9.1 한 인자 분산분석

8.1절에서는 두 모집단의 평균 비교를 t 검정을 이용하여 하는 방법에 대해 알아보았다. 이 장에서는 이를 확장하여 여러 모집단의 평균을 비교하는 방법을 살펴본다. 현실에서는 여러 모집단의 평균을 비교하는 예가 많다.

- 어느 대학에서 각 학년별 도서관 평균이용시간이 같은가?
- 세 가지 다른 품종의 벼의 수확량이 같은가?
- 어떤 화학반응에서 네 가지 다른 온도에서 반응속도가 같은가?
- 도시별 대졸자의 월평균 임금이 같은가?

학년이나 벼의 품종같이 모집단의 그룹을 구분하는데 사용되는 변량을 **인자**(factor)라고 한다. 이 절에서는 모집단의 그룹이 한 인자인 경우 모평균을 비교하는 **분산분석**(analysis of variance)을 설명하고, 9.2절에서는 표본 데이터를 추출하기 위한 실험의 계획 방법, 9.3절에서는 두 인자의 모평균을 비교하는 분산분석을 설명한다. 다음 예를 가지고 살펴보자.

[예 9.1.1] 어느 대학교에서 학년별 영어 실력을 비교하기 위하여 각 학년에서 무작위로 표본을 추출하여 같은 영어 시험을 보게 한 결과 표 9.1.1의 자료를 얻었다. 우측 열은 각 학년별 평균 $\bar{y}_{1\cdot}$, $\bar{y}_{2\cdot}$, $\bar{y}_{3\cdot}$, $\bar{y}_{4\cdot}$를 계산한 것이다.

표 9.1.1 학년별 영어 성적

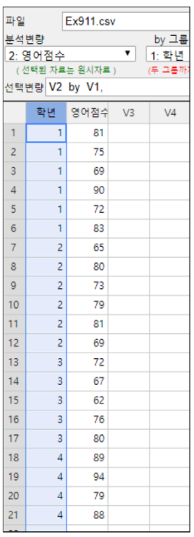
학년	영어 점수						평균
1	81	75	69	90	72	83	$\bar{y}_{1\cdot} = 78.3$
2	65	80	73	79	81	69	$\bar{y}_{2\cdot} = 74.5$
3	72	67	62	76	80		$\bar{y}_{3\cdot} = 71.4$
4	89	94	79	88			$\bar{y}_{4\cdot} = 87.5$

1) 『eStat』을 이용하여 각 학년별 영어 점수의 점그래프를 그리고 평균을 비교하라.
2) 각 학년별 영어 점수의 평균이 같은지 가설검정 하고자 한다. 귀무가설과 대립가설을 설정하라.
3) 문제 2)의 가설을 검정하기 위한 분산분석을 하라.
4) 『eStat』을 이용하여 분산분석 검정 결과를 확인하라.

<풀이>

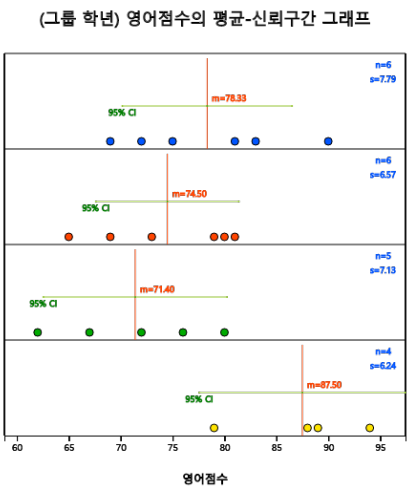
1) 각 학년별 영어점수가 비슷한지 아닌지는 점그래프를 그려 비교해보면 대략 알 수 있다. 만일 점그래프 위에 6장에서 연구한 모평균의 신뢰구간을 표시하면 더 자세한 비교가 가능하다.

표 9.1.1과 같은 데이터를 『eStat』으로 점그래프를 그리기 위해서는 <그림 9.1.1>과 같이 시트에 데이터를 입력하고 변량명을 '학년'과 '영어점수'로 설정한다. 『eStat』 주메뉴의 분산분석 아이콘 을 클릭하여 나타나는 변량선택박스에서 '분석변량'을 영어점수, 'by 그룹'을 학년으로 선택하면 <그림 9.1.2>와 같이 학년별 영어점수의 점그래프와 95% 신뢰구간이 표시된다.


파일 Ex911.csv
분석변량 by 그룹
2: 영어점수 1: 학년
(선택된 자료는 원시자료) (두 그룹까지)
선택변량 V2 by V1,

	학년	영어점수	V3	V4
1	1	81		
2	1	75		
3	1	69		
4	1	90		
5	1	72		
6	1	83		
7	2	65		
8	2	80		
9	2	73		
10	2	79		
11	2	81		
12	2	69		
13	3	72		
14	3	67		
15	3	62		
16	3	76		
17	3	80		
18	4	89		
19	4	94		
20	4	79		
21	4	88		

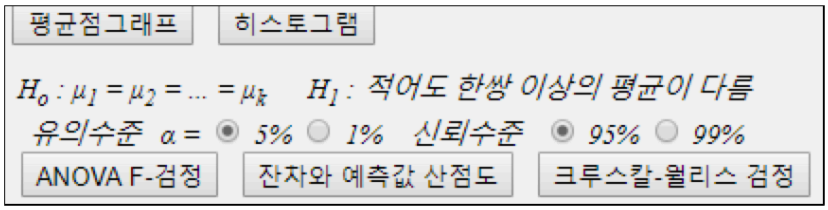
<그림 9.1.1> eStat』 데이터 입력



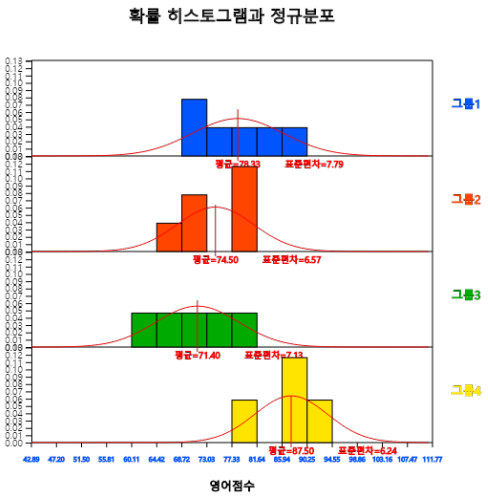
<그림 9.1.2> 영어성적 자료의 점그래프와 평균의 95% 신뢰구간 그래프

데이터의 정규성 여부를 검토하기 위해서는 이 그래프 밑의 선택사항(<그림 9.1.3>)

에서 '히스토그램' 버튼을 누르면 <그림 9.1.4>와 같이 히스토그램과 정규분포가 같이 그려진다.



<그림 9.1.3> 분산분석 선택사항



<그림 9.1.4> 각 학년별 히스토그램과 정규분포 그림

<그림 9.1.2>에서 표본평균들의 값 $\bar{y}_{1.} = 78.3$, $\bar{y}_{2.} = 74.5$, $\bar{y}_{3.} = 71.4$, $\bar{y}_{4.} = 87.5$을 비교해보면 네 번째 모집단에서 추출된 값들이 상대적으로 크다는 것을 알 수 있다. 영어점수 평균은 $\bar{y}_{3.} < \bar{y}_{2.} < \bar{y}_{1.} < \bar{y}_{4.}$ 의 순서이며, $\bar{y}_{2.}$ 와 $\bar{y}_{3.}$ 는 서로 비슷하지만 $\bar{y}_{4.}$ 는 다른 세 값보다 크다고 할 수 있다. 따라서 μ_2와 μ_3는 같을 것이고, μ_4는 다른 세 값과 차이가 있을 것이라고 예상할 수 있다. 하지만 우리는 표본에 의해 나타나는 이 차이가 과연 통계학적으로 유의한 것인지 판단할 수 있는 객관적 검정방법이 필요하게 된다.

2) 이 예에서 검정의 대상이 되는 귀무가설은 네 학년의 영어성적 평균이 모두 동일하다는 것이며, 대립가설은 영어성적 평균이 모두 동일한 것은 아니라는 것이다. 즉, μ_i $(i = 1,2,3,4)$를 각 학년 전체의 영어성적 평균이라고 하면 가설은 다음과 같이 설정된다.

귀무가설 $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

대립가설 H_1 : 적어도 한 쌍의 μ_i 들 값은 서로 같지 않다.

3) 여러 개의 표본평균값이 차이가 있는지 검정하는 기준으로 먼저 생각할 수 있는 측도는 전체평균까지의 거리일 것이다. 즉, 21명 전체의 표본평균을 $\bar{y}..$ 로 나타내면 각각의 표본평균에서 전체평균까지의 거리는 표본의 개수를 가중치로 하였을 때 다음과 같다.

$$\text{SSTr} = 6(78.3 - \bar{y}..)^2 + 6(74.5 - \bar{y}..)^2 + 5(71.4 - \bar{y}..)^2 + 4(87.5 - \bar{y}..)^2 = 643.633$$

이 거리를 급간제곱합(between sum of squares; SSB) 또는 처리제곱합(treatment sum of squares; SSTr)이라고 하는데 만일 이 거리가 0 에 가깝다면 4개 학년의 영어성적 평균은 서로 유사하다고 볼 수 있다.

그러나 이 급간제곱합은 모집단의 개수가 늘어나면 더욱 커질 수가 있어 여러 개의 모집단의 평균이 같은지 알아보기 위한 검정통계량이 되기 위해서는 수정을 필요로 한다. 각각의 관찰값에서 이 관찰값이 속한 학년의 표본평균까지의 거리를 **급내제곱합**(within sum of squares; SSW) 또는 **오차제곱합**(error sum of squares; SSE)이라 하는데 다음과 같이 정의된다.

$$\begin{aligned}\text{SSE} = &(81 - \bar{y}_1.)^2 + (75 - \bar{y}_1.)^2 + \ldots + (83 - \bar{y}_1.)^2 \\ &+ (65 - \bar{y}_2.)^2 + (80 - \bar{y}_2.)^2 + \ldots + (69 - \bar{y}_2.)^2 \\ &+ (72 - \bar{y}_3.)^2 + (67 - \bar{y}_3.)^2 + \ldots + (80 - \bar{y}_3.)^2 \\ &+ (89 - \bar{y}_4.)^2 + (94 - \bar{y}_4.)^2 + \ldots + (88 - \bar{y}_4.)^2 = 839.033\end{aligned}$$

만일 각 학년 모집단의 영어성적분포가 정규분포를 따르고 네 학년의 성적분산이 같다고 가정할 수 있다면 검정통계량

$$F_0 = \frac{\frac{SSTr}{(4-1)}}{\frac{SSE}{(21-4)}}$$

은 $F_{3,17}$ 분포를 따른다. 이것을 이용하여 4개 학년의 영어성적이 같다는 검정을 할 수 있다. 검정통계량에서 분자 $\text{SSTr}/(4-1)$ 은 **처리평균제곱**(treatment mean square, MSTr)이라고 하고, 분모 $\text{SSE}/(21-4)$ 를 **오차평균제곱**(error mean square, MSE)라고 부르는데, 일종의 분산으로 학년간의 분산과 학년내의 분산을 의미한다. 따라서 위의 검정통계량은 두 분산의 비를 구한 것으로서, 여러 개의 모평균의 검정을 **분산분석**(analysis of variance; ANOVA)이라고 부르는 이유가 여기에 있다.

각 학년별 표본의 영어성적을 가지고 검정통계량을 계산하여 보면 다음과 같다.

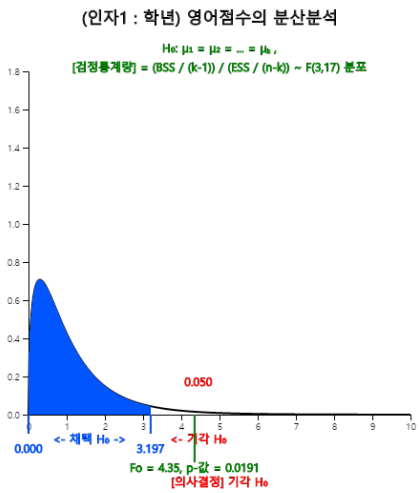
$$F_0 = \frac{\frac{SSTr}{(4-1)}}{\frac{SSE}{(21-4)}} = \frac{\frac{643.633}{3}}{\frac{839.033}{17}} = 4.347$$

$F_{3,17;0.05}$ = 3.20 이므로 표본에서 보이는 학년별 영어성적 평균의 차이가 없다는 귀무가설 $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ 은 5% 유의수준에서 기각된다. 즉, 학년별 영어성적 평균에는 차이가 있다.

위와 같은 계산을 한 눈에 볼 수 있게 만든 것이 아래의 **분산분석표**(ANOVA table)이다.

요인	제곱합	자유도	평균제곱	F 값
처리	$SSTr = 643.633$	4-1	$MSTr = 643.633/3$	$F_0 = 4.347$
오차	$SSE = 839.033$	21-4	$MSE = 839.033/17$	
전체	$SST = 1482.666$	20		

4) <그림 9.1.3>의 분산분석 선택사항에서 유의수준 5%와 신뢰수준 95%를 선택하고 'ANOVA F 검정' 버튼을 누르면 <그림 9.1.5>와 같이 F분포에서 검정통계량의 위치를 보여주는 그래프가 나타나고, 로그창에는 <그림 9.1.6>과 같은 각 학년별 평균 및 신뢰구간 표와 검정결과가 나타난다.



<그림 9.1.5> 『eStat』의 ANOVA F 검정 그래프

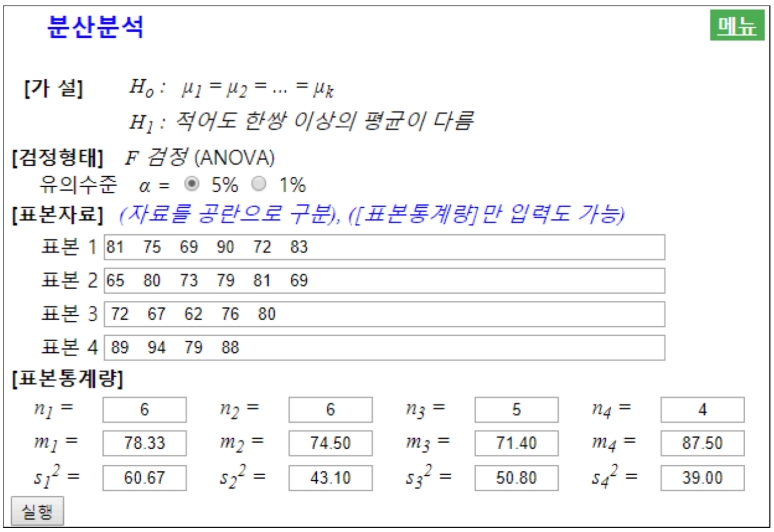
통계량	분석변량	영어점수	그룹명	학년		
그룹변량 (학년)	자료수	평균	표준편차	표준오차	모평균 95% 신뢰구간	모분산 95% 신뢰구간
1 ()	6	78.333	7.789	3.180	(70.159, 86.507)	(23.638, 364.929)
2 ()	6	74.500	6.565	2.680	(67.610, 81.390)	(16.793, 259.260)
3 ()	5	71.400	7.127	3.187	(62.550, 80.250)	(18.235, 419.472)
4 ()	4	87.500	6.245	3.122	(77.563, 97.437)	(12.516, 542.181)
전체	21	77.333	8.610	1.879	(73.414, 81.253)	(43.391, 154.593)
결측수	0					

분산분석

요인	제곱합	자유도	평균제곱	F 값	p 값
처리	643.633	3	214.544	4.347	0.0191
오차	839.033	17	49.355		
전체	1482.667	20			

<그림 9.1.6> 『eStat』의 기초통계량과 분산분석표

『eStatU』에서도 분산분석이 가능하다. 데이터를 입력하고(<그림 9.1.7>) '실행' 버튼을 누르면 위의 <그림 9.1.5>와 동일한 결과를 얻을 수 있다.



<그림 9.1.7>『eStatU』분산분석 자료 입력

위의 예제에서는 두 변량이 언급되어 있는데 영어점수와 같이 실제로 그 값이 관측되는 변량을 **분석변량** 또는 **반응변량**(response variable)이라 한다. 반응변량은 대부분 연속형 변량이다. 그리고 학년과 같이 모집단의 그룹을 구분하는데 사용되는 변량을 **그룹** 또는 **인자**(factor)라고 하는데 대부분 이산형 변량이다. 인자가 가지는 값들을 인자의 **수준**(level)이라고 하며, 이들 수준의 개수가 비교되는 모집단의 개수가 된다. 위의 영어성적 예제에서는 학년이라는 하나의 인자가 네 개의 수준 1, 2, 3, 4를 가진다. '반

응' 또는 '인자'라는 용어는 분산분석이 공학, 농학, 의학, 약학 등에서 실험을 통한 자료를 분석하는데 많이 이용되기 때문이다. 이와 같이 하나의 인자가 반응변량에 미치는 영향을 조사하는 분산분석법을 **일원분산분석**(one-way ANOVA)이라 한다. 표 9.1.2는 일원분산분석에서 수준의 개수가 k 개이고 각 수준에서의 관측값이 각각 $n_1, n_2, \cdots, n_k$ 개 있는 경우의 일반적인 자료구조를 보여준다.

표 9.1.2 일원분산분석의 자료구조와 기호

인자	관측값	평균
수준 1	$Y_{11}\ Y_{12}\ \cdots\ Y_{1n_1}$	$\bar{Y}_{1\cdot}$
수준 2	$Y_{21}\ Y_{22}\ \cdots\ Y_{2n_1}$	$\bar{Y}_{2\cdot}$
수준 k	$Y_{k1}\ Y_{k2}\ \cdots\ Y_{kn_1}$	$\bar{Y}_{k\cdot}$

일원분산분석을 위한 통계모형은 다음과 같이 주어진다.

$$\begin{aligned} Y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij},\ i = 1,2,\cdots,k;\ j = 1,2,\cdots,n_i \end{aligned} \tag{9-1}$$

위에서 Y_{ij} 는 반응변량 Y 의 i 번째 수준에서의 j 번째 관측값을 나타낸다. 이 모형에서는 i 번째 수준에서의 Y 의 모평균 μ_i 를 $\mu + \alpha_i$ 로 나타내었는데 여기에서 μ 는 Y 의 전체 모평균을 나타내며, α_i 는 $(\mu_i - \mu)$ 로 이를 반응변량에 대한 i 번째 수준의 **효과**(effect)라 한다. 오차항 ε_{ij} 는 서로 독립이며, 평균이 0 이고 분산이 σ^2 인 정규분포를 따른다고 가정한다.

오차항 ε_{ij} 는 수준간의 차이가 아닌 다른 요인에 기인하는 반응변량의 변동을 나타내는 확률변량이다. 예를 들어, 영어성적 예제에서 각 학년의 영어성적 차이는 학년이라는 변량 외에 다른 변량들, 이를테면, 개인별 영어 공부시간, 성별, IQ 등에 의해서도 발생할 수 있다. 그러나 이들 변동은 학년의 차이에 의한 변동에 비해 상대적으로 작다고 가정함으로써 다른 요인들을 일종의 오차들로 간주할 수 있는데, 오차항 ε_{ij} 는 이와 같은 여러 가지 오차들의 합으로 해석될 수 있다.

검정하고자하는 가설은 μ_i 대신에 α_i 를 사용하면 아래와 같다.

귀무가설 $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ (9-2)

대립가설 H_1 : 적어도 하나의 α_i 는 0 이 아니다. (9-3)

이 가설의 검정을 위해서는 표 9.1.3과 같은 **분산분석표**(ANOVA table)가 사용된다.

표 9.1.3 일원분산분석의 분산분석표

요인	제곱합	자유도	평균제곱	F 값
처리	$SSTr$	$k-1$	$MSTr = SSTr/(k-1)$	$F_0 = MSTr/MSE$

표 9.1.3 일원분산분석의 분산분석표

오차	SSE	$n-k$	$MSE = SSE/(n-k)$
전체	SST	$n-1$	

(여기서, $n = \sum_{i=1}^{k} n_i$)

분산분석의 세 가지 제곱합은 다음과 같이 설명될 수 있다. 설명을 위해 먼저 다음의 통계량을 정의하자.

$\bar{Y}_{i\cdot} = Y$ 의 i 번째 수준에서의 관측값들의 평균
$\bar{Y}_{\cdot\cdot} = Y$ 의 전체 관측값들의 총평균

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 \quad : \tag{9-4}$$

반응변량 Y 의 관측값들과 총평균 사이의 거리 제곱합으로 이를 총변동 또는 **총제곱합**(total sum of squares, SST)이라 한다.

$$\text{SSTr} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \quad : \tag{9-5}$$

각 수준간의 차이에 의해 발생하는 변동을 나타낸다. Y 의 i 번째 수준에서의 관측값들의 평균 $\bar{Y}_{i\cdot}$ 는 각 수준 자체내의 변동을 평균화한 것으로 그 수준의 대푯값이라 할 수 있다. 그러므로, 개개의 관측값 대신에 이 표본평균을 사용하여 총변동을 구하면(즉, 총제곱합을 구하는 공식에서 Y_{ij} 대신에 $\bar{Y}_{i\cdot}$ 를 대입하면), 이는 수준간의 차이에 의한 변동을 나타낸다. 이와 같은 수준간의 변동을 급간변동(between variation)이라 하며 이 변동을 나타내는 제곱합을 **처리제곱합**(treatment sum of squares, $SSTr$)이라 한다.

$$\text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2 \quad : \tag{9-6}$$

각 수준 내에서 발생하는 변동들의 합을 나타낸다. 각 수준 내의 변동을 급내변동(within variation)이라 하며, 이 급내변동을 나타내는 제곱합을 **오차제곱합**(error sum of squares, SSE)이라 한다.

각 제곱합이 가지는 자유도는 다음과 같은 논리에 의해 구해진다. 즉, SST 를 계산하기 위해서는 n 개의 Y_{ij} 값이 있지만, 먼저 전체평균의 추정량인 $\bar{Y}_{\cdot\cdot}$ 을 계산해야하므로 SST 는 자유도 $(n-1)$ 을 가지며, 오차제곱합 SSE 의 계산을 위해서는 k 개의 값 $\bar{Y}_{1\cdot}, \cdots, \bar{Y}_{k\cdot}$ 이 먼저 계산되므로 SSE 는 $(n-k)$ 의 자유도, 처리제곱합 SSTr 은 SST 의 자유도에서 SSE 의 자유도를 뺀 나머지 $(k-1)$ 의 자유도를 가진다. 일원분산분석에서는 항상 다음 사실이 성립한다.

제곱합과 자유도의 분할:

제곱합 : $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSTr}$ (9-7)

자유도 : $(n-1) = (n-k) + (k-1)$ (9-8)

제곱합을 자유도로 나눈 값을 평균제곱(mean square)이라 하는데 표 9.1.3에는 **처리평균제곱**(treatment mean square, MSTr)과 오차평균제곱(error mean square, MSE)이 각각 정의되어 있다. 제곱합의 의미와 같이 처리평균제곱 MSTr은 인자의 각 수준간의 변동을 나타내는 평균제곱이고, 오차평균제곱 MSE는 각 수준내의 변동을 나타내는 평균제곱이다. 그러므로, MSTr이 MSE보다 상대적으로 훨씬 크면 인자의 각 수준간의 평균값, 즉, μ_i들이 같지 않다고 결론을 내릴 수 있다. 그러면 어떠한 기준에 의하여 상대적으로 훨씬 크다고 할 수 있을까?

분산분석표의 마지막 열에 계산된 F값이 MSTr의 MSE에 대한 상대적 크기를 나타내는데, 통계이론에 의하면 ε_{ij}에 대한 가정이 성립할 때, 귀무가설 $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ 의 가정 하에서

$$F_0 = \frac{\mathrm{MSTr}}{\mathrm{MSE}} = \frac{SSTr/(k-1)}{SSE/(n-k)} \tag{9-9}$$

는 자유도 $(k-1)$과 $(n-k)$인 F 분포를 따른다. 그러므로 유의수준 α인 검정에서, 계산된 F_0의 값이 F 분포에서 구한 값 $F_{k-1,n-k;\alpha}$보다 크면 MSTr이 MSE보다 충분히 크다고 판정하여 귀무가설 H_0를 기각한다. 즉, 인자의 각 수준 간 평균들이 모두 같은 것은 아니라고 결정한다.

일원분산분석에서의 F 검정:

귀무가설 $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

대립가설 H_1 : 적어도 한 α_1는 0 이 아니다.

검정통계량: $F_0 = \frac{\mathrm{MSTr}}{\mathrm{MSE}}$

기각역: $F_0 > F_{k-1,n-k;\alpha}$ 이면 H_0를 기각

(참고: 『eStat』에서는 이 검정에 대한 p-값을 계산하여 주므로 계산된 p-값이 사용자가 사전에 정한 유의수준보다 작으면 귀무가설을 기각한다.)

9.1.1 다중비교

평균들에 대한 F 검정에서 인자의 각 수준간의 차이가 나타나지 않는 경우에는 차이

가 없다고 결론을 내리거나, 아니면 수준간의 차이를 찾기 위한 다른 실험을 시도 할 수 있다. 그러나 [예 9.1.1]에서와 같이 각 수준간의 차이가 있다는 결론을 얻은 경우에는 그 차이가 어느 수준간에서 나타나는지 살펴볼 필요가 있다.

분산분석 후의 평균 차이에 관한 분석에서는 평균차의 검정을 두 개 이상 동시에 하여야 하는데, 이때 사용하는 통계적 검정방법을 **다중비교**(multiple comparison)라 하며 여러 가지 다중비교방법들 중에서 Tukey의 HSD검정(honestly significant difference test)이 가장 많이 이용된다. 평균 μ_i와 μ_j에 대한 Tukey의 HSD검정에서는 먼저

$$\mathrm{HSD}_{ij} = q_{k,n-k;\alpha} \cdot \sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\mathrm{MSE}} \tag{9-10}$$

를 계산하여 표본평균들이 부등식

$$|\bar{y}_i - \bar{y}_j| > HSD_{ij} \tag{9-11}$$

를 만족하면 가설 $H_0 : \mu_i = \mu_j$를 기각하여 μ_i와 μ_j는 유의수준 α하에서 다르다고 결정한다. 위의 식에서 n_i와 n_j는 i와 j 수준에서의 반복수, MSE는 오차평균제곱이다. 그리고 $q_{k,n-k;\alpha}$는 모수 k와 자유도 $n-k$를 갖는 표준화범위분포(studentized range distribution)의 우측 $100 \times \alpha\%$ 백분위수로 『eStatU』에 수록되어 있다(<그림 9.1.8>).

HSD 표준화범위분포 메뉴

α = 5% 1%

백분위수표 테이블 저장

HSD 표준화범위분포	df1 =								
P(X ≥ x) = 0.05	2	3	4	5	6	7	8	9	10
df2 = 1	17.970	26.980	32.820	37.080	40.410	43.120	45.400	47.360	49.070
df2 = 2	6.080	8.330	9.800	10.880	11.740	12.440	13.030	13.540	13.990
df2 = 3	4.500	5.910	6.820	7.500	8.040	8.480	8.850	9.180	9.460
df2 = 4	3.930	5.040	5.760	6.290	6.710	7.050	7.350	7.600	7.830
df2 = 5	3.640	4.600	5.220	5.670	6.030	6.330	6.580	6.800	6.990
df2 = 6	3.460	4.340	4.900	5.300	5.630	5.900	6.120	6.320	6.490
df2 = 7	3.340	4.160	4.680	5.060	5.360	5.610	5.820	6.000	6.160
df2 = 8	3.260	4.040	4.530	4.890	5.170	5.400	5.600	5.770	5.920
df2 = 9	3.200	3.950	4.410	4.760	5.020	5.240	5.430	5.590	5.740
df2 = 10	3.150	3.880	4.330	4.650	4.910	5.120	5.300	5.460	5.600

<그림 9.1.8>『eStatU』HSD 분포표

[예 9.1.2] 학년별 영어 성적자료에 대한 분산분석에서는 귀무가설을 기각하여 각 학년의 영어평균성적이 모두 같은 것은 아니라는 결론을 얻었다. 이제 차이가 어느 학년간에 존재하는지를 유의수준 5%로 각 학년간의 평균을 다중비교하여 밝혀보자.

『eStat』을 이용하여 결과를 확인하자.

<풀이>

이 문제에서 검정하고자하는 가설은 $H_0 : \mu_i = \mu_j \quad H_1 : \mu_i \neq \mu_j$ 이고 선택기준은

$|\bar{y}_i - \bar{y}_j| > HSD_{ij}$ 이면 H_0 기각

이다. 4개 학년이 있으므로 6쌍의 평균에 대한 다중비교가 가능한데 각 검정을 하여 보면 다음과 같다. 표준화범위의 우측 5% 백분위수는 다음과 같다.

$q_{k,n-k;\alpha} = q_{4,21-4;0.05} = 4.02$

1) $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$

$|\bar{y}_1 - \bar{y}_2| = |78.3 - 74.5| = 3.8$

$$\mathrm{HSD}_{12} = q_{k,n-k;0.05} \cdot \sqrt{\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathrm{MSE}}$$
$$= q_{4,21-4;0.05} \cdot \sqrt{\frac{1}{2}\left(\frac{1}{6} + \frac{1}{6}\right)49.355} = 11.530$$

따라서 H_0 채택

2) $H_0 : \mu_1 = \mu_3 \quad H_1 : \mu_1 \neq \mu_3$

$|\bar{y}_1 - \bar{y}_3| = |78.3 - 71.4| = 6.9$

$$\mathrm{HSD}_{13} = q_{k,n-k;0.05} \cdot \sqrt{\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_3}\right)\mathrm{MSE}}$$
$$= q_{4,21-4;0.05} \cdot \sqrt{\frac{1}{2}\left(\frac{1}{6} + \frac{1}{5}\right)49.355} = 12.092$$

따라서 H_0 채택

3) $H_0 : \mu_1 = \mu_4 \quad H_1 : \mu_1 \neq \mu_4$

$|\bar{y}_1 - \bar{y}_4| = |78.3 - 88.5| = 10.2$

$$\mathrm{HSD}_{14} = q_{k,n-k;0.05} \cdot \sqrt{\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_4}\right)\mathrm{MSE}}$$
$$= q_{4,21-4;0.05} \cdot \sqrt{\frac{1}{2}\left(\frac{1}{6} + \frac{1}{4}\right)49.355} = 12.891$$

따라서 H_0 채택

4) $H_0 : \mu_2 = \mu_3 \quad H_1 : \mu_2 \neq \mu_3$

$|\bar{y}_2 - \bar{y}_3| = |74.5 - 71.4| = 3.1$

$$\mathrm{HSD}_{23} = q_{k,n-k;0.05} \cdot \sqrt{\frac{1}{2}(\frac{1}{n_2}+\frac{1}{n_3})\mathrm{MSE}}$$
$$= q_{4,21-4;0.05} \cdot \sqrt{\frac{1}{2}(\frac{1}{6}+\frac{1}{5})49.355} = 12.092$$

따라서 H_0 채택

5) $H_0 : \mu_2 = \mu_4 \quad H_1 : \mu_2 \neq \mu_4$

$$|\bar{y}_2 - \bar{y}_4| = |74.5 - 88.5| = 14$$
$$\mathrm{HSD}_{24} = q_{k,n-k;0.05} \cdot \sqrt{\frac{1}{2}(\frac{1}{n_2}+\frac{1}{n_4})\mathrm{MSE}}$$
$$= q_{4,21-4;0.05} \cdot \sqrt{\frac{1}{2}(\frac{1}{6}+\frac{1}{4})49.355} = 12.891$$

따라서 H_0 기각

6) $H_0 : \mu_3 = \mu_4 \quad H_1 : \mu_3 \neq \mu_4$

$$|\bar{y}_3 - \bar{y}_4| = |71.4 - 88.5| = 17.1$$
$$\mathrm{HSD}_{34} = q_{k,n-k;0.05} \cdot \sqrt{\frac{1}{2}(\frac{1}{n_3}+\frac{1}{n_4})\mathrm{MSE}}$$
$$= q_{4,21-4;0.05} \cdot \sqrt{\frac{1}{2}(\frac{1}{5}+\frac{1}{4})49.355} = 13.396$$

따라서 H_0 기각

위의 다중비교 결과는 <그림 9.1.1>의 평균-점그래프에서 짐작할 수 있었듯이 μ_3와 μ_4 간, μ_2와 μ_4 간에 차이가 존재한다는 것을 알 수 있다. 그리고 μ_1은 다른 평균들과 유의한 차이를 보이지 않는다.

<그림 9.1.3>의 분산분석 선택사항 창에서 '다중비교'를 선택하면 <그림 9.1.9>와 같은 Tukey의 HSD 다중비교 결과를 보여준다. 『eStat』에서는 표본평균 조합에 대한 평균차와 95% HSD 값을 보여준다. 이어서 표본평균들의 크기 순서대로 각 수준을 행과 열에 나열하여 두 수준들의 조합에 대해 HSD 검정 결과가 유의수준 5%로 유의하면 * 표시를 하고, 유의수준 1%로 유의하면 ** 표시, 유의하지 않으면 빈칸으로 남겨둔다.

다중비교	분석변량	(영어점수)	그룹명	(학년)
평균차 (95%HSD)	1 (그룹1) 78.33	2 (그룹2) 74.50	3 (그룹3) 71.40	4 (그룹4) 87.50
1 (그룹1) 78.33		3.83 (11.53)	6.93 (12.09)	9.17 (12.89)
2 (그룹2) 74.50	3.83 (11.53)		3.10 (12.09)	13.00 (12.89)
3 (그룹3) 71.40	6.93 (12.09)	3.10 (12.09)		16.10 (13.40)
4 (그룹4) 87.50	9.17 (12.89)	13.00 (12.89)	16.10 (13.40)	

평균차검정 * 95%, ** 99%	3 (그룹3) 71.40	2 (그룹2) 74.50	1 (그룹1) 78.33	4 (그룹4) 87.50
3 (그룹3) 71.40				*
2 (그룹2) 74.50				*
1 (그룹1) 78.33				
4 (그룹4) 87.50	*	*		

<그림 9.1.9> 『eStat』에서 일원분산분석 다중비교 HSD검정 결과표

평균 차이의 분석을 위해서는 각 수준별 평균에 대한 신뢰구간을 이용할 수도 있다. <그림 9.1.2>는 각 수준별로 평균에 대한 95% 신뢰구간을 보여주고 있다. 이 신뢰구간은 6장에서 설명된 공식을 이용하여 작성되며, 단지 차이점은 오차의 분산 σ^2 의 추정치로 각 수준에서 관측된 값들의 표본분산이 아닌 전체 관측치를 이용하는 공통분산(pooled variance)를 사용한다는 것이다. 분산분석에서는 MSE 가 공통분산이 된다.

이들 신뢰구간을 이용하는 사후분석에서는 신뢰구간이 중복되지 않으면 평균차이가 있다고 볼 수 있으므로 앞의 HSD 검정과 같은 결론을 얻을 수 있다.

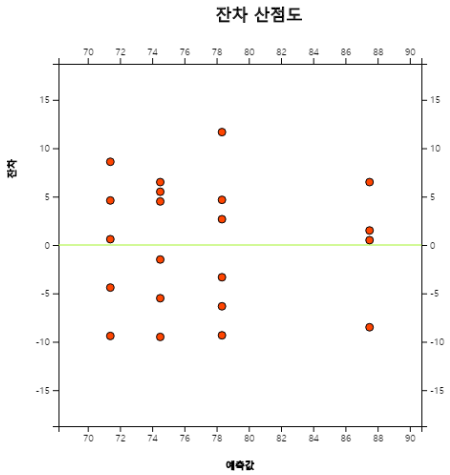
9.1.2 잔차분석

분산분석과 관련된 또 하나의 통계적 분석은 **잔차분석**(residual analysis)이다. 분산분석에서의 각종 가설검정은 오차항 ε_{ij} 에 대한 가정이 성립한다는 조건하에서 이루어진다. 오차항에 대한 가정에는 독립성(ε_{ij} 들은 서로 독립), 등분산성(각 ε_{ij} 의 분산은 σ^2 으로 일정), 정규성(각 ε_{ij} 는 정규분포) 등이 있는데, 분석의 정확성을 위해 항상 이들 가정들의 타당성을 조사하여야 한다. 그러나 ε_{ij} 는 관측될 수 없기 때문에 이들의 추정치인 잔차를 계산하여 이용한다. 분산분석에서의 잔차는 항상 오차제곱합의 공식에 사용되는 편차로 정의되는데, 이를테면, 일원분산분석에서는 $(Y_{ij} - \bar{Y}_{i\cdot})$ 이다.

[예 9.1.3] [예 9.1.1]의 학년별 영어성적 비교에 대한 잔차분석을 하라.

<풀이>

<그림 9.1.3>의 분산분석 선택사항 창에서 '잔차와 예측값 산점도'를 클릭하면 <그림 9.1.10>과 같은 잔차 대 예측치의 산점도가 나타난다. 이 산점도에서는 잔차들이 0을 중심으로 특이한 경향을 보이지 않고 랜덤하게 나타나면 각 가정이 타당함을 의미한다. 이 산점도에서는 특이한 경향이 나타나지 않는다.



<그림 9.1.10> 영어성적 예제에 대한 잔차분석

9.2 표본추출을 위한 실험계획

[예 9.1.1]의 학년별 영어 성적과 같은 데이터는 각 학년 모집단에 대한 자료를 얻기가 가능하고, 여기서 학년별로 여러 개의 표본을 임의 추출하는 것은 그리 어렵지 않다. 그러나 공학이나 의학, 농학 등의 실험을 통해 각 인자의 수준별로 표본 데이터를 얻을 때 실험 시간이 길 수도 있고 다른 외부 요인의 영향을 받아 표본의 수를 많이 추출하기 어려운 경우가 많아 표본의 추출에 주의하여야 한다.

9.2.1 완전확률화계획법

한 인자의 각 수준 간에 존재할 수 있는 차이를 정확하게 파악하기 위해서는 다른 인자들의 영향을 될 수 있는 대로 적게 해 주는 것이 좋다. 이를 위한 방법 중의 하나는 실험전체를 랜덤(random)하게 하는 것이다. 예를 들어, 세 종류의 자동차(A, B, C)의 1리터당 주행거리(연비)를 비교하는 문제를 생각해 보자. 각 종류의 자동차 5대에 대해 연비를 측정하고자 한다. 한 운전자가 모든 15대의 차를 운전해 실험해 볼 수도 있다. 하지만 하루에 5대밖에 측정할 수 없다면 총 3일에 걸쳐서 측정을 하게 된다. 이 경우 매일 날씨나 풍속, 풍향 등이 달라 질 수 있어 측정된 값이 날씨에 영향을 받게 되고 각 날짜에 어느 자동차 종류의 연비를 측정해야 되는지 문제가 된다.

하루에 모든 차의 연비를 측정하기 위하여 다섯 명의 운전자(1, 2, 3, 4, 5)가 차를 운전하려고 계획을 세웠다면 자동차의 연비는 운전자에 따라 영향을 받을 수 있는 문제가 발생한다. 한 가지 해결 방법은 15대의 차를 5명의 운전자에게 랜덤하게 3대씩 배정한 후 실험의 순서 역시 랜덤하게 하면 된다. 즉, 각 차에 1번부터 15번까지의 번호를 부여한 다음, 제비뽑기나 난수표를 사용하여 나오는 번호순서대로 연비측정 실험을 한다. 이와 같이 실험하면 운전자, 풍속, 풍향 등 외부 요인에 의한 변동이 전체 관측값에 균등하게 영향을 미치어 이 요인들로 인하여 인자의 수준 간에 차이가 발생할 가능성이 줄어든다. 이와 같은 실험방법을 **완전확률화계획법**(completely randomized design)이라 부른다. 표 9.1.4는 이 방법에 의한 배정의 한 예를 보여준다. 기호 A, B, C는 세 회사 자동차를 나타낸다.

표 9.2.1 완전랜덤 실험의 예

운전자	1	2	3	4	5
자동차	B	A	B	C	A
	B	C	A	A	C
	C	B	A	B	C

일반적으로 분산분석을 이용하는 문제에 있어서는 분석의 목적을 제대로 달성하기 위하여 자료를 얻기 위한 실험을 사전에 철저히 계획할 필요가 있는데, 위의 완전확률화계획법과 같은 실험방법은 통계학의 한 분야인 **실험계획법**(design of experiment)에서 다룬다. 실험계획의 관점에서 볼 때 일원분산분석법을 **일원배치법**(oneway layout)이라고 부른다.

9.2.2 확률화블록계획법

위에서 예로 들은 자동차 연비실험에서는 완전확률화계획법으로 15대의 자동차를 5명의 운전자에게 랜덤하게 배정하였다. 하지만 표 9.1.4의 한 배정 결과는 이와 같은 완전확률화계획법의 단점을 보여준다. 이를 테면, 운전자 1은 B와 C 회사차만, 운전자 3은 A와 B 회사차만 실험하게 되어 운전자간의 변동이 오차항에 평균화되어 포함되지를 못한다. 그러므로 운전자간의 변동이 심한 경우 오차항은 단순한 실험오차가 아닐 수 있다. 이러한 단점을 없애기 위해 각 운전자가 각 회사차를 적어도 한번씩 실험하도록 할 수 있는데 이와 같은 실험방법을 **확률화블록계획법**(randomized block design)이라 한다. 표 9.2.2는 이 경우 가능한 배정의 한 예를 보여준다. 이 표에서 괄호안의 값은 관측된 연비의 값이다.

표 9.2.2 확률화블록계획법의 예

운전자	1	2	3	5	5

표 9.2.2 확률화블록계획법의 예

연비	A(22.4)	B(12.6)	C(18.7)	A(21.1)	A(24.5)
	C(20.2)	C(15.2)	A(19.7)	B(17.8)	C(23.8)
	B(16.3)	A(16.1)	B(15.9)	C(18.9)	B(21.0)

표 9.2.2에서 전체 관측값들을 운전자에 따라 5개의 집합으로 나누는 것과 같이 동일한 성질을 갖도록 나눈 것을 **블록**(block)이라 하고, 운전자와 같이 블록을 나타내는 변량을 **블록변량**(block variable)이라 한다. 블록은 일반적으로 인자 외의 다른 요인에 의한 변동이 심할 경우 사용된다. 예를 들어, 벼품종에 따른 수확량을 조사할 때 실험에 사용되는 논의 지력이 일정하지 않은 경우에는 우선 지력이 일정하도록 몇 개의 구획(블록)을 만든 후 각 구획에 모든 품종의 벼를 심는다. 이와 같이 하면 지력의 차가 심한 경우에도 그로 인한 변동을 제거할 수 있으므로 벼품종간의 수확량 차이에 대해 더욱 정확하게 조사할 수 있다.

확률화블록계획법에서 블록이 b개 있는 경우의 통계적 모형은 다음과 같이 나타낼 수 있다.

$$Y_{ij} = \mu + \alpha_i + B_j + \varepsilon_{ij} \quad i = 1,2,\cdots,k \tag{9-12}$$

이 식에서 B_j 는 블록변량의 j 번째 수준의 반응변량에 대한 영향력을 나타낸다. 확률화블록계획법에서는 블록변량의 각 수준간의 차이에 기인하는 변동을 인자의 변동과는 독립적으로 오차항에서 분리시킬 수 있다. 이 계획법에서의 총편차는 다음과 같이 분할된다.

$$Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{j} + \bar{Y}_{..}) + (\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{j} - \bar{Y}_{..}) \tag{9-13}$$

위 식의 양변을 제곱한 다음 모든 i,j에 대해 합하면 일원분산분석에서와 같이 여러 가지의 제곱합을 얻을 수 있다. 즉,

$$\mathrm{SST} = \sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \bar{Y}_{..})^2 \quad : \text{총제곱합, 자유도} \, bk-1 \tag{9-14}$$

$$\mathrm{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{b}(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{j} + \bar{Y}_{..})^2 \quad : \text{오차제곱합,자유도} \ (b-1)(k-1) \tag{9-15}$$

$$\mathrm{SSTr} = \sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{Y}_{i.} - \bar{Y}_{..})^2 = b\sum_{i=1}^{k}(\bar{Y}_{i.} - \bar{Y}_{..})^2 \quad : \text{처리제곱합, 자유도} \ k-1 \tag{9-16}$$

$$\mathrm{SSB} = \sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{Y}_{j} - \bar{Y}_{\cdot\cdot})^2 = k\sum_{j=1}^{b}(\bar{Y}_{j} - \bar{Y}_{\cdot\cdot})^2$$

: 블록제곱합, 자유도 $b-1$ (9-17)

확률화블록계획법에서는 항상 다음 사실이 성립한다.

제곱합과 자유도의 분할:

제곱합 : $\mathrm{SST} = \mathrm{SSE} + \mathrm{SStr} + \mathrm{SSB}$ (9-18)

자유도 : $bk-1 = (b-1)(k-1) + (k-1) + (b-1)$ (9-19)

표 9.2.3은 확률화블록계획법에서의 분산분석표를 보여준다. 이 분산분석표에서 블록요인과 오차요인의 제곱합과 자유도를 합쳐 새로운 오차항의 제곱합과 자유도로 생각하면 일원분산분석의 표와 같아짐을 알 수 있다.

표 9.2.3 확률화블록계획법의 분산분석표

요인	제곱합	자유도	평균제곱합	F 값
처리	SSTr	$k-1$	$\mathrm{MSTr} = \mathrm{SSTr}/(k-1)$	$F_0 = MSTr/MSE$
블록	SSB	$b-1$	$\mathrm{MSB} = \mathrm{SSB}/(b-1)$	
오차	SSE	$(b-1)(k-1)$	$\mathrm{MSE} = \mathrm{SSE}/(b-1)(k-1)$	
전체	SST	$bk-1$		

확률화블록계획법에서는 완전확률화계획법과는 달리 전체 실험을 랜덤하게 하지 않고 각 블록안에서만 실험을 랜덤하게 하면 된다. 확률화블록계획법에서 또 유의할 점은 블록변량의 변동을 오차항에서 분리시켰지만 주관심은 일원분산분석에서와 마찬가지로 한 인자의 수준간의 차에 대한 검정이라는 것이다. 블록변량은 오차제곱합을 감소시켜 인자의 수준간의 차에 대한 검정을 더욱 정확하게 하기 위해 사용되는 것이므로 그 수준간의 차에 대한 검정은 중요하지 않다. 또한, 일원분산분석의 오차제곱합에서 블록제곱합이 분리되었으므로 오차제곱합은 감소되지만 자유도 역시 감소하므로 오차평균제곱(MSE)이 항상 감소하는 것은 아니다.

[예 9.2.1] 표 9.2.4는 세 종류의 자동차(A, B, C)를 다섯 명의 운전자가 측정한 연비 데이터(표 9.2.2)를 회사와 운전자에 따라서 정리해 적은 것이다.

표 9.2.4 세 회사 자동차의 연비

운전자	1	2	3	4	5	평균($\bar{y}_{i\cdot}$)

표 9.2.4 세 회사 자동차의 연비

	A	22.4	16.1	19.7	21.1	24.5	20.76
자동차	B	16.3	12.6	15.9	17.8	21.0	16.72
	C	20.2	15.2	18.7	18.9	23.8	19.36
평균($\bar{y}_{\cdot j}$)		19.63	14.63	18.10	19.27	23.10	18.947

1) 이 데이터가 완전확률화계획법에 의해 측정되었다고 가정할 때 『eStat』을 이용하여 세 회사 자동차의 연비가 같은지에 대한 분산분석을 하라.
2) 이 데이터가 확률화블록계획법에 의해 측정되었다고 가정할 때 『eStat』을 이용하여 세 회사 자동차의 연비가 같은지에 대한 분산분석을 하라.

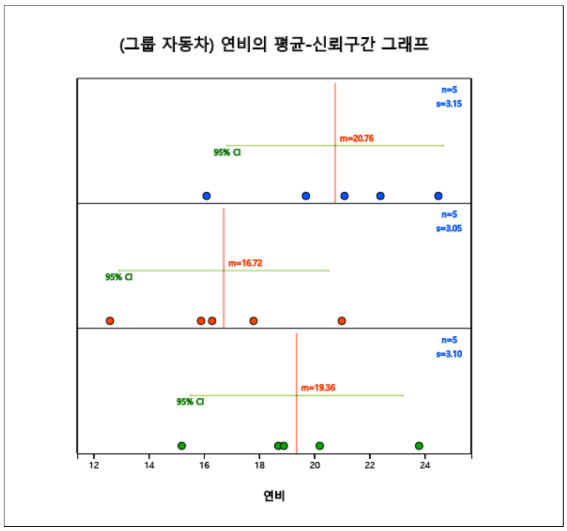
<풀이>

1) 『eStat』에서 <그림 9.2.1>과 같이 데이터 입력을 하고 분산분석 아이콘 을 클릭하여 나타나는 변량선택창에서 '분석변량'을 연비, 'by 그룹'을 자동차로 선택하면 <그림 9.2.2>와 같은 각 자동차별 평균-신뢰구간 그래프가 나타난다.

파일 Ex914.csv
분석변량 by 그룹
3: 연비 1: 자동차
(선택된 자료는 원시자료)
선택변량 V3 by V1,

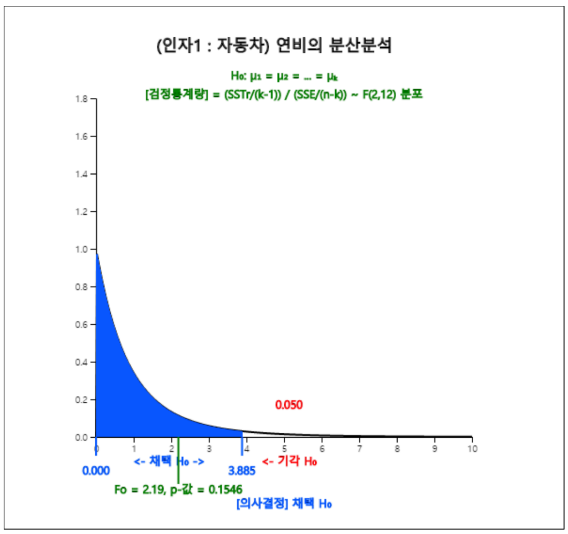
	자동차	운전자	연비	V4
1	A	1	22.4	
2	A	2	16.1	
3	A	3	19.7	
4	A	4	21.1	
5	A	5	24.5	
6	B	1	16.3	
7	B	2	12.6	
8	B	3	15.9	
9	B	4	17.8	
10	B	5	21.0	
11	C	1	20.2	
12	C	2	15.2	
13	C	3	18.7	
14	C	4	18.9	
15	C	5	23.8	

<그림 9.2.1> 확률화블록계획법 분산분석을 위한 『eStat』 데이터 입력



<그림 9.2.2> 각 자동차별 평균-신뢰구간 그래프

그래프 밑의 선택사항에서 'ANOVA F 검정' 버튼을 클릭하면 <그림 9.2.3>의 분산분석 그래프와 <그림 9.2.4>와 같은 분산분석표가 나타난다. 분산분석의 결과는 각 회사 자동차간에 연비의 차가 없다는 것이다. <그림 9.2.5>의 다중비교 결과도 마찬가지이다.



<그림 9.2.3> 각 자동차별 연비에 대한 분산분석

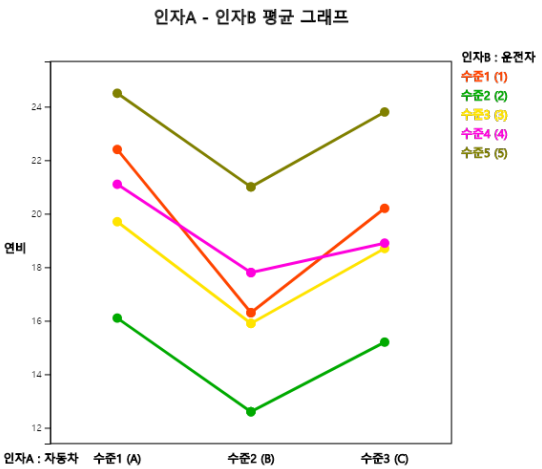
분산분석					
요인	제곱합	자유도	평균제곱	F 값	p 값
처리	42.085	2	21.043	2.190	0.1546
오차	115.312	12	9.609		
전체	157.397	14			

<그림 9.2.4> 각 자동차별 연비에 대한 분산분석

다중비교	분석변량	(연비)	그룹명	(자동차)
평균차 (95%HSD)	1 (A) 20.76	2 (B) 16.72	3 (C) 19.36	
1 (A) 20.76		4.04 (5.23)	1.40 (5.23)	
2 (B) 16.72	4.04 (5.23)		2.64 (5.23)	
3 (C) 19.36	1.40 (5.23)	2.64 (5.23)		

<그림 9.2.5> 각 자동차별 연비에 대한 다중비교

2) 이 데이터가 확률화블록계획법으로 추출되었다면 오차제곱합에서 블록제곱합을 분리하여야 한다. 『eStat』의 변량선택창에서 'by 그룹'에 운전자를 추가하면 <그림 9.2.6>과 같은 각 자동차 종류에 대해 운전자별 연비를 볼 수 있는 산점도가 나타난다. 이 산점도를 보면 운전자별 연비가 많이 차이가 있음을 볼 수 있다.



<그림 9.2.6> 자동차와 운전자에 대한 연비 산점도

그래프 밑의 선택사항에서 'ANOVA F 검정' 버튼을 클릭하면 <그림 9.2.7>과 같은 평균표와 <그림 9.2.8>과 같은 분산분석표가 나타난다. 이 분산분석표를 보면 오차제

곱합의 감소가 뚜렷하게 나타나고 오차의 평균제곱도 크게 감소되었다. 이는 운전자간의 큰 변동이 오차항에서 분리되었기 때문이다. 인자 B(운전자)가 블록에 대한 제곱합을 분리해 낸 것인데 $p-$값을 보면 매우 작아 블록(운전자) 효과가 통계적으로 의미 있는 것을 알 수 있다. 인자 A(자동차 종류)별 연비에 대한 가설 $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$을 위한 F 값이 43.447로 $F_{2,8,0.05}$ = 4.46 보다 크므로 H_0를 유의수준 0.05하에서 기각할 수 있다. 결론적으로, 오차항에서 블록의 변동을 제거함으로써 각 제품 간의 유의한 차이를 찾아낼 수 있다.

2차원 통계량						
자료수 평균 표준편차	인자 B (운전자) 수준1 (1)	인자 B (운전자) 수준2 (2)	인자 B (운전자) 수준3 (3)	인자 B (운전자) 수준4 (4)	인자 B (운전자) 수준5 (5)	인자 A 수준 i 합계
인자A (자동차) 수준1 (A)	1 22.40 NaN	1 16.10 NaN	1 19.70 NaN	1 21.10 NaN	1 24.50 NaN	5 20.76 3.15
인자A (자동차) 수준2 (B)	1 16.30 NaN	1 12.60 NaN	1 15.90 NaN	1 17.80 NaN	1 21.00 NaN	5 16.72 3.05
인자A (자동차) 수준3 (C)	1 20.20 NaN	1 15.20 NaN	1 18.70 NaN	1 18.90 NaN	1 23.80 NaN	5 19.36 3.10
인자 B 수준 j 합계	3 19.63 3.09	3 14.63 1.82	3 18.10 1.97	3 19.27 1.68	3 23.10 1.85	15 18.95 3.35
결측수	0					

<그림 9.2.7> 각 수준별 평균 (표준편차는 계산이 불가하여 NaN 표시)

분산분석					
요인	제곱합	자유도	평균제곱	F 값	p 값
인자 A (자동차)	42.085	2	21.043	43.447	< 0.0001
인자 B (운전자)	111.437	4	27.859	57.521	< 0.0001
오차	3.875	8	0.484		
전체	157.397	14			

<그림 9.2.8> 확률화블록계획법에 대한 분산분석표

평균값으로 보면 A 회사의 자동차가 연비가 가장 좋다고 할 수 있다. 다른 회사 자동차와의 차이를 더 자세히 알아보기 위해 앞 절의 다중비교방법을 적용하여보자. 이 예제에서는 각 수준에서의 반복수가 같으므로 모든 평균비교에 하나의 HSD값을 사용할 수 있다. 즉,

$$\begin{aligned} HSD &= q_{3,8\,;\,0.05}\sqrt{\frac{MSE}{r}} \\ &= (4.041)\sqrt{\frac{0.484}{5}} \\ &= 1.257 \end{aligned}$$

따라서 각 평균값의 차이(4.04, 1.40, 2.64)가 기준값 1.257보다 모두 크므로 세 회사 자동차 간에 모두 유의한 연비의 차이가 있다고 할 수 있다.

9.3 두 인자 분산분석

분산분석에서 반응변량에 영향을 주는 인자(factor)가 2개 있는 경우를 **이원분산분석** (two-way analysis of variance)이라 한다. 일반적으로 공학, 의학, 농학 등의 실험에 많이 사용되는 방법으로 반응변량은 2개 인자들의 각 수준의 조합에서 관측된다. 일반적으로 실험 결과의 신뢰도를 높이기 위하여 가능하면 각 인자(A 와 B 라 부르자) 수준들의 조합에서 2회 이상의 실험을 반복하는 것이 바람직하다.

각 인자 수준의 조합에서 여러 번 반복 실험으로 데이터를 얻는 경우 이원분산분석은 일원분산분석과 유사하게 인자 A 의 각 수준별 모평균이 같은지 검정(인자 A 의 **주효과** (main effect) 검정이라 부름)하거나, 인자 B 의 각 수준별 모평균이 같은지 검정(인자 B 의 주효과 검정이라 부름)한다. 이밖에도 이원분산분석에서는 한 인자의 효과가 다른 인자의 수준에 따라 다른가를 검정(이를 **교호작용효과**(interaction effect) 검정이라 부름)할 수 있다. 예를 들어, 어느 화학공정에서 온도가 낮을 때는 압력이 높을수록 생산량이 많고, 온도가 높을 때는 압력이 낮을수록 생산량이 많다면 온도와 압력의 두 인자 간에 교호작용효과가 있다고 할 수 있다. 교호작용효과는 한 인자의 효과가 다른 인자의 수준의 변화에 따라 변하는 모형에서 존재한다.

[예 9.3.1] 네 가지 종류의 비료와 세 가지 벼품종에 대한 쌀 수확량을 조사하기 위하여 한 농업 시험장에서 각 수준에 대한 3회 반복 실험을 한 자료가 표 9.3.1과 같다.

표 9.3.1 비료와 벼품종에 따른 수확량 (단위 kg)

비료의 종류	벼품종의 종류		
	1	2	3
1	64,66,70	72,81,64	74,51,65
2	65,63,58	57,43,52	47,58,67
3	59,68,65	66,71,59	58,45,42
4	58,50,49	57,61,53	58,45,42

1) 비료와 벼의 각 수준별 조합에 대한 수확량 평균을 구하라.

2) 『eStat』을 이용하여 벼의 품종(1, 2, 3)을 X축으로 하고 수확량을 Y축으로 하여

산점도를 그려라. 산점도의 점의 색을 비료의 종류로 구분하라. 그리고 각 수준별 조합의 평균을 산점도 위에 표시하고 비료의 종류별로 꺾은선으로 연결하여 관찰하라.

3) 비료와 벼품종의 주효과를 검정하고 두 인자의 교호작용에 대하여 검정하라.

4) 『eStat』을 이용하여 이원분산분석 검정 결과를 확인하라.

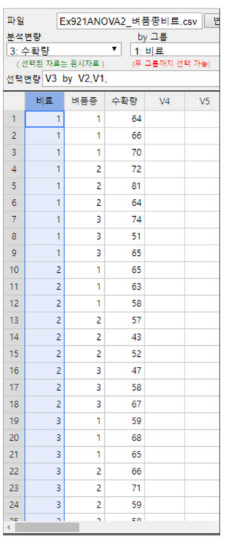
<풀이>

1) 편의상 비료를 인자A, 벼를 인자B라고 하고 주효과 및 교호작용을 알아보기 위하여 각 수준별 조합에 대한 평균 쌀 수확량을 계산하면 표 9.3.2와 같다. 실험 자료를 y_{ijk}로 표시하였을 때 각각의 비료와 벼에 대한 평균 기호 표시 $\bar{y}_{ij\cdot}$, $\bar{y}_{i\cdot\cdot}$, $\bar{y}_{\cdot j\cdot}$, $\bar{y}_{\cdots}$에 주의하라.

표 9.3.2 각각의 벼와 비료에 대한 평균 수확량

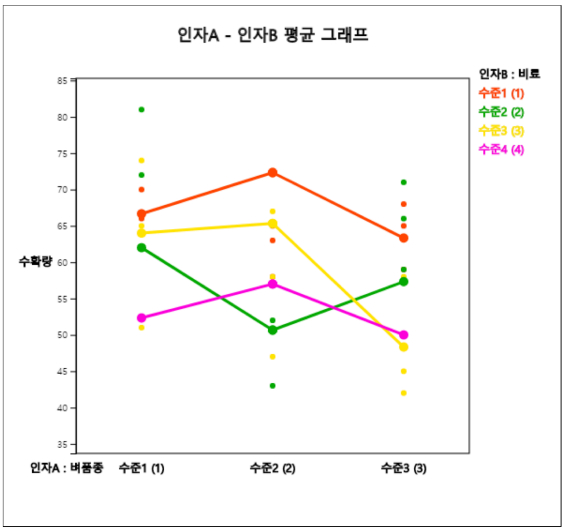
비료의 종류 (인자B)	벼의 품종 (인자A) 1	2	3	행평균
1	$\bar{y}_{11\cdot}=66.7$	$\bar{y}_{12\cdot}=72.3$	$\bar{y}_{13\cdot}=63.3$	$\bar{y}_{1\cdot\cdot}=67.4$
2	$\bar{y}_{21\cdot}=62.0$	$\bar{y}_{22\cdot}=50.7$	$\bar{y}_{23\cdot}=57.3$	$\bar{y}_{2\cdot\cdot}=56.7$
3	$\bar{y}_{31\cdot}=64.0$	$\bar{y}_{32\cdot}=65.3$	$\bar{y}_{33\cdot}=48.3$	$\bar{y}_{3\cdot\cdot}=59.2$
4	$\bar{y}_{41\cdot}=52.3$	$\bar{y}_{42\cdot}=57.0$	$\bar{y}_{43\cdot}=50.0$	$\bar{y}_{4\cdot\cdot}=53.1$
열평균	$\bar{y}_{\cdot 1\cdot}=61.3$	$\bar{y}_{\cdot 2\cdot}=61.3$	$\bar{y}_{\cdot 3\cdot}=54.8$	$\bar{y}_{\cdots}=59.1$

2) 『eStat』을 이용하여 이원분산분석용 산점도를 그리려면 먼저 자료를 <그림 9.3.1>과 같이 변량1은 비료, 변량2는 벼품종, 변량3은 수확량을 입력한다.



<그림 9.3.1> 『eStat』에서 비료와 벼품종별 수확량의 데이터 입력

주메뉴의 분산분석 아이콘 을 클릭하여 나타나는 변량선택박스에서 '분석변량'을 수확량, 'by 그룹' 변량을 벼품종과 비료를 차례로 선택하면 <그림 9.3.2>와 같은 벼품종별 쌀의 수확량에 대한 산점도가 나타난다. 그리고 각 수준별 조합의 평균을 비료별로 꺾은선으로 연결하면서 색으로 표시하여 준다. 이 그래프에서 벼품종 1은 사용하는 비료에 관계없이 벼품종 3보다 항상 수확량이 많다는 것을 보여준다. 벼품종 2는 사용하는 비료의 종류에 따라 수확량의 변화가 심하며(즉, 교호작용이 있다), 비료의 종류 1을 사용하면 대체로 벼의 품종에 관계없이 높은 수확량을 얻을 수 있다는 것을 알 수 있다.



<그림 9.3.2> 비료와 벼품종별 평균 수확량의 그래프

3) 인자 A, 즉, 벼의 주효과를 검정하는 것은 다음 귀무가설을 검정하는 것이다.

H_0 : 세 가지 벼품종에 대한 평균수확량은 같다.

이 귀무가설이 기각되면 '벼의 주효과가 있다'라고 말한다. 벼의 주효과를 검정하기 위해서는 일원분산분석에서와 같이 각각의 벼품종에 대한 평균에서 전체평균까지의 거리제곱합을 이용한다.

$$SSA = 12(61.3 - \bar{y}_{\cdots})^2 + 12(61.3 - \bar{y}_{\cdots})^2 + 12(54.8 - \bar{y}_{\cdots})^2 = 342.39$$

여기서 가중치 12는 각각의 벼에 대한 실험 자료의 개수를 의미한다.

인자 B, 즉, 비료의 주효과를 검정하는 것은 다음 귀무가설을 검정하는 것이다.

H_0 : 네 종류 비료의 평균수확량은 같다.

이 귀무가설이 기각되면 '비료의 주효과가 있다'라고 말한다. 비료의 주효과를 검정하기 위해서는 각각의 비료에 대한 평균에서 전체평균까지의 거리제곱합을 이용한다.

$$SSB = 9(67.4 - \bar{y}_{...})^2 + 9(56.7 - \bar{y}_{...})^2 + 9(59.2 - \bar{y}_{...})^2 + 9(53.1 - \bar{y}_{...})^2 = 1002.89$$

여기서 가중치 9는 각각의 비료에 대한 실험 자료의 개수를 의미한다.

벼와 비료의 교호작용 효과(인자 AB라고 표현함)를 검정하는 것은 다음 귀무가설을 검정하는 것이다.

H_0 : 벼와 비료 사이에는 교호작용이 없다

이 귀무가설이 기각되면 '벼와 비료 사이에는 교호작용 효과가 있다'라고 말한다. 이 검정을 위해서는 각각의 벼와 비료에 대한 평균에서 벼의 평균과 비료의 평균을 뺀 후 전체평균을 더한 거리제곱합을 이용한다.

$$\begin{aligned}\text{SSAB} = {} & 3(66.7 - \bar{y}_{1..} - \bar{y}_{.1.} + \bar{y}_{...})^2 + 3(72.3 - \bar{y}_{1..} - \bar{y}_{.2.} + \bar{y}_{...})^2 \\ & + 3(63.3 - \bar{y}_{1..} - \bar{y}_{.3.} + \bar{y}_{...})^2 + 3(62.0 - \bar{y}_{2..} - \bar{y}_{.1.} + \bar{y}_{...})^2 \\ & + 3(50.7 - \bar{y}_{2..} - \bar{y}_{.2.} + \bar{y}_{...})^2 + 3(57.3 - \bar{y}_{2..} - \bar{y}_{.3.} + \bar{y}_{...})^2 \\ & + 3(64.0 - \bar{y}_{3..} - \bar{y}_{.1.} + \bar{y}_{...})^2 + 3(65.3 - \bar{y}_{3..} - \bar{y}_{.2.} + \bar{y}_{...})^2 \\ & + 3(48.3 - \bar{y}_{3..} - \bar{y}_{.3.} + \bar{y}_{...})^2 + 3(52.3 - \bar{y}_{4..} - \bar{y}_{.1.} + \bar{y}_{...})^2 \\ & + 3(57.0 - \bar{y}_{4..} - \bar{y}_{.2.} + \bar{y}_{...})^2 + 3(50.0 - \bar{y}_{4..} - \bar{y}_{.3.} + \bar{y}_{...})^2 \\ = {} & 588.94\end{aligned}$$

이러한 제곱합을 이용하여 각각의 효과를 바로 검정할 수는 없고 오차제곱합(error sum of squares; SSE)을 이용하여 수정을 하는데 이를 구하기 위해서는 먼저 각각의 자료에서 전체평균까지의 거리의 합인 총제곱합(total sum of squares; SST)을 구한다.

$$\begin{aligned}\text{SST} = {} & (64 - \bar{y}_{...})^2 + (66 - \bar{y}_{...})^2 + (70 - \bar{y}_{...})^2 \\ & + \ldots + (53 - \bar{y}_{...})^2 + (59 - \bar{y}_{...})^2 + (38 - \bar{y}_{...})^2 = 3267.56\end{aligned}$$

수학적으로 이 총제곱합은 다른 제곱합들의 합이 됨을 증명할 수 있다. 즉,

$$\text{SST} = SSA + SSB + SSAB + SSE$$

따라서 오차제곱합은 다음과 같다.

$$\text{SSE} = SST - (SSA + SSB + SSAB) = 1333.33$$

만일 각각의 벼나 비료에 대한 수확량이 정규분포를 따르고 또 분산이 같다고 가정할 수 있다면 이들 제곱합을 자유도로 나눈 통계량은 F 분포를 따른다. 따라서 주효과와 교호작용효과는 F 분포를 이용하여 검정할 수 있는데 교호작용효과를 분리할 경우 이것을 먼저 검정한다. 유의수준 5%로 검정한 결과는 다음과 같다.

① 벼품종과 비료에 대한 교호작용효과의 검정

$$F_0 = \frac{\frac{\mathrm{SSAB}}{(3-1)(4-1)}}{\frac{\mathrm{SSE}}{24}} = 1.77$$

$$F_{6,24\,;\,0.05} = 2.51$$

따라서 교호작용이 있다고 할 수 없고 <그림 9.3.1>의 평균수확량의 그림에 나타난 교호작용효과는 무시할 수 있을 정도로 작거나, 다른 오차에 기인한 우연한 것일 수 있다고 해석할 수 있다. 『eStat』를 이용하여 $F_0 = 1.77$ 에 대한 p-값을 구하면 0.1488이다.

② 벼품종(A 인자)에 대한 주효과의 검정

$$F_0 = \frac{\frac{\mathrm{SSA}}{(3-1)}}{\frac{\mathrm{SSE}}{24}} = 3.08$$

$$F_{2,24;0.05} = 3.40$$

따라서 벼품종에 따라 평균수확량이 같다는 귀무가설은 기각하지 못한다. 즉, 벼의 종류에 따라 수확량의 차이가 있다는 충분한 증거가 없다. 『eStat』를 이용하여 $F_0 = 3.08$ 에 대한 p-값을 구하면 0.0644이다.

③ 비료(B 인자)에 대한 주효과의 검정

$$F_0 = \frac{\frac{\mathrm{SSB}}{(4-1)}}{\frac{\mathrm{SSE}}{24}} = 6.02$$

$$F_{3,24\,;\,0.05} = 3.01$$

따라서 비료의 종류에 따라 평균수확량이 같다는 귀무가설은 기각된다. 즉, 사용한 비료의 종류에 따라 수확량의 차이가 있다. 1)의 검정에서 교호작용효과는 없다고 가정할 수 있으므로 우리는 평균값의 산점도에서 알 수 있듯이 비료종류 1의 사용이 평균적으로 높은 수확량을 준다는 결론을 내릴 수 있다. 『eStat』를 이용하여 $F_0 = 6.02$ 에 대한 p-값을 구하면 0.0033이다.

표 9.3.3은 이 자료에 대한 이원분산분석 결과이다.

표 9.3.3 비료와 벼품종에 따른 수확량의 이원분산분석표

요인	제곱합	자유도	평균제곱	F 값	p-값

표 9.3.3 비료와 벼품종에 따른 수확량의 이원분산분석표

벼의품종	342.3889	2	171.1944	3.0815	0.0644
비료종류	1002.8889	3	334.2963	6.0173	0.0033
교호작용	588.9444	6	98.1574	1.7668	0.1488
오 차	1333.3333	24	55.5556		
전 체	3267.5556	35			

4) <그림 9.3.2> 밑의 이원분산분석 선택사항에서 'ANOVA F 검정' 버튼을 누르면 로그창에 각 수준별 조합에 대한 2차원 평균/표준편차표(<그림 9.3.3>)와 이원분산분석표(<그림 9.3.4>)가 나타난다.

2차원 통계량				
자료수 평균 표준편차	인자 A (벼품종) 수준1 (1)	인자 A (벼품종) 수준2 (2)	인자 A (벼품종) 수준3 (3)	인자 B 수준 i 합계
인자B (비료) 수준1 (1)	3 66.67 3.06	3 72.33 8.50	3 63.33 11.59	9 67.44 8.34
인자B (비료) 수준2 (2)	3 62.00 3.61	3 50.67 7.09	3 57.33 10.02	9 56.67 8.08
인자B (비료) 수준3 (3)	3 64.00 4.58	3 65.33 6.03	3 48.33 8.50	9 59.22 9.97
인자B (비료) 수준4 (4)	3 52.33 4.93	3 57.00 4.00	3 50.00 10.82	9 53.11 6.99
인자 A 수준 j 합계	12 61.25 6.65	12 61.33 10.26	12 54.75 10.79	36 59.11 9.66
결측수	0			

<그림 9.3.3> 각 수준 조합에 대한 평균/표준편차표

분산분석					
요인	제곱합	자유도	평균제곱	F 값	p 값
인자 A	342.389	2	171.194	3.082	0.0644
인자 B	1002.889	3	334.296	6.017	0.0033
교호작용	588.944	6	98.157	1.767	0.1488
오차	1333.333	24	55.556		
전체	3267.556	35			

<그림 9.3.4> 이원분산분석 결과표

위의 예제에서 알아 본 이원분산분석에 대한 이론을 일반화하여 보자. Y_{ijk}를 a개의 수준을 갖는 인자 A의 i번째 수준, b개의 수준을 갖는 인자 B의 j번째 수준에서 관측된 k번째 값을 나타내는 확률변량이라 하면 통계 모형은 다음과 같다.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, i = 1,2,\cdots,a\,;\, j = 1,2,\cdots,b\,;\, k = 1,2,\cdots,r \qquad (9\text{-}20)$$

μ : 총평균

α_i : A의 i 번째 수준의 효과
β_j : B의 j 번째 수준의 효과
γ_{ij} : A의 i 번째 수준과 B의 j 번째 수준과의 교호작용효과
ε_{ijk} : 오차항으로 서로 독립이며 $N(0,\sigma^2)$ 을 따른다.

위의 모형에서는 인자 A, B의 각 수준에서 실험이 동일하게 r 번 반복되었다고 가정한다.

분석에 필요한 제곱합은 일원분산분석에서와 같이 관측값과 총평균간의 편차를 분할한 식을 이용하여 다음과 같이 주어진다. 전체 관측값들의 개수를 $n = abr$ 이라 하면, 제곱합의 정의와 이들 사이의 관계는 다음과 같다.

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk} - \bar{Y}_{...}\right)^2 \text{ : 총제곱합, 자유도 } n-1 \tag{9-21}$$

$$\text{SSA} = br\sum_{i=1}^{a}\left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2 \text{ : 인자 A의 처리제곱합, 자유도 } a-1 \tag{9-22}$$

$$\text{SSB} = ar\sum_{j=1}^{b}\left(\bar{Y}_{.j.} - \bar{Y}_{...}\right)^2 \text{ : 인자 B의 처리제곱합, 자유도 } b-1 \tag{9-23}$$

$$\text{SSAB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}\right)^2 \text{ : 인자 A와 B의 교호작용효과를 나타내는 제곱합, 자유도 } (a-1)(b-1) \tag{9-24}$$

$$\text{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk} - \bar{Y}_{ij.}\right)^2 \text{ : 오차제곱합, 자유도 } n-ab \tag{9-25}$$

제곱합과 자유도의 분할:

제곱합: $SST = SSA + SSB + SSAB + SSE$ (9-26)

자유도: $(n-1) = (a-1) + (b-1) + (a-1)(b-1) + (n-ab)$ (9-27)

이들을 요약한 분산분석표는 표 9.3.4와 같다.

표 9.3.4 이원분산분석에서의 분산분석표

요인	제곱합	자유도	평균제곱	F값
인자 A	SSA	$a-1$	$\text{MSA} = \text{SSA}/(a-1)$	$F_1 = \text{MSA/MSE}$
인자 B	SSB	$b-1$	$\text{MSB} = \text{SSB}/(b-1)$	$F_2 = \text{MSB/MSE}$
교호작용	SSAB	$(a-1)(b-1)$	$\text{MSAB} = \text{SSAB}/((a-1)(b-1))$	$F_3 = \text{MSAB/MSE}$
오차	SSE	$n-ab$	$\text{MSE} = \text{SSE}/(n-ab)$	
전체	SST	$n-1$		

반복이 없는 이원분산분석
각 수준의 조합에서 반복 관측이 없는 경우에는 교호작용효과를 추정할 수 없어 위의 분산분석표에서 교호작용행은 삭제된다. 이 경우 확률화블록계획법의 분산분석표(<그림 9.1.16>)와 같아진다.

인자 A와 인자 B의 주효과와 교호작용에 대한 가설검정은 다음과 같다. 위와 같이 교호작용을 분리한 경우에는 이 교호작용에 대한 검정을 먼저 하는 것이 타당하다. 왜냐하면, 교호작용효과의 유의여부에 따라 각 인자의 평균검정의 결과 해석 방법이 달라지기 때문이다.

1) 교호작용효과에 대한 F 검정:

$$H_0 : \gamma_{ij} = 0, i = 1,2,\cdots,a ; j = 1,2,\cdots,b \quad (9\text{-}28)$$

$F_3 = \text{MSAB/MSE}$ 가 $F_{(a-1)(b-1),n-ab;\alpha}$ 보다 크면 귀무가설 H_0 를 기각 (9-29)

2) 인자 A의 주효과에 대한 F 검정:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \quad (9\text{-}30)$$

$F_1 = \text{MSA/MSE}$ 가 $F_{(a-1),n-ab;\alpha}$ 보다 크면 H_0 를 기각 (9-31)

3) 인자 B의 주효과에 대한 F 검정:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0 \quad (9\text{-}32)$$

$F_2 = \text{MSB/MSE}$ 가 $F_{b-1,n-ab;\alpha}$ 보다 크면 H_0 를 기각 (9-33)

(『eStat』에서는 이들 각 검정에 대한 p-값을 계산하여 주므로 이를 이용하여 검정한다. 즉, 각 검정에서 p-값이 유의수준보다 작으면 귀무가설 H_0 를 기각한다.)

교호작용효과에 대한 검정에서 유의하지 않으면 각 인자의 주효과에 대한 검정을 실시하여 수준간에 유의한 차이가 있는지를 검정하면 된다. 그러나 교호작용효과가 유의하게 있는 경우에는 각 인자의 주효과에 대한 검정은 의미가 없으므로 인자들의 어느 수준조합에서 평균들이 차이가 나타나는지를 해석하여야 한다.

일원분산분석에서와 같이 수준간의 평균차이가 있다는 결론을 얻는 경우에는 그 차이가 어느 수준간에서 나타나는지 살펴보기 위하여 각 수준에서의 신뢰구간을 비교할 수 있다. 그리고 가정의 타당성을 조사하기 위하여 잔차분석을 할 필요가 있다.

이원분산분석의 표본 추출을 위한 실험의 계획

이원분산분석에서도 공학이나 농학 등의 실험을 통해 각 인자의 수준별로 표본 데이터를 얻는 것은 다른 요인의 영향을 받을 수 있어 표본의 추출에 주의하여야 한다. 한 인자의 각 수준 간에 존재할 수 있는 차이를 정확하게 파악하기 위해서는 다른 인자들의 영향을 될 수 있는 대로 적게 해 주는 것이 좋다. 이를 위한 많이 사용되는 방법 중의 하나는 실험전체를 랜덤(random)하게 하는 완전확률화계획법(completely randomized design)이다. 이밖에도 여러 가지 실험계획법이 있는데 자세한 내용은 실험계획에 관한 참고문헌을 참조하기 바란다.