3.1: 세 이차 결합분포들

(가) $\lambda = 0.5757$, $\Phi_{12}^2 = 0.3$

| | | |
|---|---|---|
| 0.0144 | 0.0144 | 0.0712 |
| 0.0288 | 0.1136 | 0.0576 |
| 0.1492 | 0.0644 | 0.0864 |
| 0.05758 | 0.05758 | 0.2849 |

(나) $\lambda = 0.5528$, $\Phi_{13}^2 = 0.2$

| | |
|---|---|
| 0.0221 | 0.0779 |
| 0.0442 | 0.1558 |
| 0.0663 | 0.2337 |
| 0.2673 | 0.1327 |

(다) $\lambda = 0.6349$, $\Phi_{23}^2 = 0.1$

| | |
|---|---|
| 0.1183 | 0.1317 |
| 0.1548 | 0.0952 |
| 0.1270 | 0.3730 |

표 3.2: 부분연관을 만족하는 삼차원 결합분포

| | $k=1$ | | | $k=2$ | | |
|---|---|---|---|---|---|---|
| | $j=1$ | $j=2$ | $j=3$ | $j=1$ | $j=2$ | $j=3$ |
| $i=1$ | 0.010869 | 0.012775 | 0.000000 | 0.002926 | 0.001019 | 0.077345 |
| $i=2$ | 0.004447 | 0.042841 | 0.000000 | 0.023142 | 0.066068 | 0.062531 |
| $i=3$ | 0.037359 | 0.033573 | 0.000000 | 0.105675 | 0.028140 | 0.093797 |
| $i=4$ | 0.065583 | 0.065583 | 0.126970 | 0.000000 | 0.000000 | 0.139357 |

를 위하여 식 (2.2)를 확장한 다음과 같은 선형결합에 의한 결합분포를 고려하기로 한다.

$$P(\lambda) = \lambda P_{PART} + (1-\lambda)P_{MAX}, \tag{3.4}$$

여기서 $P_{PART}$는 예 3.1에서 얻은것과 같은 부분연관을 가진 결합분포를 의미하며, $P_{MAX}$는 세 변수가 서로 완전한 독립인 결합분포 $P_{IND}$로부터 가장 멀리 떨어진 최대 $\Phi^2$값을 갖는 삼차원 결합분포를 나타낸다.

식 (3.4)를 이용하여 세 변수가 완전한 연관을 갖는 삼차원 결합분포를 얻기 위해서는 먼저, 세 변수의 연관정도를 측정할 수 있는 측도가 정의되어야 한다. 이를 위해 두 분포의 적합성을 측정하기 위한 Pearson의 통계량을 직접 이용하기로 한다. 즉, $P_{MAX}$를 참 분포라고 가정하고 $P_{IND}$의 적합성을 측정하기 위한 Pearson의 통계량을 $\Phi^2_{MAX}$로 나타내기로 한다. 마찬가지로 $P_{PART}$를 참 분포라고 가정하고 $P_{PART}$의 적합성을 측정하기 위한 Pearson의 통계량을 $\Phi^2_{PART}$로 나타내기로 한다. $P_{IND}$의 Pearson의 통계량 값은 '0'이므로 $0 \le \Phi^2_{PART} \le \Phi^2_{MAX}$과 같은 관계가 성립한다. 그러므로 $\Phi^2_{PART} \le \Phi^2_0 \le \Phi^2_{MAX}$을 만족하는 $\lambda_0$ 역시 존재한다. 이러한 사실로부터 $\Phi^2_{PART} \le \Phi^2_0 \le \Phi^2_{MAX}$을 만족하는 $\Phi^2_0$을 세 변수의 완전한 연관을 측정하기 위한 측도로 이용하기로 한다. 결국 식 (3.4)에 의한 결합분포를 얻기 위해서는 $P_{IND}$로 부터 가장 멀리 떨어진 $P_{MAX}$를 구하는 문제만이 남게 된다. 이를 위해 그림 3.1과 같은 주어진 세 일차 주변분포를 만족하는 $P_{MAX}$를 얻기 위해 그림 2.1의 알고리즘을 직접 확장한 알고리즘을 제안하기로 한다.

예제 3.2: 앞의 예 3.1에서 사용하였던 일차원 주변분포들을 이용하여 그림 3.1의 알고리즘을 적용하면 표 3.3과 같은 $P_{IND}$에서 가장 멀리 떨어진 $P_{MAX}$를 얻을 수 있다. $\Phi^2_{MAX} = 3.9444$이며 예 2의 $P_{PART}$는 $\Phi^2_{PART} = 0.5755$와 같이 구해진다.

이러한 사실로부터 $\Phi^2_0 = 1.5$ 를 갖는 세 변수가 서로 완전한 연관을 갖는 삼차원 결합분포는 식 (3.4)를 이용하여 얻을 수 있다. 여기서 $\Phi^2_0 = 1.5$일 때의 $\lambda_0 = 0.3863$이다.

**Algorithm 2** : ExtremeForThree

**Require:** 세 일차 주변분포 $p_{i++}$, $p_{+j+}$, $p_{++k}$ $\forall i, j, k$

**Ensure:** 삼차원 결합분포 $P = \{p_{111}, p_{112}, \cdots, p_{I_1 I_2 I_3}\}$

$p_{ijk} = 0, \forall i, j, k$
$p'_{i++} = p_{i++}$ , $\forall i$
$p'_{+j+} = p_{+j+}$ , $\forall j$
$p'_{++k} = p_{++k}$ , $\forall k$
**for** $i = 1, \ldots, I_1$ **do**
  **for** $j = 1, \ldots, I_2$ **do**
    **for** $k = 1, \ldots, I_3$ **do**
      $q_{ijk} = \min(p'_{i++}, p'_{+j+}, p'_{++k})$, $\forall i, j, k$
      **if** ($p'_{i++} > 0$ and $p'_{+j+} > 0$ and $p'_{++k} > 0$) **then**
        $p_{ijk} = q_{ijk}$
        $p'_{i++} = p'_{i++} - q_{ijk}$
        $p'_{+j+} = p'_{+j+} - q_{ijk}$
        $p'_{++k} = p'_{++k} - q_{ijk}$
      **end if**
    **end for**
  **end for**
**end for**

그림 3.1: 최대연관을 갖는 삼차원 결합분포 생성 알고리즘

표 3.3: 세 변수의 최대연관을 갖는 삼차원 결합분포

| | $k=1$ | | | $k=2$ | | |
|---|---|---|---|---|---|---|
| | $j=1$ | $j=2$ | $j=3$ | $j=1$ | $j=2$ | $j=3$ |
| $i=1$ | 0 | 0 | 0 | 0.05 | 0 | 0 |
| $i=2$ | 0 | 0 | 0 | 0 | 0.10 | 0 |
| $i=3$ | 0 | 0.20 | 0 | 0.15 | 0 | 0 |
| $i=4$ | 0.10 | 0 | 0 | 0 | 0 | 0.40 |

표 3.4: $\Phi_0^2 = 1.5$를 갖는 삼차원 결합분포

| | $k=1$ | | | $k=2$ | | |
|---|---|---|---|---|---|---|
| | $j=1$ | $j=2$ | $j=3$ | $j=1$ | $j=2$ | $j=3$ |
| $i=1$ | 0.0041996 | 0.0000000 | 0.0165535 | 0.0451157 | 0.0000000 | 0.0253406 |
| $i=2$ | 0.0011305 | 0.0298853 | 0.0255282 | 0.0408318 | 0.0976030 | 0.0000000 |
| $i=3$ | 0.0049363 | 0.1244400 | 0.0000000 | 0.1050136 | 0.0253406 | 0.0490602 |
| $i=4$ | 0.0617547 | 0.0089418 | 0.0241614 | 0.0108732 | 0.0000000 | 0.2992901 |

## 4. 결 론

로그선형모형에 기반을 둔 다차원 임의 분할표를 생성하는 방법을 제안하였다. 제안된 방법은 로그선형모형의 형상을 이용하여 반복비율적합 방법에 의해 다차원 결합분포를 생성하며, 이렇게 얻어진 결합분포를 통해 inversion 알고리즘을 이용하여 임의 분할표를 얻을 수 있다. 모형에서 결정되는 형상들의 연관정도는 Pearson 통계량을 이용하였으며, Lee(1997)의 제안을 적용한 선형결합을 통해 적절한 연관정도를 가진 결합분포를 결정하게 된다.

이차원 임의 분할표에서 얻어진 이러한 결과를 세 변수가 적절한 완전연관을 갖는 삼차원 결합분포를 생성하는 방법으로 확장하였으며, 이를 위해 서로 독립인 분포로 부터 가장 멀리 떨어진 최대 $\Phi^2$ 값을 갖는 결합분포를 얻기 위한 greedy 알고리즘을 제안하였다. 따라서 제안된 방법을 통해 삼차원 결합분포를 쉽게 생성할 수 있으며, 제안된 방법은 사차원 이상 다차원 임의 분할표를 생성하는 방법으로 어렵지 않게 확장될 수 있다.

## 참고문헌


Davis, C. S. (1993). The computer generation of multidimensional random variates, *Computational Statistics & Data Analysis*, **16**, 205-217.

Emrich, L. J., and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates, *The American Statistician*, **45**, 302-304.

Gange, S. J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm, *The American Statistician*, **49**, 134-138.

Hu, X. and Mukerjee H. (2002). Constructing all extremes of contingency tables with given marginals, *Journal of Computational and Graphical Statistics*, **11**, 910-919.

Kalantari, B., Lari, I., Rizzi, A., and Simeone, 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*, **16**, 19-34.

Kang, S. H., and Jung, S. H. (2001). Generating correlated binary variables with complete specification of the joint distribution, *Biometrical Journal*, **43**, 263-269.

Lee, A. J. (1993). Generating random binary Deviates having fixed marginal distributions and specified degrees of association, *The American Statistician*, **47**, 209-215.

Lee, A. J. (1997). Some simple methods for generating correlated categorical variates, *Computational Statistics & Data Analysis*, **26**, 133-148.

Mirkin, B. (2001). Eleven ways to look at the chi-squared coefficient for contingency tables, *The American Statistician*, **55**, 111-120.

Park, C. G., Park, T. S., and Shin, D. W. (1996). A simple method for generating correlated binary variates, *The American Statistician*, **50**, 306-310.


[ 2006년 1월 접수, 2006년 4월 채택 ]

# Generating Multidimensional Random Tables*

Hyun Jip Choi[1)]

ABSTRACT

We suggest a method for generating multidimensional random tables based on the log-linear models. A linear combination approach by Lee(1997) is applied to get the joint distribution with the well known Pearson chi-squared statistics. We can generate completely associated joint distributions which have the fixed association among three variables by using the suggested method. Therefore the method can be extended to more higher dimension than the three dimensional tables.

*Keywords:* Random number; Pearson chi-squared statistics; Multidimensional contingency tables; Log-linear models.

* This research was supported by Kyonggi University Research Grant 2004.

1) Associate Professor, Department of Applied Information Statistics, Kyonggi University, Yeongtong-gu, Suwon, Kyonggi-Do 443-760, Korea.
E-mail: hjchoi@kyonggi.ac.kr