4장 | 일원배치법 (ONE-WAY ANOVA)

SAS를 이용한 실험 계획과 분산 분석 (자유아카데미)

서론

- 그룹을 결정하는 요인의 개수
  - 일원배치법
  - 이원배치법
  - 삼원 배치법
- 요인의 수준을 배정하는 방법
  - CRD (완전 임의화 설계)
  - RCBD (임의화 완전 블록 설계)
  - Latin Square Design (라틴 방격법)

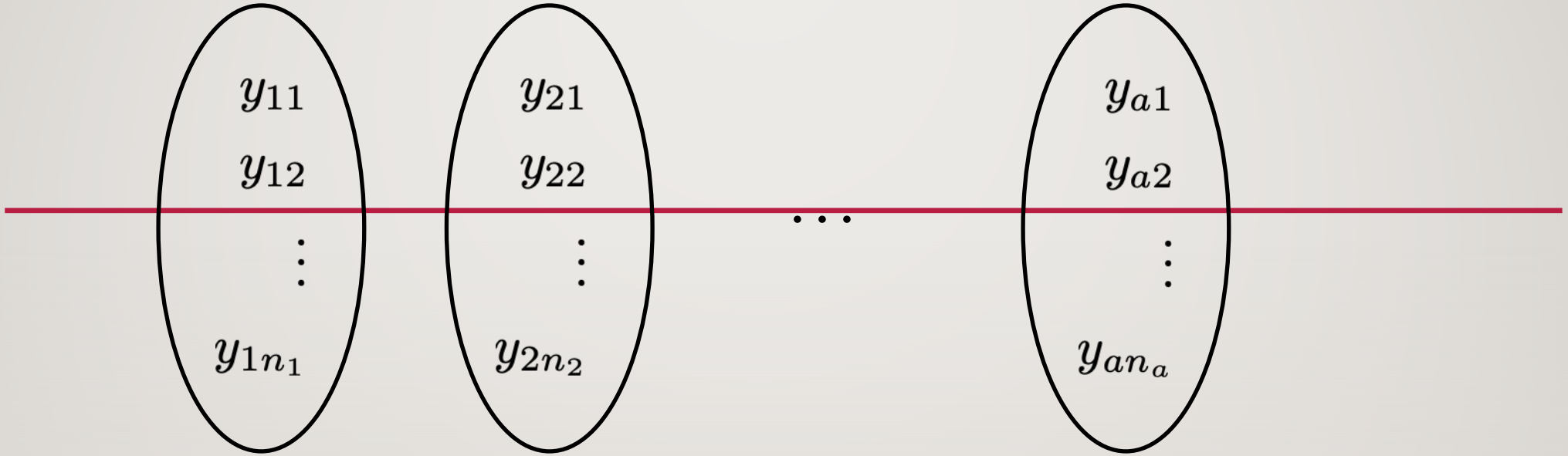
COMPLETELY RANDOMIZED DESIGN (완전 임의화 설계)

전체 $N(=\sum_{i=1}^{a} n_i)$ 개의 실험단위 중에서 임의로 n_1 개를 선택하여 처리1의 실험조건을 실시하고, 나머지 $(N - n_1)$ 개 중에서 임의로 n_2 개를 선택하여 처리2의 실험조건을 행하는 방식으로, 마지막 n_a 개까지 완전 랜덤하게(임의로) a 개의 처리조건을 실시하는 실험계획을 '완전임의화설계(Completely Randomized Design: CRD)'라고 하며 실험계획 중 가장 간단하고 기초가 되는 설계가 된다.

완전 임의화 설계의 예

송아지에게 어느 종류의 사료가 체중증가에 영향을 미치는 지 조사한다고 가정하자.

이를 위해 30마리 송아지를 실험에 참여시킬 것이고 3종류의 사료 (A1,A2, A3)를 비교할 예정이다. 따라서 우리는 10마리 송아지를 임의로 선택하여 A1사료를 먹이고, 나머지 20마리 중 10마리를 임의로 선택해서 A2사료를, 그리고 마지막으로 남은 10마리는 A3사료를 먹이는 실험을 계획한다면, 이 실험 계획은 대표적인 완전 임의화 설계이다.



$y_{1.}$ $y_{2.}$ $y_{a.}$

$\bar{y}_{1.}$ $\bar{y}_{2.}$ $\bar{y}_{a.}$

n_1 n_2 n_a

CRD (일원배치, 이원배치…)

- 일원 배치법

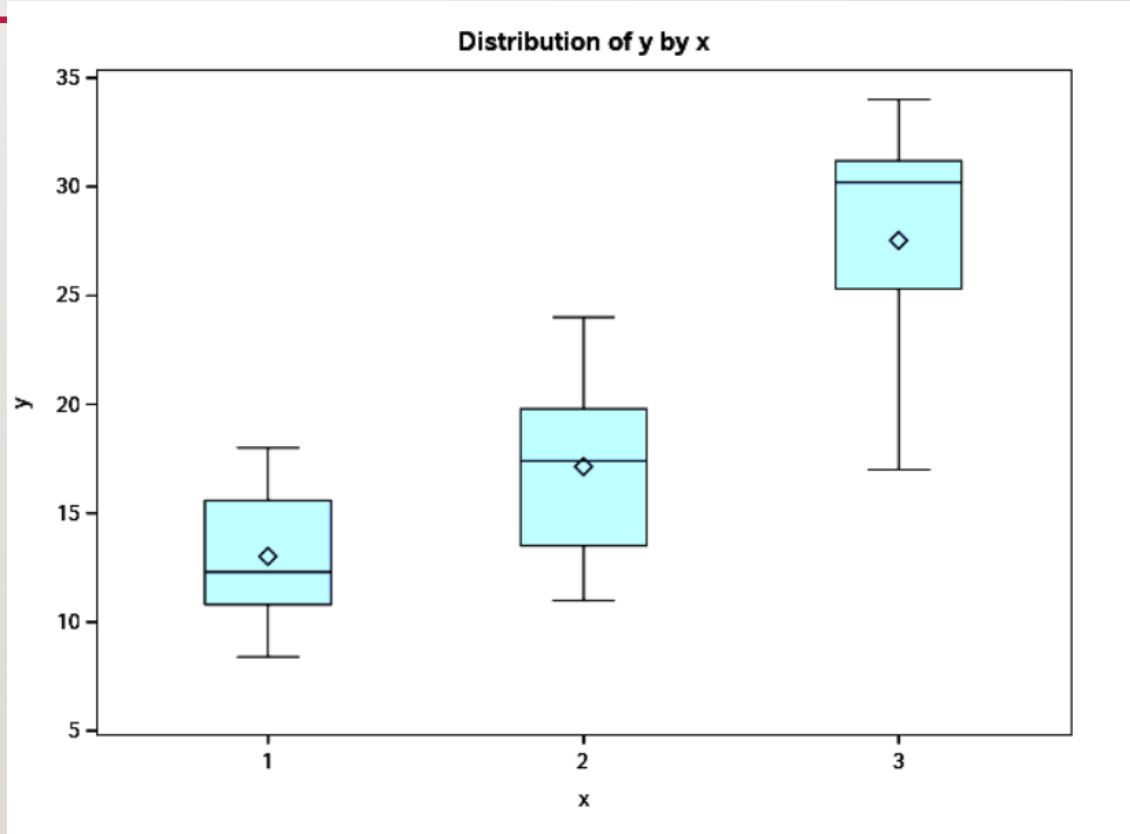
CRD에서 처리1, 처리2, … , 처리 a를 구성하는 실험 조건이 "1개의 요인의 다른 수준"으로 구성 되는 실험 계획법

- 이원배치법

CRD에서 처리1, 처리2, … , 처리 a를 구성하는 실험 조건이 "2개의 요인의 다른 수준"으로 구성 되는 실험 계획법

수준 (LEVEL) 비교

- 요인 X = 1, 2, 3 (수준)



CRD의 분산 분석

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

$$H_1 : \text{적어도 한 개의 } (i, j)\text{에 대해 } \mu_i \neq \mu_j \,, \ (1 \leq i \neq j \leq a)$$

Source	d.f.	S.S.	M.S.	F_0
Treatment	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$	MStreat	$\frac{\text{MStreat}}{\text{MSE}}$
Error	$\sum_{i=1}^{a}(n_i-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$	MSE	
Total	$\sum_{i=1}^{a}n_i-1$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2$		

표 4.1: CRD 의 분산분석표(ANOVA table)

분산 분석 제곱합의 간편한 공식

$$
\begin{aligned}
SST &= \sum_{i}^{a}\sum_{j}^{n_i}(y_{ij}-\bar{y}_{..})^2 \\
&= \sum_{i}^{a} n_i(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i}^{a}\sum_{j}^{n_i}(y_{ij}-\bar{y}_{i.})^2 \\
&= SStrt + SSE
\end{aligned}
$$

CT는 수정항

$$
SST = \sum_{i_a}^{a}\sum_{j}^{n_i} y_{ij}^2 - CT, \quad CT = \frac{y_{..}^2}{N}
$$

$$
SStrt = \sum_{i}^{} \frac{y_{i.}^2}{n_i} - CT, \quad CT = \frac{y_{..}^2}{N}
$$

$$
SSE = SST - SStrt
$$

CRD의 모형식 (I)

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

$$y_{ij} = \mu_i + \epsilon_{ij},$$

$$i = 1, 2, \cdots, a,$$

$$j = 1, 2, \cdots, n_i$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

CRD의 모형식 (2)

$$y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

$$i = 1, 2, \cdots, a,$$

$$j = 1, 2, \cdots, n_i$$

$$\sum_{i=1}^{a} \tau_i = 0$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$$

$$H_1 : \exists\, \tau_i \quad \text{s.t.} \quad \tau_i \neq 0 \quad (1 \leq i \leq a)$$

이 모형식에서 $\sum_{i=1}^{a} \tau_i = 0$ 이라는 가정은 아래 식에서 확인 가능하다.

$$\mu = \frac{1}{a}\sum_{i=1}^{a} \mu_i = \frac{1}{a}\sum_{i=1}^{a} (\mu + \tau_i)$$

코크란의 정리

아래의 조건 1), 2), 3)을 모두 만족한다면,

1) $\mathrm{SST} = \mathrm{SSA}_1 + \mathrm{SSA}_2 + \cdots + \mathrm{SSA}_t + \mathrm{SSE}$

2) $\mathrm{df}(\mathrm{SST}) = \mathrm{df}(\mathrm{SSA}_1) + \mathrm{df}(\mathrm{SSA}_2) + \cdots + \mathrm{df}(\mathrm{SSA}_t) + df\,(\mathrm{SSE})$

3) $\mathrm{E}(\mathrm{MSA}_i) = \sigma^2 + \lambda_i, \quad \mathrm{E}(\mathrm{MSE}) = \sigma^2$

$H_0 : \lambda_i = 0$을 가정할 때

$$F_0 = \frac{\mathrm{MSA}_i}{\mathrm{MSE}} \sim F_{\mathrm{df}(\mathrm{SSA}_i),\, \mathrm{df}(\mathrm{SSE})}$$

가 되며, 만일 $\lambda_i \neq 0$이면 F_0는 아래 같은 분포를 따른다.

$$F_0 = \frac{\mathrm{MSA}_i}{\mathrm{MSE}} \sim F(\lambda_i)_{\mathrm{df}(\mathrm{SSA}_i),\, \mathrm{df}(\mathrm{SSE})}$$

여기서 $F(\lambda_i)$는 비중심모수가 λ_i인 '비중심 F 분포'이다.

SUMMARY

CRD 의 모형식 $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $(i = 1, 2, \cdots, a)$ 의 가설

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$$

$$H_1 : \exists\, \tau_i \quad \text{s.t.} \quad \tau_i \neq 0\,, \quad (1 \leq i \leq a)$$

에 대해, 만일

$$F_0 = \frac{\text{MStreat}}{\text{MSE}} > F_{\alpha,\, \text{df(SStreat)},\, \text{df(SSE)}}$$

이면, 귀무가설 $H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$ 을 기각한다.

분산 분석표 작성시 유의할 점

1. 자유도(degrees of freedom)는 음수(-)가 될 수 없다.
2. 제곱합(sum of squares)은 음수(-)가 될 수 없다.
3. $\frac{y_{i\cdot}^2}{n}$과 $\left(\frac{y_{i\cdot}}{n}\right)^2$은 다르다.
4. 경우에 따라서는 SST, SStreat를 계산 한 후, SSE = SST − SStreat 를 이용하면 계산이 쉬워진다.

CRD 의 예

교육 프로그램1	교육 프로그램2	교육 프로그램3	교육 프로그램4
74	94	62	80
67	82	75	82
83	69	59	75
77	78	79	90
71	68	68	72

표 4.2: 영업교육 프로그램에 따른 잡지 판매실적

Example 어느 잡지회사에서 잡지의 판촉 활성화를 위하여 영업사원을 대상으로 하는 4종류의 영업교육 프로그램을 개발하였다. 과연 새로 개발된 4개 교육 프로그램에 따라 잡지 판매실적이 다른지 알아보려고 한다. 20명의 신입사원을 4가지 영업교육 프로그램에 5명씩 임의로 배정하여 수강시킨 후에 일정기간 동안에 걸친 각 영업사원의 판매실적을 조사하였다. 각 영업교육 프로그램 간 판매실적 차이가 존재하는지를 알아보고자 한다. 표 4.2는 사원들이 받은 교육 프로그램과 판매실적이다.

CRD 예

$$y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

$$i = 1, 2, \cdots, 4,$$

$$j = 1, 2, \cdots, 5$$

$$\sum_{i=1}^{4} \tau_i = 0$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

$$\text{CT} = \frac{y_{..}^2}{a\,n} = \frac{(1505)^2}{20} = 113251.2$$

$$\text{SST} = \sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^2 - \text{CT}$$

$$= 74^2 + 67^2 + \cdots + 72^2 - 113251.2 = 1149.75$$

$$\text{SStreat} = \sum_{i=1}^{4} \frac{y_{i\cdot}^2}{n} - \text{CT} \quad = \frac{1}{5}\left(372^2 + 391^2 + 343^2 + 399^2\right) - 113251.2 = 371.75$$

$$\text{SSE} = \text{SST} - \text{SStreat} = 1149.75 - 371.75 = 1078$$

CRD 예

Source	d.f.	S.S.	M.S.	F_0
영업교육 프로그램	3	371.75	123.916	1.839
오차	16	1078	67.375	
전체	19	1449.75		

표 4.3: 판매실적에 대한 영업교육 프로그램의 영향

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$

$$H_1 : \exists\, \tau_i \quad \text{s.t.} \quad \tau_i \neq 0\,, \quad (1 \leq i \leq 4)$$

$$F_0 = 1.839 < 3.24 = F_{0.05,\,3,16}$$

SAS CODE

일원배치법의 경우에 그룹을 규정짓는 요인이 하나 뿐이므로 Type I SS 와 Type III SS 의 값은 항상 같지만 일반적으로 실험계획에서는 Type III SS 의 값을 분산분석표에 적고 있다(Type I SS 와 Type III SS 는 회귀분석을 참조바란다).

```
data a;
input program sales @@;
cards;
1 74 1 67 1 83 1 77 1 71
:
4 80 4 82 4 75 4 90 4 72
;
proc glm data=a;
    class program;
    model sales = program;
run;
```

The GLM Procedure

Dependent Variable: sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	371.750000	123.916667	1.84	0.1807
Error	16	1078.000000	67.375000		
Corrected Total	19	1449.750000			

R-Square	Coeff Var	Root MSE	sales Mean
0.256424	10.90794	8.208228	75.25000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
program	3	371.7500000	123.9166667	1.84	0.1807

Source	DF	Type III SS	Mean Square	F Value	Pr > F
program	3	371.7500000	123.9166667	1.84	0.1807

SAS CODE

Source	d.f.	S.S.	M.S.	F_0
Program	3	371.750	123.916	1.84
Error	16	1078.000	67.375	
Total	19	1449.750		

CRD의 모수 추정

예를 들어 그룹의 수가 2이며($a = 2$) 각 그룹당 반복수를 3으로($n = 3$) 가정한 일원배치법의 모수를 추정해 보자. 일원배치법의 모형식은 식 (4.41)같이 $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ 으로 표현된다.

$$\boxed{y_{ij} = \mu + \tau_i + \epsilon_{ij}}$$

$$\Longrightarrow \underbrace{\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}}_{\boldsymbol{\epsilon}}$$

$$\Longrightarrow \quad \boxed{\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}} \qquad (4.41)$$

CRD의 모수 추정

최소제곱추정법(least squares estimation)에 의한 정규방정식 $X^TX\hat{\boldsymbol{\beta}} = X^T\boldsymbol{y}$은 다음과 같다.

$$\begin{pmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\tau_1} \\ \hat{\tau_2} \end{pmatrix} = \begin{pmatrix} y_{\cdot\cdot} \\ y_{1\cdot} \\ y_{2\cdot} \end{pmatrix} \tag{4.42}$$

여기서 X^TX는 비정칙행렬(singular matrix)이어서 역행렬이 존재하지 않으므로, 방정식의 해인 $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$는 구할 수 없게 된다.

CRD의 모수 추정

그러나 이 정규방정식을 풀어 쓰면,

$$\begin{cases} 6\,\hat{\mu} + 3\,\hat{\tau}_1 + 3\,\hat{\tau}_2 = y_{\cdot\cdot} \\ 3\,\hat{\mu} + 3\,\hat{\tau}_1 = y_{1\cdot} \\ 3\,\hat{\mu} + 3\,\hat{\tau}_2 = y_{2\cdot} \end{cases} \tag{4.43}$$

가 되어서 변수는 3 개이면서 선형독립인 식이 2 개가 존재함을 알 수 있다. 따라서 식 (4.43)의 해를 구하기 위해선 제약조건(constraint)이 하나 더 필요한데, 다음과 같은 제약조건 중 하나를 추천한다[9].

$$\hat{\mu} = 0, \qquad \hat{\tau}_1 + \hat{\tau}_2 = 0, \qquad \hat{\tau}_2 = 0 \tag{4.44}$$

CRD의 모수 추정

$\hat{\mu} = 0,$

$$\hat{\mu} = 0$$
$$\hat{\tau}_1 = \bar{y}_{1\cdot}$$
$$\hat{\tau}_2 = \bar{y}_{2\cdot}$$

$\hat{\tau}_1 + \hat{\tau}_2 = 0,$

$$\hat{\mu} = \bar{y}_{\cdot\cdot}$$
$$\hat{\tau}_1 = \bar{y}_{1\cdot} - \bar{y}_{\cdot\cdot}$$
$$\hat{\tau}_2 = \bar{y}_{2\cdot} - \bar{y}_{\cdot\cdot}$$

$\hat{\tau}_2 = 0$

$$\hat{\mu} = \bar{y}_{2\cdot}$$
$$\hat{\tau}_1 = \bar{y}_{1\cdot} - \bar{y}_{2\cdot}$$
$$\hat{\tau}_2 = 0$$

SAS CODE

```
proc glm data=a;
    class program;
    model sales = program / solution;
run;
```

$$\hat{\mu}_1 = \hat{\mu} + \hat{\tau}_1 = 79.8 - 5.4 = 74.4$$

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	79.80000000	B	3.67083097	21.74	<.0001
program 1	-5.40000000	B	5.19133894	-1.04	0.3137
program 2	-1.60000000	B	5.19133894	-0.31	0.7619
program 3	-11.20000000	B	5.19133894	-2.16	0.0465
program 4	0.00000000	B	.	.	.