정보통계학과 **백재욱** 교수

제3강 (3장)

일원배치법

3.1 일원배치의 구조모형

3.2 분산분석

3.3 분산분석 후의 추정

3.4 반복수가 같지 않은 실험

3.5 랜덤모형

정보통계학과 **백재욱** 교수

제3강 일원배치법

3.1 일원배치의 구조모형

3.1 일원배치의 구조모형

가장 중요한 요인(인자)이 반응치에 어떤 영향을 미치는지 알고 싶다.

01 모수인자(고정인자)

- 관심의 대상이 되는 반응온도가, 예를 들어 80℃, 100 ℃, 120 ℃ 로 고정된 인자
- 인자의 수준에 따라서 반응치의 모평균이 달라지는가?
- 반응치에 대한 최적조건은 무엇인가?

02 변량인자(랜덤인자)

- 선정된 반응온도, 예를 들어 80℃, 100℃, 120℃ 가 많은 반응온도 중 일부인 경우 반응온도는 변량인자이다.
- 온도(인자)의 특정한 수준은 기술적으로 의미를 가지지 못한다.
- 온도를 A라고 하는 경우 A의 분산 $\sigma^2{}_A$로 A의 효과를 파악한다.

3.1 일원배치의 구조모형

〈표 3-1〉 일원배치법의 구조모형

		실험의 반복	합계	평균
인자의 수준	A_1	$x_{11}\ x_{12} \cdots\ x_{1r}$	$T_{1.}$	$\bar{x}_{1.}$
	A_2	$x_{21}\ x_{22} \cdots\ x_{2r}$	$T_{2.}$	$\bar{x}_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	A_a	$x_{a1}\ x_{a2} \cdots\ x_{ar}$	$T_{a.}$	$\bar{x}_{a.}$
			T	$\bar{\bar{x}}$

3.1 일원배치의 구조모형

01 모형

$$x_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- 모수모형의 경우 $\sum \alpha_i = 0$ 이라고 가정함

오차 ϵ_{ij}에 대한 가정

- 정규성 : 오차의 분포는 정규분포를 따른다.
- 독립성 : 오차들은 서로 독립이다.
- 불편성 : 오차의 기대값은 0으로 치우침이 없다.
- 등분산성 : 오차의 분산은 인자의 수준과 실험의 반복에 관계없이 일정하다.

정보통계학과 **백재욱** 교수

제3강 일원배치법

3.2 분산분석

3.2 분산분석

02 귀무가설

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

- 수준 효과 간 차이가 없다

03 대립가설

$H_1: \alpha_i$ 모두 0인 것은 아니다.

- 어떤 수준 효과 간 차이가 있다

3.2 분산분석

04 변동 및 자유도의 분해

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

총 변동 α_i에 기인하는 변동 잔차변동

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

$$SS_T \quad = \quad SS_A \quad + \quad SS_E$$

자유도 : $ar - 1 \quad = \quad a - 1 \quad + \quad a(r - 1)$

3.2 분산분석

05 변동의 간단한 표현

$$SS_T = \sum_i \sum_j {x_{ij}}^2 - CT, \qquad CT = \frac{T^2}{ar}$$

$$SS_A = \sum_i \frac{{T_{i.}}^2}{r} - CT$$

$$SS_E = SS_T - SS_A$$

3.2 분산분석

〈표 3-3〉 분산분석표

요인	제곱합	자유도	평균제곱	F_0	$F(\alpha)$
A	SS_A	$a-1$	MS_A	$\frac{MS_A}{MS_E}$	$F(a-1, a(r-1); \alpha)$
E	SS_E	$a(r-1)$	MS_E		
T	SS_T	$ar-1$			

앞의 가설에 대한 검정방법

- 만약 검정통계량 '$F_0 > F(a-1, a(r-1); \alpha)$'이면 유의수준 α에서 귀무가설 기각
- 유의확률을 구하여 '유의확률 p 값 〈 유의수준α'이면 대립가설 채택

3.2 분산분석

예 3.1 납품업체 A_1, A_2 , A_3 , A_4간 마모도에 유의한 차이가 있는가?

<표 3-4> 마모도 검사자료

		반 복				$T_{i.}$	
납품업체	A_1	1.93	2.38	2.20	2.25	8.76	2.19
	A_2	2.55	2.72	2.75	2.70	10.72	2.68
	A_3	2.40	2.68	2.32	2.28	9.68	2.42
	A_4	2.33	2.38	2.28	2.25	9.24	2.31
						38.4	2.40

3.2 분산분석

풀이

《 분산분석표 》

요 인	제곱합	자유도	평균제곱	F_0	$F(0.01)$
A	0.5240	3	0.1747	8.78 **	5.95
E	0.2386	12	0.0199		
T	0.7626	15			

검정통계량 F_0의 값 $8.78 > F(3,12;0.01) = 5.95$

➡ 귀무가설 기각

(어떤 납품업체들 간에는 직물 마모도 간 차이가 있다)

3.2 분산분석

R 실습

a1 = c(1.93, 2.38, 2.20, 2.25)

a2 = c(2.55, 2.72, 2.75, 2.70)

a3 = c(2.40, 2.68, 2.32, 2.28)

a4 = c(2.33, 2.38, 2.28, 2.25)

wear = c(a1, a2, a3, a4)

group = c("a1", "a2", "a3", "a4")

group <- rep(group, c(4, 4, 4, 4))

wear.dat <- data.frame(group, wear)

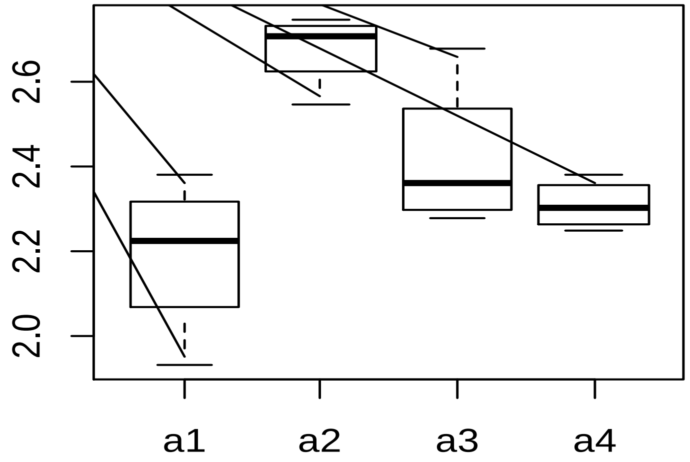
tapply(wear, group, sum)

a1 a2 a3 a4
8.76 10.72 9.68 9.24

tapply(wear, group, mean)

a1 a2 a3 a4
2.19 2.68 2.42 2.31

boxplot(wear ~ group)



aov.out = aov(wear ~ group, data=wear.dat)

summary(aov.out)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	3	0.5240	0.17467	8.785	0.00235 **
Residuals	12	0.2386	0.01988		

---

*Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

정보통계학과 **백재욱** 교수

제3강 일원배치법

3.3 분산분석 후의 추정

3.3 분산분석 후의 추정

◆ 각 수준에서 모평균 μ_i 의 추정

일원배치법의 경우 각 수준에서 μ_i 신뢰구간 : $\bar{x_{i.}} \pm t(\phi_E; \frac{\alpha}{2})\sqrt{\frac{MS_E}{r}}$

예 3.2 (예 3.1)의 직물마모도 예에서 각 수준에서의 직물마모도 모평균의 95% 신뢰구간을 구하라.

- 신뢰구간의 폭이 $t(12; 0.025)\sqrt{\frac{MS_E}{r}} = 2.179\sqrt{\frac{0.0199}{4}} = 0.1537$ 이므로 각각의 수준에서 직물 마모도의 모평균의 95% 신뢰구간은 다음과 같다.

$$\mu_1: 2.19 \pm 0.1537 = (2.0363,\ 2.3437)$$
$$\mu_2: 2.68 \pm 0.1537 = (2.5263,\ 2.8337)$$
$$\mu_3: 2.42 \pm 0.1537 = (2.2663,\ 2.5737)$$
$$\mu_4: 2.31 \pm 0.1537 = (2.1563,\ 2.4637)$$

3.3 분산분석 후의 추정

최소유의차 검정

- 일원배치에서 두 수준 A_i 와 A_j 에서 모평균의 차이인 $\mu_i - \mu_j$ 의 $100(1-\alpha)\%$ 신뢰구간 :

$$(\overline{x_{i.}} - \overline{x_{j.}}) \pm t(\phi_E; \frac{\alpha}{2}) \sqrt{MS_E \frac{2}{r}}$$

$t(\phi_E; \frac{\alpha}{2}) \sqrt{MS_E \frac{2}{r}}$ 를 최소유의차(least significant differense, LSD)라고 부름

정보통계학과 **백재욱** 교수

제3강 일원배치법

3.4 반복수가 같지 않은 실험

3.4 반복수가 같지 않은 실험

◆ 모형

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$i = 1, \ldots, a \quad j = 1, \ldots, r_i$$

$$SS_A = \sum_i \frac{{T_{i.}}^2}{r_i} - CT$$

<표 3-5> 반복수와 같지 않은 일원배치법의 분산분석표

요 인	제곱합	자유도	평균제곱	F_0	$F(\alpha)$
A	SS_A	$a-1$	MS_A	$\frac{MS_A}{MS_E}$	$F(a-1, N-a : \alpha)$
E	SS_E	$N-a$	MS_E		
T	SS_T	$N-1$			

A_i 수준에서의 모평균 μ_i 의 95% 신뢰구간 : $\bar{x}_{i.} \pm t(\phi_E; 0.025)\sqrt{\frac{MS_E}{r_i}}$

3.4 반복수가 같지 않은 실험

예 3.2 다음 4종류의 식단에 따른 식이요법이 혈액응고시간에 영향을 주는가?

<표 3-6> 혈액응고시간 자료

		실험의 반복					
인자의 수준	A_1	62	60	63	59	61	
	A_2	63	67	71	64	65	66
	A_3	68	66	71	67	68	68
	A_4	56	62	60	61	63	64

		실험의 반복						합계	평균
인자의 수준	A_1	-3	-5	-2	-6	-4		-20	-4
	A_2	-2	2	6	-1	0	1	6	1
	A_3	3	1	6	2	3	3	18	3
	A_4	-9	-3	-5	-4	-2	-1	-24	-4

자료변환

$$y_{ij} = x_{ij} - 65$$

3.4 반복수가 같지 않은 실험

풀이 가설의 설정

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$

(식이요법에 따라 혈액응고 시간에 차이가 없다)

H_1: α_i 가 모두 0인 것은 아니다.

(식이요법에 따라 혈액응고 시간이 모두 같은 것은 아니다)

3.4 반복수가 같지 않은 실험

풀이(계속) 제곱합의 계산

$$CT = \frac{T^2}{N} = \frac{(-20)^2}{23} = 17.3913$$

$$SS_T = \sum_i \sum_j x^2{}_{ij} - CT = (-3)^2 + (-5)^2 + \cdots + (-1)^2 - 17.3913$$

$$= 322.6087$$

$$SS_A = \sum_i \frac{T_{i.}{}^2}{r_i} - CT = \left[\frac{(-20)^2}{5} + \frac{6^2}{6} + \frac{18^2}{6} + \frac{(-24)^2}{6}\right] - 17.3913$$

$$= 218.6087$$

$$SS_E = SS_T - SS_A = 322.6087 - 218.6087 = 104$$

3.4 반복수가 같지 않은 실험

풀이(계속) 자유도의 계산

$$\phi_T = N - 1 = 22$$

$$\phi_A = a - 1 = 3$$

$$\phi_E = \phi_T - \phi_A = 19$$

3.4 반복수가 같지 않은 실험

풀이(계속)

<분산분석표>

요 인	제곱합	자유도	평균제곱	F_0	$F(0.01)$
A	218.6087	3	72.8696	13.31**	5.01
E	104.0000	19	5.4737		
T	322.6087	22			

- 검정통계량 F_0의 값 $13.31 > F(3,19;0.01) = 5.01$

➡ 귀무가설 기각

(식이요법에 따라 혈액응고 시간이 다 똑같은 것은 아니다)

추후분석 필요!

3.4 반복수가 같지 않은 실험

A_i수준에서의 모평균 μ_i 의 95% 신뢰구간 : $\bar{x}_{i.} \pm t(\phi_E; 0.025)\sqrt{\frac{MS_E}{r_i}}$

r_1=5, r_2=r_3=r_4=6, $t(19;\ 0.025) = 2.093$이고 MS_E=5.4737을 대입하면

- μ_1의 95% 신뢰구간 : $61 \pm 2.093 \cdot \sqrt{\frac{5.4737}{5}} \leftrightarrow 61 \pm 2.1899 = (58.811, 63.189)$
- μ_2의 95% 신뢰구간 : $66 \pm 2.093 \cdot \sqrt{\frac{5.4737}{6}} \leftrightarrow 66 \pm 1.9991 = (64.001, 67.999)$
- μ_3의 95% 신뢰구간 : $68 \pm 1.9991 = (66.001, 69.999)$
- μ_4의 95% 신뢰구간 : $61 \pm 1.9991 = (59.001, 62.999)$

➡ μ_1과 μ_2는 겹치지 않음, μ_2과 μ_3는 겹침, ...

정보통계학과 **백재욱** 교수

제3강 일원배치법

3.5 랜덤모형

3.5 랜덤모형

랜덤모형

실험일, 원료로트, 블록, 원자재의 배치$(batch)$등과 같이 요인(인자)의 수준이 고정되지 않고 랜덤하게 뽑아 시험하는 랜덤요인의 구조모형

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{..............................} \quad (3.25)$$

$\alpha_i \sim N(0, \sigma^2{}_A)$이고, 서로 독립

$\epsilon_i \sim N(0, \sigma^2{}_E)$이고, 서로 독립

α_i와 ϵ_{ij} : 서로 독립

3.5 랜덤모형

랜덤모형

*** 가설** $H_0 : \sigma^2{}_A = 0$ **(랜덤요인이 동질적이다)**

$H_1 : \sigma^2{}_A > 0$ **(랜덤요인이 이질적이다)**

$$* \; Var(x_{ij}) = Var(\alpha_i) + Var(\varepsilon_{ij})$$

$$= \sigma^2{}_A + \sigma^2{}_E \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (3.31)$$

3.5 랜덤모형

랜덤모형

기여율 $\varrho = \frac{\sigma^2{}_A}{\sigma^2{}_A+\sigma^2{}_E} \times 100\%$

$$E(MS_A) = \sigma^2{}_E + r\sigma^2{}_A$$
$$E(MS_E) = \sigma^2{}_E$$

$$\Rightarrow \quad \widehat{\sigma^2_E} = MS_E$$
$$\widehat{\sigma^2_A} = \frac{1}{r}(MS_A - MS_E)$$

총분산의 추정치 $\hat{V}ar(x_{ij}) = \hat{\sigma}_A^2 + \hat{\sigma}_E^2$

기여율 ϱ의 추정치 : $\hat{\varrho} = \frac{\hat{\sigma}_A^2}{\hat{V}ar(x_{ij})} \times 100\%$

3.5 랜덤모형

예 3.5 수백 개의 배치(batch)에서 5개의 배치를 랜덤하게 선택하고, 선택된 배치에서 3개의 시료를 채취한 후 정제하여 순도를 측정하였다. 배치(batch) 간 차이가 있는가?

<표 3-8> 배치자료에 의한 화학약품의 순도

batch	1	74	76	75
	2	68	71	72
	3	75	77	77
	4	72	74	73
	5	79	81	79

					$T_{i.}$
batch	1	-1	1	0	0
	2	-7	-4	-3	-14
	3	0	2	2	4
	4	-3	-1	-2	-6
	5	4	6	4	14

자료변환

$$y_{ij} = x_{ij} - 75$$

3.5 랜덤모형

풀이 가설의 설정

$$H_0 : \sigma^2{}_A = 0$$ (배치 간에 산포가 존재하지 않는다)

$$H_1 : \sigma^2{}_A > 0$$ (배치 간에 산포가 존재한다)

3.5 랜덤모형

풀이(계속) 제곱합의 계산

$$CT = \frac{(-2)^2}{15} = 0.27$$

$$SS_T = \sum_i \sum_j y^2{}_{ij} - CT = (-1)^2 + 1^2 + \cdots + 4^2 - 0.27 = 165.73$$

$$SS_A = \sum_i \frac{{T_i}^2}{r} - CT = \frac{1}{3}[0^2 + (-14)^2 + 4^2 + (-6)^2 + 14^2] - 0.27$$

$$= 147.74$$

$$SS_E = SS_T - SS_A = 165.73 - 147.74 = 17.99$$

3.5 랜덤모형

풀이(계속) 자유도의 계산

$$\phi_T = ar - 1 = 15 - 1 = 14$$

$$\phi_A = a - 1 = 5 - 1 = 4$$

$$\phi_E = \phi_T - \phi_A = 10$$

3.5 랜덤모형

풀이(계속)

<분산분석표>

요 인	제곱합	자유도	평균제곱	F_0	$F(0.01)$
A	147.74	4	36.94	20.52**	5.99
E	17.99	10	1.80		
T	165.73	14			

검정통계량 F_0의 값 $20.52 >> F(4,10;0.01) = 5.91$

➡ 귀무가설 기각

(배치간의 산포가 존재하여 배치가 동질적이 아님)

3.5 랜덤모형

풀이(계속) 기여율 추정치 계산

$$\widehat{\sigma_E^2} = MS_E = 1.80$$

$$\widehat{\sigma_A^2} = \frac{1}{3}(MS_A - s_E^2) = \frac{1}{3}(36.94 - 1.80) = 11.71$$

$$\widehat{Var}(x_{ij}) = \widehat{\sigma_A^2} + \widehat{\sigma_E^2} = 11.71 + 1.80 = 13.51$$

$$\hat{\rho} = \frac{\widehat{\sigma_A^2}}{\widehat{Var}(x_{ij})} \times 100\% = \frac{11.71}{13.51} \times 100\% = 86.7\%$$

➡ 데이터의 총 분산 중에서 86.7%는 순수하게 배치간의 산포 탓으로 설명할 수 있음
(13.3%만이 배치 내의 오차의 산포에 기인함)

3.5 랜덤모형

R 실습

b1 <- c(74, 76, 75)

b2 <- c(68, 71, 72)

b3 <- c(75, 77, 77)

b4 <- c(72, 74, 73)

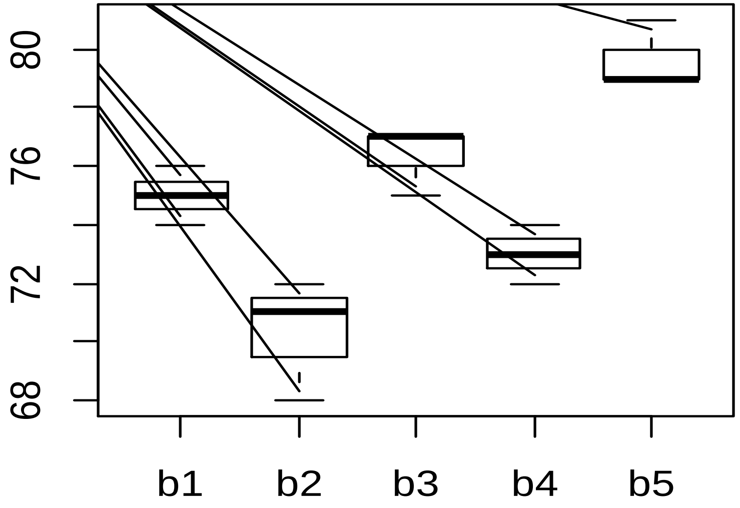
b5 <- c(79, 81, 79)

y <- c(b1, b2, b3, b4, b5)

x <- c("b1", "b2", "b3", "b4", "b5")

x <- rep(x, c(3, 3, 3, 3, 3))

boxplot(y ~ x)



anova <- aov(y ~ x)

summary(anova)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	4	147.7	36.93	20.52	8.25e-05 ***
Residuals	10	18.0	1.80		

---

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1*

다음시간 안내

제4강 (4장)

이원배치법