2005 Proceedings of the Spring Conference, Korean Statistical Society

실험계획을 평가하기 위한 측도로서의 상호정보

장대홍[1)]

요 약

통계적 품질관리나 실험계획법에서 요인의 수가 과다하게 많은 경우 주로 직교배열을 이용하여 실험을 한다. 그러나 직교배열을 쓰지 못할 때 우리는 근사직교배열을 이용하게 되는 데 이 때 근사직교배열의 직교성의 정도를 평가할 수 있는 기준으로서 상호정보를 이용할 수 있다.

주요용어 : (근사)직교배열, 상호정보

1. 서론

요인배치법을 완전하게 쓸 수 없을 때 요인배치법의 일부실시법(fractional factorial design), 포화계획(saturated design) 또는 초포화계획(supersaturated design)을 사용한다. Booth와 Cox(1962) 이후 많은 학자들이 이러한 (초)포화계획에 대하여 연구하였다(예로, 2001년 이후 Cela외 2인(2001), Li와 Lin(2002), Yamada와 Lin(2002), Yamada 와 Matsui(2002), Georgiou외 2인(2003), Fang외 2인(2003), Lu외 2인(2003), Aggarwal과 Gupta(2004), Fang외 3인(2004), Koukouvinos와 Stylianou(2004), Li외 2인(2004), Georgiou외 2인(2005), Yamada외 5인(2005) 등이 있다.). 통계적 품질관리나 실험계획법에서 요인의 수가 과다하게 많은 경우 주로 직교배열을 이용하여 실험을 한다. Plackett와 Burman(1946)이 포화계획(saturated design)인 Plackett-Burman design을 제안한 이후 직교배열이 실험계획분야에서 다양하게 쓰이고 있다. 직교배열은 다구찌방법의 중요한 수단으로 쓰이고 있다. 그러나, 직교계획을 쓰지 못할 때 우리는 근사직교배열(nearly orthogonal arrays)을 이용하게 되는 데 Wang과 Wu(1992)는 이러한 근사직교배열을 제안하였다. 우리가 실험계획으로서 직교배열이 아닌 근사직교배열을 쓰는 경우 이 근사직교배열의 직교성의 정도를 아는 것이 중요하다. Ma외 2인(2000)는 이러한 근사직교배열의 직교성의 정도를 평가하기 위한 기준을 이용하여 근사직교배열을 구하는 방법들을 제시하였다. Jang(2002, 2004)은 기존의 직교성 판정기준을 기반으로 하여 근사직교배열의 직교성의 정도를 평가하기 위한 측도들과 그래픽방법들을 제안하였다.

근사직교배열의 직교성의 정도를 평가할 수 있는 기존의 기준들 외에 또 다른 기준으로서 우리는 상호정보를 이용할 수 있다.

2. 직교성의 정도를 평가하는 기준들

우리는 강도(strenth) 2인 직교배열을 직교계획(orthogonal design)이라 부른다. r 제약을 갖으며 강도가 t이고, 크기가 N인 직교배열이란, 각 열이 2개 이상의 수준을 갖는 $N \times r$ 행렬 A에서 A의 $N \times t$ 부분행렬 각각이 모든 가능한 $1 \times t$ 행벡터들을 같은 빈도로 갖는 행렬을

1) (608-737) 부산광역시 남구 대연3동 599-1 부경대학교 수리과학부 통계학전공, 교수
E-mail: dhjang@pknu.ac.kr

말한다. $L_N(q_1^{m_1}\times q_2^{m_2}\times...\times q_k^{m_k})$, $\sum_{i=1}^{k}m_i=r$를 q_i개의 수준을 갖는 열들이 m_i개인 직교계획이라 하자. 직교계획의 정의는 다음과 같다.

1.(직교조건 1) 각 열에서 각 수준은 같은 빈도수로 나타난다.
2.(직교조건 2) 임의의 두 개의 열들에서 수준의 조합이 같은 빈도수로 나타난다.

종종 (조건 1)을 만족하는 계획을 U-형 계획이라 칭한다.

우리가 실험계획으로서 직교배열이 아닌 근사직교배열을 쓰는 경우 이 근사직교배열의 직교성의 정도를 아는 것이 중요하다. 논의의 간편성을 위하여 $N\times r$ 행렬인 배열 A는 U-형 계획이고, $A=(c_1, c_2, ..., c_r)$라고 하자. 그리고 i번째 열의 구성요소가 $1,2,...,q$로 이루어졌다고 가정하자.

근사직교배열 A의 임의의 두 개의 열 c_i와 c_j의 직교성을 평가하는 기준을 $d(c_i, c_j)$라 할 때 배열 A의 직교성의 정도를 평가할 수 있는 기준들로서 다음과 같은 기준들이 있다.

$d(c_i, c_j)=$

1. s_{ij}^2(두 개의 열 c_i, c_j의 곱의 합)

2. $\sum_{k=1}^{q}\sum_{l=1}^{q}\left|N_{c_i,c_j}(k,l)-\frac{N}{q^2}\right|$

3. $\sum_{k=1}^{q}\sum_{l=1}^{q}\left(N_{c_i,c_j}(k,l)-\frac{N}{q^2}\right)^2$

4. $\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{\left(N_{c_i,c_j}(k,l)-\frac{N}{q^2}\right)^2}{\frac{N}{q^2}}$

5. J_{ij}(두 개의 열 c_i, c_j에 대한 균등지수)

여기서, c_i, c_j는 A의 임의의 두 열이고, $N_{c_i,c_j}(k,l)$은 c_i와 c_j의 두 개의 열들에서 (k,l)쌍의 개수이고, $\frac{N}{q^2}$은 c_i와 c_j의 두 개의 열들에서 수준쌍들의 평균개수이다.

기준 1-4는 $d(c_i, c_j)$가 0에 가까울수록 직교성을 만족하게 되고 기준 5는 $d(c_i, c_j)$가 1에 가까울수록 직교성을 만족하게 된다.

$d(c_i, c_j)$를 계산하면 $d(c_i, c_j)$을 구성요소로 갖는 직교화평가행렬(장대흥(2002) 참조)인 $r\times r$ 행렬 $D_p=(d(c_i, c_j))(p=1,2,3,4,5)$를 정의할 수 있다. 여기서, p는 p번째 기준을 가리킨다. 이 행렬 D의 각 원소의 크기를 통하여 우리는 직교성의 정도를 파악할 수 있다. 또한, 가장 극단적인, 즉 가장 직교성에서 먼 경우의 $d(c_i, c_j)$값을 m_p라 하면 $I(i,j)=\frac{d(c_i, c_j)}{m_p}$를 구성요소로 갖는 비직교화지수행렬인 $r\times r$행렬 $D_p'=(I(i,j))$를 정의할 수 있다($p=5$일 때는 직교하는 경우의 $d(c_i, c_j)$의 값 1이 m_5가 되고 D_5'는 직교화지수행렬이 됨.). 이 행렬 D_p'의 각 원소의 크기를 통하여 우리는 직교성의 정도를 파악할 수 있다.

앞에서 정의한 5가지 기준들을 사용하여 구한 $I(i,j)$를 이용하여 배열 A의 직교성의 정도

를 평가할 수 있는 측도들로서 다음과 같은 두 가지의 측도들을 정의할 수 있다.

1. $ave(D'_p) = \frac{1}{\binom{r}{2}} \sum_{1 \le i < j \le r} I(i,j)$ 2. $\max(D'_p) = \max_{1 \le i < j \le r} I(i,j)$

(기준 5는 max 대신 min으로 바꿈)

기준 1-4는 $ave(D'_p)$나 $max(D'_p)$가 0에 가까울수록 직교배열에 가깝게 되고 기준 5는 $ave(D'_p)$나 $min(D'_p)$가 1에 가까울수록 직교배열에 가깝게 된다.

3. 직교성의 정도를 평가할 수 있는 기준으로서의 상호정보

상호정보(mutual information)는 다음과 같이 정의된다(Hutter와 Zaffalon(2005) 참조).

1. (연속형) 2. (이산형)

$$I_{XY} = \int\int f(x,y) log \frac{f(x,y)}{f_1(x) f_2(y)} dxdy \qquad I_{XY} = \sum_x \sum_y f(x,y) log \frac{f(x,y)}{f_1(x) f_2(y)}$$

여기서, $f(x,y)$는 확률변수(또는 확률벡터) X와 Y의 결합확률(밀도)함수이고, $f_1(x)$는 확률변수(또는 확률벡터) X의 주변확률(밀도)함수이고, $f_2(y)$는 확률변수(또는 확률벡터) Y의 주변확률(밀도)함수이다.

$I_{XY} \geq 0$이 항상 성립한다. 그리고 $I_{XY} = 0$이면 두 확률변수(확률벡터)는 서로 독립이 되고, $I_{XY} > 0$이면 두 확률변수(확률벡터)는 서로 종속이 된다. 두 변수가 범주형 변수일 때는 다음과 같이 상호정보의 추정량이 주어진다.

$$\hat{I}_{XY} = \sum_i \sum_j \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i+} n_{+j}} \tag{1}$$

여기서, n은 분할표 상의 총도수, n_{ij}는 범주형 변수 X에서 i번째 범주와 범주형 변수 Y에서 j번째 범주에 해당하는 칸(cell)의 도수, n_{i+}는 변수 X에서 i번째 범주의 도수, n_{+j}는 변수 Y에서 j번째 범주의 도수이다.

식 (1)을 이용하여 우리는 근사직교배열 A의 직교성의 정도를 평가할 수 있는 기준으로서 다음과 같은 기준을 제시할 수 있다.

$$d(\mathbf{c}_i, \mathbf{c}_j) = \sum_{k=1}^{q} \sum_{l=1}^{q} \frac{N_{\mathbf{c}_i, \mathbf{c}_j}(k,l)}{N} \log_2 \frac{N_{\mathbf{c}_i, \mathbf{c}_j}(k,l) N}{N_{\mathbf{c}_i}(k) N_{\mathbf{c}_j}(l)} \tag{2}$$

여기서, N은 직교배열의 크기, 즉 행의 수이고, $N_{\mathbf{c}_i, \mathbf{c}_j}(k,l)$은 $\mathbf{c}_i$와 $\mathbf{c}_j$의 두 개의 열들에서

(k, l)쌍의 개수이고, $N_{c_i}(k)$는 c_i에서 k번째 구성요소의 개수, $N_{c_j}(l)$은 c_j에서 l번째 구성요소의 개수이다. 이 기준은 $d(c_i, c_j)$가 0에 가까울수록 직교성을 만족하게 된다.

$d(c_i, c_j)$를 계산하면 $d(c_i, c_j)$을 구성요소로 갖는 직교화평가행렬인 $r \times r$행렬 $D_6 = (d(c_i, c_j))$를 정의할 수 있다. 이 행렬 D_6의 각 원소의 크기를 통하여 우리는 직교성의 정도를 파악할 수 있다. 또한, 가장 극단적인, 즉 가장 직교성에서 먼 경우의 $d(c_i, c_j)$값을 m_6라 하면 $I(i,j) = \dfrac{d(c_i, c_j)}{m_6}$를 구성요소로 갖는 비직교화지수행렬인 $r \times r$행렬 $D_6' = (I(i,j))$를 정의할 수 있다. 이 행렬 D_6'의 각 원소의 크기를 통하여 우리는 직교성의 정도를 파악할 수 있다.

식 (2)의 기준에 의하여 구한 $I(i,j)$를 이용하여 배열 A의 직교성의 정도를 평가할 수 있는 측도들로서 다음과 같은 두 가지의 측도들을 정의할 수 있다.

1. $ave(D_6') = \dfrac{1}{\binom{r}{2}} \sum_{1 \le i < j \le r} I(i,j)$

2. $\max(D_6') = \max_{1 \le i < j \le r} I(i,j)$

$ave(D_6')$나 $max(D_6')$가 0에 가까울수록 직교배열에 가깝게 된다.

4. 수치 예

다음 표 1과 같은 배열에 대하여 고려하여 보자. 이 배열은 Fang외 3인(2000)이 제안한 균등계획(uniform design) $U_{16}(4^5)$이다. 각 열의 구성요소가 4개로 동일하고, 각 열에서 각 구성요소의 개수가 같으므로 이 배열은 U-계획이 된다. 그러나 이 배열은 직교배열이 아닌 근사직교배열이 된다. 이 배열에 대하여 빈도행렬(장대홍(2004) 참조)을 구하면 다음 표2와 같다.

표 1. 배열 A

표 2. 배열 A에 대한 빈도행렬

$c_1 \begin{pmatrix}4\\4\\4\\4\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&0&2&1\\1&2&0&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&0&2&1\\0&2&0&2\\2&0&1&1\\1&2&1&0\end{pmatrix}$	$\begin{pmatrix}0&1&2&1\\2&0&1&1\\1&2&0&1\\1&1&1&1\end{pmatrix}$
$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$c_2 \begin{pmatrix}4\\4\\4\\4\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\2&0&1&1\\0&2&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}0&1&2&1\\1&2&0&1\\2&0&0&2\\1&1&2&0\end{pmatrix}$
$\begin{pmatrix}1&1&1&1\\1&0&2&1\\1&2&0&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$c_3 \begin{pmatrix}4\\4\\4\\4\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}0&1&1&2\\1&1&1&1\\1&1&1&1\\2&1&1&0\end{pmatrix}$
$\begin{pmatrix}1&0&2&1\\0&2&0&2\\2&0&1&1\\1&2&1&0\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\2&0&1&1\\0&2&1&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{pmatrix}$	$c_4 \begin{pmatrix}4\\4\\4\\4\end{pmatrix}$	$\begin{pmatrix}0&4&0&0\\1&0&2&1\\1&0&2&1\\2&0&0&2\end{pmatrix}$
$\begin{pmatrix}0&1&2&1\\2&0&1&1\\1&2&0&1\\1&1&1&1\end{pmatrix}$	$\begin{pmatrix}0&1&2&1\\1&2&0&1\\2&0&0&2\\1&1&2&0\end{pmatrix}$	$\begin{pmatrix}0&1&1&2\\1&1&1&1\\1&1&1&1\\2&1&1&0\end{pmatrix}$	$\begin{pmatrix}0&4&0&0\\1&0&2&1\\1&0&2&1\\2&0&0&2\end{pmatrix}$	$c_5 \begin{pmatrix}4\\4\\4\\4\end{pmatrix}$

$$A=\begin{pmatrix} 1 & 1 & 4 & 3 & 3 \\ 1 & 2 & 2 & 1 & 2 \\ 1 & 3 & 1 & 4 & 4 \\ 1 & 4 & 3 & 3 & 3 \\ 2 & 1 & 1 & 2 & 3 \\ 2 & 2 & 3 & 4 & 1 \\ 2 & 3 & 3 & 2 & 4 \\ 2 & 4 & 4 & 4 & 1 \\ 3 & 1 & 2 & 4 & 4 \\ 3 & 2 & 4 & 1 & 2 \\ 3 & 3 & 2 & 3 & 1 \\ 3 & 4 & 1 & 1 & 2 \\ 4 & 1 & 3 & 1 & 2 \\ 4 & 2 & 1 & 3 & 4 \\ 4 & 3 & 4 & 2 & 1 \\ 4 & 4 & 2 & 2 & 3 \end{pmatrix}$$

이 빈도행렬을 참조하여 식 (2)를 이용한 $d(c_i, c_j)$를 계산할 수 있다. 예로, $d(c_4, c_5)$와 가장 극단적인, 즉 가장 직교성에서 먼 경우를 계산하여 보면

c_4 \ c_5	1	2	3	4	sum
1	0	4	0	0	4
2	1	0	2	1	4
3	1	0	2	1	4
4	2	0	0	2	4
sum	4	4	4	4	16

c_i \ c_j	1	2	3	4	sum
1	4	0	0	0	4
2	0	4	0	0	4
3	0	0	4	0	4
4	0	0	0	4	4
sum	4	4	4	4	16

$$d(c_4, c_5) = 7\times\frac{0}{16}\times log_2\frac{0\times 16}{4\times 4} + 4\times\frac{1}{16}\times log_2\frac{1\times 16}{4\times 4} + 4\times\frac{2}{16}\times log_2\frac{2\times 16}{4\times 4}$$
$$+1\times\frac{4}{16}\times log_2\frac{4\times 16}{4\times 4} = 1$$
$$d(c_i, c_j) = 12\times\frac{0}{16}\times log_2\frac{0\times 16}{4\times 4} + 4\times\frac{4}{16}\times log_2\frac{4\times 16}{4\times 4} = 2 = m_6$$

이므로 $0 \leq d(c_i, c_j) \leq 2$이 성립한다. 이렇게 구한 $d(c_i, c_j)$를 이용하여 $I(i,j)$를 구한 후 비직교화지수행렬인 $r\times r$행렬 D_6'를 다음과 같이 구할 수 있다.

$$D_6' = \begin{pmatrix} - & 0 & 0.125 & 0.3125 & 0.1875 \\ 0 & - & 0 & 0.125 & 0.3125 \\ 0.125 & 0 & - & 0 & 0.125 \\ 0.3125 & 0.125 & 0 & - & 0.5 \\ 0.1875 & 0.3125 & 0.125 & 0.5 & - \end{pmatrix}$$

$I(4,5)$의 값 0.5가 제일 커 직교성에서 제일 멀고, 1열과 2열, 2열과 3열, 3열과 4열끼리는 직교함을 알 수 있다. $I(i,j)$를 이용하여 배열 A의 직교성의 정도를 평가할 수 있는 측도을 계산하면 $ave(D_6') = 0.1688$, $max(D_6') = 0.5$가 된다. 이 두개의 측도들이 모두 0이 아니므로 이 배열은 직교배열이 아님을 알 수 있다.

비교를 위하여 기존의 판정기준 중 두 번째 판정기준을 이용하여 보자. 이 두 번째 판정기준에서 가장 극단적인, 즉 가장 직교성에서 먼 경우까지를 고려하면 $0 \le d(c_i, c_j) \le 24$이 성립한다. $d(c_i, c_j)$를 이용하여 $I(i,j)$를 구한 후 비직교화지수행렬인 $r \times r$행렬 D_2'를 구하면 다음과 같다.

$$D_2' = \begin{pmatrix} - & 0 & 0.1667 & 0.4167 & 0.25 \\ 0 & - & 0 & 0.1667 & 0.4167 \\ 0.1667 & 0 & - & 0 & 0.1667 \\ 0.4167 & 0.1667 & 0 & - & 0.5833 \\ 0.25 & 0.4167 & 0.1667 & 0.5833 & - \end{pmatrix}$$

$I(i,j)$를 이용하여 배열 A의 직교성의 정도를 평가할 수 있는 측도을 계산하면 $ave(D_2') = 0.2167$, $max(D_2') = 0.5833$이 된다.이 두개의 측도들이 모두 0이 아니므로 이 배열은 직교배열이 아님을 알 수 있다.

이 두 가지 판정기준을 이용한 측도들의 값이 서로 비슷한 값을 가짐을 알 수 있다.

5. 결론

본 논문을 통하여 상호정보를 이용하여 근사직교배열의 직교성의 정도를 평가할 수 있음을 알았다. 우리는 이 기준을 기존의 다른 기준들과 비교하여 볼 수 있다.

참고문헌

장대흥 (2004), 근사직교배열의 직교성의 정도를 평가하기 위한 그래픽방법, 한국품질경영학회지, 32권 4호, 220-228.

Aggarwal, M. L. and Gupta, S. (2004), A New Method of Construction of Multi-level Supersaturated Designs, Journal of Statistical Planning and Inference, Vol. 121, 127-134.

Booth, K. H. V. and Cox, D. R. (1962), Some Systematic Supersaturated Designs, Technometrics, Vol. 4, 489-495.

Cela, R. Martinez, E. and Carro, A. M. (2001), Supersaturated Experimental Designs: New Approaches to Building and Using it Part II. Solving Supersaturated Designs by Genetic Algorithms, Chemometrics and Intelligent Laboratory Systems, Vol. 57, 75-92.

Fang, K., Ge, G., Liu, M. and Qin, H. (2004), Combinatorial Constructions for Optimal Supersaturated Designs", Discrete Mathematics, Vol. 279, 191-202.

Fang, K., Lin, D. K. J. and Liu, M. (2003), Optimal Mixed-level Supersaturated Design, Metrika, Vol. 58, 279-291.

Fang, K., Lin, D. K. J., Winker, P., and Zhang, Y. (2000), Uniform Design: Theory and Application, Technometrics, Vol. 42, 237-248.

Georgiou, S. Koukouvinos C. and Mantas, P. (2003), Construction Methods for Three-level Supersaturated Designs Based on Weighing Matrices, Statistics & Probability Letters, Vol. 63, 339-352.

Hutter, M. and Zaffalon, M. (2005), Distribution of Mutual Information from Complete and Incomplete Data, Computational Statistics and Data Analysis, Vol. 48, 633-657.

Jang, D. H.(2002), Measures for Evaluating Non-orthogonality of Experimental Designs, Communications in Statistics-Theory and Methods, Vol. 31, 249-260.

Koukouvinos, C. and Stylianou, S. (2004), Optimal Multi-level Supersaturated Designs Constructed from Linear and Quadratic Functions, Statistics & Probability Letters, Vol. 69, 199-211.

Li, P., Liu, M. and Zhang, R. (2004), Some Theory and The Construction of Mixed-level Supersaturated Designs, Statistics & Probability Letters, Vol. 69, 105-116.

Li, R. and Lin, D. K. J. (2002), Data Analysis in Supersaturated Designs, Statistics & Probability Letters, Vol. 59, 135-144.

Lu, X., Hu, W. and Zheng, Y. (2003), A Systematical Procedure in The Construction of Multi-level Supersaturated Design, Journal of Statistical Planning and Inference, Vol. 115, 287-310.

Ma, C., Fang, K., and Liski, E. (2000), A New Approach in Constructing Orthogonal and Nearly Orthogonal Arrays, Metrika, Vol. 50, 255-268.

Plackett, R. L. and Burman, J. P. (1946), The Design of Optimum Multifactorial Experiments, Biometrika, Vol. 33, 305-325.

Wang, J. C. and Wu, C. F. J. (1992), Nearly Orthogonal Arrays with Mixed Levels and Small Runs, Technometrics, Vol. 34, 409-422.

Yamada, S. and Lin, D. K. J. (2002), Construction of Mixed-level Supersaturated Design, Metrika, Vol. 56, 205-214.

Yamada, S. and Matsui T. (2002), Optimality of Mixed-level Supersaturated Design, Journal of Statistical Planning and Inference, Vol. 104, 459-468.

Yamada, S. Matsui M., Matsui T., Lin, D. K. J. and Takahashi, T. (2005), A General Construction Method for Mixed-level Supersaturated Design, Computational Statistics and Data Analysis.