제 9장. 분산분석

분산분석(analysis of variance, ANOVA) :

특성값의 변동을 나타내는 제곱합을 요인별 제곱합과 오차에 의한 제곱합으로 분해하고, 이들의 비를 통계량으로 하여 요인의 유의성을 검증하는 통계적 기법

- 인자(factor) 또는 요인 : 관측값에 영향을 주는 속성
- 인자수준 (factor level) : 인자의 여러 조건
- 처리(treatment) : 인자 수준의 조합
- 일원배치법(one-way ANOVA) : 특성값에 영향을 주는 요인이 1개만 있는 모형
- 이원배치법(two-way ANOVA) : 특성값에 영향을 주는 요인이 2개 있는 모형

9.1 일원배치법

일원배치법의 자료 구조 :

	처리1	처리2	$\cdots$	처리k	
	y_{11}	y_{21}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{1n_1}	y_{2n_2}	$\cdots$	y_{kn_k}	
평균	$\bar{y}_1.$	$\bar{y}_2.$	$\cdots$	$\bar{y}_k.$	총평균 $\bar{y}..$

일원배치법의 모형 :

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \ , \quad i=1,\cdots,k \ , \ j=1,\cdots,n_i$$

다만, μ : 총평균

α_i : 처리i 의 효과로서 $\sum n_i \alpha_i = 0$ 을 가정

ϵ_{ij} : 오차항으로 서로 독립인 $N(0,\sigma^2)$ 확률변수

제곱합의 분해 :

$$\text{SST} = \text{SSE} + \text{SS}_{\text{tr}}$$

총제곱합 $\quad \text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}..)^2,$ $\quad$ 자유도 $N-1$

잔차제곱합 $\quad \text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i.)^2,$ $\quad$ 자유도 $N-k$

처리제곱합 $\quad \text{SS}_{\text{tr}} = \sum_{i=1}^{k} n_i(y_i. - \bar{y}..)^2,$ $\quad$ 자유도 $k-1$

(단, $N = \sum n_i$: 관측값의 총개수)

평균제곱(mean square) : $\quad \mathrm{MS}_{\mathrm{tr}} = \dfrac{\mathrm{SS}_{\mathrm{tr}}}{k-1}$, $\quad \mathrm{MSE} = \dfrac{\mathrm{SSE}}{N-k}$

<u>처리효과의 유의성 검정</u> :

$$H_0 \ : \ \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \quad \text{또는} \quad H_0 \ : \ \mu_1 = \mu_2 = \cdots = \mu_k$$

- 검정통계량 : $F = \dfrac{\mathrm{MS}_{\mathrm{tr}}}{\mathrm{MSE}}$

- **분산분석표** (ANOVA table)

요인	제곱합	자유도	평균제곱	F 값	유의확률
처리	$\mathrm{SS}_{\mathrm{tr}}$	$k-1$	$\mathrm{MS}_{\mathrm{tr}}$	$f = \mathrm{MS}_{\mathrm{tr}}/\mathrm{MSE}$	$P(F \geq f)$
잔차	SSE	$N-k$	MSE		
계	SST	$N-1$			

예 1) 어느 농장에서 서식하고 있는 딱정투구벌레에 대한 연구를 하던 도중, 벌레들이 선호하는 색상이 있는가를 알아보기 위해 다음과 같은 실험을 실시하였다. 재질과 크기가 같은 네 가지 색상의 판자를 각각 여섯 개씩 준비하여, 그 위에 끈끈이를 바르고 여섯 지점에 각 네 가지 판자를 일주일 동안 설치하여 잡힌 벌레의 수를 관측하였다. 그 결과가 다음과 같다.

판자의 색	잡힌 벌레수					
레몬색 (1)	45	59	48	46	38	47
흰 색 (2)	21	12	14	17	13	17
녹 색 (3)	37	32	15	25	39	41
파란색 (4)	16	11	20	21	14	7

벌레들이 선호하는 색상에 차이가 있는지를 유의수준 5%에서 검정해보자. 주어진 자료는 일원배치모형을 적용할 수 있으며 검정하고자 하는 가설은 다음과 같다.

$$H_0 \ : \ \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0, \qquad H_1 \ : \text{적어도 한 } \alpha_i \text{ 는 0이 아니다.}$$

주어진 자료는 그룹을 나타내는 변수(color)와 관측값(num)을 나타내는 두 개의 열로 나타낼 수 있다. 자료입력의 편의를 위해 색상을 나타내는 변수를 숫자형 변수로 입력할 수 있는데, 이러한 경우 분산분석을 시행하기 전에 수치변수를 요인(factor)으로 변화하는 과정이 먼저 필요하다. 이를 위해서는 factor() 함수를 사용할 수 있고, 분산분석의 시행은 lm() 함수를 사용한다. 아래는 일원배치 분산분석을 위한 자료 입력과정과 분석 과정이다.

```
> num<-c(45,59,48,46,38,47,21,12,14,17,13,17,37,32,15,25,39,41,16,11,20,21,14,7)
> col<-factor(rep(1:4, rep(6,4)))
> bugs<-data.frame(num, col)
> fit<-lm(num ~ col, data=bugs)
> anova(fit)
Analysis of Variance Table

Response: num
          Df Sum Sq Mean Sq F value    Pr(>F)
col        3 4218.5 1406.15  30.552 1.151e-07 ***
Residuals 20  920.5   46.03
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

분산분석 결과, 검정통계량의 값은 30.55이고 유의확률은 0.001이하로 매우 작은 것으로 나타났다. 따라서 유의수준 5%에서 모평균이 모두 동일하다는 귀무가설을 기각할 수 있다. 즉, 딱정벌레들이 선호하는 색상에는 차이가 있다는 결론을 얻을 수 있다.

9.2 반복이 없는 이원배치법

반복이 없는 이원배치법의 자료구조 :

인자B / 인자A	B_1	B_2	$\cdots$	B_q	평균
A_1	y_{11}	y_{12}	$\cdots$	y_{1q}	$\bar{y}_{1\cdot}$
A_2	y_{21}	y_{22}	$\cdots$	y_{2q}	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A_p	y_{p1}	y_{p2}	$\cdots$	y_{pq}	$\bar{y}_{p\cdot}$
평균	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	$\cdots$	$\bar{y}_{\cdot q}$	$\bar{y}_{\cdot\cdot}$

반복이 없는 이원배치법의 모형 :

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}\,, \qquad i = 1,\cdots,p,\ \ j = 1,\cdots,q$$

다만, μ : 총평균

α_i : 인자 A의 i 번째 수준의 효과로서 $\sum \alpha_i = 0$

β_j : 인자 B의 j 번째 수준의 효과로서 $\sum \beta_j = 0$

ϵ_{ij} : 오차항으로 서로 독립인 $N(0,\sigma^2)$ 확률변수

제곱합의 분해 :

$$SST = SS_A + SS_B + SSE$$

총제곱합 : $\text{SST} = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \overline{y_{..}})^2$, 자유도 $N-1$

인자A 제곱합 : $\text{SS}_\text{A} = q\sum_{i=1}^{p}(\overline{y}_{i.} - \overline{y}_{..})^2$, 자유도 $p-1$

인자B 제곱합 : $\text{SS}_\text{B} = p\sum_{j=1}^{q}(\overline{y}_{.j} - \overline{y}_{..})^2$, 자유도 $q-1$

잔차제곱합 : $\text{SSE} = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2$, 자유도 $(p-1)(q-1)$

인자 A와 B의 효과의 검정:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$$
$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$$

- **분산분석표** (ANOVA table) :

요인	제곱합	자유도	평균제곱	F값	유의확률
인자A	SS_A	$p-1$	MS_A	$f_1 = \text{MS}_\text{A}/\text{MSE}$	$P(F \geq f_1)$
인자B	SS_B	$q-1$	MS_B	$f_2 = \text{MS}_\text{B}/\text{MSE}$	$P(F \geq f_2)$
잔차	SSE	$(p-1)(q-1)$	MSE		
계	SST	$pq-1$			

예 2) 어떤 금속 파이프의 부식방지를 위한 코팅방법을 고려하고 있다. 12개의 동일한 파이프에 네 가지 방법(A_1, A_2, A_3, A_4)으로 코팅을 하고 세 가지 토양(B_1, B_2, B_3)에 동일한 깊이로 묻고 일정한 시간이 지난 후에 부식정도를 측정한 결과가 다음과 같다. 코팅방법과 토양의 질에 따라 파이프의 부식(corrosion) 정도에 차이가 있는지를 유의수준 5%에서 검정해 보자.

	A_1	A_2	A_3	A_4
B_1	64	53	47	51
B_2	49	51	45	43
B_3	50	48	50	52

주어진 자료는 반복이 없는 이원배치법 모형이 적용 가능하며 코팅방법에 따른 효과를 α_i 라고 하고 토양의 질에 따른 효과를 β_i 라고 하면 검정하고자 하는 가설은 다음과 같다.

H_0 : $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, H_1 : 적어도 한 α_i 는 0이 아니다.

$$H_0 \ : \ \beta_1 = \beta_2 = \beta_3 = 0, \qquad\qquad H_1 \ : \text{ 적어도 한 } \beta_i \text{ 는 0이 아니다.}$$

주어진 이원배치법의 자료는 3개의 열(column)로 입력할 수 있다. 관측값인 파이프의 부식정도(y)와 두 개의 요인(A,B)로 입력하도록 한다. 단, 요인 A와 요인 B의 인자 수준을 (1,2,3) 등의 숫자로 입력했으므로 앞선 예제와 마찬가지로 분산분석 시행 전에 수치변수를 요인으로 변환하는 과정을 먼저 거쳐야 한다. 분산분석은 lm() 함수를 사용해서 시행할 수 있고, 두 개의 요인은 '+'기호로 연결한다.

```
> A<-factor(rep(1:4,3))
> B<-factor(rep(1:3,rep(4,3)))
> y<-c(64,53,47,51,49,51,45,43,50,48,50,52)
> cor<-data.frame(A,B,y)
> fit2<-lm(y ~ A + B, data=cor)
> anova(fit2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value Pr(>F)
A          3  83.583  27.861  1.3572 0.3422
B          2  91.500  45.750  2.2287 0.1889
Residuals  6 123.167  20.528
```

분산분석표 확인 결과, 코팅방법(A)의 효과에 대한 유의 확률은 0.3422이고 토양의 질(B)의 효과에 대한 유의 확률은 0.1889로 모두 유의수준 0.05보다 높다. 따라서 파이프의 코팅방법이나 토양의 질에 따라서 파이프의 부식정도에는 유의한 차이가 없다고 할 수 있다.

9.3 반복이 있는 이원배치법

반복이 있는 이원배치법의 자료구조 :

인자B / 인자A	B_1	B_2	$\cdots$	B_q	평균
A_1	y_{111} $\vdots$ y_{11r}	y_{121} $\vdots$ y_{12r}	$\cdots$ $\cdots$	y_{1q1} $\vdots$ y_{1qr}	$\overline{y}_{1\cdot\cdot}$
	$\overline{y}_{11\cdot}$	$\overline{y}_{12\cdot}$	$\cdots$	$\overline{y}_{1q\cdot}$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_p	y_{p11} $\vdots$ y_{p1r}	y_{p21} $\vdots$ y_{p2r}	$\cdots$ $\cdots$	y_{pq1} $\vdots$ y_{pqr}	$\overline{y}_{p\cdot\cdot}$
	$\overline{y}_{p1\cdot}$	$\overline{y}_{p2\cdot}$	$\cdots$	$\overline{y}_{pq\cdot}$	
	$\overline{y}_{\cdot 1\cdot}$	$\overline{y}_{\cdot 2\cdot}$	$\cdots$	$\overline{y}_{\cdot q\cdot}$	$\overline{y}_{\cdots}$

반복이 있는 이원배치법의 모형 :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}\,, \qquad i=1,\cdots,p\,,\; j=1,\cdots,q\,,\; k=1,\cdots,r$$

다만, μ : 총평균

α_i : 인자 A의 i 번째 수준의 효과로서 $\sum \alpha_i = 0$

β_j : 인자 B의 j 번째 수준의 효과로서 $\sum \beta_j = 0$

γ_{ij} : 인자 A의 i 번째 수준과 인자 B의 j 번째 수준의 **교호작용**

ϵ_{ij} : 오차항으로 서로 독립인 $N(0,\sigma^2)$ 확률변수

제곱합의 분해 :

$$\mathrm{SST} = \mathrm{SS_A} + \mathrm{SS_B} + \mathrm{SS_{A\times B}} + \mathrm{SSE}$$

총제곱합 : $\mathrm{SST} = \sum_i \sum_j \sum_k (y_{ijk} - \overline{y}_{...})^2$

인자A의 제곱합 : $\mathrm{SS_A} = qr\sum_i (\overline{y}_{i..} - \overline{y}_{...})^2$

인자B의 제곱합 : $\mathrm{SS_B} = pr\sum_j (\overline{y}_{.j.} - \overline{y}_{...})^2$

교호작용 제곱합 : $\mathrm{SS_{A\times B}} = r\sum_i \sum_j (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$

잔차제곱합 : $\mathrm{SSE} = \sum_i \sum_j \sum_k (y_{ijk} - \overline{y}_{ij.})^2$

인자 A와 B의 효과, 상호작용의 검정:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$$
$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$$
$$H_0 : \gamma_{ij} = 0, \qquad \forall\, i,j$$

- **분산분석표** (ANOVA table) :

요인	제곱합	자유도	평균제곱	F-값	유의확률
인자A	$\mathrm{SS_A}$	$p-1$	$\mathrm{MS_A}$	$f_1 = \mathrm{MS_A}/\mathrm{MSE}$	$P(F \ge f_1)$
인자B	$\mathrm{SS_B}$	$q-1$	$\mathrm{MS_B}$	$f_2 = \mathrm{MS_A}/\mathrm{MSE}$	$P(F \ge f_2)$
교호작용	$\mathrm{SS_{A\times B}}$	$(p-1)(q-1)$	$\mathrm{MS_{A\times B}}$	$f_3 = \mathrm{MS_{A\times B}}/\mathrm{MSE}$	$P(F \ge f_3)$
잔차	SSE	$pq(r-1)$	MSE		
계	SST	$pqr-1$			

예 3) (alzheimer.txt) 다음은 음악 감상이 알츠하이머를 겪고 있는 환자들의 불안감 정도에 어떠한 영향을 미치는지 알아보기 위한 실험의 결과이다. 초기(early stage)와 중기(middle stage)의 알츠하이머 환자들 각각에 대해 가벼운 대중음악(easy), 모차르트의 곡(Mozart),

피아노 연주곡(piano)의 세 종류의 음악을 들려주고 환자들이 느끼는 불안감의 정도를 측정하였다. 측정된 점수가 높을수록 불안감의 정도가 높다는 것을 의미한다. 주어진 자료에 대해 유의수준 5%에서 이원배치 분산분석을 시행해보자.

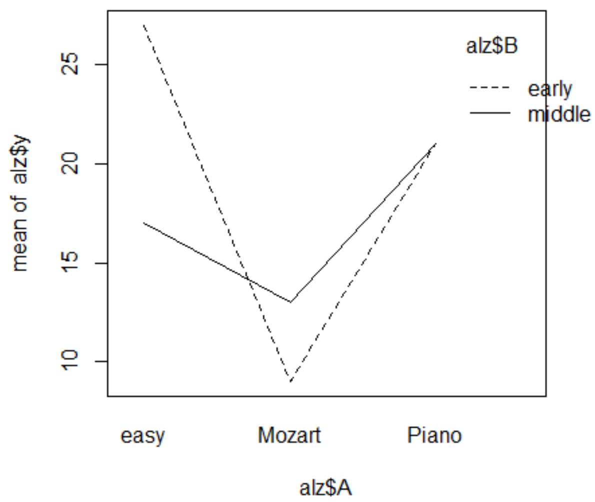
	piano	Mozart	easy
early stage Alzheimer's	21 24 22 18 20	9 12 10 5 9	29 26 30 24 26
middle stage Alzheimer's	22 20 25 18 20	14 18 11 9 13	15 18 20 13 19

검정하고자 하는 가설은 다음과 같다. (단, α_i 는 음악의 종류에 따른 효과이고 β_i는 알츠하이머 경과 정도에 따른 효과이고 γ_{ij} 는 음악과 질병의 상호작용을 의미한다)

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0, \quad H_0 : \beta_1 = \beta_2 = 0, \quad H_0 : \gamma_{ij} = 0, \quad \forall\, i, j$$

알츠하이머 병의 진행 정도(B)와 음악의 종류(A)에 대한 상호작용의 유무를 평균그림(을 통해 확인해보자. interaction.plot() 함수를 사용하면 평균 그림을 그릴 수 있고, 두 개의 요인 변수와 관측값을 나타내는 변수를 차례대로 입력해주면 다음과 같은 결과를 얻을 수 있다. 평균 그림 확인 결과, 두 요인 사이에는 상호작용이 존재하는 것으로 보인다.

```
> interaction.plot(alz$A,alz$B,alz$y)
```



상호작용을 포함하는 이원배치 분산분석은 두 개의 요인을 곱(*)으로 표현하여 다음과 같이 시행할 수 있다.

```
> fit3<-lm(y ~ A * B, data=alz)
> anova(fit3)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
A          2    740  370.00 49.8876 2.824e-09 ***
B          1     30   30.00  4.0449   0.05566 .  
A:B        2    260  130.00 17.5281 2.029e-05 ***
Residuals 24    178    7.42                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

분산분석 결과, 유의수준 5%에서 알츠하이머 병의 진행 정도(B)와 음악의 종류(A)에 따른 상호작용은 존재하는 것으로 나타났다 (F=17.5281, p-value<0.001). 이는 앞서 평균 그림을 통해서도 확인할 수 있었다. 또한 감상한 음악의 종류에 따라서 환자들의 불안감은 차이가 존재했지만 (p-value<0.001) 알츠하이머 병의 진행 정도에 따라서는 환자들이 느끼는 불안감에는 차이가 존재하지 않았다(p-value=0.056).

9.4 예제

예제1. 어느 시장 조사기관은 여러 가지 대중매체가 주는 정보의 양을 비교하기 위해 다음과 같은 실험을 계획하였다. 40명의 성인을 랜덤하게 추출하여 철저한 면접을 통해 TV, 신문, 라디오, 잡지 중 어느 매체를 많이 접하는지에 따라 분류하였다. 다음 표는 최근에 일어난 사건들에 대한 조사 대상자들의 인지도를 측정한 실험에서 얻어진 값들을 나타내고, 값이 클수록 인지도가 높은 것을 의미한다. 이 자료를 이용하여 사람들의 인지도가 대중매체에 따라 다르다고 할 수 있는지 유의수준 5%에서 검정해보자.

조사대상 대중매체			
TV	신문	라디오	잡지
16	13	18	11
19	14	18	15
25	15	15	11
22	16	14	17
21	15	14	17
15	13	10	13
16	19	18	14
22	16	15	16
21	20	15	13
18	14		11
	11		

예제2. 어느 회사의 마케팅 부서에서는 하나의 상품에 대해 세 가지 다른 디자인의 포장을 적용한 후 이 상품들을 서로 다른 5군데의 상점에서 한 달 동안 판매하였다. 그리고 그 판매 결과는 아래와 같다. 제품의 매출은 판매되는 상점과 제품의 포장 디자인에 따라 다르다고 할 수 있는가? 적절한 가설을 쓰고 유의수준 5%에서 이를 검정하시오.

	상점 1	상점 2	상점 3	상점 4	상점 5
상자 1	210	230	190	180	190
상자 2	195	170	200	190	193
상자 3	295	275	290	275	265

예제3. 남녀의 성별과 고단백질로 구성된 아침 식사의 섭취 여부가 성인의 신체적 활동 능력에 영향을 미치는지를 알아보기 위하여 랜덤하게 선택된 남녀 10명에 대해 각각 5명씩 고단백질 아침식사와 저단백질 아침식사를 섭취하게 한 후, 신체적 능력을 테스트를 통해 측정하였다. 측정된 점수가 높을수록 신체 활동 능력이 더 우수하다는 것을 의미한다. 실험 결과가 아래와 같을 때, 주어진 자료에 대해 이원배치법을 적용한 후 그 결과를 해석하여라.

	고단백질 식사	저단백질 식사
남성	10 7 9 6 8	5 4 7 4 5
여성	5 4 6 3 2	3 4 5 1 2