의학통계학

이용희

2024-09-30, 오전 01 시

Table of contents

List of Figures

List of Tables

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

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 무었인지 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 임살실험(clinical trial)이라고 부르는 실험들은 편향이 발생할 수 있는 여러 가지 요인를 제거하기 위하여 매우 정교하게 설계됩니다.

사실 의학연구에서 통계학자의 가장 중요한 역활은 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 **연구을 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)를 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 수행해야할 작업은 무었을까요? 또한 실험이 종료되면 수집한 자료들은 어떻게 분석해야 할까요?

임상실험을 계획하는 경우 통계학자의 입장에서 중요하게 고려해야할 사항은 다음과 같습니다.

- 사용할 실험 방법은 무었인가?
- 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 tesing)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 대한여 배울것 입니다.

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 Y로 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?

또는 환자의 상태(outcome)가 이항변수로 표현될 수 있다.

- 환자는 치료가 되었는가?
- 환자가 사망하였는가?

이제 코로나 19 치료제의 효과를 알아보기 위한 임상실험을 수행하는 경우를 생각해보자. 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 X로 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.

> **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, prevalnce 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.

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

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.

> **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.

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

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.

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 2가나오면 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

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.

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

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

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

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 알아보고자 임상실험을 실시하였다. 22,701명의 남성을 **임의화(randomization)** 을 통해서 두 그룹으로 나눈 후, 한 그룹은 매일 일정량의 아스피린을 복용시키고 다른 그룹은 위약(palcebo)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.

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

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 싱대구간을 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.

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

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 발병률와 약물남용의 비율을 추정하지 않고 사례-대조 연구의 자료만으로 추론이 가능하다.

다음의 가설은 두 비율의 비교를 오즈비로 표현한 것이다.

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

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

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

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

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

심작발작을 일으킨 환자와 그렇지 않은 사람들을 각각 214명씩 조사하여 과거에 약물남용을 한 경력이 있느지 조사한 **사례-대조 연구(case-control study)** 의 결과가 표 2.5 에 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(restrspective 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

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

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 `R` 의 `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 오즈비과 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 x 2 분할표

일반적으로 2×2 분할표에서 다음과 같은 두 가지 가설이 가능하다.

- 동질성 검정(homogeneuty test)

  변수 X 가 단순하게 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.

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

- 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 연과성은 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 통계량**을 사용한다.

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

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 기데 돗수 E_{ij} 는 다음과 같이 계산된다.

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

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.

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

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 아스프린과 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
```

```
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 연그ㅜ가 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 횩과를 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 x 2 분할표

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

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

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

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

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

이제 가설검정을 위한 통계량 Q_{CMH} 은 다음과 같다.

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

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

		Response				
Center	Treatment	Success	Failure	Odds Ratio	μ_{11k}	var(n_{11k})
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

```
     [,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
```

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

```
    Mantel-Haenszel chi-squared test without continuity correction

data:  beitler
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 시용한다.

이제 이산형 변수가 짝으로 나타나는 경우를 생각해보자. 예를 들어 눈병 치료에 사용되는 A약과 B약의 효과를비교하기 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.

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

분할표 그림 3.6 에서 p_1 과 p_2의 추정량은 다음과 같다.

		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.

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

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		
	Approve	Disapprove	Total
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
```

```
     [,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 덕분으로 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 종류** 가 있다.

연구자들이 진단 기법을 개발할 때 오류의 가능성이 작아지도록 노력하지만, 불행하게도 두 가지 오류의 확률을 모두 0 으로 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 A 와 B 는 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)는 바이러스에 감염되었거나 질병이 있다는 사건을 말한다. 음성(negative)는 양성의 반대 사건이다. 하지만 양성과 음성의 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.

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 ($T+$)	TP	FP
음성 ($T-$)	FN	TN

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

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

이제 분할표 4.1 에서 민감도와 특이도는 다음과 같이 정의된다.

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

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

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

Source	Test finding, No. TP	FP	FN	TN	Sensitivity (95% CrI), %	Specificity (95% CrI), %
Akgun et al[14]	30	5	25	38	60.6 (46.3-77.0)	98.8 (95.4-99.9)
Becker et al[15]	11	2	10	62	65.4 (44.3-89.8)	99.0 (96.5-99.9)
Byrne et al[16]	12	0	2	96	82.2 (64.8-95.5)	99.3 (98.0-99.9)
Caulley et al[22]	34	14	22	1869	79.5 (56.8-95.8)	99.3 (98.3-99.9)
Cheuk et al[25]	104	37	18	70	84.5 (72.4-92.0)	99.1 (97.3-99.9)
Iwasaki et al[26]	8	1	1	66	83.8 (65.8-96.1)	99.2 (97.5-99.9)
Hanson et al[23]	75	6	5	268	87.3 (74.1-96.5)	99.2 (97.7-99.9)
Kojima et al[17]	20	6	3	16	84.6 (69.9-95.0)	99.1 (97.2-99.9)
Landry et al[27]	28	2	5	89	81.8 (66.8-93.9)	99.2 (97.6-99.9)
McCormick-Baw et al[29]	47	1	2	105	86.0 (71.1-96.0)	99.2 (97.8-99.9)
Miller et al[21]	33	1	1	56	86.7 (73.1-96.6)	99.2 (97.6-99.9)
Pasomsub et al[3]	16	2	3	179	83.3 (66.1-96.0)	99.2 (97.8-99.9)
Teo et al[18]	139	70	11	116	89.4 (77.1-96.3)	99.1 (97.1-99.9)
Vogels et al[19]	32	3	2	30	86.4 (72.1-96.1)	99.1 (97.2-99.9)
Williams et al[28]	33	1	6	49	79.9 (65.0-91.6)	99.1 (97.5-99.9)
Yokata et al[20]	42	6	4	1872	85.9 (71.0-96.6)	99.2 (98.2-99.8)
Pooled	**664**	**157**	**120**	**4981**	**83.2 (74.7-91.4)**	**99.2 (98.2-99.8)**

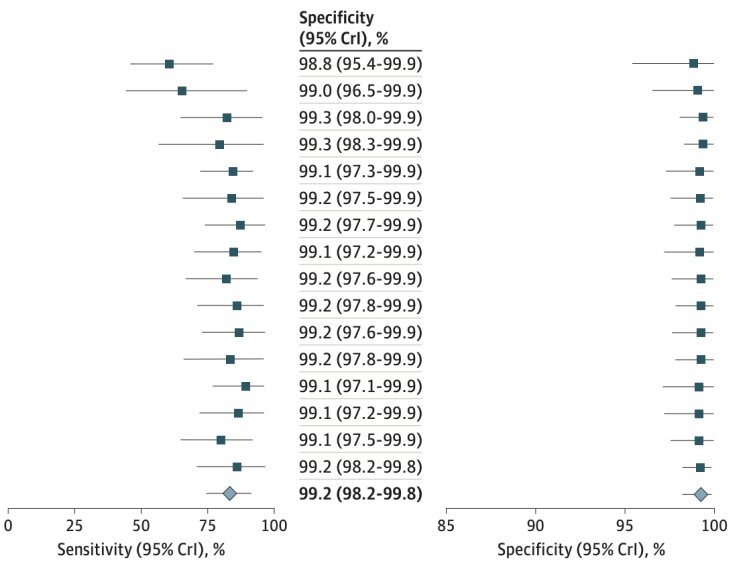


그림 4.1.: 코로나 검사의 민감도와 특이도

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 단순하게 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.

$$PV+ = P(D+|T+) \tag{4.3}$$
$$PV- = P(D-|T-) \tag{4.4}$$

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

$$\text{Sensitivity} = P(T+|D+) \tag{4.5}$$
$$\text{Specificity} = P(T-|D-) \tag{4.6}$$

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.

$$\begin{aligned} P(D+|T+) &= \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)} && (4.7)\\ &= \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + [1-P(T-|D-)][1-P(D+)]} && (4.8)\\ &= \frac{(\text{민감도})(\text{유병률})}{(\text{민감도})(\text{유병률}) + (1-\text{특이도})(1-\text{유병률})} && (4.9) \end{aligned}$$

비슷한 계산 방법으로 음성예측도는 다음과 같이 주어진다.

$$\begin{aligned} P(D-|T-) &= \frac{P(T-|D-)P(D-)}{P(T-|D-)P(D-) + P(T-|D+)P(D+)} && (4.10)\\ &= \frac{P(T-|D-)[1-P(D+)]}{P(T-|D-)[1-P(D+)] + [1-P(T+|D-)]P(D+)} && (4.11)\\ &= \frac{(\text{특이도})(1-\text{유병률})}{(\text{특이도})(1-\text{유병률}) + (1-\text{민감도})(\text{유병률})} && (4.12) \end{aligned}$$

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.

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

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

```
[1] 0.1267108
```

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

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

$$= 0.9992 \tag{4.15}$$

```
(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
```

```
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.

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

여기서 오차항 e_{ij}는 모두 독립이다.

모형 5.1 은 일반적으로 **평균모형(mean model)** 이라고 부르며 모형의 이름대로 μ_i는 i번째 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.

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

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

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

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

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

$$\sum_{i=1}^{a} \alpha_i = 0 \tag{5.3}$$

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

5.3. 가설

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

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

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

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

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \quad \text{vs.} \quad H_1 : \text{ not } H_0 \tag{5.4}$$

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)를나누어 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.

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

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_i들에 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.

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

결과적으로 다음과 같은 변동의 분해를 제곱합의 형식으로 얻을 수 있다.

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

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.

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

- SSE 를 잔차 제곱합(Residual Sum of Squares)이라고 부르며 관측 오차에 발생된 집단 내의 변동 또는 급내 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 변동으로거 집단 간의 변동 또는 급간 변동(bwtween-group variation)을 의미한다.

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

이제 분해식 5.6 을 다음과 같이 나타낼수 있다.

$$SST = SSA + SSE \tag{5.7}$$

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 상대적으 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 ϕ로 표기)를 구할 수 있다.

제곱합의 자유도 = 제곱합을 구성하는 편차의 개수 − 선형제약 조건의 개수

각 제곱합에 대한 선형제약조건은 편차들의 합이 0이 되는 조건이다. 이제 식 5.7 에 주어진 제곱합의 자유도에 대한 정보를 다음과 같이 정리할 수 있다.

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

반드시 유의해야 할 점은 총제곱합의 자유도는 처리제곱합의 자유도와 오차제곱합의 자유도의 합과 같다는 것이다. 이를 이용하면 두 개의 자유도만 구해면 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.

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

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.

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

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.

- 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_E를 가지는 F-분포를 따른다.

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

유의수준 α에서 F-통계량이 기각역을 벗어나면 귀무가설을 기각한다.

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

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

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

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

요인	제곱합	자유도	평균제곱합	F_0	p-값
처리	SSA	$\phi_A = a-1$	$MSA = SSA/\phi_A$	$F_0 = MSA/MSE$	$P[F(\phi_A, \phi_E) > F_0]$
잔차	SSE	$\phi_E = a(r-1)$	$MSE = SSE/\phi_E$		
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

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

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

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

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

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

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

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

$$(\bar{y}_{i.} - \bar{y}_{j.}) \pm t(1-\alpha/2, \phi_E)\sqrt{\frac{2MSE}{r}} \tag{5.11}$$

신뢰구간 7.1 에서 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 α에서 다음과 같은 조건을 만족하면 위의 귀무가설을 기각한다.

$$|\bar{y}_{i.} - \bar{y}_{j.}| > t(1-\alpha/2, \phi_E)\sqrt{\frac{2MSE}{r}} \tag{5.12}$$

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.

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

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.

치로 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>
```

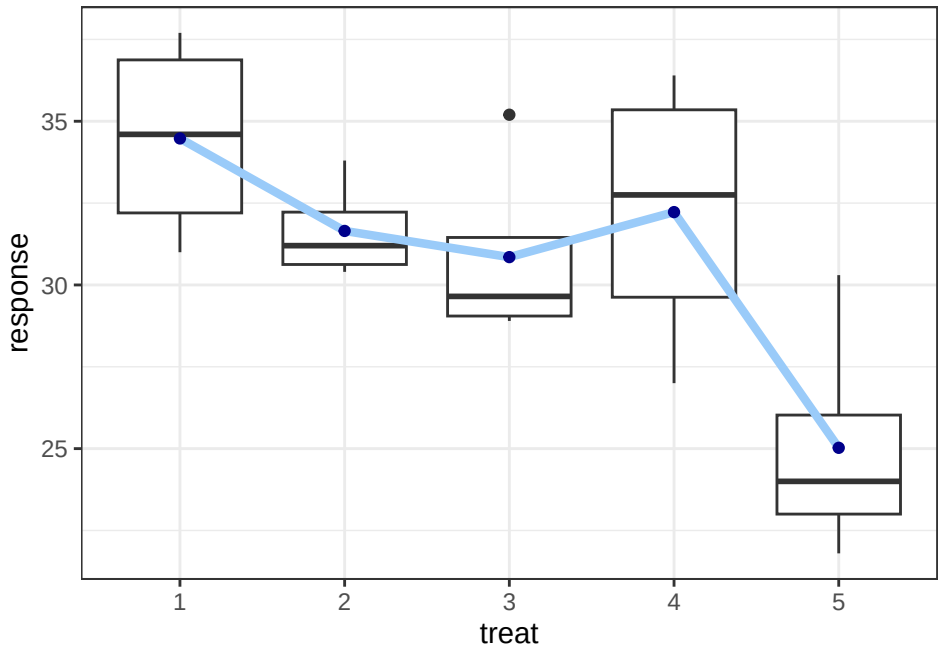
표 5.3.: 저혈당 환자에 대한 임상실험 결과

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()
```



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)
```

```
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)
```

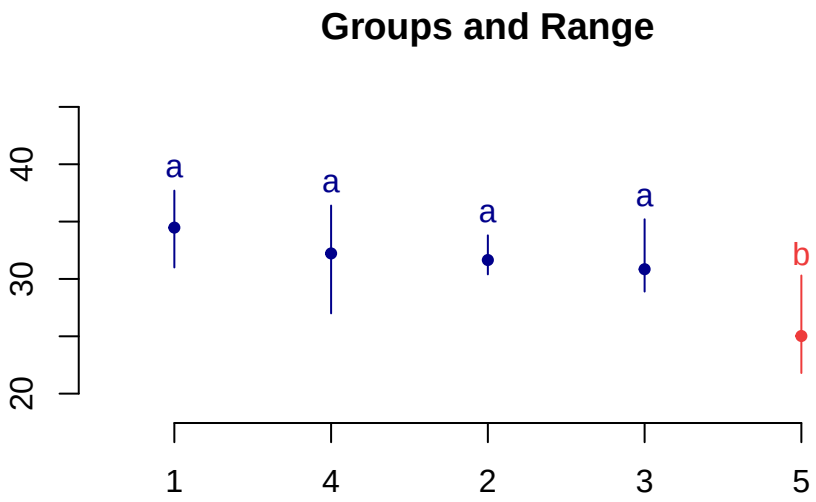
```
Study: anova.res ~ "treat"

LSD t Test for response
```

```
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 포함시켜주는 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 되야한다는 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 사용를 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 x의 효과를 다음과 같이 더해주는 것이다.

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

모형 6.1 에서 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 6.1 에서 각 모수의 추정은 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.

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

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

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

또한 공변량의 효과에 대한 검정도 할 수 있다.

$$H_0 : \beta = 0 \quad \text{vesus} \quad H_1 : \beta \neq 0 \tag{6.3}$$

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		

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 최소제곱평균(Least Square Mean)이라고 한다

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

또한 그룹의 평균들을 각각 비교할 수 있다.

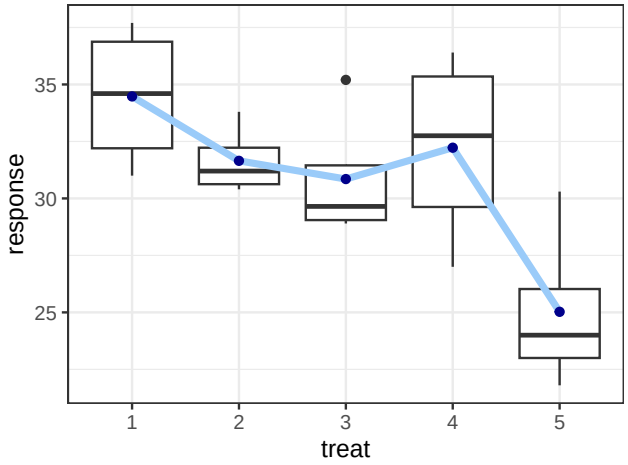
$$H_0 : \alpha_i = \alpha_j \quad \text{vs.} \quad H_1 : \alpha_i \neq \alpha_j$$

6.5. 예제: 저혈당 실험

앞에서 살펴본 예제 Section 5.9 의 저혈당 임상실험에서는 혈당을 감소시키기 위한 다섯 개의 처리(`treat`)를 비교하려고 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)) +
```

```
  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()
```

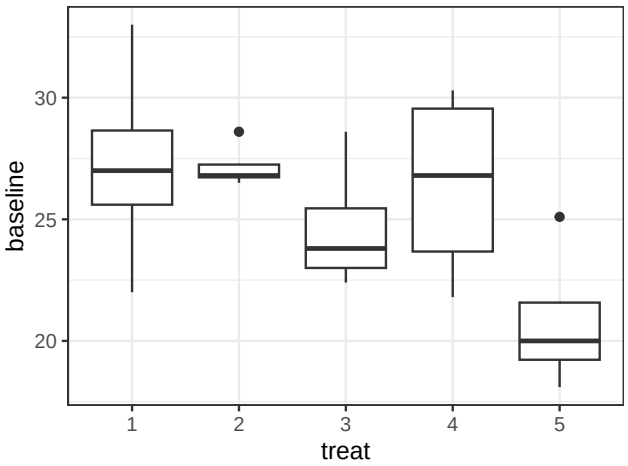


대부분의 임상실험에서는 치료를 시작하기 전에 반응변수의 값을 먼저 측정하고 치료 후의 변화를 본다. 치료를 시작하기 전에 반응변수의 값을 **기준값(baseline value)** 라고 한다. 기준값은 환자의 치료 전 상태를 나타내는 중요한 변수이며 반응값에 영향을 미칠 수 있으므로 대부분의 임상실험에서는 기준값을 공변량으로 분석에 포함한다. 저혈당 실험의 공변량 x 는 치료 전 측정한 혈당량 수치(`baseline`)이다.

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

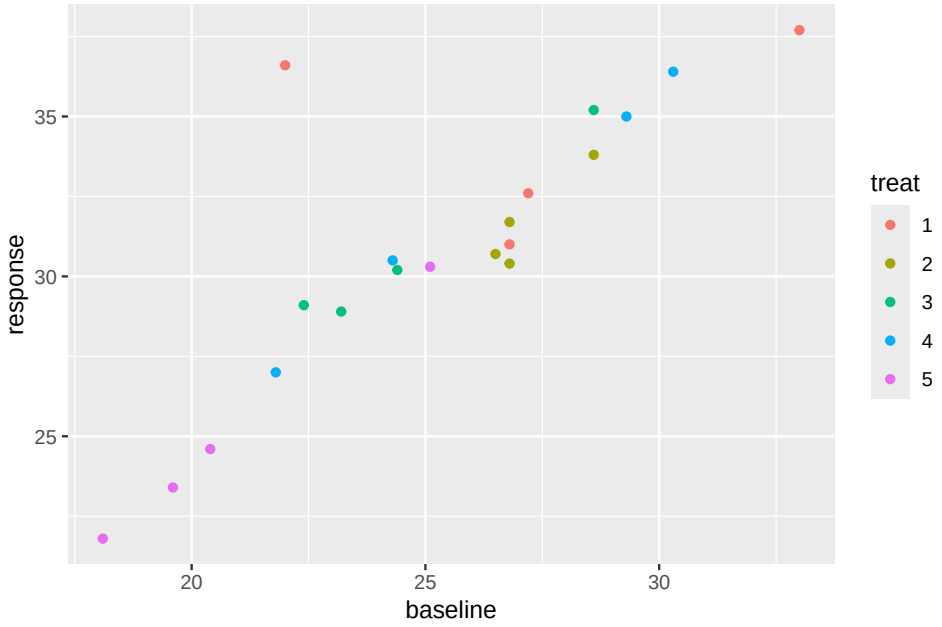
이제 치료 전 측정한 혈당량의 분포를 처리 집단별로 살펴보자. 아래 그림에서 보면 치료잡단별로 치료 전 측정한 혈당량의 분포가 다르다는 것을 알 수 있다.

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

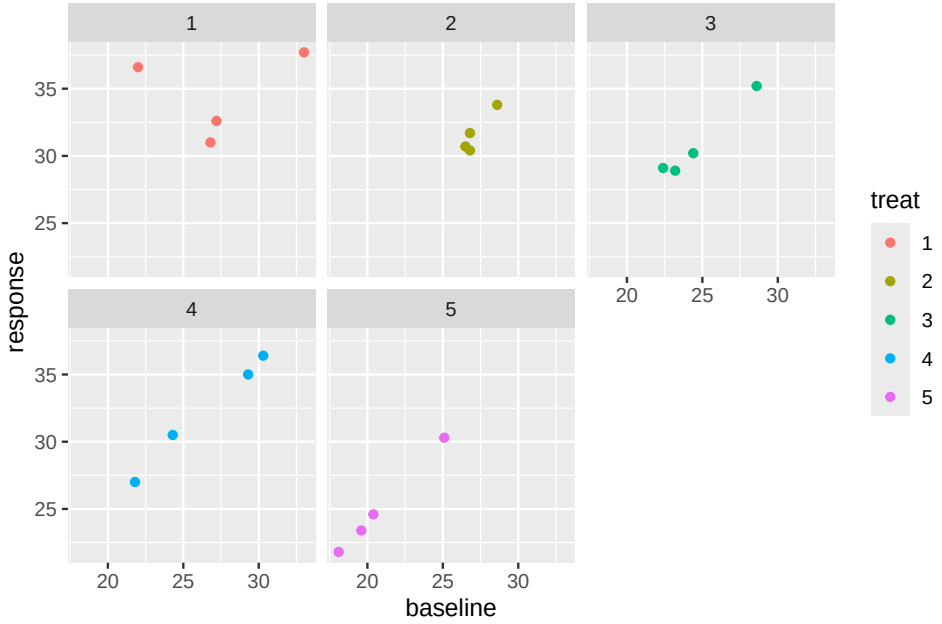


이제 치료 전 측정한 혈당량과 치료 후의 혈당량의 관계는 다음과 같은 산점도로 나타낼 수 있으며 강한 양의 선형관계가 있다는 것을 알 수 있다.즉 초기 형당량 수치가 크면 1달 후 혈당량 수치도 평균적으로 크다.

```
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.

```r
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
```

```r
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 ***
```

```
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.

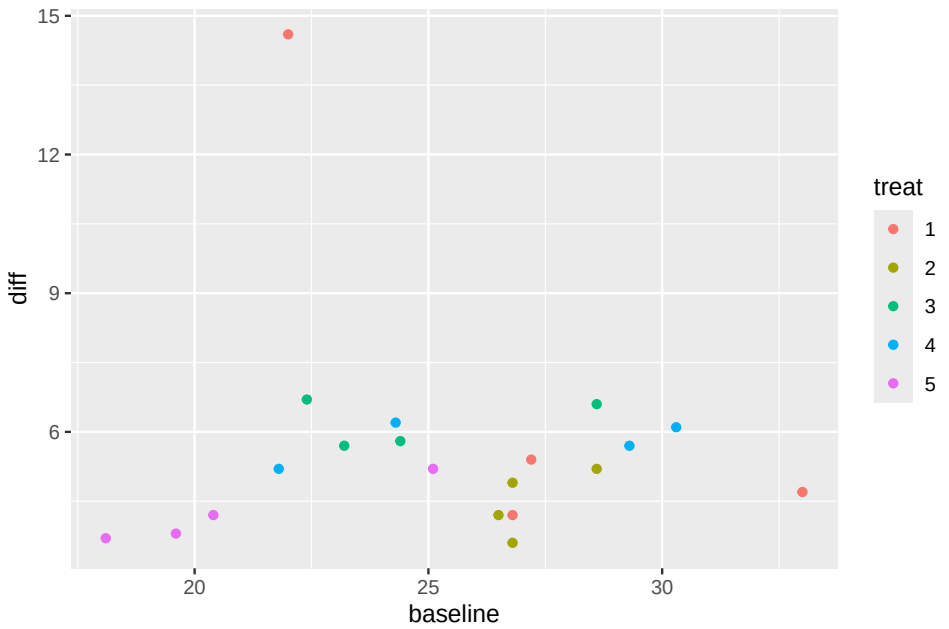
하지만 치료약의 효과를 보면 공분산분석에서는 유의하지 않게 나타났다. 이는 공변량인 치료 전의 혈당이 치료 후의 혈당을 설명하는데 매우 유의한 요인이기 때문이다. 즉 치료 후의 혈당의 집단별 변동이 큰 이유는 치료 효과때문이 아니라 치료 전의 혈당의 차이 때문으로 보여진다.

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.

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.

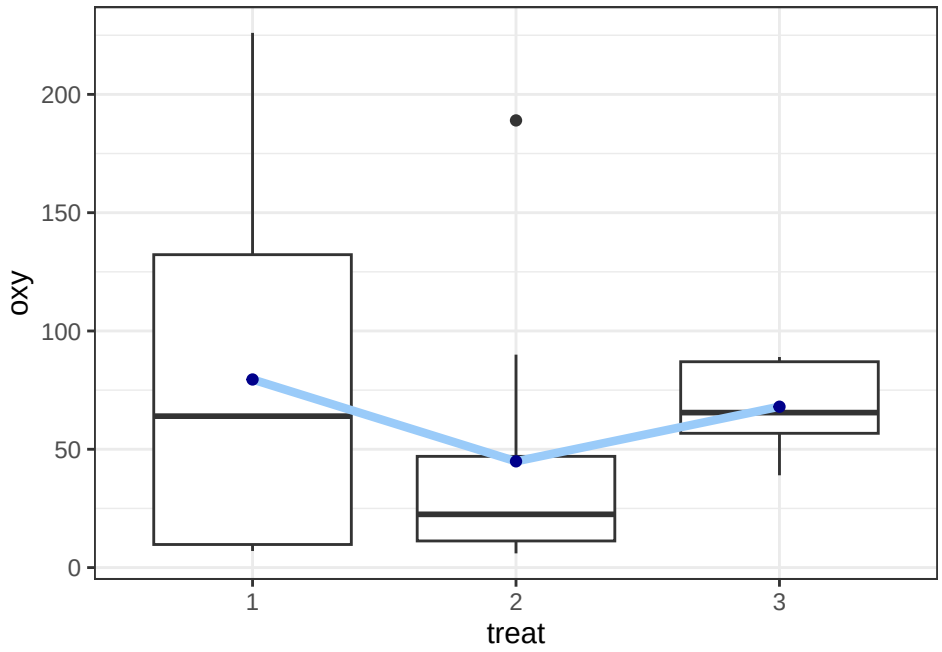
```
dfsumm2 <- oxygen %>%
    group_by(treat) %>%
    summarise(mean = mean(oxy), median = median(oxy), sd = sd(oxy), min = min(oxy),
        max = max(oxy))
dfsumm2
```

```
# 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()
```

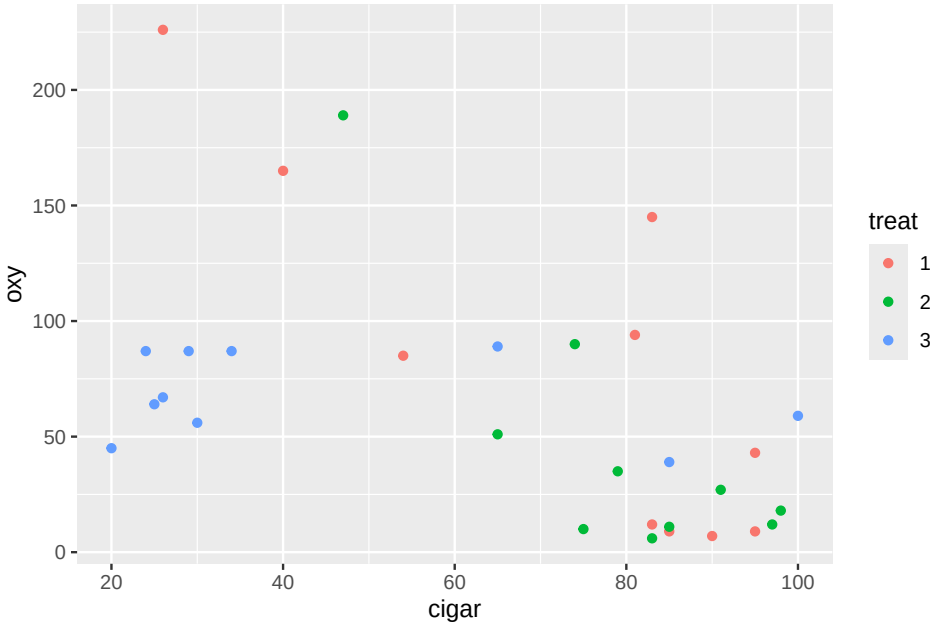
표 6.2.: 산소운반능력 임상실험 결과

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



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

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



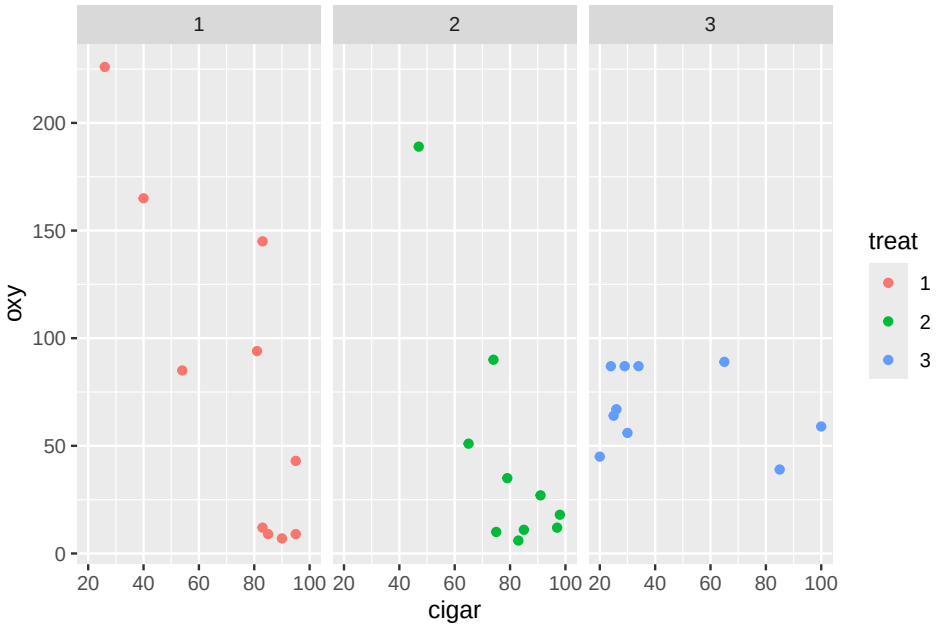
1인당 담배소비량과 산소운반 능력의 관계를 상관계수로 구해보면 음의 상관관계를 보이고 있으면 1인당 담배소비량이 증가하면 산소운반 능력이 감소하는 것을 알 수 있다.

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

```
[1] -0.5349395
```

이제 조금 더 자세하고 처리 그룹별로 산소운반 능력과 1인당 담배소비량의 관계를 살펴보자. 아래 그름을 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)
```

```
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.

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.

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

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)
```

```
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.

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

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

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

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.

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

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.

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

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 positve 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 7.3 에서 처럼 3개의 가설 검정을 동시에 실시한다면 각각의 가설검정에서 제 1 종의 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 1 종의 오류를 범하거나 두 개 또는 하나의 검정에서 제 1 종의 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 1 종의 오류를 범하는 사건
- A_2: H_{02} 검정에서 제 1 종의 오류를 범하는 사건

각 검정에서 제 1 종의 오류를 범할 확률을 α라고 가정하자.

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

이제 두개의 가설검정을 동시에 수행하는 경우 **제 1 종의 오류를 최소한 1번 범하는 사건**은 $P(A_1 \cup A_2)$ 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_1과 A_2가 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 학률은 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 1번 이라도 범하는 사건의 학률은 $1 - (1 - 0.05)^k$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 경우**를 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 개별적로 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(처이 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 k배 줄이면$(0.05/k)$ 실험단위 오류가 5%보다 작거나 같게 된다.

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

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)를 보정하는 방법을 **다중비교(mutiple comparison)** 라고 한다.

위에서 제시한 개별단위 1종 오류를 k배로 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 k개의 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.

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

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

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

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

$$\text{Bonferoni adjusted p-value } = k \times \text{ unadjusted p-value} \tag{7.6}$$

일반적으로 각 가설검정들은 완전히 독립도 아니고 또한 완전한 종속도 아니다. 따라서 실험단위 오류는 각 가설검정들이 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 결정**을 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 목적**으로 여러 개의 가설 검정을 동시에 수행하는 경우에는 다중비교를 적용하지 않거나 다중 비교보다 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 다중비교를 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)
```

```
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
```

```
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.

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

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.

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

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

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

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 $H_{(1)}, H_{(2)}, \ldots, H_{(i^*)}$ 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
```

stats 패키지에 p.adjust 함수의 BH 방법을 활용하면 FDR 제어 방법을 적용한 수정된 p-값을 계산할 수 있다. FDR 제어 방법을 적용한 수정된 p-값 은 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
```

```
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 HSD 에서는 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
```

```
  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 0.3638
4-1 -2.250  -8.558182  4.058182 0.7335
5-1 -9.450 -15.758182 -3.141818 0.0031 **

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

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.