

- 일반통계학 -
제 9장
분산분석

9-1 분산분석

1.1 분산분석(Analysis of variance: ANOVA)이란

- (1) 두 모평균의 차에 대한 검정(6장)의 확장으로 **3개 이상의 모평균의 차에 대한 비교**를 위한 대표적인 방법
- (2) 특성값의 분산 또는 변동을 분석하는 방법
- (3) 특성값의 변동을 제곱합으로 나타내고, 이 제곱합을 실험에 관련된 요인별로 분해하여, 오차에 비해 큰 영향을 주는 요인이 무엇인가를 찾아내는 분석 방법

(예제) 금속 가공품의 인장강도가 여러 공법에 따라 차이를 보이는가?

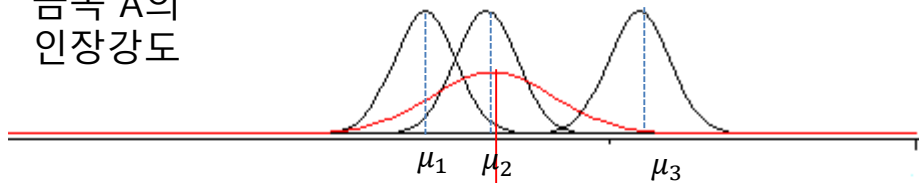
- 인장강도 : 특성값
- 공법 : 요인
- 작업자들의 능력차이와 같이 인장강도에 영향을 주지만 아직 원인이 규명되지 않은 부분 : 오차



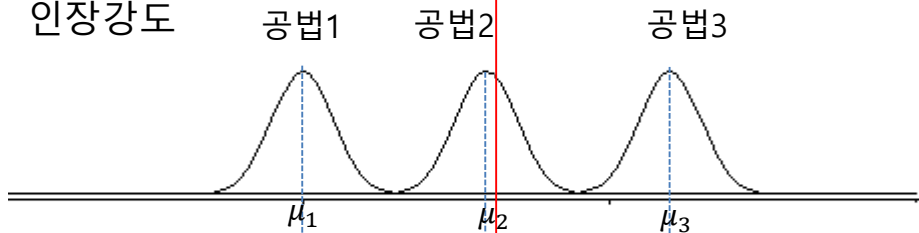
9-1 분산분석

Q. 모평균의 차이에 대한 귀무가설 가설 $H_0: \mu_1 = \mu_2 = \mu_3$ 를 검정하는데 왜 분산(변동)을 비교할까?

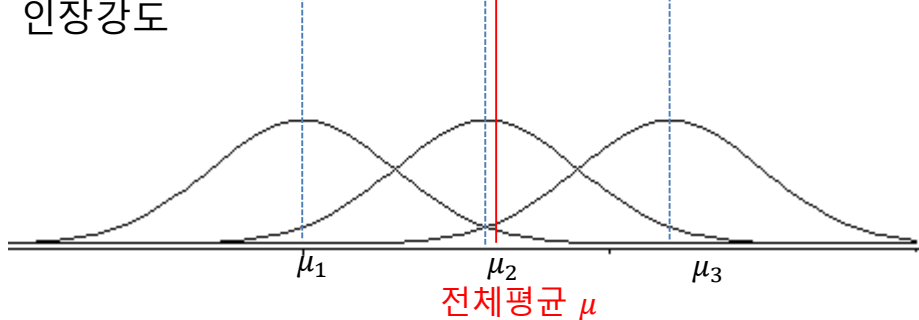
금속 A의
인장강도



금속 B의
인장강도



금속 C의
인장강도



금속 B는 공법별로 인장강도의 평균에 차이가 있다.
금속 C는 공법별로 인장강도의 평균에 차이가 있다고 단정하기 어려움.

-공법별 분포에서 겹치는 부분이 많음.

-공법 3에는 공법 2보다 인장강도가 낮은 값들이 적지 않고 공법2에도 공법1보다 인장강도가 낮은 값들이 많이 있음.

평균들간의 차이 뿐만 아니라 집단에 속한 값들의 집단내 분산이 집단 간 평균의 차이에 대한 판단에 영향을 미치고 있음.

공법 평균들 간의 분산이 크면 클수록 반면에 공법 내 분산은 작으면 작을수록, 공법 간 평균의 차이가 분명함을 알 수 있음.

금속 A는 공법내 분산은 금속B처럼 작으나 공법 평균들간의 분산 또한 작아서 공법별 분포에서 많은 부분이 서로 중복되어 공법별 평균인장강도에 대한 차이가 명확하다고 단정하기 어렵다.

출처 : 이영훈의 연구방법론

9-2 일원배치법 (One-way ANOVA)

2.1 통계적 실험

(1) 실험단위와 처리

실험이 행해지는 개체를 **실험단위**라 하고, 각각의 실험단위에 특정한 실험환경 또는 실험조건을 가하는 것을 **처리**라고 한다.

(2) 반응변수와 인자 및 인자수준

통계적 실험에서, 실험환경이나 실험조건을 나타내는 변수를 인자라 하고, 이에 대한 반응을 나타내는 변수를 **반응변수**라고 한다. 인자가 취하는 값을 그 **인자의 수준**이라고 한다.

2.2 일원배치법

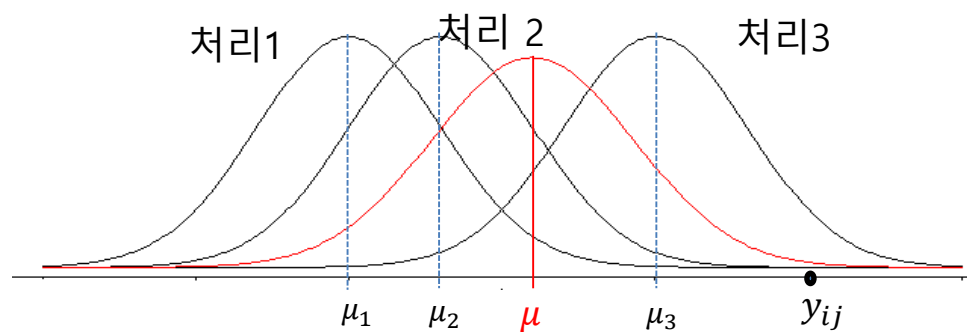
- 특성값에 대한 한 종류의 인자만의 영향을 조사하고자 할 때 사용
- 3개이상의 처리효과를 비교
- 각 수준에서의 반복수는 같지 않아도 좋으며 보통 3~5수준, 반복수 3~10을 사용
- 실험이 랜덤하게 선택된 순서에 의해 시행되어야 하므로 완전랜덤화계획이라고도 함.

(예) 금속가공품의 인장강도에 차이가 있는지의 여부를 검토하고 공법이 좋은가를 알고 싶을 때 사용함.

9-2 일원배치법

2.3 일원배치법의 자료구조

	처리1	처리2	...	처리 k	
	y_{11}	y_{21}		y_{k1}	
	y_{12}	y_{22}		y_{k2}	
	\vdots	\vdots		\vdots	
	y_{1n_1}	y_{2n_2}		y_{kn_k}	
평균	$\bar{y}_{1.}$	$\bar{y}_{2.}$		$\bar{y}_{k.}$	총평균 $\bar{y}_{..}$



모형: $Y_{ij} = \mu_i + e_{ij} = \mu + (\mu_i - \mu) + e_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \dots, k, j = 1, \dots, n_i, N = \sum_{i=1}^k n_i,$
 $\mu = \sum_{i=1}^k n_i \mu_i / N$: 처리효과 전체의 모평균, α_i : i 번째 처리효과

9-2 일원배치법

- 일원배치법의 모형

$$\begin{cases} Y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \dots, k, j = 1, \dots, n_i \\ e_{ij} \sim N(0, \sigma^2) \text{이고 서로 독립} \\ \sum_{i=1}^k n_i \alpha_i = 0 \end{cases}$$

- 총편차의 분해식

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

- 총제곱합의 분해식

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^k n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$
$$\text{SST } (N - 1) \quad = \text{SStr } (k - 1) \quad + \text{SSE } (N - k)$$

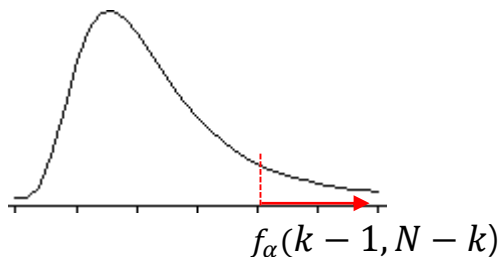
$$\text{총제곱합} = \text{처리제곱합(급간제곱합)} + \text{잔차제곱합(급내제곱합)}$$

9-2 일원배치법

- 처리효과의 유의성에 대한 가설

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \ (\Leftrightarrow \mu_1 = \mu_2 = \cdots = \mu_k = 0)$ vs H_1 : 적어도 한 α_i 는 0이 아니다.

- 처리효과가 유의하다면, 총제곱합 중에서 처리제곱합이 차지하는 비중이 커지고 잔차제곱합이 차지하는 비중이 작아질 것이다. 즉, $F = \text{MStr}/\text{MSE}$ 의 값이 커질 수록 처리효과가 유의하다는 증거가 강해지는 것이다. ($\text{MStr} = \text{SStr}/(k - 1)$, $\text{MSE} = \text{SSE}/(N - k)$)
- 귀무가설 H_0 가 사실일 때, $F = \text{MStr}/\text{MSE} \sim F(k - 1, N - k)$



9-2 일원배치법

- 분산분석표

요인	제곱합	자유도	평균제곱	F값	유의확률
처리	SStr	$k - 1$	$MStr = SStr / (k - 1)$	$f = MStr / MSE$	$P(F \geq f)$
잔차	SSE	$N - k$	$MSE = SSE / (N - k)$		
계	SST	$N - 1$			

(예제) 3곳의 자동차회사에서 생산한 경승용차의 리터 당 평균주행거리를 비교하고자 한다. 같은 조건하에서 실시한 주행거리실험 결과 다음과 같다. 3곳의 자동차회사에서 생산한 경승용차의 리터 당 평균주행거리간에 차이가 있는지 유의수준 5%에서 검정하여라.

	A자동차	B자동차	C자동차	
	16.5	15.3	19	
	18	14.8	18.4	
	14.1	16.1	15.3	
	17.8		17.3	
평균	16.6	15.4	17.5	16.6

요인	제곱합	자유도	평균제곱합	F값
처리 오차				
전체	26.02			

$$f_{0.05}(2,8) = 4.46$$

9-2 일원배치법

(예제) 어떤 직물의 가공시 처리액의 농도가 직물의 인장강도에 영향을 미치는지의 여부를 조사하기 위해, 네 가지 농도에서 반복 각 5회, 총 20회를 랜덤하게 처리한 후 인장강도를 측정한 결과가 아래와 같다. 이 자료에 대하여 일원배치법의 모형을 적용할 때, 농도에 따른 인장강도에 차이가 있는지를 알아보기 위한 분산분석표를 작성하고 적절한 가설을 유의수준 5%에서 검정하여라.

A	B	C	D	요인	제곱합	자유도	평균제곱	F값	유의확률
47	51	50	22	처리	1826.55	3			0.0045
58	62	38	23	잔차					
51	31	47	28	계	3334.55				
61	46	27	42						
46	49	23	25						
평균	52.6	47.8	37.0	28.0	총평균	41.35			

$$Y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \dots, 4, j = 1, \dots, 5$$

$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ VS H_1 : 적어도 한 α_i 는 0이 아니다.

9-2 일원배치법

- **분산분석의 기본 가정**

1. 모집단은 정규분포를 따른다.
2. 모집단은 동일한 분산을 갖는다.
3. 모든 표본은 서로 독립적으로 추출한다.

- **분산 분석을 하지 않고 두 평균에 대한 t-test를 반복 시행한다면...**

1. 두 개씩 쌍을 이루어 계산하므로 복잡해진다.
2. 제 1종 오류가 증가하게 된다.

9-3 반복이 없는 이원배치법 (요인이 2개)

3.1 반복이 없는 이원배치법의 자료구조

인자 B 인자 A	B_1	...	B_j	...	B_q	평균
A_1	y_{11}	...	y_{1j}	...	y_{1q}	$\bar{y}_{1.}$
\vdots	\vdots		\vdots		\vdots	\vdots
A_i	y_{i2}	...	y_{ij}	...	y_{iq}	$\bar{y}_{i.}$
\vdots	\vdots		\vdots		\vdots	\vdots
A_p	y_{p1}	...	y_{pj}	...	y_{pq}	$\bar{y}_{p.}$
평균	$\bar{y}_{.1}$...	$\bar{y}_{.j}$		$\bar{y}_{.q}$	총평균 $\bar{y}_{..}$

- 이원배치법에서 완전랜덤화계획 : pq 회의 실험을 랜덤하게 선택된 순서에 의하여 시행하는 것

9-3 반복이 없는 이원배치법

- 이원배치법의 모형

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, i = 1, \dots, p, j = 1, \dots, q \\ e_{ij} \sim N(0, \sigma^2) \text{이고 서로 독립} \\ \sum_{i=1}^p \alpha_i = 0, \sum_{j=1}^q \beta_j = 0 \end{cases}$$

α_i : 인자 A의 i번째 처리효과, β_j : 인자 B의 j번째 처리효과

- 총편차의 분해식

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

- 총제곱합의 분해식

$$\sum_{i=1}^p \sum_{j=1}^q (y_{ij} - \bar{y}_{..})^2 = q \sum_{i=1}^p (\bar{y}_{i.} - \bar{y}_{..})^2 + p \sum_{j=1}^q (\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^p \sum_{j=1}^q (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

$$SST (pq - 1) = SStr_A (p - 1) + SStr_B (q - 1) + SSE (p - 1) (q - 1)$$

$$\text{총제곱합} = \text{요인 A의 제곱합} + \text{요인 B의 제곱합} + \text{잔차제곱합}$$

9-3 반복이 없는 이원배치법

- 분산분석표

요인	제곱합	자유도	평균제곱	F값	유의확률
인자A	$SStr_A$	$p - 1$	$MStr_A = SStr_A / (p - 1)$	$f_1 = MStr_A / MSE$	$P(F \geq f_1)$
인자B	$SStr_B$	$q - 1$	$MStr_B = SStr_B / (q - 1)$	$f_2 = MStr_B / MSE$	$P(F \geq f_2)$
잔차	SSE	$(p - 1)(q - 1)$	$MSE = SSE / ((p - 1)(q - 1))$		
계	SST	$pq - 1$			

- 인자 A의 유의성 검정 - $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ vs H_1 : 적어도 한 α_i 는 0이 아니다.
- 인자 B의 유의성 검정 - $H_0 : \beta_1 = \beta_2 = \dots = \beta_q = 0$ vs H_1 : 적어도 한 β_j 는 0이 아니다.
- 귀무가설 H_0 하에서

$$F = MStr_A / MSE \sim F(p - 1, (p - 1)(q - 1)), \quad F = MStr_B / MSE \sim F(q - 1, (p - 1)(q - 1))$$

9-3 반복이 없는 이원배치법

(예제) 지역과 비료의 종류에 따라 토마토 생산량에 차이가 있는지를 확인하기 위하여 4개 지역에서 각각 A,B,C 세 종류의 비료를 적용시킨 후에 생산량을 조사한 결과 다음의 자료를 얻었다. 지역과 비료의 종류에 따라 토마토 생산량에 차이가 있는지를 유의수준 5%에서 검정하여라

지역 \ 비료			
	A	B	C
지역1	42.8	52.3	48.2
지역2	38.6	43.5	40.3
지역3	50.2	58.7	53.5
지역4	48.2	50.8	51.2

요인	제곱합	자유도	평균제곱	F 값
인자A	280.909		93.636	30.42
인자B	81.352		40.676	13.21
잔차	18.468		3.078	
계	380.729	11		

$$F(0.05, 2, 6)=5.14, F(0.05, 3, 6)=4.76$$

9-4 반복이 있는 이원배치법

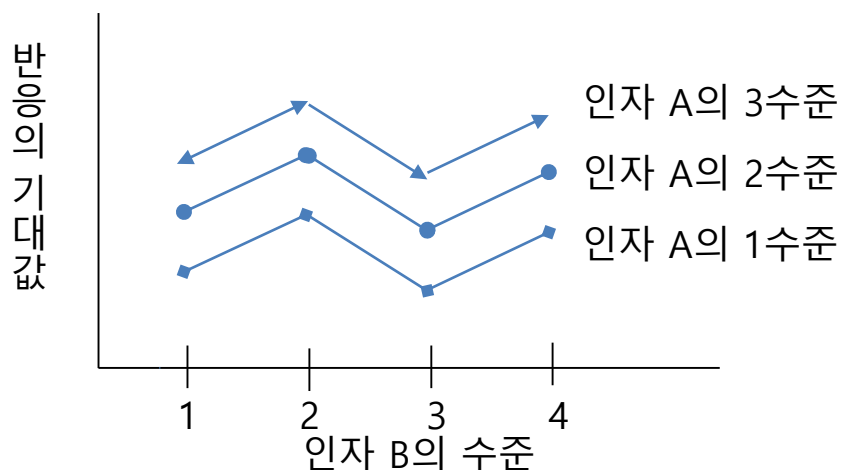
4.1 반복이 있는 이원배치법의 자료구조

인자 B 인자 A	B_1	B_2	...	B_q	평균
A_1	y_{111}	y_{121}	...	y_{1q1}	$\bar{y}_{1..}$
	y_{112}	y_{122}	...	y_{1q2}	
	\vdots	\vdots		\vdots	
	y_{11r}	y_{12r}	...	y_{1qr}	
	$\bar{y}_{11.}$	$\bar{y}_{12.}$		$\bar{y}_{1q.}$	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_p	y_{p11}	y_{p21}	...	y_{pq1}	$\bar{y}_{p..}$
	y_{p12}	y_{p22}	...	y_{pq2}	
	\vdots	\vdots		\vdots	
	y_{p1r}	y_{p2r}	...	y_{pqr}	
	$\bar{y}_{p1.}$	$\bar{y}_{p2.}$		$\bar{y}_{pq.}$	
평균	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$		$\bar{y}_{.q.}$	총평균 $\bar{y}_{...}$

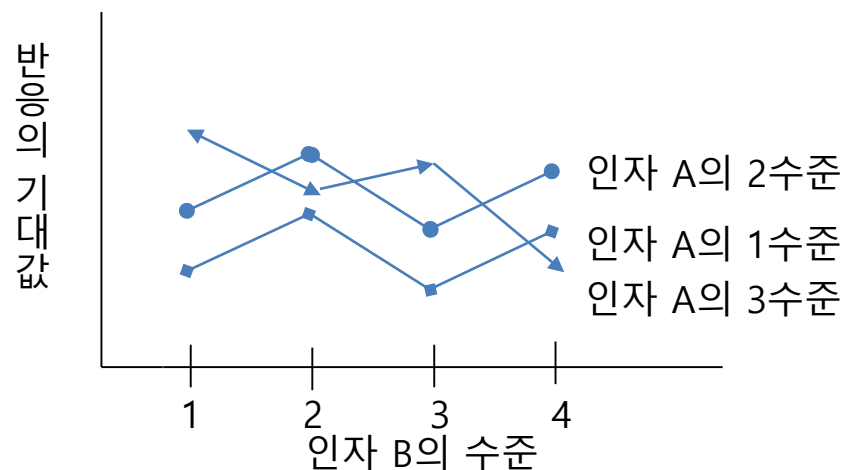
- 반복이 있는 이원배치법에서 완전랜덤화계획: pqr 회의 실험을 랜덤하게 선택된 순서에 의하여 시행하는 것
- 인자수준의 조합에서 생기는 효과인 교호작용을 분리하여 구할 수 있다. 교호작용은 인자 A의 효과가 인자 B의 수준에 따라 변하는 모형에서 존재한다.

9-4 반복이 있는 이원배치법

- 반복이 없는 이원배치법의 모집단 모형에서 $E(e_{ij}) = 0$ 이므로 $E(Y_{ij}) = \mu + \alpha_i + \beta_j$ 가 된다. $(\mu + \alpha_1 + \beta_j) - (\mu + \alpha_2 + \beta_j) = \alpha_1 - \alpha_2$ 가 되어 인자 A의 각 수준에서의 기댓값의 차이는 인자 B의 수준에 무관하게 된다. 이를 그림으로 나타내면 A의 각 수준을 연결하는 선분들이 평행하면 두 인자 A, B 사이에 교호작용이 존재하지 않음을 뜻하게 된다. A의 각 수준에서의 기대값의 차이가 B의 수준에 따라 변하게 될 때 A와 B는 교호작용이 있다고 하며 이 경우에는 반복이 없는 이원배치법을 사용해서는 안된다.



교호작용이 없는 경우



교호작용이 있는 경우

9-4 반복이 있는 이원배치법

- 이원배치법의 모형 (반복이 있는 경우)

$$\begin{cases} Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, i = 1, \dots, p, j = 1, \dots, q, k = 1, \dots, r \\ e_{ijk} \sim N(0, \sigma^2) \text{이고 서로 독립} \\ \sum_{i=1}^p \alpha_i = 0, \sum_{j=1}^q \beta_j = 0, \sum_{i=1}^p \gamma_{ij} = 0, \sum_{j=1}^q \gamma_{ij} = 0 \end{cases}$$

α_i : 인자 A의 i번째 처리효과, β_j : 인자 B의 j번째 처리효과, γ_{ij} : 인자 A의 i번째 수준과 인자 B의 j번째 수준의 교호작용 효과

- 총편차의 분해식

$$y_{ijk} - \bar{y}_{...} = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

- 총제곱합의 분해식

$$\sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^r (y_{ijk} - \bar{y}_{...})^2 = qr \sum_{i=1}^p (\bar{y}_{i..} - \bar{y}_{...})^2 + pr \sum_{j=1}^q (\bar{y}_{.j.} - \bar{y}_{...})^2 +$$

$$r \sum_{i=1}^p \sum_{j=1}^q (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^r (y_{ijk} - \bar{y}_{ij.})^2$$

$$SST (pqr - 1) = SStr_A (p - 1) + SStr_B (q - 1) + SStr_{A \times B} (p - 1) (q - 1) + SSE pq(r - 1)$$

총제곱합 = 요인 A의 제곱합 + 요인 B의 제곱합 + **교호작용 제곱합** + 오차제곱합

9-4 반복이 있는 이원배치법

- 분산분석표

요인	제곱합	자유도	평균제곱	F값	유의확률
인자A	$SStr_A$	$p - 1$	$MStr_A = SStr_A / (p - 1)$	$f_1 = MStr_A / MSE$	$P(F \geq f_1)$
인자B	$SStr_B$	$q - 1$	$MStr_B = SStr_B / (q - 1)$	$f_2 = MStr_B / MSE$	$P(F \geq f_2)$
교호작용	$SStr_{A \times B}$	$(p - 1)(q - 1)$	$MStr_{A \times B} = SStr_{A \times B} / ((p - 1)(q - 1))$	$f_3 = MStr_{A \times B} / MSE$	$P(F \geq f_3)$
잔차	SSE	$pq(r - 1)$	$MSE = SSE / pq(r - 1)$		
계	SST	$pqr - 1$			

- 교호작용의 유의성 검정 - $H_0 : r_{ij} = 0, i = 1, \dots, p, j = 1, \dots, q$ vs H_1 : 적어도 한 r_{ij} 는 0이 아니다.
- 귀무가설 H_0 하에서

$$F = MStr_{A \times B} / MSE \sim F((p - 1)(q - 1), pq(r - 1))$$

- 교호작용이 존재하지 않는다면 요인 A와 요인 B의 효과를 검정
- 교호작용이 존재한다면 각 요인별 추가적인 검정은 실시하지 않음

9-4 반복이 있는 이원배치법

(예) 리탈린(retalin)이 정상아동과 과잉운동아동에 미치는 영향을 조사

아동 \ 투여약	투여약		요인	제곱합	자유도	평균제곱	F 값
	위약	리탈린					
정상아	50 45	67 60	인자A	121	1	121	8
	55 52	58 65	인자B	42.25	1	42.25	2.79
과잉운동아	70 68	51 57	교호작용	930.25	1	930.25	61.50
	72 75	48 55	잔차	181.5	12	15.125	
			계	1275	15		

$F(0.05, 1, 12) = 4.747$

- $H_0: r_{ij} = 0, i = 1, 2, j = 1, 2$ vs H_1 : 적어도 한 r_{ij} 는 0이 아니다.
- $H_0: \alpha_1 = \alpha_2 = 0$ vs H_1 : 적어도 한 α_i 는 0이 아니다.
- $H_0: \beta_1 = \beta_2 = 0$ vs H_1 : 적어도 한 β_j 는 0이 아니다

9-4 반복이 있는 이원배치법

(예제) 세 종류의 기계와 세 사람의 기능공이 제품품질에 미치는 영향을 조사하고자 하여 2회 반복이 있는 이원배치법에 의해 생산성을 측정한 결과 다음의 자료를 얻게 되었다. 이 자료에 대하여 이원배치법의 모형을 적용할 때, 각 인자와 교호작용의 효과에 대하여 유의수준 5%에서 검정하여라

	B_1	B_2	B_3	평균
A_1	9, 14	14,16	19,22	15.67
A_2	13,16	18,26	14,18	17.5
A_3	11,12	11,17	15,16	13.67
평균	12.5	17	17.33	총평균15.61

요인	제곱합	자유도	평균제곱	F값	유의확률
인자A	44.11				0.1455
인자B	87.44				0.0387
교호작용	74.22				0.1744
잔차	82.51				
계	288.28				

- $H_0: r_{ij} = 0, i = 1, \dots, 3, j = 1, \dots, 3$ vs H_1 : 적어도 한 r_{ij} 는 0이 아니다.
- $H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$ vs H_1 : 적어도 한 α_i 는 0이 아니다.
- $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs H_1 : 적어도 한 β_j 는 0이 아니다