실험계획법 2024 강의 노트

서울시립대학교 통계학과 이용희

2024-06-05

목차

서론

이 온라인 교과서는 2024년 실험계획법 강의의 보조 교재입니다.

강의교재는 임용빈 (2020) 를 참고하시기 바랍니다.

필요한 R 라이브러리

```
library(here)           # file pathways
library(tidyverse)      # data management, summary, and visualization
library(MASS)
library(knitr)
library(kableExtra)

library(agricolae)
library(emmeans)

# 변량모형(혼합모형)
library(lme4)
library(lmerTest)

library(SixSigma)
library(FrF2)
library(unrepx)

library(rsm)
library(DoE.wrapper)
library(scatterplot3d)


# ggplot 그래프에서 한글 사용
library(showtext)
font_add_google("Nanum Pen Script", "gl")
showtext_auto()

# 참고도서 데이터
library(MontgomeryDAE)
```

1. 일원배치법

1.1. 두 집단의 평균 비교

1.1.1. t-검정

기초통계학에서 나오는 가장 기본적이고 자주 쓰이는 가설검정 방법은 두 집단의 평균의 차이를 검정하는 t-검정(t-test)이다.

교과서 2장 예제 2.2 를 다시 보자. 공장의 두 개 라인에서 생산되는 시멘트의 인장강도에 유의한 차이가 있는지 통계적 가설검정을 수행하려고 한다. 첫 번째 생산라인을 1, 두 번째 생산라인을 2 라고 했을 때 각각의 라인에서 시멘트 인장강도의 평균을 μ_1, μ_2이라고 하자.

여기서 고려해야할 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2$$

두 집단이 분산이 동일한 정규분포 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$를 따른다고 가정하고 다음과 같이 각각 n_1, n_2개의 독립 표본을 얻었다고 하자.

$$y_{11}, y_{12}, \cdots, y_{1n_1} \sim N(\mu_1, \sigma^2), \quad y_{21}, y_{22}, \cdots, y_{2n_2} \sim N(\mu_2, \sigma^2)$$

위의 가설을 다음과 같은 t-통계량을 이용하여 검정할 수 있다.

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{1/n_1 + 1/n_2}}$$

여기서 $\bar{y}_1$, $\bar{y}_2$은 두 생산라인에서 추출된 표본의 평균을 나타내고 n_1, n_2 은 각 집단의 표본 개수를 나타낸다. 또한 s_p^2 은 두 집단의 공통분산 추정량이며 다음과 같이 계산한다.

$$\hat{\sigma}^2 = s_p^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

위에서 구한 t-통계량의 절대값이 크다면 귀무가설 H_0 에 반대되는 증거이다. 유의수준을 α라고 했을 때 t-통계량 t_0의 절대값이 자유도 $df = n_1 + n_2 - 2$ 를 가지는 t-분포의 상위 $\alpha/2$ 분위수보다 크면 귀무가설을 기각하고 대립가설 H_1을 채택한다.

$$\text{Reject } H_0 \quad \text{if} \quad |t_0| > t(1 - \alpha/2, n_1 + n_2 - 2)$$

가설 검정은 p-값(p-value)을 구하고 그 값이 α보다 작으면 귀무가설을 기각하는 방법을 사용할 수 있다.

$$\text{Reject } H_0 \quad \text{if p-value } < \alpha$$

여기서 $p - value$ 는 다음과 같이 계산할 수 있으며 $t(n_1 + n_1 - 2)$는 자유도가 $n_1 + n_2 - 2$을 가지는 t-분포를 따르는 확률변수이다. .

$$\text{p-value } = 2P[t(n_1 + n_2 - 2) > |t_0|]$$

R 에서 함수 t.test를 이용하여 두 집단에 대한 t-검정을 실시해 보자.

```
line1 <- c(16.9, 16.4, 17.2, 16.4, 16.5, 17.0, 17.0, 17.2, 16.6, 16.6)
line2 <- c(16.6, 16.8, 17.4, 17.1, 17.0, 16.9, 17.3, 17.0, 17.1, 17.3)
df220 <- data.frame(line1, line2)

df220
```

```
   line1 line2
1   16.9  16.6
2   16.4  16.8
3   17.2  17.4
4   16.4  17.1
5   16.5  17.0
6   17.0  16.9
7   17.0  17.3
8   17.2  17.0
9   16.6  17.1
10  16.6  17.3
```

```
df22<- df220 %>% pivot_longer(cols = everything(), names_to = "line", values_to = "strength") %>% dplyr

df22
```

```
# A tibble: 20 x 2
   line  strength
   <chr>    <dbl>
 1 line1     16.9
 2 line1     16.4
 3 line1     17.2
 4 line1     16.4
 5 line1     16.5
 6 line1     17
 7 line1     17
 8 line1     17.2
 9 line1     16.6
10 line1     16.6
11 line2     16.6
12 line2     16.8
13 line2     17.4
14 line2     17.1
15 line2     17
16 line2     16.9
17 line2     17.3
18 line2     17
19 line2     17.1
20 line2     17.3
```

```
t.test(strength~line, df22, paired = FALSE, var.equal = TRUE, conf.level = 0.95)
```

```
    Two Sample t-test

data:  strength by line
t = -2.1338, df = 18, p-value = 0.04687
alternative hypothesis: true difference in means between group line1 and group line2 is not equal to 0
95 percent confidence interval:
 -0.535840211 -0.004159789
sample estimates:
mean in group line1 mean in group line2
              16.78               17.05
```

유의수준을 0.05로 정하면 t-검정의 결과 p-값이 유의수준 보다 작아서 귀무가설을 기각하고 대립가설 H_1을 채택한다. 즉, 두 라인의 시멘트 인장강도 평균은 유의하게 다르다.

1.1.2. t-검정의 재구성

이제 두 집단에 대한 가설 검정을 세 개 이상인 여러 개의 집단으로 확장하는 경우를 생각해보자. 여러 개의 집단에 대한 가설 검정을 고려하기 전에 두 집단에 대한 t-검정을 약간 재구성하여 여러 평균들의 차이를 비교하는 검정법에 대한 일반적인 개념을 제시해 보려고 한다. 이제 t-검정에서 검정 통계량의 분자와 분모를 따로 살펴보자

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{1/n_1 + 1/n_2}}$$

t-검정 통계량의 분자는 집단 간의 평균의 차이를 나타낸다. 즉 $\bar{y}_1 - \bar{y}_2$는 두 집단의 표본 평균의 차이를 추정하는 양이고 그 차이가 크면 클수록 두 집단의 모평균의 차이 $\mu_1 - \mu_2$가 크다는 것을 의미한다.

t-검정 통계량의 분모는 두 집단의 공통분산 추정량 $\hat{\sigma}^2 = s_p^2$에 비례한다. 즉 집단 내의 변동을 반영하는 s_p^2 이 크면 클수록 t-검정 통계량은 그 크기가 작아져서 귀무가설의 기각을 어렵게 한다.

$$s_p^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

또한 t-검정 통계량는 표본의 수(n_1과 n_2)에 비례한다. 즉 표본의 수가 증가햐면 t-검정 통계량이 커지게 된다.

정리해보면 t-검정 통계량은 **집단 간의 변동**(between-group variation)을 **집단 내의 변동**(within-group variation) 으로 나누어준 값이다. 다른 말로 급간 변동과 급내 변동을 사용하기도 한다.

이제 t-검정 통계량을 **제곱하면** 다음과 같이 표현할 수 있다.

$$t_0^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_p^2(1/n_1 + 1/n_2)} = \frac{\text{between-group variation}}{\text{within-group variation}}$$

두 집단의 평균을 비교하는 t-검정 통계량은 집단 간의 변동(집단 간의 평균들의 차이)과 집단 내의 변동(집단 내 관측치들의 퍼진 정도)의 비율로 구성된 통계량으로 생각할 수 있으며 이러한 개념을 3개 이상의 집단을 비교하는 경우로 쉽게 확장할 수 있다.

1.2. 일원배치법

1.2.1. 일원배치을 이용한 랜덤화 실험계획법

- 일원배치법(one-way randomization design)은 관심있는 중요한 한 개 요인이 반응변수에 어떠한 영향을 미치는지 알아보는 실험법이다.
- 반응값에 영향을 주는 다른 요인들에 대한 정보가 많고 사전 실험이 많이 이루어져서 가장 중요한 요인의 미세한 영향을 조사하고자 할 때 유용하다.
- 처리를 제외한 다른 요인들의 영향을 적절하게 통제할 수 있어야 한다.

1.2.2. 실험배정의 랜덤화

실험배정의 랜덤화 방법은 교과서 38-41 페이지 참조

■ 요인 수준 별로 실험 실시 순서가 랜덤한 메카니즘에 의해 결정 (4수준 5반복)

실험의 반복	요인 수준			
	A_1	A_2	A_3	A_4
1	1	6	11	16
2	2	7 ⑤	12	17
3	3 ①	8	13	18
4	4 ②	9	14 ④	19
5	5	10	15	20 ③

단계 1 : 각 실험 조건에 일련번호를 할당 (std order)

단계 2 : 1에서 20까지의 20개 숫자의 랜덤한 배열 구하기 (run order)
3, 4, 20, 14, 7, 16, …

단계 3 : 나온 순서대로 실험 실시 (원형 숫자의 순서대로)
실험 순서: $A_1, A_1, A_4, A_3, A_2, A_4, \cdots$

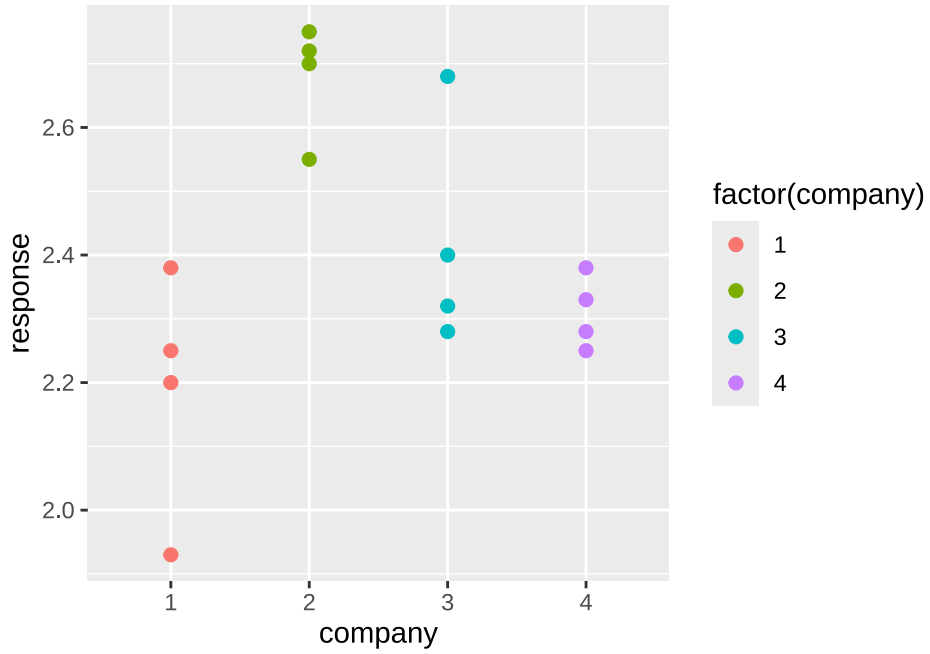
1.2.3. 예제 3.1 - 자료

이 실험에서 요인은 직물이며 4개 수준은 4개의 납품업체에서 공급한 서로 다른 직물이다. 실험 목적은 4개의 직물의 긁힘에 대한 저항력을 비교하는 실험이다. 각 업체마다 4개의 제품을 랜덤하게 선택하여 일원배치법으로 마모도 검사를 실시하였다.

```
company<- as.factor(rep(c(1:4), each=4))
response<- c(1.93, 2.38, 2.20, 2.25,
             2.55, 2.72, 2.75, 2.70,
             2.40, 2.68, 2.32, 2.28,
             2.33, 2.38, 2.28, 2.25)
df31 <- data.frame(company=company, response= response)

df31
```

```
   company response
1        1     1.93
2        1     2.38
3        1     2.20
4        1     2.25
5        2     2.55
6        2     2.72
7        2     2.75
8        2     2.70
9        3     2.40
10       3     2.68
11       3     2.32
12       3     2.28
13       4     2.33
14       4     2.38
15       4     2.28
16       4     2.25
```

```
ggplot(df31, aes(company, response)) +   geom_point(aes(colour = factor(company)), size = 2)
```



실험에서는 처리 이외의 다른 요인들이 적절하게 통제되는 것이 매우 중요하다. 4개의 처리 외에 마모도 검사의 결과에 영향을 미칠 수 있는 다른 요인을 생각해 보자.

- 검사를 수행하는 사람
- 마모도를 검사하는 도구 또는 기계
- 검사를 실시하는 환경 (측정 시간, 장소 등)
- 마모도 검사의 배정을 완전 임의(completely randomized)로 할 수 있는지?

일원배치법으로 실험을 진행할 때 다음과 같은 상항들을 고려해야 한다.

- 처리 이외의 다른 요인들을 적절하게 통제할 수 있는가?
- 어떤 경우에 완전한 랜덤화가 불가능한가? 이러한 경우 실험의 배정을 어떻게 해야 할까?

1.2.4. 일원배치법의 자료 구조와 모형

- 일원배치법에서의 자료 구조는 교과서 41-44 페이지 참조

일원배치법 실험에서는 하나의 요인 A 의 효과를 측정한다. 요인 A 에 대하여 서로 다른 a 개의 수준$(A_1, A_2, \dots, A_a)$의 효과를 비교한다고 가정하자. 각 수준에 대하여 r_i 개의 반응값을 반복 측정한다.

이제 i번 수준에서 측정된 j번째 반응변수의 값을 x_{ij}라고 하자. 일원배치법에서 측정된 자료들은 다음과 같은 모형을 가진다고 가정한다.

$$x_{ij} = \mu_i + e_{ij} \text{ where } e_{ij} \sim N(0, \sigma_E^2) \tag{1.1}$$

여기서 오차항 e_{ij}는 모두 독립이다.

첨자 i는 실험의 수준에 나타낸다 ($i = 1, 2, \cdots, a$). 균형자료의 경우는 모든 수준에 대하여 반복수가 같은 경우이다 ($j = 1, 2, \cdots, r$). 불균형자료의 경우는 수준에 대하여 반복수가 다른 경우이다 ($j = 1, 2, \cdots, r_i$).

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

식 1.1 은 일반적으로 **평균모형(mean model)** 이라고 부르며 모형의 이름대로 모두 μ_i는 i번째 수준의 평균을 의미한다.

$$E(x_{ij}) = E(\mu_i + e_{ij}) = \mu_i$$

이제 식 1.1 을 약각 변형하여 다른 형식의 모형을 만들어 보자.

$$\begin{aligned} x_{ij} &= \mu_i + e_{ij} \\ &= \mu + (\mu_i - \mu) + e_{ij} \\ &= \mu + \alpha_i + e_{ij} \end{aligned}$$

위의 모형에서 모수 μ는 반응값의 전체 평균을 의미하며 $\alpha_i = \mu_i - \mu$는 i번째 수준의 평균이 전체 평균과 어떻게 다른지 나타내는 수준의 상대적 효과를 의미한다.

다음의 식으로 정의된 일원배치 모형을 **주효과모형(main effect model)** 이라고 부른다. 모수 α_i는 i 번째 집단의 효과(처리 효과; treatment effect)를 나타낸다고 할 수 있다.

$$x_{ij} = \mu + \alpha_i + e_{ij} \text{ where } e_{ij} \sim N(0, \sigma_E^2) \tag{1.2}$$

여기서 오차항 e_{ij}는 모두 독립이며 다음과 같은 제약조건이 있다.

$$\sum_{i=1}^{a} \alpha_i = 0 \tag{1.3}$$

식 1.3 의 제약조건은 모수의 개수$(a+1)$가 그룹의 개수(a)보다 많아서 발생하는 문제를 해결하기 위하여 모수에 대한 제약조건 1개를 고려해서 모수의 개수와 그룹의 개수를 맞추어준 것이다. 나중에 이러한 제약조건에 대한 이론을 자세히 다루기로 한다.

식 1.3 의 제약조건은 **sum to zero**조건이라고 부르며 문제를 해결하는 유일한 조건은 아니다. 예를 들어서 조건 식 1.3 의 제약조건을 대신하여 $\alpha_1 = 0$ 인 **set to zero** 조건을 사용할 수 있다.

1.3. 분산분석

1.3.1. 모형과 가설

집단의 모평균을 편의상 $\mu_1, \mu_2, \cdots, \mu_a$ 이라고 하자. 평균모형 식 1.1 을 가정하고 집단들 사이에 차이가 있는지에 대한 가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \quad \text{vs.} \quad H_1 : \text{ not } H_0$$

위의 가설에서 주의할 점은 대립가설 H_1의 경우애 평균들이 서로 다른 경우가 매우 다양하다는 것이다. 예를 들어 집단이 3개 인 경우 $\mu_1 = \mu_2 < \mu_3$일 수 도 있고 $\mu_1 < \mu_2 < \mu_3$ 있으며 이 외에 매우 다양한 경우들이 있다.

이제 효과모형 식 1.2 을 고려하면 집단들 사이에 차이가 있는지에 대한 가설을 다음과 같이 바꿀수 있다. 집단에 대한 효과가 모두 0이 되면 집단 간의 평균에 대한 차이는 없다.

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \quad \text{vs.} \quad H_1 : \text{ not } H_0 \tag{1.4}$$

1.3.2. 변동의 분해

이제 앞 절에서 생각해본 t-검정의 재구성처럼 집단 간의 변동(각 집단의 평균의 차이가 얼마나 나는지에 대한 통계량)과 집단 내의 변동(각 집단내에서 관측값들의 퍼진 정도)를 측정하는 통계량을 찾아서 검정 통계량을 구성해 보자.

일단 다음과 같이 전체 평균과 집단의 평균을 정의하자.

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{r} x_{ij}}{ar} = \frac{T}{ar}, \quad \bar{x}_{i.} = \frac{\sum_{j=1}^{r} x_{ij}}{r} = \frac{T_{i.}}{r}$$

이제 하나의 관측값 x_{ij}과 전체 평균 $\bar{\bar{x}}$ 간의 편차(deviation)를 다음과 같이 분해해 보자.

$$\underbrace{x_{ij} - \bar{\bar{x}}}_{\text{total deviation}} = \underbrace{(x_{ij} - \bar{x}_{i.})}_{\text{within-group deviation}} + \underbrace{(\bar{x}_{i.} - \bar{\bar{x}})}_{\text{between-group deviation}} \tag{1.5}$$

식 1.5 에서 집단 평균과 총 평균의 편차 $(\bar{x}_{i.} - \bar{\bar{x}})$는 처리의 효과를 측정할 수 있는 통계량이다. 집단 간의 차이를 반영하는 양으로 처리 효과 α_i들에 의하여 발생한다.

집단 내의 관측값과 집단 평균의 차이 $(x_{ij} - \bar{x}_{i.})$는 집단 내의 변동을 나타내는 통계량으로 측정 오차 e_{ij}에 의하여 발생한다.

식 1.5 의 각 편차들은 양수와 음수로서 부호를 가지기 때문에 이를 변동으로 표현하기 위하여 차이를 제곱하여 합친 제곱합(sum of squares)을 고려해 보자.

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{\bar{x}})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{r}\left[(x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{\bar{x}})\right]^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i.})^2 + \sum_{i=1}^{a}\sum_{j=1}^{r}(\bar{x}_{i.} - \bar{\bar{x}})^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{\bar{x}}) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i.})^2 + \sum_{i=1}^{a} r(\bar{x}_{i.} - \bar{\bar{x}})^2 + 0(why?)
\end{aligned}$$

결과적으로 다음과 같은 변동의 분해를 제곱합의 형식으로 얻을 수 있다.

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{\bar{x}})^2}_{\text{total variation}} = \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i.})^2}_{\text{within-group variation}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{r}(\bar{x}_{i.} - \bar{\bar{x}})^2}_{\text{between-group variation}} \tag{1.6}$$

분해 식 1.6 에서 나타난 각 제곱합에 대한 이름과 의미를 살펴보자.

- SS_T를 총 제곱합(Total Sum of Squares)이라고 부르며 자료의 전체 변동을 의미한다.

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{\bar{x}})^2$$

- SS_E를 잔차 제곱합(Residual Sum of Squares)이라고 부르며 관측 오차에 발생된 집단 내의 변동 또는 급내 변동(within-group variation)을 의미한다.

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i.})^2$$

- SS_A를 처리 제곱합(Treatment Sum of Squares)이라고 부르며 처리들의 차이로 발생하는 변동으로거 집단 간의 변동 또는 급간 변동(bwtween-group variation)을 의미한다.

$$SS_A = \sum_{i=1}^{a}\sum_{j=1}^{r}(\bar{x}_{i.} - \bar{\bar{x}})^2 = \sum_{i=1}^{a} r(\bar{x}_{i.} - \bar{\bar{x}})^2$$

이제 분해 식 1.6 을 다음과 같이 나타낼수 있다.

$$SS_T = SS_A + SS_E \tag{1.7}$$

위의 분해식에서 볼 수 있듯이 집단 간의 변동의 크기를 나타내는 처리제곱합이 커질수록, 또는 집단내의 변동의 크기를 나타내는 오차제곱합이 작아질수록 귀무가설에 반대되는(즉, 집단 간의 평균이 유의한 차이가 난다는) 증거가 강해진다.

1.3.3. 자유도

제곱합은 편차(deviation)의 제곱들을 더한 형태로서 각 제곱합들에 대하여 해당하는 자유도(degrees of freedom; df 또는 ϕ로 표기)를 구할 수 있다.

제곱합의 자유도 = 제곱합을 구성하는 편차의 개수 - 선형제약 조건의 개수

각 제곱합에 대한 선형제약조건은 편차들의 합이 0이 되는 조건이다. 이제 식 식 1.7 에 주어진 제곱합의 자유도에 대한 정보를 다음과 같이 정리할 수 있다.

제곱합	편차의 개수	제약조건	제약조건의 수	자유도
SS_T	ar	$\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{\bar{x}}) = 0$	1	$\phi_T = ar - 1$
SS_A	a	$\sum_{i=1}^{a}(\bar{x}_{i.} - \bar{\bar{x}}) = 0$	1	$\phi_A = a - 1$
SS_E	ar	$\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i.}) = 0, i = 1, 2, .., a$	a	$\phi_E = ar - a$

1.3.4. 평균제곱합과 F-통계량

이제 가설 식 4.6 을 검정하기 위한 통계량을 구성해 보자. 먼저 다음과 같은 제곱합들을 각 자유도로 나눈 평균제곱합(Mean Sum of Squares)를 정의한다.

$$MS_A = \frac{SS_A}{\phi_A}, \quad MS_E = \frac{SS_E}{\phi_E} \tag{1.8}$$

앞 절에서 t-검정을 재구성하면서 알아본 통계량의 의미를 다시 생각해 보자. 집단 간의 변동과 집단 내의 변동의 상대적 비율로 그룹 간의 차이를 검정할 수 있다는 개념을 확장하여 다음과 같은 F-통계량 F_0 를 만들어 보자.

$$F_0 = \frac{MS_A}{MS_E} = \frac{\text{between-group variation}}{\text{within-group variation}} \tag{1.9}$$

위 식 1.9 에서 정의된 F-통계량은 그룹 간에 평균의 차이가 클수록, 그룹 내의 차이가 작을 수록 그 값이 커진다. 따라서 F-통계량의 값이 크면 클수록 귀무가설에 반대되는 증거가 강해진다.

이렇게 전체의 변동을 집단 간의 변동과 집단 내의 변동으로 나누어 집단 간의 평균의 차이를 추론하는 방법을 **분산분석** (Analysis of Variance, **ANOVA**)이라고 한다.

1.3.5. 분산분석을 이용한 F-검정

이제 식 1.9 에서 정의된 F-통계량을 이용하여 가설 식 4.6 를 검정하는 통계적 방법을 만들어 보자. 일단 두 제곱합의 통계적 성질은 다음과 같다.

- 잔차 제곱합을 오차항의 분산으로 나눈 통계량은 자유도가 ϕ_E 를 가지는 카이제곱 분포를 따른다.

$$\frac{SS_E}{\sigma_E^2} \sim \chi^2(\phi_E)$$

- **귀무가설이 참인 경우** 처리 제곱합을 오차항의 분산으로 나눈 통계량은 자유도가 ϕ_A 를 가지는 카이제곱 분포를 따른다.

$$\frac{SS_A}{\sigma_E^2} \sim \chi^2(\phi_A) \quad \text{under } H_0$$

- 잔차 제곱합과 처리 제곱합은 서로 독립이다.

따라서 **귀무가설이 참인 경우** F-통계량은 자유도가 ϕ_A, ϕ_E를 가지는 F-분포를 따른다.

$$F_0 = \frac{MS_A}{MS_E} = \frac{\frac{SS_A/\sigma_E^2}{\phi_A}}{\frac{SS_E/\sigma_E^2}{\phi_E}} \sim F(\phi_A, \phi_E) \quad \text{under } H_0 \tag{1.10}$$

유의수준 α에서 F-통계량이 기각역을 벗어나면 귀무가설을 기각한다.

$$\text{Reject } H_0 \text{ if } F_0 > F(1-\alpha, \phi_A, \phi_E)$$

또는 다음과 같이 계산된 p-값이 유의수준 α 보다 작으면 귀무가설을 기각한다.

$$p - value = P[F(\phi_A, \phi_E) > F_0]$$

F-통계량을 정의할 때 편리하고 유용하게 사용되는 것이 다음과 같은 분산분석표(ANOVA table)이다.

요인	제곱합	자유도	평균제곱합	F_0	p-값
처리	SS_A	$\phi_A = a-1$	$MS_A = SS_A/\phi_A$	$F_0 = MS_A/MS_E$	$P[F(\phi_A, \phi_E) > F_0]$
잔차	SS_E	$\phi_E = a(r-1)$	$MS_E = SS_E/\phi_E$		
총합	SS_T	$\phi_T = ar-1$			

요인	제곱합	자유도	평균제곱합	F_0	p-값

1.3.6. 분산분석 후의 추정

분산분석에서 고려한 요인 A의 수준에 따라서 반응값의 평균에 유의한 차이가 있다고 결론이 나면 그룹 간의 모평균을 차이에 대한 더 자세한 정보가 필요하다. 즉 집단들의 평균이 서로 유의하게 다르거나 같은지에 대한 정보를 얻는 것이 중요하다.

일단 모집단의 분산 σ_E^2 에 대한 추정은 잔차제곱합의 분포를 이용하면 다음과 같은 불편추정량을 얻을 수 있다.

$$\hat{\sigma}_E^2 = MS_E, \quad E(MS_E) = \sigma_E^2$$

다음으로 각 수준(집단)에 대한 평균에 대한 추정량은 표본평균 $\bar{x}_{i.}$ 이며

$$\hat{\mu}_i = \widehat{\mu + \alpha}_i = \bar{x}_{i.} \quad E(\bar{x}_{i.}) = \mu_i$$

$100(1-\alpha)$ % 신뢰구간(confidence interval)은 다음과 같이 주어진다.

$$\bar{x}_{i.} \pm t(1-\alpha/2, \phi_E)\sqrt{\frac{MS_E}{r}}$$

야기서 $t(1-\alpha/2, \phi_E)$는 자유도 ϕ_E를 가지는 t-분포의 $1-\alpha/2$ 분위수를 의미한다.

이제 두 개의 수준에 대한 평균의 차이에 대한 통계적 추론을 생각해 보자. 수준 A_i와 A_j의 평균의 차이에 대한 추정과 검정을 하려고 한다.

$$\delta_{ij} = \mu_i - \mu_j = \alpha_i - \alpha_j$$

두 평균의 차이 δ_{ij} 에 대한 $100(1-\alpha)$ % 신뢰구간은 다음과 같이 주어진다.

$$(\bar{x}_{i.} - \bar{x}_{j.}) \pm t(1-\alpha/2, \phi_E)\sqrt{\frac{2MS_E}{r}} \tag{1.11}$$

신뢰구간 식 1.11 에서 두 개의 표본 평균 $\bar{x}_{i.}$와 $\bar{x}_{j.}$ 은 서로 독립인 것에 유의하자.

이제 마지막으로 두 평균의 차이 δ_{ij} 에 대한 가설을 검정하여고 한다.

$$H_0 : \alpha_i = \alpha_j \quad \text{vs.} \quad H_1 : \alpha_i \neq \alpha_j$$

유의 수준 α에서 다음과 같은 조건을 만족하면 위의 귀무가설을 기각한다.

$$|\bar{x}_{i.} - \bar{x}_{j.}| > t(1-\alpha/2, \phi_E)\sqrt{\frac{2MS_E}{r}} \tag{1.12}$$

식 1.12 에서 주어진 귀무 가설 $\delta_{ij} = 0$ 을 기각하는 조건은 식 1.11 에 주어진 신뢰구간이 0 을 포함하지 않는 조건과 동일하다.

식 1.12 에서 검정을 위한 조건의 우변을 최소유의차(least significant difference; LSD) 라고 부른다. 두 수준의 차이가 유의하려면 두 평균 차이의 절대값이 최소한 최소유의차의 값보다 커야한다.

$$\text{LSD} = t(1-\alpha/2, \phi_E)\sqrt{\frac{2MS_E}{r}}$$

1.3.7. 예제 3.1 - ANOVA F-검정과 사후 추정

다시 예제 3.1의 실험 자료를 고려한다.

```
df31
```

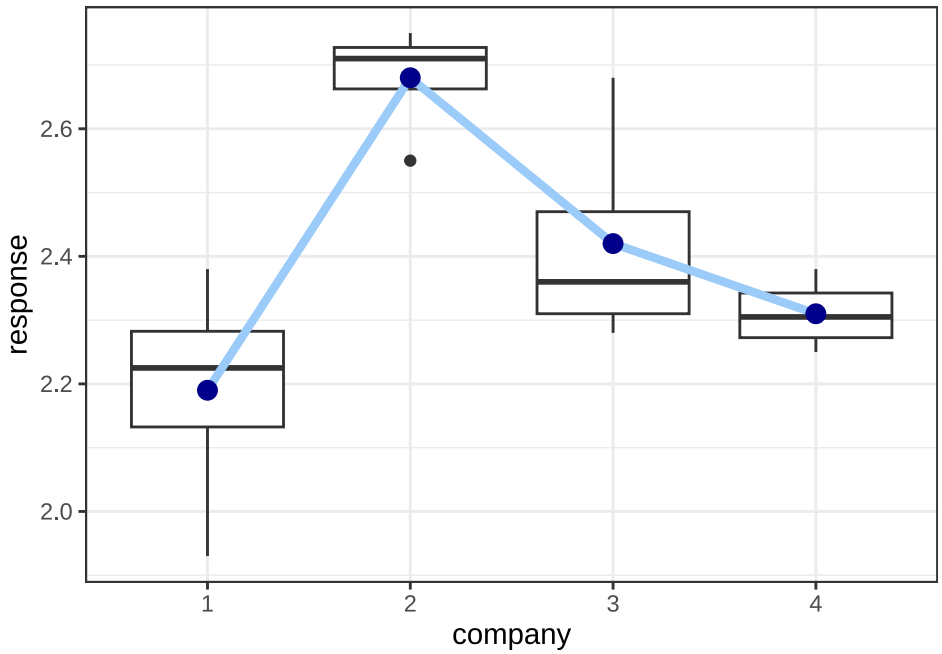
```
   company response
1        1     1.93
2        1     2.38
3        1     2.20
4        1     2.25
5        2     2.55
6        2     2.72
7        2     2.75
8        2     2.70
9        3     2.40
10       3     2.68
11       3     2.32
12       3     2.28
13       4     2.33
14       4     2.38
15       4     2.28
16       4     2.25
```

```
df31s <- df31 %>% group_by(company)  %>%  summarise(mean=mean(response), median= median(response), sd=s

df31s
```

```
# A tibble: 4 x 6
  company  mean median     sd   min   max
  <fct>   <dbl>  <dbl>  <dbl> <dbl> <dbl>
1 1        2.19   2.22 0.189   1.93  2.38
2 2        2.68   2.71 0.0891  2.55  2.75
3 3        2.42   2.36 0.180   2.28  2.68
4 4        2.31   2.30 0.0572  2.25  2.38
```

예제 3.1에서 실험의 목적은 4개의 직물의 굵힘에 대한 저항력을 비교하는 실험이다.

```
ggplot(df31, aes(company, response)) +
  geom_boxplot() +
  geom_line(data=df31s, aes(x=company, y=mean, group=1), size=1.5, col="#9ACBF9") +
  geom_point(data=df31s, aes(x=company, y=mean), col="darkblue", size=3) +
  theme_bw()
```



이제 위에서 제시한 F-검정을 이용하여 납품 업체 간에 직물 마모도에 차이가 있는지 검정해보자.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs.} \quad H_1 : \text{ not } H_0$$

```
anova.res <- aov(response~company,data=df31)
summary(anova.res)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
company      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
```

위의 분산분석표에서 p-값이 유의수준 5% 보다 매우 작으므로 네 개의 직물에 대한 평균이 같다는 귀무가설을 기각한다. 따라서 4개의 납품업체에서 받은 직물의 저항력이 유의하게 다르다고 할 수 있다. 여기서 유의할 점은 ANOVA를 이용한 F-검정은 그룹 간의 차이가 있다는 것을 의미하지만 어떻게 다른지에 대한 정보를 주지 않는다.

최소유의차(LSD) 방법에 의하여 처리 간의 평균을 신뢰구간을 구하고 차이가 있는지 검정할 수 있다.

```
### Mean of response by factor
result1 <- LSD.test(anova.res, "company", group=FALSE, console = TRUE)
```

```
Study: anova.res ~ "company"

LSD t Test for response

Mean Square Error:  0.01988333

company,  means and individual ( 95 %) CI

  response        std r         se      LCL      UCL  Min  Max    Q25   Q50
1     2.19 0.18920888 4 0.07050414 2.036385 2.343615 1.93 2.38 2.1325 2.225
2     2.68 0.08906926 4 0.07050414 2.526385 2.833615 2.55 2.75 2.6625 2.710
3     2.42 0.18036999 4 0.07050414 2.266385 2.573615 2.28 2.68 2.3100 2.360
4     2.31 0.05715476 4 0.07050414 2.156385 2.463615 2.25 2.38 2.2725 2.305
     Q75
1 2.2825
2 2.7275
3 2.4700
4 2.3425

Alpha: 0.05 ; DF Error: 12
Critical Value of t: 2.178813

Comparison between treatments means

      difference pvalue signif.         LCL         UCL
1 - 2      -0.49 0.0004     *** -0.70724487 -0.27275513
1 - 3      -0.23 0.0397       * -0.44724487 -0.01275513
1 - 4      -0.12 0.2520         -0.33724487  0.09724487
2 - 3       0.26 0.0229       *  0.04275513  0.47724487
2 - 4       0.37 0.0030      **  0.15275513  0.58724487
3 - 4       0.11 0.2916         -0.10724487  0.32724487
```

```
result1
```

```
$statistics
     MSerror Df Mean       CV  t.value       LSD
  0.01988333 12  2.4 5.875345 2.178813 0.2172449

$parameters
        test p.ajusted  name.t ntr alpha
  Fisher-LSD      none company   4  0.05

$means
  response        std r         se      LCL      UCL  Min  Max    Q25   Q50
1     2.19 0.18920888 4 0.07050414 2.036385 2.343615 1.93 2.38 2.1325 2.225
2     2.68 0.08906926 4 0.07050414 2.526385 2.833615 2.55 2.75 2.6625 2.710
```

```
3     2.42 0.18036999 4 0.07050414 2.266385 2.573615 2.28 2.68 2.3100 2.360
4     2.31 0.05715476 4 0.07050414 2.156385 2.463615 2.25 2.38 2.2725 2.305
     Q75
1 2.2825
2 2.7275
3 2.4700
4 2.3425

$comparison
      difference pvalue signif.         LCL         UCL
1 - 2      -0.49 0.0004     *** -0.70724487 -0.27275513
1 - 3      -0.23 0.0397       * -0.44724487 -0.01275513
1 - 4      -0.12 0.2520         -0.33724487  0.09724487
2 - 3       0.26 0.0229       *  0.04275513  0.47724487
2 - 4       0.37 0.0030      **  0.15275513  0.58724487
3 - 4       0.11 0.2916         -0.10724487  0.32724487

$groups
NULL

attr(,"class")
[1] "group"
```

최소유의차(LSD) 방법에 의한 평균의 차이에 대한 결과를 이용하여 처리를 다음과 같이 그룹화 하여 보여줄 수 있다.

```
result2 <- LSD.test(anova.res, "company", group=TRUE, console = TRUE)
```

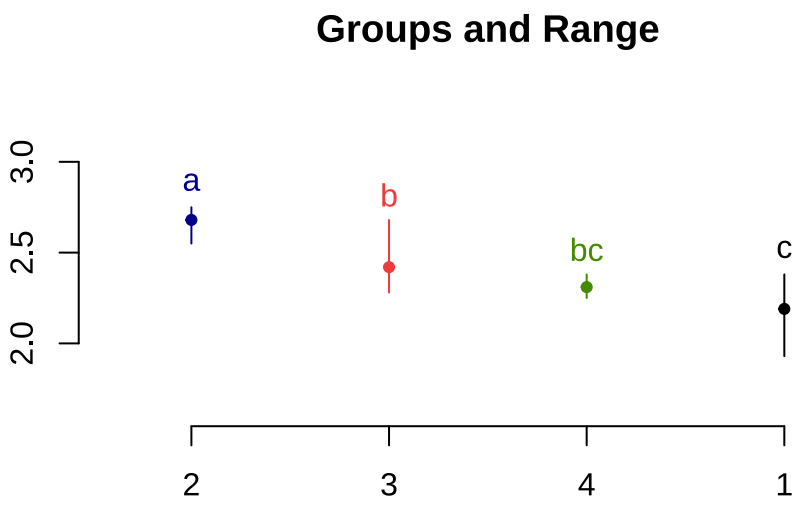
```
Study: anova.res ~ "company"

LSD t Test for response

Mean Square Error:  0.01988333

company,  means and individual ( 95 %) CI

  response        std r         se      LCL      UCL  Min  Max    Q25   Q50
1     2.19 0.18920888 4 0.07050414 2.036385 2.343615 1.93 2.38 2.1325 2.225
2     2.68 0.08906926 4 0.07050414 2.526385 2.833615 2.55 2.75 2.6625 2.710
3     2.42 0.18036999 4 0.07050414 2.266385 2.573615 2.28 2.68 2.3100 2.360
4     2.31 0.05715476 4 0.07050414 2.156385 2.463615 2.25 2.38 2.2725 2.305
     Q75
1 2.2825
2 2.7275
3 2.4700
4 2.3425
```

```
Alpha: 0.05 ; DF Error: 12
Critical Value of t: 2.178813

least Significant Difference: 0.2172449

Treatments with the same letter are not significantly different.

  response groups
2     2.68      a
3     2.42      b
4     2.31     bc
1     2.19      c
```

```
result2$groups
```

```
  response groups
2     2.68      a
3     2.42      b
4     2.31     bc
1     2.19      c
```

```
plot(result2)
```



2. 이원배치법

2.1. 예제 4.1

예제 4.1 은 교과서 89 페이지에 나온 분석 예제이다. 4종류의 사료(A)와 3종류의 돼지품종(B)이 체중 증가에 미치는 영향을 조사한 실험이다. 각 처리 조합마다 3회 반복실험하여 총 36개의 관측값을 얻었다.

2.1.1. 자료 읽기

다음과 같은 순서로 자료를 가진 데이터프레임 `df2`을 만들어 보자.

```
response<- c(64, 66, 70, 72, 81, 64,
             74, 51, 65, 65, 63, 58,
             57, 43, 52, 47, 58, 67,
             59, 68, 65, 66, 71, 59,
             58, 39, 42, 58, 41, 46,
             57, 61, 53, 53, 59, 38)
response
```

```
 [1] 64 66 70 72 81 64 74 51 65 65 63 58 57 43 52 47 58 67 59 68 65 66 71 59 58
[26] 39 42 58 41 46 57 61 53 53 59 38
```

```
food<- factor(rep(c(1:4), each=9))
breed<- factor(rep(c(1:3), each=3))
food
```

```
 [1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4
Levels: 1 2 3 4
```

```
breed
```

```
[1] 1 1 1 2 2 2 3 3 3
Levels: 1 2 3
```

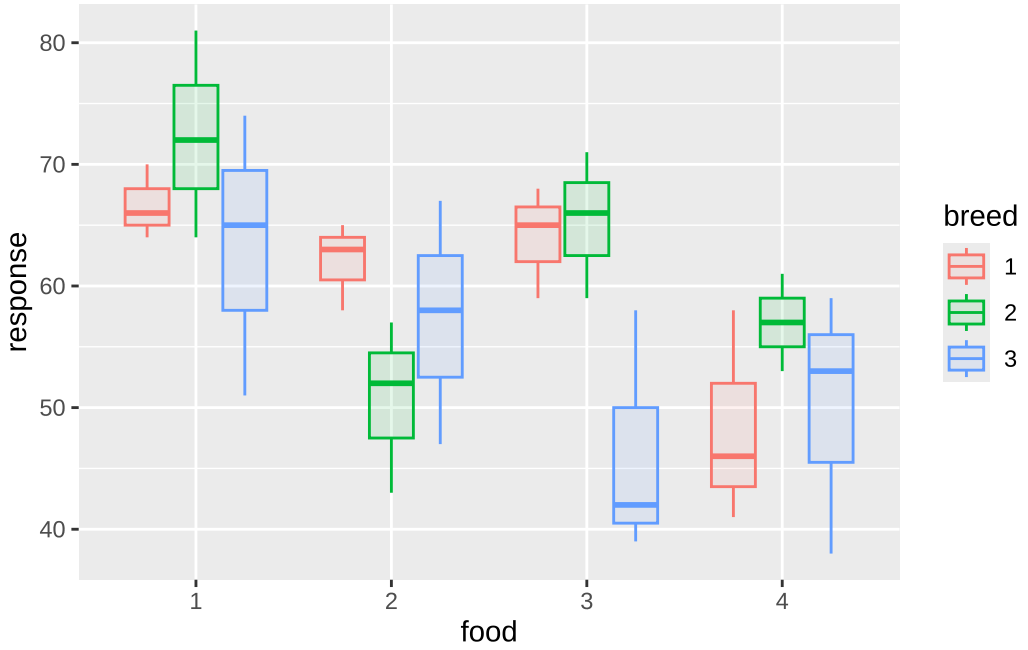
```
df2<- data.frame(food, breed, response)
head(df2)
```

```
  food breed response
1    1     1       64
2    1     1       66
3    1     1       70
4    1     2       72
5    1     2       81
6    1     2       64
```

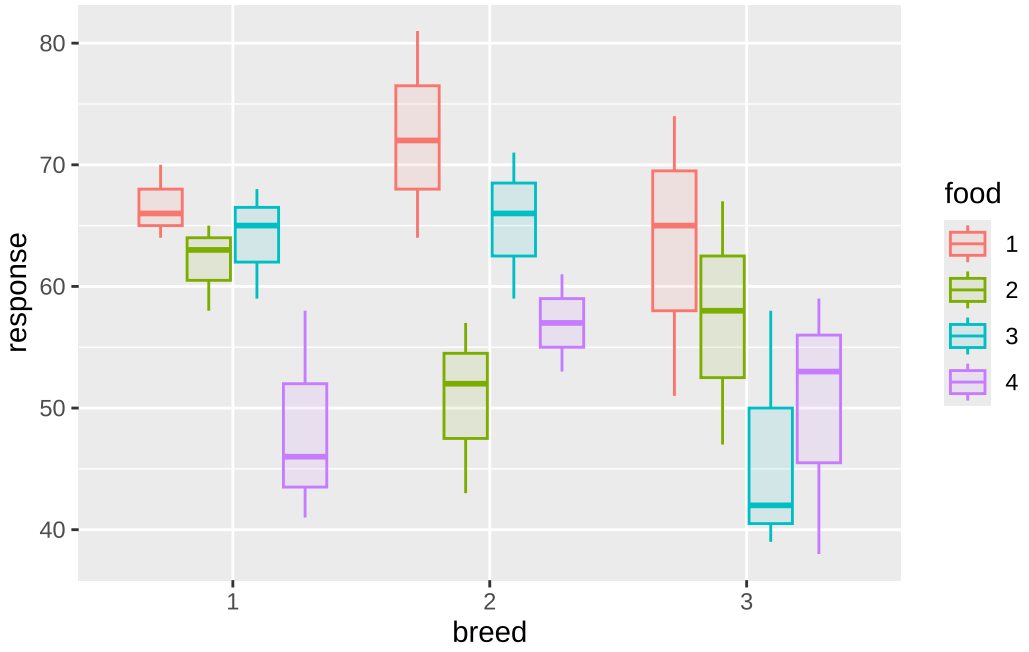
2.1.2. 자료의 시각화와 기초 통계량

이제 처리별로 효과를 시각적으로 비교하기 위하여 자료들에 대한 산점도와 상자그림을 그려보자

```
df2 %>%
  ggplot() +
  aes(x = food , y = response, fill=breed, color=breed) +
  geom_boxplot(alpha = 0.1, width = 0.75)
```

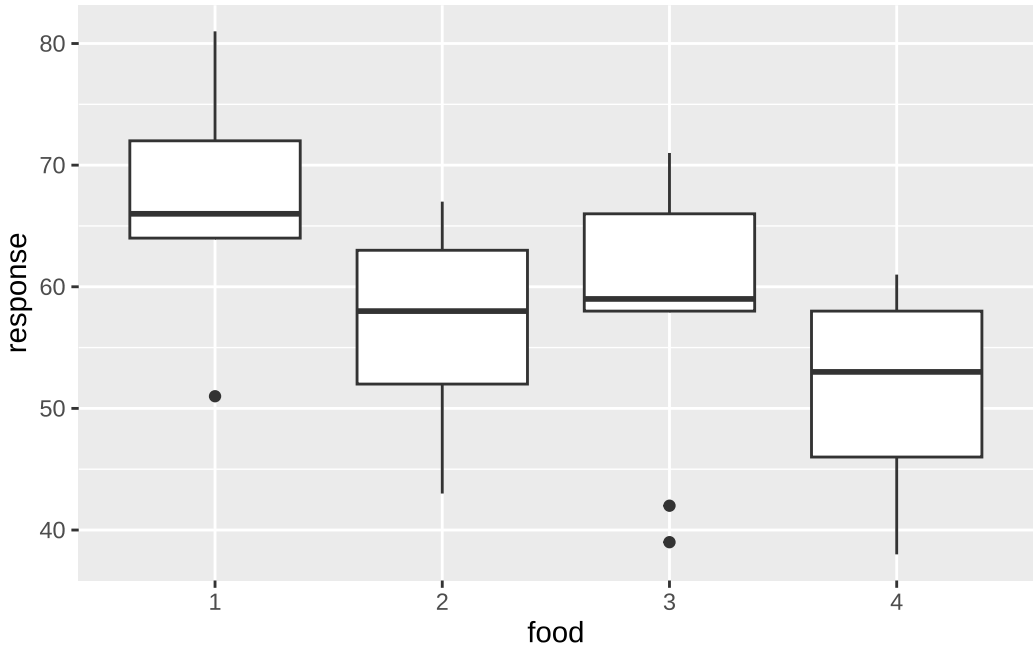



```
df2 %>%
  ggplot() +
  aes(x = breed  , y = response, fill=food, color=food) +
  geom_boxplot(alpha = 0.1, width = 0.75)
```

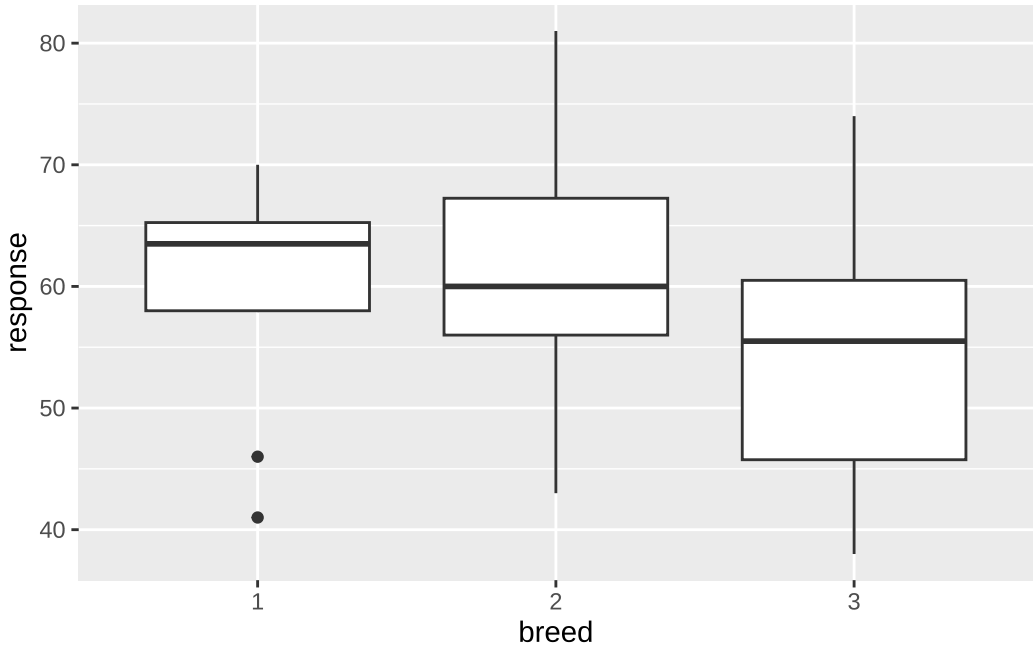


위와 같이 두 요인의 조합으로 그림을 보는 것보다 각 요인별로 요약하여 보는 것도 유용하다.

```
df2 %>%
  ggplot( aes(x = food , y = response) ) +
  geom_boxplot()
```




```
df2 %>%
  ggplot( aes(x = breed , y = response))  +
  geom_boxplot()
```



다음으로 12개의 처리 조합에 대한 체중의 기초통계량(평균과 표준편차)을 구해보자.

```
df2s <- df2 %>% group_by(food, breed)  %>%  summarise(mean=mean(response),  sd=sd(response))
df2s
```

```
# A tibble: 12 x 4
# Groups:   food [4]
   food  breed  mean    sd
   <fct> <fct> <dbl> <dbl>
 1 1     1      66.7  3.06
 2 1     2      72.3  8.50
 3 1     3      63.3 11.6
 4 2     1      62    3.61
 5 2     2      50.7  7.09
 6 2     3      57.3 10.0
 7 3     1      64    4.58
 8 3     2      65.3  6.03
 9 3     3      46.3 10.2
10 4     1      48.3  8.74
11 4     2      57    4
12 4     3      50   10.8
```

또한 각 요인에 대한 기초통계량도 구해보자.

```
df2s_food<- df2 %>% group_by(food)  %>%  summarise(mean=mean(response),  sd=sd(response))
df2s_food
```

```
# A tibble: 4 x 3
  food   mean    sd
```

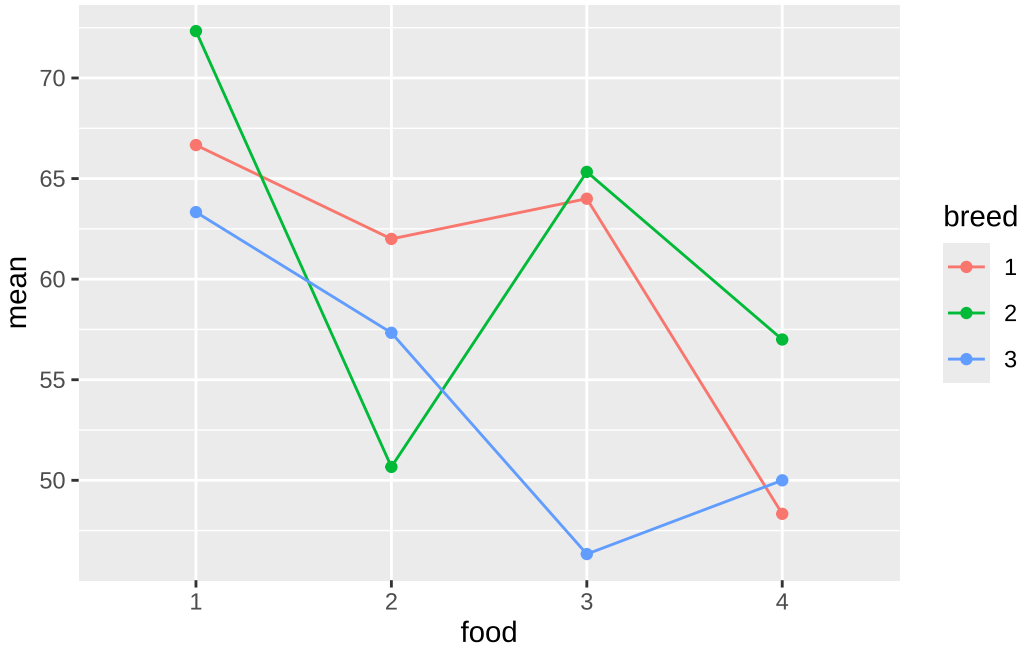
```
  <fct> <dbl> <dbl>
1 1      67.4  8.34
2 2      56.7  8.08
3 3      58.6 11.2
4 4      51.8  8.26
```

```
df2s_breed <- df2 %>% group_by(breed)  %>%  summarise(mean=mean(response),  sd=sd(response))
df2s_breed
```

```
# A tibble: 3 x 3
  breed  mean    sd
  <fct> <dbl> <dbl>
1 1      60.2  8.74
2 2      61.3 10.3
3 3      54.2 11.4
```

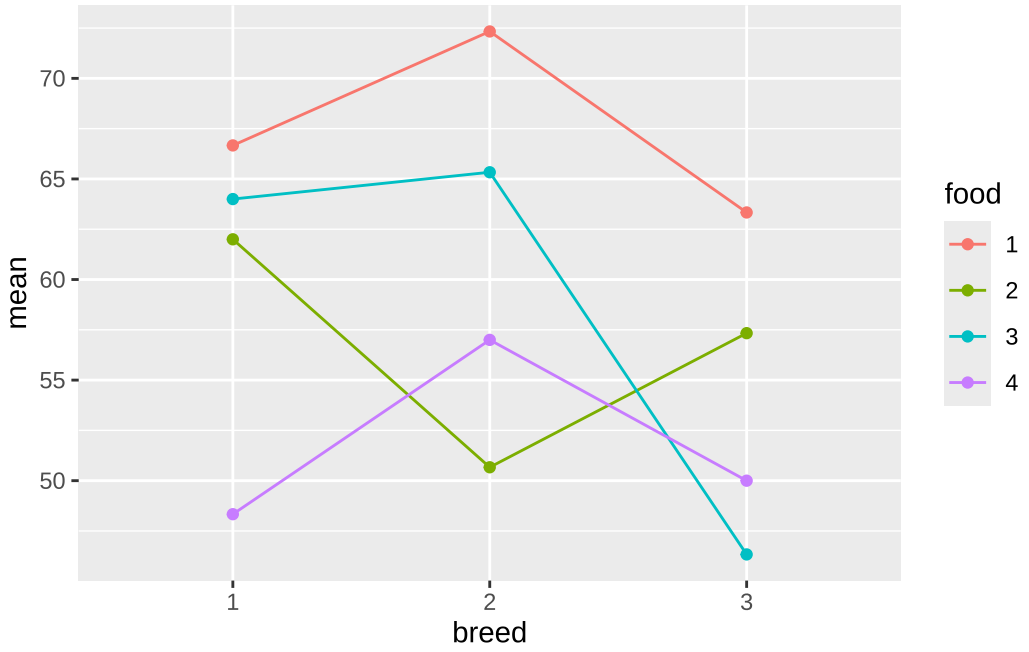
이제 위에서 계산된 처리 그룹에 대한 평균으로 상호작용 그림을 그려보자. 아래 그림에서 사료의 종류에 따라서 체중의 변화를 본 그림이다. 사료 1번에서 체중이 가장 크게 나타났고 다른 사료에 대해서는 체중이 줄어드는데 품종에 따라서 그 크기가 서로 다르다.

```
df2s %>%
  ggplot() +
  aes(x = food , y = mean, color =breed) +
  geom_line(aes(group = breed)) +
  geom_point()
```



아래 그림은 아래 그림에서 품종의 종류에 따라서 체중의 변화를 본 그림이다.

```
df2s %>%
  ggplot() +
  aes(x = breed , y = mean, color =food) +
  geom_line(aes(group = food)) +
  geom_point()
```



사료와 품종간에 상호 작용이 그림으로 나타나고 있지만 뚜렷하지 않고 해석하기도 힘들다.

2.1.3. 분산분석표와 가설검정

이제 이원배치법에서의 가설검정을 수행하기 위하여 분산분석 표를 구해보자.

```
df2aov <- aov(response ~ food*breed, data=df2)
summary(df2aov)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
food         3 1156.6   385.5   6.163 0.00294 **
breed        2  349.4   174.7   2.793 0.08121 .
food:breed   6  771.3   128.5   2.055 0.09712 .
Residuals   24 1501.3    62.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 상호작용에 대한 가설 검정

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_6 = 0 \quad \text{vs.} \quad H_1 : \text{ not } H_0$$

분산분석표에서 상호작용에 대한 가설 검정을 위한 F-통계량은 2.055이고 p-값은 0.097으로 유의수준 0.05보다 크므로 귀무가설 H_0를 기각할 수 있다. 따라서 사료와 품종 간의 상호작용은 유의하지 않다. 하지만 p-값이 0.1 미만이므로 품종에 따라서 사료가 주는 효과가 약간은 다를 가능성이 존재한다.

> **노트**
>
> 상호작용에 대한 p-값이 0.25 보다 작으므로 상호작용에 대한 모수를 가진 모형을 그대로 사용한다. (교과서 88 페이지 참조)
> 상호작용을 모형에서 제외하는 기준을 일반적으로 정하는 방법은 매우 어려우며 실험계획의 적용되는 분야와 문제에 따라 달리질 수 있다. 또한 기준을 설병할 때는 고유 분야에 대한 지식과 경험이 필요하다.
> 본 강의에서는 학생들이 상호작용을 모형에서 제외하는판단은 하지 않으며 과제나 시험에서 상호작용을 모형에서 제외하는 판단을 요구하지 않는다.

- 주효과에 대한 가설 검정

  주효과에 대한 검정에서 품종에 대한 검정은 p-값이 0.081로서 유의수준 5%에서 귀무가설을 기각할 수 없으므로 돼지 품종에 따라서는 유의한 차이가 없다. 다만 유의수준 1%에서는 유의하므로 약간의 차이는 있다고 말할 수 있다.

  사료에 대한 검정은 p-값이 0.003로서 유의수준 5%에서 귀무가설을 기각할 수 있어서 사료에 따라서는 유의한 차이가 있다.

> **노트**
>
> 보통 유의수준 1%에서 유의하면 "**제한적으로 유의하다**"(marginally significant)라고 말한다.

2.1.4. 분산분석 후의 추정

2.1.4.1. 모평균에 대한 추론

이원배치에서 유의한 상호작용이 있는 경우 처리수준 A_iB_j에 대한 모평균 μ_{ij} 에 대한 추정량은 처리수준 A_iB_j에서의 관측값들의 평균 $\bar{x}_{ij.}$ 이며 오차항의 분산 σ_E^2는 분산분석표에서 MS_E로 추정할 수 있다.

$$\hat{\sigma}_E^2 = MS_E = \frac{SS_E}{ab(r-1)} = \frac{1501.3}{24} = 62.6$$

위의 결과를 이용하면 처리수준 A_iB_j에 대한 모평균 μ_{ij}에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같이 주어진다.

$$\bar{x}_{ij.} \pm t(1-\alpha/2, ab[r-1])\sqrt{\frac{MS_E}{r}}$$

예를 들어 사료가 1 이고($i=1$) 품종이 1인 경우($j=1$) 체중의 평균 μ_{11} 에 대한 95% 신뢰 구간을 구해보자. 일단 위의 기초 통계량에서 $\bar{x}_{11.} = 66.7$ 이고 분산분석표에서 $MS_E = 62.6$, $r=3$ 그리고 t-분포의 백분위수 $t(0.975, 24)$ 은 다음과 같이 주어진다.

```
qt(0.975, 24)
```

```
[1] 2.063899
```

따라서 μ_{11} 에 대한 95% 신뢰 구간은 다음과 같다.

$$\bar{x}_{11.} \pm t(1-\alpha/2, 24)\sqrt{\frac{MS_E}{r}} = 66.7 \pm (2.06)\sqrt{\frac{62.6}{3}} = 66.7 \pm (2.06)(4.56) = (57, 76) \tag{2.1}$$

패키지 emmeans에 있는 함수 emmeans()를 다음과 같이 사용하면 각 처리에 대한 평균의 95% 신뢰구간을 쉽게 구할 수 있다.

```
emmeans(df2aov, "food", "breed")
```

```
breed = 1:
 food emmean   SE df lower.CL upper.CL
 1      66.7 4.57 24     57.2     76.1
 2      62.0 4.57 24     52.6     71.4
 3      64.0 4.57 24     54.6     73.4
 4      48.3 4.57 24     38.9     57.8

breed = 2:
 food emmean   SE df lower.CL upper.CL
 1      72.3 4.57 24     62.9     81.8
 2      50.7 4.57 24     41.2     60.1
 3      65.3 4.57 24     55.9     74.8
 4      57.0 4.57 24     47.6     66.4

breed = 3:
 food emmean   SE df lower.CL upper.CL
 1      63.3 4.57 24     53.9     72.8
 2      57.3 4.57 24     47.9     66.8
 3      46.3 4.57 24     36.9     55.8
 4      50.0 4.57 24     40.6     59.4

Confidence level used: 0.95
```

2.2. 반복이 있는 이원배치에서 상호작용이 없는 경우의 추론

교과서에서 상호작용의 유의성에 따라서 모형을 축소하는 기준을 다음과 같이 제시하고 있다.

상호작용에 대한 p-값이 0.25보다 큰 경우 상호작용이 존재하지 않는다고 판단하고 오차항에 풀링한다. 상호작용을 오차항에 풀링한다는 것은 다음과 같은 모형을 사용한다는 의미이다.

$$x_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \tag{2.2}$$

만약 예제 4.1에 대한 반복이 있는 자료에서 위와 같이 오차항을 풀링한 모형을 적합해 보면 아래와 같은 분산분석표를 얻는다.

```r
df2aov2 <- aov(response ~ food + breed, data=df2)
summary(df2aov2)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
food         3 1156.6   385.5   5.089 0.00575 **
breed        2  349.4   174.7   2.306 0.11705
Residuals   30 2272.6    75.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

만약 **반복이 있는 이원배치 모형**에서 상호작용 $A \times B$가 존재하지 않고 주효과만 유의한 경우, 즉 모형 식 2.2 을 가정한 경우 모평균 μ_{ij}에 대한 모수는 다음과 같다.

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

이러한 경우 모평균 μ_{ij}에 대한 최소제곱 추정량(least square estimator)은 표본 평균 $\bar{x}_{ij.}$이 아니라 다음과 같은 추정량이 주어진다.

$$\begin{aligned}\hat{\mu}_{ij} &= \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j \\ &= (\bar{\bar{x}}) + (\bar{x}_{i..} - \bar{\bar{x}}) + (\bar{x}_{.j.} - \bar{\bar{x}}) \\ &= \bar{x}_{i..} + \bar{x}_{.j.} - \bar{\bar{x}}\end{aligned}$$

위에서 주어진 $\hat{\mu}_{ij}$ 는 모평균 μ_{ij}의 불편 추정량이며 상호작용 $A \times B$ 이 없는 모형 @ref(eq:nointer) 에서 표본 평균 $\bar{x}_{ij.}$ 보다 분산이 작은 추정량이다. 즉,

$$Var\left(\hat{\mu}_{ij}\right) = \frac{\sigma_E^2}{n_e} \leq \frac{\sigma_E^2}{r} = Var(\bar{x}_{ij.})$$

위의 식에서 유효 반복수 n_e 는 다음과 같이 정의된다.

$$\frac{1}{n_e} = \frac{1}{br} + \frac{1}{ar} - \frac{1}{abr}, \quad n_e = \frac{abr}{a+b-1}$$

따라서 이 경우 모평균 μ_{ij}에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같은 주어진다.

$$\hat{\mu}_{ij} \pm t(1-\alpha/2, \phi_E)\sqrt{\frac{MS_E}{n_e}}$$

주의할 점은 위의 신뢰구간에서 MS_E는 상호작용이 없는 모형 식 2.2 으로 유도된 분산분석표에 나타난 MS_E 이며 자유도는 $\phi_E = abr - a - b + 1$ 이다.

참고로 예제 4.1 경우 $a = 4$, $b = 3$, $r = 3$이므로 유효 반복수 n_e 는 다음과 같이 주어진다.

$$n_e = \frac{abr}{a+b-1} = \frac{(4)(3)(3)}{4+3-1} = 6$$

상호작용 $A \times B$ 이 없는 모형 식 2.2 에서 적용한 분산분석 결과 df2aov2 에 대하여 모형 식 2.2 에서 각 처리에 대한 평균 μ_{ij}에 대한 최소제곱 추정량 $\hat{\mu}_{ij} = \bar{x}_{i..} + \bar{x}_{.j.} - \bar{\bar{x}}$ 과 95% 신뢰구간을 다음과 같이 구할 수 있다.

```
df2s_food$mean[1]
```

```
[1] 67.44444
```

```
df2s_breed$mean[1]
```

```
[1] 60.25
```

```
mean(df2$response)
```

```
[1] 58.61111
```

$$\hat{\mu}_{ij} = \bar{x}_{i..} + \bar{x}_{.j.} - \bar{\bar{x}} = 67.4 + 60.3 - 58.6 = 69.1$$

이제 상호작용 $A \times B$ 이 없는 모형 식 2.2 에서 μ_{11} 에 대한 95% 신뢰 구간은 다음과 같다.

$$\begin{aligned} \hat{\mu}_{11} \pm t(1-\alpha/2, 30)\sqrt{\frac{MS_E}{n_e}} &= 69.1 \pm (2.04)\sqrt{\frac{75.8}{6}} \\ &= 69.1 \pm (2.04)(3.55) \\ &= (61.8, 76.3) \end{aligned}$$

위의 신뢰구간 $(61.8, 76.3)$은 상호 작용이 포함된 모형에서 유도한 신뢰구간 식 2.1 에서 구한 $(57, 76)$과 다르다.

함수 emmeans()를 분산분석 결과df2aov2 에 대하여 다음과 같이 사용하면 상호작용 $A \times B$ 이 없는 모형 식 2.2 에서 각 처리에 대한 평균 μ_{ij}에 대한 최소제곱 추정량 $\hat{\mu}_{ij}$ 과 95% 신뢰구간을 다음과 같이 구할 수 있다.

```
emmeans(df2aov2, "food", "breed")
```

```
breed = 1:
 food emmean   SE df lower.CL upper.CL
 1      69.1 3.55 30     61.8     76.3
 2      58.3 3.55 30     51.0     65.6
 3      60.2 3.55 30     52.9     67.5
 4      53.4 3.55 30     46.2     60.7

breed = 2:
 food emmean   SE df lower.CL upper.CL
 1      70.2 3.55 30     62.9     77.4
 2      59.4 3.55 30     52.1     66.6
 3      61.3 3.55 30     54.0     68.5
 4      54.5 3.55 30     47.2     61.8
```

```
breed = 3:
 food emmean   SE df lower.CL upper.CL
 1      63.1 3.55 30     55.8     70.3
 2      52.3 3.55 30     45.0     59.6
 3      54.2 3.55 30     46.9     61.5
 4      47.4 3.55 30     40.2     54.7

Confidence level used: 0.95
```

위에서 나타난 emmean 은 μ_{ij}에 대한 최소제곱 추정량 $\bar{x}_{i..} + \bar{x}_{.j.} - \bar{\bar{x}}$ 으로서 아래 주어진 표본평균 $\bar{x}_{ij.}$ 과 다른 값으로 나타남을 알 수 있다.

```
df2s
```

```
# A tibble: 12 x 4
# Groups:   food [4]
   food  breed  mean    sd
   <fct> <fct> <dbl> <dbl>
 1 1     1      66.7  3.06
 2 1     2      72.3  8.50
 3 1     3      63.3 11.6
 4 2     1      62    3.61
 5 2     2      50.7  7.09
 6 2     3      57.3 10.0
 7 3     1      64    4.58
 8 3     2      65.3  6.03
 9 3     3      46.3 10.2
10 4     1      48.3  8.74
11 4     2      57    4
12 4     3      50   10.8
```

2.3. 전지의 수명 실험

전지(battery)를 제조하는 회사의 기술자들이 전지의 수명(BatteryLife)에 영향을 미치는 두 요인, 온도(Temperature)와 재료(MaterialType)의 효과를 알아보기 위해서 실행한 실험입니다.

기술자들은 온도가 크게 변할 때 전지의 수명에 어떤 영향을 미치는지 알아보기 위하여 실험을 실시하였다. 온도는 3개의 수준(15도, 70도, 125도)을 고려하였다. 전지를 생산하는 재료가 3개이므로 재료는 3개의 수준(type 1,2,3)으로 구성되어 있다. 이 실험은 9 개의 처리($ab = 3 \times 3 = 9$)에 대하여 각각 4번의 반복 측정($r = 4$)을 실시하였다.

자료의 출처는 (Montgomery 2017) 에 나와 있다

자료를 얻기 위해서는 다음과 같은 R 프로그램을 실행하여 패키지 MontgomeryDAE를 설치하고 실행해야 한다.

```
install.packages("remotes")
remotes::install_github("ehassler/MontgomeryDAE")
library(MontgomeryDAE)
```

2.3.1. 자료 읽기

이제 전지의 수명 실험 자료를 읽어 오자. 전지의 수명 실험에 대한 자료는 데이터프레임 Table5.1에 있다.

```
df <- Table5.1
head(df) # 자료의 앞부분만 보기
```

```
  MaterialType Temperature BatteryLife
1            1          15         130
2            1          15          74
3            2          15         150
4            2          15         159
5            3          15         138
6            3          15         168
```

함수 str()은 자료의 구조와 자료 안에 있는 변수의 형식을 보여준다.

```
str(df)  # 자료의 구조를 알아보는 명령
```

```
'data.frame':   36 obs. of  3 variables:
 $ MaterialType: chr  "1" "1" "2" "2" ...
 $ Temperature : num  15 15 15 15 15 15 15 15 15 15 ...
 $ BatteryLife : num  130 74 150 159 138 168 155 180 188 126 ...
```

위의 결과를 보면 데이터프레임 df에 있는 변수 MaterialType은 문자형 변수(chr)이고 나머지는 숫자형 변수(num)이다. 두 요인에 대한 변수인 MaterialType와 Temperature를 함수 factor()를 이용하여 범주형 변수로 만들어 주자 .

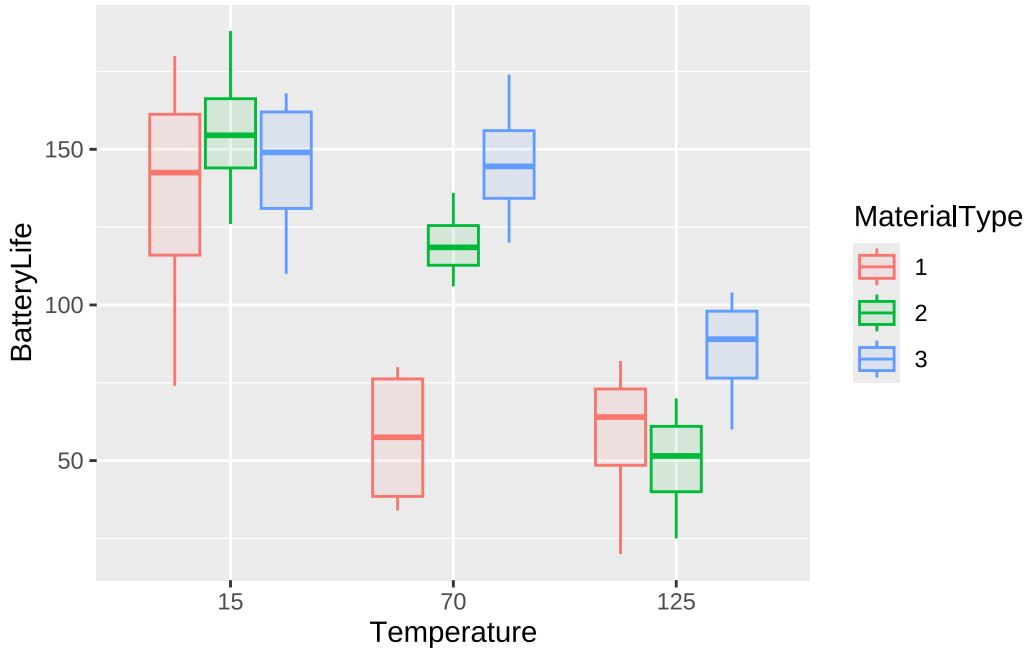
```
df$MaterialType <- factor(df$MaterialType)
df$Temperature <- factor(df$Temperature)
str(df)
```

```
'data.frame':   36 obs. of  3 variables:
 $ MaterialType: Factor w/ 3 levels "1","2","3": 1 1 2 2 3 3 1 1 2 2 ...
 $ Temperature : Factor w/ 3 levels "15","70","125": 1 1 1 1 1 1 1 1 1 1 ...
 $ BatteryLife : num  130 74 150 159 138 168 155 180 188 126 ...
```

2.3.2. 자료의 시각화와 기초 통계량

이제 처리별로 효과를 시각적으로 비교하기 위하여 자료들에 대한 산점도와 상자그림을 그려보자

```
df %>%
  ggplot() +
  aes(x = Temperature , y = BatteryLife, fill=MaterialType, color=MaterialType, group = interaction(Tem
  geom_boxplot(alpha = 0.1, width = 0.75)
```



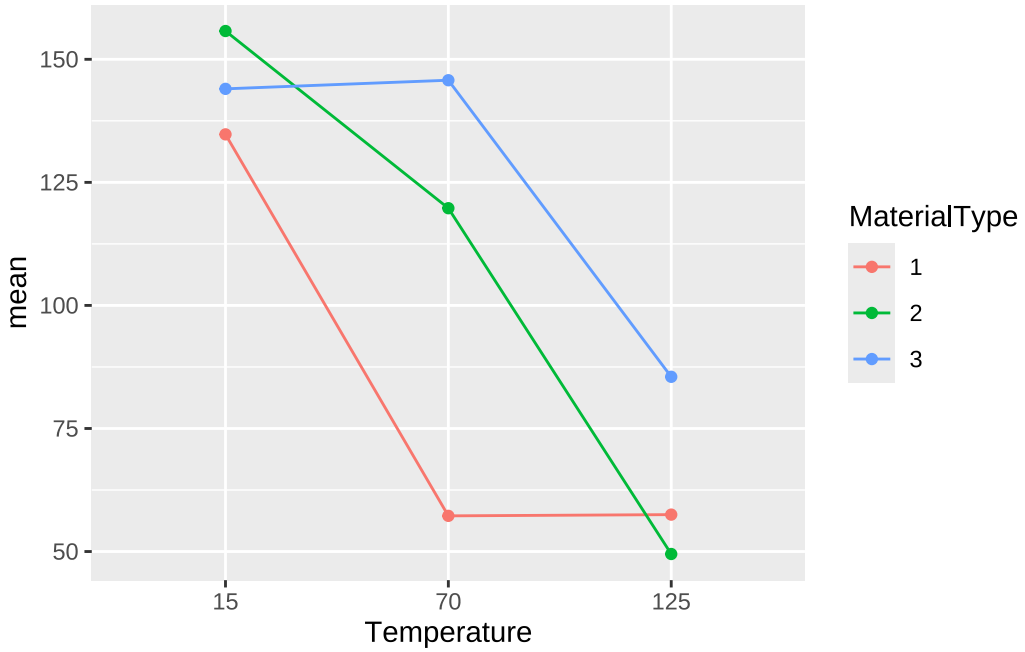
다음으로 6개의 처리 조합에 대한 전지 수명의 기초통계량(평균과 표준편차)을 구해보자.

```
dfs <- df %>% group_by(MaterialType, Temperature)  %>%  summarise(mean=mean(BatteryLife),  sd=sd(Batter
dfs
```

```
# A tibble: 9 x 4
# Groups:   MaterialType [3]
  MaterialType Temperature  mean    sd
  <fct>        <fct>       <dbl> <dbl>
1 1            15          135.   45.4
2 1            70           57.2  23.6
3 1            125          57.5  26.9
4 2            15          156.   25.6
5 2            70          120.   12.7
6 2            125          49.5  19.3
7 3            15          144    26.0
8 3            70          146.   22.5
9 3            125          85.5  19.3
```

이제 위에서 계산된 처리 그룹에 대한 평균으로 상호작용 그림을 그려보자. 아래 그림에서 온도가 증가할 수록 전지의 수명이 감소하는 경향을 보이고 있다. 또한 각 재료에 따른 온도의 변화가 수평으로 나타나지 않고 있음을 알 수 있다. 이러한 점은 온도와 재료 사이에 유의한 상호작용이 있다고 예측할 수 있다.

```
dfs %>%
  ggplot() +
  aes(x = Temperature , y = mean, color =MaterialType) +
  geom_line(aes(group = MaterialType)) +
  geom_point()
```



2.3.3. 분산분석표와 가설검정

이제 다음과 같은 모형에서 이원배치법에서의 가설검정을 수행하기 위하여 분산분석 표를 구해보자.

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

요인	제곱합	자유도	평균제곱합	F_0
요인 A	SS_A	$a-1$	MS_A	MS_A/MS_E
요인 B	SS_B	$b-1$	MS_B	MS_B/MS_E
상호작용 $A\times B$	$SS_{A\times B}$	$(a-1)(b-1)$	$MS_{A\times B}$	$MS_{A\times B}/MS_E$
잔차 E	SS_E	$ab(r-1)$	MS_E	
총합	SS_T	$abr-1$		

```
dfaov <- aov(BatteryLife~ MaterialType + Temperature + MaterialType:Temperature, data=df)
# This is equivalent to aov(BatteryLife~ MaterialType *Temperature , data=df)
summary(dfaov)
```

```
                          Df Sum Sq Mean Sq F value   Pr(>F)
MaterialType               2  10684    5342   7.911  0.00198 **
Temperature                2  39119   19559  28.968 1.91e-07 ***
MaterialType:Temperature   4   9614    2403   3.560  0.01861 *
Residuals                 27  18231     675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 상호작용에 대한 가설 검정

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{3,3} = 0 \quad \text{vs.} \quad H_1 : \text{ not } H_0$$

분산분석표에서 상호작용에 대한 가설 검정을 위한 F-통계량은 다음과 같다.

$$F_0 = \frac{MS_{A\times B}}{MS_E} = \frac{SS_{A\times B}/\phi_{AB}}{SS_E/\phi_E} = \frac{9614/4}{18231/27} = 3.560$$

위의 F-통계량에 대한 p-값은 0.0186으로 유의수준 0.05보다 작으므로 귀무가설 H_0를 기각한다. 따라서 온도와 재료의 상호작용은 유의하다.

- 주효과에 대한 가설 검정

위에서 유의한 상호작용이 있다고 판단하였기 때문에 주효과에 대한 가설검정은 기술적 의미가 없다. 기술적으로 의미가 없다는 것은 유의한 상호작용이 있으면 이미 주효과 A 의크기가 B 의 수준에 따라서 다르므로 주효과가 유의하게 있다는 것을 뜻한다.

2.3.4. 분산분석 후의 추정

2.3.5. 모평균에 대한 추론

이원배치에서 유의한 상호작용이 있는 경우 처리수준 A_iB_j에 대한 모평균 μ_{ij} 은 다음과 같다.

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} = \mu + \tau_{ij}$$

이때 μ_{ij} 에 대한 추정량은 처리수준 A_iB_j에서의 관측값들의 평균 $\bar{x}_{ij.}$으로 다음과 같은 분포를 따른다.

$$\bar{x}_{ij.} \sim N(\mu_{ij}, \sigma_E^2/r)$$

오차항의 분산 σ_E^2는 분산분석표에서 MS_E로 추정할 수 있다.

$$\hat{\sigma}_E^2 = MS_E = \frac{SS_E}{ab(r-1)} = \frac{18231}{27} = 675$$

위의 결과를 이용하면 처리수준 A_iB_j에 대한 모평균 μ_{ij}에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같이 주어진다.

$$\bar{x}_{ij.} \pm t(1-\alpha/2, ab[r-1])\sqrt{\frac{MS_E}{r}}$$

예를 들어 전지의 수명실험에서 온도가 70도이고($i=2$) 재료의 형태가 3인 경우($j=3$) 수명 시간의 평균 μ_{23} 에 대한 95% 신뢰 구간을 구해보자. 일단 위의 기초 통계량에서 $\bar{x}_{13.} = 146$ 이고 분산분석표에서 $MS_E = 675$, $r = 4$ 그리고 t-분포의 백분위수 $t(0.975, 27)$ 은 다음과 같이 주어진다.

```
qt(0.975, 27)
```

```
[1] 2.051831
```

따라서 μ_{23} 에 대한 95% 신뢰 구간은 다음과 같다.

$$\bar{x}_{23.} \pm t(1-\alpha/2, ab[r-1])\sqrt{\frac{MS_E}{r}} = 146 \pm (2.05)\sqrt{\frac{675}{4}} = (119, 172)$$

패키지 emmeans에 있는 함수 emmeans()를 다음과 같이 사용하면 각 처리에 대한 평균의 95% 신뢰구간을 쉽게 구할 수 있다. 함수 emmeans()의 첫 번째 인자는 분산분석의 결과(aov()의 결과)이며 다음의 인자들은 요인에 대한 변수명을 써주면 된다.

```
emmeans(dfaov, "MaterialType","Temperature")
```

```
Temperature = 15:
 MaterialType emmean SE df lower.CL upper.CL
 1             134.8 13 27    108.1    161.4
 2             155.8 13 27    129.1    182.4
 3             144.0 13 27    117.3    170.7

Temperature = 70:
 MaterialType emmean SE df lower.CL upper.CL
 1              57.2 13 27     30.6     83.9
 2             119.8 13 27     93.1    146.4
 3             145.8 13 27    119.1    172.4

Temperature = 125:
 MaterialType emmean SE df lower.CL upper.CL
 1              57.5 13 27     30.8     84.2
 2              49.5 13 27     22.8     76.2
 3              85.5 13 27     58.8    112.2

Confidence level used: 0.95
```

함수 emmeans()에서 출력되는 SE는 표분오차(standard error)를 의미하며 이는 평균의 추정량 $\bar{x}_{ij.}$ 의 표준편차(standard deviation)이다.

$$\hat{\mathrm{SE}}(\bar{x}_{ij.}) = \hat{sd}(\bar{x}_{ij.}) = \sqrt{\hat{Var}(\bar{x}_{ij.})} = \sqrt{\frac{MS_E}{r}} = \sqrt{675/4} = 13.0$$

2.3.6. 미래의 관측값에 대한 추론

처리수준 A_iB_j에 대한 미래의 관측값에 대한 신뢰구간을 구하는 경우 관측 오차에 의한 불확실성을 반영하기 때문에 그 신뢰구간은 다음과 같이 주어진다.

$$\bar{x}_{ij.} \pm t(1-\alpha/2, ab[r-1])\sqrt{\frac{MS_E}{r} + MS_E}$$

참고로 다른 교과서에서는 관측값에 대한 신뢰구간을 예측구간(prediction interval)이라고 부른다. 이는 모수는 추정(estimation)하지만 관측값은 예측(prediction)한다고 말하기 때문이다.

3. 블록설계, 라틴정방설계와 분할법

3.1. 블럭설계 예제

다음은 교과서 예제 5.1 -플라스틱 강도 실험을 분석하는 예제이다.

플라스틱 제품의 강도를 측정하는 것이 실험의 목적이다. 랜덤하게 4일을 택해서 각 일마다 온도를 3개 수준으로 랜덤하게 변화시켜서 제품의 강도(intensity)를 측정하였다.

여기서 온도(temp)는 고정효과(τ)이며 선택된 일(day)는 블럭(ρ)에 따른 효과이다.

$$x_{ij} = \mu + \tau_i + \rho_j + e_{ij}$$

3.1.1. 자료의 구성

이제 실험자료를 입력하여 데이터프레임으로 만들어 보자

```
intensity<- c(98.0, 97.7, 96.5,
              99.0, 98.0, 97.9,
              98.6, 98.2, 96.9,
              97.6, 97.3, 96.7)

temp <- factor(rep(c(70, 80, 90), times=4))
day <- as.factor(rep(c(1:4), each=3))

df<- data.frame(intensity=intensity, temp=temp, day=day)
df
```

```
   intensity temp day
1       98.0   70   1
2       97.7   80   1
3       96.5   90   1
4       99.0   70   2
5       98.0   80   2
6       97.9   90   2
7       98.6   70   3
8       98.2   80   3
9       96.9   90   3
10      97.6   70   4
11      97.3   80   4
```

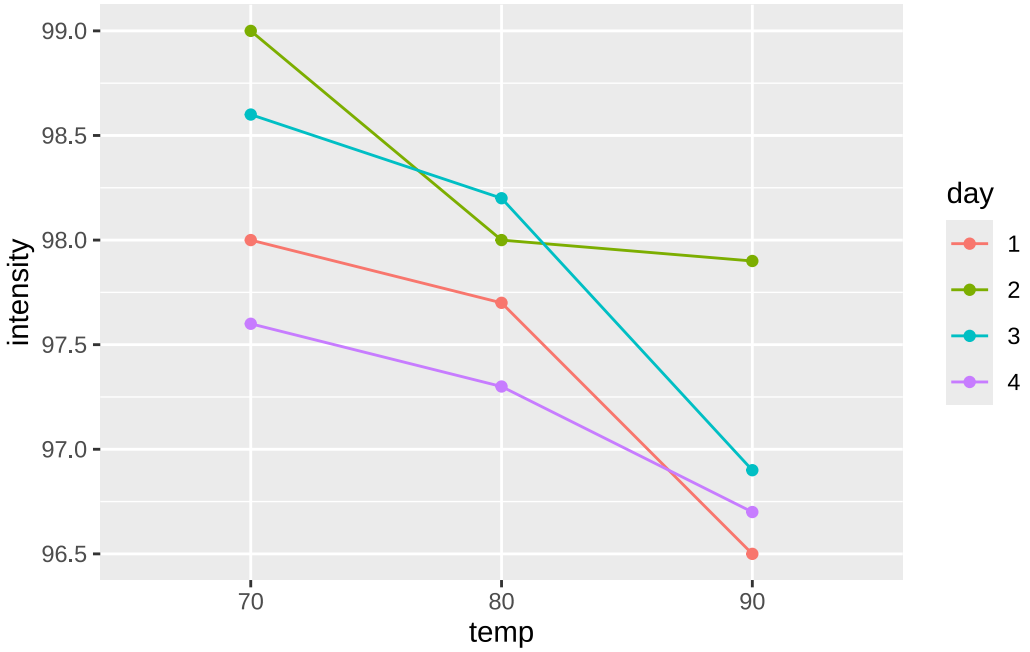
```
12       96.7   90   4
```

벡터를 범주형 변수로 만들어 줄때 두 함수 `as.factor()` 와 `factor()` 모두 사용 가능하다.

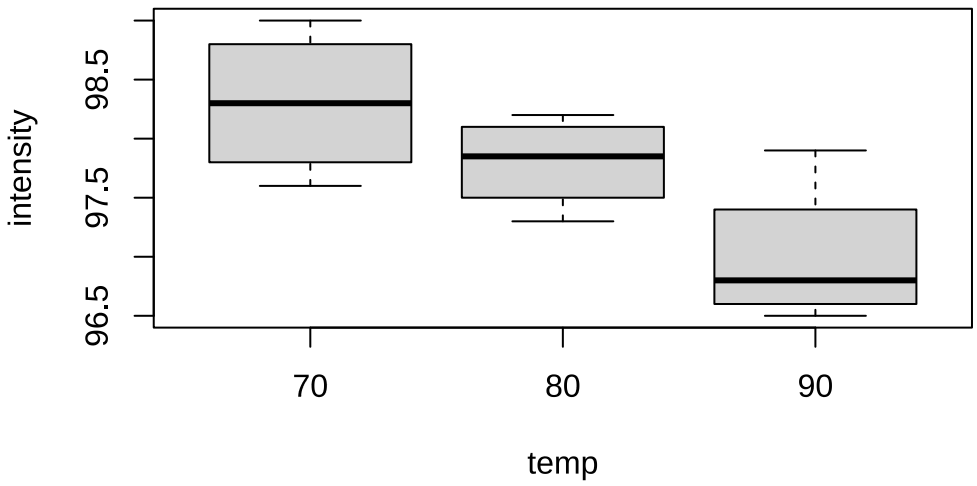
3.1.2. 시각적 분석

이제 온도의 수준에 따른 변화를 볼 수 있는 그림을 그려보자. 온도가 올라가면 강도가 떨어지는 경향을 볼 수 있다.

```
df %>%
  ggplot(aes(x = temp   , y = intensity,  color=day)) +
   geom_line(aes(group = day)) +   geom_point()
```

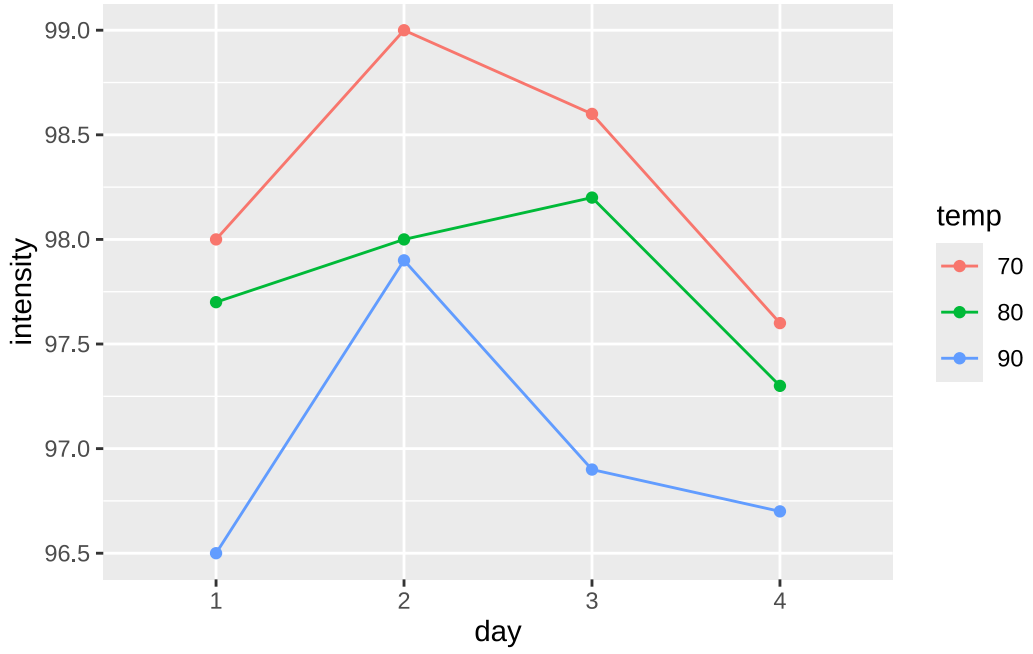



```
plot(intensity ~ temp, data=df)
```



이제 실험일에 따른 변동을 살펴보자. 실험일에 따라서 온도의 효과가 변하는 것을 볼 수 있다. 단 실험일과 온도의 상호작용은 크게 나타나지 않는다. 유의할 점은 반복이 없기 때문에 상호작용에 대한 추론은 불가능하다

```
df %>%
  ggplot(aes(x = day  , y = intensity,  color=temp)) +
   geom_line(aes(group = temp)) +   geom_point()
```



3.1.3. 분산분석

블럭 효과인 실험일(day)를 고정효과로 놓았을 경우 분산분석표는 다음과 같다.

$$\rho_j : \text{ fixed effect}, \quad e_{ij} \sim N(0, \sigma_E^2)$$

```
model<- aov(intensity ~ temp + day, data=df)
summary(model)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
temp         2   3.44  1.7200  18.429 0.00274 **
day          3   2.22  0.7400   7.929 0.01647 *
Residuals    6   0.56  0.0933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 분산분석표에서 온도의 효과를 검정하는 F-통계량의 값은 18.4285714 이고 p-값은 0.002744이다. 따라서 5% 유의수준으로 귀무가설을 기각하며 온도에 따라서 강도는 유의하게 다르다.

일반적으로 블럭효과에 대해서는 검정하지 않지만 그래도 p-값이 0.0164702 로서 매우 작으므로 실험일에 따른 변동이 크다는 것을 알 수 있다. 이는 실험을 수행하는 날에 따라서 관측값에 변동이 크다는 것이다. 단 상호작용이 그림으로 볼 때 나타나지 않기 때문에 온도의 효과는 적절하게 추정할 수 있다.

3.2. 혼합모형

고정효과와 임의효과(변량)가 동시에 모형식에 나타나는 모형을 혼합모형(mixed models)이라고 부른다. 교과서에서는 변량모형이라고 부른다. 혼합모형에 대한 자세한 기초이론은 부록 C 에서 찾아볼 수 있다.

- 혼합모형을 적합시키는 패키지는 `lme4` 이며 모형을 적합시키는 함수는 `lmer`이다.

```
library(lme4)
library(lmerTest)
```

- 혼합모형으로 부터 얻은 분산분석표에서 p-값을 보려면 패키지 `lmerTest`를 사용해야 한다.
- 함수 `lmer` 에서 고정효과에 대한 모형식은 함수 `anova`와 같다.
- 함수 `lmer` 에서 만약 변수 `var` 을 임의효과로 고려하려면 `(1|var)` 으로 쓰면 된다.

다음은 플라스틱 강도 자료 실험에서 **블럭 효과인 실험일(day, ρ)를 임의효과**로 놓았을 경우 분석결과이다. 즉

$$\rho_j \sim N(0, \sigma_B^2), \quad e_{ij} \sim N(0, \sigma_E^2)$$

```
fit <- lmer(intensity ~ temp + (1|day), data=df)
summary(fit)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: intensity ~ temp + (1 | day)
   Data: df

REML criterion at convergence: 14.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.0616 -0.7992  0.1430  0.5419  1.2297

Random effects:
 Groups   Name        Variance Std.Dev.
 day      (Intercept) 0.21556  0.4643
 Residual             0.09333  0.3055
Number of obs: 12, groups:  day, 4

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  98.3000     0.2779  4.5593 353.739  2.7e-11 ***
temp80       -0.5000     0.2160  6.0000  -2.315  0.05989 .
temp90       -1.3000     0.2160  6.0000  -6.018  0.00095 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
       (Intr) temp80
temp80 -0.389
temp90 -0.389  0.500
```

위의 결과에서 블럭효과(day) 를 나타내는 분산 성분 σ_B^2의 추정치는 0.2155556 이며 오차항(Residual)의 분산 σ_E^2의 추정치는 0.0933333 이다. 이는 급내상관 계수(ICC)는 0.6978417 로서 매우 크다는 것을 의미한다.

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_E^2} = 0.6978417$$

다음은 플라스틱 강도 자료 실험에서 블럭 효과를 임의효과로 놓았을 경우 분산분석표이다. 함수 lmer 에 의해 생성된 결과를 함수 anova에 적용하면 고정효과에 대한 분산분석과 F-검정만 보여준다. 앞에서 블럭을 고정효과로 놓았을 때 분산분석의 검정 결과와 같다.

```
anova(fit)
```

```
Type III Analysis of Variance Table with Satterthwaite's method
     Sum Sq Mean Sq NumDF DenDF F value   Pr(>F)
temp   3.44    1.72     2     6  18.429 0.002744 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3. 라틴정방설계

3.3.1. 로켓 추진체

다음은 교재 예제 5.2 - 로켓 추진체 실험을 분석하는 예제이다.

5가지의 로켓 추진체(A, B, C, D, E)의 성능을 비교하기 위하여 라틴정방계획을 사용한 실험이다.

- 행블럭: 5개의 연료 (R, ρ)
- 열블럭: 5명의 기사 (C, γ)
- 처리: 5가지의 로켓 추진체 (trt, τ)

[x_{ijk} = + _i + _j + k + e{ijk }]

3.3.2. 자료의 구성

예제 5.2에 있는 자료를 분석을 위하여 데이터프레임으로 만들어 보자.

```
trt <- c("A", "B", "C", "D", "E",
         "B", "C", "D", "E", "A",
         "C", "D", "E", "A", "B",
         "D", "E", "A", "B", "C",
         "E", "A", "B", "C", "D" )
trt <- factor(trt)
R <- factor(rep(1:5, each=5))
C <- factor(rep(1:5, times=5))
y <- c( -1,-5, -6, -1, -1,
        -8, -1, 5, 2, 11,
        -7, 13, 1, 2, -4,
        1, 6, 1, -2, -3,
        -3, 5, -5, 4, 6)
df<- data.frame(trt, R, C, y)
df
```

```
   trt R C  y
1    A 1 1 -1
2    B 1 2 -5
3    C 1 3 -6
4    D 1 4 -1
5    E 1 5 -1
6    B 2 1 -8
7    C 2 2 -1
8    D 2 3  5
9    E 2 4  2
10   A 2 5 11
11   C 3 1 -7
12   D 3 2 13
```

```
13   E 3 3  1
14   A 3 4  2
15   B 3 5 -4
16   D 4 1  1
17   E 4 2  6
18   A 4 3  1
19   B 4 4 -2
20   C 4 5 -3
21   E 5 1 -3
22   A 5 2  5
23   B 5 3 -5
24   C 5 4  4
25   D 5 5  6
```

함수 `xtabs()`는 모형식을 이용하여 다음과 같이 열과 행으로 구성된 자료를 보여줄 수 있다.

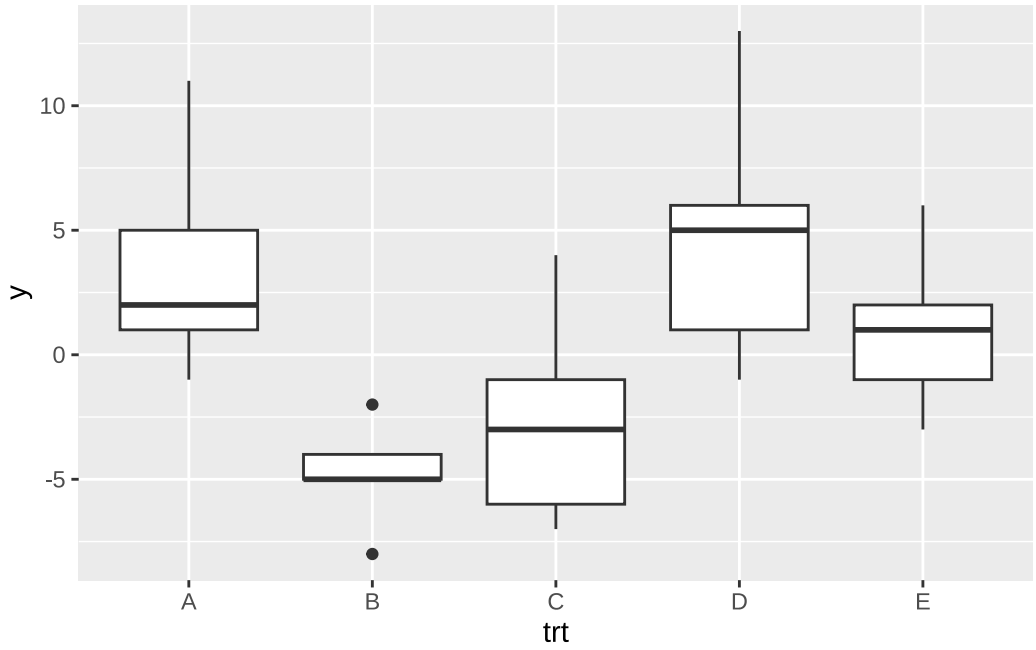
```
xtabs(y~ R + C, data = df)
```

```
   C
R    1  2  3  4  5
  1 -1 -5 -6 -1 -1
  2 -8 -1  5  2 11
  3 -7 13  1  2 -4
  4  1  6  1 -2 -3
  5 -3  5 -5  4  6
```

3.3.3. 시각적 분석

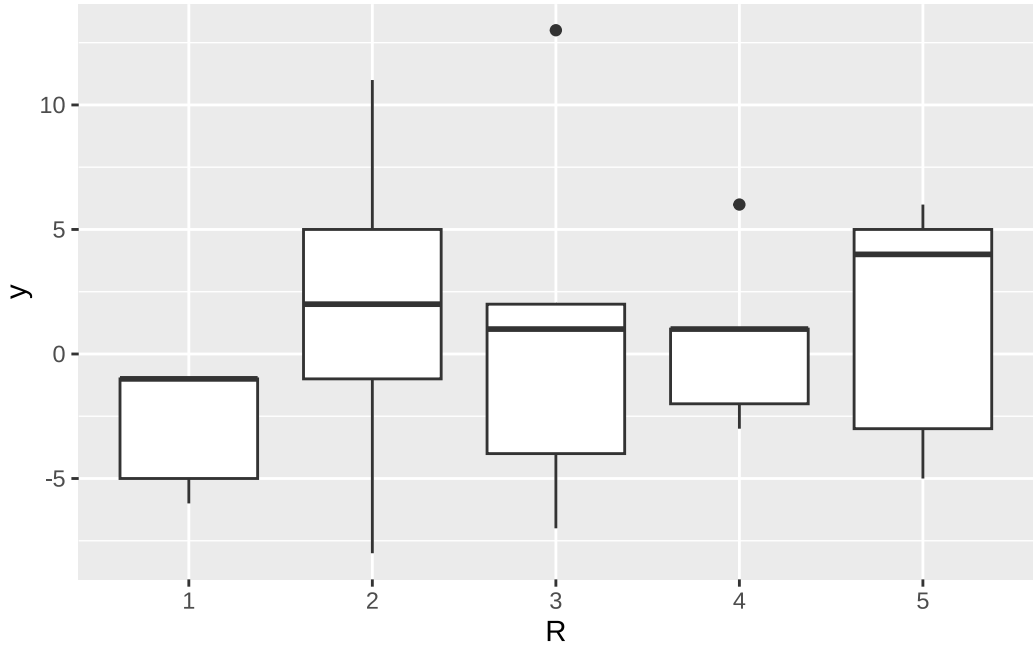
먼저 로켓 추진체, 즉 처리별로 자료의 분포를 보자. 추진체 B 와 C 가 다른 추진체들 보다 관측값이 작게 나오는 것을 알 수 있다.

```
df %>%
  ggplot() +
  aes(x = trt , y = y) +
  geom_boxplot()
```



원료(R) 뭉치별로 자료의 분포를 보면 큰 차이는 보이지 않는다.

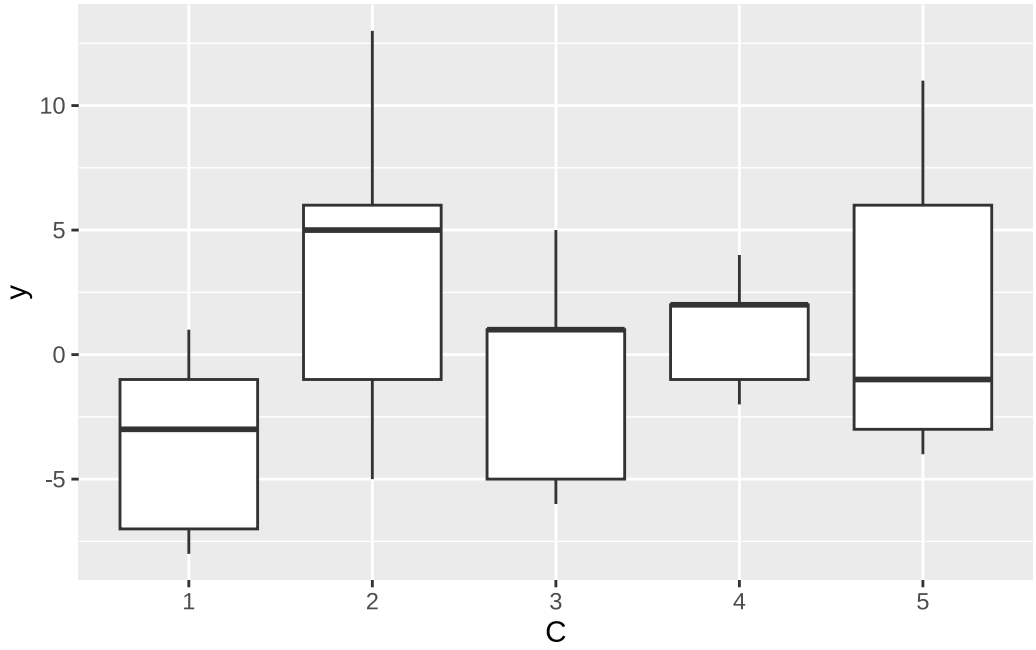
```
df %>%
  ggplot() +
  aes(x = R , y = y) +
  geom_boxplot()
```



기사(C) 별로 자료의 분포를 보면 약간의 차이가 보인다.

```
df %>%
  ggplot() +
```

```
  aes(x = C , y = y) +
  geom_boxplot()
```



3.3.4. 분산분석

이제 라틴정방계획법으로 얻은 자료에 대래 분산분석을 적용해 보자.

```
model<- aov(y ~ trt + R + C, data=df)
summary(model)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
trt          4    330   82.50   7.734 0.00254 **
R            4     68   17.00   1.594 0.23906
C            4    150   37.50   3.516 0.04037 *
Residuals   12    128   10.67
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 분산분석표에서 추진체(처리)의 효과를 검정하는 F-통계량의 값은 7.734375 이고 p-값은 0.0025365이다. 따라서 5% 유의수준으로 귀무가설을 기각하며 추진체에 따라서 성능이 유의하게 다르다.

3.3.5. 라틴정방의 구축

교과서 5.3절에서는 라틴정방 계획으로 실험을 하는 경우 처리를 랜덤하게 배정하는 방법을 설명하고 있다.

패키지 `agricolae` 에 포함된 함수 `design.lsd()`를 이용하면 다음과 같이 처리를 랜덤하게 배정해준다.

```
mytrt <- factor(c("A", "B", "C", "D", "E"))
mytrt
```

```
[1] A B C D E
Levels: A B C D E
```

```
design.lsd(mytrt)$sketch
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,] "B"  "A"  "D"  "C"  "E"
[2,] "A"  "E"  "C"  "B"  "D"
[3,] "D"  "C"  "A"  "E"  "B"
[4,] "E"  "D"  "B"  "A"  "C"
[5,] "C"  "B"  "E"  "D"  "A"
```

함수 `design.lsd()`는 실행할 때마다 랜덤하게 배정하기 때문에 기록을 위해서 랜덤 seed 를 지정하면 나중에도 동일한 계획을 얻을 수 있다.

```
design.lsd(mytrt, seed = 1234 )$sketch
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,] "C"  "B"  "E"  "A"  "D"
[2,] "A"  "E"  "C"  "D"  "B"
[3,] "B"  "A"  "D"  "E"  "C"
[4,] "D"  "C"  "A"  "B"  "E"
[5,] "E"  "D"  "B"  "C"  "A"
```

3.4. 처리 조합의 블럭

3.4.1. 화학약품의 생성률

다음은 교재 분할법 I - 예제 5.3 - 화학약품의 생성률 실험을 분석하는 예제이다.

이 실험에서는 화학약품의 생성률에 영향을 미치는 두 요인을 고려한 실험이다.

- 반응온도(`temp`, α) 3개의 수준
- 중간원료 제조회사 (`company`, β) 3개의 수준

이 실험에서는 9개의 처리를 먼저 랜덤하게 선택하고 선택된 처리 하에서 실험을 2번 반복하였다. 따라서 처리의 조합이 블럭효과(`block`, ρ)로 나타난다.

$$x_{ijk} = \mu + \alpha_i + \beta_j + \rho_{ij} + e_{2(ijk)}$$

위의 모형식에서 상호작용 효과 $(\alpha\beta)_{ij}$ 와 1차 랜덤화에 의한 오차 $e_{1(ij)}$ 는 교락되어 블럭효과 ρ_{ij}에 합쳐져서 나타난다.

$$\rho_{ij} = e_{1(ij)} + (\alpha\beta)_{ij}$$

이러한 경우 블럭효과 ρ_{ij}는 임의효과가 된다.

$$\rho_{ij} \sim N(0, \sigma_1^2), \quad e_{2(ijk)} \sim N(0, \sigma_2^2) \tag{3.1}$$

3.4.2. 자료의 구성

이제 실험자료를 입력하여 데이터프레임으로 만들어 보자

```
temp<- as.factor(rep(c("A1","A2", "A3"), each=2, times=3))
company<- as.factor(rep(c("B1", "B2", "B3"), each=6))

y <-c( 81.0, 80.2, 84.1, 83.2, 85.2, 86.1,
       83.3, 82.7, 86.2, 85.4, 86.6, 87.2,
       81.3, 81.9, 83.2, 84.2, 86.0, 86.4)

df<- data.frame(temp, company, y)
df
```

```
   temp company    y
1    A1      B1 81.0
2    A1      B1 80.2
3    A2      B1 84.1
4    A2      B1 83.2
5    A3      B1 85.2
6    A3      B1 86.1
7    A1      B2 83.3
8    A1      B2 82.7
9    A2      B2 86.2
10   A2      B2 85.4
11   A3      B2 86.6
12   A3      B2 87.2
13   A1      B3 81.3
14   A1      B3 81.9
15   A2      B3 83.2
16   A2      B3 84.2
17   A3      B3 86.0
18   A3      B3 86.4
```

3.4.3. 시각적 분석

일단 각 처리에 대한 관측값의 평균을 구해보자.

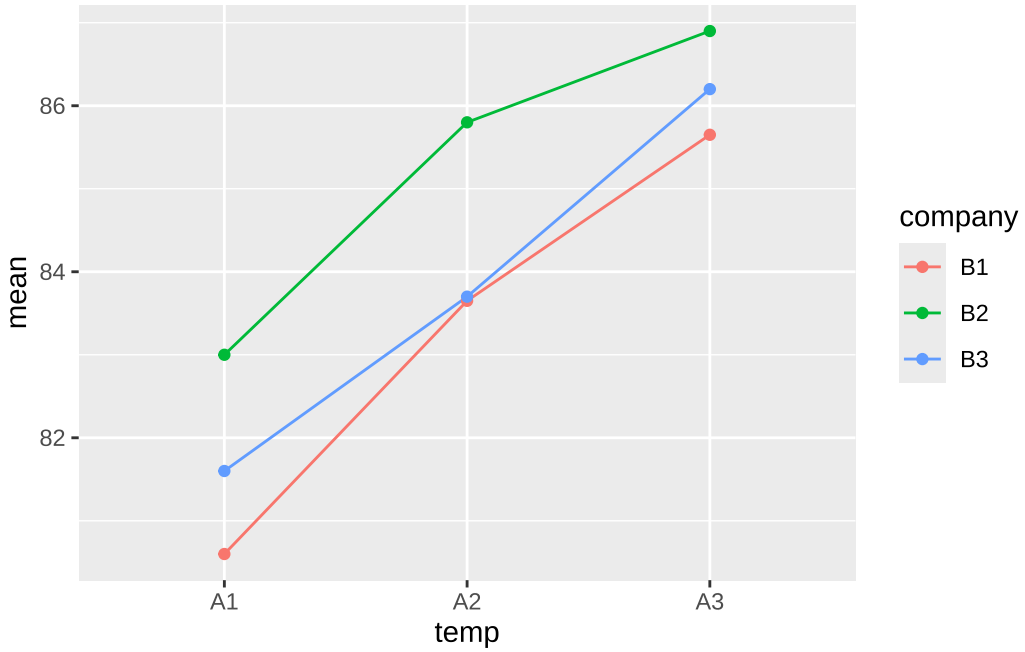
```
dfsum <- df %>% group_by(temp, company)  %>%  summarise(mean=mean(y),  sd=sd(y))
dfsum
```

```
# A tibble: 9 x 4
# Groups:   temp [3]
  temp  company  mean    sd
  <fct> <fct>   <dbl> <dbl>
1 A1    B1       80.6 0.566
2 A1    B2       83   0.424
3 A1    B3       81.6 0.424
4 A2    B1       83.6 0.636
5 A2    B2       85.8 0.566
6 A2    B3       83.7 0.707
7 A3    B1       85.6 0.636
8 A3    B2       86.9 0.424
9 A3    B3       86.2 0.283
```

이제 처리의 평균값을 가지고 온도에 따른 변화를 살펴보자. 이 경우 제조회사 원료에 대해서는 색깔을 다르게 하여 상호작용 효과도 볼 수 있다.

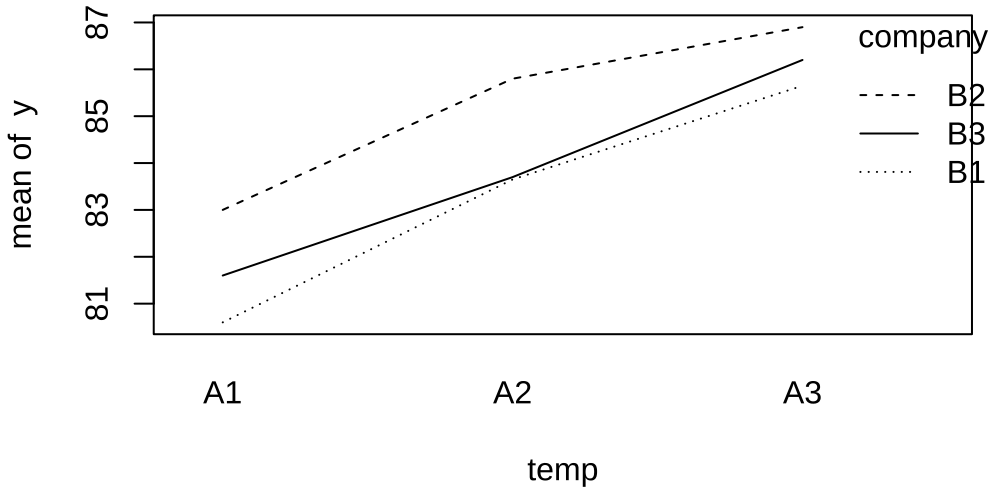
아래 상호작용 그림을 보면 온도에 따라서 화학약품의 생성률이 크게 변하는 것을 알 수 있다. 유의한 상호작용은 관측되지 않는다.

```
dfsum %>%
  ggplot(aes(x = temp  , y = mean,  color=company)) +
   geom_line(aes(group = company)) +   geom_point()
```



함수 `interaction.plot()`은상호작용 그림을 평균값을 계산하지 않고 원래 자료를 이용하여 다음과 같이 그릴 수 있다.

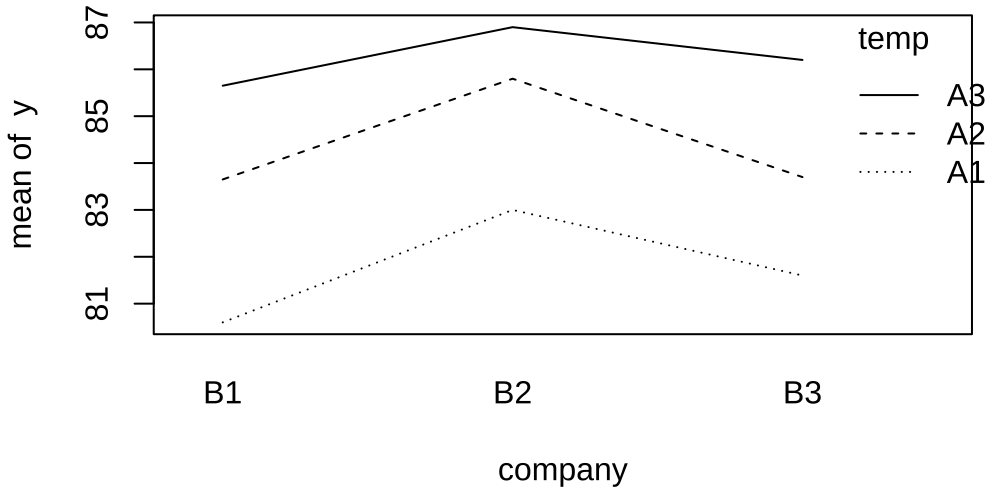
```
with(df, interaction.plot(x.factor = temp, trace.factor = company,  response = y))
```



> **노트**
>
> 위에서 함수 `with()` 은 이용하고자 하는 변수가 있는 데이터프레임을 지정하는데사용한다. 함수 `with()`의 첫 번째 인자는 앞의 예제와 같이 `df` 와 같은 데이터 프레임을 지정한다. 두 번째 인자에는 함수를 이용한 명령문을 넣어준다. 앞의 프로그램에서 함수 `interaction.plot()` 안에서 사용된 변수들(`temp`,`company`,`y`)들은 데이터프레임 `df`에 있는 변수들이다.

이제 제조회사에 따른 변화를 살펴보자. 제조회사에 따른 생성률의 변화는 크지 않다.

```
with(df, interaction.plot(x.factor = company, trace.factor =temp ,  response = y))
```



3.4.4. 분산분석

이제 분산분석을 하여 처리의 효과에 대한 검정을 해보자. 실험에서 각 처리의 조합을 블럭으로 해주어야 한다.

다음 anova 함수에서 두 처리의 조합을 temp:company 로 표시한다. 사실 temp:company는 두 처리 temp와 company의 상호작용(interaction)을 의미한다. 다음으로 처리의 조합 temp:company 이 임의효과라는 것을 Error(temp:company) 와 같이 지정해 준다.

```
model<- aov(y ~ temp + company + Error(temp:company), data=df)
summary(model)
```

```
Error: temp:company
          Df Sum Sq Mean Sq F value  Pr(>F)
temp       2  61.81  30.907   85.72 0.00052 ***
company    2  11.96   5.982   16.59 0.01157 *
Residuals  4   1.44   0.361
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9   2.57  0.2856
```

위의 분산분석표에서 온도의 효과를 검정하는 F-통계량의 값은 85.7211094 이고 p-값은 5.1981853×10^{-4} 이다. 따라서 5% 유의수준으로 귀무가설을 기각하며 온도에 따라서 생성률이 매우 유의하게 다르다.

온도의 효과를 검정하는 F-통계량의 값은 16.5916795 이고 p-값은 0.0115724이다. 따라서 5% 유의수준으로 귀무가설을 기각하며 원료 제조회사에 따라서도 생성률이 유의하게 다르다.

3.4.5. 블럭을 고려하지 않는 경우

만약에 처리 조합으로 생긴 블럭효과를 고려하지 않으면 어떤 일이 일어날까?

만약 생성률 실험자료를 **완전 랜덤화 이원배치법**에 의하여 얻은 자료라고 생각한다면 반복이 있으므로 상호작용 효과를 추론할 수 있다. 따라서 상호작용 효과를 고정효과로 놓고 분산분석을 적용할 것이다.

$$\rho_{ij} = (\alpha\beta)_{ij} : \text{ fixed effect }, \quad e_{2(ijk)} \sim N(0, \sigma_2^2) \tag{3.2}$$

아래 프로그램은 상호작용 효과를 고정효과로 생각한 것이다.

```
model2<- aov(y ~ temp + company + temp:company, data=df)
summary(model2)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
temp          2  61.81  30.907 108.235 5.07e-07 ***
company       2  11.96   5.982  20.949 0.000411 ***
temp:company  4   1.44   0.361   1.263 0.352665
Residuals     9   2.57   0.286
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

분산분석의 결과는 위와 같으며 온도와 제조회사에 대한 F-검정 통계량을 보면 임의효과 모형에서 나온 것보다 크다. 이는 F-검정 통계량을 만들 때 분모에 사용된 평균 오차제곱합 MS_E와 자유도가 달라서 나타나는 현상이다. 또한 자유도도

두 모형에서 온도에 대한 F-검정의 차이를 보자.

모형	anova 항	MS_A	MS_E	F_0
임의효과 모형 식 3.1	Error(temp:company)	30.9072222	0.3605556	85.7211094
고정효과 모형 식 3.2	temp:company	30.9072222	0.2855556	108.2354086

위의 표에서와 같이 실험계획에 따라서 나누어 주는 평균 오차제곱합 MS_E와 자유도가 다르기 때문에 검정의 결과가 다르게 나타난다.

> **노트**
>
> 실험계획에서 통계적 추론을 하는 경우 자료의 구조는 같아도 실험의 방법(랜덤화의 방법)이 다르면 가설검정의 방법이 다르다.
> 따라서 실험의 방법에 따른 적절한 통계적 추론 방법을 선택하는 것이 중요하다.

3.4.6. 혼합모형

처리들의 조합을 임의효과로 보는 모형 식 3.1 을 `lmer`로 적합시키는 프로그램은 다음과 같다.

분산분석 결과는 `anova()` 에서 임의효과 `Error(temp:company)`를 사용하는 결과와 동일하다.

```
fit <- lmer(y ~ temp + company + (1 | temp:company ), data = df)
summary(fit)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: y ~ temp + company + (1 | temp:company)
   Data: df

REML criterion at convergence: 29.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.52027 -0.46728 -0.07111  0.77604  1.20140
```

```
Random effects:
 Groups       Name        Variance Std.Dev.
 temp:company (Intercept) 0.0375   0.1936
 Residual                 0.2856   0.5344
Number of obs: 18, groups:  temp:company, 9

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  80.9111     0.3165  4.0000 255.666  1.4e-09 ***
tempA2        2.6500     0.3467  4.0000   7.644  0.00157 **
tempA3        4.5167     0.3467  4.0000  13.028  0.00020 ***
companyB2     1.9333     0.3467  4.0000   5.577  0.00507 **
companyB3     0.5333     0.3467  4.0000   1.538  0.19877
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
          (Intr) tempA2 tempA3 cmpnB2
tempA2    -0.548
tempA3    -0.548  0.500
companyB2 -0.548  0.000  0.000
companyB3 -0.548  0.000  0.000  0.500
```

```
anova(fit)
```

```
Type III Analysis of Variance Table with Satterthwaite's method
        Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
temp    48.956 24.4782     2     4  85.721 0.0005198 ***
company  9.476  4.7379     2     4  16.592 0.0115724 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.5. 분할법

3.5.1. 전자제품 수명

다음은 교재 분할법 II - 예제 5.4 - 전자제품 수명 실험을 분석하는 예제이다.

전자부품의 수명이 온도(580, 600, 620, 640도)와 시간(5, 10, 15분)에 의해 어떤 영향을 받는지에 대한 실험이다.

이 실험은 split-plot 설계를 적용하여 관측값을 얻었다. 온도를 먼저 랜덤하게 선택하고 선택된 온도에서 3개의 가열 시간에 대한 실험을 임의 순서로 진행하였다. 또한 각 실험은 3번 반복 하였다.

- 온도 (`temp`, α) : 주구, main plot - 1차 랜덤화 요인
- 시간 (`time`, β) : 분할구, split-plot, sub-plot - 2차 랜덤화 요인
- 반복 (`rep`, r) : 반복 요인

$$x_{ijk} = \mu + r_k + \alpha_i + \gamma_{ik} + \beta_j + (\alpha\beta)_{ij} + e_{2(ijk)} \tag{3.3}$$

위의 모형식에서 반복과 온도의 상호작용 효과 $(\alpha r)_{ik}$ 와 1차 랜덤화에 의한 오차 $e_{1(ik)}$ 는 교락되어 블럭효과 γ_{ik} 에 합쳐져서 나타난다.

$$\gamma_{ik} = (\alpha r)_{ik} + e_{1(ik)}$$

3.5.2. 자료의 구성

이제 실험자료를 입력하여 데이터프레임으로 만들어 보자

```
rep<- as.factor(rep(c(1:3), each=12))
temp<- as.factor(rep(c(580, 600, 620, 640), each=3, times=3))
time<- as.factor(rep(c(5, 10, 15), times=12))

y <-c(217, 233, 175, 158, 138, 152, 229, 186, 155, 223, 227, 156,
        188, 201, 195, 126, 130, 147, 160, 170, 161, 201, 181, 172,
        162, 170, 213, 122, 185, 180, 167, 181, 182, 182, 201, 199)

df <- data.frame(rep, temp, time, y)
```

함수 xtab 을 이용하면 반복에 따라서 자료 구조를 쉽게 볼 수 있다.

```
xtabs( y ~time + temp + rep, df)
```

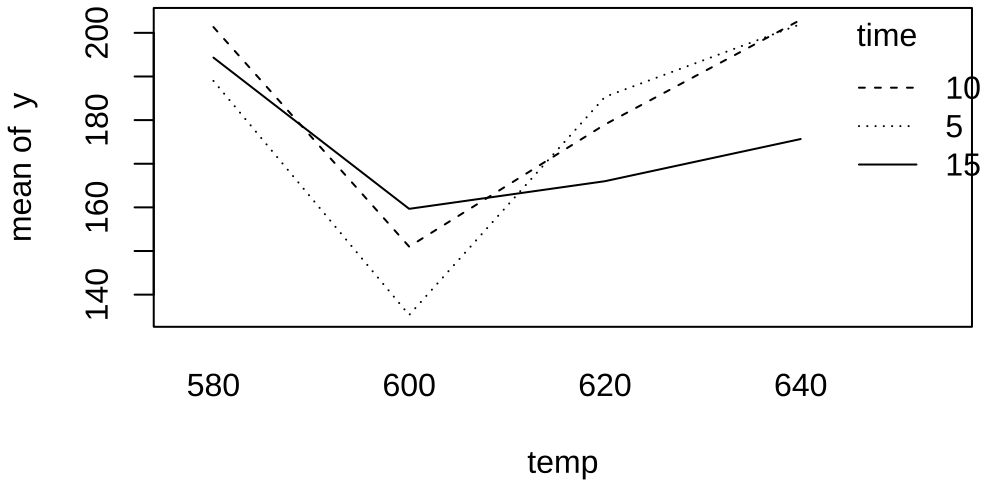
```
, , rep = 1

    temp
time 580 600 620 640
  5  217 158 229 223
  10 233 138 186 227
  15 175 152 155 156

, , rep = 2

    temp
time 580 600 620 640
  5  188 126 160 201
  10 201 130 170 181
  15 195 147 161 172

, , rep = 3

    temp
```

```
time 580 600 620 640
  5  162 122 167 182
  10 170 185 181 201
  15 213 180 182 199
```

3.5.3. 시각적 분석

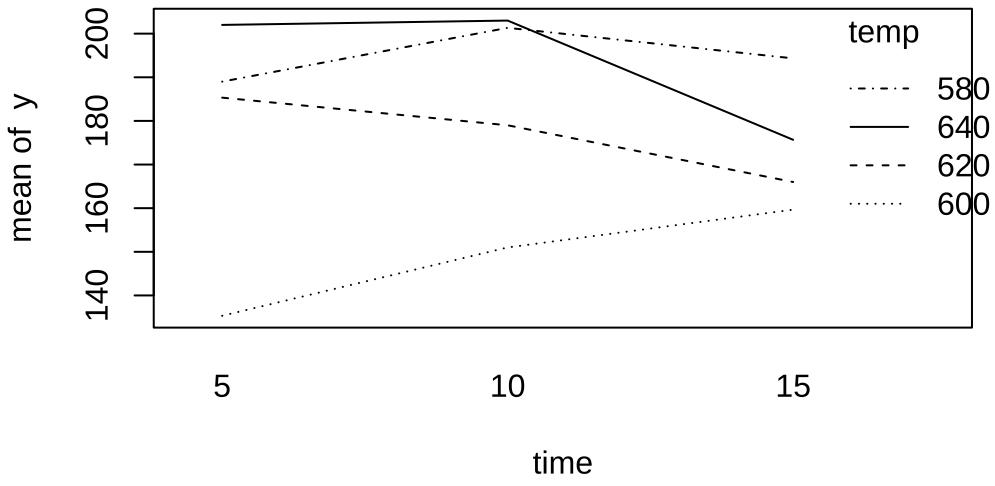
이제 온도의 수준에 따른 변화를 볼 수 있는 그림을 그려보자. 온도가 증가하면서 수명이 줄어들었다가 다시 늘어나는 현상을 볼 수 있다.

```
with(df, interaction.plot(x.factor = temp, trace.factor = time, response = y))
```



가열시간의 수준에 따른 변화를 볼 수 있는 그림을 그려보자. 가열시간이 증가하더러도 수명이 크게 변하지 않는 것을 알 수 있다.

```
with(df, interaction.plot(x.factor = time, trace.factor = temp, response = y))
```



3.5.4. 분산분석

이제 모형식 식 3.3 에 대한 분산분석을 실시해 보자.

여기서 유의할 점은 모형식 식 3.3 에서 블럭효과 γ_{ik} 는 임의효과로 생각하며 반복 수준과 온도 수준의 조합이다. 따라서 블럭효과 γ_{ik} 에 대한 항을 `Error(rep:temp)` 로 사용한다.

$$\gamma_{ik} \sim N(0, \sigma_1^2), \quad e_{2(ijk)} \sim N(0, \sigma_E^2)$$

```
model<- aov(y ~ rep + temp*time + Error(rep:temp), data=df)
summary(model)
```

```
Error: rep:temp
          Df Sum Sq Mean Sq F value Pr(>F)
rep        2   1963     981   3.319  0.107
temp       3  12494    4165  14.086  0.004 **
Residuals  6   1774     296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
time       2    566   283.1   0.456  0.642
temp:time  6   2600   433.4   0.698  0.655
Residuals 16   9933   620.8
```

분산분석표에서 온도의 효과를 검정하는 F-통계량의 값은 14.0864677 이고 p-값은 0.0040028이다. 따라서 5% 유의수준으로 귀무가설을 기각하며 온도에 따라서 제품의 수명이 유의하게 다르다.

가열시간의 효과를 검정하는 F-통계량의 값은 0.4560179 이고 p-값은 0.6417897이다. 따라서 5% 유의수준으로 귀무가설을 기각할 수 없으며 가열시간에 따라서 제품의 수명이 다르지 않다.

온도와 가열시간의 상호작용 효과를 검정하는 F-통계량의 값은 0.6981059 이고 p-값은 0.655133이다. 따라서 5% 유의수준으로 귀무가설을 기각할 수 없으며 상호작용은 유의하지 않다.

4. 대비

4.1. 카이제곱 분포

자유도가 1인 카이제곱 분포 (χ^2-distribution)은 평균이 0이고 분산이 1인 정규분포(표준 정규분포)를 따르는 확률변수의 제곱이 따르는 분포이다.

$$z \sim N(0,1) \quad \rightarrow \quad z^2 \sim \chi^2(1) \tag{4.1}$$

만약 k 개의 확률변수 $z_1, z_2, \dots, z_k$가 서로 독립이고 각각 표준 정규분포 $N(0,1)$를 따른다면 확률변수의 제곱들의 합은 자유도가 k인 카이제곱 분포 $\chi^2(k)$ 를 따른다.

$$z_1, z_2, \dots, z_k \sim_{ind} N(0,1) \quad \rightarrow \quad \sum_{i=1}^{k} z_i^2 \sim \chi^2(k) \tag{4.2}$$

만약 k 개의 확률변수 $z_1, z_2, \dots, z_k$가 서로 독립이고 각각 **정규분포** $N(\mu_i, \sigma^2)$를 따른다면 표준화 확률변수의 제곱들의 합은 자유도가 k인 카이제곱 분포 $\chi^2(k)$ 를 따른다.

$$\sum_{i=1}^{k} \left[\frac{z_i - \mu_i}{\sigma}\right]^2 \sim \chi^2(k) \tag{4.3}$$

4.2. 대비

4.2.1. 대비의 정의

다음과 같은 균형자료를 가지는 일원배치 모형을 고려하자.

$$x_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2, \dots, a, \;\; j = 1, 2, \dots, r \tag{4.4}$$

위의 일원배치 모형 식 4.4 에서 오차항 e_{ij}가 정규분포 $N(0, \sigma^2)$ 을 따른다고 가정하자.

지금까지 우리는 다음과 같은 가설검정에 대한 통계적 추론을 배웠다.

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a \quad \text{vs.} \quad H_1 : \text{ not } H_0$$

$$H_0 : \alpha_i - \alpha_j = 0 \quad \text{vs.} \quad H_1 : \alpha_i - \alpha_j \neq 0 \tag{4.5}$$

위의 첫 번째 가설은 요인 A의 효과가 있는지에 대한 검정이며 분산분석표를 이용한 F-통계량으로 검정한다. 두 번째 가설은 각 처리 수준의 차이에 대한 검정이며 평균의 차이를 이용한 t-통계량으로 검정한다.

이제 조금 더 복잡한 가설검정을 고려해보자.

만약 수준이 3개인 경우($a = 3$) 첫 번째 수준과 두 번째 수준의 평균이 세 번째 수준과 같은지 검정하고 싶다고 하자 [교과서 예제 7.1 (2)]

$$H_0 : \frac{\alpha_1 + \alpha_2}{2} = \alpha_3 \quad \text{vs.} \quad H_1 : \text{not } H_0 \tag{4.6}$$

또는 만약 요인이 온도인 경우 3개의 수준을 각각 100, 110, 120도로 같은 간격으로 증가시켰다. 반응변수의 평균이 일차적인 추세(linear trend)를 보이고 변화하는지 검정하고 싶을 경우가 있다. 또는 반응변수의 평균이 이차적인 추세(quadratic trend)를 가지는지 확인하고 싶은 경우도 있을 것이다.

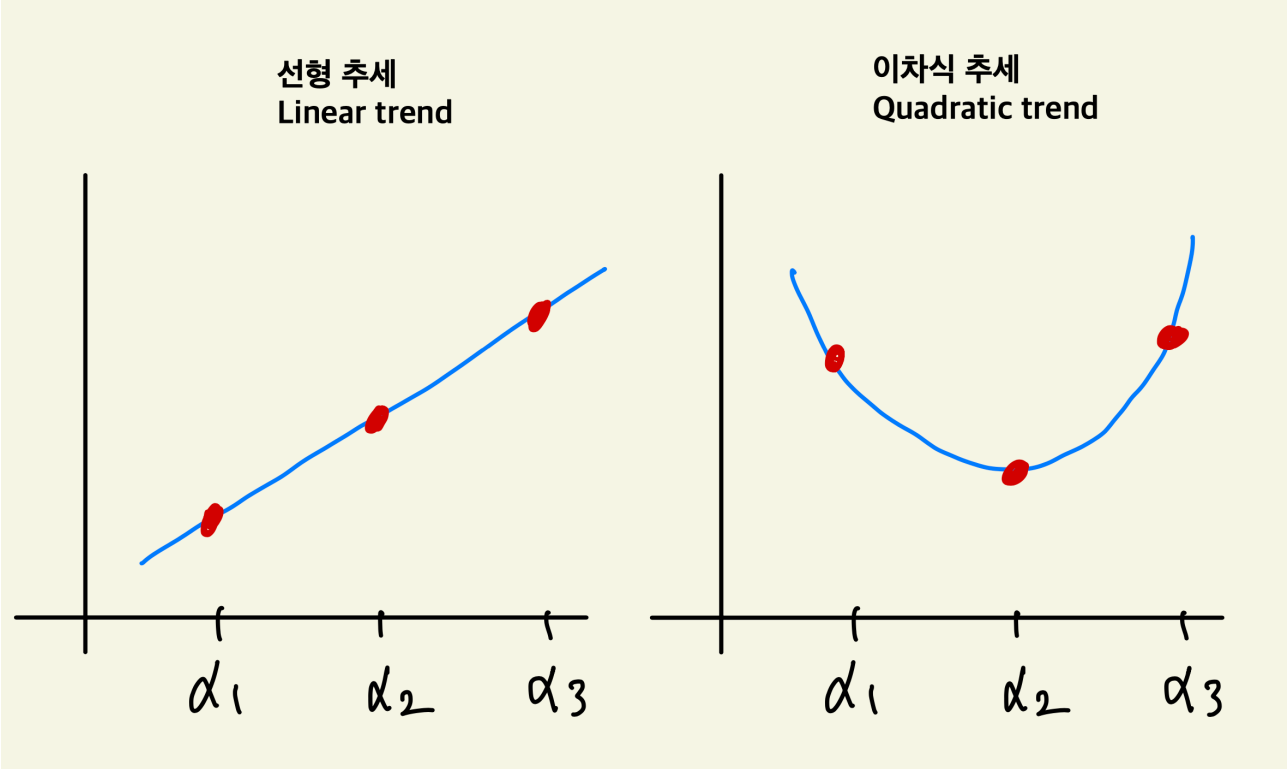


그림 4.1.: 3개의 수준이 있는 경우 선형 추세와 이차식 추세

만약 반응변수의 평균이 일차적인 추세(linear trend)를 보이면 수준의 순서에 따라서 평균이 일차적으로 증가 또는 감소하므로 두 평균의 변화를 합한 값, 즉 $(\alpha_2 - \alpha_1) + (\alpha_3 - \alpha_2)$ 이 0 과 차이가 날 것이다.

$$|(\alpha_2 - \alpha_1) + (\alpha_3 - \alpha_2)| = |\alpha_1 - \alpha_3| > 0$$

반면 반응변수의 평균이 이차적인 추세(quadratic trend)를 보이면 수준의 순서에 따라서 평균이 감소했다 증가하거나 또는 감소했다가 증가할 것이므로 두 평균의 변화를 뺀 값, 즉 $(\alpha_2 - \alpha_1) - (\alpha_3 - \alpha_2)$ 이 0 과 차이가 날 것이다.

$$|(\alpha_2 - \alpha_1) - (\alpha_3 - \alpha_2)| = |\alpha_1 - 2\alpha_2 + \alpha_3| > 0$$

따라서 선형식 $\psi = \sum_i c_i \alpha_i$를 다음과 같이 정의하면 선형식 ψ의 추정량 $\hat{\psi}$의 값이 클수록 반응변수의 평균이 선형적으로 변화하는 증거가 커진다.

$$\psi = (\alpha_2 - \alpha_1) + (\alpha_3 - \alpha_2) = -\alpha_1 + \alpha_3$$

따라서 가설검정을 다음과 같이 세운다. 귀무 가설을 기각하면 평균이 선형적으로 변화한다는 것을 의미한다.

$$H_0 : \alpha_1 - \alpha_3 = 0 \quad \text{equvalently} \quad H_0 : \text{ not } H_0 \tag{4.7}$$

위에서 제시한 2개의 가설 식 4.6 과 식 4.7 의 경우는 모수들의 특별한 선형조합으로 표시되는 가설이다. 이러한 모수들의 선형 조합으로 표시된 가설은 다음과 같이 일반적으로 나타낼 수 있다.

$$H_0 : \sum_{i=1}^{a} c_i \alpha_i = 0 \tag{4.8}$$

위의 가설 식 4.6 은 일반적인 가설 식 4.8 에서 계수 c_i들이 다음과 같은 경우이며

$$c_1 = c_2 = 1/2, \;\; c_3 = -1$$

가설 식 4.7 은 일반적인 가설 식 4.8 에서 계수 c_i들이 다음과 같은 경우이며

$$c_1 = -1, \; c_2 = 0, \; c_3 = 1$$

물론 두 개의 처리수준을 비교하는 가설 식 4.5 도 일반적인 가설의 범주에 속한다. 이 경우 $c_i = 1$, $c_j = -1$이고 나머지 $c_l = 0$인 경우이다.

이렇게 관심있는 모수들의 선형조합 ψ을 **선형식(linear combination)**이라고 부른다.

$$\psi = c_1 \alpha_1 + c_2 \alpha_2 + \cdots + c_a \alpha_a \tag{4.9}$$

선형식 식 4.9 에서 주어진 **계수들의 합이 0 인 선형식을 특별히 대비(contrast)**라고 한다.

$$C = c_1 \alpha_1 + c_2 \alpha_2 + \cdots + c_a \alpha_a, \quad \sum_{i=1}^{a} c_i = 0 \tag{4.10}$$

이러한 대비는 계수의 합이 0 이므로 각 처리 효과들을 다양하게 비교하는데 사용될 수 있다. 가설 식 4.5 , 식 4.6 , 식 4.7 에 나타난 계수들 c_i 들은 모두 더하면 0이므로 대비라고 부른다.

4.2.2. 추론

만약 선형식으로 주어진 가설 식 4.8 를 검정하려면 다음과 같이 각 처리의 표본 평균들의 조합으로 이루어진 통계량을 사용할 수 있다.

$$\hat{\psi} = L_* = c_1\bar{x}_{1.} + c_2\bar{x}_{2.} + \cdots + c_a\bar{x}_{a.} \tag{4.11}$$

이제 식 식 4.11 의 선형 추정량 L_* 의 평균과 분산을 구해보자

$$\begin{aligned} E(L_*) &= E(c_1\bar{x}_{1.} + c_2\bar{x}_{2.} + \cdots + c_a\bar{x}_{a.}) \\ &= c_1E(\bar{x}_{1.}) + c_2E(\bar{x}_{2.}) + \cdots + c_aE(\bar{x}_{a.}) \\ &= c_1(\mu+\alpha_1) + c_2(\mu+\alpha_2) + \cdots + c_a(\mu+\alpha_a) \\ &= \mu\sum_{i=1}^{a} c_i + c_1\alpha_1 + c_2\alpha_2 + \cdots + c_a\alpha_a \\ &= c_1\alpha_1 + c_2\alpha_2 + \cdots + c_a\alpha_a \end{aligned}$$

위의 유도에서 마지막 결과는 대비는 계수들의 합이 0 이라는 정의($\sum_{i=1}^{a} c_i = 0$)를 이용한 것이다.

$$\begin{aligned} V(L_*) &= V(c_1\bar{x}_{1.} + c_2\bar{x}_{2.} + \cdots + c_a\bar{x}_{a.}) \\ &= c_1^2V(\bar{x}_{1.}) + c_2^2V(\bar{x}_{2.}) + \cdots + c_a^2V(\bar{x}_{a.}) \\ &= c_1^2\frac{\sigma^2}{r} + c_2^2\frac{\sigma^2}{r} + \cdots + c_a^2\frac{\sigma^2}{r} \\ &= \frac{\sigma^2}{r}\sum_{i=1}^{a} c_i^2 \end{aligned}$$

이제 일원배치 모형 식 4.4 에서 오차항 e_{ij}가 정규분포 를 따른 다고 가정하였으므로 관측값들의 선형조합은 정규분포를 따른다. 따라서 식 식 4.11 의 선형 추정량 L_*은 다음 같이 정규분포를 따른다.

$$L_* \sim N\left(\sum_i c_i\alpha_i, \frac{\sigma^2}{r}\sum_{i=1}^{a} c_i^2\right) \tag{4.12}$$

만약 선형식에 대한 가설 식 4.8 이 참이라면 선형 추정량 L_*의 평균은 0이 되고

$$L_* \sim N\left(0, \frac{\sigma^2}{r}\sum_{i=1}^{a} c_i^2\right) \quad \text{under } H_0 : \sum_i c_i\alpha_i = 0 \tag{4.13}$$

위에서 배운 카이제곱 분포에 대한 결과 식 4.3 에 따라서 다음과 같은 결과를 얻는다.

$$\frac{L_*^2}{\frac{\sigma^2}{r}\sum_{i=1}^{a} c_i^2} \sim \chi^2(1) \quad \text{under } H_0 : \sum_i c_i\alpha_i = 0 \tag{4.14}$$

분산분석에서 구한 오차제곱합 SSE는 또한 다음과 같이 자유도가 $n-a$인 카이제곱 분포를 따르며 선형식 L_*과 독립이다 (독립의 이유에 대한 설명은 본 강의의 수준을 넘어서므로 생략한다.)

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-a) \tag{4.15}$$

따라서 분포에 대한 두개의 결과 식 4.14 과 식 4.15 를 이용해 보자. 귀무가설 식 4.8 이 참인 경우 카이제곱 분포를 따르는 서로 독립인 독립변수의 비는 F-분포를 따른다.

$$F = \frac{\left[L_*^2/\frac{\sigma^2}{r}\sum_{i=1}^a c_i^2\right]/1}{\left[SSE/\sigma^2\right]/(n-a)} = \left[\frac{L_*^2}{\sum_{i=1}^a c_i^2/r}\right]/MSE \sim F(1, n-a) \quad \text{under } H_0 \tag{4.16}$$

따라서 대비로 표현되는 산형식에 대한 가설검정 식 4.8 은 식 식 4.16 의 F 통계량이 자유도 $(1, n-a)$를 가진 F-분포의 상위 5% 백분위수보다 크면 귀무가설을 기각한다.

4.2.3. 표본합

위 식 4.11 의 선형 추정량 L_* 은 각 처리에 대한 자료 합 $T_1, T_2, \ldots, T_a$로 다음과 같이 나타낼 수 있다. 교과서 7.1 식은 선형 추정량 L 을 평균이 아닌 합으로 표시하고 있다.

각 처리에 대한 평균은 $\bar{x}_{i.} = T_i/r$이므로

$$\begin{aligned} L &= c_1T_1 + c_2T_2 + \cdots + c_aT_a \\ &= r(c_1\bar{x}_{1.} + c_2\bar{x}_{2.} + \cdots + c_a\bar{x}_{a.}) \\ &= rL_* \end{aligned}$$

따라서 합으로 표시한 선형식 L의 평균과 분산은 다음과 같다

$$\begin{aligned} E(L) &= E(rL) \\ &= rE(L) \\ &= r\sum_{i=1}^a c_i\alpha_i \\ V(L) &= V(rL_*) \\ &= r^2V(L_*) \\ &= r\sigma^2\sum_{i=1}^a c_i^2 \end{aligned}$$

카이제곱 분포에 대한 결과 식 4.3 에 따라서 다음과 같은 결과를 얻는다.

$$\frac{L^2}{r\sigma^2\sum_{i=1}^a c_i^2} \sim \chi^2(1) \quad \text{under } H_0 : \sum_i c_i\alpha_i = 0 \tag{4.17}$$

여기서 유의할 점은 평균으로 구성된 선형식과 합으로 구성된 선형식으로 유도된 카이제곱 통계량은 동일하다. 따라서 대비에 나타나는 계수들에 상수를 곱해줘도 검정통계량의 변화는 없다.

$$\frac{L^2}{r\sigma^2\sum_{i=1}^a c_i^2} = \frac{r^2L_*^2}{r\sigma^2\sum_{i=1}^a c_i^2} = \frac{L_*^2}{\frac{\sigma^2}{r}\sum_{i=1}^a c_i^2}$$

이제 평균의 선형식과 같은 방법으로 합으로 표시된 선형식 L에 대하여 다음과 같은 결과를 구할 수 있다.

$$\begin{aligned} F &= \frac{\left[L^2/r\sigma^2 \sum_{i=1}^{a} c_i^2\right]/1}{\left[SSE/\sigma^2\right]/(n-a)} \\ &= \left[\frac{L^2}{r\sum_{i=1}^{a} c_i^2}\right]/MSE \\ &= SS_L/MSE \sim F(1, n-a) \quad \text{under } H_0 \end{aligned}$$

위에서 SS_L은 교과서 7.2 식에서 정의된 통계량과 같다.

4.3. 직교 대비

4.3.1. 직교 대비의 정의

다음과 같이 처리그룹의 합으로 표시된 2개의 서로 다른 대비 C_1과 C_2를 고려하자.

$$C_1 = c_1T_1 + c_2T_2 + \cdots + c_aT_a, \quad \sum_{i=1}^{a} c_i = 0 \tag{4.18}$$

$$C_2 = d_1T_1 + d_2T_2 + \cdots + d_aT_a, \quad \sum_{i=1}^{a} d_i = 0 \tag{4.19}$$

서로 다른 두 대비에서 계수들의 내적이 0 이 되는 경우 **두 대비가 직교(orthogonal)**한다고 말한다.

$$\sum_{i=1}^{a} c_i d_i = 0 \quad \rightarrow \text{ orthogonal constrast} \tag{4.20}$$

대비가 서로 직교하면 그에 따른 두 제곱합 SS_{C_1}과 SS_{C_2}는 서로 독립이다.

$$SS_{C_1} = \frac{C_1^2}{r\sum_i c_i^2} \quad \sim_{indep.} \quad SS_{C_2} = \frac{C_2^2}{r\sum_i d_i^2}$$

따라서 앞 절에서 배운 각 대비에 대한 가설을 검정할 수 있는 F-통계량도 독립이다.

$$F_1 = \frac{SS_{C_1}}{MSE} \quad \sim_{indep.} \quad F_2 = \frac{SS_{C_2}}{MSE}$$

4.3.2. 처리 제곱합의 분해

만약 요인 A가 a개의 수준을 가지면 이 요인에 대한 직교하는 대비를 $a-1$개 만들 수 있다. 주의할 점은 직교하는 대비들는 유일하지 않다.

또한 각 대비 C_i에 대한 제곱합은 자유도가 1인 카이제곱 분포를 따르며 서로 독립이다. 더 나아가 분산분석에서 요인 A에 대한 처리제곱합 SS_A가 다은과 같이 분해된다.

$$SS_A = SS_{C_1} + SS_{C_2} + SS_{C_3} + \cdots + SS_{C_{a-1}}$$

4.3.3. 대표적인 대비

4.3.3.1. 다항 대비

직교하는 대비들 중에 대표적인 예로 다항 대비(polynomial contrasts)가 있다. 다항 대비는 처리 수준의 간격이 일정한 경우 평균의 변화가 선형(linear)인지, 이차적(quadratic)인지, 더 나아가 k차 다항식의 변화를 가지는지 검정할 수 있다.

다항대비의 계수들은 검정하고자 하는 변화의 추세가 강할 수록 대응하는 제곱합이 크게 되도록 설계되어 있다. 따라서 귀무가설에 대한 p-값이 크면 변화의 추세가 강하게 나타난다고 말할 수 있다.

예를 들어 3개의 수준에서 다음과 같이 2개의 다항 대비를 구할 수 있다. 아래 R 출력에 나오는 행렬의 각 열이 서로 직교하는 대비이다. 대비들의 계수의 제곱의 합이 1이 되도록($\sum_i c_i^2 = 1$) 정규화한 결과이다.

```
contr.poly(3)
```

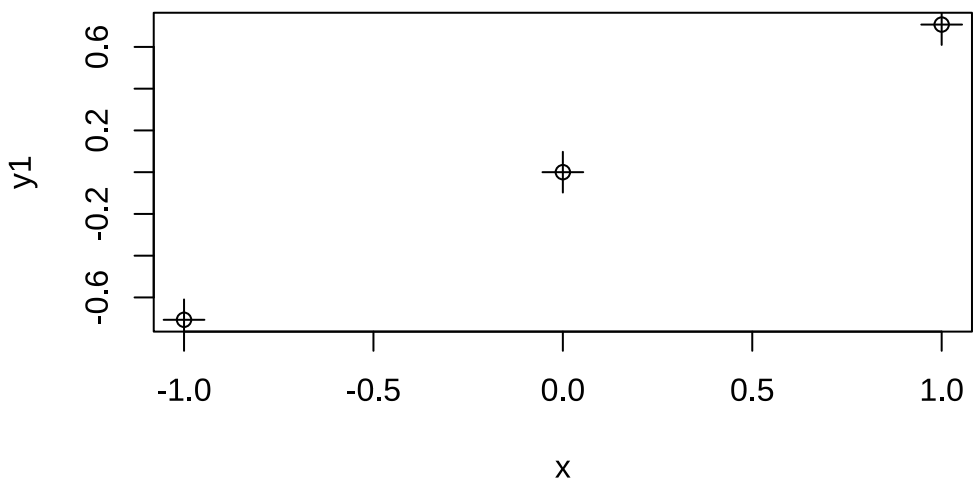
```
                .L         .Q
[1,] -7.071068e-01  0.4082483
[2,] -9.073800e-17 -0.8164966
[3,]  7.071068e-01  0.4082483
```

첫 번째 열이 선형 대비(linear contrast)로 계수는 다음과 같다.

$$c_1 = -\frac{1}{\sqrt{2}}, \quad c_2 = 0 \quad c_3 = \frac{1}{\sqrt{2}}$$

선형 대비를 그림으로 그려보면 다음과 같다.

```
x <- c(-1,0,1)
y1 <- contr.poly(3)[,1]
plot(x,y1 )
points(x,y1,cex=2, pch =3)
```

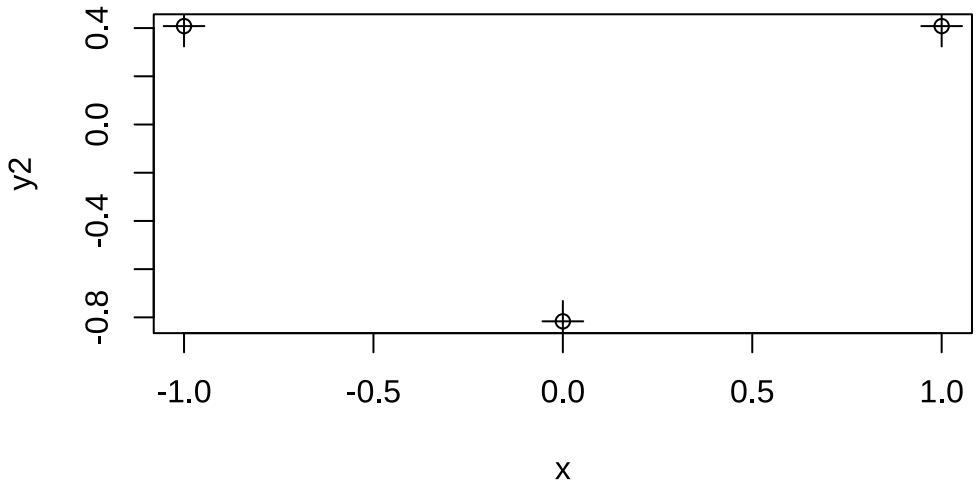


두 번째 열은 이차 대비(quadratic contrast)로 계수는 다음과 같다.

$$c_1 = \frac{1}{\sqrt{6}}, \quad c_2 = -\frac{2}{\sqrt{6}} \quad c_3 = \frac{1}{\sqrt{6}}$$

이차 대비를 그림으로 그려보면 다음과 같다.

```
y2 <- contr.poly(3)[,2]
plot(x,y2 )
points(x,y2,cex=2, pch =3)
```



다음과 같이 수준의 개수가 5인 경우 4차 다항대비를 구해준다. 함수 `contr.poly(k)`는 $k-1$차 다항 대비까지 구해준다.

```
contr.poly(5)
```

```
               .L         .Q            .C         ^4
[1,] -6.324555e-01  0.5345225 -3.162278e-01  0.1195229
[2,] -3.162278e-01 -0.2672612  6.324555e-01 -0.4780914
[3,] -3.288380e-17 -0.5345225  9.637305e-17  0.7171372
[4,]  3.162278e-01 -0.2672612 -6.324555e-01 -0.4780914
[5,]  6.324555e-01  0.5345225  3.162278e-01  0.1195229
```

4.4. 교과서 예제 7.1

교과서 예제 7.1 에서 제조회사에 대한 비교를 하는 경우 요인의 개수가 3개이므로 2개의 직교 대비를 이용한다.

예제 7.1 에서 제조회사에 대한 비교를 하는 경우 이용한 대비의 계수는 다음과 같다.

$$c_1 = 1, \;\; c_2 = 1, \;\; c_3 = -2$$

$$d_1 = 1, \;\; d_2 = -1, \;\; d_3 = 0$$

4.4.1. 이원배치 자료

예제 7.1 원료의 제조회사 A (A_0:자회사, A_1:국내 타회사, A_2:외국회사) 와 성형온도 B(B_0: 100℃, B_1: 110℃, B_2: 120℃)가 플라스틱강도에 미치는 영향은?

< 플라스틱 제품의 강도 >

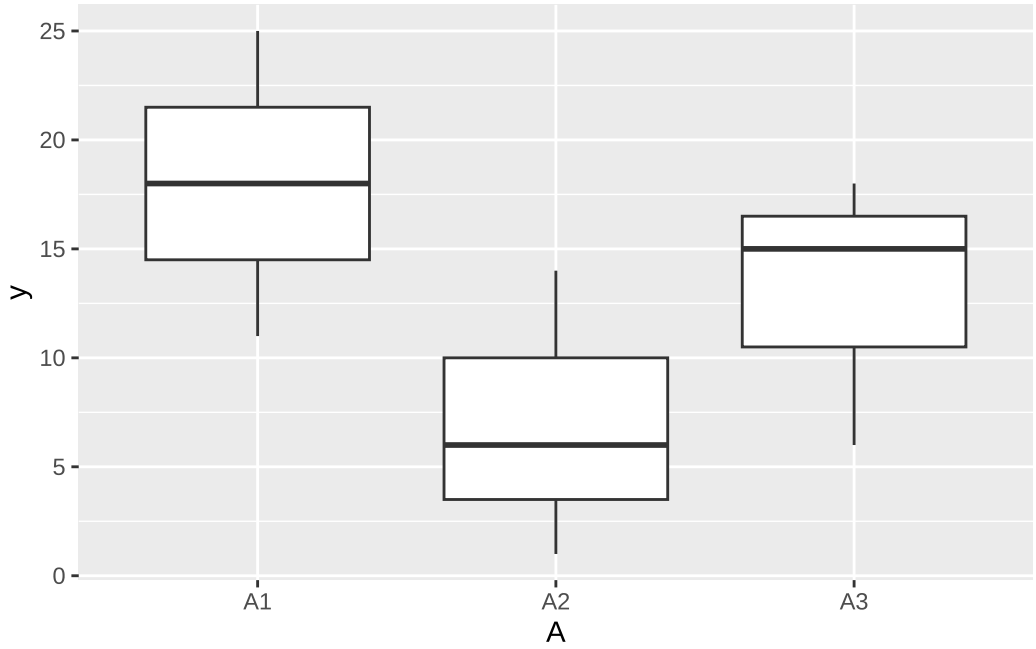
A \ B	B_0	B_1	B_2	$T_{i.}$
A_0	11	18	25	54
A_1	1	6	14	21
A_2	6	15	18	39
$T_{.j}$	18	39	57	T=114

그림 4.2.: 이원배치 자료

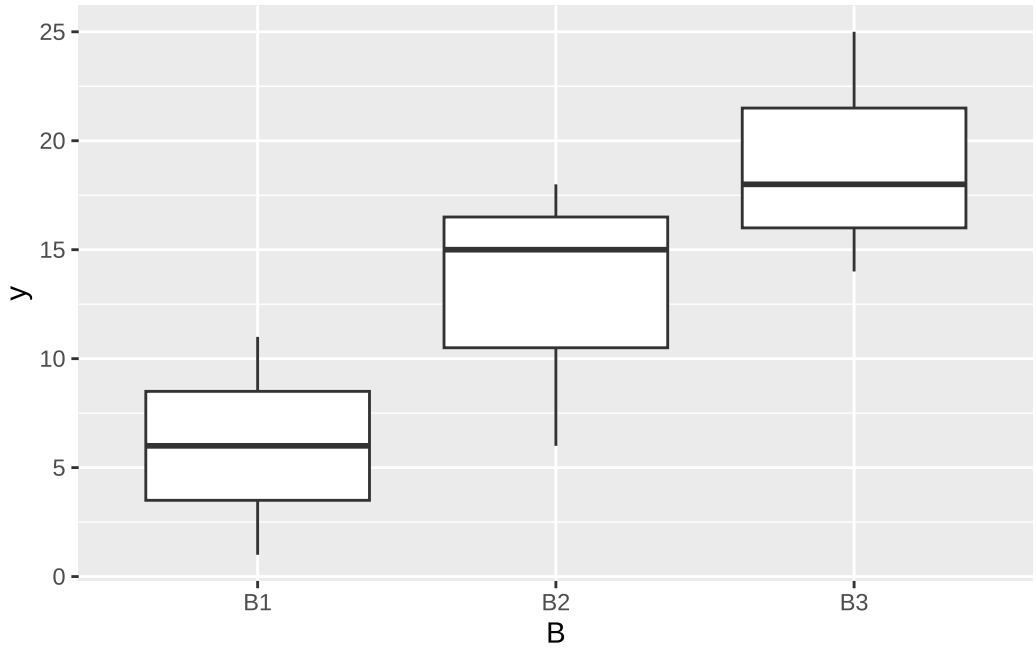
```
y <- c(11,18,25,1,6,14,6,15,18)
A <- factor(c(rep(c("A1", "A2","A3"), each=3)))
B <- factor(c(rep(c("B1", "B2","B3"), 3)))
df <- data.frame(A,B,y)
df
```

```
   A  B  y
1 A1 B1 11
2 A1 B2 18
3 A1 B3 25
4 A2 B1  1
5 A2 B2  6
6 A2 B3 14
7 A3 B1  6
8 A3 B2 15
9 A3 B3 18
```

```
df %>%  ggplot( aes(x = A , y = y) ) +  geom_boxplot()
```




```
df %>%  ggplot( aes(x = B , y = y) ) +  geom_boxplot()
```



4.4.2. 분산분석표

(1) 분산분석표의 작성

요인	제곱합	자유도	평균제곱	F_0
A	182	2	91	45.5**
B	254	2	127	63.5**
E	8	4	2	
T	444	8		

그림 4.3.: 분산분석표

```
fm1 <- aov(y~A+B, data=df)
summary(fm1)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
A            2    182      91    45.5 0.001773 **
B            2    254     127    63.5 0.000932 ***
Residuals    4      8       2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4.3. 직교대비에 대한 제곱합의 분해

(2) 직교대비에 의한 요인 A의 변동의 분해

L_1 = 국산과 외제와의 차이
$= (T_{0.} + T_{1.}) - 2T_{2.} = -3$

L_2 = 자회사와 국내 타회사와의 차이
$= (T_{0.} - T_{1.}) + 0 \times T_{2.} = 33$

선형식 L_1과 L_2는 직교대비 &

$$SS_{L_1} = \frac{(L_1)^2}{(\sum c_i^2) \times r} = \frac{(-3)^2}{6 \times 3} = 0.5$$

$$SS_{L_2} = \frac{(L_2)^2}{(\sum d_i^2) \times r} = \frac{(33)^2}{2 \times 3} = 181.5$$

여기서

$$SS_A\ (= 182) = SS_{L_1} + SS_{L_2} \text{ 확인}$$

(3) 직교대비에 의한 요인 B의 변동의 분해

L_l : 일차효과 대비 & L_q: 이차효과 대비

$$L_l = (T_{.1} + T_{.0}) + (T_{.2} - T_{.1}) = T_{.2} - T_{.0} = 39$$
$$L_q = (T_{.2} - T_{.1}) - (T_{.1} - T_{.0}) = T_{.2} - 2T_{.1} + T_{.0} = -3$$

$\Rightarrow L_l$과 L_q는 서로 직교

$$SS_l = \frac{(39)^2}{2 \times 3} = 253.5$$

$$SS_q = \frac{(-3)^2}{6 \times 3} = 0.5$$

여기서 $SS_b = SS_l + SS_q = 254$

그림 4.4.: 제곱합의 분해

4.4.4. 직교대비에 대한 검정

4.4.4.1. 요인 A: 제조 회사에 대한 직교 대비

```
# 직교 대비 계수 설정
c1 <- c(1, 1, -2) # 국산 대 외제
c2 <- c(1, -1, 0) # 자사 대 국내사

#직교대비 행렬 생성
matA <- cbind(c1,c2)
matA
```

```
     c1 c2
[1,]  1  1
```

```
[2,]  1 -1
[3,] -2  0
```

```
# 요인에 대한 대비 지정
contrasts(df$A) <- matA

# 직교대비에 대한 검정
fm1 <- aov(y~A+B, data=df) # 직교대비 지정한 후에 다시 분산분석을 해주어야 한다.
summary.aov(fm1 , split=list(A=list("국산 대 외제"=1, "자사 대 국내사" = 2)))
```

```
                    Df Sum Sq Mean Sq F value   Pr(>F)
A                    2  182.0    91.0   45.50 0.001773 **
  A: 국산 대 외제    1    0.5     0.5    0.25 0.643330
  A: 자사 대 국내사  1  181.5   181.5   90.75 0.000678 ***
B                    2  254.0   127.0   63.50 0.000932 ***
Residuals            4    8.0     2.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4.4.2. 요인 B: 성형온도에 대한 직교 대비

```
# 요인에 대한 대비 지정
contrasts(df$B) <- contr.poly(3)

# 직교대비에 대한 검정
fm1 <- aov(y~A+B, data=df) # 직교대비 지정한 후에 다시 분산분석을 해주어야 한다.
summary(fm1 , split=list(B=list("선형"=1, "이차" = 2)))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
A            2  182.0    91.0   45.50 0.001773 **
B            2  254.0   127.0   63.50 0.000932 ***
  B: 선형    1  253.5   253.5  126.75 0.000355 ***
  B: 이차    1    0.5     0.5    0.25 0.643330
Residuals    4    8.0     2.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4.5. 두 요인을 모두 나타내는 분산분석

```
fm1 <- aov(y~A+B, data=df) # 직교대비 지정한 후에 다시 분산분석을 해주어야 한다.
summary(fm1 , split=list( A=list("국산 대 외제"=1, "자사 대 국내사" = 2), B=list("선형"=1, "이차" = 2)))
```

```
                    Df Sum Sq Mean Sq F value   Pr(>F)
A                    2  182.0    91.0   45.50 0.001773 **
  A: 국산 대 외제  1    0.5     0.5    0.25 0.643330
  A: 자사 대 국내사 1  181.5   181.5   90.75 0.000678 ***
B                    2  254.0   127.0   63.50 0.000932 ***
  B: 선형       1  253.5   253.5  126.75 0.000355 ***
  B: 이차       1    0.5     0.5    0.25 0.643330
Residuals            4    8.0     2.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(4) 대비의 변동을 포함한 분산분석표의 작성

요인	제곱합	자유도	평균제곱	F_0
A	182	2	91	45.5**
L_1	0.5	1	0.5	0.25
L_2	181.5	1	181.5	90.75**
B	254	2	127	63.5**
L_l	253.5	1	253.5	117.75**
L_q	0.5	1	0.5	0.25
E	8	4	2	
T	444	8		

그림 4.5.: 직교대비에 대한 분산분석표

5. 2수준 요인배치법

5.1. 반복이 없는 2^3 요인배치법

먼저 반복이 없는 2^3 요인배치법이 적용되는 교과서 예제 7.3 에 대하여 논의합니다.

5.1.1. 처리조합 자료의 생성

먼저 R 프로그림을 이용하여 분석을 하기 위해서는 실험 자료를 `data.frame` 형식으로 만들어야 한다.

먼저 각 요인의 수준을 조합하여 처리의 조합을 만들어 보자. 처리의 조합을 만드는 일은 일일이 손으로 처리 조합을 만들 수 있지만 패키지 `FrF2` 에 있는 함수 `FrF2()`를 사용하면 처리 조합에 대한 데이터프레임을 쉽게 만들 수 있다.

이제 2^3 요인배치법의 처리 조합은 다음과 같이 만들 수 있다.

```
X <- FrF2(nruns=8, nfactors=3, randomize = FALSE)
X
```

```
  A  B  C
1 -1 -1 -1
2  1 -1 -1
3 -1  1 -1
4  1  1 -1
5 -1 -1  1
6  1 -1  1
7 -1  1  1
8  1  1  1
class=design, type= full factorial
```

위에서 함수 `FrF2()`는 다음과 같은 인자를 가진다.

- `nruns` : 처리 조합의 개수
- `nfactors` : 요인의 개수
- `randomize=TRUE` : 처리조합의 순서를 임의로 바꾸는 명령

```
FrF2(nruns, nfactors, randomize=TRUE)
```

우리는 지금 3개의 요인을 가진 2요인 배치법을 고려하므로 위와 같이 `nruns=8`, `nfactors=3` 으로 지정해야 한다. 또한 실험을 직접 실행하는 것이 아니므로 실험의 순서는 임의화 하지 않는다 (`randomize = FALSE`). 유의할 점은 요인의 이름은 다른 지정을 하지 않으면 알파벳 대문자 순서(`A`, `B`, `C`,…)로 지정해 준다.

이제 반응값을 위에서 생성한 처리의 조합순서대로 생성하고 함수 `cbind()` 를 이용하여 실험자료를 만들어 보자.

A	B	C	y
-1	-1	-1	2
1	-1	-1	-5
-1	1	-1	15
1	1	-1	13
-1	-1	1	-12
1	-1	1	-17
-1	1	1	-2
1	1	1	-7

```
y <- c(2,-5,15,13,-12,-17,-2,-7)
df <- cbind(X,y)
df %>%  kbl() %>%   kable_paper("hover", full_width = F)
```

주의

위에서 작성한 3 요인의 2수준 배치법의 자료에서 처리의 순서는 **표준형 순서(standard order)**로 정렬되어 있다. 표준형 순서는 요인의 순서를 A, B, C 로 고려한다면 제일 먼저 나오는 요인의 수준이 가장 빨리 변하고 다음 요인의 순서가 그 보다 느리게 변하며 가장 마지막의 요인에 대한 수준의 순서가 가장 느리게 변하는 것을 의미한다.
즉 요인 A 의 순서는 -+-+-+-+, 요인 B의 순서는 --++--++ 이며 마지막 요인 C의 순서는 ----++++ 이다.
함수 FrF2() 는 randomize = FALSE 로 지정해 주면 처리의 순서를 표준형 순서로 생성한다.

5.1.2. 처리효과의 계산

5.1.2.1. 인수분해법

이제 교과서에서 나오는 방법으로 처리 효과를 계산해 보자.

요인 A 에 대한 주 효과는 인수분해 방법을 통해서 다음과 같이 계산할 수 있다.

$$A = \frac{1}{4}(a-1)(b+1)(c+1) = \frac{1}{4}[(a+ac+ab+abs)-((1)-c-b-bd)]$$

따라서 A 에 대한 주효과는 다음과 같이 계산된다.

$$A = \frac{1}{4}[(-5+13-17-7)-(2+15-12-2)] = -4.75$$

다른 모든 요인들의 주효과와 상호작용 효과는 교과서 181-182에 나오는 인수분해법으로 구할 수 있다.

5.1.2.2. 함수 yates()

패키지 unrepx 에 나오는 함수 yates 를 이용하면 손쉽게 처리 효과를 계산할 수 있다.

```
yates(df$y, labels = c("A", "B", "C"))
```

A	B	C	D	yy
-1	-1	-1	-1	1
1	-1	-1	-1	2
-1	1	-1	-1	3
1	1	-1	-1	4
-1	-1	1	-1	5
1	-1	1	-1	6
-1	1	1	-1	7
1	1	1	-1	8
-1	-1	-1	1	9
1	-1	-1	1	10
-1	1	-1	1	11
1	1	-1	1	12
-1	-1	1	1	13
1	-1	1	1	14
-1	1	1	1	15
1	1	1	1	16

```
     A      B     AB      C     AC     BC    ABC
 -4.75  12.75   1.25 -15.75  -0.25  -2.75  -1.25
attr(,"mean")

-1.625
```

함수 yates 는 첫 번째 인자로 표준형 순서로 정렬되어 있는 반응값의 벡터를 넣어주고 두 번째 인자 labels 로 요인의 이름으로 구성된 문자 벡터를 넣어준다. 함수 yates 의 결과는 각 요인의 효과를 계산해 주고 마지막으로 전체 평균 $\bar{y}$ 를 생성한다.

함수 yates() 를 이용하면 표준형 순서로서 영문 소문자 표기법으로 표시된 처리조합을 구할 수 있다. 예를 들어서 2^4 실험법에 대하여 abcd 표기법으로 표준형 순서로 정렬된 처리 조합을 아래와 같이 구할 수 있다,

물론 가장 처음의 처리 (1) 은 결과에 나타나지 않는다.

```
XX <- FrF2(16, 4, randomize=FALSE)
yy <- 1:16
df4 <- cbind(XX,yy)
df4 %>%  kbl() %>%   kable_paper("hover", full_width = F)
```

```
yates(df4$yy , c("A", "B", "C", "D"))
```

```
   A    B   AB    C   AC   BC  ABC    D   AD   BD  ABD   CD  ACD  BCD ABCD
   1    2    0    4    0    0    0    8    0    0    0    0    0    0    0
attr(,"mean")

8.5
```

5.1.2.3. R 프로그램을 이용

R 프로그램을 이용하여 Yates 방법으로 처리 효과를 계산해주는 함수 myyates 를 만들어 보자.

```
# yates 방법으로 처리 효과를 계산해주는 함수
myyates <- function(y) {
   n <- length(y)  #자료의 수
   k = round(log(n)/log(2))  # 요인의 수
   nhalf <- n/2  # (자료의 수)/2

   res<- rep(0,n)

   for ( i in 1:k ){
     for (j in 1:nhalf) {
       res[j] <- y[2*j-1] + y[2*j]
     }
     for (j in 1:nhalf) {
       res[j+nhalf] <- -y[2*j-1] + y[2*j]
     }
     y <-res
   }
   res <- res/nhalf
   res[1] <- res[1]/2
   res
}

myyates(df$y)
```

```
[1]  -1.625  -4.750  12.750   1.250 -15.750  -0.250  -2.750  -1.250
```

함수 myyates 를 이용하여 얻은 결과에서 처음 나온 수는 전체 평균의 2 배이며 두 번째 수부터 표준 효과의 추정값이다.

5.1.2.4. 회귀식의 이용

이제 위에서 고려한 데이터프레임 df 에 대한 회귀식을 적합시키자.

아래 회귀식에서 y~A*B*C 는 변수 y 를 반응변수로 하고 3개의 요인 A, B, C 의 각 개별 효과와 모든 상호작용 효과를 고려한 선형 모형이다.

즉, 아래 적합한 모형은 요인의 수준이 모두 범주형인 경우이므로 다음과 같은 3원배치 분산분석 모형을 적합하는 것이다.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma_k) + e_{ijk}, \quad i = 1, 2,\ j = 1, 2,\ k = 1, 2 \tag{5.1}$$

```
fm1 <- lm (y~ A*B*C, data=df)
summary(fm1)
```

```
Call:
lm.default(formula = y ~ A * B * C, data = df)
```

```
Residuals:
ALL 8 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.625        NaN     NaN      NaN
A1            -2.375        NaN     NaN      NaN
B1             6.375        NaN     NaN      NaN
C1            -7.875        NaN     NaN      NaN
A1:B1          0.625        NaN     NaN      NaN
A1:C1         -0.125        NaN     NaN      NaN
B1:C1         -1.375        NaN     NaN      NaN
A1:B1:C1      -0.625        NaN     NaN      NaN

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:    NaN
F-statistic:   NaN on 7 and 0 DF,  p-value: NA
```

위의 모형 식 5.1 은 모수의 개수가 $1+2+2+2+4+4+4+8=27$이다. 하지만 관측값이 8개이기 때문에 정규방정식의 해가 유일하게 존재하지 않는다. 따라서 최소한 19개의 제약 조건이 필요하다. 고차원 배치법에 대한 회귀모형에서 제약조건을 주는 방법은 과목의 범위를 벗어나므로 생략한다.

위의 추정 결과는 8개의 관측값을 가지고 8개의 모수를 가진 모형을 적합하는 경우이며 이렇게 관측값의 개수와 모수의 개수가 같은 모형을 **포화모형(saturated model)**이라고 부른다. 포화모형에서는 오차항의 분산을 추정할 수 있는 잔차가 모두 0이기 때문에 σ^2 을 추정할 수 없다. 따라서 표준오차도 구할 수 없다.

회귀식의 계수에 대한 추정치에서 절편을 제외한 추정치는 대응하는 효과 추정치의 절반인 것을 알 수 있다. 절편의 주정치는 전체 평균 $\bar{y}$이다.

```
coef(fm1)[-1]*2
```

```
      A1       B1       C1    A1:B1    A1:C1    B1:C1 A1:B1:C1
  -4.750   12.750  -15.750    1.250   -0.250   -2.750   -1.250
```

위에서 적합한 회귀식을 선형모형 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{e}$ 로 보면 모형의 계획행렬 $\boldsymbol{X}$ 는 다음과 같이 나오며 열이 각 효과의 대비인 것을 알 수 있다.

```
X <- model.matrix(fm1)
X
```

```
  (Intercept) A1 B1 C1 A1:B1 A1:C1 B1:C1 A1:B1:C1
1           1 -1 -1 -1     1     1     1       -1
2           1  1 -1 -1    -1    -1     1        1
3           1 -1  1 -1    -1     1    -1        1
4           1  1  1 -1     1    -1    -1       -1
5           1 -1 -1  1     1    -1    -1        1
6           1  1 -1  1    -1     1    -1       -1
```

```
7           1 -1  1  1    -1    -1     1       -1
8           1  1  1  1     1     1     1        1
attr(,"assign")
[1] 0 1 2 3 4 5 6 7
attr(,"contrasts")
attr(,"contrasts")$A
   [,1]
-1   -1
1     1

attr(,"contrasts")$B
   [,1]
-1   -1
1     1

attr(,"contrasts")$C
   [,1]
-1   -1
1     1
```

```
yvec <- matrix(df$y, 8, 1)
yvec
```

```
     [,1]
[1,]    2
[2,]   -5
[3,]   15
[4,]   13
[5,]  -12
[6,]  -17
[7,]   -2
[8,]   -7
```

따라서 위에서 함수 `model.matrix` 로 구한 행렬 $\boldsymbol{X}$ 의 전치 $\boldsymbol{X}^t$ 에 반응 변수 벡터 $\boldsymbol{y}$ 를 곱해주면, 즉 $\boldsymbol{X}^t\boldsymbol{y}$ 는 각 효과에 대하여 합으로 구한 대비를 얻을 수 있다. 이렇게 합으로 구한 대비를 적절한 수로 나누어 주면 평균의 대비도 얻을 수 있다. 이 예제와 같은 반복이 없는 2^3 배치법은 4로 나누어 주면 된다. 주의할 점은 앞에서 효과를 회귀계수로 구하는 경우와 마찬가지로 $\boldsymbol{X}^t\boldsymbol{y}$ 의 첫 번째 원소는 모든 반응값의 총합 $T_{\cdots}$ 인 것에 유의하자.

$$\boldsymbol{X}^t\boldsymbol{y} = \begin{bmatrix} T_{\cdots} \\ T_{1..} - T_{0..} \\ \vdots \end{bmatrix}$$

예를 들어서 A 에 대한 주효과는 다음과 같이 구할 수 있으며 아래에서 `total_effect` 가 $T_{1..} - T_{0..}$ 이고 `mean_effect` 는 $(T_{1..} - T_{0..})/4$ 이다.

$$A = \frac{1}{4}(T_{1..} - T_{0..}) = \frac{1}{4}\begin{bmatrix}-1 & 1 & -1 & 1 & -1 & 1 & -1 & 1\end{bmatrix}\begin{bmatrix}2\\-5\\15\\13\\-12\\-17\\-2\\-7\end{bmatrix} = -4.75$$

```
total_effect <- t(X) %*% yvec
total_effect <- total_effect[-1]
total_effect
```

```
[1] -19  51 -63   5  -1 -11  -5
```

```
mean_effect <- total_effect/4
mean_effect
```

```
[1]  -4.75  12.75 -15.75   1.25  -0.25  -2.75  -1.25
```

앞에서 회귀모형의 계수가 각 효과의 2 배로 나타나는 이유는 다음과 같이 회귀식의 계수를 구하는 정규방정식에서 $\boldsymbol{X}^t\boldsymbol{X}$ 가 대각행렬이며 대각원소의 값이 자료의 개수 $(2^3) = 8$ 로 나타나기 때문이다. 효과를 구하기 구하는 때는 합으로 이루어진 대비를 $2^2 = 4$ 로 나누기 때문에 회귀계수의 추정값은 효과의 추정값의 절반으로 나타난다.

$$\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{y} \quad \rightarrow \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

```
t(X) %*% X
```

```
            (Intercept) A1 B1 C1 A1:B1 A1:C1 B1:C1 A1:B1:C1
(Intercept)           8  0  0  0     0     0     0        0
A1                    0  8  0  0     0     0     0        0
B1                    0  0  8  0     0     0     0        0
C1                    0  0  0  8     0     0     0        0
A1:B1                 0  0  0  0     8     0     0        0
A1:C1                 0  0  0  0     0     8     0        0
B1:C1                 0  0  0  0     0     0     8        0
A1:B1:C1              0  0  0  0     0     0     0        8
```

```
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% yvec
beta_hat
```

```
              [,1]
(Intercept) -1.625
A1          -2.375
B1           6.375
```

```
C1          -7.875
A1:B1        0.625
A1:C1       -0.125
B1:C1       -1.375
A1:B1:C1    -0.625
```

5.1.3. 분산분석

분산분석표는 앞에서 적합한 모형 식 5.1 을 적합한 후 `aov()` 또는 `anova()` 함수를 적용하면 구할 수 있다.

앞에서 언급하였듯이 모형 식 5.1 은 포화모형이므로 제곱합은 구할 수 있지만 잔차제곱합을 구할 수 없으므로 가설 검정은 할 수 없다.

```
anova(fm1)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
A          1  45.13   45.13     NaN    NaN
B          1 325.12  325.12     NaN    NaN
C          1 496.13  496.13     NaN    NaN
A:B        1   3.13    3.13     NaN    NaN
A:C        1   0.13    0.13     NaN    NaN
B:C        1  15.12   15.12     NaN    NaN
A:B:C      1   3.13    3.13     NaN    NaN
Residuals  0   0.00     NaN
```

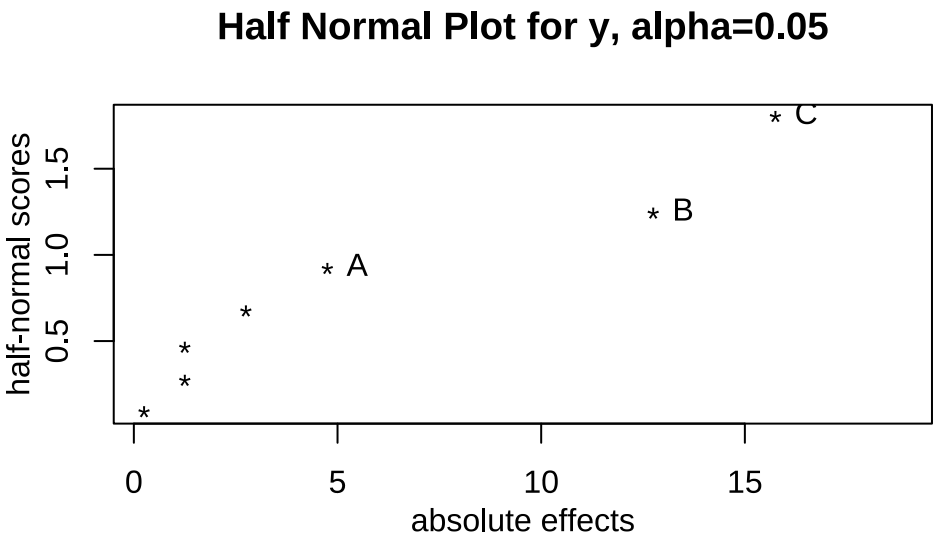
위의 분산분석표에서 효과 $A \times B$, $A \times C$, $A \times B \times C$ 에 대한 제곱합의 크기가 다른 효과에 비하여 상대적으로 매우 작다.

5.1.4. 핵심 요인효과의 선별

핵심요인 효과는 효과 추정치의 절대값 $|\bar{y}_1 - \bar{y}_0|$ 들을 가지고 반정규확률 그림을 그려서 선별할 수 있다.

반정규확률 그림은 패키지 `FrF2`에 있는 함수 `DanielPlot()` 를 사용하여 구할 수 있다.

```
DanielPlot(fm1, half=TRUE)
```



위의 반정규확률 그림을 보면 주요인 B 와 C 가 핵심 요인임을 알 수 있다.

교과서 예제의 결론과 같이 제곱합이 작은 3개의 효과 $A \times B$, $A \times C$, $A \times B \times C$ 를 풀링하여 모형을 다시 적합해 보자.

이제 회귀식을 적합시키자.

```
fm11 <- lm (y~ A+B+C+B*C, data=df)
summary(fm11)
```

```
Call:
lm.default(formula = y ~ A + B + C + B * C, data = df)

Residuals:
     1      2      3      4      5      6      7      8
 1.125 -1.125 -1.375  1.375  0.125 -0.125  0.125 -0.125

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6250     0.5154  -3.153 0.051148 .
A1           -2.3750     0.5154  -4.608 0.019220 *
B1            6.3750     0.5154  12.369 0.001138 **
C1           -7.8750     0.5154 -15.280 0.000609 ***
B1:C1        -1.3750     0.5154  -2.668 0.075826 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.458 on 3 degrees of freedom
Multiple R-squared:  0.9928,    Adjusted R-squared:  0.9832
F-statistic: 103.7 on 4 and 3 DF,  p-value: 0.001514
```

```
anova(fm11)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
A          1  45.13   45.13  21.2353 0.0192201 *
B          1 325.12  325.12 153.0000 0.0011384 **
C          1 496.13  496.13 233.4706 0.0006088 ***
B:C        1  15.12   15.12   7.1176 0.0758265 .
Residuals  3   6.38    2.13
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.1.5. 부록: 처리 조합을 만드는 다른 방법

참고로 처리 조합의 데이터프레임을 만드는 다른 방법을 알아보자.

함수 `expand.grid()` 는 인자로 주어진 벡터들의 원소들로 구성된 모든 조합을 만들어 주는 함수이다.

```
X1 <-  expand.grid(A = gl(2, 1, labels = c("-1", "1")),
                   B = gl(2, 1, labels = c("-1", "1")),
                   C = gl(2, 1, labels = c("-1", "1")))
X1
```

```
   A  B  C
1 -1 -1 -1
2  1 -1 -1
3 -1  1 -1
4  1  1 -1
5 -1 -1  1
6  1 -1  1
7 -1  1  1
8  1  1  1
```

함수 `expand.grid()` 의 인자에 대한 설명은 다음과 같다.

```
gl(n, k, length = n*k, labels = seq_len(n), ordered = FALSE)

- n: an integer giving the number of levels.
- k: an integer giving the number of replications.
- labels: an optional vector of labels for the resulting factor levels.
- ordered: a logical indicating whether the result should be ordered or not.
```

만약 반복이 있다면 데이터프레임을 함수 `rbind()` 를 이용하여 붙이면 된다.

```
X2 <- rbind(X1, X1)
X2
```

```
    A  B  C
1  -1 -1 -1
2   1 -1 -1
3  -1  1 -1
4   1  1 -1
5  -1 -1  1
6   1 -1  1
7  -1  1  1
8   1  1  1
9  -1 -1 -1
10  1 -1 -1
11 -1  1 -1
12  1  1 -1
13 -1 -1  1
14  1 -1  1
15 -1  1  1
16  1  1  1
```

5.2. 반복이 없는 2^4 요인배치법

먼저 반복이 없는 2^4 요인배치법이 적용되는 교과서 예제 7.4 에 대하여 논의합니다.

5.2.1. 처리조합 자료의 생성

```
X <- FrF2(nruns=16, nfactors=4, randomize = FALSE)
X
```

```
    A  B  C  D
1  -1 -1 -1 -1
2   1 -1 -1 -1
3  -1  1 -1 -1
4   1  1 -1 -1
5  -1 -1  1 -1
6   1 -1  1 -1
7  -1  1  1 -1
8   1  1  1 -1
9  -1 -1 -1  1
10  1 -1 -1  1
11 -1  1 -1  1
12  1  1 -1  1
```

A	B	C	D	y
-1	-1	-1	-1	-1
1	-1	-1	-1	0
-1	1	-1	-1	9
1	1	-1	-1	4
-1	-1	1	-1	5
1	-1	1	-1	3
-1	1	1	-1	11
1	1	1	-1	8
-1	-1	-1	1	-1
1	-1	-1	1	-9
-1	1	-1	1	1
1	1	-1	1	5
-1	-1	1	1	-9
1	-1	1	1	-13
-1	1	1	1	-5
1	1	1	1	-4

```
13 -1 -1  1  1
14  1 -1  1  1
15 -1  1  1  1
16  1  1  1  1
class=design, type= full factorial
```

```
y<- c(-1, 0, 9, 4, 5, 3, 11, 8,-1, -9, 1, 5, -9, -13, -5, -4)
df2 <- cbind(X,y)
df2 %>%  kbl() %>%   kable_paper("hover", full_width = F)
```

5.2.2. 처리효과의 계산

```
yates(df2$y, c("A", "B", "C", "D"))
```

```
    A     B    AB     C    AC    BC   ABC     D    AD    BD   ABD    CD   ACD
-2.00  6.75  1.25 -1.50  0.00 -0.75 -0.25 -9.25  0.25  0.50  3.00 -5.25  0.25
  BCD  ABCD
 0.00 -1.50
attr(,"mean")

0.25
```

5.2.2.1. 포화모형의 적합

이제 포화모형인 회귀식을 적합시키자.

```
fm2 <- lm (y~ A*B*C*D, data=df2)
summary(fm2)
```

```
Call:
lm.default(formula = y ~ A * B * C * D, data = df2)

Residuals:
ALL 16 residuals are 0: no residual degrees of freedom!

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.500e-01        NaN     NaN      NaN
A1          -1.000e+00        NaN     NaN      NaN
B1           3.375e+00        NaN     NaN      NaN
C1          -7.500e-01        NaN     NaN      NaN
D1          -4.625e+00        NaN     NaN      NaN
A1:B1        6.250e-01        NaN     NaN      NaN
A1:C1       -1.095e-15        NaN     NaN      NaN
B1:C1       -3.750e-01        NaN     NaN      NaN
A1:D1        1.250e-01        NaN     NaN      NaN
B1:D1        2.500e-01        NaN     NaN      NaN
C1:D1       -2.625e+00        NaN     NaN      NaN
A1:B1:C1    -1.250e-01        NaN     NaN      NaN
A1:B1:D1     1.500e+00        NaN     NaN      NaN
A1:C1:D1     1.250e-01        NaN     NaN      NaN
B1:C1:D1     1.045e-15        NaN     NaN      NaN
A1:B1:C1:D1 -7.500e-01        NaN     NaN      NaN

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:    NaN
F-statistic:   NaN on 15 and 0 DF,  p-value: NA
```

```
coef(fm2)[-1]*2
```

```
           A1            B1            C1            D1         A1:B1
-2.000000e+00  6.750000e+00 -1.500000e+00 -9.250000e+00  1.250000e+00
        A1:C1         B1:C1         A1:D1         B1:D1         C1:D1
-2.190503e-15 -7.500000e-01  2.500000e-01  5.000000e-01 -5.250000e+00
     A1:B1:C1      A1:B1:D1      A1:C1:D1      B1:C1:D1   A1:B1:C1:D1
-2.500000e-01  3.000000e+00  2.500000e-01  2.090373e-15 -1.500000e+00
```

```
anova(fm2)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
A          1  16.00   16.00     NaN    NaN
```

```
B           1 182.25  182.25     NaN    NaN
C           1   9.00    9.00     NaN    NaN
D           1 342.25  342.25     NaN    NaN
A:B         1   6.25    6.25     NaN    NaN
A:C         1   0.00    0.00     NaN    NaN
B:C         1   2.25    2.25     NaN    NaN
A:D         1   0.25    0.25     NaN    NaN
B:D         1   1.00    1.00     NaN    NaN
C:D         1 110.25  110.25     NaN    NaN
A:B:C       1   0.25    0.25     NaN    NaN
A:B:D       1  36.00   36.00     NaN    NaN
A:C:D       1   0.25    0.25     NaN    NaN
B:C:D       1   0.00    0.00     NaN    NaN
A:B:C:D     1   9.00    9.00     NaN    NaN
Residuals   0   0.00     NaN
```

5.2.2.2. 3차 이상의 상호작용을 풀링

이제 3차 이상의 상호적용을 풀링한 모형을 적합시키자.

```
fm21 <- lm (y~ A + B + C+ D+ A*B + A*C + A*D + B*C + B*D + C*D, data=df2)
summary(fm21)
```

```
Call:
lm.default(formula = y ~ A + B + C + D + A * B + A * C + A *
    D + B * C + B * D + C * D, data = df2)

Residuals:
    1     2     3     4     5     6     7     8     9    10    11    12    13
-2.25  2.25  2.00 -2.00 -0.75  0.75  1.00 -1.00  2.50 -2.50 -2.25  2.25  0.50
   14    15    16
-0.50 -0.75  0.75

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.500e-01  7.542e-01   0.331  0.75371
A1          -1.000e+00  7.542e-01  -1.326  0.24219
B1           3.375e+00  7.542e-01   4.475  0.00655 **
C1          -7.500e-01  7.542e-01  -0.994  0.36564
D1          -4.625e+00  7.542e-01  -6.133  0.00167 **
A1:B1        6.250e-01  7.542e-01   0.829  0.44500
A1:C1       -1.352e-15  7.542e-01   0.000  1.00000
A1:D1        1.250e-01  7.542e-01   0.166  0.87485
B1:C1       -3.750e-01  7.542e-01  -0.497  0.64011
B1:D1        2.500e-01  7.542e-01   0.331  0.75371
```

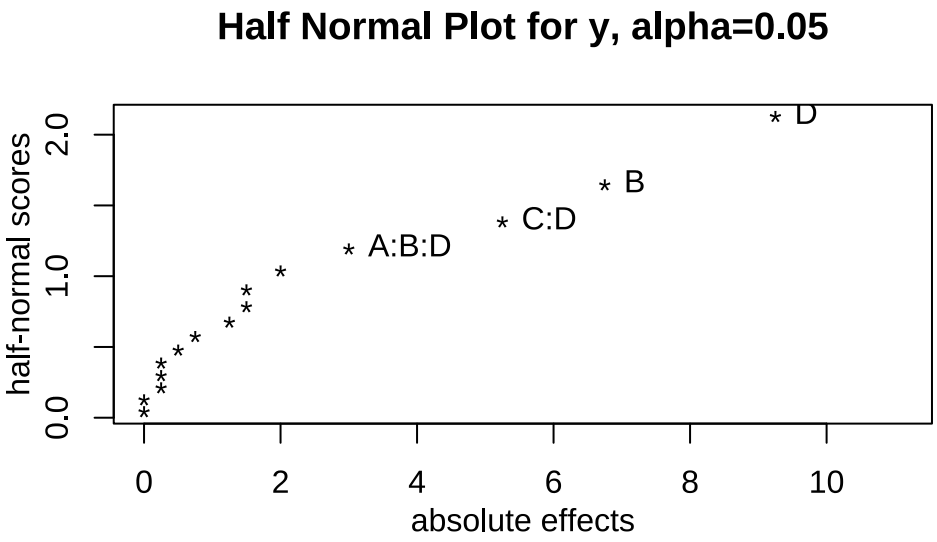
```
C1:D1       -2.625e+00  7.542e-01  -3.481  0.01765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.017 on 5 degrees of freedom
Multiple R-squared:  0.9364,    Adjusted R-squared:  0.8091
F-statistic: 7.357 on 10 and 5 DF,  p-value: 0.01992
```

```
anova(fm21)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
A          1  16.00   16.00  1.7582 0.242188
B          1 182.25  182.25 20.0275 0.006548 **
C          1   9.00    9.00  0.9890 0.365645
D          1 342.25  342.25 37.6099 0.001674 **
A:B        1   6.25    6.25  0.6868 0.444996
A:C        1   0.00    0.00  0.0000 1.000000
A:D        1   0.25    0.25  0.0275 0.874848
B:C        1   2.25    2.25  0.2473 0.640107
B:D        1   1.00    1.00  0.1099 0.753712
C:D        1 110.25  110.25 12.1154 0.017645 *
Residuals  5  45.50    9.10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.2.3. 핵심 요인효과의 선별

```
DanielPlot(fm2, half=TRUE)
```




```
fm22 <- lm (y~ A+B+C+D+ C*D, data=df2)
summary(fm22)
```

```
Call:
lm.default(formula = y ~ A + B + C + D + C * D, data = df2)

Residuals:
   Min     1Q Median     3Q    Max
-3.625 -1.438  0.250  1.375  3.625

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2500     0.5876   0.425 0.679529
A1           -1.0000     0.5876  -1.702 0.119634
B1            3.3750     0.5876   5.743 0.000187 ***
C1           -0.7500     0.5876  -1.276 0.230691
D1           -4.6250     0.5876  -7.871 1.36e-05 ***
C1:D1        -2.6250     0.5876  -4.467 0.001203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.351 on 10 degrees of freedom
Multiple R-squared:  0.9227,    Adjusted R-squared:  0.8841
F-statistic: 23.88 on 5 and 10 DF,  p-value: 2.925e-05
```

```
anova(fm22)
```

```
Analysis of Variance Table
```

```
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          1  16.00   16.00  2.8959 0.1196341
B          1 182.25  182.25 32.9864 0.0001869 ***
C          1   9.00    9.00  1.6290 0.2306913
D          1 342.25  342.25 61.9457 1.358e-05 ***
C:D        1 110.25  110.25 19.9548 0.0012029 **
Residuals 10  55.25    5.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.3. 반복이 있는 2^3 요인배치법

이제 반복이 있는 2^3 요인배치법이 적용되는 예제에 대하여 논의합니다.

자료는 (Montgomery 2017) 이 나온 예제 6.1 을 사용하였다. 반도체 공정에서 웨이퍼를 만드는 공정에서 다음과 같은 2개의 수준을 가진 3개의 요인과 반응변수을 고려한 실험을 실시하였다.

- 요인 A: electrodes
- 요인 B: the gas flow
- 요인 C: RF power applied to the cathode
- 반응변수 : etch rate for silicon nitride

각 처리 조합마다 2개의 반복 측정값을 얻었다. 을 얻었다.

5.3.1. 처리조합 자료의 생성

패키지 `FrF2` 에 있는 함수 `FrF2()`에서 `replications=2`를 지정하면 2개의 반복이 있는 2^3 요인배치법의 처리 조합을 생성해준다.

이제 2^3 요인배치법의 처리 조합은 다음과 같이 만들 수 있다.

```
X <- FrF2(nruns=8, nfactors=3, randomize = FALSE, replications=2)
X <- as.data.frame(X)
X
```

```
    A  B  C Blocks
1  -1 -1 -1     .1
2   1 -1 -1     .1
3  -1  1 -1     .1
4   1  1 -1     .1
5  -1 -1  1     .1
6   1 -1  1     .1
7  -1  1  1     .1
8   1  1  1     .1
9  -1 -1 -1     .2
```

```
10  1 -1 -1     .2
11 -1  1 -1     .2
12  1  1 -1     .2
13 -1 -1  1     .2
14  1 -1  1     .2
15 -1  1  1     .2
16  1  1  1     .2
```

```
X <- X[,-4]  # 블럭변수가 필요없으모 제외
X
```

```
    A  B  C
1  -1 -1 -1
2   1 -1 -1
3  -1  1 -1
4   1  1 -1
5  -1 -1  1
6   1 -1  1
7  -1  1  1
8   1  1  1
9  -1 -1 -1
10  1 -1 -1
11 -1  1 -1
12  1  1 -1
13 -1 -1  1
14  1 -1  1
15 -1  1  1
16  1  1  1
```

이제 반응값을 처리조합과 결합하여 실험자료를 만들어 보자.

```
y <- c(550, 669, 633, 642, 1037, 749, 1075, 729, 604, 650, 601, 635, 1052, 868, 1063, 860)
df3 <- cbind(X,y)
df3 %>%  kbl() %>%   kable_paper("hover", full_width = F)
```

5.3.2. 처리효과의 계산

5.3.2.1. 회귀식의 이용

이제 회귀식을 적합시키자.

```
fm3 <- lm (y~ A*B*C, data=df3)
summary(fm3)
```

A	B	C	y
-1	-1	-1	550
1	-1	-1	669
-1	1	-1	633
1	1	-1	642
-1	-1	1	1037
1	-1	1	749
-1	1	1	1075
1	1	1	729
-1	-1	-1	604
1	-1	-1	650
-1	1	-1	601
1	1	-1	635
-1	-1	1	1052
1	-1	1	868
-1	1	1	1063
1	1	1	860

```
Call:
lm.default(formula = y ~ A * B * C, data = df3)

Residuals:
   Min     1Q Median     3Q    Max
-65.50 -11.12   0.00  11.12  65.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  776.062     11.865  65.406 3.32e-12 ***
A1           -50.813     11.865  -4.282 0.002679 **
B1             3.688     11.865   0.311 0.763911
C1           153.062     11.865  12.900 1.23e-06 ***
A1:B1        -12.437     11.865  -1.048 0.325168
A1:C1        -76.813     11.865  -6.474 0.000193 ***
B1:C1         -1.062     11.865  -0.090 0.930849
A1:B1:C1       2.812     11.865   0.237 0.818586
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.46 on 8 degrees of freedom
Multiple R-squared:  0.9661,    Adjusted R-squared:  0.9364
F-statistic: 32.56 on 7 and 8 DF,  p-value: 2.896e-05
```

위의 추정 결과는 16개의 관측값을 가지고 8개의 모수를 가진 모형을 적합하는 경우이므로 σ^2을 추정할 수 있다.

회귀식의 계수에 대한 추정치에서 절편을 제외한 추정치는 대응하는 효과 추정치의 절반인 것을 알 수 있다. 절편의 주정치는 전체 평균 $\bar{y}$이다.

```
coef(fm3)[-1]*2
```

```
      A1       B1       C1    A1:B1    A1:C1    B1:C1 A1:B1:C1
-101.625    7.375  306.125  -24.875 -153.625   -2.125    5.625
```

5.3.3. 분산분석

```
anova(fm3)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
A          1  41311   41311  18.3394 0.0026786 **
B          1    218     218   0.0966 0.7639107
C          1 374850  374850 166.4105 1.233e-06 ***
A:B        1   2475    2475   1.0988 0.3251679
A:C        1  94403   94403  41.9090 0.0001934 ***
B:C        1     18      18   0.0080 0.9308486
A:B:C      1    127     127   0.0562 0.8185861
Residuals  8  18020    2253
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

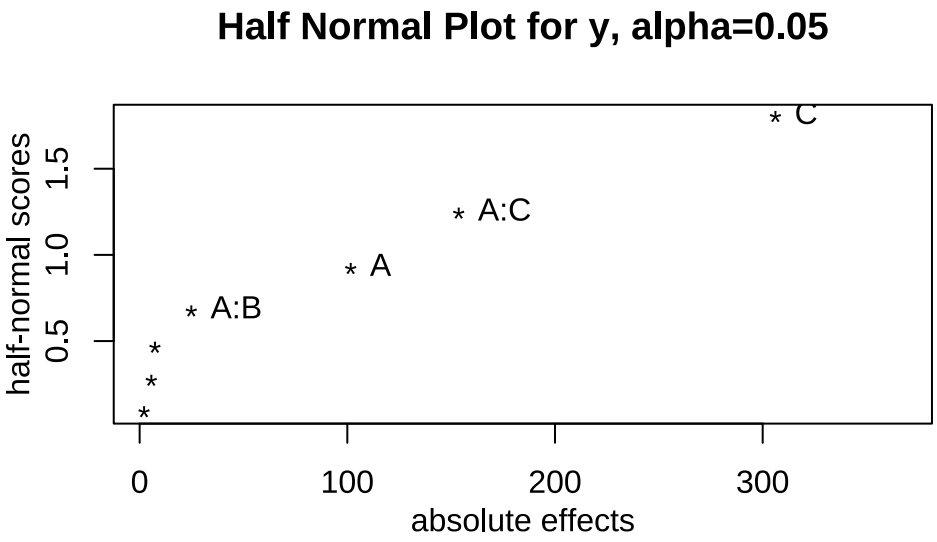
위의 분산분석표에서 효과 $B \times C$, $A \times B \times C$ 에 대한 제곱합의 크기가 다른 효과에 비하여 상대적으로 매우 작다.

5.3.4. 핵심 요인효과의 선별

핵심요인 효과는 효과 추정치의 절대값 $|\bar{y}_1 - \bar{y}_0|$ 들을 가지고 반정규확률 그림을 그려서 선별할 수 있다.

반정규확률 그림은 패키지 `FrF2`에 있는 함수 `DanielPlot()` 를 사용하여 구할 수 있다.

```
DanielPlot(fm3, half=TRUE)
```



위의 반정규확률 그림을 보면 주요인 A, C 와 상호작용 $A \times C$ 가 핵심 요인임을 알 수 있다.

이제 핵심 요인으로 판단되는 주요인 A, C 와 상호작용 $A \times C$ 만을 포함한 모형을 다시 적합해 보자.

이제 회귀식을 적합시키자.

```
fm31 <- lm (y~ A+C+A*C, data=df3)
summary(fm31)
```

```
Call:
lm.default(formula = y ~ A + C + A * C, data = df3)

Residuals:
   Min     1Q Median     3Q    Max
-72.50 -15.44   2.50  18.69  66.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   776.06      10.42  74.458  < 2e-16 ***
A1            -50.81      10.42  -4.875 0.000382 ***
C1            153.06      10.42  14.685 4.95e-09 ***
A1:C1         -76.81      10.42  -7.370 8.62e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.69 on 12 degrees of freedom
Multiple R-squared:  0.9608,    Adjusted R-squared:  0.9509
F-statistic: 97.91 on 3 and 12 DF,  p-value: 1.054e-08
```

```
anova(fm31)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
A          1  41311   41311  23.767 0.0003816 ***
C          1 374850  374850 215.661 4.951e-09 ***
A:C        1  94403   94403  54.312 8.621e-06 ***
Residuals 12  20858    1738
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. 2수준 요인배치법 - 교락법

6.1. 교과서 예제 8.2

반복이 없는 2^4 요인배치법에서 16회의 실험을 동일한 환경에서 실시할 수 없어서 상호작용효과 ACD 와 BCD 를 블럭과 교락시켜서 실험하였다 (교과서 예제 8.2, 230 페이지)

6.1.1. 실험자료의 생성

먼저 16개의 실험 처리 조합을 표준 순서로서 생성하자. `default.level=c(0,1)` 은 요인의 수준을 0과 1로 표시하는 옵션이다.

```
X <- FrF2(nruns=16, nfactors=4, randomize = FALSE,  default.level=c(0,1))
X
```

```
   A B C D
1  0 0 0 0
2  1 0 0 0
3  0 1 0 0
4  1 1 0 0
5  0 0 1 0
6  1 0 1 0
7  0 1 1 0
8  1 1 1 0
9  0 0 0 1
10 1 0 0 1
11 0 1 0 1
12 1 1 0 1
13 0 0 1 1
14 1 0 1 1
15 0 1 1 1
16 1 1 1 1
class=design, type= full factorial
```

실험의 표준 순서는 다음과 같이 `yates()` 함수를 이용하여 알 수 있다.

```
yates(rep(0,16))
```

```
   A    B   AB    C   AC   BC  ABC    D   AD   BD  ABD   CD  ACD  BCD ABCD
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
attr(,"mean")
```

```
0
```

그리고 처리의 표준 순서로 실험값과 블럭변수를 입력한다. 그리고 처리조합 정보 X 와 결합하여 최종 자료인 df 를 만들자.

예제 8.2 2^4요인배치법을 2^2개의 블록으로 나누어 실시. 상호작용효과 ACD, BCD를 블록과 교락. 정의대비 $I = ACD = BCD = AB$ 선형표현식을 이용하여 배치가 올바른지 확인 & 분산분석표 작성.

블록1	블록2	블록3	블록4
(1)=45	b=48	a=71	ab=65
abc=65	ac=60	bc=80	c=68
abd=104	ad=100	bd=45	d=43
cd=75	bcd=70	acd=86	abcd=96
$(L_1 = L_2 = 0)$	$(L_1 = 0, L_2 = 1)$	$(L_1 = 1, L_2 = 0)$	$(L_1 = L_2 = 1)$

< 그림 8.5 > ACD, BCD를 교락시킨 2^4요인배치법의 교락법

그림 6.1.: 예제 8.2 자료

```
y <- c(45,71, 48, 65, 68, 60, 80, 65, 43, 100, 45, 104, 75, 86, 70, 96 )
block <- factor(c(1, 3, 2, 4, 4, 2, 3, 1, 4, 2, 3, 1, 1, 3, 2, 4))
treat <- c("0", names(yates(rep(0,16))))
df0 <- data.frame(block=block, treat=treat,  y=y )
df <- cbind(X, df0)
```

```
df %>%  kbl() %>%   kable_paper("hover", full_width = F)
```

위의 실험 자료를 블럭과 처리 순으로 정렬해 보자.

```
df %>% arrange( block, treat)
```

```
   A B C D block treat   y
1  0 0 0 0     1     0  45
2  1 1 1 0     1   ABC  65
3  1 1 0 1     1   ABD 104
4  0 0 1 1     1    CD  75
5  1 0 1 0     2    AC  60
6  1 0 0 1     2    AD 100
7  0 1 0 0     2     B  48
```

A	B	C	D	block	treat	y
0	0	0	0	1	0	45
1	0	0	0	3	A	71
0	1	0	0	2	B	48
1	1	0	0	4	AB	65
0	0	1	0	4	C	68
1	0	1	0	2	AC	60
0	1	1	0	3	BC	80
1	1	1	0	1	ABC	65
0	0	0	1	4	D	43
1	0	0	1	2	AD	100
0	1	0	1	3	BD	45
1	1	0	1	1	ABD	104
0	0	1	1	1	CD	75
1	0	1	1	3	ACD	86
0	1	1	1	2	BCD	70
1	1	1	1	4	ABCD	96

```
8  0 1 1 1     2   BCD  70
9  1 0 0 0     3     A  71
10 1 0 1 1     3   ACD  86
11 0 1 1 0     3    BC  80
12 0 1 0 1     3    BD  45
13 1 1 0 0     4    AB  65
14 1 1 1 1     4  ABCD  96
15 0 0 1 0     4     C  68
16 0 0 0 1     4     D  43
```

6.1.2. 선형표현식

이제 위에서 생성된 자료에서 상호작용효과 ACD 와 BCD 가 블록과 교락되어 있는지 선형표현식을 이용하여 확인해 보자.

$$L_1 = x_1 + x_3 + x_4 (\text{mod} 2)$$
$$L_2 = x_2 + x_3 + x_4 (\text{mod} 2)$$

예를 들어 블럭 1 ($L_1 = 0, L_2 = 0$) 에 베치된 처리에 대하여 선형식의 값을 구해보자.

블럭	처리	$L_1(ACD)$	$L_2(BCD)$
1	(0)	$\texttt{MOD}(0+0+0,2) = 0$	$\texttt{MOD}(0+0+0,2) = 0$
1	ABC	$\texttt{MOD}(1+1+0,2) = 0$	$\texttt{MOD}(1+1+0,2) = 0$
1	ABD	$\texttt{MOD}(1+0+1,2) = 0$	$\texttt{MOD}(1+0+1,2) = 0$
1	CD	$\texttt{MOD}(0+1+1,2) = 0$	$\texttt{MOD}(0+1+1,2) = 0$

예를 들어 블럭 2 ($L_1 = 0, L_2 = 1$) 에 베치된 처리에 대하여 선형식의 값을 구해보자.

블럭	처리	$L_1(ACD)$	$L_2(BCD)$
2	AC	$\texttt{MOD}(1+1+0,2) = 0$	$\texttt{MOD}(0+1+0,2) = 1$

블럭	처리	$L_1(ACD)$	$L_2(BCD)$
2	AD	$\texttt{MOD}(1+0+1,2)=0$	$\texttt{MOD}(0+0+1,2)=1$
2	B	$\texttt{MOD}(0+0+0,2)=0$	$\texttt{MOD}(1+0+0,2)=1$
2	BCD	$\texttt{MOD}(0+1+1,2)=0$	$\texttt{MOD}(1+1+1,2)=1$

이렇게 모든 처리에 대하여 구한 선형식의 값은 다음과 같이 함수 `conf.design()` 로 구할 수 있다. 아래 주어진 블럭배치의 결과는 자료 `df` 에서 처리들이 블럭에 배치된 것과 일치함을 확인할 수 있다.

먼저 상호작용효과 ACD 와 BCD 가 블록과 교락되도록 정의할 수 있는 행렬을 만들어 보자. 아래 0 과 1 로 구성된 4×2 행렬을 보면, 먼저 4개의 인자를 나타내는 4개의 열로 구성되어 있으며 각 행은 블록과 교락된 상호작용효과에 해당되는 요인을 1로 표시한 것이다. 따라서 첫 행은 ACD 에 해당되는 요인들을 1로 표시하고 나머지 행은 BCD 에 해당되는 요인들을 1로 표시한 것이다.

A	B	C	D
1	0	1	1
0	1	1	1

```
Def.contrast <-   matrix(c(1,0,1,1, 0,1,1,1), 2,4, byrow=TRUE)
Def.contrast
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    0    1    1
[2,]    0    1    1    1
```

다음으로 위에서 정의한 행렬을 이용하여 상호작용효과 ACD 와 BCD 가 블록과 교락되도록 실험을 설계해 보자. 아래 함수 `conf.design()` 를 사용하는 경우 인자들의 역활은 다음과 같다.

- `G` : 상호작용효과를 정의하는 행렬
- `p` : 실험의 수준
- `block.name` : 블럭의 이름
- `treatment.names` : 처리의 이름

```
df2 <- conf.design(G=Def.contrast , p=2, block.name = "블럭", treatment.names=c("A", "B", "C", "D"))
df2
```

```
   블럭 A B C D
1    00 0 0 0 0
2    00 1 1 1 0
3    00 1 1 0 1
4    00 0 0 1 1
5    01 0 1 0 0
6    01 1 0 1 0
7    01 1 0 0 1
8    01 0 1 1 1
9    10 1 0 0 0
```

```
10   10 0 1 1 0
11   10 0 1 0 1
12   10 1 0 1 1
13   11 1 1 0 0
14   11 0 0 1 0
15   11 0 0 0 1
16   11 1 1 1 1
```

아래 함수 `conf.design()` 의 결과 `df2` 는 앞에서 만든 실험자료 `df` 와 처리와 블럭의배정이 일치하는 것을 확인할 수 있다.

6.1.3. 결합요인

상호작용효과 ACD 와 BCD 가 블록과 교락되어 있을 경우 발생하는 **결합요인**은 AB 이다. 따라서 상호작용효과 AD 도 블록 효과와 교락된다.

$$ACD \times BCD = ABC^2D^2 = AB$$

6.1.4. Yates 계산법

이제 자료 `df` 에 함수 `yates()` 를 다음과 같이 적용하여 각 효과의 추정치(effect)를 계산해 보자. 참고로 `attr(a, "mean")` 는 Yates 추정치가 저장된 `a` 에서 반응값의 전체 평균 $\bar{y}_{....}$ 을 구하는 함수이다.

```
a <- yates(df$y, c("A", "B", "C", "D"))
a
```

```
      A       B      AB       C      AC      BC     ABC       D      AD      BD
 21.625   3.125   0.125   9.875 -18.125   2.375   1.875  14.625  16.625  -0.375
    ABD      CD     ACD     BCD    ABCD
  4.125  -1.125  -1.625  -2.625   1.375
attr(,"mean")

70.0625
```

```
attr(a, "mean")
```

```
70.0625
```

이제 위의 결과를 이용하여 교과서 표 8.4 와 동일한 Yates 계산의 결과를 구해보자.

```
yates_effect <- data.frame(treat = names(a), effect= a)
yates_effect
```

```
     treat  effect
A        A  21.625
B        B   3.125
AB      AB   0.125
C        C   9.875
AC      AC -18.125
BC      BC   2.375
ABC    ABC   1.875
D        D  14.625
AD      AD  16.625
BD      BD  -0.375
ABD    ABD   4.125
CD      CD  -1.125
ACD    ACD  -1.625
BCD    BCD  -2.625
ABCD  ABCD   1.375
```

```
totalmean <-  data.frame(treat="(0)", effect = attr(a, "mean"))
totalmean
```

```
  treat  effect
1   (0) 70.0625
```

```
yates_effect <- rbind(totalmean, yates_effect)
yates_effect
```

```
     treat   effect
1      (0)  70.0625
A        A  21.6250
B        B   3.1250
AB      AB   0.1250
C        C   9.8750
AC      AC -18.1250
BC      BC   2.3750
ABC    ABC   1.8750
D        D  14.6250
AD      AD  16.6250
BD      BD  -0.3750
ABD    ABD   4.1250
CD      CD  -1.1250
ACD    ACD  -1.6250
BCD    BCD  -2.6250
ABCD  ABCD   1.3750
```

위에서 구한 데이터프레임의 `effect` 는 평균 효과를 의미한다. 예를 들어서 처리 A 에 대한 효과는 다음과 같이 구한다.

$$
\begin{aligned}
A &= \frac{1}{8}(a + ab + ac + abc + ad + abd + acd + abcd - (0) - b - c - bc - d - bd - cd - bcd) \\
&= \frac{1}{8}(T_{1\cdots} - T_{0\cdots}) \\
&= \bar{y}_{1\cdots} - \bar{y}_{0\cdots} \\
&= 21.625
\end{aligned}
$$

따라서 제곱합을 구하는 방법은 처리합의 차를 제곱한 값 $(T_{1\cdots} - T_{0\cdots})^2$ 을 총 실험의 크기 $n = 16$ 으로 나눈다. 이는 평균처리 효과를 제곱한 값에 4를 곱해주는 양과 같다.

$$
SS_A = \frac{(T_{1\cdots} - T_{0\cdots})^2}{16} = 4(\bar{y}_{1\cdots} - \bar{y}_{0\cdots})^2
$$

이제 위에서 구한 평균 처리 효과를 이용하여 제곱합을 구해보자. 주의할 점은 평균 처리 효과가 저장된 `yates_effect` 의 첫 행은 전체평균 $\bar{y}_{\cdots\cdot} = T_{\cdots\cdot}/16$ 이 저장되어 있기 때문에 4를 한번 더 곱해주어야 한다.

$$
CT = \frac{T_{\cdots\cdot}^2}{16} = 16(\bar{y}_{\cdots\cdot})^2
$$

```
yates_effect$SS <- 4*yates_effect$effect^2
yates_effect$SS[1] <- yates_effect$SS[1] *4
yates_effect
```

```
     treat   effect         SS
1      (0)  70.0625 78540.0625
A        A  21.6250  1870.5625
B        B   3.1250    39.0625
AB      AB   0.1250     0.0625
C        C   9.8750   390.0625
AC      AC -18.1250  1314.0625
BC      BC   2.3750    22.5625
ABC    ABC   1.8750    14.0625
D        D  14.6250   855.5625
AD      AD  16.6250  1105.5625
BD      BD  -0.3750     0.5625
ABD    ABD   4.1250    68.0625
CD      CD  -1.1250     5.0625
ACD    ACD  -1.6250    10.5625
BCD    BCD  -2.6250    27.5625
ABCD  ABCD   1.3750     7.5625
```

<표 8.4> Yastes 계산법

처리조합	자 료	(1)	(2)	(3)	(4)	요인효과	변 동
(1)	45	116	229	502	1121	70.063=M	78540.06
a	71	113	273	619	173	21.625=A	1870.56
b	48	128	292	20	25	3.125=B	39.06
ab	65	145	327	153	1	0.125=AB	0.06*
c	68	143	43	14	79	9.875=C	390.06
ac	60	149	-23	11	-145	-18.125=AC	1314.06
bc	80	161	116	-16	19	2.375=BC	22.56
abc	65	166	37	17	15	1.875=ABC	14.06
d	43	26	-3	44	117	14.625=D	855.56
ad	100	17	17	35	133	16.625=AD	1105.56
bd	45	-8	6	-66	-3	-0.375=BD	0.56
abd	104	-15	5	-79	33	4.125=ABD	68.06
cd	75	57	-9	20	-9	-1.125=CD	5.06
acd	86	59	-7	-1	-13	-1.625=ACD	10.56*
bcd	70	11	2	2	-21	-2.625=BCD	27.56*
abcd	96	26	15	13	11	1.375=ABCD	7.56

*표시된 것은 블록과 교락된 것임

그림 6.2.: 예제 8.2 Yates 계산

6.1.5. 블럭변동

자료에서 블럭의 변동을 구하는 방법은 교과서 231 에 나온 것처럼 각 블럭에 대한 관측값의 합을 구해서 변동의 공식을 이용할 수 있다. 각 블럭 안의 관측값들 합을 T_i 라고 하면

$$SS_{block} = \frac{1}{4}\sum_{i=1}^{4} T_i^2 - CT \quad \text{where } CT = \frac{T^2}{(4)(4)}$$

SS_{bbdk} 구하기: 각 블록에서 자료의 합

	블록1	블록2	블록3	블록4
합	289	278	282	272

$$SS_{bbdk} = \frac{1}{4}[(289)^2 + (278)^2 + (282)^2 + (272)^2] - 78540.0625$$
$$= 38.1875$$

그림 6.3.: 예제 8.2 블럭 제곱합

또한 다음과 같은 선형식을 R 록 적합식키고 분산분석을 이용하여 블럭의 변동(SS_{block})을 구할 수 있다. 아래 함수 `anova()` 결과에 의하면 $SS_{block} = 38.19$ 이다.

$$y_{ij} = \mu + (\text{block})_i + e_{ij}$$

```
res_block <- lm(y~block, data=df)
anova(res_block)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
block      3   38.2   12.73  0.0268 0.9937
Residuals 12 5692.8  474.40
```

블럭의 변동은 상호작용 효과 ACD, BCD, AB 에 대한 제곱합들의 합과 같다.

$$SS_{block} = SS_{ACD} + SS_{BCD} + SS_{AB} = 10.5625 + 27.5625 + 0.0625 = 38.19$$

```
yates_effect %>% filter(treat == "ACD" | treat == "BCD" | treat == "AB")
```

```
    treat effect      SS
AB     AB  0.125  0.0625
ACD   ACD -1.625 10.5625
BCD   BCD -2.625 27.5625
```

6.1.6. 핵심요인의 선별

핵심요인의 선별하기 위하여 먼저 각 처리의 제곱합을 순서대로 나열해 보자. 주요인 A, C, D 와 상호작용 효과 AC 와 AD 의 제곱합이 다른 것보다 크게 나타나는 것을 볼 수 있다.

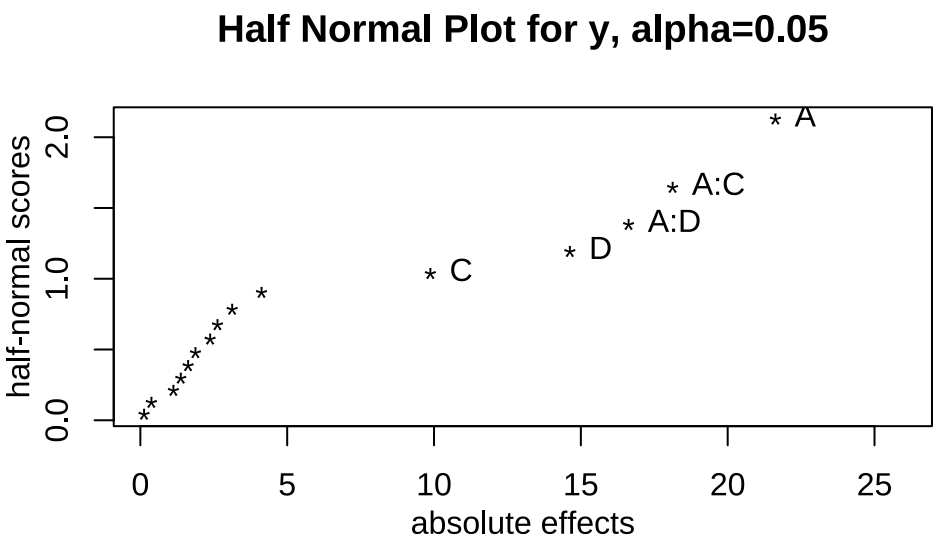
```
yates_effect %>%  arrange(desc(SS))
```

```
     treat   effect         SS
1      (0)  70.0625 78540.0625
A        A  21.6250  1870.5625
AC      AC -18.1250  1314.0625
AD      AD  16.6250  1105.5625
D        D  14.6250   855.5625
C        C   9.8750   390.0625
ABD    ABD   4.1250    68.0625
B        B   3.1250    39.0625
BCD    BCD  -2.6250    27.5625
BC      BC   2.3750    22.5625
ABC    ABC   1.8750    14.0625
ACD    ACD  -1.6250    10.5625
ABCD  ABCD   1.3750     7.5625
```

```
CD      CD  -1.1250     5.0625
BD      BD  -0.3750     0.5625
AB      AB   0.1250     0.0625
```

이제 모든 효과가 포함된 완전모형(full model)을 적합시키고 반정규확률 그림을 그려서 핵심효인을 다시 찾아보자. 제곱합을 비교할 때와 같이 주요인 A, C, D 와 상호작용 효과 AC 와 AD이 핵심 요인으로 보여진다.

```
fullmodel <- lm (y~ A*B*C*D, data=df)
DanielPlot(fullmodel, half=TRUE)
```



6.1.7. 최종 모형

위의 핵심요인의 선별 결과를 고려하여 주요인 A, B, C, D 와 상호작용 효과 AC 와 AD 를 포함하는 축소된 모형을 최종모형으로 적합해 보자. 축소모형에는 당연히 블럭효과도 포함해야 한다. 또한 오차항에는 블럭과 교럭된 상호작용 효과들과 축소모형에 포함된 효과들을 제외한 다른 효과들이 풀링된다.

$$SS_E = SS_{B\times C} + SS_{B\times D} + SS_{C\times D} + SS_{A\times B\times C} + SS_{A\times B\times D} + SS_{A\times B\times C\times D}$$

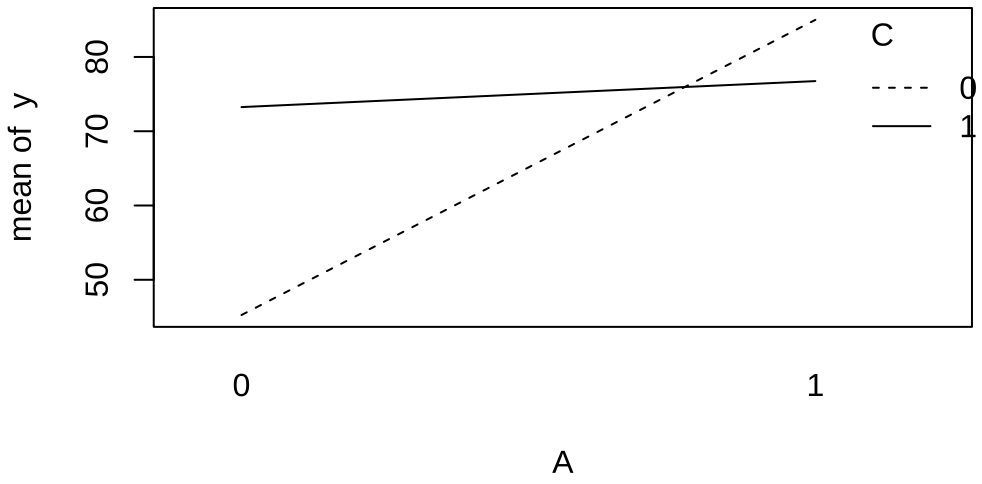
```
finalmodel <- lm(y ~ block + A +B + C+ D + A:C + A:D, data=df)
anova(finalmodel)
```

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
block      3   38.19   12.73  0.6479 0.6123498
A          1 1870.56 1870.56 95.2142  6.66e-05 ***
```

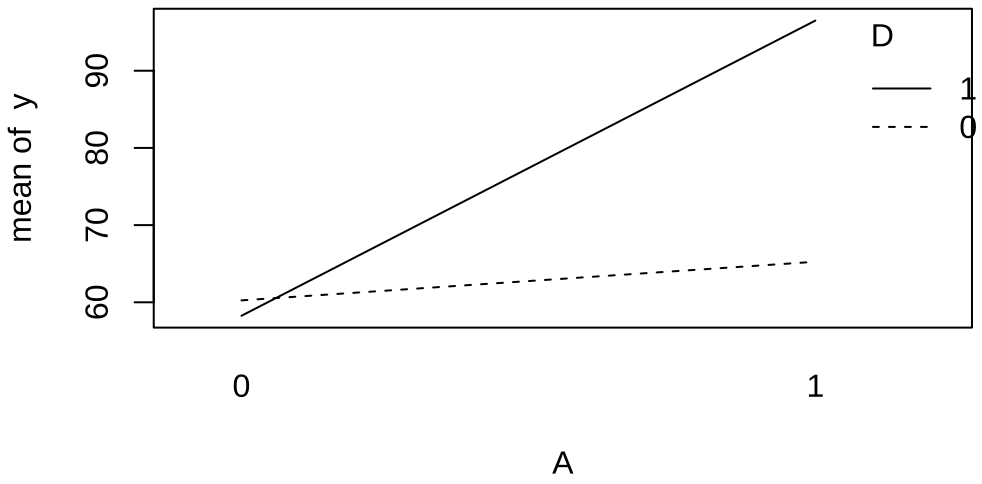
```
B           1   39.06   39.06  1.9883 0.2081893
C           1  390.06  390.06 19.8547 0.0043020 **
D           1  855.56  855.56 43.5493 0.0005821 ***
A:C         1 1314.06 1314.06 66.8876 0.0001800 ***
A:D         1 1105.56 1105.56 56.2747 0.0002902 ***
Residuals   6  117.88   19.65
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.2. 상호작용 그림

```
with(df, interaction.plot(x.factor = A, trace.factor = C,  response = y))
```




```
with(df, interaction.plot(x.factor = A, trace.factor = D,  response = y))
```



6.3. 8장 연습문제

다음은 8장의 연습문제를 풀 때 도움이 되는 R 프로그램이다.

6.3.1. 연습문제 1

```
X1 <- FrF2(nruns=16, nfactors=4, randomize = FALSE,  default.level=c(0,1))
X1
```

```
   A B C D
1  0 0 0 0
2  1 0 0 0
3  0 1 0 0
4  1 1 0 0
5  0 0 1 0
6  1 0 1 0
7  0 1 1 0
8  1 1 1 0
9  0 0 0 1
10 1 0 0 1
11 0 1 0 1
12 1 1 0 1
13 0 0 1 1
14 1 0 1 1
15 0 1 1 1
16 1 1 1 1
```

```
class=design, type= full factorial
```

```
yates(rep(0,16))
```

```
   A    B   AB    C   AC   BC  ABC    D   AD   BD  ABD   CD  ACD  BCD ABCD
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
attr(,"mean")

0
```

```
Def.contrast1 <- c(1,1,1,1)
conf.design(G=Def.contrast1 , p=2, block.name = "블럭", treatment.names=c("A", "B", "C", "D"))
```

```
   블럭 A B C D
1     0 0 0 0 0
2     0 1 1 0 0
3     0 1 0 1 0
4     0 0 1 1 0
5     0 1 0 0 1
6     0 0 1 0 1
7     0 0 0 1 1
8     0 1 1 1 1
9     1 1 0 0 0
10    1 0 1 0 0
11    1 0 0 1 0
12    1 1 1 1 0
13    1 0 0 0 1
14    1 1 1 0 1
15    1 1 0 1 1
16    1 0 1 1 1
```

6.3.2. 연습문제 4

```
X2 <- FrF2(nruns=8, nfactors=3, randomize = FALSE,  default.level=c(0,1))
X2
```

```
  A B C
1 0 0 0
2 1 0 0
3 0 1 0
4 1 1 0
5 0 0 1
6 1 0 1
7 0 1 1
8 1 1 1
class=design, type= full factorial
```

```
yates(rep(0,8))
```

```
  A   B  AB   C  AC  BC ABC
  0   0   0   0   0   0   0
attr(,"mean")

0
```

```
Def.contrast2 <- c(1,1,1)
conf.design(G=Def.contrast2 , p=2, block.name = "블럭", treatment.names=c("A", "B", "C"))
```

```
  블럭 A B C
1    0 0 0 0
2    0 1 1 0
3    0 1 0 1
4    0 0 1 1
5    1 1 0 0
6    1 0 1 0
7    1 0 0 1
8    1 1 1 1
```

6.3.3. 연습문제 10

```
X3 <- FrF2(nruns=64, nfactors=6, randomize = FALSE,  default.level=c(0,1))
X3
```

```
   A B C D E F
1  0 0 0 0 0 0
2  1 0 0 0 0 0
3  0 1 0 0 0 0
4  1 1 0 0 0 0
5  0 0 1 0 0 0
6  1 0 1 0 0 0
7  0 1 1 0 0 0
8  1 1 1 0 0 0
9  0 0 0 1 0 0
10 1 0 0 1 0 0
11 0 1 0 1 0 0
12 1 1 0 1 0 0
13 0 0 1 1 0 0
14 1 0 1 1 0 0
15 0 1 1 1 0 0
16 1 1 1 1 0 0
17 0 0 0 0 1 0
```

18 1 0 0 0 1 0
19 0 1 0 0 1 0
20 1 1 0 0 1 0
21 0 0 1 0 1 0
22 1 0 1 0 1 0
23 0 1 1 0 1 0
24 1 1 1 0 1 0
25 0 0 0 1 1 0
26 1 0 0 1 1 0
27 0 1 0 1 1 0
28 1 1 0 1 1 0
29 0 0 1 1 1 0
30 1 0 1 1 1 0
31 0 1 1 1 1 0
32 1 1 1 1 1 0
33 0 0 0 0 0 1
34 1 0 0 0 0 1
35 0 1 0 0 0 1
36 1 1 0 0 0 1
37 0 0 1 0 0 1
38 1 0 1 0 0 1
39 0 1 1 0 0 1
40 1 1 1 0 0 1
41 0 0 0 1 0 1
42 1 0 0 1 0 1
43 0 1 0 1 0 1
44 1 1 0 1 0 1
45 0 0 1 1 0 1
46 1 0 1 1 0 1
47 0 1 1 1 0 1
48 1 1 1 1 0 1
49 0 0 0 0 1 1
50 1 0 0 0 1 1
51 0 1 0 0 1 1
52 1 1 0 0 1 1
53 0 0 1 0 1 1
54 1 0 1 0 1 1
55 0 1 1 0 1 1
56 1 1 1 0 1 1
57 0 0 0 1 1 1
58 1 0 0 1 1 1
59 0 1 0 1 1 1
60 1 1 0 1 1 1
61 0 0 1 1 1 1
62 1 0 1 1 1 1
63 0 1 1 1 1 1
64 1 1 1 1 1 1

```
class=design, type= full factorial
```

```
yates(rep(0,64))
```

```
     A      B     AB      C     AC     BC    ABC      D     AD     BD    ABD
     0      0      0      0      0      0      0      0      0      0      0
    CD    ACD    BCD   ABCD      E     AE     BE    ABE     CE    ACE    BCE
     0      0      0      0      0      0      0      0      0      0      0
  ABCE     DE    ADE    BDE   ABDE    CDE   ACDE   BCDE  ABCDE      F     AF
     0      0      0      0      0      0      0      0      0      0      0
    BF    ABF     CF    ACF    BCF   ABCF     DF    ADF    BDF   ABDF    CDF
     0      0      0      0      0      0      0      0      0      0      0
  ACDF   BCDF  ABCDF     EF    AEF    BEF   ABEF    CEF   ACEF   BCEF  ABCEF
     0      0      0      0      0      0      0      0      0      0      0
   DEF   ADEF   BDEF  ABDEF   CDEF  ACDEF  BCDEF ABCDEF
     0      0      0      0      0      0      0      0
attr(,"mean")

0
```

```
Def.contrast3 <-   matrix(c(1,1,1,1,0,0, 1,1,0,0,1,1), 2,6, byrow=TRUE)
conf.design(G=Def.contrast3 , p=2, block.name = "블럭", treatment.names=c("A", "B", "C", "D", "E", "F"))
```

```
   블럭 A B C D E F
1    00 0 0 0 0 0 0
2    00 1 1 0 0 0 0
3    00 0 0 1 1 0 0
4    00 1 1 1 1 0 0
5    00 1 0 1 0 1 0
6    00 0 1 1 0 1 0
7    00 1 0 0 1 1 0
8    00 0 1 0 1 1 0
9    00 1 0 1 0 0 1
10   00 0 1 1 0 0 1
11   00 1 0 0 1 0 1
12   00 0 1 0 1 0 1
13   00 0 0 0 0 1 1
14   00 1 1 0 0 1 1
15   00 0 0 1 1 1 1
16   00 1 1 1 1 1 1
17   01 1 0 1 0 0 0
18   01 0 1 1 0 0 0
19   01 1 0 0 1 0 0
20   01 0 1 0 1 0 0
21   01 0 0 0 0 1 0
22   01 1 1 0 0 1 0
```

23	01	0	0	1	1	1	0
24	01	1	1	1	1	1	0
25	01	0	0	0	0	0	1
26	01	1	1	0	0	0	1
27	01	0	0	1	1	0	1
28	01	1	1	1	1	0	1
29	01	1	0	1	0	1	1
30	01	0	1	1	0	1	1
31	01	1	0	0	1	1	1
32	01	0	1	0	1	1	1
33	10	0	0	1	0	0	0
34	10	1	1	1	0	0	0
35	10	0	0	0	1	0	0
36	10	1	1	0	1	0	0
37	10	1	0	0	0	1	0
38	10	0	1	0	0	1	0
39	10	1	0	1	1	1	0
40	10	0	1	1	1	1	0
41	10	1	0	0	0	0	1
42	10	0	1	0	0	0	1
43	10	1	0	1	1	0	1
44	10	0	1	1	1	0	1
45	10	0	0	1	0	1	1
46	10	1	1	1	0	1	1
47	10	0	0	0	1	1	1
48	10	1	1	0	1	1	1
49	11	1	0	0	0	0	0
50	11	0	1	0	0	0	0
51	11	1	0	1	1	0	0
52	11	0	1	1	1	0	0
53	11	0	0	1	0	1	0
54	11	1	1	1	0	1	0
55	11	0	0	0	1	1	0
56	11	1	1	0	1	1	0
57	11	0	0	1	0	0	1
58	11	1	1	1	0	0	1
59	11	0	0	0	1	0	1
60	11	1	1	0	1	0	1
61	11	1	0	0	0	1	1
62	11	0	1	0	0	1	1
63	11	1	0	1	1	1	1
64	11	0	1	1	1	1	1

7. 반응표면 분석

7.1. 반응표면 분석 개요

7.1.1. 실험계획의 절차

- 요인배치법의 실험 자료의 분석은 실험조건인 처리에서의 모평균 비교.
- 분산분석을 통하여 유의한 요인효과들을 선별
- 각 처리에서 예측치 구하기.
- 실험에서 고려된 처리조건들 중에서 **모평균의 값을 최적으로 하는 실험조건 찾기** 및 재현성 검토

7.1.2. 반응표면분석의 목적

반응표면분석(response surface method)에서는 관심영역에 속한 임의의 계량인자(quantatitative factor)들의 값에서 수율(반응변수)의 예측이 실험 목적이다.

일단 반응표변분석에서는 두 개의 계량인자 x_1, x_2 와 반응변수 y 의 모평균 $\eta = E(y)$이 다음과 같은 함수관계를 가진다고 가정한다.

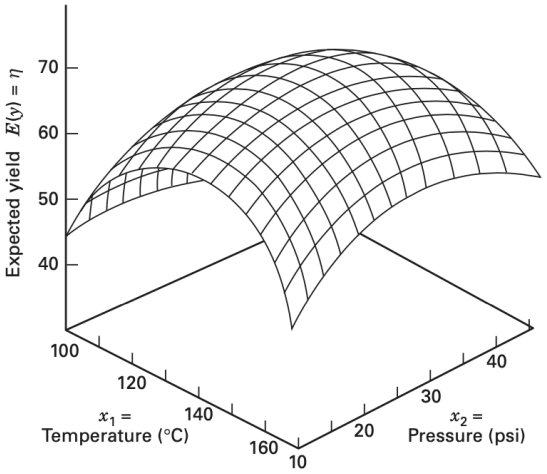
$$\eta = f(x_1, x_2)$$

위의 관계에서 함수 $f(x_1, x_2)$는 모르는 함수이다.

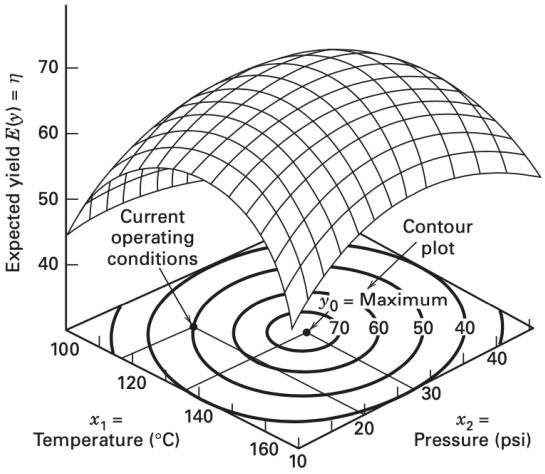
반응표면분석의 목적은 반응변수의 모평균 η의 최대값 η^* 가 나타나는 계량인자의 수준 x_1^* 와 x_2^* 를 주어진 영역에서 찾는 것이다.

$$\eta^* = \max_{x_1, x_2} f(x_1, x_2) = f(x_1^*, x_2^*)$$

함수 $f(x_1, x_2)$는 복잡한 형태를 가질 수 있지만 반응표변분석에서는 최대값의 주위에서 함수 $f(x_1, x_2)$를 **이차함수로 근사**하여 최적점을 찾는다.



■ **FIGURE 11.1** **A three-dimensional response surface showing the expected yield (η) as a function of temperature (x_1) and pressure (x_2)**



■ **FIGURE 11.2** **A contour plot of a response surface**

그림 7.1.: 이차함수 근사 (Montgomery 2017)

7.1.3. 반응표면분석의 절차

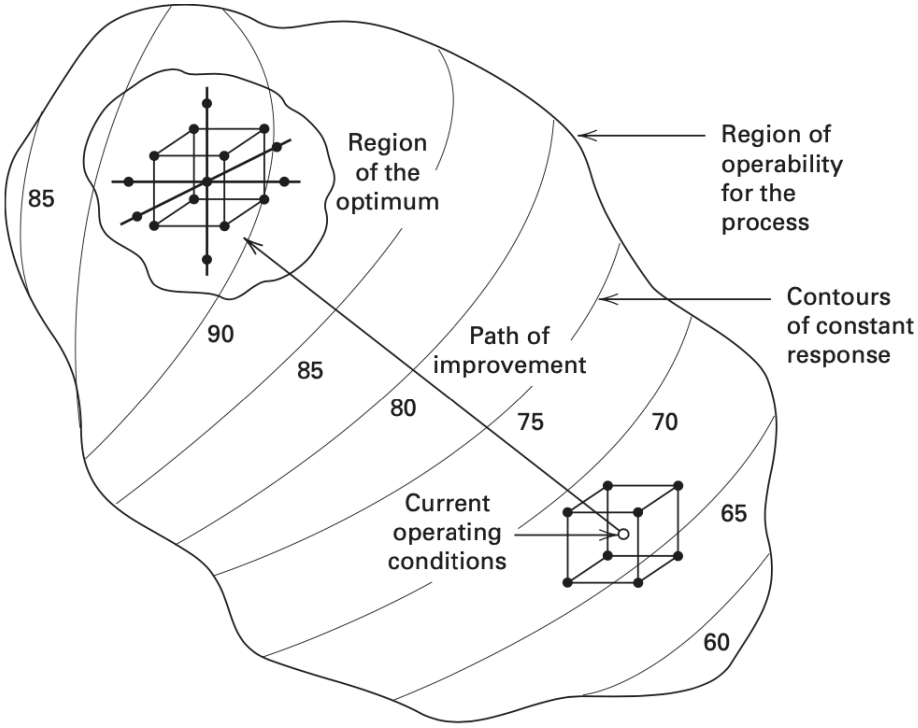
- 반응값을 나타내는 변수를 반응변수로, 계량인자를 설명변수로 간주하여 실험 자료를 이용하여 회귀분석을 실시.
  - 관심영역의 최적조건 근처에서 곡선효과 존재
  - 곡선효과를 반영하는 가장 간결한 모형인 이차 다항모형 가정
  - 적절한 모형 찾기.
- 예측치를 최적으로 하는 최적조건을 관심영역에서 찾고, 최적조건에서의 재현성 검토하기.
- 반응표면분석의 절차는 다음과 같이 크게 3단계로 구성된다.

1. 2수준 일부실시법에 의해서 핵심인자들을 선별하기.
2. 선별된 핵심인자들에 대한 축차적인 실험 설계(중심점을 갖는 2수준 요인 배치법)와 분석에 의해서 최적조건 근처의 설명변수들의 영역 으로 이동하기. (최대경사법 적용)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$$

3. 최적조건 근처에서 2차 모형을 가정, ccd(중심합성설계)에 의한 실험설계 및 실험자료의 회귀분석을 통한 적절한 모형 찾기 및 최적조 건 찾기와 재현성 검토.

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum\sum_{i<j} \beta_{ij} x_i x_j + e$$



■ **FIGURE 11.3 The sequential nature of RSM**

그림 7.2.: 반응표면분석의 순차적인 특성 (Montgomery 2017)

7.2. 최대경사법

7.2.1. 개요

- 초기 실험에서는 실험에서 고려하는 계량인자들의 관심 영역이 반응변수의 최적값을 가지는 영역에서 멀리 떨어져 있는 경우가 흔하다.
- 따라서 초기 실험에서는 최적의 영역으로 이동하기 위한 계량인자들의 방향을 알아내는 것이 중요하다.
- **간단한 초기 실험의 결과를 이용하여 계량인자들의 값을 최적점 근처로 순차적으로 이동시킬 수 있는 쉽고 경제적인 절차가 필요하다.**
- 주어진 선형모형에서 계량인자들의 값을 이동시킬 수 있는 간단한 방법들 중의 하나는 **최대경사법(method of steepest ascent)** 이다.

7.2.2. 9.2절 자료와 변환

교과서 9.2 절에 나오는 실험 자료를 고려해 보자. 화학공장에서 공정의 수율(y)을 최적화하는 공정조건을 찾는 실험자가 관심이 있는 시간(x_1, `time`)의 범위는 30-40분 이고 온도(x_2, `temp`)의 범위는 160-180도 라고 하자. 이제 주어진 시간과 온도에 따라서 실험을 7번 수행하였으며 결과 자료는 다음과 같다.

```
time <- c(30,30,40,40,35,35,35)
temp <- c(160, 180, 160, 180, 170, 170, 170)
y <- c(72.5, 74.2, 76.3, 77.0, 74.8, 75.6, 75.2)
df1 <- data.frame(time=time, temp=temp, y=y)
df1
```

```
  time temp    y
1   30  160 72.5
2   30  180 74.2
3   40  160 76.3
4   40  180 77.0
5   35  170 74.8
6   35  170 75.6
7   35  170 75.2
```

반응표면분석에서는 고려하는 변량의 범위를 $(-1, 1)$ 로 변환하는 작업을 먼저 수행한다. 위의 자료에서 시간과 온도의 범위를 $(-1, 1)$로 변환하기 위하여 다음과 같은 식을 적용한다.

$$x_1 = \frac{\text{time} - 35}{5}, \quad x_2 = \frac{\text{temp} - 170}{10}$$

위와 같은 변환은 패키지 `rsm` 에 있는 함수 `coded.data`를 사용해서 쉽게 수행할 수 있다.

```
df11 <- coded.data(df1, x1 ~ (time - 35)/5, x2 ~ (temp - 170)/10)
class(df11)
```

```
[1] "coded.data" "data.frame"
```

```
df11
```

```
  time temp    y
1   30  160 72.5
2   30  180 74.2
3   40  160 76.3
4   40  180 77.0
5   35  170 74.8
6   35  170 75.6
7   35  170 75.2

Data are stored in coded form using these coding formulas ...
x1 ~ (time - 35)/5
x2 ~ (temp - 170)/10
```

위에서 변환된 자료 `df11` 는 변환된 이름 `x1` 과 `x2` 로 자료가 저장되어 있지만 원래 변수의 이름 `time` 과 `temp` 도 특정한 함수에서 사용이 가능한다.

9.2 장 자료에서 7번 실험을 수행하는 경우 고려한 실험점(experiment point)은 정사각형의 각 꼭지점에서 하나의 관측값을 얻고 중심점 $(0, 0)$ 에서 3개의 실험값을 얻었다.

```
df11 %>% ggplot(aes(time, temp)) + geom_point(colour = "red", size = 2)  + theme_bw()
df11 %>% ggplot(aes(x1, x2)) + geom_point(colour = "red", size = 2)  + theme_bw()
```

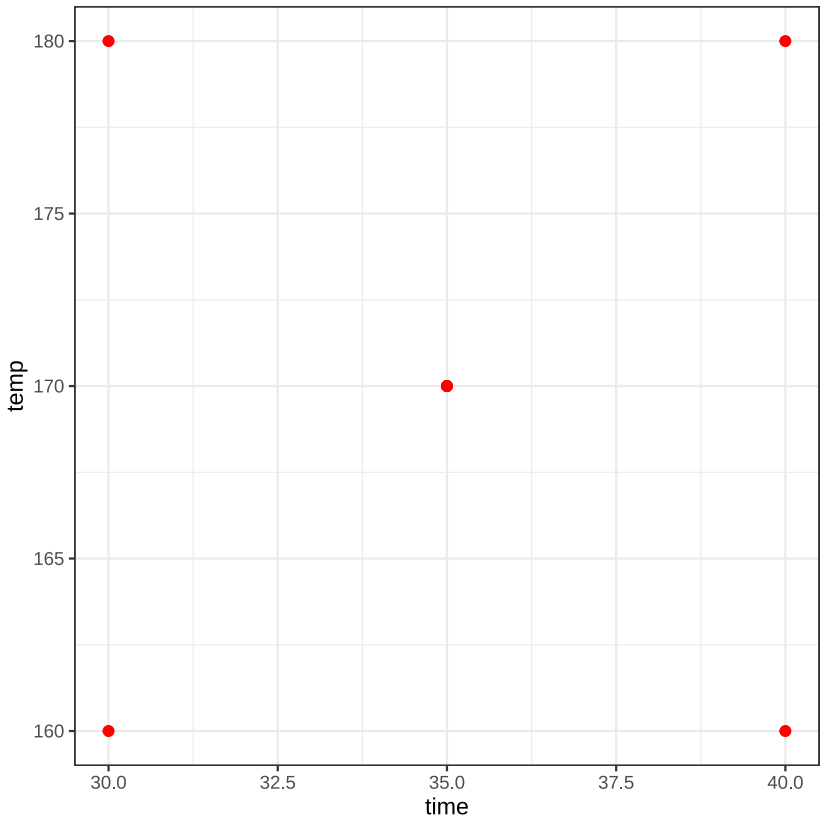


그림 7.3.: 실험점의 배치

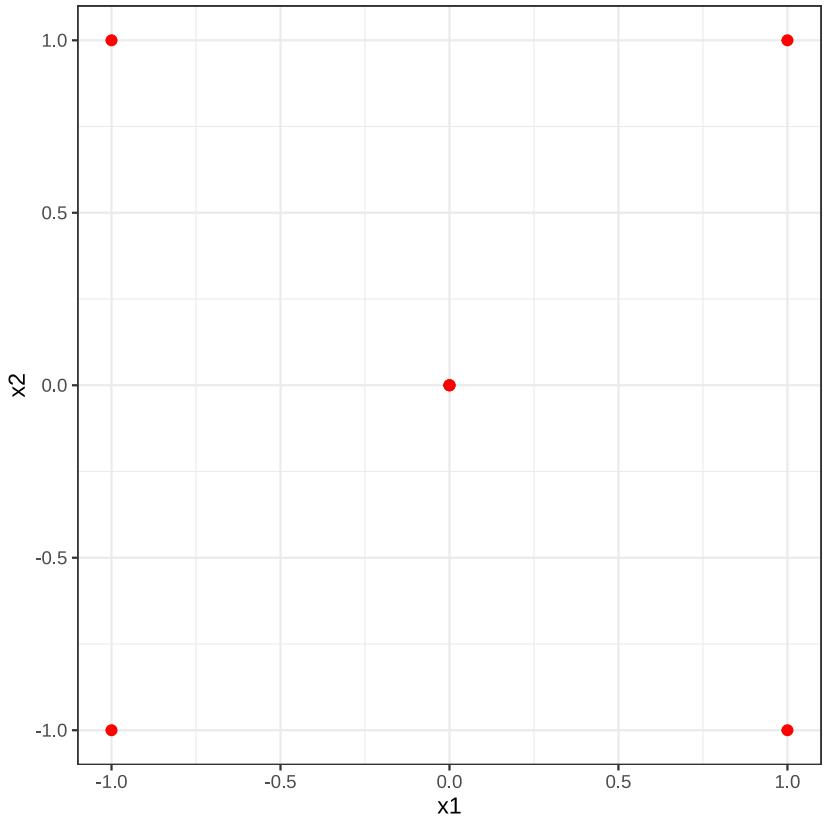


그림 7.4.: 실험점의 배치

변환된 자료 `x1` 과 `x2` 의 값으로 부터 원래 자료의 값을 얻는 방법은 함수 `code2val()`을 사용한다. 예를 들어 `x1 = 0.5` , `x2 = 0.0` 으로 코딩된 원래 자료 `time` 과 `temp` 의 값은 다음과 같이 얻을 수 있다.

```
code2val(data.frame(x1 = c(0.5), x2 = c(0.0)), codings(df11))
```

```
  time temp
1 37.5  170
```

7.2.3. 선형회귀식

다음은 반응변수 y 를 예측하는 모형으로 두 개의 변량인자 x_1 과 x_2 를 가지는 다음과 같은 일차선형 모형을 고려하자.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{7.1}$$

주어진 자료를 이용하여 위의 일차선형 모형의 회귀계수들을 추정하면 반응변수의 평균 $E(y|x)$ 에 대한 다음과 같은 예측식을 얻을 수 있다.

$$\widehat{E(y|x)} = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \tag{7.2}$$

변환된 x_1 과 x_2 가 저장된 데이터 프레임 `df11` 에 있는 자료를 이용하여 일차선형 모형을 적합한 결과는 다음과 같다.

```
lmdf11 <- lm(y ~ x1 + x2, data=df11)
summary(lmdf11)
```

```
Call:
lm.default(formula = y ~ x1 + x2, data = df11)

Residuals:
      1       2       3       4       5       6       7
-0.3357  0.1643  0.1643 -0.3357 -0.2857  0.5143  0.1143

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.0857     0.1510 497.201 9.82e-11 ***
x1            1.6500     0.1998   8.259  0.00117 **
x2            0.6000     0.1998   3.003  0.03981 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3996 on 4 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9261
F-statistic: 38.62 on 2 and 4 DF,  p-value: 0.002425
```

예측식 식 7.2 를 3차원 공간에 표시하면 다음과 같은 평면으로 나타나게 된다. 패키지 scatterplot3d 의 함수 scatterplot3d()를 이용하여 3차원 산포도와 추정된 선형식을 그릴 수 있다.

```
s3d <- scatterplot3d(as.data.frame(df11), type = "h", color = "blue", angle=55, pch = 16, xlab="x1", yl
s3d$plane3d(lmdf11)
```

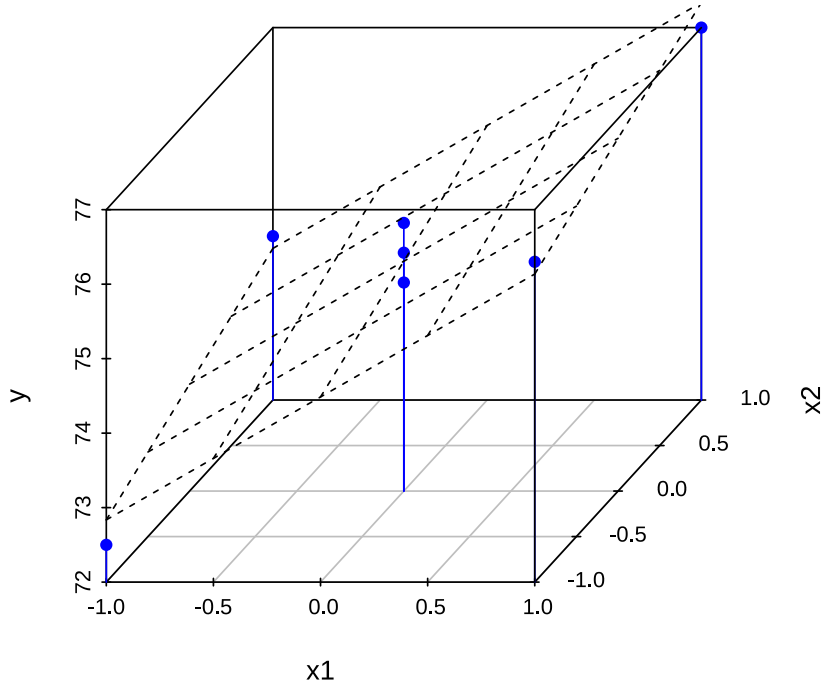


그림 7.5.: 일차 선형식

위에서 적합된 예측식 식 7.2 을 함수 `contour.lm()`을 이용하여 등고선 그림으로 나타내면 다음 그림과 같다.

```
contour.lm(lmdf11, x2~x1, labcex=1.4 )
points(c(-1,-1,1,1, 0), c(-1,1,-1,1,0), col="blue", cex=1.3, pch=19)
```

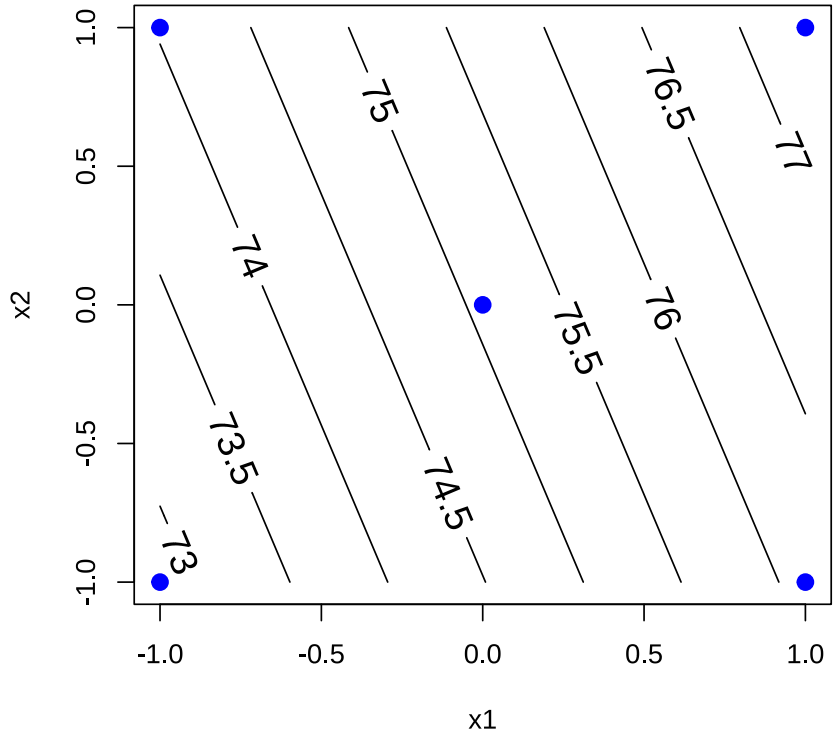


그림 7.6.: 일차 선형식

7.2.4. 최대경사법

이제 반응변수의 변화에 대한 선형 예측식이 식 7.2 로 주어진 경우 반응변수가 가장 크게 증가하는 방향을 2차원 공간 (x_1, x_2) 에서 찾아보자. 이렇게 공간에서 다변량 함수의 변화가 가장 크게 변하는 방향을 찾는 방법을 최대경사법(method of steepest ascent)이라고 부르며 이는 기울기 하강법(gradient descent method)의 반대 방법이다.

이변량 함수 $f(x_1, x_2)$ 에 대한 기울기 벡터(gradient) ∇f는 다음과 같이 각 축에 대한 부분 미분(partial derivative)로 이루어진 벡터이다.

$$\nabla f = \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f(x_1, x_2)}{\partial x_1} \\ \frac{\partial f(x_1, x_2)}{\partial x_2} \end{bmatrix}$$

기울기 벡터 ∇f 에 대한 기하학적 의미는 다음과 같은 그림으로 나타낼 수 있으며 ∇f는 주어진 점에서 함수 f가 가장 빨리 증가하는 방향을 의미한다. 따라서 기울기 벡터 ∇f 를 최대 경사(steepest ascent) 벡터라고 부른다.

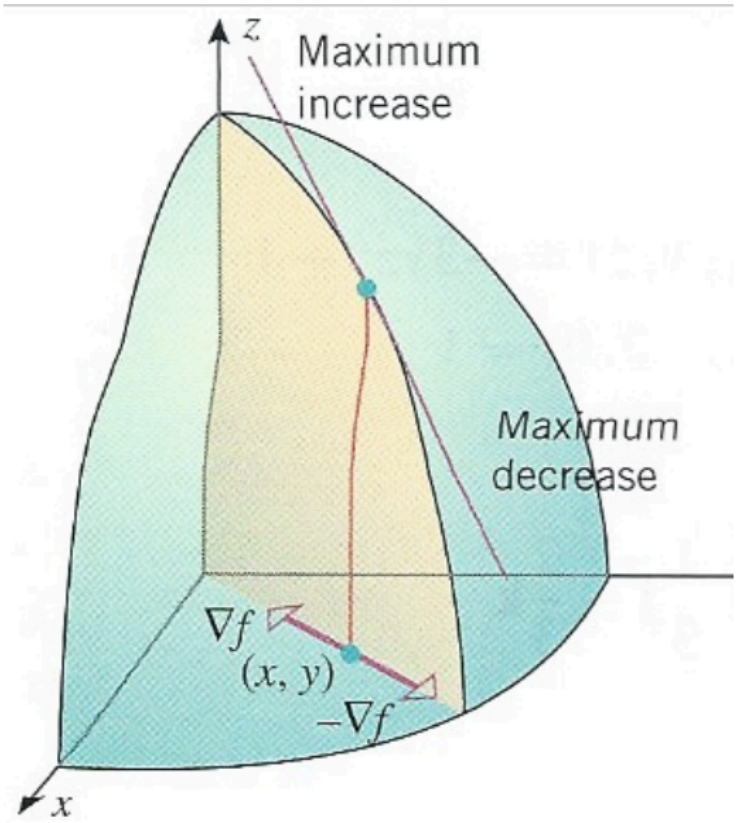


그림 7.7.: 기울기 벡터의 기하학적 의미

선형 예측식 식 7.2 이 주어진 경우 2차원 공간 (x_1, x_2)에서 반응변수가 가장 크게 증가하는 방향, 즉 최대경사 방향 ∇f 은 다음과 같이 주어진다.

$$\nabla f = \begin{bmatrix} \frac{\partial f(x_1,x_2)}{\partial x_1} \\ \frac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}{\partial x_1} \\ \frac{\partial(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \tag{7.3}$$

따라서 9.2절의 실험자료에 대하여 적합한 모형식을 이용하면 최대경사 벡터 ∇f 는 다음과 같이 주어진다.

$$\nabla f = \begin{bmatrix} 1.65 \\ 0.6 \end{bmatrix} \tag{7.4}$$

위의 최대경사 벡터 ∇f은 적합한 모형식에서 x_1과 x_2 가 1 단위 만큼 증가할 때 반응변수의 변화를 나타내는 회귀계수 $\hat{\beta}_1$, $\hat{\beta}_2$의 방향을 의미한다.

```
coef(lmdf11)
```

```
(Intercept)          x1          x2
   75.08571     1.65000     0.60000
```

식 식 7.4 에 나타난 최대경사 벡터 ∇f 을 등고선 그림에 나타내면 다음과 같다. 아래의 등고선에서 최대경사 방향은 기울기가 $0.6/1.65 = 0.364$ 를 가진 원점을 지나는 직선 방향이다.

```
contour.lm(lmdf11, x2~x1, xlabs=c( "x1", "x2"), bounds = list(x1 = c(-3, 3), x2 = c(-3, 3)))
points(c(-1,-1,1,1, 0), c(-1,1,-1,1,0), col="blue", cex=1.3, pch=19)
slope <- as.numeric( coef(lmdf11)["x2"] / coef(lmdf11)["x1"])
slope
```

```
[1] 0.3636364
```

```
lines(x=c(0,1), y=c(0,slope ), col="red")
abline(h=0, lty=2)
abline(v=0, lty=2)
```

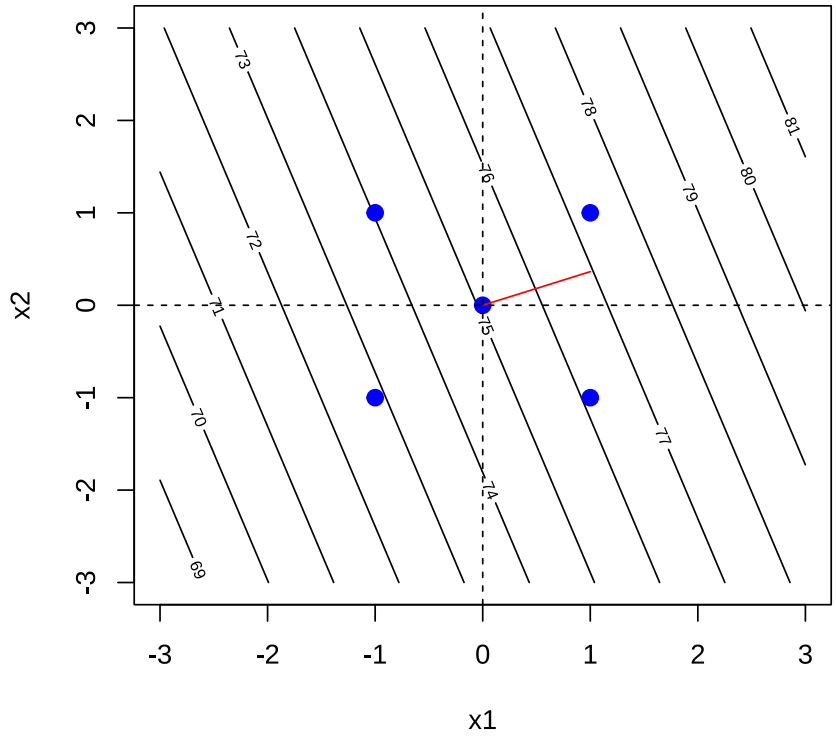


그림 7.8.: 일차 선형식(빨간선이 최대경사 방향)

최대경사 벡터를 구하는 경우 벡터의 길이가 1 이 되도록 하는 경우도 있다. 이러한 경우의 최대경사 방향의 벡터는 다음과 같다.

$$\begin{bmatrix} 1.65/\sqrt{1.65^2+0.6^2} \\ 0.60/\sqrt{1.65^2+0.6^2} \end{bmatrix} = \begin{bmatrix} 0.9398 \\ 0.3417 \end{bmatrix} \tag{7.5}$$

```
radi <- sqrt(sum(c(coef(lmdf11)["x1"], coef(lmdf11)["x2"])^2))
c(coef(lmdf11)["x1"], coef(lmdf11)["x2"])/radi
```

```
       x1        x2
0.9397934 0.3417431
```

7.2.5. 패키지 rsm

위에서 언급한 분석을 포함한 다양한 반응표면 분석은 패키지 rsm 에 있는 여러 가지 함수를 이용하여 쉽게 수행할 수 있다.

7.2.5.1. 일차 선형식의 적합

먼저 일차 선형식 식 7.1 을 적합하는 경우 함수 rsm() 을 다음과 같이 사용할 수 있다. 모형식에서 FO(x1, x2) 는 2개의 변수 x1 과 x2 로 구성된 일차선형식(FO; First Order function)를 사용하는 것을 의미한다.

```
rsmdf1 <- rsm(y ~ FO(x1, x2), data = df11)
summary(rsmdf1)
```

```
Call:
rsm(formula = y ~ FO(x1, x2), data = df11)

            Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 75.08571    0.15102 497.2005 9.818e-11 ***
x1           1.65000    0.19978   8.2592  0.001172 **
x2           0.60000    0.19978   3.0034  0.039810 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9261
F-statistic: 38.62 on 2 and 4 DF,  p-value: 0.002425

Analysis of Variance Table

Response: y
            Df  Sum Sq Mean Sq F value   Pr(>F)
FO(x1, x2)   2 12.3300  6.1650 38.6174 0.002425
Residuals    4  0.6386  0.1596
Lack of fit  2  0.3186  0.1593  0.9955 0.501119
Pure error   2  0.3200  0.1600

Direction of steepest ascent (at radius 1):
       x1        x2
0.9397934 0.3417431
```

```
Corresponding increment in original units:
    time     temp
4.698967 3.417431
```

위의 함수 `rsm()`에서 얻은 결과는 선형모형을 적합하는 함수 `lm()` 의 결과에 추가적은 분석 결과가 포함된다. 추가 분석에 대한 설명은 다음과 같다.

- Analysis of Variance Table 는 일반적인 분산분석표에 2차 반응표면 함수의 각 부분에 대한 유의성 F-검정이 주어진다. 반복이 있는 경우 적합성결여 검정(lack of fit test)이 주어지며 해당하는 p-값이 유의수준 α 보다 작으면 모형의 적합성에 문제가 있다는 의미이다. 위의 결과에서는 분산분석에서 `Lack of fit` 에 대한 p-값이 0.5011 이므로 선형모형의 적합성에는 큰 문제가 없다.
- Direction of steepest ascent (at radius 1) 부분에서는 길이가 1 인 최대경사 벡터 ∇f를 구해준다.
- Corresponding increment in original units 부분에서는 최대경사 벡터를 원래 변환되기 전의 자료의 단위로 구해준다.

7.2.5.2. 등고선 그림

반응 표면의 등고선 그림은 함수 `rsm()` 의 적합 결과를 이용하여 `contour()` 함수를 사용하면 쉽게 그릴 수 있다.

```
contour(rsmdf1, ~ x1 + x2)
contour(rsmdf1, ~ x1 + x2, image = TRUE)
```

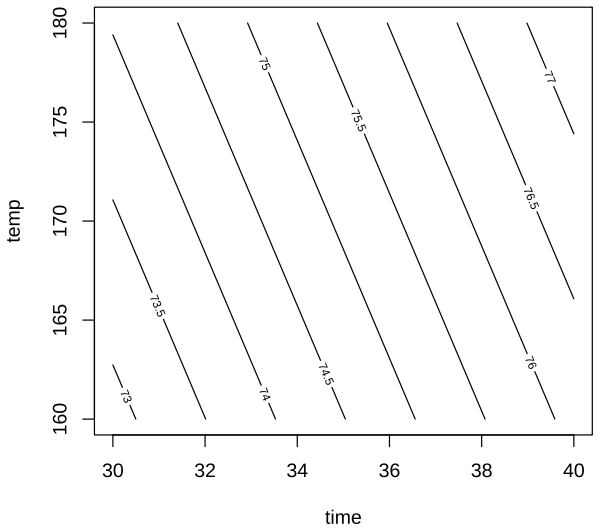


그림 7.9.: rsm 결과를 이용한 등고선 그림

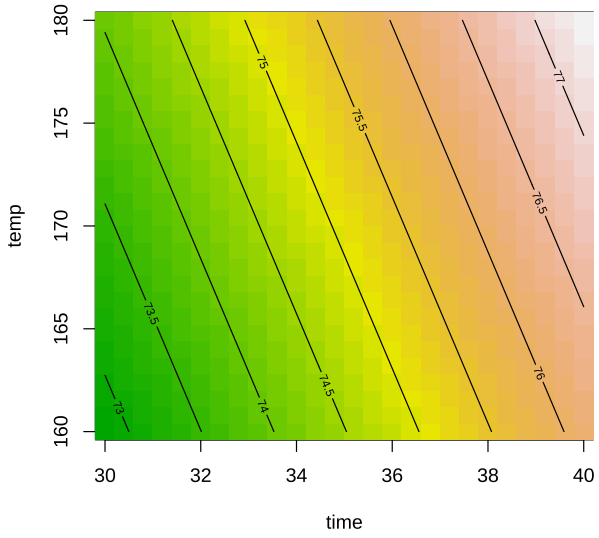


그림 7.10.: rsm 결과를 이용한 등고선 그림

7.2.5.3. 최적 실험점 탐색

이제 실험 원점 $(0, 0)$ 에서 최대경사 벡터 ∇f 방향으로 여러 개의 실험점을 계산하려면 함수 `steepest()`를 사용한다. 아래의 함수에서 `dist=seq(0,5, by=1)` 는 최대경사 벡터 방향으로 길이가 1 만큼 증가하면서 길이가 5인 점까지 실험점을 구하라는 의미이다.

```
deltapoints <- steepest(rsmdf1, dist=seq(0,5, by=1))
```

```
Path of steepest ascent from ridge analysis:
```

```
deltapoints
```

```
  dist    x1    x2 |   time   temp |   yhat
1    0 0.000 0.000 | 35.000 170.00 | 75.086
2    1 0.940 0.342 | 39.700 173.42 | 76.842
3    2 1.880 0.683 | 44.400 176.83 | 78.598
4    3 2.819 1.025 | 49.095 180.25 | 80.352
5    4 3.759 1.367 | 53.795 183.67 | 82.108
6    5 4.699 1.709 | 58.495 187.09 | 83.864
```

위에서 구한 실험점들을 등고선 그림에 표시하면 다음과 같다.

```
contour(rsmdf1, ~ x1 + x2,  bounds = list(x1 = c(-3, 3), x2 = c(-3, 3)))
points(deltapoints$time, deltapoints$temp, cex=1.2, col="red", pch=19)
```

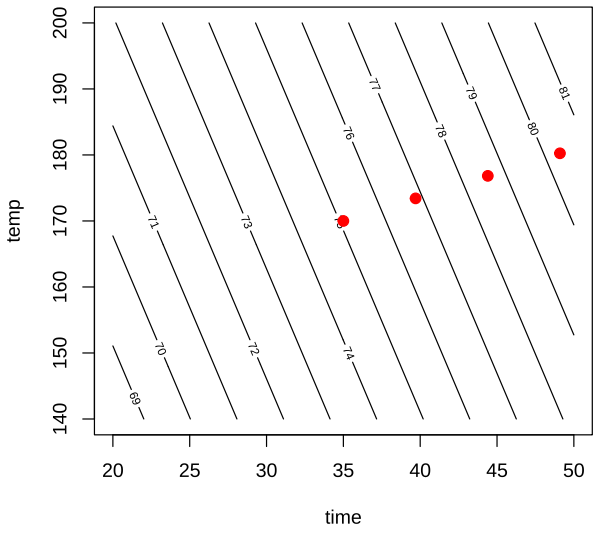


그림 7.11.: 최대경사 벡터 방향의 실험점

만약 아래 그림과 같이 최대경사 방향으로 4번의 실험을 수행하였다면 다음 실험의 중심점은 $(50, 180.9)$를 선택한다. 왜냐하면 더 경사면으로 올라간 실험점 $(55, 184.6)$ 에서는 반응값이 감소하기 때문이다.

< 표 9.3 > 최대경사법 실험 결과

(x_1, x_2)	$(time, Temp)$	y
Δ	(40, 173.6)	77.1
2Δ	(45, 177.3)	77.9
3Δ	(50, 180.9)	**80.9**
4Δ	(55, 184.6)	78.0

그림 7.12.: 최대경사법을 이용한 실험점 찾기

7.2.6. 변량이 3개 이상인 경우

```
heli
```

```
   block    A    R    W   L ave logSD
1      1 11.8 2.26 1.00 1.5 367    72
2      1 13.0 2.26 1.00 1.5 369    72
3      1 11.8 2.78 1.00 1.5 374    74
4      1 13.0 2.78 1.00 1.5 370    79
5      1 11.8 2.26 1.50 1.5 372    72
```

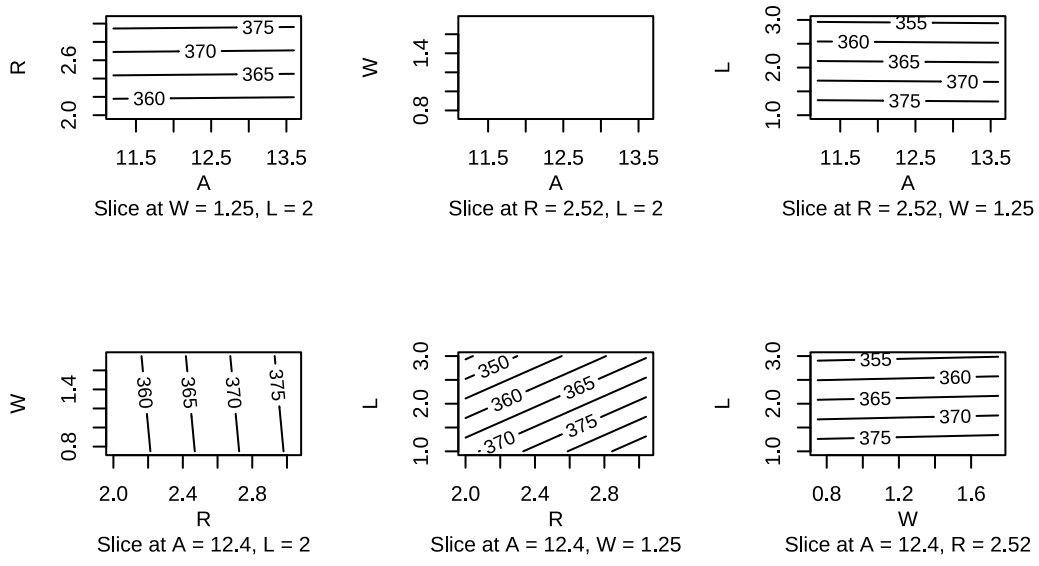
```
6      1 13.0 2.26 1.50 1.5 355    81
7      1 11.8 2.78 1.50 1.5 397    72
8      1 13.0 2.78 1.50 1.5 377    99
9      1 11.8 2.26 1.00 2.5 350    90
10     1 13.0 2.26 1.00 2.5 373    86
11     1 11.8 2.78 1.00 2.5 358    92
12     1 13.0 2.78 1.00 2.5 363   112
13     1 11.8 2.26 1.50 2.5 344    76
14     1 13.0 2.26 1.50 2.5 355    69
15     1 11.8 2.78 1.50 2.5 370    91
16     1 13.0 2.78 1.50 2.5 362    71
17     1 12.4 2.52 1.25 2.0 377    51
18     1 12.4 2.52 1.25 2.0 375    74
19     2 11.2 2.52 1.25 2.0 361   111
20     2 13.6 2.52 1.25 2.0 364    93
21     2 12.4 2.00 1.25 2.0 355   100
22     2 12.4 3.04 1.25 2.0 373    80
23     2 12.4 2.52 0.75 2.0 361    71
24     2 12.4 2.52 1.75 2.0 360    98
25     2 12.4 2.52 1.25 1.0 380    69
26     2 12.4 2.52 1.25 3.0 360    74
27     2 12.4 2.52 1.25 2.0 370    86
28     2 12.4 2.52 1.25 2.0 368    74
29     2 12.4 2.52 1.25 2.0 369    89
30     2 12.4 2.52 1.25 2.0 366    76

Data are stored in coded form using these coding formulas ...
x1 ~ (A - 12.4)/0.6
x2 ~ (R - 2.52)/0.26
x3 ~ (W - 1.25)/0.25
x4 ~ (L - 2)/0.5
```

```
heli.rsm <- rsm(ave ~  FO(x1, x2, x3, x4), data = heli)
summary(heli.rsm)
```

```
Call:
rsm(formula = ave ~ FO(x1, x2, x3, x4), data = heli)

             Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 366.500000   1.437359 254.9816 < 2.2e-16 ***
x1           -0.083333   1.607016  -0.0519 0.9590556
x2            5.083333   1.607016   3.1632 0.0040659 **
x3            0.250000   1.607016   0.1556 0.8776230
x4           -6.083333   1.607016  -3.7855 0.0008577 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Multiple R-squared:  0.4935,    Adjusted R-squared:  0.4125
F-statistic: 6.091 on 4 and 25 DF,  p-value: 0.001453

Analysis of Variance Table

Response: ave
                   Df  Sum Sq Mean Sq F value   Pr(>F)
FO(x1, x2, x3, x4)  4 1510.00  377.50  6.0907 0.001453
Residuals          25 1549.50   61.98
Lack of fit        20 1458.67   72.93  4.0147 0.064646
Pure error          5   90.83   18.17

Direction of steepest ascent (at radius 1):
         x1          x2          x3          x4
-0.01050596  0.64086379  0.03151789 -0.76693536

Corresponding increment in original units:
           A            R            W            L
-0.006303578  0.166624586  0.007879473 -0.383467680
```

```r
 par(mfrow = c(2, 3))
contour(heli.rsm, ~ x1 + x2 + x3 + x4)
```




```r
steepest(heli.rsm)
```

```
Path of steepest ascent from ridge analysis:

   dist      x1     x2     x3      x4 |        A        R        W       L |     yhat
```

```
1   0.0  0.000 0.000 0.000  0.000 | 12.4000 2.52000 1.25000 2.0000 | 366.500
2   0.5 -0.005 0.320 0.016 -0.383 | 12.3970 2.60320 1.25400 1.8085 | 370.461
3   1.0 -0.011 0.641 0.032 -0.767 | 12.3934 2.68666 1.25800 1.6165 | 374.433
4   1.5 -0.016 0.961 0.047 -1.150 | 12.3904 2.76986 1.26175 1.4250 | 378.394
5   2.0 -0.021 1.282 0.063 -1.534 | 12.3874 2.85332 1.26575 1.2330 | 382.366
6   2.5 -0.026 1.602 0.079 -1.917 | 12.3844 2.93652 1.26975 1.0415 | 386.327
7   3.0 -0.032 1.923 0.095 -2.301 | 12.3808 3.01998 1.27375 0.8495 | 390.299
8   3.5 -0.037 2.243 0.110 -2.684 | 12.3778 3.10318 1.27750 0.6580 | 394.260
9   4.0 -0.042 2.563 0.126 -3.068 | 12.3748 3.18638 1.28150 0.4660 | 398.227
10  4.5 -0.047 2.884 0.142 -3.451 | 12.3718 3.26984 1.28550 0.2745 | 402.193
11  5.0 -0.053 3.204 0.158 -3.835 | 12.3682 3.35304 1.28950 0.0825 | 406.160
```

7.3. 2차 반응표면

7.3.1. 개요

- 여러번의 간단한 실험을 순차적으로 수행하면서 1차모형과 최대경사법을 이용하여 최적점 근처로 실험점을 이동한다.
- 최적조건 근처의 영역에서는 반응표면모형의 모형으로 곡선효과가 고려된 2차 다항 모형을 가정하고 최적점을 찾는다.

7.3.2. 2차 다항 모형

이제 반응변수 y의 변화(반응표면; response surface)를 k 개의 독립변수 $x_1, x_2, \cdots, x_k$ 로 이루어진 2차 다항식(second polynomial function)으로 적합하는 2차 다항 모형을 고려하자.

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + e \tag{7.6}$$

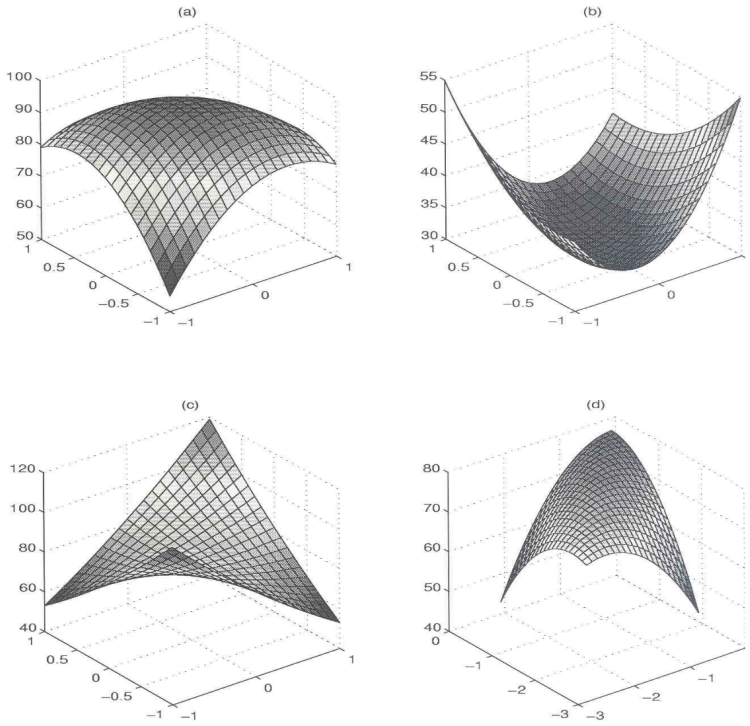
> **노트**
>
> 반응표면분석에서 고려한 2차 다항 모형에서는 독립변수의 값들이 모두 $[-1, 1]$ 사이에 있다고 가정하자.

7.3.3. 이차 반응표면의 모양

예를 들어 독립변수가 두 개인 경우 다음과 같은 2차 다항 모형이 된다.

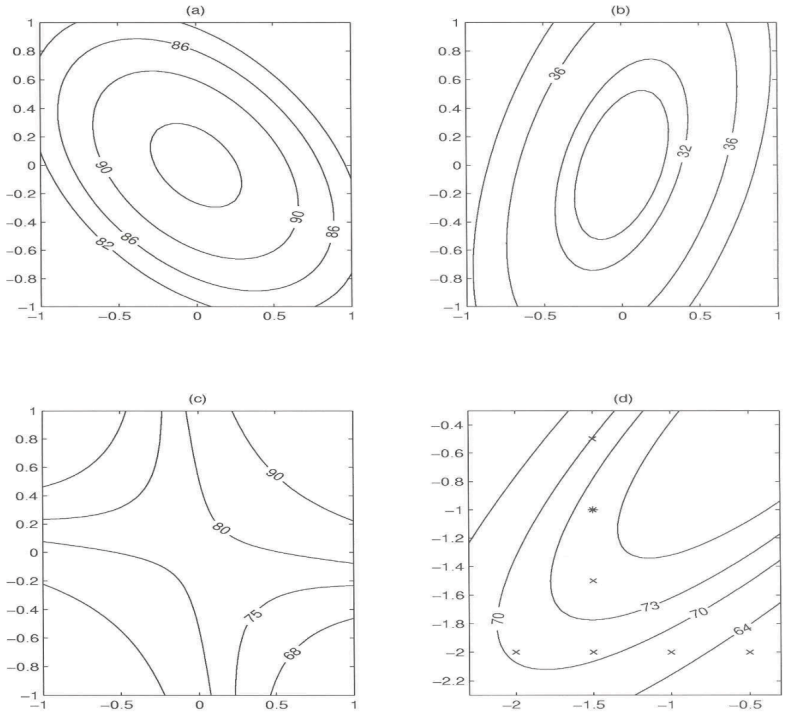
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + e \tag{7.7}$$

독립변수가 두 개인 경우 2차 다항 모형 식 7.7 은 다음 그림과 같이 다양한 반응 표면을 가질 수 있다.



< 그림 9.1 > 반응표면 그림의 4가지 형태
(a) 산 (b) 사발 (c) 말안장 (d) 올라가는 능선

그림 7.13.: 이차 반응표면의 모양- 표면그림과 등고선



< 그림 9.2 > 등고선 그림의 4가지 형태
(a) 산 (b) 사발 (c) 말안장 (d) 올라가는 능선

그림 7.14.: 이차 반응표면의 모양- 표면그림과 등고선

7.3.4. 최적점과 정상점

2차 다항모형 식 7.6 을 반응표면 분석에 사용하는 이유는 반응변수의 값이 최대가 되는 독립변수들의 값을 구하기 위함이다. 이렇게 반응변수의 값이 최대가 되는 독립변수들의 값을 **최적점(optimum point)** 라고 부른다.

최적점을 찾는 방법은 2차 다항모형 식 7.6 을 각 독립변수로 미분한 식을 0으로 놓고 방정식을 푸는 것이다. 최적점을 찾는 방법을 좀 더 체계적으로 구성하기 위하여 2차 다항모형 식 7.6 을 벡터와 행렬로 나타내어 보자.

먼저 자료를 이용하여 2차 다항모형 식 7.6 에 나타난 계수들을 추정한 후 반응변수의 예측식을 다음과 같이 나타낼 수 있다. 아래의 식에서 b_0, b_i, b_{ii}, b_{ij} 는 각각 회귀계수 $\beta_0, \beta_i, \beta_{ii}, \beta_{ij}$ 의 추정값이라고 하자.

$$\hat{y} = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i=1}^{k} b_{ii} x_i^2 + \sum\sum_{i<j} b_{ij} x_i x_j \tag{7.8}$$

이제 추정식 식 7.8 를 다음과 같이 벡터와 행렬로 나타내자.

$$\hat{y} = b_0 + \boldsymbol{x}^t \boldsymbol{b} + \boldsymbol{x}^t \boldsymbol{B} \boldsymbol{x} \tag{7.9}$$

여기서 $k \times 1$ 차원의 독립변수 벡터 $\boldsymbol{x}$, $k \times 1$ 차원의 1차 계수 벡터 $\boldsymbol{b}$, $k \times K$ 차원의 2차 계수 행렬 $\boldsymbol{B}$ 는 다음과 같이 주어진다. 여기서 2차 계수 행렬 $\boldsymbol{B}$ 는 대칭행렬이며 비대각 원소는 계수 b_{ij} 의 반 값임을 유의하자.

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} b_{11} & \frac{b_{12}}{2} & \frac{b_{13}}{2} & \cdots & \frac{b_{1k}}{2} \\ \frac{b_{12}}{2} & b_{22} & \frac{b_{23}}{2} & \cdots & \frac{b_{2k}}{2} \\ \frac{b_{13}}{2} & \frac{b_{23}}{2} & b_{33} & \cdots & \frac{b_{3k}}{2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{b_{1k}}{2} & \frac{b_{2k}}{2} & \frac{b_{3k}}{2} & \cdots & b_{kk} \end{bmatrix} \tag{7.10}$$

이제 정삼점을 찾는 방법은 2차 다항 예측식 식 7.8 를 각 독립변수 x_i 로 미분한 식을 0으로 놓은 방정식을 풀면 된다.

$$\frac{\partial \hat{y}}{\partial x_i} = b_i + 2b_{ii}x_i + \sum_{j \neq i} b_{ij}x_j = 0, \quad i = 1, 2, \cdots, k \tag{7.11}$$

방정식 식 7.11 을 벡터식으로 표시하면 다음과 같은 벡터 방정식을 얻는다.

$$\frac{\partial \hat{y}}{\partial \boldsymbol{x}} = \boldsymbol{b} + 2\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0} \tag{7.12}$$

벡터 방정식 식 7.12 을 만족하는 벡터 $\boldsymbol{x}^*$ 를 **정상점(stationary point)** 라고 부르며 정상점 $\boldsymbol{x}^* = (x_1^*, x_2^*, \cdots, x_k^*)^t$ 는 다음과 같이 얻어진다.

$$\boldsymbol{x}^* = -\frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b} \tag{7.13}$$

식 식 7.13 에 주어진 정상점 $\boldsymbol{x}^*$ 은 행렬 $\boldsymbol{B}$ 의 성질에 따라서 반응값을 최대로 하는 최적점일 수도 있고 아닐 수도 있다. 다음 절에서는 정상점이 최적점일 조건을 살펴보기로 하자.

7.3.5. 2차 다항식의 표현

앞 절에서 보았듯이 정상점 $\boldsymbol{x}^*$ 은 추정된 2차 다항식에 대하여 미분방정식 식 7.13 을 만족하는 점이다.

2차 다항식은 그림 @ref(fig:plot3) 처럼 다양한 모양을 가진다. 실험의 목적이 반응변수를 최대로 하는 최적점을 찾는 것이기 때문에 2차 다항식의 모양이 산 모양이면 정상점이 최적점이 되지만 다른 형태이면 최적점이 아니다.

정상점 $\boldsymbol{x}^*$ 이 최적점이 될 조건은 행렬 $\boldsymbol{B}$에 대한 정준분석(canonical analysis)를 통하여 파악할 수 있다. 정준분석은 행령의 고유값(eigen value)과 고유벡터(eigen vector) 를 통하여 이루어진다. 이 절에서는 정분분석을 하기 위하여 2차 다항식을 다루기 쉬운 형식으로 표현하고자 한다.

식 식 7.13 의 정상점 $\boldsymbol{x}^*$ 을 중심으로 하는 축 $\boldsymbol{z}$ 를 다음과 같이 고려하고

$$\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}^* \quad \text{equivalently} \quad \boldsymbol{x} = \boldsymbol{x}^* + \boldsymbol{z} \tag{7.14}$$

2차 다항식 식 7.9 를 다음과 같이 $\boldsymbol{z}$ 의 함수로 변환해 보자. 아래 식에서 정상점 $\boldsymbol{x}^* = -\boldsymbol{B}^{-1}\boldsymbol{b}/2$ 이다.

$$
\begin{aligned}
\hat{y} &= b_0 + \boldsymbol{x}^t\boldsymbol{b} + \boldsymbol{x}^t\boldsymbol{B}\boldsymbol{x} \\
&= b_0 + \boldsymbol{x}^t\boldsymbol{b} + (\boldsymbol{x}^* + \boldsymbol{z})^t\boldsymbol{B}(\boldsymbol{x}^* + \boldsymbol{z}) \\
&= b_0 + \boldsymbol{x}^t\boldsymbol{b} + 2\boldsymbol{x}^{*t}\boldsymbol{B}\boldsymbol{z} + \boldsymbol{x}^{*t}\boldsymbol{B}\boldsymbol{x}^* + \boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z} \\
&= b_0 + \boldsymbol{x}^t\boldsymbol{b} - \boldsymbol{b}^t\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{z} + \frac{1}{4}\boldsymbol{b}^t\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z} \\
&= b_0 + \boldsymbol{x}^t\boldsymbol{b} - \boldsymbol{b}^t\boldsymbol{z} + \frac{1}{4}\boldsymbol{b}^t\boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z} \\
&= b_0 + (\boldsymbol{x}^* + \boldsymbol{z})^t\boldsymbol{b} - \boldsymbol{b}^t\boldsymbol{z} + \frac{1}{4}\boldsymbol{b}^t\boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z} \\
&= b_0 + (\boldsymbol{z}^t\boldsymbol{b} - \boldsymbol{b}^t\boldsymbol{z}) + \boldsymbol{x}^{*t}\boldsymbol{b} + \frac{1}{4}\boldsymbol{b}^t\boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z} \\
&= b_0 + 0 + \left[-\frac{1}{2}\boldsymbol{b}^t\boldsymbol{B}^{-1}\right]\boldsymbol{b} + \frac{1}{4}\boldsymbol{b}^t\boldsymbol{B}^{-1}\boldsymbol{b} + \boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z} \\
&= \left[b_0 - \frac{1}{4}\boldsymbol{b}^t\boldsymbol{B}^{-1}\boldsymbol{b}\right] + \boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z}
\end{aligned}
\tag{7.15}
$$

위의 식에서 $\boldsymbol{z} = 0$ 인 점은 원래의 측으로 보면 정상점 $\boldsymbol{x} = \boldsymbol{x}^*$ 이다. 정상점에서의 반응변수의 예측값을 $\hat{y}_s$ 라고 하면

$$
\hat{y}_s = b_0 + \boldsymbol{x}^{*t}\boldsymbol{b} + \boldsymbol{x}^{*t}\boldsymbol{B}\boldsymbol{x}^* = b_0 - \frac{1}{4}\boldsymbol{b}^t\boldsymbol{B}^{-1}\boldsymbol{b}
$$

따라서 2차 다항식의 예측식 식 7.9 은 변수 $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}^*$ 으로 다음과 같이 나타낼 수 있다.

$$
\hat{y} = \hat{y}_s + \boldsymbol{z}^t\boldsymbol{B}\boldsymbol{z} \tag{7.16}
$$

7.4. 2차모형의 정준분석

7.4.1. 개요

- 여러번의 간단한 실험을 순차적으로 수행하면서 1차모형과 최대경사법을 이용하여 최적점 근처로 실험점을 이동한다.
- 최적조건 근처의 영역에서는 반응표면모형의 모형으로 곡선효과가 고려된 2차 다항 모형을 가정하고 최적점을 찾는다.
- 2차 다항 모형은 계수의 추정치에 따라서 반응표면의 모양을 4가지 모양으로 크게 분류하할 수 있다.
- 추정된 2차 모형이 어떤 모양에 해당하는지 판단할 수 있는 기법인 정준분석(canonocal analysis)을 적용한다.

7.4.2. 이차형식

임의의 벡터 $\boldsymbol{x}$에 대하여 차원이 $k \times k$ 인 대칭 행렬 $\boldsymbol{B}$ 이 주어진 경우 이차형식(quadratic form) $s(\boldsymbol{x}, \boldsymbol{B})$는 다음과 같이 정의된다.

$$
s(\boldsymbol{x}, \boldsymbol{B}) = \boldsymbol{x}^t\boldsymbol{B}\boldsymbol{x} \tag{7.17}
$$

만약 행렬 $\boldsymbol{B}$의 고유값이 다음과 같고

$$
\lambda_1, \lambda_2, \ldots, \lambda_k
$$

이에 대응하는 고유벡터가 다음과 같다고 하자.

$$\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_k$$

자세한 고유값과 고유젝터에 대한 이론은 부록을 참조하자.

대칭 행렬 $\boldsymbol{B}$의 고유값과 고유벡터를 이용하면 다음과 같은 스펙트럴 분해(spectral decomposition)가 가능하다 (부록 참조).

$$\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^t = \lambda_1 \boldsymbol{q}_1 \boldsymbol{q}_1^t + \lambda_2 \boldsymbol{q}_2 \boldsymbol{q}_2^t + \cdots + \lambda_k \boldsymbol{q}_k \boldsymbol{q}_k^t \tag{7.18}$$

이제 식 식 7.17 에서 정의된 이차형식은 스펙트럴 분해를 이용하여 다음과 같은 분해가 가능하다.

$$\begin{aligned} s(\boldsymbol{x}, \boldsymbol{B}) &= \boldsymbol{x}^t \boldsymbol{B} \boldsymbol{x} \\ &= \boldsymbol{x}^t \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^t \boldsymbol{x} \\ &= \boldsymbol{x}^t \left[\lambda_1 \boldsymbol{q}_1 \boldsymbol{q}_1^t + \lambda_2 \boldsymbol{q}_2 \boldsymbol{q}_2^t + \cdots + \lambda_k \boldsymbol{q}_k \boldsymbol{q}_k^t\right] \boldsymbol{x} \\ &= \sum_{i=1}^{k} \lambda_i (\boldsymbol{x}^t \boldsymbol{q}_i)(\boldsymbol{q}_i^t \boldsymbol{x}) \\ &= \sum_{i=1}^{k} \lambda_i w_i^2 \end{aligned}$$

여기서

$$w_i = \boldsymbol{x}^t \boldsymbol{q}_i = \boldsymbol{q}_i^t \boldsymbol{x}, \quad i = 1, 2, \ldots, k$$

7.5. 2차 다항식의 정준형식

앞 절에서 논의한 것을 정리하면 2차 다항식의 예측식은 식 식 7.14 에서 정의된 변수 $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}^*$ 를 이용하여 식 식 7.16 과 같이 표현할 수 있으므로 다음과 같이 정준형식으로 나타낼 수 있다.

$$\begin{aligned} \hat{y} &= \hat{y}_s + \boldsymbol{z}^t \boldsymbol{B} \boldsymbol{z} \\ &= \hat{y}_s + \lambda_1 w_1^2 + \lambda_2 w_2^2 + \cdots + \lambda_k w_k^2 \end{aligned}$$

여기서

$$w_i = \boldsymbol{z}^t \boldsymbol{q}_i = \boldsymbol{q}_i^t \boldsymbol{z}, \quad i = 1, 2, \ldots, k$$

$$\hat{y}_s = b_0 - \frac{1}{4} \boldsymbol{b}^t \boldsymbol{B}^{-1} \boldsymbol{b}$$

위의 식 식 7.19 을 2차 다항식의 **정준형식(canonical form)** 이라고 부른다.

이제 행렬 $\boldsymbol{B}$ 의 고유값을 이용한 정준분석을 이용하여 다음과 같이 정상점의 형태와 최적점의 유무를 알아낼 수 있다.

표 7.1.: 정준분석을 이용한 최적점의 판단

행렬 $\boldsymbol{B}$ 의 고유값	정상점은
모두 음수이면	최대점(최적점)
모두 양수 이면	최저점
양수와 음수가 섞여있으면	안장점

7.5.1. 변환된 변수

2차 다항식의 정준형식 식 7.19 에 나타난 변환된 변수 w_i 와 원래 사용된 변수 x_i 의 관계를 알아보자.

변수 z_i 는 원래 변수 x_i 에서 정상점 x_i^* 를 빼서 만든 변수이다 식 7.14. 또한 $w_i = \boldsymbol{q}_i^t \boldsymbol{z}$ 이므로 다음과 같은 변환식이 얻어진다.

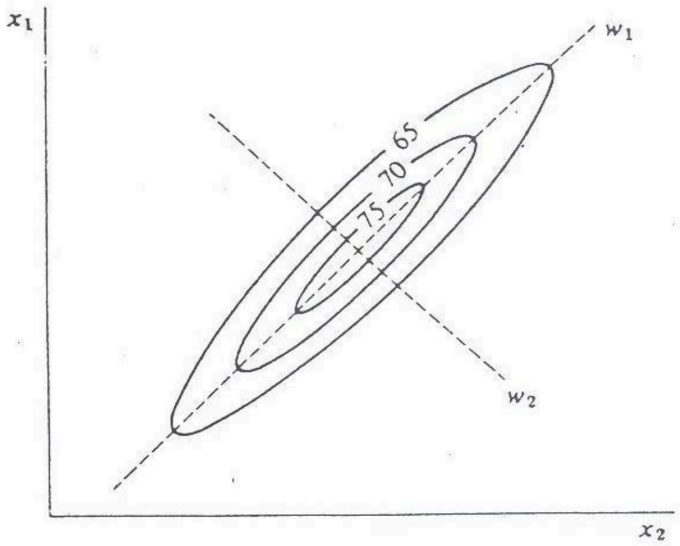
$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} = \boldsymbol{Q}^t \boldsymbol{z} = \boldsymbol{Q}^t (\boldsymbol{x} - \boldsymbol{x}^*) \tag{7.19}$$

여기서 행렬 $\boldsymbol{Q}$ 는 행렬 $\boldsymbol{B}$ 의 고유벡터들로 이루진다. 따라서 행렬 $\boldsymbol{Q}$ 는 직교행렬이다.

$$\boldsymbol{Q} = [\boldsymbol{q}_1 \ \boldsymbol{q}_2 \ \cdots \boldsymbol{q}_k], \quad \boldsymbol{Q}^t \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^t = \boldsymbol{I}$$

따라서 식 식 7.19 에서 정의된 변환은 벡터 $\boldsymbol{z}$를 행렬 $\boldsymbol{Q}$ 를 이용하여 회전하는 변환이다.

$$\boldsymbol{w}^t \boldsymbol{w} = \boldsymbol{z}^t \boldsymbol{Q}\boldsymbol{Q}^t \boldsymbol{z} = \boldsymbol{z}^t \boldsymbol{z}$$



< 그림 9.7 > 이차 모형의 정준 형식

7.5.2. 예제: 2개의 독립 변수

만약 독립변수가 2개 있는 2차 다항 모형에서는 행렬 $\boldsymbol{B}$의 차원은 2×2이다. 따라서 행렬 $\boldsymbol{B}$의 고유치는 λ_1 과 λ_2 라고 하자.

이 경우는 2차 형식이 다음과 같이 분해될 수 있다.

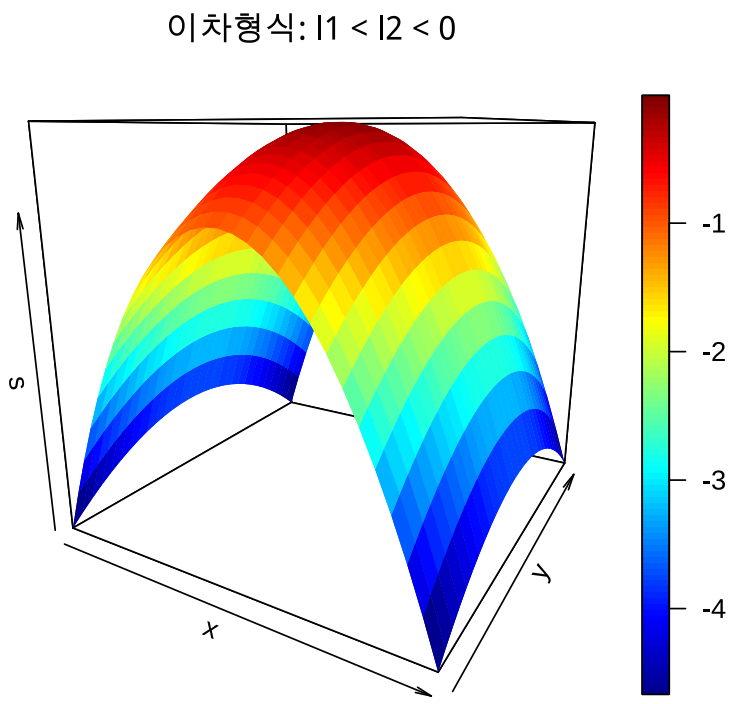
$$s(\boldsymbol{x}, \boldsymbol{B}) = \boldsymbol{x}^t \boldsymbol{B} \boldsymbol{x} = \lambda_1 w_1^2 + \lambda_2 w_2^2$$

위의 2차 형식을 계산하는 함수를 R 로 만들어 보자.

```
quad <- function(w1, w2, l1, l2){
  l1*w1^2 + l2*w2^2
}
```

1. $\lambda_1 < \lambda_2 < 0$ 의 경우
    - 이차형식 $s(\boldsymbol{x}, \boldsymbol{B})$ 는 정상점이 **최대 반응점**이 되며 w_2 에 비하여 w_1 축으로 반응의 변화가 급하다.
    - $\lambda_1 = -4, \lambda_2 = -1$ 인 경우 2차 형식의 형태는 다음과 같다.

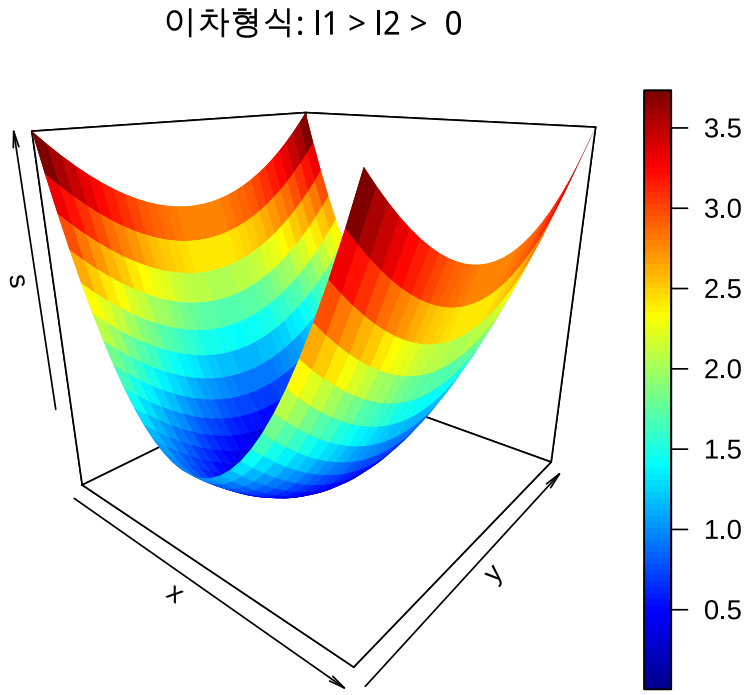
```
w1 <- w2 <- seq(-1, 1, length= 30)
s <- outer(w1, w2 ,quad, -4, -1)
persp3D(w1, w2, s,
      main="이차형식: l1 < l2 < 0",
      zlab = "s",
      theta = 30, phi = 15)
```



2. $\lambda_1 > \lambda_2 > 0$ 의 경우

- 이차형식 $s(\boldsymbol{x}, \boldsymbol{B})$ 는 정상점이 최소 반응점이 되며 w_2 에 비하여 w_1 축으로 반응의 변화가 급하다.
- $\lambda_1 = 2, \lambda_2 = 1$ 인 경우 2차 형식의 형태는 다음과 같다.

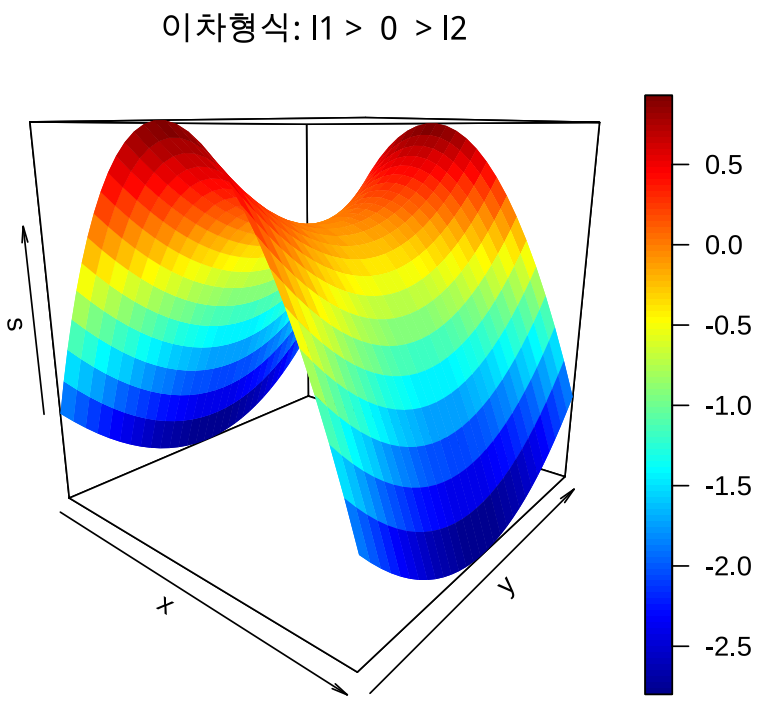
```
w1 <- w2 <- seq(-1, 1, length= 30)
s <- outer(w1, w2 ,quad, 3, 1)
persp3D(w1, w2, s,
      main="이차형식: l1 > l2 >  0",
      zlab = "s",
      theta = 40, phi = 20)
```



3. $\lambda_1 > 0, \lambda_2 < 0, |\lambda_1| > |\lambda_2$ 의 경우

   - 이차형식 $s(\boldsymbol{x}, \boldsymbol{B})$ 는 정상점이 안장점(saddle point)이 되며 w_2 축으로는 반응값이 감소하게 되고 w_1 축으로는 반응값이 증가하게 된다. 또한 w_2 에 비하여 w_1 축으로 반응의 변화가 급하다.
   - $\lambda_1 = -3, \lambda_2 = 1$ 인 경우 2차 형식의 형태는 다음과 같다.

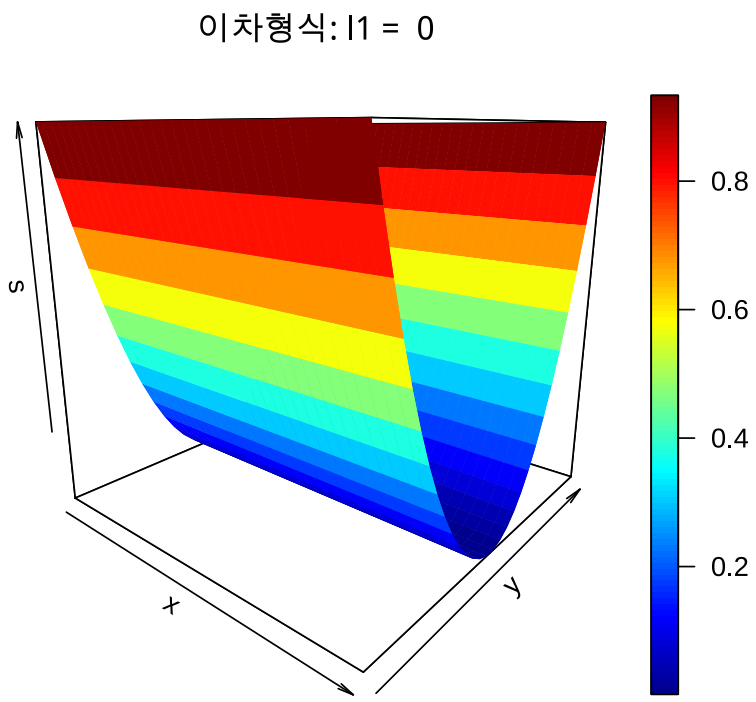
```
w1 <- w2 <- seq(-1, 1, length= 30)
s <- outer(w1, w2 ,quad, -3, 1)
persp3D(w1, w2, s,
      main="이차형식: l1 >  0  > l2",
      zlab = "s",
      theta = 40, phi = 15)
```



4. $\lambda_i \approx 0$ 인 경우

- 반응표면 체계가 능선 체계(ridge system)이라고 한다. 능선 체계에서는 최적점이 무수히 많은 정상 능선 체계(stationary ridge system)와 능선이 한 방향으로 계속 증가하는 형태(rising ridge system)으로 나뉜다. 올라가는 능선 체계에서는 현제의 실험 영역에서 최적점이 없기 때문에 실헌 영역을 변경하여 다시 실험을 진행해야 한다.
- $\lambda_1 = 0, \lambda_2 = 1$ 인 경우 2차 형식의 형태는 다음과 같다.

```
w1 <- w2 <- seq(-1, 1, length= 30)
s <- outer(w1, w2 ,quad, 0, 1)
persp3D(w1, w2, s,
      main="이차형식: l1 =  0 ",
      zlab = "s",
      theta = 40, phi = 15)
```



7.6. 최적점 탐색을 위한 실험계획

7.6.1. 개요

- 최적조건 근처의 영역에서는 반응표면모형의 모형으로 곡선효과가 고려된 2차 다항 모형을 가정하고 최적점을 찾는다.
- 최적 조건을 찾는 실험에서는 실험점을 효율적으로 배치해야 한다.
- 2차다항식을 적합해야 하기 때문에 각 요인에 대하여 최소한 3개의 수준이 필요하다.

7.6.2. 중심합성설계

반응표면 분석에서 고려하는 변량(독립변수)의 개수가 k 개 이면 한 요인에 대해서 최소한 3개의 수준이 필요하다. 3개의 수준이 필요한 이유는 2차 다항식을 고려해야 하기 때문이다.

따라서 요인의 개수가 k 이면 실험점의 개수가 3^k 가 필요하며 실험점의 개수는 요인의 수에 따라서 기하급수적으로 늘어나게 된다. 예를 들어 $k = 3$ 인 경우에는 실험점이 $3^3 = 27$ 개, 경우에는 실험점이 $3^4 = 81$ 개로 늘어나는데 이는 현실적으로 감당하기 어려운 경우일 수 있다.

따라서 반응표면 방법에서는 실험을 축차적으로(sequentially) 실행하면서 최적점 근처에서 더 많은 실험을 수행할 수 있는 효율적인 실험 계획을 고려해야 한다. 아래는 반응표면 실험의 축차적인 절차를 설명한다.

(1) 1차 모형 적합을 위한 1단계 실험

- 먼저 **중심점(center points)에서** n_0 **개**의 실험을 수행하고 각 요인의 **상자점들(factorial points)에서** 2^k 실험을 실시한다.
- 이러한 기초 실험은 최대경사법을 이용하는 중간 과정에서 언제나 수행해야 하는 실험이다.
- 기초 실험은 1차 모형을 적합하기 위한 실험계획이다.
- 기초 실험에서 실험 구간에 최적점이 포함되어 있다고 판단되면 2차 모형 적합을 위한 2단계실험을 실시한다.
- 1단계 실험의 개수는 상자점 2^k 개 + 중심점 n_0 개

(2) 2차 모형 적합을 위한 2단계 실험

- 최적점이 가까워진 경우 2차 모형 적합을 위해서 1단계 실험의 실험점들에 추가적인 실험점을 더해서 2단계 실험을 진행한다.
- 2차 모형을 적합하기 위한 실험으로 **각 요인에 대하여 2개의 축점(axial points, star points)을 더 추가한다.** 기초실험에서 수행한 실험점들과 합쳐서 각 요인마다 3개의 수준을 가지는 실험점을 만든다.
- 2단계 실험에서 추가한 축점은 2차 모형의 효율적인 추정을 고려하여 선택한다.
- 2단계 실험의 개수는 축점 $2k$ 개

반응표면방법에서 위와 같이 최적점 탐색을 하는 경우 사용되는 대표적인 실험설계는 **중심합성설계(CCD: central composite design)** 이다.

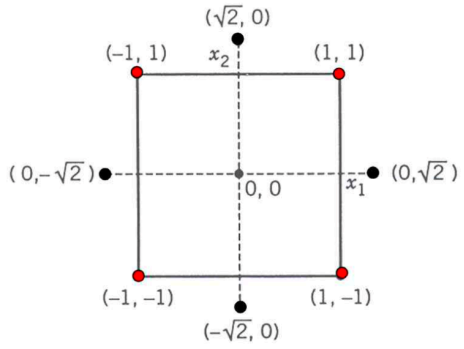


Figure 7.4 Central composite design for $k = 2$ and $\alpha = \sqrt{2}$.

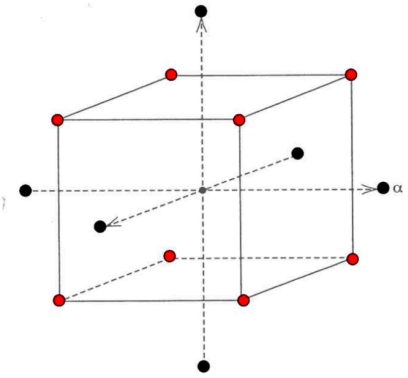


Figure 7.5 Central composite design for $k = 3$ and $\alpha = \sqrt{3}$.

그림 7.15.: 구형 계획법을 이용한 중심합성설계

중심합성설계에서 축차적으로 실험을 실시한 경우 k개의 요인을 고려하면 실험점들은 그 특성에 따라서 다음과 같은 나타난다.

실험점	목적	개수
상자점(factorial points)	1차 모형 적합을 위한 2수준 배치법에서의 실험점	$F = 2^k$
중심점(center points)	관심 구간의 중심에 위치한 실험점	n_0
축점(axial points)	2차 모형 적합을 위해 추가된 실험점	$n_a = 2k$

7.6.3. 회전가능 중심합성설계

중심합성설계에서 2차 다항식을 추정하기 위한 2단계 실험에서 추가하는 축점(axial points)을 어떻게 선택하는 것이 좋을까?

축점을 배치할 때 중요한 고려사항은 2차 다항식에서 얻은 추정치들의 분산이 각 실험점들에서 동일하게 나타나게 하는 것이다. 이러한 성질을 **회전 가능성(Rotatibility)** 라고 부른다. 이러한 회전 가능성이 만족하면 고려한 모든 실험점들에서 구한 예측값들의 정도(precision)가 같다는 의미이다.

회전가능성이 중요한 이유는 반응표면분석이 반응의 최적점을 찾는 실험이고 최적점의 위치는 알 수 없으므로 모든 방향에 대한 예측값의 정도를 동일하게 설정하는 것이 합리적이기 때문이다.

2차 다항식도 선형모형에 속하므로 다음과 같은 선형모형을 고려할 때

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

최소제곱법으로 얻은 회귀계수의 추정량을 $\hat{\boldsymbol{\beta}}$ 이라고 하자. 만약 관심이 있는 실험점 $\boldsymbol{x}$ 에서 반응값의 예측치는 $\hat{y} = \boldsymbol{x}^t\hat{\boldsymbol{\beta}}$ 이다. 또한 예측치의 분산은 다음과 같이 주어진다.

$$Var(\hat{y}) = Var(\hat{y}|\boldsymbol{x}) = \sigma^2\boldsymbol{x}^t(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{x}$$

"중심합성설계가 회전 가능하다"는 의미는 실험에서 사용한 모든 실험점들에서 반응변수 예측치의 분산이 동일하다는 것을 의미한다.

$$Var(\hat{y}|\boldsymbol{x}_i) = Var(\hat{y}|\boldsymbol{x}_j) \quad \text{for all } i, j \tag{7.20}$$

이렇게 실험에서 고려한 모든 실험점에서 예측치의 분산이 같은 실험계획을 '회전가능하다'(rotatable) 라고 말하며 일반적으로 **균등 정밀 중심합성설계(uniform precision CCD)**라고 부른다.

중심합성설계가 회전가능하게 되는 조건은 무었일까? 일반적으로 2차 실험에서 추가되는 축점의 길이 α에 따라서 회전가능성이 결정된다. 상자점의 수가 F개 인 경우 회전 가능한 계획을 얻기 위해서는 **축점의 길이를 $\alpha = F^{1/4}$ 로 설정**하면 된다.

다음은 요인의 개수 k 에 대한 균등 정밀 중심합성설계의 각 실험점의 개수와 축점의 길이 α를 나타낸 표이다.

표 7.3.: 표 9.4. 균등 정밀 중심합성설계의 각 실험점의 개수와 축점의 거리

k	2	3	4	5	6	7
$F = 2^k$	4	8	16	32	64	128
$n_a = 2k$	4	6	8	10	12	14
n_0	n_0	n_0	n_0	n_0	n_0	n_0
$\alpha = F^{1/4}$	1.414	1.682	2.000	2.378	2.828	3.363

교과서 표 9.5 에 오타가 있습니다. 위의 표가 수정된 실험점의 개수와 축점의 거리입니다.

참고로 중심점 $(0, 0)$ 에서의 실험의 수 n_0 는 일반적으로 3개에서 5개이다.

> **노트**
>
> 구형 계획법
> 반응표면분석에서 회전 가능한 계획법은 실험에서 고려한 모든 실험점들의 분산을 같게 하는 실험계획이다. 하지만 실제 현장에서는 분산이 모두 동일한 경우 보다는 **모든 실험점이 원점에서 같은 거리에 있는 경우**를 선호하는 경우도 있다. 고려한 모든 실험점이 중심점으로부터 거리가 동일한 계획을 **구형 계획법(spherical design)**이라고 한다. 요인의 개수가 k 개인 경우, 구형 계획법에서 축점의 거리는 $\alpha = \sqrt{k}$ 이다.

7.6.4. Box-Benken 설계

Box-Benken 설계는 2차 모형 적합을 위한 효율적인 3수준 실험설계로 3수준 요인배치법의 일부 실험조건에서만 실험을 실시하는 것이다.

예를 들어 3개의 요인가 있다면 관심영역 상자점에서 2개의 요인를 먼저 선택하고 2^2 개의 $(\pm 1, \pm 1)$ 수준에서 실험을 하고 나머지 인자는 수준의 중앙값인 0으로 고정한다.

이렇게 고려하는 인자의 개수가 k 개인 경우, 2개의 인자를 선택하는 조합의 수는 $k(k-1)/2$ 개이다. 또한 각 조합마다 4개의 실험점$(\pm 1, \pm 1)$ 이 추가되므로 총 실험점의 개수는 다음과 같다.

$$4k(k-1)/2 + n_0$$

Box-Benken 설계는 선택된 2개의 조합이 블럭으로 나타나는 블럭 일부실시법이다.

< 표 9.6 > k=3인 경우의 Box-Benken 설계 실험점

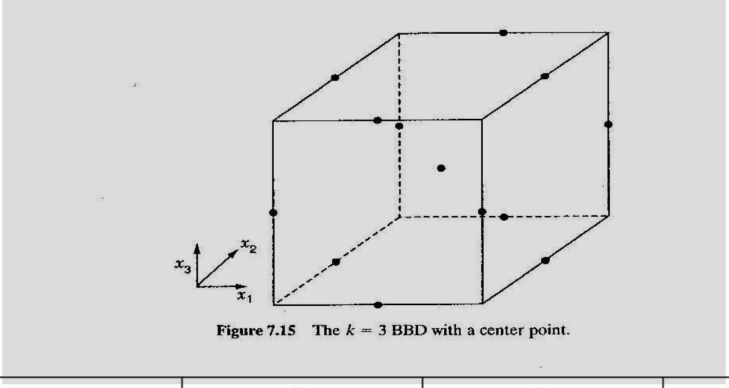


Figure 7.15 The $k = 3$ BBD with a center point.

	x_1	x_2	x_3
블록 1	±1	±1	0
블록 2	±1	0	±1
블록 3	0	±1	±1
중앙점	0	0	0

그림 7.16.: 요인의 개수가 3개 인 경우 Box-Benken 설계

7.6.5. 계획법에 따른 실험점의 개수

요인의 개수가 k 개인 경우 각 실험계획법에 대한 실험점의 개수는 다음의 표에 있는 공식으로 계산할 수 있다. n_c 는 중심점 $(0, 0)$ 에서 실험의 개수이다(앞 절에서는 중심점을 n_0 로 표시).

요인의 수	Box-Benken 설계	중심합성설계(CCD)	3수준 요인 배치법
k	$4k(k-1)/2 + n_c$	$2^k + 2k + n_c$	3^k

< 표 9.7 > 2차 반응표면모형 적합을 위한 실험 설계의 실험 크기 비교

k	Box-Benken 설계	중심합성설계(C.C.D.)	3수준 요인배치법
3	$12+n_c$	$14+n_c$	27
4	$24+n_c$	$24+n_c$	81
5	$40+n_c$	$26(\text{또는 } 42)+n_c$	243

그림 7.17.: 계획법에 따른 실험점의 개수 예제

7.6.6. R 을 이용한 실험계획

7.6.6.1. 회전가능 중심합성설계

요인의 개수가 $k = 3$ 이고 중심점에서 실험의 개수가 $n_0 = 2$ 인 회전가능 중심합성설계의 실험점을 패키지 `rsm` 의 함수 `ccd()` 를 통하여 구해보자.

함수 `ccd()`를 사용하는 경우 인자는 다음과 같다.

- `basis = 3` : 요인의 개수(k)를 3개로 지정
- `n0 = 2` : 중심점의 수 (n_0) 를 2개로 지정
- `alpha = "rotatable"` : 회전 가능한 계획의 축점을 지정
- `randomize = F` : 실험의 순서를 임의화하지 않는다.

함수 `ccd()`의 결과를 볼 때 주의할 점은 다음과 같다.

- 실험점들은 2개의 블럭(`Block`)으로 나누어 표시된다.
- 첫 번째 블럭(`Block=1`)은 1차 모형을 적합하기 위한 상자점과 중심점으로 구성된다.
- 두 번째 블럭(`Block=2`)은 2차 모형을 적합하기 위한 축점과 중심점으로 구성된다.
- 따라서 중심점의 총 개수는 지정한 개수 `n0 = 2` 의 2배가 된다.

```
designR <- ccd (basis = 3, n0 = 2, alpha = "rotatable", randomize = F)
designR
```

```
   run.order std.order  x1.as.is  x2.as.is  x3.as.is Block
1          1         1 -1.000000 -1.000000 -1.000000     1
2          2         2  1.000000 -1.000000 -1.000000     1
3          3         3 -1.000000  1.000000 -1.000000     1
4          4         4  1.000000  1.000000 -1.000000     1
5          5         5 -1.000000 -1.000000  1.000000     1
6          6         6  1.000000 -1.000000  1.000000     1
7          7         7 -1.000000  1.000000  1.000000     1
8          8         8  1.000000  1.000000  1.000000     1
9          9         9  0.000000  0.000000  0.000000     1
10        10        10  0.000000  0.000000  0.000000     1
11         1         1 -1.681793  0.000000  0.000000     2
12         2         2  1.681793  0.000000  0.000000     2
13         3         3  0.000000 -1.681793  0.000000     2
```

```
14         4         4  0.000000  1.681793  0.000000     2
15         5         5  0.000000  0.000000 -1.681793     2
16         6         6  0.000000  0.000000  1.681793     2
17         7         7  0.000000  0.000000  0.000000     2
18         8         8  0.000000  0.000000  0.000000     2

Data are stored in coded form using these coding formulas ...
x1 ~ x1.as.is
x2 ~ x2.as.is
x3 ~ x3.as.is
```

교과서 표 9.4에 나타난 실험계획에 대한 데이터프레임을 만드는 방법은 아래와 같다. 참고할 점은 중심점의 수를 6개로 지정하면(n0=6) 모두 12개의 중심점이 생성되므로 마지막 6개의 중심점은 자료에 포함시키지 않는다.

```
design94 <- ccd (basis = 3, n0 = 6, alpha = "rotatable", randomize = F)
design94
```

```
   run.order std.order  x1.as.is  x2.as.is  x3.as.is Block
1          1         1 -1.000000 -1.000000 -1.000000     1
2          2         2  1.000000 -1.000000 -1.000000     1
3          3         3 -1.000000  1.000000 -1.000000     1
4          4         4  1.000000  1.000000 -1.000000     1
5          5         5 -1.000000 -1.000000  1.000000     1
6          6         6  1.000000 -1.000000  1.000000     1
7          7         7 -1.000000  1.000000  1.000000     1
8          8         8  1.000000  1.000000  1.000000     1
9          9         9  0.000000  0.000000  0.000000     1
10        10        10  0.000000  0.000000  0.000000     1
11        11        11  0.000000  0.000000  0.000000     1
12        12        12  0.000000  0.000000  0.000000     1
13        13        13  0.000000  0.000000  0.000000     1
14        14        14  0.000000  0.000000  0.000000     1
15         1         1 -1.681793  0.000000  0.000000     2
16         2         2  1.681793  0.000000  0.000000     2
17         3         3  0.000000 -1.681793  0.000000     2
18         4         4  0.000000  1.681793  0.000000     2
19         5         5  0.000000  0.000000 -1.681793     2
20         6         6  0.000000  0.000000  1.681793     2
21         7         7  0.000000  0.000000  0.000000     2
22         8         8  0.000000  0.000000  0.000000     2
23         9         9  0.000000  0.000000  0.000000     2
24        10        10  0.000000  0.000000  0.000000     2
25        11        11  0.000000  0.000000  0.000000     2
26        12        12  0.000000  0.000000  0.000000     2

Data are stored in coded form using these coding formulas ...
x1 ~ x1.as.is
```

```
x2 ~ x2.as.is
x3 ~ x3.as.is
```

```
y<-c(7.6,7.9,8.9,7.1,10.2,7.8,11.9,8.3,11.5,11.2,13.8,10.7,11,10.9,10.8,6,7.9,
     7.3,5,9.8)
data94 <- cbind(data.frame(design94)[1:20,3:5], y)
data94
```

```
          x1        x2        x3    y
1  -1.000000 -1.000000 -1.000000  7.6
2   1.000000 -1.000000 -1.000000  7.9
3  -1.000000  1.000000 -1.000000  8.9
4   1.000000  1.000000 -1.000000  7.1
5  -1.000000 -1.000000  1.000000 10.2
6   1.000000 -1.000000  1.000000  7.8
7  -1.000000  1.000000  1.000000 11.9
8   1.000000  1.000000  1.000000  8.3
9   0.000000  0.000000  0.000000 11.5
10  0.000000  0.000000  0.000000 11.2
11  0.000000  0.000000  0.000000 13.8
12  0.000000  0.000000  0.000000 10.7
13  0.000000  0.000000  0.000000 11.0
14  0.000000  0.000000  0.000000 10.9
15 -1.681793  0.000000  0.000000 10.8
16  1.681793  0.000000  0.000000  6.0
17  0.000000 -1.681793  0.000000  7.9
18  0.000000  1.681793  0.000000  7.3
19  0.000000  0.000000 -1.681793  5.0
20  0.000000  0.000000  1.681793  9.8
```

7.6.6.2. Box-Benken 설계

요인의 개수가 $k = 3$ 이고 중심점에서 실험의 개수가 $n_0 = 2$ 인 Box-Benken 설계의 실험점을 패키지 `rsm` 의 함수 `bbd()` 를 통하여 구해보자 (교과서 표 9.6).

함수 `bbd()`를 사용하는 경우 인자는 다음과 같다.

- `k = 3` : 요인의 개수(k)를 3개로 지정
- `n0 = 2` : 중심점의 수 (n_0) 를 2개로 지정
- `randomize = F` : 실험의 순서를 임의화하지 않는다.

```
designBB <- bbd ( k= 3, n0 = 2,  randomize = F)
designBB
```

```
   run.order std.order x1.as.is x2.as.is x3.as.is
1          1         1       -1       -1        0
2          2         2        1       -1        0
```

```
3          3          3         -1          1          0
4          4          4          1          1          0
5          5          5         -1          0         -1
6          6          6          1          0         -1
7          7          7         -1          0          1
8          8          8          1          0          1
9          9          9          0         -1         -1
10        10         10          0          1         -1
11        11         11          0         -1          1
12        12         12          0          1          1
13        13         13          0          0          0
14        14         14          0          0          0

Data are stored in coded form using these coding formulas ...
x1 ~ x1.as.is
x2 ~ x2.as.is
x3 ~ x3.as.is
```

7.7. 이차반응표면분석 사례

7.7.1. 개요

- 교과서 9.6 절의 반응표면분석 사례 예제
- 교과서에서는 Design Expert 프로그램을 사용하였지만 본 강의노트에서는 R 의 rsm 패키지를 사용하여 사례 분석

7.7.2. 실험의 목적과 개요

빵을 포장하는 비닐 봉지의 접착력 y 을 가장 크게 하는 마감 공정의 조건을 찾는 실험을 수행하려고 한다. 실험에서 고려하는 반응변수 y 는 접착력이고 설명변수는 3개를 고려하는데 변수의 정의, 고려하는 범위와 변환식는 다음과 같다.

표 7.5.: 실험에서 고려하는 설명변수

독립변수 이름	설명(단위)	관심 범위	상자점 $(-1, 1)$ 변환식
T	마감온도(섭씨 온도)	(100, 140)	$x_1 = (T - 120)/20$
C	냉각온도(섭씨 온도)	(5, 15)	$x_2 = (C - 10)/5$
P	폴리 에틸렌 첨가제의 양(%)	(0.5, 1.7)	$x_2 = (P - 1.1)/(0.6)$

반응표면분석은 다음과 같은 2차 모형을 고려하여 접착력을 최대로 하는 최적점을 찾으려고 한다.

$$y = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \beta_{ii} x_i^2 + \sum_{i=1}^{2} \sum_{j=i+1}^{3} \beta_{ij} x_i x_j + e \tag{7.21}$$

다음은 반응표면분석을 위한 패키지 rsm 에서 함수 rsm() 을 이용하여 1차 또는 2차 다항모형을 적합하는 경우 사용되는 모형식의 종류와 설명이다.

표 7.6.: 함수 rsm() 의 모형식

모형식	설명	사용의 예	식
FO	first-order, 1차 다항식	FO(x1, x2, x3)	$\sum_{i=1}^{3} \beta_i x_i$
TWI	two-way interaction, 두 변수의 상호작용	TWI(x1,x2)	$\sum_{i=1}^{2} \sum_{j=i+1}^{3} \beta_{ij} x_i x_j$
PQ	pure quadratic , 제곱항	PQ(x1)	$\sum_{i=1}^{3} \beta_{ii} x_i^2$
SO	second-order, 2차 다항식	SO(x1, x2, x3)	all terms

7.7.3. 중심합성설계의 실험점 생성

설명변수(요인)의 개수가 3개이며 회전가능한 중심합성설계를 사용한 실험점은 다음과 같이 ccd() 함수로 구할 수 있다.

- ccd() 함수에서 독립 변수의 변환식을 지정할 수 있으며 실험점이 원래 변수(T, C, P)의 값으로 표시된다.
- 중심점에서 6개의 실험을 하려고 한다. ccd() 함수는 지정된 n0 값보다 2배의 실험점을 생성하기 때문에 n0 = 6 를 사용하고 마지막 6개의 중심점을 제거하였다.

```
mydesign0 <- ccd(3, n0 = 6, alpha = "rotatable", randomize = F, coding =list(x1 ~ (T - 120)/20, x2 ~ (C
mydesign0
```

```
   run.order std.order         T         C         P Block
1          1         1 100.00000  5.000000 0.5000000     1
2          2         2 140.00000  5.000000 0.5000000     1
3          3         3 100.00000 15.000000 0.5000000     1
4          4         4 140.00000 15.000000 0.5000000     1
5          5         5 100.00000  5.000000 1.7000000     1
6          6         6 140.00000  5.000000 1.7000000     1
7          7         7 100.00000 15.000000 1.7000000     1
8          8         8 140.00000 15.000000 1.7000000     1
9          9         9 120.00000 10.000000 1.1000000     1
10        10        10 120.00000 10.000000 1.1000000     1
11        11        11 120.00000 10.000000 1.1000000     1
12        12        12 120.00000 10.000000 1.1000000     1
13        13        13 120.00000 10.000000 1.1000000     1
14        14        14 120.00000 10.000000 1.1000000     1
15         1         1  86.36414 10.000000 1.1000000     2
16         2         2 153.63586 10.000000 1.1000000     2
17         3         3 120.00000  1.591036 1.1000000     2
18         4         4 120.00000 18.408964 1.1000000     2
19         5         5 120.00000 10.000000 0.0909243     2
20         6         6 120.00000 10.000000 2.1090757     2
21         7         7 120.00000 10.000000 1.1000000     2
22         8         8 120.00000 10.000000 1.1000000     2
23         9         9 120.00000 10.000000 1.1000000     2
24        10        10 120.00000 10.000000 1.1000000     2
25        11        11 120.00000 10.000000 1.1000000     2
```

```
26        12        12 120.00000 10.000000 1.1000000     2

Data are stored in coded form using these coding formulas ...
x1 ~ (T - 120)/20
x2 ~ (C - 10)/5
x3 ~ (P - 1.1)/0.6
```

```
mydesign <- mydesign0[1:20,3:5]
mydesign
```

```
           T         C         P
1  100.00000  5.000000 0.5000000
2  140.00000  5.000000 0.5000000
3  100.00000 15.000000 0.5000000
4  140.00000 15.000000 0.5000000
5  100.00000  5.000000 1.7000000
6  140.00000  5.000000 1.7000000
7  100.00000 15.000000 1.7000000
8  140.00000 15.000000 1.7000000
9  120.00000 10.000000 1.1000000
10 120.00000 10.000000 1.1000000
11 120.00000 10.000000 1.1000000
12 120.00000 10.000000 1.1000000
13 120.00000 10.000000 1.1000000
14 120.00000 10.000000 1.1000000
15  86.36414 10.000000 1.1000000
16 153.63586 10.000000 1.1000000
17 120.00000  1.591036 1.1000000
18 120.00000 18.408964 1.1000000
19 120.00000 10.000000 0.0909243
20 120.00000 10.000000 2.1090757

Data are stored in coded form using these coding formulas ...
x1 ~ (T - 120)/20
x2 ~ (C - 10)/5
x3 ~ (P - 1.1)/0.6
```

7.7.4. 실험자료 읽어오기

이제 실제 실험을 수행하고 반응값인 접착력을 측정한 자료를 읽어보자. 측정한 자료는 화일 chap9_rsm.csv 에 저장되어 있으며 다음과 같이 자료를 읽어서 데이터프레임`rsm_data` 를 생성할 수 있다.

> **!** 중요
>
> 자료 화일은 온라인강의실 14주차 섹션에서 화일로 다운로드받을 수 있다.

주의할 점은 화일 `chap9_rsm.csv` 이 현재 작업 경로(working directory)에 있다고 가정한다. 현재 작업 경로에 화일이 없으면 화일의 전체 경로를 지정해주어야 한다.

```
rsm_data0 <- read.csv(here::here("data","chap9_rsm.csv"), header = T)
rsm_data0
```

```
           T         C         P    y
1  100.00000  5.000000 0.5000000  7.6
2  140.00000  5.000000 0.5000000  7.9
3  100.00000 15.000000 0.5000000  8.9
4  140.00000 15.000000 0.5000000  7.1
5  100.00000  5.000000 1.7000000 10.2
6  140.00000  5.000000 1.7000000  7.8
7  100.00000 15.000000 1.7000000 11.9
8  140.00000 15.000000 1.7000000  8.3
9   86.36414 10.000000 1.1000000 10.8
10 153.63586 10.000000 1.1000000  6.0
11 120.00000  1.591036 1.1000000  7.9
12 120.00000 18.408964 1.1000000  7.3
13 120.00000 10.000000 0.0909243  5.0
14 120.00000 10.000000 2.1090757  9.8
15 120.00000 10.000000 1.1000000 11.5
16 120.00000 10.000000 1.1000000 11.2
17 120.00000 10.000000 1.1000000 13.8
18 120.00000 10.000000 1.1000000 10.7
19 120.00000 10.000000 1.1000000 11.0
20 120.00000 10.000000 1.1000000 10.9
```

이제 위의 표에서 제시된 변환식이 적용된 실험자료의 데이터프레임을 만들자.

```
rsm_data <- coded.data(rsm_data0, x1 ~ (T - 120)/20, x2 ~ (C - 10)/5, x3 ~ (P-1.1)/0.6)
rsm_data
```

```
           T         C         P    y
1  100.00000  5.000000 0.5000000  7.6
2  140.00000  5.000000 0.5000000  7.9
3  100.00000 15.000000 0.5000000  8.9
4  140.00000 15.000000 0.5000000  7.1
5  100.00000  5.000000 1.7000000 10.2
6  140.00000  5.000000 1.7000000  7.8
7  100.00000 15.000000 1.7000000 11.9
8  140.00000 15.000000 1.7000000  8.3
9   86.36414 10.000000 1.1000000 10.8
10 153.63586 10.000000 1.1000000  6.0
11 120.00000  1.591036 1.1000000  7.9
12 120.00000 18.408964 1.1000000  7.3
```

```
13 120.00000 10.000000 0.0909243  5.0
14 120.00000 10.000000 2.1090757  9.8
15 120.00000 10.000000 1.1000000 11.5
16 120.00000 10.000000 1.1000000 11.2
17 120.00000 10.000000 1.1000000 13.8
18 120.00000 10.000000 1.1000000 10.7
19 120.00000 10.000000 1.1000000 11.0
20 120.00000 10.000000 1.1000000 10.9

Data are stored in coded form using these coding formulas ...
x1 ~ (T - 120)/20
x2 ~ (C - 10)/5
x3 ~ (P - 1.1)/0.6
```

7.7.5. 2차 다항식 모형의 적합

7.7.5.1. 모형의 적합

이제 자료 `rsm_data` 에 대하여 식 식 7.21 으로 표현된 2차 선형모형을 적합해보자. 2차 다항식의 적합은 함수 `rsm()` 에서 `SO(x1+x2+x3)`를 사용한다.

아래의 적합한 결과를 요약하면 다음과 같다.

- 1차항(`FO`)은 `x1` 과 `x3` 가 유의하다.
- 상호작용(`TWI`)은 모두 유의하지 않다.
- 2차항(`PQ`)은 모두 유의하다.
- 2차 다항식의 정준분석으로 3개의 고유값이 모두 음수이다 $(-0.6064, -1.2442, -1.3711)$
- 따라서 최적점은 반응변수가 최대가 되는 실험점이다.
- 최적점은 원자료의 단위로 $T = 100.86$, $C = 11.45$, $P = 1.52$이다.

```
res2 <- rsm(y ~ SO(x1, x2, x3), data = rsm_data)
summary(res2)
```

```
Call:
rsm(formula = y ~ SO(x1, x2, x3), data = rsm_data)

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 11.47992    0.47725 24.0542 3.508e-10 ***
x1          -1.14028    0.31665 -3.6011  0.004839 **
x2           0.12382    0.31665  0.3910  0.703984
x3           1.08170    0.31665  3.4161  0.006590 **
x1:x2       -0.41250    0.41372 -0.9971  0.342252
x1:x3       -0.56250    0.41372 -1.3596  0.203816
x2:x3        0.21250    0.41372  0.5136  0.618665
x1^2        -0.86177    0.30825 -2.7957  0.018933 *
x2^2        -1.14462    0.30825 -3.7133  0.004019 **
```

```
x3^2        -1.21533    0.30825 -3.9427  0.002763 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared:  0.8553,    Adjusted R-squared:  0.725
F-statistic: 6.566 on 9 and 10 DF,  p-value: 0.003469

Analysis of Variance Table

Response: y
                Df Sum Sq Mean Sq F value   Pr(>F)
FO(x1, x2, x3)   3 33.946 11.3153  8.2636 0.004630
TWI(x1, x2, x3)  3  4.254  1.4179  1.0355 0.418243
PQ(x1, x2, x3)   3 42.719 14.2397 10.3992 0.002037
Residuals       10 13.693  1.3693
Lack of fit      5  7.065  1.4129  1.0658 0.472962
Pure error       5  6.628  1.3257

Stationary point of response surface:
        x1         x2         x3
-0.9569885  0.2907543  0.6919092

Stationary point in original units:
         T          C          P
100.860230  11.453771   1.515146

Eigenanalysis:
eigen() decomposition
$values
[1] -0.6064157 -1.2442428 -1.3710551

$vectors
         [,1]       [,2]        [,3]
x1  0.8054829 -0.3224169 -0.49723696
x2 -0.3957602 -0.9171818 -0.04638263
x3 -0.4411022  0.2341470 -0.86637408
```

7.7.5.2. 등고선과 3차원 그림

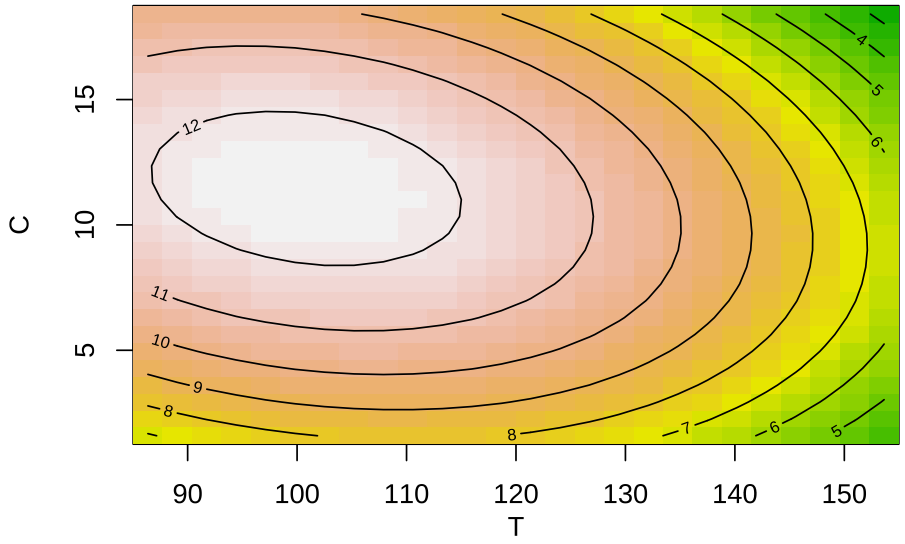
이제 각 두 개의 요인에 대하여 적합된 2차 다항식 반응표면의 예측값을 등고선 그림과 3차원 그림으로 나타내어 보자.

이 실험에서는 3개의 독립변수를 사용하였기 때문에 등고선 그림을 그리는 경우 2개의 변수를 사용해서 그려야 한다. 이 경우 나머지 독립변수의 값은 주어진 값으로 고정시켜야 한다. 함수 `contour()` 에서 `at=summary(res2)$canonical$xs` 는 두 독립변수를 축으로 그림을 그릴때 나머지 독립변수의 값을 최적점으로 지정해주는 옵션이다.

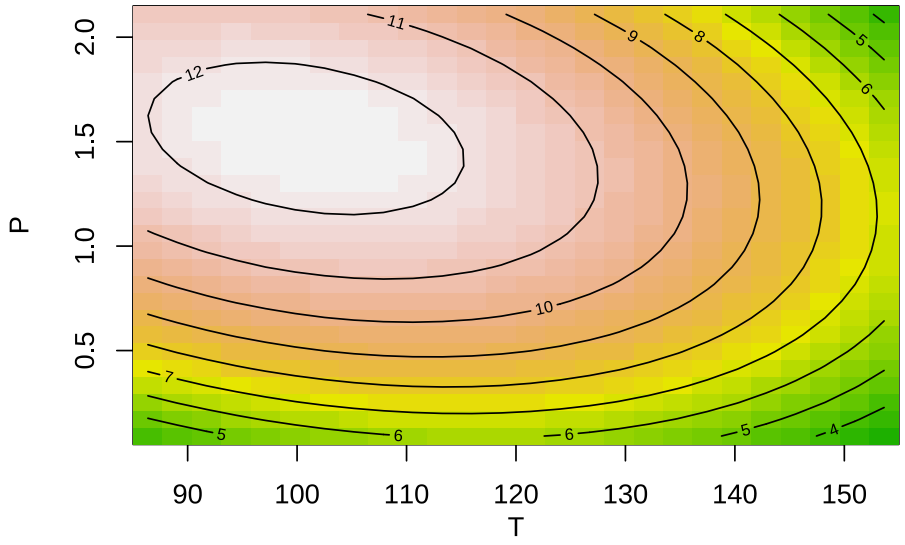
```
summary(res2)$canonical$xs # 최적실험점
```

```
        x1         x2         x3
-0.9569885  0.2907543  0.6919092
```

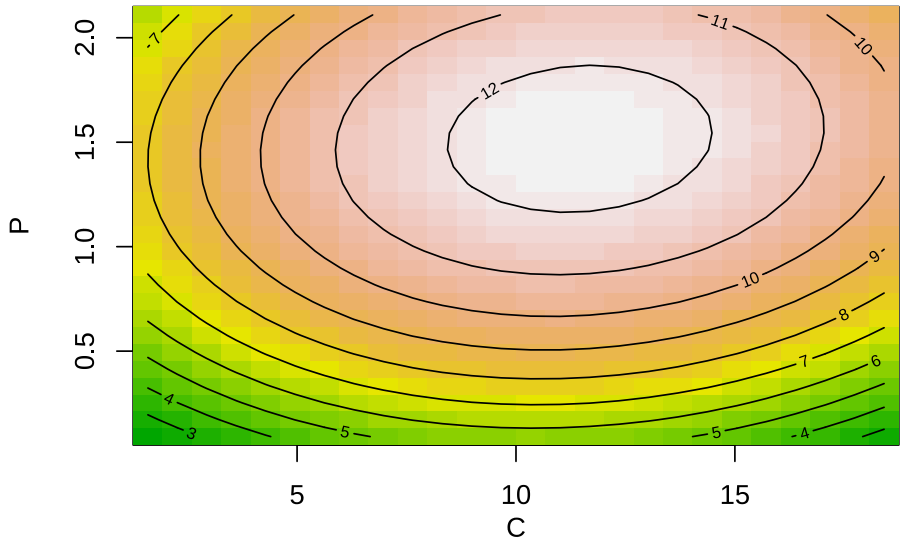
```
par(mar = c(4, 4, .1, .1))
contour (res2, ~ x1+x2+x3, image = TRUE, at=summary(res2)$canonical$xs)
```



Slice at P = 1.52, x1 = -0.956988519983455, x2 = 0.290754285136

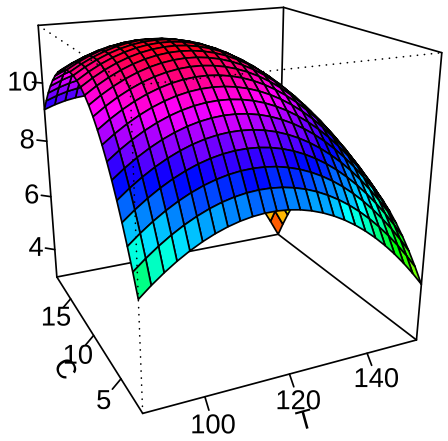


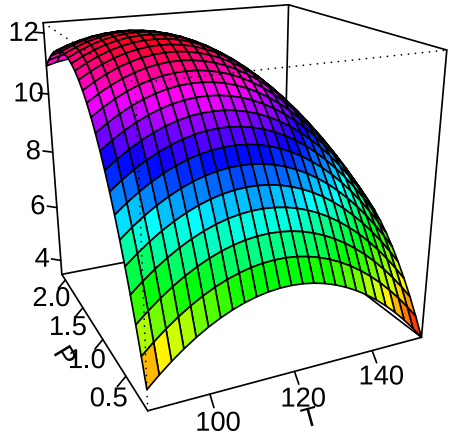
Slice at C = 11.45, x1 = -0.956988519983455, x3 = 0.691909216502

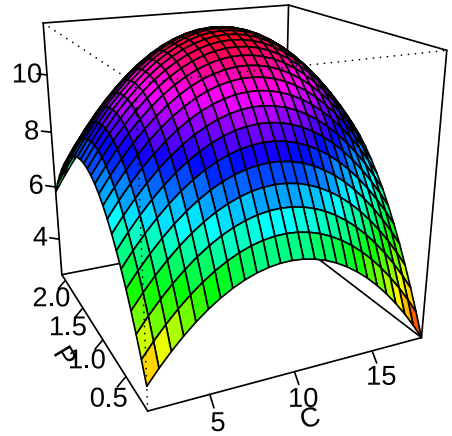


Slice at T = 100.86, x2 = 0.290754285136889, x3 = 0.691909216502

```
par(mar = c(4, 4, .1, .1))
persp(res2, x2~x1, col = rainbow(50))
persp(res2, x3~x1, col = rainbow(50))
persp(res2, x3~x2, col = rainbow(50))
```







7.7.6. 변수 선택

앞에서 2차 다항식의 모형을 적합한 경우 상호작용 효과가 유의하지 않았다. 이제 2차 다항식 모형을 `lm()` 함수를 이용하여 적합한 후에 함수 `step()`을 이용하여 후방제거법(backward elimination)으로 변수 선택을 실행해 보자. `step()` 에서는 별도로 지정하지 않으면 AIC(Akaike Information Criteria)에 의한 최적 모형을 선택해 준다.

2차 다항식 모형을 `lm()` 함수로 적합하는 경우는 상호작용과 2차식을 모두 모형식에 아래와 같이 포함시켜주어야 한다.

후방제거법(backward elimination)을 실행한 결과 다음과 같은 모형이 최종적으로 선택되었다.

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + e$$

```
res21 <- lm(y~x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + I(x1^2) + I(x2^2) + I(x3^2), data=rsm_data)
summary(res21)
```

```
Call:
lm.default(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 +
    I(x1^2) + I(x2^2) + I(x3^2), data = rsm_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.22327 -0.50492 -0.07776  0.31191  2.32008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.4799     0.4773  24.054 3.51e-10 ***
x1           -1.1403     0.3166  -3.601  0.00484 **
x2            0.1238     0.3166   0.391  0.70398
x3            1.0817     0.3166   3.416  0.00659 **
I(x1^2)      -0.8618     0.3082  -2.796  0.01893 *
I(x2^2)      -1.1446     0.3082  -3.713  0.00402 **
I(x3^2)      -1.2153     0.3082  -3.943  0.00276 **
x1:x2        -0.4125     0.4137  -0.997  0.34225
x1:x3        -0.5625     0.4137  -1.360  0.20382
x2:x3         0.2125     0.4137   0.514  0.61866
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.17 on 10 degrees of freedom
Multiple R-squared:  0.8553,    Adjusted R-squared:  0.725
F-statistic: 6.566 on 9 and 10 DF,  p-value: 0.003469
```

```
step(res21 , direction = "backward")
```

```
Start:  AIC=12.42
y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + I(x1^2) + I(x2^2) +
    I(x3^2)

          Df Sum of Sq    RSS    AIC
- x2:x3    1    0.3613 14.054 10.944
- x1:x2    1    1.3613 15.054 12.319
<none>                 13.693 12.423
- x1:x3    1    2.5313 16.224 13.816
- I(x1^2)  1   10.7026 24.396 21.973
- I(x2^2)  1   18.8809 32.574 27.756
- I(x3^2)  1   21.2857 34.979 29.180

Step:  AIC=10.94
y ~ x1 + x2 + x3 + I(x1^2) + I(x2^2) + I(x3^2) + x1:x2 + x1:x3

          Df Sum of Sq    RSS    AIC
```

```
- x1:x2    1    1.3613 15.416 10.793
<none>                  14.054 10.944
- x1:x3    1    2.5313 16.586 12.256
- I(x1^2)  1   10.7026 24.757 20.267
- I(x2^2)  1   18.8809 32.935 25.976
- I(x3^2)  1   21.2857 35.340 27.386

Step:  AIC=10.79
y ~ x1 + x2 + x3 + I(x1^2) + I(x2^2) + I(x3^2) + x1:x3

          Df Sum of Sq    RSS     AIC
- x2       1    0.2094 15.625  9.0627
<none>                  15.416 10.7929
- x1:x3    1    2.5312 17.947 11.8336
- I(x1^2)  1   10.7026 26.118 19.3379
- I(x2^2)  1   18.8809 34.296 24.7862
- I(x3^2)  1   21.2857 36.701 26.1416

Step:  AIC=9.06
y ~ x1 + x3 + I(x1^2) + I(x2^2) + I(x3^2) + x1:x3

          Df Sum of Sq    RSS     AIC
<none>                  15.625  9.0627
- x1:x3    1    2.5312 18.156 10.0656
- I(x1^2)  1   10.7026 26.327 17.4976
- I(x2^2)  1   18.8809 34.506 22.9079
- I(x3^2)  1   21.2857 36.911 24.2553



Call:
lm.default(formula = y ~ x1 + x3 + I(x1^2) + I(x2^2) + I(x3^2) +
    x1:x3, data = rsm_data)

Coefficients:
(Intercept)           x1           x3      I(x1^2)      I(x2^2)      I(x3^2)
    11.4799      -1.1403       1.0817      -0.8618      -1.1446      -1.2153
      x1:x3
    -0.5625
```

7.7.7. 최종모형 선택

7.7.7.1. 최종모형의 선택과 적합

앞에서 후방제거법(backward elimination)을 실행한 결과에서 주 효과 x_2 가 제외되었지만 2차항 x_2^2 이 선택되었으므로 주 효과 x_2 도 최종 모형에 포함시킨다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + e \tag{7.22}$$

따라서 위에서 고려한 최종모형 식 7.22 을 선택하여 반응표면 분석을 다시 실행해 보자.

- 최종모형의 정준분석으로 3개의 고유값이 모두 음수이다 $(-0.7064, -1.1446, -1.3707)$
- 따라서 최적점은 반응변수가 최대가 되는 실험점이다.
- 최종모형에서 최적점은 원자료의 단위로 $T = 102.55$, $C = 10.27$, $P = 1.49$이다.

```
finalres <- rsm(y ~ FO(x1, x2, x3) + TWI(x1, x3) + PQ(x1,x2,x3), data = rsm_data)
summary(finalres)
```

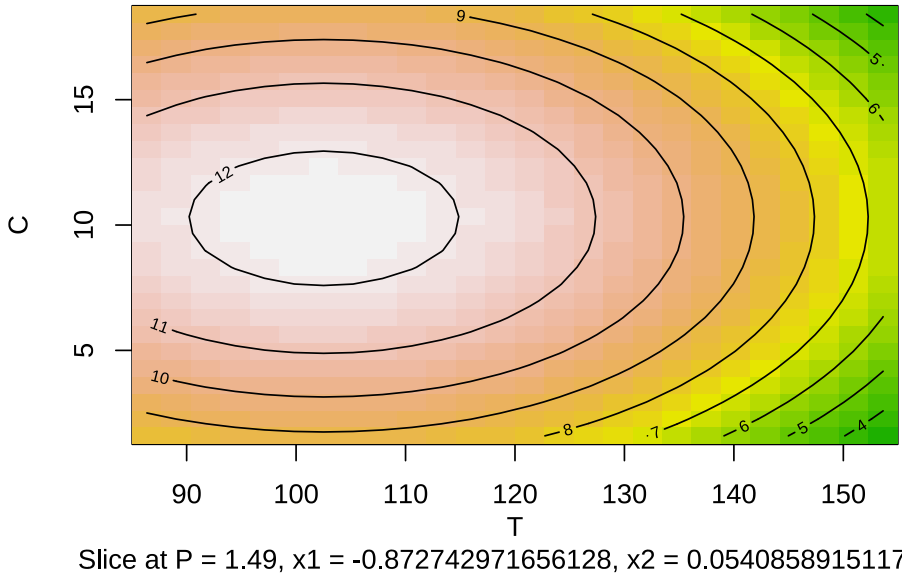
```
Call:
rsm(formula = y ~ FO(x1, x2, x3) + TWI(x1, x3) + PQ(x1, x2, x3),
    data = rsm_data)

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 11.47992    0.46226 24.8343 1.099e-11 ***
x1          -1.14028    0.30670 -3.7179  0.002938 **
x2           0.12382    0.30670  0.4037  0.693534
x3           1.08170    0.30670  3.5269  0.004170 **
x1:x3       -0.56250    0.40072 -1.4037  0.185750
x1^2        -0.86177    0.29856 -2.8864  0.013667 *
x2^2        -1.14462    0.29856 -3.8337  0.002379 **
x3^2        -1.21533    0.29856 -4.0706  0.001552 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R-squared:  0.8371,    Adjusted R-squared:  0.742
F-statistic: 8.807 on 7 and 12 DF,  p-value: 0.0006411

Analysis of Variance Table

Response: y
               Df Sum Sq Mean Sq F value   Pr(>F)
FO(x1, x2, x3)  3 33.946 11.3153  8.8082 0.002327
TWI(x1, x3)     1  2.531  2.5313  1.9704 0.185750
PQ(x1, x2, x3)  3 42.719 14.2397 11.0847 0.000896
Residuals      12 15.416  1.2846
Lack of fit     7  8.787  1.2553  0.9469 0.543851
Pure error      5  6.628  1.3257

Stationary point of response surface:
         x1          x2          x3
-0.87274297  0.05408589  0.64699393
```

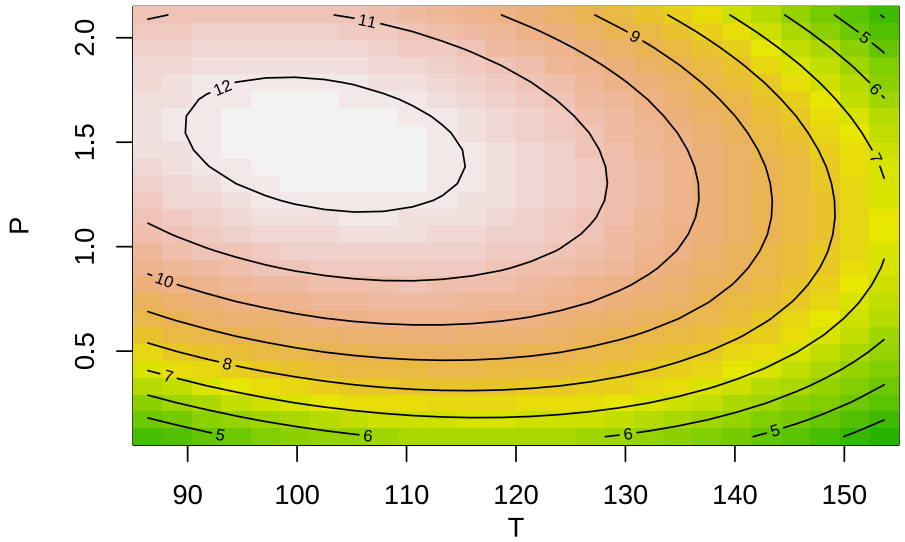
```
Stationary point in original units:
         T          C          P
102.545141  10.270429   1.488196

Eigenanalysis:
eigen() decomposition
$values
[1] -0.7063571 -1.1446152 -1.3707412

$vectors
         [,1] [,2]      [,3]
x1  0.8752577    0 0.4836569
x2  0.0000000    1 0.0000000
x3 -0.4836569    0 0.8752577
```

7.7.7.2. 등고선과 3차원 그림

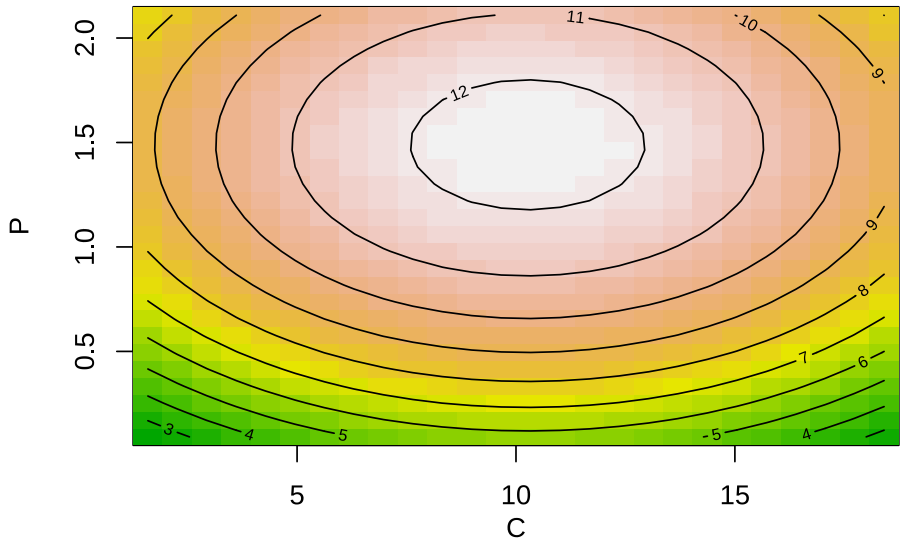
이제 적합된 최종모형에서 각 두 개의 요인에 대하여 반응표면의 예측값을 등고선 그림과 3차원 그림으로 나타내어 보자.

```
par(mar = c(4, 4, .1, .1))
contour (finalres, ~ x1+x2+x3, image = TRUE, at=summary(finalres)$canonical$xs)
```



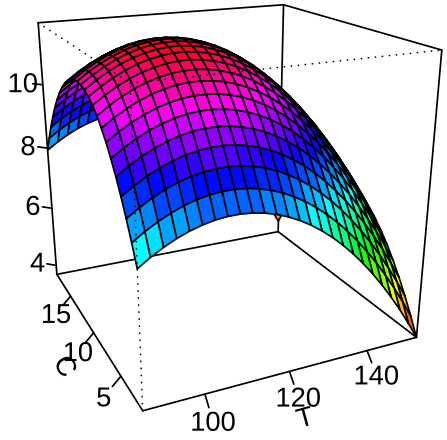


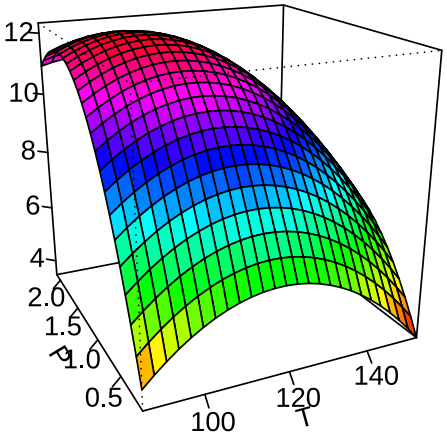
Slice at C = 10.27, x1 = -0.872742971656128, x3 = 0.646993934911

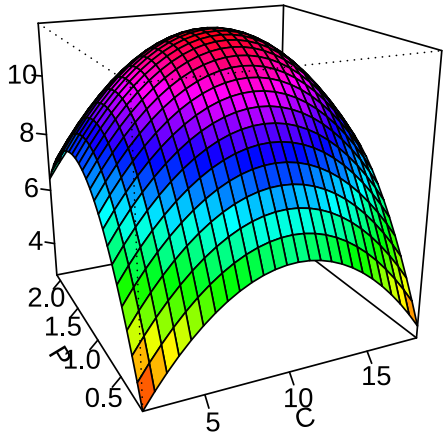


Slice at T = 102.55, x2 = 0.0540858915117846, x3 = 0.64699393491

```
par(mar = c(4, 4, .1, .1))
persp(finalres, x2~x1, col = rainbow(50))
persp(finalres, x3~x1, col = rainbow(50))
persp(finalres, x3~x2, col = rainbow(50))
```


10
8
6
4
15
10
C
5
100
120
140
T


12
10
8
6
4
2.0
1.5
P
1.0
0.5
100
120
140
T


10
8
6
4
2.0
1.5
P
1.0
0.5
5
10
15
C

8. 의학연구에서의 실험계획법

8.1. 공분산분석

의학연구에서는 치료법이나 약물의 효과를 알아보기 위한 다양한 실험이 진행된다.

치료법이나 약품의 효과에 대하여 기초 연구(화학실험, 동물실험 등)가 어느 정도 진행되어 기대되는 효능이 있으며 큰 부작용이 없다고 판단되면 인간에게 치료법을 적용하거나 약품을 사용하는 임상실험(clinical trial)을 진행한다.

임상실험은 많은 경우 **병렬 계획(parallel design)**에 의거한 일원배치법을 사용한다.

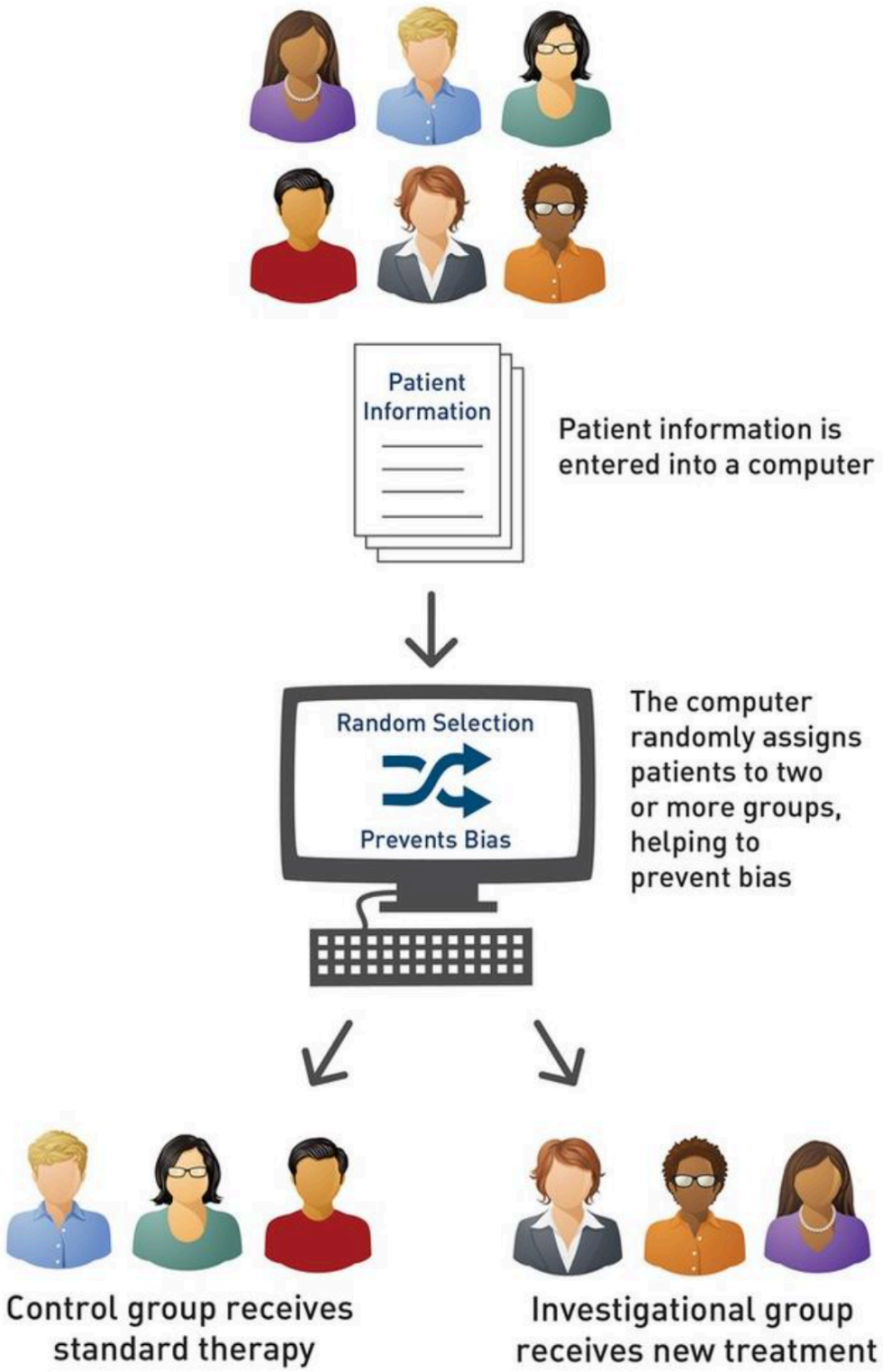


그림 출처

임상실험에서 요인은 치료 방법(treatmanet)이며 보통의 경우 2개의 수준를 가진다. 두 개의 수준 중 하나는 실험자가 연구 대상으로 고려한 효과가 기대되는 치료/약품(active)이며 다른 하나의 수준은 제어 수준(control)이다. 임상실험에서는 대부분의 경우 효과가 있는 치료나 약품을 적용하지 않아도 **위약효과(placebo effect)**가 나타나기 때문에 실험자가 사용하려고 하는 치료법의 효과는 언제나 제어군에서 나타난 효과와의 차이로 파악해야 한다.

병렬 계획은 환자가 임의로 선택된 하나의 치료/약품만 받는 실험을 말한다. 참고로 뒤에서 살펴보겠지만 환자가 2개 이상의

치료을 받는 교차실험(crossover design)도 있다.

8.1.1. 공분산 분석의 개요

많은 임상 실험에서는 실험의 질과 수준이 실제 환자를 치료하는 환경과 동일하게 유지하는 것을 원칙으로 한다. 따라서 임상실험은 동장이나 연구실에서 수행하는 매우 정교하게 통제된 실험과 다르게 치료방법 외의 다양한 요인들이 영향을 미치게 된다. 이러한 다양한 요인들은 맹검화(blinding) 등 다양한 실험 기법을 사용하여 통제된다.

다양한 요인의 영향을 통제하려는 시도에도 불구하고 대표적으로 실험의 결과에 영향을 미치는 요인은 환자의 초기 상태(baseline)과 기관/병원(center/hospital effect)이다. 임상실험에 참가하는 환자들은 약품을 처리받기 전의 상태가 모두 다르기 때문에 처리의 효과뿐만이 아니라 환자의 초기 상태도 최종 반응값에 영향을 미친다. 또한 대부분의 임상실험은 여러 개의 병원(또는 지역, 나라)에서 동시에 실행되므로 병원, 지역, 국가의 특성에 따라서 임상시험의 결과에 영향을 미친다.

이렇게 실험에서 주요하게 고려하는 요인인 아닌 다른 요인이 영향을 미친다고 판단될 때 그 요인을 **공변량 (covariate)** 라고 부르며 공변량을 모형에 포함시키는 분석을 **공분산분석(Analysis of Covariance; ANCOVA)**라고 부른다.

공변량의 형태는 보통 실험 단위가 가지고 있는 특성이나 실험자가 가진 특성을 반영한다. 예를 들어 다음과 같은 공변량의 형태가 있다.

- 교육 방법을 비교하는 실험에서 학생들이 가진 학습 역량
- 혈압 강하를 위한 약품에 대한 실험에서 임상 참가 전 환자의 혈압과 나이
- 천식에 대한 약품 실험이 여러 국가에서 실행될때 국가의 효과
- 암을 진단하는 방법에 대한 임상 실험이 다수의 병원에서 진행될 때 병원의 효과

일반적으로 임상실험이나 관측연구에서는 관심이 있는 처리(treatment)나 요인(factor)뿐만 아니라 다른 요인들도 반응변수에 영향을 미친다. 이러한 다른 요인들의 영향을 제거하기 위한 방법은 여러가지가 있지만 실험인 경우 임의화 방법(randomization)으로 그 영향을 상쇄시킬 수 도 있고 관측연구인 경우에는 사례-대조연구 방법을 이용하여 그 영향을 최소화하려고 노력을 한다. 하지만 다양한 통제 방법에도 불구하고 여러 가지 변수들이 반응변수에 영향을 미친다. 이러한 경우에 중요한 요인을 모형에 포함시켜서 그 영향을 반영하고 동시에 자료의 변동을 부가적으로 설명해주는 방법이 공분산 분석이다.

8.1.2. 공분산분석의 모형

실험에서 공변량은 연속형 변수일 수도 있고 범주형일 수 도 있다. 만약 공변량이 범주형 변수인 경우 분석의 방법은 이원배치 분산분석과 매우 유사하다. 이 절에서는 공변량은 연속형 변수라고 가정하고 분석 방법을 논의할 것이다. 또한 공변량이 2개 이상인 경우도 있지만 이 절에서는 공변량이 하나인 경우만 고려한다.

공분산분석의 모형은 일원배치 분산분석 모형(처리의 수는 a개, 반복수는 r)에 공변량 x의 효과를 다음과 같이 더해주는 것이다.

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + e_{ij}, \quad i = 1, 2, \dots, a, \;\; j = 1, 2, \dots, r \tag{8.1}$$

모형 식 8.1 에서 x_{ij}는 관측값 y_{ij}의 공변량이며 이를 중심화(centering)하여 회귀모형의 독립변수로 표현한다. 모형 식 8.1 에서 $\bar{x}_{..} = sum_i sum_j x_{ij}/(ar)$ 로 공변량 값의 전체 평균이다.

참고로 공변량을 중심화 하지 않는 모형은 다음과 같이 표현할 수 있다.

$$
\begin{aligned}
y_{ij} &= \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + e_{ij} \\
&= \beta_0 + \alpha_i + \beta x_{ij} + e_{ij} \\
&= \beta_{0i} + \beta x_{ij} + e_{ij}
\end{aligned}
$$

8.1.3. 모수의 추정과 가설 검정

모형 식 8.1 에서 각 모수의 추정은 최소제곱법을 이용하여 추정하며 부가조건 $\sum_i \alpha_i = 0$ 을 이용하면 다음과 같은 추정량을 얻을 수 있다

$$
\begin{aligned}
\hat{\mu} &= \bar{y}_{..} \\
\hat{\alpha}_i &= \bar{y}_{i.} - \bar{y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \\
\hat{\beta} &= \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2}
\end{aligned}
$$

ANCOVA 모형에서는 다음과 같은 두 가지 가설을 검정할 수 있다.

ANOVA 모형에서와 같이 각 그룹의 평균에 대한 검정을 할 수 있고

$$
H_0 : \alpha_1 = \alpha_2 = ... = \alpha_a = 0 \quad \text{vesus} \quad H_1 : \text{ not } H_0
$$

또한 공변량의 효과에 대한 검정도 할 수 있다.

$$
H_0 : \beta = 0 \quad \text{vesus} \quad H_1 : \beta \neq 0 \tag{8.2}
$$

8.1.4. 예제: 혈압 강하를 위한 임상 실험

공분산분석의 개념을 이해하기 위한 예제로서 혈압 강하를 위한 임상 실험에서 얻는 자료를 분석해 보자.

자료는 (Chen, Peace, 와/과 Zhang 2017) 에 나온 자료이며 화일 dbp.txt 에 저장되어 있다.

혈압 강하를 위한 임상실험은 2개의 처리 집단(`A` 와 `B`) 로 각각 20명이 실험에 참가하였다. 약을 복용하기 전에 혈압을 측정하고(`DBP1`) 약을 복용한 후 한 달 간격으로 4번 측정을 하였다 (`DBP2-DBP5`).

최종적으로 관심있는 반응변수는 약품을 복용하기 전 혈압에서 4개월 후 혈압이 변화한 차이이다. 따라서 반응변수 `diff` 는 `DBP5`에서 `DBP1`을 뺀 값이다.

또한 공변량으로서 성별(`Sex`)과 연령(`Age`)를 측정하였다. 공분산분석에서는 공변량으로 연령을 고려할 것이다.

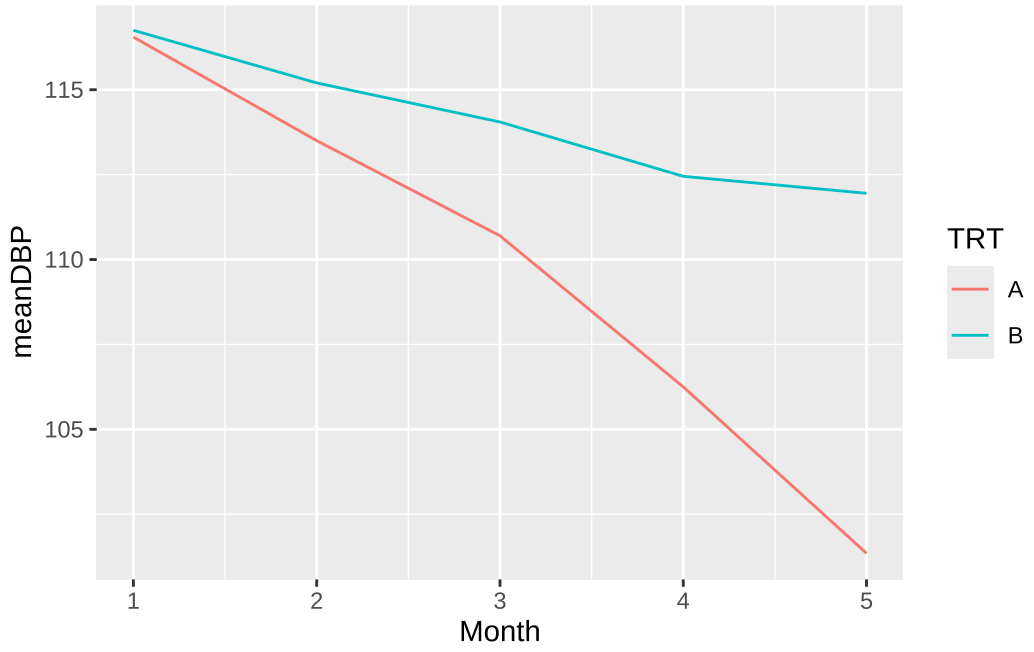
다음은 화일에서 자료를 읽고 정리하는 프로그램이다.

```
dpb <- read.csv(here::here("data","dbp.txt"), sep="", header=TRUE)
df <- dpb %>% mutate(diff = DBP5 -DBP1)
head(df)
```

```
  Subject TRT DBP1 DBP2 DBP3 DBP4 DBP5 Age Sex diff
1       1   A  114  115  113  109  105  43   F   -9
2       2   A  116  113  112  103  101  51   M  -15
3       3   A  119  115  113  104   98  48   F  -21
4       4   A  115  113  112  109  101  42   F  -14
5       5   A  116  112  107  104  105  49   M  -11
6       6   A  117  112  113  104  102  47   M  -15
```

처리 그룹간에 시간에 따른 혈압의 평균적인 변화를 그름으로 살펴보자.

```
df1 <- df %>% select(Subject, TRT, DBP1, DBP2, DBP3, DBP4, DBP5) %>% gather(TimePoint, dbp,DBP1, DBP2,
df1s <- df1 %>% group_by(TRT, TimePoint)  %>%  summarise(meanDBP=mean(dbp))
df1s$Month <- rep(1:5,2)
df1s %>% ggplot(aes(x = Month,  y = meanDBP, colour = TRT)) + geom_line()
```



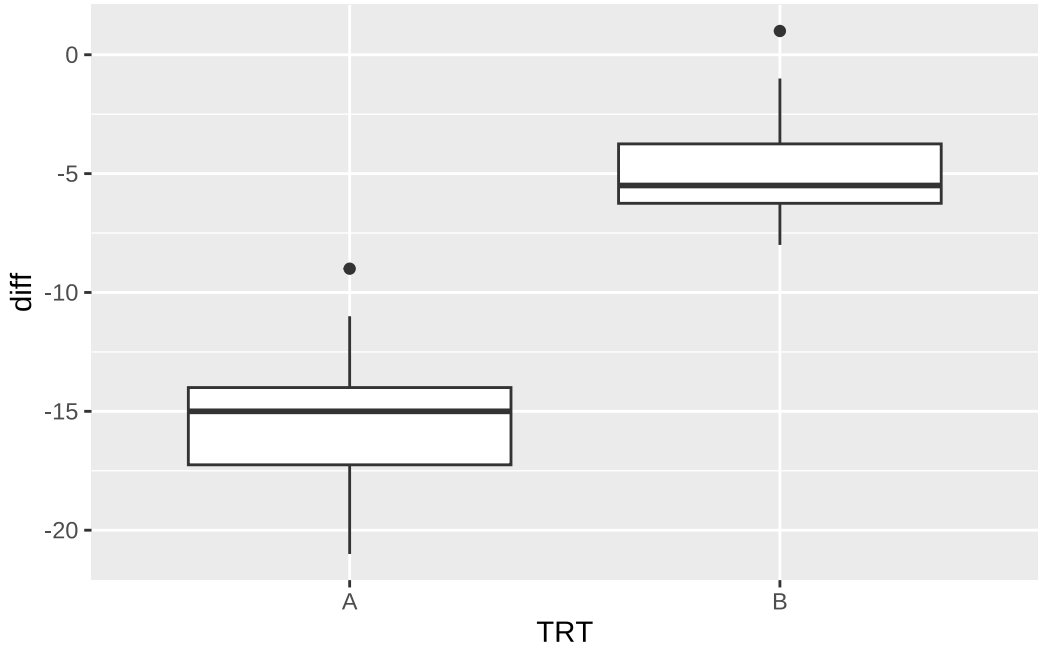
이제 두 처리 그룹간에 혈압의 변화 `diff` 에 대하여 기초통계량으로 살펴보자.

```
dfs2 <- df %>% group_by(TRT)  %>%  summarise(mean=mean(diff), median= median(diff), sd=sd(diff), min=mi
dfs2
```

```
# A tibble: 2 x 6
  TRT    mean median    sd   min   max
  <chr> <dbl>  <dbl> <dbl> <int> <int>
1 A     -15.2  -15    2.97   -21    -9
2 B      -4.8   -5.5  2.42    -8     1
```

두 처리 그룹간에 혈압의 변화에 대하여 그림으로 살펴보자.

```
ggplot(df, aes(TRT, diff)) +  geom_boxplot()
```



A 그룹이 B 그룹보다 평균적으로 혈압이 약 10 mmHg 더 감소하였다.

8.1.4.1. 분산분석

8.1.4.1.1. 공변량이 없는 일원배치법에서의 분산분석

우리는 아래와 같은 일원배치 실험계획에서 처리 효과에 대한 검정을 위한 분산분석표가 아래와 같이 주어지는 것을 배웠다.

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

요인	제곱합	자유도	평균제곱합	F_0	p-값
처리	SS_A	$\phi_A = a - 1$	$MS_A = SS_A/\phi_A$	$F_0 = MS_A/MS_E$	$P[F(\phi_A, \phi_E) > F_0]$
잔차	SS_E	$\phi_E = a(r-1)$	$MS_E = SS_E/\phi_E$		
총합	SS_T	$\phi_T = ar - 1$			

위의 분산분석표에서 다음과 같이 제곱합의 분해가 얻어진다.

$$SS_T = SS_A + SS_E \tag{8.3}$$

혈압 자료에 처리 효과만 있는 일원배치법으로 분산분석표를 구해보자. 두 처리 집단 사이에 혈압의 변화에 대한 차이는 매우 유의하다. 참고로 가설 검정에 이용되는 F-값은 147.63 이다.

```
lmres1 <- lm(diff~TRT, data=df)
lmaov1 <- anova(lmres1)
lmaov1
```

```
Analysis of Variance Table

Response: diff
          Df Sum Sq Mean Sq F value    Pr(>F)
TRT        1 1081.6 1081.60  147.63 1.169e-14 ***
Residuals 38  278.4    7.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8.1.4.1.2. 공변량이 있는 일원배치법에서의 분산분석

공변량이 포함된 모형 식 8.1 에 대한 분산분석표는 다음과 같이 주어진다.

요인	제곱합	자유도	평균제곱합	F_0	p-값
변량	SS_X	$\phi_X = 1$	$MS_X = SS_X/\phi_X$	$F_X = MS_X/MS_E$	
처리	SS_A	$\phi_A = a-1$	$MS_A = SS_A/\phi_A$	$F_0 = MS_A/MS_E$	$P[F(\phi_A,\phi_E) > F_0]$
잔차	SS'_E	$\phi_E = a(r-1)-1$	$MS'_E = SS'_E/\phi_E$		
총합	SS_T	$\phi_T = ar-1$			

이제 혈압 자료에 연령을 공변량으로 포함한 일원배치법으로 공분산분석표를 구해보자. 두 처리 집단 사이에 혈압의 변화에 대한 차이는 매우 유의하다. 참고로 가설 검정에 이용되는 F-값은 176.03 로서 공변량이 없는 경우(147.63)보다 크다.

```
lmres2 <- lm(diff~TRT + Age, data=df)
lmaov2 <- anova(lmres2)
lmaov2
```

```
Analysis of Variance Table

Response: diff
          Df  Sum Sq Mean Sq  F value    Pr(>F)
TRT        1 1081.60 1081.60 176.0395 1.228e-15 ***
Age        1   51.07   51.07   8.3119  0.006525 **
Residuals 37  227.33    6.14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

공변량이 있는 경우 분산분석표에서 다음과 같이 제곱합의 분해가 얻어진다.

$$SS_T = SS_A + SS_X + SS'_E \tag{8.4}$$

이제 공변량이 없는 경우의 제곱합의 분해 식 8.3 과 있는 경우의 분해 식 8.3 을 보면 공변량이 없는 경우의 오차제곱합이 두 개의 제곱합으로 분해되는 것을 알 수 있다.

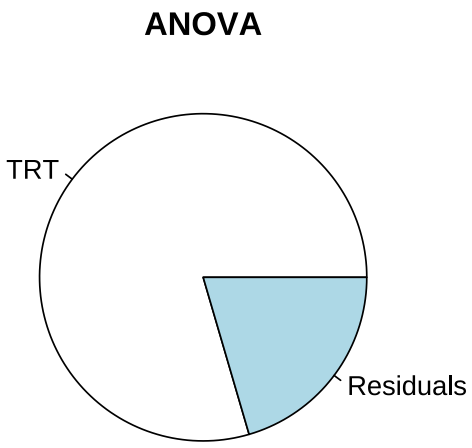
$$SS_E = SS_X + SS'_E \tag{8.5}$$

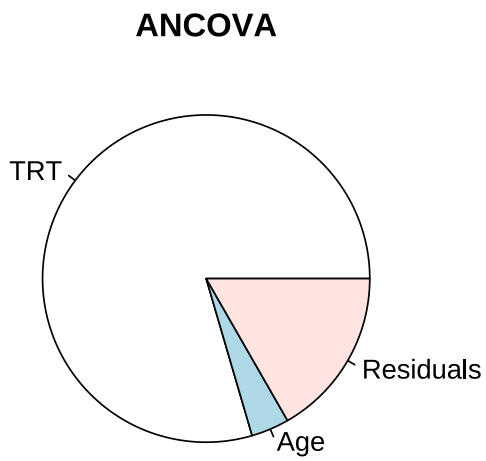
즉, 만약 반응변수와 공변량의 상관관계가 크면 공변량에 대한 제곱합 SS_X가 커질 것이며 이는 공변량이 있는 모형에서 처리효과를 검정하는 경우 사용되는 오차제곱합 SS'_E 가 공변량이 없는 경우의 SS_E 보다 작아지는 것을 알 수 있다.

결론적으로 반응변수와 공변량의 상관관계가 크면, 공변량을 포함하는 모형에서 처리 효과를 검정하는 F-값이 공변량을 포함하지 않는 것보다 일반적으로 커지게 된다. 이는 공변량이 처리효과로 설명하지 못하는 변동 중의 일부를 설명하기 때문에 처리효과에 대한 검정력이 높아지게 된다.

공변량의 유무에 따른 제곱합의 분해의 차이를 그림으로 나타내면 다음과 같다.

```
ssq1 <- data.frame(effect=c("TRT", "Residuals"), sumsquare=lmaov1$"Sum Sq")
ssq2 <- data.frame(effect=c("TRT", "Age", "Residuals"), sumsquare=lmaov2$"Sum Sq")

par(mar = c(2, 2, 2, 2))
pie(ssq1$sumsquare, labels = ssq1$effect, main="ANOVA")
pie(ssq2$sumsquare, labels = ssq2$effect, main = "ANCOVA")
```





참고로 공변량을 포함하는 모형에서 오차제곱합의 자유도는 포함하지 않는 모형보다 1개가 줄어든다. 이는 공변량에 의한 제곱합의 자유도 1개가 추가되기 때문이다.

8.1.4.2. 공분산분석 모형의 해석

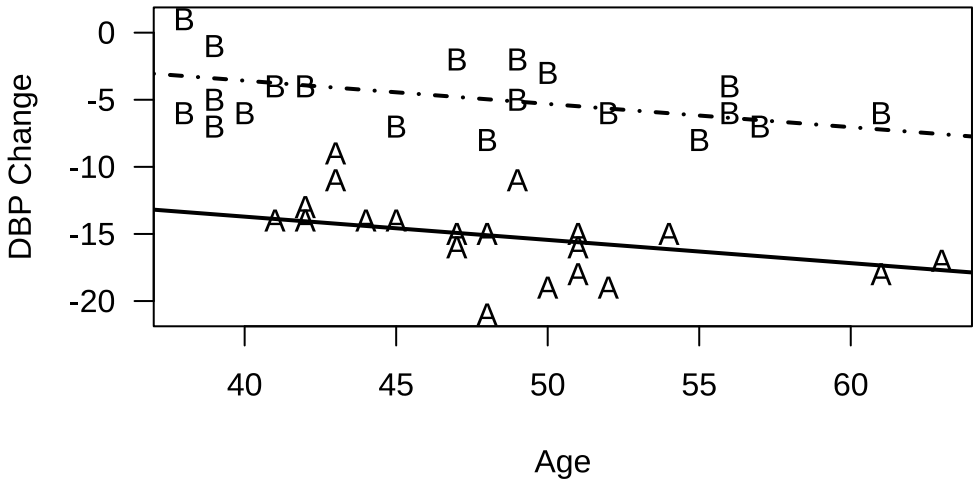
이제 위에서 구한 공분산분석 모형 식 8.1 에 대한 추정식의 계수를 살펴보자.

```
summary(lmres2)
```

```
Call:
lm.default(formula = diff ~ TRT + Age, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9039 -1.6516 -0.0091  1.1557  5.2299

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.78086    2.97236  -2.281  0.02838 *
TRTB        10.13149    0.78936  12.835 3.38e-15 ***
Age         -0.17323    0.06009  -2.883  0.00653 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.479 on 37 degrees of freedom
Multiple R-squared:  0.8328,    Adjusted R-squared:  0.8238
F-statistic: 92.18 on 2 and 37 DF,  p-value: 4.243e-15
```

이제 처리 그룹(A, B)과 연령 간의 관계를 그림으로 그려보면 다음과 같이 나타낼 수 있다. 연령이 증가하면 혈압의 강화 효과가 점점 더 커지는 것을 알 수 있으며 통계적으로도 유의하다.

```
plot(diff~Age,las=1,pch=as.character(TRT), df, xlab="Age", ylab="DBP Change")
abline(lmres2$coef[1], lmres2$coef[3],lwd=2, lty=1)
abline(lmres2$coef[1]+lmres2$coef[2], lmres2$coef[3],lwd=2, lty=4)
```



8.2. 임상살험의 목적

8.2.1. 임상실험의 목적: 우월성, 동등성, 비열등성

- 우월성 (superiority)
  - T is superior to S
  - Treatment T has more therapeutic effect than S.
- 동등성(equivalence)
  - T is equivalent to S
  - Two treatments T and S have equal therapeutic effect.
- 비열등성(noninferiority)
  - T is noninferior to S
  - Treatment T is not inferior to S (T is as effective as S)

Table 1. Hypotheses Associated with the Different Types of Studies when Comparing a New Therapy Against a Current Therapy with Respect to Efficacy

Type of study	Null hypotheses	Research hypothesis
Traditional comparative	There is no difference between the therapies	There is a difference between the therapies
Equivalence	The therapies are not equivalent	The new therapy is equivalent to current therapy
Noninferiority	The new therapy is inferior to the current therapy	The new therapy is not inferior to the current therapy

출처: Walker 와/과 Nowacki (2011)

8.2.2. 통계적 가설

평균이 큰 것이 좋다고 가정하자.

- 우월성 연구의 가설

$$H_0 : \mu_T = \mu_S \quad \text{vs.} \quad H_1 : \mu_T \neq \mu_S$$

- 사실상 우월성 실험의 대립 가설은 $H_1 : \mu_T > \mu_S$ 이다.
- 우월성에 대한 가설을 아래와 같이 세울수도 있지만 거의 사용하지 않는다.

$$H_0 : \mu_T \leq \mu_S + \delta \quad \text{vs.} \quad H_1 : \mu_T > \mu_S + \delta$$

- δ는 margin 이라고 부르며 의학적(임상적)으로 우월한 차이를 보이는 치료 효과의 차이를 말한다.
- 동등성(equivalence)

$$H_0 : |\mu_T - \mu_S| \geq \delta \quad \text{vs.} \quad H_1 : |\mu_T - \mu_S| < \delta$$

- 비열등성(noninferiority)

$$H_0 : \mu_T \leq \mu_S - \delta \quad \text{vs.} \quad H_1 : \mu_T > \mu_S - \delta$$

8.2.3. 통계적 검정 절차 - 우월성

- 우월성 연구의 가설

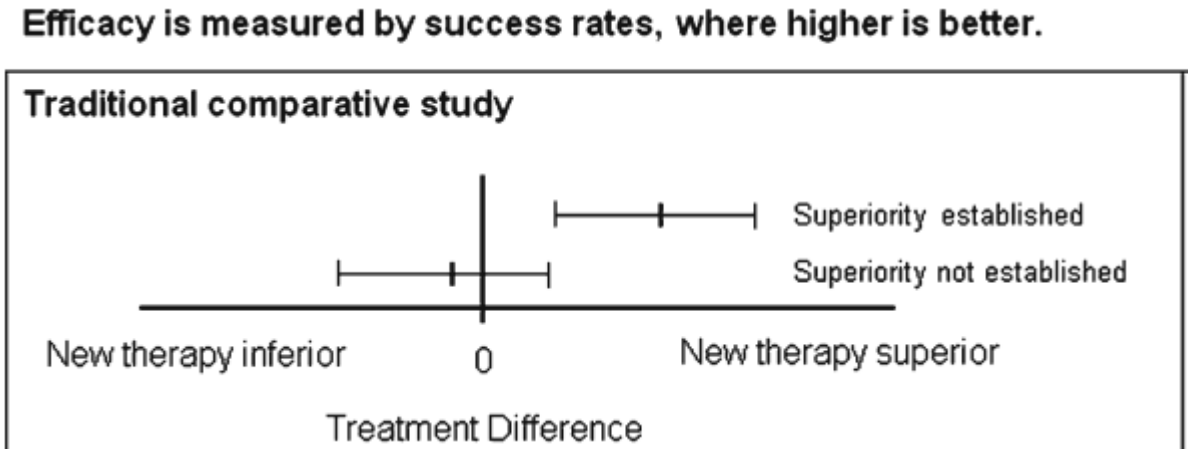
$$H_0 : \mu_T = \mu_S \quad \text{vs.} \quad H_1 : \mu_T \neq \mu_S$$

- 일반적인 t-검정 또는 z-검정을 사용
- 귀무가설의 기각 조건

$$\frac{\hat{\mu}_T - \hat{\mu}_S}{se(\hat{\mu}_T - \hat{\mu}_S)} > c_{\alpha/2}$$

또는

$$(\hat{\mu}_T - \hat{\mu}_S) - c_{\alpha/2} se(\hat{\mu}_T - \hat{\mu}_S) > 0$$



출처: Walker 와/과 Nowacki (2011)

8.2.4. 통계적 검정 절차 - 비열등성

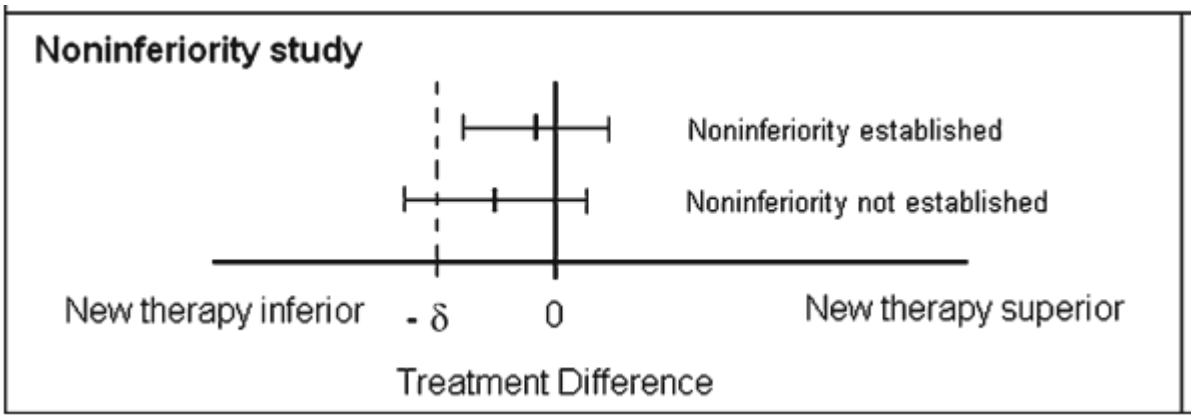
- 비열등성의 가설

$$H_0 : H_0 : \mu_T \leq \mu_S - \delta \quad \text{vs.} \quad H_1 : \mu_T > \mu_S - \delta$$

#{eq-noninf}

- 일반적인 t-검정 또는 z-검정을 사용
- 귀무가설의 기각 조건

$$(\hat{\mu}_T - \hat{\mu}_S) - c_\alpha \; se(\hat{\mu}_T - \hat{\mu}_S) > -\delta$$



출처: Walker 와/과 Nowacki (2011)

8.2.5. 통계적 검정 절차 - 동등성

-동등성 가설

$$H_0 : |\mu_T - \mu_S| \geq \delta \quad \text{vs.} \quad H_1 : |\mu_T - \mu_S| < \delta \tag{8.6}$$

- 일반적인 t-검정 또는 z-검정이 아닌 **the two one-sided test**를 사용
- 가설 식 8.6 의 귀무가설은 다음 두 개의 단측 귀무 가설의 합집합니다.

$$H_{01} : \mu_T - \mu_S \geq \delta \qquad H_{02} : \mu_T - \mu_S \leq -\delta \tag{8.7}$$

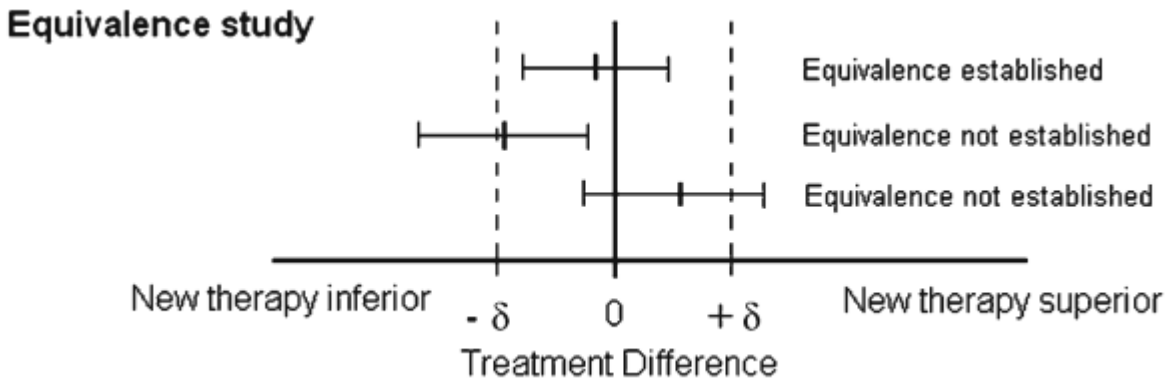
따라서 식 8.7 에 있는 두 개의 가설이 모두 기각되면 식 8.6 에 있는 동등성 가설을 기각할 수 있다.

- **the two one-sided test**는 다음과 같은 두 조건이 만족되면 귀무가설을 기각한다.

$$\frac{\hat{\mu}_T - \hat{\mu}_S - \delta}{se(\hat{\mu}_T - \hat{\mu}_S)} < -c_\alpha \quad \text{and} \quad \frac{\hat{\mu}_T - \hat{\mu}_S + \delta}{se(\hat{\mu}_T - \hat{\mu}_S)} > c_\alpha$$

위의 귀무가설 기각 조건은 다음과 동일하다. 즉 $100(1-2\alpha)\%$ 신뢰구간이 $(-\delta, \delta)$ 안에 존재하면 귀무가설을 기각한다.

$$-\delta < (\hat{\mu}_T - \hat{\mu}_S) - c_\alpha \;\; se(\hat{\mu}_T - \hat{\mu}_S) < (\hat{\mu}_T - \hat{\mu}_S) + c_\alpha \;\; se(\hat{\mu}_T - \hat{\mu}_S) < \delta$$



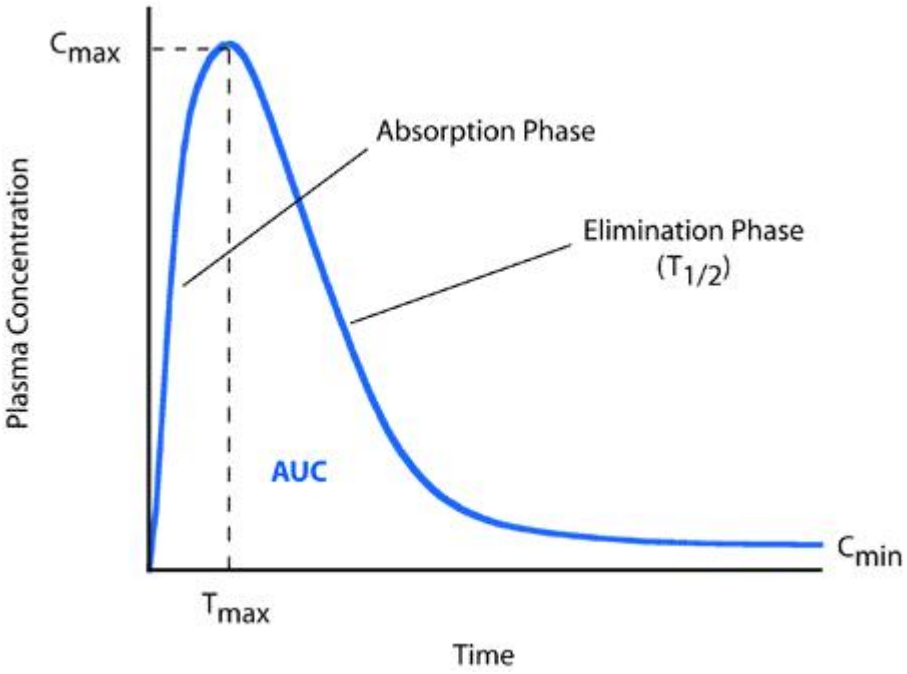
8.2.6. 동등성 검정의 Size 와 유의수준

- $100(1-2\alpha)\%$ 신뢰구간을 이용하는 경우 검정의 Size 와 유의수준은 ? (Kang (2008) 참조)

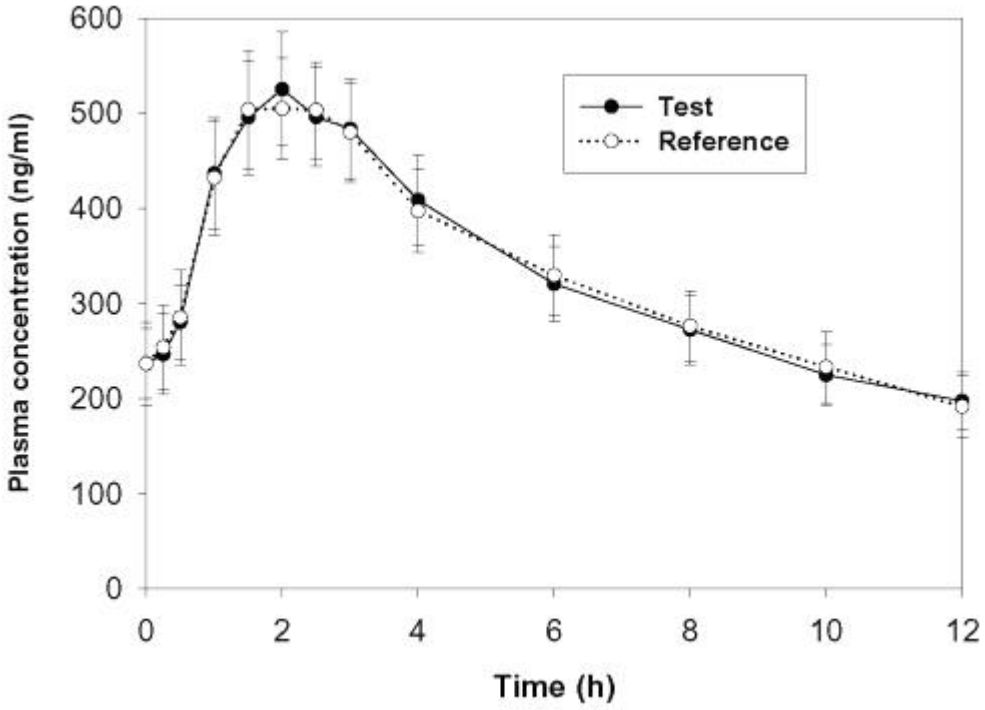
8.3. 교차실험과 동등성 검정

8.3.1. 생체이용률(Bioavailability)

- the rate and extent to which the active ingredient is absorbed from a drug product and becomes available at the site of action
- 주성분 또는 그 활성대사체가 제제로부터 전신순환혈로 흡수되는 속도와 양의 비율
- Pharmacokinetic (PK) measures (평가항목) of bioavailability
  - AUC_t: Area under the blood or plasma concentration-time curve; 일정시간까지 혈중농도-시간곡선하면적
  - C_{max}: Maximum Concentration; 최고혈중농도
  - T_{max}: Time to Maximum Concentration; 최고혈중농도 도달시간



8.3.2. 약품 주성분의 생체이용률의 평균적 변화



8.3.3. 생물학적동등성의 정의: FDA 과 KFDA

- Bioequivalence by FDA

  absence of a significant difference in Bioavailability between two formulations…. when administered at the same molar dose under similar conditions in an appropriately designed study

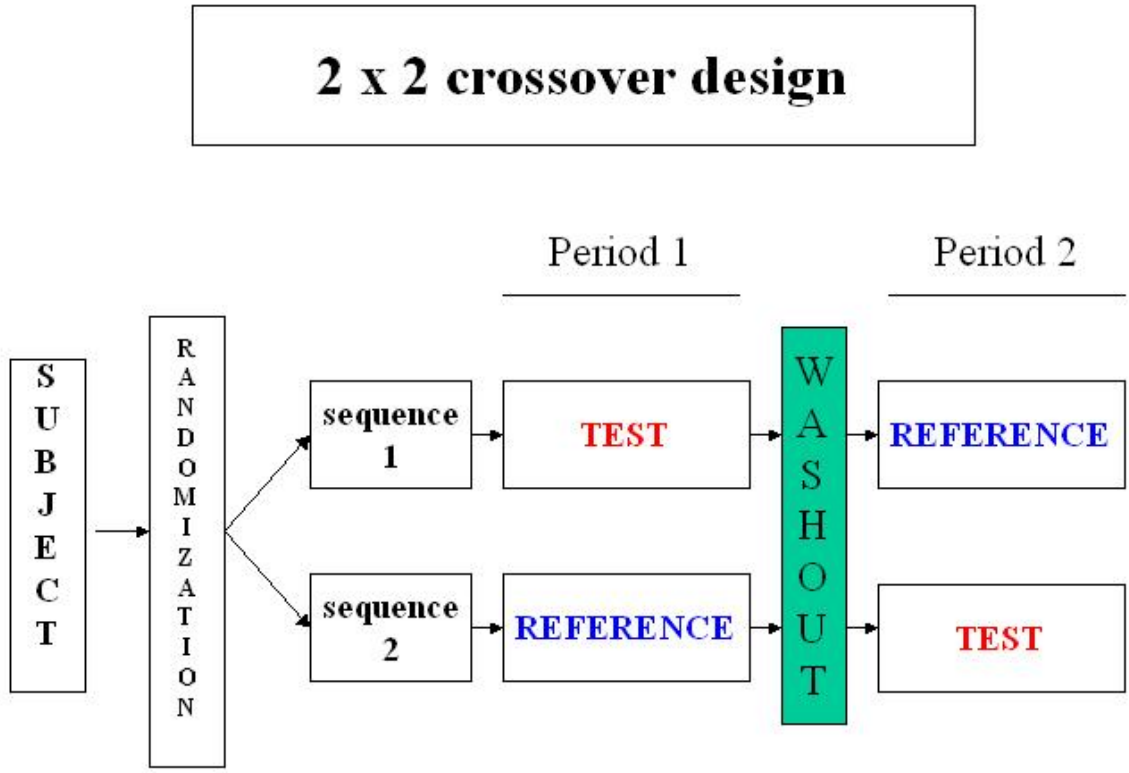
  - in vivo: Bioequivalence
  - in vitro: Bioequivalence

- KFDA

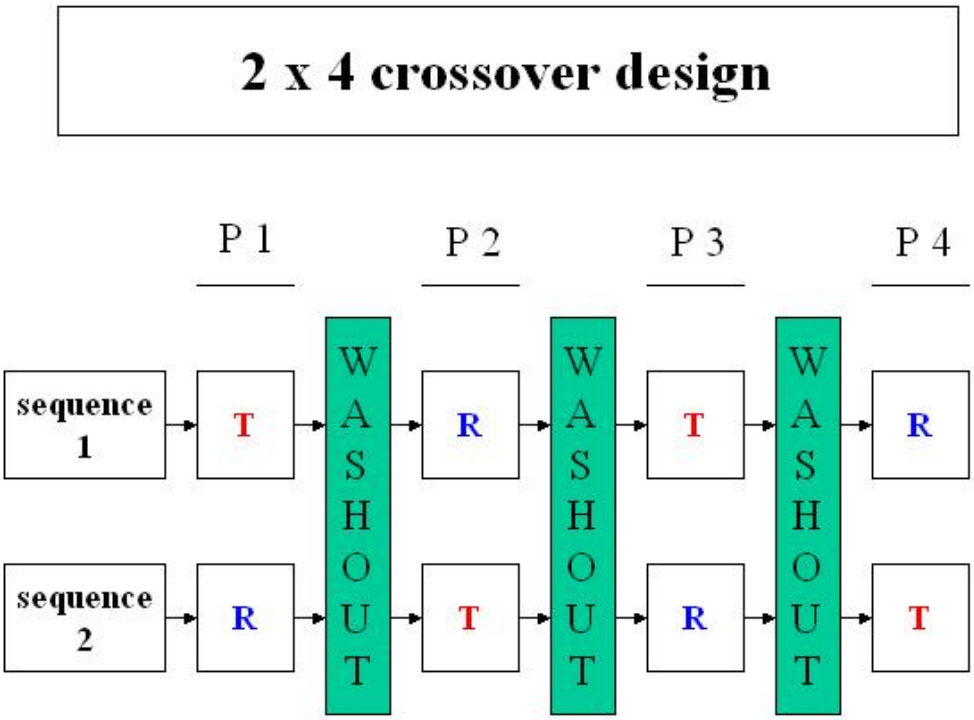
의약품동등성시험이란 그 주성분 · 함량 및 제형이 동일한 두 제제에 대한 의약품동등성을 입증하기 위해 실시하는 생물학적 동등성시험, 비교용출시험, 비교붕해등 기타시험의 생체내 · 외 시험을 말한다.

8.3.4. 생물학적동등성 실험의 설계

- 생체이용률(bioavailibility)은 개인간에 변동이 크다
- 개인효과(individual effect)를 제거하기 위한 쌍비교 t-검정 (paired t-test)의 개념을 도입
- 실험자가 두 개의 처리를 모두 받는다.
- 생동성실험은 주로 **교차시험(crossover design)**을 이용한다.
- 제재의 반감기가 긴 경우 등 특수한 경우는 병렬계획(Parallel design) 실험도 가능하다.
- 2x2 교차실험



- 2x4 교차시험



8.3.5. 교차실험에 대한 통계적 모형

- 보통 10-20명의 실험 대상자
- 각 실험 대상자가 2개(3개 또는 4개)의 반응값(PK responses)을 가진다.
- 각 실험대상자의 반응값은 독립이 아니다 (correlated response; repeated measurements)
- 실험대상자 간의 변이가 크다 (large between-subject variation)
- 시험약과 대조약간의 (로그)반응값의 평균의 차이가 주 검토대상이다.
- 정규분포를 가정한 선형혼합모형(linear mixed model)

8.3.6. 평균적 생물학적동등성에 대한 가설

- The absence of a significant difference (중대한 차이가 없다) in two population means between two formulations .

= 시험약(T)과 대조약(R)간의 반응값의 평균의 차이: $\mu_T - \mu_R$:

- 보통 반응변수(PK response)에 로그를 취한 뒤 통계분석
- $\delta = \mu_T - \mu_R$: 시험약(T)과 대조약(R)간의 로그 반응값의 평균의 차이

- 평균적 생물학적동등성에 대한 가설

$$H_0 : \delta \leq \delta_L \quad or \quad \delta \geq \delta_U \quad vs. \quad H_1 : \delta_L < \delta < \delta_U$$

- 동등성 한계 (bioequivalence limit)

$$\delta_L = -0.223 = log(0.8) \quad and \quad \delta_U = 0.223 = log(1.25)$$

- 평균적 생물학적동등성에 대한 가설(로그변환 전)

$$H_1 : 0.8 < \frac{\mu_T'}{\mu_R'} < 1.25$$

- 평균적 생물학적동등성을 어떤 통계적 방법으로 검정할 것인가?

$$H_0 : \delta \leq \delta_L \quad or \quad \delta \geq \delta_U \quad vs. \quad H_1 : \delta_L < \delta < \delta_U$$

- Historical development of statistical tests for ABE
  - Westlake (1976), Hsu (1984), Bofinger (1985, 1992), Schuirmann (1987), Liu(1990)
  - Berger and Hsu (1996), Brown, Hwang, and Munk (1997), Perlman and Wu (1999), Welleck (2003), Romano (2005)
  - FDA guidance: 1992, 1997, 1999, 2000 and some drafts
- 가설

$$H_0 : |\mu_T - \mu_R| \geq \delta \quad vs. \quad H_1 : |\mu_T - \mu_R| < \delta$$

- 신뢰구간을 이용한 방법
- 2개의 단측검정을 결합한 방법 (Two ones-sided tests; TOST)

$$H_{01} : \mu_T - \mu_R < -\delta \text{ and } H_{02} : \mu_T - \mu_R > \delta$$

8.3.7. 신뢰구간을 이용한 평균적 생물학적동등성 검정

- 가설

$$H_0 : |\mu_T - \mu_R| \geq \delta \quad vs. \quad H_1 : |\mu_T - \mu_R| < \delta$$

- $\mu_T - \mu_R$ 에 대한 신뢰구간 $C(Y)$ 를 구한다.
- 신뢰구간이 동등성 한계안에 포함되면 평균적 생물학적동등성 선언!

$$C(y) \subset (-\delta, \delta)$$

8.3.8. 2개의 단측검정을 이용한 방법

- 정규분포 가정
- 각 처리에 대한 평균 $\bar{y}_T$ 과 $\bar{y}_R$ 는 μ_T 과 μ_R의 추정량
- SE 를 $\bar{y}_T - \bar{y}_R$의 표준 오차(standard error)라고 하자
- 다음을 만족하면 귀무가설 H_0를 기각 (생물학적 동등성을 선언)

$$\frac{\bar{y}_T - \bar{y}_R + \delta}{SE} > t_\alpha \quad \text{and} \quad \frac{\bar{y}_T - \bar{y}_R - \delta}{SE} < -t_\alpha$$

- 위의 귀무가설 기각조건은 아래와 동일한다 (90% 신뢰구간이 동등성 한계안에 있다)

$$[(\bar{y}_T - \bar{y}_R - t_\alpha(SE),\ (\bar{y}_T - \bar{y}_R + t_\alpha(SE))] \subseteq (-\delta,\ \delta)$$

- TOST is a special case of intersection-union test (Berger and Hsu, 1996)
- TOST is level 2α test, but its size is actually α.

$$\text{size of test } = \sup_{H_0} P(\text{ test rejects } H_0)$$

- Improved tests are proposed by Berger and Hsu (1996), Brown, Hwang, and Munk (1997), Perlman and Wu (1999), Welleck (2003), Romano (2005)
- But, still TOST is widely used because of its validity and simplicity

References

Chen, Ding-Geng Din, Karl E Peace, 와/과 Pinggao Zhang. 2017. *Clinical trial data analysis using R and SAS*. CRC Press.

Kang, Seung-Ho. 2008. "동등성 시험을 신뢰구간을 사용하여 검정하는 경우 왜 신뢰도 90". 응용통계연구 21 (5): 867--73.

Montgomery, Douglas C. 2017. *Design and analysis of experiments*. John wiley & sons.

Schabenberger, Oliver, 와/과 Francis J Pierce. 2001. *Contemporary statistical models for the plant and soil sciences*. CRC press.

Walker, Esteban, 와/과 Amy S Nowacki. 2011. "Understanding Equivalence and Noninferiority Testing". *Journal of general internal medicine* 26 (2): 192–96.

임용빈. 2020. *Design Expert, Minitab* 과 *R*을 활용한 실험계획법. 자유아카데미.

A. R을 이용한 자료의 시각화 비교

대부분의 연구나 실험의 목적은 집단(group, 그룹)간의 유의한 차이가 있는지 검증하는 것이다. 집단의 차이는 집단의 특성를 파악할 수 있는 기술 통계량(descriptive statistics)를 사용하여 비교 할 수 있다. 즉 각 집단에 대한 관심변수의 평균, 중앙값 등으로 집단 간 중심의 차이를 비교할 수 있고 표준편차, 사분위범위(Inter Quartile Range; IQR) 등을 사용하면 퍼진 정도도 비교할 수 있다.

이러한 기술 통계량을 이용한 비교도 의미가 있지만 그림을 통하여 집단 간의 차이를 나타내는 것이 자료의 특성을 이해하는데 더 큰 도움이 된다. 그림을 이용하면 자료의 전체적인 퍼진 정도를 파악하기 쉽고 이상치(outlier) 등을 알아내는데 도움이 된다.

이 장에서는 교과서에 제시된 예제 자료를 R 프로그램을 이용하여 분석할 것이다. 기술 통계량과 그림을 이용하여 집단을 비교하는 방법을 알아보고자 한다.

A.1. 두 개 모집단의 비교

A.1.1. 예제 2.2 자료

교재 2장의 예제 2.2 에서 소개된 인장 강도의 자료는 시멘트 공장의 2개의 생산라인에서 생산된 시멘트의 인장 강도를 측정한 것이다. 분석의 목적은 2개의 생산라인의 분포가 동일한지를 비교하는 것이다.

먼저 R로 데이터프레임(data.frame)으로 만들어 보자. 예제 자료를 line1 과 line2 의 벡터 형식으로 만들고 data.frame의 형식인 df0에 저장하려면 다음과 같은 명령어를 사용하면 된다.

```
line1 <- c(16.9, 16.4, 17.2, 16.4, 16.5, 17.0, 17.0, 17.2, 16.6, 16.6)
line2 <- c(16.6, 16.8, 17.4, 17.1, 17.0, 16.9, 17.3, 17.0, 17.1, 17.3)
df220 <- data.frame(line1, line2)
df220
```

```
   line1 line2
1   16.9  16.6
2   16.4  16.8
3   17.2  17.4
4   16.4  17.1
5   16.5  17.0
6   17.0  16.9
7   17.0  17.3
8   17.2  17.0
9   16.6  17.1
10  16.6  17.3
```

data.frame 인 df0에는 각 그룹(line1과 line2)에 대한 10개의 자료가 2개의 열(column)에 각각 저장되어 있다. 이러한 자료의 형태를 넓은 형태의 자료(wide-format data)라고 부른다.

위에서 만든 데이터프레임 df0 를 변형하여 **반응값**들을 하나의 변수(strength)로 합치고, 집단을 나타내는 변수 line를 생성하여 다른 형태의 데이터프레임 df를 다음과 같이 만들어 보자. 아래와 같은 형태의 자료를 좁은 형태의 자료(narrow-format data)라고 부른다. 넓은 형태보다 좁은 형태의 자료가 통계적 분석을 적용하기 편하다.

```
# convert wide format to long format
df22<- df220 %>% pivot_longer(cols = everything(), names_to = "line", values_to = "strength") %>% dplyr


df22
```

```
# A tibble: 20 x 2
   line  strength
   <chr>    <dbl>
 1 line1     16.9
 2 line1     16.4
 3 line1     17.2
 4 line1     16.4
 5 line1     16.5
 6 line1     17
 7 line1     17
 8 line1     17.2
 9 line1     16.6
10 line1     16.6
11 line2     16.6
12 line2     16.8
13 line2     17.4
14 line2     17.1
15 line2     17
16 line2     16.9
17 line2     17.3
18 line2     17
19 line2     17.1
20 line2     17.3
```

A.1.2. 기술 통계량에 의한 요약 - 넓은 형태의 자료

넓은 형태의 자료 df0에 대한 요약통계(평균, 중앙값, 사분위수, 최소, 최대 등)를 다음과 같이 summary 함수를 이용하여 구하고 집단간의 차이를 비교할 수 있다.

```
summary(df220)
```

```
     line1           line2
 Min.   :16.40   Min.   :16.60
 1st Qu.:16.52   1st Qu.:16.93
 Median :16.75   Median :17.05
```

```
 Mean   :16.78   Mean   :17.05
 3rd Qu.:17.00   3rd Qu.:17.25
 Max.   :17.20   Max.   :17.40
```

A.1.3. 기술 통계량에 의한 요약 - 좁은 형태의 자료

좁은 형태의 자료 df에 대해서는 다음과 같이 먼저 group_by함수로 집단을 구별하는 변수를 지정한다. 그 다음으로 summarise함수를 이용하여 여러 가지 통계량를 집단별로 계산할 수 있다.

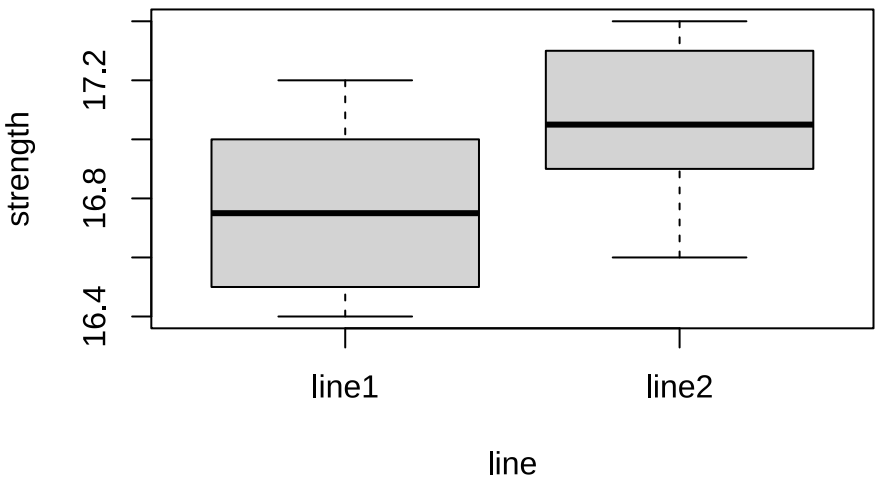
```
df22 %>% group_by(line)  %>%  summarise(mean=mean(strength), median= median(strength), sd=sd(strength),
```

```
# A tibble: 2 x 6
  line   mean median    sd   min   max
  <chr> <dbl>  <dbl> <dbl> <dbl> <dbl>
1 line1  16.8   16.8 0.316  16.4  17.2
2 line2  17.0   17.0 0.246  16.6  17.4
```

A.1.4. 집단 자료에 대한 시각화

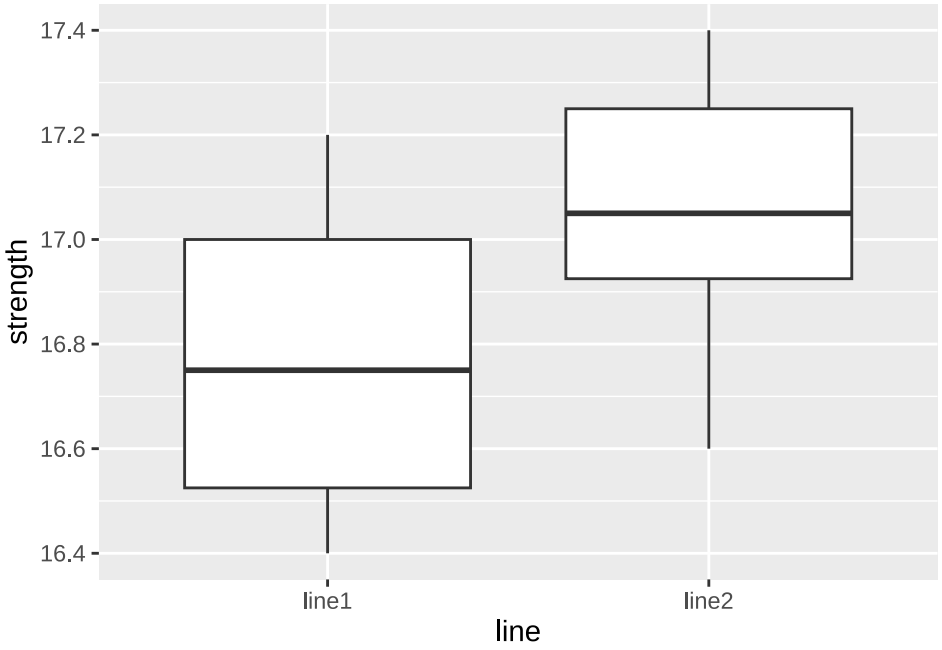
다음으로 각 집단별로 상자그림(boxplot)을 그려서 자료의 중심과 퍼진 정도를 그림으로 비교해 보자. 위에서 좁은 형태로 구성된 자료에 대하여 다음과 같은 명령어로 상자그림을 집단별로 그릴 수 있다.

```
with(df22, boxplot(strength~line))
```



패키지 ggplot2를 사용하면 좀 더 멋진 상자그림으로 시각화를 할 수 있다.

```
ggplot(df22, aes(line, strength)) +  geom_boxplot()
```



A.2. 세 개 이상의 모집단의 비교

A.2.1. 예제 3.1 자료

4개의 서로 다른 원단업체에서 직물을 공급받고 있다. 공급한 직물의 긁힘에 대한 저항력을 알아보기 위하여 각 업체마다 4개의 제품을 랜덤하게 선택하여 일원배치법에 의하여 마모도 검사를 실시하였다. 자료는 다음과 같다.

```
company<- as.factor(rep(c(1:4), each=4))
response<- c(1.93, 2.38, 2.20, 2.25,
             2.55, 2.72, 2.75, 2.70,
             2.40, 2.68, 2.32, 2.28,
             2.33, 2.38, 2.28, 2.25)
df31<- data.frame(company=company, response= response)
df31
```

```
   company response
1        1     1.93
2        1     2.38
3        1     2.20
4        1     2.25
5        2     2.55
6        2     2.72
7        2     2.75
8        2     2.70
9        3     2.40
```

```
10       3     2.68
11       3     2.32
12       3     2.28
13       4     2.33
14       4     2.38
15       4     2.28
16       4     2.25
```

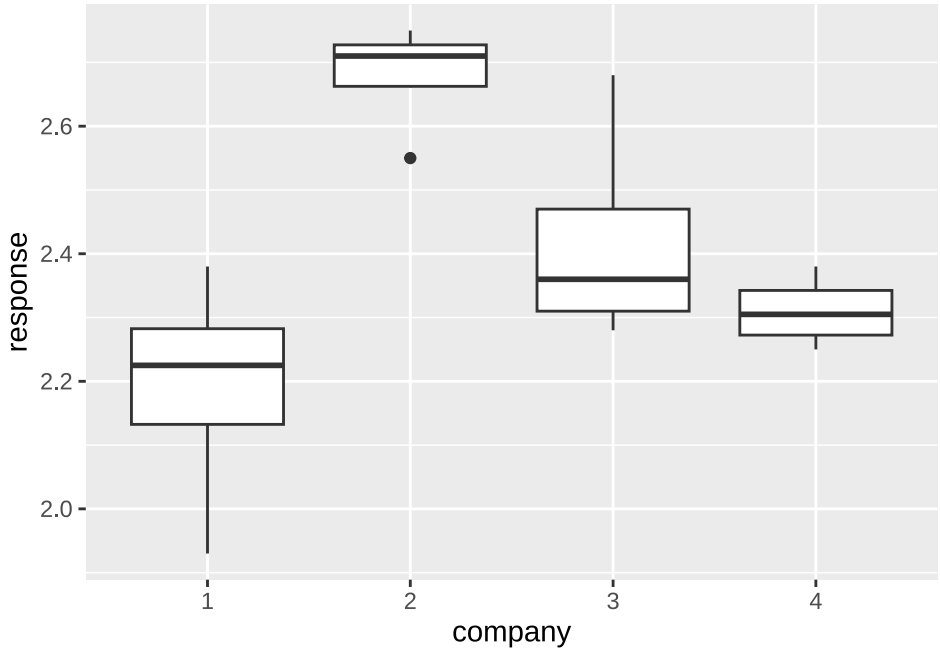
A.2.2. 기술 통계량에 의한 요약

```
df31s <- df31 %>% group_by(company)  %>%  summarise(mean=mean(response), median= median(response), sd=s
df31s
```

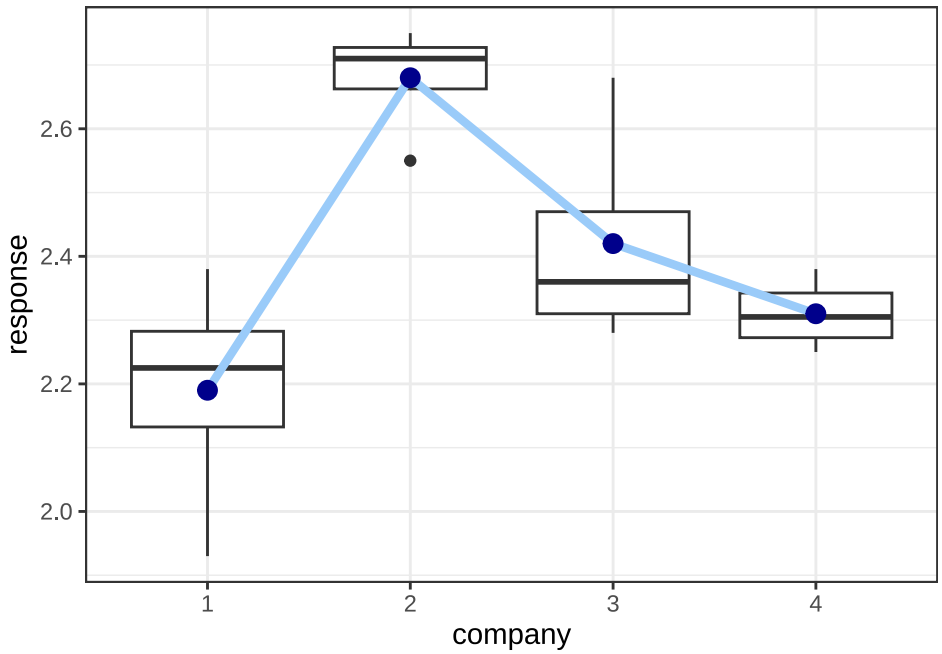
```
# A tibble: 4 x 6
  company  mean median     sd   min   max
  <fct>   <dbl>  <dbl>  <dbl> <dbl> <dbl>
1 1        2.19   2.22 0.189   1.93  2.38
2 2        2.68   2.71 0.0891  2.55  2.75
3 3        2.42   2.36 0.180   2.28  2.68
4 4        2.31   2.30 0.0572  2.25  2.38
```

A.2.3. 집단 자료에 대한 시각화

```
ggplot(df31, aes(company, response)) +  geom_boxplot()
```



```
ggplot(df31, aes(company, response)) +
  geom_boxplot() +
  geom_line(data=df31s, aes(x=company, y=mean, group=1), size=1.5, col="#9ACBF9") +
  geom_point(data=df31s, aes(x=company, y=mean), col="darkblue", size=3) +
  theme_bw()
```



B. 일원배치 모형과 최소제곱법

B.1. 최소제곱법과 제약조건

이제 일원배치법에 대한 통계적 모형에서 모수에 대한 추정을 생각해 보자.

$$y_{ij} = \mu + \alpha_i + e_{ij} \tag{B.1}$$

추정해야할 모수는 전체 평균 μ와 각 그룹의 처리 효과 α_i 그리고 분산 σ_E^2이다. 전체 평균과 그룹의 효과는 오차제곱합(Sum of Square Error; SSE)을 최소로 하는 모수를 추정하는 최소제곱법(Least Square method; LS)으로 구할 수 있다.

$$\min_{\mu,\alpha_1,\cdots\alpha_a} \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij} - \mu - \alpha_i)^2 = \min_{\mu,\alpha_1,\cdots\alpha_a} SSE \tag{B.2}$$

위의 오차제곱합이 모든 모수에 대하여 미분가능한 이차식으므로 최소제곱 추정량은 제곱합을 모수에 대하여 미분하고 0 으로 놓아 방정식을 풀어서 얻을 수 있다.

오차제곱합을 모수 μ와 $\alpha_1, \alpha_2, \ldots, \alpha_a$ 로 미분하여 0 으로 놓은 방정식은 다음과 같다.

$$\begin{aligned}
\frac{\partial}{\partial \mu} SSE &= -2\sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij} - \mu - \alpha_i) = 0 \\
\frac{\partial}{\partial \alpha_i} SSE &= -2\sum_{j=1}^{r}(y_{ij} - \mu - \alpha_i) = 0, \quad i = 1, 2, \ldots, a
\end{aligned}$$

위의 방정식을 정리하면 다음과 같은 $a+1$개의 방정식을 얻는다.

$$\begin{aligned}
\mu + \frac{\sum_{i=1}^{a}\alpha_i}{a} &= \bar{\bar{y}} \\
\mu + \alpha_1 &= \bar{y}_{1.} \\
\mu + \alpha_2 &= \bar{y}_{2.} \\
&\cdots\cdots \\
\mu + \alpha_a &= \bar{y}_{a.}
\end{aligned} \tag{B.3}$$

위의 방정식에서 첫 번째 방정식은 다른 a개의 방정식을 모두 합한 방정식과 같다. 따라서 모수는 $a+1$개이지만 실제 방정식의 개수는 a개이므로 유일한 해가 얻어지지 않는다. 따라서 유일한 해를 구하려면 하나의 제약조건이 필요하며 일반적으로 다음과 같은 두 개의 조건 중 하나를 사용한다.

B.1.1. set-to-zero condition

첫 번째 효과 α_1를 0으로 놓는 조건을 주는 것이다 ($\alpha_1 = 0$). set-to-zero 조건 하에서는 다음과 같은 추정량이 얻어진다.

$$\hat{\mu} = \bar{y}_{1.}, \quad \hat{\alpha}_1 = 0, \;\; \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{1.}, \;\; i = 2, \dots, a \tag{B.4}$$

B.1.2. sum-to-zero condition

처리들의 효과의 합은 0이라는 조건을 주는 것이다 ($\sum_{i=1}^{a} \alpha_i = 0$). sum-to-zero 조건에서는 계수의 추정치가 다음과 같이 주어진다.

$$\hat{\mu} = \bar{\bar{y}}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{\bar{y}}, \;\; i = 1, 2, \dots, a \tag{B.5}$$

여기서 유의할 점은 **개별 모수들의 추정량은 조건에 따라서 달라지지만 집단의 평균을 나타내는 모수 $\mu + \alpha_i$ 에 대한 추정량은 언제나 같다.**

$$\widehat{\mu + \alpha}_i = \hat{\mu} + \hat{\alpha}_i = \bar{y}_{i.}$$

만약에 자료를 아래와 같은 평균 모형으로 나타낼 경우에는 각 평균 μ_i 는 각 그룹의 표본 평균으로 추정된다.

$$y_{ij} = \mu_i + e_{ij}$$

평균 모형에서 각 그룹의 모평균에 대한 최소제곱 추정량은 $\hat{\mu}_i = \bar{y}_{i.}$ 이며 이는 주효과 모형에서의 추정량과 동일하다.

또한 모형에 관계없이 오차항의 분산 σ_E^2 에 대한 추정량은 다음과 같이 주어진다.

$$\hat{\sigma}_E^2 = \frac{\sum_i \sum_j (y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2}{a(r-1)}$$

B.2. 선형모형과 제약 조건

일원배치 모형 식 B.1 를 다음과 같은 벡터를 이용한 선형모형(linear regression model) 형태로 나타내고자 한다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{B.6}$$

위의 선형모형식의 요소 $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{e}$는 다음과 같은 벡터와 행렬로 표현된다.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1r} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2r} \\ \vdots \\ y_{a1} \\ y_{a2} \\ \vdots \\ y_{ar} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & . & . & 0 \\ 1 & 1 & 0 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & . & . & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \\ 1 & 0 & 0 & . & . & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1r} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2r} \\ \vdots \\ e_{a1} \\ e_{a2} \\ \vdots \\ e_{ar} \end{bmatrix} \tag{B.7}$$

이제 위에서 논의한 최소제곱법을 선형 모형 식 B.6 에 적용하면 다음과 같이 표현할 수 있다.

$$\min_{\mu,\alpha_1,\cdots\alpha_a} \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij}-\mu-\alpha_i)^2 = \min_{\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X\beta})^t(\boldsymbol{y}-\boldsymbol{X\beta}) \tag{B.8}$$

최소제곱법의 기준을 만족하는 계수 $\boldsymbol{\beta}$는 다음과 같은 정규방정식(normal equation)의 해(solution)이다.

$$\boldsymbol{X}^t\boldsymbol{X\beta} = \boldsymbol{X}^t\boldsymbol{y} \tag{B.9}$$

정규방정식 식 B.9 을 일워배치의 선형모형식 식 B.7 에 나타난 $\boldsymbol{y}$, $\boldsymbol{X}$로 이용하여 나타내면 다음과 같다.

$$\begin{bmatrix} ar & r & r & \cdot & \cdot & r \\ r & r & 0 & \cdot & \cdot & 0 \\ r & 0 & r & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r & 0 & 0 & \cdot & \cdot & r \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_a \end{bmatrix} = \begin{bmatrix} ar\bar{\bar{y}} \\ r\bar{y}_{1.} \\ r\bar{y}_{2.} \\ \cdot \\ \cdot \\ r\bar{y}_{a.} \end{bmatrix} \tag{B.10}$$

정규방정식 식 B.10 는 위에서 구한 최소제곱법에서 유도된 방정식 식 B.3 과 같다.

여기서 유의할 점은 선형모형식 식 B.7 의 계획행렬 $\boldsymbol{X}$ 가 완전 계수(full rank) 행렬이 아니다. 계획행렬 $\boldsymbol{X}$의 첫 번째 열은 다른 열을 합한 것과 같다. 또한 정규 방정식 식 B.10 에서 $\boldsymbol{X}^t\boldsymbol{X}$ 행렬도 완전계수 행렬이 아니다. 따라서 $\boldsymbol{X}^t\boldsymbol{X}$ 행렬의 역행렬은 존재하지 않는다.

이러한 이유로 모수에 대한 유일한 추정량이 존재하지 않기 때문에 앞에서 언급한 제약 조건을 고려해야 정규방정식을 풀 수 있다.

B.2.1. Set-to-zero 조건에서의 모형과 최소제곱 추정량

만약 Set-to-zero 조건을 가정한다면 모수에서 α_1을 제외하고 선형모형식 식 B.7 를 다음과 같이 다시 표현할 수 있다.
효과 α_1을 0 으로 놓는다는 것은 α_1을 추정할 필요가 없으므로 모수벡터 $\boldsymbol{\beta}$ 에서 α_1를 빼고 계획행렬에서도 대응하는 열을 제거하는 것이다.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1r} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2r} \\ \vdots \\ y_{a1} \\ y_{a2} \\ \vdots \\ y_{ar} \end{bmatrix} = \begin{bmatrix} 1 & 0 & . & . & 0 \\ 1 & 0 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & . & . & 0 \\ 1 & 1 & . & . & 0 \\ 1 & 1 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & . & . & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & . & . & 1 \\ 1 & 0 & . & . & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & . & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_a \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1r} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2r} \\ \vdots \\ e_{a1} \\ e_{a2} \\ \vdots \\ e_{ar} \end{bmatrix} \tag{B.11}$$

이제 수정된 모형식 식 B.11 에 최소제곱법을 적용하여 정규방정식을 구하면 다음과 같은 방정식을 얻는다.

$$\begin{bmatrix} ar & r & r & \cdot & \cdot & r \\ r & r & 0 & \cdot & \cdot & 0 \\ r & 0 & r & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r & 0 & 0 & \cdot & \cdot & r \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \cdot \\ \cdot \\ \alpha_a \end{bmatrix} = \begin{bmatrix} ar\bar{\bar{y}} \\ r\bar{y}_{2.} \\ r\bar{y}_{3.} \\ \cdot \\ \cdot \\ r\bar{y}_{a.} \end{bmatrix} \tag{B.12}$$

위의 정규방정 식 B.12 를 풀면 위에서 언급한 sum-to-zero 조건에서 구해지는 모수의 추정량 식 B.4 를 얻을 수 있다.

B.2.2. Sum-to-zero 조건에서의 모형과 최소제곱 추정량

이제 Sum-to-zero 조건에서 모수의 추정에 대해 알아보자. 조건 $\sum_{i=1}^{a} \alpha_i = 0$ 조건을 마지막 모수 α_a 에 대하여 표현하면 다음과 같다.

$$\alpha_a = -\alpha_1 - \alpha_2 - \cdots - \alpha_{a-1}$$

따라서 마지막 처리 α_a 에 대한 관측값에 대한 모형은 다음과 같아 쓸 수 있다.

$$y_{aj} = \mu + \alpha_a + e_{aj} = \mu + (-\alpha_1 - \alpha_2 - \cdots - \alpha_{a-1}) + e_{ij}$$

이러한 결과를 모형방정식에 반영한다. 즉, 모수벡터 $\boldsymbol{\beta}$ 에서 α_a 를 제거하고 계획행렬에 위의 마지막 처리에 대한 효과식을 반영하면 다음과 같은 선형모형식을 얻는다.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1r} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2r} \\ \vdots \\ y_{a1} \\ y_{a2} \\ \vdots \\ y_{ar} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & . & . & 0 \\ 1 & 1 & 0 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & . & . & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \\ 1 & 0 & 0 & . & . & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \\ 1 & 0 & 0 & . & . & 1 \\ 1 & -1 & -1 & . & . & -1 \\ 1 & -1 & -1 & . & . & -1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & . & . & -1 \\ 1 & -1 & -1 & . & . & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{a-1} \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1r} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2r} \\ \vdots \\ e_{a1} \\ e_{a2} \\ \vdots \\ e_{ar} \end{bmatrix} \tag{B.13}$$

이제 수정된 모형식 식 B.13 에 최소제곱법을 적용하여 정규방정식을 구하면 다음과 같은 방정식을 얻는다.

$$\begin{bmatrix} ar & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 2r & r & \cdot & \cdot & r \\ 0 & r & 2r & \cdot & \cdot & r \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & r & r & \cdot & \cdot & 2r \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_{a-1} \end{bmatrix} = \begin{bmatrix} ar\bar{\bar{y}} \\ r\bar{y}_{1.} - r\bar{y}_{a.} \\ r\bar{y}_{2.} - r\bar{y}_{a.} \\ \cdot \\ \cdot \\ r\bar{y}_{a-1,.} - r\bar{y}_{a.} \end{bmatrix} \tag{B.14}$$

위의 정규방정 식 B.14 를 풀면 위에서 언급한 sum-to-zero 조건에서 구해지는 모수의 추정량 식 B.5 를 얻을 수 있다.

B.3. 추정 가능한 함수

B.3.1. 일원배치법에 추정가능한 모수

앞 절에서 보았듯이 일원배치법을 선형 모형식으로 표현하는 경우 평균에 대한 모수는 모두 $a+1$ 개가 있다.

$$\mu, \alpha_1, \alpha_2, \cdots, \alpha_a$$

하지만 모형식에서 계획행렬 $\boldsymbol{X}$가 완전 계수 행렬이 아니기 때문에 1개의 제약 조건을 가정하고 모수를 추정하였다. 하지만 제약 조건이 달라지면 각 모수의 추정량이 달라지기 때문에 각 모수는 유일한 값으로 추정이 불가능하다.

이렇게 각 모수들은 제약 조건에 따라서 유일하게 추정이 불가능하지만 앞 절에서 보았듯이 $\mu + \alpha_i$ 에 대한 추정량은 제약조건에 관계없이 표본 평균 $\bar{y}_{i.}$ 으로 동일하게 추정되어 진다.

그러면 어떤 모수들은 유일하게 추정이 불가능하고 어떤 모수들이 유일하게 추정이 가능할까?

이제 제약조건이 달라도 유일하게 추정이 가능한 모수들의 형태를 살펴보자.

B.3.2. 추정가능한 모수의 함수

선형모형 $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ 에서 계획행렬 $\boldsymbol{X}$의 계수가 완전하지 않으면 모수 벡터 $\boldsymbol{\beta}$는 유일한 값으로 추정할 수 없다.

이제 임의의 벡터 $\boldsymbol{c}$가 있을 때 모수들의 선형결합 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$를 고려하자.

예를 들어 일원배치 모형에서는 다음과 같은 모수들의 선형결합을 고려하는 것이다.

$$\psi = \boldsymbol{c}^t\boldsymbol{\beta} = [c_0\ c_1\ c_2\ \cdots\ c_a] \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{bmatrix} = c_0\mu + c_1\alpha_1 + c_2\alpha_2 + \cdots + c_a\alpha_a$$

위에서 본 것처럼 하나의 모수 α_1에 대한 유일한 추정은 불가능하다.

$$\alpha_1 = (0)\mu + (1)\alpha_1 + (0)\alpha_2 + \cdots + (0)\alpha_a$$

하지만 모수의 조합 $\mu + \alpha_2$ 은 유일한 추정이 가능하다.

$$\mu + \alpha_1 = (1)\mu + (1)\alpha_1 + (0)\alpha_2 + \cdots + (0)\alpha_a$$

이제 문제는 선형조합 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$ 에서 계수들 $c_0, c_1, \ldots, c_a$가 어떤 값을 가지는 경우 유일한 추정이 가능한 지 알아내는 것이다.

이제 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$ 에 대한 유일한 추정량 $\hat{\psi}$ 이 있다고 가정하자. 선형 모형에서 추정량 $\hat{\psi}$의 형태는 관측값에 대한 선형함수가 되어야 한다. 따라서 추정량을 $\hat{\psi} = \boldsymbol{a}^t\boldsymbol{y}$ 로 나타낼 수 있다. 이제 추정량 $\hat{\psi}$의 기대값은 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$이어야 하므로 다음이 성립해야 한다.

$$E(\hat{\psi}|\boldsymbol{X}) = E(\boldsymbol{a}^t\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{a}^t E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{a}^t\boldsymbol{X\beta} = \boldsymbol{c}^t\boldsymbol{\beta}$$

위의 식에서 가장 마지막 두 항의 관계를 보면 다음이 성립해야 한다.

$$\boldsymbol{a}^t\boldsymbol{X} = \boldsymbol{c}^t \quad \text{equivalently} \quad \boldsymbol{c} = \boldsymbol{X}^t\boldsymbol{a} \tag{B.15}$$

즉 추정가능한 모수의 조합 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$에서 **계수 벡터 $\boldsymbol{c}$ 는 계획행렬에 있는 행들의 선형 조합**으로 표시되어야 한다는 것이다. 이렇게 유일하게 추정이 가능한 모수의 조합을 **추정가능한 함수(estimable function)**이라고 한다.

B.3.3. 예제

2개의 수준이 있고 반복이 2번 있는 일원배치 ($a=2, r=2$) 에 대한 선형모형 **?@eq-lm2을** 생각해보자. 이 경우 계획행렬 $\boldsymbol{X}$ 과 모수벡터 $\boldsymbol{\beta}$ 는 다음과 같다.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

이제 유일하게 추정 가능한 모수 조합 ψ 은 어떤 형태일까?

$$\psi = \boldsymbol{c}^t\boldsymbol{\beta} = c_0\mu + c_1\alpha_1 + c_2\alpha_2$$

위의 식 식 B.15 에서 추정가능한 모수의 조합에 대한 계수 벡터 $\boldsymbol{c}$ 는 다음과 같은 조건을 만족해야 한다.

$$\boldsymbol{c} = \boldsymbol{X}^t\boldsymbol{a}$$

이제 임의의 벡터 $\boldsymbol{a}$ 에 대하여 $\boldsymbol{c} = \boldsymbol{X}^t\boldsymbol{a}$의 형태를 보자.

$$\begin{aligned} \boldsymbol{c} &= \boldsymbol{X}^t\boldsymbol{a} \\ &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \\ &= a_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + a_3 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + a_4 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \\ &= (a_1 + a_2) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (a_3 + a_4) \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \\ &= b_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + b_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \end{aligned} \tag{B.16}$$

이제 $\boldsymbol{X}^t\boldsymbol{a}$ **는 계획행렬** $\boldsymbol{X}$**에 있는 유일한 행들의 선형조합**임을 알 수 있다.

위의 식 식 B.16 에서 유의할 점은 벡터 $\boldsymbol{a} = [a_1\ a_2\ a_3\ a_4]^t$는 임의로 주어진 벡터이다.

식 식 B.16 에서 $a_1 = 1, a_2 = 1$ 인 경우는 $a_1 = 2, a_2 = 0$ 인 경우와 동일하다.

따라서 유일하게 추정 가능한 모수의 선형조합 $\psi = \boldsymbol{c}^t\boldsymbol{\beta}$ 에 대한 계수 벡터 $\boldsymbol{c} = [c_0\ c_1\ c_2]^t$ 는 계획행렬 $\boldsymbol{X}$의 유일한 행들의 선형 조합으로 구성되어야 한다.

$$\boldsymbol{c} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = b_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + b_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \tag{B.17}$$

- 처리의 효과를 나타내는 모수 α_i는 추정이 불가능하다.

첫 번째 처리에 대한 효과 모수 α_1 를 선형조합으로 나타내면

$$\alpha_1 = c_0\mu + c_1\alpha_1 + c_2\alpha_2 = (0)\mu + (1)\alpha_1 + (0)\alpha_2$$

따라서 조건 식 B.17 에서 $\boldsymbol{c}^t = [0\ 1\ 0]$을 만들수 있는 계수 b_1과 b_2를 찾아야 하는데 이는 불가능하다. 따라서 모수 α_1 은 추정 불가능하다.

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = b_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + b_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

- 처리의 평균을 나타내는 모수의 조합 $\mu + \alpha_i$는 추정이 가능하다.

모수 조합 $\mu + \alpha_1$ 를 선형조합으로 나타내면

$$\mu + \alpha_1 = c_0\mu + c_1\alpha_1 + c_2\alpha_2 = (1)\mu + (1)\alpha_1 + (0)\alpha_2$$

따라서 조건 식 B.17 에서 $\boldsymbol{c}^t = [1\ 1\ 0]$을 만들수 있는 계수는 $b_1 = 1$과 $b_2 = 0$ 이므로 추정이 가능하다.

$$\boldsymbol{c} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = (1) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (0) \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

- 처리 효과의 차이를 나타내는 모수의 조합 $\alpha_1 - \alpha_2$는 추정이 가능하다.

$$\alpha_1 - \alpha_2 = c_0\alpha_0 + c_1\alpha_1 + c_2\alpha_2 = (0)\boldsymbol{u} + (1)\alpha_1 + (-1)\alpha_2$$

따라서 조건 식 B.17 에서 $\boldsymbol{c}^t = [0\ 1\ -1]$을 만들수 있는 계수는 $b_1 = 1$과 $b_2 = -1$ 이므로 추정이 가능하다.

$$\boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = (1) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + (-1) \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

B.4. R 실습

B.4.1. 예제 3.1

4개의 서로 다른 원단업체에서 직물을 공급받고 있다. 공급한 직물의 긁힘에 대한 저항력을 알아보기 위하여 각 업체마다 4개의 제품을 랜덤하게 선택하여 ($a = 4$, $r = 4$) 일원배치법에 의하여 마모도 검사을 실시하였다.

B.4.2. 자료의 생성

```
company<- as.factor(rep(c(1:4), each=4))
response<- c(1.93, 2.38, 2.20, 2.25,
             2.55, 2.72, 2.75, 2.70,
             2.40, 2.68, 2.32, 2.28,
             2.33, 2.38, 2.28, 2.25)
df31<- data.frame(company=company, response= response)
df31
```

```
   company response
1        1     1.93
2        1     2.38
3        1     2.20
4        1     2.25
5        2     2.55
6        2     2.72
7        2     2.75
8        2     2.70
9        3     2.40
10       3     2.68
11       3     2.32
12       3     2.28
13       4     2.33
14       4     2.38
15       4     2.28
16       4     2.25
```

각 수준에 대한 표보 평균을 구해보자.

```
df31s <- df31 %>% group_by(company)  %>%  summarise(mean=mean(response), median= median(response), sd=s
df31s
```

```
# A tibble: 4 x 6
  company  mean median     sd   min   max
  <fct>   <dbl>  <dbl>  <dbl> <dbl> <dbl>
1 1        2.19   2.22 0.189   1.93  2.38
2 2        2.68   2.71 0.0891  2.55  2.75
3 3        2.42   2.36 0.180   2.28  2.68
4 4        2.31   2.30 0.0572  2.25  2.38
```

B.4.3. 선형모형의 적합(set-to-zero)

이제 자료를 다음과 같은 선형 모형으로 적합해 보자. 선형 모형의 적합은 `lm()` 함수를 사용한다.

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

여기서 선형식의 모수와 R의 변수는 다음과 같은 관계를 가진다,

선형식의 모수	R의 변수
μ	(Intercept)
α_1	company1
α_2	company2
α_3	company3
α_4	company4

```
fit1 <- lm(response~company,data=df31)
summary(fit1)
```

```
Call:
lm.default(formula = response ~ company, data = df31)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2600 -0.0700  0.0150  0.0625  0.2600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.19000    0.07050  31.062 7.79e-13 ***
company2     0.49000    0.09971   4.914 0.000357 ***
company3     0.23000    0.09971   2.307 0.039710 *
company4     0.12000    0.09971   1.204 0.251982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.141 on 12 degrees of freedom
Multiple R-squared:  0.6871,    Adjusted R-squared:  0.6089
F-statistic: 8.785 on 3 and 12 DF,  p-value: 0.002353
```

위에서 적합한 결과를 보면 평균 μ와 4개의 처리 α_1, α_2, α_3, α_4 가 모형에 있지만 모수의 추정량은 평균(intercept)과 3개의 모수(company2, company3, company4)만 추정량이 주어진다.

R 에서 옵션을 지정하지 않고 함수 lm()으로 선형모형을 적합하는 경우 set-to-zero 조건을 적용하며 자료에 나타난 처리의 수준들 중 순위가 가장 낮은 수준의 효과를 0으로 지정한다 (company1=0). set-to-zero 조건을 강제로 지정하려면 다음과 같은 명령문을 먼저 실행한다.

```
options(contrasts=c("contr.treatment", "contr.poly"))
```

위의 결과를 보면 (Intercept)에 대한 추정량이 첫 번째 처리 company1의 평균과 같은 것을 알 수 있다.

set-to-zero 조건에서의 계획행렬은 다음과 같이 볼 수 있다.

```
model.matrix(fit1)
```

```
   (Intercept) company2 company3 company4
1            1        0        0        0
2            1        0        0        0
3            1        0        0        0
4            1        0        0        0
5            1        1        0        0
6            1        1        0        0
7            1        1        0        0
8            1        1        0        0
9            1        0        1        0
10           1        0        1        0
11           1        0        1        0
12           1        0        1        0
13           1        0        0        1
14           1        0        0        1
15           1        0        0        1
16           1        0        0        1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$company
[1] "contr.treatment"
```

이제 각 처리 평균에 대한 추정값 $\widehat{\mu + \alpha_i}$을 구해보자.

```
emmeans(fit1, "company")
```

```
 company emmean     SE df lower.CL upper.CL
 1          2.19 0.0705 12     2.04     2.34
 2          2.68 0.0705 12     2.53     2.83
 3          2.42 0.0705 12     2.27     2.57
 4          2.31 0.0705 12     2.16     2.46

Confidence level used: 0.95
```

이 경우 처리 평균에 대한 추정값은 산술 평균과 동일하게 나온다.

B.4.4. 선형모형의 적합 (sum-to-zero)

이제 일원배치 모형에서 sum-to-zero 조건을 적용하여 모수를 추정해 보자. sum-to-zero 조건을 적용하려면 다음과 같은 명령어를 실행해야 한다.

```
options(contrasts=c("contr.sum", "contr.poly"))
```

이제 다시 선형모형을 적합하고 추정결과를 보자.

```
fit2 <- lm(response~company,data=df31)
summary(fit2)
```

```
Call:
lm.default(formula = response ~ company, data = df31)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2600 -0.0700  0.0150  0.0625  0.2600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.40000    0.03525  68.081  < 2e-16 ***
company1    -0.21000    0.06106  -3.439 0.004901 **
company2     0.28000    0.06106   4.586 0.000626 ***
company3     0.02000    0.06106   0.328 0.748892
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.141 on 12 degrees of freedom
Multiple R-squared:  0.6871,    Adjusted R-squared:  0.6089
F-statistic: 8.785 on 3 and 12 DF,  p-value: 0.002353
```

이제 sum-to-zero 조건에 따라서 위의 set-to-zero 결과와 모수의 추정값이 다르게 나타나는 것을 알 수 있다. 마지막 모수 company4(α_4)는 sum-to-zero 조건을 이용하여 다음과 같은 관계를 이용하여 구할 수 있다.

$$\alpha_4 = -(\alpha_1 + \alpha_2 + \alpha_3)$$

sum-to-zero 조건에서의 계획행렬은 다음과 같이 볼 수 있다.

```
model.matrix(fit2)
```

```
   (Intercept) company1 company2 company3
1            1        1        0        0
2            1        1        0        0
3            1        1        0        0
```

```
4            1        1        0        0
5            1        0        1        0
6            1        0        1        0
7            1        0        1        0
8            1        0        1        0
9            1        0        0        1
10           1        0        0        1
11           1        0        0        1
12           1        0        0        1
13           1       -1       -1       -1
14           1       -1       -1       -1
15           1       -1       -1       -1
16           1       -1       -1       -1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$company
[1] "contr.sum"
```

이제 각 처리 평균에 대한 추정값 $\widehat{\mu + \alpha_i}$을 구해보면 set-to-zero 조건에서의 추정값과 동일함을 알 수 있다.

```
emmeans(fit2, "company")
```

```
 company emmean     SE df lower.CL upper.CL
 1          2.19 0.0705 12     2.04     2.34
 2          2.68 0.0705 12     2.53     2.83
 3          2.42 0.0705 12     2.27     2.57
 4          2.31 0.0705 12     2.16     2.46

Confidence level used: 0.95
```

B.4.5. 분산분석

분산분석의 결과는 어떠한 제약 조건에서도 동일하다.

```
res1 <- anova(fit1)
res1
```

```
Analysis of Variance Table

Response: response
          Df Sum Sq  Mean Sq F value   Pr(>F)
company    3 0.5240 0.174667  8.7846 0.002353 **
Residuals 12 0.2386 0.019883
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res2<- anova(fit2)
res2
```

```
Analysis of Variance Table

Response: response
          Df Sum Sq  Mean Sq F value   Pr(>F)
company    3 0.5240 0.174667  8.7846 0.002353 **
Residuals 12 0.2386 0.019883
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

C. 혼합 모형

교과서에서 **변량모형**으로 불리는 모형으로 흔히 **임의효과 모형(random effect model)** 또는 **혼합모형(mixed effects model)** 이라고 부른다.

C.1. 고정효과

앞 장에서 하나의 요인있는 일원배치 모형에 대한 추론에 대하여 알아보았다.

$$x_{ij} = \mu + \alpha_i + e_{ij} \tag{C.1}$$

여기서 오차항 e_{ij}는 모두 독립이며 $N(0, \sigma_E^2)$를 따른다.

일원배치 모형 식 C.1 에서 전체 평균 μ 와 처리수준의 효과를 나타내는 $\alpha_1, \alpha_2, \ldots, \alpha_a$는 모두 고정된 값을 가지는 모수(parameter)이다. 식 C.1 의 오른쪽 항들 중에서 확률변수는 오차항 e_{ij}이 유일하다.

처리수준의 효과 α_i들이 모수이라는 것은 의미는 만약 새로운 실험에서 실험단위(experiment unit)에 동일한 처리를 적용하면 평균 처리 효과는 α_i로 일정하다는 의미이다.

예를 들어 예제 3.1에서 수행한 실험을 다른 회사에서 동일한 납품업체의 원단(동일한 실험 단위와 처리)을 가지고 새로운 실험을 하면 평균적인 효과는 예제 3.1과 동일하다는 가정을 할 수 있다. 물론 흑과는 동일하지만 설명할 수 없는오차때문에 관측값은 다를 수 있다.

또한 예제 4.1 에 대한 실험에서도 만약 동일한 돼지 품종과 사료를 사용하여 새로운 실험을 수행할 때 처리 효과는 원래 실험과 같다고 가정할 수 있다. 즉, 처리라는 것이 기술적인 의미를 지니고 있어 반복하여 재현할 수 있는 효과이다. 이러한 고정된 모수로서의 효과를 **고정 효과(fixed effect)**라고 부른다.

더 나아가 고정효과를 가지는 모형에서는 고정효과를 추정하고 처리 수준간의 차이가 있는지 추론하는 것이 실험의 주요 목적이다.

C.2. 임의효과

이제 고정효과와는 다른 의미를 가지는 몇 가지 실험들을 생각해 보자.

보기 C.1 (화학약품 회사:교과서예제 3.3). 화학약품 회사에서는 매년 원자재의 수백 개의 배치(batch)를 정제하여 순도가 높은 화학약품을 만든다. 품질 관리를 위하여 수백 개의 배치들 중에서 5개를 랜덤하게 선택하고 배치당 3개의 시료를 채취한 후에 순도를 측정하였다.

배치마다 순도가 크게 다르면 품질을 일정하게 유지할 수 없눈 문제가 생긴다. 따라서 실험의 목적은 품질 관리이며 배치 간의 변동과 배치 내의 변동을 알아보는 것이다.

보기 C.2 (학교간의 성적 비교). 학교 간에 성적의 차이를 알아보기 위하여 서울에 있는 603개의 학교에서 20개의 학교를을 임의로 추출하고 추출된 학교에 속한 6학년 학생들 10명을 임의로 추출하여 과학시험을 보게 하여 점수를 얻었다.

이러한 자료에서 학생들의 성적은 가장 점수가 낮은 학생부터 매우 우수한 성적을 낸 학생까지 점수의 변동(variation)이 존재한다. 변동의 요인은 무었일까? 학생의 개인의 차이(예:학생의 지능, 노력 정도, 학습 환경)도 변동의 요인이지만 또한 학교의 차이(예: 교사, 거주 환경)도 변동의 요인이 될 수 있다.

보기 C.3 (Test-ReTest). 새로 개발된 CT 로 만든 영상에 근거하여 의사들이 암의 단계를 점수로 파악하는 방법이 제안되었다. 제안된 방법의 유료성과 안정성을 알아보기 위하여 실험을 진행하였다. 일단 5명의 암환자들에서 CT 영상을 촬영하였다. 다음으로 15명의 의사를 임의로 추출하고 5명의 CT 영상을 본 후 암의 진행 단계를 판단할 수 있는 점수를 매기도록 하였다.

실험의 목적은 CT 영상에 근거한 진단이 의사들간에 잘 일치하는지를 알아보는 실험이다. 이 실험에서는 의사와 환자라는 두 가지 요인이 존재한다.

위의 예제에서 **배치, 의사, 학교**는 고정 효과를 가정한 실험에서 고려하는 요인과는 성격이 틀리다. 5개의 배치들은 수백 개의 배치들에서 임의로 추출 되었으며 5명의 의사들은 다수의 의사들 중 임의로 추출되었다. 603개 초등 학교의 모집단에서 20개의 학교가 임의로 추출되었다.

배치, 의사 또는 학교 간의 차이는 잘 설계된 실험의 처리에 대한 고정 효과와는 다르다. 동일한 배치, 학교 또는 의사으로 부터 나온 관측값들은 동일한 처릴 받은 값들이라기 보다는 동일한 **집단(group, cluster)**에서 나온 관측값으로 볼 수 있다. 위의 예제들에서는 식 C.2 에서 효과 α_i 의 변동은 모집단을 구성하는 집단들의 변동이라고 할 수 있다.

위에서 언급한 3개 예제는 실험의 목적이 선택된 수준들의 효과의 기술적인 비교가 아니라 모집단이 가지고 있는 여러 가지 변동(variance)에 대하여 추론하는 것이다.

> **노트**
>
> 같은 학교에 다니는 학생들은 주거 환경, 교사 등 공통적인 요인에 의하여 영향을 받는다고 가정할 수 있다. 따라서 같은 학교에 다는 학생들의 성적이 독립이 아닐 수도 있다.
> 같은 의사가 5명의 환자에 대한 평가하여 진단을 한 경우 5개의 진단결과는 다른 환자에 대항 결과임에도 불구하고 서로 독립이라고 가정하지 않을 수 있다. 의사의 역량, 경험, 성향에 따라서 환자에 대한 진단에 공통적인 영향을 미칠 수 있기 때문이다.

고정효과처럼 기술적인 처리효과가 아니라 모집단의 구성 단위들의 변동을 기술하는 효과를 **임의효과(random effect, 변량)** 라고 한다. 임의효과를 가진 일원배치 모형을 **변량모형**(random models) 또는 임의효과 모형(random effect models)이라고 부르며 다음과 같이 나타낼 수 있다.

$$x_{ij} = \mu + \alpha_i + e_{ij} \quad \text{where } \alpha_i \sim N(0, \sigma_A^2), \;\; e_{ij} \sim N(0, \sigma_E^2) \tag{C.2}$$

위의 식에서 $\alpha_1, \alpha_2, \ldots, \alpha_a$를 임의 효과라고 부르며 서로 독립인 확률 변수로서 분포는 $N(0, \sigma_A^2)$을 따른다. 또한 임의 효과 α_i와 오차항 e_{ij}은 서로 독립이다.

임의효과가 가지는 분산을 σ_A^2을 분산성분(variance component)라고 하며 집단 간의 변동을 의미한다. σ_A^2이 크면 모집단을 구성하고 있는 단위들의 변동이 크다고 할 수 있다. 반면 σ_A^2이 작으면 단위들간의 변동이 작아진다.

그럼 효과를 어느 경우에 고정효과로 가정하는지? 또 임의효과로 가정하는 경우는 언제인지? 이러한 질문에 대하여 간단하고 명료한 대답은 없다. 많은 학자들이 이 문제에 대하여 다양한 설명을 내놓았는데 정답은 없다.

심지어 다음과 같이 말한 학자도 있어요

> **노트**
>
> Before proceeding further with random field linear models, we need to remind the reader of the adage that **one modeler's random effect is another modeler's fixed effect.**
> Schabenberger 와/과 Pierce (2001) (627 page)

모형은 실제 현상이 어떻게 작동되는지 인간이 가진 제한적인 지식으로 간단한 수식과 분포 가정을 사용하여 기술하는 것이기 때문에 가정한 모형이 옳다 그르다를 판별하기 어렵다. G.P. Box 가 말했듯이 모형을 평가하는 가장 중요한 요소는 **모형의 유용성**일 것이다. 즉, 유용하지 않는 모형은 사람들이 금방 외면해 버릴 것이고 유용한 모형은 실제 자료를 예측하는데 도움을 주니 많은 사람들이 이용할 것이다.

아직도 같은 자료에 대하여 고정효과와 임의효과 모형이 동시에 사용되고 있으니 두 모형 모두 유용하다고 할 수 있다. 하지만 두 효과에 대한 어느 정도 차이점은 알아야 한다. 지금까지 경험으로 고정효과와 임의효과의 대략적 의미와 차이점은 다음과 같습니다.

- 고정효과
  - 기술적인 효과(technical effect)
  - 실험자가 기술적으로 반복하여 적용할 수 있는 효과
  - 평균 효과의 비교가 주 목적인 경우
  - 예를 들어 온도, 사료, 비료, 촉매 등등
- 임의효과
  - 효과가 있는 것 같은데 기술적으로 명확한 설명이 어려운 효과 (Unobservable heterogeneity)
  - 숨겨진 변수 (latent variable)
  - 모집단에서 추출된 집단(group, cluster, repeated menasure)에 속하여 나타나는 효과 - 급내상관계수
  - 효과들의 변동에 관심있는 경우
  - 예를 들어 학교, 병원, 재배단위(plot) 등등

C.3. 변량모형의 성질

C.3.1. 총변동의 분해

일원배치 변량 모형 식 C.2 을 따르는는 반응변수 x_{ij}의 평균과 분산은 다음과 같다.

$$
\begin{aligned}
E(x_{ij}) &= E(\mu + \alpha_i + e_{ij}) \\
&= E(\mu) + E(\alpha_i) + E(e_{ij}) \\
&= \mu + 0 + 0 \\
&= \mu
\end{aligned}
$$

$$
\begin{aligned}
V(x_{ij}) &= Var(\mu + \alpha_i + e_{ij}) \\
&= V(\alpha_i) + V(e_{ij}) \\
&= \sigma_A^2 + \sigma_E^2
\end{aligned}
\tag{C.3}
$$

식 C.3 에서 나타난 분해는 다음과 같이 의미로 표현할 수 있다.

$$
\underbrace{V(x_{ij})}_{\text{total variation}} = \underbrace{\sigma_A^2}_{\text{variation between groups}} + \underbrace{\sigma_E^2}_{\text{variation within group}}
$$

C.3.2. 관측값의 종속성

식 C.2 로 표현된 변량모형의 가장 큰 특징 중에 하나는 같은 집단에 속하는 관측치들은 서로 독립이 아니며 양의 상관관계가 있는 것이다. 예를 들어 위의 학교간의 성적 비교 예제에서 두 학생 x_{ij}와 x_{ik}이 같은 학교 i에 속한다면

$$
\begin{aligned}
Cov(x_{ij}, x_{ik}) &= Cov(\mu + \alpha_i + e_{ij}, \mu + \alpha_i + e_{ik}) \\
&= Cov(\alpha_i, \alpha_i) + Cov(\alpha_i, e_{ik}) + Cov(e_{ij}, \alpha_i) + Cov(e_{ij}, e_{ik}) \\
&= Cov(\alpha_i, \alpha_i) + 0 + 0 + 0 \\
&= V(\alpha_i, \alpha_i) \\
&= \sigma_A^2
\end{aligned}
$$

따라서

$$
\begin{aligned}
corr(x_{ij}, x_{ik}) &= \frac{Cov(x_{ij}, x_{ik})}{\sqrt{V(x_{ij})\, V(x_{ik})}} \\
&= \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} \\
&= \rho
\end{aligned}
$$

위의 상관계수(교과서애서 기여율)는 보통 급내 **상관계수(Intra Class Correlation, ICC)**라고 부른다. 그룹 변동의 크기를 나타내는 분산성분 σ_A^2가 그룹 내 변동을 나타내는 오차항의 분산 σ_E^2보다 상대적으로 클수록 급내 상관계수가 1에 가까와진다.

보통 σ_A^2을 집단간 변동(between-group variance)라 하고 σ_E^2를 집단내 변동(within-group variance)라고 한다. 따라서 σ_A^2와 σ_E^2의 상대적인 크기의 차이에 따라 그룹내 관측값의 상관관계가 달라진다.

C.3.3. 제곱합의 기대값

일원배치 변량 모형 식 C.2 은 고정효과 모형 식 C.1 과 동일한 분산분석(ANOVA) 표를 사용한다. 분산분석 표의 제곱합을 이용하여 σ_A^2와 σ_E^2에 대한 추정량을 얻을 수 있다.

첫 째로 분산분석 표에서 SS_E의 기대값을 구해보자.

먼저 다음과 같은 분해를 고려하자.

$$\begin{aligned} x_{ij} - \bar{x}_{i.} &= (\mu + \alpha_i + e_{ij}) - \frac{\sum_{j=1}^{r}(\mu + \alpha_i + e_{ij})}{r} \\ &= (\mu + \alpha_i + e_{ij}) - \left(\mu + \alpha_i + \frac{\sum_{j=1}^{r} e_{ij}}{r}\right) \\ &= (\mu + \alpha_i + e_{ij}) - (\mu + \alpha_i + \bar{e}_{i.}) \\ &= e_{ij} - \bar{e}_i \end{aligned}$$

이므로 오차제곱합 SS_E의 기대값은 다음과 같이 구해진다.

$$\begin{aligned} E\left[\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{x}_{i.})^2\right] &= E\left[\sum_{i=1}^{a}\sum_{j=1}^{r}(e_{ij} - \bar{e}_{i.}))^2\right] \\ &= (r-1)\sum_{i=1}^{a} E\left[\frac{\sum_{j=1}^{r}(e_{ij} - \bar{e}_{i.}))^2}{r-1}\right] \\ &= (r-1)\sum_{i=1}^{a}\sigma_E^2 \\ &= a(r-1)\sigma_E^2 \end{aligned}$$

또한 SS_A의 기대값을 구하기 위하여

$$\begin{aligned} \bar{x}_{i.} - \bar{\bar{x}} &= (\mu + \alpha_i + \bar{e}_{i.}) - (\mu + \bar{\alpha} + \bar{\bar{e}}) \\ &= (\alpha_i - \bar{\alpha}) + (\bar{e}_{i.} - \bar{\bar{e}}) \end{aligned}$$

이므로 SS_A의 기대값은 다음과 같이 구해진다.

$$\begin{aligned} E\left[\sum_{i=1}^{a}\sum_{j=1}^{r}(\bar{x}_{i.} - \bar{\bar{x}})^2\right] &= E\left[\sum_{i=1}^{a}\sum_{j=1}^{r}\{(\alpha_i - \bar{\alpha}) + (\bar{e}_{i.} - \bar{\bar{e}})\}^2\right] \\ &= \sum_{i=1}^{a}\sum_{j=1}^{r} E[(\alpha_i - \bar{\alpha})^2] + \sum_{i=1}^{a}\sum_{j=1}^{r} E[(\bar{e}_{i.} - \bar{\bar{e}})^2] \\ &= r(a-1)E\left[\frac{\sum_{i=1}^{a}(\alpha_i - \bar{\alpha})^2}{a-1}\right] + r(a-1)E\left[\frac{\sum_{i=1}^{a}(\bar{e}_{i.} - \bar{\bar{e}})^2}{a-1}\right] \\ &= r(a-1)\sigma_A^2 + r(a-1)\frac{\sigma_E^2}{r} \\ &= (a-1)(r\sigma_A^2 + \sigma_E^2) \end{aligned}$$

위의 계산에서 이용한 사실은 α_i는 서로 독립으로 $N(0, \sigma_A^2)$를 따르고 $\bar{e}_{i.}$는 서로 독립으로 $N(0, \sigma_E^2/r)$를 따른다는 것이다.

위의 제곱합의 기대값을 정리해보면 다음과 같은 두 방정식을 얻는다.

$$E(SS_A) = (a-1)(r\sigma_A^2 + \sigma_E^2), \quad E(SS_E) = a(r-1)\sigma_E^2$$

위의 모수 방정식에 적률추정법(methods of moment)을 적용하면 다음과 같은 방정식을 얻고

$$SS_A = (a-1)(r\hat{\sigma}_A^2 + \hat{\sigma}_E^2), \quad SS_E = a(r-1)\hat{\sigma}_E^2$$

위의 방정식을 풀면 σ_A^2와 σ_E^2의 불편 추정량을 구할 수 있다. 여기서 유의할 사항은 σ_A^2에 대한 추정량은 0보다 작은 값이 나올 수 있으므로 이럴 경우 0으로 지정한다.

$$\begin{aligned} s_E^2 = \hat{\sigma}_E^2 &= \frac{SS_E}{a(r-1)} = MS_E \\ s_A^2 = \hat{\sigma}_A^2 &= \max\left[0, \frac{SS_A/(a-1) - \hat{\sigma}_E^2}{r}\right] = \max\left[0, \frac{MS_A - MS_E}{r}\right] \end{aligned}$$

C.3.4. 가설 검정

고정효과 모형에서 요인 A의 수준간에 차이가 있는 지를 검정하는 경우 귀무가설은 $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ 이었다. 이제 변량 모형에서는 집단 간의 변동이 없는지 검정하는 것이므로 다음과 같은 가설을 고려한다.

$$H_0 : \sigma_A^2 = 0 \quad \text{vs.} \quad H_1 : \sigma_A^2 > 0 \tag{C.4}$$

분산성분 σ_A^2가 0 이라는 의미는 모든 α_i 가 0이고 이는 집단 간의 차이가 없는 상황을 의미한다. 위의 **가설을 검정하는 방법은 고정효과 모형과 동일하다.** 즉 다음과 같은 조건이 만족되면 귀무가설을 기각한다.

$$\text{reject } H_0 \text{ if } F_0 = \frac{MS_A}{MS_E} > F[1-\alpha, a-1, a(r-1)]$$

C.4. 예제 3.3

교과서 59 페이지에 있는 예제를 R 프호그램을 사용하여 풀어보자.

화학약품 회사에서는 매년 원자재의 수백 개의 배치(batch)를 정제하여 순도가 높은 화학약품을 만든다. 품질 관리를 위하여 수백 개의 배치들 중에서 5개를 랜덤하게 선택하고 배치당 3개의 시료를 채취한 후에 순도를 측정하였다.

배치마다 순도가 크게 다르면 품질을 일정하게 유지할 수 없눈 문제가 생긴다. 따라서 실험의 목적은 품질 관리이며 배치 간의 변동과 배치 내의 변동을 알아보는 것이다.

C.4.1. 자료

다음과 같이 자료를 만들자

```
response <- c( 74, 76, 75,
               68, 71, 72,
               75, 77, 77,
               72, 74, 73,
               79, 81, 79)
batch <- factor(rep(1:5, each=3))
df <- data.frame(batch, response)
df
```

```
   batch response
1      1       74
2      1       76
3      1       75
4      2       68
5      2       71
6      2       72
7      3       75
8      3       77
9      3       77
10     4       72
11     4       74
12     4       73
13     5       79
14     5       81
15     5       79
```

C.4.2. 추정과 가설검정

변량모형을 적합시키기 위해서는 `lme4` 패키지 가 필요하다. 일원배치 변량모형을 적합시키는 함수는 `lmer`이며 다음과 같이 사용한다. 아래 모형식에서 `1`은 평균 μ를 나타내고 `(1|batch)`는 배치에 대한 임의 효과 α_i를 나타낸다.

```
res <- lmer(response ~ 1 + (1|batch), data=df )
summary(res)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: response ~ 1 + (1 | batch)
   Data: df

REML criterion at convergence: 62.8

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.90384 -0.53153  0.00484  0.61386  1.16817

Random effects:
 Groups   Name        Variance Std.Dev.
 batch    (Intercept) 11.71    3.422
 Residual              1.80    1.342
Number of obs: 15, groups:  batch, 5

Fixed effects:
            Estimate Std. Error     df t value Pr(>|t|)
(Intercept)   74.867      1.569  4.000   47.71 1.15e-06 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 결과에서 다음과 같은 추정값을 얻는다.

- $\hat{\mu} = 74.867$
- $\hat{\sigma}_A^2 = 11.71$
- $\hat{\sigma}_E^2 = 1.80$

따라서 급내 상관계수(기여율)의 추정값은 다음과 같다.

$$\hat{\rho} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_E^2} = \frac{11.7}{11.7 + 1.8} = 0.867$$

위의 $\hat{\rho} = 0.867$을 기여율로 해석하면 총변동 중에서 배치 간의 변동이 차지하는 비율이 86.7% 이라는 것이다.

또한 $H_0 : \sigma_A^2 = 0$ 에 대한 검정은 다음과 같이 aov함수를 사용하여 수행할 수 있다.

```
summary(aov(response ~ batch, data=df ))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
batch        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 ' ' 1
```

p-값이 유의수준 5% 보다 매우 작으므로 H_0를 기각한다. 배치 간 변동이 유의하다고 할 수 있다. 따라서 품질이 배치 간에 따라서 크게 다르다.

D. 교락

교락(confounding)은 실험 또는 표본 추출의 방법에서 서로 다른 두 효과가 섞여서 자료를 통하여 구별할 수 없는 경우를 말한다. 실험계획에서 교락은 대부분 처리 효과와 오차/임의효과를 구별할 수 없는 경우에 발생한다.

여러분이 반복이 없는 이원배치법에서는 상호작용과 오차가 교락되어 상호작용에 대한 추론을 할 수 없다고 배웠다.

D.0.1. 일원배치

이러한 교락의 개념을 이해하기 위하여 가장 간단한 실험 계획인 일원배치를 생각하고 반복이 있는 경우와 없는 경우를 생각해 보자.

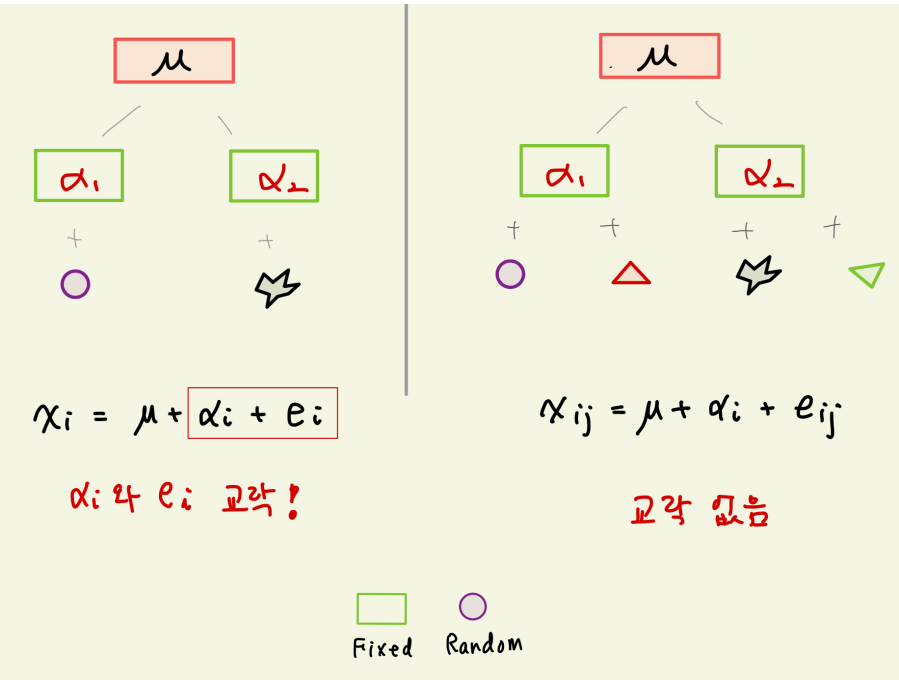


그림 D.1.: 일원배치: 반복이 없는 경우와 있는 경우

- 반복이 없는 일원배치

위의 그림에서 반복이 없는 일원배치에서는 처리효과와 실험단위(오차항)가 교락되어 구별할 수 없다.

예를 들어 철수에게는 A 약을 처방하고 영이에게는 B 약을 처방한 경우, 만약 철수의 치료 효과가 영이보다 좋으면 A 약의 효과가 더 좋다고 말할 수 있는가? 이런 경우 약의 효과인지 실험 대상인 개인의 특성인지 알 수 없다.

반복이 없는 일원배치에서는 효과의 차이를 알 수 있는 통계량이 두 관측값의 차이 $x_1 - x_2$ 밖에 없으며 이를 모형식으로 보면 다음과 같다.

$$x_1 - x_2 = \alpha_1 - \alpha_2 + e_1 - e_2$$

즉 처리 효과 α_i와 오차 e_i의 효과를 분리해야 하는데 사용할 수 있는 통계량이 하나 밖에 없어서 처리효과에 대한 추론이 불가능하다.

여기서 유의할 점은 두 관측값의 차이 $x_1 - x_2$ 와 평균으로부터 편차 $x_1 - \bar{x}$는 기본적으로 같은 정보를 가진 통계량이다.

$$x_1 - \bar{x} = \frac{x_2 - x_1}{2}$$

- 반복이 있는 일원배치

반복이 있는 일원배치의 경우 우리는 2개의 편차를 만들 수 있으며 두 편차가 가지고 정보에서 처리 효과에 대한 정보를 분리해 낼수 있다.

$$\begin{aligned} x_{11} - \bar{\bar{x}} &= (x_{11} - \bar{x}_{1\cdot}) + (\bar{x}_{1\cdot} - \bar{\bar{x}}) \\ &= \frac{1}{2}\left[(-1)x_{12} + (1)x_{11} + (0)x_{21} + (0)x_{22}\right] \\ &\quad + \frac{1}{4}\left[(1)x_{12} + (1)x_{11} + (-1)x_{21} + (-1)x_{22}\right] \\ &= (e_{11} - \bar{e}_{1\cdot}) + ([\alpha_1 - \bar{\alpha}] - [\bar{e}_{1\cdot} - \bar{\bar{e}}]) \end{aligned}$$

반복이 있는 일원배치에서 잔차제곱합 MS_E는 $x_{ij} - \bar{x}_{i\cdot}$가 지닌 정보, 즉 오차항의 분산에 대한 정보를 가지고 있다. 또한 MS_A는 $\bar{x}_{i\cdot} - \bar{\bar{x}}$ 가 지닌 정보, 즉 오차항의 분산과 처리 효과의 정보 모두 가지고 있다. 이러한 사실은 각 평균제곱합의 기대값을 보면 알 수 있다.

제곱합의 기대값을 구하는 방법은 섹션 C.3.3 을 참조하자.

$$E(MS_E) = \sigma_E^2, \quad E(MS_A) = \sigma_E^2 + r\frac{\sum_i^a (\alpha_i - \bar{\alpha})^2}{a-1}$$

따라서 처리효과가 있는지에 대한 검정은 MS_E와 MS_E의 비(ratio)를 이용하여 검정한다(F-검정).

D.0.2. 완전 랜덤화 이원배치

이제 이원배치에서 반복이 없는 경우와 있는 경우를 살펴보자.

D.0.2.1. 반복이 없는 이원배치

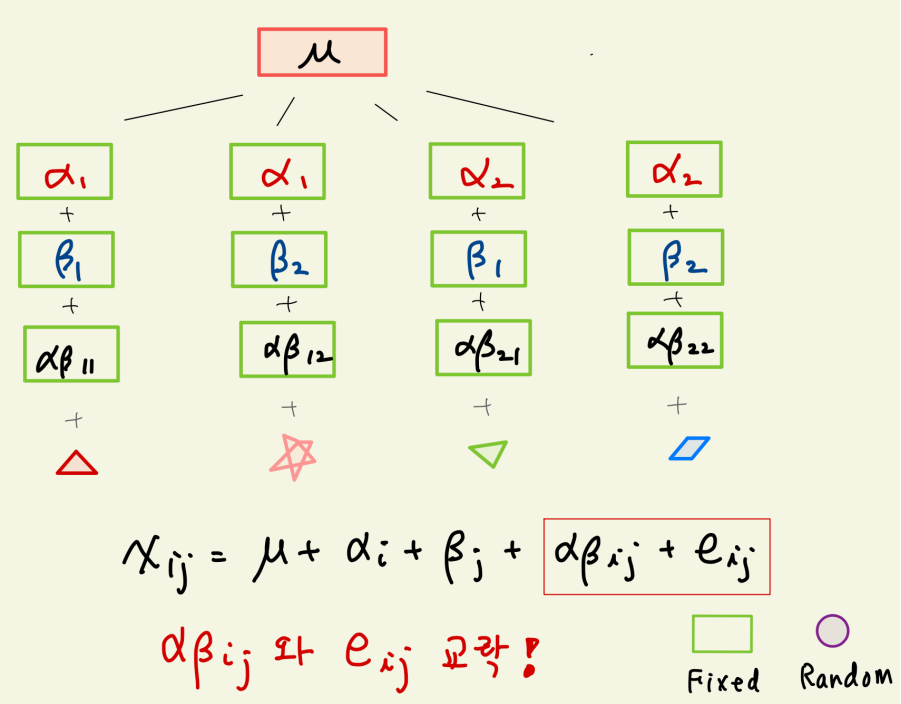


그림 D.2.: 이원배치: 반복이 없는 경우

반복이 없는 이원배치는 관측자료의 편차를 각 효과에 대한 편차들로 다음과 같이 분해할 수 있다.

$$\underbrace{(x_{ij} - \bar{\bar{x}})}_{\text{total deviation}} = \underbrace{(\bar{x}_{i.} - \bar{\bar{x}})}_{\text{A effect}} + \underbrace{(\bar{x}_{.j} - \bar{\bar{x}})}_{\text{B effect}} + \underbrace{(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})}_{\text{(A x B) + residual}}$$

위의 분해에서 이원배치 모형식을 이용하여 마지막 항 $x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}$ 을 모수와 오차로 표현해보면 다음과 같다.

$$x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}} = [(\alpha\beta)_{ij} - (\bar{\alpha\beta})_{i.} - (\bar{\alpha\beta})_{.j} + (\bar{\bar{\alpha\beta}})] + [e_{ij} - \bar{e}_{i.} - \bar{e}_{.j} + \bar{\bar{e}}] \tag{D.1}$$

위의 식을 보면 편차 $x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}$ 는 상호작용에 대한 정보와 오차항의 정보가 섞여 있고 더 이상 분리할 수 없음을 알 수 있다. 따라서 상호작용과 오차항은 교락되어 있다.

D.0.2.2. 반복이 있는 이원배치

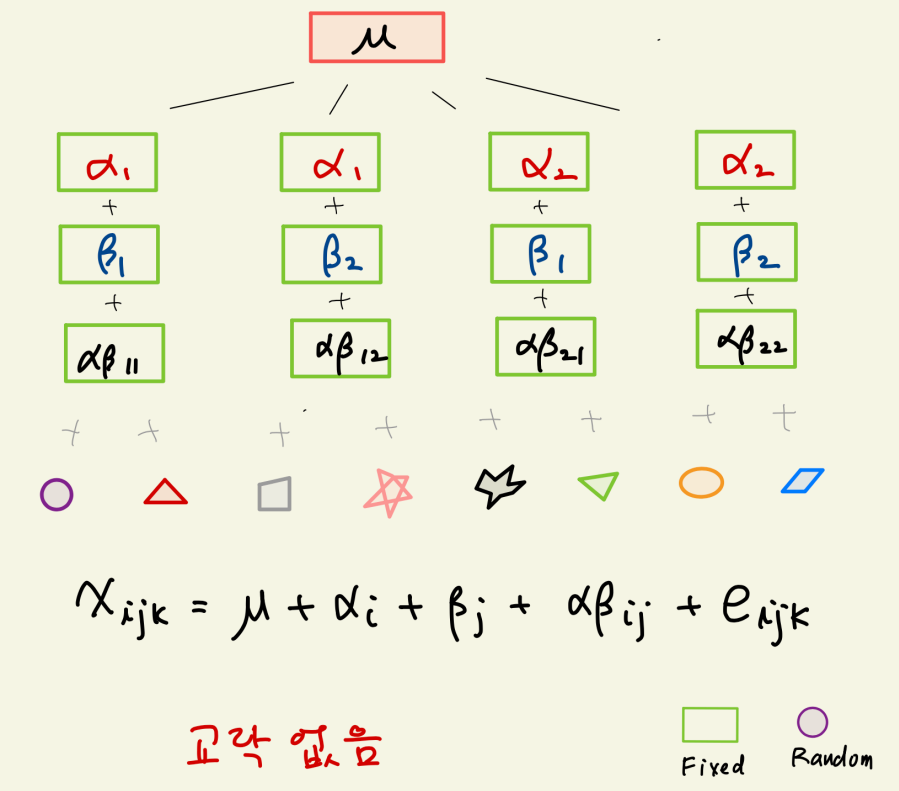


그림 D.3.: 이원배치: 반복이 있는 경우

반복이 있는 이원배치는 관측자료의 편차를 각 효과에 대한 편차들로 다음과 같이 분해할 수 있다. 주목할 점은 반복이 있기 때문에 반복이 없는 경우보다 하나의 항 $x_{ijk} - \bar{x}_{ij.}$ 이 추가된다.

$$\underbrace{(x_{ijk} - \bar{\bar{x}})}_{\text{total deviation}} = \underbrace{(\bar{x}_{i..} - \bar{\bar{x}})}_{\text{A effect}} + \underbrace{(\bar{x}_{.j.} - \bar{\bar{x}})}_{\text{B effect}} + \underbrace{(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{\bar{x}})}_{\text{A x B}} + \underbrace{(x_{ijk} - \bar{x}_{ij.})}_{\text{residual}}$$

반복이 있는 이원배치 모형에서 상호작용에 대한 편차는 반복이 없는 경우의 식 @ref(eq:inter)과 유사하게 다음과 같이 표시할 수 있다.

$$x_{ij.} - \bar{x}_{i..} - \bar{x}_{.j} + \bar{\bar{x}} = [(\alpha\beta)_{ij.} - (\bar{\alpha\beta})_{i.} - (\bar{\alpha\beta})_{.j} + (\bar{\bar{\alpha\beta}})] + [e_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{\bar{e}}]$$

또한 잔차에 대한 편차는 다음과 같이 표시된다.

$$x_{ijk} - \bar{x}_{ij.} = e_{ijk} - \bar{e}_{ij.}$$

이제 잔차에 대한 편차는 순수하게 오차항만의 정보를 가지고 있고 상호작용에 대한 편차는 상호작용과 오차에 대한 정보를 가지고 있다. 따라서 두 편차로 만든 두 개의 제곱합을 이용하면 상호작용에 대한 효과를 분리해낼 수 있다.

따라서 상호작용 효과가 있는지에 대한 검정은 $x_{ij.} - \bar{x}_{i..} - \bar{x}_{.j} + \bar{\bar{x}}$ 로 계산된 $MS_{(A\times B)}$와 $x_{ijk} - \bar{x}_{ij.}$ 로 만들어진 MS_E 의 비(ratio)를 이용하여 검정한다(F-검정).