

의학통계학

이용희

2024-11-03, 오후 11 시

Table of contents

Preface	1
개요	2
1. 서론	3
1.1. 공정한 실험	1
1.2. 의학연구의 특성	1
1.3. 통계적 주제	2
1.4. 강의의 구성	2
I. 분할표와 연관성	3
2. 연관성의 측도	4
2.1. 필요한 패키지	4
2.2. 이항변수	4
2.3. 분할표와 연관성의 측도	5
2.3.1. 분할표	5
2.3.2. 상대위험	6
2.3.3. 기여위험과 백신효과	7
2.3.4. 오즈비	8
2.4. 신뢰구간	9
2.4.1. 예제: 아스피린 임상실험	10
2.5. 사례-대조 연구	12
2.5.1. 사례대조 연구의 목표와 가설	12
2.5.2. 오즈비의 비교	13
2.5.3. 예제: 약물남용 사례-대조 연구	14
3. 연관성의 검정	16
3.1. 필요한 패키지	16
3.2. 카이제곱 검정	16
3.3. 코크란-멘텔-헨젤 검정	19
3.4. 맥나마 검정	23
4. 진단의 평가	26
4.1. 민감도와 특이도	26
4.2. 양성예측도와 음성예측도	28

II. 집단의 비교	32
5. 분산분석	33
5.1. 필요한 패키지	33
5.2. 일원배치법	33
5.3. 가설	34
5.4. 변동의 분해	35
5.5. 자유도	36
5.6. 평균제곱합과 F-통계량	37
5.7. 분산분석을 이용한 F-검정	37
5.8. 분산분석 후의 추정	39
5.9. 예제: 저혈당 임상실험	40
6. 공분산분석	46
6.1. 필요한 패키지	46
6.2. 공분산분석 개요	46
6.3. 공분산분석의 모형	47
6.4. 가설검정	47
6.4.1. 최소제곱평균과 각 평균의 비교	49
6.5. 예제: 저혈당 실험	49
6.6. 예제: 산소운반능력	55
7. 다중비교	63
7.1. 필요한 패키지	63
7.2. 일원배치에서 평균의 비교	63
7.3. 두 개 이상의 가설	64
7.4. 실험단위 오류	64
7.5. 다중비교	66
7.6. 다중비교 방법	67
7.6.1. 다중비교 방법을 적용하지 않는 경우	67
7.6.2. 본페로니 수정(Bonferroni correction)	68
7.6.3. Tukey의 HSD	69
7.6.4. FDR 방법	70
7.6.5. Dunnett 비교	73
8. 반복측정자료	75
8.1. 필요한 패키지	75
8.2. 독립표본과 쌍표본	75
8.3. 일원배치 모형	78
8.3.1. 고정효과 모형	78
8.3.2. 혼합효과 모형	79
8.4. 반복측정자료	81
8.4.1. 개체들의 선형 회귀모형	83
8.4.2. 임의계수 모형	87

8.5. 예제: 고혈압 강하제 실험	97
8.5.1. 혼합효과 모형	98
8.5.2. 임의효과가 없는 모형	101
8.5.3. 공분산분석	103
III. 생존분석	106
9. 생존함수와 위험함수	107
9.1. 필요한 패키지	107
9.2. 생존분석 소개	107
9.2.1. 생존 시간의 정의	107
9.2.2. 생존 시간의 특징	108
9.3. 생존함수와 위험함수	109
9.4. 위험함수의 형태	111
9.5. 평균 생존시간과 중위 생존시간	118
9.6. 임의중도절단	118
9.6.1. 중도 절단의 이유	118
9.6.2. 임의중도절단	119
9.6.3. 최대가능도 추정	120
10. 생존함수의 추정	124
10.1. 필요한 패키지	124
10.2. 생명표 방법	124
10.2.1. 생명표 개요	124
10.2.2. 생명표 방법을 이용한 생존함수 추정	125
10.3. 비모수적 방법	131
10.3.1. 누적한계추정량	131
10.3.2. Nelson-Aalen 추정량	134
10.4. 생존함수 추정의 예제	137
10.4.1. 예제 1	137
10.4.2. 예제 2	140
Appendices	144
A. 다변량 확률변수의 성질	144
A.1. 일변량분포	144
A.2. 확률벡터의 분포	145
A.3. 다변량 정규분포	146
A.3.1. 일변량 정규분포	146
A.3.2. 이변량 정규분포	146
A.3.3. 다변량 정규분포	147

B. 선형혼합모형의 이론	148
B.1. 독립군집 모형	148
B.1.1. 하나의 군집에 대한 모형	148
B.1.2. 전체 군집에 대한 모형	150
B.2. 반응값의 분포	152
B.3. 예측	154
B.3.1. 예제: IQ 의 예측	154
B.3.2. 임의효과와의 예측	157
C. 과제 자료 화일	159
C.1. 과제 4	159
References	160

List of Figures

3.1. 2 x 2 분할표	16
3.2. 2 x 2 분할표: 관측 도수	17
3.3. K 개의 2 x 2 분할표	19
3.4. 8개 병원의 임상실험 결과	21
3.5. 짝표본 실험에 의한 2 x 2 분할표	23
3.6. 짝표본 실험에 의한 2 x 2 분할표	24
3.7. 영국시민의 수상에 대한 지지도 조사 자료	24
4.1. 코로나 검사의 민감도와 특이도	27
8.1. 생물학적동등성 실험에서 사용되는 교차실험(crossover design) - 개체 당 2개의 관측치 (각 처리에 대하여 한 개의 관측값)	76
9.1. 중도 절단 자료	109
9.2. 위험함수와 생존함수의 관계	110
10.1. 누적한계추정법(파랑선)과 Nelson-Aalen 추정법(빨간선)의 비교	137

List of Tables

2.1. 2×2 분할표	5
2.2. 코로나 치료제 실험 결과	6
2.3. 2×2 분할표 예제	8
2.4. 아스피린 임상실험 결과	10
2.5. 약물 남용 사례-대조 연구 결과	12
4.1. 진단 기법의 실험 결과	27
4.2. 코로나 바이러스 검사법의 결과	28
5.3. 저혈당 환자에 대한 임상실험 결과	41
6.1. 공분산분석 모형의 분산분석표	48
6.2. 산소운반능력 임상실험 결과	57
10.2. 생명표 계산표	128
10.3. 누적한계추정법	132
10.4. Nelson-Aalen 추정법	135

Preface

이 사이트는 2024년 의학통계학 강의를 위한 온라인 노트입니다.

개요

학부생을 위한 의학통계학 강의를 위하여 제공는 온라인 교과서입니다.

Caution

이 강의노트는 지속적으로 업데이트될 예정입니다. 따라서 노트의 내용이 중간에 변경될 수도 있으니 주의하기 바랍니다.

의학연구와 생명과학에 주로 사용되는 통계적 방법의 이론과 적용을 배우는 과목으로서 수강을 원하는 학생들은 기초통계학 또는 유사한 과목을 먼저 수강할 것을 추천합니다.

이 교과서에서는 통계 방법들의 실습을 위하여 R 프로그램을 사용합니다. R 프로그램이 익숙하지 않는 학생들은 R 프로그램에 대한 기초적인 내용을 먼저 숙지하는 것을 추천합니다.

이 교과서에서 사용한 일부 예제는 Jaewon Lee (2005) 와 Agresti (2003) 를 참조하였습니다.

이 강의에서 사용하는 R 패키지는 다음과 같다.

```
library(tidyverse)
library(ggplot2)
library(epiR)
library(faraway)
```

1. 서론

1. 서론

이 교과서는 의학연구와 생명과학에 관련된 연구에서 사용되는 다양한 통계적 방법들의 이론과 응용을 살펴보기 위한 것입니다.

이 교과서에서는 다음과 같은 주제를 다룰 것입니다.

- 연관성의 검정
- 검정 및 평가
- 분산분석과 공분산분석
- 로지스틱 회귀분석
- 반복측정자료
- 생존분석
- 임상시험연구의 개념 및 설계
- 실험 대상자수의 결정

1.1. 공정한 실험

20세기가 시작되면서 통계적 방법은 모든 과학분야의 연구에 적용되고 있습니다. 현대과학은 공정하고 합리적인 실험(또는 관측연구)에서 나오는 결론을 선호합니다. 이를 증거에 기반한 연구(evidence-based research)라고 합니다.

증거에 기반한 연구를 수행하려면 공정한 실험을 수행할 수 있는 구체적인 방법을 결정해야 합니다. 또한 실험 대상자의 수와 자료를 분석하는 방법도 적절하게 정해야 합니다. 즉 실험을 계획하는 방법이 매우 중요합니다.

그럼 공정하고 합리적인 실험은 어떤 것일까요? 이러한 질문에 대하여 실험을 수행할 때 발생할 수 있는 공정하지 않고 비합리적인 요소가 무엇인지 찾아보면 답이 나오게 됩니다.

이 과목을 수강할 때 여러분은 이러한 질문을 수시로 던지면서 생각해 보기를 강력하게 추천합니다.

1.2. 의학연구의 특성

사람을 대상으로 하는 의학연구, 즉 우리가 흔히 임상실험(clinical trial)이라고 부르는 실험들은 편향이 발생할 수 있는 여러 가지 요인을 제거하기 위하여 매우 정교하게 설계됩니다.

사실 의학연구에서 통계학자의 가장 중요한 역할은 자료의 분석이 아니라 실험의 계획과 수행에 있습니다. 물론 다양한 방법을 사용하여 유용한 결론을 이끌어 내는 탐색적인 연구에서는 주어진 자료를 분석하는 능력도 매우 중요합니다. 하지만 치료법의 효과와 안전성을 확인하는 주요한 임상실험에서는 실험을 적절하고 효과적으로 계획하는 능력이 더 중요합니다.

통상적으로 중요한 임상실험은 그 특성 상 연구를 시작하게 되면 최종 분석에 사용될 주요 변수를 바꿀수 없을 뿐만 아니라 실험의 방법과 대상자의 수도 원칙적으로 변경할 수 없습니다. 이렇게 실험을 수행할 때 강력한 제한을 두는 것은 실험에서 발생할 수 있는 여러가지 편향(bias)을 줄이기 위한 것입니다. 또한 유리한 실험만 선택하여 사용하거나 실험이 처리(treatment)외의 다른 요인에 의하여 영향을 받는 것을 차단하기 위해서입니다.

1.3. 통계적 주제

예를 들어서 최근에 유행하고 있는 코로나 19에 대한 치료제의 효과와 안전성을 확인하는 임상실험을 상상해 봅시다. 이런 임상실험을 수행하기 위해서는 실험의 시작부터 종료까지 어떤 절차를 거쳐야 하며 각 절차마다 수행해야 할 작업은 무엇일까요? 또한 실험이 종료되면 수집한 자료들은 어떻게 분석해야 할까요?

임상실험을 계획하는 경우 통계학자의 입장에서 중요하게 고려해야 할 사항은 다음과 같습니다.

- 사용할 실험 방법은 무엇인가?
- 실험에 참가하는 환자들을 어떻게 치료 집단(treatment groups)에 배정해야 하는가?
- 얼마나 많은 환자들이 실험에 참가해야 하는가?
- 실험에서 수집되는 자료들은 어떤 특성을 가지는가?
- 실험에서 수집된 자료를 어떤 방법을 적용하여 분석해야 하는가?

1.4. 강의의 구성

이 과목에서는 먼저 의학과 생명과학의 연구에서 수집된 자료를 분석할 수 있는 통계적 방법들은 배울 것입니다. 이러한 통계적 방법들은 주로 통계적 가설검정(statistical hypothesis testing)과 선형모형(linear model) 또는 회귀모형(regression model)에 의거한 방법들입니다.

두번째로 배우게 될 내용을 임상시험에서 사용되는 실험계획에 관련된 개념과 방법들입니다. 임의화(randomization)이 왜 실험에서 중요한 개념인지를 배울 것입니다. 유의수준과 검정력에 기반하여 실험 대상자의 수를 산정하는 법을 배울 것입니다. 또한 임상실험에서 사용되는 다양한 계획법에 대하여 배울 것입니다.

Part I.

분할표와 연관성

2. 연관성의 측도

2.1. 필요한 패키지

```
library(tidyverse)
library(ggplot2)
library(epiR)
library(faraway)
```

2.2. 이항변수

통계학에서 관측값은 값이 가지는 특성에 따라서 연속형 변수(continuous variable)과 범주형 변수(categorical variable)로 나눈다.

결과가 2개인 범주형 변수인 이항변수(binary variable)는 매우 중요한 역할을 한다. 그 이유는 두 개의 선택 중에서 하나를 선택해 야할 의사결정이 실제로 대부분을 차지하고 있기 때문이다.

예를 들어서 코로나 19에 감염된 환자가 병원에서 치료를 받고 있다고 가정해보자. 환자는 병원에서 여러 가지 검사를 수행하면서 다양한 자료를 수집한다. 예를 들어 환자는 수시로 체온을 재고 항체검사, 혈액검사 등을 받을 것이다. 다양한 검사 등에서 나온 자료는 연속형 또는 범주형 자료로 구성될 것이다.

하지만 의사가 가장 중요하게 결정할 사항은 환자가 계속 치료를 필요로 하는지 아닌지 결정해야 한다. 즉, 여러 가지 검사를 고려하여 최종적으로 의사는 환자가 더 치료가 필요한지 아닌 지를 결정해야 한다. 의사의 결정을 이항변수 Y 로 다음과 같이 표현할 수 있다..

$$Y = \begin{cases} 1 & \text{if patient still needs treatment} \\ 0 & \text{if patient dose not need treatment any more (GO HOME!)} \end{cases}$$

실제 임상에서는 이러한 두 개의 가능한 선택 중에 하나를 선택하는 결정이 빈번하게 일어나며 이러한 결정은 대부분 중요한 임상적 결정이다. 예를 들어 다음과 같은 의사결정들은 이항변수로 표현할 수 있다.

- 환자는 약을 복용해야 하는가?
- 환자는 입원을 해야 하는가?
- 환자는 중환자실로 가야 하는가?
- 환자는 퇴원해도 되는가?

2. 연관성의 측도

또는 환자의 상태(outcome)가 이항변수로 표현될 수 있다.

- 환자는 치료가 되었는가?
- 환자가 사망하였는가?

이제 코로나 19 치료제의 효과를 알아보기 위한 임상실험을 수행하는 경우를 생각해 보자. 통상적으로 임상실험에서는 두 개의 집단을 비교하며 가장 많이 사용하는 두 개의 집단은 실제 치료(drug)를 받은 사람들과 위약(placebo)을 받은 사람들이다. 즉 치료를 받은 사람과 받지 않는 사람들의 효과를 비교하는 것이 임상실험의 목적이다. 이러한 경우 앞에서 논의한 의사 결정과 마찬가지로 한 환자가 받은 치료의 종류를 이항변수 X 로 나타낼 수 있다.

$$X = \begin{cases} 1 & \text{if patient receives drug} \\ 0 & \text{if treatment receives placebo} \end{cases}$$

2.3. 분할표와 연관성의 측도

2.3.1. 분할표

이제 앞에서 말한 두 개의 변수 X 와 Y 의 관계에 대해서 생각해 보자. 실험에서 사람들은 코로나 19에 대한 치료약의 효과에 관심이 있다. 코로나 19 환자가 치료약을 처치 받으면 치료약을 이용하지 않는 환자보다 빨리 치료되거나 사망할 가능성이 낮은 지가 주요 관심사이다. 즉, 치료약이 환자의 회복 속도나 사망과 연관(association)이 있는지 알고 싶은 것이며, 특히 실험이 매우 정교하게 설계된 경우는 치료약이 환자의 회복이나 사망에 영향을 미치는 원인이 되는지(cause-effect relation) 파악하고 싶은 것이다.

- 먼저 코로나 19에 대한 치료약의 효과에 대한 임상실험에 n 명의 환자들이 실험에 참가 했다고 가정하자.
- 치료약이 효과가 있는지에 대한 결과(Y)는 치료를 시작하여 정해진 기간 내에 사망하였는지에 대한 사건으로 결정하였다.

$$Y = \begin{cases} 1 & \text{if patient is dead within D days} \\ 0 & \text{otherwise} \end{cases}$$

코로나 19에 대한 치료약의 효과에 대한 임상실험의 결과를 다음과 같은 분할표(contingency table)로 요약할 수 있다.

표 2.1.: 2×2 분할표

치료/결과	사망 ($Y = 1$)	생존 ($Y = 0$)	합계
위약 ($X = 0$)	n_{11}	n_{12}	n_{1+}
치료약 ($X = 1$)	n_{21}	n_{22}	n_{2+}
합계	n_{+1}	n_{+2}	n

많은 임상실험이나 의학연구의 결과들을 위와 같은 2×2 분할표로 요약할 수 있다. 이제 우리의 관심은 분할표를 통해서 임상실험의 결과를 어떻게 통계적으로 추론할 수 있는지이다.

i Note

분할표에서 연관성의 측도를 계산하는 경우 성공의 기준(이항변수로 표현하면 $Y = 1$)에 따라서 계산을 수행해야 한다. 어떤 경우는 사망이나 악화와 같은 위험한 사건이 성공 사건이 될 수 있으며 어떤 경우는 생존이나 회복과 같은 좋은 사건이 성공이 될 수 있다.

또한 기준이 되는 그룹(이항변수 X)에 따라서 연관성의 측도 계산할 때 분자와 분모에 해당하는 그룹을 적절하게 선택해야 한다.

분할표에서 연관성의 측도를 계산하는 경우 분석의 의도와 목적에 맞게 성공 사건과 기준그룹을 정의하고 그에 따라서 연관성의 측도를 계산해야 한다.

2.3.2. 상대위험

2×2 분할표 2.1 에서 두 개의 처리군, 즉 치료약을 받은 집단과 위약을 받은 집단의 효과를 비교할 때 가장 많이 사용되는 측도(measure)는 **상대위험(relative risk, risk ratio, prevalence ratio;RR)**이다.

주어진 집단의 위험율을 그 집단에 속한 환자의 수에서 사망한 사람의 비율이다. 분할표 2.1 에서 위약 집단의 위험율은 n_{11}/n_{1+} 이며 이는 치료를 받지 않는 경우에 나타나는 기준점인 위험율(baseline risk)을 의미한다. 치료약 집단의 위험율은 n_{21}/n_{2+} 이다. 통상적으로 위험율은 비율(proportion, percent)로 나타내며 발생률(rate, 예를 들어 인구 1000명당 X명)로 나타내기도 한다.

상대위험은 두 위험율의 비율로서 다음과 같이 정의한다.

$$RR = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}} = \frac{\frac{n_{11}}{n_{11}+n_{12}}}{\frac{n_{21}}{n_{21}+n_{22}}} \quad (2.1)$$

상대위험이 1보다 크면 분자에 위치한 집단이 위험(위의 예제에서는 위험이 사망을 의미한다)에 처할 가능성이 분모에 위치한 집단보다 RR 배 높다는 것을 의미한다. 상대위험이 1이면 두 집단에 대한 위험이 동일하다는 것을 의미한다.

예를 들어 특정한 코로나 치료제의 효과를 실험하는 임상실험에서 다음과 결과를 얻었다.

표 2.2.: 코로나 치료제 실험 결과

치료/결과	사망 ($Y = 1$)	생존 ($Y = 0$)	합계
위약 ($X = 0$)	10	1212	1222
치료약 ($X = 1$)	5	2355	2360
합계	15	3567	3582

상대위험은 다음과 같이 계산된다.

$$RR = \frac{10/1222}{5/2360} = 3.8625 \approx 4$$

상대위험이 약 4 배란 의미는 치료약을 받은 집단보다 위약집단이 사망할 가능성이 약 4배 높다는 것이다.

i Note

우리는 두 집단의 비율을 비교할 때 두 비율의 차이를 이용하는 방법을 자주 사용한다. 두 집단의 비율이 각각 p_1, p_2 라면 두 비율의 차이는 $p_1 - p_2$ 이며 이는 우리가 평상 적으로 사용하는 비율의 비교 측도이다. 예를 들어 대통령 후보들의 지지율과 차이는 많은 언론에서 사용하고 있으며 기초 통계학에서 두 모집단의 비교를 위한 가설 검정에서도 비율의 차이를 이용하였다. 위의 코로나 치료제의 효과를 비교하는 실험에서 치료집단과 위약집단의 사망률 차이를 측도로 사용하면 어떨까?

2.3.3. 기여위험과 백신효과

기여위험(attributable proportion, attributable risk percent, AR)은 두 그룹의 위험에 대한 비교를 위한 다른 측도이다. 기여위험은 특정한 성격을 가진 집단(exposed group)이 위험에 처한 전체 집단에서 차지하는 비율을 백분율로 나타낸다.

$$AR = \frac{(n_{11}/n_{1+}) - (n_{21}/n_{2+})}{n_{11}/n_{1+}} \times 100 \quad (2.2)$$

예를 들어 비흡연자(unexposed group)와 흡연자(exposed group)의 폐암에 대한 위험을 비교하는 경우를 생각해 보자. 비흡연자의 폐암으로 인한 사망률이 연간 1000명 당 0.07명이고 흡연자는 1000명당 0.57명이라고 하면 일단 상대위험은 약 8배이다.

$$RR = 0.57/0.07 = 8.1428$$

두 집단의 비교를 기여위험으로 나타내면 다음과 같다.

$$AR = \frac{0.57 - 0.07}{0.57} \times (100) = 87.7\% \approx 88\%$$

만약 흡연이 폐암을 일으키는 원인이고 두 집단의 다른 요인이 유사하다고 가정하면, 기여위험이 약 88% 라는 것은 모든 폐암 환자(위험에 처한 전체 집단)의 88% 가 흡연에 의한 것이라고 해석할 수 있다.

최근에 코로나 19에 대한 백신과 치료제의 임상실험에서 효과를 발표하는 경우 위에서 언급한 상대위험을 사용하지 않고 백신효과(Vaccine efficacy, vaccine effectiveness; VE) 라는 백분율을 사용한다. 백신효과는 기본적으로 기여위험과 동일한 측도이다.

예를 들어 위의 예제에서 치료제의 효과를 백신효과(VE)로 계산하면 다음과 같다.

$$VE = \left[\frac{10/1222 - 5/2360}{10/1222} \right] \times 100 = 74.1101\%$$

백신효과가 74% 란 의미는 치료제를 사용하면 사용하지 않는 경우보다 사망을 74% 줄일 수 있다고 해석할 수 있다.

2. 연관성의 측도

간단한 예로서 코로나19로 인한 치명율(사망자/확진자)을 비교한다고 가정하자. 백신을 맞은 그룹의 치명율이 1% 이고 백신을 맞지 않는 그룹의 치명율이 2% 백신효과는 50%이다.

2.3.4. 오즈비

오드(odd)는 가능성을 나타내는 측도로서 전통적으로 도박에서 유래된 측도이다.

우리가 주사위를 던져서 1과 2가 나오면 성공, 다른 숫자가 나오면 실패라고 하는 경우 성공의 확률은 $2/6 = 0.3333$ 으로 계산한다. 확률을 계산하는 경우는 분모에 전체 사건의 수를 사용한다.

위의 주사위 예제로 오드를 계산하면 $2/4 = 0.5$ 가 된다. 즉, 오드는 분모에 성공을 제외한 실패의 사건을 수를 사용한다. 만약 오드가 1이면 무슨 의미인가? 오드가 1이면 성공하는 사건의 수가 실패하는 사건의 수가 동일하다는 의미이다. 게임에서 이길 확률이 $1/2$ 이면 공정한 게임이며 이 경우 오드는 1 이다.

전통적으로 오드는 확률의 개념이 나오기 전에 가능성의 측도로 오랫동안 사용되어 왔으며 도박에서 상대방이 1번 이길 때 내가 이기는 평균적인 횟수를 의미한다.

$$odd = \frac{\text{number of events for success}}{\text{number of events for failure}}$$

예를 들어 위의 코로나 치료제 실험에서 성공을 사망할 사건이라고 하면 위약군의 오드는 $n_{11}/n_{12} = 10/1212$ 이고 치료군의 오드는 $n_{21}/n_{22} = 5/2355$ 이다.

두 집단을 비교하는 측도 중 하나는 **오즈비(odds ratio; OR)**가 있다. 오즈비는 두 그룹의 오드들의 비율로 정의된다. 오즈비가 1이면 두 그룹에서 성공 사건의 가능성이 같다는 것이다.

$$OR = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

코로나 치료제 실험에서의 오즈비는 $(10/1212)/(5/2355) = 3.8861$ 이다.

오즈비는 상대위험이나 기여위험에 비하여 의미 있는 해석이 어렵다. 오즈비가 1이면 두 집단이 성공의 가능성이 같다(또는 두 요인의 연관성이 없다)는 것으로 해석이 쉽다. 하지만 예를 들어 오즈비가 1 보다 큰 경우(또는 작은 경우) 두 집단의 차이를 의미 있게 해석하는 것이 어렵다.

오즈비는 향후 학습할 통계적 가설검정에서 중요한 모수(parameter)로 사용되며 특히 실험의 방법이 사례-대조 연구와 같은 특별한 방법을 사용하는 경우 오즈비가 중요한 역할을 하게 된다.

예를 들어 다음과 같은 분할표에서 비율의 차이, 상대위험, 오즈비를 구하여 비교해 보자.

표 2.3.: 2×2 분할표 예제			
처리 / 결과	성공 ($Y = 1$)	실패 ($Y = 0$)	합계
0 ($X = 0$)	6	4	10

2. 연관성의 측도

처리 /결과	성공 ($Y = 1$)	실패 ($Y = 0$)	합계
1 ($X = 1$)	4	6	10
합계	10	10	20

비율의 차이(DP)은 다음과 같이 계산된다.

$$DP(0/1) = 6/10 - 4/10 = 0.2$$

상대위험은 다음과 같이 계산된다.

$$RR(0/1) = \frac{6/10}{4/10} = \frac{6}{4} = 1.5$$

오즈비는 다음과 같이 계산된다.

$$OR(0/1) = \frac{6/4}{4/6} = \frac{(6)(6)}{(4)(4)} = 2.25$$

2.4. 신뢰구간

상대위험과 오즈비는 분할표에서 연관성을 나타내는 하나의 측도, 즉 점추정량(point estimation)이다. 하나의 숫자로 표현되는 점추정은 표본으로 부터 발생한 불확실성을 반영하지 못한다. 따라서 점추정량을 보완하기 위하여 신뢰구간(confidence interval)을 제시할 수 있다.

상대위험과 오즈비는 표본비율 또는 셀 도수의 함수로 나타난다. 하지만 함수의 형태가 비율로서 비선형이기 때문에 상대위험과 오즈비의 근사적인 표준오차(standard error)는 쉽게 구할 수 없다.

다항분포를 가정하고 로그 오즈비의 점근적 분산을 다음과 같이 유도할 수 있다.

$$v_1 = V(\log OR) \approx \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

따라서 로그 오즈비의 $100(1 - \alpha) \%$ 근사 신뢰구간을 다음과 같이 구할 수 있다.

$$\log OR \pm z_{\alpha/2} \sqrt{v_1}$$

위의 신뢰구간을 오즈비로 역변환하면 오즈비 OR 의 $100(1 - \alpha) \%$ 근사 신뢰구간을 다음과 같다.

$$(OR \times \exp[-z_{\alpha/2} \sqrt{v_1}], OR \times \exp[z_{\alpha/2} \sqrt{v_1}]) \quad (2.3)$$

2. 연관성의 측도

상대위험(RR)의 신뢰구간도 오즈비의 신뢰구간을 유도하는 방법과 유사하게 델타 방법을 사용하며 다음과 같이 구할 수 있다.

$$(RR \times \exp[-z_{\alpha/2}\sqrt{v_2}], OR \times \exp[z_{\alpha/2}\sqrt{v_2}]) \quad (2.4)$$

위의 식 2.4 에서 v_2 는 다음과 같이 계산한다.

$$v_2 = V(\log RR) \approx \frac{1 - n_{11}/n_{1+}}{n_{11}} + \frac{1 - n_{21}/n_{2+}}{n_{21}}$$

2.4.1. 예제: 아스피린 임상실험

소량의 아스피린 복용이 심장병으로 인한 위험을 줄이는데 효과가 있는지 알아보려고 임상실험을 실시하였다. 22,701명의 남성을 임의화(randomization) 을 통해서 두 그룹으로 나눈 후, 한 그룹은 매일 일정량의 아스피린을 복용시키고 다른 그룹은 위약(placebo)를 복용하게 한 후 약 5년간 심근경색이 일어나는지 알아보았다. 임상실험의 결과는 아래 표와 같다.

	표 2.4.: 아스피린 임상실험 결과		
	심근경색 발생	심근경색 없음	합
아스피린	139	10,898	11,037
위약	239	10,795	11,034

위약 집단과 아스피린 집단의 상대위험은 다음과 같다.

$$RR = \frac{139/11037}{239/11034} = 0.581$$

상대위험을 보면 1보다 작으므로 아스피린을 복용한 집단이 위약 집단에 비해서 심근 경색이 일어날 위험이 적어진다는 것을 알 수 있다.

상대위험의 95% 근사 신뢰구간은 다음과 같이 계산한다.

먼저 다음 v_2 를 계산하면

$$v_2 = \frac{1 - n_{11}/n_{1+}}{n_{11}} + \frac{1 - n_{21}/n_{2+}}{n_{21}} = \frac{1 - 139/11037}{139} + \frac{1 - 239/11034}{239} = 0.011$$

상대위험의 신뢰구간은 다음과 같다.

$$(0.581 \times \exp[-1.96\sqrt{0.011}], 0.581 \times \exp[1.96\sqrt{0.011}]) = (0.473, 0.715)$$

2. 연관성의 측도

위의 신뢰구간은 1을 포함하지 않으므로 상대위험이 1 과 유의한 차이가 있다고 할 수 있다. 결론적으로 아스피린의 복용은 심근경색의 발생을 감소시킨다고 할 수 있다.

이제 `epiR` 패키지를 사용하여 위에서 분석한 내용을 다시 구해보자.

먼저 위의 임상실험 자료를 R 의 `matrix` 형태로 저장한다.

```
ex1dat <- matrix( c(139, 10898, 239, 10795), 2, 2, byrow=TRUE)
ex1dat
```

```
      [,1] [,2]
[1,]  139 10898
[2,]  239 10795
```

이제 함수 `epi.2by2`를 이용하여 상대위험과 상대구간을 구해보자. 임의화를 사용한 임상실험 자료인 경우 `method = "cross.sectional"` 으로 지정한다. 관심이 있는 사건(심근경색, outcome)의 도수가 첫 번째 열(column)에 있으니 `outcome = "as.columns"`이라고 지정한다.

아래 결과에 `Prevalence ratio`라고 나오는 것이 상대위험이다.

```
epi.2by2(dat = ex1dat, method = "cross.sectional", conf.level = 0.95, units = 100,
  interpret = FALSE, outcome = "as.columns")
```

	Outcome +	Outcome -	Total	Prev risk *
Exposed +	139	10898	11037	1.26 (1.06 to 1.49)
Exposed -	239	10795	11034	2.17 (1.90 to 2.46)
Total	378	21693	22071	1.71 (1.55 to 1.89)

Point estimates and 95% CIs:

```
-----
Prev risk ratio                0.58 (0.47, 0.72)
Prev odds ratio                0.58 (0.47, 0.71)
Attrib prev in the exposed *   -0.91 (-1.25, -0.56)
Attrib fraction in the exposed (%) -71.99 (-111.63, -39.78)
Attrib prev in the population * -0.45 (-0.77, -0.13)
Attrib fraction in the population (%) -26.47 (-36.51, -17.18)
-----
```

Uncorrected chi2 test that OR = 1: `chi2(1) = 26.944 Pr>chi2 = <0.001`

Fisher exact test that OR = 1: `Pr>chi2 = <0.001`

Wald confidence limits

CI: confidence interval

* Outcomes per 100 population units

2.5. 사례-대조 연구

심장발작을 일으킨 환자와 그렇지 않은 사람들을 각각 214명씩 조사하여 과거에 약물남용을 한 경력이 있는지 조사한 사례-대조 연구의 자료이다.

표 2.5.: 약물 남용 사례-대조 연구 결과

	심장 발작 발생	심장발작 없음
약물남용 유	73	18
약물남용 무	141	196
합	214	214

이 연구의 목표는 약물남용과 심장발작의 연관성이 있는지를 알아보는 것이다. 이제 다음과 같은 사건들을 정의해 보자.

- $H+$: 심장발작이 발생했다.
- $H-$: 심장발작이 발생하지 않았다.
- $D+$: 약물남용을 했다.
- $D-$: 약물남용을 하지 않았다.

위에서 정의된 사건들을 고려할 때 사례-대조 연구의 자료에서 다음과 같은 조건부 확률에 대한 추정값을 구할 수 있다.

$$P(\text{약물남용을 했다}|\text{심장발작이 발생했다}) = P(D+|H+) = \frac{73}{214}$$

$$P(\text{약물남용을 하지 않았다}|\text{심장발작이 발생했다}) = P(D-|H+) = 1 - P(D+|H+) = \frac{141}{214}$$

$$P(\text{약물남용을 했다}|\text{심장발작이 발생하지 않았다}) = P(D+|H-) = \frac{18}{214}$$

$$P(\text{약물남용을 하지 않았다}|\text{심장발작이 발생하지 않았다}) = P(D-|H-) = 1 - P(D+|H-) = \frac{196}{214}$$

2.5.1. 사례대조 연구의 목표와 가설

연구에서 비교하고 싶은 비율은 위에서 추정한 확률이 아니고 조건과 결과가 바뀐 다음과 같은 조건부 확률이다.

$$P(\text{심장발작이 발생했다}|\text{약물남용을 했다}) = P(H+|D+)$$

$$P(\text{심장발작이 발생했다}|\text{약물남용을 하지 않았다}) = P(H+|D-)$$

즉 연구의 목표는 다음과 같은 가설을 검정하는 것이다.

2. 연관성의 측도

$$H_0 : P(H+|D+) = P(H+|D-) \quad \text{vs} \quad H_1 : P(H+|D+) \neq P(H+|D-) \quad (2.5)$$

전체 모집단을 약물남용을 한 사람들과 하지 않은 사람들로 두 집단으로 나누었을 때 두 집단에 대한 심장발작의 확률이 같은지 다른지 비교하고 싶은 것이다.

위의 식에서 보듯이 추정하고 싶은 확률인 $P(H+|D+)$ 와 $P(H+|D-)$ 를 추정하려면 전체 모집단에 대한 심장발작 발병률 $P(H+)$ 와 약물남용의 비율 $P(D+)$ 를 알아야 한다. 즉

$$\begin{aligned} P(H+|D+) &= \frac{P(H+ \cap D+)}{P(D+)} \\ &= \frac{P(D+|H+)P(H+)}{P(D+)} \\ &\approx (73/214) \frac{P(H+)}{P(D+)} \end{aligned}$$

위의 식은 다음의 조건부 확률 공식을 각 단계마다 적용한 결과이다.

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

사례-대조 연구의 자료만으로는 모집단에 대한 심장발작 발병률 $P(H+)$ 와 약물남용의 비율 $P(D+)$ 을 구할 수 없다. 또한 다른 외부의 자료가 있다 하더라도 약물남용의 비율을 정확하게 추정하는 것은 매우 어렵다.

2.5.2. 오즈비의 비교

이러한 문제는 두 집단의 비율의 차이나 상대위험을 비교하지 않고 오즈비를 구하여 비교하면 심장발작 발병률과 약물남용의 비율을 추정하지 않고 사례-대조 연구의 자료만으로 추론이 가능하다.

다음의 가설은 두 비율의 비교를 오즈비로 표현한 것이다.

$$H_0 : \frac{P(H+|D+)/P(H-|D+)}{P(H+|D-)/P(H-|D-)} = 1 \quad \text{vs} \quad H_1 : \frac{P(H+|D+)/P(H-|D+)}{P(H+|D-)/P(H-|D-)} \neq 1 \quad (2.6)$$

위의 가설 2.6 는 단순한 비율을 비교하는 가설 2.5 과 동일한 가설이다.

가설 2.6 에서 나타는 오즈비는 심장발작 발병률과 약물남용의 비율을 이용하지 않고 사례-대조 연구에서 추정할 수 있는 조건부 확률만으로 추정할 수 있다.

2. 연관성의 측도

$$\begin{aligned}
 \frac{P(H+|D+)/P(H-|D+)}{P(H+|D-)/P(H-|D-)} &= \frac{[P(H+|D+)P(D+)]/[P(H-|D+)P(D+)]}{[P(H+|D-)P(D-)]/[P(H-|D-)P(D-)]} \\
 &= \frac{P(H+ \cap D+)/P(H- \cap D+)}{P(H+ \cap D-)/P(H- \cap D-)} \\
 &= \frac{[P(D+|H+)P(H+)]/[P(D+|H-)P(H-)]}{[P(D-|H+)P(H+)]/[P(D-|H-)P(H-)]} \\
 &= \frac{P(D+|H+)/P(D+|H-)}{P(D-|H+)/P(D-|H-)} \\
 &= \frac{(73/214)/(142/214)}{(18/214)/(196/214)} \\
 &= \frac{(73)(196)}{(141)(18)} \\
 &= 5.64
 \end{aligned}$$

결론적으로 사례-대조 연구에서는 연구의 목표에 대한 가설 검정을 비율의 차이나 상대위험으로 표현하여 수행할 수 없다. 하지만 오즈비를 검정하는 것으로 가설을 세우면 자료에서 쉽게 유도할 수 있는 오즈비로 가설 검정을 쉽게 수행할 수 있다.

2.5.3. 예제: 약물남용 사례-대조 연구

심장발작을 일으킨 환자와 그렇지 않은 사람들을 각각 214명씩 조사하여 과거에 약물남용을 한 경력이 있는지 조사한 사례-대조 연구(case-control study)의 결과가 표 2.5에 있다.

사례-대조 연구는 사례(case)가 발견되면, 즉 위의 연구와 같이 심장발작이 일어난 환자가 발생하면 그 환자와 유사한 나이와 성별 등을 가진 일반사람을 찾아 매칭하여 환자와 일반인의 과거 경력을 조사하는 후향적인 연구(retrospective study)이다. 반대로 앞의 예제에서 본 임의화를 이용한 임상실험은 전향적 연구(prospective study)이다.

이러한 사례-대조 연구에서는 상대위험을 이용하여 연관성을 알아낼 수 없다. 하지만 사례-대조 연구에서 상대위험 대신 오즈비를 이용하여 연관성을 추론할 수 있다.

위의 심장발작에 대한 사례-대조 연구의 결과에서 오즈비와 그 신뢰구간을 구해보자.

먼저 오즈비는 다음과 같다.

$$OR = \frac{(73)(196)}{(18)(141)} = 5.64$$

위의 결과는 심장발작이 일어난 집단에서 약물남용을 한 환자들의 오즈가 심장발작이 일어나지 않은 집단에서 약물남용을 한 사람들의 오즈에 비해 5.6배 크다는 것을 알 수 있으며 이는 1보다 상당히 크다.

오즈비의 95% 근사 신뢰구간은 다음과 같이 계산한다.

먼저 다음 v_1 를 계산하면

2. 연관성의 측도

$$v_1 = V(\log OR) \approx \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} = \frac{1}{73} + \frac{1}{18} + \frac{1}{141} + \frac{1}{196} = 0.08$$

오즈비의 신뢰구간은 다음과 같다.

$$(5.64 \times \exp[-1.96\sqrt{0.08}], 5.64 \times \exp[1.96\sqrt{0.08}]) = (3.222, 9.863)$$

위의 신뢰구간을 보면 1을 포함하지 않으므로 약물남용이 심장발작의 위험을 높인다고 말할 수 있다.

이제 `epiR` 패키지를 사용하여 위에서 분석한 내용을 다시 구해보자.

먼저 위의 사례-대조 연구 자료를 R의 `matrix` 형태로 저장한다.

```
ex2dat <- matrix( c(73,18,141,196), 2, 2, byrow=TRUE)
ex2dat
```

```
      [,1] [,2]
[1,]   73   18
[2,]  141  196
```

이제 함수 `epi.2by2`를 이용하여 오즈비와 신뢰구간을 구해보자. 사례-대조 연구의 자료인 경우 `method = "case.control"`으로 저장한다. 사례-대조 연구로 지정하면 상대위험이 출력되지 않는다. 관심이 있는 사건(심장발작, outcome)의 도수가 첫 번째 열(column)에 있으니 `outcome = "as.columns"`이라고 지정한다.

```
epi.2by2(dat = ex2dat, method = "case.control", conf.level = 0.95, units = 100,
  interpret = FALSE, outcome = "as.columns")
```

	Outcome +	Outcome -	Total	Odds
Exposed +	73	18	91	4.06 (2.50 to 7.27)
Exposed -	141	196	337	0.72 (0.57 to 0.89)
Total	214	214	428	1.00 (0.83 to 1.21)

Point estimates and 95% CIs:

```
-----
Exposure odds ratio                    5.64 (3.22, 9.86)
Attrib fraction (est) in the exposed (%) 82.19 (68.26, 90.44)
Attrib fraction (est) in the population (%) 28.06 (20.13, 35.21)
-----
```

Uncorrected chi2 test that OR = 1: `chi2(1) = 42.218 Pr>chi2 = <0.001`

Fisher exact test that OR = 1: `Pr>chi2 = <0.001`

Wald confidence limits

CI: confidence interval

3. 연관성의 검정

이 절에서는 두 변수의 연관성에 통계적 가설 검정 방법을 살펴보자.

3.1. 필요한 패키지

```
library(tidyverse)
library(ggplot2)
library(epiR)
library(faraway)
```

3.2. 카이제곱 검정

일단 2개의 이항변수 X 와 Y 를 고려하고 가능한 결과의 조합과 그 확률은 다음과 같은 2×2 분할표로 나타낼 수 있다.

		Y		
		0	1	Total
X	0	p_{11}	p_{12}	p_{1+}
	1	p_{21}	p_{22}	p_{2+}
Total		p_{+1}	p_{+2}	1

그림 3.1.: 2×2 분할표

일반적으로 2×2 분할표에서 다음과 같은 두 가지 가설이 가능하다.

- 동질성 검정(homogeneity test)

변수 X 가 단순히 독립 집단을 나누는 변수인 경우 (예를 들어 실험약 집단과 위약 집단) 두 그룹 간에 이항변수 Y 의 성공확률이 같은지 검정하는 경우이다. 실험약 집단과 위약 집단에서 심장병이 발병할 확률이 같은지 검정을 수행할 때 귀무가설은 다음과 같다.

3. 연관성의 검정

$$H_0 : p_{1j} = p_{2j} = p_j$$

- 독립성 검정(independent test)

변수 X 와 Y 가 모두 확률변수인 경우 두 변수가 독립인지 검정하는 경우이다. 예를 들어 흡연(X)과 심근경색(Y)의 관계를 연구하는 경우 두 사건이 모두 확률적인 사건이라고 보고 다음과 같이 독립에 대한 가설을 고려한다.

$$H_0 : p_{ij} = p_{i+}p_{+j}$$

다음과 같이 n 개의 관측값으로 구성된 2×2 분할표에서 동질성과 독립성 가설을 검정하는 방법은 동일하며 따라서 굳이 두 가지 가설을 엄격하게 구별할 이유는 없다. 만약 귀무가설이 기각되면 두 변수의 연관성은 유의하다고 결론을 내린다.

		Y		
		0	1	Total
X	0	n_{11}	n_{12}	n_{1+}
	1	n_{21}	n_{22}	n_{2+}
Total		n_{+1}	n_{+2}	n

그림 3.2.: 2 x 2 분할표: 관측 도수

동질성과 독립성에 대한 검정은 다음과 같은 카이제곱 통계량을 사용한다.

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (3.1)$$

위의 카이제곱 통계량에서 $O_{ij} = n_{ij}$ 는 각 셀의 관측도수이며 E_{ij} 는 귀무가설 하에서의 셀 도수의 예측값이다. 동질성 검정을 고려할 때 만약 귀무가설이 참이라면 확률 $p_{1j} = p_{2j} = p_j$ 는 다음과 같이 추정할 수 있다.

$$\hat{p}_j = \frac{n_{+j}}{n}$$

따라서 셀 (i, j) 에 대한 기대 도수 E_{ij} 는 다음과 같이 계산된다.

$$E_{ij} = n_{i+}\hat{p}_j = \frac{n_{i+}n_{+j}}{n} \quad (3.2)$$

3. 연관성의 검정

귀무가설 하에서 표본의 크기가 충분히 크면 식 3.1 의 카이제곱 검정통계량 χ^2 는 자유도가 1인 카이제곱 분포를 따른다. 그러므로 이 사실을 이용하여 p-값을 계산하거나 기각역을 구하여 검정한다.

일반적인 $I \times J$ 분할표도 동일한 방법으로 가설검정을 할 수 있다. 카이제곱 통계량을 구하는 방법은 2×2 분할표와 유사하다. 다만 귀무가설이 참인 경우 검정통계량은 자유도가 $(I - 1)(J - 1)$ 인 카이제곱 분포를 따른다.

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

이제 실제 분할표에서 카이제곱 검정을 수행해 보자. 아스피린 임상실험 결과가 주어진 표 2.4 에서 아스피린의 효과사 없는 경우, 즉 귀무가설이 참인 경우 다음과 같이 심근경색의 유무에 대한 예측 확률을 구할 수 있다.

$$\hat{p}_1 = \frac{n_{+1}}{n} = \frac{139 + 239}{22071} = 0.0171$$

$$\hat{p}_2 = \frac{n_{+2}}{n} = \frac{10898 + 10795}{22071} = 0.9829$$

이제 각 셀의 기대도수를 식 3.2 에 의하여 계산할 수 있다. 예를 들어 E_{11} 은 다음과 같이 계산된다.

$$E_{11} = \frac{n_{1+}n_{+1}}{n} = n_{1+}\hat{p}_1 = (11037)(0.0171) = 189.03$$

각 셀에 대한 기대도수 E_{ij} 를 구하고 식 3.1 의 카이제곱 통계량을 구하면 다음과 같다.

$$\begin{aligned} \chi^2 &= \frac{(139 - 189.03)^2}{189.03} + \frac{(10898 - 10848.00)^2}{10848.00} \\ &\quad + \frac{(239 - 188.97)^2}{188.97} + \frac{(10795 - 10845.03)^2}{10845.03} \\ &= 26.94 \end{aligned}$$

자유도가 1인 카이제곱 분포의 상위 5% 백분위수 3.84 이다. 위에서 구한 카이제곱 통계량의 값이 26.94 로서 3.84 보다 크므로 귀무가설을 기각한다. 즉 아스피린과 위약을 복용한 두 그룹 사이에는 심근경색이 일어날 비율에 유의한 차이가 있다.

R 에서도 카이제곱 검정을 쉽게 수행할 수 있다. 앞에서 표 2.4 의 자료를 행렬의 형태로 저장하였는데 함수 `chisq.test()` 를 사용하면 결과를 쉽게 구할 수 있다.

```
ex1dat <- matrix( c(139, 10898, 239, 10795), 2, 2, byrow=TRUE)
ex1dat
```

```
      [,1] [,2]
[1,]  139 10898
[2,]  239 10795
```

3. 연관성의 검정

```
chisq.test(ex1dat)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: ex1dat
```

```
X-squared = 26.408, df = 1, p-value = 2.764e-07
```

분할표에서의 기대도수 E_{ij} 는 다음과 같이 얻을 수 있다.

```
chisq.test(ex1dat)$expected
```

```
      [,1]      [,2]
[1,] 189.0257 10847.97
[2,] 188.9743 10845.03
```

3.3. 코크란-맨텔-헨젤 검정

임상실험이나 의학연구는 여러 나라 또는 여러 병원들에서 진행되는 경우가 있다. 이러한 경우 국가나 병원의 고유한 특성에 따라서 실험의 결과가 다르게 나타날 수 있다. 이렇게 그룹에 의한 효과를 그룹 효과 또는 층(strata)에 의한 효과라고 한다. 예를 들어 진통제에 대한 효과는 그 나라의 문화나 관습에 따라서 효과의 차이가 나타날 수 있다. 또한 여러 개의 변원에서 연그루가 동시에 진행된다면 병원의 규모, 위치, 환자들의 특성에 따라서 치료 효과의 차이가 나타날 수 있다.

이렇게 그룹에 따른 차이가 예상되는 경우 그룹의 효과를 제어하면서 처리 효과의 차이를 검정하는 방법이 필요하다. 이렇게 여러 개의 층으로 구성된 독립집단에서 얻은 자료에서 층에 의한 효과를 통제하면서 동질성 또는 독립성 검정을 수행하는 방법을 코크란-맨텔-헨젤 검정 (Cochran-Mantel-Haenzel test)라고 한다.

아래와 같이 K 개의 독립집단(또는 층)에서 각각 얻은 K 개의 2×2 분할표가 있다고 하자.

		Y		
		0	1	Total
X	0	n_{k11}	n_{k12}	n_{k1+}
	1	n_{k21}	n_{k22}	n_{k2+}
Total		n_{k+1}	n_{k+2}	n_k

그림 3.3.: K 개의 2×2 분할표

3. 연관성의 검정

K 개의 독립집단이 있고 성공의 확률이 p_1 , 실패의 확률이 p_2 라고 한다면 처리의 효과를 전체적으로 비교하는 가설은 다음과 같다.

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 \neq p_2$$

이제 귀무가설의 가정 하에서 각 분할표에서 n_{k11} 에 대한 기대도수 μ_{k11} 와 그 분산 v_{k11} 을 다음과 같이 계산한다.

$$\mu_{k11} = E(n_{k11}|H_0) = \frac{n_{k1+}n_{+1}}{n_k}$$

$$v_{k11} = V(n_{k11}|H_0) = \frac{n_{k1+}n_{k2+}n_{+1}n_{+2}}{n_k^2(n_k - 1)}$$

이제 가설검정을 위한 통계량 Q_{CMH} 은 다음과 같다.

$$Q_{CMH} = \frac{\left[\sum_{k=1}^K (n_{k11} - \mu_{k11}) \right]^2}{\sum_{k=1}^K v_{k11}} \quad (3.3)$$

귀무가설이 참인 경우 검정통계량 Q_{CMH} 은 자유도가 1 인 카이제곱 분포를 따른다.

이제 Agresti (2012) 의 6.3절에 있는 다기관 임상시험(**multi-center clinical trial**) 의 예제를 살펴보자. 아래 표는 모두 8개의 독립적인 병원에서 감염 치료제에 대한 효과에 대한 실험을 실시하여 얻은 자료이다.

마지막 병원을 제외한 7개의 병원에서 치료제의 효과가 긍정적으로 나타났다. 여기서 주목할 점은 병원에 따라서 연관성의 강도가 매우 다르게 나타날 수 있다는 것이다.

이제 각 병원을 층(strata)로 고려하고 병원의 효과를 제어하면서 식 3.3 의 검정 통계량 Q_{CMH} 를 이용하여 치료제의 효과가 있는지 검정해보자. 검정은 아래와 같이 R 프로그램을 이용한다. 함수 `mantelhaen.test()` 는 코크란-맨텔-헨젤 검정을 수행하는 함수이다.

```
beitler <- c(11,10,25,27,16,22,4,10,14,7,5,12,2,1,14,16,6,0,11,12,1,0,10,10,1,1,4,8,4,6,2,1)
beitler <- array(beitler, dim=c(2,2,8))
beitler
```

```
, , 1
```

```
 [,1] [,2]
```

```
[1,]   11   25
```

```
[2,]   10   27
```

```
, , 2
```

TABLE 6.9 Clinical Trial Relating Treatment to Response for Eight Centers

Center	Treatment	Response		Odds Ratio	μ_{11k}	$\text{var}(n_{11k})$
		Success	Failure			
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

그림 3.4.: 8개 병원의 임상실험 결과

3. 연관성의 검정

	[,1]	[,2]
[1,]	16	4
[2,]	22	10

, , 3

	[,1]	[,2]
[1,]	14	5
[2,]	7	12

, , 4

	[,1]	[,2]
[1,]	2	14
[2,]	1	16

, , 5

	[,1]	[,2]
[1,]	6	11
[2,]	0	12

, , 6

	[,1]	[,2]
[1,]	1	10
[2,]	0	10

, , 7

	[,1]	[,2]
[1,]	1	4
[2,]	1	8

, , 8

	[,1]	[,2]
[1,]	4	2
[2,]	6	1

```
mantelhaen.test(beitler, correct=FALSE)
```


3. 연관성의 검정

Mantel-Haenszel chi-squared test without continuity correction

```
data: beittler
Mantel-Haenszel X-squared = 6.3841, df = 1, p-value = 0.01151
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.177590 3.869174
sample estimates:
common odds ratio
      2.134549
```

검정 통계량 Q_{CMH} 의 값이 6.3841 이고 p-값은 0.0115 이므로 귀무가설을 기각한다.

3.4. 맥나마 검정

연속형 변수에서 짝지은 자료를 비교할 때 사용하는 방법이 대응 t-검정(paired t-test) 또는 짝표본 t-검정이다. 예를 들어 천식환자가 A약을 먹고 폐활량을 측정하고 일정 기간이 지나서 같은 환자가 B약을 먹고 폐활량을 측정하면 두 관측값은 독립이 아니다. 따라서 이러한 경우 독립 t-검정이 아닌 대응 t-검정을 사용한다.

이제 이산형 변수가 짝으로 나타나는 경우를 생각해보자. 예를 들어 눈병 치료에 사용되는 A약과 B약의 효과를 비교하기 위하여 각각의 약을 환자의 오른쪽 눈과 왼쪽 눈에 처치를 하고 치료의 여부를 관측하였다고 하자.

		Right eye		
		cured	not cured	Total
Left eye	cured	n_{11}	n_{12}	n_{1+}
	not cured	n_{21}	n_{22}	n_{2+}
Total		n_{+1}	n_{+2}	n

그림 3.5.: 짝표본 실험에 의한 2 x 2 분할표

위의 표에서 n_{11} 은 A약과 B약의 효과가 모두 나타난 환자의 도수이다. n_{12} 은 A약은 효과가 있고 B약은 효과가 없는 환자의 도수이다. 이러한 자료는 앞에서 배운 카이제곱 검정을 적용할 수 없다.

이제 일반적으로 짝표본에서 나온 자료가 다음 표와 같이 얻어졌다고 가정하자.

이제 조건 1 에서 성공의 확률을 p_1 이라고 하고 조건 2에서 성공의 확률을 p_2 라고 하면 짝표본에서 얻어진 분할표 그림 3.6 에서 관심있는 가설은 다음과 같다.

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 \neq p_2$$

분할표 그림 3.6 에서 p_1 과 p_2 의 추정량은 다음과 같다.

3. 연관성의 검정

		조건 2		
		예	아니오	Total
조건 1	예	n_{11}	n_{12}	n_{1+}
	아니오	n_{21}	n_{22}	n_{2+}
Total		n_{+1}	n_{+2}	n

그림 3.6.: 짝표본 실험에 의한 2 x 2 분할표

$$\hat{p}_1 = \frac{n_{1+}}{n}, \quad \hat{p}_2 = \frac{n_{+1}}{n}$$

p_1 과 p_2 의 추정량의 차이는 두 조건에 따른 결과가 일치하지 않는 도수 n_{12} 와 n_{21} 의 차이에 의존한다.

$$\hat{p}_1 - \hat{p}_2 = \frac{n_{1+}}{n} - \frac{n_{+1}}{n} = \frac{n_{11} + n_{12}}{n} - \frac{n_{11} + n_{21}}{n} = \frac{n_{12} - n_{21}}{n}$$

맥나마 검정(McNemar Test)는 도수 n_{12} 와 n_{21} 에 의거하여 두 확률이 같은지 검정하는 방법을 제시하였다. 맥나마 검정을 위한 통계량은 다음과 같다.

$$Q_M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \quad (3.4)$$

맥나마 검정 통계량 Q_M 은 귀무가설 하에서 근사적으로 자유도가 1인 카이제곱 분포를 따른다.

다음은 1600명 영국 시민들의 수상에 대한 지지 여부를 두 개의 연속된 여론 조사에서 수집한 자료이다 (Agresti 2012 의 10장 참조). 이제 두 시점에서 수상에 대한 지지율이 같은지 아닌지 R 을 이용하여 맥나마 검정을 해보자. 맥나마 검정은 함수 `mcnemar.test()`를 사용하여 수행할 수 있다.

TABLE 10.1 Rating of Performance of Prime Minister

First Survey	Second Survey		Total
	Approve	Disapprove	
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

그림 3.7.: 영국시민의 수상에 대한 지지도 조사 자료

```
ex3dat <- matrix(c(794,150,86,570),byrow=T,ncol=2)
ex3dat
```

3. 연관성의 검정

```
      [,1] [,2]  
[1,]  794  150  
[2,]   86  570
```

```
mcnemar.test(ex3dat ,correct=F)
```

McNemar's Chi-squared test

data: ex3dat

McNemar's chi-squared = 17.356, df = 1, p-value = 3.099e-05

검정의 p-값이 매우 작으므로 귀무가설을 기각한다. 두 시점에서 수상에 대한 지지율이 하락했다고 할 수 있다. 참고로 첫 번째 조사에서의 지지율의 추정치는 $\hat{p}_1 = 944/1600 = 0.59$ 이고 두 번째 조사에서의 지지율의 추정치는 $\hat{p}_2 = 880/1600 = 0.55$ 이다. 또한 의견을 바꾸지 않은 사람의 비율은 $(794 + 570)/1600 = 0.8225$ 로 대부분의 시민들이 지지 의견을 바꾸지 않았다.

4. 진단의 평가

의학에서 진단은 환자의 상태나 질병의 징후를 판단하는 일이다. 진단을 수행하기 위해서 의사는 전통적인 진단법도 사용하지만 다양한 계측 기계를 이용하는 진단 기법도 사용한다. 최근에는 첨단 분석 장비를 이용하여 다양한 질병에 대한 진단을 정확하고 쉽게 할 수 있다. 특히 최근 코로나 시대에 들면서 일반인도 여러 가지 이유로 진단 검사를 받는 경우가 자주 일어난다.

진단 기법을 사용하여 감염 여부 등을 판단하는 경우 언제나 오류가 발생한다. 최근에는 첨단 기술 덕분에 이러한 오류율이 많이 줄어 들었지만 오류가 없는 진단 기법은 개발하기 힘들다.

코로나 검사를 받고 음성 판정을 받아도 실제 양성인 경우가 나타나며, 반대로 양성 판정을 받아도 음성이 경우가 나타난다. 이렇게 진단에서 발생하는 오류는 두 가지 종류가 있다.

연구자들이 진단 기법을 개발할 때 오류의 가능성이 작아지도록 노력하지만, 불행하게도 두 가지 오류의 확률을 모두 0으로 만들 수 없다.

극단적인 예를 들어보자. 코로나 바이러스 감염의 유무를 판단하는 진단 기법 A는 검사를 받는 사람을 모두 양성이라고 판단한다고 하자. 이 경우 양성인 사람이 음성으로 잘못 판단되는 오류의 확률은 0이다. 반대로 진단 기법 B는 검사를 받는 사람을 모두 음성이라고 판단한다면 음성인 사람이 양성으로 잘못 판단되는 오류의 확률은 0이다. 여기서 진단 기법 A와 B는 모두 쓸모없는 검사라는 것을 우리는 잘 알고 있다. 양성인 사람과 음성인 사람을 잘 구별할 수 있는 진단 기법이 좋은 방법이다.

이제 우리는 진단 기법을 평가할 때 사용되는 확률의 측도에 대하여 알아보자.

Note

일반적으로 양성(positive)은 바이러스에 감염되었거나 질병이 있다는 사건을 말한다. 음성(negative)은 양성(positive)의 반대 사건이다. 하지만 양성(positive)과 음성(negative)의 의미가 바뀌는 경우도 종종 있다.

4.1. 민감도와 특이도

진단 기법을 평가하는 경우 다음과 같은 두 질문에 대해서 생각해 보아야 한다.

- 양성인 사람을 얼마나 잘 양성으로 판단하는가?
- 음성인 사람을 얼마나 잘 음성으로 판단하는가?

양성인 사람을 얼마나 잘 양성으로 판단하는지에 대한 평가 기준이 **민감도(sensitivity)**이고 음성인 사람을 얼마나 잘 음성으로 판단하는지에 대한 평가 기준이 **특이도(specificity)**이다. 민감도와 특이도의 정도는 확률로서 나타낼 수 있다.

4. 진단의 평가

진단 기법에 대한 실험 연구를 수행하면 그 결과는 2×2 분할표로 다음과 같이 요약할 수 있다. 일반적으로 진단 기법의 효과를 측정하는 실험은 대상자에 대한 질병의 유무를 알고 시작한다.

표 4.1.: 진단 기법의 실험 결과

진단(T) / 질병(D)	양성 (D+)	음성 (D-)
양성 (T+)	<i>TP</i>	<i>FP</i>
음성 (T-)	<i>FN</i>	<i>TN</i>

위의 표에서 각 셀에 해당하는 진단 결과는 다음과 같이 나타낼 수 있다.

- *TP* : True Positive
- *FP* : False Positive
- *FN* : False Negative
- *TN* : True Negative

이제 분할표 4.1 에서 민감도와 특이도는 다음과 같이 정의된다.

$$\text{Sensitivity(민감도)} = \frac{TP}{TP + FN} \quad (4.1)$$

$$\text{Specificity(특이도)} = \frac{TN}{FP + TN} \quad (4.2)$$

다음은 코로나 바이러스 검사법에 대한 여러 연구에서 나온 민감도와 특이도 결과를 보여 준다 (Butler-Laporte et al. (2021)).

Figure 3. Primary Meta-analysis Results for the Detection of Severe Acute Respiratory Syndrome Coronavirus 2 in Saliva Samples

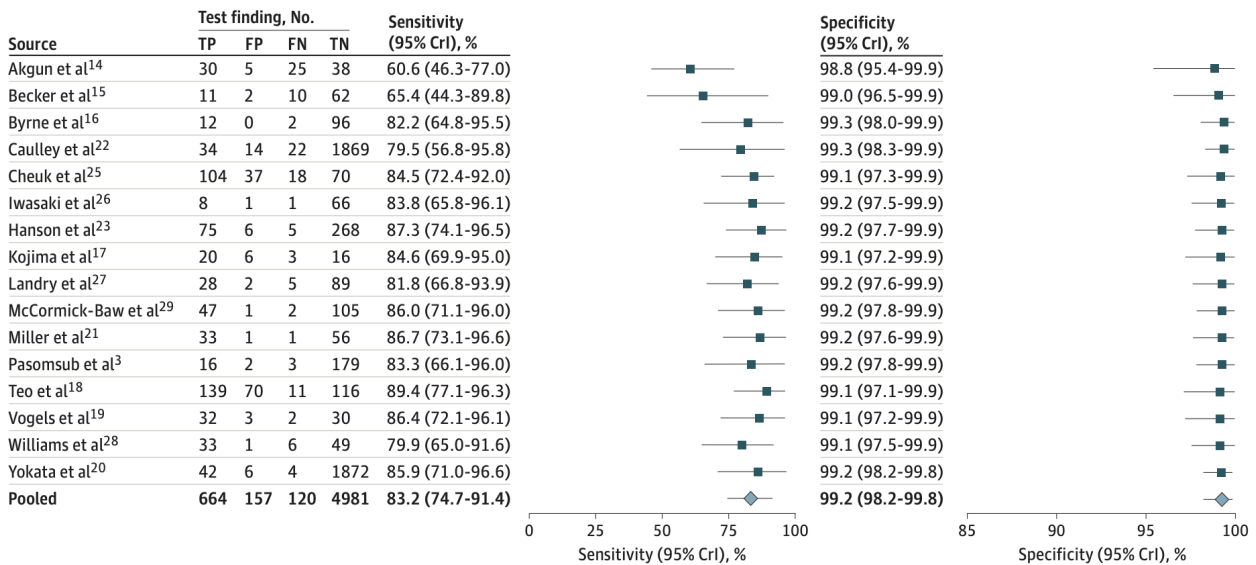


그림 4.1.: 코로나 검사의 민감도와 특이도

4. 진단의 평가

예제로서 그림 그림 4.1 에 제시한 종합적인 결과(pooled counts)를 이용하여 민감도와 특이도를 구해보자.

표 4.2.: 코로나 바이러스 검사법의 결과		
진단(T) / 질병(D)	양성 (D+)	음성 (D-)
양성 (T+)	664	157
음성 (T-)	120	4981

민감도와 특이도는 다음과 같이 구할 수 있다.

$$\text{Sensitivity} = \frac{664}{664 + 120} = 0.8469$$

$$\text{Specificity} = \frac{4891}{157 + 4891} = 0.9689$$

위에서 구한 민감도와 특이도는 Butler-Laporte et al. (2021) 에서 제시한 민감도(83.2%), 특이도(99.2%) 와 유사하지만 약간의 차이가 있다. 그 이유는 Butler-Laporte et al. (2021) 는 모든 실험 결과를 단순히 더한 것이 아니라 메타분석(meta analysis)을 사용하여 얻은 결과이기 때문이다. 메타분석은 같은 주제에 대한 여러 개의 독립적인 연구 결과들을 결합하여 결론을 추론하는 연구 방법이다.

4.2. 양성예측도와 음성예측도

앞에서 살펴본 민감도와 특이도를 구하는 실험에서는 실험 대상자가 질병이 있는지 없는지 알고 있다. 하지만 실제 검사는 진단을 받는 사람이 질병이 있는지 모르는 상태에서 진행된다.

따라서 우리가 정말 관심 있는 확률은 양성으로 진단된 사람이 실제로 양성인지?에 대한 확률이다.

양성으로 판정되었을 때 실제로 병에 걸렸을 확률을 양성예측도(PV+) (predicted value of positive test, predictive value positive) 라고 부르며 음성으로 판정되었을 때 실제로 병에 걸리지 않았을 확률을 음성예측도(PV-) (predicted value of negative test, predicted value negative) 라고 부른다. 양성예측도와 음성예측도는 조건부 확률로 표현할 수 있다.

$$PV+ = P(D+ | T+) \quad (4.3)$$

$$PV- = P(D- | T-) \quad (4.4)$$

이제 앞에서 살펴본 민감도와 특이도도 다음과 같이 조건부 확률로 나타낼 수 있다.

$$\text{Sensitivity} = P(T+ | D+) \quad (4.5)$$

$$\text{Specificity} = P(T- | D-) \quad (4.6)$$

4. 진단의 평가

이제 실제로 중요한 양성예측도와 음성예측도를 민감도와 특이도를 이용하여 유도해 보자. 두 확률은 사건과 조건이 바뀐 확률이기 때문에 베이즈 정리(Bayes' Theorem)을 이용하여 구할 수 있다.

일단 양성예측도를 구하는 식을 베이즈 정리를 적용하여 유도해 보자.

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

위의 식에서 나타나는 확률 $P(D+)$ 는 모집단에서 질병에 걸린 사람들의 비율을 의미하며 이를 **유병률(prevalence)** 이라고 부른다. 즉 양성예측도를 구하려면 질병의 유병률을 알아야 한다.

다시 식을 정리해 보면 양성예측도에 대한 공식은 다음과 같다.

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)} \quad (4.7)$$

$$= \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + [1 - P(T-|D-)][1 - P(D+)]} \quad (4.8)$$

$$= \frac{(\text{민감도})(\text{유병률})}{(\text{민감도})(\text{유병률}) + (1 - \text{특이도})(1 - \text{유병률})} \quad (4.9)$$

비슷한 계산 방법으로 음성예측도는 다음과 같이 주어진다.

$$P(D-|T-) = \frac{P(T-|D-)P(D-)}{P(T-|D-)P(D-) + P(T-|D+)P(D+)} \quad (4.10)$$

$$= \frac{P(T-|D-)[1 - P(D+)]}{P(T-|D-)[1 - P(D+)] + [1 - P(T+|D-)]P(D+)} \quad (4.11)$$

$$= \frac{(\text{특이도})(1 - \text{유병률})}{(\text{특이도})(1 - \text{유병률}) + (1 - \text{민감도})(\text{유병률})} \quad (4.12)$$

이제 표 4.2 의 결과를 이용하여 코로나 검사의 양성예측도와 음성예측도를 구해보자.

코로나 유병률은 나라마다 다르고 추정하기도 힘들다. 따라서 쉽게 현재 까지 누적환자수를 전체 인구로 나눈 단순한 비율을 유병률로 사용해 보자(주의! 우리가 여기서 사용한 비율은 실제 유병률을 계산하는 방법과 다르다). 2021년 현재 누적 환자 수가 274,415 명이고 2020년 기준 총인구는 51,829,136 명이므로 유병률을 $274415/51829136 = 0.0053$ 이라고 하자.

이제 표 4.2 의 결과를 이용하면 코로나 검사의 양성예측도와 음성예측도는 다음과 같이 추정할 수 있다.

$$\begin{aligned} P(D+|T+) &= \frac{(\text{민감도})(\text{유병률})}{(\text{민감도})(\text{유병률}) + (1 - \text{특이도})(1 - \text{유병률})} \\ &= \frac{(0.8469)(0.0053)}{(0.8469)(0.0053) + (1 - 0.9689)(1 - 0.0053)} \\ &= 0.1267 \end{aligned}$$

4. 진단의 평가

```
(0.8469)*(0.0053)/((0.8469)*(0.0053) + (1-0.9689)*(1-0.0053))
```

```
[1] 0.1267108
```

$$P(D-|T-) = \frac{(\text{특이도})(1 - \text{유병률})}{(\text{특이도})(1 - \text{유병률}) + (1 - \text{민감도})(\text{유병률})} \quad (4.13)$$

$$= \frac{(0.9689)(1 - 0.0053)}{(0.9689)(1 - 0.0053) + (1 - 0.8469)(0.0053)} \quad (4.14)$$

$$= 0.9992 \quad (4.15)$$

```
(0.9689)*(1- 0.0053)/((0.9689)*(1-0.0053) + (1-0.8469)*(0.0053))
```

```
[1] 0.9991588
```

사실 코로나 유병률은 정확하게 알 수도 없고 시간에 따라 변할 것이다. 이제 다양한 유병률에 따라서 양성예측도와 음성예측도가 어떻게 변하는지 계산해 보자.

```
calpred <- function(prev, sen, spe){
  pred.pos <- sen*prev/(sen*prev + (1-spe)*(1-prev))
  pred.neg <- spe*(1-prev)/(spe*(1-prev) + (1-sen)*(prev))
  res <- data.frame(sen, spe, prev, pred.pos, pred.neg)
  colnames(res) <- c("Sensitivity", "SPecificity", "Prevalnce", "Pred. Post.", "Pred. Nega.")
  res
}

preval.range <- seq(0, 0.02, 0.002)
calpred(preval.range, 0.8469, 0.9689)
```

	Sensitivity	SPecificity	Prevalnce	Pred. Post.	Pred. Nega.
1	0.8469	0.9689	0.000	0.00000000	1.0000000
2	0.8469	0.9689	0.002	0.05174816	0.9996834
3	0.8469	0.9689	0.004	0.09858220	0.9993658
4	0.8469	0.9689	0.006	0.14117039	0.9990471
5	0.8469	0.9689	0.008	0.18006506	0.9987273
6	0.8469	0.9689	0.010	0.21572673	0.9984064
7	0.8469	0.9689	0.012	0.24854242	0.9980845
8	0.8469	0.9689	0.014	0.27883973	0.9977614
9	0.8469	0.9689	0.016	0.30689786	0.9974372
10	0.8469	0.9689	0.018	0.33295620	0.9971120
11	0.8469	0.9689	0.020	0.35722119	0.9967856

4. 진단의 평가

```
calpred(0.0053, 0.8469, 0.9689 )
```

	Sensitivity	SPecificity	Prevalnce	Pred. Post.	Pred. Nega.
1	0.8469	0.9689	0.0053	0.1267108	0.9991588

Part II.

집단의 비교

5. 분산분석

5.1. 필요한 패키지

```
library(tidyverse)
library(kableExtra)
library(ggplot2)
library(here)

library(agricolae)
```

5.2. 일원배치법

일원배치법 실험(one-way factor design)에서는 하나의 요인(factor) A 의 효과를 측정한다. 요인 A 에 대하여 서로 다른 a 개의 수준(level), A_1, A_2, \dots, A_a 의 효과를 비교한다고 가정하자. 각 수준에 대하여 r_i 개의 반응값을 반복 측정한다.

이제 i 번째 수준에서 측정된 j 번째 반응변수의 값을 y_{ij} 라고 하자. 일원배치법에서 측정된 자료들은 다음과 같은 모형을 가진다고 가정한다.

$$y_{ij} = \mu_i + e_{ij} \quad \text{where} \quad e_{ij} \sim N(0, \sigma_E^2) \quad (5.1)$$

여기서 오차항 e_{ij} 는 모두 독립이다.

모형 5.1 은 일반적으로 **평균모형(mean model)** 이라고 부르며 모형의 이름대로 μ_i 는 i 번째 수준에서 반응변수의 평균을 의미한다.

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

이제 모형 5.1 을 변형하여 다른 형식의 모형을 만들어 보자.

5. 분산분석

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

위의 모형에서 모수 μ 는 반응변수의 전체 평균을 의미하며 $\alpha_i = \mu_i - \mu$ 는 i 번째 수준의 평균이 전체 평균과 어떻게 다른지 나타내는 수준의 상대적 효과를 의미한다.

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

$$y_{ij} = \mu + \alpha_i + e_{ij} \text{ where } e_{ij} \sim N(0, \sigma_E^2) \quad (5.2)$$

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

$$\sum_{i=1}^a \alpha_i = 0 \quad (5.3)$$

제약조건 5.3는 모수의 개수($a + 1$)가 그룹의 개수(a)보다 많아서 발생하는 문제를 해결하기 위하여 모수에 대한 제약 조건 1개를 고려해서 모수의 개수와 그룹의 개수를 맞추어준 것이다.

제약조건 5.3은 **sum to zero**조건이라고 부르며 문제를 해결하는 유일한 조건은 아니다. 예를 들어서 조건 5.3을 대신하여 $\alpha_1 = 0$ 인 **set to zero** 조건을 사용할 수 있다.

5.3. 가설

집단의 모평균을 편의상 $\mu_1, \mu_2, \dots, \mu_a$ 이라고 하자. 평균모형 5.1을 가정하고 집단들 사이에 차이가 있는지에 대한 가설은 다음과 같다.

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

위의 가설에서 주의할 점은 대립가설 H_1 의 경우를 고려하면 평균들이 서로 다른 경우가 매우 다양하다는 것이다. 예를 들어 집단이 3개 인 경우 $\mu_1 = \mu_2 < \mu_3$ 일 수도 있고 $\mu_1 < \mu_2 < \mu_3$ 있으며 이 외에 매우 다양한 경우들이 있다.

이제 효과모형 5.2을 고려하면 집단들 사이에 차이가 있는지에 대한 가설을 다음과 같이 바꿀수 있다. 집단에 대한 효과가 모두 0이 되면 집단 간의 평균에 대한 차이는 없다.

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_a = 0 \quad \text{vs.} \quad H_1 : \text{not } H_0 \quad (5.4)$$

5.4. 변동의 분해

이제 관측값이 가진 모든 변동을 집단 간의 변동(각 집단의 평균의 차이가 얼마나 나는지에 대한 통계량)과 집단 내의 변동(각 집단내에서 관측값들의 퍼진 정도)를 나누어 파악할 수 있는 통계량을 찾아서 검정 통계량을 구성해 보자.

일단 각 집단의 반복 측정값의 횟수는 모두 같다고 가정하자($r_i = r$). 전체 평균과 집단의 평균을 다음과 같이 정의하자.

$$\bar{\bar{y}} = \frac{\sum_{i=1}^a \sum_{j=1}^r y_{ij}}{ar} = \frac{T}{ar}, \quad \bar{y}_{i.} = \frac{\sum_{j=1}^r y_{ij}}{r} = \frac{T_i}{r}$$

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

$$\underbrace{y_{ij} - \bar{\bar{y}}}_{\text{total deviation}} = \underbrace{(y_{ij} - \bar{y}_{i.})}_{\text{within-group deviation}} + \underbrace{(\bar{y}_{i.} - \bar{\bar{y}})}_{\text{between-group deviation}} \quad (5.5)$$

식 5.5 에서 집단 평균과 총 평균의 편차 ($\bar{y}_{i.} - \bar{\bar{y}}$)는 처리의 효과를 측정할 수 있는 통계량이다. 집단 간의 차이를 반영하는 양으로 처리 효과 α_i 들에 의하여 발생한다.

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

식 5.5 의 각 편차들은 양수와 음수로서 부호를 가지기 때문에 이를 변동으로 표현하기 위하여 차이를 제곱하여 합친 제곱합(sum of squares)을 고려해 보자.

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

결과적으로 다음과 같은 변동의 분해를 제곱합의 형식으로 얻을 수 있다.

$$\underbrace{\sum_{i=1}^a \sum_{j=1}^r (y_{ij} - \bar{\bar{y}})^2}_{\text{total variation}} = \underbrace{\sum_{i=1}^a \sum_{j=1}^r (y_{ij} - \bar{y}_{i.})^2}_{\text{within-group variation}} + \underbrace{\sum_{i=1}^a \sum_{j=1}^r (\bar{y}_{i.} - \bar{\bar{y}})^2}_{\text{between-group variation}} \quad (5.6)$$

분해식 5.6 에서 나타난 각 제곱합에 대한 이름과 의미를 살펴보자.

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

$$SST = \sum_{i=1}^a \sum_{j=1}^r (y_{ij} - \bar{y})^2$$

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

$$SSE = \sum_{i=1}^a \sum_{j=1}^r (y_{ij} - \bar{y}_{i.})^2$$

- SSA 를 처리 제곱합(Treatment Sum of Squares)이라고 부르며 처리들의 차이로 발생하는 변동으로 집단 간의 변동 또는 집단 변동(between-group variation)을 의미한다.

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

이제 분해식 5.6 을 다음과 같이 나타낼수 있다.

$$SST = SSA + SSE \quad (5.7)$$

위의 분해식의 통계량들을 보면 다음과 같은 사실을 알수 있다.

- 집단 간의 변동의 크기를 나타내는 처리제곱합(SSA)이 커질수록 집단 간의 평균이 유의한 차이가 난다는 증거가 강해진다.
- 위의 현상을 다시 말하면, 집단내의 변동의 크기를 나타내는 오차제곱합(SSE)이 작아질수록 집단 간의 평균이 유의한 차이가 증거가 강해진다.
- 총제곱합은 자료의 전체 변동을 나타내며 그룹을 어떻게 나누건 그 값은 언제나 일정하다.
- 총제곱합(SST)은 처리제곱합(SSA)과 오차제곱합(SSE)의 합이다.
- 따라서 처리제곱합이 커지면 오차제곱합이 상대적으로 작아지는 현상을 나타낸다. 또한 처리제곱합이 작아지면 오차제곱합이 상대적으로 커지는 현상을 나타낸다.
- 처리제곱합과 오차제곱합의 비율로 집단 간의 차이를 추론할 수 있다.

5.5. 자유도

제곱합은 편차(deviation)의 제곱들을 더한 형태로서 각 제곱합들에 대하여 해당하는 자유도(degrees of freedom; df 또는 ϕ 로 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 5.7 에 주어진 제곱합의 자유도에 대한 정보를 다음과 같이 정리할 수 있다.

제곱합	편차의 개수	제약조건	제약조건에 의한	
			수	자유도
SST	ar	$\sum_{i=1}^a \sum_{j=1}^r (y_{ij} - \bar{y}) = 0$	1	$\phi_T = ar - 1$
SSA	a	$\sum_{i=1}^a (\bar{y}_i - \bar{y}) = 0$	1	$\phi_A = a - 1$
SSE	ar	$\sum_{j=1}^r (y_{ij} - \bar{y}_{i.}) = 0, i = 1, 2, \dots, a$	a	$\phi_E = ar - a$

만드시 유의해야 할 점은 총제곱합의 자유도는 처리제곱합의 자유도와 오차제곱합의 자유도의 합과 같다는 것이다. 이를 이용하면 두 개의 자유도만 구하면 나머지 하나의 자유도를 쉽게 구할 수 있다.

5.6. 평균제곱합과 F-통계량

이제 가설 5.4 을 검정하기 위한 통계량을 구성해 보자. 먼저 다음과 같은 제곱합들을 각 자유도로 나눈 평균제곱합 (Mean Sum of Squares)를 정의한다.

$$MSA = \frac{SSA}{\phi_A}, \quad MSE = \frac{SSE}{\phi_E} \quad (5.8)$$

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.

$$F_0 = \frac{MSA}{MSE} = \frac{\text{between-group variation}}{\text{within-group variation}} \quad (5.9)$$

위 식 5.9 에서 정의된 F-통계량은 그룹 간에 평균의 차이가 클수록, 그룹 내의 변동이 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.

5.7. 분산분석을 이용한 F-검정

이제 식 5.9 에서 정의된 F-통계량을 이용하여 가설 5.4 를 검정하는 통계적 방법을 만들어 보자. 일단 두 제곱합의 통계적 성질은 다음과 같다.

5. 분산분석

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

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

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

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

- 잔차 제곱합과 처리 제곱합은 서로 독립이다.

따라서 귀무가설이 참인 경우 F-통계량은 자유도가 ϕ_A, ϕ_E 를 가지는 F-분포를 따른다.

$$F_0 = \frac{MSA}{MSE} = \frac{\frac{SSA/\sigma_E^2}{\phi_A}}{\frac{SSE/\sigma_E^2}{\phi_E}} \sim F(\phi_A, \phi_E) \quad \text{under } H_0 \quad (5.10)$$

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.

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

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

요인	제곱합	자유도	평균제곱합	F_0	p-값
처리	SSA	$\phi_A = a - 1$	$MSA = \frac{SSA}{\phi_A}$	$F_0 = \frac{MSA}{MSE}$	$P[F(\phi_A, \phi_E) > F_0]$
잔차	SSE	$\phi_E = a(r - 1)$	$MSE = \frac{SSE}{\phi_E}$		
총합	SST	$\phi_T = ar - 1$			

5.8. 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.

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

다음으로 각 수준(집단)에 대한 평균에 대한 추정량은 표본평균 \bar{y}_i 이며

$$\hat{\mu}_i = \widehat{\mu + \alpha_i} = \bar{y}_i, \quad E(\bar{y}_i) = \mu_i$$

100(1 - α) % 신뢰구간(confidence interval)은 다음과 같이 주어진다.

$$\bar{y}_i. \pm t(1 - \alpha/2, \phi_E) \sqrt{\frac{MSE}{r}}$$

여기서 $t(1 - \alpha/2, \phi_E)$ 는 자유도 ϕ_E 를 가지는 t-분포의 $1 - \alpha/2$ 분위수를 의미한다.

이제 두 개의 수준에 대한 평균의 차이에 대한 통계적 추론을 생각해 보자. 수준 A_i 와 A_j 의 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 - α) % 신뢰구간은 다음과 같이 주어진다.

$$(\bar{y}_i. - \bar{y}_j.) \pm t(1 - \alpha/2, \phi_E) \sqrt{\frac{2MSE}{r}} \quad (5.11)$$

신뢰구간 7.1에서 두 개의 표본 평균 \bar{y}_i 와 \bar{y}_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 α 에서 다음과 같은 조건을 만족하면 위의 귀무가설을 기각한다.

$$|\bar{y}_i. - \bar{y}_j.| > t(1 - \alpha/2, \phi_E) \sqrt{\frac{2MSE}{r}} \quad (5.12)$$

식 7.2 에서 주어진 귀무 가설 $\delta_{ij} = 0$ 을 기각하는 조건은 식 7.1 에 주어진 신뢰구간이 0 을 포함하지 않는 조건과 동일하다.

식 7.2 에서 검정을 위한 조건의 우변을 최소유의차(least significant difference; LSD) 라고 부른다. 두 수준의 차이가 유의하려면 두 평균 차이의 절대값이 최소한 최소유의차의 값보다 커야한다.

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

5.9. 예제: 저혈당 임상실험

부교재 Jaewon Lee (2005) 의 5장에 나오는 당뇨병 환자에 대한 예제를 이용하여 치료군 간의 차이를 살펴보자. 한 연구자는 저혈당에 걸린 20명의 환자에 대하여 혈당을 높이는 서로 다른 5개의 약품 효능을 비교하고자 한다. 환자 20명을 5개의 그룹(treat)으로 나누어 각각의 약품을 이용하여 1개월 간 치료를 실시하였다.

치료 시작전에 모든 환자의 혈당을 측정하고(baseline) 1개월의 치료 기간이 지난 후 혈당을 측정하였다(response).

임상시험자료를 다음과 같이 읽어서 data.frame 형식으로 저장한다.

```
diabetes <- read.csv(here("data", "chapter-5-data.txt"), sep=' ', header = F)
colnames(diabetes) <- c("treat", "baseline", "response")
diabetes$treat <- factor(diabetes$treat)
diabetes <- diabetes %>% arrange(treat)
```

저혈당 환자에 대한 임상실험에서 얻은 자료는 다음과 같다.

```
diabetes %>%
  kbl(caption = "저혈당 환자에 대한 임상실험 결과") %>%
  kable_styling(bootstrap_options = "striped", full_width = F, position = "center",
    font_size = 12)
```

각 치료그룹의 치료 후 혈당(response)의 치료집단 별 기초통계량을 구해보자.

```
dfsumm <- diabetes %>%
  group_by(treat) %>%
  summarise(mean = mean(response), median = median(response), sd = sd(response),
    min = min(response), max = max(response))
dfsumm

# A tibble: 5 x 6
  treat mean median    sd  min  max
<fct> <dbl>  <dbl> <dbl> <dbl> <dbl>
1     1  10.5    10.5  1.5  8.5  12.5
2     2  12.0    12.0  1.0  10.0  14.0
3     3  11.0    11.0  1.0  9.0  13.0
4     4  13.0    13.0  1.0  11.0  15.0
5     5  14.0    14.0  1.0  12.0  16.0
```

5. 분산분석

표 5.3.: 저혈당 환자에 대한 임상실험 결과

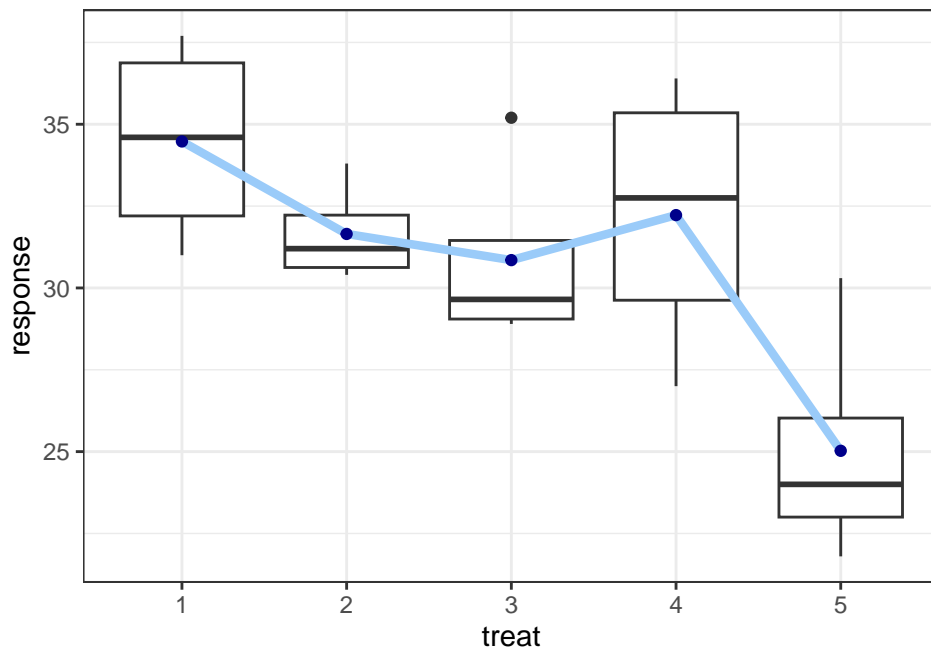
treat	baseline	response
1	27.2	32.6
1	33.0	37.7
1	22.0	36.6
1	26.8	31.0
2	28.6	33.8
2	26.5	30.7
2	26.8	31.7
2	26.8	30.4
3	28.6	35.2
3	23.2	28.9
3	22.4	29.1
3	24.4	30.2
4	29.3	35.0
4	30.3	36.4
4	21.8	27.0
4	24.3	30.5
5	20.4	24.6
5	25.1	30.3
5	19.6	23.4
5	18.1	21.8

1	1	34.5	34.6	3.19	31	37.7
2	2	31.6	31.2	1.54	30.4	33.8
3	3	30.8	29.6	2.96	28.9	35.2
4	4	32.2	32.8	4.30	27	36.4
5	5	25.0	24	3.70	21.8	30.3

치료집단 별로 치료그룹의 치료 후 혈당(response)의 분포를 다음과 상자그림으로 비교해보자.

```
ggplot(diabetes, aes(treat, response)) +
  geom_boxplot() +
  geom_line(data=dfsumm, aes(x=treat, y=mean, group=1), linewidth=1.5, col="#9ACBF9") +
  geom_point(data=dfsumm, aes(x=treat, y=mean), col="darkblue") +
  theme_bw()
```

5. 분산분석



이제 위에서 제시한 F-검정을 이용하여 약품별로 치료 후 혈당의 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~treat,data=diabetes)
summary(anova.res)
```

```

      Df Sum Sq Mean Sq F value Pr(>F)
treat    4  198.4   49.60    4.643  0.0122 *
Residuals 15  160.3   10.68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 분산분석표에서 p-값이 유의수준 5% 보다 매우 작으므로 약품별로 치료 후 혈당의 평균이 같다는 귀무가설을 기각한다. 따라서 5개의 집단의 치료 후 혈당 평균은 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. `LSD.test()` 함수는 `agricolae` 패키지에서 제공한다.

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

5. 분산분석

Study: anova.res ~ "treat"

LSD t Test for response

Mean Square Error: 10.68417

treat, means and individual (95 %) CI

	response	std r	se	LCL	UCL	Min	Max	Q25	Q50	Q75
1	34.475	3.188913	4 1.634332	30.9915	37.9585	31.0	37.7	32.200	34.60	36.875
2	31.650	1.537314	4 1.634332	28.1665	35.1335	30.4	33.8	30.625	31.20	32.225
3	30.850	2.955785	4 1.634332	27.3665	34.3335	28.9	35.2	29.050	29.65	31.450
4	32.225	4.297577	4 1.634332	28.7415	35.7085	27.0	36.4	29.625	32.75	35.350
5	25.025	3.698986	4 1.634332	21.5415	28.5085	21.8	30.3	23.000	24.00	26.025

Alpha: 0.05 ; DF Error: 15

Critical Value of t: 2.13145

Comparison between treatments means

	difference	pvalue	signif.	LCL	UCL
1 - 2	2.825	0.2405		-2.1014081	7.751408
1 - 3	3.625	0.1376		-1.3014081	8.551408
1 - 4	2.250	0.3458		-2.6764081	7.176408
1 - 5	9.450	0.0010	***	4.5235919	14.376408
2 - 3	0.800	0.7341		-4.1264081	5.726408
2 - 4	-0.575	0.8069		-5.5014081	4.351408
2 - 5	6.625	0.0118	*	1.6985919	11.551408
3 - 4	-1.375	0.5608		-6.3014081	3.551408
3 - 5	5.825	0.0235	*	0.8985919	10.751408
4 - 5	7.200	0.0071	**	2.2735919	12.126408

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

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

Study: anova.res ~ "treat"

LSD t Test for response

5. 분산분석

Mean Square Error: 10.68417

treat, means and individual (95 %) CI

	response	std r	se	LCL	UCL	Min	Max	Q25	Q50	Q75
1	34.475	3.188913	4 1.634332	30.9915	37.9585	31.0	37.7	32.200	34.60	36.875
2	31.650	1.537314	4 1.634332	28.1665	35.1335	30.4	33.8	30.625	31.20	32.225
3	30.850	2.955785	4 1.634332	27.3665	34.3335	28.9	35.2	29.050	29.65	31.450
4	32.225	4.297577	4 1.634332	28.7415	35.7085	27.0	36.4	29.625	32.75	35.350
5	25.025	3.698986	4 1.634332	21.5415	28.5085	21.8	30.3	23.000	24.00	26.025

Alpha: 0.05 ; DF Error: 15

Critical Value of t: 2.13145

least Significant Difference: 4.926408

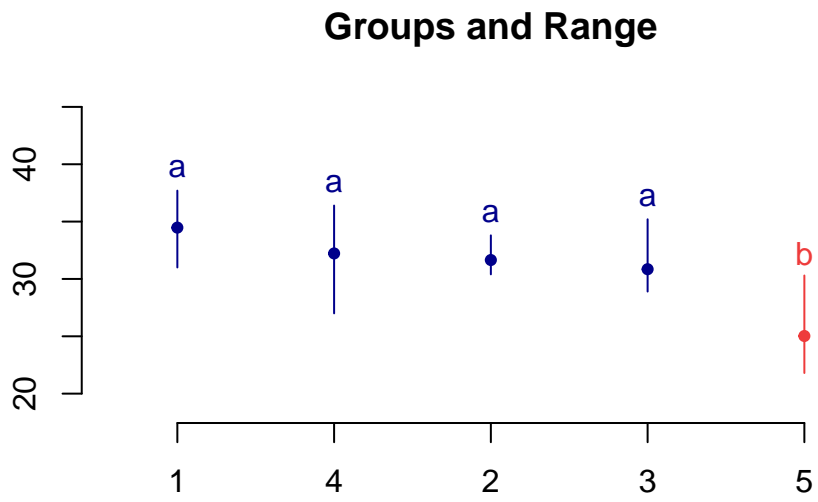
Treatments with the same letter are not significantly different.

	response	groups
1	34.475	a
4	32.225	a
2	31.650	a
3	30.850	a
5	25.025	b

result2\$groups

	response	groups
1	34.475	a
4	32.225	a
2	31.650	a
3	30.850	a
5	25.025	b

```
plot(result2)
```



6. 공분산분석

6.1. 필요한 패키지

```
library(tidyverse)
library(kableExtra)
library(ggplot2)
library(here)

library(car)
library(agricolae)
```

6.2. 공분산분석 개요

서로 다른 집단을 비교하는 실험이나 관측연구에서 관심이 있는 처리(treatment)나 요인(factor)뿐만 아니라 다른 예측변수들도 반응변수에 영향을 미친다. 이러한 예측변수들의 영향을 제거하기 위한 방법은 여러가지가 있지만 실험인 경우 임의화 방법(randomization)으로 그 영향을 상쇄시킬 수 도 있고 관측연구인 경우에는 사례-대조연구 방법을 이용하여 그 영향을 최소화하려고 노력을 한다.

하지만 많은 경우에 여러 가지 변수들이 반응변수에 영향을 미친다. 이러한 경우에 이러한 독립변수(또는 예측변수)를 모형에 포함시켜서 그 영향을 반영하고 동시에 자료의 변동을 부가적으로 설명해주는 방법이 공분산 분석(analysis of covariance; ANCOVA)이다.

공분산 분석에서 고려되는 예측변수를 **공변량(covariate)** 이라고 부른다.

대부분의 실험연구에는 실험 전에 여러 가지 점수를 측정하는데 이 경우 이러한 점수를 공변량으로 모형에 포함시켜 주는 것이 좋다 (예: 실험 전 상태에 대한 점수, 시험점수, IQ 점수). 또한 임상실험을 여러 개의 병원에서 진행하는 경우 병원 효과를 공변량으로 자주 사용한다.

공분산 모형의 주요한 장점은 반응변수에 대해 설명력이 높은 공변량을 사용하게 되면 잔차제곱합이 감소하여 처리의 효과에 대한 검정력을 높일 수 있다.

여기서 주의해야 할 점은 공변량과 처리는 독립이 되어야한다는 점이다. 만약 처리의 결과가 공변량에 영향을 미치게 되면 이러한 공변량은 모형에 포함시키는 것이 부적절하다.

예를 들어 자동차정비 교육을 위한 두 가지 학습법을 비교하는 실험을 생각해 보자. 학생들을 임의로 두 가지 학습법 중 하나를 선택하여 3개월 동안 교육을 받게 하고 시험을 보아 평균 점수의 차이를 알아보았다. 이 때 공변량으로 총 학습시간을 고려하였는데 학습법의 선택이 총 학습시간에 영향을 줄 수 있다. 즉 고려된 학습법 중 하나는

컴퓨터를 이용한 학습법이며 이 학습법에 배정된 학생들은 컴퓨터 사용을 익히는 시간까지 학습시간에 포함되는 것이 나타났다. 이렇게 공변량이 처리에 의해 영향을 받는 경우(교호작용이 있는 경우)는 이를 모형에 포함시키는 것은 위험하다.

6.3. 공분산분석의 모형

이제 일원배치에서 하나의 공변량이 있는 공분산분석의 모형은 일원배치 모형에 공변량 x 의 효과를 다음과 같이 더해주는 것이다.

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + e_{ij}, \quad i = 1, 2, \dots, a, \quad j = 1, 2, \dots, r \quad (6.1)$$

모형 6.1에서 x_{ij} 는 관측값 y_{ij} 의 공변량이며 $\bar{x}_{..} = \sum_{i=1}^a \sum_{j=1}^r x_{ij}$ 로 공변량의 전체 평균이다. 위의 효과모형은 다음과 같이 평균모형으로 나타낼 수 있다. 어떤 모형이든 모수에 대한 가설 검정의 결과는 동일하다.

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + e_{ij} \\ &= \beta_0 + \alpha_i + \beta x_{ij} + e_{ij} \\ &= \beta_{0i} + \beta x_{ij} + e_{ij} \end{aligned}$$

모형 6.1에서 각 모수의 추정치는 ANOVA 모형에서와 같이 최소제곱법을 이용하여 추정하며 부가조건 $\sum_i \alpha_i = 0$ 을 이용하면 다음과 같은 추정량을 얻을 수 있다

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\alpha}_i &= \bar{y}_{i.} - \bar{y}_{..} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \\ \hat{\beta} &= \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2} \end{aligned}$$

각 집단의 차이와 공변량의 기울기에 대한 통계적 가설 검정에 관심이 있는 경우 공분산분석은 다음과 같이 공변량의 평균으로 보정하지 않는 모형을 사용해도 무방하다.

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + e_{ij}, \quad i = 1, 2, \dots, a, \quad j = 1, 2, \dots, r \quad (6.2)$$

6.4. 가설검정

공분산 분석 모형에서는 다음과 같은 두 가지 가설을 검정할 수 있다. 분산 분석 모형에서와 같이 각 그룹의 평균에 대한 검정을 할 수 있고

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_a = 0 \quad \text{versus} \quad H_1 : \text{not } H_0$$

또한 공변량의 효과에 대한 검정도 할 수 있다.

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0 \quad (6.3)$$

가설검정을 위한 제곱합들을 다음과 같이 정의하자.

$$\begin{aligned} S_{xx(i)} &= \sum_{j=1}^r (x_{ij} - \bar{x}_{i.})^2 \\ S_{yy(i)} &= \sum_{j=1}^r (y_{ij} - \bar{y}_{i.})^2 \\ S_{xy(i)} &= \sum_{j=1}^r (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.}) \\ S_{xx} &= \sum_{i=1}^a \sum_{j=1}^r (x_{ij} - \bar{x})^2 \\ S_{yy} &= \sum_{i=1}^a \sum_{j=1}^r (y_{ij} - \bar{y})^2 \\ S_{xy} &= \sum_{i=1}^a \sum_{j=1}^r (x_{ij} - \bar{x})(y_{ij} - \bar{y}) \\ SST &= \sum_{i=1}^a \sum_{j=1}^r (y_{ij} - \bar{y})^2 \\ SS_T &= \frac{(S_{xx})(S_{yy}) - (S_{xy})^2}{S_{xx}} - SSE \\ SS_X &= \sum_{i=1}^a S_{yy(i)} - SSE \\ SSE &= \frac{(\sum_i S_{xx(i)})(\sum_i S_{yy(i)}) - (\sum_i S_{xy(i)})^2}{\sum_i S_{xx(i)}} \end{aligned}$$

이제 위의 두 검정은 다음과 같은 분산분석표를 이용한 F-검정법으로 수행할 수 있다. $n = ar$ 으로 총 관측값의 개수이다.

표 6.1.: 공분산분석 모형의 분산분석표

요인	제곱합	자유도	평균제곱합	F
공변량	SS_X	1	$MS_X = SS_X/1$	$F_1 = MS_X/MSE$
처리	SS_T	a-1	$MS_T = SS_T/(a-1)$	$F_2 = MS_T/MSE$
오차	SSE	n-a-1	$MSE = SSE/(n-a-1)$	
총합	SST	n-1		

6. 공분산분석

위의 분산분석표에서 공변량 효과에 대한 가설 6.3 은 다음과 같이 p-값을 계산하여 검정할 수 있다.

$$p - value = P[F(1, n - a - 1) > F_1]$$

또한 그룹의 평균에 대한 검정은 ANOVA 검정과 유사하게 p-값을 계산하여 검정할 수 있다.

$$p - value = P[F(a - 1, n - a - 1) > F_2]$$

위의 두 F-검정에 쓰이는 F-분포의 두 번째 자유도가 ANOVA 검정에서 사용되는 자유도($n - a$)보다 하나가 작음($n - a - 1$)을 유의하자.

6.4.1. 최소제곱평균과 각 평균의 비교

공분산 분석 모형 6.1 에서 각 처리에 대한 평균을 구할 때 공변량의 값에 따라서 그 값이 변한다. 따라서 각 그룹의 평균을 비교하는 경우에는 모형의 공변량에 공변량의 전체 평균을 넣어 사용한다. 이러한 평균을 보정된 최소제곱 평균(Least Square Mean)이라고 한다

$$\bar{y}_i = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}\bar{x}_{..}$$

또한 그룹의 평균들을 각각 비교할 수 있다.

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

6.5. 예제: 저혈당 실험

앞에서 살펴본 예제 Section 5.9 의 저혈당 임상실험에서는 혈당을 감소시키기 위한 다섯 개의 처리(**treat**)를 비교 하려고 한다. 반응 변수 **response** 는 치료 적용 1달 후 혈당량 수치이며 각 처리그룹에 대한 자료와 상자그림은 다음과 같다.

```
diabetes <- read.csv(here("data", "chapter-5-data.txt"), sep=' ', header = F)

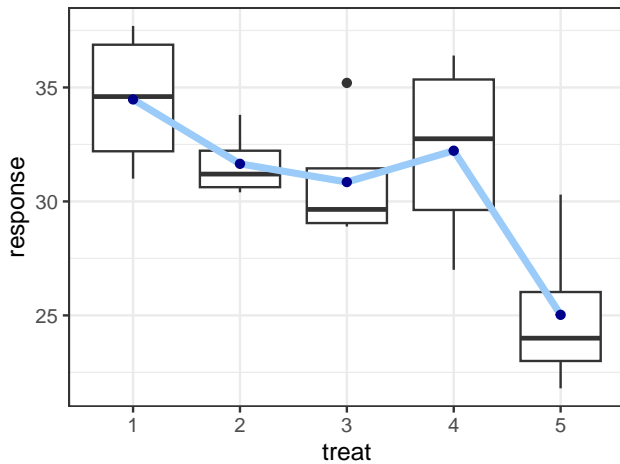
diabetes <- read.csv(here("data", "chapter-5-data.txt"), sep=' ', header = F)
colnames(diabetes) <- c("treat", "baseline", "response")
diabetes$treat <- factor(diabetes$treat)
diabetes <- diabetes %>% arrange(treat)

dfsumm <- diabetes %>% group_by(treat) %>% summarise(mean=mean(response), median= median(res

ggplot(diabetes, aes(treat, response)) +
```

6. 공분산분석

```
geom_boxplot() +
geom_line(data=dfsumm, aes(x=treat, y=mean, group=1), linewidth=1.5, col="#9ACBF9") +
geom_point(data=dfsumm, aes(x=treat, y=mean), col="darkblue") +
theme_bw()
```

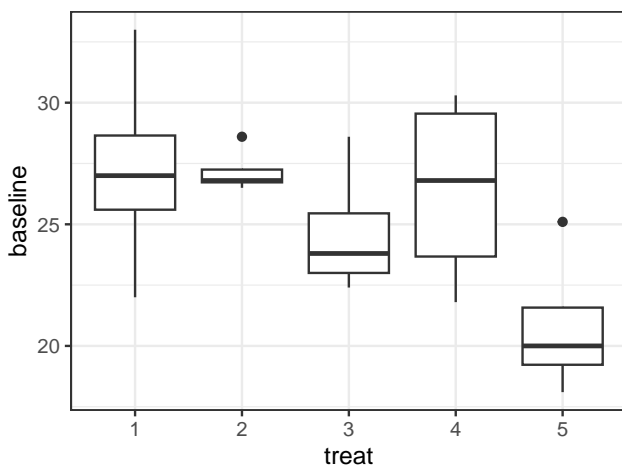


대부분의 임상실험에서는 치료를 시작하기 전에 반응변수의 값을 먼저 측정하고 치료 후의 변화를 본다. 치료를 시작하기 전에 반응변수의 값을 **기준값(baseline value)** 라고 한다. 기준값은 환자의 치료 전 상태를 나타내는 중요한 변수이며 반응값에 영향을 미칠 수 있으므로 대부분의 임상실험에서는 기준값을 공변량으로 분석에 포함한다. 저혈당 실험의 공변량 x 는 치료 전 측정한 혈당량 수치(**baseline**)이다.

이러한 기준값의 분포가 각 처리집단 별로 크게 다르면 실험의 공정성에 문제가 생긴다.

이제 치료 전 측정한 혈당량의 분포를 처리 집단별로 살펴보자. 아래 그림에서 보면 치료집단별로 치료 전 측정한 혈당량의 분포가 다르다는 것을 알 수 있다.

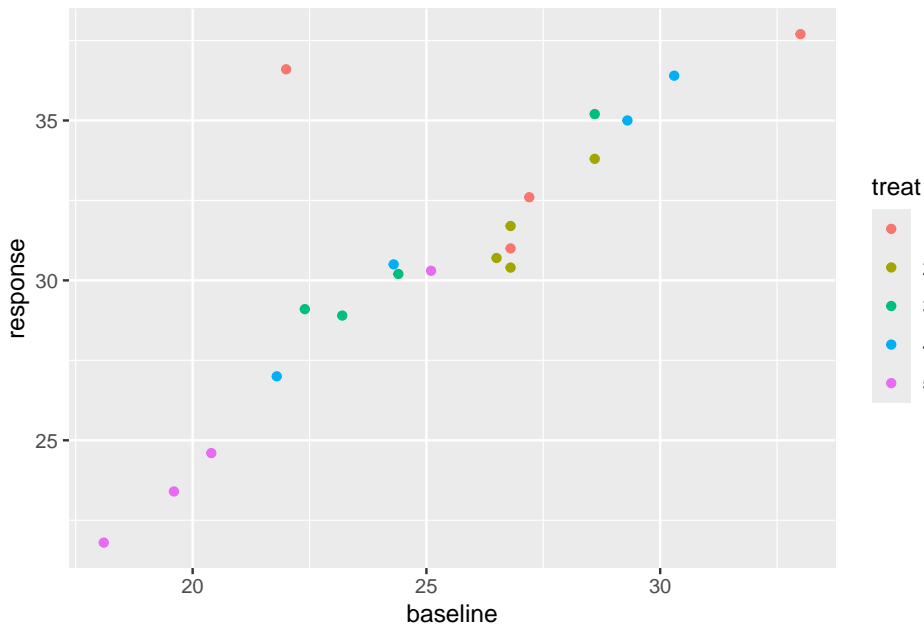
```
ggplot(diabetes, aes(treat, baseline)) +
geom_boxplot() +
theme_bw()
```



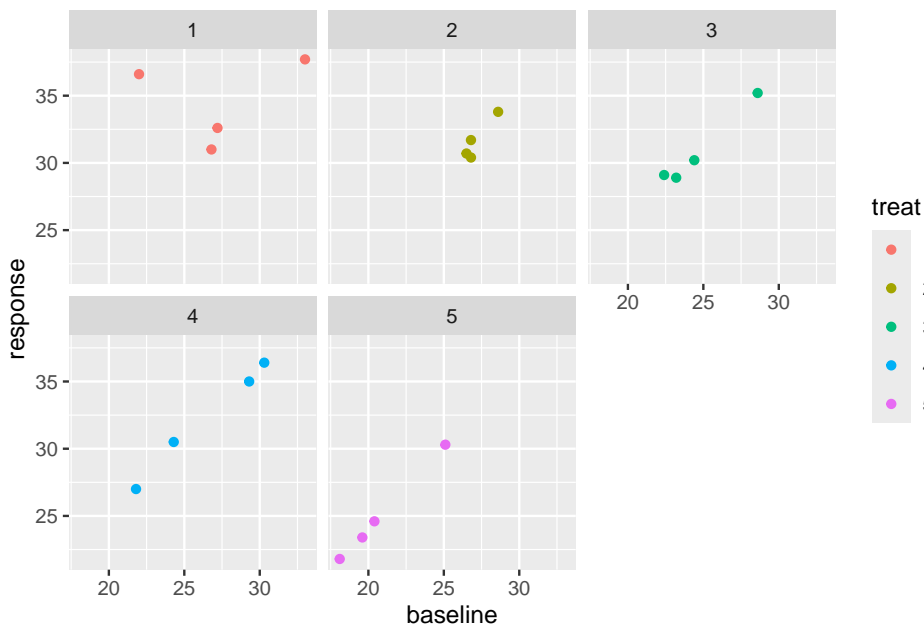
6. 공분산분석

이제 치료 전 측정한 혈당량과 치료 후의 혈당량의 관계는 다음과 같은 산점도로 나타낼 수 있으며 강한 양의 선형관계가 있다는 것을 알 수 있다. 즉 초기 혈당량 수치가 크면 1달 후 혈당량 수치도 평균적으로 크다.

```
ggplot(diabetes, aes(baseline, response))+geom_point(aes(colour = treat))
```



```
ggplot(diabetes, aes(baseline, response))+geom_point(aes(colour = treat)) +  
  facet_wrap("treat")
```



이제 공변량을 사용하지 않는 분산분석 모형을 적합해 보자. 처리 간에 혈당의 평균은 유의한 차이가 없다.

```
diab1 <- lm(response~treat, data=diabetes )
summary(diab1)
```

Call:

```
lm(formula = response ~ treat, data = diabetes)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.225	-1.781	-0.800	2.306	5.275

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.475	1.634	21.094	1.46e-12 ***
treat2	-2.825	2.311	-1.222	0.240469
treat3	-3.625	2.311	-1.568	0.137642
treat4	-2.250	2.311	-0.973	0.345753
treat5	-9.450	2.311	-4.089	0.000968 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.269 on 15 degrees of freedom

Multiple R-squared: 0.5532, Adjusted R-squared: 0.434

F-statistic: 4.643 on 4 and 15 DF, p-value: 0.01224

```
anova(diab1)
```

Analysis of Variance Table

Response: response

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treat	4	198.41	49.602	4.6425	0.01224 *
Residuals	15	160.26	10.684		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

다음으로 치료 전의 혈당을 공변량으로 포함한 공분석 결과를 적합해 보자.

Caution

우리가 지금까지 사용한 `anova()` 함수를 이용하면 소위 Type I 제곱합(type I sum of squares)을 이용한 분산분석을 제공한다.

모형이 두 개 이상의 요인을 가지는 경우에 분산분석을 적용하는 경우 `anova()` 함수의 Type I 제곱합은 고려하는 요인의 순서에 따라서 제곱합의 값이 달라지기 때문에 우리가 원하는 검정을 실시할 수 없다.

모형이 두 개 이상의 요인을 가지는 경우는 패키지 `car` 의 `Anova()` 함수를 이용해야 하며, 선택문으로 `type="III"` 를 사용해야 한다. 이렇게 선택문 `type="III"` 으로 `Anova()` 함수를 이용해야만 요인의 순서에 관계없이 각 요인의 순수 효과만을 이용하여 가설 검정을 할 수 있다.

요인이 하나인 경우는 요인의 순서에 상관이 없으므로 `anova()` 함수를 사용해도 무방하다.

```
diab2 <- lm(response ~ baseline + treat, data=diabetes )
summary(diab2)
```

Call:

```
lm(formula = response ~ baseline + treat, data = diabetes)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.1360	-1.0024	-0.2827	0.7257	6.0806

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.9437	4.8219	2.892	0.011834 *
baseline	0.7534	0.1723	4.373	0.000637 ***
treat2	-2.7685	1.5554	-1.780	0.096793 .
treat3	-1.6660	1.6186	-1.029	0.320776
treat4	-1.6284	1.5618	-1.043	0.314787
treat5	-4.5903	1.9115	-2.401	0.030788 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.2 on 14 degrees of freedom

Multiple R-squared: 0.8112, Adjusted R-squared: 0.7437

F-statistic: 12.03 on 5 and 14 DF, p-value: 0.0001164

```
Anova(diab2, type="III")
```

Anova Table (Type III tests)

Response: response

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	40.457	1	8.3621	0.0118339 *
baseline	92.528	1	19.1248	0.0006369 ***

6. 공분산분석

```
treat      34.188  4  1.7666 0.1916720
Residuals  67.734 14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

분산분석 모형에서는 평균 잔차제곱합이 $MSE = 10.684$ 이지만 공분산분석에서는 $MSE = 67.734/14 = 4.838$ 로 감소하였다.

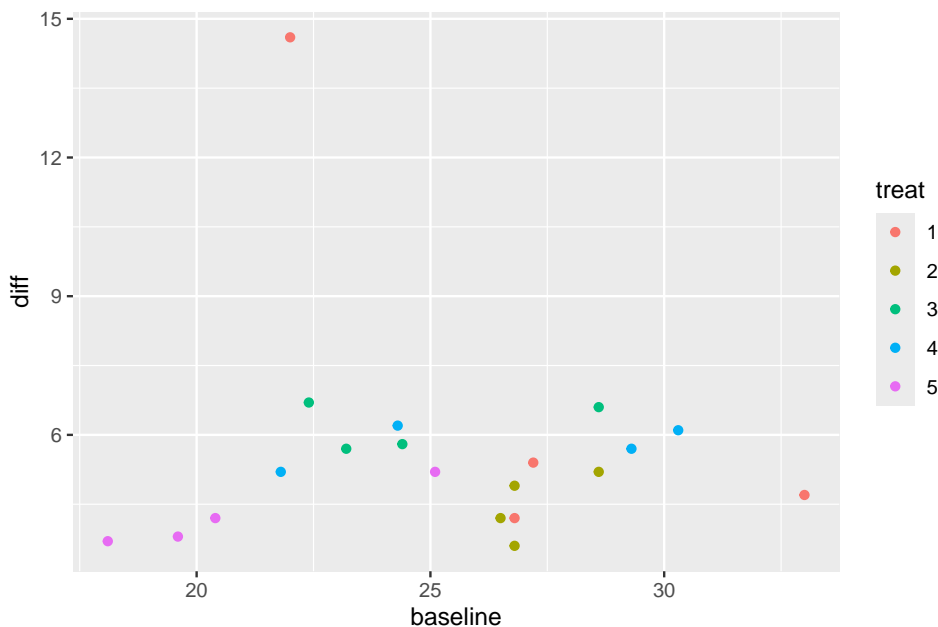
하지만 치료약의 효과를 보면 공분산분석에서는 유의하지 않게 나타났다. 이는 공변량인 치료 전의 혈당이 치료 후의 혈당을 설명하는데 매우 유의한 요인이기 때문이다. 즉 치료 후의 혈당의 집단별 변동이 큰 이유는 치료 효과때문이 아니라 치료 전의 혈당의 차이 때문으로 보여진다.

이러한 결과의 교훈은 여러 개의 집단을 비교하는 경우, 처리를 배정할 때 임의화 방법(randomization)을 사용하여 실험 전 환자들의 인구적 특성과 증상의 정도가 처리그룹간에 큰 차이가 없이 만들어야 한다. 우리가 살펴본 저혈당 임상실험은 처리를 비교하기 위한 공정한 실험이라고 보기 힘들다.

마지막으로 분석에 사용하는 반응변수를 치료 후 반응값이 아닌 기준값에서의 변화량(**change from baseline**)으로 사용할 수 있다.

이제 다음과 치료 전과 후의 변화량을 반응변수로 하고 공분산분석을 적용해보자.

```
diabetes2 <- diabetes %>% dplyr::mutate(diff = response - baseline)
ggplot(diabetes2, aes(baseline, diff)) +
  geom_point(aes(colour = treat))
```



```
diab3 <- lm(diff~ baseline + treat, data=diabetes2 )
summary(diab3)
```


Call:

```
lm(formula = diff ~ baseline + treat, data = diabetes2)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.1360	-1.0024	-0.2827	0.7257	6.0806

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.9437	4.8219	2.892	0.0118 *
baseline	-0.2466	0.1723	-1.431	0.1743
treat2	-2.7685	1.5554	-1.780	0.0968 .
treat3	-1.6660	1.6186	-1.029	0.3208
treat4	-1.6284	1.5618	-1.043	0.3148
treat5	-4.5903	1.9115	-2.401	0.0308 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.2 on 14 degrees of freedom

Multiple R-squared: 0.3387, Adjusted R-squared: 0.1025

F-statistic: 1.434 on 5 and 14 DF, p-value: 0.2723

```
Anova(diab3, type="III")
```

Anova Table (Type III tests)

Response: diff

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	40.457	1	8.3621	0.01183 *
baseline	9.908	1	2.0480	0.17435
treat	34.188	4	1.7666	0.19167
Residuals	67.734	14		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

위의 분석을 보면 변화량을 반응변수로 하면 기준값과 처리 모두 유의하지 않음을 알 수 있다.

6.6. 예제: 산소운반능력

부교재 Jaewon Lee (2005) 의 5장에 나오는 산소운반 능력에 대한 실험 자료를 분석해 보자.

6. 공분산분석

흡연자를 대상으로 산소운반 능력을 개선하는 서로 다른 치료제의 효과를 비교하는 실험이다. 산소운반 능력에 영향을 미치는 중요한 변수로 1인당 담배소비량이 고려되어 이를 공변량으로 포함하여 공분산분석을 실시하려고 한다.

다음과 같이 자료를 읽으며 변수의 설명은 다음과 같다.

- `treat` : 치료법
- `cigar` : 1인당 담배 소비량
- `oxy` : 산소운반능력

```
oxygen <- read.csv(here("data","chapter-5-data-2.txt"), sep=' ', header = T)
oxygen$treat<- factor(oxygen$treat)
oxygen <- oxygen %>% dplyr::rename(cigar = x, oxy = y)
```

산소운반능력에 대한 임상실험에서 얻은 자료는 다음과 같다.

```
oxygen %>%
  kbl(caption = "산소운반능력 임상실험 결과") %>%
  kable_styling(bootstrap_options = "striped", full_width = F, position = "center",
    font_size = 12)
```

치료집단 별로 치료 후 산소운반능력의 분포를 다음과 상자그림으로 비교해보자.

```
dfsumm2 <- oxygen %>%
  group_by(treat) %>%
  summarise(mean = mean(oxy), median = median(oxy), sd = sd(oxy), min = min(oxy),
    max = max(oxy))
dfsumm2
```

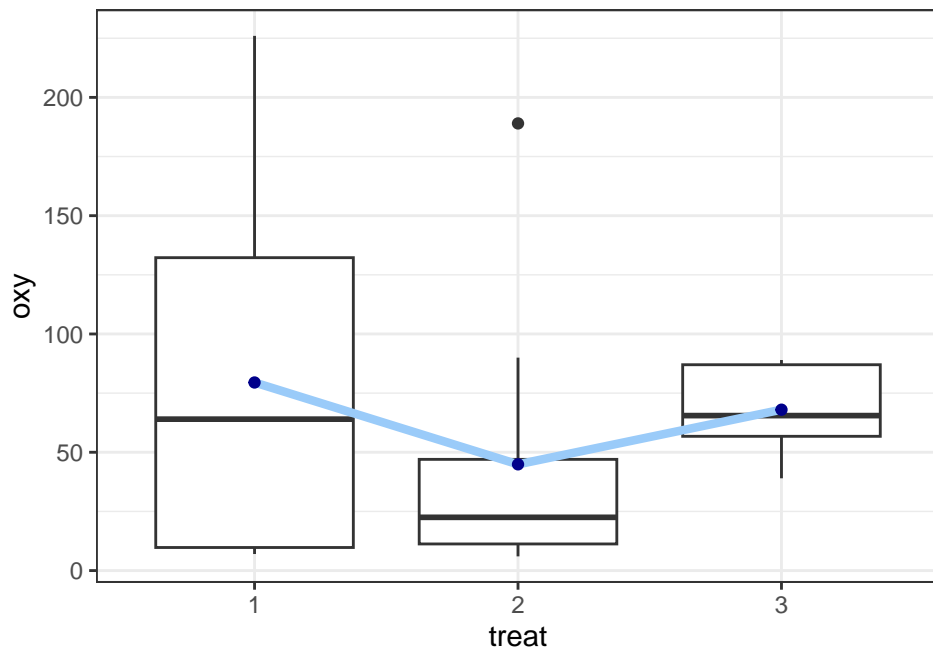
```
# A tibble: 3 x 6
  treat mean median    sd  min  max
  <fct> <dbl>  <dbl> <dbl> <int> <int>
1 1      79.5    64   77.7     7  226
2 2      44.9   22.5  56.7     6  189
3 3       68   65.5  18.7    39   89
```

```
ggplot(oxygen, aes(treat, oxy)) +
  geom_boxplot() +
  geom_line(data=dfsumm2, aes(x=treat, y=mean, group=1), linewidth=1.5, col="#9ACBF9") +
  geom_point(data=dfsumm2, aes(x=treat, y=mean), col="darkblue") +
  theme_bw()
```

표 6.2.: 산소운반능력 임상실험 결과

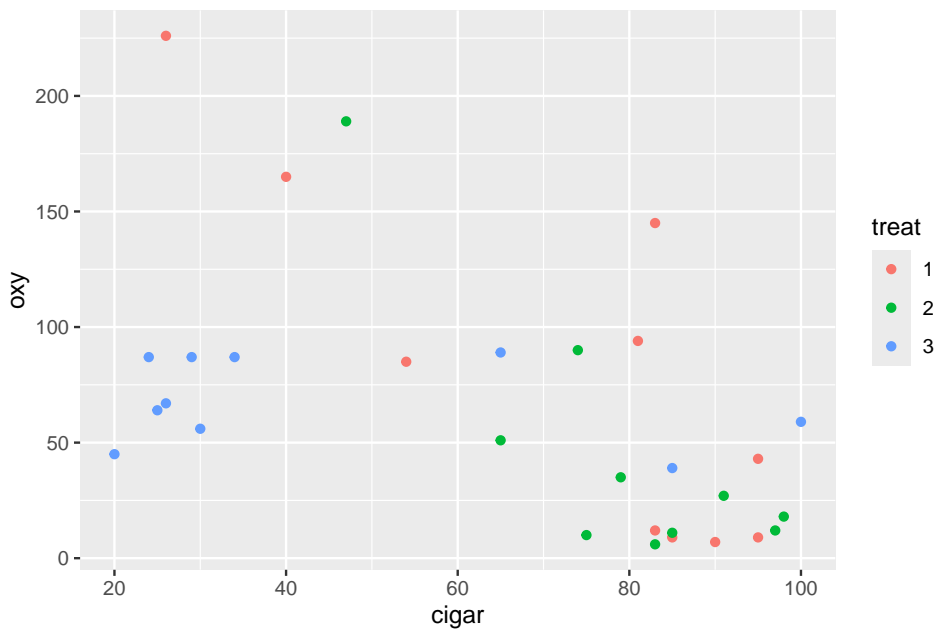
treat	cigar	oxy
1	40	165
1	54	85
1	85	9
1	95	43
1	81	94
1	26	226
1	90	7
1	95	9
1	83	12
1	83	145
2	85	11
2	83	6
2	65	51
2	98	18
2	47	189
2	74	90
2	75	10
2	97	12
2	79	35
2	91	27
3	65	89
3	25	64
3	34	87
3	20	45
3	30	56
3	29	87
3	100	59
3	85	39
3	24	87
3	26	67

6. 공분산분석



이제 1인당 담배소비량과 산소운반 능력의 관계는 다음과 같은 산점도로 나타낼 수 있다.

```
ggplot(oxygen, aes(cigar, oxy))+
  geom_point(aes(colour = treat))
```



1인당 담배소비량과 산소운반 능력의 관계를 상관계수로 구해보면 음의 상관관계를 보이고 있으면 1인당 담배소비량이 증가하면 산소운반 능력이 감소하는 것을 알 수 있다.

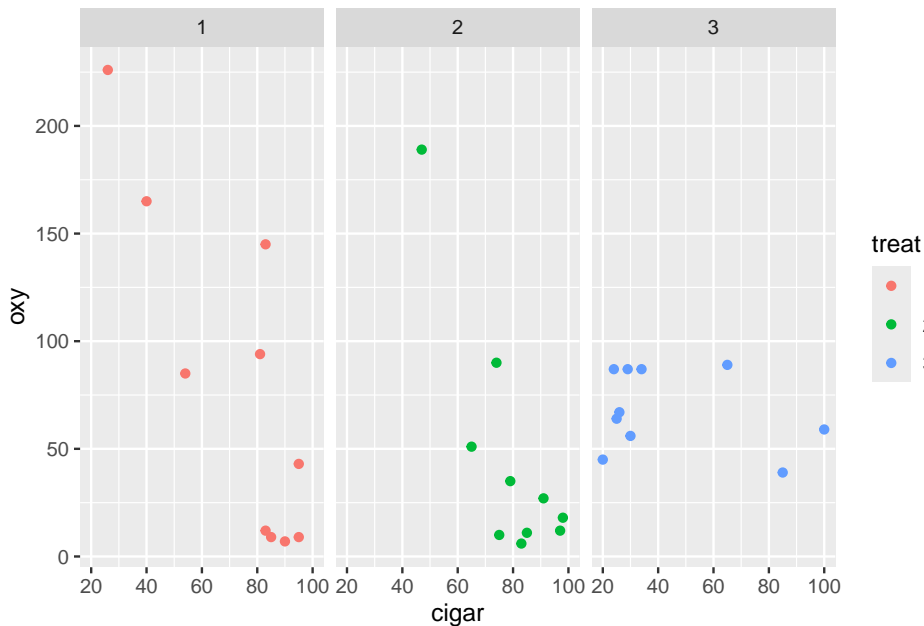
```
cor(oxygen$cigar, oxygen$oxy)
```

```
[1] -0.5349395
```

6. 공분산분석

이제 조금 더 자세하고 처리 그룹별로 산소운반 능력과 1인당 담배소비량의 관계를 살펴보자. 아래 그림을 보면 산소운반 능력과 1인당 담배소비량의 관계가 처리 그룹별로 다르게 나타나는 것을 알 수 있다. 처리 그룹 1 과 2에서는 1인당 담배소비량이 증가하면 산소운반 능력이 감소하는 것을 알 수 있지만 처리 그룹 3에서는 그렇지 않다.

```
ggplot(oxygen, aes(cigar, oxy)) +  
  geom_point(aes(colour = treat)) +  
  facet_wrap("treat")
```



이제 공변량을 사용하지 않는 분산분석 모형을 적합해 보자. 다음 결과를 보면 처리 간에 산소운반 능력이 유의한 차이가 없다.

```
oxy1 <- lm(oxy~treat, data=oxygen )  
anova(oxy1)
```

Analysis of Variance Table

Response: oxy

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treat	2	6210	3105.0	0.9696	0.3921
Residuals	27	86465	3202.4		

다음으로 1인당 담배소비량을 공변량으로 포함한 공분석 결과를 적합해 보자.

```
oxy2 <- lm(oxy ~ treat + cigar, data=oxygen )  
summary(oxy2)
```

6. 공분산분석

Call:

```
lm(formula = oxy ~ treat + cigar, data = oxygen)
```

Residuals:

Min	1Q	Median	3Q	Max
-57.974	-32.116	-8.566	30.148	96.489

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	187.0662	30.9742	6.039	2.22e-06 ***
treat2	-25.4892	20.5656	-1.239	0.226266
treat3	-54.7028	23.2087	-2.357	0.026234 *
cigar	-1.4695	0.3743	-3.926	0.000567 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.69 on 26 degrees of freedom

Multiple R-squared: 0.4143, Adjusted R-squared: 0.3467

F-statistic: 6.13 on 3 and 26 DF, p-value: 0.002697

```
Anova(oxy2, type="III")
```

Anova Table (Type III tests)

Response: oxy

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	76152	1	36.4746	2.224e-06 ***
treat	11873	2	2.8433	0.0764293 .
cigar	32183	1	15.4146	0.0005667 ***
Residuals	54283	26		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

위에서 공변량을 포함한 공분산분석 결과를 보면 담배소비량이 산소운반 능력을 설명하는 유의한 변수이며 음의 기울기를 나타내는 것을 알 수 있다.

처리 간에 산소운반 능력이 유의 수준 5% 로는 유의한 차이가 없다고 나타났지만 공변량이 없는 위의 모형보다 p-값이 크게 감소한것을 알 수 있다 .

유의할 점은 공변량이 없는 경우 모분산의 추정값인 평균 잔차제곱합이 $MSE = 3202.4$ 로 추정되고 공변량이 포함되면 $MSE = 54283/26 = 2087.8$ 이다. 이는 1인당 담배소비량이 산소운반 능력을 설명하는 유의한 변수이기 때문이다. 따라서 공변량이 포함된 모형에서 처리에 대한 F -검정통계량의 값이 커지는 결과가 나타난다.

6. 공분산분석

이렇게 반응변수를 설명하는데 있어서 유의한 공변량을 포함시키면 일반적으로 처리에 대한 검정력이 높아진다.

1인당 담배소비량과 산소운반 능력의 관계가 각 그룹마다 차이가 있다면 공변량의 효과가 처리집단에 따라서 달라지는 다음 모형을 고려해 보자. 다음 모형에서 공변량에 대한 회귀 계수의 값이 처리집단마다 다른 것($\beta_1, \beta_2, \beta_3$)을 알 수 있다.

$$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + e_{ij}, \quad i = 1, 2, \dots, a, \quad j = 1, 2, \dots, r \quad (6.4)$$

이제 식 6.4 를 이용하여 공분산 분석을 실행해 보자.

공변량의 효과가 처리 집단마다 다르게 놓은 모형 6.4 에서는 처리집단 간의 유의한 차이가 나타난다 (p-값 = 8.272e-05). 평균 잔차제곱합도 $MSE = 29146/24 = 1214.4$ 으로 줄어든 것을 알 수 있다.

```
oxy3 <- lm(oxy ~ treat + treat*cigar, data=oxygen )
summary(oxy3)
```

Call:

```
lm(formula = oxy ~ treat + treat * cigar, data = oxygen)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-48.179	-25.812	-3.413	20.850	91.544

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	274.0363	36.6575	7.476	1.03e-07 ***
treat2	10.4882	71.0085	0.148	0.883811
treat3	-198.4239	42.1708	-4.705	8.79e-05 ***
cigar	-2.6576	0.4776	-5.564	1.00e-05 ***
treat2:cigar	-0.3603	0.8919	-0.404	0.689783
treat3:cigar	2.4838	0.6256	3.970	0.000568 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.85 on 24 degrees of freedom

Multiple R-squared: 0.6855, Adjusted R-squared: 0.62

F-statistic: 10.46 on 5 and 24 DF, p-value: 2.031e-05

```
Anova(oxy3, type="III")
```

Anova Table (Type III tests)

6. 공분산분석

Response: oxy

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	67868	1	55.884	1.026e-07	***
treat	34648	2	14.265	8.272e-05	***
cigar	37600	1	30.961	1.004e-05	***
treat:cigar	25137	2	10.349	0.0005742	***
Residuals	29146	24			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

7. 다중비교

7.1. 필요한 패키지

```
library(tidyverse)
library(kableExtra)
library(ggplot2)
library(here)

library(car)
library(agricolae)
library(DescTools)
```

7.2. 일원배치에서 평균의 비교

분산분석표를 이용한 F-검정으로 귀무가설을 기각하면 모든 처리 수준의 평균이 같지 않다는 결론을 내리고 어떤 집단 간에 평균의 차이가 유의한지 더 분석해야 한다. 평균 차이에 대한 신뢰구간과 가설 검정은 아래와 같이 주어진다.

두 수준 평균의 차이 $\delta_{ij} = \mu_i - \mu_j$ 에 대한 $100(1 - \alpha) \%$ 신뢰구간은 다음과 같이 주어진다.

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm t(1 - \alpha/2, \phi_E) \sqrt{\frac{2MS_E}{r}} \quad (7.1)$$

두 평균의 차이 δ_{ij} 에 대한 가설을 검정하는 유의 수준 α 에서 다음과 같은 조건을 만족하면 위의 귀무가설을 기각한다.

$$|\bar{y}_{i.} - \bar{y}_{j.}| > t(1 - \alpha/2, \phi_E) \sqrt{\frac{2MS_E}{r}} \quad (7.2)$$

식 7.2에서 검정을 위한 조건의 우변을 최소유의차(least significant difference; LSD)라고 부른다. 두 수준의 차이가 유의하려면 두 평균 차이의 절대값이 최소한 최소유의차의 값보다 커야한다.

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

7.3. 두 개 이상의 가설

일원배치 계획에서 수준의 개수가 a 개 인 경우 처리 수준들의 차이에 대하여 비교를 한다면 $\binom{a}{2}$ 개의 가설검정을 수행해야 한다. 예를 들어 처리 수준이 3개 있는 경우 다음과 같이 3개의 조합에 대하여 가설 검정을 수행할 수 있다.

$$H_{01} : \mu_1 = \mu_2, \quad H_{02} : \mu_2 = \mu_3, \quad H_{03} : \mu_3 = \mu_1 \quad (7.3)$$

가설검정에서 사용되는 유의수준(significance level, α)에 대하여 생각해 보자. 지금까지 가설검정을 수행할 때 **유의수준 5%** 라는 말을 사용해 왔는데 이것이 무슨 의미를 가지는지 알아보자.

유의수준 5%라는 것은 수행하는 가설검정에서 귀무가설이 옳은 경우에 기각하는 확률을 말한다. 예를 들어 7.3의 3개의 검정에 대하여 각각 t-검정을 수행하는 경우 귀무가설이 옳은데 우연하게 자료가 극단적으로 나와서 귀무가설을 기각하고 대립가설을 채택하는 확률이 유의수준이며 보통 5%를 사용한다. 이러한 오류를 제 1종의 오류(Type I error; false discovery error; false positive error)라고 한다.

Note

제 1종의 오류(Type I Error): 실제로는 유의하지 않지만 검정 결과 유의하다고 판단하는 경우이며 다음과 같은 다른 이름으로는 불린다.

- False discovery error (FDE, 거짓 양성 오류)
- False positive error (FPE)

위 7.3에서 처럼 3개의 가설 검정을 동시에 실시한다면 각각의 가설검정에서 제 1종의 오류를 범할 확률은 5%이다. 그런데 3개의 가설 검정을 동시에 실행하므로 다음과 같이 3개의 검정을 합쳐서 다음과 같은 확률에 관심이 있을 수 있다.

3개의 가설검정을 동시에 수행할 때 제 1종의 오류가 최소한 1번 발생할 확률은 얼마인가?

세 개의 가설검정을 동시에 수행하는 경우 세 검정 모두 제 1종의 오류를 범하거나 두 개 또는 하나의 검정에서 제 1종의 오류를 범할 사건의 확률은 얼마나 될까? 5%보다 작을까 아니면 클까? 또는 5%인가? 간단한 확률 공식을 이용하여 알아보자.

7.4. 실험단위 오류

일단 두 개의 검정 H_{01} 과 H_{02} 을 각각 유의수준 $\alpha = 0.05$ 로서 동시에 수행 한다고 가정하고 다음과 같은 사건을 정의한다.

- A_1 : H_{01} 검정에서 제 1종의 오류를 범하는 사건
- A_2 : H_{02} 검정에서 제 1종의 오류를 범하는 사건

7. 다중비교

각 검정에서 제 1 종의 오류를 범할 확률을 α 라고 가정하자.

$$P(A_1) = P(A_2) = \alpha = 0.05$$

이제 두개의 가설검정을 동시에 수행하는 경우 제 1 종의 오류를 최소한 1번 범하는 사건은 $P(A_1 \cup A_2)$ 이며 여사건의 확률공식을 이용하면 다음과 같이 나타낼 수 있다.

$$P(A_1 \cup A_2) = 1 - P(A_1^c \cap A_2^c)$$

여기서 우리는 $P(A_1^c) = P(A_2^c) = 1 - 0.05 = 0.95$ 를 알 수 있지만 두 사건의 교집합에 대한 확률은 계산하기 쉽지 않다. 왜냐하면 두 사건 A_1 과 A_2 가 일반적으로 독립이 아니어서 두 확률의 곱으로 쉽게 나타낼 수 없다.

만약에 두 사건이 독립이라면 다음과 같은 결과가 나온다. 즉 두 개의 독립인 가설검정을 동시에 수행하는 경우 최소한 1번의 제 1 종의 오류를 범하는 사건의 확률은 0.0975로 5%의 두 배 정도가 된다.

$$P(A_1 \cup A_2) = 1 - P(A_1^c \cap A_2^c) = 1 - P(A_1^c)P(A_2^c) = 1 - (1 - 0.05)^2 = 0.0975 > 0.05$$

만약 k 개의 독립인 가설검정을 동시에 수행하는 경우 제 1 종의 오류를 최소한 1번이라도 범하는 사건의 확률은 $1 - (1 - 0.05)^k$ 으로 급격하게 증가한다. 예를 들어 $k = 6$ 인 경우 26.5%로 5%의 5 배가 된다. 여기서 유의할 점은 이러한 결과는 모든 가설검정이 독립이고 여러 개의 가설검정들을 동시에 고려하는 경우이다.

즉, 두 개 이상의 가설검정을 동시에 고려해서 제 1 종의 오류를 최소한 1번 범할 경우를 오류라고 한다면 그 확률은 고려하는 검정의 개수가 증가함에 따라 빠르게 커진다.

이렇게 두 개 이상의 가설검정을 동시에 고려해서 계산하는 오류의 확률을 실험단위 오류(**Experiment-wise error** 또는 **Family-wise error**)라고 하며 반대로 가설검정을 동시에 고려하지 않고 개별적으로 생각하는 오류를 개별단위 오류(**Individual-wise error**)라고 한다.

Example 7.1 (제어집단이 있는 임상실험). 임상실험에서 신약(처리 1)의 효과가 위약(처리 2)보다는 우월하다는 사실을 입증하는 것이 일반적이다. 그런데 기존의 약(처리 3)보다 우월하다는 사실을 동시에 입증하려고 하는 경우도 있다. 이러한 경우 다음과 같은 두 개의 가설을 동시에 수행해야 한다.

$$H_{01} : \mu_1 = \mu_2, \quad H_{02} : \mu_1 = \mu_3$$

이러한 경우 신약(처리 1)을 제어 집단이라고 부르며 다른 두 그룹들(처리 2,3)과 각각 비교해야 한다.

3개의 집단(신약, 위약, 기존의 약)을 가진 일원배치법으로 실험을 수행한 경우 첫 번째 가설 H_{01} 은 $\bar{y}_1 - \bar{y}_2$ 를 이용하고 두 번째 가설 H_{02} 은 $\bar{y}_1 - \bar{y}_3$ 을 이용하여 가설검정을 한다.

이러한 경우 각 검정에 대하여 유의 수준을 5% (개별단위 오류를 범할 확률이 5%)라고 해도 실험단위 오류를 범할 확률은 5%보다 크다.

7.5. 다중비교

다시 실험 단위 오류의 계산으로 돌아가서 만약에 두 사건이 독립이 아닌 경우에 실험적 오류를 통제할 수 있는, 즉 5%보다 작거나 같게 하는 방법에 대해서 알아보자 두 사건이 독립이 아닌 일반적인 경우에 확률 공식을 이용하여 다음과 같은 부등식을 얻을 수 있다.

$$P(A_1 \cup A_2) \leq P(A_1) + P(A_2) = (2)(0.05) = 0.1$$

위의 결과를 보면 만약에 두 개의 가설검정을 동시에 수행하는 경우 각 가설검정에 대한 개별단위의 제 1 종 오류에 대한 확률을 반으로 줄이면 $(0.05/2=0.025)$ 실험적 오류가 5%보다 작거나 같게 된다.

$$P(A_1 \cup A_2) \leq P(A_1) + P(A_2) = (2)(0.05/2) = 0.05$$

위에서 보인 같은 논리로서 k 개의 가설검정을 동시에 수행하는 경우 각 가설검정에 대한 개별적 1종 오류의 확률을 k 배 줄이면 $(0.05/k)$ 실험단위 오류가 5%보다 작거나 같게 된다.

$$P(A_1 \cup A_2 \cup \dots \cup A_k) \leq (k)(0.05/k) = 0.05$$

여기서 한 가지 유의할 점은 만약 두 개의 가설이 완전히 종속이거나 $(A_1 = A_2)$ 거의 종속이면 실험적 오류는 거의 변하지 않는다. 따라서 개별단위 1종 오류에 대한 수정은 거의 필요하지 않다.

$$P(A_1 \cup A_2) = 1 - P(A_1^c \cap A_2^c) \approx 1 - P(A_1^c) = 0.05$$

이렇게 실험단위 오류를 통제하기 위하여 (5%보다 작거나 같게) 각 가설에 대한 개별단위 1 종 오류의 확률(유의수준)을 보정하는 방법을 **다중비교(multiple comparison)** 라고 한다.

위에서 제시한 개별단위 1종 오류를 k 배로 줄이는 $(0.05/k)$ 방법을 특별하게 본페로니 수정(Bonferroni correction) 이라고 부른다. 본페로니 수정은 가장 보수적인 수정(most conservative correction)이라고 불리는데 그 이유는 실험적 오류가 가질 수 있는 가장 큰 값을 가정하고 보정하기 때문에 각각 수정한 개별단위 오류에 대한 유의수준이 너무 작게 되어 $(0.05/k)$ 귀무가설의 기각이 매우 힘들기 때문이다.

만약 k 개의 가설 검정에 본페로니 수정을 적용한다면 신뢰구간과 가설검정은 다음과 같이 수정된다.

두 수준 평균의 차이 $\delta_{ij} = \mu_i - \mu_j$ 에 대한 본페로니 수정 신뢰구간은 다음과 같이 주어진다.

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm t(1 - \alpha/(2k), \phi_E) \sqrt{\frac{2MS_E}{r}} \quad (7.4)$$

두 평균의 차이 δ_{ij} 에 대한 가설을 본페로니 수정 검정은 다음과 같은 조건을 만족하면 귀무가설을 기각한다.

$$|\bar{y}_{i.} - \bar{y}_{j.}| > t(1 - \alpha/(2k), \phi_E) \sqrt{\frac{2MS_E}{r}} \quad (7.5)$$

7. 다중비교

기각역에 본페로니 수정을 하는 것은 원래의 p-값에 가설의 개수 k 를 곱하여 수정 p-값을 사용하는 것과 같다.

$$\text{Bonferoni adjusted p-value} = k \times \text{unadjusted p-value} \quad (7.6)$$

일반적으로 각 가설검정들은 완전히 독립도 아니고 또한 완전한 종속도 아니다. 따라서 실험단위 오류는 각 가설 검정들이 어떻게 확률적으로 관련되어 있느냐에 따라 매우 달라진다. 이러한 이유로 인하여 다중비교의 방법은 매우 다양하며, 선택한 방법에 따라서 검정의 결과도 매우 달라질 수있는 사실에 유의해야 한다. 다중비교의 방법을 선택하는 것은 매우 어려운 일이다.

가설이 2개 이상 있는 경우 실험단위의 오류의 확률을 제어해야 하는지에 대한 판단은 상황에 따라서 달라진다.

앞에서 살펴본 임상실험의 예와 같이 **중요한 의사 결정을** 동시에 수행하는 2개 이상의 검정 결과에 따라서 해야할 경우 주로 다중 비교를 적용한다.

또한 다중 비교 방법은 실험의 설계와 목적에 따라서 많은 방법들이 존재한다. 주어진 실험 계획과 목적에 부합하는 다중 비교법을 선택해야 한다.

반면 탐색적인 목적으로 여러 개의 가설 검정을 동시에 수행하는 경우에는 다중비교를 적용하지 않거나 다중 비교 보다 더 유연한 False Discovery Rate 방법(참조) 을 사용한다.

7.6. 다중비교 방법

이제 다중비교를 수행할 수 있는 중요한 방법들을 알아보고 저혈당 임상실험 예제에 적용해 보자.

앞에서 살펴본 저혈당 임상실험은 5개의 처리가 있다. 따라서 $\binom{5}{2} = 10$ 개의 가설 검정(또는 신뢰구간)을 수행해야 한다.

5개 치료집단(`treat`)이 처리 수준이며 각 처리수준 은 1, 2, 3, 4, 5 로 표시된다.

7.6.1. 다중비교 방법을 적용하지 않는 경우

먼저 다중비교 방법을 적용하지 않는 경우 결과를 보자. 함수 `LSD.test` 에서 `p.adj=c("none")` 를 지정하면 다중 비교를 적용하지 않는다. 명령문 `p.adj` 를 지정하지 않으면 수정을 하지 않는 LSD 방법에 의한 신뢰 구간 7.1 와 검정 방법 7.2 로 구한 결과를 준다.

LSD 방법을 적용하는 경우 유의한 차이를 보이는 조합이 4개로 나타났다 (1-5,2-5,3-5,4-5).

```
diabetes <- read.csv(here("data", "chapter-5-data.txt"), sep=' ', header = F)
colnames(diabetes) <- c("treat", "baseline", "response")
diabetes$treat <- factor(diabetes$treat)
diabetes <- diabetes %>% arrange(treat)
```

7. 다중비교

```
anova.res <- aov(response~treat,data=diabetes) #일원배치
test1 <- LSD.test(anova.res, "treat", alpha = 0.05, group = FALSE, console = FALSE, p.adj=c("n
test1$comparison
```

	difference	pvalue	signif.	LCL	UCL
1 - 2	2.825	0.2405		-2.1014081	7.751408
1 - 3	3.625	0.1376		-1.3014081	8.551408
1 - 4	2.250	0.3458		-2.6764081	7.176408
1 - 5	9.450	0.0010	***	4.5235919	14.376408
2 - 3	0.800	0.7341		-4.1264081	5.726408
2 - 4	-0.575	0.8069		-5.5014081	4.351408
2 - 5	6.625	0.0118	*	1.6985919	11.551408
3 - 4	-1.375	0.5608		-6.3014081	3.551408
3 - 5	5.825	0.0235	*	0.8985919	10.751408
4 - 5	7.200	0.0071	**	2.2735919	12.126408

7.6.2. 본페로니 수정(Bonferroni correction)

이제 다중비교 방법 중에 가장 보수적인 본페로니 수정(Bonferroni correction)을 적용해 보자. 함수 `LSD.test` 에서 `p.adj=c("bonferroni")`를 이용한다.

아래의 결과는 본페로니 수정 방법에 의한 신뢰 구간 7.4 와 검정 방법 7.5 으로 구한 결과이다.

본페로니 수정이 적용된 신뢰구간은 LSD 방법의 신뢰구간보다 길며 수정된 p-값 7.6 은 LSD 방법으로 구한 값의 10배이다.

LSD 방법을 적용하는 경우 유의한 차이를 보이는 조합이 4개로 나타났는데(1-5,2-5,3-5,4-5) 본페로니 수정을 적용한 경우에는 1개로 줄어 들었다(1-5).

수정한 p-값이 1이 초과하면 확률이기 때문에 1로 주어진다.

```
test2 <- LSD.test(anova.res, "treat", alpha = 0.05, group = FALSE, console = FALSE, p.adj=c("b
test2$comparison
```

	difference	pvalue	signif.	LCL	UCL
1 - 2	2.825	1.0000		-4.7700036	10.420004
1 - 3	3.625	1.0000		-3.9700036	11.220004
1 - 4	2.250	1.0000		-5.3450036	9.845004
1 - 5	9.450	0.0097	**	1.8549964	17.045004
2 - 3	0.800	1.0000		-6.7950036	8.395004
2 - 4	-0.575	1.0000		-8.1700036	7.020004
2 - 5	6.625	0.1177		-0.9700036	14.220004
3 - 4	-1.375	1.0000		-8.9700036	6.220004

7. 다중비교

```
3 - 5      5.825 0.2355      -1.7700036 13.420004
4 - 5      7.200 0.0709      . -0.3950036 14.795004
```

7.6.3. Tukey의 HSD

함수TukeyHSD는 분산분석을 실행한 결과를 이용하여 다중비교 방법 중 가장 많이 이용되는 Tukey's Honest Significant Difference (HSD) 방법으로 다중비교를 제공한다.

Tukey의 HSD는 너무 보수적인 결과를 주는 본페로니 수정을 개선한 것이다. 따라서 Tukey의 HSD 에서 얻은 결과는 수정하지 않는 LDS 의 결과와 Bonferoni 방법의 중간에 있다고 할 수 있다.

Tukey의 HSD 에서는 본페로니와 유사하게 2개의 조합(1-5,4-5)만이 유의한 차이가 있다고 나타난다.

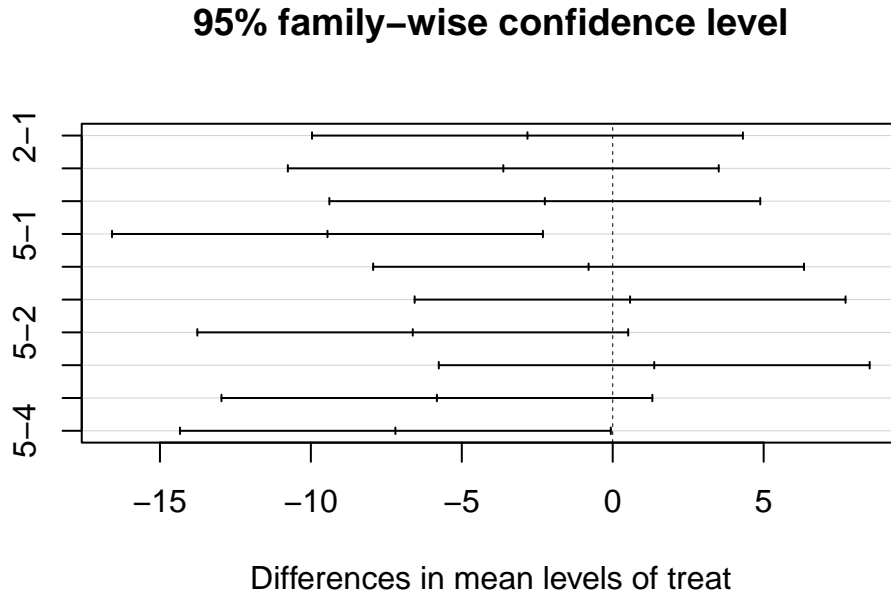
```
test3 <- TukeyHSD(anova.res, conf.level = 0.95, ordered=FALSE)
test3
```

```
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = response ~ treat, data = diabetes)
```

```
$treat
      diff      lwr      upr    p adj
2-1 -2.825  -9.962104  4.31210359 0.7391923
3-1 -3.625 -10.762104  3.51210359 0.5376665
4-1 -2.250  -9.387104  4.88710359 0.8628887
5-1 -9.450 -16.587104 -2.31289641 0.0073287
3-2 -0.800  -7.937104  6.33710359 0.9965642
4-2  0.575  -6.562104  7.71210359 0.9990516
5-2 -6.625 -13.762104  0.51210359 0.0751342
4-3  1.375  -5.762104  8.51210359 0.9737551
5-3 -5.825 -12.962104  1.31210359 0.1380280
5-4 -7.200 -14.337104 -0.06289641 0.0475257
```

```
plot(test3)
```



7.6.4. FDR 방법

False Discovery Rate (FDR)는 다중 비교에서 기각된 가설 중 실제로 제 1 종 오류가 일어난 비율, 즉 실제 거짓 양성인 비율을 의미한다.

수식적으로 FDR은 다음과 같이 표현할 수 있다.

$$\text{FDR} = \mathbb{E} \left[\frac{V}{R} \right] \quad (R > 0), \quad (7.7)$$

위의 7.7 에서 V 와 R 은 다음과 같이 정의되며

- V : 거짓 양성(False Positives)의 수
- R : 유의하다고 판단된 전체 가설의 수

FDR은 기대값을 의미하며, 여기서는 비율의 평균을 의미한다.

FDR을 제어하기 위해 여러 가지 방법이 제안되었으며 이 중 가장 널리 사용되는 방법은 Benjamini-Hochberg 절차(BH FDR) 이고 검정의 절차는 다음과 같다.

먼저 m 개의 가설 검정을 한다고 가정하자.

1. 유의 수준 α 에 대해, 모든 p-값을 오름차순으로 정렬한다. 정렬된 p-값 $p_{(i)}$ 에 해당하는 귀무가설을 $H_{(i)}$ 라고 하자.

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$$

2. 오름차순으로 정렬한 경우 i 번째 p-값 $p_{(i)}$ 에 해당하는 새로운 유의수준 α_i^* 을 다음과 같이 계산한다.

7. 다중비교

$$\alpha_i^* = \frac{i}{m}\alpha$$

3. $p_{(i)} \leq \alpha_i^*$ 를 만족하는 가장 큰 $i = i^*$ 를 찾는다.

$$i^* = \max\{i : p_{(i)} \leq \alpha_i^*\}$$

4. i^* 보다 작거나 같은 p-값을 유의하다고 판단한다. 즉, 가설 $H_{(1)}, H_{(2)}, \dots, H_{(i^*)}$ 를 기각한다.

정렬된 p-값 중에서 $p_{(i^*)}$ 까지의 가설을 기각하고 나머지 가설은 기각하지 않는다.

이러한 BH 절차는 FDR을 유의수준 α 이하로 유지하도록 설계된 방법이다. BH 절차는 본페로니 수정보다 덜 보수적이며, 거짓 발견의 비율을 직접 제어하므로 검정력을 유지할 수 있다는 장점이 있다.

FDR 조정은 생물학, 유전학, 심리학 등 여러 분야에서 사용되며, 특히 유전자 발현 연구와 같은 대량 데이터 분석에서 주로 사용된다. 예를 들어 유전자 발현 데이터에서 수천 개의 유전자에 대해 차이가 있는지 검정하는 경우 FDR 조정을 통해 거짓 양성을 제어할 수 있다. 하지만, 가설들 간의 상관관계에 민감하게 작동할 수 있다.

이제 위에서 설명한 방법으로 저혈당 임상실험에 FDR을 적용해 보자.

아래 코드는 LSD 방법으로 계산된 p-값을 작은 순서대로 정렬하고 BH 절차를 통하여 새로운 유의수준 α_i (BH_alpha) 를 계산하여 FDR을 적용한 결과를 보여준다.

10개의 가설 검정 중 BH 절차를 적용한 결과는 3개의 가설이 유의하다고 판단되었다.

```
ordinary_p <- test1$comparison %>% dplyr::select(pvalue)

ordered_p <- ordinary_p %>% arrange(pvalue)
ordered_p <- ordered_p %>% dplyr::mutate(BH_alpha = 0.05* (1:length(pvalue))/length(pvalue))

ordered_p
```

	pvalue	BH_alpha
1 - 5	0.0010	0.005
4 - 5	0.0071	0.010
2 - 5	0.0118	0.015
3 - 5	0.0235	0.020
1 - 3	0.1376	0.025
1 - 2	0.2405	0.030
1 - 4	0.3458	0.035
3 - 4	0.5608	0.040
2 - 3	0.7341	0.045
2 - 4	0.8069	0.050

7. 다중비교

stats 패키지에 p.adjust 함수의 BH 방법을 활용하면 FDR 제어 방법을 적용한 수정된 p-값을 계산할 수 있다. FDR 제어 방법을 적용한 수정된 p-값은 원래 p-값보다 크며 0.05 보다 작으면 가설을 기각할 수 있다.

```
FDR_BH_p <- p.adjust(ordinary_p$pvalue, method = "BH")
FDR_BH_p
```

```
[1] 0.40083333 0.27520000 0.49400000 0.01000000 0.80690000 0.80690000
[7] 0.03933333 0.70100000 0.05875000 0.03550000
```

```
new_p <- ordinary_p %>% dplyr::mutate(pvalue_FDR = FDR_BH_p)
new_p
```

		pvalue	pvalue_FDR
1	- 2	0.2405	0.40083333
1	- 3	0.1376	0.27520000
1	- 4	0.3458	0.49400000
1	- 5	0.0010	0.01000000
2	- 3	0.7341	0.80690000
2	- 4	0.8069	0.80690000
2	- 5	0.0118	0.03933333
3	- 4	0.5608	0.70100000
3	- 5	0.0235	0.05875000
4	- 5	0.0071	0.03550000

위에서 살펴본 수정을 하지 않은 LSD 방법, Tukey의 HSD, FDR의 BH 절차와 본페로니 수정 방법에서 계산된 p-값을 아래 표에서 비교하였다. 각 수정된 p-값이 유의수준 (0.05) 보다 작으면 유의한 차이가 있다고 판단한다.

```
comp_pval <- data.frame(comp_group=rownames(test1$comparison),
LSD = test1$comparison$pvalue,
FDR_BH = p.adjust(test1$comparison$pvalue, method = "BH"),
HSD = round(as.numeric(test3$treat[,4]),4),
Bonf = test2$comparison$pvalue
)
comp_pval
```

	comp_group	LSD	FDR_BH	HSD	Bonf
1	1 - 2	0.2405	0.40083333	0.7392	1.0000
2	1 - 3	0.1376	0.27520000	0.5377	1.0000
3	1 - 4	0.3458	0.49400000	0.8629	1.0000
4	1 - 5	0.0010	0.01000000	0.0073	0.0097
5	2 - 3	0.7341	0.80690000	0.9966	1.0000

7. 다중비교

6	2 - 4	0.8069	0.80690000	0.9991	1.0000
7	2 - 5	0.0118	0.03933333	0.0751	0.1177
8	3 - 4	0.5608	0.70100000	0.9738	1.0000
9	3 - 5	0.0235	0.05875000	0.1380	0.2355
10	4 - 5	0.0071	0.03550000	0.0475	0.0709

위의 표를 수정이 되지 않은 LSD 방법으로 계산된 p-값을 기준으로 정렬하면 다음과 같다.

```
comp_pval %>% dplyr::arrange(LSD)
```

	comp_group	LSD	FDR_BH	HSD	Bonf
1	1 - 5	0.0010	0.01000000	0.0073	0.0097
2	4 - 5	0.0071	0.03550000	0.0475	0.0709
3	2 - 5	0.0118	0.03933333	0.0751	0.1177
4	3 - 5	0.0235	0.05875000	0.1380	0.2355
5	1 - 3	0.1376	0.27520000	0.5377	1.0000
6	1 - 2	0.2405	0.40083333	0.7392	1.0000
7	1 - 4	0.3458	0.49400000	0.8629	1.0000
8	3 - 4	0.5608	0.70100000	0.9738	1.0000
9	2 - 3	0.7341	0.80690000	0.9966	1.0000
10	2 - 4	0.8069	0.80690000	0.9991	1.0000

위 표에서 볼 수 있듯이 다중비교 수정을 적용하지 않는 LSD 방법에서는 4개의 가설이 유의하다고 나타났다. FDR의 BH 절차를 적용하면 3개의 가설이 유의하다고 나타났으며 Tukey의 HSD에서는 2개의 가설이 유의하다고 나타났다. 본페로니 수정 방법에서는 유의한 차이가 하나의 가설에서만 나타났다.

본페로니 수정은 가장 보수적이며 FDR의 BH 절차는 다중비교 방법 중 상당히 유연한 방법이다.

7.6.5. Dunnett 비교

Example 7.1 에서 설명하듯이 임상실험에서는 하나의 기준집단 또는 제어집단(control group)을 다른 여러 개의 집단과 비교하는 경우가 흔하다. 이러한 경우 사용할 수 있는 다중비교 방법이 Dunnett의 방법이다.

Dunnett의 방법은 패키지 DescTools 의 `DunnettTest()` 함수로 실행할 수 있다.

이제 예를 들어 저혈당 실험에서 그룹 1 을 기준 집단이라고 하고 나머지 4개의 집단과 평균이 다른지 검정하고 싶다고 하자. 다음과 같이 반응변수, 그룹변수 그리고 기준집단(control)의 값을 지정해 주면 된다.

첫 번째 기준집단(control)은 5번째 집단과 유의한 차이가 있다.

```
test4 <- DunnettTest(diabetes$response, diabetes$treat, control = 1, conf.level = 0.95)
test4
```

7. 다중비교

Dunnett's test for comparing several treatments with a control :
95% family-wise confidence level

\$`1`

	diff	lwr.ci	upr.ci	pval
2-1	-2.825	-9.133182	3.483182	0.5694
3-1	-3.625	-9.933182	2.683182	0.3640
4-1	-2.250	-8.558182	4.058182	0.7335
5-1	-9.450	-15.758182	-3.141818	0.0033 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

8. 반복측정자료

8.1. 필요한 패키지

```
library(tidyverse)
library(kableExtra)
library(ggplot2)
library(here)

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

8.2. 독립표본과 쌍표본

서로 다른 두 처리의 효과(treatment effect)를 비교하기 위하여 주로 사용되는 방법은 두 개의 독립표본(independent sample)을 비교하는 t-검정법이다. 서로 독립인 두개의 집단(구성원이 겹치지 않는 집단)에 서로 다른 처리를 적용한 뒤에 관측된 자료의 표본 평균을 비교하여 두 개의 처리 효과의 차이를 통계적으로 검정하는 방법이다. t-검정을 위한 분포 가정은 다음과 같다.

$$x_1, x_2, \dots, x_n \sim_{iid} N(\mu_1, \sigma^2) \quad y_1, y_2, \dots, y_n \sim_{iid} N(\mu_2, \sigma^2)$$

위의 가정을 다음과 같은 평균모형(mean model)로 표현할 수 있다.

$$x_i = \mu_1 + e_{i1}, \quad y_i = \mu_2 + e_{i2} \quad (8.1)$$

여기서 e_{i1} 와 e_{i2} 들은 모두 독립이며 $N(0, \sigma^2)$ 을 따르는 오차들이다.

여기서 확률변수 x 와 y 는 서로 독립이고 각 관측값 x_1, x_2, \dots, x_{n_1} 과 y_1, y_2, \dots, y_{n_2} 들도 각각 모두 독립이다. 분포 가정에서 다른 것은 확률변수 x 와 y 의 평균이 다르다.

이러한 가정 하에서 다음의 두개의 가설검정을 할 수 있는 방법이 t-검정법이며

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2 \quad (8.2)$$

검정통계량은 다음과 같이 주어진다.

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{n}}} \quad (8.3)$$

여기서 \bar{x} 와 \bar{y} 는 각 집단의 표본 평균이고 s_p^2 은 합동분산추정량(pooled variance estimator)이다.

$$s_p^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{2n - 2}$$

이제 두 개의 처리를 비교하는 경우, 독립 표본이 아닌 경우를 고려해 보자.

독립 표본이 아닌 대표적인 경우가 쌍표본(또는 대응표본)에 대한 t-검정이다(paired t-test). 쌍표본 검정에서는 하나의 개체에 두 개의 처리를 모두 적용하여 각 처리에 대한 반응값을 쌍 (x_i, y_i) 으로 얻는다. 예를 들어 최초로 허가 받은 약품과 복제약의 생물학적동등성(bioequivalence)을 입증하는 실험에서는 한 사람에게 최초허가약을 투여하여 약의 효과를 보고 일정 시간이 지난 뒤 복제약을 같은 사람에게 투여하여 그 효과를 측정한다. 다른 예로서 두 개의 눈병 치료제를 각각 눈에 투여하여 효과를 비교하는 경우도 이러한 쌍표본에 속한다. 넓은 의미에서 일관성 쌍둥이에게 각각 다른 처리를 하여 비교하는 것도 대응비교라고 할 수 있다.

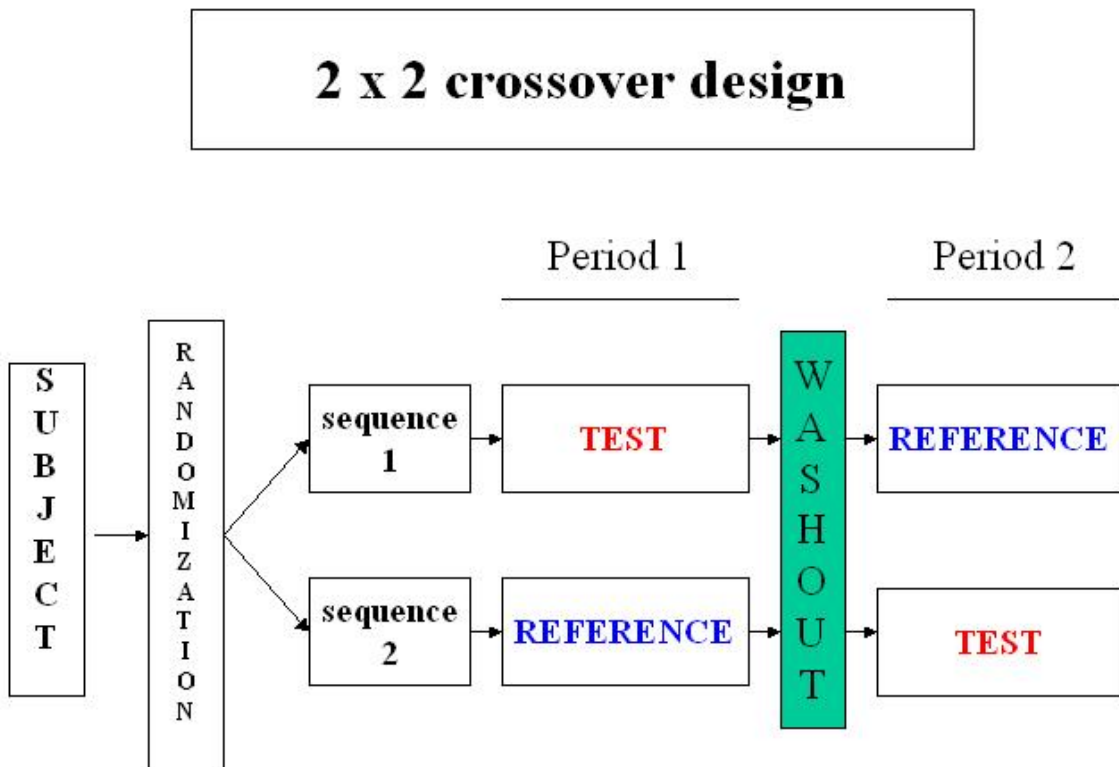


그림 8.1.: 생물학적동등성 실험에서 사용되는 교차실험(crossover design) - 개체 당 2개의 관측치 (각 처리에 대하여 한 개의 관측값)

8. 반복측정자료

가장 단순한 대응비교로서 각 개체에 대하여 두 개의 처리에 대한 쌍표본 (x_i, y_i) 를 관측한다고 가정하자. 이에 대한 분포 모형은 다음과 같다.

$$d_i = x_i - y_i \sim_{iid} N(\delta, \sigma^2) \text{ where } \delta = \mu_1 - \mu_2 = E(x) - E(y)$$

대응비교에서 사용되는 t-통계량은 다음과 같다.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{\bar{x} - \bar{y}}{s_d/\sqrt{n}} \quad (8.4)$$

여기서 s_d^2 는 d_1, d_2, \dots, d_n 의 표본분산이다.

$$s_d^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}$$

위에서 알아본 독립표본에 의한 비교와 쌍표본에 의한 비교가 다른 점은 무엇일까 생각해 보자. 표본의 비교가 다른 개체에서 추출된 독립인 관측치를 이용하는지 또는 같은 개체에서 추출된 대응하는(독립이 아닌) 관측치를 이용하는지에 따라서 서로 다른 t-통계량을 사용한다. 식 8.3 과 식 8.4 에 나타난 t-통계량을 비교하면 분자에 나타난 통계량은 효과의 차이를 나타내는 두 개의 평균의 차이로서 기본적으로 동일하다($\bar{d} = \bar{x} - \bar{y}$). 하지만 분모에서는 분자에 나타난 통계량의 표본오차(standard error)를 나타내는 양으로서 서로 다르다. 독립표본에서는 표본의 평균이 서로 독립이므로 다음과 같이 평균의 차이에 대한 분산이 각각의 분산의 합과 같으므로 이에 대한 추정량으로서 합동분산추정량을 이용하였다.

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) \quad (8.5)$$

쌍표본에서는 위의 식 8.5 을 적용할 수 없다. 왜냐하면 두개의 표본 평균이 서로 독립이 아닐 가능성이 매우 높기 때문이다. 같은 개체에서 나온 관측치는 어떠한 형태로든 서로 관계가 있을 가능성이 높기 때문에 독립을 함부로 가정할 수 없다. 예를 들어 생물학적동등성 실험에서는 약 효과의 차이보다 약이 몸에 흡수되는 개인적인 체질이 관측값에 더 큰 영향을 줄 수 있다.

확률변수 x 와 y 가 독립이 아닌 경우 두 모형군의 차이를 비교하기 위하여 비교에 사용된 통계량은 두 확률변수의 차이이다.

$$d_i = x_i - y_i$$

여기서 두 개의 확률변수의 차이를 이용할 때 암시적인 가정은 두 개의 확률변수의 차이를 내면 두 변수에 공통적으로 포함된 개인의 특성이 서로 상쇄되어 처리의 차이만이 확률변수 d_i 에 존재한다는 것이다. 지금 설명한 대응비교 모형의 합리적인 가정을 요약하면 다음과 같다.

- 개인의 특성을 반영하는 공통 요인이 두 변수에 모두 영향을 미친다.
- 따라서 두 관측값 (x_i, y_i) 가 독립이 아니다
- 두 관측값의 차이를 내면 공통요인이 서로 상쇄되어 처리효과만 남는다.

$$E(d_i) = \mu_1 - \mu_2$$

위의 가정을 구현할 수 있는 대응비교 모형을 다음과 같은 가법모형(additive models)로 표현할 수 있다.

$$x_i = \mu_1 + a_i + e_{i1}, \quad y_i = \mu_2 + a_i + e_{i2} \quad (8.6)$$

여기서 a_i 는 두 확률변수 (x_i, y_i) 에 공통으로 포함된 개인적인 특성을 나타내는 요인이며 위의 식 8.6는 식 8.1에 공통요인 a_i 가 추가된 형태이다.

두 확률변수 (x_i, y_i) 가 종속이기 위해서는 다음과 같은 가정을 이용할 수 있다.

$$a_i \sim N(0, \sigma_a^2), \quad e_{i1} \sim_{iid} N(0, \sigma_e^2) \quad e_{i2} \sim_{iid} N(0, \sigma_e^2)$$

여기서 a_i 가 평균이 0이고 분산이 σ_a^2 인 확률변수이다. 이러한 요인을 임의효과(random effect)라고 하며 모수(parameter)인 평균 μ_i 은 고정효과(fixed effect)라고 부른다. e_{i1} 와 e_{i2} 들은 모두 독립이며 평균이 0이고 분산이 σ_e^2 인 정규분포를 따르는 오차들이다. 또한 a_i 와 (e_{i1}, e_{i2}) 도 독립이다.

위와 같은 가정에서 두 변수의 차이를 내면 공통요인인 a_i 가 제거되어 두 처리의 차이만이 남게되며 x_i 와 y_i 는 분포의 가정상 독립이 아니다.

$$\begin{aligned} d_i &= x_i - y_i \\ &= \mu_1 + a_i + e_{i1} - (\mu_2 + a_i + e_{i2}) \\ &= \mu_1 - \mu_2 + (e_{i1} - e_{i2}) \\ &= \mu_1 - \mu_2 + e_i^* \\ \text{Cov}(x_i, y_i) &= \text{Cov}(\mu_1 + a_i + e_{i1}, \mu_2 + a_i + e_{i2}) \\ &= \text{Cov}(a_i, a_i) \\ &= \text{Var}(a_i) \\ &= \sigma_a^2 \end{aligned}$$

다음 절에서는 여기서 논의한 독립표본과 쌍표본의 개념 및 추정법을 일반적인 선형모형으로 확장하여 체계적인 비교를 해볼것이다.

8.3. 일원배치 모형

8.3.1. 고정효과 모형

먼저 일원배치 요인계획(one-way factor design)이용한 실험을 생각해 보자. 고려하는 요인의 수준의 개수를 I 라고 하면 I 개의 수준 중에 하나를 임의로 선택하여 실험대상에 적용하는 임의화 방법으로 각 수준마다 J 의 관측값을 얻어다고 하자. 다음과 같은 ANOVA모형을 고려하여 분석을 할 수 있다.

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2, \dots, I \text{ and } j = 1, 2, \dots, J \quad (8.7)$$

여기서 e_{ij} 는 서로 독립이며 $N(0, \sigma_e^2)$ 를 따르는 오차항이다.

ANOVA 모형 8.7에서 μ 와 α_i 는 고정효과(fixed effect)라고 부르며 추정해야 할 모수(parameter)이다. 세심하게 설계된 실험에서는 수준에 대한 효과 α_i 의 값이 변하지 않게 통제할 수 있는 실험 환경이 가능하다고 생각할 수 있으므로 α_i 의 값을 변하지 않는 고정효과로 보는 것이 합리적이다.

일원배치 요인계획을 이용한 실험에서는 주요 관심사가 수준간의 차이가 있는지에 대한 것이며 이는 제곱합을 이용한 ANOVA table에서 F-test를 이용하여 검정할 수 있다.

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_I$$

Source	Sum of Square	DF	Mean Square	F
Treatment	$SSA = \sum_{i=1}^I J(\bar{y}_i - \bar{y})^2$	$I - 1$	$MSA = SS_{reg}/(I - 1)$	$F = MSA/MSE$
Error	$SSE = \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y}_i)^2$	$I(J - 1)$	$MSE = SSE/(I(J - 1))$	
total	$SST = \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y})^2$	$IJ - 1$		

8.3.2. 혼합효과 모형

이제 다음과 같은 자료의 추출을 생각해 보자. 서울시 A구에 초등학교가 20개 있다고 하자. 20개의 학교중 6개의 학교를 임의로 추출하고 추출된 학교에 속한 모든 6학년 학생들에게 과학시험을 보게하여 점수를 얻었다. 이러한 자료에서 학생들의 성적은 모두 같지 않을 것이 당연하며 가장 점수가 낮은 학생부터 높은 학생까지 점수의 변동(variation)이 존재한다. 변동의 요인은 무엇일까? 학생의 개인의 차이(예: 학생의 지능, 노력 정도, 학습 환경)도 변동의 요인이지만 또한 학교의 차이도 변동의 요인이 될 수 있다.

여기서 학교에 대한 요인은 앞 절에서 본 실험 자료에서 나타나는 고정 효과에 의한 요인과는 성격이 틀리다. 20개의 학교라는 모집단에서 6개의 학교가 추출되었으며 이 때 학교의 차이는 설계된 실험에서는 수준에 대한 효과와는 다르게 표본 추출 때문에 생기는 변동이라고 할 수 있다. 또한 같은 학교에 다니는 학생들은 같은 지역과 교사 등 공통적인 요인에 의하여 영향을 받는다고 가정할 수 있다. 따라서 같은 학교에 다니는 학생들의 성적이 독립이 아닐 수도 있다.

이러한 효과를 임의효과(random effect)라고 부르며 학생들의 과학점수에 대한 모형을 다음과 같은 혼합효과모형(mixed effects model)으로 설명할 수 있다.

$$y_{ij} = \mu + a_i + e_{ij}, \quad i = 1, 2, \dots, I \text{ and } j = 1, 2, \dots, J \quad (8.8)$$

여기서 a_i 는 학교의 효과를 나타내는 임의효과이며 서로 독립이고 $N(0, \sigma_a^2)$ 를 따른다. 또한 개인에 대한 효과 또는 오차항 e_{ij} 는 서로 독립이며 $N(0, \sigma_e^2)$ 를 따른다. 임의효과 a_i 와 오차항 e_{ij} 는 서로 독립이다. 고정효과 μ 는 전체 평균을 나타내는 모수이다. 학교를 추출할 때 그 효과를 분산이 σ_a^2 를 가지는 정규모집단에서 추출한다고 가정하는

것이다. 학교를 하나의 군집(cluster)로 생각하고 학교의 효과를 군집효과로 보고 총변동을 설명하고 나머지 변동은 오차의 변동으로 개인의 효과 등으로 설명한다.

$$Var(y_{ij}) = Var(\mu + a_i + e_{ij}) = Var(a_i) + Var(e_{ij}) = \sigma_a^2 + \sigma_e^2$$

식 8.8 로 표현된 임의효과를 포함한 혼합모형(일원배치 임의효과 모형; one-way random effect models)의 가장 큰 특징 중에 하나는 같은 군집에 속하는 관측치들은 서로 독립이 아니며 양의 상관관계가 있다. 위의 예제에서 두 학생 j 와 k 가 같은 학교 i 에 속한다면

$$Cov(y_{ij}, y_{ik}) = Cov(\mu + a_i + e_{ij}, \mu + a_i + e_{ik}) = Cov(a_i, a_i) = \sigma_a^2$$

따라서

$$corr(y_{ij}, y_{ik}) = \frac{Cov(y_{ij}, y_{ik})}{\sqrt{Var(y_{ij})Var(y_{ik})}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

위의 상관계수를 보면 학교간의 변동의 크기를 나타내는 σ_a^2 각 개인간의 변동을 나타내는 σ_e^2 보다 상대적으로 클수록 상관계수가 1에 가까워진다. 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 군집내 관측값의 상관관계가 달라진다.

식 8.8의 임의효과 모형은 임의효과가 2개 이상인 모형으로 확장될 수 있다. 예제에서 학교를 추출하고 학교내에서 학급을 추출하여 추출된 학급내의 학생들이 시험을 보면 다음과 같은 모형을 적용할 수 있다.

$$y_{ijk} = \mu + a_i + b_{ij} + e_{ijk}$$

위의 모형에서 a_i 는 $N(0, \sigma_a^2)$ 를 따르는 학교에 대한 임의효과, b_{ij} 는 $N(0, \sigma_b^2)$ 를 따르는 학급에 대한 임의효과, e_{ijk} 는 $N(0, \sigma_e^2)$ 를 따르는 학생에 대한 임의효과 또는 오차항으로 생각할 수 있다.

임의효과 모형에 고정효과가 같이 포함되어 있는 모형을 혼합모형(mixed model)이라고 부르며 반응변수의 변동에 영향을 미치는 요인들 또는 예측변수 중에서 임의효과와 고정효과를 구별하여 정하고 동시에 모형에 포함시킬 수 있다. 예를 들어서 학교를 선택하여 과학시험을 볼 경우에 학기가 시작할 때 과학교수법 두가지 중 하나를 임의화해서 각 학교에 배정하여 배정된 교수법으로 학생을 가르친 후에 시험을 보았다면 교수법에 대한 효과는 고정효과로 볼 수 있다. 따라서 다음과 같은 모형을 고려할 수 있다.

$$y_{ij} = \mu + \tau_{k(i)} + a_i + e_{ij} \quad (8.9)$$

여기서 $\tau_{k(i)}$ 는 i 번째 학교의 교수법에 대한 고정효과이다(교수법이 배정된 결과에 의하여 $k(i) = 1$ 또는 $k(i) = 2$ 이 된다). 이 교수법에 대한 고정효과에 대해서는 두 교수법이 차이가 있는 지에 대한 검정이 주요한 관심사일 것이다 ($H_0 : \tau_1 = \tau_2$)

고정효과와 임의효과의 표기

고정효과는 보통 모형식에서 그리스 문자(α , β)로 표시하고 임의효과는 모형식에서 영어 알파벳 문자(a , b)로 표시한다

8.4. 반복측정자료

반복측정자료(longitudinal data, repeated measurements)는 관측단위안에서 여러 개의 관측값을 측정한 자료의 형식을 말한다. 예를 들어 환자가 병원을 여러 번 방문하고 방문시마다 혈압을 측정하였다면 한 명의 환자에서 반복 측정된 자료는 서로 독립이 아니다. 또한 가구조사(household survey)에서 가구원의 취업 여부, 건강 상태등을 여러 해동안 매년 측정하는 경우 이러한 자료를 패널자료(panel data) 또는 longitudinal 자료라고 한다. 이렇게 하나의 관측단위 안에서 측정한 자료들은 서로 독립이 아닌 특징이 있고 자료를 분석하는 경우 이러한 자료들의 종속구조를 고려하는 모형을 사용하는 것이 적절하다. 이렇게 반복측정자료에서 반복자료들의 공분산구조를 설정하는 통계적 방법들은 다양하지만 대표적으로 쉽게 사용할 수 있는 방법이 임의효과를 포함한 혼합모형을 사용하는 방법이다.

Example 8.1 (트럭운전사의 수면시간). lme4 패키지에 자료인 `sleepstudy`는 화물트럭 운전자들에 대한 수면부족 현상에 대하여 연구한 자료이다. 18명의 운전자들이 매일 3시간의 수면(부족한 수면)을 하면서 매일 일정한 동작의 반응시간을 10일동안 반복적으로 측정한 자료가 있다. 한명의 운전자에게 10일 동안의 반응에 대한 측정자료 10개가 존재하므로 이는 반복측정 자료이며 이러한 10개의 자료는 독립이 아니다. 일단 자료의 구조를 살펴보자. 반응변수 'Reaction'은 반응시간(ms)를 나타내며 설명변수로서 Days는 날짜($t = 0, 1, 2, \dots, 9$), Subject는 운전자의 고유번호를 나타낸다.

```
library(lme4)
library(lmerTest)
str(sleepstudy)
```

```
'data.frame':  180 obs. of  3 variables:
 $ Reaction: num  250 259 251 321 357 ...
 $ Days    : num   0  1  2  3  4  5  6  7  8  9 ...
 $ Subject : Factor w/ 18 levels "308","309","310",...: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(sleepstudy, n = 20)
```

	Reaction	Days	Subject
1	249.5600	0	308
2	258.7047	1	308
3	250.8006	2	308
4	321.4398	3	308
5	356.8519	4	308
6	414.6901	5	308
7	382.2038	6	308

8. 반복측정자료

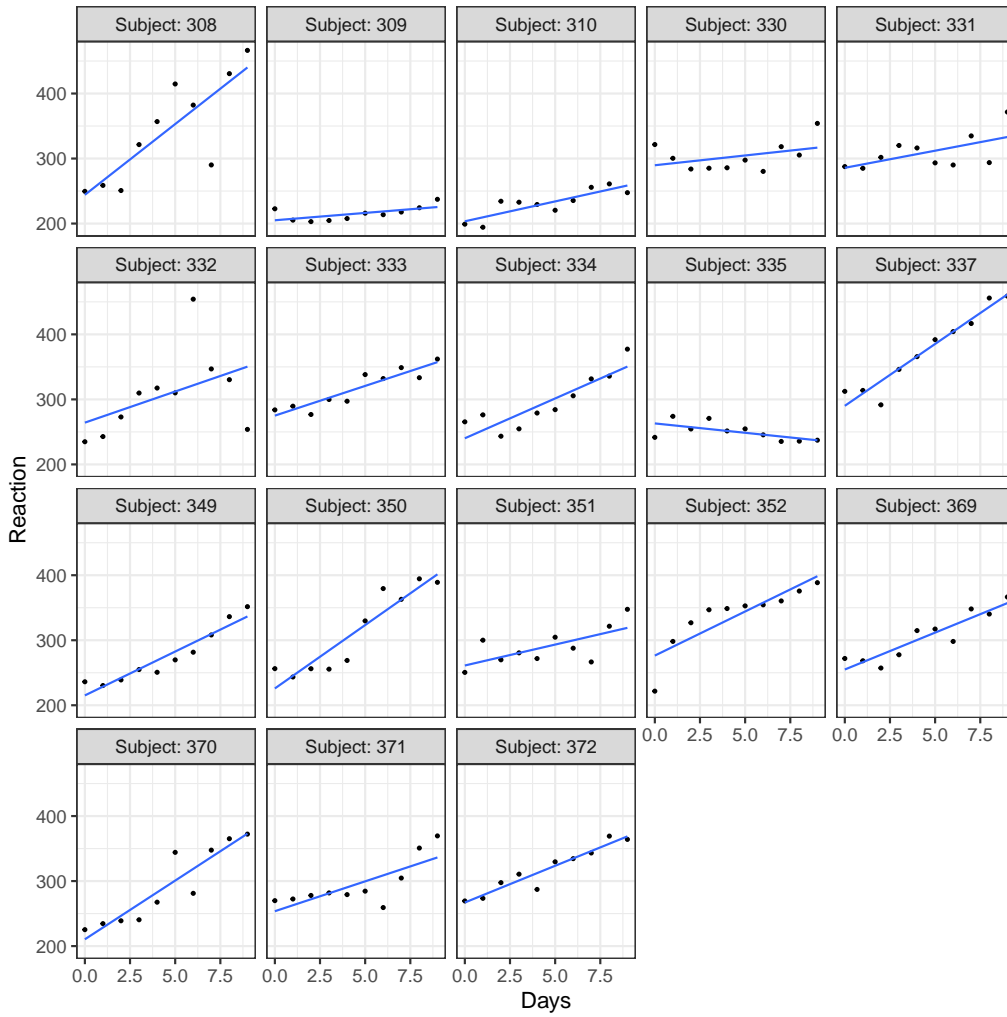
8	290.1486	7	308
9	430.5853	8	308
10	466.3535	9	308
11	222.7339	0	309
12	205.2658	1	309
13	202.9778	2	309
14	204.7070	3	309
15	207.7161	4	309
16	215.9618	5	309
17	213.6303	6	309
18	217.7272	7	309
19	224.2957	8	309
20	237.3142	9	309

각 운전자에 대한 10일 간의 반응속도가 시간에 따라 어떻게 변하는 가를 알아보자. 전반적으로 시간이 지나면서 운전자들의 반응시간이 증가하고 있음을 알 수 있다. 또한 개인 별로 반응 시간의 변화와 패턴이 다르다는 것을 알 수 있다.

```
library(ggplot2)
ggplot(sleepstudy, aes(x = Days, y = Reaction)) +
  geom_point(size = 0.5) +
  geom_smooth(method = lm, se = FALSE, linewidth = 0.5) +
  facet_wrap("Subject", labeller = label_both) +
  theme_bw()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

8. 반복측정자료



8.4.1. 개체들의 선형 회귀모형

각 운전자 i 에 대하여 10일간 측정한 반응속도 y_{ij} 를 시간에 대하여 선형모형으로 적합하면 개인별 회귀직선을 다음과 같이 표시할 수 있다.

$$y_{ij} = \beta_{0i} + \beta_{1i}t_j + e_{ij}, \quad i = 1, 2, \dots, 18, \quad j = 1, 2, \dots, 10 \quad (8.10)$$

여기서 오차항 e_{ij} 은 서로 독립이며 $N(0, \sigma_e^2)$ 를 따른다고 가정한다.

행렬식으로는 다음과 같이 나타낼 수 있다.

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \mathbf{e}_i$$

여기서

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i,10} \end{bmatrix}, \mathbf{X}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix}, \boldsymbol{\beta}_i = \begin{bmatrix} \beta_{0i} \\ \beta_{1i} \end{bmatrix}, \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{i,10} \end{bmatrix}$$

위의 식에서 β_{0i} 와 β_{1i} 는 i 번째 운전사의 반응속도를 설명내는 회귀직선의 절편과 기울기이다. 절편 β_{0i} 는 실험 시작때 반응속도를 의미하고 기울기 β_{1i} 는 실험이 진행되는 동안 반응속도가 어떻게 변하는 지 변화의 방향과 크기를 보여준다. 함수 `lmList`를 아래와 같이 이용하면 식 8.10 을 각 운전사마다 적합시켜 각각의 절편과 기울기를 구할 수 있다.

```
lmf1 <- lmList(Reaction ~ Days | Subject, sleepstudy)
lmf1
```

Call: `lmList(formula = Reaction ~ Days | Subject, data = sleepstudy)`

Coefficients:

	(Intercept)	Days
308	244.1927	21.764702
309	205.0549	2.261785
310	203.4842	6.114899
330	289.6851	3.008073
331	285.7390	5.266019
332	264.2516	9.566768
333	275.0191	9.142045
334	240.1629	12.253141
335	263.0347	-2.881034
337	290.1041	19.025974
349	215.1118	13.493933
350	225.8346	19.504017
351	261.1470	6.433498
352	276.3721	13.566549
369	254.9681	11.348109
370	210.4491	18.056151
371	253.6360	9.188445
372	267.0448	11.298073

Degrees of freedom: 180 total; 144 residual

Residual standard error: 25.59182

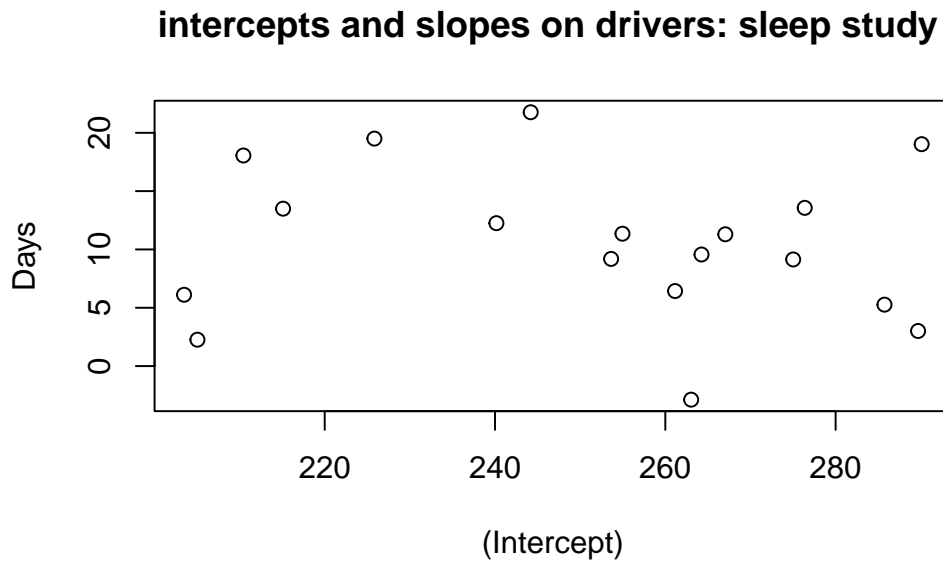
```
cor(coef(lmf1))
```

(Intercept)	Days
-------------	------

8. 반복측정자료

```
(Intercept)  1.0000000 -0.1375534  
Days         -0.1375534  1.0000000
```

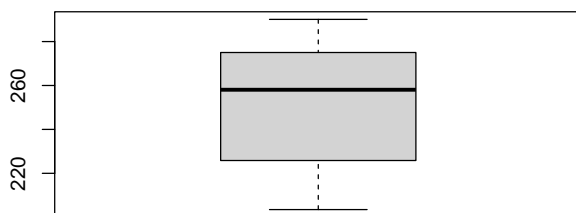
```
plot(coef(lmf1), main = "intercepts and slopes on drivers: sleep study ")
```



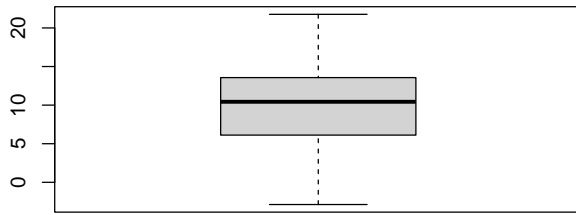
18개의 절편과 기울기는 큰 상관관계는 없는것으로 보이지만 약한 음의 상관계수가 나타났다.

절편과 기울기에 대한 분포를 보기 위하여 상자그림을 그려보면 평균을 중심으로 대칭인 분포를 보이고 있다.

```
boxplot(coef(lmf1)[1])  
boxplot(coef(lmf1)[2])
```



8. 반복측정자료



이제 각 운전사에 대하여 회귀식을 따로 적합하지 않고 전체 운전자들의 자료를 모두 합쳐서 하나의 회귀식을 고려할 수 있다. 개체의 특성을 반영하는 모형이 아닌 전체 집단에 대한 평균적인 모형(population model)을 고려하는 것이다.

$$y_{ij} = \beta_0 + \beta_1 t_j + e_{ij}, \quad i = 1, 2, \dots, 18, j = 1, 2, \dots, 10 \quad (8.11)$$

여기서 오차항은 서로 독립이며 $N(0, \sigma_e^2)$ 를 따른다고 가정한다.

위와 같은 전체 운전자 집단의 관측값을 운전자의 특성을 고려하지 않고 세운 모형으로서 시간에 따른 반응시간에 대한 모집단의 전체적인 평균적 함수 관계를 파악하는 모형이라고 할 수 있다.

```
lmpop <- lm(Reaction ~ Days, sleepstudy)
summary(lmpop)
```

Call:

```
lm(formula = Reaction ~ Days, data = sleepstudy)
```

Residuals:

Min	1Q	Median	3Q	Max
-110.848	-27.483	1.546	26.142	139.953

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	251.405	6.610	38.033	< 2e-16 ***
Days	10.467	1.238	8.454	9.89e-15 ***

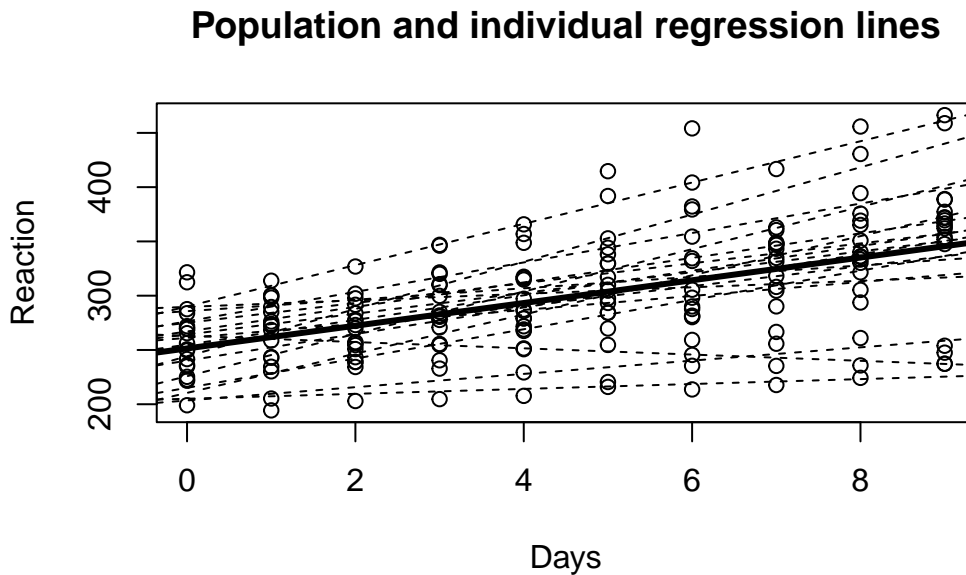
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.71 on 178 degrees of freedom

Multiple R-squared: 0.2865, Adjusted R-squared: 0.2825

F-statistic: 71.46 on 1 and 178 DF, p-value: 9.894e-15


```
with(sleepstudy, plot(Days, Reaction, main = "Population and individual regression lines"))
abline(a = coef(lmpop)[1], b = coef(lmpop)[2], lwd = 3)
for (i in 1:18) {
  xx <- as.numeric(coef(lmf1)[i, ])
  abline(a = xx[1], b = xx[2], lty = 2)
}
```



이제 각 운전사에 대하여 개체별로 적합한 회귀식의 계수들($\hat{\beta}_{0i}, \hat{\beta}_{1i}$)와 전체집단에 적합한 회귀식의 계수($\hat{\beta}_0, \hat{\beta}_1$)의 관계를 보면 개체별로 회귀 계수들의 평균이 전체에 적용한 모형의 계수와 매우 가까운 사실을 알 수 있다.

```
apply(coef(lmf1), 2, mean)
```

(Intercept)	Days
251.40510	10.46729

```
coef(lmpop)
```

(Intercept)	Days
251.40510	10.46729

8.4.2. 임의계수 모형

앞 절의 모형과 분석에서 알 수 있듯이 한 개체에 대하여 여러 개의 관측값을 측정한 자료에 회귀방정식을 각각 적합시켜보고 또한 개체의 특성을 고려하지않은 전체 모형을 적합해보면 다음과 같은 두 가지 결과를 볼 수 있다.

- 각 개체별 회귀식은 개인의 특성을 반영한다. 즉, 개체에 따라 시간에 따른 반응시간의 변화가 다르게 나타난다.
- 하지만 개인별로 볼 때도 전체적으로는 시간에 따라서 반응시간이 증가하는 경향이 있음을 알 수 있다.
- 전체 자료에 적합한 모형을 보면 개인별로 적합한 모형의 공통적인 성격, 즉 시간에 따른 반응시간의 증가를 알 수 있다.
- 이러한 결과를 보고 각 개인의 변화는 전체적인 변화를 따르면서 각 개인의 특성이 반영되었다고 가정할 수 있다.

위에서 논의하였듯이 전체적인 경향과 개인의 특성을 동시에 고려할 수 있는 모형이 생각할 수 있고 이러한 모형이 다음과 같은 모형이다.

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j + e_{ij}, \quad i = 1, 2, \dots, 18, \quad j = 1, 2, \dots, 10 \quad (8.12)$$

모형 8.12 는 절편과 기울기가 두 개의 구성 요소로 더해져서 표현된다. 기울기는 $\beta_1 + b_{1i}$ 로써 나타내어지며 β_1 은 모집단이 가지는 공통적인 경향을 반영하는 모수이고 b_{1i} 는 i 번째 개체의 특성을 반영한 확률변수이다. 절편도 유사한 형식으로 구성된다.

각 개인에 대한 특성을 나타내는 변수 (b_{0i}, b_{1i})을 확률변수로 설정하고 이를 모수(β_0, β_1) (parameter or fixed effect)와 구별하여 임의효과(random effect)라고 한다.

임의효과는 모집단을 구성하는 개인이 표본에 추출되었다고 생각하며 확률분포를 따른다고 가정한다. 반복측정자료에서 임의효과를 공통으로 가지고 있는 관측치는 독립이 아니게 되며 따라서 같은 개체에서 나온 관측값은 독립이 아니다.

모형 8.12 처럼 각 개인의 임의 효과가 포함된 선형모형을 임의계수모형(random coefficient model) 이라고 한다. 이러한 모형은 반복측정자료에서 개인의 특성을 고려하여 모형을 설정하는데 유용하다.

18명에 대한 회귀직선의 절편과 기울기를 보면 개인의 차이에 따른 변동을 볼 수 있으며 이러한 각 개인간의 변동을 임의효과를 이용하여 다음과 같은 모형을 생각해보자.

$$\beta_i = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix}, \quad \mathbf{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b1}^2 & \rho\sigma_{b1}\sigma_{b2} \\ \rho\sigma_{b1}\sigma_{b2} & \sigma_{b2}^2 \end{bmatrix} \right)$$

위의 모형은 각 개인의 회귀직선에서 각 절편과 기울기가 전체평균 β_0 와 β_1 를 따르며 각 개인의 차이는 전체평균에 임의효과인 b_{0i} 와 b_{1i} 가 더해져서 나타난다는 것을 의미한다. 이변량 임의효과 b_{0i} 와 b_{1i} 는 이변량 정규분포를 따르며 각각의 분산과 상관계수가 $\sigma_{b1}^2, \sigma_{b2}^2, \rho$ 이다.

다른 개체에 대한 임의효과는 서로 독립이며 임의 효과와 오차항은 독립이다. 여기서 오차항은 서로 독립이며 $N(0, \sigma_e^2)$ 를 따른다고 가정한다.

$$Cov(\mathbf{b}_i, \mathbf{b}_j) = \mathbf{0} \text{ when } i \neq j, \quad Cov(\mathbf{b}_i, e_{jk}) = \mathbf{0} \text{ for all } i, j, k$$

8. 반복측정자료

위 8.12 같은 혼합효과모형(mixed effects model)을 각 개인 i 에 대하여 행렬식으로 표시하면 다음과 같다. 아래 모형식은 일반적으로 독립군집 모형에서 하나의 군집에 대한 반복된 관측값들에 대한 모형이며 독립군집 모형의 일반적인 이론은 Appendix B 에서 더 논의할 것이다.

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i, \quad i = 1, 2, \dots, 18$$

여기서

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i,10} \end{bmatrix}, \quad \mathbf{X}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{Z}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix}, \quad \mathbf{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix}, \quad \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{i,10} \end{bmatrix}$$

위의 각 개인에 대한 모형을 모두 합쳐서 하나의 혼합효과모형으로 나타내면 다음과 같이 표현할 수 있다.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \quad (8.13)$$

여기서 반응변수벡터 \mathbf{y} 와 고정효과 $\boldsymbol{\beta}$ 에 대한 계획행렬 \mathbf{X} 는 각 개인의 반응변수벡터 \mathbf{y}_i 와 \mathbf{X}_i 를 행으로 쌓아놓은 것으로 표현된다. 오차항에 대한 벡터 \mathbf{e} 도 동일한 형식의 벡터이다.

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{18} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_{18} \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_{18} \end{bmatrix}$$

임의효과 벡터 \mathbf{b} 는 각 개인에 대한 임의효과벡터 \mathbf{b}_i 를 행으로 쌓아놓은 것과 같고 임의효과에 대한 계획행렬 \mathbf{Z} 는 각 개인의 계획행렬 \mathbf{Z}_i 를 대각원소로 구성하는 행렬이다.

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_{18} \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Z}_{18} \end{bmatrix}$$

혼합모형 8.13를 `lmer()` 함수를 이용하여 적합시켜보자. 모형에서 `(1 + Days|Subject)`이 개체에 대하여 절편과 기울기에 대한 임의효과를 지정한다.

```
fm1 <- lmer(Reaction ~ 1 + Days + (1 + Days|Subject), sleepstudy)
summary(fm1)
```

8. 반복측정자료

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
```

```
Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
```

```
Data: sleepstudy
```

```
REML criterion at convergence: 1743.6
```

```
Scaled residuals:
```

Min	1Q	Median	3Q	Max
-3.9536	-0.4634	0.0231	0.4634	5.1793

```
Random effects:
```

Groups	Name	Variance	Std.Dev.	Corr
Subject	(Intercept)	612.10	24.741	
	Days	35.07	5.922	0.07
Residual		654.94	25.592	

```
Number of obs: 180, groups: Subject, 18
```

```
Fixed effects:
```

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	251.405	6.825	17.000	36.838	< 2e-16 ***
Days	10.467	1.546	17.000	6.771	3.26e-06 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Correlation of Fixed Effects:
```

```
(Intr)
```

```
Days -0.138
```

위의 혼합모형 적합결과를 살펴보자. 첫째로 고정효과에 대한 추정식은 다음과 같다

```
fixef(fm1)
```

(Intercept)	Days
251.40510	10.46729

또한 오차항에 대한 분산 및 임의효과의 분산성분과 상관계수는 다음과 같이 나타난다.

```
VarCorr(fm1)
```

Groups	Name	Std.Dev.	Corr
Subject	(Intercept)	24.7407	

8. 반복측정자료

Days	5.9221	0.066
Residual	25.5918	

위의 고정효과와 임의효과의 분산성분에 대한 추정치를 이용하여 모형의 적합의 결과를 다음과 같이 나타낼 수 있다.

$$\text{Reaction Time} = N(251.405 + 10.467 \text{ Days}, 25.6^2)$$

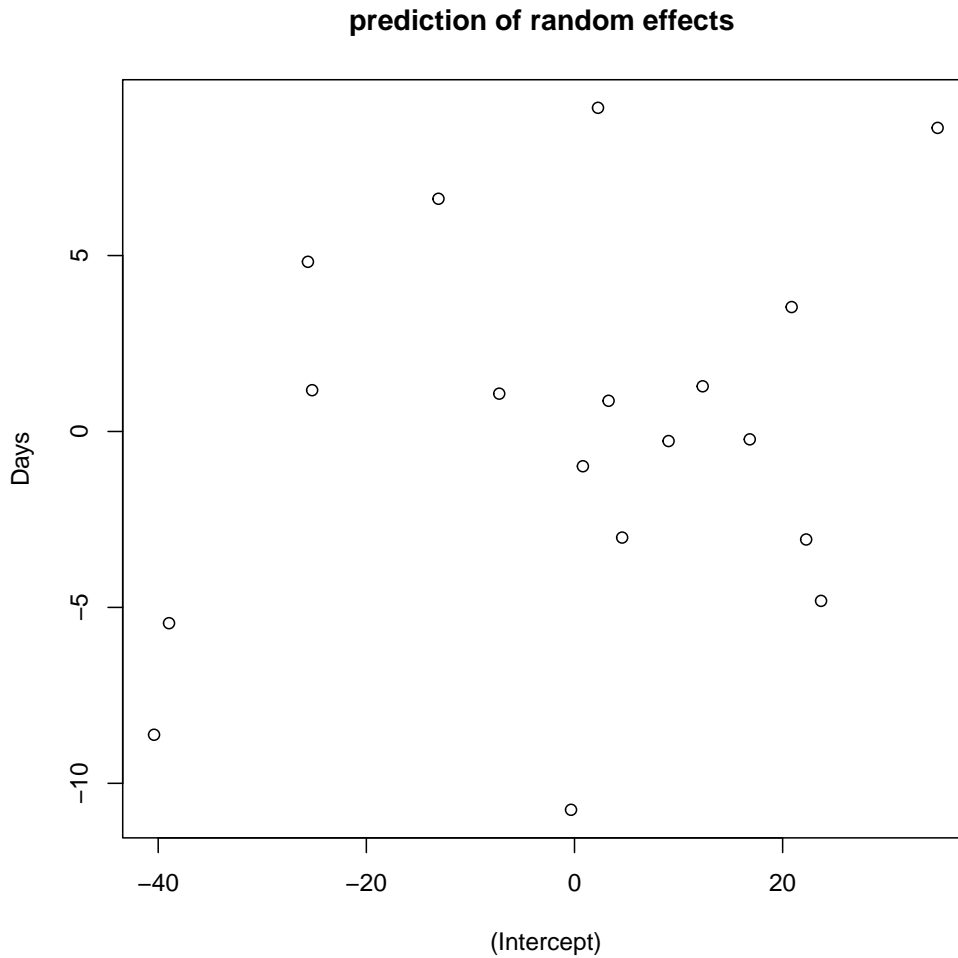
$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 24.74^2 & (0.07)(24.74)(5.92) \\ (0.07)(24.74)(5.92) & 5.92^2 \end{bmatrix} \right)$$

이제 임의효과 \mathbf{b} 에 대한 예측(prediction)을 생각해보자. 우리는 오직 관측벡터 \mathbf{y} 만을 관측하고 임의효과 \mathbf{b} 는 관측을 할 수 없는 확률변수이다. 하지만 주어진 관측벡터와 추정된 분산으로 임의효과의 값을 예측할 수있으며 그 결과는 다음과 같다.

```
re <- ranef(fm1)$Subject
re
```

	(Intercept)	Days
308	2.2585509	9.1989758
309	-40.3987381	-8.6196806
310	-38.9604090	-5.4488565
330	23.6906196	-4.8143503
331	22.2603126	-3.0699116
332	9.0395679	-0.2721770
333	16.8405086	-0.2236361
334	-7.2326151	1.0745816
335	-0.3336684	-10.7521652
337	34.8904868	8.6282652
349	-25.2102286	1.1734322
350	-13.0700342	6.6142178
351	4.5778642	-3.0152621
352	20.8636782	3.5360011
369	3.2754656	0.8722149
370	-25.6129993	4.8224850
371	0.8070461	-0.9881562
372	12.3145921	1.2840221

```
plot(re, main ="prediction of random effects ")
```



예측된 각 개인의 절편과 기울기에 대한 임의효과 b_{0i} 과 b_{1i} 에 고정효과 추정량 $\hat{\beta}$ 를 더해주면 각 개인의 절편과 기울기에 대한 예측값을 구할 수 있다.

```
beta <- matrix(as.numeric(fixef(fm1)),18,2,byrow=T)
beta + re
```

	(Intercept)	Days
308	253.6637	19.6662617
309	211.0064	1.8476053
310	212.4447	5.0184295
330	275.0957	5.6529356
331	273.6654	7.3973743
332	260.4447	10.1951090
333	268.2456	10.2436499
334	244.1725	11.5418676
335	251.0714	-0.2848792
337	286.2956	19.0955511
349	226.1949	11.6407181

8. 반복측정자료

```
350    238.3351 17.0815038
351    255.9830  7.4520239
352    272.2688 14.0032871
369    254.6806 11.3395008
370    225.7921 15.2897709
371    252.2122  9.4791297
372    263.7197 11.7513080
```

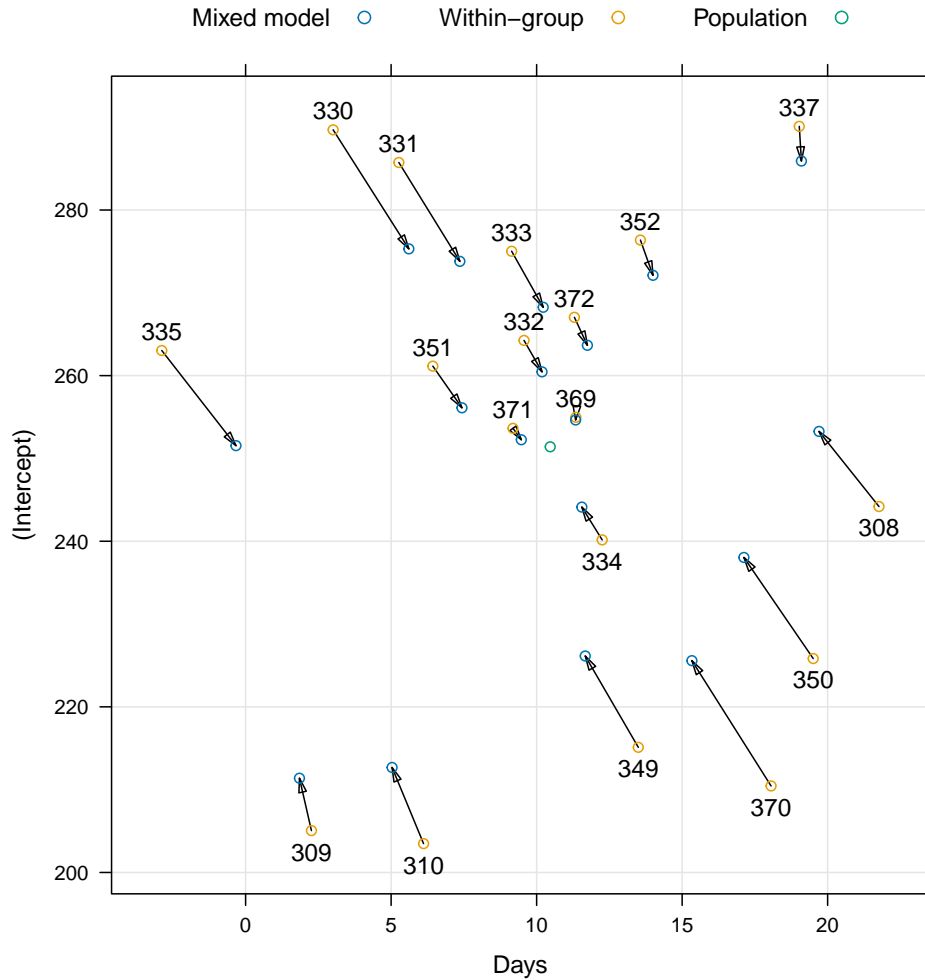
위의 결과를 각 개체에 대해 별도의 회귀직선을 적합시켜서 얻은 18개의 절편과 기울기와 비교해보자.

```
coef(lmf1)
```

	(Intercept)	Days
308	244.1927	21.764702
309	205.0549	2.261785
310	203.4842	6.114899
330	289.6851	3.008073
331	285.7390	5.266019
332	264.2516	9.566768
333	275.0191	9.142045
334	240.1629	12.253141
335	263.0347	-2.881034
337	290.1041	19.025974
349	215.1118	13.493933
350	225.8346	19.504017
351	261.1470	6.433498
352	276.3721	13.566549
369	254.9681	11.348109
370	210.4491	18.056151
371	253.6360	9.188445
372	267.0448	11.298073

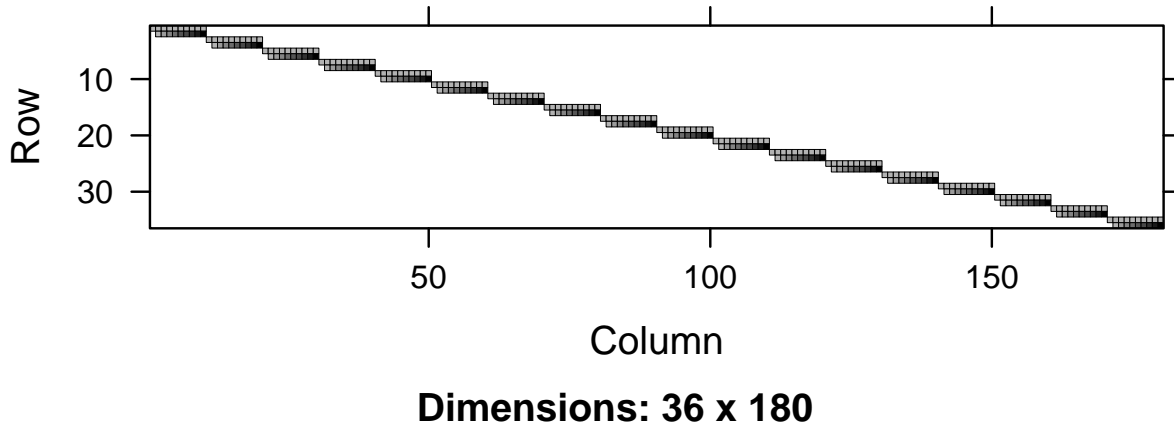
이렇게 혼합모형을 통해서 얻은 각 개인의 절편과 기울기에 대한 예측값과 각각의 개인에 대해서 회귀직선을 따로 적합하여 얻은 절편과 기울기의 관계를 그림으로 그려보면 다음과 같다. 즉 혼합모형을 통해서 얻은 각 개인의 절편과 기울기는 절편과 기울기의 전체평균값 방향으로 축소되는 경향(shrinkage)을 볼수있다.

8. 반복측정자료



여기서 혼합모형의 식 8.13의 임의효과에 대한 계획행렬 Z 의 구조를 살펴보자. R의 적합된 결과에서 `getME` 함수를 이용하여 계획행렬 Z 의 전치행렬(transpose matrix, Z^t)을 얻을 수 있다. 계획행렬 Z 는 그 값의 많은 부분이 0으로 구성되어 있어서 성감행렬(sparse matrix)라고 부르며 이런 행렬은 특별한 형식으로 저장되어 있다.

```
image(getME(fm1, "Zt"))
```

또한 혼합모형의 식 8.13 에서 임의효과의 상관계수가 $\rho = 0$ 인 경우의 모형을 고려해 보자

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b1}^2 & 0 \\ 0 & \sigma_{b2}^2 \end{bmatrix} \right)$$

임의효과의 상관계수가 $\rho = 0$ 인 모형은 아래와 같이 (1+Days | Subject) 를 사용하는 대신 두 개의 바 || 를 사용한 (1+Days || Subject)으로 적합시키면 결과를 다음과 같이 얻을 수 있다.

```
fm2 <- lmer(Reaction ~ 1 + Days + (1+Days||Subject), sleepstudy)
# 다음 명령어는 위의 명령어와 같은 모형을 적합한다.
# fm2 <- lmer(Reaction ~ 1 + Days + (1|Subject) + (0+Days|Subject), sleepstudy)

summary(fm2)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
```

```
Formula: Reaction ~ 1 + Days + (1 + Days || Subject)
Data: sleepstudy
```

```
REML criterion at convergence: 1743.7
```

```
Scaled residuals:
```

```
      Min       1Q   Median       3Q      Max
```

8. 반복측정자료

```
-3.9626 -0.4625 0.0204 0.4653 5.1860
```

Random effects:

Groups	Name	Variance	Std.Dev.
Subject	(Intercept)	627.57	25.051
Subject.1	Days	35.86	5.988
Residual		653.58	25.565

Number of obs: 180, groups: Subject, 18

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	251.405	6.885	18.156	36.513	< 2e-16 ***
Days	10.467	1.560	18.156	6.712	2.59e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

(Intr)

Days -0.184

위의 고정효과와 임의효과의 분산성분에 대한 추정치를 이용하여 모형의 적합의 결과를 다음과 같이 나타낼 수 있다.

$$\text{Reaction Time} = N(251.405 + 10.467 \text{ Days}, 25.6^2)$$

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 25.1^2 & 0 \\ 0 & 6.0^2 \end{bmatrix} \right)$$

두 모형, 즉 절편과 기울기에 대한 두 임의효과가 종속인지 또는 독립인지에 대한 두 모형을 AIC(Akaike Information Criteris)로 비교하해 보자. 두 모형은 AIC의 값은 거의 동일하지만 임의효과의 상관계수가 0인 모형이 AIC 값이 약간 작으며 상관계수의 추정치가 매우 작으므로 임의효과의 상관계수가 0인 모형을 선택하는 것이 합리적이라고 판단된다.

```
AIC(fm1)
```

```
[1] 1755.628
```

```
AIC(fm2)
```

```
[1] 1753.669
```

8.5. 예제: 고혈압 강하제 실험

여기서 다룰 예제는 Jaewon Lee (2005) 6장에 나오는 예제이다.

고혈압 환자 66명을 대상으로 32명에게는 기존의 혈압 강하제(C)를 투여하고, 나머지 34명에게는 새로운 혈압 강하제(E)를 투여하였다. 환자가 치료를 시작하기 전에 혈압을 측정하고(0주) 치료를 받은 후 4주, 8주 후에 2번 반복 측정하였다. 이 실험의 목표는 다음과 같다.

1. 기존의 약과 새로운 혈압 강하제 간의 유의한 차이가 있는가?
2. 방문 시간(time)에 따라서 혈압(bp)의 차이가 있는가?
3. 약(trt)과 방문 시간(time) 간에 상호작용(interaction)이 있는가?

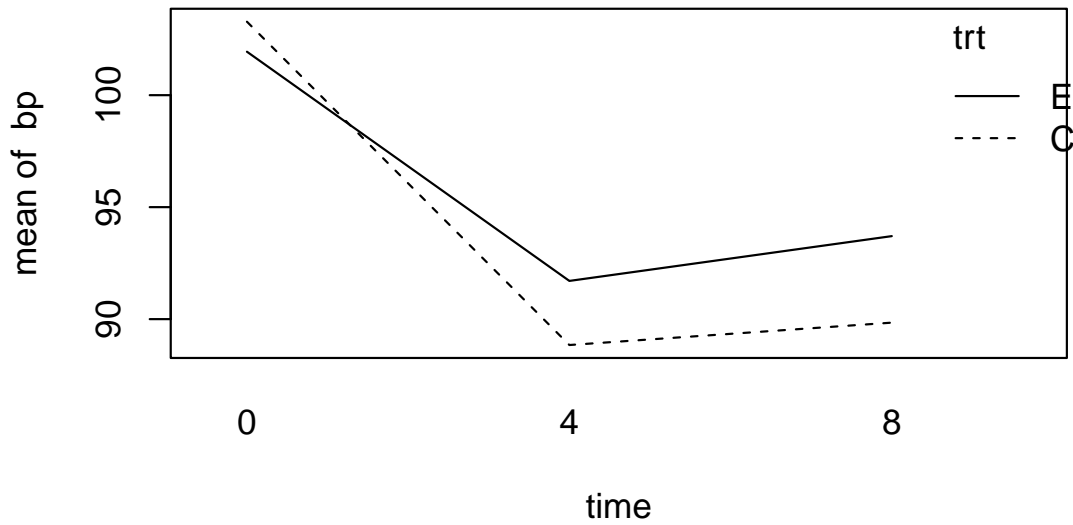
이제 자료를 읽어서 `blood_data`에 저장하자.

```
blood_data<- read.csv(here("data", "ex0601.txt"), header=T, sep="")
head(blood_data,5)
```

	id	trt	time	bp
1	1	C	0	95
2	2	C	0	100
3	3	C	0	98
4	4	C	0	114
5	5	C	0	110

모형식을 고려하기 전에 각 처리 그룹에서 시간에 따른 혈압의 변화를 살펴보자. 강하제를 복용한 후에 평균 혈압은 크게 내려갔지만 8주 후에는 다소 올라가는 경향을 보인다.

```
with(blood_data, interaction.plot(x.factor = time, trace.factor = trt, response = bp))
```



8.5.1. 혼합효과 모형

반응값 y_{ijk} 를 i 번째 강하제($i = 1, 2$)를 복용한 j 번째 환자의 k 시간에 측정한 혈압이라고 하자($j = 1, 2, \dots, n_i$, $k = 0, 1, 2$).

먼저 약을 복용하기 전의 혈압 y_{ij0} 은 처리 전의 기준이 되는 값이다. 따라서 우리는 4주와 8주에 측정한 혈압, y_{ij1} 와 y_{ij2} 에서 0주에 측정한 혈압 y_{ij0} 을 뺀 두 개의 차이, d_{ij1} 과 d_{ij2} 를 반응 변수로 사용할 것이다. 각 환자는 2개의 관측값을 가지며 이는 반복측정 자료이다.

$$d_{ij1} = y_{ij1} - y_{ij0}, \quad d_{ij2} = y_{ij2} - y_{ij0}$$

혈압의 차이를 계산하여 다음과 같이 새로운 데이터프레임을 만들자.

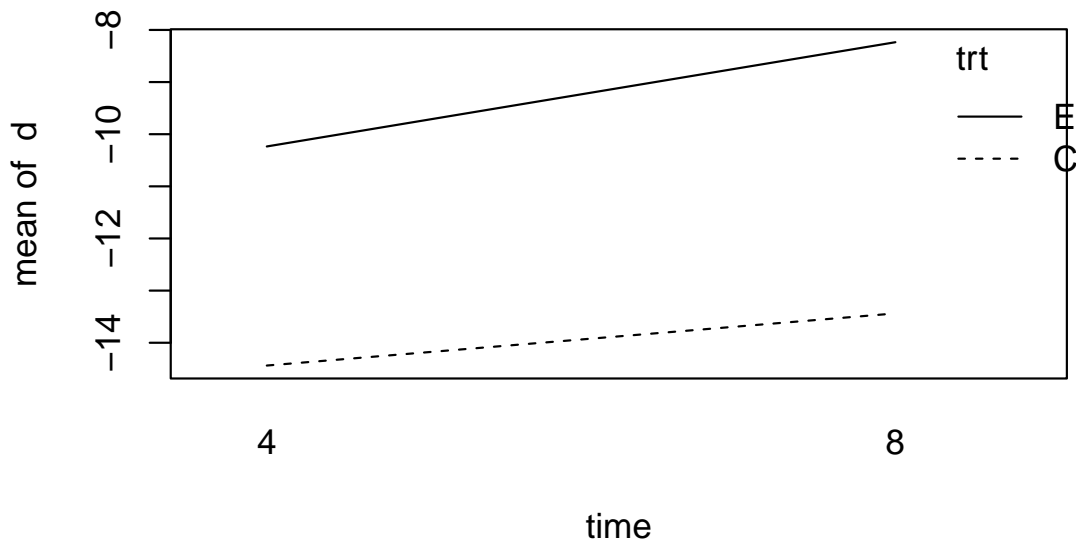
```
df0 <- blood_data %>% filter(time == 0) %>% select(id, bp) %>% rename( bp0 = bp)
df1 <- blood_data %>% filter(time != 0)
blood_data1 <- left_join(df1, df0, by='id') %>% mutate(d = bp - bp0)
head(blood_data1 ,10)
```

	id	trt	time	bp	bp0	d
1	1	C	4	88	95	-7
2	2	C	4	95	100	-5
3	3	C	4	90	98	-8
4	4	C	4	98	114	-16
5	5	C	4	81	110	-29

8. 반복측정자료

6	6	C	4	98	100	-2
7	7	C	4	92	100	-8
8	8	C	4	79	100	-21
9	9	C	4	81	107	-26
10	10	C	4	85	101	-16

```
with(blood_data1, interaction.plot(x.factor = time, trace.factor = trt, response = d))
```



이제 다음과 같은 모형을 고려하자.

$$d_{ijk} = \mu + \alpha_i + \beta_k + (\alpha\beta)_{ik} + b_{ij} + e_{ijk}, \quad i = 1, 2 \quad j = 1, \dots, n_i \quad k = 1, 2 \quad (8.14)$$

일단 반복 측정한 시점이 2번이므로 측정 시간은 연속형 변수가 아닌 범주형 변수로 지정한다.

. 위의 모형식 8.14 에서 각 항의 의미는 다음과 같다.

- μ : 전체 평균 혈압
- α_i : 혈압 강하제의 효과 ($i = 1, 2$)
- β_k : 시간 효과($k = 1, 2$)
- $(\alpha\beta)_{ik}$: 강하제와 시간의 상호작용
- $b_{ij} \sim_{iid} N(0, \sigma_b^2)$: 환자의 효과를 나타내는 임의효과 ($i = 1, 2, j = 1, 2, \dots, n_i$)
- $e_{ijk} \sim_{iid} N(0, \sigma_e^2)$: 오차항

위에서 환자의 효과를 나타내는 임의효과와 오차항은 독립이라고 가정한다.

이제 자료에서 치료약(trt), 시간(time), 환자의 ID(id)를 범주형 변수로 변환하자.

8. 반복측정자료

```
blood_data1$id <- factor(blood_data1$id)
blood_data1$trt <- factor(blood_data1$trt)
blood_data1$time <- factor(blood_data1$time)
```

이제 다음과 같은 R 명령문을 이용하여 모형 8.14 을 적합하고 추정 결과를 살펴 보자.

```
blood_fit <- lmer(d ~ trt + time + trt*time +(1|id), data=blood_data1)
summary(blood_fit)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
```

```
Formula: d ~ trt + time + trt * time + (1 | id)
```

```
Data: blood_data1
```

```
REML criterion at convergence: 862
```

```
Scaled residuals:
```

Min	1Q	Median	3Q	Max
-1.92294	-0.44315	0.05264	0.36653	2.14310

```
Random effects:
```

Groups	Name	Variance	Std.Dev.
id	(Intercept)	58.00	7.616
Residual		14.87	3.857

```
Number of obs: 132, groups: id, 66
```

```
Fixed effects:
```

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	-14.4375	1.5091	78.3627	-9.567	8.24e-15 ***
trtE	4.2022	2.1026	78.3627	1.999	0.0491 *
time8	1.0000	0.9642	64.0000	1.037	0.3036
trtE:time8	1.0000	1.3434	64.0000	0.744	0.4594

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Correlation of Fixed Effects:
```

	(Intr)	trtE	time8
trtE		-0.718	
time8	-0.319	0.229	
trtE:time8	0.229	-0.319	-0.718

위의 적합 결과를 보면 다음과 같은 해석이 가능하다.

8. 반복측정자료

1. 혈압 강하제 간의 유의한 차이는 있다. 기존의 혈압 강하제(C)가 더 효과적이다
2. 시간에 따른 효과의 차이는 유의하지 않다.
3. 강하제와 시간 간의 상호작용도 유의하지 않다.

또한 환자의 개인적인 효과를 반영한 임의효과의 분산이 오차항의 분산의 거의 4배이다. 이는 개인 간의 변동이 개인 내의 변동보다 매우 크다는 것을 의미한다.

$$\hat{\sigma}_b^2 = 58.00 \quad \hat{\sigma}_e^2 = 14.87$$

반복측정한 자료의 상관계수, 즉 그룹 내의 관측값의 상관계수는 0.7956 으로 매우 높다.

$$\hat{\rho}(d_{ij1}, d_{ij2}) = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_e^2} = \frac{58.0}{58.0 + 14.9} = 0.7956$$

이제 각 요인에 대한 유의성을 보기 위하여 분산분석을 실시해 보자. 선택 명령문으로 `ddf = "Kenward-Roger"` 를 사용하는 것이 좋으며 이는 혼합효과 모형에서 잔차제곱합의 자유도를 적절하게 추정해 준다. 위의 분산분석 결과를 보면 약의 효과와 상호작용은 앞에서 내린 결론과 동일하지만 시간에 따른 효과가 유의하게 나타난다.

```
anova(blood_fit, ddf = "Kenward-Roger")
```

Type III Analysis of Variance Table with Kenward-Roger's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
trt	82.855	82.855	1	64	5.5701	0.02133 *
time	74.182	74.182	1	64	4.9870	0.02904 *
trt:time	8.242	8.242	1	64	0.5541	0.45937

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

8.5.2. 임의효과가 없는 모형

만약 반복측정 자료의 구조를 무시하고 개인에 대한 임의효과를 고려하지 않으면 어떻게 될까? 이러한 가정은 실험에서 얻은 모든 반응값들이 독립이라는 것이다.

다음과 같이 개인에 대한 임의효과가 없는 모형을 고려하고 적합해 보자.

$$d_{ijk} = \mu + \alpha_i + \beta_k + (\alpha\beta)_{ik} + e_{ijk}, \quad i = 1, 2 \quad j = 1, \dots, n_i \quad k = 1, 2 \quad (8.15)$$

```
blood_fit1 <- lm(d ~ trt + time + trt*time, data=blood_data1)
summary(blood_fit1)
```

Call:

```
lm(formula = d ~ trt + time + trt * time, data = blood_data1)
```

Residuals:

Min	1Q	Median	3Q	Max
-17.765	-6.562	-1.562	6.286	23.235

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-14.438	1.509	-9.567	<2e-16 ***
trtE	4.202	2.103	1.999	0.0478 *
time8	1.000	2.134	0.469	0.6402
trtE:time8	1.000	2.974	0.336	0.7372

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.537 on 128 degrees of freedom

Multiple R-squared: 0.08017, Adjusted R-squared: 0.05861

F-statistic: 3.719 on 3 and 128 DF, p-value: 0.01323

```
Anova(blood_fit1, type="III")
```

Anova Table (Type III tests)

Response: d

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	6670.1	1	91.5284	< 2e-16 ***
trt	291.1	1	3.9945	0.04777 *
time	16.0	1	0.2196	0.64018
trt:time	8.2	1	0.1131	0.73719
Residuals	9328.0	128		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

혈압 강하제의 효과가 임의효과를 포함한 모형과 같이 유의하게 나타난다. 그런데 여기서 임의효과를 포함 하지 않는 모형의 분석 결과에 유의해야 할 점은 다음과 같다.

- 임의효과를 고려하지 않으면 마치 66명의 두 배인 132 명으로 구성된 독립 표본으로 실험한 것처럼 가정하는 것이다. 따라서 표본의 수가 늘어나는 효과가 생긴다.
- 반면 임의 효과로 설명되는 그룹의 변동이 오차의 변동에 합쳐지기 때문에 오차항의 분산이 크게 추정된다.

8. 반복측정자료

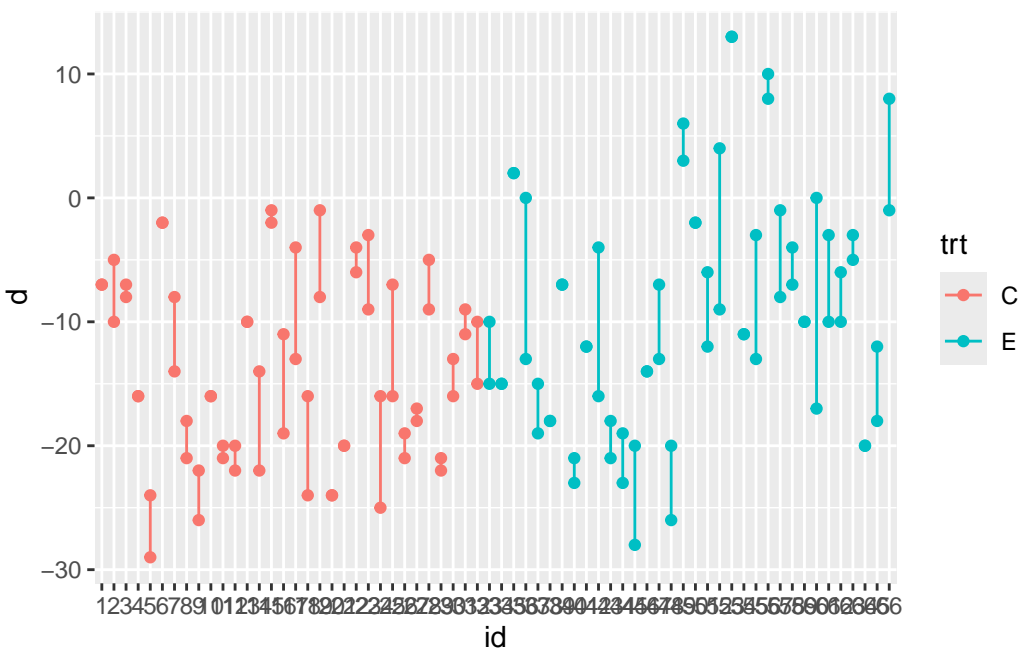
위의 두 분석 결과를 비교해보면 혼합효과 모형에서 오차항의 분산에 대한 추정치는 $\hat{\sigma}_e^2 = 14.87$ 이지만 임의효과를 고려하지 않은 모형에서는 $\hat{\sigma}_e^2 = (8.537)^2 = 72.88$ 로 추정된다.

하지만 임의효과를 고려하지 않은 모형에서는 표본의 수가 늘어나므로, 오차항의 분산이 크게 추정되더라도 추정량에 대한 표준오차(standard error)는 혼합모형과 유사하게 나타난다.

혼합모형으로 추정한 그룹 내 상관관계수($\hat{\rho} = 0.8$)가 매우 크게 나타났으므로 관측값들이 모두 독립이라는 가정은 적절하지 않다.

각 환자에 대하여 두 반응값을 연결하여 그리면 환자들의 두 반응값들이 매우 가깝게 나타나는 것을 알 수 있다.

```
ggplot(blood_data1, aes(x=id, y=d, color=trt)) +
  geom_point() +
  geom_line()
```



8.5.3. 공분산분석

실제로 많은 임상 실험에서는 여러 번 반복 측정을 실시해도 결과를 분석하는 경우 최종 측정값만을 고려한다. 이제 약을 복용하기 전과 8주의 혈압의 차이를 반응변수로 하는 공분산 분석을 실시해 보자.

약을 복용하기 전과 8주의 혈압의 차이 $d_{ij2} = y_{ij2} - y_{ij0}$ 를 반응변수로 하고 공변량을 실험 전 혈압 y_{ij0} 으로 하여 분석을 실시한다. 모형은 공분산분석 모형으로 다음과 같다.

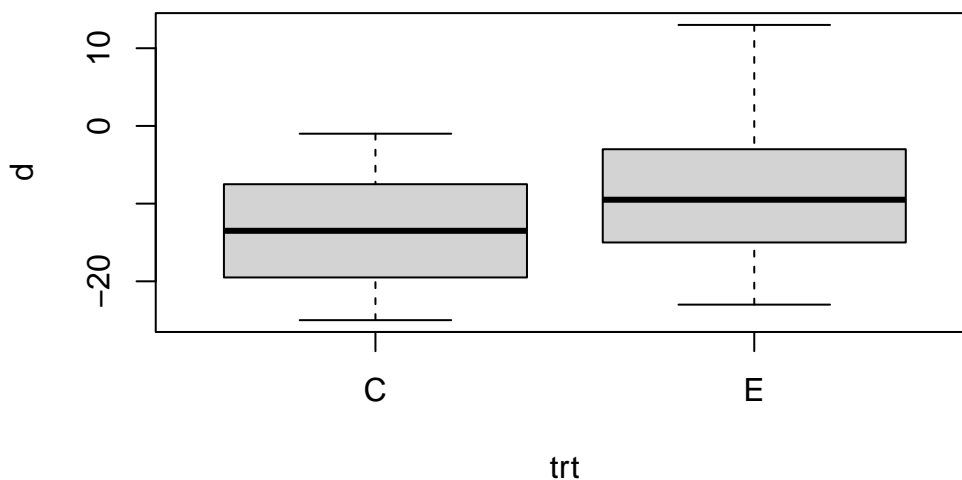
$$d_{ij2} = \mu + \alpha_i + \beta y_{ij0} + e_{ij}$$

```
blood_data2 <- blood_data1 %>% filter(time == 8)
head(blood_data2,5)
```

8. 반복측정자료

	id	trt	time	bp	bp0	d
1	1	C	8	88	95	-7
2	2	C	8	90	100	-10
3	3	C	8	91	98	-7
4	4	C	8	98	114	-16
5	5	C	8	86	110	-24

```
plot(d ~ trt, data=blood_data2)
```



이제 공분산 모형을 적합해 보자. 처리 간의 차이도 매우 유의하고 실험 전의 혈압도 유의하다. 주목할 점은 실험 전의 혈압에 대한 회귀계수의 부호가 음수이므로 약을 먹기 전에 혈압이 높을수록 약을 먹은 후의 혈압이 더 많이 감소한다. 이러한 현상은 임상실험에서 자주 나타나는 현상이다. 일반적으로 질병의 정도가 심한 환자일 수록 평균적으로 치료 효과가 더 커진다.

```
blood_ancova <- lm( d ~ trt + bp0, data = blood_data2)
summary(blood_ancova)
```

Call:

```
lm(formula = d ~ trt + bp0, data = blood_data2)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.982	-6.485	-0.014	5.300	23.505

Coefficients:

8. 반복측정자료

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.8989    21.3060   1.544   0.1276
trtE         4.6010     1.9778   2.326   0.0232 *
bp0        -0.4486     0.2058  -2.180   0.0330 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 7.952 on 63 degrees of freedom
Multiple R-squared:  0.1578,    Adjusted R-squared:  0.1311
F-statistic: 5.903 on 2 and 63 DF,  p-value: 0.004469

```

```
Anova(blood_ancova, type="III")
```

Anova Table (Type III tests)

Response: d

```

              Sum Sq Df F value  Pr(>F)
(Intercept)  150.8   1   2.3843 0.12757
trt          342.2   1   5.4115 0.02323 *
bp0          300.4   1   4.7505 0.03304 *
Residuals    3983.6  63
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Part III.

생존분석

9. 생존함수와 위험함수

9.1. 필요한 패키지

```
library(tidyverse)
library(kableExtra)
library(ggplot2)
library(here)

library(survival)
```

9.2. 생존분석 소개

생존분석(Survival Analysis)은 특정 사건(event)이 발생하기까지의 시간을 분석하는 통계적 방법이다. 여기서 생존이란 단어는 좁은 의미에서 생존 시간을 의미하지만, 실제로는 질병 발생, 제품 고장, 사건 재발 등 다양한 사건에 적용될 수 있다.

9.2.1. 생존 시간의 정의

생존 시간(Survival Time)은 관심 사건(event)이 일어날 때까지의 시간 간격(time interval)을 말한다. 생존 시간이 발생하는 중요한 분야와 예는 다음과 같다.

- 의학 및 생명 과학 분야
 - 암 환자의 생존 시간 : 암 치료 후 환자가 재발이나 사망할 때까지의 시간. 생존 시간은 종종 치료 효과나 병의 진행 상태를 평가하는 데 중요한 지표로 사용된다.
 - 수술 후 회복 기간 : 수술 후 환자가 회복되어 퇴원할 때까지 걸리는 시간. 이는 회복에 영향을 미치는 요인(예: 연령, 수술 종류 등)을 평가할 때 유용하다.
 - 감염 질환의 치료 시간 : 환자가 감염 질환을 치료받고 증상이 완화되기까지의 시간. 특정 치료법의 효과를 비교하는 데 사용된다.
- 공학 및 제조 분야
 - 기계 부품의 고장 시간 : 기계나 전자 장비의 부품이 처음 사용된 후 고장 나기까지 걸리는 시간. 이는 제품의 내구성 및 품질 관리에 필수적인 데이터로 사용된다.

9. 생존함수와 위험함수

- 선박 엔진 수명: 엔진이 처음 가동되어 고장 나거나 수리/교체가 필요할 때까지의 시간. 각 제조사나 모델에 따라 엔진의 수명이 달라지므로, 이를 비교해 품질과 성능을 평가할 수 있다.
- 가전제품의 고장 시간: TV, 냉장고, 핸드폰 등의 가전제품이 처음 설치된 후 고장 나기까지 걸리는 시간. 이 정보는 고객 보증 기간 설정과 결함 분석에 중요한 역할을 한다.

• 사회 및 경제학 분야

- 고용 지속 기간: 신규 채용된 직원이 특정 회사에서 근무를 지속하는 기간. 이는 이직률, 근무 환경의 질, 직무 만족도 등을 분석하는 데 활용된다.
- 구독 서비스 유지 시간: 고객이 스트리밍 서비스, 구독 서비스 등 서비스에 가입하여 중단하기까지 걸리는 시간. 이를 통해 구독 모델의 효율성과 마케팅 전략을 평가할 수 있다.
- 대출 상환 기간: 고객이 금융 대출을 받고 완전히 상환하기까지의 기간. 이는 채무 불이행 가능성을 예측하고 대출 조건을 설정하는 데 중요하다.

위의 예제에서 알 수 있듯이 생존분석은 시간에 따라서 특정사건이 발생할 가능성이 어떻게 변하는지에 대한 방법을 연구하는 분야이다.

9.2.2. 생존 시간의 특징

생존 시간은 다음과 같은 특별한 성질을 가진다.

1. 생존시간은 0을 포함하는 양수이다. 즉, 생존시간은 음수로 나타낼 수 없으며 생존시간에 대한 분포를 고려하는 경우 0보다 큰 값만 가질 수 있는 분포를 고려해야 한다. 예를 들어 정규분포는 생존시간의 분포로 사용하기 적합하지 않다.
2. 많은 연구에서 생존 시간을 관측하려면 일정한 기간을 정하고 시간이 흐르면서 관측값을 얻는 경우가 많다. 만약 생존시간이 연구기간보다 길면 생존시간을 정확하게 관측할 수 있다. 이러한 현상을 중도절단(censoring)이라 하며, 생존분석에서 매우 중요한 개념이다.
 - 예를 들어 임상실험은 일반적으로 실험에 참가할 환자를 모집하고(accural) 일정 기간의 치료기간(treatment) 또는 관찰기간(follow-up)을 정하여 실시한다. 만약 관찰하려는 특정사건(event)가 실험의 종료 이후에 발생하면(censoring) 생존시간에 대한 중도절단자료(censored data)가 발생한다.
 - 환자가 치료 실험에 참여하고 있는데, 치료 효과를 보지 못하거나 다른 이유로 실험에서 중도에 이탈할 수 있어. 이 경우, 연구자들은 환자가 실험을 중단한 시점까지만 관찰할 수 있게 되므로, 정확한 사건 발생 시간을 알 수 없게 된다.
 - 또는 실험에 참가한 환자의 특정 시간 발생이 실험의 종료 후에 발생하면 이 환자의 생존시간은 정확하게 알 수 없다. 다만 종료된 시점 이후에 특정사건이 일어난다는 사실은 명확하므로 중도절단 자료는 불완전한 정보(incomplete information)를 알려준다.
 - 아래 그림은 임상실험에서 중도절단 자료가 발생하는 상황을 단순하게 나타낸 그림이다. 3명의 환자(patient 1,3,4)의 사건이 임상실험 종료후 발생하면 이 환자들의 생존시간은 정확하게 알 수 없다.

Fig. 1.1 Clinical trial accrual and follow-up periods. The *vertical dashed lines* indicate the trial start, end of accrual, and end of follow-up. The X's denote deaths and the *open circles* denote censoring events

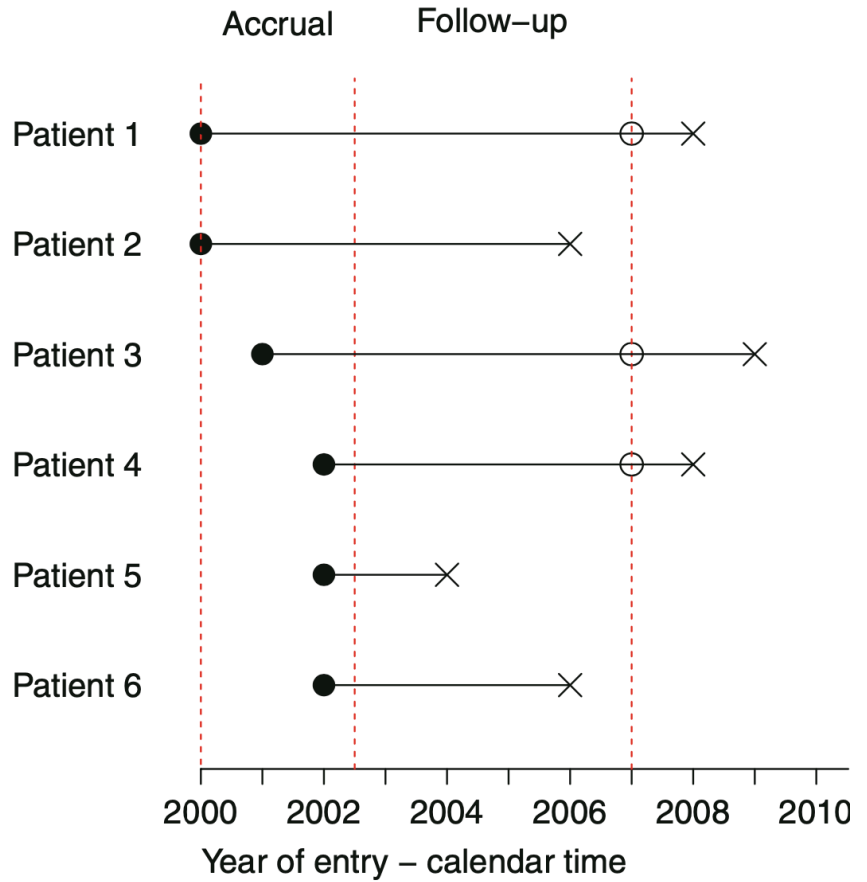


그림 9.1.: 중도 절단 자료

9.3. 생존함수와 위험함수

확률변수 $T > 0$ 를 생존시간이라고 하고 $f(t)$ 를 확률밀도함수라고 하자. T 의 누적분포함수(cumulative distribution function; CDF)는 다음과 같이 정의된다.

$$F(t) = P(T \leq t) = \int_0^t f(t)dt \quad (9.1)$$

또한 생존함수(Survival function)은 다음과 같이 정의된다.

$$S(t) = P(T > t) = 1 - F(t) \quad (9.2)$$

위험함수(hazrd function)의 정의는 다음과 같으며 만약 생존시간이 t 보다 클때 바로 사망할 확률을 의미하며 순간위험률(instantaneous failure rate)이다.

9. 생존함수와 위험함수

$$\begin{aligned}
 h(t) &= \lim_{dt \rightarrow 0} \frac{P(t < T \leq t + dt | T > t)}{dt} \\
 &= \lim_{dt \rightarrow 0} \frac{1}{dt} \frac{P(P(t < T \leq t + dt))}{P(T > t)} \\
 &= \frac{f(t)}{S(t)}
 \end{aligned}$$

위험함수의 의미를 좀더 자세히 알아보자.

- 위험함수 9.3 는 아주 짧은 시간에 일어날 사건의 확률로서 순간적인(instantaneous)인 의미를 가진다. 정의에서 보듯이 아주 짧은 시간에 일어나는 극한($\lim_{dt \rightarrow 0}$)의 의미이다.
- 위험함수는 확률로서 정의되지만 사건 발생률(rate)로 이해할 수 있다. 위험함수의 정의에서 분모는 확률이고 분자는 시간으로 표시되기 때문에 마치 단위 시간당 일어나는 발생률로 이해할 수 있다. 또한 위험함수의 값은 $[0,1]$ 사이에 있는 것이 아니기 때문에 단위 시간당 발생률로 이해하는 것이 편리하다 ($h(t) \geq 0$).
- 사건 발생률은 분야에 따라서 사망률, 실패율, 위험률 등으로 불린다.
- 위험함수는 조건부 확률로 정의되어 있다. 즉, 모든 개체에 대한 사건이 아니라 시간 t 까지 살아있는 개체들에 대한 위험률이다.
- 생존분석에서는 생존함수보다 위험함수가 더 큰 의미를 가진다. 위험함수가 시간에 따라서 변하는 추세는 시간에 따른 위험의 변화를 나타내기 때문에 생존함수보다 위험의 변화에 대한 정보를 잘 파악할 수 있다.

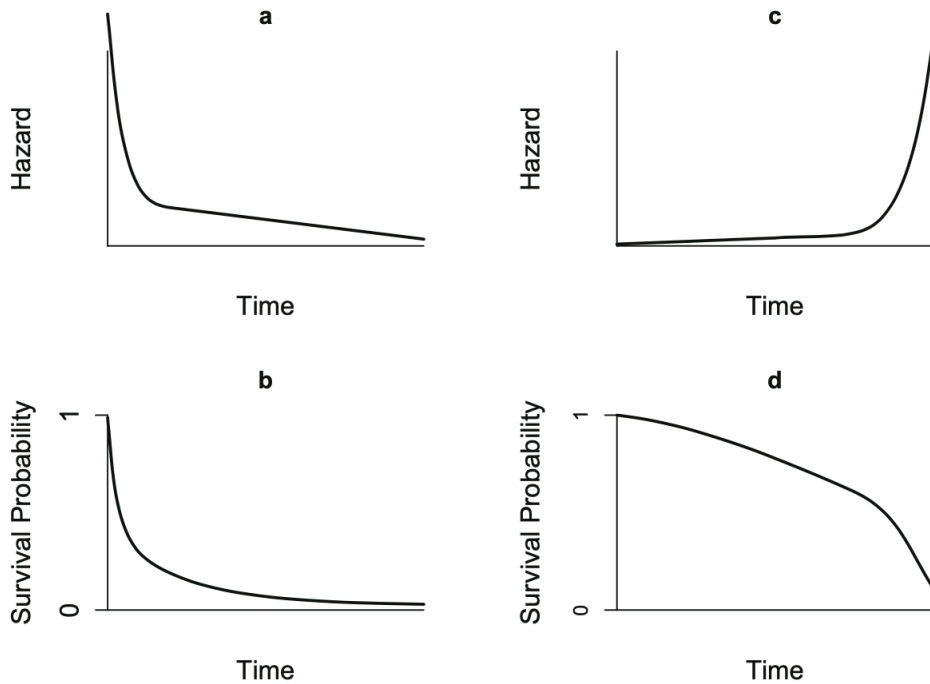


Fig. 2.1 Hazard and survival functions with high initial hazard (a and b) and low initial hazard (c and d)

그림 9.2.: 위험함수와 생존함수의 관계

9. 생존함수와 위험함수

생존함수와 위험함수는 다음과 같은 관계를 가지고 있다.

$$\begin{aligned}h(t) &= \frac{f(t)}{S(t)} \\&= \frac{f(t)}{1 - F(t)} \\&= -\frac{\partial}{\partial t} \log[1 - F(t)] \\&= -\frac{\partial}{\partial t} \log S(t)\end{aligned}$$

따라서 다음의 관계가 성립한다.

$$S(t) = \exp[-H(t)]$$

여기서

$$H(t) = \int_0^t h(t)dt \quad (9.3)$$

이며 $H(t)$ 를 누적 위험함수(cumulative hazard function)라고 한다.

위의 식에서 볼 수 있듯이 위험함수를 알면 생존함수를 쉽게 구할 수 있고 또한 반대로 생존함수를 알면 위험함수를 자동적으로 알게된다.

9.4. 위험함수의 형태

앞에서 언급했듯이 위험함수는 시간에 따른 위험의 변화를 말해준다. 생존분석에서 중요한 분포들과 예제를 가지고 다양한 위험함수의 형태를 알아보자.

Example 9.1 (지수분포). 만약 생존시간 T 가 지수분포(Exponential distribution)을 따른다고 하자.

$$f(t; \lambda) = \lambda e^{-\lambda t}$$

위에서 λ 는 흔히 단위시간 당 사망률, 실패율(failure rate)라고 부르며 $E(T) = 1/\lambda$ 이다.

생존함수(survaival function)와 위험함수(hazard function)은 다음과 같이 주어진다.

9. 생존함수와 위험함수

$$\begin{aligned}
 S(t) &= P(T > t) = 1 - P(T \leq t) \\
 &= 1 - \int_0^t \lambda e^{-\lambda t} \\
 &= 1 - (1 - e^{-\lambda t}) \\
 &= e^{-\lambda t} \\
 h(t) &= \frac{f(t)}{S(t)} \\
 &= \lambda e^{-\lambda t} / e^{-\lambda t} \\
 &= \lambda
 \end{aligned}$$

누적위험함수는 아래와 같다.

$$H(t) = \int_0^t h(t)dt = \lambda t$$

생존시간이 지수분포를 따른다면 위험함수는 시간에 관계없이 상수 λ 이다. 이렇게 시간에 따라서 위험함수가 상수인 경우는 현실에서는 거의 나타나지 않는다. 위험함수가 상수라는 것은 시간이 지나도 (1) 제품이 고정날 또는 (2) 사람이 사망할 위험성이 변하지 않는다는 의미이다.

지수분포는 생존분석에서 중요한 의미를 가진다. 즉, 생존시간이 지수분포를 따르면 개체는 얼마나 시간이 경과 했는지 정보가 없다는 것이며 주어진 시각에 사건이 일어나지 않았다면 바로 다음 시간에 사건이 일어날 확률은 언제나 동일하다. ■

Example 9.2 (와이블분포). 만약 생존시간이 와이블 분포(Weibull distribution)를 따른다면 확률밀도함수는 다음과 같다.

$$f(t) = \frac{\lambda t^{\lambda-1}}{\theta^\lambda} \exp \left[- \left(\frac{t}{\theta} \right)^\lambda \right], \quad \theta, \lambda > 0, t \geq 0.$$

평균은 $E(T) = \theta \Gamma(1 + 1/\lambda)$ 이다. 여기서 $\lambda = 1$ 이면 지수분포가 된다.

이때 생존 함수는 다음과 같다.

$$\begin{aligned}
 S(t) &= 1 - \int_0^t \frac{\lambda t^{\lambda-1}}{\theta^\lambda} \exp \left[- \left(\frac{t}{\theta} \right)^\lambda \right] dt \\
 &= \exp \left[- \left(\frac{t}{\theta} \right)^\lambda \right].
 \end{aligned}$$

또한 위험함수는 다음과 같다.

9. 생존함수와 위험함수

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\frac{\lambda t^{\lambda-1}}{\theta^\lambda} \exp\left[-\left(\frac{t}{\theta}\right)^\lambda\right]}{\exp\left[-\left(\frac{t}{\theta}\right)^\lambda\right]} \\ &= \left(\frac{\lambda}{\theta^\lambda}\right) t^{\lambda-1}. \end{aligned}$$

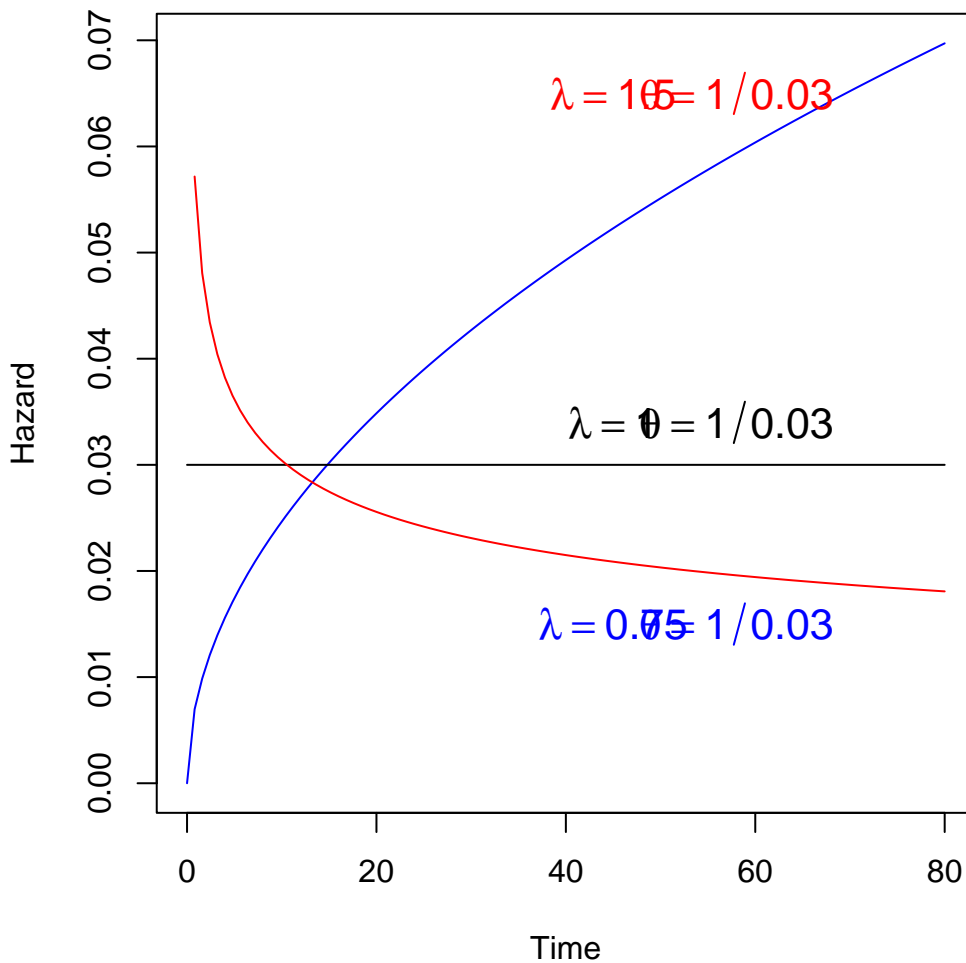
위에서 볼 수 있듯이 $\lambda = 1$ 이면 지수분포가 되고 위험함수는 상수이다. 만약에 $\lambda > 1$ 이면 위험함수는 시간에 따라 증가하며 $\lambda < 1$ 이면 위험함수는 감소한다.

누적위험함수는 아래와 같다.

$$H(t) = \int_0^t h(t)dt = (t/\theta)^\lambda$$

다음은 모수의 변화에 따른 와이블 분포의 위험함수이다.

```
weibHaz <- {function(x, shape, scale) dweibull(x, shape=shape,
  scale=scale)/pweibull(x, shape=shape, scale=scale, lower.tail=F)}
curve(weibHaz(x, shape=1.5, scale=1/0.03), from=0, to=80,
  ylab='Hazard', xlab='Time', col="blue")
curve(weibHaz(x, shape=1.0, scale=1/0.03), from=0, to=80,
  ylab='Hazard', xlab='Time', add=T, col="black")
curve(weibHaz(x, shape=0.75, scale=1/0.03), from=0, to=80,
  ylab='Hazard', xlab='Time', add=T, col="red")
text(45, 0.065, expression(lambda == 1.50), col="red", cex=1.3)
text(58, 0.065, expression(theta==1/0.03), col="red", cex=1.3)
text(45, 0.015, expression(lambda == 0.75), col="blue", cex=1.3)
text(58, 0.015, expression(theta==1/0.03), col="blue", cex=1.3)
text(45, 0.034, expression(lambda== 1.00), col="black", cex=1.3)
text(58, 0.034, expression(theta==1/0.03), col="black", cex=1.3)
```



■

Example 9.3 (미국 여자와 남자의 연령에 따른 사망위험률). 다음 그림은 미국 여자와 남자의 연령에 따른 사망위험률을 나타내는 위험함수를 보여주는 그림이다. Y 축은 로그 스케일로 그린 그림이며 2004년의 자료를 이용하여 그린 그림이다. 태어난 날과 그 후 몇 주동안은 매우 사망위험이 높다가 급격하게 떨어진다. 그리고 10대에서 약간의 증가 현상을 보이며 조금 감소하다가 시간이 지날수록 위험률이 완만하게 증가한다. 주목할 점은 남자가 여자보다 사망의 위험이 언제나 크다는 것이며 이는 잘 알려진 사실이다.

```
tm <- c(0, # birth
        1/365, # first day of life
        7/365, # seventh day of life
        28/365, # fourth week of life
        1:107) # subsequent years
hazMale <- as.numeric(survexp.us[, "male", "2004"]) # 2004 males
hazFemale <- as.numeric(survexp.us[, "female", "2004"]) # 2004 females
tm.diff <- diff(tm)
```

```

survMale <- exp(-cumsum(hazMale*tm.diff)*365.24)
survFemale <- exp(-cumsum(hazFemale*tm.diff)*365.24)

# Figure 2.1.2 log hazard and survival for US males and females in 2004
par(mfrow=c(2,1),      # two rows and one column of plots
    mar=c(4.2,5,2,2)) # set margins for the lower, left, top, and right of each plot

logHazMale <- log(hazMale)
logHazFemale <- log(hazFemale)

plot(logHazMale ~ tm[-1], type="l",
     xlab="Age in years",          # x axis label
     ylab="Hazard",col="blue",    # y axis label
     lwd=2,                       # double line width
     las=1,                      # make y axis labels perpendicular to axis
     axes=F, cex.lab=1.3, cex.axis=1.3) # make blue line solid
lines(logHazFemale ~ tm[-1],type="l",
     col="red",lwd=2, lty=2)      # add a red dashed line to the plot

yyLabs <- c(1e-07, 1e-06, 1e-05, 1e-04, 1e-03, 1e-02)
yyLabsLog <- log(yyLabs)
axis(2, at=yyLabsLog, labels=c(expression(10^-7), expression(10^-6),
  expression(10^-5), expression(10^-4), expression(10^-3), expression(10^-2)), las=1)
axis(1, cex.axis=1.3)
legend("bottomright", legend=c("males","females"),
      lty=c(1,2), col=c("blue","red"), lwd=2, cex=1.3)
title("Hazards for US males and females in 2004")

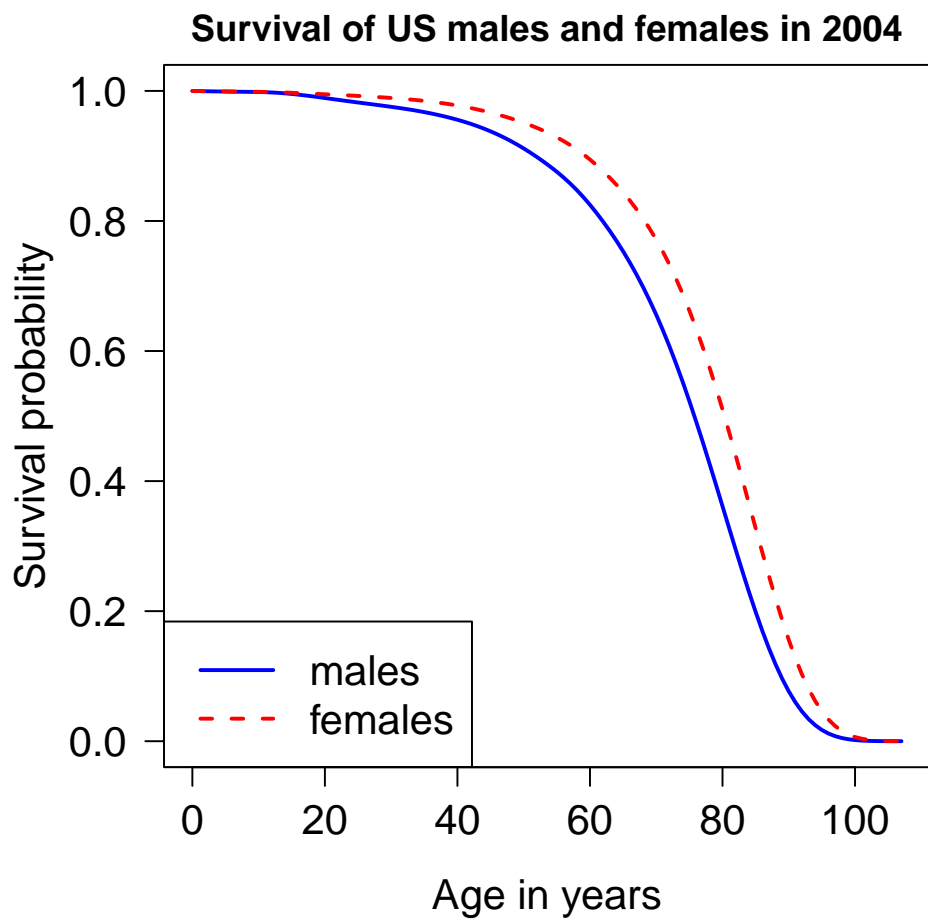
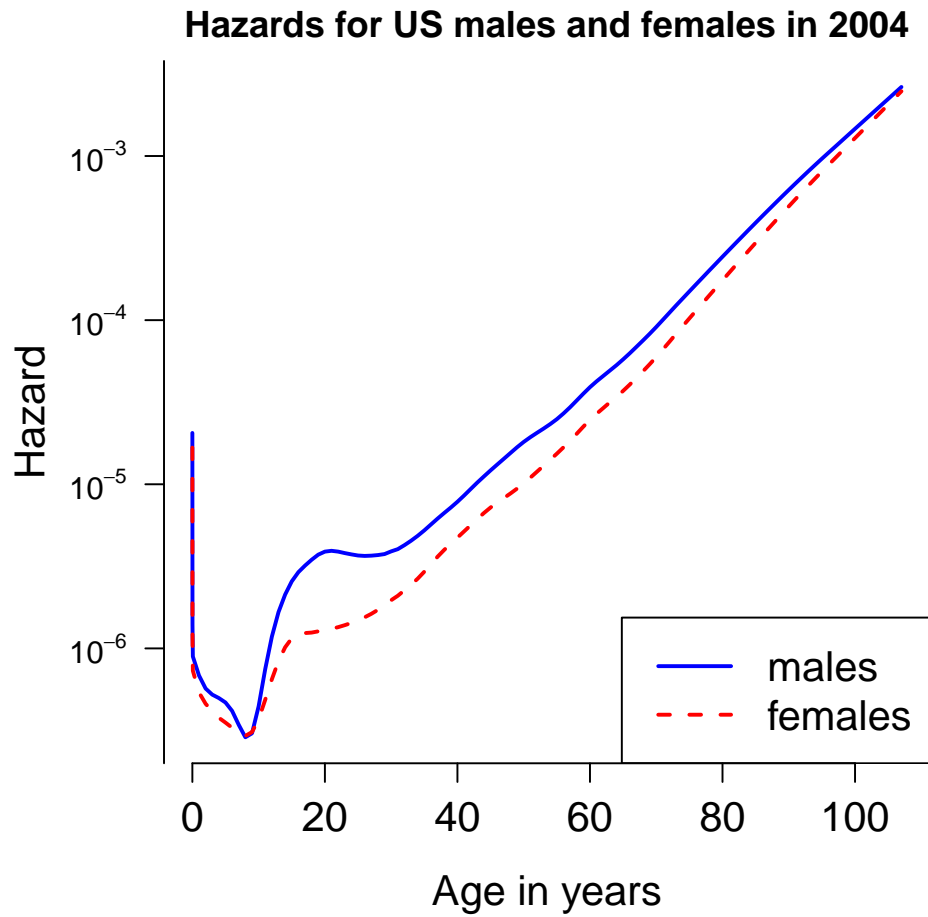
tm.diff <- diff(tm)          # same length as "tm"
survMale <- exp(-cumsum(hazMale*tm.diff)*365.24)          # survival probs for males
survFemale <- exp(-cumsum(hazFemale*tm.diff)*365.24)      # survival probs for females
#windows(width=7,height=5)
plot(survMale ~ tm[-1],type="l",          # lower case "l" indicates line plot
     xlab="Age in years",                # x axis label
     ylab="Survival probability",        # y axis label
     col="blue",                        # line color
     lwd=2,                            # double line width
     las=1,                            # make y axis labels perpendicular to axis
     ylim=c(0,1), cex.lab=1.3, cex.axis=1.3) # y axis limit ranges from 0 to 1

lines(survFemale ~ tm[-1], col="red", lwd=2, lty=2)      # add a red dashed line to the plot
legend("bottomleft", legend=c("males","females"),

```

9. 생존함수와 위험함수

```
lty=c(1,2), col=c("blue","red"), lwd=2, cex=1.3)  
title("Survival of US males and females in 2004")
```





9.5. 평균 생존시간과 중위 생존시간

생존시간의 평균은 다음과 같이 정의된다.

$$E(T) = \int_0^{\infty} t f(t) dt$$

잘 알려진 생존함수의 공식

$$f(t) = -\frac{dS(t)}{dt}$$

와 부분 적분 공식(integration by part)를 이용하면

평균 생존시간 μ 다음과 같이 구할 수 있다.

$$\begin{aligned} E(T) &= \int_0^{\infty} t \left(-\frac{dS(t)}{dt} \right) dt \\ &= [tS(t)]_0^{\infty} - \int_0^{\infty} S(t) dt \\ &= \int_0^{\infty} S(t) dt \end{aligned}$$

위 식에서 $[tS(t)]_0^{\infty}$ 부분을 계산해보면 $t \rightarrow \infty$ 일 때 $S(t) \rightarrow 0$ 이기 때문에 첫 번째 항은 0이 된다.

중위 생존시간은 T_{med} 다음 방정식을 만족하는 t 의 값이다.

$$T_{med} = \inf \{t \mid S(t) \leq 1/2\}$$

9.6. 임의중도절단

9.6.1. 중도 절단의 이유

의학연구에서 중도절단이 일어나는 이유는 다양하다. 예를 들어 환자가 치료를 받는 도중에 치료효과가 없다고 판단되거나 약물의 부작용 또는 건강상태의 악화로 인하여 실험을 중단하는 경우가 있을 수 있다.

제품 내구성 시험에서 제품의 특성상 특정 부품이 고장이 나면 다른 부품의 내구성 시험을 이어갈 수 없게 되는 경우가 있다. 예를 들어, 자동차 엔진 내구성 시험 중 주요 부품이 고장 나서 전체 엔진 테스트가 중단될 수 있다.

또는 의학 실험의 종료일이 되어 실험이 종료되는 경우, 환자들의 특정 사건이 아직 일어나지 않아서 중도절단이 발생하는 경우도 있다.

많은 이유에서 중도 절단이 발생할 수 있고 이러한 중도절단에 영향을 미치는 요인도 다양하므로 이러한 관계를 통계적 모형으로 구현하는 것은 쉽지 않은 일이다.

특히 생존시간에 영향을 미치는 요인들(예를 들어 환자 질병의 심각성 등)이 중도절단에 영향을 미치는 경우에는 복잡한 모형이 필요하다.

중도절단에 대한 확률적 모형에서 가장 단순한 모형은 중도절단이 무작위로 발생한다고 가정하는 것이다. 이러한 가정을 임의 중도절단(random censoring)이라고 한다.

9.6.2. 임의중도절단

생존시간 T_1, T_2, \dots, T_n 을 독립적으로 분포 $f_\theta(t)$ 에서 추출하였다고 가정하고 중도절단시간 C_1, C_2, \dots, C_n 도 독립적으로 분포 $g(c)$ 에서 추출하였다고 가정하자.

생존시간 T 와 중도절단시간 C 가 서로 독립이라고 가정하며 이러한 가정을 임의 중도절단(random censoring)이라고 한다.

이제 관측한 생존시간 X_i 는 다음과 같이 정의한다.

$$X_i = \min(T_i, C_i) \quad i = 1, 2, \dots, n$$

관측한 생존시간 X_i 은 중도절단이 발생하지 않는다면 실제로 특정사건이 발생할 때까지의 시간 $X_i = T_i$ 이며, 또는 중도절단이 발생하면 중도절단시간까지의 시간 $X_i = C_i$ 이 된다.

더 나아가 중도절단이 발생하였는지 여부를 나타내는 절단 표시변수 δ_i 는 다음과 같이 정의한다. 표시변수 δ_i 가 1의 값을 가지면 실제 생존시간을 관측한 것이고 0이면 절단된 자료이다.

$$\delta_i = \begin{cases} 1 & T_i < C_i \\ 0 & T_i > C_i \end{cases}$$

중도절단이 있는 자료는 각 개체에 대하여 쌍으로 구성된 관측값이 만들수 있다. 아래 코드에서 **Surv** 함수는 생존시간 **t** 와 중도절단 표시변수 **cens** 을 입력으로 받아서 생존시간과 중도절단 표시변수를 쌍으로 구성된 자료를 생성한다. + 표시가 된 값이 중도절단된 자료이다.

```
t <- c(9, 13, 13, 18, 23, 28, 31, 34, 45, 48, 161)
cens <- c(1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0)
survival_data <- Surv(t, cens)
survival_data
```

```
[1] 9 13 13+ 18 23 28+ 31 34 45+ 48 161+
```

Surv 함수로 생성된 자료에서 [,1] 은 생존시간을 나타내고 [,2] 는 중도절단 표시변수를 나타낸다.

```
survival_data[,1]
```

```
[1] 9 13 13 18 23 28 31 34 45 48 161
```

```
survival_data[,2]
```

```
[1] 1 1 0 1 1 0 1 1 0 1 0
```

9.6.3. 최대가능도 추정

이제 생존시간 T 의 분포 $F_\theta(t)$ 가 주어졌을 때 표본 (x_i, δ_i) 의 가능도 함수 L 는 다음과 같이 나타낼 수 있다.

$$L(\theta; x, \delta) = \prod_{i \in UC} P_\theta(T_i = x_i) \prod_{i \in C} P_\theta(T_i > x_i) = \prod_{i \in UC} f_\theta(x_i) \prod_{i \in C} [1 - F_\theta(x_i)]$$

여기서 집합 UC 는 실제 생존시간이 관측된 자료들, 집합 C 는 중도절단된 자료를 의미한다.

모수의 최대가능도 추정은 $L(\theta; X, \delta)$ 를 최대로 하는 모수 θ 의 추정량을 찾는 방법이다.

$$\max_{\theta} L(\theta; X, \delta)$$

Example 9.4 (지수분포). 이제 예제로서 생존시간의 분포가 지수분포를 따른다고 가정하고 가능도함수를 구해보자.

$$\begin{aligned} L(\lambda; x, \delta) &= \prod_{i \in UC} P_\lambda(T_i = x_i) \prod_{i \in C} P_\lambda(T_i > x_i) \\ &= \prod_{i \in UC} f_\lambda(x_i) \prod_{i \in C} [1 - F_\lambda(x_i)] \\ &= \prod_{i=1}^n [\lambda e^{-\lambda x_i}]^{\delta_i} [e^{-\lambda x_i}]^{1-\delta_i} \\ &= \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda x_i} \\ &= \lambda^{\sum_{i=1}^n \delta_i} e^{-\lambda \sum_{i=1}^n x_i} \end{aligned}$$

이제 위의 가능도함수를 최대화하는 모수를 찾기위하여 로그가능도함수를 고려하고

$$\ell(\lambda; x, \delta) = \log L(\lambda; x, \delta) = \log \lambda \sum_{i=1}^n \delta_i - \lambda \sum_{i=1}^n x_i$$

로그가능도함수를 모수 λ 에 대하여 미분하고 0으로 놓고 풀면 최대가능도 추정량을 구할 수 있다.

9. 생존함수와 위험함수

$$\frac{\partial}{\partial \lambda} \ell(\lambda; x, \delta) = \frac{\sum_{i=1}^n \delta_i}{\lambda} - \sum_{i=1}^n x_i = 0$$

따라서 최대가능도 추정량은 다음과 같이 주어진다.

$$\hat{\lambda}_{ML} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n x_i} \quad (9.4)$$

만약 중도절단자료가 없다면 최대가능도 추정량은 일반적인 경우와 같이 다음과 같다.

$$\hat{\lambda}_{ML} = \frac{n}{\sum_{i=1}^n x_i}$$

위에서 만든 생존시간 자료 `survival_data`에 대하여 지수분포에서 추출한 표본이라고 가정하고 최대가능도 추정량을 구해보자. 여기서 주의할 점은 9.4에서 제시된 식은 모든 자료의 합 ($\sum x_i$)을 중도절단된 자료의 수를 뺀 수 ($\sum_i \delta_i$)로 나눈 값의 역수이다. 생존시간 자료인 `survival_data`에 `mean`를 적용하면 중도절단된 자료를 고려하지 않고 평균을 계산한다. 따라서 중도절단된 자료를 고려하여 평균을 계산하려면 다음과 같이 계산해야 한다.

```
mean_survival <- sum(survival_data[,1])/sum(survival_data[,2])
mean_survival
```

```
[1] 60.42857
```

```
lambda_ml <- 1/mean_survival
lambda_ml
```

```
[1] 0.01654846
```

참고로 중도절단된 자료를 모두 포함하여 일반적인 평균을 구하여 모수 λ 를 추정하면 다음과 같이 얻어진다. 위의 결과에서 볼수 있듯이 중도절단된 자료를 고려하지 않으면 추정량이 과대 추정되는 것을 알 수 있다. 즉, 중도절단된 자료를 고려하지 않으면 사건 발생률이 과대평가되는 것이다.

$$\frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n x_i} \leq \frac{n}{\sum_{i=1}^n x_i}$$

```
mean_all <- mean(survival_data[,1])
mean_all
```

```
[1] 38.45455
```

```
1/mean_all
```

```
[1] 0.02600473
```

```
■
```

이제 모의실험을 이용하여 중도절단이 있는 자료에서 최대가능도 추정량과 중도절단을 고려하지 않는 추정량을 비교해보자. 아래 모의실험은 지수분포에서 추출한 자료를 이용하여 중도절단이 있는 자료를 생성하고 최대가능도 추정량과 중도절단을 고려하지 않는 추정량을 비교한다.

```
set.seed(12221) # 재현성을 위해 시드 설정

# 데이터 생성
B <- 1000 # 시뮬레이션 횟수
n <- 11 # 표본의 개수
lambda <- 1 # 실제 생존시간(T_i) - 지수분포의 모수
cens_lambda <- 0.3 # 중도절단 시간(C_i) - 지수분포의 모수

sam_df <- data.frame(lambda_ml = rep(0, B), lambda_all = rep(0, B))

for ( i in 1:B ) {
  t <- rexp(n, rate = lambda) # 지수분포에서 표본 추출
  c <- rexp(n, rate = cens_lambda) # 중도절단 시간 추출
  x <- pmin(t, c) # 중도절단된 자료 생성
  cens <- t < c # 중도절단 표시변수 생성
  survival_data <- Surv(x, cens) # 생존시간과 중도절단 표시변수로 데이터 생성
  mean_survival <- sum(survival_data[,1])/sum(survival_data[,2]) # 중도절단을 고려한 평균
  lambda_ml <- 1/mean_survival # 중도절단을 고려한 최대 가능도 추정량
  mean_all <- mean(survival_data[,1]) # 중도절단을 고려하지 않은 평균
  lambda_all <- 1/mean_all # 중도절단을 고려하지 않은 추정량
  sam_df[i,] <- c(lambda_ml, lambda_all) # 결과 저장
}

# 관측자료의 예
survival_data
```

```
[1] 0.5272836 0.1539508 1.7449258 0.1629189+ 0.8230156 0.5750349
[7] 0.3123141+ 0.3258889 0.1040071 1.9326109 0.0152728
```

```
# 결과 요약
summary(sam_df)
```

9. 생존함수와 위험함수

<code>lambda_ml</code>	<code>lambda_all</code>
Min. :0.3043	Min. :0.6099
1st Qu.:0.8278	1st Qu.:1.1137
Median :1.0461	Median :1.3537
Mean :1.1210	Mean :1.4413
3rd Qu.:1.3250	3rd Qu.:1.6769
Max. :3.6322	Max. :4.4394

위의 결과를 보면 중도절단을 고려하지 않은 일반적인 추정량(`lambda_all`)이 발생율을 과대평가하는 것을 알 수 있다. 중도절단을 고려한 최대가능도 추정량(`lambda_ml`)이 실제 모수에 가까운 추정량을 제공하는 것을 알 수 있다.

10. 생존함수의 추정

10.1. 필요한 패키지

```
library(tidyverse)
library(kableExtra)
library(ggplot2)
library(here)

library(survival)
library(KMsurv)
library(asauro)
library(survminer)
```

이 장에서는 자료를 이용하여 생존함수를 추정하는 방법에 대하여 알아본다. 생존함수를 추정하는 대표적인 두 방법은 다음과 같다.

- 생명표 방법을 이용한 생존함수 추정
- 비모수적 방법을 이용한 생존함수 추정

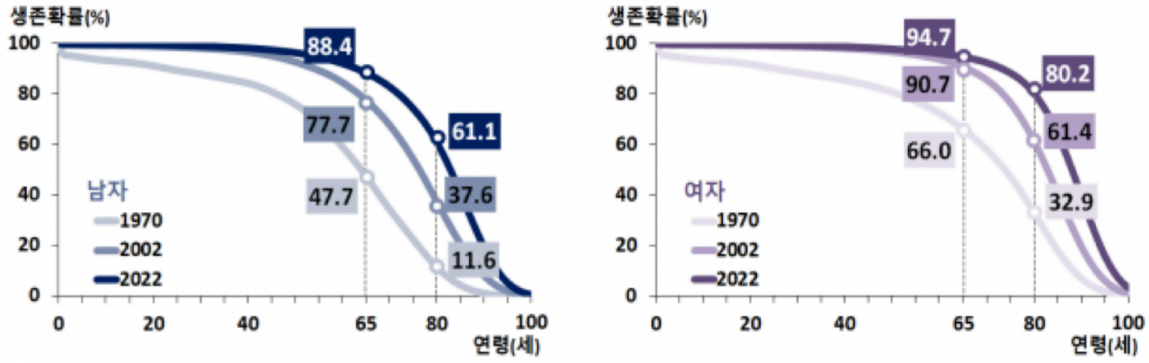
10.2. 생명표 방법

10.2.1. 생명표 개요

생명표(life table)는 특정 집단의 사망률, 생존율, 평균 수명 등을 시간에 따라 체계적으로 정리한 표로, 인구의 생존 패턴을 분석하거나 예측하기 위해 사용된다. 보통 생명표는 인구 집단이 특정 연령대에서 얼마나 오래 생존할 수 있는지를 평가하는 데 쓰이며, 각 연령대까지의 생존 확률, 즉 생존함수를 보여준다.

아래 그림은 통계청에서 작성한 한국인의 생명표로 2007, 2020, 2022년의 한국인 남자와 여자의 생명표이다.

[그림 3] 0세에서 특정 연령까지의 생존확률, 1970년, 2002년, 2022년



2007, 2020, 2022년의 한국인 남자와 여자의 생명표(출처: 통계청, 2022년 생명표, 2023년 12월 1일 보도자료)

위의 한국인의 생명표를 보면 다음과 같은 사실을 알 수 있다.

- 모든 연령층에서 여자의 생존확률이 남자의 생존확률보다 높음.
- 2022년 출생아가 65세까지 생존할 확률은 남자 88.4%, 여자 94.7%임.
 - 2002년 출생아가 65세까지 생존할 확률은 남자 77.7%, 여자 90.7%임.
 - 2007년 출생아가 65세까지 생존할 확률은 남자 47.7%, 여자 66.0%임.
- 2022년 출생아가 80세까지 생존할 확률은 남자 61.1%, 여자 80.2%임.
- 2022년 출생아가 100세까지 생존할 확률은 남자 0.7%, 여자 3.1%임.
- 2007년, 2020년, 2022년의 생명표를 비교하면 한국인의 평균수명이 증가하고 있음을 알 수 있다.

10.2.2. 생명표 방법을 이용한 생존함수 추정

시간 $(0, \infty)$ 구간을 다음과 같은 경계선을 이용하여 k 개의 구간으로 나누었다고 가정하자. ($t_0 = 0$)

$$(0, t_1] \quad (t_1, t_2] \quad \dots \quad (t_{k-1}, t_k]$$

생존함수는 다음과 같은 조건부 확률의 축차식으로 구할 수 있다.

$$\begin{aligned}
S(t_i) &= P(T > t_i) \\
&= P(T > t_i | T > t_{i-1}) P(T > t_{i-1}) \\
&= P(T > t_i | T > t_{i-1}) P(T > t_{i-1} | T > t_{i-2}) P(T > t_{i-2}) \\
&= \dots \\
&= P(T > t_i | T > t_{i-1}) P(T > t_{i-1} | T > t_{i-2}) \times \\
&\quad \dots \times P(T > t_2 | T > t_1) P(T > t_1 | T > t_0) P(T > t_0) \\
&= P(T > t_i | T > t_{i-1}) P(T > t_{i-1} | T > t_{i-2}) \times \\
&\quad \dots \times P(T > t_2 | T > t_1) P(T > t_1) \\
&= \prod_{k=1}^i P(T > t_k | T > t_{k-1})
\end{aligned}$$

참고로 위의 식에서 $P(T > t_0) = P(T > 0) = 1$ 이다.

이제 예를 들어 다음과 같이 5개 각 구간에서 사망자의 수 d_i 와 중도절단자의 수 c_i 가 주어졌다고 가정하자. 이 자료는 전체 115명의 환자에 대하여 5개 구간에서 사망자와 중도절단자의 수를 보여준다.

구간 I_i	i	위험그룹 인원수 n_i	사망자 수 d_i	중도절단 수 c_i
$(0, t_1]$	1	115	47	19
$(t_1, t_2]$	2	49	5	17
$(t_2, t_3]$	3	27	2	15
$(t_3, t_4]$	4	10	2	2
$(t_4, t_5]$	5	6	0	6

위의 표에서 i 번째 구간의 위험그룹 인원수 n_i 는 다음과 같이 주어진다.

$$n_i = n_{i-1} - d_{i-1} - c_{i-1}$$

위험그룹 인원수(**number at risk**) n_i 은 i 번째 구간의 시작 시점까지 살아있는 사람의 수이다. 이는 바로 전 구간의 위험 그룹 인원수 n_{i-1} 에서 사망한 사람의 수 d_{i-1} 와 중도절단된 사람의 수 c_{i-1} 를 뺀 값이다. 분석에 처음으로 고려되는 전체 집단의 수를 n 이라고 하면 $n_0 = n$ 이다.

조건부 생존 확률 $P(T > t_i | T > t_{i-1})$ 는 다음과 같이 추정할 수 있다. 이는 i 번째 구간을 시작할 때 살아있는 사람들(위험 그룹)을 고려했을 때, i 번째 구간에서 사망한 사람의 비율 d_i/n'_i 을 1에서 뺀 확률이다.

$$\begin{aligned}
P(T > t_i | T > t_{i-1}) &= 1 - P(T \leq t_i | T > t_{i-1}) \\
&= 1 - \frac{d_i}{n'_i}
\end{aligned} \tag{10.1}$$

위의 10.1 의 계산에서 위험그룹 인원수(**number at risk**) n_i 대신 보정된 위험그룹 인원수(**corrected number at risk**) 또는 유효인원수(**effective sample size**) n'_i 를 사용하였으며 다음과 같이 계산한다.

10. 생존함수의 추정

$$n'_i = n_i - \frac{c_i}{2}$$

보정된 위험그룹 인원수(유효인원수) n'_i 는 중도절단된 사람이 실제로 사망한 시점을 모르기 때문에 위험그룹 인원수 를 보정한 값(corrected value)이다. 중도절단된 사람이 사망한 시점을 모르기 때문에 중도절단된 사람의 수를 반으로 줄여서 보정된 위험그룹 인원수를 계산한다.

조건부 생존 함수 10.1 를 계산하는 예를 들어보자. 일단 첫 번째 $i = 1$ 인 경우를 살펴보자.

$$\begin{aligned} P(T > t_1 | T > t_0) &= 1 - P(T \leq t_1 | T > 0) \\ &= 1 - P(T \leq t_1) \\ &= 1 - \frac{d_1}{n'_1} \\ &= 1 - \frac{47}{115 - 19/2} \\ &= 1 - \frac{47}{105.5} \\ &= 1 - 0.4455 \\ &= 0.5545 \end{aligned}$$

다음으로 세 번째 구간 ($i = 3$) 인 경우를 살펴보자.

$$\begin{aligned} P(T > t_3 | T > t_2) &= 1 - P(T \leq t_3 | T > t_2) \\ &= 1 - \frac{d_3}{n'_3} \\ &= 1 - \frac{2}{27 - 17/2} \\ &= 1 - \frac{2}{19.5} \\ &= 1 - 0.1026 \\ &= 0.8974 \end{aligned}$$

생존함수의 추정은 조건부 확률의 축차식 10.1 과 조건부 생존 확률 10.1 을 이용하여 다음과 같이 계산할 수 있다.

$$\hat{S}(t_i) = \prod_{k=1}^i \left(1 - \frac{d_k}{n'_k}\right)$$

예를 들어 위의 예제에서 생존함수 $S(t_3) = P(T > t_3)$ 의 추정은 다음과 같이 계산할 수 있다. 참고로 $S(t_0) = 1$ 이다.

$$\begin{aligned}
\hat{S}(t_3) &= P(T > t_3 | T > t_2) P(T > 2 | T > t_1) P(T > t_1 | T > t_0) \\
&= \prod_{k=1}^3 P(T > t_k | T > t_{k-1}) \\
&= \prod_{k=1}^3 \left(1 - \frac{d_k}{n_k}\right) \\
&= \left(1 - \frac{d_1}{n_1}\right) \times \left(1 - \frac{d_2}{n_2}\right) \times \left(1 - \frac{d_3}{n_3}\right) \\
&= 0.5545 \times 0.8765 \times 0.8974 \\
&= 0.4860
\end{aligned}$$

생명표방법을 이용한 생존함수 추정은 표를 이용하면 편리하다. 다음은 위에서 고려한 예제 표에 대하여 각 구간에서 사망자수와 중도절단수가 주어진 경우 생명표방법을 이용하여 생존함수를 추정하는 예를 보여준다.

표 10.2.: 생명표 계산표

구간	i	n_i	d_i	c_i	n'_i	d_i/n'_i	$1 - d_i/n'_i$	$\hat{S}(t_{i-1})$
$(0, t_1]$	1	115	47	19	105.5	0.4455	0.5545	1.000
$(t_1, t_2]$	2	49	5	17	40.5	0.1235	0.8765	0.5545
$(t_2, t_3]$	3	27	2	15	19.5	0.1026	0.8974	0.4860
$(t_3, t_4]$	4	10	2	2	9.0	0.2222	0.7778	0.4362
$(t_4, t_5]$	5	6	0	6	3.0	0.0000	1.00	0.3393

Caution

위의 생명표 계산표 10.2 에서 각 구간에 대한 생존함수 추정은 구간의 시작점에 대한 생존함수 추정값이다. 즉 i 번째 구간 $(t_{i-1}, t_i]$ 에 대한 생존함수 추정값은 t_{i-1} 에 대한 생존함수 $S(t_{i-1}) = P(T > t_{i-1})$ 의 추정값이다.

생명표 방법에서 각 구간의 중간 시점에 대한 위험함수(hazard function)을 다음과 같이 추정할 수 있다.

$$\hat{h}(t_{im}) = \frac{d_i}{b_i(n'_i - d_i/2)}, \text{ where } t_{im} = \frac{t_{i-1} + t_i}{2} \quad (10.2)$$

식 10.2 에서 $b_i = t_i - t_{i-1}$ 는 i 번째 구간의 길이이다.

예를 들어, 구간의 길이가 $b_2 = 1$ 이라면, 두 번째 구간의 중간 시점에서 위험함수 $h(t_{2m})$ 의 추정은 다음과 같이 계산할 수 있다.

$$\hat{h}(t_{2m}) = \frac{d_2}{b_2(n'_2 - d_2/2)} = \frac{5}{(1)(40.5 - 5/2)} = 0.1316$$

참고로 추정된 생존함수의 분산은 다음과 같은 그린우드의 공식(Greenwood's formula)으로 구할 수 있다.

$$\widehat{Var}(\hat{S}(t_i)) = [\hat{S}(t_i)]^2 \sum_{k=1}^i \frac{d_k}{n'_k(n'_k - d_k)}$$

R에서는 패키지 `KMsurv`의 함수 `lifeTable()`를 이용하면 생명표방법을 이용한 생존함수 추정을 할 수 있다. 함수 `lifeTable()`의 각 인자들은 다음과 같이 정의된다.

- `tis`: 구간의 끝점 벡터로, 길이가 `nlost` 및 `nevent` 보다 개수가 1개 많은 벡터
- `ninit`: 처음에 연구에 참여한 전체 집단의 수
- `nlost`: 각 구간에 대한 중도절단된 사람의 수를 나타내는 벡터
- `nevent`: 각 구간에 대한 사건의 수(number of events)를 나타내는 벡터

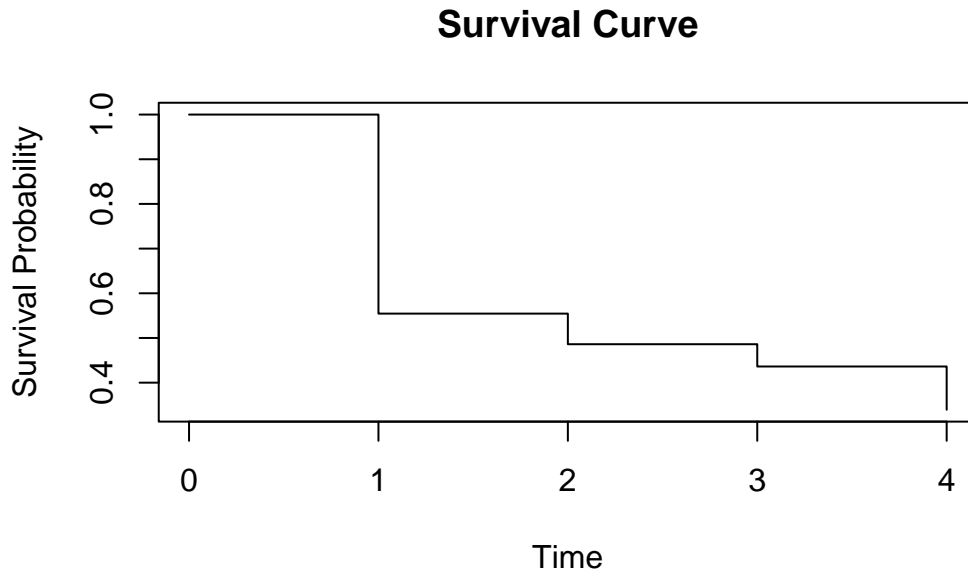
위의 표에 나타난 자료를 이용하여 생명표방법을 이용한 생존함수 추정값을 구해보자.

```
death1 <- c(47, 5, 2, 2, 0)
censor1 <- c(19, 17, 15, 2, 6)
intEndpts <- 0:(length(death1))
N <- sum(censor1) + sum(death1)
lfres0 <- lifetab(tis = intEndpts, ninit=N, nlost=censor1, nevent=death1)
lfres0[c("nsubs", "nevent", "nlost", "nrisk", "hazard", "surv", "se.surv")]
```

	nsubs	nevent	nlost	nrisk	hazard	surv	se.surv
0-1	115	47	19	105.5	0.5731707	1.0000000	0.0000000
1-2	49	5	17	40.5	0.1315789	0.5545024	0.04838917
2-3	27	2	15	19.5	0.1081081	0.4860453	0.05119190
3-4	10	2	2	9.0	0.2500000	0.4361945	0.05679548
4-5	6	0	6	3.0	NA	0.3392624	0.07486856

함수 `lifetab()`의 결과를 이용하여 패키지 `survminer`의 함수 `ggsurvplot()`을 이용하여 생존함수를 그래프로 그릴 수 있다.

```
intTime <- 0:(dim(lfres0)[1]-1) # 구간의 시작점
plot(intTime, lfres0$surv, type = "s", xlab = "Time", ylab = "Survival Probability", main = "S
```

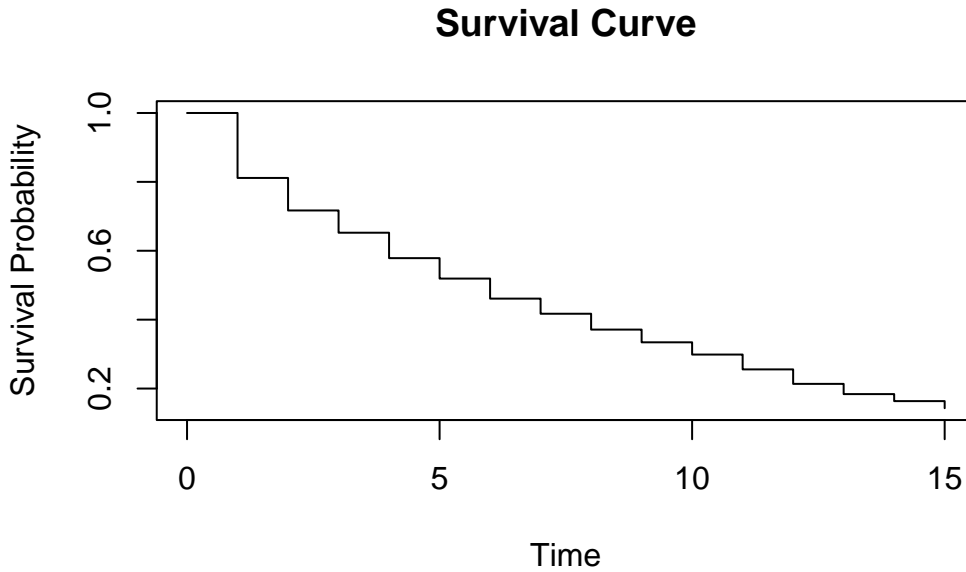


Jaewon Lee (2005) 의 예제 7-1 에 나타난 자료를 이용하여 생명표방법을 이용한 생존함수 추정값을 구해보자.

```
death2 <- c(456, 226, 152, 171, 135, 125, 83, 74, 51, 42, 43, 34, 18, 9, 6, 0)
censor2 <- c(0, 39, 22, 23, 24, 107, 133, 102, 68, 64, 45, 53, 33, 27, 23, 30)
intEndpts = 0:(length(death2))
N <- sum(censor2) + sum(death2)
lfres71 <- lifetab(tis = intEndpts, ninit=N, nlost=censor2, nevent=death2)
lfres71[c("nsubs", "nevent", "nlost", "nrisk", "hazard", "surv", "se.surv")]
```

	nsubs	nevent	nlost	nrisk	hazard	surv	se.surv
0-1	2418	456	0	2418.0	0.20821918	1.00000000	0.000000000
1-2	1962	226	39	1942.5	0.12353102	0.8114144	0.007955134
2-3	1697	152	22	1686.0	0.09440994	0.7170105	0.009179397
3-4	1523	171	23	1511.5	0.11991585	0.6523689	0.009734736
4-5	1329	135	24	1317.0	0.10804322	0.5785647	0.010138361
5-6	1170	125	107	1116.5	0.11859583	0.5192585	0.010304216
6-7	938	83	133	871.5	0.10000000	0.4611239	0.010379949
7-8	722	74	102	671.0	0.11671924	0.4172073	0.010450930
8-9	546	51	68	512.0	0.10483042	0.3711964	0.010578887
9-10	427	42	64	395.0	0.11229947	0.3342218	0.010717477
10-11	321	43	45	298.5	0.15523466	0.2986843	0.010890741
11-12	233	34	53	206.5	0.17941953	0.2556577	0.011124244
12-13	146	18	33	129.5	0.14937759	0.2135639	0.011396799
13-14	95	9	27	81.5	0.11688312	0.1838794	0.011765989
14-15	59	6	23	47.5	0.13483146	0.1635737	0.012259921
15-16	30	0	30	15.0	NA	0.1429117	0.013300258

```
intTime <- 0:(dim(lfres71)[1]-1) # 구간의 시작점
plot(intTime, lfres71$surv, type = "s", xlab = "Time", ylab = "Survival Probability", main = "Survival Curve")
```



10.3. 비모수적 방법

의학연구에서는 사람이나 동물을 대상으로 연구를 진행하는데 이러한 경우 실험 대상자의 수도 많지 않으며 생존함수가 어떤 분포를 따르는지 미리 알기 힘들다. 앞에서 살펴본 생명표방법은 각 구간에서 사망자수와 중도절단수가 주어진 경우 생존함수를 추정하는 방법이므로 실험대상자의 수가 많지 않은 경우에는 적용하기 어렵다.

따라서 생존함수를 추정할 때 비모수적 방법(nonparametric methods)를 이용한다. 비모수적 방법에서는 생명표방법에서 고려한 생존함수를 축차적인 조건부 확률의 곱으로 나타내는 공식 10.1 을 이용하는 것은 동일하지만 구간을 나누는 방법이 다르다.

10.3.1. 누적한계추정량

비모수적 방법으로 생존함수를 추정하는 경우 시간에 따른 생존함수의 변화가 각 관측점에서 일어나게 된다. 즉, 구간을 나누는 값이 자료의 값이 되며, 이러한 방법을 비모수적 방법이라고 한다.

표본에서 관측된 생존시간들을 순서대로 $t_1 < t_2 < \dots < t_n$ 나열한 다음 누적한계추정법(product limit estimator; Kaplan-Meier estimator)은 생존함수를 다음의 식으로 추정한다.

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i} = \prod_{t_i \leq t} \left(\frac{n_i - d_k}{n_i}\right)^{\delta_i} \quad (10.3)$$

- n_i : 시간 t_i 바로 직전까지 중도절단되지 않고 생존한 사람의 수 (위험집단의 수; number at risk)

- d_i : 시간 t_i 에 사망한 사람의 수
- δ_i : 중도절단을 나타내는 지시변수로, $\delta_i = 1$ 이면 실제 사건이 일어난 것이고, $\delta_i = 0$ 이면 중도절단을 나타낸다.

Caution

누적한계추정법에서 구간을 구성하는 시점은 실제 사건이 일어난 시점만 고려한다. 중도절단된 자료만 발생한 시점은 구간을 구성하는 시점으로 고려하지 않는다.

다음 주어진 12명에 대한 생존자료에 대하여 누적한계추정법으로 생존함수를 추정하는 방법을 알아보자. +는 중도절단된 자료를 표시한다.

9, 13, 13+, 18, 23, 28+, 31, 31, 31+, 45+, 48, 161+

위의 자료에서 실제로 사건이 일어난 시점만 고려하여 구간을 만들면 다음과 같다.

$(0, 9], (9, 13], (13, 18], (18, 23], (23, 31], (31, 48], (48, \infty)$

이제 위에서 구한 구간에 대하여 생존함수를 추정해보자.

표 10.3.: 누적한계추정법

구간	t	n_i	d_i	d_i/n_i	$1 - d_i/n_i$	$\hat{S}(t)$
$(0, 9]$	0	12	0	$0/12$	$1 - 0/12$	1.000
$(9, 13]$	9	12	1	$1/12$	$1 - 1/12$	0.917
$(13, 18]$	13	11	1	$1/11$	$1 - 1/11$	0.833
$(18, 23]$	18	9	1	$1/9$	$1 - 1/9$	0.741
$(23, 31]$	23	8	1	$1/8$	$1 - 1/8$	0.648
$(31, 48]$	31	6	2	$2/6$	$1 - 2/6$	0.432
$(48, \infty]$	48	2	1	$1/2$	$1 - 1/2$	0.216

누적한계추정법은 생존함수를 다음과 같이 축차적으로 구할 수 있다.

$$\begin{aligned}
\hat{S}(0) &= 1 \\
\hat{S}(9) &= \hat{S}(0) \times \left[1 - \frac{1}{12}\right] = 0.917 \\
\hat{S}(13) &= \hat{S}(9) \times \left[1 - \frac{1}{11}\right] = 0.833 \\
\hat{S}(18) &= \hat{S}(13) \times \left[1 - \frac{1}{9}\right] = 0.741 \\
\hat{S}(23) &= \hat{S}(18) \times \left[1 - \frac{1}{8}\right] = 0.648 \\
\hat{S}(31) &= \hat{S}(23) \times \left[1 - \frac{2}{6}\right] = 0.432 \\
\hat{S}(48) &= \hat{S}(31) \times \left[1 - \frac{1}{2}\right] = 0.216
\end{aligned}$$

생존함수 추정량 $\hat{S}(t)$ 의 분산은 다음과 같이 주어진다.

$$Var(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

위의 분산추정량으로 생존함수의 신뢰구간을 구하면 $[0,1]$ 의 범위를 벗어나는 경우가 생긴다. 이러한 단점을 보완하기 위하여 생존함수의 변환(log-log transformation)을 이용한 신뢰구간을 구하는 방법을 주로 사용한다.

$$Var(\log[-\log \hat{S}(t)]) \approx \frac{1}{[\hat{S}(t)]^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

다음은 예제 자료에 대하여 누적한계추정법으로 생존함수를 추정하는 R 프로그램이다.

```
t <- c(9, 13, 13, 18, 23, 28, 31, 31, 31, 45, 48, 161)
cens <- c(1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0)
df <- Surv(t, cens)
df
```

```
[1] 9 13 13+ 18 23 28+ 31 31 31+ 45+ 48 161+
```

```
res.km <- survfit(Surv(t, cens)~1, conf.type="log-log")
res.km
```

```
Call: survfit(formula = Surv(t, cens) ~ 1, conf.type = "log-log")
```

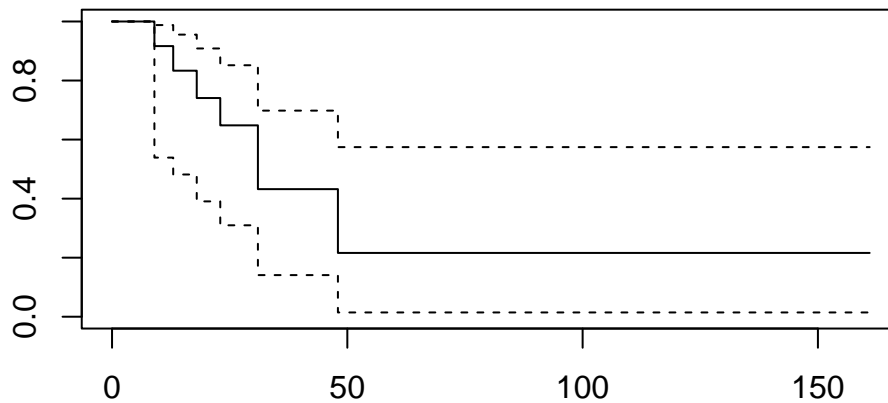
```
      n events median 0.95LCL 0.95UCL
[1,] 12      7     31     13     NA
```

```
summary(res.km)
```

```
Call: survfit(formula = Surv(t, cens) ~ 1, conf.type = "log-log")
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
9	12	1	0.917	0.0798	0.5390	0.988
13	11	1	0.833	0.1076	0.4817	0.956
18	9	1	0.741	0.1295	0.3907	0.909
23	8	1	0.648	0.1426	0.3097	0.852
31	6	2	0.432	0.1568	0.1410	0.698
48	2	1	0.216	0.1717	0.0145	0.574

```
plot(res.km)
```



10.3.2. Nelson-Aalen 추정량

Nelson-Aalen 추정량은 생존함수와 위험함수의 관계를 이용하는 추정량이다. 누적 위험함수 $H(t)$ 는 시간 t 까지 위험함수의 합이므로 다음과 같이 추정할 수 있다.

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

따라서 생존함수과의 관계를 이용하면 Nelson-Aalen 추정량은 다음과 같이 주어진다.

$$\hat{S}(t) = e^{-\hat{H}(t)} \quad (10.4)$$

10. 생존함수의 추정

위에서 고려한 동일한 12명에 대한 생존자료에 대하여 Nelson-Aalen 추정법으로 생존함수를 추정하는 방법을 알아보자. +는 중도절단된 자료를 표시한다.

9, 13, 13+, 18, 23, 28+, 31, 31, 31+, 45+, 48, 161+

이제 누적한계추정법과 동일한 방법으로 구간을 설정하고 10.4 에 주어진 추정식으로 다음과 같은 표를 만들어 생존함수를 추정할 수 있다.

표 10.4.: Nelson-Aalen 추정법

구간	t	n_i	d_i	d_i/n_i	$\sum_{t_i \leq t} d_i/n_i$	$\hat{S}(t)$
(0, 9]	0	12	0	0/12	0.0000	1.00
(9, 13]	9	12	1	1/12	0.0833	0.920
(13, 18]	13	11	1	1/11	0.1742	0.840
(18, 23]	18	9	1	1/9	0.2853	0.752
(23, 31]	23	8	1	1/8	0.4104	0.663
(31, 48]	31	6	2	2/6	0.7437	0.475
(48, ∞]	48	2	1	1/2	1.2439	0.288

아래는 예제 자료에 대하여 Nelson-Aalen 추정량(type="fh")으로 생존함수를 추정하는 R 프로그램이다.

```
t <- c(9, 13, 13, 18, 23, 28, 31, 31, 31, 45, 48, 161)
cens <- c(1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0)
res.fh <- survfit(Surv(t, cens)~1, conf.type="log-log", type="fh")
res.fh
```

Call: survfit(formula = Surv(t, cens) ~ 1, conf.type = "log-log", type = "fh")

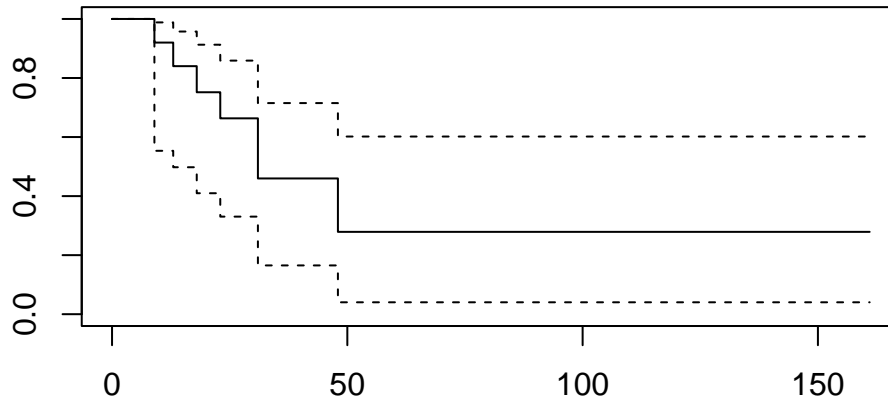
```
      n events median 0.95LCL 0.95UCL
[1,] 12      7     31     13     NA
```

```
summary(res.fh)
```

Call: survfit(formula = Surv(t, cens) ~ 1, conf.type = "log-log", type = "fh")

```
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  9     12      1    0.920  0.0767    0.5534    0.988
 13     11      1    0.840  0.1036    0.4978    0.957
 18      9      1    0.752  0.1248    0.4097    0.913
 23      8      1    0.663  0.1379    0.3305    0.859
 31      6      2    0.460  0.1532    0.1653    0.715
 48      2      1    0.279  0.1675    0.0403    0.602
```

```
plot(res.fh)
```



누적한계추정법과 Nelson-Aalen 추정법으로 구한 생존함수의 추정치를 비교해보면 매우 유사하다.

```
compsurv <- data.frame(time = res.km$time, estKM = res.km$surv, estNA = res.fh$surv)
compsurv
```

	time	estKM	estNA
1	9	0.9166667	0.9200444
2	13	0.8333333	0.8400932
3	18	0.7407407	0.7517484
4	23	0.6481481	0.6634157
5	28	0.6481481	0.6634157
6	31	0.4320988	0.4597740
7	45	0.4320988	0.4597740
8	48	0.2160494	0.2788670
9	161	0.2160494	0.2788670

```
plot(res.fh, col = "blue")
lines(res.km, col = "red")
```

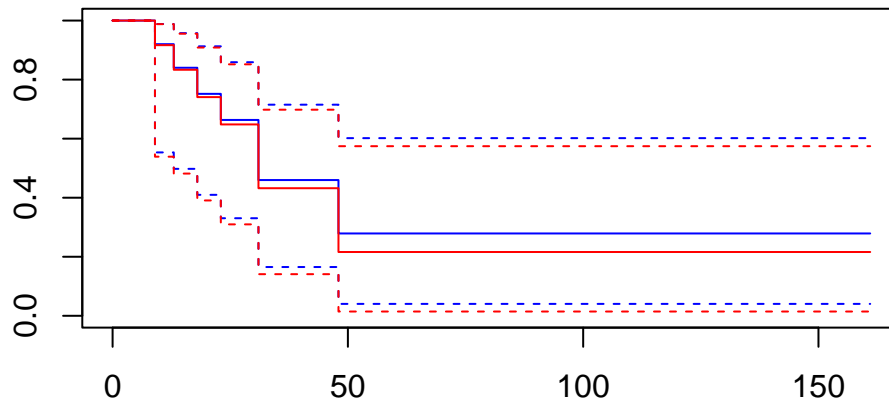


그림 10.1.: 누적한계추정법(파랑선)과 Nelson-Aalen 추정법(빨간선)의 비교

i 중위 생존시간

중위 생존시간의 추정량은 \hat{t}_{med} 다음 방정식을 만족하는 t 의 값이다.

$$\hat{t}_{med} = \inf \{t \mid \hat{S}(t) \leq 1/2\}$$

```
res.km
```

```
Call: survfit(formula = Surv(t, cens) ~ 1, conf.type = "log-log")
```

	n	events	median	0.95LCL	0.95UCL
[1,]	12	7	31	13	NA



10.4. 생존함수 추정의 예제

10.4.1. 예제 1

참고도서 Jaewon Lee (2005) 의 예제 7-2 에 나온 자료에 대하여 누적한계추정법으로 구한 생존함수의 추정결과는 다음과 같다.

10. 생존함수의 추정

```
t <- c(3, 4, 4.5, 5.5, 6.0, 6.4, 6.5, 7.0, 7.5, 8.4, 10, 10, 12, 15)
cens <- c(1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1)
Surv(t, cens)
```

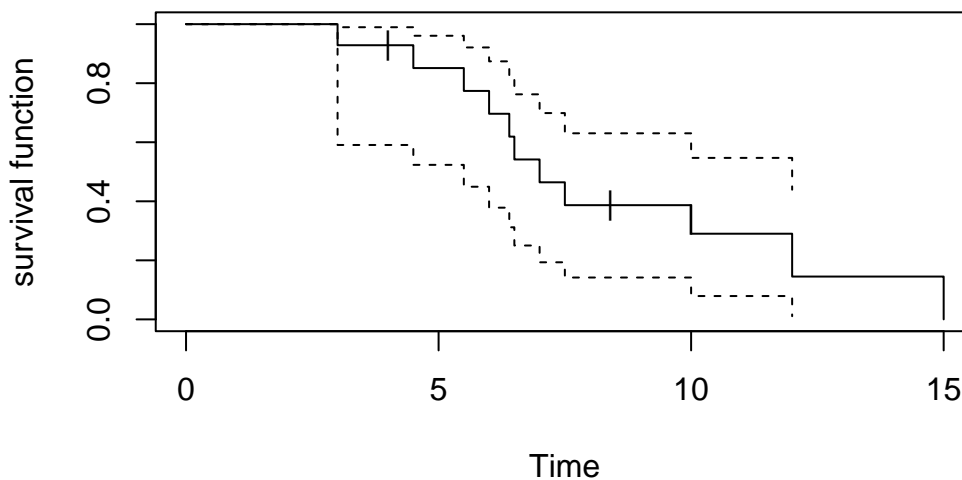
```
[1] 3.0 4.0+ 4.5 5.5 6.0 6.4 6.5 7.0 7.5 8.4+ 10.0 10.0+
[13] 12.0 15.0
```

```
res.km72 <- survfit(Surv(t, cens)~1, conf.type="log-log")
summary(res.km72)
```

Call: survfit(formula = Surv(t, cens) ~ 1, conf.type = "log-log")

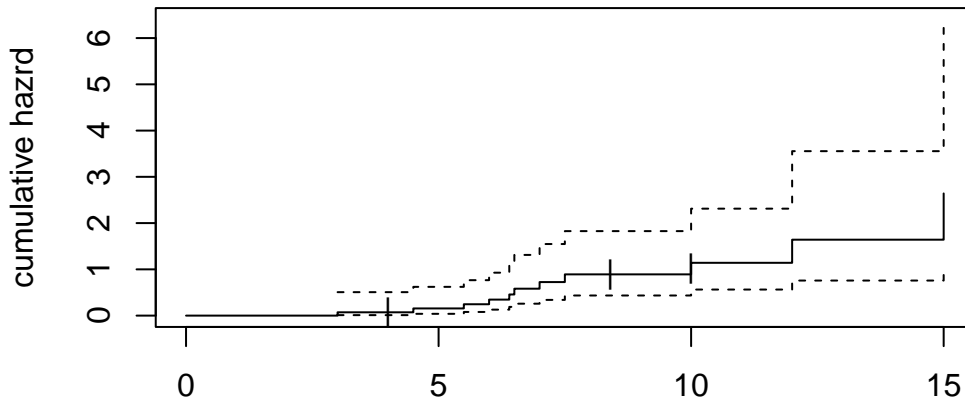
time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3.0	14	1	0.929	0.0688	0.5908	0.990
4.5	12	1	0.851	0.0973	0.5234	0.961
5.5	11	1	0.774	0.1152	0.4493	0.921
6.0	10	1	0.696	0.1270	0.3784	0.874
6.4	9	1	0.619	0.1344	0.3119	0.821
6.5	8	1	0.542	0.1381	0.2502	0.762
7.0	7	1	0.464	0.1384	0.1934	0.699
7.5	6	1	0.387	0.1352	0.1418	0.630
10.0	4	1	0.290	0.1315	0.0791	0.547
12.0	2	1	0.145	0.1219	0.0108	0.439
15.0	1	1	0.000	NaN	NA	NA

```
plot(res.km72, mark="|", ylab="survival function", xlab="Time")
```



다음은 누적한계추정법으로 구한 누적위험함수 $H(t)$ 이다.

```
plot(res.km72, cumhaz=TRUE, mark="|", ylab="cumulative hazrd")
```

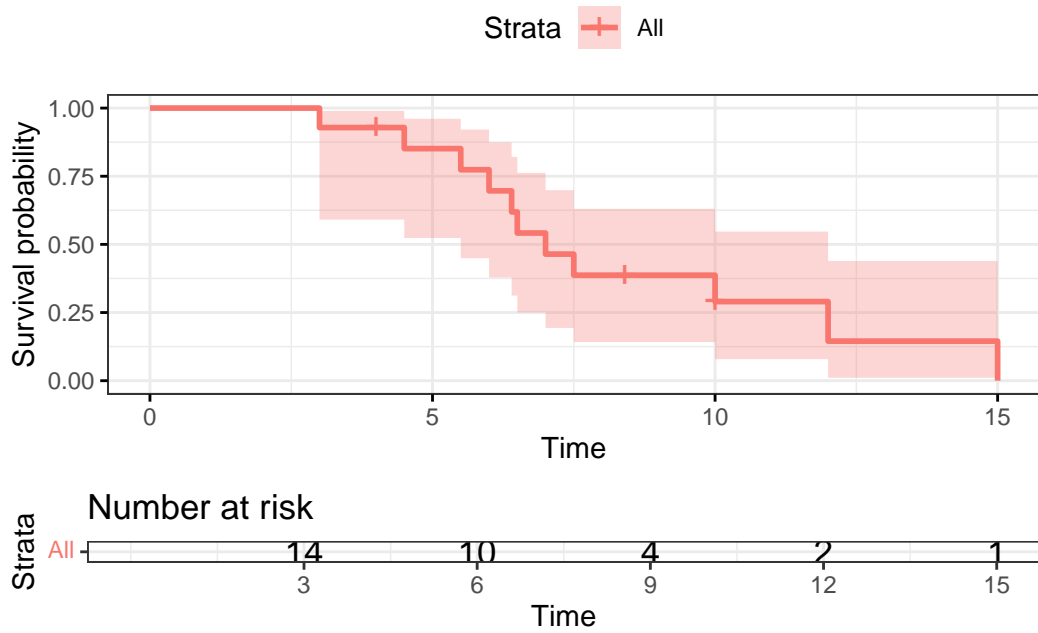


다음은 `survminer` 패키지의 `ggsurvplot()` 함수를 이용하여 생존함수와 누적위험함수를 그래프로 그린 결과이다.

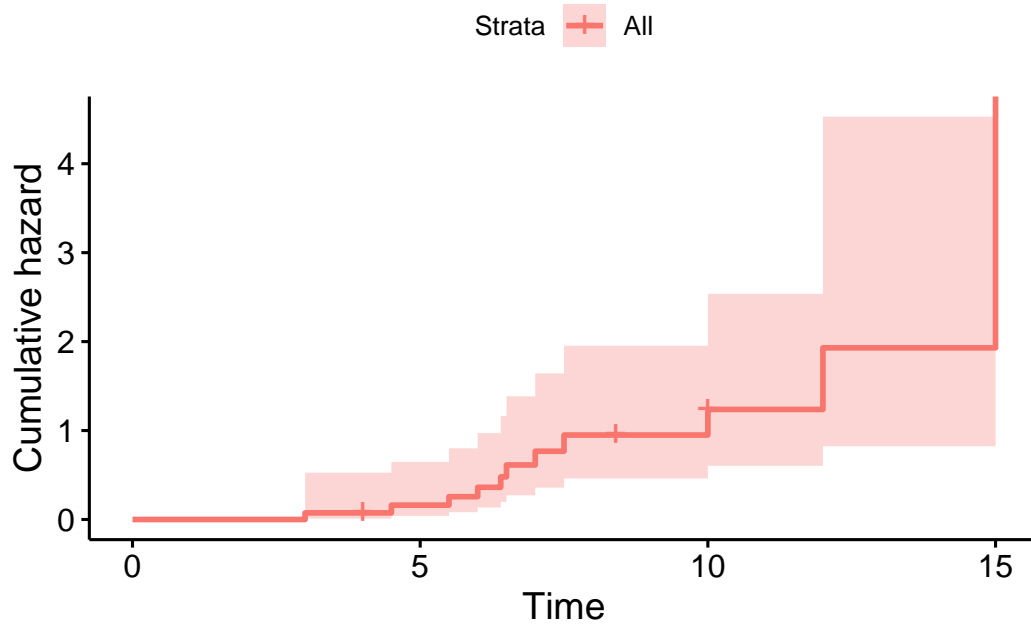
```
t <- c(3, 4, 4.5, 5.5, 6.0, 6.4, 6.5, 7.0, 7.5, 8.4, 10, 10, 12, 15)
cens <- c(1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1)
example1 <- data.frame(t = t, cens = cens)

res.km721 <- survfit(Surv(t, cens)~1, data=example1, conf.type="log-log")
ggsurvplot(res.km721, data=example1, risk.table = TRUE, risk.table.height = 0.25, ggtheme = t
```

10. 생존함수의 추정



```
ggsurvplot(res.km721, data=example1, fun = "cumhaz")
```



10.4.2. 예제 2

참고도서 Moore (2016) 의 29페이지에 나오는 예제를 소개한다. 위암환자들에게 치료제인 Xelox를 투약하는 실제 2상 임상실험의 자료를 분석한 예이다. 관심있는 변수는 “progression-free time” 으로서 임상실험에 참가한 시점부터 회복이 되거나 사망하거나 둘 중에 빠르게 나타난 시간이다.

아래는 누적한계추정법으로 구한 생존함수의 추정치와 그림(생존함수와 누적위험함수)이다.

10. 생존함수의 추정

```
example2 <- gastricXelox %>%
  dplyr::mutate(timeMonths = timeWeeks*7/30.25) # convert weeks to months
Surv(example2$timeMonths, example2$delta)
```

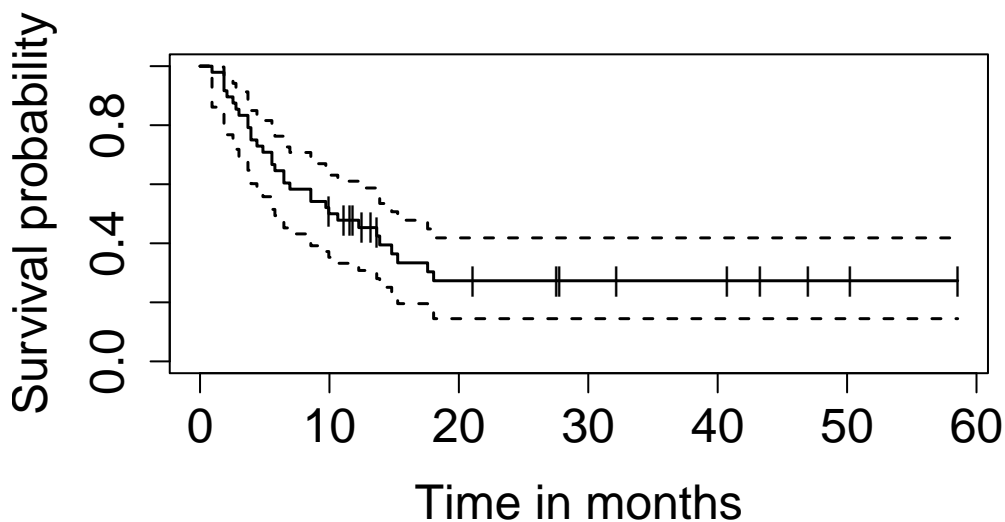
```
[1] 0.9256198 1.8512397 1.8512397 1.8512397 2.0826446 2.5454545
[7] 2.7768595 3.0082645 3.7024793 3.7024793 3.9338843 3.9338843
[13] 4.3966942 4.8595041 5.5537190 5.5537190 5.7851240 6.4793388
[19] 6.4793388 6.9421488 8.5619835 8.5619835 9.7190083 9.9504132
[25] 9.9504132+ 10.6446281 11.1074380+ 11.5702479+ 11.8016529+ 12.2644628
[31] 12.4958678+ 13.1900826+ 13.6528926+ 13.6528926 13.8842975 14.8099174
[37] 15.2727273 17.5867769 18.0495868 21.0578512+ 27.5371901+ 27.7685950+
[43] 32.1652893+ 40.7272727+ 43.2727273+ 46.9752066+ 50.2148760+ 58.5454545+
```

```
result.km <- survfit(Surv(timeMonths, delta) ~ 1, conf.type="log-log", data=example2)
# median survival and 95% confidence interval is printed as follows:
result.km
```

```
Call: survfit(formula = Surv(timeMonths, delta) ~ 1, data = example2,
  conf.type = "log-log")
```

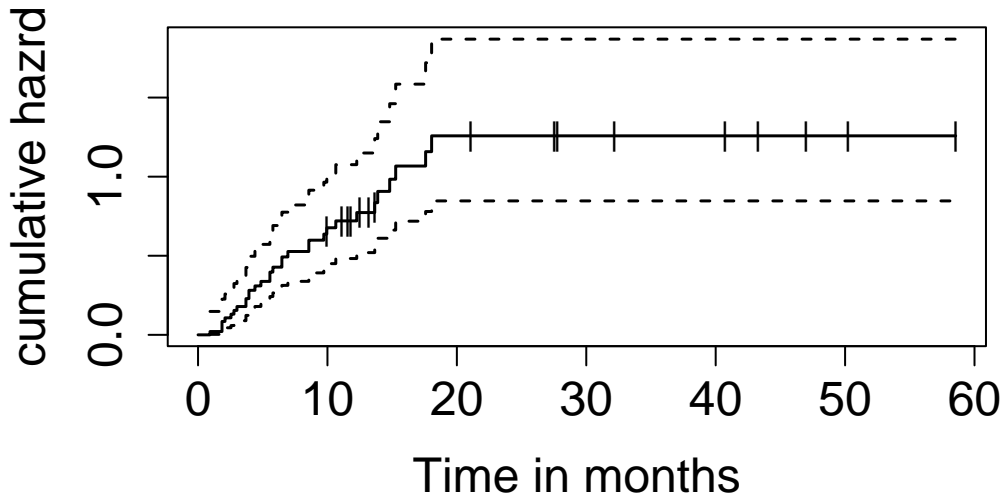
```
      n events median 0.95LCL 0.95UCL
[1,] 48      32  10.3    5.79   15.3
```

```
plot(result.km, mark="|", ylab="Survival probability", xlab="Time in months",
  cex.axis=1.5, cex.lab=1.5, lwd=1.5)
```



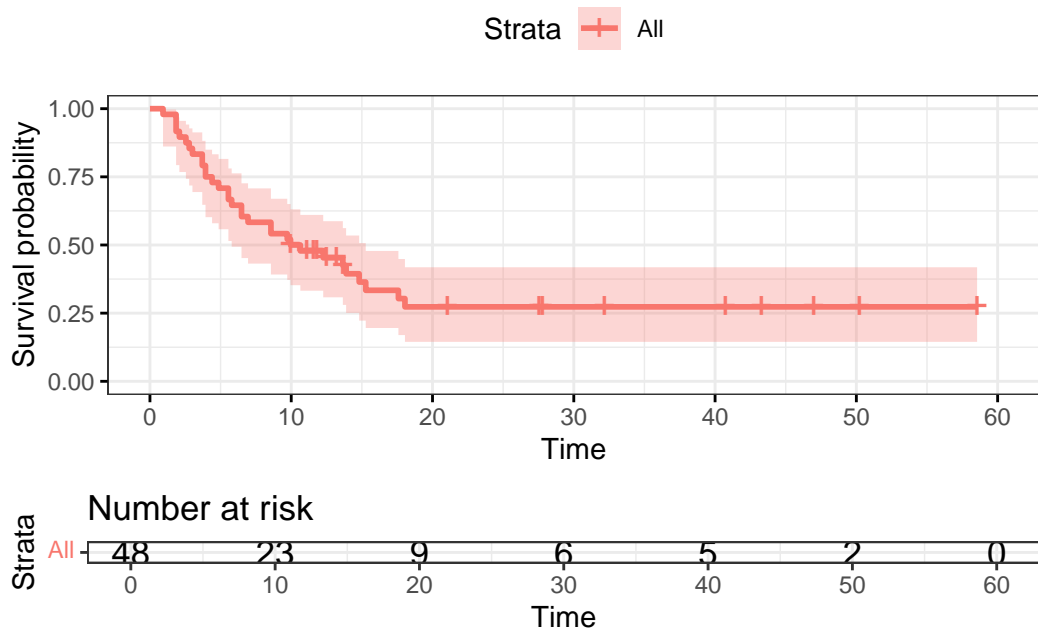
10. 생존함수의 추정

```
plot(result.km, cumhaz=TRUE, mark="|", ylab="cumulative hazrd", xlab="Time in months", cex.ax
```



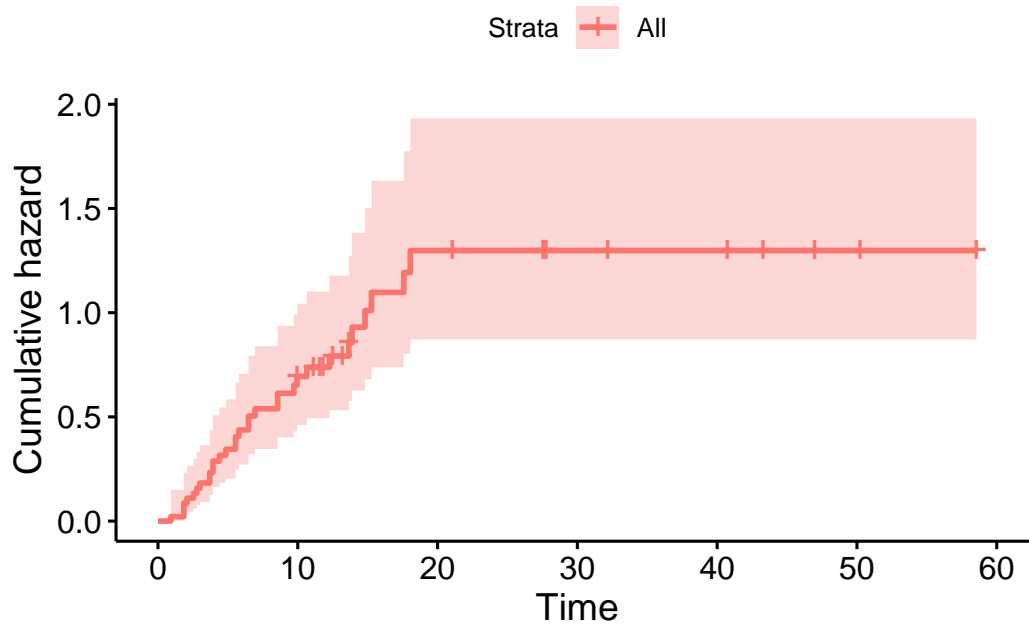
다음은 `survminer` 패키지의 `ggsurvplot()` 함수를 이용하여 생존함수와 누적위험함수를 그래프로 그린 결과이다.

```
ggsurvplot(result.km, data = example2, risk.table = TRUE, risk.table.height = 0.25, ggtheme =
```



```
ggsurvplot(result.km, data = example2, fun = "cumhaz")
```


10. 생존함수의 추정



A. 다변량 확률변수의 성질

A.1. 일변량분포

일변량 확률변수 x 가 확률밀도함수 $f(x)$ 를 가지는 분포를 따를 때 기대값과 분산은 다음과 같이 정의된다.

$$E(x) = \int x f(x) dx = \mu$$

$$V(x) = E[x - E(x)]^2 = \int (x - \mu)^2 f(x) dx = \sigma^2$$

새로운 확률변수 y 가 확률변수 x 의 선형변환으로 표시된다면 (a 와 b 는 실수)

$$y = ax + b$$

그 기대값(평균)과 분산은 다음과 같이 계산된다.

$$\begin{aligned} E(y) &= E(ax + b) \\ &= \int (ax + b) f(x) dx \\ &= a \int x f(x) dx + b \\ &= aE(x) + b \\ &= a\mu + b \end{aligned}$$

$$\begin{aligned} V(y) &= Var(ax + b) \\ &= E[ax + b - E(ax + b)]^2 \\ &= E[a(x - \mu)]^2 \\ &= a^2 E(x - \mu)^2 \\ &= a^2 \sigma^2 \end{aligned}$$

A.2. 확률벡터의 분포

다음과 같은 확률벡터 \mathbf{x} 가 p 차원의 다변량분포를 따른다고 하자.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix}$$

다변량 확률벡터 \mathbf{x} 의 평균벡터 $\boldsymbol{\mu}$ 과 공분산 행렬 $\boldsymbol{\Sigma}$ 은 다음과 같이 계산된다.

$$E(\mathbf{x}) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ E(x_3) \\ \vdots \\ E(x_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}$$

$$V(\mathbf{x}) = Cov(\mathbf{x}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ & \cdots & \cdots & \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix} = \boldsymbol{\Sigma}$$

여기서

$$\sigma_{ii} = V(x_i), \quad \sigma_{ij} = Cov(x_i, x_j)$$

두 확률변수의 상관계수 ρ_{ij} 는 다음과 같이 정의된다.

$$\rho_{ij} = \frac{Cov(x_i, x_j)}{\sqrt{V(x_i)V(x_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

새로운 확률벡터 \mathbf{y} 가 확률벡터 \mathbf{x} 의 선형변환이라고 하자.

$$\mathbf{y} = \mathbf{Ax} + \mathbf{b}$$

단 여기서 \mathbf{A} 는 $p \times p$ 실수 행렬이고 \mathbf{b} 는 $p \times 1$ 실수 벡터이다.

이제 확률벡터 \mathbf{y} 의 기대값(평균벡터)과 공분산은 다음과 같이 계산된다.

$$E(\mathbf{y}) = E(\mathbf{Ax} + \mathbf{b}) = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$V(\mathbf{y}) = \text{Var}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\Sigma\mathbf{A}^t$$

A.3. 다변량 정규분포

A.3.1. 일변량 정규분포

일변량 확률변수 x 가 평균이 μ 이고 분산이 σ^2 인 정규분포를 따른다면 다음과 같이 나타내고

$$\mathbf{x} \sim N(\mu, \sigma^2)$$

확률밀도함수 $f(x)$ 는 다음과 같이 주어진다.

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$

A.3.2. 이변량 정규분포

이제 2개의 확률변수로 이루어진 2-차원 확률벡터 $\mathbf{x} = (x_1, x_2)^t$ 가 다음과 같은 평균 벡터 $\boldsymbol{\mu}$ 과 공분산 행렬 Σ 을 가지고 정규분포를 따른다면 이변량 정규분포(bivariate normal distribution)을 따른다고 하며

$$E(\mathbf{x}) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \boldsymbol{\mu}, \quad \text{Cov}(\mathbf{x}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \Sigma$$

다음과 같이 표기한다.

$$\mathbf{x} \sim N_2(\boldsymbol{\mu}, \Sigma)$$

이변량 정규분포의 확률밀도함수 $f(\mathbf{x})$ 에서 다음과 같이 주어진다.

$$f(\mathbf{x}) = (2\pi)^{-1} |\Sigma|^{-1/2} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})^t}{2}\right) \quad (\text{A.1})$$

! 이변량 정규분포에서 조건부 분포

만약 x_2 가 주어졌을 때 x_1 의 조건부 분포는 정규분포이고 평균과 분산은 다음과 같이 주어진다.

$$E(x_1|x_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2) \quad (\text{A.2})$$

$$V(x_1|x_2) = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}$$

따라서 x_2 가 주어졌을 때 x_1 의 조건부 분포는 다음과 같은 정규분포를 따른다.

$$x_1|x_2 \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2), \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}\right)$$

x_1 과 x_2 의 상관계수는 $\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})$ 이다.

A.3.3. 다변량 정규분포

앞에서 살펴본 이변량 정규분포를 p -차원 확률벡터로 쉽게 확장할 수 있다. p -차원 확률벡터 \mathbf{x} 가 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 다변량 정규분포(p-dimensional normal distribution)를 따른다면 다음과 같이 나타내고

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

확률밀도함수 $f(\mathbf{x})$ 는 A.1 에서 평균 벡터 $\boldsymbol{\mu}$ 와 공분산 행렬 $\boldsymbol{\Sigma}$ 의 차원이 p 차원으로 주어진 식으로 나타낼 수 있다.

- 다변량 정규분포에서, 두 변수의 공분산이 0 인 경우와 두 변수의 독립은 동치이다.
- 다변량 정규분포의 주변 분포(marginal distribution)는 여전히 정규분포를 따른다. 즉, 다변량 정규분포에서 일부 변수만 따로 떼어내더라도 그 변수들 역시 다변량 정규분포를 따른다.
- 다변량 정규분포에서 공분산 행렬 $\boldsymbol{\Sigma}$ 는 대칭행렬(symmetric matrix)이며 양정치 행렬(positive definite matrix)이다.

다변량 정규분포를 따르는 확률벡터 \mathbf{x} 를 다음과 같이 두 부분으로 나누면

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$$

확률벡터 \mathbf{x}_2 가 주어진 경우 \mathbf{x}_1 의 조건부 분포는 다변량 정규분포를 따르고 평균과 공분산은 다음과 같다.

$$E(\mathbf{x}_1|\mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

$$V(\mathbf{x}_1|\mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t$$

B. 선형혼합모형의 이론

i 고정효과와 임의효과의 표기

고정효과는 보통 모형식에서 그리스 문자(α, β)로 표시하고 임의효과는 모형식에서 영어 알파벳 문자(a, b)로 표시한다

B.1. 독립군집 모형

선형혼합모형을 적용할 수 있는 자료의 종류와 구조는 다양하지만 가장 흔하게 나타나는 자료 구조는 독립 군집(independent cluster)의 형태로 n 개의 독립적인 군집(개체, 그룹 등)들에서 각각 m_i 개의 반응값이 나타나는 구조이다.

독립적인 군집은 여러 개의 내포되는 층으로 구성될 수도 있고(예를 들어 학교, 교사, 학생) 개체(반복측정자료) 또는 공간적/시간적인 군집으로도 나타날 수 있다. 이렇게 독립적 군집에 대한 자료를 설명하는 혼합모형을 독립군집 모형(independent cluster model) 이라고 부른다.

앞의 Section 8.4.2 에서 논의한 임의계수모형은 회귀계수에 임의효과가 더해진 독립군집 혼합모형의 특수한 경우로 볼 수 있다.

예를 들어 Example 8.1 에서 살펴본 트럭 운전자 자료에서 트럭 운전사는 하나의 군집이 되고 각 운전자마다 반복한 측정치가 있는 자료가 대표적인 예이다. 또한 학교(군집)를 독립적으로 추출하고 각 학교에서 다시 학생들을 추출하는 경우도 독립 군집의 형태라고 할 수 있다.

B.1.1. 하나의 군집에 대한 모형

먼저 자료가 n 개의 독립적인 군집으로 나누어진 경우를 생각해 보자. 먼저 i 번째 군집에서 측정된 m_i 개의 반응값을 모아 놓은 반응벡터를 \mathbf{y}_i 라고 하고

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{im_i} \end{bmatrix}$$

다음과 같은 독립군집 모형을 따른다고 가정한다.

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, 2, \dots, n \quad (\text{B.1})$$

- \mathbf{y}_i : 반응벡터 \mathbf{y}_i 는 $m_i \times 1$ 벡터로 i 번째 군집에 속한 관측값들을 모아 놓은 벡터이다.
- \mathbf{X}_i : 차원이 $m_i \times p$ 인 행렬로 i 번째 군집의 고정 효과에 대한 계획행렬이다.
- $\boldsymbol{\beta}$: $p \times 1$ 의 모수벡터로서 고정 효과를 나타낸다.
- \mathbf{Z}_i : 차원이 $m_i \times q$ 인 행렬 i 번째 군집의 임의 효과에 대한 계획행렬이다.
- \mathbf{b}_i : i 번째 군집에 대한 q 개의 임의효과를 모아 놓은 벡터로 $q \times 1$ 차원 벡터이다.
- \mathbf{e}_i : m_i 개의 오차를 모아놓은 오차 벡터이다.

또한 군집에 대한 임의효과 벡터와 오차항 벡터의 분포는 각각 다음과 같이 주어진다.

$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G}), \quad \mathbf{e}_i \sim N(0, \sigma_e^2 \mathbf{I}), \quad i = 1, 2, \dots, n \quad (\text{B.2})$$

개체에 대한 임의효과 벡터 \mathbf{b}_i 는 서로 독립적으로 평균이 0 공분산이 \mathbf{G} 인 다변량 정규분포를 따른다. 임의효과들의 공분산 행렬 \mathbf{G} 는 양반정치 행렬(semi-positive definite matrix)이다. 여기서 양반정치행렬이라는 의미는 분산의 0이 될 수 있는 양정치 행렬을 말한다.

모든 오차항은 서로 독립이고 분산은 σ_e^2 으로 동일하다.

$$V(e_{ik}) = \sigma_e^2 \quad \text{for all } i, k$$

$$\text{Cov}(e_{ik}, e_{jl}) = 0 \quad \text{for all } i, j, k, l$$

다른 개체에 대한 임의효과 벡터는 서로 독립이며 임의 효과 벡터와 오차항 벡터도 독립이다.

$$\text{Cov}(\mathbf{b}_i, \mathbf{b}_j) = \mathbf{0} \quad \text{when } i \neq j$$

$$\text{Cov}(\mathbf{a}_i, \mathbf{e}_i) = \mathbf{0}$$

Example B.1. 일원배치모형 8.8 은

$$y_{ij} = \mu + a_i + e_{ij}, \quad i = 1, 2, \dots, I, \quad j = 1, 2, \dots, J$$

독립군집 모형의 가장 단순한 구조로서 I 개의 군집으로 이루어져 있으며 각 군집에서 J 개의 관측값이 나타난 형태이다. 주어진 그룹 i 에 대하여 모형을 다음과 같이 행렬식으로 표시한다면

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, 2, \dots, I$$

위의 식에서 나타난 각 항은 다음과 같다.

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iJ} \end{bmatrix}, \mathbf{X}_i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \boldsymbol{\beta} = [\mu], \mathbf{Z}_i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \mathbf{b}_i = [a_i], \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{iJ} \end{bmatrix}$$

Example B.2. Example 8.1 에서 살펴본 트럭 운전자 자료에서 임의계수 모형 8.12 을 다음과 같이 나타낼 수 있다.

$$\begin{aligned} y_{ij} &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j + e_{ij} \\ &= (\beta_0 + \beta_1 t_j) + (b_{0i} + b_{1i} t_j) + e_{ij} \end{aligned}$$

위의 임의계수 모형을 독립군집 모형 B.1 의 형태로 기술하면 계획행렬 \mathbf{X}_i 와 \mathbf{Z}_i 는 다음과 같다.

$$\mathbf{y} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i,10} \end{bmatrix}, \mathbf{X}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \mathbf{Z}_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix}, \mathbf{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix}, \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{i,10} \end{bmatrix}$$

위의 모형에서 고정효과에 대한 계획행렬 \mathbf{X}_i 와 임의효과에 대한 계획행렬 \mathbf{Z}_i 이 같은 것에 유의하자.

B.1.2. 전체 군집에 대한 모형

위의 각 그룹 i 에 대한 혼합효과모형 B.1 을 모두 합쳐서 아래와 같은 행렬식으로 표현할 수 있다.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \quad (\text{B.3})$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}, \boldsymbol{\beta} = \boldsymbol{\beta}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{Z}_I \end{bmatrix}, \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}, \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_n \end{bmatrix}$$

여기서 반응변수 벡터 \mathbf{y} 와 고정효과 $\boldsymbol{\beta}$ 에 대한 계획행렬 \mathbf{X} 는 각 군집의 반응변수 벡터 \mathbf{y}_i 와 \mathbf{X}_i 를 행으로 쌓아놓은 것으로 표현된다. 오차항에 대한 벡터 \mathbf{e} 도 동일한 형식의 벡터이다.

임의효과 벡터 \mathbf{b} 는 각 군집에 대한 임의효과 벡터 \mathbf{b}_i 를 행으로 쌓아놓은 것과 같고 임의효과에 대한 계획행렬 \mathbf{Z} 는 각 군집의 계획행렬 \mathbf{Z}_i 를 대각원소로 구성하는 행렬이다.

다음은 모형 B.3 에 나타난 벡터와 행렬의 차원이다. 이 경우 전체 반응변수의 수를 N 이라고 하자.

$$N = m_1 + m_2 + \cdots + m_n = \sum_{i=1}^n m_i$$

- $\mathbf{y} : N \times 1$
- $\mathbf{X} : N \times p$
- $\boldsymbol{\beta} : p \times 1$
- $\mathbf{Z} : N \times qn$
- $\mathbf{a} : qn \times 1$ 로 나타낸다.
- $\mathbf{e} : N \times 1$

여기서 유의할 점은 군집의개수 n 이 증가해도 고정효과를 나타내는 회귀계수의 벡터 $\boldsymbol{\beta}$ 의 차원은 변하지 않지만 임의효과에 대한 벡터 \mathbf{b} 의 차원은 증가한다는 사실이다.

또한 임의효과에 대한 계획행렬 \mathbf{Z} 는 원소에 0 이 많은 희소행렬(sparse matrix)로 나타난다.

임의효과 벡터 \mathbf{b} 는 n 개의 임의효과 벡터를 가지는 세로로 쌓은 벡터이고 평균은 $\mathbf{0}$ 이다. 또한 서로 다른 군집에 대한 임의효과 벡터는 서로 독립이므로 임의효과 벡터 \mathbf{b} 의 공분산 행렬은 차원이 $nq \times nq$ 인 다음과 같은 행렬로 나타난다.

$$Cov(\mathbf{b}) = \begin{bmatrix} \mathbf{G} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G} \end{bmatrix}$$

오차항의 벡터 \mathbf{e} 는 일반적인 선형모형과 같이 평균이 0 이고 분산이 $\sigma_e^2 \mathbf{I}$ 인 다변량정규분포를 따른다고 가정한다.

$$\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$$

Example B.3. Example B.1 에서 살펴본 일원배치모형 8.8 을 전체 군집에 대한 혼합모형 B.3 의 형태로 나타내면 다음과 같다.

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1J} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2J} \\ \vdots \\ y_{I1} \\ y_{I2} \\ \vdots \\ y_{IJ} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \boldsymbol{\beta} = [\mu], \quad \mathbf{Z} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_I \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1J} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2J} \\ \vdots \\ e_{I1} \\ e_{I2} \\ \vdots \\ e_{IJ} \end{bmatrix}$$

임의효과 벡터 \mathbf{b} 는 I 개의 그룹에 대한 임의효과를 가지는 벡터로서 공분산행렬 \mathbf{G} 는 다음과 같다.

$$\mathbf{G} = Cov(\mathbf{b}) = \begin{bmatrix} \sigma_a^2 & 0 & \dots & 0 \\ 0 & \sigma_a^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_a^2 \end{bmatrix} = \sigma_a^2 \mathbf{I}_I$$

따라서 임의효과 벡터 \mathbf{b} 는 다음과 같은 다변량 정규분포를 따른다.

$$\mathbf{b} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{I}_I)$$

오차항의 벡터는 일반적인 선형모형과 같이 평균이 0 이고 분산이 $\sigma_e^2 \mathbf{I}$ 인 다변량정규분포를 따른다.

$$\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_{IJ})$$

B.2. 반응값의 분포

이제 선형혼합모형 B.3 에서 관측값의 분포에 대하여 살펴보자. 먼저 반응변수 벡터 \mathbf{y} 는 다변량 정규분포를 따르며 기대값 $\boldsymbol{\mu}$ 과 공분산 행렬 \mathbf{V} 는 다음과 같이 주어진다.

$$\begin{aligned}
\boldsymbol{\mu} &= E(\mathbf{y}) \\
&= E(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}) \\
&= E(\mathbf{X}\boldsymbol{\beta}) + \mathbf{Z}E(\mathbf{b}) + E(\mathbf{e}) \\
&= \mathbf{X}\boldsymbol{\beta} + \mathbf{0} + \mathbf{0} \\
&= \mathbf{X}\boldsymbol{\beta} \\
\mathbf{V} &= Cov(\mathbf{y}) \\
&= Cov(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}) \\
&= Cov(\mathbf{Z}\mathbf{b} + \mathbf{e}) \\
&= Cov(\mathbf{Z}\mathbf{b}) + Cov(\mathbf{e}) \\
&= \mathbf{Z}Cov(\mathbf{b})\mathbf{Z}^t + \sigma_e^2\mathbf{I} \\
&= \mathbf{Z}\mathbf{G}\mathbf{Z}^t + \sigma_e^2\mathbf{I}
\end{aligned}$$

Example B.4. 일원배치모형에서 층(그룹)의 수가 3개이고($I = 3$) 각 층에서 관측치가 2개($J = 2$)인 경우 독립군집 모형 B.1 에서 각 항은 다음과 같다.

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix}, \mathbf{X}_i = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{\beta} = [\mu], \mathbf{Z}_i = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{b}_i = [a_i], \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \end{bmatrix}$$

3개의 그룹을 합친 전체 모형 B.3 에서 각 항은 다음과 같다.

또한 각 개체에 대한 임의효과에 대한 계획행렬 \mathbf{Z}_i 와 전체 계획행렬 \mathbf{Z} 는 다음과 같이 나타나고

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

임의효과 \mathbf{b} 의 공분산 행렬 \mathbf{G} 이 다음과 같다.

$$\mathbf{G} = Cov(\mathbf{b}) = \begin{bmatrix} \sigma_a^2 & 0 & 0 \\ 0 & \sigma_a^2 & 0 \\ 0 & 0 & \sigma_a^2 \end{bmatrix}$$

이제 반응 벡터 \mathbf{y} 의 공분산 행렬 $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^t + \sigma^2\mathbf{I}$ 를 구해보자.

먼저 $\mathbf{Z}\mathbf{Z}^t$ 은 다음과 같이 나타난다.

$$\begin{aligned}
\mathbf{ZGZ}^t &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_a^2 & 0 & 0 \\ 0 & \sigma_a^2 & 0 \\ 0 & 0 & \sigma_a^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 & 0 & 0 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 \\ 0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 \end{bmatrix}
\end{aligned}$$

반응값 벡터 \mathbf{y} 의 공분산 행렬 $\mathbf{V} = \mathbf{ZGZ}^t + \sigma_e^2 \mathbf{I}$ 는 다음과 같이 주어진다.

$$\text{Cov}(\mathbf{y}) = \mathbf{V} = \begin{bmatrix} \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & 0 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & 0 & 0 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 \\ 0 & 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 \end{bmatrix}$$

같은 그룹에 속한 관측값의 공분산은 σ_a^2 이고 같은 그룹에 속한 두 개의 관측값들에 대한 상관계수는 다음과 같다.

$$\text{cor}(y_{i1}, y_{i2}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

B.3. 예측

B.3.1. 예제: IQ의 예측

이 예제는 Searle, Casella, and McCulloch (2009)의 7장에 나온 예제입니다.

인간의 지능을 정의하고 측정하려는 노력은 오래된 연구의 주제이다. 인간의 지능을 IQ(Intelligence Quotient)로 수량화하고 이를 측정하기 위한 여러 가지 시험들이 제시되었다.

먼저 문제를 단순하게 하기 위하여 학생들로 구성된 집단을 고려하고 IQ가 정규분포를 따른다고 가정하자. 이제 IQ를 측정하기 위하여 시험을 실시하는데 여러 개의 문항으로 구성된 시험을 보고 평균 성적(score)을 IQ 측정에 사용하기로 한다. 이 경우 각 문항의 점수도 정규분포를 따른다고 가정하자.

만약 한 명의 학생의 평균 성적이 130인 경우 이 학생의 진짜 IQ에 대한 예측값은 얼마인가?

이러한 문제는 관측된 정보를 이용하여 관측되지 않는 임의효과를 예측하는 유형의 문제로 볼 수 있으며 많은 연구자들이 이러한 문제의 해법을 주는데 많은 노력을 기울여 왔다.

B.3.1.1. 혼합모형

위의 IQ 문제를 푸는 여러 가지 방법 중 하나는 i 번째 학생의 j 번째 문항에 대한 점수 y_{ij} 를 다음과 같이 같은 혼합효과 모형을 따른다고 가정한다.

일원배치 모형을 이용하여 학생의 실제 IQ 를 $\mu + a_i$ 로 보고 J 개의 다양한 문항에 대한 관측값 y_{ij} 을 통하여 IQ 를 예측하고자 한다.

$$y_{ij} = \mu + a_i + e_{ij}, \quad i = 1, 2, \dots, I \text{ and } j = 1, 2, \dots, J$$

위의 모형에서 μ 는 전체 모집단에 대한 IQ 의 평균으로 생각하면 되고 임의효과 a_i 는 i 번째 학생의 지적 능력과 관련된 효과로 생각할 수 있다.

따라서 $\mu + a_i$ 를 i 번째 학생의 지적 능력 IQ 로 생각할 수 있다.

여기서 평균 성적 a_i 는 서로 독립이고 $N(0, \sigma_a^2)$ 를 따른다고 가정하고 오차항 e_{ij} 도 서로 독립이며 $N(0, \sigma_e^2)$ 를 따른다고 가정하자.

B.3.1.2. 평균 점수의 분포

이제 IQ 테스트에서 i 번째 학생의 평균 점수 \bar{y}_i 는 다음과 같이 계산되며

$$\bar{y}_i = \frac{\sum_j y_{ij}}{J} = \mu + a_i + \frac{\sum_j e_{ij}}{J}$$

또한 다음과 같은 통계적 사실과

$$E(\bar{y}_i) = E\left(\mu + b_i + \frac{\sum_j e_{ij}}{J}\right) = \mu$$

$$V(\bar{y}_i) = V\left(\mu + b_i + \frac{\sum_j e_{ij}}{J}\right) = V(b_i) + V\left(\frac{\sum_j e_{ij}}{J}\right) = \sigma_a^2 + \frac{\sigma_e^2}{J}$$

정규분포의 성질에 의하여 \bar{y}_i 는 다음과 같은 정규분포를 따르게 된다.

$$\bar{y}_i \sim N\left(\mu, \sigma_a^2 + \frac{\sigma_e^2}{J}\right)$$

B.3.1.3. 성적과 IQ 의 결합 분포

또한 i 번째 학생의 실제 IQ 인 $\mu + a_i$ 의 분포는 다음과 같이 주어진다.

$$\mu + a_i \sim N(\mu, \sigma_a^2)$$

더 나아가 i 번째 학생의 평균 성적 \bar{y}_i 와 실제 IQ 인 $\mu + a_i$ 의 공분산은 다음과 같이 주어진다.

$$\begin{aligned} \text{Cov}(\bar{y}_i, \mu + a_i) &= \text{Cov}\left(\mu + a_i + \frac{\sum_j e_{ij}}{J}, \mu + a_i\right) \\ &= \text{Cov}(a_i, a_i) = \sigma_a^2 \end{aligned}$$

위와 같은 결과와 다변량 정규분포의 성질을 이용하면 i 번째 학생의 평균 성적 \bar{y}_i 와 실제 IQ 인 $\mu + a_i$ 의 결합분포는 다음과 같이 유도할 수 있다.

$$\begin{bmatrix} \text{IQ} \\ \text{test score} \end{bmatrix} = \begin{bmatrix} \mu + a_i \\ \bar{y}_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2/J \end{bmatrix}\right) \quad (\text{B.4})$$

B.3.1.4. IQ 의 최적 예측

i 번째 학생의 평균 성적 \bar{y}_i 가 주어진 경우 실제 IQ 인 $\mu + a_i$ 어떻게 예측할 수 있을까?

! Important

통계 이론에 의하면 실제 IQ 인 $\mu + a_i$ 에 대한 최적 예측값(best prediction) 은 학생의 평균 성적 \bar{y}_i 가 주어졌을 때 $\mu + a_i$ 에 대한 조건부 기대값(conditiona expectation)이다.

위에서 주어진 결과 B.4 에서 이변량 정규분포의 조건부 분포에 대한 성질 A.2 을 이용하면 \bar{y}_i 가 주어진 경우 $\mu + a_i$ 의 조건부 기대값은 다음과 같이 주어진다.

$$\begin{aligned} E(\mu + a_i | \bar{y}_i) &= E(\mu + a_i) + \frac{\text{Cov}(\mu + a_i, \bar{y}_i)}{V(\bar{y}_i)}[\bar{y}_i - E(\bar{y}_i)] \\ &= \mu + \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/J}(\bar{y}_i - \mu) \end{aligned} \quad (\text{B.5})$$

모집단의 IQ 의 평균이 $\mu = 100$ 이라고 가정하자. 또한 임의효과와 오차항의 분산이 각각 10^2 와 10^2 라고 가정하자.

$$\mu = 100, \quad \sigma_a^2 = 10^2, \quad \sigma_e^2 = 10^2$$

이제 10개의 문항으로 이루어진 테스트를 통하여 i 번째 학생의 평균 점수 \bar{y}_i 가 130 점으로 나타났다고 하자.

$$\bar{y}_i = 130, \quad J = 10$$

학생의 실제 IQ 인 $\mu + a_i$ 의 예측값은 얼마인가?

위의 결과 B.5 에서 주어진 값들을 대입하면 다음과 같이 주어진다.

$$\begin{aligned}\widehat{IQ} &= E(\mu + a_i | \bar{y}_i = 130) \\ &= \mu + \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/J} (\bar{y}_i - \mu) \\ &= 100 + \frac{10^2}{10^2 + 10^2/10} (130 - 100) \\ &= 127.27\end{aligned}$$

위의 식에서 유도한 것은 정규분포를 가정하고 $\mu + a_i$ 에 대한 조건부 기대값을 구한 것이며 이는 $y_{ij} = 130$ 이 주어진 경우 IQ 의 $\mu + a_i$ 에 대한 최적 예측값은 127.27 로 나타난 것이다.

B.3.2. 임의효과의 예측

임의효과는 관측할 수 없는 확률변수이지만 관측값이 주어진 경우 임의효과에 대한 기대값으로 그 값을 예측할 수 있다. 임의효과는 모수가 아니라 확률변수이므로 이에 대한 추론을 추정(estimation)이라고 하지 않고 예측(prediction)이라고 한다.

앞에서 공부한 일원배치모형 8.8 에서 임의효과 a_i 와 그룹의 평균 \bar{y}_i 의 공분산 구해보자.

$$\begin{aligned}\text{cov}(a_i, \bar{y}_{i.}) &= \text{cov}(a_i, \mu + a_i + \bar{e}_{i.}) \\ &= \text{cov}(a_i, a_i) \\ &= \text{var}(a_i) \\ &= \sigma_a^2\end{aligned}$$

임의효과 a_i 와 그룹의 평균 \bar{y}_i 는 각각 다음과 같은 정규분포를 따르므로

$$a_i \sim N(0, \sigma_a^2), \quad \bar{y}_{i.} \sim N\left(\mu, \sigma_a^2 + \frac{\sigma_e^2}{J}\right)$$

임의효과 a_i 와 그룹의 평균 \bar{y}_i 는 다음과 같은 이변량 정규분포를 따른다.

$$\begin{bmatrix} a_i \\ \bar{y}_{i.} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2/J \end{bmatrix}\right)$$

따라서 그룹의 평균 \bar{y}_i 이 주어진 경우 임의효과 a_i 의 조건부 분포는 정규분포이며 조건부 기대값은 다음과 같이 주어진다 (강의노트에서 A.2 참조)

$$\begin{aligned}
E(a_i|\bar{y}_{i.}) &= E(a_i) + \frac{\text{cov}(a_i, \bar{y}_{i.})}{\text{var}(\bar{y}_{i.})}(\bar{y}_{i.} - \mu) \\
&= 0 + \frac{\text{cov}(a_i, \bar{y}_{i.})}{\text{var}(\bar{y}_{i.})}(\bar{y}_{i.} - \mu) \\
&= \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/J}(\bar{y}_{i.} - \mu)
\end{aligned}$$

위의 식에서 모르는 모수 μ , σ_a^2 , σ_e^2 를 최대가능도 추정량으로 대체해주면 임의효과 a_i 에 대한 예측값을 구할 수 있다.

$$\begin{aligned}
\hat{a}_i &= \hat{E}(a_i|\bar{y}_{i.}) \\
&= \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2/J}(\bar{y}_{i.} - \hat{\mu}) \\
&= \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2/J}(\bar{y}_{i.} - \bar{y}_{..})
\end{aligned} \tag{B.6}$$

이제 임의효과 모형에서 그룹의 평균 $\mu + a_i$ 의 예측값을 유도하면 다음과 같다.

$$\begin{aligned}
\widehat{\mu + a_i} &= \hat{\mu} + \hat{a}_i \\
&= \bar{y}_{..} + \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2/J}(\bar{y}_{i.} - \bar{y}_{..})
\end{aligned}$$

i 고정효과의 그룹 평균 예측

위의 B.6 을 보면 고정효과를 고려한 일원배치모형 8.7 에서 각 그룹의 평균 $\mu + \alpha_i$ 의 추정식과 차이가 나는 것을 볼 수 있다.

$$\widehat{\mu + \alpha_i} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) = \bar{y}_{i.}$$

임의효과를 이용하면 그룹의 효과가 고정효과를 이용한 모형보다 절대값이 작게 예측되는 것을 알 수 있다.

$$\left| \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2/J}(\bar{y}_{i.} - \bar{y}_{..}) \right| \leq |\bar{y}_{i.} - \bar{y}_{..}|$$

두 모형에서 비교하면 임의효과 모형에서 그룹 평균에 대한 예측값이 고정효과 모형에서 보다 전체 평균에 더 가깝게 나타나는 것을 알 수 있다. 이러한 현상을 전체 평균으로의 축소현상(shrinkage to grand mean) 이라고 부른다.

예제 Example 8.1 에서 공부한 반복측정자료 `sleepstudy` 에서 혼합모형을 통해서 얻은 각 개인의 절편과 기울기에 대한 예측값이 개체의 개별적인 회귀직선에서 얻은 절편과 기울기보다 전체평균값 방향으로 축소되는 경향(shrinkage)을 볼 수 있다.

C. 과제 자료 화일

이 장은 과제에 사용되는 자료 화일의 다운로드 링크를 제공한다.

C.1. 과제 4

- 문제 2 자료

References

- Agresti, Alan. 2003. *Categorical Data Analysis*. Vol. 482. John Wiley & Sons.
- . 2012. *Categorical Data Analysis*. Vol. 792. John Wiley & Sons.
- Butler-Laporte, Guillaume, Alexander Lawandi, Ian Schiller, Mandy Yao, Nandini Dendukuri, Emily G McDonald, and Todd C Lee. 2021. “Comparison of Saliva and Nasopharyngeal Swab Nucleic Acid Amplification Testing for Detection of SARS-CoV-2: A Systematic Review and Meta-Analysis.” *JAMA Intern Med* 181 (3): 353–58.
- Jaewon Lee, Hanna Yu, Mira Park. 2005. 생명과학연구를 위한 통계적 방법. 1st ed. 자유아카데미.
- Moore, Dirk F. 2016. *Applied Survival Analysis Using r*. Springer.
- Searle, Shayle R, George Casella, and Charles E McCulloch. 2009. *Variance Components*. John Wiley & Sons.