Youngnam Statistical Letters
June 1993, Vol. 4, 121-130

불균형 일원 변량모형에서 추정방법에 따른 분산성분의 추정량이 음이 될 확률의 계산[1)]

송 규 문[2)]

요 약

불균형 일원 변량모형에서 AOV추정량과 사전값이 0, 1, ∞인 MINQUE에 국한하여 정규분포를 가정할 때 분산성분의 추정량이 음이 될 이론적 확률을 구하고, 비정규분포에 대해서는 모의실험을 통해 추정량이 음이 될 확률을 구하였다. 이 때 정규분포에서의 이론적 확률과 모의실험에 의해 계산된 확률간에 유의한 차이가 없고, 표본수, 수준수 그리고 ρ가 커지면 각 추정량은 음이 될 확률이 작아지며, 고려된 추정량 중에서 AOV추정량이 대부분의 경우에 음이 될 확률이 가장 작게 나타났다.

1. 서 론

인자의 각 수준에서 추출된 표본수가 $n_1, n_2, \cdots, n_t$인 불균형 일원 변량모형(unbalanced one-way random effects model)은 다음과 같이 나타낼 수 있다.

$$y_{ij}=\mu+\alpha_i+\varepsilon_{ij} \qquad (1.1)$$
$$i=1, \cdots, t;\ j=1, \cdots, n_i;\ n=\sum n_i$$

여기서 μ는 고정된 모수이고, $\alpha_i, \varepsilon_{ij}$는 독립인 확률변수로 평균이 각각 0이고, 분산은 $\mathrm{Var}(\alpha_i)=\sigma_\alpha^2\geq 0$, $\mathrm{Var}(\varepsilon_{ij})=\sigma_e^2\geq 0$ 이다.

식 (1.1)를 행렬기호로 나타내면 다음과 같다.

$$y=1\mu+U\alpha+\varepsilon \qquad (1.2)$$

---

1)본 연구는 계명대학교 1992년도 갑종연구비에 의하여 지원되었음.

2)계명대학교 통계학과 교수

여기서 y는 n×1 벡터로 평균 1μ, 분산 $\sigma_a^2UU' + \sigma_e^2I$를 가지며, 1은 모든 원소가 1인 n×1 벡터이다. 또한 U는 대각선으로 1_{n_i}(모든 원소가 1인 n_i×1인 벡터)이고, 비 대각선으로 **0** 인 행렬이다.

주어진 모형에 대하여 분산성분을 추정하는 방법에는 여러가지가 있으며, 이와같은 추정방법에 의하여 분산성분을 추정할 때 음추정값을 경험하는 것은 흔한 일이다. 따라서 분산성분의 음추정값에 관한 연구는 폭넓게 연구되어 왔으며, Searle(1971)은 균형 일원 변량모형에서 AOV추정량의 분포를 구하여 $\hat{\sigma}_a^2$이 음이 될 확률을 나타냈고, Verdooren(1982)는 균형 일원 변량모형에서 수준수, 각 수준의 표본수, 그리고 분산성분의 비율이 주어질 때 $\hat{\sigma}_a^2$이 음이 될 확률을 구하였다. 본 연구에서는 분산성분을 추정하는 가장 보편적인 방법인 분산분석(analysis of variance, AOV)방법과 최소 노음 이차 불편추정(minimum norm quadratic unbiased estimation or estimator, MINQUE)방법에 대하여 분포, 분산성분의 비율 $\rho(=\sigma_a^2/\sigma_e^2)$ 및 디자인 형태에 따른 분산성분의 추정량이 음이 될 확률을 구하고, 그 결과를 분석하고자 한다.

2. 정규분포에서 분산성분의 추정량의 분포와 음이 될 확률

2.1 AOV추정량의 분포와 음이 될 확률

모형 (1.1)에서 분산성분의 AOV추정량은 다음으로 부터 구할 수 있다.

$$\begin{bmatrix} n-\sum n_i^2 & t-1 \\ 0 & n-t \end{bmatrix} \begin{bmatrix} \hat{\sigma}_a^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} y'Ay \\ y'By \end{bmatrix} \tag{2.1}$$

여기서 $A=U(U'U)^{-1}U-(1/n)J$, $B=I-U(U'U)^{-1}U'$로 U는 식 (1.2)에 주어진 것과 같고, J는 모든 원소가 1인 n×n인 행렬이다.

모형 (1.2)에서 y가 정규분포를 따른다고 가정하면 다음이 성립한다.

$$y \sim N(1\mu, V) \tag{2.2}$$

여기서 $V= \sigma_a^2UU' + \sigma_e^2I= \sigma_e^2(I+\rho UU')$, $\rho= \sigma_a^2/\sigma_e^2$ 이다. 따라서

$$z=(1/\sigma_e)T(y-1\mu) \sim N(0, I_n) \tag{2.3}$$

이 된다. 여기서 V는 비정칙 값 분해(singular value decomposition, SVD)에 의해

$V=P\Lambda P'$ 이 되고, T는 $TVT'=I$가 되는 $T=P\Lambda^{-\frac{1}{2}}P'$ 이다. 이때 $y=\sigma_e T^{-1}z+1\mu$이고, $1'A=0$, $1'B=0$이 되며, Johnson과 Kotz(1970)에 의하여 다음이 성립한다.

$$y'Ay=z'Dz \sim \sum_{i=1}^{t-1}\lambda_i \chi^2(1)$$
$$y'By=z'Ez \sim \sum_{i=1}^{n-t}\lambda_i^* \chi^2(1) \qquad (2.4)$$

여기서 $D=\sigma_e^2 T^{-1}AT^{-1}$, $E=\sigma_e^2 T^{-1}BT^{-1}$이고, λ_i는 D의 고유값 $(\lambda_1>\lambda_2>\cdots>\lambda_{t-1})$이고, λ_i^*는 E의 고유값 $(\lambda_1^*>\lambda_2^*>\cdots>\lambda_{t-1}^*)$ 이다. 또한 $AVB=0$이므로 $y'Ay$와 $y'By$는 독립이다. 따라서 $\hat{\sigma}_a^2$의 분포는 다음과 같다.

$$\begin{aligned}\hat{\sigma}_a^2 &= (1/k)[y'Ay-\{(t-1)/(n-t)\}y'By] \\ &\sim (1/k)[\sum\lambda_i \chi^2(1)-\{(t-1)/(n-t)\}\sum\lambda_i^* \chi^2(1)] \\ &\sim \sum_1^{n-1} a_i \chi^2(1) \qquad (2.5)\end{aligned}$$

여기서 $k=n-\sum(n_i^2/n)$, $a_1=(1/k)\lambda_1$, $\cdots$, $a_{t-1}=(1/k)\lambda_{t-1}$, $a_t=\{(t-1/n-t)\}\lambda_i^*$, $\cdots$, $a_{n-1}=\{(t-1)/(n-t)\}\lambda_{n-t}^*$ 이다. 이 때 $\hat{\sigma}_a^2$가 음이 될 확률은 다음과 같다.

$$P(\hat{\sigma}_a^2<0) = P(\sum_i^{n-1} a_i \chi^2(1)<0) \qquad (2.6)$$

식 (2.6)의 계산은 Farebrother(1984)의 알고리즘에 의해 구할 수 있다. 그러나 $\rho=0$인 경우 $\hat{\sigma}_a^2$의 분포는 다음과 같이 된다.

$$\hat{\sigma}_a^2 \sim k'\chi^2(t-1)-k'\{(t-1)/(n-t)\}\chi^2(n-t) \qquad (2.7)$$

여기서 $k'=\sigma_e^2/k$, k는 식(2.5)와 같다. 따라서 $\rho=0$ 인 경우 $\hat{\sigma}_a^2$ 가 음이 될 확률은 다음과 같다.

$$P(\hat{\sigma}_a^2<0) = P[F(t-1,\ n-t)<1] \qquad (2.8)$$

이것은 Searle(1971)의 균형 일원 변량모형에서 AOV추정량이 음이 될 확률을 구하는 것과 같음을 알 수 있다.

2.2 MINQUE의 분포와 음이 될 확률

MINQUE는 변수변환에 대하여 불변이므로 모형 (1.2) 대신에 다음과 같은 변환모형을 생각할 수 있다.

$CC' = I_{n-1}$, $C1 = 0$이 되는 (n-1)×n 직교대비행렬 C는 $CC' = I_n - (1/n)J_n$의 성질을 갖는다. 따라서 식 (1.2)의 양변에 직교대비행렬 C를 곱하면 다음과 같은 변환모형을 얻을 수 있다.

$$t = Cy = CU\alpha + C\varepsilon \tag{2.9}$$

여기서 t는 (n-1)x1 벡터로 평균 0 이고, 분산은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} Var(t) &= \sigma_a^2 CUU'C' + \sigma_e^2 I_{n-1} = \sigma_e^2(\rho W + I_{n-1}) \\ &= \sigma_e^2 H \end{aligned}$$

여기서 $H = \rho W + I_{n-1}$, $\rho = \sigma_a^2/\sigma_e^2$, $W = CUU'C'$이다. 이 때 r를 ρ의 사전값이라 하고, H에서 ρ 대신 r로 대치한 H^{-1}를 R이라 하면 R은 다음과 같다.

$$R = (I_{n-1} + rW)^{-1} \tag{2.10}$$

그러면 σ_a^2과 σ_e^2의 MINQUE는 다음으로 부터 구할 수 있다.

$$S \cdot \sigma = q \tag{2.11}$$

여기서 $\sigma' = [\sigma_a^2, \sigma_e^2]$이고, s_{ij}를 S의 (i, j)번째 원소, q_i를 q의 i번째 원소라 하면

$$\begin{aligned} s_{ij} &= tr(RV_iRV_j), \quad i,j = 1,2 \\ q_i &= t'RV_iRt, \quad i = 1,2 \end{aligned} \tag{2.12}$$

이고, $V_1 = CUU'C'$, $V_2 = CC' = I_{n-1}$이다. SVD에 의해 $W = CUU'C' = P\Lambda P'$으로 나타낼 수 있고, W는 양반정치 행렬이므로 $\Lambda = diag(\lambda_1, \cdots, \lambda_{t-1}, 0, \cdots, 0)_{n-1}$이고, P는 고유벡터로 이루어진 (n-1)x(n-1) 행렬이다. 따라서 식 (2.10)의 R은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} R &= (I_{n-1} + rW)^{-1} \\ &= (PP' + rP\Lambda P')^{-1} \\ &= P(I + r\Lambda)^{-1}P' \\ &= PD^{-1}P' \end{aligned}$$

여기서 $D = diag(1 + r\lambda_1, \cdots, 1 + r\lambda_{t-1}, 1, \cdots, 1)_{n-1}$ 이다. 이상으로 부터 식 (2.12)의 s_{ij}와 q_i는

다음과 같이 나타낼 수 있다.

$$
\begin{aligned}
s_{11} &= tr(RV_1RV_1) = \sum_{i=1}^{n-1}\{\lambda_i/(1+r\lambda_i)\}^2 \\
s_{12} &= s_{21} = \sum_{i=1}^{n-1}\{\lambda_i/(1+r\lambda_i)^2\} \\
s_{22} &= (n-t) + \sum_{i=1}^{t-1}\{1/(1+r\lambda_i)^2\} \\
q_1 &= t' PD^{-2}\Lambda P' t \\
q_2 &= t' PD^{-2}P' t
\end{aligned}
\tag{2.13}
$$

식 (2.9)에서 t가 정규분포를 따른다면 t의 분포는 다음과 같다.

$$t \sim N(0,\ \sigma_e^2(I_{n-1}+\rho P\Lambda P'))$$

그러면,

$$u=(1/\sigma_e)(I+\rho\Lambda)^{-1/2}P' t \sim N(0,\ I_{n-1}) \tag{2.14}$$

이 되어

$$t=\sigma_e P(I+\rho\Lambda)^{1/2}u$$

로 나타낼 수 있다. 그러므로

$$
\begin{aligned}
\hat{\sigma}_a^2 &= (1/|S|)(s_{22}q_1 - s_{12}q_2) \\
&= u' Mu \\
&\sim \sum_{i=1}^{n-1} b_i\ \chi^2(1)
\end{aligned}
\tag{2.15}
$$

여기서 $|S| = s_{11}s_{22} - s_{12}s_{21}$, $M=(\sigma_e^2/|S|)(1+\rho\Lambda)D^{-2}(s_{22}\Lambda - s_{12}I)$이고, b_i는 M의 고유값이다. 따라서 σ_a^2 의 MINQUE $\hat{\sigma}_a^2$ 의 분포도 상호독립인 $\chi^2(1)$의 선형결합으로 나타나고, 음이 될 확률도 Farebrother(1984)의 알고리즘을 적용하여 구할 수 있다.

σ_a^2 의 MINQUE를 적용할 때는 ρ에 대한 사전정보가 필요하며 보통 0, 1, ∞ 등을 적용한다. 사전값이 0, 1 일때 MINQUE는 식 (2.13)에 r=0 또는 r=1을 대입하여 식 (2.15)로 부터 구할 수 있다.

사전값이 ∞인 MINQUE는 모형 (1.2)에서 Kaplan(1982)과 Westfall(1987a) 등이 구하였

으며, 이때도 σ_a^2 의 MINQUE $\hat{\sigma}_a^2$ 의 분포 또한 상호독립인 $\chi^2(1)$의 선형결합으로 나타낼 수 있고, Farebrother(1984)의 알고리즘에 의해 추정량이 음이 될 확률을 구할 수 있다.

3. 비정규분포에서 분산성분의 추정량이 음이 될 확률

정규분포를 가정하지 않는 경우에 AOV추정량과 MINQUE의 분포를 구하는 것은 매우 어려우며, 추정량이 음이 될 확률을 구하기도 쉽지 않다. 그러므로 본 장에서는 정규분포를 따르지 않는 경우에 추정량이 음이 될 확률에 대해 모의 실험을 통하여 알아보고자 한다.

3.1 모의실험을 위한 실험설계

본 연구에서 고려한 추정방법은 AOV방법과 사전값이 0, 1, ∞인 MINQUE방법으로 4가지이다.

분포 형태는 Westfall(1987b)에서와 같이 첨도가 -1.2, 0, 6인 대상으로 한다. 첨도가 -1.2인 분포는 균일분포, 첨도가 0인 분포는 정규분포, 그리고 첨도가 6인 분포는 두 정규분포의 비율을 가중치로 하는 혼합분포를 이용한다. 이 때 α_i와 ε_{ij} 의 첨도를 각각 k_a, k_e라 하면 $(k_a, k_e)=(0, 0)$를 중심으로한 (k_a, k_e)의 4가지 조합 (-1.2, 0), (6, 0), (0, -1.2), (0, 6)을 분포형태로 택한다.

다음에 고려한 요인은 Westfall(1987b)과 같이 급내상관 θ로 θ를 10가지 값으로 세분한 θ=0, 0.1, ⋯ , 0.9로 하여, 이에 대응하는 $\rho(=\sigma_a^2/\sigma_e^2)$가 0, 1/9, 1/4, 3/7, 2/3, 1, 3/2, 7/3, 4, 9인 값을 대상으로 한다. 그러면 이 3요인에 의한 실험조건은 4×4×10=160가지가 된다.

실험조건의 각 경우에 대한 디자인 형태는 표본수, 수준수, 불균형성을 고려하여 각각 3수준으로 3×3 라틴방격법에 의해 실험설계를 한다.

Swallow와 Searle(1978)에 의하여 표본수는 적은, 보통, 많은 경우로 나누어 각각 15, 30, 75으로 하고, 수준수는 적은, 보통, 많은 경우로 나누어 각각 3, 6, 9로 한다. 또한 불균형성은 다음의 불균형 측도 v에 의해 결정한다.

$$v = (1/n)\Sigma(n_i-\bar{n})^2/\ \bar{n} \qquad (3.1)$$

여기서 $n=\Sigma n_i$, $\bar{n}=(1/t)\Sigma n_i$ 이다.

불균형 정도는 v의 값이 0.3이하, 0.3에서 0.6, 0.6이상으로 나누어 각 경우를 약한, 보통, 심한 불균형으로 구분한다.

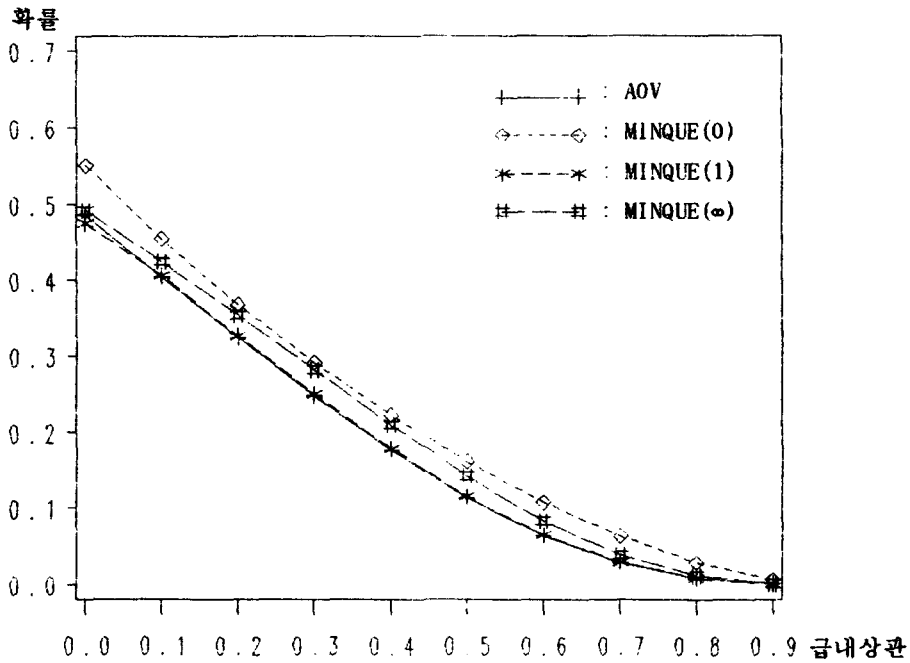
이상의 표본수, 수준수, 불균형성에 대하여 모의실험에 사용한 디자인 형태를 나타내면 다음과 같다.

(표) 모의실험에 사용한 디자인 형태

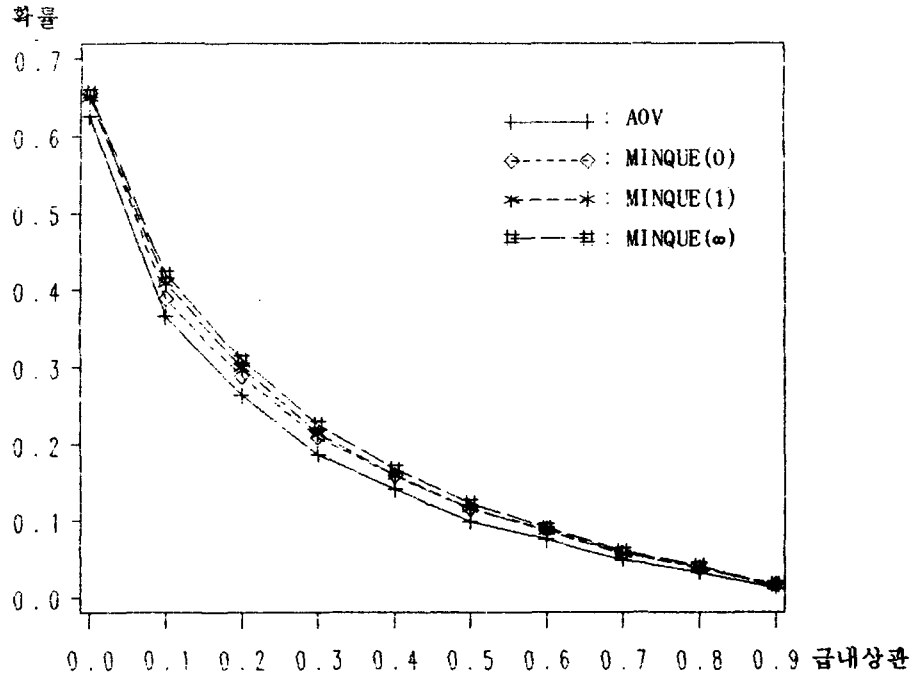
디자인 형태	수준수	불균형성	표본수
(3,5,7)	3	약한	15
(2,10,18)	3	중간	30
(5,15,55)	3	심한	75
(3,3,5,5,7,7)	6	약한	30
(5,5,10,10,15,30)	6	보통	75
(1,1,1,2,3,7)	6	심한	15
(5,5,5,5,10,10,10,10)	9	약한	75
(1,1,1,1,1,1,2,3,4)	9	보통	15
(2,2,2,2,2,2,2,4,12)	9	심한	30

3.2 실험결과 분석

각 분포에서 디자인 형태에 대하여 ρ에 대한 각 추정량이 음이 될 확률을 계산한 것 중에서 2가지 경우를 택해 그림으로 나타내면 다음과 같다.



(그림1) $(k_a, k_e)=(0,0)$일때 (1, 1, 1, 1, 1, 1, 2, 3, 4)에서
ρ에 대한 $\hat{\sigma}_a^2$이 음이 될 확률



(그림2) $(k_a, k_e)=(6, 0)$일때 (5, 15, 55)에서 ρ에 대한 $\hat{\sigma}_a^2$이 음이 될 확률

(그림1)은 $(k_a, k_e)=(0, 0)$일때 표본수가 적고, 수준수가 많고, 불균형이 보통인 경우로 ρ에 따라 AOV추정량과 MINQUE(1)이 낮은 확률을 갖고, MINQUE(0)가 높은 확률을 보이고 있다. (그림2)는 $(k_a, k_e)=(6, 0)$일때 표본수가 많고, 수준수가 많고, 불균형성이 심한 경우로 ρ에 따라 AOV추정량이 낮은 확률을 갖고, MINQUE(∞)가 높은 확률을 보이고 있다.

정규분포에서의 추정량이 음이 될 이론적확률과 모의실험 결과에 대하여 분산분석 기법을 적용하고, 각 분포에서 ρ에 따른 실험 분석의 결과에 의하여 다음과 같은 결과를 얻을 수 있다.

1. AOV추정량과 MINQUE가 음이 될 확률은 분포간에 차이가 없어 정규분포를 가정할 때 나타나는 이론적 확률로 구할 수 있다.
2. ρ, 수준수, 표본수가 커지면 각 추정량은 음이 될 확률이 작아진다.
3. ρ, 수준수, 표본수가 적고, 불균형이 심하면 각 추정량은 커진다.
4. 표본수가 적고, 수준수가 많으면 추정량이 음이 될 확률은 각 방법간에 차이가 심해진다.
5. AOV추정량과 MINQUE 중에서 대부분의 경우 AOV추정량이 음이 될 확률이 가장 작다.

4. 결론

불균형 일원 변량모형에서 분산성분을 추정하는 방법에는 여러가지가 있으나 본 연구에서는 AOV방법과 MINQUE방법에 국한하여 고찰하였다. 그런데 정규분포를 가정할 때 AOV추정량과 MINQUE는 분포를 구할 수 있어 추정량이 음이 될 이론적 확률을 얻을 수 있었다. 그러나 비정규분포에 대하여 AOV추정량과 MINQUE는 그 분포와 이론적 확률을 구하기가 어려워 모의실험을 통하여 추정량이 음이 될 확률을 구하였다.

이와 같은 확률의 결과로 부터 정규분포를 가정할 때 AOV추정량과 MINQUE의 음이 될 이론적 확률은 비정규분포인 경우 모의실험에 의해 구한 음이 될 확률과 유의한 차이가 없는 것으로 나타나 정규분포를 가정하여 추정량이 음이 될 이론적 확률로 비정규분포에 적용해도 큰 무리가 없음을 알 수 있었다. 또한 고려된 추정량 중에서 AOV추정량이 대부분의 경우 음이 될 확률이 가장 작게 나타났고, 표본수, 수준수, ρ가 클 때 각 방법은 음이 될 확률이 작아지며, 반면에 표본수, 수준수, ρ가 적고 불균형이 심해지면 음이 될 확률이 커짐을 알 수 있었다. 각 디자인 형태에서 표본수, 수준수, ρ가 크면 음이 될 확률이 작아짐을 보였다.

5. 참고문헌

1. Farebrother, R. W. (1984), "The Distribution of a Linear Combination of central χ^2 Random Variables," Applied Statistics, 363-396.
2. Johnson and Kotz. (1970), Continuous Univariate Distribution-2. Houghton Mifflin Company.
3. Kaplan, J. (1982), "A Theorem Relating MINQUE and Unweighted Means Estimators of Variance Components in the One-Way Design," Communications in Statistics - Theory and Methods, 11(4), 423-428.
4. Searle, S. R. (1971), Linear Models. New York: Wiley.
5. Swallow, W. H. and Sezarle, S. R. (1978), "Minimum Variance Quadratic Unbiased Estimation of Variance Components," Technometrics, 20, 265-272.
6. Verdooren, L. R. (1982), "How large is the probability for the estimate of a variance component to be negative?," Biometrical Journal, 24, 339-360.
7. Westfall, P. H. (1987a), "Computable MINQUE-Type Estimates of Variance Components," Journal of the American Statistical Association, 82, 586-589.
8. Westfall, P. H. (1987b), "A Comparison of Variance Component Estimates for Arbitrary Underlying Distributions," Journal of the American Statistical Association, 82, 866-874.

On the Probability of the Estimate of Variance Components that is Negative in Unbalanced One-Way Random Model[1]

Gyu Moon Song[2]

ABSTRACT

For the One-way random effects model with unbalanced data, the AOV and MINQUE estimates of variance components are frequently found to be negative. The primary objective of present study is placed on the computation of the probability of the main effect variance component, , being negative. The probability of negative estimators from AOV and MINQUE can be obtained by theoretical computation under the normality assumption. It is, however, difficult to compute the probability of negative estimates for these estimators under arbitrary distributions, and hence their probabilities of being negative were computed by simulation experiment in this study. It was shown that there was no significant difference between the theoretical probability under normality and calculated probability by simulation experiment, and that probability of negative estimates decreases as sample size, number of levels and the value of increase.

[1] This research was supported by the Research Foundation of Keimyung University 1992

[2] Department of Statistics, Keimyung University, Taegu 704-701, Korea.