범주형자료분석팀

3팀

이정민
이경미
이현진
윤예빈
김준영

CONTENTS

1

범주형 자료분석

범주형 자료분석

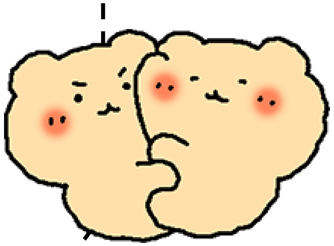
범주형 자료분석 *(Categorical Data Analysis, CDA)*

반응변수가 범주형인 자료에 대한 분석

변수 : 모집단의 특징을 나타내는 것

자료 : 관측치들의 집합

관측치 : 측정한 값



1 범주형 자료분석

변수의 구분

X 변수

독립변수

설명변수

Y 변수

종속변수

반응변수

반응변수가 범주형인 자료를 분석 = Y변수가 범주형인 자료를 분석

1 범주형 자료분석

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

1 범주형 자료분석

자료의 형태

관측값이 **수치로 측정**되는 자료

<table>
<tr><td rowspan="4">자료</td><td rowspan="2">양적 (Quantitative) 자료</td><td>이산형 (Discrete) 자료</td></tr>
<tr><td>연속형(Continuous) 자료</td></tr>
<tr><td rowspan="2">질적 (Qualitative) 자료</td><td>명목형 (Nominal) 자료</td></tr>
<tr><td>순서형 (Ordinal) 자료</td></tr>
</table>

이산형 자료

값을 셀 수 있는 자료

Ex) 나이, 신생아 수...

연속형 자료

연속인 어떤 구간에서

값을 취하는 자료

Ex) 키, 몸무게...

1 범주형 자료분석

자료의 형태

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

특징

- (1) 공분산, 상관계수 등의 수리적인 계산 가능
- (2) 정규분포의 가정이 있다면, 회귀분석 가능

1 범주형 자료분석

자료의 형태

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

관측 결과가 여러 개의 **범주의 집합**으로 나타나는 자료

명목형 자료

범주 간에 순서의 의미가 **없는** 자료

Ex) 성별, 혈액형...

순서형 자료

범주 간에 순서의 의미가 **있는** 자료

Ex) 선호도 (좋음/보통/싫음)

1 범주형 자료분석

자료의 형태

특징

- (1) **순서형 자료에 명목형 자료분석 방법을 적용**할 수 있음
- (2) **분할표를 작성**할 수 있음
- (3) 각 범주에 **특정 점수를 할당하여 양적자료로 활용**할 수 있음
- (4) 일반적으로 **사칙 연산 불가능**

1 범주형 자료분석

자료의 형태

(1) 순서형 자료에 명목형 자료분석 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

자료의 형태

(2) 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_{++}

자료의 형태

(3) 각 범주에 특정 점수를 할당하여 양적자료로 활용할 수 있음

코드값	코드값 의미	서열
001	대통령	1
002	부통령	2
…	…	…
019	9급	31
020	기능 1급	128

코드값이 변수의 범주를 의미
→ **범주형 변수**로 생각할 수 있음

코드값의 크기가 서열을 의미
→ **연속형 변수**로 생각할 수 있음

1 범주형 자료분석

자료의 형태

올바른 범주형 자료 구분의 필요성

(3) 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

↓

분석해야 하는 자료를 **범주형 변수로 간주할지,**

수치형 자료로 간주할지 판단 후 자료통계분석 수행 !

코드값	코드값 의미	서열
001	대통령	1
002	부통령	2
…	…	…
019	[illegible]	[illegible]
020	기능 1급	128

코드값이 단지 범주를 의미

→ **범주형 변수**로 생각할 수 있음

코드값의 크기가 서열을 의미

→ **연속형 변수**로 생각할 수 있음

자료의 형태

(4) 일반적으로 사칙연산이 불가능

범주형 변수에 대한 분석은 통계량보다는

범주별 빈도에 관심을 가지고,

이를 일반화할 때는 **특정 범주가 발생할 확률**에 관심을 가짐

2

분할표

분할표

분할표 *(Contingency Table)*

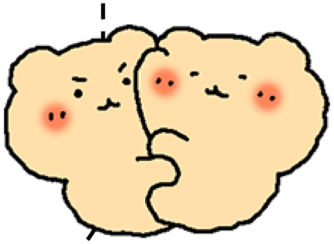
각 범주형 변수에 대한 **결과의 도수(frequency)**를 각 칸에 정리한 표

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

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Y</td></tr>
<tr><td>1</td><td>…</td><td>J</td></tr>
<tr><td rowspan="3">X</td><td>1</td><td colspan="3" rowspan="3">I *J 개 칸</td></tr>
<tr><td>…</td></tr>
<tr><td>I</td></tr>
</table>

수준(Level): 각 변수의 **카테고리 개수**

X 변수는 I개의 수준, Y 변수는 J개의 수준을 갖고 있음

I×J 크기의 행렬

분할표

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Y</td></tr>
<tr><td>1</td><td>…</td><td>J</td></tr>
<tr><td rowspan="3">X</td><td>1</td><td colspan="3" rowspan="3">I *J 개 칸</td></tr>
<tr><td>…</td></tr>
<tr><td>I</td></tr>
</table>

범주형 자료를 분할표로 표현하는 이유

(1) **예측 검정력**에 대한 요약 가능

(2) **독립성 검정** 실시 가능

여러 차원의 분할표

수준에 따라 무한가지 형태의 분할표를 만들 수 있음

BUT! 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

수준에 따라 무한가지 형태의 분할표를 만들 수 있음

BUT! 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!

2차원 분할표 (I x J)

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

2차원 분할표 (I x J)

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

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

3차원 분할표 (I x J x K)

3차원 분할표

기존의 설명변수와 반응변수에 K개의 수준을 가진

제어변수(제한변수, Control Variable) Z가 추가된 형태

부분분할표

Z변수(제어변수)의

수준에 따라 X변수와

Y변수가 분류된 분할표

주변분할표

Z변수(제어변수)의

모든 수준을 결합하여 만든

2차원 분할표

3차원 분할표 (I x J x K)

3차원 분할표

기존의 설명변수와 반응변수에 K개의 수준을 가진

제어변수(제한변수, Control Variable) Z가 추가 된 형태

부분분할표

Z변수(제어변수)의

수준에 따라 X변수와

Y변수가 분류된 분할표

주변분할표

Z변수(제어변수)의

모든 수준을 결합하여 만든

2차원 분할표

3차원 분할표 (I x J x K)

부분분할표 *(Partial Table)*

고정된 Z변수의 각 수준에서 반응변수(Y)에 미치는
설명변수(X)의 효과 확인 가능

		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_{++2}

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

3차원 분할표 (I x J x K)

주변분할표 *(Marginal Table)*

X변수와 Y변수 간의 관계에서 **Z변수의 영향을 무시**한 형태

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

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

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)

비율에 대한 분할표

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

결합확률 *(Joint Probability)*

모집단에서부터 임의로 추출된 표본이 X 변수의 I번째 수준과

Y 변수의 J번째 수준에 **동시에 속할 확률**

$\pi_{ij} = P(X = i,\ \ Y = j)$, 위 표에서 각 칸의 확률

항상 $\sum \pi_{ij} = 1$ 을 만족!

2 분할표

비율에 대한 분할표

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

주변확률 (Marginal Probability)

결합분포의 **행과 열의 합**

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

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

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

비율에 대한 분할표

조건부 확률 *(Conditional Probability)*

X가 주어졌을 때 Y에 대한 확률

즉 **X 변수의 각 수준에서의 Y 변수의 값**

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

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 추출된 표본들이
하나의 특성에 대해 동일한 분포를 가지는지 검정

독립성 검정

두 범주형 변수가 통계적으로 관계가 있는지 확인하기 위한 검정

독립성 검정의 목적

두 변수 간 **연관성 유무** 확인

분석 가치 판단

↓

두 범주형 변수가 **통계적으로 관계가 있는지**를 확인

독립성 검정의 목적

두 변수 간 **연관성 유무** 확인

분석 가치 판단

결과가 **독립**이라면 두 변수에 대해 더 이상 분석을 진행할 필요가 없음

3 독립성 검정

독립성 검정의 가설

귀무가설 H_0 : 두 범주형 변수는 독립이다. $(\pi_{ij} = \pi_{i+} \cdot \pi_{+j})$

대립가설 H_1 : 두 범주형 변수는 독립이 아니다. $(\pi_{ij} \neq \pi_{i+} \cdot \pi_{+j})$

통계학에서 **X와 Y가 서로 독립**이라는 것은 P(Y|X) = P(Y),
즉 P(X∩Y) = P(X)*P(Y)가 **성립함**을 의미함

분할표상에서 두 변수가 독립이라는 것은
모든 결합확률이 행과 열 주변확률의 곱과 동일하다는 의미

3 독립성 검정

독립성 검정의 가설

귀무가설 H_0 : 두 범주형 변수는 독립이다. $(\pi_{ij} = \pi_{i+} \cdot \pi_{+j})$

대립가설 H_1 : 두 범주형 변수는 독립이 아니다. $(\pi_{ij} \neq \pi_{i+} \cdot \pi_{+j})$

통계학에서 **X와 Y가 서로 독립**이라는 것은 P(Y|X) = P(Y),
즉 P(X∩Y) = P(X)*P(Y)가 **성립함**을 의미함

분할표상에서 **두 변수가 독립**이라는 것은
모든 **결합확률이 행과 열 주변확률의 곱과 동일**하다는 의미

독립성 검정의 가설

귀무가설 H_0 : 두 범주형 변수는 독립이다. $(\pi_{ij} = \pi_{i+} \cdot \pi_{+j})$

대립가설 H_1 : 두 범주형 변수는 독립이 아니다. $(\pi_{ij} \neq \pi_{i+} \cdot \pi_{+j})$

통계학에서 **X와 Y가 서로 독립**이라는 것은 P(Y|X) = P(Y),
즉 P(X∩Y) = P(X)*P(Y)가 **성립함**을 의미함

분할표상에서 **두 변수가 독립**이라는 것은
모든 **결합확률이 행과 열 주변확률의 곱과 동일**하다는 의미

관측도수와 기대도수

관측도수 (Observed Frequency)	기대도수 (Expected Frequency)
실제 **관측값** 분할표의 각 칸의 **도수**	귀무가설 하에 각 칸의 도수에 대한 **기댓값**
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

관측도수 (Observed Frequency)	기대도수 (Expected Frequency)
실 … 측값 분할표의 각 칸의 **도수**	… 하에 … 각 칸의 도수에 대한 **기댓값**
각 칸의 결합확률 × n $n_{ij} = n \cdot \pi_{ij}$	전체 표본 n × 행과 열의 주변확률 $\mu_{ij} = n \cdot \pi_{i+} \cdot \pi_{+j}$

앞선 가설을 **기대도수와 관측도수**를 이용하여

다시 표현한다면?

3 독립성 검정

관측도수와 기대도수

독립성 검정의 가설 다시 표현하기

귀무가설 H_0 : 두 범주형 변수는 **독립**이다.

$$\left(\pi_{ij} = \pi_{i+} \times \pi_{+j}\right)$$

$$\left(n \times \pi_{ij} = n \times \pi_{i+}\pi_{+j}\right)$$

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

즉, 귀무가설 하에서 **기대도수 = 관측도수**와 같은 의미

관측도수 (Observed Frequency)	기대도수 (Expected Frequency)
실제 **관측값** 분할표의 각 칸의 **도수**	귀무가설 하에 각 칸의 도수에 대한 **기댓값**
각 칸의 결합확률 × n $n_{ij} = n \cdot \pi_{ij}$	전체 표본 n × 행과 열의 주변확률 $\mu_{ij} = n \cdot \pi_{i+} \cdot \pi_{+j}$

독립성 검정의 종류

기대도수가 5 이상이면 대표본으로 구분

<table>
<tr><th colspan="3">2차원 분할표 독립성 검정</th></tr>
<tr><td rowspan="3">대표본</td><td rowspan="2">명목형</td><td>피어슨 카이제곱 검정 (Pearson's chi-squared test)</td></tr>
<tr><td>가능도비 검정 (Likelihood-ratio test)</td></tr>
<tr><td>순서형</td><td>MH 검정 (Mantel-Haenszel test)</td></tr>
<tr><td colspan="2">소표본</td><td>피셔의 정확검정 (Fisher's Exact test)</td></tr>
</table>

3 독립성 검정

대표본 + 명목형 자료 독립성 검정

피어슨 카이제곱 검정

검정통계량

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

기각역

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

검정 과정

관측도수 & 기대도수의 차이가 크다 → 검정통계량이 크다 → p-value가 작다 → 귀무가설 기각 → 변수 간 연관성이 존재

대표본 + 명목형 자료 독립성 검정

피어슨 카이제곱 검정

검정통계량

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

기각역

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

검정 과정

관측도수 & 기대도수의 **차이가 크다** → **검정통계량이 크다** → p-value가 작다 → **귀무가설 기각** → 변수 간 **연관성이 존재**

대표본 + 명목형 자료 독립성 검정

피어슨 카이제곱 검정

검정통계량

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

기각역

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

귀무가설 하에서는 $\mu_{ij} = n_{ij}$ 이므로 **검정통계량** X^2은 **최솟값인 0**이 됨

하지만 n_{ij}**와** μ_{ij}**의 차이가 커질수록** 검정통계량이 커져 **기각**됨

범주형 자료의 통계적 검정

검정통계량이 카이제곱분포를 따르는 이유

범주형 자료에 대한 **포아송분포** 가정

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

포아송분포의 정규근사

$$Poisson(\mu) \sim N(\mu, \mu)$$

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

표준정규분포의 합이 카이제곱분포를 따르는 성질

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

범주형 자료의 통계적 검정

검정통계량이 카이제곱분포를 따르는 이유

범주형 자료에 대한 포아송분포 가정

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

포아송분포의 **정규근사**

$$Poisson(\mu) \sim N(\mu, \mu)$$

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

표준정규분포의 합이 카이제곱분포를 따르는 성질

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

범주형 자료의 통계적 검정

검정통계량이 카이제곱분포를 따르는 이유

범주형 자료에 대한 포아송분포 가정

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

포아송분포의 정규근사

$Poisson(\mu) \sim N(\mu, \mu)$

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

표준정규분포의 합이 **카이제곱분포를** 따르는 성질

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

범주형 자료의 통계적 검정

검정통계량이 카이제곱분포를 따르는 이유

범주형 자료에 대한 포아송분포 가정

포아송분포의 정규근사 $Poisson(\mu) \sim N(\mu, \mu)$

표준정규분포의 합이 **카이제곱분포**를 따르는 성질

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

3 독립성 검정

대표본 + 명목형 자료 독립성 검정

가능도비 검정

검정통계량

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

기각역

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

귀무가설 하에서는 $\mu_{ij} = n_{ij}$ 이므로 $\log\frac{n_{ij}}{\mu_{ij}} = \log n_{ij} - \log\mu_{ij} = 0$

따라서, **검정통계량** $G^2 = 0$

대표본 + 명목형 자료 독립성 검정

가능도[illegible]검정

검정통계량

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

기각역

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

카이제곱 검정과 가능도비 검정 흐름

두 검정은 같은 검정 흐름을 가짐!

관측도수와 기대도수의 **차이가 크다** → **검정통계량이 크다** → p-value가 작다 → **귀무가설 기각** → 변수 간 **연관성이 존재**

귀무가설 하에서는 $\mu_{ij} = n_{ij}$ 이므로 $\log\frac{n_{ij}}{\mu_{ij}} = \log n_{ij} - \log\mu_{ij} = 0$

따라서, **검정통계량** $G^2 = 0$

대표본 + 순서형 자료 독립성 검정

MH 검정

검정통계량

$$M^2 = (n-1)r^2 \sim {\chi_1}^2$$

기각역

$$M^2 \geq {\chi^2}_{\alpha,1}$$

각 변수의 수준에 차등적인 점수 할당

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

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

각 수준 간 점수 차이는 동등할 필요 없음

대표본 + 순서형 자료 독립성 검정

피어슨 교차적률 상관계수

검정통계량에서의 r

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

공분산을 두 표준편차의 곱으로 나눈 상관계수의 형태

MH 검정

검정통계량

$$M^2 = (n-1)r^2 \sim \chi_1{}^2$$

기각역

$$M^2 \geq \chi^2_{\alpha,1}$$

각 변수의 수준에 차등적인 점수 할당

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

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

각 수준 간 점수 차이는 동등할 필요 없음

대표본 + 순서형 자료 독립성 검정

MH 검정

검정통계량

$$M^2 = (n-1)r^2 \sim {\chi_1}^2$$

기각역

$$M^2 \geq {\chi^2}_{\alpha,1}$$

귀무가설 하에서는 $\mu_{ij} = n_{ij}$ 이므로 $\mathrm{r} = 0$

따라서, 검정통계량 $M^2 = 0$

대표본 + 순서형 자료 독립성 검정

MH 검정

검정통계량

$$M^2 = (n-1)r^2 \sim {\chi_1}^2$$

기각역

$$M^2 \geq {\chi^2}_{\alpha,1}$$

검정 과정

상관계수 |r|이 크다 → 검정통계량 M^2가 크다 → p-value가 작다 → **귀무가설 기각** → 변수 간 **연관성이 존재**

소표본 독립성 검정

피셔의 정확검정 *(Fisher's Exact test)*

하나 이상의 기대도수가 5이하인 경우 사용

카이제곱 검정법과 달리 p-value를 이산적으로 정확히 구함

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

분할표에서 행과 열의 **주변합들이 고정**되어 있을 때,

각 도수의 분포는 **초기하분포**를 따름

$$P(n_{11} = 1) = \frac{\binom{482}{1}\binom{518}{5}}{\binom{1000}{6}} = \text{약 } 0.1074$$

소표본 독립성 검정

피셔의 정확검정 *(Fisher's Exact test)*

하나 이상의 기대도수가 5이하인 경우 사용

카이제곱 검정법과 달리 p-value를 이산적으로 정확히 구함

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

열의 주변합을 기준으로 계산한

이 확률분포는 미지의 모수를

포함하고 있지 않기 때문에,

p-value를 정확하게 구할 수 있음!

소표본 독립성 검정

피셔의 정확검정에서 *P-value* 의 의미

성별과 RH 혈액형 사이에는 관련성이 없다(H_0)는 진실 하에서, **여성일수록 RH-일 확률이 높아지거나 낮아지는 관련성이 있다(H_1)고 결론**을 내릴 확률

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

즉, 현재 관측된 결과보다 **대립가설을 더 지지하는 결과들에 대한 초기하분포의 확률** 합

여기서는 **도수가 1인 경우와 같거나 극단적인 경우**를 모두 더한 **0.2192**

소표본 독립성 검정

피셔의 정확검정에서 *P-value* 의 의미

성별과 RH 혈액형 사이에는 관련성이 없다(H_0)는 진실 하에서, **여성일수록 RH-일 확률이 높아지거나 낮아지는 관련성이 있다(H_1)고 결론**을 내릴 확률

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

즉, 현재 관측된 결과보다 **대립가설을 더 지지하는 결과들에 대한 초기하분포의 확률** 합

여기서는 **도수가 1인 경우와 같거나 극단적인 경우**를 모두 더한 **0.2192**

소표본 독립성 검정

분석 흐름 정리

성별과 RH 혈액형 사이에 관련성이 없다고 가정

이 때, RH- 6명 중 1명이 여성일 확률은 0.1074이고,

이 확률보다 일어나기 힘든 경우(0.1.5.6)의 합은 0.2192

유의성 기준이 0.3일 때, 이는 기준을 넘지 못하므로

현실적으로 일어나기 어려운 현상

따라서, **여성과 RH 혈액형 사이에 연관성이 있다**고 결론

독립성 검정의 한계

독립성 검정은 **두 범주형 변수의 연관성 유무만 판단**할 뿐,

구체적으로 어떻게 연관이 있는지는 알 수 없음

(검정통계량이 크다 ≠ 두 변수 간 연관성이 크다)

독립성 검정의 한계

변수 간 연관성의 성질을 파악하기 위해 **연관성 척도**를 알아야 함

독립성 검정은 두 범주ㅎ ... 연관성 유무만 판단할 뿐,

구체적으로 어떻 ... 는지는 알 수 없음

(검정통계량이 크 ... 간 연관성이 크다)

4

연관성 측도

비율의 비교 척도

두 범주형 변수가 모두 2가지 수준만을 갖는 **이항변수**일 때

아래 표의 3가지 척도를 통해 2×2 분할표에서 **변수 간 연관성**을 파악할 수 있음

비율은 각 행에 따른 조건부확률!

비율의 비교 척도		
비율의 차이	상대 위험도	**오즈비**

비율의 비교 척도

비율의 차이 *(Difference of Proportions)*

각 행의 조건부 확률 간 차이

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

여성이 연인이 있을 조건부 확률

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

남성이 연인이 있을 조건부 확률

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

비율의 비교 척도

비율의 차이 *(Difference of Proportions)*

각 행의 조건부 확률 간 차이

성별	연인 유무	
	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 비교 척도

비율의 차이 *(Difference of Proportions)*

각 행의 조건부 확률 간 차이

성별	연인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

비율의 차이 : 0.4 - 0.4 = 0

성별과 연인 유무라는
두 변수가 **독립**임을 알 수 있음

비율의 비교 척도

상대위험도 *(Relative Risk)*

두 집단 간 조건부 확률의 비

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

상대위험도 : $\frac{0.814}{0.793} = 1.027$

여성일 때 연인이 있을 확률이

남성일 때보다 약 **1.027배** 높음

비율의 비교 척도

상대위험도 *(Relative Risk)*

두 집단 간 조건부 확률의 비

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

상대위험도 : $\frac{0.814}{0.793} = 1.027$

여성일 때 연인이 있을 확률이 남성일 때보다 약 1.027배 높음

상대위험도 값($\frac{\pi_1}{\pi_2}$)은 ≥ 0이고

$\frac{\pi_1}{\pi_2} = 1$일 때, 두 변수가 **독립의 관계**라고 해석할 수 있음

4 연관성 척도

비율의 비교 척도

비율의 차이 vs 상대위험도

성별	연인 유무	
	있음	없음
여성	0.05	0.95
남성	0.01	0.99

성별	연인 유무	
	있음	없음
여성	0.55	0.45
남성	0.51	0.49

조건부 확률이 0 혹은 1에 가까운 경우와 0.5에 가까운 경우의

비율의 차이와 **상대위험도** 비교

4 연관성 측도

비율의 비교 척도

비율의 차이 vs 상대위험도

성별	연인 유무	
	있음	없음
여성	0.05	0.95
남성	0.01	0.99

성별	연인 유무	
	있음	없음
여성	0.55	0.45
남성	0.51	0.49

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 비교 척도

비율의 차이 vs 상대위험도

성별	연인 유무	
	있음	없음
여성	0.05	0.95
남성	0.01	0.99

성별	연인 유무	
	있음	없음
여성	0.55	0.45
남성	0.51	0.49

비율의 차이는 모두 0.04로 동일하지만

상대위험도는 5와 약 1.078로 차이가 큼

따라서 조건부 확률이 **0 혹은 1**에 가까운 경우에는

비율의 차이만을 이용해 연관성을 판단하는 것은 매우 위험!

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

후향적 연구에서의 한계

후향적 연구

이미 나온 결과를 바탕으로 과거 기록을 관찰하는 연구

비율의 차이와 **상대위험도**는 두 변수 간 연관성을 파악하는
직관적인 척도이나,
후향적 연구처럼 **반응변수(Y)를 고정**시킨 연구에서는
사용할 수 없다는 한계를 지님

후향적 연구에서의 한계

후향적 연구

이미 나온 결과를 바탕으로 과거 기록을 관찰하는 연구

비율의 차이와 **상대위험도**는 두 변수 간 연관성을 파악하는 직관적인 척도이나,
후향적 연구처럼 **반응변수(Y)를 고정**시킨 연구에서는 사용할 수 없다는 한계를 지님

후향적 연구에서의 한계

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

연구자가 대조군 (Y=0)의 합을 변경한다면

그에 따른 비율의 차이와 상대위험도도 달라지기 때문에

후향적 연구에서는 비율의 차이와 상대위험도를 사용하기 어려움

4 연관성 측도

후향적 연구에서의 한계

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

연구자가 대조군 [illegible] 합을 변경한다면
그에 따른 비율의 차이와 상대위험도도 달라지기 때문에
후향적 연구에서는 비율의 차이와 상대위험도를 사용하기 어려움

오즈비를 통해 해결 가능!

오즈비 *(Odds Ratio)*

오즈 *(Odds)*

성공확률 / 실패확률

π 를 어떤 사건에서의 성공확률이라고 정의한다면,

오즈는 다음과 같이 표현

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

오즈비 *(Odds Ratio)*

<table>
<tr><td rowspan="2">성별</td><td colspan="2">연인 유무</td></tr>
<tr><td>있음</td><td>없음</td></tr>
<tr><td rowspan="2">여성</td><td>509 (0.814)</td><td>116 (0.186)</td></tr>
<tr><td colspan="2">0.814/0.186 = 4.388…</td></tr>
<tr><td rowspan="2">남성</td><td>398 (0.793)</td><td>104 (0.207)</td></tr>
<tr><td colspan="2">0.793/0.207 = 3.826…</td></tr>
</table>

여성이 연인이 있을 오즈 : **약 4.388**

남성이 연인이 있을 오즈 : **약 3.826**

오즈는 **성공확률**이 **실패확률**의 몇 배인지 나타내며

오즈비는 이렇게 계산된 오즈의 비를 의미

오즈비 *(Odds Ratio)*

성별	연인 유무	
	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⋯	

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

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

즉, 여성이 연인이 있을 오즈가 남성의 오즈보다 **약 1.147배 높음**

4 연관성 측도

오즈비 *(Odds Ratio)*

오즈비 값에 따른 의미

$\theta = 1$: 두 행에서 성공의 오즈가 같음, **독립!**

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

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

서로 **역수 관계**에 있는 오즈비는

방향만 반대이고 두 변수간 **연관성의 정도**는 같음

오즈비 *(Odds Ratio)*

오즈비 값에 따른 의미

$\theta = 1$: 두 행에서 성공의 오즈가 같음, **독립!**

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

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

서로 **역수 관계**에 있는 오즈비는

방향만 반대이고 두 변수간 **연관성의 정도는 같음**

오즈비 *(Odds Ratio)*

오즈비 값에 따른 의미

$\theta = 1$: 두 행에서 성공의 오즈가 같음, **독립!**

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

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

그러나, 기존 오즈비는 분자의 오즈가 더 큰 경우(1 ~ ∞)와

분모의 오즈가 더 큰 경우(0 ~ 1)가 서로 **비대칭**인 범위를 가지고 있음

서로 **역수 관계**에 있는 오즈비는

방향만 반대이고 두 변수간 **연관성의 정도**는 같음

오즈비 *(Odds Ratio)*

오즈비 값에 따른 의미

$\theta = 1$: 두 행에서 성공의 오즈가 같음, 독립!

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

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

그러나, 기존 오즈비는 분자의 오즈가 더 큰 경우(1 ~ ∞)와

분모의 오즈가 더 큰 경우(0 ~ 1)가 서로 **비대칭**인 범위를 가지고 있음

서로 **역수 관계**에 있는 오즈비는

방향만 반대이고 두 변수간 **연관성의 정도**는 같음

로그 오즈비를 통해 해결!

4 연관성 측도

로그 오즈비 *(Log Odds Ratio)*

로그 오즈비

오즈비에 **log**를 씌운 형태

로그 오즈비는 0을 기준으로 (-∞ ~ ∞) 값을 지니게 되며

분모의 값이 더 큰 경우와 분자의 값이 더 큰 경우가 **대칭적**인 범위를 가짐!

오즈비의 장점

후향적 연구처럼 한 변수가 **고정**되어 있는 경우에도 사용 가능

<table>
<tr><th rowspan="2">알코올 중독</th><th colspan="2">심장 질환 발병 여부</th><th rowspan="2">합</th></tr>
<tr><th>심장 질환 환자</th><th>건강한 사람</th></tr>
<tr><td rowspan="2">O</td><td>4 (4/6)</td><td>2 (2/6)</td><td rowspan="2">6</td></tr>
<tr><td colspan="2">4/2</td></tr>
<tr><td rowspan="2">X</td><td>46 (46/144)</td><td>98 (98/144)</td><td rowspan="2">144</td></tr>
<tr><td colspan="2">46/98</td></tr>
<tr><td>합</td><td>50</td><td>100</td><td>150</td></tr>
</table>

<table>
<tr><th rowspan="2">알코올 중독</th><th colspan="2">심장 질환 발병 여부</th><th rowspan="2">합</th></tr>
<tr><th>심장 질환 환자</th><th>건강한 사람</th></tr>
<tr><td rowspan="2">O</td><td>4 (4/10)</td><td>6 (6/10)</td><td rowspan="2">10</td></tr>
<tr><td colspan="2">4/2</td></tr>
<tr><td rowspan="2">X</td><td>46 (46/340)</td><td>294 (294/340)</td><td rowspan="2">340</td></tr>
<tr><td colspan="2">46/98</td></tr>
<tr><td>합</td><td>50</td><td>300</td><td>350</td></tr>
</table>

오즈비의 장점

	왼쪽 분할표	오른쪽 분할표	변화
비율의 차이 $(\pi_1 - \pi_2)$	$\frac{4}{6} - \frac{46}{144}$ $= 0.347$	$\frac{4}{10} - \frac{46}{340}$ $= 0.265$	있음
상대위험도 (π_1/π_2)	$\frac{4/6}{46/144} = 2.087$	$\frac{4/10}{46/340} = 2.956$	있음
오즈비 $(odds1/odds2)$	$\frac{4/2}{46/98} = 4.26$	$\frac{4/6}{46/294} = 4.26$	없음

오즈비 값은 **대조군의 크기에 관계없이 동일한 값**을 가짐

후향적 연구에서는 오즈비만 사용 가능

오즈비의 장점

행과 열의 순서가 바뀌어도 값이 동일

알코올 중독	심장 질환 유무		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

심장 질환 유무가 행일 때의 오즈비

$$\frac{odds1}{odds2} = \frac{4/2}{46/98} = 4.26$$

오즈비의 장점

행과 열의 순서가 바뀌어도 값이 동일

알코올 중독이 행일 때의 오즈비

$$\frac{odds1}{odds2} = \frac{4/46}{2/98} = 4.26$$

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

4 연관성 척도

오즈비의 장점

행과 열의 순서가 바뀌어도 값이 동일

✓ 오즈비 값이 $P(Y|X), P(X|Y)$ 중 어떤 식으로 정의하든
동일한 값을 가지므로 **반응변수 구분이 불필요**

알코올 중독이 행일 때의 오즈비

$$\frac{odds1}{odds2} = \frac{4/46}{2/98} = 4.26$$

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

오즈비의 장점

증명 과정은 다음과 같음

$$\text{오즈비} = \frac{odds1}{odds2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=2)/P(Y=0|X=2)}$$

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

오즈비의 장점

오즈비가 이러한 장점을 가질 수 있는 이유는 무엇일까?

오즈비 값이 $P(Y|X), P(X|Y)$ 중 어떤 식으로 정의하든

동일한 값을 가지므로 반응변수 구분이 불필요

$$\text{오즈비} = \frac{odds1}{odds2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=2)/P(Y=0|X=2)}$$

오즈비가 교차적비이기 때문!

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

교차적비 *(Cross-Product Ratio)*

교차적비

분할표에서 대각선에 위치한 값끼리 곱한 수 간의 비율

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

따라서 **오즈비**는 대각성분의 곱은 분자로,

비대각성분의 곱은 분모로 된 **교차적비의 형태 !**

4 연관성 측도

교차적비 *(Cross-Product Ratio)*

교차적비

분할표에서 대각선에 위치한 값끼리 곱한 수 간의 비율

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

한 변수가 고정된 상태에서 **대조군의 크기**가 변하거나

분할표에서 **행과 열의 위치**가 바뀌더라도 **같은 값 유지 !**

오즈비와 상대위험도

오즈비와 상대위험도의 관계

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

성별	연인 유무	
	있음	없음
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⋯	

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

성공확률 p_1, p_2가 0에 가까우면

오즈비 ≅ 상대위험도 × 1

오즈비와 상대위험도

오즈비와 상대위험도의 관계

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

상대위험도와 **오즈비**가 **근사한 값**을 가진다면,

복잡한 오즈보다 비교적 간단한 상대위험도를 사용하여 **해석이 용이**

성별		
	있음	없음
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…	

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

성공확률 p_1, p_2가 0에 가까우면 오즈비 $\cong$ 상대위험도 $\times 1$

오즈비와 상대위험도

오즈비와 상대위험도의 관계

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

또한, 상대위험도를 계산할 수 없는 자료에 오즈비를 추정하여

상대위험도를 근사시키는 데 오즈비를 사용할 수 있음

성별	이음	없음
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⋯	

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

성공확률 p_1, p_2가 0에 가까우면 오즈비 ≃ 상대위험도 × 1

3차원 분할표에서의 오즈비

조건부 연관성 *(Conditional Association)*

부분분할표에서의 연관성

즉, 제어변수(Z)가 고정되어 있을 때 X와 Y 간의 연관성을 의미

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	

제어변수(Z)의 각 수준별로 교차적비 계산

3차원 분할표에서의 오즈비

동질 연관성 *(Homogeneous Association)*

제어변수의 각 수준별 조건부 오즈비가 모두 같은 경우

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

조건부 독립성 *(Conditional Independence)*

X와 Y가 서로 독립인 상태!

제어변수의 각 수준별 조건부 오즈비가 모두 1로 같은 경우

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

3차원 분할표에서의 오즈비

동질 연관성 *(Homogeneous Association)*

제어변수의 각 수준별 조건부 오즈비가 모두 같은 경우

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

동질연관성은 대칭적이므로 **X와 Y 간에 동질연관성이 존재**한다면

XZ, YZ 간에도 동질 연관성이 존재

$$(\theta_{XZ(1)} = \theta_{XZ(2)} = \ldots = \theta_{XZ(J)}, \theta_{YZ(1)} = \theta_{YZ(2)} = \ldots = \theta_{YZ(I)})$$

3차원 분할표에서의 오즈비

주변 연관성 *(Conditional Association)*

제어변수(Z)의 모든 수준을 합친 **주변분할표**에서의 연관성

주변분할표에서의 오즈비인 **주변 오즈비**를 통해 파악 가능

주변분할표

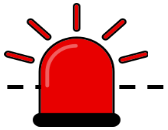
성별(X)	통학 여부(Y)		주변 오즈비
	O	X	
남자	11+14 = 25	25+5 = 30	θ_{XY+}
여자	10+7 = 17	27+12 = 39	= 1.911⋯

학과에 관계없이 남자일 때 통학할 오즈가 여자보다 **약 1.911배** 높음

3차원 분할표에서의 오즈비

주변 독립성 *(Marginal Independence)*

주변 오즈비가 1일 때($\theta_{XY+} = 1$), **주변 독립성**을 가짐



분할표에서 조건부 독립성이 성립하더라도

주변 독립성이 성립하지 않을 수 있음

조건부 오즈비와 주변 오즈비의 방향성이 항상 같지는 않기 때문!

3차원 분할표에서의 오즈비

주변 독립성 *(Marginal Independence)*

주변 오즈비가 1일 때($\theta_{XY+} = 1$), **주변 독립성**을 가짐

분할표에서 조 성립하더라도

주변 독립성이 성립하지 않을 수 있음

심슨의 역설을 통해 확인!

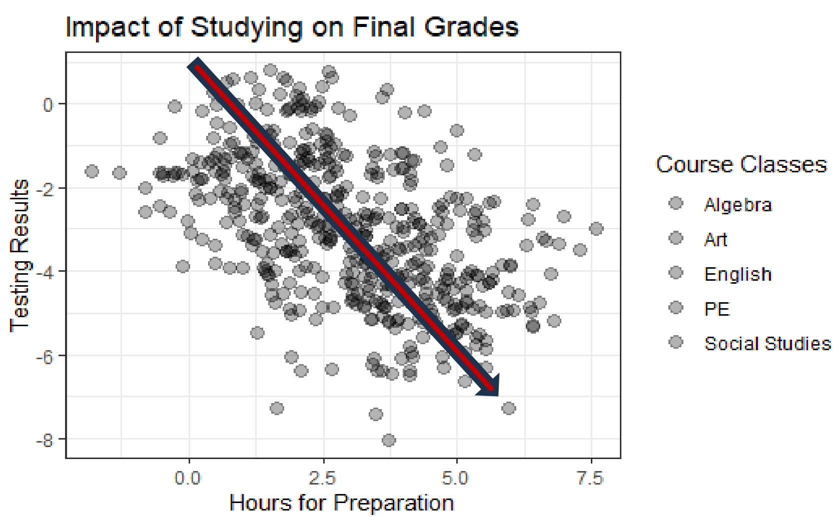
조건부 오즈비와 주변 오즈비의 방향성이 항상 같지는 않기 때문!

4 연관성 측도

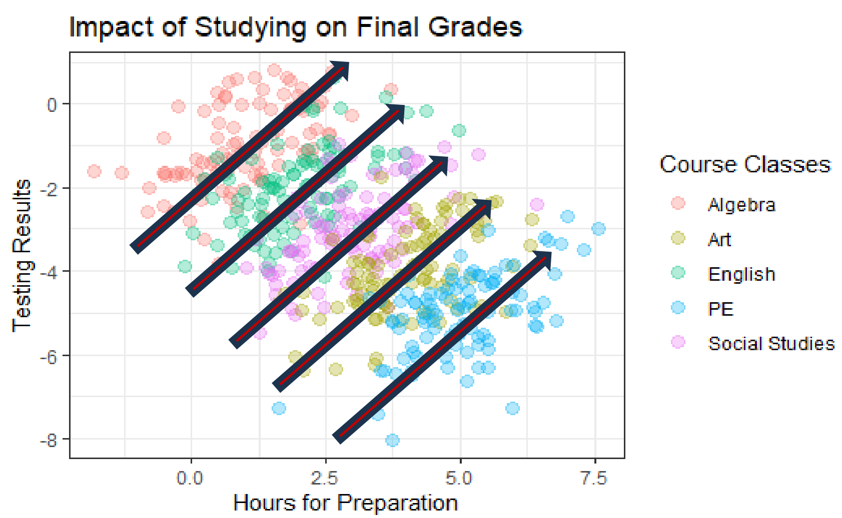
심슨의 역설 *(Simpson's Paradox)*

심슨의 역설

전반적인 추세가 경향성이 존재하는 것처럼 보이지만
세부 그룹으로 나눠서 살펴볼 경우
앞선 경향성이 사라지거나 **반대로 해석**되는 경우



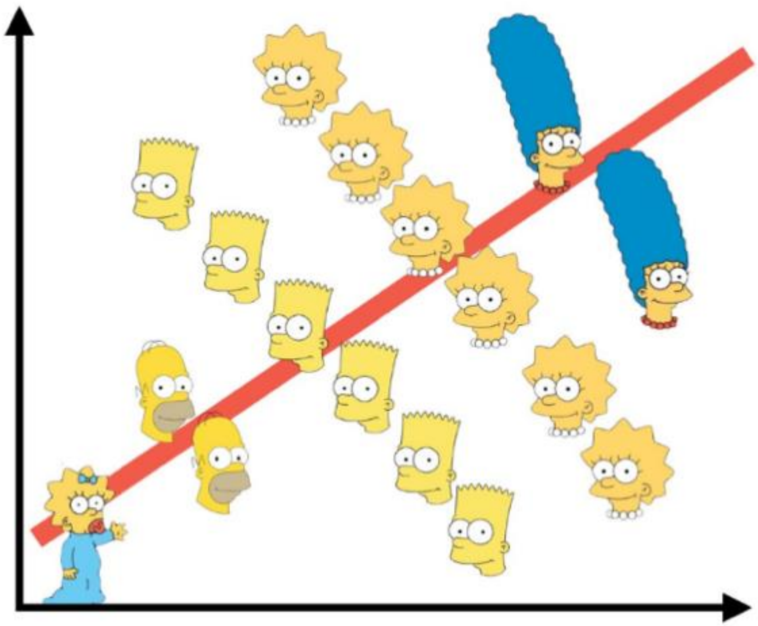
전체적으로 **우하향**하는 추세선



각 그룹별로 **우상향**하는 추세선

심슨의 역설 *(Simpson's Paradox)*

즉, 조건부 오즈비와 주변 오즈비의 연관성 방향이 다르게 나타나는 경우



심슨 가족과 함께 알아보는 심슨의 역설 ^_^

전체 심슨 가족은
우상향하는 추세를 보이지만
가족 구성원 각각을 살펴보면
우하향하는 추세를 보임

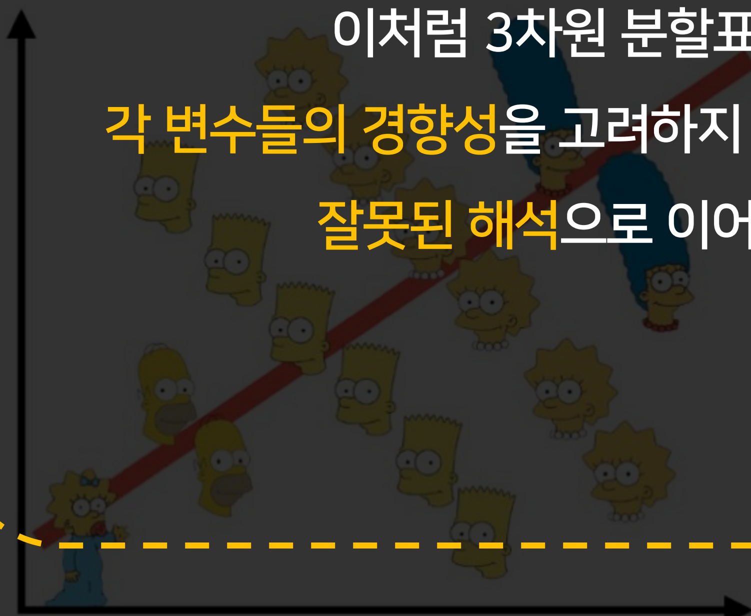
심슨의 역설 *(Simpson's Paradox)*

즉, 조건부 오즈비와 주변 오즈비의 연관성 방향이 다르게 나타나는 경우

이처럼 3차원 분할표나 플랏을 해석할 때

각 변수들의 경향성을 고려하지 않고 **전체의 추세**만 확인한다면

잘못된 해석으로 이어질 수 있으므로 주의!



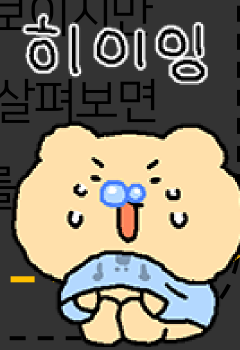
심슨 가족과 함께 알아보는 심슨의 역설 ^_^

전체 심슨 가족은

우상향하는 추세를 보이지만

가족 구성원 각각을 살펴보면

우하향하는 추세를



심슨의 역설 *(Simpson's Paradox)*

부분분할표

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

부분분할표

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

조건부 오즈비와 **주변 오즈비**가

정반대의 연관성 방향을 보이고 있음

오즈비의 기준 : 1

심슨의 역설 *(Simpson's Paradox)*

부분분할표

학과 (Z)	성별 (X)	통학 여부 (Y) O	통학 여부 (Y) X	합
경영	남자	40	5	195
	여자	130	20	
통계	남자	15	5	27
	여자	5	2	

제어변수인 학과(Z)에 따라
전체 도수의 차이가 크게 남!

제어변수(Z)가 연관성을 해석하는 데 **큰 영향을 끼치는 변수로 작용**했기 때문에
조건부 오즈비와 주변 오즈비 간에 **서로 다른 결과**가 도출

심슨의 역설 (*Simpson's Paradox*)

부분분할표

학과 (Z)	성별 (X)	통학 여부 (Y) O	통학 여부 (Y) X	합
경영	남자	40	5	195
	여자	130	20	
통계	남자	15	5	27
	여자	5	2	

제어변수인 학과(Z)에 따라

전체 도수의 차이가 크게 남!

정리하자면,

심슨의 역설은 도수의 크기에 따른 영향력 차이로 발생!

따라서 조건부 오즈비와 주변 오즈비의 방향성 차이에 유의하여 분석해야 함

제어변수(Z)가 연관성을 해석하는 데 큰 영향을 끼치는 변수로 작용했기 때문에

조건부 오즈비와 주변 오즈비 간에 서로 다른 결과가 도출

THANK YOU