실험계획법 2024 강의 노트

서울시립대학교 통계학과 이용희

2024-04-08

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)

# 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_2 이라고 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_2 개의 독립 표본을 얻었다고 하자.

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

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_0 의 절대값이 자유도 $df = n_1 + n_2 - 2$ 를 가지는 t-분포의 상위 $\alpha/2$ 분위수보다 크면 귀무가설을 기각하고 대립가설 H_1 을 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
```

```r
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 증가하면 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: $A_1, A_1, A_4, A_3, A_2, A_4, \cdots$

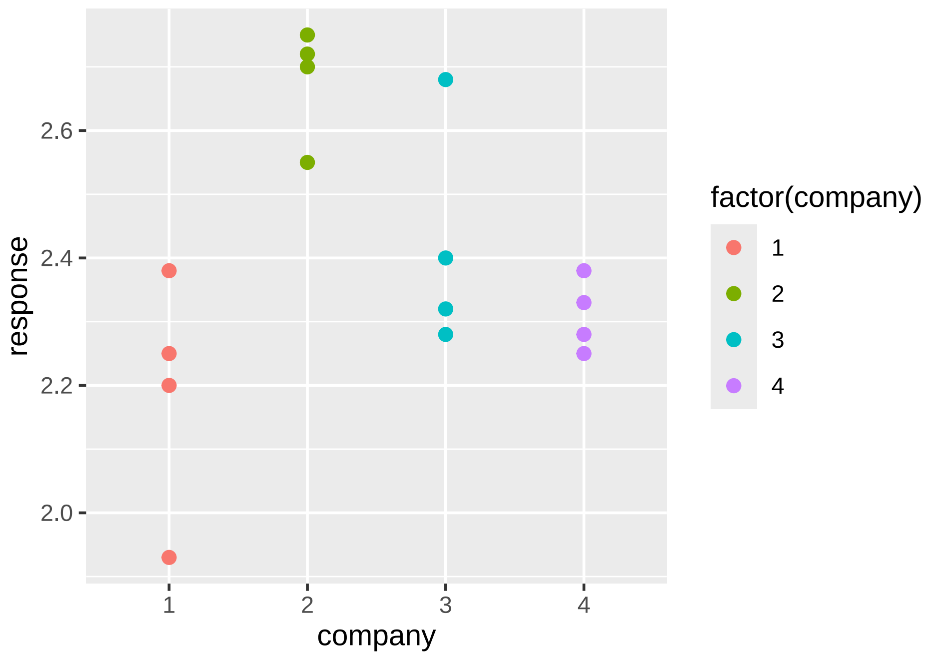
1.2.3. 예제 3.1 - 자료

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.

```r
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
```

```r
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 i 번 수준에서 측정된 j 번째 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 i 는 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_i 는 i 번째 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_A 를 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_E 를 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 α 에서 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.

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

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
```

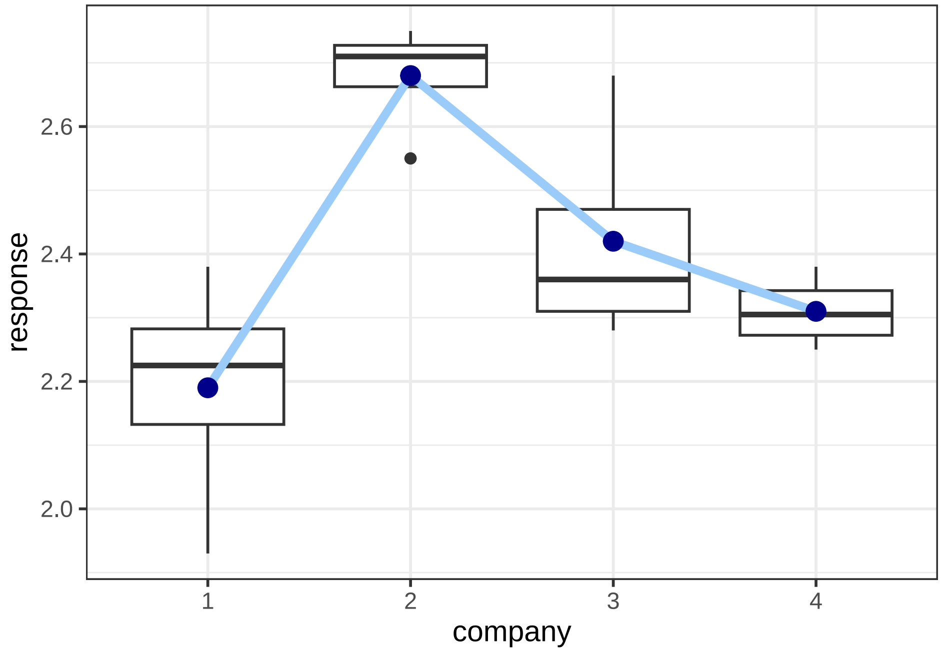
```
   company response
1        1     1.93
2        1     2.38
3        1     2.20
4        1     2.25
5        2     2.55
6        2     2.72
7        2     2.75
8        2     2.70
9        3     2.40
10       3     2.68
11       3     2.32
12       3     2.28
13       4     2.33
14       4     2.38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

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

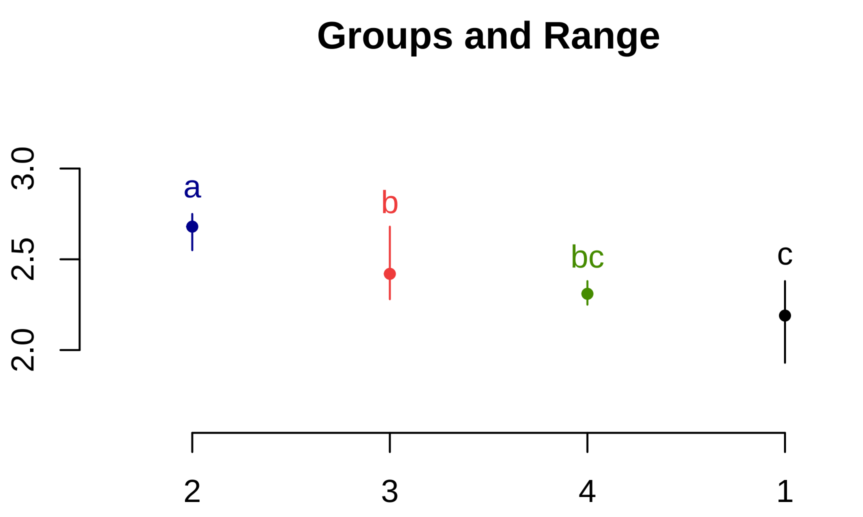
```
Study: anova.res ~ "company"

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
```

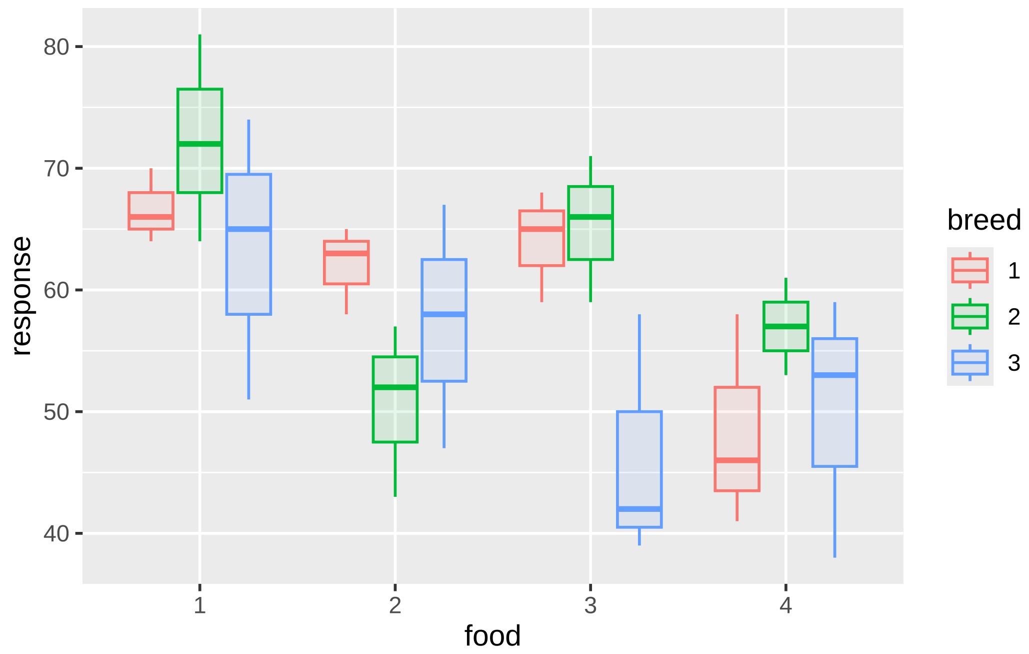
```
df2<- data.frame(food, breed, response)
head(df2)
```

```
  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
```

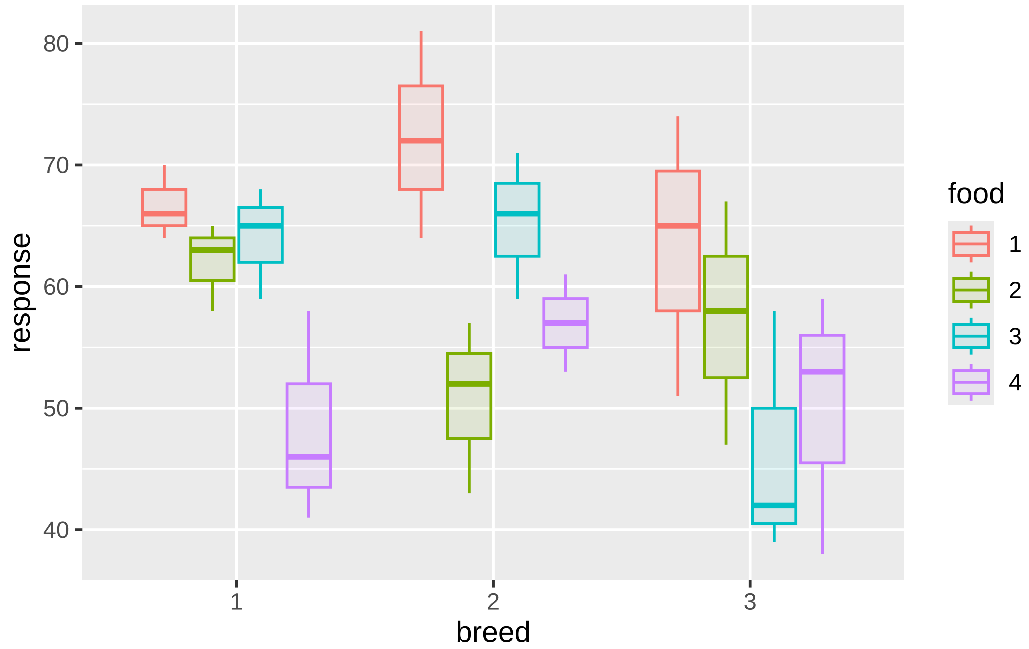
2.1.2. 자료의 시각화와 기초 통계량

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)
```

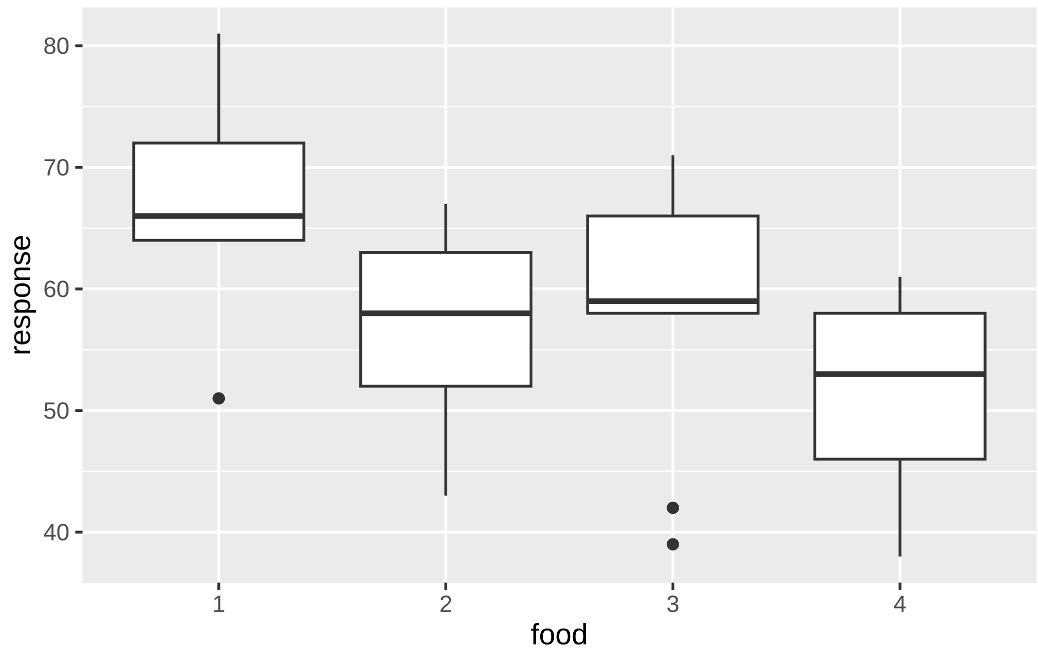



```
df2 %>%
  ggplot() +
  aes(x = breed  , y = response, fill=food, color=food) +
  geom_boxplot(alpha = 0.1, width = 0.75)
```

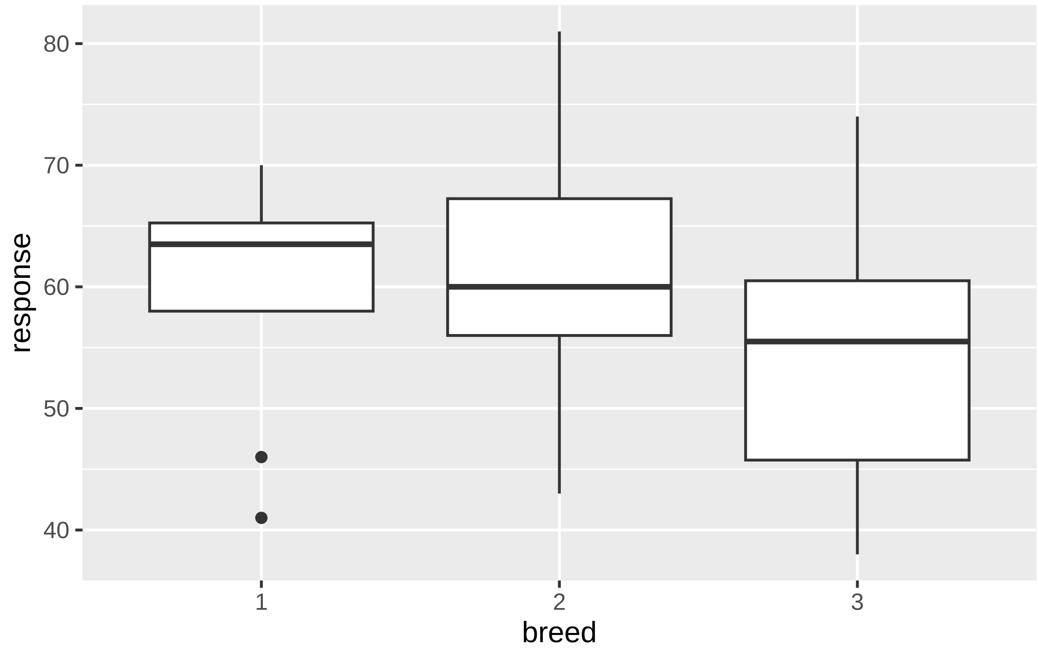


위와 같이 두 요인의 조합으로 그림을 보는 것보다 각 요인별로 요약하여 보는 것도 유용하다.

```
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
```

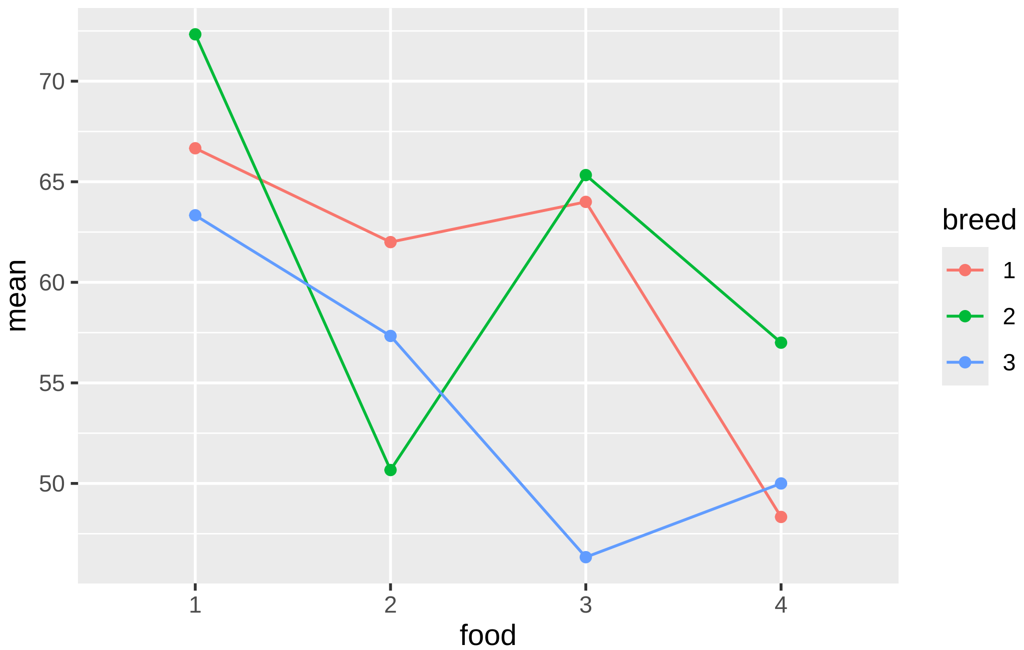
```
  <fct> <dbl> <dbl>
1 1      67.4  8.34
2 2      56.7  8.08
3 3      58.6 11.2
4 4      51.8  8.26
```

```
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
```

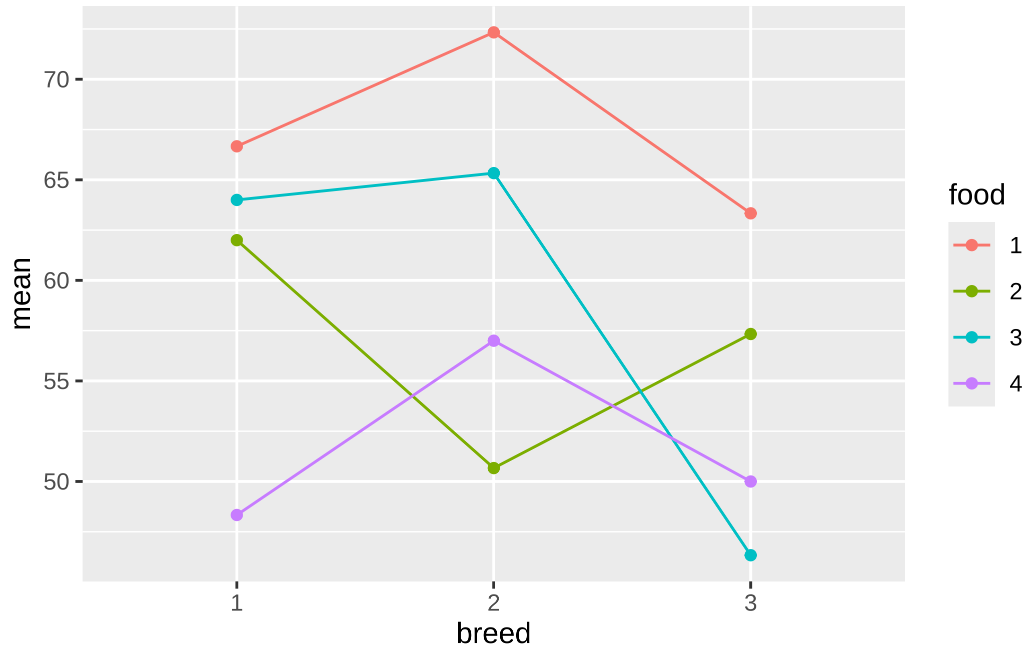
이제 위에서 계산된 처리 그룹에 대한 평균으로 상호작용 그림을 그려보자. 아래 그림에서 사료의 종류에 따라서 체중의 변화를 본 그림이다. 사료 1번에서 체중이 가장 크게 나타났고 다른 사료에 대해서는 체중이 줄어드는데 품종에 따라서 그 크기가 서로 다르다.

```
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 $A_i B_j$에 대한 모평균 μ_{ij} 에 대한 추정량은 처리수준 $A_i B_j$에서의 관측값들의 평균 $\bar{x}_{ij.}$ 이며 오차항의 분산 σ^2_E는 분산분석표에서 MS_E로 추정할 수 있다.

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

위의 결과를 이용하면 처리수준 $A_i B_j$에 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.

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

```
[1] 2.063899
```

따라서 μ_{11} 에 대한 95% 신뢰 구간은 다음과 같다.

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

패키지 `emmeans`에 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.

$$x_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \quad (2.2)$$

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

```
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,

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

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 유도된 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.

```r
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.

```r
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 `Table5.1`에 있다.

```
df <- Table5.1
head(df) # 자료의 앞부부만 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 `df`에 있는 변수 `MaterialType`은 문자형 변수(`chr`)이고 나머지는 숫자형 변수(`num`)이다. 두 요인에 대한 변수인 `MaterialType`와 `Temperature`를 함수 `factor()`를 이용하여 범주형 변수로 만들어 주자 .

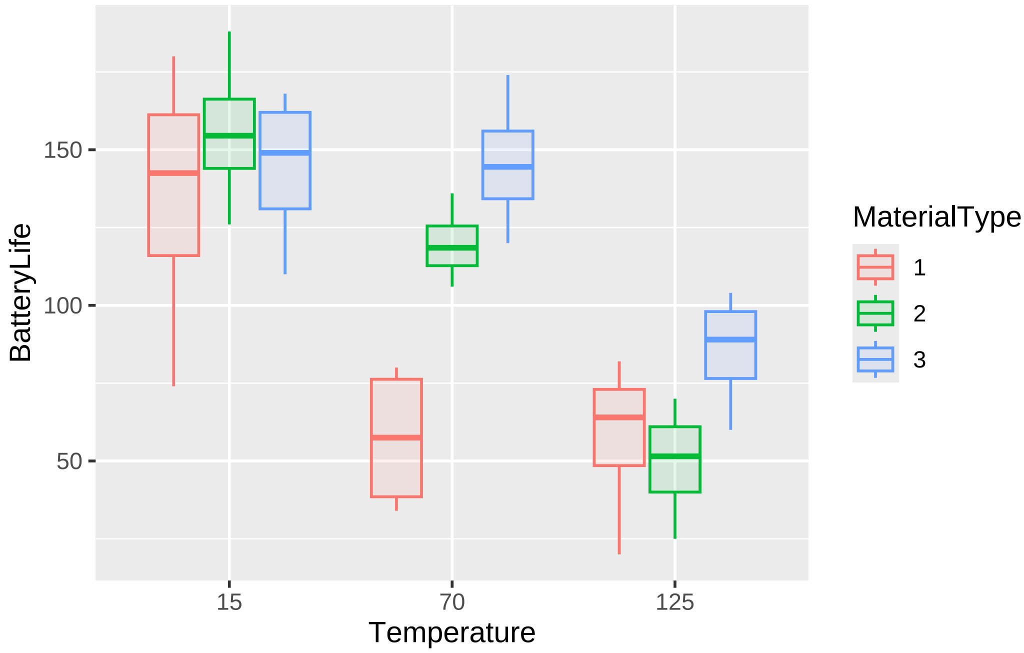
```
df$MaterialType <- factor(df$MaterialType)
df$Temperature <- factor(df$Temperature)
str(df)
```

```
'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)
```



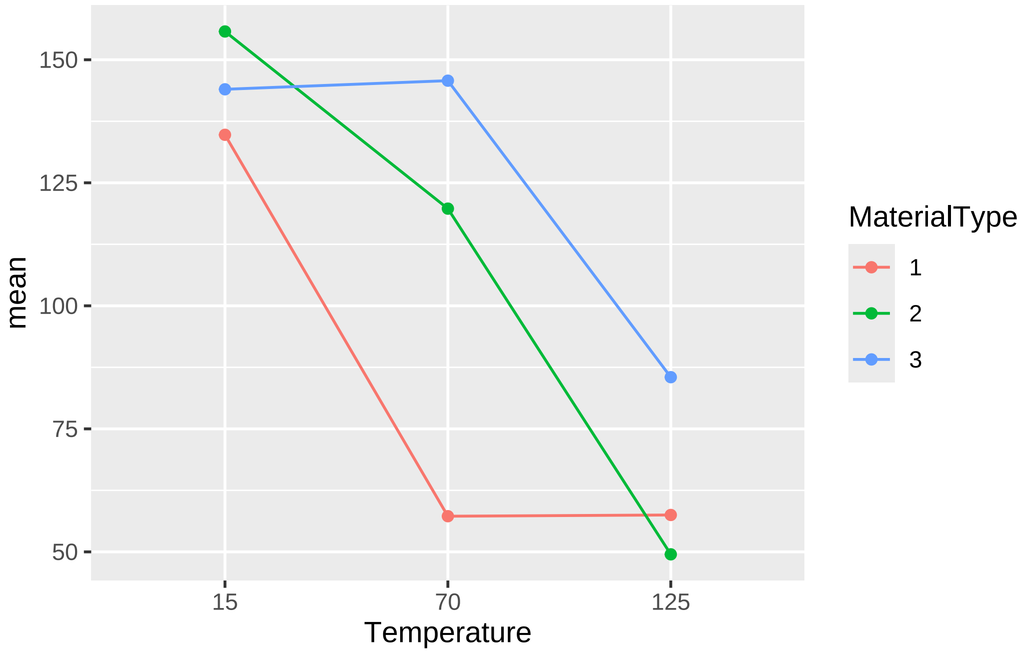
다음으로 6개의 처리 조합에 대한 전지 수명의 기초통계량(평균과 표준편차)을 구해보자.

```
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 형테가 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.

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

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
```

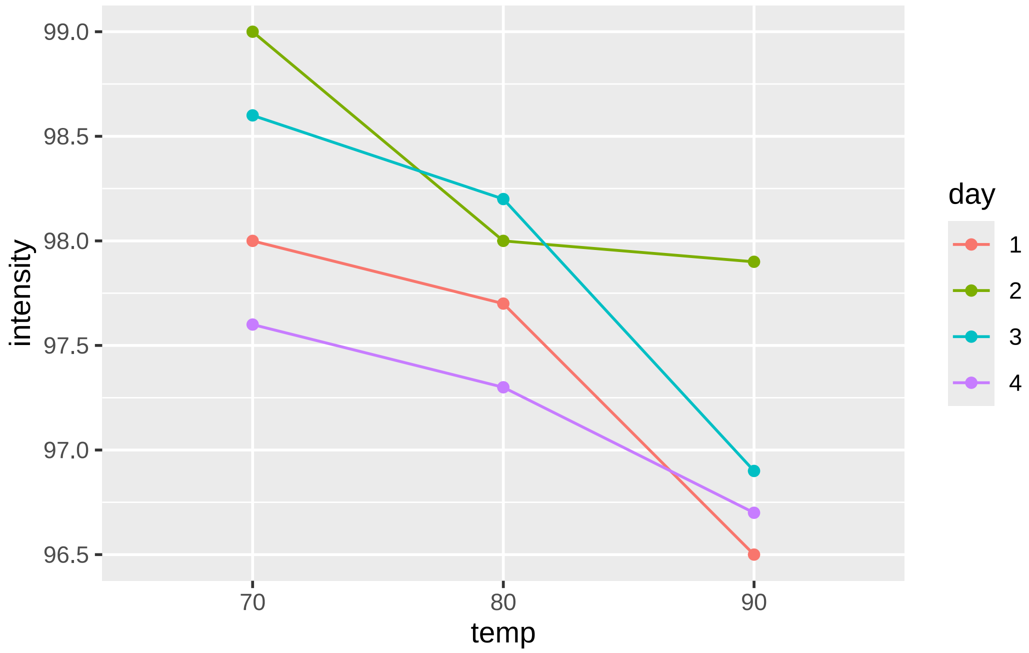
```
12        96.7    90    4
```

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

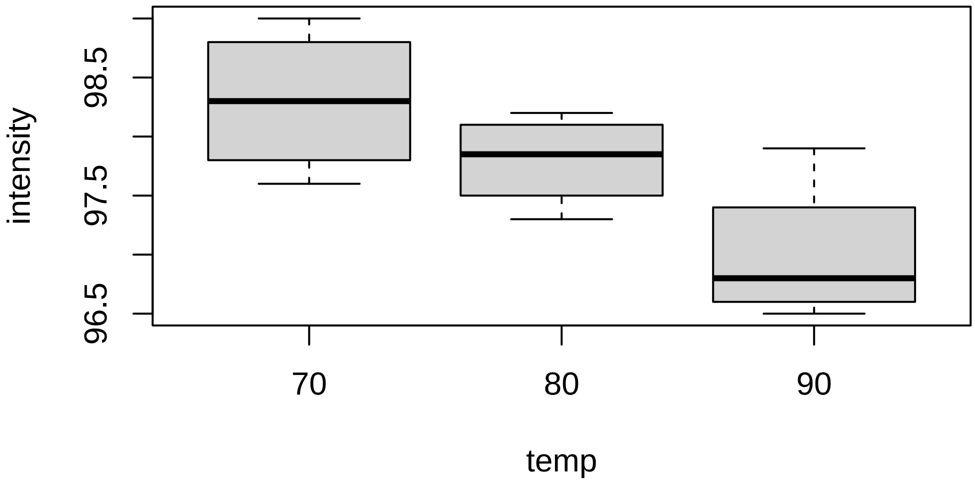
3.1.2. 시각적 분석

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()
```

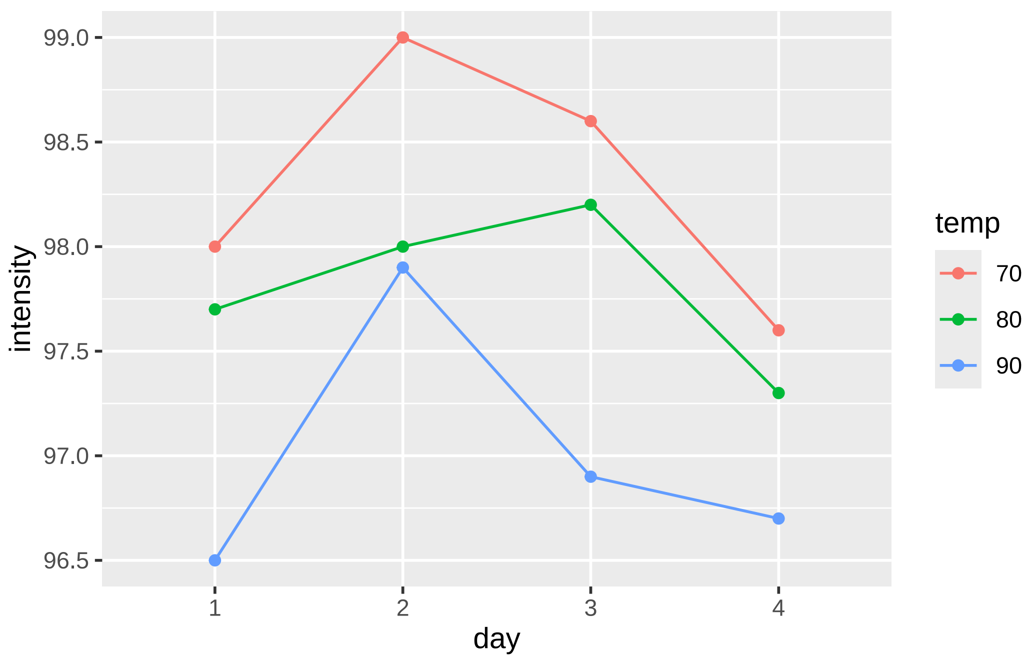



```
plot(intensity ~ temp, data=df)
```



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 실험울 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^2 의 추정치는 0.2155556 이며 오차항(`Residual`)의 분산 σ_E^2 의 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 `anova`에 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.

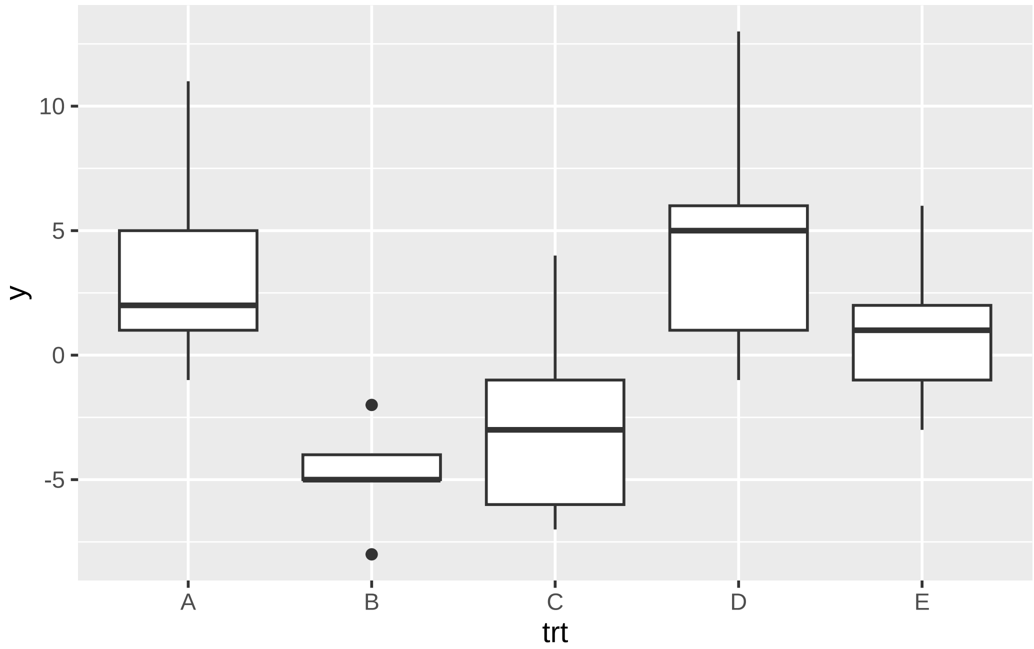
```r
xtabs(y~ R + C, data = df)
```

```
   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

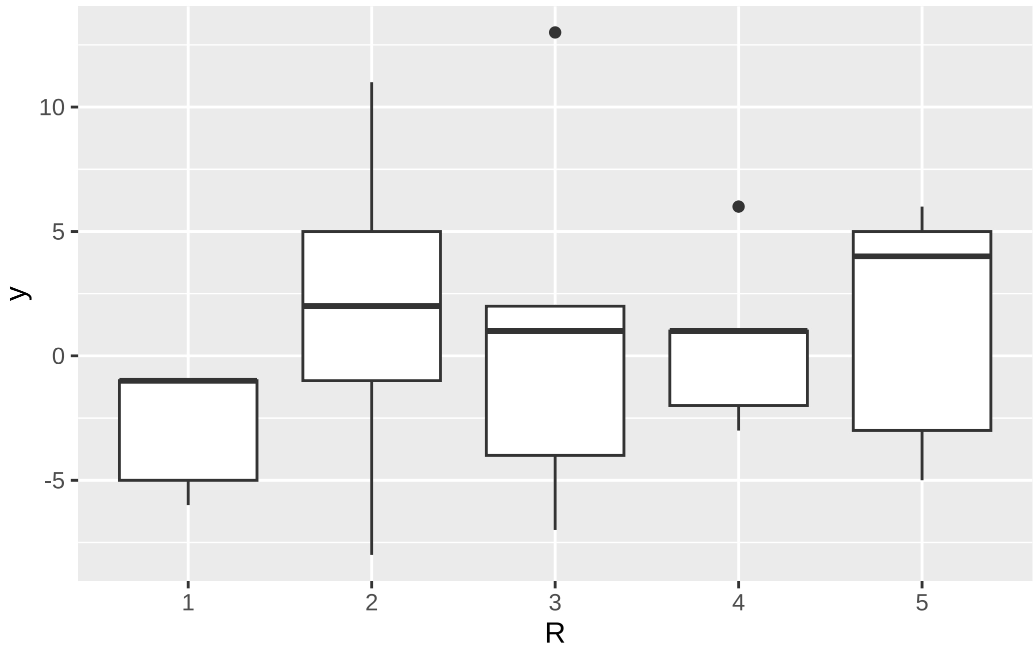
먼저 로켓 추진체, 즉 처리별로 자료의 분포를 보자. 추진체 B 와 C 가 다른 추진체들 보다 관측값이 작게 나오는 것을 알 수 있다.

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



원료(R) 뭉치별로 자료의 분포를 보면 큰 차이는 보이지 않는다.

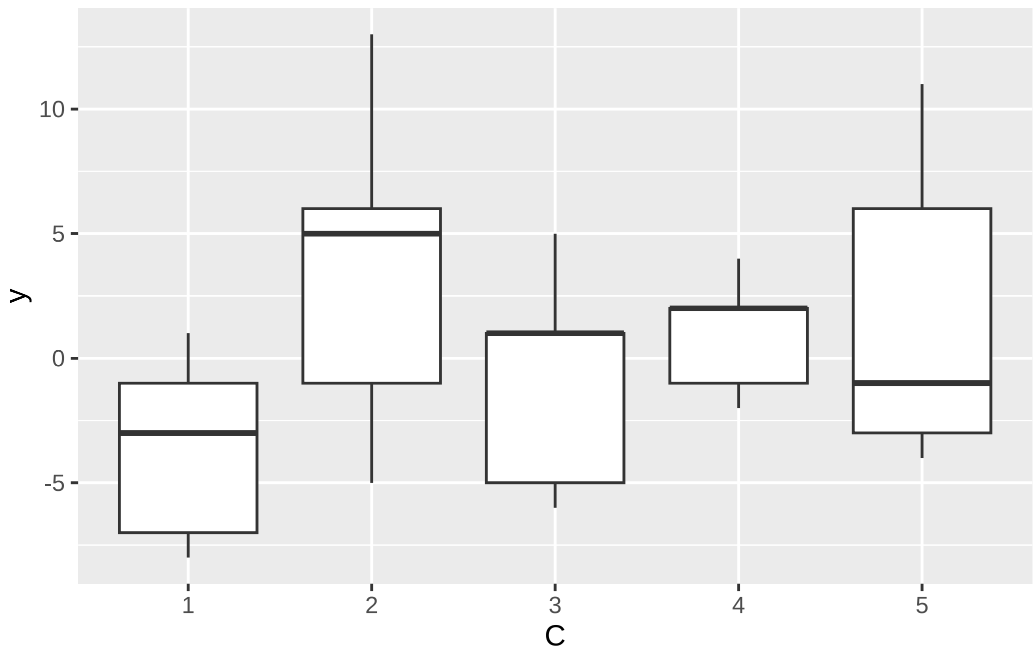
```
df %>%
  ggplot() +
  aes(x = R , y = y) +
  geom_boxplot()
```



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 대해 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.

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

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

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

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

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

```r
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.

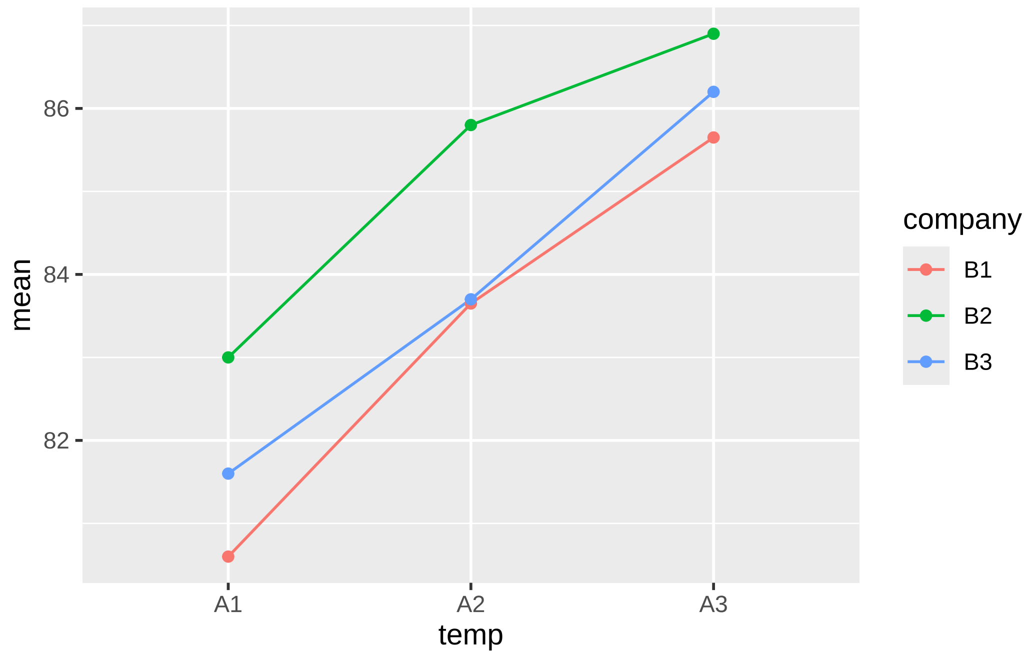
```
dfsum <- df %>% group_by(temp, company)  %>%  summarise(mean=mean(y),  sd=sd(y))
dfsum
```

```
# 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.

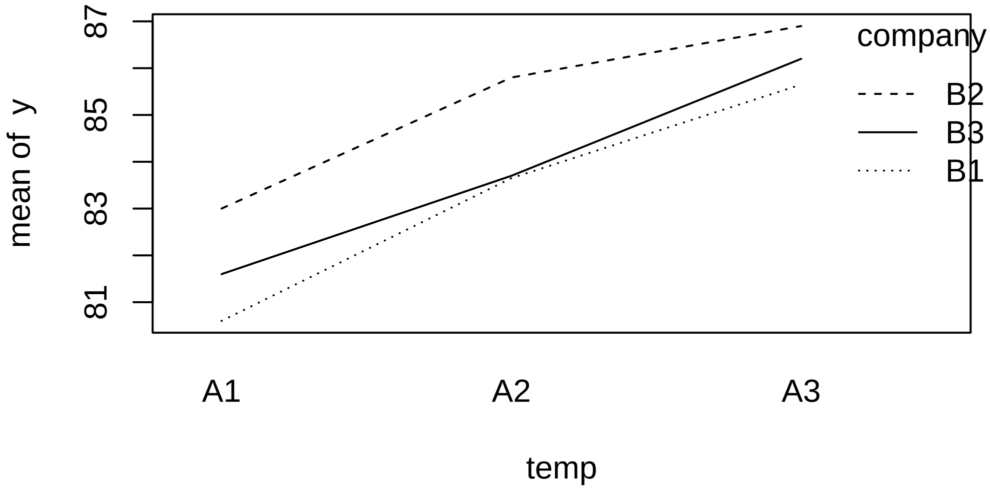
아래 상호작용 그림을 보면 온도에 따라서 화학약품의 생성률이 크게 변하는 것을 알 수 있다. 유의한 상호작용은 관측되지 않는다.

```
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))
```

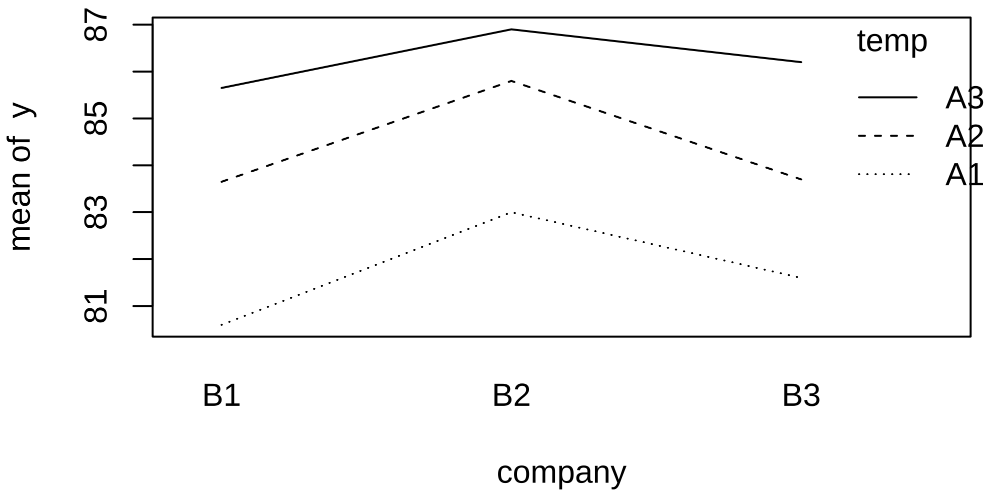


ℹ 노트

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:company`는 두 처리 `temp`와 `company`의 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.

> **i 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.

```r
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
```

```r
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)
```

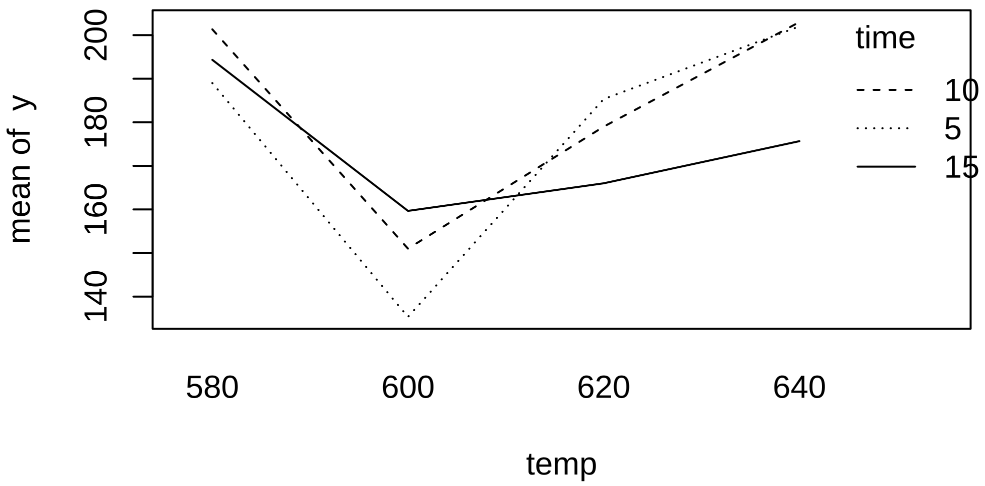
```
, , rep = 1

    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

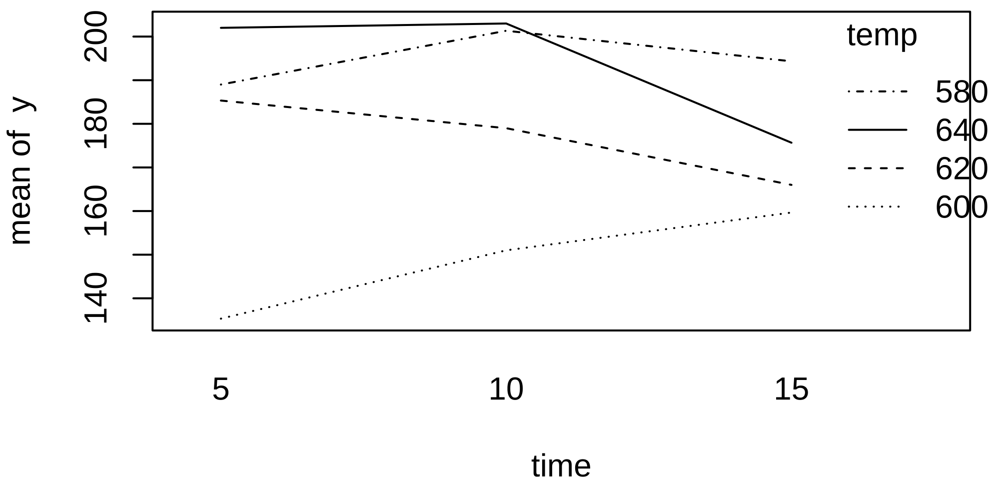
이제 온도의 수준에 따른 변화를 볼 수 있는 그림을 그려보자. 온도가 증가하면서 수명이 줄어들었다가 다시 늘어나는 현상을 볼 수 있다.

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



가열시간의 수준에 따른 변화를 볼 수 있는 그림을 그려보자. 가열시간이 증가하더라도 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.

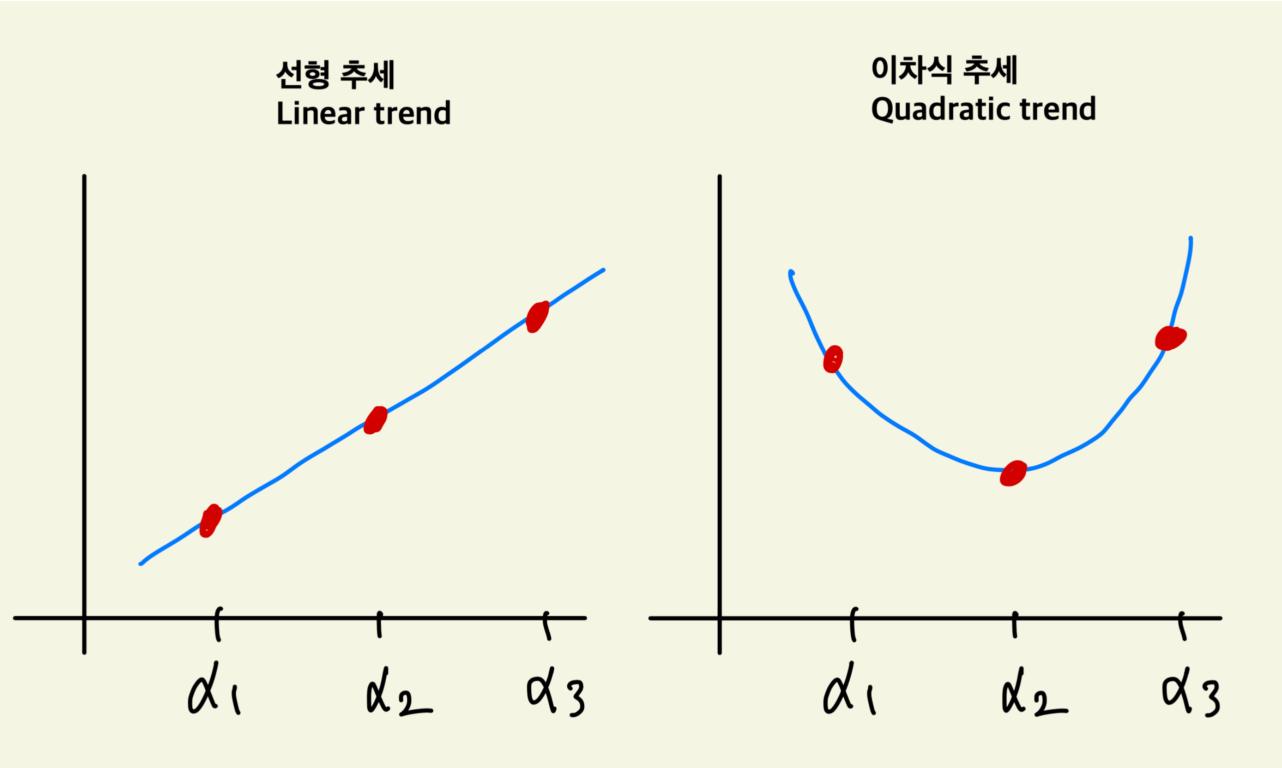


그림 4.1.: 3개의 수준이 있는 경우 선형 추세와 이차식 추세

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 ψ 의 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_i 들이 다음과 같은 경우이며

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

가설 식 4.7 은 일반적인 가설 식 4.8 에서 계수 c_i 들이 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 ψ 을 **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

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

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

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

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 $T_1, T_2, \dots, T_a$ 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

$$\begin{aligned} L &= c_1 T_1 + c_2 T_2 + \dots + c_a T_a \\ &= r(c_1 \bar{x}_{1.} + c_2 \bar{x}_{2.} + \dots + c_a \bar{x}_{a.}) \\ &= rL_* \end{aligned}$$

따라서 합으로 표시한 선형식 L 의 평균과 분산은 다음과 같다

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

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 게수들에 상수를 곱해줘도 검정통계량의 변화는 없다.

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

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

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

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 다음과 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)
```

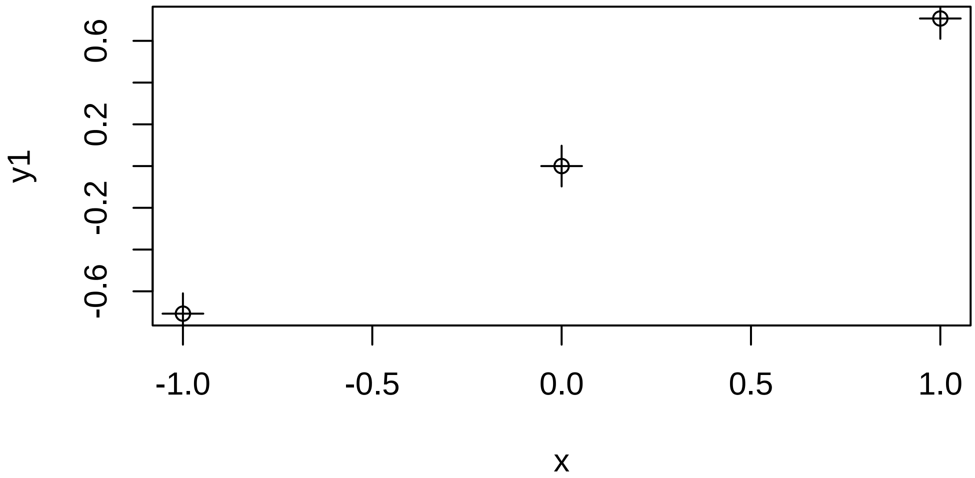
```
                .L         .Q
[1,] -7.071068e-01  0.4082483
[2,] -9.073800e-17 -0.8164966
[3,]  7.071068e-01  0.4082483
```

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)
```

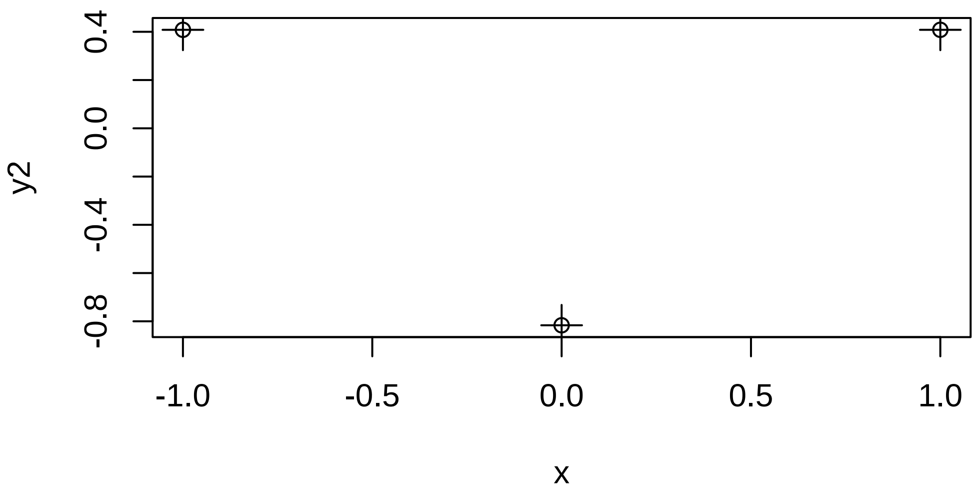


두 번째 열은 이차 대비(quadratic contrast)로 계수는 다음과 같다.

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

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.

```r
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 $\boldsymbol{A}$ ($\boldsymbol{A_0}$:자회사, $\boldsymbol{A_1}$:국내 타회사, $\boldsymbol{A_2}$:외국회사) 와 성형온도 $\boldsymbol{B}$($\boldsymbol{B_0}$: 100℃, $\boldsymbol{B_1}$: 110℃, $\boldsymbol{B_2}$: 120℃)가 플라스틱강도에 미치는 영향은?

< 플라스틱 제품의 강도 >

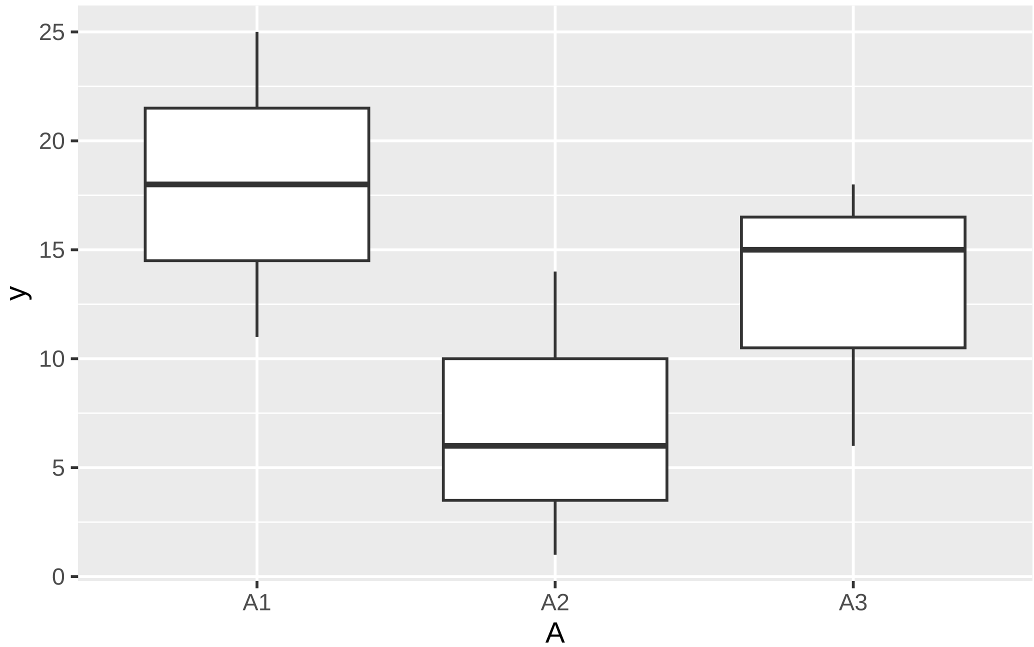
A \ B	B_0	B_1	B_2	$T_{i\cdot}$
A_0	11	18	25	54
A_1	1	6	14	21
A_2	6	15	18	39
$T_{\cdot j}$	18	39	57	T=114

그림 4.2.: 이원배치 자료

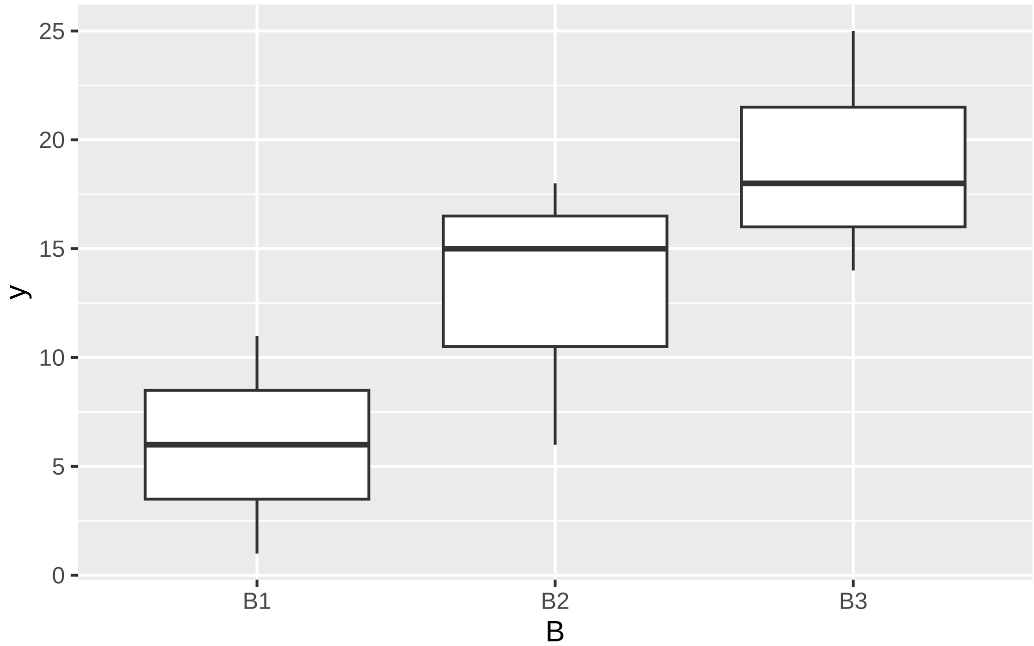
```r
y <- c(11,18,25,1,6,14,6,15,18)
A <- factor(c(rep(c("A1", "A2","A3"), each=3)))
B <- factor(c(rep(c("B1", "B2","B3"), 3)))
df <- data.frame(A,B,y)
df
```

```
   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

```r
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 *A*의 변동의 분해

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

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

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 *B*의 변동의 분해

L_l : 일차효과 대비 & L_q: 이차효과 대비

$$L_l = (T_{\cdot 1} + T_{\cdot 0}) + (T_{\cdot 2} - T_{\cdot 1}) = T_{\cdot 2} - T_{\cdot 0} = 39$$

$$L_q = (T_{\cdot 2} - T_{\cdot 1}) - (T_{\cdot 1} - T_{\cdot 0}) = T_{\cdot 2} - 2T_{\cdot 1} + T_{\cdot 0} = -3$$

$\Rightarrow L_l$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

References

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.

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.frame` 의 형식인 `df0`에 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 `df0`에는 각 그룹(`line1`과 `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 `line`를 생성하여 다른 형태의 데이터프레임 `df`를 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 `df0`에 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.

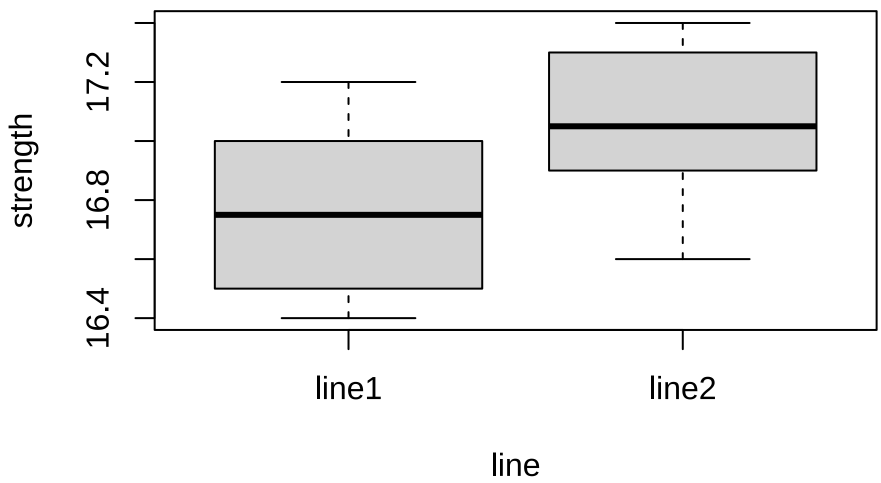
```
df22 %>% group_by(line)  %>%  summarise(mean=mean(strength), median= median(strength), sd=sd(strength),
```

```
# 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

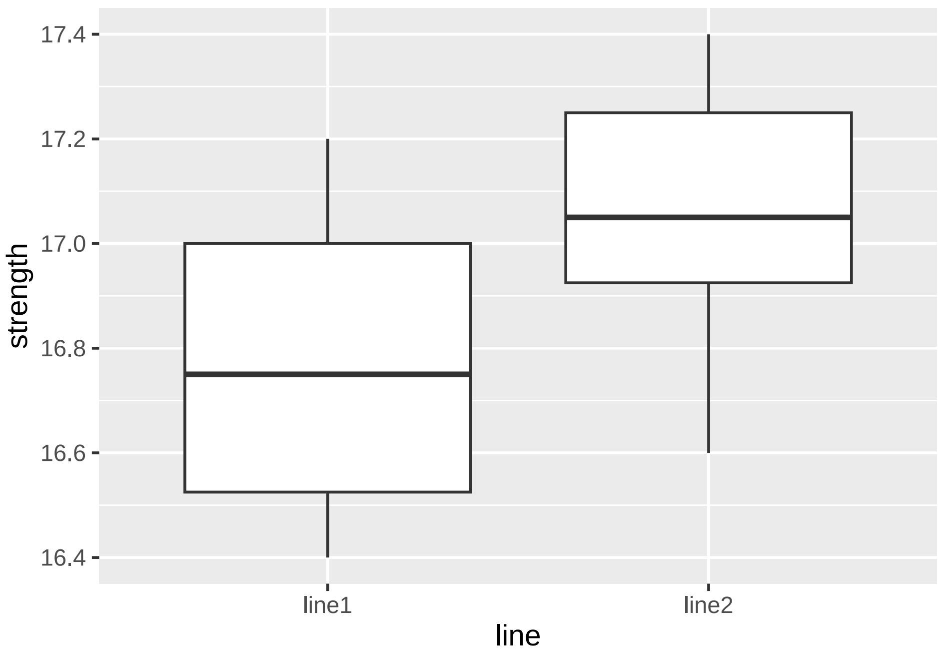
다음으로 각 집단별로 상자그림(boxplot)을 그려서 자료의 중심과 퍼진 정도를 그림으로 비교해 보자. 위에서 좁은 형태로 구성된 자료에 대하여 다음과 같은 명령어로 상자그림을 집단별로 그릴 수 있다.

```
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 검사을 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
```

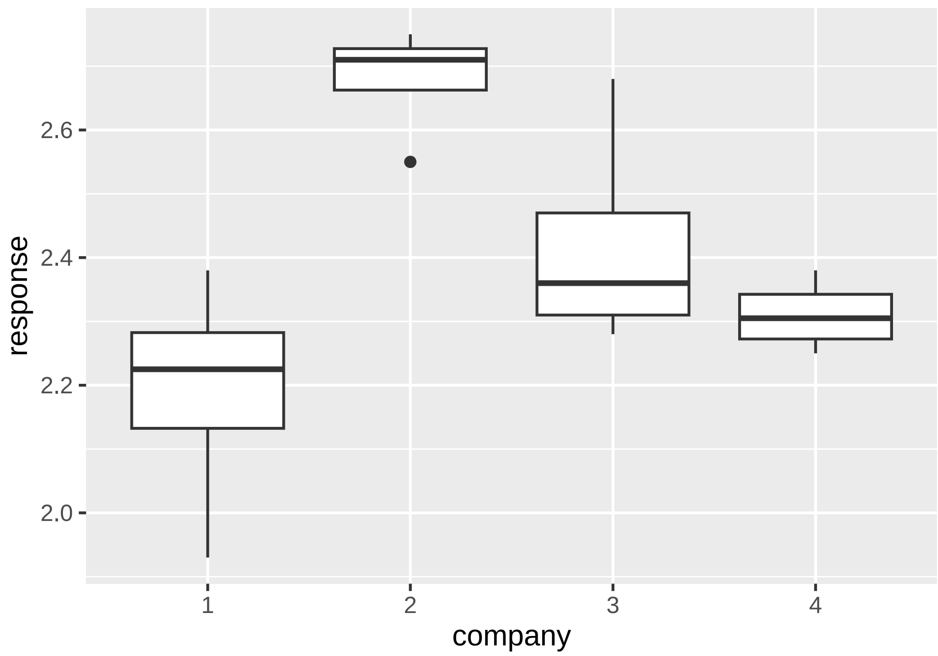
A.2.2. 기술 통계량에 의한 요약

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

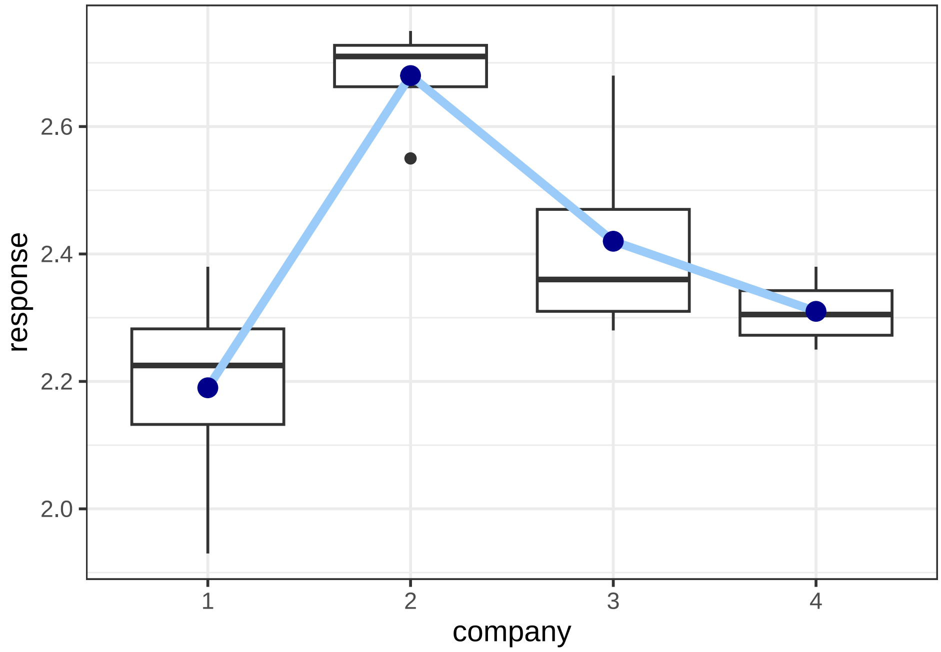
```
# A tibble: 4 x 6
  company  mean median     sd   min   max
  <fct>   <dbl>  <dbl>  <dbl> <dbl> <dbl>
1 1        2.19   2.22 0.189   1.93  2.38
2 2        2.68   2.71 0.0891  2.55  2.75
3 3        2.42   2.36 0.180   2.28  2.68
4 4        2.31   2.30 0.0572  2.25  2.38
```

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.

$$\min_{\mu,\alpha_1,\dots\alpha_a} \sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij}-\mu-\alpha_i)^2 = \min_{\mu,\alpha_1,\dots\alpha_a} SSE \tag{B.2}$$

위의 오차제곱합이 모든 모수에 대하여 미분가능한 이차식으므로 최소제곱 추정량은 제곱합을 모수에 대하여 미분하고 0 으로 놓아 방정식을 풀어서 얻을 수 있다.

오차제곱합을 모수 μ와 $\alpha_1, \alpha_2, \dots, \alpha_a$ 로 미분하여 0 으로 놓은 방정식은 다음과 같다.

$$
\begin{aligned}
\frac{\partial}{\partial \mu} SSE &= -2\sum_{i=1}^{a}\sum_{j=1}^{r}(y_{ij}-\mu-\alpha_i) = 0 \\
\frac{\partial}{\partial \alpha_i} SSE &= -2\sum_{j=1}^{r}(y_{ij}-\mu-\alpha_i) = 0, \quad i = 1, 2, \dots, a
\end{aligned}
$$

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_1 를 0으로 놓는 조건을 주는 것이다 ($\alpha_1 = 0$). set-to-zero 조건 하에서는 다음과 같은 추정량이 얻어진다.

$$\hat{\mu} = \bar{y}_{1.}, \quad \hat{\alpha}_1 = 0, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{1.}, \quad i = 2, \dots, a \tag{B.4}$$

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.

$$\hat{\mu} = \bar{\bar{y}}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{\bar{y}}, \quad i = 1, 2, \dots, a \tag{B.5}$$

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.

$$
\min_{\mu,\alpha_1,\dots\alpha_a} \sum_{i=1}^{a} \sum_{j=1}^{r} (y_{ij} - \mu - \alpha_i)^2 = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X\beta})^t (\boldsymbol{y} - \boldsymbol{X\beta}) \tag{B.8}
$$

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.

$$
\begin{bmatrix} ar & r & r & \cdot & \cdot & r \\ r & r & 0 & \cdot & \cdot & 0 \\ r & 0 & r & \cdot & \cdot & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ r & 0 & 0 & \cdot & \cdot & r \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_a \end{bmatrix} = \begin{bmatrix} ar\bar{\bar{y}} \\ r\bar{y}_{1.} \\ r\bar{y}_{2.} \\ . \\ . \\ r\bar{y}_{a.} \end{bmatrix} \tag{B.10}
$$

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_1 을 제외하고 선형모형식 식 B.7 를 다음과 같이 다시 표현할 수 있다.
효과 α_1 을 0 으로 놓는다는 것은 α_1 을 추정할 필요가 없으므로 모수벡터 $\boldsymbol{\beta}$ 에서 α_1 를 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 게획행렬에 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 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ 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_1 에 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 $c_0, c_1, \dots, c_a$ 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.

$$E(\hat{\psi}|\boldsymbol{X}) = E(\boldsymbol{a}^t\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{a}^t E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{a}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{c}^t\boldsymbol{\beta}$$

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-lm2**을 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_i 는 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(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(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
```

```r
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 알아보있다.

$$x_{ij} = \mu + \alpha_i + e_{ij} \tag{C.1}$$

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 $\alpha_1, \alpha_2, \dots, \alpha_a$ 는 모두 고정된 값을 가지는 모수(parameter)이다. 식 C.1 의 오른쪽 항들 중에서 확률변수는 오차항 e_{ij} 이 유일하다.

처리수준의 효과 α_i 들이 모수이라는 것은 의미는 만약 새로운 실험에서 실험단위(experiment unit)에 동일한 처리를 적용하면 평균 처리 효과는 α_i 로 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 없는 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.
> 길은 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 $\alpha_1, \alpha_2, \dots, \alpha_a$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

> **i** 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.

$$
\underbrace{V(x_{ij})}_{\text{total variation}} \quad = \quad \underbrace{\sigma_A^2}_{\text{variation between groups}} \quad + \quad \underbrace{\sigma_E^2}_{\text{variation within group}}
$$

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^2 가 그룹 내 변동을 나타내는 오차항의 분산 σ_E^2보다 상대적으로 클수록 급내 상관계수가 1에 가까와진다.

보통 σ_A^2 을 집단간 변동(between-group variance)라 하고 σ_E^2 를 집단내 변동(within-group variance)라고 한다. 따라서 σ_A^2 와 σ_E^2 의 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^2 와 σ_E^2 에 대한 추정량을 얻을 수 있다.

첫 째로 분산분석 표에서 SS_E 의 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_E 의 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_A 의 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_A 의 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_i 는 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^2 와 σ_E^2 의 불편 추정량을 구할 수 있다. 여기서 유의할 사항은 σ_A^2 에 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^2 가 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 프로그램을 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

```r
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.

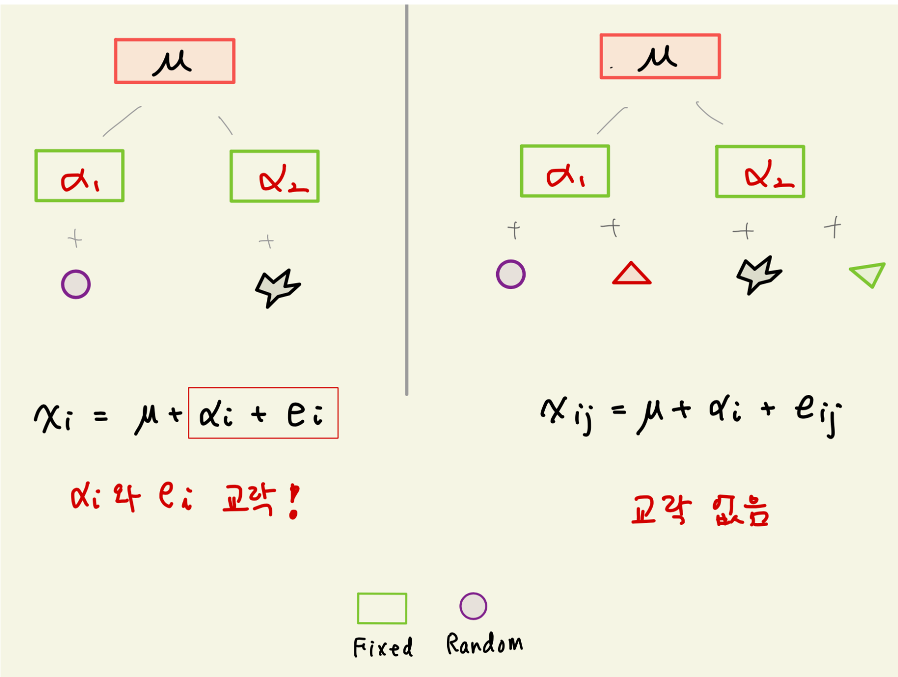


그림 D.1.: 일원배치: 반복이 없는 경우와 있는 경우

- 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_i 와 오차 e_i 의 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.

$$\begin{aligned} x_{11} - \bar{\bar{x}} &= (x_{11} - \bar{x}_{1.}) + (\bar{x}_{1.} - \bar{\bar{x}}) \\ &= \frac{1}{2}\left[(-1)x_{12} + (1)x_{11} + (0)x_{21} + (0)x_{22}\right] \\ &\quad + \frac{1}{4}\left[(1)x_{12} + (1)x_{11} + (-1)x_{21} + (-1)x_{22}\right] \\ &= (e_{11} - \bar{e}_{1.}) + ([\alpha_1 - \bar{\alpha}] - [\bar{e}_{1.} - \bar{\bar{e}}]) \end{aligned}$$

반복이 있는 일원배치에서 잔차제곱합 MS_E 는 $x_{ij} - \bar{x}_{i.}$ 가 지닌 정보, 즉 오차항의 분산에 대한 정보를 가지고 있다. 또한 MS_A 는 $\bar{x}_{i.} - \bar{\bar{x}}$ 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_E 와 MS_E 의 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

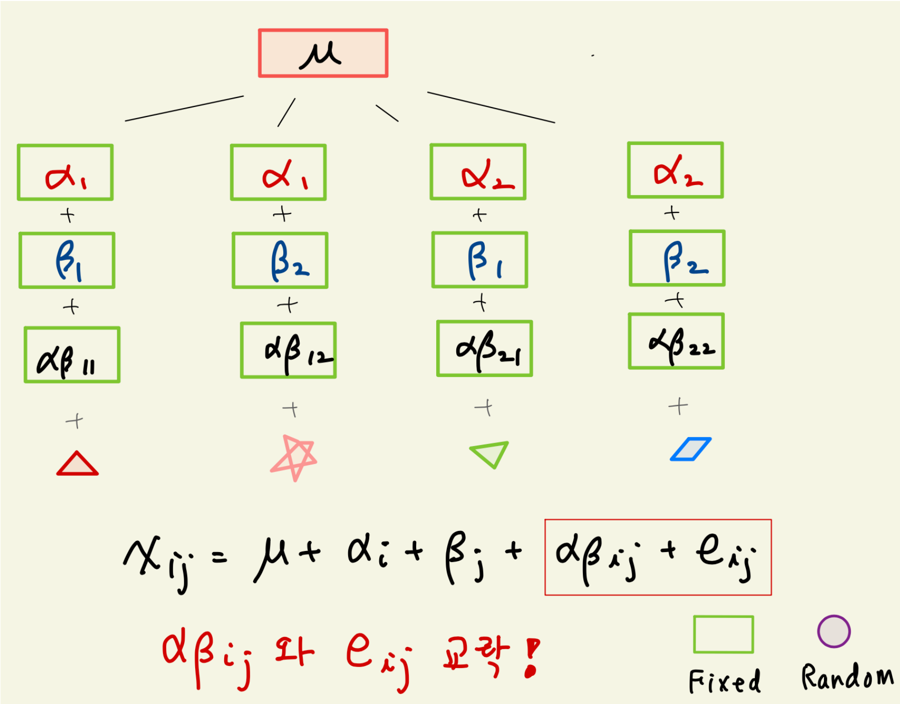


그림 D.2.: 이원배치: 반복이 없는 경우

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

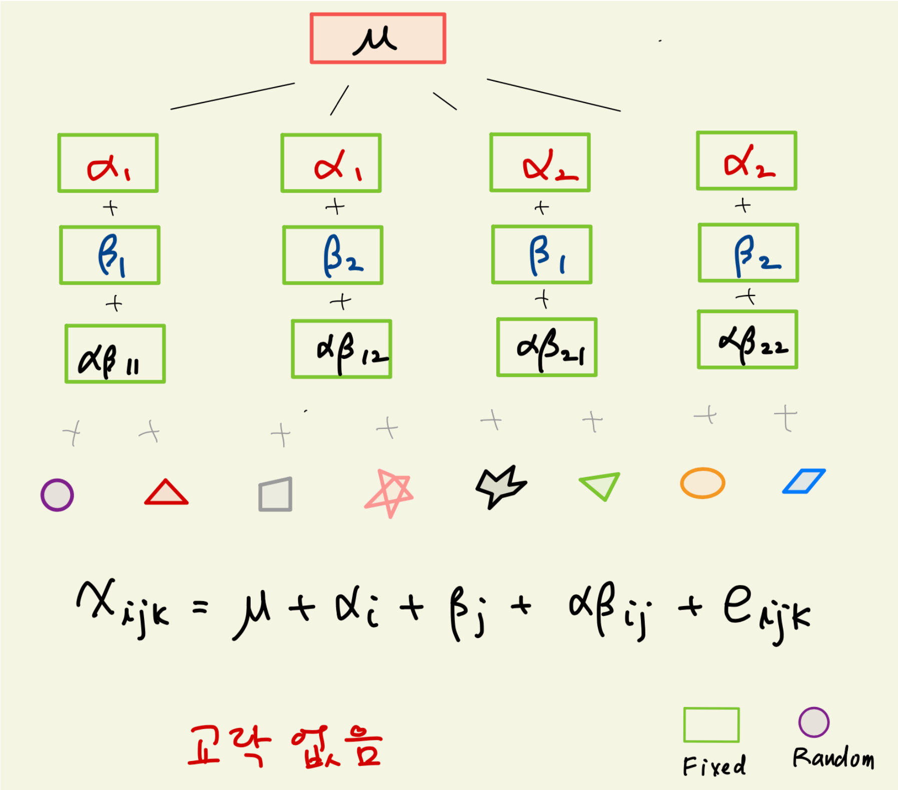


그림 D.3.: 이원배치: 반복이 있는 경우

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.

$$x_{ij.} - \bar{x}_{i..} - \bar{x}_{..j} + \bar{\bar{x}} = [(\alpha\beta)_{ij.} - (\bar{\alpha\beta})_{i.} - (\bar{\alpha\beta})_{.j} + (\bar{\bar{\alpha\beta}})] + [e_{ij.} - \bar{e}_{i..} - \bar{e}_{.j.} + \bar{e}]$$

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).