

# 데이터 분석 방법론

이용희

2023-09-06



# Table of contents

<b>Preface</b>	<b>1</b>
<b>1. 서론</b>	<b>3</b>
1.1. 학습 내용 . . . . .	1
1.2. R 언어 . . . . .	1
1.3. 참고도서 . . . . .	1
1.4. 유용한 사이트 . . . . .	1
<b>I. 분할표의 분석</b>	<b>3</b>
<b>2. 연관성의 측도</b>	<b>5</b>
2.1. 필요한 패키지 . . . . .	5
2.2. 이항변수 . . . . .	5
2.3. 분할표와 연관성의 측도 . . . . .	6
2.3.1. 분할표 . . . . .	6
2.3.2. 상대위험 . . . . .	7
2.3.3. 기여위험과 백신효과 . . . . .	8
2.3.4. 오즈비 . . . . .	8
2.4. 신뢰구간 . . . . .	10
2.4.1. 예제: 아스피린 임상실험 . . . . .	10
2.5. 사례-대조 연구 . . . . .	12
2.5.1. 사례대조 연구의 목표와 가설 . . . . .	13
2.5.2. 오즈비의 비교 . . . . .	13
2.5.3. 예제: 약물남용 사례-대조 연구 . . . . .	14
<b>3. 연관성의 검정</b>	<b>17</b>
3.1. 필요한 패키지 . . . . .	17
3.2. 카이제곱 검정 . . . . .	17
3.3. 코크란-멘텔-헨젤 검정 . . . . .	20
3.4. 맥나마 검정 . . . . .	23
<b>4. 진단의 평가</b>	<b>27</b>
4.1. 민감도와 특이도 . . . . .	27
4.2. 양성예측도와 음성예측도 . . . . .	29
<b>References</b>	<b>33</b>



# List of Figures

3.1. 2 x 2 분할표 . . . . .	17
3.2. 2 x 2 분할표: 관측 도수 . . . . .	18
3.3. K 개의 2 x 2 분할표 . . . . .	20
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 . . . . .	28



# List of Tables

2.1. $2 \times 2$ 분할표 . . . . .	6
2.2. 코로나 치료제 실험 결과 . . . . .	7
2.3. $2 \times 2$ 분할표 예제 . . . . .	9
2.4. 아스피린 임상실험 결과 . . . . .	10
2.5. 약물 남용 사례-대조 연구 결과 . . . . .	12
4.1. 진단 기법의 실험 결과 . . . . .	28
4.2. 코로나 바이러스 검사법의 결과 . . . . .	29





# Preface

이 사이트는 데이터 분석 방법론 강의 온라인 강의 노트입니다.



## 1. 서론



## 1.1. 학습 내용

이 교과서는 다양한 형태를 가진 자료를 분석하는 통계적 방법들의 이론과 응용을 살펴보기 위한 것입니다.

이 교과서에서는 다음과 같은 주제를 다룰 것입니다.

- 교차표에서의 통계적 분석방법
- 범주형 자료와 발생횟수를 따르는 자료에 대한 모형 구축과 추론
- 일반화 선형모형에서의 추론
- 반복측정자료와 군집자료에 대한 분석 방법

## 1.2. R 언어

이 교과서에서는 통계 방법들의 실습을 위하여 R 프로그램을 사용합니다. R 프로그램이 익숙하지 않는 학생들은 R 프로그램에 대한 기초적인 내용을 먼저 숙지하는 것을 추천합니다. 참고로 저자의 R 기초 강의 사이트에서 R 프로그램에 대한 기초적인 내용을 배울 수 있습니다.

이 강의에서 사용하는 R 패키지는 다음과 같다.

## 1.3. 참고도서

- Faraway (2016)
- Agresti (2007)
- Agresti (2012)

## 1.4. 유용한 사이트

- Data sets for “An Introduction to Categorical Data Analysis”
- R codes for “An Introduction to Categorical Data Analysis”



## Part I.

### 분할표의 분석





## 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)와 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?

또는 환자의 상태(outcome)가 이항변수로 표현될 수 있다.

- 환자는 치료가 되었는가?
- 환자가 사망하였는가?

## 2. 연관성의 측도

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.

Table 2.1.: $2 \times 2$ 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  $2 \times 2$  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 Table 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 Table 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.

Table 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. 연관성의 측도

### 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.

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.

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

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

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 Equation 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.

	Table 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.2by2를 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 ratio라고 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)
```

## 2. 연관성의 측도

```
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.

Table 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.

$$\begin{aligned} P(\text{심장발작이 발생했다}|\text{약물남용을 했다}) &= P(H+|D+) \\ P(\text{심장발작이 발생했다}|\text{약물남용을 하지 않았다}) &= P(H+|D-) \end{aligned}$$

즉 연구의 목표는 다음과 같은 가설을 검정하는 것이다.

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

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

## 2. 연관성의 측도

가설 Equation 2.6 에서 나타는 오즈비는 심장발작 발병률과 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

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

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

상대위험의 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 상대구간을 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  $2 \times 2$  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

Figure 3.1.:  $2 \times 2$  분할표

일반적으로  $2 \times 2$  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.

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

### 3. 연관성의 검정

- 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  $2 \times 2$  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		$n_{+1}$	$n_{+2}$	$n$

Figure 3.2.:  $2 \times 2$  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 기댓도수  $E_{ij}$  는 다음과 같이 계산된다.

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

귀무가설 하에서 표본의 크기가 충분히 크면 식 Equation 3.1 의 카이제곱 검정통계량  $\chi^2$  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  $2 \times 2$  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 Table 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 Equation 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 Equation 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 Table 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
```

### 3. 연관성의 검정

[,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 병원에서 연구가 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  $2 \times 2$  분할표가 있다고 하자.

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

Figure 3.3.:  $K$  개의  $2 \times 2$  분할표

$K$  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  $\mu_{k11}$  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.



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

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.

**TABLE 6.9 Clinical Trial Relating Treatment to Response for Eight Centers**

Center	Treatment	Response		Odds Ratio	$\mu_{11k}$	$\text{var}(n_{11k})$
		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	$\infty$	3.52	1.20
	Control	0	12			
6	Drug	1	10	$\infty$	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: Beitzler and Landis (1985).

Figure 3.4.: 8개 병원의 임상실험 결과

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 Equation 3.3 의 검정 통계량  $Q_{CMH}$  를 이용하여 치료제의 효과가 있는지 검정해보자. 검정은 아래와 같이 R 프로그램을 이용한다. 함수 `mantelhaen.test()` 는 코크란-멘텔-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

### 3. 연관성의 검정

	[,1]	[,2]
[1,]	11	25
[2,]	10	27

, , 2

	[,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)
```

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

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

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

분할표 Figure 3.6에서  $p_1$ 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		$n_{+1}$	$n_{+2}$	$n$

Figure 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 10 장 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

Figure 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 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 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.

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

#### 4. 진단의 평가

Table 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 Negative
- $TN$  : True Negative

이제 분할표 Table 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

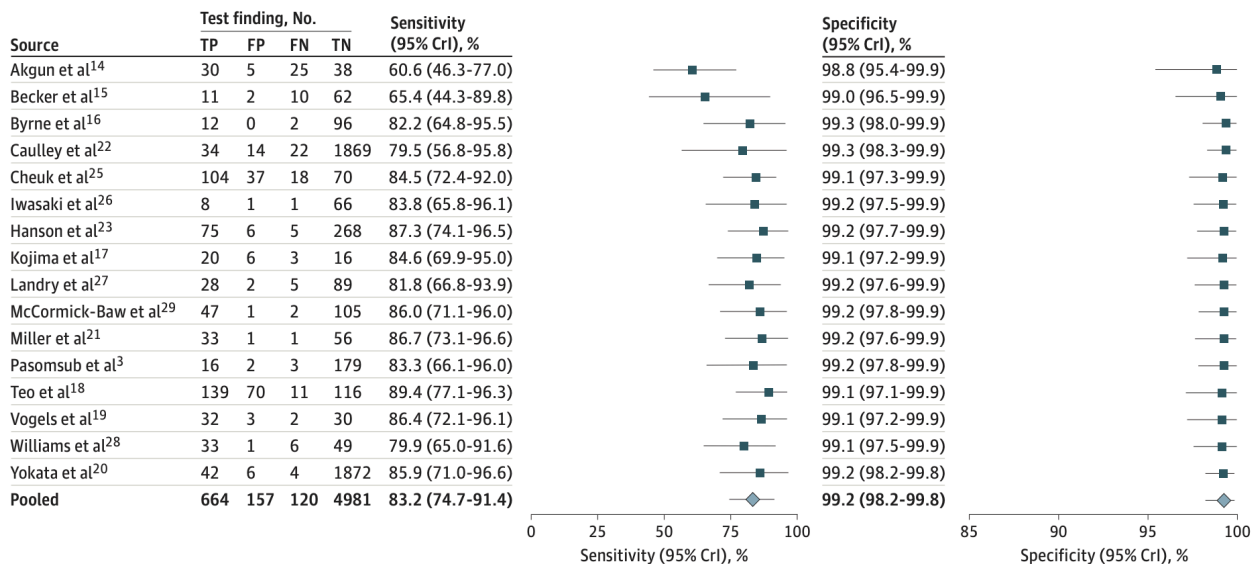


Figure 4.1.: 코로나 검사의 민감도와 특이도

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



Table 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.

$$\text{Sensitivity} = \frac{664}{664 + 157} = 0.8087$$

$$\text{Specificity} = \frac{4891}{120 + 4891} = 0.9760$$

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.

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

$$PV- = P(D- | T-)(\#eq : pv) \quad (4.4)$$

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

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

$$\text{Specificity} = P(T- | D-)(\#eq : sensitivity2) \quad (4.6)$$

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.

#### 4. 진단의 평가

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

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

$$\begin{aligned} P(D+|T+) &= \frac{(\text{민감도})(\text{유병률})}{(\text{민감도})(\text{유병률}) + (1 - \text{특이도})(1 - \text{유병률})} \\ &= \frac{(0.8087)(0.0053)}{(0.8087)(0.0053) + (1 - 0.9760)(1 - 0.0053)} \\ &= 0.1522 \end{aligned}$$

$$(0.8087) * (0.0053) / ((0.8087) * (0.0053) + (1 - 0.9760) * (1 - 0.0053))$$

[1] 0.1522115

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

$$= \frac{(0.9760)(1 - 0.0053)}{(0.9760)(1 - 0.0053) + (1 - 0.8087)(0.0053)} \quad (4.14)$$

$$= 0.9989 \quad (4.15)$$

```
(0.9760)*(1- 0.0053)/((0.9760)*(1-0.0053) + (1-0.8087)*(0.0053))
```

```
[1] 0.9989567
```

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 ,0.8087, 0.9760 )
```

	Sensitivity	SPecificity	Prevalnce	Pred. Post.	Pred. Nega.
1	0.8087	0.976	0.000	0.0000000	1.0000000
2	0.8087	0.976	0.002	0.0632553	0.9996074
3	0.8087	0.976	0.004	0.1191947	0.9992135
4	0.8087	0.976	0.006	0.1690179	0.9988183
5	0.8087	0.976	0.008	0.2136761	0.9984218
6	0.8087	0.976	0.010	0.2539329	0.9980241
7	0.8087	0.976	0.012	0.2904083	0.9976250
8	0.8087	0.976	0.014	0.3236113	0.9972247
9	0.8087	0.976	0.016	0.3539633	0.9968231
10	0.8087	0.976	0.018	0.3818165	0.9964201
11	0.8087	0.976	0.020	0.4074671	0.9960159

```
calpred(preval.range ,0.8320, 0.9920 )
```

	Sensitivity	SPecificity	Prevalnce	Pred. Post.	Pred. Nega.
1	0.832	0.992	0.000	0.0000000	1.0000000
2	0.832	0.992	0.002	0.1724710	0.9996607
3	0.832	0.992	0.004	0.2946176	0.9993203

#### 4. 진단의 평가

4	0.832	0.992	0.006	0.3856613	0.9989788
5	0.832	0.992	0.008	0.4561404	0.9986361
6	0.832	0.992	0.010	0.5123153	0.9982923
7	0.832	0.992	0.012	0.5581395	0.9979473
8	0.832	0.992	0.014	0.5962326	0.9976011
9	0.832	0.992	0.016	0.6283988	0.9972538
10	0.832	0.992	0.018	0.6559215	0.9969053
11	0.832	0.992	0.020	0.6797386	0.9965557

# References

- Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Ltd.
- . 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.
- Faraway, Julian J. 2016. *Extending the Linear Model with r: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC press.

