범주형자료분석팀

3팀

권가민
김수인
김준령
박윤아
이정민

INDEX

1. 범주형 자료 분석

2. 분할표

3. 독립성 검정

4. 연관성 측도

1

범주형 자료 분석

범주형 자료 분석

범주형 자료 분석 (Categorical Data Analysis, CDA)

반응변수가 범주형인 자료에 대한 분석

자료(Data)란 변수에 대한 **관측치의 집합**

이 때 **모집단의 특징을 나타내는 변수**를 측정한 값이 관측치

각 변수를 열(column)로, 변수 별 관측치를 행(row)으로 나열한 행렬 형태

변수의 구분

X 변수

독립변수(independent variable),
설명변수(explanatory variable)

Y 변수

종속변수(dependent variable),
반응변수(response variable)

범주형 자료분석은 **Y 변수가 범주형**인 자료를 분석

자료의 형태

자료는 크게 양적 자료(=수치형 자료)와

질적 자료(=범주형 자료)로 나뉨

우리는 질적자료에 집중!

자료	양적 (Quantitative) 자료	이산형 (Discrete) 자료
		연속형 (Continuous) 자료
	질적 (Qualitative) 자료	명목형 (Nominal) 자료
		순서형 (Ordinal) 자료

자료의 형태

양적 자료 (수치형 자료)

관측된 값이 **수치**로 측정되는 자료

이산형 자료와 연속형 자료로 구분됨

이산형 자료: 값을 셀 수 있는 자료

Ex) 나이, 신생아 수

연속형 자료: 연속인 어떤 구간에서 값을 취하는 자료

Ex) 키, 몸무게

자료의 형태

양적 자료 (수치형 자료)

관측된 값이 **수치**로 측정되는 자료

이산형 자료와 연속형 자료로 구분됨

공분산, 상관계수 등의 수리적인 계산 가능

정규분포의 가정이 있다면, 회귀분석 가능

자료의 형태

집단, 그룹, 카테고리

질적 자료 (범주형 자료)

관측 결과가 여러 개의 **범주의 집합**으로 나타나는 자료

범주 간 순서의 유무에 따라 명목형 자료와 순서형 자료로 구분됨

명목형 자료: 범주간에 순서의 의미가 없는 자료

Ex) 성별, 혈액형

순서형 자료: 범주간에 순서의 의미가 있는 자료

Ex) 선호도 (좋음/보통/싫음)

질적 자료의 특징

- 순서형 자료에 명목형 자료 분석 방법을 적용할 수 있음
- 분할표를 작성할 수 있음
- 각 범주에 특정 점수를 할당하여 **양적자료로 활용할 수 있음**
- 일반적으로 **사칙연산이 불가능**

질적 자료의 특징

순서형 자료에 명목형 자료 분석 방법을 적용할 수 있음

그러나 분석 과정에서 순서에 대한 정보가 무시되어

검정력에 심각한 손실을 가져옴

반대로 명목형 자료에는 순서에 관한 정보가 없으므로

순서형 자료에 대한 분석법을 적용할 수 없음

질적 자료의 특징

분할표를 작성할 수 있음

		Y		
		1	⋯	J
X	1	I*J 개 칸		
	⋯			
	I			

	Y			합계
X	n_{11}	⋯	n_{1j}	n_{1+}
	⋯	⋯	⋯	⋯
	n_{i1}	⋯	n_{ij}	n_{i+}
합계	n_{+1}	⋯	n_{+j}	n_{++}

질적 자료의 특징

각 범주에 특정 점수를 할당하여 **양적자료로 활용할 수 있음**

코드값	코드값 의미	서열
001	대통령	1
002	부통령	2
003	국무총리	3
004	부총리	4
…	…	…
019	9급	31
020	기능 1급	128

코드값이 변수의 범주를 의미

→ **범주형 변수**로 생각할 수 있음

코드값의 크기가 서열을 나타냄

→ **연속형 변수**로 생각할 수 있음

범주형 자료 분석

질적 자료의 특징 범주형 자료를 잘 구분해야 함

각 범주에 특정 점수를 할당하여 **양적자료로 활용할 수 있음**

자료가 숫자로 표현되어 있다고

반드시 **수치형 자료**인 것은 아님!

범주형 변수가 문자형 변수에 국한되어 있지 않음

코드값	코드값 의미	서열
001	대통령	1
002	부통령	2
003	국무총리	3
004	부총리	4
…	…	…
019	9급	31
020	기능 1급	128

변수의 범주를 의미

→ **범주형 변수**로 생각할 수 있음

코드값의 크기가 서열을 나타냄

→ **연속형 변수**로 생각할 수 있음

범주형 자료 분석

질적 자료의 특징

범주형 자료를 잘 구분해야 함

각 범주에 특정 점수를 할당하여 양적자료로 활용할 수 있음

코드값	코드값 의미	서열
001	대통령	1
002	부통령	2
003	국무총리	3
004	부총리	4
…	…	…
019	9급	31
020	기능 1급	128

변수의 범주를 의미

→ 범주형 변수로 생각할 수 있음

코드값의 크기가 서열을 나타냄

→ 연속형 변수로 생각할 수 있음

자료가 숫자로 표현되어 있다고

반드시 수치형 자료인 것은 아님!

범주형 변수가 문자형 변수에 국한되어 있지 않음

분석해야 하는 자료를 범주형 변수로 간주할지,

수치형 자료로 간주할지 주의하여 판단 후 자료통계분석 수행!

질적 자료의 특징

일반적으로 **사칙연산이 불가능**

범주형 변수에 대한 분석은 통계량보다는 **범주별 빈도**에 관심을 갖고,

이를 일반화할 때는 **특정 범주가 발생할 확률**에 관심을 가짐

2

분할표

분할표

분할표 (Contingency Table)

각 범주형 변수에 대한 **결과의 도수**(frequency)를 각 칸에 정리한 표

즉, 범주형 변수들에 대한 관측값을 일목요연하게 **도표로 요약한 자료**

중심(평균, 중앙값)이나 산포도(분산, 표준편차) 등의

기술통계(descriptive)를 진행하는

수치형 변수 분석과는 차이가 있음

분할표

		Y		
		1	···	J
X	1	I×J 개 칸		
	···			
	I			

수준(Level): 각 변수의 **카테고리 개수**

X 변수는 I개의 수준, Y 변수는 J개의 수준을 갖고 있음

→ I×J 크기의 행렬

분할표

		Y		
		1	···	J
X	1	I×J 개 칸		
	···			
	I			

범주형 자료를 분할표로 표현하는 이유

예측 검정력에 대한 요약 가능

독립성 검정 실시할 수 있음

여러 차원의 분할표

수준에 따라 **무한가지 형태**의 분할표를 만들 수 있음
3차원 이상의 고차원일 경우, 분할표보다는
모델링 등의 방식을 통한 분석이 더 큰 편의성을 가짐

2차원 분할표(two-way table)와
3차원 분할표(three-way table)에 집중!

여러 차원의 분할표

2차원 분할표 (I × J)

	Y			합계
X	n_{11}	$\cdots$	n_{1j}	n_{1+}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	n_{i1}	$\cdots$	n_{ij}	n_{i+}
합계	n_{+1}	$\cdots$	n_{+j}	n_{++}

- 일반적으로 X 변수가 행에, Y 변수가 열에 위치
- n_{ij}는 각 칸의 도수를, n_{i+}, n_{+j}는 각 열과 행의 주변(marginal) 도수를 표현

여러 차원의 분할표

2차원 분할표 (I × J)

	Y			합계
X	n_{11}	$\cdots$	n_{1j}	n_{1+}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	n_{i1}	$\cdots$	n_{ij}	n_{i+}
합계	n_{+1}	$\cdots$	n_{+j}	n_{++}

- 일반적으로 X 변수가 행에, Y 변수가 열에 위치
- n_{ij}는 각 칸의 도수를, n_{i+}, n_{+j}는 각 열과 행의 **주변(marginal) 도수**를 표현

여러 차원의 분할표

2차원 분할표 (I × J)

	Y			합계
X	n_{11}	$\cdots$	n_{1j}	n_{1+}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	n_{i1}	$\cdots$	n_{ij}	n_{i+}
합계	n_{+1}	$\cdots$	n_{+j}	$\boxed{n_{++}}$

- n_{++}는 총계
- '+'는 그 위치에 해당하는 도수를 **모두 더했다는 의미**

여러 차원의 분할표

3차원 분할표 (I × J × K)

3차원 분할표

기존의 설명변수와 반응변수에 K개의 수준을 가진

제어변수(제한변수, Control Variable) Z가 추가된 형태

		Y		합계
Z	X	n_{111}	n_{121}	n_{1+1}
		n_{211}	n_{221}	n_{2+1}
	합계	n_{+11}	n_{+21}	n_{++1}
	X	n_{112}	n_{122}	n_{1+2}
		n_{212}	n_{222}	n_{2+2}
	합계	n_{+12}	n_{+22}	n_{+22}

여러 차원의 분할표

3차원 분할표 (I × J × K)

		Y		합계
Z	X	n_{111}	n_{121}	n_{1+1}
		n_{211}	n_{221}	n_{2+1}
	합계	n_{+11}	n_{+21}	n_{++1}
	X	n_{112}	n_{122}	n_{1+2}
		n_{212}	n_{222}	n_{2+2}
	합계	n_{+12}	n_{+22}	n_{+22}

부분분할표

	Y		합계
X	n_{11+}	n_{12+}	n_{1++}
	n_{21+}	n_{22+}	n_{2++}
합계	n_{+1+}	n_{+2+}	n_{+++}

주변분할표

여러 차원의 분할표

3차원 분할표 (I × J × K)

		Y		합계
Z	X	n_{111}	n_{121}	n_{1+1}
		n_{211}	n_{221}	n_{2+1}
	합계	n_{+11}	n_{+21}	n_{++1}
	X	n_{112}	n_{122}	n_{1+2}
		n_{212}	n_{222}	n_{2+2}
	합계	n_{+12}	n_{+22}	n_{+22}

부분분할표

	Y		합계
합계	n_{+1+}	n_{+2+}	n_{+++}

주변분할표

부분분할표

: 3차원 분할표의 형태

여러 차원의 분할표

3차원 분할표 (I × J × K)

		Y		합계
	X	n_{111}	n_{121}	n_{1+1}
		n_{212}	n_{222}	n_{2+2}
	합계	n_{+12}	n_{+22}	n_{+22}

부분분할표

주변분할표: Z 변수 각 수준에서 도수를 합친 2차원 분할표의 형태

	Y		합계
X	n_{11+}	n_{12+}	n_{1++}
	n_{21+}	n_{22+}	n_{2++}
합계	n_{+1+}	n_{+2+}	n_{+++}

주변분할표

여러 차원의 분할표

3차원 분할표 (I × J × K)

거주지	성별	통학 여부		합계
		O	X	
서울	남자	11	25	36
	여자	10	27	37
	합계	21	52	73
인천	남자	16	4	20
	여자	22	10	32
	합계	38	14	52

부분분할표

Z 변수의 수준에 따라

X 변수와 Y 변수가 분류된 분할표

고정된 Z 변수의 각 수준에서

반응변수에 미치는

설명변수의 효과 확인 가능

여러 차원의 분할표

3차원 분할표 (I × J × K)

X 변수와 Y 변수 간의 관계에서
Z 변수의 영향을 무시한 형태

거주지	통학 여부		합계
	O	X	
남자	11+16	25+4	56
여자	10+22	27+10	69
합계	59	66	125

주변분할표

여러 차원의 분할표

3차원 분할표 (I × J × K)

X 변수와 Y 변수 간의 관계에서

위 예시에서는

제어변수 Z(거주지)를 통합하여 표현

거주지와 무관하게 통학 여부(Y)에

성별(X)이 미치는 영향만을 확인

부분분할표와 주변분할표를 통해

변수 간 연관성(association)을 파악할 수 있음!

거주지	통학 여부		합계
	O	X	
남자	11+16	25+4	56
여자	10+22	27+10	69
합계	59	66	125

주변분할표

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수(frequency) 대신 **비율(ratio)**이 들어간 분할표

비율은 각 칸의 도수인 n_{ij}를 전체 도수 n_{++}으로 나눈 것

	Y		합계
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
합계	π_{+1}	π_{+2}	$\pi_{++} = 1$

π_{ij} : 전체 대비 각 칸의 비율 (= 확률)

π_{++} : 분할표 내 모든 칸의 확률의 합 (= 1)

분할표에서의 확률분포

결합확률 (Joint Probability)

모집단에서부터 임의로 추출된 표본이 X 변수의 I번째 수준과 Y 변수의 J번째 수준에 **동시에 속할 확률**

	Y		합계
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
합계	π_{+1}	π_{+2}	$\pi_{++} = 1$

$$\pi_{ij} = P(X = i, Y = j)$$

위 표에서 각 칸의 확률

분할표에서의 확률분포

결합확률 (Joint Probability)

모집단에서부터 임의로 추출된 표본이 X 변수의 I번째 수준과 Y 변수의 J번째 수준에 **동시에 속할 확률**

	Y		합계
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
합계	π_{+1}	π_{+2}	$\pi_{++} = 1$

항상 $\sum \pi_{ij} = 1$ 을 만족!

분할표에서의 확률분포

주변확률 (Marginal Probability)

결합분포의 **행과 열의 합**

Ex) X 변수의 I번째 수준이 전부 일어날 행의 확률(π_{+j})

	Y		합계
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
합계	π_{+1}	π_{+2}	$\pi_{++} = 1$

$$\pi_{i+} = P(X = i),\ \pi_{+j} = P(Y = j)$$

분할표에서의 확률분포

주변확률 (Marginal Probability)

결합분포의 **행과 열의 합**

Ex) X 변수의 I번째 수준이 전부 일어날 행의 확률(π_{+j})

	Y		합계
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
합계	π_{+1}	π_{+2}	$\pi_{++} = 1$

주변 확률 역시 분포함수이기 때문에 해당 범위 내의 확률들의 합은 1

$$\sum_I \pi_{i+} = \sum_J \pi_{+j} = 1$$

분할표에서의 확률분포

조건부 확률 (Conditional Probability)

X가 주어졌을 때 Y에 대한 확률,

즉 X 변수의 각 수준에서의 Y 변수의 값

$P(\mathrm{Y}|\mathrm{X}=\mathrm{i}) = \frac{\pi_{ij}}{\pi_{i+}}$ 로 표현

표본에 대해서는 π 대신 p를 사용하여 추정량을 나타냄

$$p_{ij} = \frac{n_{ij}}{n_{++}}$$

(p_{ij}는 각 칸에 속한 표본의 비율, n_{ij}는 각 칸 도수)

분할표에서의 확률분포

확률분포 예시

연령에 따른 선호 스포츠				
	야구 (Y=1)	축구 (Y=2)	농구 (Y=3)	합계
10대 (X=1)	78 (0.31)	23 (0.09)	29 (0.12)	130
20대 (X=2)	41 (0.16)	42 (0.17)	37 (0.15)	120
합계	119	65	66	250

20대이면서 축구를 선호할 **결합확률**

$$\pi_{22} = \frac{42}{250} = \text{약 } 0.17$$

연령대와 무관하게

농구를 선호할 **주변확률**

$$\pi_{+3} = \frac{66}{250} = \text{약 } 0.264$$

분할표에서의 확률분포

확률분포 예시

연령에 따른 선호 스포츠				
	야구 (Y=1)	축구 (Y=2)	농구 (Y=3)	합계
10대 (X=1)	78 (0.31)	23 (0.09)	29 (0.12)	130
20대 (X=2)	41 (0.16)	42 (0.17)	37 (0.15)	120
합계	119	65	66	250

20대이면서 축구를 선호할 **결합확률**

$$\pi_{22} = \frac{42}{250} = \text{약 } 0.17$$

연령대와 무관하게

농구를 선호할 **주변확률**

$$\pi_{+3} = \frac{66}{250} = \text{약 } 0.264$$

분할표에서의 확률분포

확률분포 예시

연령에 따른 선호 스포츠				
	야구 (Y=1)	축구 (Y=2)	농구 (Y=3)	합계
10대 (X=1)	78 (0.31)	23 (0.09)	29 (0.12)	130
20대 (X=2)	41 (0.16)	42 (0.17)	37 (0.15)	120
합계	119	65	66	250

20대라는 가정 하에

야구를 선호할 **조건부 확률**

$P(\mathrm{Y}=1|\mathrm{X}=2)$

$$=\frac{\pi_{21}}{\pi_{2+}}=\frac{\frac{41}{250}}{\frac{120}{250}}=\frac{41}{120}=\text{약 } 0.34$$

3

독립성 검정

범주형 자료의 통계적 검정

적합도 검정

실제로 얻어진 관측치들의 분포가
귀무가설 하에서 가정한 이론상의 분포와 같은지 검정

동질성 검정

서로 다른 모집단에서 추출한 표본들이
하나의 특성에 대해 동일한 분포를 가지는지 검정

독립성 검정

분할표에서 한 특성이 다른 특성에 영향을 미치는지 검정

범주형 자료의 통계적 검정

적합도 검정

실제로 얻어진 관측치들의 분포가
귀무가설 하에서 가정한 이론상의 분포와 같은지 검정

동질성 검정

서로 다른 모집단에서 추출한 표본들이
하나의 특성에 대해 동일한 분포를 가지는지 검정

독립성 검정

분할표에서 한 특성이 다른 특성에 영향을 미치는지 검정

독립성 검정

독립성 검정의 목적

두 변수 간 **연관성 유무** 확인

분석 가치 판단

두 범주형 변수가

통계적으로 관계가 있는지를 확인

두 변수가 독립이라면

두 변수에 대해 더 이상 분석을 진행할 필요가 없음

독립성 검정의 가설

모든 결합확률이
행과 열의 주변확률의 곱과 동일함

귀무가설 H_0: 두 범주형 변수는 **독립이다**

$$(\pi_{ij} = \pi_{i+} \cdot \pi_{+j})$$

대립가설 H_1: 두 범주형 변수는 **독립이 아니다**

$$(\pi_{ij} \neq \pi_{i+} \cdot \pi_{+j})$$

관측도수와 기대도수

관측도수 (Observed Frequency)

$[n_{ij}]$

실제 **관측값**

분할표의 각 칸의 **도수**

각 칸의 결합확률 × n

$n_{ij} = n \cdot \pi_{ij}$

기대도수 (Expected Frequency)

$[\mu_{ij} = n \times \pi_{ij}]$

귀무가설 하에

각 칸의 도수에 대한 **기댓값**

전체 표본 n × 행과 열의 주변확률

$\mu_{ij} = n \cdot \pi_{i+} \cdot \pi_{+j}$

관측도수와 기대도수

관측도수 (Observed Frequency)

$[n_{ij}]$

실제 관측값

분할표의 각 칸의 **도수**

기대도수 (Expected Frequency)

$[\mu_{ij} = n \times \pi_{ij}]$

귀무가설 하에

각 칸의 도수에 대한 **기댓값**

앞선 가설을 관측도수와 기대도수로 표현한다면?

각 칸의 결합확률 × n

$$n_{ij} = n \cdot \pi_{ij}$$

전체 표본 n × 행과 열의 주변확률

$$\mu_{ij} = n \cdot \pi_{i+} \cdot \pi_{+j}$$

관측도수와 기대도수

독립성 검정의 가설 다시 표현하기

관측도수 (Observed Frequency)

$[n_{ij}]$

실제 **관측값**

분할표의 각 칸의 **도수**

기대도수 (Expected Frequency)

$[\mu_{ij} = n \times \pi_{ij}]$

귀무가설 하에

분할표의 도수에 대한 **기댓값**

귀무가설 H_0: 두 범주형 변수는 독립이다

$$\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$$

양변 n 곱함

$$n \cdot \pi_{ij} = \pi_{i+} \cdot \pi_{+j}$$

각 칸의 결합확률 × n

$$n_{ij} = \mu_{ij}$$

전체 표본 n × 행과 열의 주변확률

즉, 귀무가설 하에서 '관측도수 = 기대도수'를 의미함

독립성 검정의 종류

모든 기대도수가 5 이상이면
대표본으로 구분

2차원 분할표 독립성 검정		
대표본	명목형	피어슨 카이제곱 검정 (Pearson's Chi-Squared Test)
		가능도비 검정 (Likelihood-Ratio Test)
	순서형	MH 검정 (Mantel-Haenszel Test)
소표본		피셔의 정확검정 (Fisher's Exact test)

독립성 검정의 종류

모든 기대도수가 5 이상이면
대표본으로 구분

2차원 분할표 독립성 검정		
대표본	명목형	피어슨 카이제곱 검정 (Pearson's Chi-Squared Test)
		가능도비 검정 (Likelihood-Ratio Test)
	순서형	MH 검정 (Mantel-Haenszel Test)
소표본		피셔의 정확검정 (Fisher's Exact test)

독립성 검정

대표본 + 명목형 자료의 독립성 검정

피어슨 카이제곱 검정

$$X^2 = \Sigma \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2_{(I-1)(J-1)}$$

$$X^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$$

가능도 비 검정

$$G^2 = 2\Sigma n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right) \sim \chi^2_{(I-1)(J-1)}$$

$$G^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$$

검정 과정

관측도수와 기대도수 간 차이가 큼 → 검정통계량의 값이 큼 → 귀무가설 기각

→ 두 변수는 독립이 아님 → 변수 간의 연관성 존재

독립성 검정

대표본 + 명목형 자료의 독립성 검정

피어슨 카이제곱 검정

$$X^2 = \Sigma \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2_{(I-1)(J-1)}$$

$$X^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$$

가능도 비 검정

$$G^2 = 2\Sigma n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right) \sim \chi^2_{(I-1)(J-1)}$$

$$G^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$$

검정 과정

관측도수와 기대도수 간 차이가 큼 → **검정통계량**의 값이 큼 → 귀무가설 기각

→ 두 변수는 독립이 아님 → 변수 간의 연관성 존재

대표본 + 명목형 자료의 독립성 검정

(1) 피어슨 카이제곱 검정 (Pearson's Chi-squared Test)

검정 통계량 : $X^2 = \Sigma \frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}} \sim \chi^2_{(I-1)(J-1)}$

기각역 : $X^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$

귀무가설을 기각하지 못한다면

두 변수 간 독립

→ 관측도수와 기대도수 간의 차이가 없음

($n_{ij} = \mu_{ij}$, for all i and j)

→ 검정통계량 $X^2 = 0$ (최솟값)

대표본 + 명목형 자료의 독립성 검정

검정통계량이 카이제곱분포를 따르는 이유

범주형 자료가

포아송 분포를 따른다는 가정

$E(n_{ij}) = V(n_{ij}) = \mu_{ij}$

포아송분포의 **정규 근사**

$\text{Poisson}(\mu) \sim N(\mu, \mu)$

표준정규분포를 따르는 확률변수

제곱의 합이 **카이제곱**을 따른다는 성질

대표본 + 명목형 자료의 독립성 검정

검정통계량이 카이제곱분포를 따르는 이유

범주형 자료가

포아송 분포를 따른다는 가정

포아송분포의 **정규 근사**

$\text{Poisson}(\mu) \sim N(\mu, \mu)$

각 도수(n_{ij})를 표준화

$$\frac{n_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}}} \sim N(0,1)$$

표준정규분포를 따르는 확률변수

제곱의 합이 **카이제곱**을 따른다는 성질

대표본 + 명목형 자료의 독립성 검정

검정통계량이 카이제곱분포를 따르는 이유

범주형 자료가

포아송 분포를 따른다는 가정

포아송분포의 **정규 근사**

$\mathrm{Poisson}(\mu) \sim N(\mu, \mu)$

표준정규분포를 따르는 확률변수

제곱의 합이 **카이제곱**을 따른다는 성질

$$\sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2_{(I-1)(J-1)}$$

대표본 + 명목형 자료의 독립성 검정

검정통계량이 카이제곱분포를 따르는 이유

범주형 자료가

포아송 분포를 따른다는 가정

포아송분포의 **정규 근사**

$\text{Poisson}(\mu) \sim N(\mu, \mu)$

표준정규분포를 따르는 확률변수

제곱의 합이 **카이제곱**을 따른다는 성질

범주형 자료의 검정통계량

$$X^2 = \Sigma \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2_{(I-1)(J-1)}$$

대표본 + 명목형 자료의 독립성 검정

(2) 가능도 비 검정 (Likelihood - Ratio Test)

검정 통계량 : $G^2 = 2\Sigma n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right) \sim \chi^2_{(I-1)(J-1)}$

기각역 : $\mathrm{X}^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$

귀무가설을 기각하지 못한다면

관측도수와 기대도수 간의 차이가 없음 ($n_{ij} = \mu_{ij}$, for all I and J)

→ $\log\left(\frac{n_{ij}}{\mu_{ij}}\right) = \log n_{ij} - \log \mu_{ij} = 0$

→ 검정통계량 $G^2 = 0$

LRT 통계량이 $-2\log\Lambda$인 이유

(Principle of LRT)

대표본 + 명목형 자료의 독립성 검정

(2) 가능도 비 검정 (Likelihood - Ratio Test)

가능도 비 (Likelihood Ratio)

검정 통계량 : $G^2 = 2\sum_{i,j} n_{ij}\log\left(\frac{n_{ij}}{\mu_{ij}}\right) \sim \chi^2_{(I-1)(J-1)}$

$$\Lambda = \frac{max_{\theta\in\Theta_0} L(\theta; x_1, \cdots, x_n)}{max_{\theta\in\Theta} L(\theta; x_1, \cdots, x_n)}$$

기각역 : $X^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$

표본크기 n이 충분히 클 때, 다음의 **점근적 성질**을 만족함

귀무가설을 기각하지 못한다면

$$-2\log\Lambda \approx \chi^2_{v-v_0} \quad (v = \dim(\Theta), v_0 = \dim(\Theta_0))$$

관측도수와 기대도수 간의 차이가 없음 ($n_{ij} = \mu_{ij}$, for all I and j)

$$\log\left(\frac{n_{ij}}{\mu_{ij}}\right) = \log n_{ij} - \log\mu_{ij} = 0$$

검정 시 카이제곱분포 이용 가능!

대표본 + 순서형 자료의 독립성 검정

MH 검정 (Mantel-Haenszel Test)

검정 통계량 : $M^2 = (n-1)r^2 \sim \chi_1^2$

기각역 : $M^2 \geq \chi_{\alpha,1}^2$

각 변수의 수준에 **차등적인 점수를** 할당하여

선형 추세를 측정

행점수 $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_I$

열점수 $v_1 \leq v_2 \leq \cdots \leq v_J$

대표본 + 순서형 자료의 독립성 검정

MH 검정 (Mantel-Haenszel Test)

피어슨 교차적률 상관계수

$$r = \frac{\sum(\mu_i - \bar{\mu})(v_i - \bar{v})P_{ij}}{\sqrt{\sum(\mu_i - \bar{\mu})^2 p_{i+} \cdot \sum(v_i - \bar{v})^2 p_{+j}}}$$

공분산을 두 표준편차의 곱으로 나눈 **상관계수**의 형태

대표본 + 순서형 자료의 독립성 검정

MH 검정 (Mantel-Haenszel Test)

검정 통계량 : $M^2 = (n-1)r^2 \sim \chi_1^2$

기각역 : $M^2 \geq \chi_{\alpha,1}^2$

귀무가설을 기각하지 못한다면

관측도수와 기대도수 간의 차이가 없음 ($n_{ij} = \mu_{ij}$, for all I and J)

→ 상관계수와 마찬가지로 $r = 0$

→ 검정통계량 $M^2 = 0$

대표본 + 순서형 자료의 독립성 검정

MH 검정 (Mantel-Haenszel Test)

검정 통계량 : $M^2 = (n-1)r^2 \sim \chi_1^2$

기각역 : $M^2 \geq \chi_{\alpha,1}^2$

검정 과정

상관계수가 큼 → 검정통계량의 값이 큼 → **귀무가설 기각**

→ 두 변수는 **독립이 아님** → 변수 간의 **연관성 존재**

독립성 검정

대표본 + 순서형 자료의 독립성 검정

MH 검정 (Mantel-Haenszel Test)

독립성 검정은 두 범주형 변수의 연관성 유무만 판단하기 때문에
구체적으로 어떻게 연관이 있는지는 파악할 수 없음

검정 통계량 : $M^2 = (n-1)r^2 \sim \chi_1^2$

기각역 : $M^2 \geq \chi_{\alpha,1}^2$

검정 과정

상관계수가 큼 → 검정통계량의 값이 큼 → 귀무가설 기각

→ 두 변수는 독립이 아님 → 변수 간의 연관성 존재

독립성 검정

대표본 + 순서형 자료의 독립성 검정

MH 검정 (Mantel-Haenszel Test)

독립성 검정은 두 범주형 변수의 연관성 유무만 판단하기 때문에
구체적으로 어떻게 연관이 있는지는 파악할 수 없음

검정 통계량 : $M^2 = (n-1)r^2 \sim \chi_1^2$

기각역 : $M^2 \geq \chi_{\alpha,1}^2$

연관성 측도를 통해 변수 간 연관성의 성질을 파악!

검정 과정

상관계수가 큼 → 검정통계량의 값이 큼 → 귀무가설 기각

→ 두 변수는 독립이 아님 → 변수 간의 연관성 존재

4

연관성 측도

비율의 비교 척도

비율의 차이 (Difference of Proportions)

각 행의 조건부 확률 간 차이

상대위험도 (Relative Risk)

두 집단 간 조건부 확률의 비

오즈비 (Odds Ratio)

각 행 별로 계산된 오즈의 비

연관성 측도

비율의 비교 척도

비율의 차이 (Difference of Proportions)

각 행의 조건부 확률 간 차이

상대위험도 (Relative Risk)

두 집단 간 조건부 확률의 비

오즈비 (Odds Ratio)

각 행 별로 계산된 오즈의 비

두 범주형 변수 모두 2가지 수준만을 갖는 **이항변수**일 때

2×2 분할표에서 **변수 간 연관성**을 파악할 수 있음

비율의 차이 (Difference of Proportions)

π_i : i번째 행의 조건부 확률

비율의 차이

각 행의 조건부 확률 간 차이

$$-1 < \pi_1 - \pi_2 < 1$$

성별	연인유무	
	있음	없음
여성	509 (0.814)	116 (0.186)
남성	398 (0.793)	104 (0.207)

여성이 연인이 있을 조건부확률

$$\frac{509}{509+116} = 0.814$$

남성이 연인이 있을 조건부확률

$$\frac{398}{398+104} = 0.793$$

비율의 차이 (Difference of Proportions)

π_i : i번째 행의 조건부 확률

비율의 차이

각 행의 조건부 확률 간 차이

$$-1 < \pi_1 - \pi_2 < 1$$

성별	연인유무	
	있음	없음
여성	509 (0.814)	116 (0.186)
남성	398 (0.793)	104 (0.207)

비율의 차이

0.814 - 0.793 = 0.021

여성일 때 연인이 있을 확률이

남성일 때보다 약 0.021 높음!

비율의 차이 (Difference of Proportions)

π_i : i번째 행의 조건부 확률

비율의 차이

각 행의 조건부 확률 간 차이

$$-1 < \pi_1 - \pi_2 < 1$$

성별	연인유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

비율의 차이

0.4 – 0.4 = 0

성별과 연인 유무가 독립

비율의 차이 (Difference of Proportions)

π_i : i번째 행의 조건부 확률

비율의 차이

각 행의 조건부 확률 간 차이

$$-1 < \pi_1 - \pi_2 < 1$$

성별	연인유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

비율의 차이

0.4 - 0.4 = 0

$\pi_1 - \pi_2 = 0$ 일 때 두 변수가 **독립**

성별과 연인 유무가 독립

비율의 차이 (Difference of Proportions)

π_i : i번째 행의 조건부 확률

비율의 차이

각 행의 조건부 확률 간 차이

$$-1 < \pi_1 - \pi_2 < 1$$

조건부 확률이 0 혹은 1에 근접했을 경우

두 집단의 영향력 차이가 크지만 이를 반영하지 못함

성별	연인유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

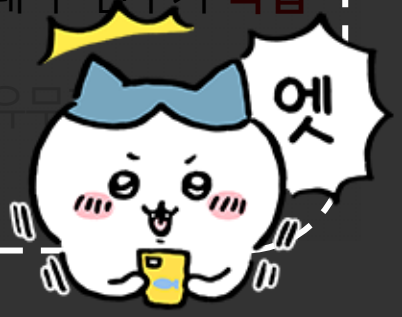
상대위험도 이용!

비율의 차이

$0.4 - 0.4 = 0$

$\pi_1 - \pi_2 = 0$ 일 때 두 변수가 독립

성별과 연인 유무



상대위험도 (Relative Risk)

상대위험도

두 집단 간 조건부 확률의 비

$$\frac{\pi_1}{\pi_2} \geq 0$$

성별	연인유무	
	있음	없음
여성	509 (0.814)	116 (0.186)
남성	398 (0.793)	104 (0.207)

상대위험도

0.814 / 0.793 = 1.027

여성일 때 연인이 있을 확률이

남성일 때보다 약 1.027배 높음!

상대위험도 (Relative Risk)

상대위험도

두 집단 간 조건부 확률의 비

$$\frac{\pi_1}{\pi_2} \geq 0$$

성별	연인유무	
	있음	없음
여성	509 (0.814)	116 (0.186)
남성	398 (0.793)	104 (0.207)

상대위험도

0.814 / 0.793 = 1.027

$\frac{\pi_1}{\pi_2} = 1$ 일 때 두 변수가 **독립**

여성일 때 연인이 있을 확률이

남성일 때보다 약 1.027배 높음!

비율의 차이 vs 상대위험도

성별	연인유무	
	있음	없음
여성	**0.05**	0.95
남성	**0.01**	0.99

각 조건부 확률이 0 혹은 1에 가까운 경우

성별	연인유무	
	있음	없음
여성	**0.55**	0.45
남성	**0.51**	0.49

각 조건부 확률이 0.5에 가까운 경우

비율의 차이

0.05 - 0.01 = 0.04

0.55 - 0.51 = 0.04

상대위험도

0.05 / 0.01 = 5

0.55 / 0.51 = 1.078

비율의 차이 vs 상대위험도

성별	연인유무	
	있음	없음
여성	**0.05**	0.95
남성	**0.01**	0.99

각 조건부 확률이 0 혹은 1에 가까운 경우

성별	연인유무	
	있음	없음
여성	**0.55**	0.45
남성	**0.51**	0.49

각 조건부 확률이 0.5에 가까운 경우

상대위험도는 서로 5배가 차이나는 반면,

비율의 차이는 두 경우 모두 0.04로 연관성이 미미하다는 결과

연관성 측도

비율의 차이 vs 상대위험도

성별	연인유무	
	있음	없음
여성	0.05	0.95
남성	0.01	0.99

각 조건부 확률이 0 혹은 1에 가까운 경우

성별	연인유무	
	있음	없음
여성	0.55	0.45
남성	0.51	0.49

각 조건부 확률이 0.5에 가까운 경우

조건부확률이 0 혹은 1의 값에 가까울 때

비율의 절대적인 차이만을 이용해 연관성을 판단하는 것은 매우 위험!

상대위험도는 상 차이나는 반면,

비율의 차이는 두 경우 모두 0.04로 연관성이 미미하다는 결과

후향적 연구에서의 한계

후향적 연구

이미 나온 결과를 바탕으로 과거 기록을 관찰하는 연구

	심장 질환 O (Y = 1)	심장 질환 X (Y = 0)	합
알코올 중독 O ($X = 1$)	4	2	6
알코올 중독 X ($X = 0$)	46	98	144
합	50	100	150

연구자가 정한 비율 혹은 숫자에 따라 열의 분포가 달라지므로,

조건부 확률을 이용하는 **상대위험도와 비율의 차이 모두에 영향 미침**!

후향적 연구에서의 한계

후향적 연구

이미 나온 결과를 바탕으로 과거 기록을 관찰하는 연구

비율의 차이와 상대위험도 모두
반응변수(Y)를 고정시킨 후향적 연구에서는 사용불가

	심장 질환 O $(Y = 1)$	심장 질환 X $(Y = 0)$	합
알코올 중독 O $(X = 1)$	4	2	6
알코올 중독 X $(X = 0)$	46	98	144
합	50	100	150

오즈비 이용!

연구자가 정한 비율 혹은 숫자에 따라 열의 분포가 달라지므

조건부 확률을 이용하는 **상대위험도와 비율의 차이 모두에 영향 ㅁ**

오즈비 (Odds Ratio)

오즈 (Odds)

성공확률 / 실패확률

$$odds = \frac{\pi}{1-\pi}, \qquad \pi = \frac{odds}{1+odds}$$

반응변수 중 하나를 성공으로 정의한 뒤 성공확률 π를 계산

즉, 오즈는 성공확률이 실패확률의 몇 배인지 나타냄

오즈비 (Odds Ratio)

오즈비

각 행 별로 계산된 오즈의 비

$$\theta = \frac{odds1}{odds2} = \frac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)}$$

성별	연인유무	
	있음	없음
여성	509 (0.814)	116 (0.186)
	0.814/0.186 = 4.388…	
남성	398 (0.793)	104 (0.207)
	0.793/0.207 = 3.826…	

여성이 연인이 있을 오즈: 4.388

남성이 연인이 있을 오즈: 3.826

$$\theta = \frac{4.388}{3.826} = 1.147$$

오즈비 (Odds Ratio)

오즈비

각 행 별로 계산된 오즈의 비

$$\theta = \frac{odds1}{odds2} = \frac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)}$$

성별	연인유무	
	있음	없음
여성	509 (0.814)	116 (0.186)
	0.814/0.186 = 4.388…	
남성	398 (0.793)	104 (0.207)
	0.793/0.207 = 3.826…	

여성이 연인이 있을 **오즈**가

남성의 **오즈**보다 약 1.147**배** 높다

오즈비 (Odds Ratio)

오즈비 값에 따른 의미

$\theta = 1$: 두 행에서 성공의 오즈가 **같다**, 즉 **독립**

$\theta > 1$: 첫 번째 행에서의 성공의 오즈가 두 번째 행보다 **높음**

$0 < \theta < 1$: 첫 번째 행에서의 성공의 오즈가 두번째 행보다 **낮음**

역수 관계에 있는 오즈비는 방향만 반대일 뿐 두 변수 간 연관성의 크기는 같음

오즈비 (Odds Ratio)

오즈비 값에 따른 의미

$\theta = 1$: 두 행에서 성공의 오즈가 **같다**, 즉 **독립**

$\theta > 1$: 첫 번째 행에서의 성공의 오즈가 두 번째 행보다 **높음**

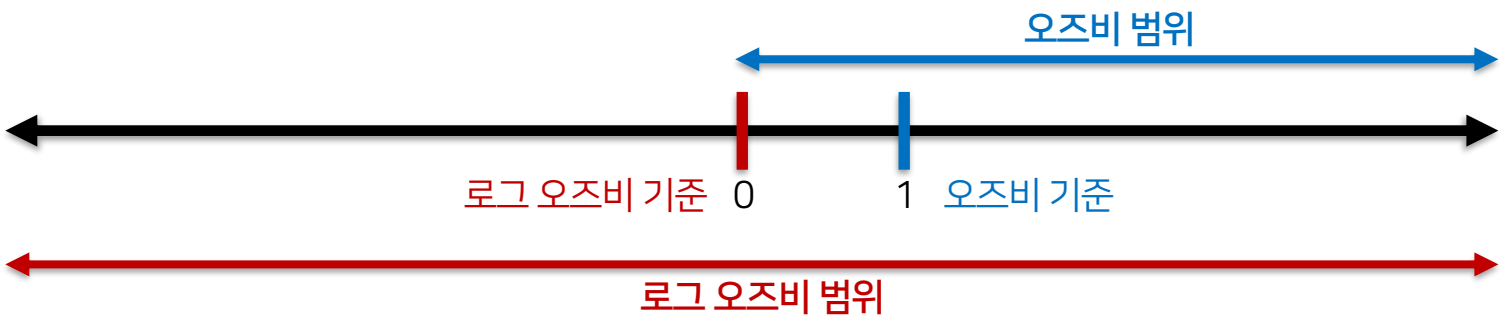
$0 < \theta < 1$: 첫 번째 행에서의 성공의 오즈가 두번째 행보다 **낮음**

역수 관계에 있는 오즈비는 방향만 반대일 뿐 두 변수 간 연관성의 크기는 같음

로그 오즈비 (Log Odds Ratio)

로그 오즈비

오즈비에 **로그**(log)를 씌운 형태



기존의 비대칭적인 오즈비의 범위를 교정해주는 역할

로그를 이용해 기준을 0, 범위를 $(-\infty, \infty)$로 교정

오즈비의 장점

후향적 연구처럼 한 변수가 **고정**되어 있는 경우에도 사용 가능!

알코올 중독	심장 질환		합
	O	X	
O	4 (4/6)	2 (2/6)	6
	4/2		
X	46 (46/144)	98 (98/144)	144
	46/98		
합	50	100	150

건강한 사람이 100명일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

건강한 사람이 300명일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

오즈비의 장점

후향적 연구처럼 한 변수가 **고정**되어 있는 경우에도 사용 가능!

알코올 중독	심장 질환		합
	O	X	
O	4 (4/10)	6(6/10)	10
	4/6		
X	46 (46/144)	294 (46/294)	340
	46/294		
합	50	300	350

건강한 사람이 100명일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

건강한 사람이 300명일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

오즈비의 장점

후향적 연구처럼 한 변수가 **고정**되어 있는 경우에도 사용 가능!

알코올 중독	심장 질환		합
	O	X	
O	4 (4/10)	6(6/10)	10
	4/6		
X	46 (46/144)	294 (46/294)	340
	46/294		
합	50	300	350

건강한 사람이 100명일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

건강한 사람이 300명일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

오즈비의 장점

후향적 연구처럼 한 변수가 **고정**되어 있는 경우에도 사용 가능!

	건강한 사람 100명	건강한 사람 300명	변화
비율의 차이	$\frac{4}{6} - \frac{46}{144} = 0.347$	$\frac{4}{10} - \frac{46}{340} = 0.265$	O
상대위험도	$\frac{4/6}{46/144} = 2.087$	$\frac{4/10}{46/340} = 2.956$	O
오즈비	$\frac{4/2}{46/98} = 4.26$	$\frac{4/6}{46/294} = 4.26$	X

오즈비의 장점

행과 열의 순서가 바뀌어도 같은 값을 지님

알코올 중독	심장 질환		합
	O	X	
O	4 (4/6)	2 (2/6)	6
	4/2		
X	46 (46/144)	98 (98/144)	144
	46/98		
합	50	100	150

알코올 중독 여부가 행일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

심장 질환 여부가 행일 때의 오즈비

$$\frac{4/46}{2/98} = 4.26$$

오즈비의 장점

행과 열의 순서가 바뀌어도 같은 값을 지님

심장 질환	알코올 중독		합
	O	X	
O	4 (4/50)	46 (46/50)	50
	4/46		
X	2 (2/100)	98 (98/100)	150
	2/98		
합	6	144	150

알코올 중독 여부가 행일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

심장 질환 여부가 행일 때의 오즈비

$$\frac{4/46}{2/98} = 4.26$$

오즈비의 장점

행과 열의 순서가 바뀌어도 같은 값을 지님

심장 질환	알코올 중독		합
	O	X	
O	4 (4/50)	46 (46/50)	50
	4/46		
X	2 (2/100)	98 (98/100)	150
	2/98		
합	6	144	150

알코올 중독 여부가 행일 때의 오즈비

$$\frac{4/2}{46/98} = 4.26$$

심장 질환 여부가 행일 때의 오즈비

$$\frac{4/46}{2/98} = 4.26$$

오즈비의 장점

행과 열의 순서가 바뀌어도 같은 값을 지님

$$\frac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)} = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=2)/P(Y=0|X=2)}$$

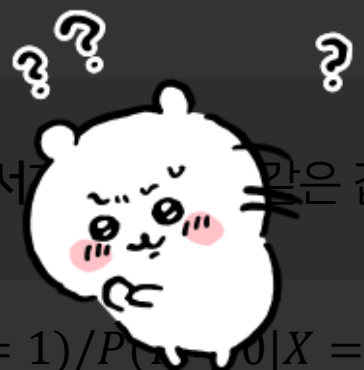
$$= \frac{\frac{P(X=1|Y=1)P(Y=1)}{P(X=1)} / \frac{P(X=1|Y=0)P(Y=0)}{P(X=1)}}{\frac{P(X=2|Y=1)P(Y=1)}{P(X=2)} / \frac{P(X=2|Y=0)P(Y=0)}{P(X=2)}}$$

$$= \frac{P(X=1|Y=1)/P(X=1|Y=0)}{P(X=2|Y=1)/P(X=2|Y=0)}$$

위와 같이 **베이즈 정리**($\mathrm{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$)를 이용해 증명 가능

오즈비의 장점

행과 열의 순서가 바뀌어도 같은 값을 지님



$$\frac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)} = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=2)/P(Y=0|X=2)}$$

오즈비가 다음과 같은 장점을 가질 수 있는 이유?

$$= \frac{\frac{P(X=1|Y=1)P(Y=1)}{P(X=1)} / \frac{P(X=1|Y=0)P(Y=0)}{P(X=1)}}{\frac{P(X=2|Y=1)P(Y=1)}{P(X=2)} / \frac{P(X=2|Y=0)P(Y=0)}{P(X=2)}}$$

교차적비이기 때문!

$$= \frac{P(X=1|Y=1)/P(X=1|Y=0)}{P(X=2|Y=1)/P(X=2|Y=0)}$$

위와 같이 **베이즈 정리**($P(A|B) = \frac{P(B|A)P(A)}{P(B)}$)를 이용해 증명 가능

교차적비 (Cross-product Ratio)

교차적비

분할표에서 **대각선에 위치한 값**끼리 곱한 수 간의 비율

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

대각성분의 곱

비대각성분의 곱

교차적비의 성질에 따라 대조군의 크기가 변하거나

분할표에서 행과 열의 위치가 바뀌더라도 같은 값을 유지!

오즈비와 상대위험도

두 그룹 모두 **성공 확률이 0**에 가까운 경우 **오즈비와 상대위험도는 근사함**

$$\text{오즈비} = \frac{p_1/(1-p_1)}{p_2(1-p_2)} = \text{상대위험도} \times \frac{(1-p_1)}{(1-p_2)}$$

성별	연인 유무	
	O	X
여성	4 (0.02)	196 (0.98)
	0.02/0.98 = 0.0204…	
남성	3 (0.01)	297 (0.99)
	0.01/0.99 = 0.0101…	

$$\frac{0.02}{0.01} \cong \frac{0.0204}{0.0101}$$

상대위험도와 오즈비가 유사함

오즈비와 상대위험도

두 그룹 모두 **성공 확률이 0**에 가까운 경우 **오즈비와 상대위험도는 근사함**

상대위험도를 계산할 수 있는 자료에서는

$$\text{오즈비} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \text{상대위험도} \times \frac{(1-p_1)}{(1-p_2)}$$

복잡한 오즈비보다는 비교적 **간단한 상대위험도** 사용 가능

상대위험도를 계산할 수 없는 자료에서는

상대 위험도를 근사시키는 데 오즈비 사용 가능

성별	연인 유무	
	O	X
여성	4 (0.02)	196 (0.98)
	0.02/0.98 = 0.0	
남성	3 (0.01)	29
	0.01/0.99 = 0.010	

$$\frac{0.02}{0.01} \cong \frac{0.0204}{0.0101}$$

상대위험도와 오즈비가 유사함

3차원 분할표에서의 오즈비

조건부 연관성 (Conditional Association)

제어변수 **Z가 고정**되어 있을 때 X와 Y간의 연관성

제어변수의 각 수준별로 교차적비를 구하면 됨

부분분할표

학과(Z)	성별(X)	통학 여부(Y)		조건부 오즈비
		O	X	
경제	남자	11	25	$\theta_{XY(1)} = 1.188$
	여자	10	27	
통계	남자	14	5	$\theta_{XY(2)} = 4.8$
	여자	7	12	

3차원 분할표에서의 오즈비

동질 연관성 (Homogeneous Association)

제어변수의 각 수준별 **오즈비가 모두 같은** 경우

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$

동질 연관성은 대칭적이므로 X와 Y 간에 동질연관성이 존재한다면 XZ, YZ 간에도 동질 연관성 존재

조건부 독립성 (Conditional Independence)

제어변수의 각 수준별 **오즈비가 모두 1로 같은** 경우

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)} = 1$$

제어변수에 관계없이 X와 Y가 서로 독립인 상태

조건부 독립성은 동질 연관성의 일종으로 더 엄격한 성립조건 가짐

3차원 분할표에서의 오즈비

동질 연관성 (Homogeneous Association)

제어변수의 각 수준별 **오즈비가 모두 같은** 경우

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$

동질 연관성은 대칭적이므로 X와 Y 간에 동질연관성이 존재한다면 XZ, YZ 간에도 동질 연관성 존재

조건부 독립성 (Conditional Independence)

제어변수의 각 수준별 **오즈비가 모두 1로 같은** 경우

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)} = 1$$

제어변수에 관계없이 X와 Y가 서로 독립인 상태

조건부 독립성은 동질 연관성의 일종으로 더 엄격한 성립조건 가짐

3차원 분할표에서의 오즈비

주변 오즈비

제어변수 Z의 모든 수준을 합친 것

주변분할표			
성별(X)	통학 여부(Y)		주변 오즈비
	O	X	
남자	11+14=25	25+5=30	$\theta_{XY+} = 1.911\cdots$
여자	10+7=17	27+12=39	

결국 2차원 분할표에서 오즈비를 구하는 것과 동일하므로 해석도 동일함

주변분할표가 3차원 분할표에서 파생되었기 때문에 별도의 표현을 사용

주변 오즈비가 1일 경우 주변 독립성을 가짐

연관성 측도

3차원 분할표에서의 오즈비

주변 오즈비

제어변수 Z의 모든 수준을 합친 것

주변분할표

성별(X)	통학 여부(Y) O	통학 여부(Y) X	주변 오즈비
남자	11+14=25	25+5=30	$\theta_{XY+} = 1.911\cdots$
여자	10+7=17	27+12=39	

조건부 독립성이 성립한다고 주변 독립성도 성립하는 것은 아님!

심슨의 역설을 통해 확인!

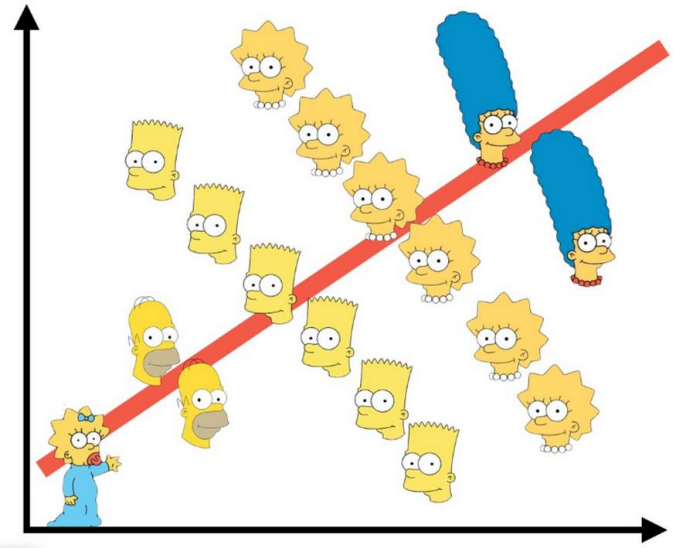
결국 2차원 분할표에서 오즈비를 구하는 것과 동일하므로 해석도 동일함

주변분할표가 3차원 분할표에서 파생되었기 때문에 별도의 표현을 사용

주변 오즈비가 1일 경우 주변 독립성을 가짐

3차원 분할표에서의 오즈비

심슨의 역설 (Simpson's Paradox)



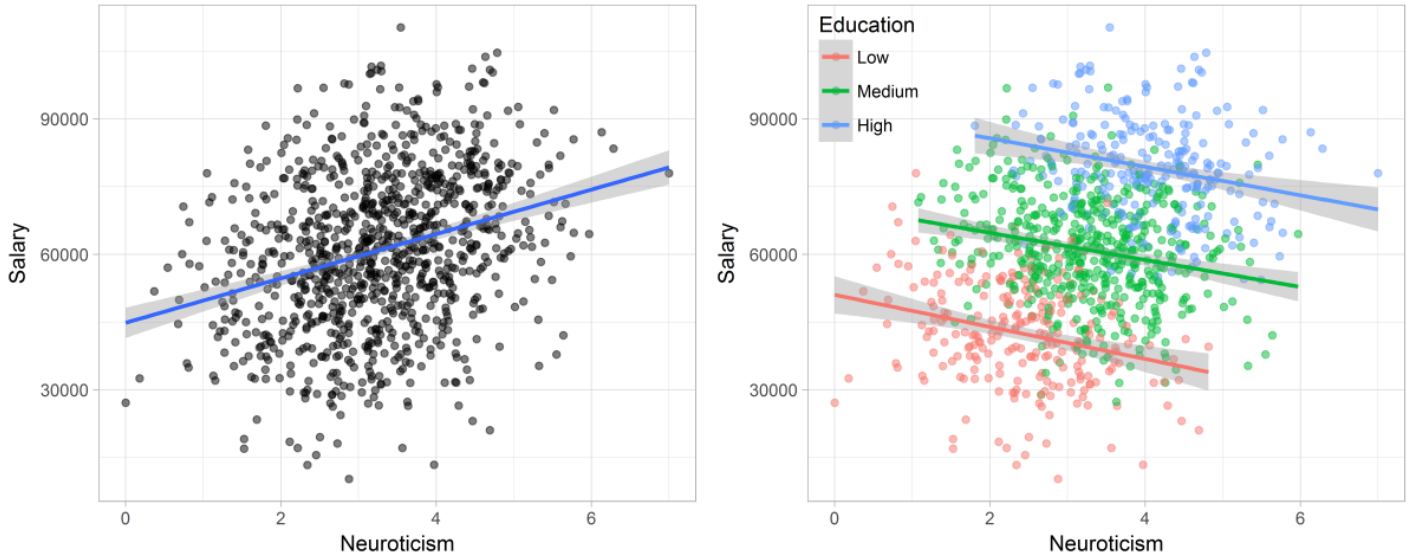
심슨의 역설

전반적인 추세가 경향성을 가지는 것처럼 보이지만

세부 그룹을 나눠서 살펴볼 경우 **앞선 경향성이 사라지거나 반대로 해석되는 경우**

3차원 분할표에서의 오즈비

심슨의 역설 (Simpson's Paradox)



전반적 추세는 **우상향**, 세부 그룹별 추세는 **우하향**하는 추세선을 가짐

즉, **조건부 오즈비와 주변 오즈비의 연관성 방향**이 다르게 나타남

3차원 분할표에서의 오즈비

심슨의 역설 (Simpson's Paradox)

부분분할표

학과 (Z)	성별 (X)	통학 여부(Y)		조건부 오즈비
		O	X	
경제	남자	40	5	$\theta_{XY(1)} = 1.23$
	여자	130	20	
통계	남자	15	5	$\theta_{XY(2)} = 1.2$
	여자	5	2	

주변분할표

성별(X)	통학 여부(Y)		주변 오즈비
	O	X	
남자	55	10	$\theta_{XY+} = 0.90$
여자	135	22	

오즈비는 **1을 기준**으로 하므로 **정반대의 연관성** 방향을 보임

3차원 분할표에서의 오즈비

심슨의 역설 (Simpson's Paradox)

부분분할표

학과 (Z)	성별 (X)	통학 여부(Y) O	통학 여부(Y) X	주변 오즈비
경제	남자	10	5	$\theta_{XY(1)} = 1.23$
	여자	130	20	
통계	남자	15	5	$\theta_{XY(2)} = 1.2$
	여자	5	2	

주변분할표

성별(X)	통학 여부(Y) O	통학 여부(Y) X	주변 오즈비
남자	145	25	$\theta_{XY+} = 0.90$
여자	135	22	

심슨의 역설이 발생한 구체적 원인

앞선 예시의 경우 경제학과는 총 195명인 반면, 통계학과는 총 27명

제어변수인 학과에 따라 도수의 크기가 크게 차이남

오즈비는 **1을 기준**으로 하므로 **정반대의 연관성** 방향을 보임

3차원 분할표에서의 오즈비

심슨의 역설 (Simpson's Paradox)

부분분할표

학과 (Z)	성별 (X)	통학 여부(Y) O	통학 여부(Y) X	조건부 오즈비
경제	남자	10	5	$\theta_{XY(1)} = 1.23$
	여자	130	20	
통계	남자	15	5	$\theta_{XY(2)} = 1.2$
	여자	5	2	

주변분할표

성별(X)	통학 여부(Y) O	통학 여부(Y) X	주변 오즈비
남자	145	25	$\theta_{XY+} = 0.90$
여자	135	22	

심슨의 역설이 발생한 구체적 원인

앞선 예시의 경우 경제학과는 총 195명인 반면, 통계학과는 총 27명

제어변수인 학과에 따라 도수의 크기가 크게 차이남

이로 인한 영향력 차이로 심슨의 역설이 발생하는 것!

오즈비는 **1을 기준**으로 하므로 **정반대의 연관성** 방향을 보임

THANK YOU

소표본 + 명목형 자료의 독립성 검정

피셔의 정확검정 (Fisher's Exact test)

	RH-	RH+	합계
여성	1	481	482
남성	5	513	518
합계	6	994	1000

분할표에서 모든 주변합들이 주어져 있음

위 도표에서 '여성이면서 RH-인 도수 n_{11}'만 주어져도

나머지 세 칸의 도수를 알 수 있음

→ 각 도수의 분포가 **초기하분포**를 따름

소표본 + 명목형 자료의 독립성 검정

피셔의 정확검정 (Fisher's Exact test)

여성, RH-인 수	0	1	2	3	4	5	6
초기하 분포	0.0191	0.1074	0.2512	0.2174	0.3122	0.0804	0.0123

도수가 1인 경우의 초기하분포 확률(0.1074)보다

확률이 더 작은 경우를 모두 더해 p-value를 계산!

검정 과정

p-value가 작음 → **귀무가설 기각**

→ 두 변수는 **독립이 아님** → 변수 간의 **연관성 존재**

3차원 분할표 독립성 검정

CMH test (Cochran-Mantel-Haenszel test)

귀무가설 H_0 : X와 Y는 제어변수의 모든 수준에서 조건부 독립이다 ($\theta = 1$)

검정 통계량 : $X^2_{CMH} = \frac{[\sum_{k=1}^{r}(n_{11k}-\mu_{11k})]^2}{\sum_{k=1}^{r} Var(n_{11k})} \sim \chi^2_1$

기각역 : $X^2 \geq \chi^2_{\alpha,1}$

3차원 분할표가 2×2×K 형태일 때 사용

BD 검정에서 **X와 Y가 동질 연관성**을 가짐을 확인한 후 시행

귀무가설을 기각하면 **공통오즈비**를 통해 X와 Y 간의 관계를 해석

3차원 분할표 독립성 검정

CMH test (Cochran-Mantel-Haenszel test)

부분분할표				
연령대(Z)	비만(X)	당뇨(Y)		조건부 오즈비
		O	X	
50대 이하	O	10	90	$\theta_{XY(1)}$ =1.476
	X	35	465	
50대 이상	O	36	164	$\theta_{XY(2)}$ =1.53
	X	25	175	

주변분할표			
비만(X)	당뇨(Y)		주변 오즈비
	O	X	
O	46	254	$\theta_{XY(+)}$ =1.93
X	60	640	

각 수준별 조건부 오즈비와 달리 주변 오즈비는 값이 과장되어 있음

→ **공통 오즈비** 이용!

3차원 분할표 독립성 검정

CMH test (Cochran-Mantel-Haenszel test)

공통 오즈비: $\widehat{OR_{CMH}} = \frac{\sum \frac{a_i d_i}{n_i}}{\sum \frac{b_i c_i}{n_i}}$

(a_i, b_i, c_i, d_i는 2×2 분할표의 도수)

검정 과정

BD test의 귀무가설(X, Y는 동질 연관성) 채택 → CMH test

→ CMH test 귀무가설(X, Y는 조건부 독립) 기각 → **공통 오즈비 계산**