The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 667-675

Why do we get Negative Variance Components in ANOVA?[1)]

Jang-Taek Lee[2)]

Abstract

The usefulness of analysis of variance(ANOVA) estimates of variance components is impaired by the frequent occurrence of negative values. The probability of such an occurrence is therefore of interest.

In this paper, we investigate a variety of reasons for negative estimates under one way random effects model. It can be shown, through simulation, that this probability increases when the number of treatments is too small for fixed total observations, unbalancedness of data is severe, ratio of variance components is too small, and data may contain many outliers.

Keywords : negative estimates; ANOVA.

1. 서론

반복수가 같지 않은 i번째 처리효과에 있어서 j번째 관측치에 대한 불균형일원변량모형은 다음과 같이 나타낼 수 있다.

$$y_{ij} = \mu + a_i + e_{ij}, \qquad i = 1, 2, \cdots, k, \ j = 1, 2, \cdots, n_i. \tag{1.1}$$

위의 모형에서 μ는 미지의 모수, a_i와 e_{ij}는 서로 독립이며, 평균이 0이고 분산이 각각 σ_a^2와 σ_e^2인 정규확률변수, n_i는 i번째 처리효과에 있어서 관측치의 개수이다. 위와 같은 불균형일원변량모형에 대한 중요한 논의의 대상은 모형의 변량인자에 대한 분산성분을 추정하는 문제인데, 이 경우 여러 가지 분산성분추정량 중에서 가장 많이 사용되는 것은 분산분석표를 이용하여 구해지는 분산분석(ANOVA) 추정량이라고 할 수 있다. 하지만 모형의 가정이 타당함에도 불구하고 σ_a^2의 분산분석 추정량은 실제 분산성분의 값이 음수로 추정될 수 있는데, 따라서 본 논문에서는 여러 가지 k, n_i, 급내상관계수 ρ의 값에 대해 σ_a^2의 분산분석추정량이 음이 될 확률을 구해 보고, 추정치가 음수가 되는 보다 구체적인 원인을 알아보고자 한다.

본 논문의 구성은 1절에서는 연구의 기본적인 배경을 설명하고, 2절에서는 균형일원변량모형에서 분산분석추정량이 음수가 될 확률을 분포함수를 이용하여 구해본다. 그리고 3절에서는 2절의

1) This work was supported by Dankook University Research Fund in 2000
2) Professor, Division of Natural Science, Dankook University, San 8, HanNam-Dong, YongSan-Gu, Seoul, Korea
E-mail : jtlee@dankook.ac.kr

내용을 불균형일원변량모형으로 확장하여 여러 가지 모양의 실험계획에 대한 모의실험을 통하여 분산분석추정치가 음수가 되는 주요원인을 살펴본다. 끝으로 4절에서는 본 연구의 결론이 주어진다.

2. 균형일원변량모형의 경우

모형(1.1)에 대한 분산분석표를 작성하면 다음 <표 1>과 같다.

<표 1> 불균형일원변량모형의 분산분석표

요인	자유도	제곱합	제곱평균	평균제곱기대값
처리	$k-1$	$SSA=\sum n_i(\bar{y}_{i.}-\bar{y}_{..})^2$	$MSA=SSA/(k-1)$	$\sigma_e^2+n_0\sigma_a^2$
오차	$N-k$	$SSE=\sum(y_{ij}-\bar{y}_{i.})^2$	$MSE=SSE/(N-k)$	σ_e^2
총합	$N-1$	$SST=\sum(y_{ij}-\bar{y}_{..})^2$		

<표 1>에서 사용된 표기는 $N=\sum_{i=1}^{k}n_i$는 총 관측치의 개수, $\bar{y}_{i.}=\sum_{j=1}^{n_i}y_{ij}/n_i$는 i번째 그룹의 평균, $\bar{y}_{..}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}/N$는 총평균, 그리고 n_0는 $n_0=(N-\sum_{i=1}^{k}n_i^2/N)/(k-1)$로 정의된 가중평균 관측개수이다. 이 경우 σ_a^2의 분산분석추정량 $\hat{\sigma}_a^2$은 <표 1>의 제곱평균이 평균제곱기대값과 같다고 두고 구한 추정량으로 정의되며, $\hat{\sigma}_a^2$ 는 다음과 같이 구할 수 있다.

$$\hat{\sigma}_a^2 = (MSA-MSE)/n_0 \tag{2.1}$$

그리고 만일 균형일원변량모형인 경우($n_i=n$)이면 σ_a^2의 분산분석추정량이 음이 될 확률 p는 다음과 같이 구할 수 있다.

$$P(\hat{\sigma}_a^2 \le 0) = P(F(k-1,k(n-1)) \le 1/(1+\theta n)),\ \theta=\sigma_a^2/\sigma_e^2. \tag{2.2}$$

다음 <표 2>는 k와 n이 2부터 1간격으로 50까지와 급내상관계수 $\rho=\theta/(1+\theta)$는 0.00부터 0.99까지 0.01 간격으로 선택되어지는 모든 경우의 음이 될 확률 p의 값 240100개를 이용하여 구한 log(p)와 k, n, ρ의 표본상관계수이다.

<표 2> log(p)와 k, n, ρ의 표본상관계수

	k	n	ρ
log(p)	-0.59403	-0.26461	-0.61837

<표 2>에서 log(p)는 k, n, ρ의 값이 증가할수록 대체적으로 감소하는 것을 알 수 있다. 그리고 변수 log(p)와 k의 상관관계가 크기 때문에 비록 ρ값을 모르더라도 k의 값을 잘 선택함으로써 음이 될 확률 p를 훨씬 작게 할 수 있다는 사실을 확인할 수 있다. 다음 <표 3>은 <표 2>를 구할 때 사용한 (k,n,ρ)의 조합에 대하여 급내상관계수 ρ의 여러 가지 값에 대한 확률 p의 최대값과 그때의 k와 n의 값이다.

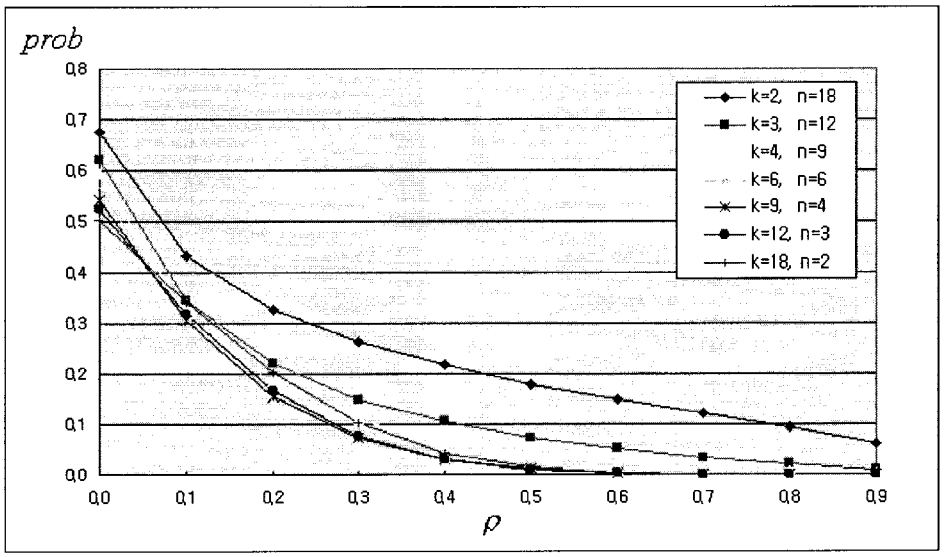
<표 3> 분산성분추정량이 음이 될 최대확률

ρ	(k,n)	확률 p	ρ	(k,n)	확률 p
0.0	(2,50)	0.6802	0.5	(2,2)	0.3780
0.1	(2,3)	0.5647	0.6	(2,2)	0.3333
0.2	(2,3)	0.5082	0.7	(2,2)	0.2847
0.3	(2,2)	0.4606	0.8	(2,2)	0.2294
0.4	(2,2)	0.4201	0.9	(2,2)	0.1601

<표 3>으로부터 분산성분추정량이 음이 될 확률은 가장 큰 경우에는 0.68정도가 되는 사실과 (k,n)의 조합이 잘못 선택되면 비록 $\rho=0.9$인 경우에도 확률이 0.16이나 된다는 사실, 그리고 여러 가지 ρ의 값에 대해 음이 될 최대확률은 모두 $k=2$에서 발생하는 것을 확인할 수 있다.

다음 <그림 1>은 총 자료개수가 $N=36$인 경우에 가능한 (k,n)의 7가지 실험계획에 대한 σ_a^2의 분산분석추정량이 음이 될 최대확률을 그래프로 표시한 결과이다. 이 경우 급내상관계수 ρ는 0.0부터 0.1 간격으로 0.9까지 선택하였다. <그림 1>로부터 모든 ρ의 값에 대하여 항상 음이 될 확률 p를 최소로 하는 실험계획은 존재하지 않으며, 처리효과의 수준의 수 k 값이 증가하고, 반복의 수 n값이 감소하면 σ_a^2의 분산분석추정량이 음이 될 최대확률은 대체적으로 감소하지만, k값이 9이상이면 확률 p는 거의 비슷하여진다는 사실을 확인할 수 있다.

<그림 1: 여러 가지 k와 n에 대한 음이 될 확률>



<표 4>는 고정된 총 관측치의 개수 N에 대한 $\widehat{\sigma_a^2}$가 음이 될 확률 p의 평균값을 최소로 하는 실험계획을 제시한다. <표 4>에서 사용된 N의 값은 12이상 100이하의 값으로 선택되어졌고, 특히 4개 이상의 (k,n) 조합이 가능한 경우만 예시되어 있다. <표 4>의 평균확률 값 $\overline{p}$는 ρ의 값을 모

른다고 가정하고, 급내상관계수를 0.00부터 0.99까지 0.01 간격으로 고려한 100개의 자료를 이용하여 구한 음이 될 확률 p의 평균값이다. 이 경우 급내상관계수의 값을 더 세분하면 확률의 평균값만 단지 바뀔 뿐이지 선택되어지는 실험계획은 동일함을 확인할 수 있었다. <표 4>에는 생략되어 있지만, k와 n의 조합이 2개가 가능한 경우에는 항상 고정된 총 관측치수 N에 대하여 $N=kn$을 만족하는 최대자연수 k를 택하면 된다. 예를 들어 $N=6$이면 $(k,n)=(3,2)$의 값을 택하면 된다.

<표 4> 총 관측치수 N에 대한 최적인 k와 n의 값

N	k	n	$\bar{p}$	N	k	n	$\bar{p}$
12	6	2	0.1729	56	8	7	0.0612
16	8	2	0.1477	60	10	6	0.0575
18	9	2	0.1388	63	9	7	0.0554
20	10	2	0.1313	64	8	8	0.0551
24	8	3	0.1139	66	11	6	0.0534
28	7	4	0.1032	70	10	7	0.0508
30	10	3	0.0983	72	9	8	0.0497
32	8	4	0.0932	75	15	5	0.0497
35	7	5	0.0883	78	13	6	0.0472
36	9	4	0.0854	80	10	8	0.0455
40	8	5	0.0791	81	9	9	0.0452
42	7	6	0.0774	84	12	7	0.0440
44	11	4	0.0740	88	11	8	0.0422
45	9	5	0.0721	90	10	9	0.0413
48	8	6	0.0689	96	12	8	0.0394
50	10	5	0.0665	98	14	7	0.0393
52	13	4	0.0660	99	11	9	0.0382
54	9	6	0.0625	100	10	10	0.0379

3. 불균형일원변량모형의 경우

불균형일원변량모형에 있어서 분산분석추정량이 음이 될 확률을 해석적으로 구하는 것은 불가능하다. 왜냐하면 불균형자료인 경우는 식(2.2)와 같은 분포함수 관계식이 존재하지 않기 때문이다. 따라서 이 절에서는 여러 가지 불균형자료에 대한 모의실험을 수행하여 분산분석추정량이 음이 될 확률을 구하고 그에 따른 추론을 하고자 한다.

모의실험에 사용되는 자료의 불균형도를 측정하기 위하여 다음과 같은 Hess(1979)의 측도를 사용하였다.

$$D=\sum_{i=1}^{k}(n_i-\bar{n})^2-\text{최소값}\left[\sum_{i=1}^{k}(t_i-\bar{n})^2\right] \tag{3.1}$$

식(3.1)에서 $\bar{n}$는 $\bar{n}=N/k$으로 정의되며, 최소값$\left[\sum_{i=1}^{n_i}(t_i-\bar{n})^2\right]$은 $\sum_{i=1}^{k}t_i=N$을 만족하며, $t_i\geq 1$인 자연수 $t_1,t_2,\cdots,t_k$을 이용하여 구해진다. 그리고 T^*를 $T^*=D/D_{max}$로 정의한다. 이 경우 D_{max}는 $\sum_{i=1}^{k}n_i=N$를 만족하고, $n_i\geq 1$인 모든 가능한 자연수의 조합을 이용하여 구한 D의 최대값이다. 그러면 자료가 균형이면 T^*는 $T^*=0$가 되며, 자료가 불균형도가 커질수록 1에 가까운

T^*가 구해지며, 불균형도가 가장 커지면 $T^* = 1$이 된다.

불균형일원변량모형에 대한 분산분석추정량이 음이 될 확률에 관한 보다 일반적인 결론을 유도하기 위하여 모의실험에서 다양한 실험계획을 고려하였으며, 자세한 실험계획은 <표 5>에 제시되어 있다. <표 5>에서 U_1, U_2, U_3는 각각 불균형도가 근사적으로 1, 0.5, 0에 가장 가까운 실험계획을 의미한다. 또한 모양에서 ( )는 반복수인데, 예를 들어 $(N, k) = (15, 3)$의 U_1인 경우, [1(2),13]은 1, 2번째 셀의 관측치 수는 각각 1개이고, 세 번째 셀은 13개임을 뜻한다.

<표 5> 모의실험에 고려된 실험계획의 종류 및 모양

(N,k)	이름	모양	T^*	(N,k)	이름	모양	T^*
(15,3)	U_1	[1(2),13]	1	(45,3)	U_1	[1(2),43]	1
	U_2	[1,4,10]	0.438		U_2	[1,10,34]	0.495
	U_3	[5(3)]	0		U_3	[15(3)]	0
(15,6)	U_1	[1(5),10]	1	(45,6)	U_1	[1(5),40]	1
	U_2	[1(4),4,7]	0.454		U_2	[1(2),2,4,7,30]	0.499
	U_3	[3(3),2(3)]	0		U_3	[7,8,7,8(2),7]	0
(15,9)	U_1	[1(8),7]	1	(45,9)	U_1	[1(8),37]	1
	U_2	[1(7),5,3]	0.468		U_2	[1,2,1,4,1,7,1(2),27]	0.502
	U_3	[1(3),2(6)]	0		U_3	[5(9)]	0
(30,3)	U_1	[1(2),28]	1	(60,3)	U_1	[1(2),58]	1
	U_2	[1,7,22]	0.482		U_2	[6,7,47]	0.505
	U_3	[10(3)]	0		U_3	[20(3)]	0
(30,6)	U_1	[1(5),25]	1	(60,6)	U_1	[1(5),55]	1
	U_2	[1(2),3(3),19]	0.5		U_2	[1(3),10,6,41]	0.499
	U_3	[5(6)]	0		U_3	[10(6)]	0
(30,9)	U_1	[1(8),22]	1	(60,9)	U_1	[1(8),52]	1
	U_2	[1(2),2,1(3),5,2,16]	0.492		U_2	[1(5),10,2,5,38]	0.509
	U_3	[3(6), 4(3)]	0		U_3	[6(3),7(6)]	0

난수생성은 통계패키지 SAS를 이용하였으며, 여러 가지 실험계획 및 급내상관계수에 대하여 일반성을 잃지 않고 두 가지 가정 $\sigma_a^2 + \sigma_e^2 = 1$과 $\mu = 0$을 사용하였다. 다음 <표 6>은 <표 5>의 실험계획을 이용하여 수행된 모의실험의 결과이다. 그리고 <표 7>은 고정된 N의 값에 대하여 k 값과 불균형도가 음이 될 확률에 미치는 영향력을 보다 상세하게 알아보기 위한 $N = 36$인 경우에 있어서의 모의실험 결과이다.

두 가지 모의실험결과 <표 6>과 <표 7>을 통하여 알 수 있는 중요한 결론들은 다음과 같다.

1. 고려된 모든 실험계획에 대하여 음이 될 확률은 $\rho = 0$를 제외하고는 불균형도의 영향을 받으

며, 불균형도가 심해질수록 음이 될 확률은 커진다.

N	k	ρ	U_1	U_2	U_3	N	k	ρ	U_1	U_2	U_3
15	3	0.0	0.6050	0.6064	0.6053	45	3	0.0	0.6272	0.6257	0.6174
		0.1	0.5460	0.5062	0.4560			0.1	0.5498	0.4150	0.3031
		0.2	0.4742	0.4139	0.3493			0.2	0.4785	0.3024	0.1834
		0.3	0.4207	0.3423	0.2654			0.3	0.4267	0.2310	0.1229
		0.4	0.3626	0.2695	0.1962			0.4	0.3646	0.1818	0.0849
		0.5	0.2961	0.2085	0.1458			0.5	0.2960	0.1359	0.0561
		0.6	0.2389	0.1615	0.1097			0.6	0.2398	0.1072	0.0424
		0.7	0.1863	0.1175	0.0752			0.7	0.1795	0.0714	0.0260
		0.8	0.1191	0.0729	0.0461			0.8	0.1109	0.0414	0.0152
		0.9	0.0616	0.0355	0.0229			0.9	0.0627	0.0250	0.0090
	6	0.0	0.5337	0.5313	0.5312		6	0.0	0.5652	0.5646	0.5699
		0.1	0.4493	0.4341	0.4154			0.1	0.4596	0.3508	0.2519
		0.2	0.3638	0.3385	0.3036			0.2	0.3656	0.2236	0.1127
		0.3	0.3006	0.2609	0.2189			0.3	0.2804	0.1408	0.0547
		0.4	0.2315	0.1931	0.1446			0.4	0.2117	0.0895	0.0260
		0.5	0.1742	0.1351	0.0964			0.5	0.1476	0.0481	0.0110
		0.6	0.1147	0.0839	0.0519			0.6	0.0926	0.0254	0.0039
		0.7	0.0632	0.0462	0.0251			0.7	0.0488	0.0113	0.0018
		0.8	0.0272	0.0181	0.0089			0.8	0.0190	0.0032	0.0005
		0.9	0.0054	0.0034	0.0014			0.9	0.0037	0.0006	0.0000
	9	0.0	0.4819	0.4808	0.4803		9	0.0	0.5455	0.5455	0.5456
		0.1	0.4072	0.4012	0.3965			0.1	0.4161	0.3376	0.2553
		0.2	0.3441	0.3299	0.3201			0.2	0.3120	0.2088	0.1140
		0.3	0.2678	0.2526	0.2376			0.3	0.2228	0.1235	0.0455
		0.4	0.2038	0.1897	0.1712			0.4	0.1462	0.0673	0.0193
		0.5	0.1387	0.1224	0.1065			0.5	0.0905	0.0379	0.0066
		0.6	0.0792	0.0694	0.0568			0.6	0.0482	0.0171	0.0018
		0.7	0.0377	0.0337	0.0224			0.7	0.0178	0.0059	0.0003
		0.8	0.0124	0.0090	0.0056			0.8	0.0055	0.0008	0.0000
		0.9	0.0013	0.0008	0.0004			0.9	0.0005	0.0000	0.0000
30	3	0.0	0.6225	0.6218	0.6160	60	3	0.0	0.6308	0.6281	0.6224
		0.1	0.5511	0.4526	0.3678			0.1	0.5527	0.3644	0.2580
		0.2	0.4783	0.3406	0.2423			0.2	0.4802	0.2370	0.1484
		0.3	0.4242	0.2720	0.1715			0.3	0.4234	0.1617	0.0952
		0.4	0.3605	0.2117	0.1187			0.4	0.3613	0.1159	0.0652
		0.5	0.2965	0.1649	0.0837			0.5	0.2953	0.0824	0.0432
		0.6	0.2377	0.1243	0.0623			0.6	0.2406	0.0633	0.0314
		0.7	0.1822	0.0851	0.0385			0.7	0.1759	0.0389	0.0195
		0.8	0.1129	0.0513	0.0238			0.8	0.1110	0.0244	0.0117
		0.9	0.0617	0.0292	0.0128			0.9	0.0613	0.0129	0.0070
	6	0.0	0.5583	0.5596	0.5577		6	0.0	0.5687	0.5662	0.5719
		0.1	0.4585	0.3899	0.3255			0.1	0.4597	0.3190	0.2000
		0.2	0.3692	0.2681	0.1770			0.2	0.3619	0.1959	0.0802
		0.3	0.2841	0.1750	0.0973			0.3	0.2758	0.1204	0.0351
		0.4	0.2149	0.1139	0.0540			0.4	0.2128	0.0756	0.0152
		0.5	0.1551	0.0686	0.0269			0.5	0.1426	0.0407	0.0070
		0.6	0.0971	0.0392	0.0118			0.6	0.0912	0.0235	0.0028
		0.7	0.0499	0.0167	0.0050			0.7	0.0491	0.0106	0.0011
		0.8	0.0220	0.0059	0.0014			0.8	0.0201	0.0027	0.0003
		0.9	0.0037	0.0010	0.0002			0.9	0.0030	0.0005	0.0000
	9	0.0	0.5338	0.5334	0.5333		9	0.0	0.5513	0.5508	0.5490
		0.1	0.4139	0.3729	0.3262			0.1	0.4170	0.3083	0.2015
		0.2	0.3189	0.2593	0.1900			0.2	0.3124	0.1827	0.0730
		0.3	0.2422	0.1710	0.1033			0.3	0.2176	0.0983	0.0241
		0.4	0.1633	0.1036	0.0483			0.4	0.1480	0.0547	0.0089
		0.5	0.0950	0.0556	0.0201			0.5	0.0872	0.0257	0.0020
		0.6	0.0534	0.0279	0.0075			0.6	0.0474	0.0136	0.0009
		0.7	0.0213	0.0091	0.0014			0.7	0.0170	0.0040	0.0001
		0.8	0.0065	0.0019	0.0005			0.8	0.0050	0.0013	0.0000
		0.9	0.0005	0.0002	0.0001			0.9	0.0004	0.0001	0.0000

<표 6> 불균형자료인 경우에 있어서 음이 될 확률

<표 7> N=36인 경우에 있어서 음이 될 확률

불균형도	ρ	(36,2)	(36,3)	(36,4)	(36,6)	(36,9)	(36,12)	(36,18)
U_1	0.0	0.6709	0.6244	0.5926	0.5607	0.5413	0.5198	0.4985
	0.1	0.6296	0.5525	0.5150	0.4605	0.4151	0.3949	0.3803
	0.2	0.5815	0.4772	0.4373	0.3682	0.3158	0.2871	0.2696
	0.3	0.5352	0.4227	0.3536	0.2794	0.2327	0.2081	0.1794
	0.4	0.4889	0.3613	0.2863	0.2140	0.1546	0.1187	0.0978
	0.5	0.4269	0.2967	0.2236	0.1522	0.0942	0.0640	0.0479
	0.6	0.3753	0.2365	0.1695	0.0953	0.0503	0.0309	0.0160
	0.7	0.3361	0.1802	0.1098	0.0508	0.0203	0.0081	0.0054
	0.8	0.2659	0.1116	0.0581	0.0203	0.0066	0.0022	0.0004
	0.9	0.1732	0.0627	0.0225	0.0032	0.0004	0.0001	0.0000
	평균	0.4484	0.3326	0.2768	0.2205	0.1831	0.1634	0.1495
U_2	0.0	0.6697	0.6260	0.5938	0.5630	0.5366	0.5193	0.4992
	0.1	0.5038	0.4345	0.4144	0.3761	0.3619	0.3509	0.3598
	0.2	0.4043	0.3234	0.3002	0.2501	0.2470	0.2317	0.2404
	0.3	0.3353	0.2578	0.2217	0.1700	0.1625	0.1460	0.1477
	0.4	0.2856	0.1975	0.1628	0.1126	0.0994	0.0752	0.0704
	0.5	0.2321	0.1512	0.1178	0.0680	0.0544	0.0352	0.0289
	0.6	0.1862	0.1178	0.0804	0.0395	0.0268	0.0147	0.0088
	0.7	0.1624	0.0797	0.0491	0.0170	0.0101	0.0035	0.0024
	0.8	0.1257	0.0463	0.0252	0.0061	0.0030	0.0012	0.0001
	0.9	0.0772	0.0286	0.0086	0.0008	0.0002	0.0001	0.0000
	평균	0.2982	0.2263	0.1974	0.1603	0.1502	0.1378	0.1358
U_3	0.0	0.6734	0.6171	0.5924	0.5663	0.5392	0.5203	0.4992
	0.1	0.4283	0.3407	0.3149	0.2962	0.2952	0.3079	0.3395
	0.2	0.3295	0.2141	0.1822	0.1475	0.1541	0.1669	0.2058
	0.3	0.2608	0.1500	0.1041	0.0759	0.0713	0.0802	0.1111
	0.4	0.2211	0.0999	0.0664	0.0391	0.0329	0.0316	0.0425
	0.5	0.1816	0.0711	0.0424	0.0173	0.0116	0.0102	0.0138
	0.6	0.1392	0.0516	0.0238	0.0077	0.0047	0.0029	0.0033
	0.7	0.1229	0.0318	0.0156	0.0028	0.0005	0.0005	0.0007
	0.8	0.0953	0.0189	0.0065	0.0007	0.0002	0.0000	0.0000
	0.9	0.0598	0.0111	0.0015	0.0001	0.0000	0.0000	0.0000
	평균	0.2512	0.1606	0.1350	0.1154	0.1110	0.1121	0.1216

2. 고정된 N의 값에 대하여 k의 값이 커질수록 음이 될 확률은 감소한다. 이러한 경향은 불균형도가 심할수록 더욱 강하게 나타난다.
3. 고정된 N의 값에 대하여 k의 값이 커질수록 불균형도의 영향을 덜 받는다.
4. 임의의 값 ρ에 대하여 음이 될 확률 p는 불균형도보다 k의 값에 더 민감하다.
5. <표 7>의 자료가 균형인 경우인 U_3을 살펴보면 <표 4>에 제시되어 있는 (N,k)=(36,9)인 경우에 음이 될 확률의 평균값이 가장 작은 것을 확인할 수 있으나, 불균형도가 심해질수록 더 큰 k 값에서 확률의 평균값이 작아지는 것을 알 수 있다.

<표 8>은 이상점이 존재하는 경우에 이상점의 유무가 음이 될 확률에 미치는 영향을 실험계획 (N,k)=(36,2)인 경우에 조사한 결과이다. 총 36개의 관측치중 고려된 이상점의 개수는 1개(2.78%)와 2개(5.56%)이고, 정규분포에서 3배의 표준편차 밖으로 존재하는 점은 거의 존재하지 않기 때문에 이상점은 평균에 3배의 표준편차와 10배의 표준편차를 각각 더한 두 가지 경우를 고려

하였다. 또한 이상점의 발생위치는 반복수가 적은 셀과 반복수가 많은 셀에서 발생하는 두 가지

<표 8> 이상점이 존재하는 경우에 음이 될 확률

크기	개수	위치	ρ	U_1	U_2	U_3	크기	개수	위치	ρ	U_1	U_2	U_3
3	1	y_{11}	0.0	0.6713	0.6731	0.6782	10	1	y_{11}	0.0	0.6763	0.6785	0.6777
			0.1	0.5006	0.5169	0.4479				0.1	0.2265	0.4430	0.4717
			0.2	0.4131	0.4203	0.3451				0.2	0.1485	0.3381	0.3939
			0.3	0.3352	0.3543	0.2798				0.3	0.1184	0.3019	0.3430
			0.4	0.2910	0.3051	0.2381				0.4	0.0986	0.2595	0.3076
			0.5	0.2355	0.2664	0.2059				0.5	0.0779	0.2370	0.2918
			0.6	0.2052	0.2205	0.1741				0.6	0.0667	0.2131	0.2690
			0.7	0.1587	0.1959	0.1495				0.7	0.0484	0.1954	0.2559
			0.8	0.1196	0.1710	0.1362				0.8	0.0412	0.1790	0.2408
			0.9	0.0840	0.1427	0.1117				0.9	0.0263	0.1636	0.2279
		y_{21}	0.0	0.6925	0.6925	0.6764			y_{21}	0.0	0.6776	0.6740	0.6729
			0.1	0.6442	0.5307	0.4345				0.1	0.6851	0.5515	0.4643
			0.2	0.6123	0.4160	0.3454				0.2	0.6819	0.4920	0.3908
			0.3	0.5705	0.3622	0.2859				0.3	0.6768	0.4455	0.3388
			0.4	0.5387	0.3047	0.2366				0.4	0.6612	0.4174	0.3174
			0.5	0.4923	0.2524	0.1976				0.5	0.6474	0.3962	0.2933
			0.6	0.4465	0.2344	0.1818				0.6	0.6467	0.3697	0.2729
			0.7	0.4048	0.2011	0.1479				0.7	0.6124	0.3525	0.2557
			0.8	0.3624	0.1660	0.1309				0.8	0.6028	0.3423	0.2461
			0.9	0.3105	0.1475	0.1047				0.9	0.5866	0.3334	0.2268
	2	y_{11} y_{21}	0.0	0.6876	0.6719	0.6724		2	y_{11} y_{21}	0.0	0.6906	0.6642	0.6739
			0.1	0.5220	0.5102	0.4588				0.1	0.2587	0.4816	0.4985
			0.2	0.4263	0.4369	0.3517				0.2	0.1933	0.4142	0.4309
			0.3	0.3526	0.3714	0.3042				0.3	0.1695	0.3714	0.4025
			0.4	0.3132	0.3214	0.2530				0.4	0.1583	0.3429	0.3731
			0.5	0.2671	0.2899	0.2353				0.5	0.1385	0.3409	0.3704
			0.6	0.2392	0.2484	0.2034				0.6	0.1330	0.3299	0.3563
			0.7	0.2004	0.2364	0.1856				0.7	0.1299	0.3145	0.3520
			0.8	0.1785	0.2181	0.1636				0.8	0.1220	0.3064	0.3432
			0.9	0.1422	0.1876	0.1478				0.9	0.1139	0.2985	0.3414
		y_{21} y_{22}	0.0	0.6715	0.6718	0.6727			y_{21} y_{22}	0.0	0.6771	0.6722	0.6760
			0.1	0.6512	0.5267	0.4546				0.1	0.7286	0.6019	0.5027
			0.2	0.6162	0.4373	0.3552				0.2	0.7567	0.5642	0.4386
			0.3	0.5885	0.3777	0.3050				0.3	0.7708	0.5311	0.4013
			0.4	0.5579	0.3246	0.2546				0.4	0.7740	0.5162	0.3960
			0.5	0.5285	0.2901	0.2302				0.5	0.7740	0.4920	0.3743
			0.6	0.4961	0.2659	0.2003				0.6	0.7755	0.4895	0.3625
			0.7	0.4709	0.2321	0.1903				0.7	0.7759	0.4903	0.3444
			0.8	0.4321	0.2154	0.1661				0.8	0.7796	0.4744	0.3437
			0.9	0.4023	0.1962	0.1444				0.9	0.7682	0.4707	0.3464

경우로 각각 나누어서 살펴보았다. 이상점이 1개 있고, 셀 도수가 작은 곳에서 발생한 y_{11}인 경우에는 <표 8>과 <표 7>의 (36,2)과 비교하여 볼 때 U_1계획에서는 $\rho=0$인 경우를 제외하고 이상점이 없는 경우보다 음이 될 확률이 작아진다. 또한 이상점의 값이 커지고 불균형도가 심해질수록 음이 될 확률은 더 줄어드는데, 이것은 y_{11}의 값이 커짐으로써 MSA의 값을 더 크게 만들 가능성이 많다는 데서 그 원인을 찾아볼 수 있다. 그리고 이상점의 값이 커지면 U_1인 경우에는 음이 될 확률이 이상점의 값이 작은 경우보다 더 작아지나, U_2와 U_3로 갈수록 점점 더 커짐을 확인할 수 있으며, 자료가 균형적인 경우에는 이상점의 크기와 상관없이 이상점이 존재하는 경우가

없는 경우보다 음이 될 확률을 더 크게 만든다고 할 수 있다.

이상점이 1개이고 셀 도수가 많은 곳에서 발생하는 y_{21}인 경우에는 불균형도가 심하고, 이상점의 값이 클수록 셀 도수가 작은 데서 발생하는 y_{11}인 경우보다 음이 될 가능성이 커진다. 이상점이 2개(5.56%)인 경우에는 U_1계획에서 이상점의 위치가 (y_{11}, y_{21})인 경우를 제외하고 모두 음이 될 확률이 이상점이 없는 경우보다 대체적으로 커진다. 이 예외는 이상점이 y_{11}만 존재할 때 음이 될 가능성이 전반적으로 더 작아지는 원인 때문으로 판단되어진다. 아울러 같은 크기의 셀에서 이상점이 발생할 때, 개수가 1개인 경우보다 2개인 경우가 음이 될 확률이 전반적으로 더 커지는 사실을 확인할 수 있으며, 셀 도수가 큰 셀에서 이상점이 많이 발생하는 (y_{21}, y_{22})인 경우에는 급내상관계수의 값이 매우 큰 값임에도 불구하고 U_1계획인 경우에는 음이 될 확률이 77%를 초과한다.

본 연구에는 결과가 나와있지 않지만, 평균에 3배의 표준편차와 10배의 표준편차를 각각 빼서 이상점을 만들고 모의실험을 행하였으나, 모두 비슷한 결과가 발생하는 것을 확인할 수 있었으며, $(N, k) = (36, 2)$대신 다른 실험계획에서도 비슷한 결론에 도달함을 확인할 수 있었다. 이상을 종합하면 이상점이 존재하는 경우에는 이상점의 개수, 발생하는 셀의 위치, 절대값의 크기가 음이 될 확률에 큰 영향을 미친다고 할 수 있다.

4. 결론

본 연구에서는 일원변량모형에서 분산성분 σ_a^2의 분산분석추정량이 음이 될 확률을 구하여 보았다. 음이 될 확률 p는 급내상관계수의 값이 매우 작은 경우, 불균형도가 심한 경우, 총관측치의 개수 N보다 수준의 수 k값이 현저하게 작은 경우, 이상점의 개수가 많고 이상점이 셀 도수가 많은 셀에 모두 존재하는 경우에 매우 커짐을 확인할 수 있었다.

하지만 일반적으로 급내상관계수의 값과 이상점의 발생여부는 알 수 없기 때문에 분산분석추정치가 음이 될 가능성을 줄이기 위해서는 균형자료인 경우에는 <표 4>의 결과를 이용하여 (k, n)의 조합을 고려하고, 불균형자료인 경우에는 <표 4>의 결과보다 더 큰 k 값을 선택하는 것이 바람직하다고 할 수 있다.

참 고 문 헌

[1] Hess, J. L. (1976). Sensitivity of MINQUE with Respect to A Prior Weights, *Biometrics,* 35, 645-649.

[2] Searle, S. R. (1971). *Linear Models,* John Wiley & Sons, New York.