실험계획법

2021-15115 권이태

September 7, 2024

Contents

1 실험계획의 개념 **3**
- 1.1 실험계획의 순서 . . . 3
- 1.2 실험계획의 기본원리 . . . 4
- 1.3 실험계획법의 분류 . . . 4

2 일원배치법 **5**
- 2.1 인자와 모형의 분류 . . . 5
  - 2.1.1 인자의 분류 . . . 5
  - 2.1.2 모형의 분류 . . . 6
- 2.2 분산분석 . . . 6
- 2.3 분산분석 후의 추정 . . . 8
  - 2.3.1 각 수준의 모평균 추론 . . . 8
  - 2.3.2 수준별 모평균차 추론 . . . 8
  - 2.3.3 오차분산 추론 . . . 9
- 2.4 변량모형에서의 추론 . . . 9

3 반복이 없는 이원배치법 **10**
- 3.1 모수모형인 경우 . . . 10
- 3.2 난괴법 . . . 12
- 3.3 결측치의 취급 . . . 13

4 반복이 있는 이원배치법 **14**
- 4.1 모수모형인 경우 . . . 14
  - 4.1.1 등분산의 검토 . . . 15
  - 4.1.2 변동의 분해와 분산분석 . . . 15
  - 4.1.3 분산분석 후의 추정 . . . 16
  - 4.1.4 오차항에의 풀링 . . . 17
- 4.2 혼합모형의 경우 . . . 18
- 4.3 결측치의 취급 . . . 19
- 4.4 대비와 직교분해 . . . 20
- 4.5 계수치 데이터의 분석 . . . 20

5 다원배치법 **21**
- 5.1 평균제곱의 기대값을 구하는 방법 . . . 21
- 5.2 분산분석과 그 이후의 추정 . . . 22

6 분할법 **23**
6.1 일차단위가 일원배치일 때의 단일분할법 . . . 23
6.1.1 랜덤화와 데이터의 구조 . . . 24
6.1.2 분산분석표의 작성 . . . 24
6.1.3 분산분석 후의 추정 . . . 24
6.2 일차단위가 이원배치일 때의 단일분할법 . . . 25
6.3 이단분할법 . . . 25
6.4 인자가 분할이 안 되는 경우 . . . 25

7 라틴방격법 **26**
7.1 라틴방격법에서의 분산분석 . . . 26
7.2 그레코라틴방격법 . . . 27
7.3 초그레코라틴방격법 . . . 28

8 요인배치법 **29**
8.1 2^2 요인실험 . . . 29
8.1.1 반복이 없는 경우 . . . 29
8.1.2 반복이 있는 경우 . . . 30
8.2 2^n 요인실험 . . . 31
8.3 Yates의 계산법 . . . 32
8.4 3^2 요인실험 . . . 33

9 교락법과 일부실시법 **34**
9.1 2^n형의 교락법 . . . 34
9.1.1 인수분해식을 이용하는 방법 . . . 34
9.1.2 합동식의 이용방법 . . . 35
9.1.3 분산분석 . . . 36
9.2 완전교락과 부분교락 . . . 36
9.3 3^n형의 교락법 . . . 37
9.3.1 $S_{A\times B}$의 분해방법 . . . 37
9.3.2 단독교락 . . . 37
9.3.3 이중교락 . . . 37
9.4 2^n형의 일부실시법 . . . 38
9.4.1 2^n형의 1/2실시 . . . 38
9.4.2 2^n형의 1/4실시 . . . 38

10 불완비블록계획법 **39**
10.1 균형 불완비블록계획법 . . . 39
10.2 Youden 방격법 . . . 41

Chapter 1

실험계획의 개념

실험계획의 목적에는 아래의 것들이 있다.

- 어떤 요인이 반응에 유의한 영향을 주고 있는가를 파악하고 그 크기를 알기 위해
- 작은 영향을 주는 요인들이나 측정오차에 의한 변동은 어느정도인지 알기 위해
- 유의한 영향을 주는 요인들이 어떤 수준이어야 최적의 **특성치**를 얻을 수 있는지

특히 데이터에 산포를 준다고 생각되는 요인들 중 실험에서 직접 취급되는 원인을 **인자(factor)**라고 부른다. 또한 실험이 이루어지는 인자의 조건을 **수준(level)**이라 부르며, 그 개수를 **인자의 수준수**라 한다. **실험계획법**은 원하는 바를 위하여 인자의 선택에서부터 실험의 실행에 이르기까지 어떤 형태의 분석을 수행해야 할지 계획하는 방법이다.

1.1 실험계획의 순서

실험계획의 과정은 아래 7개의 과정을 반복하여 이루어진다.

1. **실험목적의 설정**: 실험을 통하여 얻고자 하는 목적을 명확히 설정하여야 한다.
2. **특성치의 선택**: 실험의 목적을 달성하기 위하여 적절한 특성치를 선택한다.
3. **인자와 인자수준의 선택**: 실험의 목적 달성을 위해 적절한 인자는 모두 선택해야 하나, 과다한 인자는 실험의 정밀도를 떨어뜨리고 실험비용이 증가한다는 단점이 있어 적절한 개수의 인자만을 골라 사용하여야 한다.

   인자 수준의 선택에 있어서는 실험자의 흥미영역과 실험가능한 영역의 교집합 내에서 현재 사용되는 인자수준/최적이라 생각되는 인자수준 등을 포함해 2-5 수준으로 결정해주는 것이 좋다.

   인자는 **계량인자**와 **계수인자**로 구분될 수 있다. 계량인자의 경우 인자수준은 흥미영역의 양끝과 그를 등간격으로 나눠 포함해주는 것이 좋고, 계수인자인 경우 주요 수준을 포함해 주어야 한다.
4. **실험의 배치와 실험순서의 랜덤화**: 인자수준의 조합, 블럭의 구성, 실험순서의 랜덤화 등이 올바르고 효율적인 추론을 위해 필요하다.
5. **실험의 실시**: 정해진 실험계획에 따라 관리 하에서 실험을 수행한다.
6. **데이터의 분석**: 분산분석, 상관분석, 회귀분석 등을 수행한다.
7. **분석결과의 해석과 조치**: 작성된 분산분석표 등에서 통계적 추론을 수행하고, 이를 실험의 목적 등을 고려하여 해석한다. 이후에는 그 결과를 바탕으로 새로운 실험을 계획하거나, 작업 방식을 바꾸는 등의 과정이 이어진다.

1.2 실험계획의 기본원리

실험계획의 기본원리에는 아래 다섯 가지가 있다.

1. **랜덤화의 원리**: 뽑힌 인자 외에 기타 원인들의 영향이 실험결과에 편의를 주지 않도록, 실험 순서/환경 등이 잘 랜덤화되어야 한다.
2. **반복의 원리**: 실험이 반복될수록 오차항의 자유도가 커지고 오차분산의 추정량이 정확해지며, 실험결과의 신뢰성을 높일 수 있다. 단 이 경우 비용 역시 많이 소모되므로, 적절한 반복수의 실험계획을 구상해야 한다.
3. **블록화의 원리**: 완벽한 랜덤화가 어렵다면, 실험의 환경을 균일하게 쪼개어 여러 블록으로 만든 뒤, 블록 내에서 인자의 영향을 조사하는 것이 바람직하다.
4. **교락의 원리**: 구할 필요가 없는 고차의 교호작용을 블럭과 교호시키는 등의 방식으로 실험의 효율을 높일 수 있다.
5. **직교화의 원리**: 요인 간에 직교성이 있도록 실험을 계획하면, 같은 실험횟수라도 검정력이 더 높은 실험계획을 할 수 있다.

1.3 실험계획법의 분류

1. **요인배치법**: 인자의 각 수준의 모든 조합에 대하여 완전히 랜덤화하여 실험을 행하는 방법이다.
2. **분할법**: 요인배치법에서 실험순서가 완전히 랜덤하게 정해지지 않고, 실험 전체를 여러 단계로 나누어 단계별로 랜덤화하는 방법이다. 이는 실험마다 인자수준을 바꾸기 어려운 경우 자주 사용된다.
3. **교락법**: 검출할 필요가 없는 교호작용을 다른 요인과 교락시켜 배치함으로써 실험횟수를 줄이고도 분석을 수행한다.
4. **일부실시법**: 일부 교호작용의 검출을 포기하고, 각 인자 조합 중 일부만 선택하여 실험을 실시함으로써 실험 횟수를 줄일 수 있다.
5. **불완비블록계획법**: 블록 내에 존재하는 인자수준의 조합이 모두 들어있지 않은 실험계획법으로, 수준수/블록수가 많을 때 실험 횟수를 줄일 수 있다.
6. **반응표면계획법**, **혼합물 실험계획법**, **로버스트 실험계획법**: 특정한 목적 하에서 사용되는 실험계획법이다.

Chapter 2

일원배치법

일원배치법은 하나의 인자가 존재할 때의 실험계획법으로, 가장 단순한 실험계획법이다. 일원배치법은 수준수와 반복수에 대해서는 별 제한이 없고, 반복수가 수준마다 같지 않아도 된다는 간편함이 있다. 단, 실험의 수행에 있어서는 난수표 등을 이용하여 실험을 완전히 랜덤화하는 것이 중요하다.
일반적인 반복이 같지 않은 일원배치법의 데이터 형태는 아래와 같다.

	인자의 수준				
	A_1	A_2	$\cdots$	A_l	
실험의 반복	x_{11}	x_{21}	$\cdots$	x_{l1}	
	x_{12}	x_{22}	$\cdots$	x_{l2}	
	x_{13}	x_{23}	$\cdots$	x_{l3}	
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
	x_{1m_1}	x_{2m_2}	$\cdots$	x_{lm_l}	
합계	$T_{1\cdot}$	$T_{2\cdot}$	$\vdots$	$T_{l\cdot}$	T
평균	$\bar{x}_{1\cdot}$	$\bar{x}_{2\cdot}$	$\vdots$	$\bar{x}_{l\cdot}$	$\bar{\bar{x}}$

이 데이터 구조 하에서 데이터 구조식은

$$x_{ij} = \mu + a_i + e_{ij}$$
$$(i,j) = (1,1),(1,2),\cdots,(1,m_1),(2,1),\cdots,(l,m_l)$$

이며, 이때 a_i는 인자 A의 **주효과**라 불리며, 모평균 μ로부터 수준 i에서의 모평균 μ_i가 얼마나 치우쳐져 있는지를 의미한다. 이때의 조건은

$$e_{ij} \sim_{i.i.d.} N(0,\sigma_E^2)$$
$$\sum_{i=1}^{l} a_i = 0$$

이 될 것이다. 한편 여기에서처럼 오차항은 **정규성**, **독립성**, **불편성**, **등분산성**을 만족하게 된다.

2.1 인자와 모형의 분류

2.1.1 인자의 분류

수준을 택하는 방법에 따라 인자를 분류할 수 있다. 먼저 **모수인자**는 기술적으로 미리 정해진 수준을 사용하는 인자로, 각 수준이 기술적인 의미를 가지고 있는 경우를 의미한다. 반대로 **변량인자**는 수준이 선택이

랜덤하게 이루어지며 각 수준이 기술적으로 의미를 가지고 있지 아니한 인자를 의미한다. 실험일 등이 여기에 포함된다.

앞서 본 데이터 구조식에서는 A_i가 모수인자라고 가정한다. 모수인자인 경우에는 모평균을 μ로 하기 위하여 $\sum_{i=1}^{l} m_i a_i = 0$임이 강제되고, a_i는 상수이다. 따라서 $\mu_i = \mu + a_i$ 등도 일정한 상수이며, 모평균간의 산포를

$$\sigma_A^2 = \frac{1}{l-1}\left(\frac{\sum_i m_i a_i^2}{\bar{m}}\right)$$

으로 정의한다. 반면 A_i가 변량인자인 경우, μ_i와 a_i는 샘플링마다 변하는 값이다. 따라서 a_i 역시 분산 σ_A^2에 대하여

$$\mathbb{E}[a_i] = 0, \quad \mathrm{Var}(a_i) = \sigma_A^2 = \mathbb{E}\left[\frac{1}{l-1}\sum_{i=1}^{l}(a_i - \bar{a})^2\right]$$

인 확률변수이며, $\sum_{i=1}^{l} a_i = l\bar{a} \neq 0$일 수 있다.

2.1.2 모형의 분류

실험계획법에서 사용되는 모형은 인자의 종류에 따라 아래와 같이 구분된다.

- **모수모형**: 모수인자만으로 구성된 데이터의 구조모형
- **변량모형**: 변량인자만으로 구성된 데이터의 구조모형
- **혼합모형**: 인자가 두 개 이상이고 모수인자와 변량인자가 섞여 있는 데이터의 구조모형

2.2 분산분석

모수모형을 가정하자. 개개의 데이터 x_{ij}와 총평균 $\bar{\bar{x}}$와의 총편차를 둘로 분해하면

$$(x_{ij} - \bar{\bar{x}}) = (x_{ij} - \bar{x}_{i\cdot}) + (\bar{x}_{i\cdot} - \bar{\bar{x}})$$

으로 나누어진다. 이때 $(x_{ij} - \bar{x}_{i\cdot})$는 측정할 때 수반되는 측정오차나 부정확성을 나타내는 오차로, **잔차**라 불리기도 한다. 다음으로 $(\bar{x}_{i\cdot} - \bar{\bar{x}})$는 각 수준이 가진 효과를 묘사한다. 한편 두 항은 서로 직교하므로,

$$\sum_{i=1}^{l}\sum_{j=1}^{m_i}(x_{ij} - \bar{\bar{x}})^2 = \sum_i\sum_j(\bar{x}_{1\cdot} - \bar{\bar{x}})^2 + \sum_{i=1}^{l}\sum_{j=1}^{m_i}(x_{ij} - \bar{x}_{i\cdot})^2$$

이다. 이때 좌변을 **총제곱합** S_T로 쓰며, 첫째항은 수준에 의한 **급간변동** S_A, 둘째항은 오차에 의한 **급내변동** S_E라 한다. 따라서

$$S_T = S_A + S_E$$

가 성립한다.

한편 아래의 계산법 역시 유용하다.

Theorem 1. *$N = \sum_{i=1}^{l} m_i$라 할 때, **수정항** $CT = \dfrac{T^2}{N}$을 생각하자. 그렇다면 아래가 성립한다.*

$$S_T = \sum_i\sum_j x_{ij}^2 - CT$$

$$S_A = \sum_i \frac{T_{i\cdot}^2}{m_i} - CT$$

$$S_E = S_T - S_A$$

한편 분산분석을 위해서는 자유도를 알아야 한다. 제곱합의 자유도는 제곱을 한 편차의 개수에서 편차들의 선형제약조건 개수를 뺀 것과 동일함이 알려져 있다. 예를 들어, S_T의 계산에는 $(x_{ij}-\bar{\bar{x}})^2$이라는 제곱이 총 N개 존재하는데,

$$\sum_i \sum_j (x_{ij}-\bar{\bar{x}}) = 0$$

이라는 선형제약조건 역시 하나 존재하므로 총 자유도가 $N-1$이 된다. 같은 이유로, ϕ_A의 자유도는 $l-1$, ϕ_E의 자유도는 $N-l$이 된다. 각각의 자유도를 ϕ_T, ϕ_A, ϕ_E로 쓰면 변동에서처럼

$$\phi_T = \phi_A + \phi_E$$

역시 성립함을 알 수 있다. 또한 인자의 효과가 없다는 가설 하에서 $S_T/\sigma_E^2, S_A/\sigma_E^2, S_E/\sigma_E^2$는 각각 자유도가 ϕ_T, ϕ_A, ϕ_E인 카이제곱분포를 따른다. 또 S_A와 S_E는 독립이다.

변동과 자유도를 안다면, 이를 표준화하기 위하여 **평균제곱**을 구할 수 있다. 일반적으로 S_T에 대해서는 구하지 않고, S_A와 S_E를 ϕ_A와 ϕ_E로 나누어

$$V_A = \frac{S_A}{\phi_A}, \quad V_E = \frac{S_E}{\phi_E}$$

를 얻는다. 한편 이들의 비인

$$F_A = \frac{V_A}{V_E} = \frac{S_A/\phi_A}{S_E/\phi_E}$$

은 자유도가 ϕ_A, ϕ_E인 F분포를 따르게 된다. 이는 분산분석에 이용된다.

그 이해를 위해서는 아래의 정리를 볼 필요가 있다.

Theorem 2. *아래가 성립한다.*

$$\mathbb{E}[V_A] = \sigma_E^2 + \bar{m}\sigma_A^2$$
$$\mathbb{E}[V_E] = \sigma_E^2$$

한편 이로부터 $\hat{\sigma}_E^2 = V_E$라 쓰기도 한다. 또한 F는 기대값이 각각 $\sigma_E^2 + \bar{m}\sigma_A^2$와 σ_E^2인 확률변수의 비로, 그 값이 커질수록 σ_A^2이, 즉 a_i의 변동이 커지게 됨을 의미한다. 따라서 **수준간에 특성치의 차이가 없다**는 가설검정의 귀무가설과 대립가설

$$H_0 : a_1 = a_2 = \cdots = a_l = 0, \quad \text{v.s.} \quad H_1 : \exists i \text{ s.t. } a_i \neq 0$$

이

$$H_0 : \sigma_A^2 = 0 \quad \text{v.s.} \quad H_1 : \sigma_A^2 > 0$$

으로 써질 수도 있음을 고려하면, F가 클 경우 이 귀무가설을 기각하는 것이 바람직해진다. 따라서 F검정은 아래와 같이 써진다.

Definition 1. *일원배치법에서의 유의수준 α에서의 F검정은, 만약 F_A의 값이*

$$F_A > F(\phi_A, \phi_E; \alpha)$$

이면 귀무가설을 기각하고, 아니면 채택하는 방식으로 이루어진다. 이때 $F(\phi_A, \phi_E; \alpha)$는 자유도가 ϕ_A, ϕ_E인 F분포의 $1-\alpha$ 퀀타일을 의미한다.

한편 이를 한 눈에 보기 위하여, 아래처럼 분산분석표를 작성하기도 한다.

요인	S	ϕ	V	$\mathbb{E}[V]$	F_0	$F(\alpha)$
A	S_A	$l-1$	V_A	$\sigma_E^2+\bar{m}\sigma_A^2$	V_A/V_E	$F(l-1,N-l;\alpha)$
E	S_E	$N-l$	V_E	σ_E^2		
T	S_T	$N-1$				

실제 데이터 분석 시에는

요인	S	ϕ	V	F_0
A	S_A	$l-1$	V_A	V_A/V_E
E	S_E	$N-l$	V_E	
T	S_T	$N-1$		

정도만 작성하기도 한다.

2.3 분산분석 후의 추정

분산분석을 통해 인자의 유의성을 파악하였다면, 이를 바탕으로 모평균과 같은 다른 특성들을 추정해볼 수 있다. 따라서 분산분석 후의 추정 역시 중요한 과제 중 하나이다.

2.3.1 각 수준의 모평균 추론

모평균 μ_i의 추정에서, 점추정값은

$$\hat{\mu}_i=\widehat{\mu+a_i}=\bar{x}_{i\cdot}$$

이며, $100(1-\alpha)\%$ 신뢰구간은

$$\bar{x}_{i\cdot}\pm t(\phi_E,\alpha/2)\sqrt{\frac{V_E}{m_i}}$$

으로 주어진다.

2.3.2 수준별 모평균차 추론

모평균차에 대한 추론에서 점추정값은

$$\hat{\mu}_i-\hat{\mu}_{i'}=\bar{x}_{i\cdot}-\bar{x}_{i'\cdot}$$

으로 주어지며, $100(1-\alpha)\%$ 신뢰구간은

$$(\bar{x}_{i\cdot}-\bar{x}_{i'\cdot})\pm t(\phi_E,\alpha/2)\sqrt{\left(\frac{1}{m_i}+\frac{1}{m_{i'}}\right)V_E}$$

이다. 이때 두 수준 간에 차이가 유의하려면

$$t(\phi_E,\alpha/2)\sqrt{\left(\frac{1}{m_i}+\frac{1}{m_{i'}}\right)V_E}$$

보다는 모평균차가 커야 하므로, 이 값을 **최소유의차**(LSD)라 부르기도 한다. 특히 반복수가 같은 경우 LSD는 정해진 값이 되므로, 한 번 구해 두면 유의성을 판단하기 쉽다.

2.3.3 오차분산 추론

S_E/σ_E^2이 자유도가 ϕ_E인 카이제곱분포를 따르므로, 오차분산의 $100(1-\alpha)\%$ 신뢰구간은

$$\frac{S_E}{\chi^2(\phi_E;\alpha/2)} \leq \sigma_E^2 \leq \frac{S_E}{\chi^2(\phi_E;1-\alpha/2)}$$

이다.

2.4 변량모형에서의 추론

만약 인자 A가 변량인자인 경우에도, 분산분석표의 작성 방법은 모수인자인 경우와 동일하다. 단, 데이터의 구조식과 분산분석 이후의 추정에서 차이가 생기게 된다.

먼저 데이터 구조식은 아래처럼 표현될 수 있다.

$$\begin{aligned} x_{ij} &= \mu + a_i + e_{ij} \\ a_i &\sim_{i.i.d.} N(0,\sigma_A^2) \\ e_{ij} &\sim_{i.i.d.} N(0,\sigma_E^2) \\ \mathrm{Cov}(a_i, e_{ij}) &= 0 \end{aligned}$$

즉 변량모형임에 따라 a_i가 정해진 모수값이 아니라 확률변수이며, 따라서 $\sum_i m_i a_i$ 따위가 0이 아닐 수 있다. 한편 잘 알려진 바에 따르면,

$$\mathbb{E}[V_A] = \sigma_E^2 + \left(\frac{N^2 - \sum m_i^2}{N(l-1)}\right)\sigma_A^2$$

이다. 만약 m_i가 모두 동일하다면, 이는 $\sigma_E^2 + m\sigma_A^2$와도 같아 모수모형에서와 기대값이 같다. 한편 여기에서 F검정과 같은 인자의 유의성 검정 등은 모수모형과 동일하다.

추론에서는 A가 변량인자이므로 모평균이나 그 차에 대한 추정은 별로 의미가 없고, σ_A^2과 σ_E^2을 추정하여 각 인자/오차에 의한 변동의 크기를 확인할 수 있다.

$$\begin{aligned} \hat{\sigma}_E^2 &= V_E \\ \hat{\sigma}_A^2 &= \frac{V_A - V_E}{\frac{N^2 - \sum m_i^2}{N(l-1)}} = \frac{N(l-1)(V_A - V_E)}{N^2 - \sum m_i^2} \end{aligned}$$

인자수준에 따른 변동 σ_A^2을 확인하여 특성치가 변량인자 A에 얼마나 민감하게 반응하는지를 계측하고는 한다.

Chapter 3

반복이 없는 이원배치법

3.1 모수모형인 경우

모수모형인 경우 이원배치법의 데이터 구조식은 아래와 같다.

$$
\begin{aligned}
x_{ij} &= \mu + a_i + b_j + e_{ij} \\
e_{ij} &\sim_{i.i.d.} N(0, \sigma_E^2) \\
\sum_{i=1}^{l} a_i &= 0 \\
\sum_{j=1}^{m} b_j &= 0
\end{aligned}
$$

반복이 없으므로 실험에서 교호작용의 검출은 포기하며, 제약조건은 모수모형임에 따라 주어진다. $N = lm$ 개의 자료는 아래처럼 배열된다.

	인자의 수준					
	A_1	A_2	$\cdots$	A_l	합	평균
B_1	x_{11}	x_{21}	$\cdots$	x_{l1}	$T_{\cdot 1}$	$\bar{x}_{\cdot 1}$
B_2	x_{12}	x_{22}	$\cdots$	x_{l2}	$T_{\cdot 2}$	$\bar{x}_{\cdot 2}$
B_3	x_{13}	x_{23}	$\cdots$	x_{l3}	$T_{\cdot 3}$	$\bar{x}_{\cdot 3}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
B_m	x_{1m}	x_{2m}	$\cdots$	x_{lm}	$T_{\cdot m}$	$\bar{x}_{\cdot m}$
합계	$T_{1\cdot}$	$T_{2\cdot}$	$\cdots$	$T_{l\cdot}$	T	
평균	$\bar{x}_{1\cdot}$	$\bar{x}_{2\cdot}$	$\cdots$	$\bar{x}_{l\cdot}$		$\bar{\bar{x}}$

한편 여기에서 $x_{ij} - \bar{\bar{x}}$는 아래처럼 세 부분으로 나누어질 수 있다.

$$(x_{ij} - \bar{\bar{x}}) = (\bar{x}_{i\cdot} - \bar{\bar{x}}) + (\bar{x}_{\cdot j} - \bar{\bar{x}}) + (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})$$

또한 각 부분은 직교하므로, 제곱합 역시도

$$\sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2 = \sum_i \sum_j (\bar{x}_{i\cdot} - \bar{\bar{x}})^2 + \sum_i \sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2 + \sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$$

으로 분해될 수 있다. 여기에서 좌변은 총변동 S_T, 오른쪽 항은 차례대로 A에 의한 변동, B에 의한 변동, 오차변동 S_A, S_B, S_E가 된다.

Theorem 3. *실제로 데이터로부터 변동을 계산할 경우에는, 아래 공식을 사용하는 것이 편리하다.* $CT =$

$\frac{T^2}{N}$*이라 할 때,*

$$S_T = \sum_i \sum_j x_{ij}^2 - CT$$
$$S_A = \sum_i \sum_j \frac{T_{i\cdot}^2}{m} - CT$$
$$S_B = \sum_i \sum_j \frac{T_{\cdot j}^2}{l} - CT$$
$$S_E = S_T - S_A - S_B$$

한편 앞서와 동일한 방식으로 논의를 이어가면, S_T의 자유도 ϕ_T는 $N-1$, S_A의 자유도 ϕ_A는 $l-1$, S_B의 자유도 ϕ_B는 $m-1$, S_E의 자유도 ϕ_E는 $(N-1)-(l-1)-(m-1)=(l-1)(m-1)$이다. 따라서 이들로 나누어 평균제곱 V_A, V_B, V_E를 얻게 된다.

Theorem 4. *평균제곱의 기대값은 아래와 같이 주어진다.*

$$\mathbb{E}[V_A] = \sigma_E^2 + m\sigma_A^2$$
$$\mathbb{E}[V_B] = \sigma_E^2 + l\sigma_B^2$$
$$\mathbb{E}[V_E] = \sigma_E^2$$

이를 통해 $\hat{\sigma}_E^2 = V_E$*를 불편추정값으로 사용할 수 있다.*

F통계량은 $F_A = V_A/V_E \sim F(\phi_A, \phi_E)$, $F_B = V_B/V_E \sim F(\phi_B, \phi_E)$이며 적절한 기각역과의 비교를 통해 가설

$$H_0 : a_1 = a_2 = \cdots = a_l$$

또는

$$H_0 : b_1 = b_2 = \cdots = b_m$$

을 검정하게 된다. 따라서 분산분석표라 함은 아래와 같은 형태가 된다.

요인	S	ϕ	V	F_0
A	S_A	$l-1$	V_A	$F_A = V_A/V_E$
B	S_B	$m-1$	V_B	$F_B = V_B/V_E$
E	S_E	$(l-1)(m-1)$	V_E	
T	S_T	$lm-1$		

분산분석 이후의 추정에서는 다양한 추정을 해볼 수 있다. 먼저 인자 A의 i수준에서의 모평균 $\mu(A_i)$의 점추정값은

$$\hat{\mu}(A_i) = \widehat{\mu + a_i} = \bar{x}_{i\cdot}$$

이며, 그 신뢰구간은

$$\bar{x}_{i\cdot} \pm t(\phi_E; \alpha/2)\sqrt{\frac{V_E}{m}}$$

으로 주어진다. 동일한 방식으로 인자 B의 j수준에서의 모평균 $\mu(B_j)$에 대해서는 점추정값과 신뢰구간이 각각

$$\bar{x}_{\cdot j}, \quad \bar{x}_{\cdot j} \pm t(\phi_E; \alpha/2)\sqrt{\frac{V_E}{l}}$$

이다.

한편 2인자 A, B를 조합한 조건에 있어서 모평균을 추정할 수 있다. 일반적으로 모평균의 추정은 인자가 유의할 때만 수행하나, 특정 상황에서는 A나 B를 어떤 수준에 조합하고 모평균의 추정을 수행해볼 수 있다.

A인자의 i수준과 B인자의 j수준에서의 모평균의 점추정값과 구간추정값은

$$\bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{\bar{x}}, \quad \bar{x}_{i\cdot} + \bar{x}_{\cdot j} - \bar{\bar{x}} \pm t(\phi_E; \alpha/2)\sqrt{\frac{V_E}{n_e}}$$

이다. 이때 n_e는 유효반복수

$$n_e = \frac{lm}{l+m-1} = \left(\frac{1}{l} + \frac{1}{m} - \frac{1}{lm}\right)^{-1}$$

이다.

인자수준에 따른 차이를 보고자 하는 경우에는, A의 i수준과 i' 수준의 차이는 점/구간 추정값이

$$\bar{x}_{i\cdot} - \bar{x}_{i'\cdot}, \quad \bar{x}_{i\cdot} - \bar{x}_{i'\cdot} \pm t(\phi_E; \alpha/2)\sqrt{\frac{2V_E}{m}}$$

이며, B의 j수준과 j' 수준의 차이는 점/구간 추정값이

$$\bar{x}_{\cdot j} - \bar{x}_{\cdot j'}, \quad \bar{x}_{\cdot j} - \bar{x}_{\cdot j'} \pm t(\phi_E; \alpha/2)\sqrt{\frac{2V_E}{l}}$$

이 된다.

3.2 난괴법

한 인자는 모수이고 한 인자는 변량인 경우에는 해당 이원배치법을 **난괴법**이라 부른다. 일반적인 일원배치법에 비하여, 난괴법을 이용하면 블록의 영향이 큰 경우 오차분산이 작아지는 장점이 있다. 단 블록간 차이가 크지 않은 경우 오차항의 자유도가 증가하는 것을 막기 위해 풀링하여 일원배치를 사용하는 경우도 있다. 이때의 데이터 구조식은 아래와 같다. A를 모수인자, B를 변량인자라 하자.

$$\begin{aligned}
x_{ij} &= \mu + a_i + b_j + e_{ij} \\
e_{ij} &\sim_{i.i.d.} N(0, \sigma_E^2) \\
b_j &\sim_{i.i.d.} N(0, \sigma_B^2) \\
\mathrm{Cov}(e_{ij}, b_j) &= 0 \\
\sum_{i=1}^{l} a_i &= 0 \\
\sum_{j=1}^{m} b_j &\neq 0
\end{aligned}$$

난괴법에서도 데이터 구조식만 다를 뿐 분산분석의 과정은 앞선 모수모형에서와 같다. 난괴법인 경우에는 분산분석 이후의 추정 과정이 달라진다. B가 변량인자이므로 $\mu(b_j)$ 등의 추론은 의미가 없다. 단 그 산포를 보기 위하여 σ_B^2을 보는 것만 의미가 있다. 물론 B인자가 유의한 경우에만 이를 수행한다. 그 추정량은

$$\hat{\sigma}_B^2 = \frac{V_B - V_E}{l}$$

으로 주어진다. $\mu(A_i)$에 대한 추론에서는, b 역시 확률변수이므로 그 추론 방법이 달라진다. 점추정값과 구간추정값은

$$\bar{x}_{i\cdot}, \quad \bar{x}_{i\cdot} \pm t(\phi*; \alpha/2)\sqrt{\frac{V_B + (l-1)V_E}{lm}}$$

이며, 이때 $\phi*$는 Satterthwaite의 자유도

$$\phi* = \frac{(V_B + (l-1)V_E)^2}{\frac{V_B^2}{\phi_B} + \frac{((l-1)V_E)^2}{\phi_E}}$$

이다. 이는 B인자가 유의하지 않더라도 그런대로 수행할 수 있다. 한편 모평균차에 대한 추론에서는 점추정값과 구간추정값이

$$\bar{x}_{i\cdot} - \bar{x}_{i'\cdot}, \quad (\bar{x}_{i\cdot} - \bar{x}_{i'\cdot}) \pm t(\phi_E; \alpha/2)\sqrt{\frac{2V_E}{m}}$$

으로 주어진다.

3.3 결측치의 취급

실험 중 사고가 있었거나 하는 등의 이유로 특성치와는 무관하게 데이터가 랜덤하게 결측될 수 있다. 이 경우 이원배치법의 형태로 자료를 만들기 위하여 결측치를 채워 줄 필요가 있다. Yates의 결측치 대체법에서는 A_iB_j 수준에서의 특성치 x_{ij}가 결측된 경우, 오차변동 S_E를 최소화하는 y를 선택한다. 이원배치법에서는, 그러한 y가

$$y = \frac{lT'_{i\cdot} + mT'_{\cdot j} - T'}{(l-1)(m-1)}$$

이다. 이때 $T'_{i\cdot} = T_i - x_{ij}$, $T'_{\cdot j} = T_j - x_{ij}, T' = T - x_{ij}$로 x_{ij}를 제외한 열별/행별/전체합을 의미한다. 만약 두 개 이상의 결측치가 있는 경우에도 동일하게 S_E를 각각에 대한 함수로 쓴 뒤 최소화 문제를 풀어 결측치를 대체할 수 있다. 다만 결측치가 많은 경우 실험 자체의 신빙성이 낮아지므로 실험 자체를 다시 하는 것이 올바르다.

한편 그 이후의 추정에 있어서는 기존의 이원배치법/난괴법에서와 유사하나, 결측치를 대체함에 따라 오차변동과 총변동의 자유도 ϕ_E와 ϕ_T가 기존보다 결측치의 개수만큼 빼져야 한다. 예를 들어 결측치가 하나 있는 경우, $\phi_E = (l-1)(m-1) - 1$이 되고 $\phi_T = lm - 2$가 될 것이다. 이에 따라 V_E가 변하고 F_A, F_B도 달라질 수 있음에 따라 인자의 유의성 역시 달라질 수 있다.

추후 다룰 삼원배치법에서도 이처럼 S_E를 최소화하는 값으로 결측치를 대체할 수 있다. $A_iB_jC_k$ 수준에서 결측치가 나타난 경우, 올바른 대체값은

$$y = \frac{T' - lT'_{i\cdot\cdot} - mT'_{\cdot j\cdot} - nT'_{\cdot\cdot k} + lmT'_{ij\cdot} + lnT'_{i\cdot k} + mnT'_{\cdot jk}}{(l-1)(m-1)(n-1)}$$

이 된다.

Chapter 4

반복이 있는 이원배치법

이원배치법에서 반복이 존재하는 경우, 인자와 조합의 효과, 즉 교호작용을 분리할 수 있다는 장점이 있다. 또한 그 분리가 가능하기에 주효과에 대한 추론 역시 가능해지며, 반복수가 증가함에 따라 검정력도 높아지는 장점이 있다. 여기에서는 반복수가 r로 모두 동일한 이원배치법을 생각하여 보도록 하자.

4.1 모수모형인 경우

데이터의 배열은 아래와 같다.

	인자의 수준					
	A_1	A_2	$\cdots$	A_l	합	평균
B_1	x_{111} x_{112} $\vdots$ x_{11r}	x_{211} x_{212} $\vdots$ x_{21r}	$\cdots$ $\cdots$ $\ddots$ $\cdots$	x_{l11} x_{l12} $\vdots$ x_{l1r}	$T_{\cdot 1\cdot}$	$\bar{x}_{\cdot 1\cdot}$
B_2	x_{121} x_{122} $\vdots$ x_{12r}	x_{221} x_{222} $\vdots$ x_{22r}	$\cdots$ $\cdots$ $\ddots$ $\cdots$	x_{l21} x_{l22} $\vdots$ x_{l2r}	$T_{\cdot 2\cdot}$	$\bar{x}_{\cdot 2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_m	x_{1m1} x_{1m2} $\vdots$ x_{1mr}	x_{2m1} x_{2m2} $\vdots$ x_{2mr}	$\cdots$ $\cdots$ $\ddots$ $\cdots$	x_{lm1} x_{lm2} $\vdots$ x_{lmr}	$T_{\cdot m\cdot}$	$\bar{x}_{\cdot m\cdot}$
합계	$T_{1\cdot\cdot}$	$T_{2\cdot\cdot}$	$\cdots$	$T_{l\cdot\cdot}$	T	
평균	$\bar{x}_{1\cdot\cdot}$	$\bar{x}_{2\cdot\cdot}$	$\cdots$	$\bar{x}_{l\cdot\cdot}$		$\bar{\bar{x}}$

한편 보조용으로는 $T_{ij\cdot}$들의 표

	A_1	A_2	$\cdots$	A_l
B_1	$T_{11\cdot}$ $T_{12\cdot}$ $\vdots$ $T_{1m\cdot}$	$T_{21\cdot}$ $T_{22\cdot}$ $\vdots$ $T_{2m\cdot}$	$\cdots$ $\cdots$ $\ddots$ $\cdots$	$T_{l1\cdot}$ $T_{l2\cdot}$ $\vdots$ $T_{lm\cdot}$

를 사용하기도 한다.

한편 모수모형 하에서 데이터의 구조는

$$\begin{aligned} x_{ijk} &= \mu + a_i + b_j + (ab)_{ij} + e_{ijk} \\ e_{ijk} &\sim_{i.i.d.} N(0, \sigma_E^2) \\ \sum_{i=1}^{l} a_i &= 0 \\ \sum_{j=1}^{m} b_j &= 0 \\ \sum_i (ab)_{ij}, \quad \sum_j (ab)_{ij} &= 0 \end{aligned}$$

으로 주어진다.

4.1.1 등분산의 검토

A, B 인자의 조합조건마다 실험오차가 등분산인가를 검토해 보는 것이 구조식의 정당성을 확인하기 위해 필요하다. 반복이 있기 때문에 이것이 가능해진다. 만약 등분산 가정이 성립한다면, 실험 전체가 **관리상태하**에 있다고 말한다.

대표적으로, **R-관리도**를 이용한 등분산 검정의 과정은 아래와 같다.

1. 각 조합조건의 반복 데이터로부터, 데이터가 분포하는 범위 R_{ij}를 구한다.
2. lm개의 범위 R_{ij}들을 이용하여 범위평균 $\bar{R}$을 구한다.
3. 반복수 r에 따라 달라지는 미리 정해진 값 D_4를 곱해 $D_4\bar{R}$을 계산한다.
4. R_{ij}들 중 $D_4\bar{R}$보다 큰 것이 있는지 확인한다. 만약 큰 것이 있다면, 등분산의 가정이 틀렸음을 의미한다.

D_4는 $r = 2, 3, 4, 5, 6, 7, 8$에서 각각 $3.267, 2.575, 2.282, 2.115, 2.004, 1.924, 1.864$ 정도로 주어진다.

4.1.2 변동의 분해와 분산분석

앞서와 동일하지만 반복에 따른 변동까지 포함하여 아래처럼 변동을 분해할 수 있다.

$$(x_{ijk} - \bar{\bar{x}}) = (\bar{x}_{i\cdot\cdot} - \bar{\bar{x}}) + (\bar{x}_{\cdot j\cdot} - \bar{\bar{x}}) + (\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot} - \bar{x}_{j\cdot\cdot} + \bar{\bar{x}}) + (x_{ijk} - \bar{x}_{ij\cdot})$$

또한 이들 변동은 모두 직교하므로, 제곱합 역시

$$\sum_{i,j,k}(x_{ijk} - \bar{\bar{x}})^2 = \sum_{i,j,k}(\bar{x}_{i\cdot\cdot} - \bar{\bar{x}}) + \sum_{i,j,k}(\bar{x}_{\cdot j\cdot} - \bar{\bar{x}})^2 + \sum_{i,j,k}(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot} - \bar{x}_{j\cdot\cdot} + \bar{\bar{x}})^2 + \sum_{i,j,k}(x_{ijk} - \bar{x}_{ij\cdot})^2$$

로 분해할 수 있다. 왼쪽부터 순서대로 $S_T, S_A, S_B, S_{A\times B}, S_E$가 될 것이다.

Theorem 5. *$CT = \dfrac{T^2}{lmr}$에 대하여, $S_T, S_A, S_B, S_{A\times B}, S_E$는 아래처럼 쉽게 계산할 수 있다.*

$$\begin{aligned} S_T &= \sum_{i,j,k} x_{ijk}^2 - CT \\ S_A &= \sum_{i,j,k} \frac{T_{i\cdot\cdot}^2}{mr} - CT \\ S_B &= \sum_{i,j,k} \frac{T_{\cdot j\cdot}^2}{lr} - CT \end{aligned}$$

$$S_{AB} = \sum_{i,j,k} \frac{T_{ij\cdot}^2}{r} - CT$$
$$S_{A\times B} = S_{AB} - S_A - S_B$$
$$S_E = S_T - S_{AB}$$

또한 각각의 자유도는 $\phi_T = lmr - 1, \phi_A = l - 1, \phi_B = m - 1, \phi_{A\times B} = (l-1)(m-1), \phi_E = lm(r-1)$*으로 주어진다.*

앞서와 동일하게, 주어진 자유도로 제곱합들을 나누어 평균제곱 역시 얻어낼 수 있다.

Theorem 6. *평균제곱들의 기대값은 아래와 같이 주어진다.*

$$\mathbb{E}[V_A] = \sigma_E^2 + mr\sigma_A^2$$
$$\mathbb{E}[V_B] = \sigma_E^2 + lr\sigma_B^2$$
$$\mathbb{E}[V_{A\times B}] = \sigma_E^2 + r\sigma_{A\times B}^2$$
$$\mathbb{E}[V_E] = \sigma_E^2$$

이로부터 얻어지는 분산분석표의 형태는 아래와 같다.

요인	S	ϕ	V	F_0
A	S_A	$l-1$	V_A	$F_A = V_A/V_E$
B	S_B	$m-1$	V_B	$F_B = V_B/V_E$
$A\times B$	$S_{A\times B}$	$(l-1)(m-1)$	$V_{A\times B}$	$F_{A\times B} = V_{A\times B}/V_E$
E	S_E	$lm(r-1)$	V_E	
T	S_T	$lmr-1$		

4.1.3 분산분석 후의 추정

A_i 수준에서의 모평균에 대해서는 점추정값과 구간추정값을

$$\bar{x}_{i\cdot\cdot}, \quad \bar{x}_{i\cdot\cdot} \pm t(\phi_E; \alpha/2)\sqrt{\frac{V_E}{mr}}$$

으로, B_j 수준에서의 모평균에 대해서는 점추정값과 구간추정값을

$$\bar{x}_{\cdot j\cdot}, \quad \bar{x}_{\cdot j\cdot} \pm t(\phi_E; \alpha/2)\sqrt{\frac{V_E}{lr}}$$

으로 한다. 모평균차에 대한 추정에서는 A_i와 $A_{i'}$ 수준 간에 대하여

$$\bar{x}_{i\cdot\cdot} - \bar{x}_{i'\cdot\cdot}, \quad (\bar{x}_{i\cdot\cdot} - \bar{x}_{i'\cdot\cdot}) \pm t(\phi_E; \alpha/2)\sqrt{\frac{2V_E}{mr}}$$

으로, B_j와 $B_{j'}$ 수준 간에 대하여

$$\bar{x}_{\cdot j\cdot} - \bar{x}_{\cdot j'\cdot}, \quad (\bar{x}_{\cdot j\cdot} - \bar{x}_{\cdot j'\cdot}) \pm t(\phi_E; \alpha/2)\sqrt{\frac{2V_E}{lr}}$$

으로 한다.

다음으로 두 인자의 수준조합 A_iB_j에서의 모평균 추론에서는 교호작용이 유의하였는지 여부에 따라 방식이

달라진다. 먼저 교호작용이 무시되지 않는 경우, 점추정값과 구간추정값은

$$\bar{x}_{ij\cdot}, \quad \bar{x}_{ij\cdot} \pm t(\phi_E;\alpha/2)\sqrt{\frac{V_E}{r}}$$

으로 해당 수준 내에서의 모평균을 추론에 이용할 수 있다. 한편 이 때문에 최적수준을 구하는 경우 교호작용이 존재한다면 그냥 반복끼리 평균을 구해 가장 높은 특성치를 갖는 수준을 최적수준이라 볼 수 있다.
반면 교호작용이 무시되는 경우에는 $(ab)_{ij}$를 무시하므로, 추정하고자 하는 모수 $\mu(A_iB_j)$는 $\mu + a_i + b_j$가 되며 그 점추정값으로는 $\bar{x}_{i\cdot\cdot} + \bar{x}_{j\cdot\cdot} - \bar{\bar{x}}$를 사용하게 된다. 앞에서와 같이, 여기에서의 유효반복수는

$$n_e = \left(\frac{1}{mr} + \frac{1}{lr} - \frac{1}{lmr}\right)^{-1} = \frac{lmr}{l+m-1}$$

이다. 따라서 점추정값과 구간추정값은

$$\bar{x}_{i\cdot\cdot} + \bar{x}_{j\cdot\cdot} - \bar{\bar{x}}, \quad (\bar{x}_{i\cdot\cdot} + \bar{x}_{j\cdot\cdot} - \bar{\bar{x}}) \pm t(\phi_E;\alpha/2)\sqrt{\frac{V_E}{n_e}}$$

가 될 것이다. 교호작용을 무시하는 경우 최적수준은 여러 번의 조합을 통해 직접 찾아 내야 한다.

4.1.4 오차항에의 풀링

교호작용을 무시하는 경우, 유의하지 않은 교호작용을 오차항에 넣어 새로운 오차항으로 만드는 **풀링**을 수행해볼 수도 있다. 만약 정말로 교호작용이 유의하지 않은 경우에는, 새로운 오차변동 $S_{E'}$을 $S_{E'} = S_E + S_{A\times B}$로 정의할 수 있다. $\sigma^2_{A\times B} = 0$일 때 그 기대값은

$$\mathbb{E}[S_{E'}] = \phi_E\sigma^2_E + \phi_{A\times B}\sigma^2_E = (\phi_E + \phi_{A\times B})\sigma^2_E$$

이므로, 새로운 오차변동 $S_{E'}$은 원래 오차항에 유의하지 않은 교호작용의 오차항을 더해 만들며, 그 자유도는 둘의 자유도의 합인 $\phi_E + \phi_{A\times B}$가 된다. $V_{E'}$ 역시 새롭게 구할 수 있다.
교호작용을 오차항에 풀링할 것인지에 대한 일반적인 원칙은 없으나 아래를 고려한다.

1. 실험의 목적을 고려하여 결정한다. 교호작용의 존재여부가 중요하거나, 존재한다고 알려져 있거나 하는 등의 경우 풀링하지 않을 수 있다.

2. 기술적/통계적인 면을 고려하여 결정한다. 만약 오차분산의 자유도가 $\phi_E > 20$처럼 큰 경우에는 풀링이 실질적인 큰 변화를 주지 않으므로, 풀링을 하는 편이 옳다. 반면 $\phi_E \leq 20$인 경우에는 $F_{A\times B} \leq 1$이면 풀링하고, $1 < FA\times B \leq F(\phi_{A\times B}, \phi_E; 0.1)$ 정도로 애매할 경우 r이 큰 경우에만 풀링한다. 그 외에는 기술적인 면을 고려하여 결정한다.

3. 교호작용에 대한 제2종의 과오를 고려하여 결정한다. 만약 실험상 제2종의 과오를 범하는 것이 큰 잘못이라면, $F_0 \leq 1$인 정도가 아니면 풀링하지 않는다.

4.2 혼합모형의 경우

인자 A가 모수인자이고 인자 B가 변량인자인 경우를 고려하여 보자. 이 경우 데이터 구조식에 변화가 있다.

$$\begin{aligned} x_{ijk} &= \mu + a_i + b_j + (ab)_{ij} + e_{ijk} \\ e_{ijk} &\sim_{i.i.d.} N(0, \sigma_E^2) \\ b_j &\sim_{i.i.d.} N(0, \sigma_B^2) \\ \mathrm{Cov}(b_j, e_{ijk}) &= 0 \\ \sum_{i=1}^{l} a_i &= 0 \\ \sum_{j=1}^{m} b_j &\neq 0 \\ \sum_{i=1}^{l} (ab)_{ij} &= 0 \\ \sum_{j=1}^{m} (ab)_{ij} &\neq 0 \end{aligned}$$

분산분석표를 작성하는 과정은 유사하다. 특히 $V_A, V_B, V_{A\times B}, V_E$를 구하는 과정까지는 동일하다. 다만 문제가 생기는 것은 아래처럼 그 기대값이 달라질 수 있다.

Theorem 7. $\sigma^2_{A\times B}$*를*

$$\sigma^2_{A\times B} = \mathbb{E}\left[\frac{\sum_i \sum_j (ab)_{ij}^2}{m(l-1)}\right]$$

으로 정의할 때(그 분모는 총 제곱합 ml개에서 제약조건 m개 만큼을 뺀 $m(l-1)$임이 합당하다),

$$\begin{aligned} \mathbb{E}[V_A] &= \sigma_E^2 + r\sigma^2_{A\times B} + mr\sigma_A^2 \\ \mathbb{E}[V_B] &= \sigma_E^2 + lr\sigma_B^2 \\ \mathbb{E}[V_{A\times B}] &= \sigma_E^2 + r\sigma^2_{A\times B} \\ \mathbb{E}[V_E] &= \sigma_E^2 \end{aligned}$$

사실 달라진 것은 $\mathbb{E}[V_A]$ 뿐이다. 그러나 이 때문에, F통계량을 구함에 있어 $\sigma_A^2 = 0$을 검정하려면 V_A/V_E 대신 $V_A/V_{A\times B}$를 사용하는 것이 올바르게 된다. 이는 변량모형에서 $\sum_j (ab)_{ij} \neq 0$임에 따라 나타난다. 따라서 혼합모형에서의 분산분석표는 모수모형에서와 살짝 다른

요인	S	ϕ	V	F_0
A	S_A	$l-1$	V_A	$F_A = V_A/V_{A\times B}$
B	S_B	$m-1$	V_B	$F_B = V_B/V_E$
$A\times B$	$S_{A\times B}$	$(l-1)(m-1)$	$V_{A\times B}$	$F_{A\times B} = V_{A\times B}/V_E$
E	S_E	$lm(r-1)$	V_E	
T	S_T	$lmr-1$		

이 될 것이다.

한편 여기에서 분산분석 후 추정은 변량인자 B의 산포 $\hat{\sigma}_B^2$에 대한 논의로부터 시작한다. 만약 B의 효과가 유의하지 않은 것으로 드러났다면 딱히 수행할 필요는 없다. 만약 B의 효과가 유의하였다면,

$$\hat{\sigma}_B^2 = \frac{V_B - V_E}{lr}$$

으로 구하며,

$$\hat{\sigma}^2_{A\times B} = \frac{V_{A\times B} - V_E}{r}$$

으로 추정한다.

$\mu(A_i)$에 대한 추정에서는 점추정값은 동일하게 $\bar{x}_{i\cdot\cdot}$을 사용할 수 있다. 그러나 이 표본평균에서 b_j 역시 확률변수임에 따라,

$$\mathrm{Var}(\bar{x}_{i\cdot\cdot}) = \frac{\sigma^2_B}{m} + \frac{\sigma^2_{A\times B}}{m} + \frac{\sigma^2_E}{mr}$$

으로 복잡해진다. 이들의 추정값으로 대체하면, 교호작용이 유의한 경우,

$$\widehat{\mathrm{Var}}(\bar{x}_{i\cdot\cdot}) = \frac{1}{lmr}(V_B + lV_{A\times B} - V_E)$$

이다. 따라서 그 구간추정값은

$$\bar{x}_{i\cdot\cdot} \pm t(\phi*; \alpha/2)\sqrt{\frac{V_B + lV_{A\times B} - V_E}{lmr}}$$

으로 나타나며, $\phi*$는 Satterthwaite의 자유도

$$\phi* = \frac{(V_B + lV_{A\times B} - V_E)^2}{\frac{V_B^2}{\phi_B} + \frac{(lV_{A\times B})^2}{\phi_{A\times B}} + \frac{V_E^2}{\phi_E}}$$

이다. 반면 교호작용이 유의하지 않은 경우에는 $\sigma^2_{A\times B} = 0$이므로 분산추정량이

$$\widehat{\mathrm{Var}}(\bar{x}_{i\cdot\cdot}) = \frac{1}{lmr}(V_B + (l-1)V_E)$$

가 될 것이며, 구간추정값은

$$\bar{x}_{i\cdot\cdot} \pm t(\phi*; \alpha/2)\sqrt{\frac{V_B + (l-1)V_E}{lmr}}$$

이다. $\phi* = \frac{(V_B + (l-1)V_{A\times B})^2}{\frac{V_B^2}{\phi_B} + \frac{((l-1)V_E)^2}{\phi_E}}$이다.

마지막으로 두 수준 간의 모평균차 $\mu(A_i) - \mu(A_{i'})$에 대해서는, 간단하게 점추정값과 구간추정값이

$$\bar{x}_{i\cdot\cdot} - \bar{x}_{i'\cdot\cdot}, \quad (\bar{x}_{i\cdot\cdot} - \bar{x}_{i'\cdot\cdot}) \pm t(\phi_{A\times B}; \alpha/2)\sqrt{\frac{2V_{A\times B}}{mr}}$$

으로 나타날 것이다.

4.3 결측치의 취급

모수모형이든 변량모형이든, 결측치가 발생하였을 경우에는 앞서와 마찬가지로 S_E를 최소화하는 Yates의 대체법을 사용할 수 있다. 반복이 있는 경우, A_i, B_j 수준의 s번째 반복에서 발생한 결측치의 값이

$$\hat{y} = \bar{x}'_{ij} = \frac{\sum_{k=1}^{r} x_{ijk} - x_{ijs}}{r-1}$$

로 나머지 $r-1$개 반복의 특성치 평균으로 대체되는 것이 적당함을 알 수 있다. 두 개 이상인 경우에도 동일한 방식으로 수행하면 된다. 단, 결측치의 수만큼 총변동와 오차변동의 자유도는 감소하게 됨을 유의하자.

4.4 대비와 직교분해

Definition 2. *n개의 측정치 $x_1, \cdots, x_n$의 정수계수 1차식*

$$L = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

*을 **선형식**이라 부르고, 계수들의 제곱합을 선형식 L의 **단위수** D라고 한다. $D > 0$인 경우를 선형식으로 한정하기도 한다. 이때 **선형식의 변동**은*

$$S_L = \frac{L^2}{D}$$

로 주어지며, 자유도 1을 갖는다.
한편 측정치 $m_1, \cdots, m_l$개의 합 $T_{1\cdot}, T_{2\cdot}, \cdots, T_{l\cdot}$을 고려할 때, 선형식

$$L = c_1T_{1\cdot} + c_2T_{2\cdot} + \cdots + c_lT_{l\cdot}$$

*이 만약 $m_1c_1 + m_2c_2 + \cdots + m_lc_l = 0$을 만족한다면 이 선형식을 **대비**라 부르며, 그 변동은*

$$S_L = \frac{L^2}{(\sum m_ic_i^2)}$$

으로 자유도는 1이다.
*2개의 대비 L_1와 L_2에 대하여 그들의 계수 c와 c'이 벡터 m에 대해 서로 직교한다면, 두 대비는 **직교**한다고 말한다. 특히 그들이 직교하면 S_{L_1}과 S_{L_2}는 자유도 1의 변동이다. 특히 l개 수준이 존재하는 인자 A에서, 대비 $L_1, L_2, \cdots, L_{l-1}$이 모두 직교하도록 설정하면*

$$S_A = S_{L_1} + \cdots + S_{L_{l-1}}$$

*과 같이 **직교분해**할 수도 있다.*
한편 대비를 구한 뒤에는 $\mathbb{E}[L] = 0$을 가설검정할 수도 있고, S_A를 대비들의 제곱합으로 분해하여 어떤 대비의 변동이 S_A의 큰 부분을 차지하고 있는지 알아 볼 수 있다. 그 유의성에 대한 가설검정은 $V_{L_1} = S_{L_1}$을 구한 뒤 V_E로 나누어 F통계량을 만들고, 그것이 귀무가설 하에서 $F(1, \phi_E)$를 따름을 이용한다. 특정한 형태의 가설을 가지고 있는 경우, 그 목적을 이루기 위해서는 적절한 대비와 적절한 직교분해를 수행할 필요가 있다.

4.5 계수치 데이터의 분석

만약 특성치가 계량변수가 아니라 성공/실패 여부 등의 계수변수라도 동일한 방법으로 분석할 수 있다. 데이터가 충분히 많다면, 즉 반복수 r와 실제 성공 확률 p에 대하여 $rp > 5, r(1-p) > 5$ 정도의 조건이 성립하여 정규분포로의 근사가 가능하다면, 분산분석이 여전히 유효하다. 따라서 데이터를 모두 1과 0으로만 가지고 있다고 생각하고 앞선 논의를 그대로 수행하면 된다.

Chapter 5

다원배치법

이원배치법을 더욱 확장하여, 취급하고 싶은 인자가 3개 이상인 경우에도 모든 인자의 수준조합 하에서 실험을 수행할 수 있다. 이러한 **다원배치법**에서는 인자의 수가 늘어날수록 실험의 횟수가 증가하고 랜덤화가 어려워지는 등의 단점이 있다.

5.1 평균제곱의 기대값을 구하는 방법

다원배치법에서 분산분석표를 얻고 인자의 유의성을 검정할 때, 평균제곱의 기대값을 알아야 올바른 F통계량을 만들고 검정을 수행할 수 있다. 시험에 나올 수 있는 가장 복잡한 형태인 반복수가 r인 삼원배치법 하에서 A, B가 모수인자이고 C인자가 변량인자일 때의 상황을 고려하여 보자. A인자가 i수준, B인자가 j수준, C인자가 k수준, 반복수가 p일 때의 특성치를 x_{ijkp}라 하자. 완전히 랜덤화는 되지 않고 ijk를 고정한 뒤 반복 r번을 수행한다고 생각하자.

1. 아래 표와 같이 행에 순서대로 $A_i, B_j, C_k, (AB)_{ij}, (AC)_{ik}, (BC)_{jk}, (ABC)_{ijk}, E_{p(ijk)}$를 기입하고, 열에는 각 인자의 수준수, 인자의 종류(F/R), 첨자를 기입한다. 반복의 경우 변량인자로 취급한다.

	l	m	n	r
	F	F	R	R
A_i				
B_j				
C_k				
$(AB)_{ij}$				
$(AC)_{ik}$				
$(BC)_{jk}$				
$(ABC)_{ijk}$				
$E_{p(ijk)}$				

2. 각 행에 대하여 같은 첨자가 있는 곳은 제외하고 나머지 장소에 수준수를 기입한다.

	l	m	n	r
	F	F	R	R
A_i		m	n	r
B_j	l		n	r
C_k	l	m		r
$(AB)_{ij}$			n	r
$(AC)_{ik}$		m		r
$(BC)_{jk}$	l			r
$(ABC)_{ijk}$				r
$E_{p(ijk)}$				

3. 행에 있는 첨자 중 괄호에 들어 있는 것이 있다면, 그 행에서 괄호에 들어 있는 첨자에 해당하는 열을 1로 채운다.

	l	m	n	r
	F	F	R	R
A_i		m	n	r
B_j	l		n	r
C_k	l	m		r
$(AB)_{ij}$			n	r
$(AC)_{ik}$		m		r
$(BC)_{jk}$	l			r
$(ABC)_{ijk}$				r
$E_{p(ijk)}$	1	1	1	

4. 빈 곳을 찾아서 열이 F면 0을, R이면 1을 기입한다.

	l	m	n	r
	F	F	R	R
A_i	0	m	n	r
B_j	l	0	n	r
C_k	l	m	1	r
$(AB)_{ij}$	0	0	n	r
$(AC)_{ik}$	0	m	1	r
$(BC)_{jk}$	l	0	1	r
$(ABC)_{ijk}$	0	0	1	r
$E_{p(ijk)}$	1	1	1	1

5. 각 행별로 포함된 첨자가 있는 열을 가리고, 행별로 적혀진 숫자를 곱한 뒤, 첨자가 포함된 행의 분산을 차례로 곱하여 합한다. 이것이 $\mathbb{E}[V]$가 된다.

	l	m	n	r	
	F	F	R	R	$\mathbb{E}[V]$
A_i	0	m	n	r	$\sigma_E^2 + mr\sigma_{A\times C}^2 + mnr\sigma_A^2$
B_j	l	0	n	r	$\sigma_E^2 + lr\sigma_{B\times C}^2 + lnr\sigma_B^2$
C_k	l	m	1	r	$\sigma_E^2 + lmr\sigma_C^2$
$(AB)_{ij}$	0	0	n	r	$\sigma_E^2 + r\sigma_{A\times B\times C}^2 + nr\sigma_{A\times B}^2$
$(AC)_{ik}$	0	m	1	r	$\sigma_E^2 + mr\sigma_{A\times C}^2$
$(BC)_{jk}$	l	0	1	r	$\sigma_E^2 + lr\sigma_{B\times C}^2$
$(ABC)_{ijk}$	0	0	1	r	$\sigma_E^2 + r\sigma_{A\times B\times C}^2$
$E_{p(ijk)}$	1	1	1	1	σ_E^2

이후에는 검정하고자 하는 가설에 맞게 F통계량을 정의함으로써 적절한 검정통계량을 얻을 수 있다. 예를 들어, $\sigma_A^2 = 0$을 검정하고자 하는 경우 올바른 비교를 위해서는 F통계량을 $V_A/V_{A\times C}$로 정의해야 함을 알 수 있다.

5.2 분산분석과 그 이후의 추정

분산분석과 그 이후의 추정은 $\mathbb{E}[V]$ 자료를 바탕으로 수행된다. 다만 주의할 것은 분산분석 이후의 추정방법이 주효과만 유의한 경우, 일부 교호작용만 유의한 경우, 모두 유의한 경우 등에 따라, 그리고 어떤 인자가 변량인자인지에 따라 달라지는데, 이는 상황을 바탕으로 어떤 추정량을 점추정량으로 사용할지와 그 점추정량의 분산이 어떻게 결정되는지에 따라 유효반복수 n_e 혹은 자유도 $\phi*$를 직접 계산해야 하는 복잡함이 있다. 따라서 이는 문제풀이 상황마다 매우 달라지므로, 여기에서는 생략한다.

Chapter 6

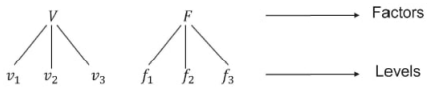
분할법

분할법은 실험의 랜덤화가 곤란한 경우 자주 사용하는 실험 방법으로 **일차단위**를 먼저 랜덤화한 뒤 그를 고정하고 다시 **이차단위**의 랜덤화를 수행하는 상황을 이야기한다. 분할법이 가지는 특징은 아래와 같다.

- 실험을 실시할 때 대개 일차단위를 수준변경이 어려운 것으로, 이차단위를 수준변경이 용이한 것으로 설정한다. 삼차단위 이상이 존재하는 경우, 수준변경이 쉬울수록 고차단위에서 랜덤화한다.
- 오차가 일차단위의 오차, 이차단위의 오차 등으로 분할되는 특성을 가진다. 일차단위의 오차의 기대치가 이차단위의 오차의 기대치보다 크며, 자유도는 일차오차가 더 작다.
- 이차인자 B나 교호작용 $A \times B$의 효과가 일차인자 A의 효과보다 더 정도 좋게 추정된다.
- 일차오차를 무시할 수 있다면 그 오차를 이차오차에 풀링할 수 있으며, 그 경우 보통의 다원배치법과 동일한 구성을 가진다.
- 분할에 따라 다원배치법에 비해 추정과 검정이 더욱 복잡하다.
- 일차단위 인자와 이차단위 인자의 교호작용은 이차단위에 속하는 요인이 된다.

6.1 일차단위가 일원배치일 때의 단일분할법

단일분할법은 한 번만 분할이 일어나는 것으로, 일차단위와 이차단위로만 분할되는 가장 쉬운 분할법 중 하나이다. 특히 일차단위가 일원배치라 함은 일차단위에 인자가 하나만 있는 경우를 의미한다. 예를 들어 아래 그림에서는 일차단위가 V, 이차단위가 f가 된다.



Block I		Block II		Block III	
V_2	f_1	V_3	f_1	V_1	f_1
	f_2		f_3		f_2
	f_3		f_2		f_3
V_1	f_2	V_2	f_2	V_3	f_3
	f_3		f_3		f_2
	f_1		f_1		f_1
V_3	f_1	V_1	f_1	V_2	f_2
	f_2		f_3		f_1
	f_3		f_2		f_3

6.1.1 랜덤화와 데이터의 구조

일차단위가 A, 이차단위가 B이고 반복이 r번 존재하는 분할법에서는 아래의 데이터 구조식을 가정한다.

$$x_{ijk} = \mu + \underbrace{a_i + r_k + e_{(1)ik}}_{\text{일차단위}} + \underbrace{b_j + (ab)_{ij} + e_{(2)ijk}}_{\text{이차단위}}$$

여기서 **일차단위오차** $e_{(1)ik}$는 $(ar)_{ik}$와 교락되어 있고, 이에 따라 $A \times R$은 검출이 불가능하다. 즉 반복 $R_1, \cdots, R_r$에서 각각 $A_1, \cdots, A_l$을 랜덤화하고, 각 A_i에서 다시 이차적으로 $B_1, \cdots, B_m$을 랜덤화하는 상황을 고려하는 것이다. **이차단위오차** $e_{(2)ijk}$는 $(br)_{jk}, (abr)_{ijk}$와 교락되어 있다. 이러한 분할법은 결국 $(ab)_{ij}$라는 교호작용만을 정도 있게 검출할 수 있으며, 실험 목적이 그러한 경우에만 사용할 수 있다.

6.1.2 분산분석표의 작성

$S_T, S_A, S_B, S_{A\times B}$를 구하는 과정은 한 인자가 변량인자인 삼원배치법에서와 동일하다. 단, S_E를 일차단위오차에 의한 것 S_{E_1}과 이차단위오차에 의한 것 S_{E_2}로 분해할 수 있게 된다.
S_{E_1}은 삼원배치법에서 $S_{A\times R}$를 구하는 것과 동일하게,

$$S_{E_1} = S_{AR} - S_A - S_R$$

으로 구할 수 있다. 그 다음으로 S_{E_2}는 S_E에서 S_{E_1}을 뺌으로써 구한다. 한편 $S_{E_2} = S_{B\times R} + S_{A\times B\times R}$을 통해 구해도 괜찮다.

Definition 3. *일차단위가 일원배치일 때의 단일분할법에서, 분산분석표는 아래의 형태를 따른다.*

요인	S	ϕ	V	F_0
A	S_A	$l-1$	V_A	$F_A = V_A/V_{E_1}$
R	S_R	$r-1$	V_R	$F_R = V_R/V_{E_1}$
E_1	S_{E_1}	$(l-1)(r-1)$	V_{E_1}	$F_{E_1} = V_{E_1}/V_{E_2}$
B	S_B	$m-1$	V_B	$F_B = V_B/V_{E_2}$
$A\times B$	$S_{A\times B}$	$(l-1)(m-1)$	$V_{A\times B}$	$F_{A\times B} = V_{A\times B}/V_{E_2}$
E_2	S_{E_2}	$l(m-1)(r-1)$	V_E	
T	S_T	$lmr-1$		

즉 이는 복잡해 보이지만 사실 삼원배치법의 특별한 한 형태와 동일하며, 다원배치법에서의 분석법을 그대로 이용하여 분석할 수 있다.

6.1.3 분산분석 후의 추정

만약 R과 E_1이 유의하지 않다면, E_2에 풀링할 수 있다. 이 경우 분산분석 후의 추정은 반복이 있는 이원배치법과 동일하게 된다. 그러나 R과 E_1이 유의하다면, 그에 맞게 추론을 해야만 한다. 예를 들어 A_i 수준에서의 $\mu(A_i)$에 대한 추론을 하는 경우, 점추정값은 $\bar{x}_{i\cdot\cdot}$인 것이 자연스럽다. 그런데

$$\bar{x}_{i\cdot\cdot} = \mu + a_i + \bar{r} + \bar{e}_{(1)i\cdot} + \bar{e}_{(2)i\cdot\cdot}$$

이므로 그 분산이

$$\mathrm{Var}(\bar{x}_{i\cdot\cdot}) = \frac{\sigma_R^2}{r} + \frac{\sigma_{E_1}^2}{r} + \frac{\sigma_{E_2}^2}{mr}$$

이다. 만약 R이나 E_1 중 유의하지 않은 것이 있다면 그 항은 제거된다. 그런데 다원배치법에서 V의 기대값을 구하는 방법에 의하여

$$\hat{\sigma}_R^2 = \frac{V_R - V_{E_1}}{lm}, \quad \hat{\sigma}_{E_1}^2 = \frac{V_{E_1} - V_{E_2}}{m}, \quad \hat{\sigma}_{E_2}^2 = V_{E_2}$$

이므로 분산추정값이

$$\widehat{\mathrm{Var}}(\bar{x}_{i\cdot\cdot}) = \frac{1}{lmr}[V_R + (l-1)V_{E_1}]$$

이 된다. 따라서 Satterhwaite의 자유도 $\phi*$를 기반으로 한 추론을 해야만 한다. 이처럼 분할법에서의 추론은 인자의 유의성을 파악한 뒤 적절한 점추정값을 구하고, 해당 점추정값의 분산을 구한 뒤, 이를 분산추정량으로 대체하고, 그에 상응하는 유효반복수 n_e/자유도 $\phi*$를 구해 추론하는 방식으로 이루어진다. 따라서 다원배치법과 마찬가지로 일반적인 이야기를 하기 어렵다.

6.2 일차단위가 이원배치일 때의 단일분할법

인자 A, B, C가 있는 3인자의 실험에서 C만 랜덤화가 용이하다면, A와 B를 일차단위의 인자로 두는 방법을 생각해볼 수 있다. 일차단위가 반복이 없다고 해보자. 이때 데이터의 구조식은

$$x_{ijk} = \mu + \underbrace{a_i + b_j + e_{(1)ij}}_{\text{1차단위}} + \underbrace{c_k + (ac)_{ik} + (bc)_{jk} + e_{(2)ijk}}_{\text{2차단위}}$$

으로 주어진다. 일차단위의 반복이 없으므로 $(ab)_{ij}$는 $e_{(1)ij}$와 교락되어 있고, $(abc)_{ijk}$는 $e_{(2)ijk}$와 교락된다. 이후에는 앞서 소개한 것처럼 $\mathbb{E}[V]$를 구하고 그를 통해 분산분석표를 작성하면 된다. 분산분석 이후의 추정도 마찬가지로 인자의 유의성에 따라 적절한 점추정량을 구하고, 그 분산추정량으로써 추론하면 된다.

6.3 이단분할법

인자가 세 개 이상일 때에 랜덤화가 어려운 정도에 따라 두 번 순차적으로 분할실험할 수 있다. 이러한 **이단분할법**의 대표적인 예시는 아래와 같다. 인자 A, B, C가 있고 A를 일차단위, B를 이차단위, C를 삼차단위로 하며 r번의 반복실험을 했다고 하자. 그렇다면 데이터의 구조는 아래와 같다.

$$x_{ijkp} = \mu + \underbrace{a_i + r_p + e_{(1)ip}}_{\text{일차단위}} + \underbrace{b_j + (ab)_{ij} + e_{(2)ijp}}_{\text{이차단위}} + \underbrace{c_k + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + e_{(3)ijkp}}_{\text{삼차단위}}$$

이에 맞게 추론을 계속적으로 수행하면 된다. 이러한 복잡한 계산에서 $\phi_{A\times B} = \phi_A \times \phi_B$, $\phi_{E_2} = \phi_{B\times R} + \phi_{A\times B\times R}$와 같은 테크닉들을 잘 사용할 수 있다. R이 e들과 교락됨에 따라, E_1, E_2, E_3가 $(ar)_{ip}$ 따위들과 교락 불가능하게 되고, 이에 E_s가 s차단위의 인자들과 반복의 교락들의 합들로 이루어지기 때문이다.

6.4 인자가 분할이 안 되는 경우

인자가 분할이 되지 않는 경우, 인자수준을 고정시켜 두고 반복만 해당 수준 내에서 할 때가 있다. 이 경우에는 반복 자체를 이차단위로, 인자들을 일차단위로 두는 분할법으로 생각할 수 있다. 예를 들어 A, B인자를 일차단위로 하고 반복을 이차단위로 하는 경우 데이터의 구조는

$$x_{ijk} = \mu + a_i + b_j + e_{(1)ij} + e_{(2)ijk}$$

처럼 나타나며 $e_{(1)ij}$는 $(ab)_{ij}$와 교락된다. 여기에서 동일하게 추론을 수행하면 된다.

그 외에도 인자수준의 의미가 분할의 가지마다 달라지는 **지분실험법**과 같이 분할법의 특정한 예시들이 있는데, 이는 데이터 구조식을 적절히 세운 뒤 앞에서처럼 수행하면 해결된다.

Chapter 7

라틴방격법

m개의 숫자 혹은 글자를 어느 행, 어느 열에도 하나씩만 있게끔 나열하는 것을 $m \times m$ **라틴방격**이라 한다. 예를 들어, 아래는 3×3 라틴방격이 된다.

$$C_1\ C_2\ C_3$$
$$C_2\ C_3\ C_1$$
$$C_3\ C_1\ C_2$$

특히 1행과 1열이 순서대로 나열된 경우 **표준라틴방격**이라 부르며, 일반적으로 $m \times m$ 라틴방격에서 가능한 배열방법의 수는 표준라틴방격의 수에 $m! \times (m-1)!$을 곱한 만큼이라고 알려져 있다.
이제 수준이 3개씩 있는 인자 A, B, C에 대한 라틴방격법을 고려하여 보자. 그렇다면 아래와 같이 인자수준을 배치하는 상황을 고려해 볼 수 있다.

	A_1	A_2	A_3
B_1	C_1	C_2	C_3
B_2	C_2	C_3	C_1
B_3	C_3	C_1	C_2

그렇다면 모든 인자수준에서 실험을 수행하는 것이 아니라, 9번의 수준 $A_1B_1C_1, A_1B_2C_2, A_1B_3C_3, A_2B_1C_2, A_2B_2C_3, A_2B_3C_1, A_3B_1C_3, A_3B_2C_1, A_3B_3C_2$에서만 실험을 수행하게 된다. 한 인자의 각 수준에서 수행한 실험에는 다른 인자의 각 수준 조건이 모두 균등하게 들어 있고, 이에 따라 각 수준 간의 산포를 구하면 다른 요인에 영향을 받지 않고 인자의 효과를 볼 수 있게 된다. 라틴방격은 대개 3인자 실험에서 각 인자의 수준수가 m로 모두 동일한 상황에서만 사용할 수 있다. 실험 횟수가 m^2번으로 삼원배치법에서의 m^3번보다 매우 감소한다는 장점이 있지만, 교호작용을 검출할 수 없다는 단점을 동반한다.

7.1 라틴방격법에서의 분산분석

라틴방격법은 주로 모수모형에서 이용되며, 데이터 구조식은

$$x_{ijk} = \mu + a_i + b_j + c_k + e_{ijk}$$

형태이다. 한편 여기에서 분산분석표의 작성은 삼원배치법에서 S_A, S_B, S_C를 구하는 방법과 완전히 동일하며, 각각의 자유도는 $m-1$로 주어질 것이다. 한편 오차항의 자유도는 총 자유도 m^2-1에서 $3(m-1)$을 뺀 $(m-1)(m-2)$이 된다. 따라서 $m \geq 3$일 때에만 유효하다. 분산분석표의 형태는
위처럼 되며, $\mathbb{E}[V_A] = \sigma_E^2 + m\sigma_A^2$으로 나타난다. V_B, V_C의 경우 σ_A^2의 위치만 σ_B^2과 σ_C^2으로 대체하면 된다. F검정의 검정통계량 역시 매우 합당한 방식으로 얻어진다.

요인	S	ϕ	V	F_0
A	S_A	$m-1$	V_A	$F_A = V_A/V_E$
B	S_B	$m-1$	V_B	$F_B = V_B/V_E$
C	S_C	$m-1$	V_C	$F_C = V_C/V_E$
E	S_E	$(m-1)(m-2)$	V_E	
T	S_T	m^2-1		

한편 분산분석 이후의 추정에서는 각 인자 수준에서 특성치의 모평균 추정이 의미가 있다. A가 유의한 경우 $\mu(A_i)$에 대한 점추정량과 구간추정량은

$$\bar{x}_{i\cdot\cdot}, \quad \bar{x}_{i\cdot\cdot} \pm t(\phi_E;\alpha/2)\sqrt{\frac{V_E}{m}}$$

이 될 것이다. 그 다음으로 두 인자가 유의한 경우, 두 인자의 수준조합에서 모평균을 추정하는 것 역시 가능하다. A_iB_j 수준에서의 점추정량과 구간추정량은

$$\bar{x}_{i\cdot\cdot} + \bar{\cdot j\cdot} - \bar{\bar{x}}, \quad (\bar{x}_{i\cdot\cdot} + \bar{\cdot j\cdot} - \bar{\bar{x}}) \pm t(\phi_E;\alpha/2)\sqrt{\frac{V_E}{n_e}}$$

으로 주어지며, 유효반복수 n_e는 $\frac{1}{n_e} = \frac{1}{m} + \frac{1}{m} - \frac{1}{m^2}$으로부터 $\frac{m^2}{2m-1}$이 된다. 세 인자가 유의한 경우 $A_iB_jC_k$ 수준에서의 점추정량은

$$\bar{x}_{i\cdot\cdot} + \bar{x}_{\cdot j\cdot} + \bar{x}_{\cdot\cdot k} - 2\bar{\bar{x}}$$

이 될 것이고, 유효반복수 $n_e = \frac{m^2}{3m-2}$ 하에서 마찬가지로 신뢰구간을 구할 수 있을 것이다.

7.2 그레코라틴방격법

두 개의 3×3 라틴방격

1 2 3 1 3 2
2 3 1 2 1 3
3 1 2 3 2 1

을 고려하자. 이 둘을 겹치어 합하면 아래의 격자가 될 것이다.

11 23 32
22 31 13
33 12 21

이는 1, 2, 3의 2개씩의 모든 조합을 포함하고 있고, 중복도 없고, 이러한 경우 두 개의 방격이 서로 직교한다고 말한다. 또한, 직교하는 두 개의 라틴방격을 종합한 방격을 **그레코라틴방격**이라 한다.

그레코라틴방격을 이용하면 4인자로 라틴방격법을 확장하여 아래처럼 그레코라틴방격법을 얻는다.

	A_1	A_2	A_3
B_1	C_1D_1	C_2D_3	C_3D_2
B_2	C_2D_2	C_3D_1	C_1D_3
B_3	C_3D_3	C_1D_2	C_2D_1

이때의 구조식은

$$x_{ijkp} = \mu + a_i + b_j + c_k + d_p + e_{ijkp}$$

가 되며, 라틴방격에서와 동일한 방식으로 분산분석표를 작성해줄 수 있다. 결과적으로 얻는 분산분석표의

형태는 아래와 같게 된다.

요인	S	ϕ	V	F_0
A	S_A	$m-1$	V_A	$F_A = V_A/V_E$
B	S_B	$m-1$	V_B	$F_B = V_B/V_E$
C	S_C	$m-1$	V_C	$F_C = V_C/V_E$
D	S_D	$m-1$	V_D	$F_D = V_D/V_E$
E	S_E	$(m-1)(m-3)$	V_E	
T	S_T	m^2-1		

각 수준조합에서의 모평균 추정은 역시 앞에서와 마찬가지로 적절한 점추정량을 구하고 해당 점추정량의 분산을 구하고, 유효반복수 n_e를 계산하여 신뢰구간을 제작함으로써 만든다. 이때 주의할 것은 인자의 어떤 조합에서 실제로는 실험이 수행되지 않았더라도, 모평균의 추정은 가능하다는 것이다. 예를 들어 $A_1B_1C_2D_3$ 수준에서의 실험은 실제 행해지지 못했지만, 교호작용을 검출하지 않고 주효과만으로 a_1, b_1, c_2, d_3을 추정하였기에 그 합으로써 모평균의 추정이 가능하다는 것이다. 따라서 최적조건은 단지 데이터 중 최대의 특성치를 보이는 수준조합이 아니라, 각 인자별로 최대인 수준들을 조합하여 만들어진다.

7.3 초그레코라틴방격법

서로 직교하는 라틴방격을 3개 조합하여 만든 방격을 **초그레코라틴방격**이라 하며, 이를 이용해 5인자 실험을 라틴방격법을 통해 수행할 수 있다. 여기에서도 마찬가지 방법으로 데이터 구조식을 세운 뒤 분산분석표를 작성하여 분석을 진행한다.

Chapter 8

요인배치법

일반적인 m^n **요인배치법**은 인자의 수가 n이고 각 인자의 수준이 m인 실험계획법으로, 모든 인자간의 수준의 조합에서 실험이 이루어지는 실험이다. 즉 이는 다원배치법의 일종이라고 할 수 있다. 이는 모든 요인효과를 추정할 수 있다는 장점이 있으나, 실험횟수가 굉장히 커 $m = 2, 3$에 주로 제한된다는 특징이 있다.

8.1 2^2 요인실험

8.1.1 반복이 없는 경우

인자 A, B에 수준수가 각각 2인 실험을 고려하자. 각각의 수준이 A_0, A_1, B_0, B_1이며 데이터는 y로 나타내자. 그렇다면 데이터의 구조식은

$$y_{ij} = \mu + a_i + b_j + (ab)_{ij}$$

으로 표현되며, 반복이 없기에 $(ab)_{ij}$는 오차항과 교락되어 있다. 따라서 오차항을 쓰지 않는다. 한편 A와 B가 모수인자이기에 아래의 제약조건이 존재한다.

$$\begin{aligned} a_0 + a_1 &= 0 \\ b_0 + b_1 &= 0 \\ (ab)_{00} + (ab)_{10} &= 0 \\ (ab)_{01} + (ab)_{11} &= 0 \\ (ab)_{00} + (ab)_{01} &= 0 \\ (ab)_{10} + (ab)_{11} &= 0 \end{aligned}$$

따라서 이로부터 얻은 정보를 대입하면, 아래의 데이터 구조식을 얻을 수 있다.

$$\begin{aligned} y_{00} &= \mu + a_0 + b_0 + (ab)_{00} \\ y_{01} &= \mu + a_0 - b_0 - (ab)_{00} \\ y_{10} &= \mu - a_0 + b_0 - (ab)_{00} \\ y_{11} &= \mu - a_0 - b_0 + (ab)_{00} \end{aligned}$$

따라서 위의 연립방정식을 풀면 $\mu, a_0, b_0, (ab)_{00}$를 y_{ij}의 함수로서 얻을 수 있다. 특히 y_{ij}의 값을 $a^i b^j$처럼 쓰면, 인자 A의 두 수준 간의 효과의 차이 $a_1 - a_0$를 A로 쓸 때

$$A = a_1 - a_0 = -2a_0 = \frac{1}{2}(ab + a - b - (1))$$

으로 인자 A의 **주효과**가 A_1일 때의 특성치 평균에서 A_0일 때의 특성치 평균을 뺀 값이 됨을 알 수 있다. 동일한 이유로, B의 주효과는

$$B = \frac{1}{2}(ab + b - a - (1))$$

이며, 교호작용 AB는

$$AB = \frac{1}{2}(ab + (1) - a - b)$$

가 된다. 추후에도 이는 확장되어, A나 B가 포함된 수가 홀수인지 짝수인지에 따라 부호가 바뀌어 그 선형합이 추정치가 됨을 알 수 있다.

한편 데이터의 배열이 아래와 같을 때, 분산분석을 수행하여 보자.

	A_1	A_1	합계
B_1	y_{00}	y_{10}	$T_{\cdot 0}$
B_2	y_{01}	y_{11}	$T_{\cdot 1}$
합계	$T_{0\cdot}$	$T_{1\cdot}$	T

Theorem 8. *변동의 계산은 아래와 같이 수행된다.*

$$S_T = \sum_i \sum_j y_{ij}^2 - CT = A^2 + B^2 + (AB)^2$$

$$S_A = \sum_i \frac{T_{i\cdot}^2}{2} - CT = \frac{1}{4}(T_{1\cdot} - T_{0\cdot})^2 = A^2$$

$$S_B = \sum_j \frac{T_{\cdot j}^2}{2} - CT - \frac{1}{4}(T_{\cdot 1} - T_{\cdot 0})^2 = B^2$$

$$S_{A\times B} = S_T - S_A - S_B = (AB)^2$$

분산분석표는 아래와 같다.

요인	S	ϕ	V
A	S_A	1	V_A
B	S_B	1	V_B
$A \times B$	$S_{A\times B}$	1	$V_{A\times B}$
T	S_T	3	

단 여기에서 교호작용 $A \times B$에는 오차가 교락되어 있어, 교호작용이 무시되지 않는다면 주효과에 대한 검정이 어렵다. 따라서 반복을 수행해야만 한다.

8.1.2 반복이 있는 경우

반복이 있다면 교호작용의 검출이 가능하다. 이때 데이터의 구조식은

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}$$

이 되며, $T_{00\cdot}, T_{01\cdot}, T_{10\cdot}, T_{11\cdot}$을 각각 A_iB_j 수준에서 r개의 반복된 데이터합이라고 하자. 그렇다면 이들의 평균을 각각 $(1), a, b, ab$로 표시한다면, 반복이 없을 때와 동일하게 A, B, AB를 추정할 수 있다. 즉

$$A = \frac{1}{2}(ab + a - b - (1)) = \frac{1}{2r}(T_{1\cdot\cdot} - T_{0\cdot\cdot})$$

$$B = \frac{1}{2}(ab + b - a - (1)) = \frac{1}{2r}(T_{\cdot 1\cdot} - T_{\cdot 0\cdot})$$

$$AB = \frac{1}{2}(ab + (1) - a - b) = \frac{1}{2r}(T_{11\cdot} + T_{00\cdot} - T_{10\cdot} - T_{01\cdot})$$

이다.

Theorem 9. *분산분석에서의 변동 계산은 아래와 같다.*

$$S_T = \sum_i \sum_j \sum_k y_{ijk}^2 - CT$$
$$S_A = \sum_i \frac{T_{i\cdot\cdot}^2}{2r} - CT = \frac{1}{4r}(T_{1\cdot\cdot} - T_{0\cdot\cdot})^2 = rA^2$$
$$S_B = \sum_j \frac{T_{\cdot j\cdot}^2}{2r} - CT = \frac{1}{4r}(T_{\cdot 1\cdot} - T_{\cdot 0\cdot})^2 = rB^2$$
$$S_{A\times B} = \frac{1}{4r}(T_{11\cdot} + T_{00\cdot} - T_{10\cdot} - T_{01\cdot})^2 = r(AB)^2$$
$$S_E = S_T - (S_A + S_B - S_{A\times B})$$

이로부터 분산분석표는 아래처럼 만들 수 있다.

요인	S	ϕ	V	F
A	S_A	1	V_A	$F_A = V_A/V_E$
B	S_B	1	V_B	$F_B = V_B/V_E$
$A\times B$	$S_{A\times B}$	1	$V_{A\times B}$	$F_{A\times B} = V_{A\times B}/V_E$
E	S_E	$4(r-1)$	V_E	
T	S_T	3		

한편 분산분석 이후의 추정은 사실 반복이 있는 이원배치법과 동일하다. 요인배치법이 이와 다른 것은 단지 효과의 추정과 제곱합의 계산이 쉽다는 것뿐이다.

8.2 2^n 요인실험

이를 더 확장하여 반복이 r회 있는 2^n 요인실험에 대해 라아 보자. 주효과 A, B, C, D 등은 아래의 공식에 따라 얻어진다.

$$\text{주효과} = \frac{1}{2^{n-1}r}((1\text{ 수준의 데이터합}) - (0\text{ 수준의 데이터합}))$$

2인자 교호작용 $A\times B, C\times D$ 등은 아래의 공식에 따라 얻어진다.

$$2\text{인자 교호작용} = \frac{1}{2^{n-1}r}((\text{두 인자 수준의 합이 짝수인 데이터합}) - (\text{두 인자 수준의 합이 홀수인 데이터합})$$

이를 더욱 확장하면, k인자 교호작용은 해당인자의 수준의 합이 k와 2로 나눈 나머지가 같은 데이터합에서 그렇지 않은 데이터합을 뺀 뒤 $2^{n-1}r$으로 나눔으로써 쉽게 구해줄 수 있다.
한편 각 인자의 변동은 아래처럼 구해진다.

$$\text{각 인자의 변동} = 2^{n-2}r(\text{주효과})^2 = \frac{1}{2^n r}(\text{주효과 대비})^2$$

2인자 교호작용의 변동은 아래처럼 구해진다.

$$\text{각 2인자 교호작용의 변동} = 2^{n-2}r(2\text{인자 교호작용})^2 = \frac{1}{2^n r}(2\text{인자 교호작용 대비}^2)$$

따라서 이를 일반화하면 k인자 교호작용의 변동은 $2^{n-2}r(k\text{인자 교호작용})^2$ 혹은 $\frac{1}{2^n r}(k\text{인자 교호작용 대비})^2$으로써 구할 수 있다. 또한 앞서 다루었듯 각 변동은 한 대비의 제곱으로 이루어지기에, 모두 자유도가 1이다. 따라서 주효과의 자유도는 1씩 n개, 2인자 교호작용의 자유도는 1씩 $\binom{n}{2}$개, k인자 교호작용의 자유도는 1씩 $\binom{n}{k}$개 존재하게 되며, 오차의 자유도는 $(2^n r - 1) - (2^n - 1) = 2^n(r-1)$이 될 것이다. 다만 3인자 이상의 교호작용은 별 의미가 없어 풀링시키기도 한다.

8.3 Yates의 계산법

2^n 요인실험에서 Yates의 계산법이라 함은 대비, 요인의 효과, 변동을 쉽게 구해내는 방법이다. 아래에서는 2^3 요인실험을 기반으로 예시를 들어 설명한다.

1. 먼저 n개의 인자를 순서대로 늘어놓고, 이들 수준을 0과 1로 늘어 놓는다. 예를 들어 처리수준 $A_0B_1C_1$에서는 011이다. 그 다음 이를 이진법 수로 보고, 그 순서대로 늘어 놓는다. 그리고 각 처리 수준에서의 데이터합을 계산하여 기입한다. 여기에서는 편의상 $T_{ijk\cdot} = a^i b^j c^k$로 표기한다.

처리조합	데이터합	(1)	(2)	(3)	요인의 효과	요인의 변동
000	(1)					
001	c					
010	b					
011	bc					
100	a					
101	ac					
110	ab					
111	abc					

2. (1) 열에서 상반부 2^{n-1}개는 2^n개의 데이터합을 위로부터 두개씩 짝지어 이 짝을 합하여 쓰고, 하반부 2^{n-1}개는 2^n개의 데이터합을 위로부터 두개씩 짝지어 아래 것에서 위의 것을 빼어 쓴다.

처리조합	데이터합	(1)	(2)	(3)	요인의 효과	요인의 변동
000	(1)	$c+(1)$				
001	c	$bc+b$				
010	b	$ac+a$				
011	bc	$abc+ab$				
100	a	$c-(1)$				
101	ac	$bc-b$				
110	ab	$ac-a$				
111	abc	$abc-ab$				

3. (2)와 (3), 더 나아가서는 (n)열까지, 인자의 개수만큼 이 과정을 반복한다.

처리조합	데이터합	(1)	(2)	(3)
000	(1)	$c+(1)$	$bc+b+c+(1)$	$abc+ab+ac+a+bc+b+c+(1)$
001	c	$bc+b$	$abc+ab+ac+a$	$abc-ab+ac-a+bc-b+c-(1)$
010	b	$ac+a$	$bc-b+c-(1)$	$abc+ab-ac-a+bc+b-c-(1)$
011	bc	$abc+ab$	$abc-ab+ac-a$	$abc-ab-ac+a+bc-b-c+(1)$
100	a	$c-(1)$	$bc+b-c-(1)$	$abc+ab+ac+a-bc-b-c-(1)$
101	ac	$bc-b$	$abc+ab-ac-a$	$abc-ab+ac-a-bc+b-c+(1)$
110	ab	$ac-a$	$bc-b-c+(1)$	$abc+ab-ac-a-bc-b+c+(1)$
111	abc	$abc-ab$	$abc-ab-ac+a$	$abc-ab-ac+a-bc+b+c-(1)$

4. (n) 열의 원소가 대비를 의미한다. 요인의 효과는 이를 $2^{n-1}r$로 나누어, 요인의 변동은 그 제곱을 $2^n r$으로 나누어 계산한다. 단, μ에 대한 추론을 수행하게 되는 가장 위의 항(모든 부호가 +인 항)에서는 요인의 효과 μ를 구할 때 $2^{n-1}r$ 대신 $2^n r$으로 나눈다. 변동을 구할 때에는 다른 것들과 동일하게 $2^n r$으로 나눈다. 이들이 처리조합에 해당하는 주효과 혹은 교호작용의 효과와 변동이 된다.

처리조합	데이터합	(3)	요인의 효과	요인의 변동
000	(1)	$abc+ab+ac+a+bc+b+c+(1)$	$(3)/2^3r=M$	$(3)^2/2^3r=CT$
001	c	$abc-ab+ac-a+bc-b+c-(1)$	$(3)/2^2r=C$	$(3)^2/2^3r=S_C$
010	b	$abc+ab-ac-a+bc+b-c-(1)$	$(3)/2^2r=B$	$(3)^2/2^3r=S_{B\times C}$
011	bc	$abc-ab-ac+a+bc-b-c+(1)$	$(3)/2^2r=BC$	$(3)^2/2^3r=S_{B\times C}$
100	a	$abc+ab+ac+a-bc-b-c-(1)$	$(3)/2^2r=A$	$(3)^2/2^3r=S_A$
101	ac	$abc-ab+ac-a-bc+b-c+(1)$	$(3)/2^2r=AC$	$(3)^2/2^3r=S_{A\times C}$
110	ab	$abc+ab-ac-a-bc-b+c+(1)$	$(3)/2^2r=AB$	$(3)^2/2^3r=S_{A\times B}$
111	abc	$abc-ab-ac+a-bc+b+c-(1)$	$(3)/2^2r=ABC$	$(3)^2/2^3r=S_{A\times B\times C}$

하나하나 계산하는 것보다 표를 통해 계산량을 줄일 수 있다는 점에서, Yates의 계산법은 전통적이면서도 간편한 계산법이다.

8.4 3^2 요인실험

인자가 두 개이면서 수준이 3이면, 처리조합 A_iB_j를 ij로 대체함으로써 표현할 수 있다. 다만 이 요인실험의 분석은 이원배치법에서와 동일하게 수행 가능하다. 이 경우 인자가 계량인자면서 수준 간의 간격이 일정한 경우, 인자의 **일차효과**와 **이차효과**를 대비로써 분리해낼 수 있다. 대표적으로 A의 일차효과는

$$C_L = T_{2\cdot\cdot} - T_{0\cdot\cdot}$$

이 되며, 이차효과는

$$C_Q = T_{2\cdot\cdot} - 2T_{1\cdot\cdot} + T_{0\cdot\cdot}$$

이다. 한편 4수준 이상인 경우 삼차효과 등의 고차효과까지 마찬가지로 구할 수 있다.

한편 이 대비들은 서로 직교하므로, 각각이 갖는 변동 S_L과 S_Q 역시 구할 수 있을 것이며, 이는 반복이 r 회일 때

$$S_L = \frac{1}{6r}(T_{2\cdot} - T_{0\cdot})^2$$
$$S_Q = \frac{1}{18r}(T_{2\cdot} - 2T_{1\cdot} + T_{0\cdot})^2$$

으로 계산가능하고, 그 합이 S_A가 된다. 따라서 일차효과와 이차효과 등이 얼마나 유효한지 분산분석표를 만들어 분산분석을 할 수 있다.

교호작용 $S_{A\times B}$ 역시 분해가 가능한데, 이 경우에는 인자 B의 각 수준에서 인자 A의 일차효과

$$B_j : L_j = -T_{0j\cdot} + T_{2j\cdot}$$

을 구하면

$$S_{A_L\times B} = \frac{L_0^2+L_1^2+L_2^2}{2r} - \frac{(L_0+L_1+L_2)^2}{6r}$$

로 계산할 수 있고, 이는 B의 수준이 바뀜에 따라 A의 일차효과가 어떻게 달라지는지를 나타내는 변동이 된다. 마찬가지로 이차효과 역시 각 수준에서 Q_j를 구하고,

$$S_{A_Q\times B} = \frac{Q_0^2+Q_1^2+Q_2^2}{6r} - \frac{(Q_0+Q_1+Q_2)^2}{18r}$$

으로써 구한다. 이들의 합은 $S_{A\times B}$가 되며, 그 자유도는 3-1 = 2씩이다.

두 인자 모두가 계량인자인 경우에는 교호작용을 더 세분화하여 4개로 나눌 수 있고, 각각은 역시 대비를 구한 뒤 적절한 수로 나누어 자유도가 1인 제곱합을 만듦으로써 얻는다.

Chapter 9

교락법과 일부실시법

요인배치법 혹은 다원배치법에서는 인자의 수나 인자의 수준수가 늘어나면 실험횟수가 너무 늘어나는 단점이 있다. 이처럼 실험횟수가 늘어나면 실험비용이나 실험시간이 증가하기 때문에, 통계적 관리상태를 유지하기에도 어렵다. **교락법**은 실험 전체를 몇 개의 블록으로 나누어 배치함으로서 실험의 관리를 쉽게 하는 방법이다. 한편 고차 이상의 교호작용이 대개 유의하지 않음을 이용하여, 이들을 블록과 교락시킴으로써 인자의 조합 중 일부만을 실행하는 **일부실시법** 역시 사용된다.

9.1 2^n형의 교락법

교락법의 기본 아이디어는 높은 차수의 교호작용을 블록과 교락시키는 것이다. 예를 들어 2^3요인 실험에서는 3인자 교호작용 ABC가 최고차이다. 또한 ABC는

$$ABC = \frac{1}{4}((a+b+c+abc)-((1)+ab+ac+bc))$$

이다. 따라서 이를 블록과 교락시키기 위해서는 앞선 a, b, c, abc를 블록 2에서, $(1), ab, ac, bc$는 블록 1에서 수행하여 ABC와 블록을 교락시킬 수 있다. 이렇게 하면 주효과나 2인자 교호작용을 추정하는 데에는 문제가 없다. 예를 들어 블록 2에서 α만큼 블록으로 인해 더 높은 특성치$a+\alpha, b+\alpha, c+\alpha, abc+\alpha$가 등장한다면,

$$\begin{aligned} A &= \frac{1}{4}(((a+\alpha)+ab+ac+(abc+\alpha))-((b+\alpha)+(c+\alpha)+bc+(1))) \\ &= \frac{1}{4}((a+ab+ac+abc)-((1)+b+c+bc)) \end{aligned}$$

로 α의 효과가 사라짐을 알 수 있다. 따라서 ABC만이 교락될 뿐, 주효과의 추정은 여전히 가능함을 알 수 있다.

한편 더욱 일반적으로는 교호작용이 없다고 알려진 항을 블록과 교락시키는 등의 더욱 복잡한 교락법을 해볼 수 있다. 그 배치 방법에는 인수분해식을 이용하는 방법과 합동식을 이용하는 방법이 있다.

9.1.1 인수분해식을 이용하는 방법

A의 주효과를 교락시키는 상황을 고려하여 보자. A의 주효과는

$$\begin{aligned} A &= \frac{1}{4}((a+ab+ac+abc)-((1)+b+c+bc)) \\ &= \frac{1}{4}(a-1)(b+1)(c+1) \end{aligned}$$

으로 쓸 수 있다. 따라서 일반적으로는 인수분해의 형태에서 교락시키려는 효과에는 -1을, 그렇지 않은 경우 1을 더한 뒤 인수분해를 풀어 계수가 $+$인 효과는 블록 2에, 그렇지 않은 효과는 블록 1에 넣는 식으로 블록을

배치하면 교락시킬 수 있다.

일반적으로는 2개의 블록으로 나눌 때 최고차의 교호작용을 교락시킨다. 예를 들어 4인자인 경우 $ABCD$를 교락시키기 위하여

$$\begin{aligned} ABCD &= \frac{1}{8}(a-1)(b-1)(c-1)(d-1) \\ &= \frac{1}{8}((abcd+ab+ac+ad+bc+bd+cd+(1))-(abc+abd+acd+bcd+a+b+c+d)) \end{aligned}$$

을 계산하고, $abcd, ab, ac, ad, bc, bd, cd, (1)$을 블록 2에, $abc, abd, acd, bcd, a, b, c, d$를 블록 1에 넣는 식으로 교락시킬 수 있다.

한편 블록이 2개로 나누어지는 **단독교락** 이외에도 블록을 4개로 나누는 **2중교락** 역시 이용할 수 있다. 이러한 경우 (1)을 포함하는 블록이 **주블록**이 된다. 단, 이 경우에는 3개의 교호작용을 교락시켜주어야만 한다. 예를 들어 2^4 요인실험에서 ABC, BCD를 교락시키는 경우를 생각하여 보자. 그렇다면 ABC를 교락시키는 경우

$$ABC = \frac{1}{8}((a+b+c+ad+bd+cd+abc+abcd)-((1)+d+ab+ac+bc+abd+acd+bcd))$$

이므로 $+$와 $-$군으로 나눌 수 있고, BCD를 교락시키는 경우

$$BCD = \frac{1}{8}((b+c+d+ab+ac+ad+bcd+abcd)-((1)+a+bc+bd+cd+abc+acd+abd))$$

으로 $+$군과 $-$군으로 나눌 수 있다. 그렇다면 4개의 블록을 둘 모두에서 $+$인 것들, 앞에서 $+$고 뒤에서 $-$인 것들과 같이 4개로 나눌 수 있고, 실제로 아래와 같다.

$$\begin{aligned} &\text{블록 } 1 : c, b, ad, abcd \\ &\text{블록 } 2 : a, bd, cd, abc \\ &\text{블록 } 3 : d, ab, ac, bcd \\ &\text{블곡 } 4 : (1), bc, abd, acd \end{aligned}$$

한편 이 경우에는 교락시킨 두 요인을 곱하여 생기는 요인

$$AD = AB^2C^2D = ABC \times BCD$$

역시 교락되어 있다. 이를 고려하여, 추정하고자 하는 주효과와 같은 효과가 교락되지 않도록 적절한 기술적 고려와 함께 교락시킬 교호작용을 결정해 주어야만 한다.

9.1.2 합동식의 이용방법

2^4 요인실험에서 두 개의 블록으로 나누면서 $ABCD$를 블록과 교락시킨다면,

$$I = ABCD$$

와 같이 표현하고, 이를 **정의대비**라 한다. 이 정의대비는

$$L = x_1 + x_2 + x_3 + x_4 \ (\mathrm{mod} 2)$$

로 표현할 수도 있다. 이로부터 특정 수준, 예를 들어 ab라면, $x_1 = x_2 = 1$이고 $x_3 = x_4 = 0$이도록 하여 $L = 0 \ (\mathrm{mod} 2)$를 계산한다. L이 0이라면 주블록에, 1이라면 주블록과 다른 블록에 넣는 방식으로 교락시킨다.

한편 4개의 블록을 만드는 경우, 동일하게 정의대비 세 개

$$I = ABC$$
$$I = BCD$$
$$I = AD$$

를 만든 뒤, 대응되는 선형표현식

$$L_1 = x_1 + x_2 + x_3 \ (\text{mod} 2)$$
$$L_2 = x_2 + x_3 + x_4 \ (\text{mod} 2)$$
$$L_3 = x_1 + x_4 \ (\text{mod} 2)$$

을 만들고 셋 중 둘을 골라 그 4가지 경우의 수를 기반으로 블록화하면 된다.

9.1.3 분산분석

분산분석은 일반적인 2^n요인실험과 동일하게 수행하면 된다. 단 여기서 다른 점은, 랜덤화가 블록을 먼저 결정한 뒤 해당 블록 내에서 뒤섞기가 되었기 때문에, 블록과 교락시킨 교호작용이 존재한다는 점이다. 예를 들어, ABC, BCD, AD를 블록과 교락시킨 경우 블록간의 변동이

$$S_{Block} = S_{A\times B\times C} + S_{B\times C\times D} + S_{A\times D}$$

가 된다.
따라서 먼저 2^n 요인실험에서처럼 Yates의 계산법을 이용해 교호작용들을 추정하고 난 뒤, 그들을 더함으로써 블록에 의한 변동 S_{Block}을 계산할 수 있다. 또한 고차의 교호작용을 오차에 풀링시켜 오차변동 S_E를 구할 수 있다. 그 다음 그들을 이용하여 V를 얻고, 다시 이에 기반하여 F통계량을 얻고 추론을 수행하면 된다.

9.2 완전교락과 부분교락

교락법에서 반복을 수행할 때, 어떤 반복에서나 동일한 요인효과가 교락되어 있는 경우를 **완전교락**이라 한다. 완전교락을 수행하는 경우 교락된 요인에 대한 정보는 여전히 얻을 수 없으나, 기타 다른 요인들에 대한 더 많은 정보를 얻을 수 있다는 장점이 있다. 이는 반복을 인자 R로 하고 이를 일차단위로 하는 분할법에서와 유사하다. 따라서 일차오차 E_1은 블록과 반복의 교락 $Block \times R$의 변동으로 하고 R과 $Block$의 유의성 검정에 사용하고, 이차오차 E_2는 주효과와 교호작용의 검정에 사용한다.
부분교락은 각 반복마다 블록효과와 교락시키는 요인이 다른 경우를 의미한다. 부분교락된 2^3 요인실험으로 예를 들어 보면,

- 반복1: $(1), ab, ac, bc/a, b, c, abc \Rightarrow$ 교락된 교호작용: ABC
- 반복2: $a, b, ac, bc/(1), ab, c, abc \Rightarrow$ 교락된 교호작용: AB
- 반복3: $ac, abc, (1), b/a, bc, ab, c \Rightarrow$ 교락된 교호작용: AC
- 반복4: $b, c, ac, ab/bc, (1), a, abc \Rightarrow$ 교락된 교호작용: BC

으로 된다. 그렇다면 여기에서 A, B, C는 반복을 4번 하였다고 보면 되고, AB, AC, BC, ABC의 추론에 있어서는 반복을 3번 하였다고 보고 교락되지 않은 경우의 반복들만 추론에 이용하면 되는 것이다. 이러한 경우, ABC, AB, AC, BC의 **상대정보**가 $\frac{3}{4}$가 된다고 말한다.

그 분산분석에서는 반복에 의한 변동 S_R의 자유도는 $\phi_R = r-1$이 되며, 반복내 블록에 의한 변동 $S_{Block/R}$의 자유도가 $\phi_{Block/R} = r$이 된다. 이 변동은 각 반복에서의 자유도가 1인 블록과 교락된 교호작용의 변동 r개의 합이다. 이는 기존과 구하는 방법이 동일하므로 생략한다. 둘을 합쳐 블록에 의한 변동 $S_{Block} = S_R + S_{Block/R}$을 계산할 수도 있으며, 그 자유도는 $2r-1$이 된다.

한편 한 번이라도 교락된 교호작용의 추론에선, 해당 교호작용이 교락되지 않은 반복의 개수를 r'로 대신 사용해야 한다. 또한 그 계산에서도 r'개의 반복에서 얻은 데이터만을 사용해야 한다. 위의 예시에서는 BC에 대한 교호작용을 추론하고자 할 때는 반복 1, 반복 2, 반복 3에서의 자료만 사용해야 하며, 효과와 변동을 구할 때 나누는 값도 $2^2 \times 4 = 16, 2^3 \times 4 = 32$가 아니라 $2^2 \times 3 = 12, 2^3 \times 3 = 24$이 되어야 할 것이다.

분산분석 과정에서는 효과들에 대한 추론에서는 S_E로부터 얻은 V_E를 통하여 수행하면 된다. 단, 반복내 블록 효과의 유의성에 대한 F검정은 수행하기 어려우며, 반복 S_R에 대한 F검정만 가능하다.

9.3 3^n형의 교락법

9.3.1 $S_{A\times B}$의 분해방법

반복이 없는 3^2 요인실험에서 교호작용 $S_{A\times B}$는 자유도로 4를 가진다. 이는 각각 자유도가 2인 두 제곱합

$$S_{AB} = S_{AB(J)} = \frac{1}{3}((y_{00}+y_{21}+y_{12})^2 + (y_{10}+y_{01}+y_{22})^2 + (y_{20}+y_{11}+y_{02})^2) - CT$$
$$S_{AB^2} = S_{AB(I)} = \frac{1}{3}((y_{00}+y_{11}+y_{22})^2 + (y_{20}+y_{01}+y_{12})^2 + (y_{10}+y_{21}+y_{02})^2) - CT$$

으로 분해될 수 있다. 반복이 r번 있는 경우 저들 대신 그 합을 이용하고 3 대신 $3r$을 이용하면 될 것이다. 한편 동일한 논리로 Levi-Civita Tensor를 이용하면 3^n형의 경우에도 이들을 자유도 2를 갖는 2^{n-1}개의 변동으로 분해할 수 있음이 잘 알려져 있다.

9.3.2 단독교락

3^2 요인실험에서, 이를 3개의 블록으로 나누어 실험하는 상황을 고려해볼 수 있다. 이 경우 AB 혹은 AB^2을 교락시킨다. 예를 들어 AB^2을 교락시킨다면 정의대비는 $I = AB^2$이고, 상응하는 선형식은

$$L = x_1 + 2x_2 \ (\mathrm{mod} 3)$$

이다. 따라서 9개의 실험은 L의 값에 따라

$$00, 11, 22/10, 21, 02/20, 01, 12$$

로 나누어 실험될 수 있다.

한편 분산분석에서는 일반적인 3^n 요인실험과 동일하게 분산분석표를 만들되, 블록과 교락시킨 교호작용에 해당하는 부분을 블록에 의한 변동으로 취급하면 된다. 앞서 배운 분해법은 $S_{A\times B}$의 변동을 교락되지 않은 S_{AB}와 블록변동에 해당하는 S_{AB^2}으로 나누는 데 도움이 된다.

9.3.3 이중교락

3^3 요인실험에서, 9개의 블록으로 나누어 실험하는 상황을 고려해볼 수 있다. 이 경우에는 $A^3 = B^3 = C^3 = \cdots = 1$과 요인 X에 대해 $X = X^2$임을 이용하여 교락을 진행한다. 3^3 요인실험에서 9개의 블록으로 나누려면 8개의 자유도를 블록과 교락시켜야 하고, 각각이 2개의 자유도를 가지므로 4개의 요인을 교락시켜야 한다. 그러나 4개의 요인, 예를 들어

$$AB^2C^2, AB, BC^2, AC$$

이면 $(AB^2C^2)(AB) = AC$, $(AB)(BC^2) = AB^2C^2$이므로 2개만이 독립이다. 따라서 적당히 2개를 선택한 뒤, 그들에 대한 L표현식을 이용하여 블록화를 수행하면 된다. 분산분석 시에는 고차 교호작용을 오차로 취급하고, 교락된 요인들을 블럭에 의한 변동으로 취급하고, 나머지 교호작용과 주효과에 대한 추론만 수행한다. 따라서 이는 실험 환경에 제약이 있을 때, 주효과에 대한 정보를 얻기 위해 수행하는 교락법이라고 볼 수 있다.

9.4 2^n형의 일부실시법

9.4.1 2^n형의 1/2실시

2^n형의 1/2 실시라 함은, 전체 2^n번의 실험을 모두 수행하는 대신 적절히 정해진 2^{n-1}번의 실험만 수행함으로써 실험의 횟수를 줄이는 실험배치를 의미한다. 이는 교락법의 배치로부터 쉽게 유도할 수 있다. 예를 들어 2^3 형에서 정의대비를 $I = ABC$로 할 때 교락법은

$$a, b, c, abc/(1), ab, ac, bc$$

로 블록을 나누었다. 여기에서 $\frac{1}{2}$ 반복을 하려는 경우, 하나의 블록만 택하여 해당 블록의 실험만 수행한다. 예를 들어 첫째 블록에 대해 실험하는 경우, A의 주효과는

$$A = \frac{1}{2}((a + abc) - (b + c))$$

으로 추정할 수 있다. 그런데 BC 교호작용 역시

$$BC = \frac{1}{2}((a + abc) - (b + c))$$

로 추정된다. 즉 둘은 완전히 교락한다. 이와 같이 일부실시법에서 동일한 추정식으로 표시되는 요인효과를 서로 **별명**관계에 있다고 한다. 어느 요인효과끼리 별명되어 있는지는 정의대비에 요인효과를 곱하여 알 수 있다. A의 별명은 $I = ABC$의 양변에 A를 곱하여 $A = A^2BC = BC$를 얻기에, A의 별명이 BC인 것이다. 즉 일부실시법에서 BC가 존재하지 않는다고 가정할 수 있다면, 우리는 해당 효과가 A의 주효과라고 주장할 수 있다. 단, BC가 0이 아니면 A의 주효과가 과대/과소평가될 수 있다.

만약 블록 2를 사용하는 경우,

$$A = \frac{1}{2}((ab + ac) - ((1) + bc))$$

로 A는 정확히 말하면

$$-BC = \frac{1}{2}((ab + ac) - ((1) + bc))$$

와 교락되어 있다. 따라서 이를 구분하기 위하여, 일부실시법에서 어떤 블록을 사용하냐에 따라 별명관계를 $A = -BC$ 처럼 부호를 조정해 표현하기도 한다.

9.4.2 2^n형의 1/4실시

$\frac{1}{4}$실시는 교락법에서 이중교락을 수행하여 총 3개 요인을 블록과 교락시킬 때 4개의 블록이 만들어지면, 그 중 하나만 선택하여 실시함으로써 이루어진다. 여기에서는 별명관계인 인자들이 2개가 한 그룹을 이루는 것이 아니라, 4개가 한 그룹을 이루게 된다. 이는 각 요인에 3개의 정의대비를 각각 곱함으로써 얻어진다. 예를 들어 $I = ABCD = ACE = BDE$가 정의대비라면, A의 별명은 $A, BCD, CD, ABDE$가 되는 것이다. 따라서 고차 교호작용의 부재를 가정하면 A의 주효과를 추정하는 데 일부실시법을 이용할 수 있음을 안다.

이러한 것을 확장하면 3^n 요인실험에서 $\frac{1}{3}$실시, 직교배열표를 이용한 일부실시법도 쉽게 할 수 있다.

Chapter 10

불완비블록계획법

A인자가 모수인자이고 B인자가 블록인자인 난괴법에서, 어떤 블록에서 모든 A 수준의 실험을 하지 못하는 상황도 있을 수 있다. 이와 같은 블록을 **불완비블록**이라 하며, 이에 기반한 실험계획법을 **불완비블록계획법**이라 한다.

10.1 균형 불완비블록계획법

블완비블록 계획에 대한 데이터의 구조식을

$$\begin{aligned} y_{ijk} &= \mu + a_i + b_j + e_{ijk} \\ k &= n_{ij} \\ e_{ijk} &\sim_{i.i.d.} N(0, \sigma_E^2) \end{aligned}$$

이라 하자. 이때 n_{ij}는 0 또는 1이고 $k = n_{ij} = 0$인 경우에는 데이터가 구해지지 않아 존재하지 않고, 1인 경우에는 y_{ij1}이 구해졌다고 생각한다. 처리가 l개, 블록은 m개 있다고 생각하자. 만약 아래의 조건들이 만족된다면, 이를 **균형 불완비블록계획법**이라 한다.

- 모든 블록에서 p개의 처리가 이루어진다.
- 각 처리는 r개의 블록에 나타난다.
- 어느 두 처리를 보더라도 이 두 처리가 동시에 이루어지는 블록의 수는 λ이다.
- 처리의 수가 블록의 수보다 많지는 않다.

	처리번호						
	1	2	3	4	5	6	7
B_1	O	O		O			
B_2		O	O		O		
B_3			O	O		O	
B_4				O	O		O
B_5	O				O	O	
B_6		O				O	O
B_7	O		O				O

위 테이블이 대표적인 $l = 7, m = 7, p = 3, r = 3, \lambda = 1$일 때의 불완비블록계획이다. 한편 λ는

$$\lambda = \frac{r(p-1)}{l-1}$$

으로써 쉽게 구할 수 있다. 또한 $rl = mp$ 역시 성립함을 자연스럽게 알 수 있다. 따라서 l, m과 λ를 알면 적절한 p, r을 결정해 불완비블록계획을 짤 수 있게 된다.

균형 불완비블록계획에서 데이터가 얻어졌을 때 분산분석을 행하는 방법은 다원배치법에서와 약간 다르다. 총변동은 처리간 변동, 블록간 변동, 오차항으로 나누어지며, 블록간 변동은

$$S_{Block} = \sum_{j=1}^{m} \frac{T_{\cdot j\cdot}^2}{p} - CT$$

로 기존과 동일하다. 이를 **미수정된 블록변동**이라 한다. 이는 사실 블록별로 처리가 다른 것을 고려하지 않는다는 점에서 완전한 블록변동은 아니다.

수정된 처리변동은 아래와 같이 주어진다.

$$S_{Trt(adj.)} = \frac{p}{\lambda l} \sum_{i=1}^{l} \left(T_{i\cdot\cdot} - \frac{1}{p} B_i \right)^2$$

이때 $B_i = \sum_{j=1}^{m} n_{ij} T_{\cdot j\cdot}$는 i번째 처리가 수행되는 모든 블록에 있는 데이터의 합을 의미한다. 이를 p로 나누어서 보정하는 것은 해당 처리가 포함된 블록의 모든 관측값들의 평균을 빼주는 의미를 가진다. 총변동을 구하는 방법은 기존과 같고, 오차변동은 총변동에서 앞서 구한 미수정된 블록변동과 수정된 처리변동을 빼어 구한다.

이로써 얻어지는 분산분석표는 아래와 같다.

요인	S	ϕ	V	F_0
블록(미수정)	S_{Block}	$m-1$	V_{Block}	
처리(수정)	$S_{Trt(adj.)}$	$l-1$	V_{Trt}	$F_{Trt} = V_{Trt}/V_E$
오차	S_E	$N-l-m+1$	V_E	
T	S_T	$N-1$		

블록간 변동 유무에 대한 검정은 어려우며, 수정된 처리에 의한 변동에 대해서만 $H_0 : a_1 = a_2 = \cdots = a_l$ 검정을 F 검정으로써 수행하게 된다.

한편 분산분석 이후에는, 모수모형에서 $\sum_{i=1}^{l} a_i = 0$으로 가정한다면 a_i의 불편추정량이

$$\hat{a}_i = \frac{p}{\lambda l} Q_i = \frac{p}{\lambda l} \left(T_{i\cdot\cdot} - \frac{1}{p} B_i \right)$$

임이 잘 알려져 있다. $\mu(A_i)$에 대해서는

$$\hat{\mu}(A_i) = \bar{y} + \frac{p}{\lambda l} Q_i$$

를 사용할 수 있다.

처리효과의 대비 $\sum_{i=1}^{l} c_i a_i$의 유의성에 대해서는 그 추정량이

$$\frac{p}{\lambda l} \sum_{i=1}^{l} c_i Q_i$$

이고 그 분산이

$$\frac{p\sigma_E^2}{\lambda l} \sum_i c_i^2$$

임을 이용하여, t통계량

$$t_0 = \frac{\sum_i c_i \hat{a}_i}{\sqrt{\frac{pV_E}{\lambda l} \sum_i c_i^2}}$$

이 자유도 $N-l-m+1$인 t분포임을 통해 검정한다. 신뢰구간 역시 이와 유사한 방식으로 만들 수 있다.

10.2 Youden 방격법

라틴방격에서 하나의 열 이상이 제거된 형태를 **불완비라틴방격**이라 부르고, 이를 이용한 실험계획법을 **Youden 방격법**이라 부른다. 4×4 라틴방격에서 열 하나를 제거한 Youden 방격은

	A_1	A_2	A_3
B_1	C_1	C_2	C_3
B_2	C_4	C_1	C_2
B_3	C_2	C_3	C_4
B_4	C_3	C_4	C_1

와 같다. 블록 4개가 3종류의 인자를 한 번에 실험할 수 있을 때, 4종류의 인자 C의 특성을 실험하는 상황이 이러한 상황이다. 이러한 경우 데이터의 구조식은

$$y = \mu + a_i + b_j + c_k + e_{ijk}$$

가 될 것이며, 이러한 Youden 방격은 불완비블록으로 바꾸어 쓸 수 있다. 아래처럼 말이다.

	처리번호			
	1	2	3	4
B_1	O	O	O	
B_2	O	O		O
B_3		O	O	O
B_4	O		O	O

$l = m = 4$, $p = 3$, $r = 3$, $\lambda = 2$인 균형 블완비블록계획이기도 하다.

허나 모든 Youden 방격이 불완비블록으로 변환될 때 균형을 이루는 것은 아니다. 다만 불완비블록에서 $l = m, p = r$인 경우 Youden 방격으로 변환하는 것은 항상 가능하다. 어찌하였든 Youden 방격법은 균형 불완비블록계획법으로 몇몇 경우 변환이 가능하며, 그에 맞게 B의 미수정된 변동과 C의 수정된 변동을 구할 수 있다. A에 의한 변동은 별도로 Youden 방격법에서 바로 구하면 된다. 그렇다면 아래와 같은 분산분석표가 만들어진다.

요인	S	ϕ	V	F_0
위치(A)	S_A	$3-1$	V_A	$F_A = V_A/V_E$
블록(B)	S_B	$4-1$	V_B	
수정된 인자(C)	S_C	$4-1$	V_C	$F_C = V_C/V_E$
오차(E)	S_E	$12-2-3-3$	V_E	
T	S_T	$12-1$		

따라서 여기에서 A와 C에 의한 효과를 검정할 수 있다. 균형 불완비블록계획에서와 동일하게, B의 효과는 미수정되어 F검정을 바로 할 수 없고, 수정시켜 수행해야 하나 여기에서는 생략한다.