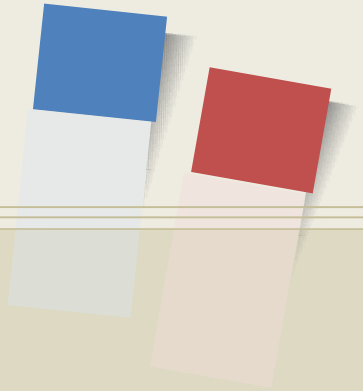


13주1강 분산분석2

9.2절

9.2 일원분류 분산분석

일원분류 분산분석 (one way ANOVA)

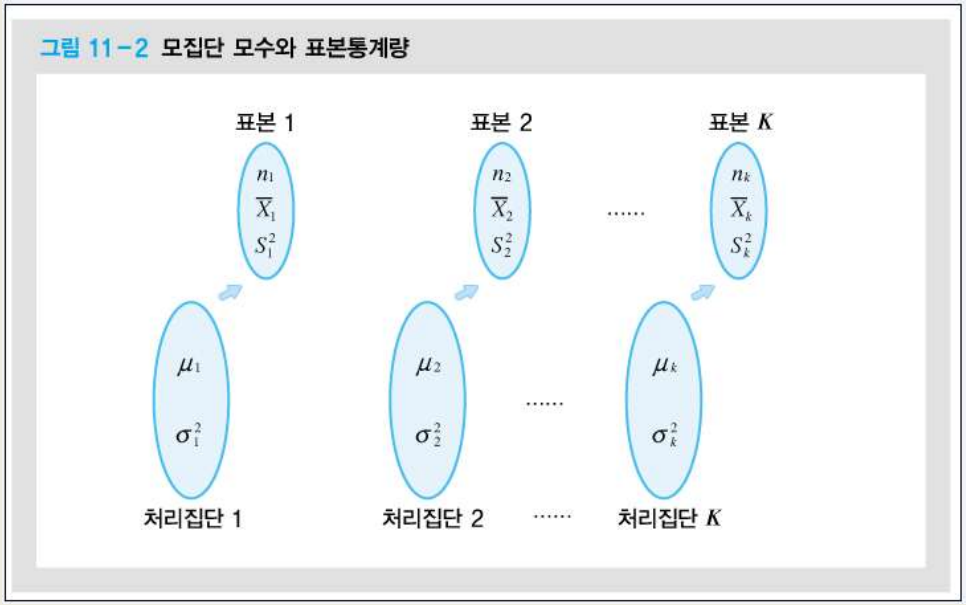
- 통계적 실험에서 두 개 이상의 집단으로부터 자료를 관측하였을 때 각 집단의 반응에 차이가 있는가를 분석하는 통계적 분석방법을 **분산분석 (ANOVA: analysis of variance)** 이라고 한다.

- 그 중 **요인이 하나인 실험**에 대한 분산분석법을 **일원분류 분산분석 (one way ANOVA)** 이라고 한다.

- 일원분류 분산분석은 실험단위들이 처리의 각 수준에 랜덤하게 배정되는 완전확률화 계획법에 의하여 실험이 실시된 것을 전제로 한다.

일원분류 분산분석의 자료 형태

- 실험에서 비교하고자 하는 K개의 처리집단이 있을 때, 처리집단의 모평균이 모두 같은지에 대한 검정을 실시하기 위하여 각 집단에서 n_i개, $(i = 1, 2, \ldots, K)$ 의 표본을 추출하여 실험을 실시한 후에 관측한 반응의 표본평균과 분산을 각각 μ_i, σ_i^2, $(i = 1, 2, \ldots, K)$ 로 표현할 수 있다.



① 분석의 전제조건

* 각 처리집단의 모분산은 동일하다. [$\sigma_i^2 = \sigma^2, \ i = 1, 2, \dots, K$]
* 각 처리집단의 관찰값은 모두 정규분포를 따른다. [$N(\mu_i, \sigma^2), i = 1,2, \dots, K$]

② 가설 설정

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

→ 모든 처리집단의 모평균이 동일하다는 것을 의미

$H_1 : H_0$가 사실이 아니다.

→ 모평균 모두가 동일하지는 않다는 것, 즉 최소한 하나는 다른 값과 다르다는 것을 의미

검정과정

- 확률변수 X 를 K개의 처리집단에서 관측된 반응이라고 할 때 다음과 같이 표현할 수 있다.

$$\boldsymbol{X_{ij}, i = 1, 2, \ldots, K, j = 1, 2, \ldots, n_i}$$

첫 번째 첨자 i : i번째 처리집단을 의미한다.

두 번째 첨자 j : i번째 처리집단에서 j번째 관측값임을 나타낸다.

모든 관측값을 표로 나타내면 다음과 같다.

표 11-1 일원분류 분산분석의 자료

처리집단	1	2	…	K
관측값	X_{11} X_{12} ⋮ X_{1n_1}	X_{21} X_{22} ⋮ X_{2n_2}	…	X_{K1} X_{K2} ⋮ X_{Kn_K}
합	X_1	X_2	…	X_K
평균	$\overline{X}_1$	$\overline{X}_2$	…	$\overline{X}_K$

① 제곱합 계산

- i번째 처리집단의 평균과 전체관측값의 평균은 다음과 같이 나타낼 수 있다.

$$\overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \qquad \overline{X}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$$

개별관측값과 전체평균과의 편차는 다음과 같이 나타낼 수 있다.

$$X_{ij} - \overline{X}_{\cdot\cdot} = (X_{ij} - \overline{X}_{i\cdot}) + (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})$$

- $X_{ij} - \overline{X}_{i\cdot}$: 각 관측값과 그 관측값이 속한 처리집단의 평균과의 차이로 반응에 대한 **오차효과를 측정하는 값**이다.
- $\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}$: 각 처리의 평균과 전체평균과의 차이로 반응에 대한 **처리효과를 측정하는 값**이다.

- 편차의 모든 관측값에 대한 제곱합은 다음과 같이 나타낼 수 있다.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{..})^2=\sum_{i=1}^{k}n_i(\overline{X}_{i.}-\overline{X}_{..})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i.})^2$$

총제곱합 (TSS) = 처리제곱합 (SS_T) + 오차제곱합 (SS_E)

$$(X_{ij}-\overline{X}_{..})^2=(X_{ij}-\overline{X}_{i.}+\overline{X}_{i.}-\overline{X}_{..})^2=(X_{ij}-\overline{X}_{i.})^2+2(X_{ij}-\overline{X}_{i.})(\overline{X}_{i.}-\overline{X}_{..})+(\overline{X}_{i.}-\overline{X}_{..})^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{..})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i.})^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}[X_{ij}(\overline{X}_{i.}-\overline{X}_{..})-\overline{X}_{i.}(\overline{X}_{i.}-\overline{X}_{..})]+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{X}_{i.}-\overline{X}_{..})^2$$

한편, $$\sum_{i=1}^{k}\sum_{j=1}^{n_i}[X_{ij}(\overline{X}_{i.}-\overline{X}_{..})-\overline{X}_{i.}(\overline{X}_{i.}-\overline{X}_{..})]=\sum_{i=1}^{k}(\overline{X}_{i.}-\overline{X}_{..})\sum_{j=1}^{n_i}X_{ij}-\sum_{i=1}^{k}n_i\overline{X}_{i.}(\overline{X}_{i.}-\overline{X}_{..})=0$$

* 처리제곱합 (SS_T)

- K개 집단의 평균과 전체평균과의 편차의 제곱합으로 자유도가 $K-1$ 이다.

* 오차제곱합 (SS_E)

- 서로 독립인 K개의 처리집단에서 각 관측값과 처리집단평균과의 편차의 제곱합으로 자유도는 $N-K$ 이다.

* 평균제곱합

- 각각의 제곱합을 자유도로 나눈 값을 평균제곱합이라고 하며 다음과 같이 한다.

$$\boldsymbol{MS_T = SST / (K-1)}$$

$$\boldsymbol{MS_E = SSE / (N-K)}$$

* 검정통계량 F

- 귀무가설 하에서의 검정통계량의 값은 다음과 같다.

$$\boldsymbol{F = MST / MSE}$$

- 검정통계량 F는 자유도가 $(K-1, N-K)$인 F-분포를 따른다.
- 따라서 유의수준 α 하에서의 검정은 위와 같이 구한 F 값이 $F_{(K-1,N-K)}$보다 크면 귀무가설을 기각하게 된다.

② 분산분석표의 작성

- 지금까지의 제곱합과 자유도, 평균제곱합을 이용하여 **분산분석표(ANOVA table)**를 다음과 같이 작성할 수 있다.

표 11-2 일원분류 분산분석표

변인(source)	자유도(*d.f.*)	제곱합(*SS*)	평균제곱합(*MS*)
처리(treatment)	$K-1$	$\sum_{i=1}^{K} n_i(\overline{X}_i - \overline{X}_{..})^2$	$SS_T/(K-1)$
오차(error)	$N-K$	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i.})^2$	$SS_E/(N-K)$
전체(total)	$N-1$	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{..})^2$	

- $n_1 = n_2 = \cdots = n_k = n$이라 두면
- $N = nK$
- $N - K = nK - K = (n-1)K$
- $N - 1 = nK - 1$

$$\sum_{i=1}^{k} n_i \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2 = n \sum_{i=1}^{k} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - \overline{X}_{\cdot\cdot}\right)^2$$

- 각 처리집단의 관측값의 수가 같은 경우 각 처리집단의 관측치의 수가 같으면 (즉 $n_1 = n_2 = \cdots = n_k = n$) 분산분석표의 작성과정이 좀더 간단해진다. 즉, 처리집단의 수를 K라 하고 각 처리집단의 관측값의 수를 n이라고 할 때 분산분석표는 다음과 같다.

표 11-7 각 처리집단의 관측값의 수가 동일한 경우의 분산분석표

변인	*d.f.*	*SS*	*MS*	*F*
처리	$K-1$	$n\sum_{i=1}^{K}(\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot})^2$	$SS_T/(K-1)$	$\frac{MS_T}{MS_E}$
오차	$K(n-1)$	$\sum_{i=1}^{K}\sum_{j=1}^{n}(X_{ij}-\overline{X}_{i\cdot})^2$	$SS_E/K(n-1)$	
전체	$K\cdot n-1$	$\sum_{i=1}^{K}\sum_{j=1}^{n}(X_{ij}-\overline{X}_{\cdot\cdot})^2$		

- $X_{i.} = \sum_{j=1}^{n_i} X_{ij}$

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left({X_{ij}}^2 - 2X_{ij}\overline{X}_{..} + {\overline{X}_{..}}^2\right)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}{X_{ij}}^2 - 2\overline{X}_{..}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij} + N{\overline{X}_{..}}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}{X_{ij}}^2 - N{\overline{X}_{..}}^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}{X_{ij}}^2 - N\left(\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}{X_{ij}}^2 - \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}\right)^2$$

- $CM = \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}\right)^2$

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}{X_{ij}}^2 - CM$$

$$SST = \sum_{i=1}^{k} n_i \left(\overline{X}_{i.} - \overline{X}_{..}\right)^2 = \sum_{i=1}^{k} n_i \left(\overline{X}_{i.}^{\,2} - 2\overline{X}_{i.}\,\overline{X}_{..} + \overline{X}_{..}^{\,2}\right)$$

$$= \sum_{i=1}^{k} n_i \left[\left(\frac{X_{i.}}{n_i}\right)^2 - 2\overline{X}_{i.}\,\overline{X}_{..} + \overline{X}_{..}^{\,2}\right]$$

$$= \sum_{i=1}^{k} \frac{X_{i.}^{\,2}}{n_i} - 2\overline{X}_{..} \sum_{i=1}^{k} n_i \overline{X}_{i.} + N\overline{X}_{..}^{\,2}$$

$$= \sum_{i=1}^{k} \frac{X_{i.}^{\,2}}{n_i} - 2\overline{X}_{..}\left(N\,\overline{X}_{..}\right) + N\overline{X}_{..}^{\,2}$$

$$= \sum_{i=1}^{k} \frac{X_{i.}^{\,2}}{n_i} - N\overline{X}_{..}^{\,2} = \sum_{i=1}^{k} \frac{X_{i.}^{\,2}}{n_i} - CM$$

예9-2.

- 4가지 교육방법의 효과를 비교분석하기 위하여 학생 40명을 랜덤하게 10명씩 4개 집단으로 나누고 한 학기 동안 각 교육방법으로 교육을 실시한 후에 치른 학기말 시험성적이 다음과 같다. 학기 중에 질병이나 전학 등으로 인하여 학기말 시험을 치른 학생의 수가 같지 않은데, 다음에 주어진 자료에 의할 때 4가지 교육방법의 효과가 다르다고 할 수 있는가를 분석하라.

표 11-3 4가지 교육방법에 의한 학기말시험 성적

교육방법	1	2	3	4
시험 성적	65	75	59	94
	87	69	78	89
	73	83	67	80
	79	81	62	88
	81	72	83	
	69	79	76	
		90		
시험 성적($X_{\cdot j}$)	454	549	425	351

- **(sol)**

$$CM = \frac{\left(\sum_{i=1}^{4}\sum_{j=1}^{n_i} X_{ij}\right)^2}{N} = \frac{(1779)^2}{23} = 137,601.8$$

$$TSS = \sum_{i=1}^{4}\sum_{j=1}^{n_i} X_{ij}^2 - CM$$
$$= (65)^2 + (87)^2 + (73)^2 + \cdots + (88)^2 - CM$$
$$= 139,511 - 137,601.8$$
$$= 1,909.2$$

$$SS_T = \sum_{i=1}^{4} \frac{X_i^2}{n_i} - CM$$
$$= \frac{(454)^2}{6} + \frac{(549)^2}{7} + \frac{(425)^2}{6} + \frac{(351)^2}{4} - CM$$
$$= 138,314.4 - 137,601.8$$
$$= 712.6$$

$$SS_E = TSS - SS_T = 1,909.2 - 712.6 = 1,996.6$$

$$MS_T = \frac{SS_T}{4-1} = \frac{712.6}{3} = 237.5$$

$$MS_E = \frac{SS_E}{N-4} = \frac{1,196.6}{23-4} = 63.0$$

$$F = \frac{MS_T}{MS_E} = \frac{237.5}{63.0} = 3.77$$

- 계산한 값을 가지고 분산분석표를 작성해 보면 다음과 같은 표를 만들 수 있다.

표 11-4 교육효과 분석자료의 분산분석표

변인	*d.f.*	*SS*	*MS*	*F*
처리	3	712.6	237.5	3.77
오차	19	1,196.6	63.0	
전체	22	1,909.2		

- $\mu_1, \mu_2, \mu_3, \mu_4$를 각각의 교육방법에 있어서 모집단의 평균성적이라고 할 때, 검정하고자 하는 가설은 다음과 같다.

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

  $H_1 : H_0$이 사실이 아니다.

- 귀무가설하에서의 검정통계량의 값 $F = 3.77$은 분포를 따른다. $\alpha = 0.05$일 때 $F_{3,19} = 3.13$ 이므로 $F = 3.77 > F_{0.05,3,19}$가 되어 귀무가설은 기각된다.
- 즉, 주어진 자료에 의할 때 5% 유의수준 하에서 4가지 교육방법의 효과가 동일하다고 볼 수 없다.

(완전 확률화 계획법에서 모수의 추정)

$100(1-\alpha)\%$ 신뢰구간

① 단일 모평균 μ : $\overline{X}_i \pm t_{\alpha/2} \cdot \frac{S}{\sqrt{n_i}}$ (여기서 $S = \sqrt{MS_E} = \sqrt{\frac{SS_E}{N-K}}$)

② 두 모평균의 차 $\mu_i - \mu_j$: $\left(\overline{X}_{i\cdot} - \overline{X}_{j\cdot}\right) \pm t_{\alpha/2} \cdot S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ (여기서 $S = \sqrt{MS_E} = \sqrt{\frac{SS_E}{N-K}}$)

예9-3.

- [예 9.2]에 주어진 자료를 이용한 분석에서 첫째 교육방법의 효과(μ_1)에 대한 95% 신뢰구간과, 첫째 교육방법과 넷째 교육방법의 효과의 차이($\mu_1-\mu_4$)에 대한 95% 신뢰구간을 구하라.
- (sol)

$$\overline{X}_{1\cdot} = \frac{454}{6} = 75.67$$
$$\overline{X}_{4\cdot} = \frac{351}{4} = 87.75$$
$$S = \sqrt{MS_E} = \sqrt{63.0} = 7.94$$

S의 $d.f. = 19$

- 자유도가 19인 t-분포에서 $t_{0.025,19} = 2.093$ 이다.
- 따라서 μ_1의 95% 신뢰구간은 다음과 같다.

$$\overline{X}_{1\cdot} \pm t_{(0.025\,:\,19)}\frac{S}{\sqrt{n_1}} = 75.67 \pm 2.093\frac{7.94}{\sqrt{6}} = 75.67 \pm 6.78 = (68.89, 82.45)$$

- $\mu_1 - \mu_4$의 95% 신뢰구간은 다음과 같다.

$$(\overline{X}_{1\cdot} - \overline{X}_{4\cdot}) \pm t_{(0.025\,:\,19)} \cdot S\sqrt{\frac{1}{n_1}+\frac{1}{n_4}} = (75.67 - 87.75) \pm (2.093)\cdot(7.94)\sqrt{\frac{1}{6}+\frac{1}{4}} = -12.08 \pm 10.73 = (-22.81, -1.35)$$

〈자료 1〉

처리	1	2	3
관측값	60	67	72
	65	71	75
	70	72	78
평균	65	70	75

〈자료 2〉

처리	1	2	3
관측값	40	50	50
	60	70	80
	95	90	95
평균	65	70	75

그림 9-4 〈자료 1〉의 관측값

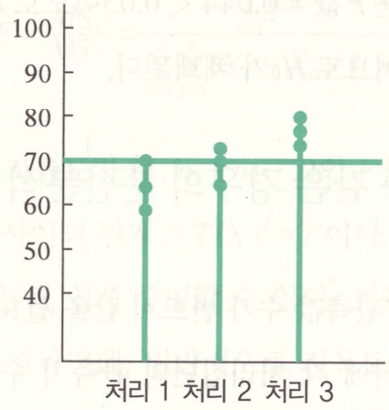


그림 9-5 〈자료 2〉의 관측값

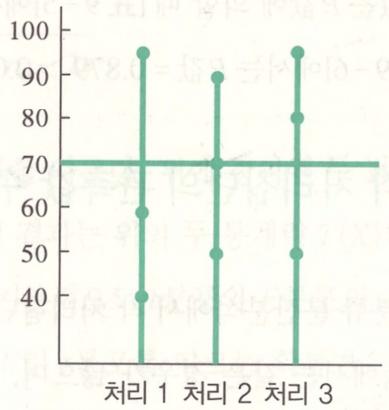


표 9-5 〈자료 1〉에 대한 분산분석표

변인	*d.f.*	*SS*	*MS*	*F*	*P*
처리	2	150.0	75.0	5.49	0.044
오차	6	82.0	13.7		
전체	8	232.0			

표 9-6 〈자료 2〉에 대한 분산분석표

변인	*d.f.*	*SS*	*MS*	*F*	*P*
처리	2	150	75	0.13	0.879
오차	6	3,400	567		
전체	8	3,550			

- <표9-5>에서 $F = 5.49 > F_{(0.05:2,6)} = 5.14$이므로
  $H_0 : \mu_1 = \mu_2 = \mu_3$가 기각
- <표9-5>에서 $F = 0.13 < F_{(0.05:2,6)} = 5.14$이므로
  $H_0 : \mu_1 = \mu_2 = \mu_3$가 채택
- <자료1>보다 <자료2>가 각 처리집단 내에서 관측값의 폭이 크다.

- 두 개의 독립집단에 대한 모평균의 동일성($\mu_1 = \mu_2$)의 검증

$$T(X) = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$F = \frac{MS_T}{MS_E} = \frac{\sum_{i=1}^{2} n_i\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\overline{X}_{ij} - \overline{X}_{i\cdot}\right)^2/(n_1 + n_2 - 2)} \sim F(1, n_1 + n_2 - 2)$$

$$\Longrightarrow \quad T(X)^2 = F$$

끝~~ ♥♥