분산 분석

윤 정 훈

들어가기

- 지금까지 두 모집단의 평균을 비교하는 문제를 이표본 검정으로 다루었다. **만약 비교하고자 하는 모집단이 세 개 이상**이라면 어떤 방법을 써야 할까?
- 표본이 세 개 이상인 경우, 정규분포의 모평균들을 어떻게 비교/검정 할 것인지에 대해
  - ✓ 집단 사이의 흩어짐과 집단 내의 흩어짐을 측정하여 비교하는 **분산비 추정량**(estimator of variance ratio)을 사용한다.
  - ✓ 검정을 위한 검정통계량은 F분포를 사용한다.
  - ✓ 집단 내의 흩어짐에 비해 집단 사이의 흩어짐의 정도가 큰 경우, 집단 사이에 차이가 있다고 결론을 내리게 된다.

분산분석

- 분산분석은 3개 이상의 모평균에 대한 분석으로, 실험치(측정치)의 변동을 총 제곱 합(total sum of squares)으로 나타내고 이 총 제곱 합을 실험과 관련된 독립변인(요인 또는 인자)의 작용에 대한 각자의 제곱 합으로 분해한 후, 나머지를 오차변동으로 해석하는 검정법이다.
- 각 독립변인마다 분해한 분산을 오차분산과 비교하여 특히 큰 영향을 주는 독립변인이 무엇인가를 검정하고, 그 결과 실험치가 있으면 요인마다 효과 추정을 행한다.
- 또한, 분산분석은 측정치의 변동을 독립변인별로 분해하여 어느 독립변인이 실험치(종속변인)에 어느 정도 영향을 주는지를 파악한다.

분산분석의 종류

- 요인(인자) 수에 의한 분류
  - 일원분산분석(one-way ANOVA)
  - 이원분산분석(two-way ANOVA)
- 요인(인자)의 모형에 의한 분류
  - 모수효과 모형(Fixed effect model)
  - 변량효과 모형(Random effect model)
  - 혼합효과 모형(Mixed effect model)

일원분산분석

- 독립변인(인자 또는 요인)은 하나이고 이 독립변인을 k개 집단(수준)으로 나뉜 표본들을 서로 비교 검정하는 것이다. 독립변인의 집단(수준)은 3개 이상 ($k \geq 3$), 종속변인(실험치 또는 나타나는 결과)는 하나인 것을 **일원분산분석** 또는 **oneway ANOVA**라고 한다.
  - ✓ ex1) 가구소득에 따른 식료품소비 정도의 차이
    - 독립변인 : 가구소득
    - 독립변인의 집단 : 가구소득집단의 구분 – 저소득층, 중산층, 고소득층
    - 종속변인 : 식료품소비
  - ✓ ex2) 10세 남아 체중의 한/중/일 국가간 차이
    - 독립변인 : 10세 남아
    - 독립변인의 집단 : 한/중/일
    - 종속변인 : 체중

일원분산분석 예제

- 3대의 기계에서 생산되는 공구들의 파괴강도를 측정한 후, 각 기계가 생산한 공구의 파괴강도간에 차이가 있는지를 일원분산분석을 이용하여 검정한다고 가정하자.
  - 독립변인 : 각각의 기계들
  - 독립변인에 따른 집단수 : $k = 3$
  - 종속변인(실험치) : 파괴강도
  - 파괴강도의 효과를 조사하는 것이다. 다른 독립변인은 영향이 거의 없거나 일정하게 유지할 수 있을 때 하는 분석이다. 집단 수가 3개(3대의 기계에서 생산된 제품) 이상으로 평균치 사이에 차이가 있는가를 검정한다. 반복 수에는 제한이 없으며 실험순서는 임의로 선택하여 실시한다.
  - **총변동**을 **집단간 변동(기계 간)**과 **집단내 변동(같은 기계안에서의 공구들 간)**으로 나누어서 분산비를 검정한다.

확률화 실험계획

- 실험단위들을 각 처리 집단에 무작위로 배정하는 계획이다.
- 일원배치법에 적용되는 방법에 대한 예제
  - 위암에 대한 서로 다른 치료제 A_1, A_2, A_3가 있다고 할 때, 이 세 가지 치료제의 효과를 비교하기 위한 실험계획을 세워보자.
  - 여기서 실험치는 위암의 치료이고, 집단수는 서로 다른 치료제 A_1, A_2, A_3 3개 이다. 실험계획은 위암에 걸린 입원한 환자 30명을 대상으로 한다. (독립변인 : 각 치료제)
  - 10명씩 A, B, C의 세 집단으로 무작위로 나눔.
  - A집단에 치료제 A_1, B집단에 치료제 A_2, C집단에 치료제 A_3를 각각 투여
  - 한 종류 암을 치료하는 데까지 걸린 시간을 관측하여 각 약에 대한 치료효과를 분석.

일원배치법

- 독립변인 A의 수준(집단)수가 k개$(A_1, \ldots, A_k)$이며, 각 수준마다 n번씩 동일하게 반복 실험을 할 경우, 일원배치법의 데이터는 다음과 같이 배열된다. $N = k \times n$ 총 실험횟수이다.

처리군(A_i) / 구 분	인자의 수준(요인의 집단)						평균(합계)
	A_1	A_2	$\cdots$	A_i	$\cdots$	A_p	
수준 별 실험의 반복 수: n	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{1j}	y_{2j}	$\cdots$	y_{ij}	$\cdots$	y_{kj}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}	
평균[합계]	$\bar{y}_1[T_1]$	$\bar{y}_2[T_2]$	$\cdots$	$\bar{y}_i[T_2]$	$\cdots$	$\bar{y}_p[T_p]$	$\bar{y}[T]$

일원배치법

처리군(A_i) / 구 분	인자의 수준(요인의 집단)						평균(합계)
	A_1	A_2	$\cdots$	A_i	$\cdots$	A_p	
수준 별 실험의 반복 수: n	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{1j}	y_{2j}	$\cdots$	y_{ij}	$\cdots$	y_{kj}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}	
평균[합계]	$\bar{y}_1[T_1]$	$\bar{y}_2[T_2]$	$\cdots$	$\bar{y}_i[T_2]$	$\cdots$	$\bar{y}_p[T_p]$	$\bar{y}[T]$

$T_i = \sum_{j=1}^{n} y_{ij}$, $\bar{y}_i = \frac{T_i}{n}$

$T = \sum_{i=1}^{k} T_i$, $\bar{y} = \frac{T}{N}$

① $y_{ij} = \mu_i + \epsilon_{ij}\ (i = 1, \dots, k), (j = 1, \dots, n)$: 집단의 평균과 집단에 속한 데이터의 관계

② $\mu_i = \mu + \alpha_i$, $\sum_{i=1}^{k} \alpha_k = 0$: 전체 평균과 집단의 평균과의 관계

③ $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $\mu = \frac{1}{k}\sum_{i=1}^{k} \mu_i$

ϵ_{ij} : 관찰 오차, 독립적이고 $N(0, \sigma^2)$를 따른다.

μ : 전체적인 처리효과

α_i : i번째 처리방법의 순수효과(i번째 모평균의 오차)

일원분산분석의 모수 추정

- $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$의 모수들의 추정을 고려하자.
  - 최소제곱추정법을 사용할 것이다. 특정모형에서 모형과 관찰값의 차이를 잰 오차제곱합을 생각하고, 이를 최소화하는 값을 모수의 추정값으로 삼는 방법이다.
  - i번째 처리집단의 j번째 관찰값 y_{ij}은 다음과 같이 표현될 수 있다. $y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$ 즉 전체 표본평균 + 집단 i의 처리효과 + 표본내 오차 로 나타낼 수 있다.
  - 모수 $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$, $\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}} = \bar{\boldsymbol{y}}_{\boldsymbol{i}} - \bar{\boldsymbol{y}} (i = 1, \dots, k)$
    - ✓ $\epsilon_{ij} = y_{ij} - \mu - \alpha_i$이므로, $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2$를 최소화하면 된다.

일원분산분석의 변동

- 자료 전체의 변동, 즉 흩어진 정도는 처리방법간과 처리방법내의 것으로 구성된다.
  - $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left((\bar{y}_i-\bar{y})+(y_{ij}-\bar{y}_i)\right)^2$
    $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$
  - ✓ Total Sum of Squares(전체 제곱합) : $TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$
  - ✓ Sum of Squares Between(처리방법간 제곱합) : $SSB = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2$
  - ✓ Sum of Squares Within(처리방법내 제곱합) : $SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$
- $TSS = SSB + SSW$

일원분산분석의 변동

- 각 제곱합들을 해당하는 자유도로 나눔으로써 척도화하여 평균제곱합(Mean Sum of Squares)이라고 한다.

- 처리간 제곱합은 k개의 변수 $(\bar{y}_i - \bar{y})$들로 구성되어 있으나 여기에는 합이 0이라는 제약조건이 있다. 즉 $\sum_{i=1}^{k}(\bar{y}_i - \bar{y}) = 0$이다. 이 제약조건에 의해 $(k-1)$개의 변수만이 "자유"로우며 그렇기에 처리간 제곱합의 자유도는 $(k-1)$이 된다는 것이다.

- 처리내 제곱합, 전체 제곱합도 마찬가지로 각각의 수에서 1를 뺀 수만큼의 변수만이 "자유"롭다.

일원분류 분산분석표

(One-way ANOVA Table)

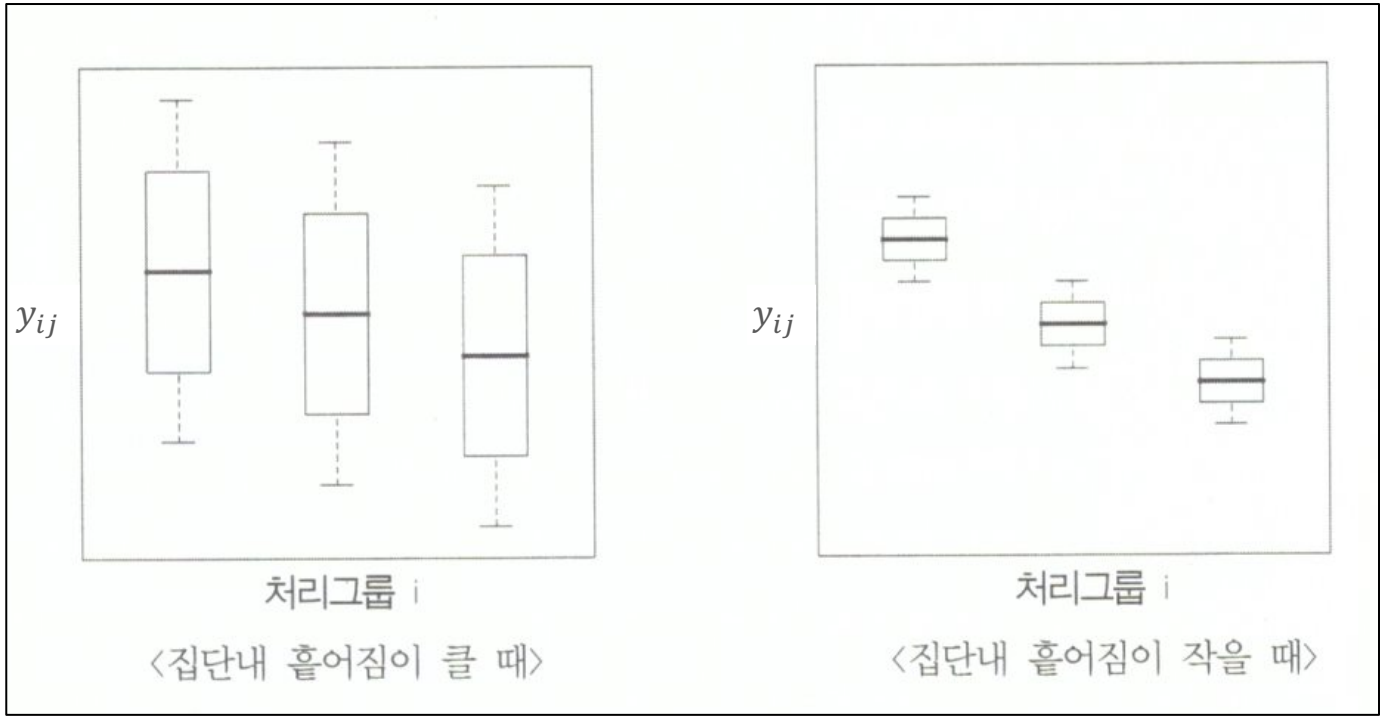
요인	제곱합	자유도	평균제곱합
처리간	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2$	$I-1$	$MSB = \frac{SSB}{I-1}$
처리내	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$n-I$	$MSW = \frac{SSW}{n-1}$
전체	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$n-1$	

일원분산분석의 가설검정

- k개의 처리효과들 간에 차이가 없다는 가설을 검정하자.
  - $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$
  - $H_1 : \alpha_i$들 중 하나는 0이 아니다.

일원분산분석의 가설검정

- 가설검정을 위한 검정통계량으로 집단간의 흩어짐과 집단내의 흩어짐에 대한 비를 고려할 수 있다(F검정). 집단내의 흩어짐이 클 때는 집단간의 차이를 보기가 힘들어지지만 집단내의 흩어짐이 작을 때는 집단간의 차이가 쉽게 드러난다.



일원분산분석의 가설검정

- 집단간의 흩어짐을 측정하는 SSB와 집단내에서의 흩어짐을 측정하는 SSW의 비를 고려한다.
  - 집단내의 흩어짐에 비하여 집단간의 흩어짐이 큰 것이 귀무가설을 기각하는 증거가 된다.
- 모형 $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$에서
  - 귀무가설 $H_0 : \alpha_1 = \cdots = \alpha_k = 0$ 이 참이면 :

    $\frac{SSB}{\sigma^2}$은 자유도가 $I-1$인 카이제곱분포를 따르며, $E\left[\frac{SSB}{I-1}\right] = \sigma^2$이다.
  - 귀무가설 $H_0 = \alpha_1 = \cdots = \alpha_I$이 참이 아니면 :

    $E\left[\frac{SSB}{I-1}\right] = \sigma^2 + \sum_{i=1}^{I} \frac{\alpha_i^2}{I-1}$을 만족한다.
  - α_i의 값에 관계없이 $\frac{SSW}{\sigma^2}$는 자유도가 $(n-I)$인 카이제곱분포를 가지며 $E\left[\frac{SSW}{(n-I)}\right] = \sigma^2$를 만족한다.
  - SSB와 SSW는 서로 독립이다.

일원분산분석의 가설검정

- 위의 정리를 보면, $MSB = \frac{SSB}{I-1}$의 기대값은 귀무가설이 틀릴때, σ^2 보다 더욱 커진다. 따라서 분산비의 추정에 의한 검정통계량 $\frac{MSB}{MSW}$의 값이 큰 경우, 즉 집단들 사이의 흩어짐이 집단내의 흩어짐에 비해 상대적으로 큰 경우 이는 귀무가설을 기각하는 근거가 된다.

- 검정통계량 $F = \frac{MSB}{MSW}$는 귀무가설 H_0하에서 자유도가 $(I-1, n-I)$인 F분포를 따르며, 따라서 가설에 대한 유의수준 α인 기각영역은 $F \geq F_\alpha(I-1, n-I)$로 구해진다.

일원분산분석의 가설검정 예제 1

- 5종류의 수면제 A, B, C, D, E가 있는데, 수면제를 복용한 후 졸음을 느끼기 시작한 시간을 관찰하였다. 다음의 자료에 근거하여 수면제 사이에 졸음을 느끼는 시간이 다르다고 할 수 있는가를 알아보자.

수면제 종류	A	B	C	D	E
반복 (단위 : 분)	9	11	13	18	17
	10	16	18	20	12
	15	12	18	22	13
	14	18	19	19	15
	13	18	19	23	11

일원분산분석의 가설검정 예제 1

- 5종류의 수면제 효과들 간의 차이가 없다는 가설
  - $H_0 : \alpha_1 = \cdots = \alpha_5 = 0$
  - $H_1 : \alpha_i$ 들 중 적어도 하나는 0이 아님
  - $k = 5$, $n_1 = \cdots = n_5 = 5$

요인	자유도	제곱합	평균제곱합	F값
처리 간	4	210.640	52.660	7.72
처리 내	20	136.400	6.820	
전체	24	347.040		

  - $F = 7.72 > F_{0.05}(4,20) = 2.87$ 이므로 유의수준 $\alpha = 0.05$에서 귀무가설 H_0을 기각한다.
  - 즉, 5종류의 수면제 효과들 간에 차이가 없다고 볼 수 없다.

이원분류 분산분석을 들어가며

- No treatment, treatment A, treatment B 이렇게 3 그룹을 비교할 때 one factor ANOVA, one way ANOVA, 일원분산분석 이라고 한다는 것을 배웠다.
- 남성과 여성에게 위의 테스트(No treatment, treatment A, treatment B)를 각각 진행하면, 우리는 이것을 two factor ANOVA, two way ANOVA, 이원분산분석 이라고 한다.

이원분류 분산분석을 들어가며

- 완전임의배치법에 대한 일원분류 분산분석에서는 각 처리방법에 대한 관찰값들이 균일(homogeneous)하다는 가정을 전제로 하였다. 그러나 때로는 이러한 균일성을 가질 수 없는 경우가 있다. 이런 경우는 성질이 유사한 실험단위들끼리 묶어서 균일한 그룹, '블록(block)'을 만들고, 각 블록 안에서 모든 처리방법들을 랜덤배치하는 실험방법이 더 타당하다. 이 과정을 통해 오차의 분산을 줄이고 실험결과의 분석을 더 유효하게 할 수 있다.

이원분류 분산분석을 들어가며

- 몇 종류의 비료가 농작물의 수확량에 미치는 영향을 비교하는 실험이 있다고 하자. 완전 임의 배치법을 쓰기 위해서는 실험에 사용되는 토지가 모두 균일해야 하지만, 때로는 비옥도가 서로 다른 토지를 모두 이용해야 하는 경우가 있다. 이 때 실험이 효과적이기 위해서는 농작물을 재배하는 토지의 비옥도에 따라 유사한 것들을 모아 몇 개의 블록을 만들고, 각 블록내에서 실험단위인 농작물에 사용될 비료를 랜덤으로 배치 할 수 있다. 이러한 실험계획을 randomized block design 이라고 한다.

Randomized block design

- randomized block design에 의하여 i번째 블록에서 $(i = 1, \dots, I)$ j번째 처리방법을 적용한 $(j = 1, \dots, J)$ 관찰값을 y_{ij}라고 하면 다음 표와 같은 결과를 얻을 수 있다.

구분		처리방법				블록평균
		1	2	…	J	
블록	1	y_{11}	y_{12}	…	y_{1J}	$\bar{y}_{1.}$
	2	y_{21}	y_{22}	…	y_{2J}	$\bar{y}_{2.}$
	…	…	…	…	…	
	I	y_{I1}	y_{I2}	…	y_{IJ}	$\bar{y}_{I.}$
처리평균		$\bar{y}_{.1}$	$\bar{y}_{.2}$		$\bar{y}_{.J}$	$\bar{y}$

i번째 블록평균 : $\bar{y}_{i.} = \left(\frac{1}{J}\right)\sum_{j=1}^{J} y_{ij}$ j번째 처리평균 : $\bar{y}_{.j} = \left(\frac{1}{i}\right)\sum_{i=1}^{I} y_{ij}$ 전체표본 평균 : $\bar{y} = \left(\frac{1}{n}\right)\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}$

Randomized block design

- $y_{ij} = \mu_{ij} + \epsilon_{ij}$ $(i = 1, \ldots, I \ , \ j = 1, \ldots, J)$
  - ✓ μ_{ij}는 알려져 있지 않은 상수이고 ϵ_{ij}는 서로 독립인 $N(0, \sigma^2)$을 따르는 확률변수이며 $n = I \times J$이다.
  - ✓ $\mu_{ij} = \mu + \alpha_i + \beta_j$
- $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$

μ : 전체적인 처리효과

α_i : i번째 블록의 순수효과

ϵ_{ij} : 관찰 오차, 독립적이고 $N(0, \sigma^2)$를 따른다.

β_j : j번째 처리방법에 대한 처리효과

  - ✓ $\mu = \left(\frac{1}{I \times J}\right) \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}$
  - ✓ $\mu_{i.} = \left(\frac{1}{J}\right) \sum_{j=1}^{J} \mu_{ij}$
  - ✓ $\mu_{.j} = \left(\frac{1}{I}\right) \sum_{i=1}^{I} \mu_{ij}$
  - ✓ $\alpha_i = \mu_{i.} - \mu$ (i번째 블록의 평균과 전체평균의 차이 : i번째 블록효과)
  - ✓ $\beta = \mu_{.j} - \mu$ (j번째 처리방법의 효과)
  - ✓ $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0$
  - ✓ $\epsilon_{ij} \sim N(0, \sigma^2)$
  - ✓ $\mu_{ij} = \mu_{i.} + \mu_{.j} - \mu$

이원분류 분산분석

- i번째 블록과 j번째 처리방법의 관찰값 y_{ij}는 다음과 같이 분해
  - $y_{ij} = \bar{y} + (\bar{y}_{i.} - \bar{y}) + (\bar{y}_{.j} - \bar{y}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})$
  - 관찰값=전체평균+블록효과+처리효과+잔차
  - $(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}) = y_{ij} - (\bar{y}_{i.} - \bar{y}) - (\bar{y}_{.j} - \bar{y}) - \bar{y}$
  - 잔차 = 관찰값 - 블록효과 - 처리효과 - 전체효과

이원분류 분산분석의 모수추정

- $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$에서의 모수 μ , α_i , β_j에 대한 최소제곱 추정량은
  - $\hat{\mu} = \bar{y}$
  - $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}$
  - $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}$
    - ✓ $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \mu - \alpha_i - \beta_j)^2$를 최소화하여 모수를 추정한다.
    - ✓ 라그랑지 승수(Lagrange multiplier)를 도입하여 풀 수 있다.

이원분산분석의 변동

- 자료 전체의 흩어진 정도는 블록간의 것, 처리방법간의 것, 그리고 오차에 의한 부분으로 구성된다.
  - $\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i.}-\bar{y})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{.j}-\bar{y})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2$
    - ✓ $TSS = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y})^2$
    - ✓ $SSB = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i.}-\bar{y})^2$
    - ✓ $SST = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{.j}-\bar{y})^2$
    - ✓ $SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2$
  - $TSS = SSB + SST + SSE$

이원분류 분산분석표

(Two-way ANOVA Table)

요인	제곱합	자유도	평균제곱합
블록	$\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i.}-\bar{y})^2$	$I-1$	$MSB=\frac{SSB}{I-1}$
처리방법	$\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{.j}-\bar{y})^2$	$J-I$	$MST=\frac{SST}{n-1}$
오차	$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2$	$(I-I)(J-I)$	$MSE=\frac{SSE}{(I-1)(J-1)}$
전체	$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y})^2$	$I\times J-1$	

이원분류 분산분석의 가설검정

- 블록효과 또는 처리효과의 차이에 관한 가설에 대한 검정통계량으로 오차 제곱합에 대한 해당요인의 제곱합 비를 고려할 수 있다.
  - H_0
    - I개의 블록효과들간에 차이가 없다. $(H_0 : \alpha_1 = \cdots = \alpha_I = 0)$
    - J개의 처리효과들간에 차이가 없다. $(H_0 : \beta_1 = \cdots = \beta_J = 0)$
  - H_1
    - I개의 블록효과들간에 차이가 있다.
    - J개의 처리효과들간에 차이가 있다.

이원분류 분산분석의 가설검정

- $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$에서
  - 귀무가설 $H_0 : \alpha_0 = \cdots = \alpha_I = 0$ 이 참이면 :

    통계량 $\frac{SSB}{\sigma^2}$는 자유도가 $I-1$인 카이제곱분포를 따르며 $E\left[\frac{SSB}{(I-1)}\right] = \sigma^2$ 이다.
  - 만일 귀무가설 $H_0 : \alpha_1 = \cdots = \alpha_I = 0$ 이 참이 아니면 :

    $\frac{SSB}{\sigma^2}$ 는 비중심 카이제곱분포를 가지며 $E\left[\frac{SSB}{(I-1)}\right] = \sigma^2 + \frac{J}{(I-1)}\sum_{i=1}^{I}\alpha_i^2$ 을 만족한다.
  - 귀무가설 $H_0 : \beta_1 = \cdots = \beta_J = 0$ 이 참이면 :

    통계량 $\frac{SST}{\sigma^2}$는 자유도가 $J-1$인 카이제곱분포를 따르며 $E\left[\frac{SST}{J-1}\right] = \sigma^2$ 이다.
  - 귀무가설이 참이 아니면 :

    $\frac{SST}{\sigma^2}$ 는 비중심 카이제곱분포를 가지며 $E\left[\frac{SST}{J-1}\right] = \sigma^2 + \frac{I}{J-1}\sum_{j=1}^{J}\beta_j^2$ 을 만족한다.

이원분류 분산분석의 가설검정

- α_i 또는 β_j의 값에 관계없이 SSE/σ^2은 자유도가 $(I-1)(J-1)$인 카이제곱분포를 가지며 $E\left[\frac{SSE}{(I-1)(J-1)}\right] = \sigma^2$를 만족한다.
- SSB , SST 그리고 SSE는 서로 독립이다.
- 블록효과와 처리효과에 대한 F 검정
  - ✓ SSB/σ^2 와 SSE/σ^2 는 서로 독립이므로, 귀무가설 $H_0 : \alpha_1 = \cdots = \alpha_I = 0$ 이 참일 때 :
    - 통계량 $F = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i.}-\bar{y})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2/(I-1)(J-1)}$ 은 자유도가 $\big((I-1),(I-1)(J-1)\big)$ 인 F 분포를 따른다.
    - 따라서, $F \geq F_\alpha\big(I-1,(I-1)(J-1)\big)$ 이면 유의수준 α에서 귀무가설 $H_0 : \alpha_1 = \cdots = \alpha_I = 0$ 을 기각한다.
  - ✓ SST/σ^2 와 SSE/σ^2 도 서로 독립이므로 귀무가설 $H_0 : \beta_1 = \cdots = \beta_J = 0$ 이 참일 때 :
    - $F = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{.j}-\bar{y})^2/(J-1)}{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y})^2/(I-1)(J-1)}$ 은 자유도가 $\big((J-1),(I-1)(J-1)\big)$ 인 F 분포를 따른다.
    - 따라서, $F \geq F_\alpha\big(J-1,(I-1)(J-1)\big)$ 이면 유의수준 α에서 귀무가설 $H_0 : \beta_1 = \cdots = \beta_J = 0$ 을 기각한다.

이원분류 분산분석의 가설검정 예제

- 다음은 5종류의 비료가 생산량에 끼치는 효과를 비교하기 위하여 3개의 블록(논)을 이용한 난괴법에 의한 실험결과이다. 즉, 각 논을 5개의 균일한 구역으로 나누어 5종류의 비료를 랜덤하게 배치하였다. 다음의 자료는 논의 수확량을 나타낸 것이다. 자료에 근거하여 블록들 간과 비료들 간에 차이가 있다고 할 수 있는지 검정해보자.

비료의 종류	블록		
	1	2	3
A	15.9	15.2	16.0
B	13.3	12.2	10.1
C	12.5	11.1	9.3
D	13.2	14.3	11.0
E	16.5	19.5	14.2

이원분류 분산분석의 가설검정 예제

- 이제 3개의 블록들 간에 차이가 없다는 가설은 이원분류 분산분석에 대한 모형을 사용했을 때,
  - $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$, $H_1 : \alpha_i$ 들 중 적어도 하나는 0이 아니다.
  - $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, $H_1 : \beta_j$ 들 중 적어도 하나는 0이 아니다.
- 분산분석 결과표

요인	자유도	제곱합	평균제곱합	F값	p-value
블록	2	16.956	8.478	4.98	0.0394
처리방법	4	74.257	18.564	10.90	0.0025
오차	8	13.631	1.704		
전체	14	104.844			

이원분류 분산분석의 가설검정 예제

- 블록간의 분산비(F)의 값이 $4.98 > F_{0.05}(2,8) = 4.46$ 이므로 유의수준 $\alpha = 0.05$에서 블록간에 수확량의 차이가 있다고 할 수 있다.
- 처리방법간의 분산비(F)의 값이 $10.90 > F_{0.05}(4,8) = 3.84$ 이므로 유의수준 $\alpha = 0.05$에서 비료의 종류 사이에도 차이가 있다고 할 수 있다.
- p-value는 블록간이 0.0394, 처리방법인 비료간이 0.0025이다. 이로부터 처리방법간에 차이 있음이 블록간에 차이 있음보다 통계적으로 더 유의함(significant)을 알 수 있다.