불균형 이원분류자료 분석과 가설검정+

장 석 환*

〈요 약〉

불균형자료의 분석과 가설검정을 위하여 표본수 n_{ij}가 $n_{ij}>0$인 경우와 $n_{ij}\geqq 0$인 경우의 인위적인 이원분류자료를 써서 여러가지 모형하에서 주효과와 교호작용에 대한 변동을 계산하는 방법을 Searle의 R(|)표기법을 이용하여 검토하였고 이들 변동과 일반적으로 널리 쓰이는 가설과의 관계를 실증적으로 재검토하였다.

1. 서 론

불균형자료의 분산분석에 대하여는 Herr(1986)가 최근 30년간의 연구사를 구체적으로 고찰하였으나 Yates(1934)의 연구로 거슬러 올라가지 않을 수 없다. 그 후 Patterson(1946), Finney(1948), Stevens(1948), Henderson(1953), Kramer(1955), Gossless와 Lucas(1965) 등 많은 사람들이 주로 분산분석에 관련한 변동의 계산에 관하여 연구하였으며 최근에 Searle (1971, 1972, 1977), Searle 등 (1979), Kutner(1974), Hocking과 Speed(1975), Speed와 Hocking(1976) 등도 분산분석에서의 변동을 검정하고져하는 귀무가설과 연관하여 설명하였고 특히 Searle은 R(|)표기법에 의한 변동계산의 알고리즘을 이용하여 가설을 설명하고 있다. 또한 Burdick(1974, 1980)은 각 변동을 기하학적으로 설명하였고 여러가지 통계 Package별로 검토하였으며 白(1987)은 4가지 형태의 불균형자료를 SAS Package를 이용하여 분석하여 보인바 있다.

본 연구에서는 간단한 경우로써 二元分類자료에서 조사자료수를 n_{ij}라 할 때 (1) $n_{ij}\geqq 0$인 경우와 (2) $n_{ij}>0$ 일 때 분산분석에서 각 변동이 갖는 의미와 이들 변동과 관련하여 실제로 검정하는 가설이 갖는 성질을 파악하여 불균형 자료분석의 이해를 돕고져 한다.

2. 선형모형

인자 A의 수준을 l, 인자 B의 수준을 m으로 하는 二元分類表(two-way classification table)에서 각 조합에 대한 자료수를 n_{ij}라 하고 전체자료수를 $N=\sum\sum n_{ij}$라 하면 반응변수 y는

+ 본 연구는 계명대학교 1989년도 비사교수연구비에 의하여 수행되었음.
* 계명대학교 통계학과, 대구직할시 달서구 신당동 1000번지

다음과 같이 완전계수 모형(full rank model)으로 나타낼 수 있다.

$$y_{ijk} = \mu_{ij} + e_{ijk}, \tag{1}$$

여기서 $i=1, 2, \cdots, l$, $j=1, 2, \cdots, m$, $k=1, 2, \cdots, n_{ij}$이며 μ_{ij}는 (i, j)칸의 모평균, e_{ijk}는 통상적인 분산분석에서와 같이 $iid\ N(0, \sigma^2)$의 분포를 하는 확률변수라고 가정한다. (1)에서 因子의 효과가 相加的(additive)이면 인자들 간의 교호작용 γ_{ij}는

$$\gamma_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu = 0 \tag{2}$$

이라는 조건하에 (1)은

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \tag{3}$$

로 쓸 수 있다. (3)은 전형적이 二元分類에 대한 모형식으로 μ는 전체평균, $\alpha_i = \mu_{i\cdot} - \mu$, $\beta_j = \mu_{\cdot j} - \mu$로 각각 인자 A와 인자 B의 주효과를 나타낸다. 만약 (2)의 γ_{ij}가 $\gamma_{ij} \neq 0$이면 (1)은 교호작용이 있는 이원분류모형식

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \tag{4}$$

로 될 것이다. (3)과 (4)는 (1)에서 과대조정(over-parameterization)되어 만들어진 모형으로 유일한 해(unique solution)를 기대할 수 없을 뿐만 아니라 자료의 불균형에 의한 직교성의 결여로 인자의 주효과와 교호작용에 대한 변동의 계산과 가설이 복잡 다양해진다.

3. 분산분석과 가설

Speed(1978) 등은 Searle의 R(|)표기법이 변동의 내용을 표현하는데 편리하기는 하나 검정하려는 가설을 명확하게 나타내지는 못하는 결점을 지적하였으나 변동을 표현하는 편리한 방법이므로 본 연구에서도 Searle의 R(|)표기법을 쓰기로 한다. 分散分析에서 검정되는 가설은 Searle(1971, 1977, 1979), Kutner(1974), Hocking과 Speed(1975), Hocking과 Speed (1976), Speed-Hocking-Hackney(1976) 등 많은 사람들에 의하여 여러가지 변동과 관련하여 설명하고 있으나 Speed 등 (1976, 1978)의 가설이 보편적으로 통용되므로 이들 가설을 인용하기로 한다(표 1-1, 1-2).

〈표 1-1〉 μ_{ij}모형에 대한 가설

인 자 A	인 자 B
$H_1 : \mu_{i\cdot} = \mu_{i'\cdot}$	$H_5 : \mu_{\cdot j} = \mu_{\cdot j'}$
$H_2 : \sum_j n_{ij}\mu_{ij}/n_{i\cdot} = \sum_j n_{i'j}\mu_{i'j}/n_{i'\cdot}$	$H_6 : \sum_i n_{ij}\mu_{ij}/n_{\cdot j} = \sum_i n_{ij}\cdot\mu_{ij'}/n_{\cdot j}$
$H_3 : \sum_j (N_{ij})\mu_{ij} = \sum_i \sum_j n_{ij}n_{i'j}\mu_{ij}/n_{\cdot j}$	$H_7 : \sum_i (n_{ij})\mu_{ij} = \sum_j \sum_i n_{ij}n_{ij'}/n_{i\cdot}$
$H_4 : \mu_{ij} = \mu_{i'j}$	$H_8 : \mu_{ij} = \mu_{ij'}$

〈표 1-2〉 완전 모형에 대한 가설

$H_1^*: \alpha_i + \sum_j \gamma_{ij} = \alpha_{i'} + \sum_j \gamma_{i'j}$

$H_2^*: \beta_j + \sum_i \gamma_{ij} = \beta_{j'} + \sum_i \gamma_{ij'}$

$H_3^*: \alpha_i + \sum_j n_{ij}\beta_j/n_{i.} + \sum_j n_{i'j}\gamma_{i'j}/n_{i'.}$

$= \alpha_{i'} + \sum_j n_{i'j}\beta_j/n_{i'.} + \sum_j n_{i'j}\gamma_{i'j}/n_{i'.}$

$H_4^*: \beta_{j.} + \sum_i n_{ij}\alpha_i/n_{.j} + \sum_i n_{ij}\gamma_{ij}/n_{.j}$

$= \beta_{j'.} + \sum_i n_{ij'}\alpha_i/n_{.j'} + \sum_i n_{ij'}\gamma_{ij'}/n_{.j'}$

$H_5^*: (n_{i.} - \sum_j n_{ij}^2/n_{.j})\alpha_i + \sum_j (n_{ij} - n_{ij}^2/n_{.j})\gamma_{ij}$

$= \sum_{i=i'} (\sum_j n_{ij}n_{ij}/n_{.j})\alpha_{i'} + \sum_{i=i'} \sum_j (n_{ij}n_{i'j}/n_{.j})\gamma_{i'j}$

$H_6^*: (n_{.j} - \sum_i n_{ij}^2/n_{i.})\beta_j + \sum_i (n_{ij} - n_{ij}^2/n_{i.})\gamma_{ij}$

$= \sum_{j=j'} (\sum_i n_{ij}n_{ij'}/n_{i.})\beta_{j'} = \sum_{j=j'} \sum_i (n_{ij}n_{ij'}/n_{i.})\gamma_{ij'}$

$H_7^*: \gamma_{ij} - \gamma_{i'j} - \gamma_{ij'} + \gamma_{i'j'} = 0$

$H_8^*: \alpha_1 + \gamma_{i1} = \alpha_i + \gamma_{i1}$

$H_9^*: \beta_1 + \gamma_{i1} + \beta_j + \gamma_{ij}$

표 1-1, 1-2의 가설을 변동과 관련시켜 실증적으로 검토하기 위하여 $n_{ij}>0$인 경우와 $n_{ij}\geqq 0$인 경우를 다음 표 2-1과 표 2-2와 같은 가상적인 자료를 이용하기로 한다.

〈표 2-1〉 $n_{ij}>0$인 자료

	B_1	B_2	B_3	합계($n_{i.}$)
A_1	4	9	8	
	6	6	10	
		6		
	10(2)	21(3)	18(2)	49(7)
A_2	5	8	10	
	7			
	12(2)	8(1)	10(1)	30(4)
A_3	8	10	4	
	7		7	
	11			
	9			
	35(4)	10(1)	11(2)	56(6)
합계 ($n_{.j}$)	57(8)	39(5)	39(5)	$y_{...}=35$ $N=18$

〈표 2-2〉 $n_{ij}\geqq 0$인 자료

	B_1	B_2	B_3	B_4	합계($n_{i.}$)
A_1	7	14		7	
	6			10	
	11				
	24(3)	14(1)		17(2)	55(6)
A_2	14		18		
	12		22		
	26(2)		40(2)		66(4)
A_3		14	16	11	
			18	15	
				13	
		14(1)	34(2)	39(3)	87(6)
합계	50(5)	28(2)	74(4)	56(5)	$y_{...}=208$
($n_{.j}$)					$N=16$

3. 1 교호작용이 없는 경우

표 2-1은 $n_{ij} \geq 1$이고 $n_{ij} \neq n_{i.} n_{.j}/N$ 이며 표 2-2는 $n_{ij} \geq 0$인 경우로서 연결된 자료이므로 변동의 계산이 가능하다. 모형식 (1)을 행렬로 나타내면 일반적으로

$$Y = X\mu + e \tag{5}$$

와 같고 Y는 $N \times 1$열 벡터, X는 $N \times s$계획 행렬 (s는 자료가 있는 칸의 수), $\mu' = (\mu_{ij},\ i = 1, 2, \cdots, 1,\ j = 1, 2, \cdots m)$는 $s \times 1$의 모수벡터이며 e는 $N \times 1$오차벡터로 $N(0, I\sigma^2)$의 분포를 한다고 가정한다.

(5)에서 μ의 불편추정량 $\hat{\mu}$는

$$\hat{\mu} = (X'X)^{-1} X'Y \tag{6}$$

로 추정되며 표 2-1의 자료에서

$$\begin{aligned}(X'X)^{-1} &= \text{diag}\,[1/2\ \ 1/3\ \ 1/2\ \ 1/2\ \ 1\ \ 1\ \ 1/4\ \ 1\ \ 1/2] \\ &= D_1^{-1}\end{aligned} \tag{7}$$

이고 따라서

$$\begin{aligned}\hat{\mu}_1' &= [\mu_{11}\ \ \mu_{122}\ \ \mu_{13}\ \cdots\ \mu_{31}\ \ \mu_{32}\ \ \mu_{33}] \\ &= \left[5\ \ 7\ \ 9\ \ 6\ \ 8\ \ 10\ \ 8\frac{3}{4}\ \ 10\frac{1}{2}\right]\end{aligned} \tag{8}$$

이다. (8)에 의한 변동은

$$\begin{aligned}R(\mu_1 \mid \mu) &= R(\mu_1) - R(\mu) \\ &= 49.25\end{aligned} \tag{9}$$

이고 총변동과 오차변동은 다음과 같다.

$$\begin{aligned}\text{SSTO} &= \Sigma\Sigma\Sigma\, y_{ijk}^2 - y^2_{\ldots}/N \\ &= 74.5 \\ \text{SSE} &= 25.25\end{aligned}$$

또 표 2-2의 자료에 대한 $(X'X)^{-1}$와 μ_2는

$$\begin{aligned}(X'X)^{-1} &= \text{diag}\,[1/3\ \ 1\ \ 1/2\ \ 1/2\ \ 1/2\ \ 1\ \ 1/2\ \ 1/3] \\ &= D_2^{-1}\end{aligned} \tag{10}$$

$$\hat{\mu}_2' z = \left[8\ \ 14\ \ 8\frac{1}{2}\ \ 13\ \ 20\ \ 14\ \ 17\ \ 13\right] \tag{11}$$

이며 따라서 (11)에 의하여 설명되는 변동은

$$R(\mu_2 \mid \mu) = 247.5 \tag{12}$$

이고 총변동과 오차변동은

$$\begin{aligned} SSTO &= \sum\sum\sum y_{ijk}^2 \quad y^2 \ldots / N \\ &= 286 \\ SSE &= 38.5 \end{aligned}$$

로 추정된다.

(9)와 (12)에서 $R(\mu \mid \mu)$는 처리에 의한 변동과 같으며 처리의 자유도는 $s-1$이고 분산분석표에서 "Ho : μ_{ij}는 모두 같다"는 일반적인 귀무가설은 μ_{ij}로 구성되는 $s-1$개의 독립적인 대비(contrasts)를 동시에 검정하는 것이므로 K'를 이들 대비로 만들어진 행렬이라고 하면 K'는 완전 행계수 행렬(full row rank matrix)이다. 모수를 적합시킴으로써 설명되는 변동을 가설에 관련시켜 생각하면 (Searle 1971)

$$Q_{HO} = (Kb)' (K'GK)^{-1}(K'b) \tag{13}$$

로 나타낼 수 있으며 표 2-1과 표 2-2에 대한 Q_{HO}는 각각 (9)과 (12)의 값과 같다. (3)을 참모형이라고 가정하면 정규방정식은

$$X'X\theta^0 = X'Y \tag{14}$$

이고 X는 $N\times(l+m+1)$계획 행렬, $\theta^{0\prime} = [\mu\ \alpha_1 \cdots \alpha_l\ \beta_1 \cdots \beta_m]$이다. $r(X) = p$라고 하고 $p<(l+m+1)$이므로 $G = (X'X)^-$를 $(X'X)(X'X)(X'X) = X'X$를 만족시키는 $X'X$의 한 일반화 역행렬(generalized inverse matrix ; g.i.m.)이라고 하면 (14)의 해는

$$\theta^0 = G\,X'Y \tag{15}$$

이고 인자 A와 인자 B에 의하여 설명되는 변동은

$$\begin{aligned} R(\alpha, \beta \mid \mu) &= \theta^{0\prime}x'y \quad - y^2 \ldots / N \\ &= R(\mu, \alpha, \beta) - R(\mu) \end{aligned} \tag{16}$$

이므로 표 2-1의 자료에 대한 g.i.m.을 $Go = (X'X)^-$라고 하면

$$\theta^0 = G_0X'Y \tag{17}$$

이고 θ^0에 의하여 설명되는 변동은 (16)에 의하여

$$\begin{aligned} R(\alpha, \beta \mid \mu) &= 1020.02998 - 1012.5 \\ &= 7.52998 \end{aligned} \tag{18}$$

이다.

표 2-1의 자료에서 $H_0 : \mu_{11} = \mu_{21} = \mu_{31}$, $H_0 : \mu_{12} = \mu_{22} = \mu_{32}$에서 (3)을 적용하여 구해지는 K'의 하나는 다음과 같이 생각할 수 있다.

$$K' = \begin{bmatrix} 0 & 1 & -1 & 0 & -3/14 & 5/28 & 1/28 \\ 0 & 1 & 0 & -1 & -2/7 & 2/7 & 0 \\ 0 & -7/20 & 1/20 & 3/20 & 1 & -1 & 0 \\ 0 & -3/20 & 1/20 & 1/10 & 1 & 0 & -1 \end{bmatrix}$$

따라서 (13)에 의하여

$$Q_{HO} = (K'\theta^{\circ})' (K' G_{\circ} K)^{-1} (K'\theta^{\circ})$$
$$= 7.52998$$

로 (18)과 같은 결과를 보인다.

(3)에서 $\beta_j = 0$, $j=1, 2, \cdots, m$, 이면

$$y_{ijk} = \mu + \alpha_i + e_{ijk} \tag{19}$$

로 축소되며 이는 一元分類에 대한 모형식과 같다. 표 2-1에서 (19)로 설명되는 변동은 같은 방법으로

$$R(\alpha \mid \mu) = \theta_1^{\circ\prime} X' Y - y^2_{\cdots}/N$$
$$= 3.5 \tag{20}$$

이고 여기서 $\theta_1^{\circ\prime} = [\mu, \alpha_1, \alpha_2, \alpha_3]$는 (19)의 정규방정식에 대한 한 해이며 이 때 $g.i.m.$은 $G_1 = (X'X)^-$이다.

$H_{\circ}: \mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot}$에서 K'를 유도하면

$$K' = \begin{bmatrix} 0 & 1 & -1 & 0 & -3/4 & 5/28 & 1/28 \\ 0 & 1 & 0 & -1 & -2/7 & 2/7 & 0 \end{bmatrix}$$

이고

$$Q_{HO} = (K'\theta_1^{\circ})' (K' G_1 K)^{-1} (K\theta_1^{\circ})$$
$$= 3.5$$

로 (20)과 같다. Searle(1971), Speed 등(1978)은 (20)의 결과는 因子 B를 무시한 A의 변동이라 하였고 이는 또

$$SSA(unadj) = \sum \eta_{i\cdot} (y_{i\cdot\cdot} - y_{\cdots})^2$$

과 같다.

모형 (3)에서 $\alpha_i = 0$, $i=1, 2, \cdots, l$이면

$$y_{ijk} = \mu + \beta_j = e_{ijk} \tag{21}$$

로 되며 因子 B에 의한 변동은

$$R(\beta \mid \mu) = \text{SSB(unadj)} = 2.025 \tag{22}$$

(21)에서 모수벡터 $\theta_2^{o\prime} = [\mu\ \beta_1\ \beta_2\ \beta_3]$, $G_2 = (X'X)^-$는 $X'X$의 한 $g.i.m$이다. $H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3}$에서

$$K' = \begin{bmatrix} 0 & -7/20 & 1/20 & 3/20 & 1 & -1 & 0 \\ 0 & -3/20 & 1/20 & 1/10 & 1 & 0 & -1 \end{bmatrix}$$

과 같다. 따라서 $H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3}$에 의한 변동은

$$Q_{H_0} = (K'\theta_2^{o\prime})'(K'G_2K)^{-1}(K'\theta_{2_0}') = 2.025$$

로 $SSB = \sum n_{.j}(y_{.j} - y_{..})^2$과 같은 결과를 보인다.

위의 결과 (18), (20), (22)에서

$$R(\alpha \mid \mu, \beta) = R(\alpha, \beta \mid \mu) - R(\beta \mid \mu) = 5.50498 \tag{23}$$

$$R(\beta \mid \mu, \alpha) = R(\alpha, \beta \mid \mu) - R(\alpha \mid \mu) = 4.02998 \tag{24}$$

로 계산된다. (23)은 因子 A의 수정된 변동(adj. SS)이고 (24)는 因子 B의 수정된 변동(adj. SS)이다.

표 2-2의 자료를 이용하여 같은 방법으로 (3)을 적합시켜 변동을 계산하면 (16)에 의하여

$$R(\alpha, \beta \mid \mu) = \theta^{o\prime}X'Y = 239.7814 \tag{25}$$

$\theta^{o\prime} = [\mu\ \alpha'\ ,\ \beta']$, $G_3 = (X'X)^-$를 표 2-2에 의한 $X'X$의 한 g. i. 행력이라 하고 $H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.}$, $H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$에서 K'를 구하면 다음과 같다.

$$K' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 1/6 & -1/2 & 1/3 \\ 0 & 1 & 0 & -1 & 1/2 & 0 & -1/3 & -1/6 \\ 0 & 1/10 & 2/5 & -1/2 & 1 & -1 & 0 & 0 \\ 0 & 3/5 & -1/10 & -1/2 & 1 & 0 & -1 & 0 \\ 0 & 1/5 & 2/5 & -3/5 & 1 & 0 & 0 & -1 \end{bmatrix}$$

또 (19)와 (21)에 의하여 설명되는 변동은

$$R(\alpha \mid \mu) = 150.6667 \tag{26}$$

$$R(\beta \mid \mu) = 184.2 \tag{27}$$

이므로 (25), (26), (27)에서

$$R(\alpha \mid \mu, \beta) = 55.5814 \tag{28}$$

$$R(\beta \mid \mu, \alpha) = 89.1147 \tag{29}$$

를 얻는다.

3. 2 교호작용이 있을 경우

인자들 간에 교호작용이 존재하면 모형식은 (4)와 같고 이 때의 X는 $N \times (l+m+s-1)$의 계획행렬이다. 정규방정식은

$$X'X\theta^* = X'Y \tag{30}$$

에서 $\theta^* = [\mu\ \alpha_i\ \beta_j\ \gamma_{ij}]$, $i=1, 2, \cdots, 1$, $j=1, 2, \cdots, m$이고 γ_{ij}는 $n_{ij}>0$인 칸에만 나타난다. (30)의 한 해는

$$\theta^* = G_4 X'Y \tag{31}$$

로 쓸 수 있고, 이 때 G_4는 (7)의 D_1^{-1}과 (10)의 D_2^{-1}를 이용하면 쉽게 구할 수 있다. 즉

$$G_{4i} = \begin{vmatrix} O & O \\ O & D_i^{-1} \end{vmatrix}, \quad i = 1, 2, \tag{32}$$

이며 G_{41}은 표 2-1, G_{42}는 표 2-2에 대한 $g.i.m.$을 의미한다. 따라서 표 2-1은 θ_1^*는 (8)의 결과에서

$$\theta_1^{*\prime} = [O\ \mu_1'] \tag{33}$$

이고 표 2-2의 $\theta_2^{*\prime}$는 (11)의 결과에서

$$\theta_2^{*\prime} = [O\ \mu_1'] \tag{34}$$

이고 (4)에 의하여 설명되는 변동은

$$R(\alpha, \beta, \gamma \mid \mu) = \theta_1^{*\prime} X'Y - y\cdots^2/N = 49.25$$

$$R(\alpha, \beta, \gamma \mid \mu) = \theta_2^{*\prime} X'Y - y\cdots^2/N = 247.5$$

와 같다.

식(5)에서 $\beta_j=0$, $j=1, 2, \cdots, m$이면 모형식은

$$y_{ijk} = \mu + \alpha_i + \gamma_{ij} + e_{ijk} \tag{35}$$

로 되며 정규방정식

$$X'X\theta^*_{11} = X'Y \tag{36}$$

의 한 해는

$$\theta^*_{11} = G_5X'Y \tag{37}$$

이다. G_5는 (35)의 모형식하에서 표 2-1에 대한 $(X'X)$의 $g.i.m.$이다. 따라서 (37)에 의한 변동은

$$R(\alpha, \gamma \mid \mu) = 49.25$$

로서 $R(\alpha, \beta, \gamma \mid \mu)$, (9)의 $R(\mu_1 \mid \mu)$와 같다.

(35)의 모형식하에서 표 2-1자료에 대한 가설 $H_0 : \mu_1 = \mu_2 = \mu_3$로부터 유도한 K'의 하나는

$$K' = \begin{bmatrix} 0 & 1 & -1 & 0 & 2/7 & 3/7 & 2/7 & -1/2 & -1/4 & -1/4 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 2/7 & 3/7 & 2/7 & 0 & 0 & 0 & -4/7 & -1/7 & -1/7 \end{bmatrix}$$

로 생각할 수 있으며 (37)의 θ^*_{11}와 K'를 이용한 $Q_{H0}=3.5$임을 알 수 있다.

표 2-2의 X에 대한 $X'X$의 한 $g.i.m$을 G_6라 하면 정규방정식의 해는

$$\theta^*_{12}{}' = [\mu\ \alpha'\ \gamma'] \tag{38}$$

이다. 따라서

$$R(\alpha, \gamma \mid \mu) = 247.5$$

로 (34)에 의한 $R(\alpha, \beta, \gamma \mid \mu)$와 (12)의 $R(\mu_2 \mid \mu)$와 같다.

표 2-2에 대한 가설 $H_0 : \mu_1 = \mu_2 = \mu_3$에서 K'는

$$K' = \begin{bmatrix} 0 & 1 & -1 & 0 & 1/2 & 1/6 & 1/3 & -1/2 & -1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1/2 & 1/6 & 1/3 & 0 & 0 & -1/6 & -1/3 & -1/2 \end{bmatrix}$$

이고

$$Q_{H0} = (K'\theta^*_{12})'(K'G_6K)^{-1}(K'\theta^*_{12}) = 150 \cdot \frac{2}{3}$$

로 (26)의 $R(\alpha \mid \mu)$와 같은 결과를 보인다.

式 (5)에서 $\alpha_i = 0$, $i = 1, 2, \cdots, 1$, 이면 다시

$$y_{ijk} = \mu + \beta_j + \gamma_{ij} + e_{ijk} \tag{40}$$

로 축소되며 표 2-1에 대한 $X'X$의 한 $g.i.m.$행렬을 G_7라 하면 G_7에 의한 정규방정식의 해는

$$\theta^*_{21}{}' = [\mu\ \beta'\ \gamma'] \tag{40}$$

이고 역시 $H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3}$에서 K'를 구하면

$$K' = \begin{bmatrix} 0 & 1 & -1 & 0 & 1/4 & -3/5 & 0 & 1/4 & -1/5 & 0 & 1/2 & 1/5 & 0 \\ 0 & 1 & 0 & -1 & 1/4 & 0 & -2/5 & 1/4 & 0 & -1/5 & 1/2 & 0 & -2/5 \end{bmatrix}$$

이다. G_7을 이용한 $Q_{HO}=2.025$로 (22)의 $R(\beta \mid \mu)$와 같다. 같은 방법으로 표 2-2에 대한 $(X'X)$의 한 $g.i.m.$을 G_8이라 하면 G_8에 의한 정규방정식의 해는

$$\theta^*_{22}{}' = [\mu\ \beta'\ \gamma'] \tag{41}$$

이고 $H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4}$에서 K'는

$$K' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 3/5 & -1/2 & 0 & 2/5 & 0 & -1/2 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 3/5 & 0 & 0 & 2/5 & -1/2 & 0 & -1/2 & 0 \\ 0 & 1 & 0 & 0 & -1 & 3/5 & 0 & -2/5 & 2/5 & 0 & 0 & 0 & -3/5 \end{bmatrix}$$

이므로 G_8 이용한 $Q_{HO}=184.2$로 역시 $R(\beta \mid \mu)$와 같다.

교호작용에 의하여 설명되는 변동은

$$R(\mu, \alpha, \beta, \gamma) - R(\mu, \alpha, \beta) = R(\gamma \mid \mu, \alpha, \beta) \tag{42}$$

로 표현되며 표 2-1과 표 2-2의 자료에 대한 $R(\gamma \mid \mu, \alpha, \beta)$는 다음과 같다. 즉

$$\begin{aligned} &\text{표 } 2-1 : R(\gamma \mid \mu, \alpha, \beta) = 41.72 \\ &\text{표 } 2-2 : R(\gamma \mid \mu, \alpha, \beta) = 7.7186 \end{aligned} \tag{43}$$

이는 또한 $R(\mu, \alpha, \gamma) - R(\mu, \alpha)$와 $R(\mu, \beta, \gamma) - R(\mu, \beta)$에서도 같은 결과를 보인다. 표 2-1은 $n_{ij}>0$이므로 교호작용에 대한 자유도는 $df=4$이고 γ_{ij}에 의한 4개의 독립적인 대비를 (44)와 같이

$$H_0 : \begin{cases} \gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22} = 0 \\ \gamma_{12} - \gamma_{13} - \gamma_{22} + \gamma_{23} = 0 \\ \gamma_{21} - \gamma_{22} - \gamma_{31} + \gamma_{32} = 0 \\ \gamma_{22} - \gamma_{23} - \gamma_{32} + \gamma_{33} = 0 \end{cases} \tag{44}$$

놓으면 L'는

$$L' = \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 & 1 \end{bmatrix}$$

과 같고 (7)의 D_1^{-1}과 (8)의 μ_1에 의하여

$$Q_{H0} = (L'\mu_1)'(L'D_1^{-1}L)^{-1}(L'\mu_1)$$
$$= 41.72$$

로 (43)과 같으며 (44)와 다른 H_0를 생각할 수 있으나 그 결과는 같다.

표 2-2에있어서 교호작용의 자유도는 $df = s-l-m-1=2$이므로 γ_{ij}의 독립적인 대비는

$$H_0: \begin{matrix} \gamma_{12}-\gamma_{14}-\gamma_{32}+\gamma_{34}=0 \\ \gamma_{11}-\gamma_{14}-\gamma_{21}-\gamma_{23}+\gamma_{33}+\gamma_{34}=0 \end{matrix} \quad (45)$$

로 놓을 수 있고, 따라서 L'는

$$L' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & -1 & 1 & 0 & -1 & 1 \end{bmatrix}$$

과 같다. 그러므로 (10)의 D_2^{-1}와 (11)의 μ_2에 의하여

$$Q_{H0} = (L'\mu_2)'(L'D_2^{-1}L)^{-1}(L'\mu_2)$$
$$= 7.7186$$

로 역시 (43)과 같다.

이상의 결과를 종합하여 분산분석표를 만들어 보면 표 3과 같다.

표 3 분산분석표

S.V.	$n_{ij}>0$			$n_{ij}\geqq 0$		
	d.f.	SS	MS	d.f.	SS	MS
$R(\alpha \mid \mu, \beta)$	2	5.505	2.752	2	55.581	27.791
$R(\beta \mid \mu, \alpha)$	2	4.030	2.015	3	89.115	29.705
$R(\gamma \mid \mu, \alpha, \beta)$	4	41.720	10.430	2	7.719	33.860
Error	9	25.250	2.806	8	38.500	4.813
Total	17	74.50	–	15	286.00	–

4. 고찰 및 결론

二元分類자료에 대한 가설은 일반적으로 가정된 모형에 따라 결정된다. (1)을 참모형이라 하면 표 1-1에서와 같이 대부분의 가설은 μ_{ij}, $i=1, 2, \cdots, l$, $j=1, 2, \cdots m$, 의 평균 또는 가중평균 개념으로 표현되어 평균간의 비교를 할 수 있으나 모형식 (4)에 근거하면 α_i, β_j, γ_{ij} 등의 성분으로 표현되기 때문에 불균형 자료에서 $H_0: \mu_i = \mu_{i'}$, $i \neq i'$의 가설은 직교성의 결여로 β_j와 γ_{ij} 成分이 포함되기 때문에 순수하게 $\alpha_i = \alpha_{i'}$를 검정할 수 없다. (Searle 1971)

표 2-1과 표 2-2에 대한 참모형식을 (1)로 가정하면 (5)는 완전계수모형이고 따라서 μ의 불편추정량 $\mu' = [y_{ij}, i=1, 2, \cdots, l, j=1, 2, \cdots, m]$은 자료가 있는 칸의 평균의 벡터이다. 이 모형하에서는 모든 μ_{ij}가 추정가능하고 μ_{ij}의 선형조합 역시 추정가능하므로 μ_{ij}의 선형조

합에 의한 가설을 쉽게 검정할 수 있다. 표 1-1의 H_1과 H_5는 Speed 등(1978)이 지적했듯이 수준평균을 비교하는 가설이므로 이해하기 쉬우나 H_2와 H_6, H_3과 H_7은 표본수(n_{ij})의 가중치의 영향을 받는 가설이므로 특히 H_3과 H_7은 가설이 갖는 의미와 변동계산이 복잡하다.

모형식 (3)에 대한 상수 적합에서 (16)에 의한 $R(\alpha, \beta \mid \mu)$는 $H_0 : \mu_{i\cdot} = \mu_{i'\cdot}$, $i \neq i'$ 과 $H_0 : \mu_{\cdot j} = \mu_{\cdot j'}$, $j \neq j'$ 를 동시에 검정하는 변동으로 Kutner(1974)와 Hocking&Speed(1975)는 (2)의 조건하에서 n_{ij}/N이 모집단의 비율을 잘 나타낼 때 H_2와 H_6이 주효과를 검정하는 적합한 가설이고 그렇지 않으면 H_1과 H_5가 적합하다고 하였다. 실제로 이들을 (3)을 토대로 하면

$$H_0 : \alpha_i + (1/n_{i\cdot}) \sum_j \beta_j = \alpha_{i'} + (1/n_{i'\cdot}) \sum_j \beta_j \tag{46}$$

$$H_0 : \beta_j + (1/n_{\cdot j}) \sum_i \alpha_i = \beta_{j'} + (1/n_{\cdot j'}) \sum_i \alpha_i \tag{47}$$

과 같다. 따라서 (19)에서와 같이 $\beta_j = 0$, $j = 1, 2, \cdots, m$이면 (46)은 $H_0 : \alpha_i = \alpha_{i'}$로 되며 H_2는 (20), (26)의 $R(\alpha \mid \mu)$에 의해 검정되고 $\alpha_i = 0$, $i = 1, 2, \cdots, l$이면 (47)은 $H_0 : \beta_j = \beta_{j'}$로 되며 H_6은 (22), (27)의 $R(\beta \mid \mu)$에 의하여 검정된다. (23)과 (28)의 $R(\alpha \mid \mu, \beta)$는 실제로 β에 의해 수정된 SSA 즉 SSA(adj.)이며 Searle(1971)은 회귀모형식을 이용하여 β를 먼저 모형식에 적합시킨 후에 α에 의해 추가적으로 설명되는 변동이라고 해석하며 H_3과 H_5^*를 검정하는 변동이다. (24)와 (29)에의한 $R(\beta \mid \mu, \alpha)$는 SSB(adj.)이며 α를 적합시키고 β를 추가적으로 적합시킬 때 β에 의해 설명되는 변동으로 H_7과 H_6^*를 검정하는 변동이다.

교호작용이 존재하면 즉 $\gamma_{ij} \neq 0$일 때 모형(3)에 의하여 설명되는 변동은 $R(\alpha, \beta, \gamma \mid \mu)$이고 $R(\alpha, \beta, \gamma \mid \mu) = R(\alpha, \gamma \mid \mu) = R(\beta, \gamma \mid \mu) = \sum\sum (y_{ij} - y_{\cdot\cdot})^2$이며 이에 대한 이론적 배경은 Searle (1971)이 잘 설명하고 있다. H_1^*는 모형식 (35)하에서 $H_0 : \mu_{i\cdot} = \mu_{i'\cdot}$, $i \neq i'$와 같으며 $Q_{H0} = R(\alpha \mid \mu)$이므로 결국 H_1^*를 검정하고 H_2^*는 모형식 (39)하에서 $H_0 : \mu_{\cdot j} = \mu_{\cdot j'}$, $j \neq j'$과 같고 이에 대한 변동은 $Q_{H0} = R(\beta \mid \mu)$이다. H_3^*와 H_4^*는 모형식 (4)에 의하여 유도될 수 있고 같은 방법으로 H_3^*는 $Q_{H0} = R(\alpha \mid \mu)$에 의하여, 또 H_4^*는 $R(\beta \mid \mu)$에 의하여 검정된다는 것을 알 수 있다. H_5^*와 H_6^*는 $R(\alpha \mid \mu, \beta)$에 의해서 검정되고 H_6^*는 $R(\beta \mid \mu, \alpha)$에 의하여 검정된다. H_7^*는 교호작용에 대한 가설로 (42)에 의한 $R(\gamma \mid \mu, \alpha, \beta)$에 의하여 검정된다. $R(\gamma \mid \mu, \alpha, \beta)$는 교호작용에 대한 자유도 d.f.$= s - l - m + 1$에 해당하는 γ_{ij}의독립적인 대비에 대한 변동이므로 $n_{ij} > 0$인 자료에서 의미있고 이해하기 쉬운 대비를 검정할 수 있겠으나 $n_{ij} \geq 0$인 자료에서는 Searle(1971)이 말한바와 같이 γ_{ij}의 대비는 추정가능한 함수에 의하여 유도될 수 있다. 예를 들면 표 2-2에서 μ_{ij}는 추정가능한 함수이므로 μ_{ij}에 의한 대비를 다음과 같이 만들 수 있다.

$$\begin{aligned}
&(1)\ \mu_{12} - \mu_{32} - \mu_{14} + \mu_{34} = 0 \\
&(2)\ \mu_{11} - \mu_{21} - \mu_{14} + \mu_{24} = 0 \\
&(3)\ \mu_{23} - \mu_{33} - \mu_{24} + \mu_{34} = 0
\end{aligned} \tag{48}$$

대비 (2)와 (3)을 조합하여 새로운 대비를 형성하고 μ_{ij}를 γ_{ij}로 나타내면

$$\begin{aligned}
&(1)\ \gamma_{12} - \gamma_{14} - \gamma_{32} + \gamma_{34} = 0 \\
&(2)\ \gamma_{11} - \gamma_{14} - \gamma_{21} + \gamma_{23} - \gamma_{33} + \gamma_{34} = 0
\end{aligned} \tag{49}$$

와 같다. (49)에서 두번째 대비는 별의미가 없는것 같으나 (48)에서 (2)번과 (3)번 대비는 분명히 의미 있는 대비들이다. H_8*와 H_9*는 특수한 경우로서 $n_{ij}>0$인 자료에서 가능하며 특정한 수준을 비교하는 가설이다.

불균형 자료를 여러가지 모형하에서 모수를 적합시킴으로써 설명되는 변동과 가설의 성질은 참모형의 구조에 달려 있다고 생각된다. 가정된 모형하에서 변동을 계산할 때 실험계획 모형에 있어서 실험에 포함된 인자의 중요도에 따라서 효과를 평가하거나 사전에 검정하고자 하는 순서가 정해 졌다면 회귀분석에서와 같이 $R(\alpha|\mu)$, $R(\beta|\mu,\alpha)$, $R(\gamma|\mu,\alpha,\beta)$ 순서대로 모형식에 적합시켜 각 인자가 추가적으로 설명하는 변동을 검토할 필요가 있지만 (Overall과 Klett 1972) 실제로 가설검정은 가설이 보여주듯이 모수적합 순서에 관계없이 검정되기 때문에 별의미가 없다고 본다.(Speed등 1978) 자료가 불균형일 때 K'에 따라 동일한 변동이 서로 다른 형태의 가설을 검정하게 되므로 혼동하기 쉽다. 따라서 가설에 적합한 변동을 구하여 검정해야 할 것이며 실제로 검정되는 가설의 의미를 잘 파악해야 할 것이다.

〈참 고 문 헌〉

(1) Burdick, D. S., Herr, D. G., O'Fallon, W. M., and O'Nell B. V. (1974). "Exact Methods in the Unbalanced, Two-way Analysis of Variance-A Geometric View", *Communications in Statistics*, Part A-Theory and Methods, 3, 581 - 595.

(2) Gosslee, D. G. and Lucas, H. L. (1965). "Analysis of Variances of Disproportionate Data When Interaction is Present", *Biometrics*, 21, 115 - 133.

(3) Henderson, C. R. (1953). "Estimation of Variance and Covariance Components", *Biometrics*, 9, 226 - 252

(4) Herr, D. G. (1986). "On the History of ANOVA in Unbalanced, Factorial Designs : The First 30 Years", *The American Statistician*, 40, 265 - 270.

(5) Hocking. R. R. and Speed, F, M, (1975). "A Full Rank Analysis of Some Linear Model Problems", *Journal of the American Statistical Association*, 70, 706 - 712.

(6) Kramer, C. Y (1955). "On the Analysis of Variance of a Two-way Classification With Unequal Subclass Numbers", *Biometrics*, 11, 441 - 452.

(7) Kutner, M. H. (1974). "Hypothesis Testing in Linear Models (Eisenhart Model I)". *The American Statistician*, 28, 98 100.

(8) 白雲鵬. (1978). "反復數가 같지 않은 二元表의 分析", 應用統計, 高麗大學敎 統計硏究所, 2, 7 - 28.

(9) Patterson, R. E. (1946). "The Use of Adjusting Factors in the Analysis of Data With Disproportionate Subclass Numbers", *Journal of the American Statistical Association*, 41, 334 - 346.

(10) Searle, S. R. (1971A). *Linear Models*, John Wiley.

(11) ——— (1977). *Illustrative Calculations of Sums of Squares in the 2-way Crossed Classification, Unbalanced Data, All Cell filled*, Paper No. BU-608-M in the Biometrics Unit Mimeo Series, Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853.

(12) ———, Speed, F. M., and Henderson, H. V. (1979). *Some Computational and Model Equivalences in Aanalyses of Variannce of Unequal-Subclass-Numbers Data*, Paper No. BU-668-M in the Biometrics Unit Mimeo Series, Cornell University.

(13) Speed, F. M., and Hocking, R. R. (1976). "The Use of R ( )-Notation with Unbalanced Data", *The American Statistician*, 30, 30 - 33.

(14) ———, Hocking R. R., and Hackney, O. P. (1978). "Methods of Analysis of Linear Models with Unbalanced Data", *Journal of the American Statistical Association*, 73, 105 - 112.

(15) Stevens, W. L. (1948). "Statistical Analysis of a Nonorthogonal Tri-factorial Experiment", *Biometrika*, 35, 346 - 367.

(16) Yates, F. (1934). "The Analysis of Multiple Classifications With Unequal Numbers in the Different Classes". *Journal of the American Statistical Association*, 29,, 51 - 66.

Analysis of Variance and Hypothesis Testing With Unbalanced Data

Suk H. Chang*

〈Abstract〉

For the present study two sets of artificially unbalanced data of being $n_{ij}>0$ and $n_{ij}\geqq 0$ were used. The Hypotheses that are commonly used in ANOVA were examined by computing the sums of squares associated with the hypotheses under various postulated models, using Searle's R(|)-notation.

* Dept. of Statistics, Keimyung University, Taegu, Dalsogu, Sindang-dong, 1000