

범주형자료분석팀

2팀

김보현
송승현
최용원
김동희
오주원

INDEX

1. 범주형 자료분석
2. 분할표
3. 독립성 검정
4. 연관성 측도

1

범주형 자료분석

1

범주형 자료분석

변수의 구분

범주형 자료분석



반응변수(Y 변수)가 범주형인 자료의 분석



자료 (Data)

분석 대상인 모집단의 특성

변수를 열(column) 로,

변수 별 관측치를 행(row) 으로 나열해서 만들어진 행렬

변수를 측정한 값

1

범주형 자료분석

변수의 구분

범주형 자료분석



반응변수(Y 변수)가 범주형인 자료의 분석

Y 변수

종속변수, 반응변수,
목적변수, 결과변수,
표적변수

X 변수

독립변수, 설명변수,
예측변수, 위험인자,
요인 (범주형)

변수의 구분

자료

양적(수치형) 자료

Quantitative data

이산형 자료

Discrete data

연속형 자료

Continuous data

질적(범주형) 자료

Qualitative data

명목형 자료

Nominal data

순서형 자료

Ordinal data



자료의 형태

양적 자료

관측값이 수치로 측정되는 자료,
측이나 셈과 같은 수량의 형태를 지님

이산형 자료

Discrete data

값을 셀 수 있는 자료

정수 형태

ex) 나이, 신생아 수 등

연속형 자료

Continuous data

연속인 구간에서 값을 취하는 자료

실수 형태

ex) 키, 몸무게 등

자료의 형태

양적 자료

관측값이 수치로 측정되는 자료,
측이나 셈과 같은 수량의 형태를 지님



양적 자료의 특징

공분산, 상관계수 등의 수리적 계산 가능
정규분포 가정 하에서 회귀분석 가능

회귀팀 클린업 참고!

자료의 형태

질적 자료

관측 결과가 여러 개의 범주의 집합으로 나타나는 자료
순서 유무에 따라 명목형과 순서형으로 구분

명목형 자료

Nominal data

범주간에 순서의 의미가 없는 자료

ex) 성별, 혈액형 등

순서형 자료

Ordinal data

범주간에 순서의 의미가 있는 자료

ex) 별점, 순위, 선호도 등

자료의 형태

질적 자료

관측 결과가 여러 개의 범주의 집합으로 나타나는 자료
순서 유무에 따라 명목형과 순서형으로 구분

명목형 자료

Nominal data

범주간에 순서의 의미가 없는 자료

ex) 성별, 혈액형 등

순서형 자료

Ordinal data

프로야구 팀

LG	기아	SSG	두산	키움
----	----	-----	----	----

ex) 별점, 순위, 선호도 등

1

범주형 자료분석

자료의 형태

질적 자료

관측 결과가 여러 개의 범주의 집합으로 나타나는 자료
순서 유무에 따라 명목형과 순서형으로 구분

명목형 자료
Nominal data
프로야구 순위

1위	2위	3위	4위	5위
----	----	----	----	----

ex) 성별, 혈액형 등

순서형 자료
Ordinal data

범주간에 순서의 의미가 있는 자료
ex) 별점, 순위, 선호도 등

질적 자료의 특징



순서형 자료에 명목형 자료 분석방법을 적용할 수는 있으나,
순서에 대한 정보가 무시되어 검정력에 심각한 손실 발생



분할표 작성 가능



각 범주에 특정 점수를 할당하여 양적자료로 활용

질적 자료의 특징



수치형 자료처럼 표현된 범주형 자료에 주의!

순서에 대한 정보가 무시되어 검정력에 심각한 손실 발생



수치의 형태로 표현된 범주형 자료에
수치적인 의미가 있다고 보는 것은 옳지 않음

분류표 작성 가능



분석의 용이성을 위해

숫자의 형태를 빌려 범주형 변수를 나타낸 것 일뿐!

각 범주에 특정 점수를 할당하여 양적자료로 활용

2

분할표

분할표

분할표

범주형 변수들에 대한 관측값을 일목요연하게 도표로 요약한 자료

수치형 자료

중심, 산포도 등의 기술통계를 통한 자료 분석

범주형 자료

분할표를 통한 자료 분석

분할표

분할표

범주형 변수들에 대한 관측값을 일목요연하게 도표로 요약한 자료

		Y		
		1	...	J
X	1			
	...			
	I			

I * J개 칸

분할표

분할표

범주형 변수들에 대한 관측값을 일목요연하게 도표로 요약한 자료

	Y		
	1	...	J
X	1		
	...		
	I		

J개의 수준

I개의 수준

수준 (level)

각 변수의 카테고리 개수

EX) 성별 : 남/여 총 2개의 수준

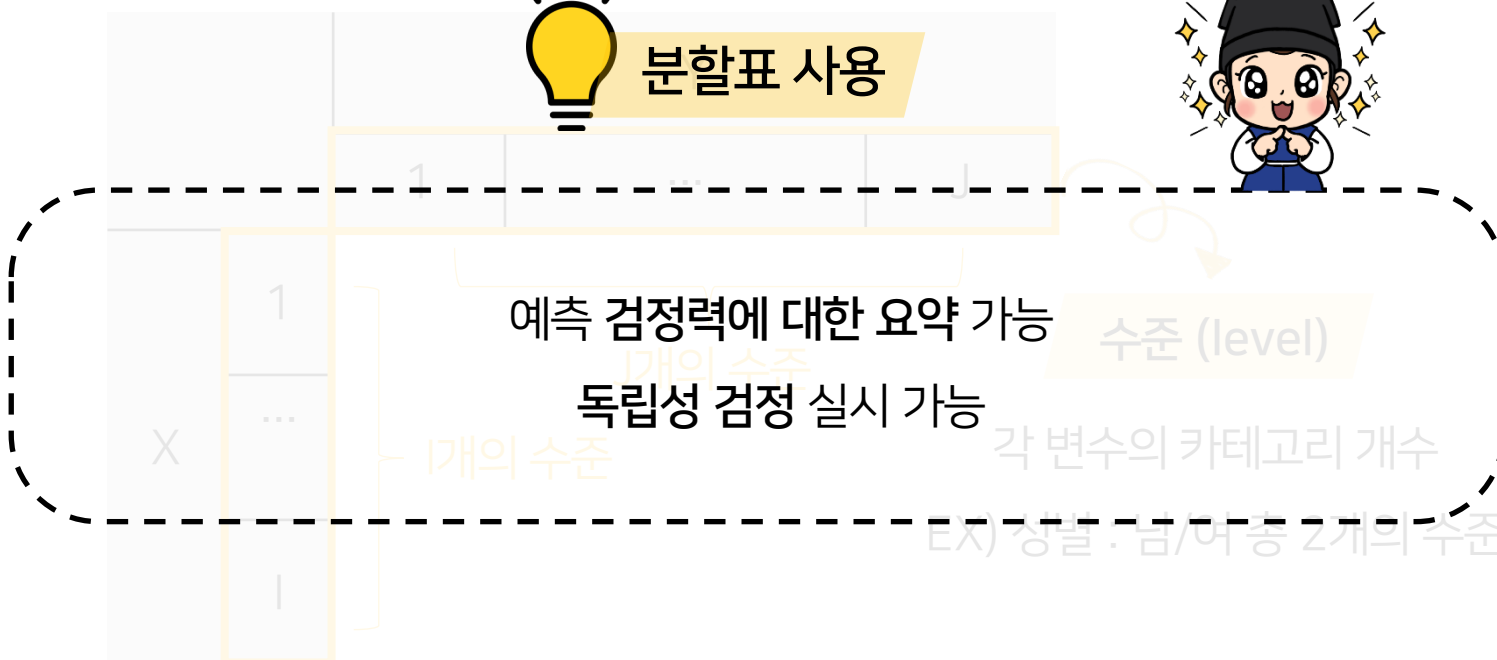
분할표

분할표

범주형 변수들에 대한 관측값을 일목요연하게 도표로 요약한 자료



분할표 사용



분할표



분할표

범주형 변수들에 대한 관측값을 일목요연하게 도표로 요약한 자료

분할표는 수준에 따라 **무한가지** 경우의 형태로 만들 수 있음

하지만 **3차원 이상의 고차원**일 경우,

모델링 등의 방식을 통한 분석이 더 큰 **편의성**을 가짐



따라서 본 클린업에서는 **2차원 분할표**와 **3차원 분할표**를
중점적으로 다룰 예정!

여러 차원의 분할표

2차원 분할표

두 개의 범주형 변수를 분류한 분할표

	Y			합계
X	n_{11}	...	n_{1j}	n_{1+}

	n_{i1}	...	n_{ij}	n_{i+}
합계	n_{+1}	...	n_{+j}	n_{++}

X : 설명변수

Y : 반응변수

보통 설명변수를 행에,
반응변수를 열에 위치

여러 차원의 분할표

2차원 분할표

두 개의 범주형 변수를 분류한 분할표

	Y			합계
X	n_{11}	...	n_{1j}	n_{1+}

	n_{i1}	...	n_{ij}	n_{i+}
합계	n_{+1}	...	n_{+j}	n_{++}

 n_{ij} : 각 칸의 도수 n_{i+} : 각 행의 주변 도수 n_{+j} : 각 열의 주변 도수 n_{++} : 총계

'+' 첨자는 그 위치에 해당하는
도수를 모두 더했다는 의미!

여러 차원의 분할표

3차원 분할표

세 가지 범주형 변수를 분류한 분할표

기존의 X와 Y에 K개의 수준을 가진 **제어변수 Z**가 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_{++2}

X : 설명변수

Y : 반응변수

Z : 제어변수

부분분할표

거주지	성별	통학 여부		합계
		0	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

제어변수 Z의 수준에 따라
X, Y 변수가 분류된 분할표



고정된 제어변수의 각 수준에서
반응변수에 미치는 설명변수의 효과
확인 가능!

부분분할표

거주지	성별	통학 여부		합계
		0	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

제어변수 수준에 따라



'거주지'를 **제어변수**로 설정
거주지마다 통학 여부에 대한
성별의 영향 파악 가능

고정된 제어변수의 각 수준에서
반응변수에 미치는 **설명변수의 효과**
확인 가능!

주변분할표

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

주변분할표

제어변수 **Z**의 모든 수준을
결합하여 만든 2차원 분할표



X 변수와 Y 변수 간의 관계에서
Z 변수를 무시한 형태

주변분할표

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

주변분할표



제어변수인 '거주지'가 통합
거주지와 무관하게 통학 여부에
대한 성별의 영향 파악 가능

X 변수와 Y 변수 간의 관계에서

Z 변수를 무시한 형태

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

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

π_{ij} : 전체 대비 각 칸의 비율

π_{++} : 모든 칸의 확률의 합

이때 π_{ij} 은 각 칸의 도수인 n_{ij} 를
전체 도수 n_{++} 으로 나눠서 구함

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}

결합확률 (Joint Probability)



각 칸의 확률



모집단에서 추출된 표본이

X 변수의 I번째 수준과

Y 변수의 J번째 수준에

동시에 속할 확률

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}

결합확률 (Joint Probability)



각 칸의 확률



모집단에서 추출된 표본이

X 변수의 I번째 수준과

Y 변수의 J번째 수준에



결합확률의 합 $\sum \pi_{ij} = 1$

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}

주변확률 (Marginal Probability)



X 변수의 I번째 수준이
전부 일어날 행의 확률



Y 변수의 J번째 수준이
전부 일어날 열의 확률

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}

주변확률 (Marginal Probability)



X 변수의 I번째 수준이

전부 일어날 행의 확률

행의 주변확률



Y 변수의 J번째 수준이

전부 일어날 열의 확률

열의 주변확률

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}

주변확률 (Marginal Probability)



X 변수의 I번째 수준이



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



Y 변수의 J번째 수준이

전부 일어날 열의 확률

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}

조건부 확률

(Conditional Probability)



X 변수의 각 수준에서의
Y 변수의 값



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

비율에 대한 분할표

연령대에 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

$$\text{결합확률} : \pi_{22} = \frac{42}{250} \approx 0.17$$

연령대에 무관하게 농구를 선호할

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

20대라는 가정 하에 야구를 선호할

$$\text{조건부 확률} : \frac{\pi_{21}}{\pi_{2+}} = \frac{42}{250} \approx 0.34$$

비율에 대한 분할표

연령대에 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

$$\text{결합확률} : \pi_{22} = \frac{42}{250} \approx 0.17$$

연령대에 무관하게 농구를 선호할

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

20대라는 가정 하에 야구를 선호할

$$\text{조건부 확률} : \frac{\pi_{21}}{\pi_{2+}} = \frac{42}{250} \approx 0.34$$

비율에 대한 분할표

연령대에 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

$$\text{결합확률} : \pi_{22} = \frac{42}{250} \approx 0.17$$

연령대에 무관하게 농구를 선호할

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

20대라는 가정 하에 야구를 선호할

$$\text{조건부 확률} : \frac{\pi_{21}}{\pi_{2+}} = \frac{41}{120} \approx 0.34$$

3

독립성 검정

독립성 검정

독립성 검정

두 범주형 변수가 통계적으로 관계가 있는지 확인하기 위한 검정



독립성 검정의 목적

- 1) 두 변수 간 연관성 유무 판단
- 2) 분석 가치 판단

독립성 검정의 목적

독립성 검정

두 범주형 변수가 통계적으로 관계가 있는지 확인하기 위한 검정



독립성 검정의 목적

- 1) 두 변수 간 연관성 유무 판단
- 2) 분석 가치 판단

독립성 검정의 목적



독립성 검정

두 범주형 변수가 통계적으로 관계가 있는지 확인하기 위한 검정

독립성 검정의 결과, 두 변수가 독립으로 도출된다면



반응변수에 대해 설명변수가 어떠한 영향도 끼치지 못한다는 뜻이므로

독립성 검정의 목적

분석 가치가 없다고 판단 가능

1) 두 변수 간 연관성 유무 판단

2) 분석 가치 판단



독립성 검정의 가설

독립성 검정

두 범주형 변수가 통계적으로 관계가 있는지 확인하기 위한 검정

모든 결합확률이 행과 열 주변 확률의 곱과 동일

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

관측도수와 기대도수

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



관측도수와 기대도수

독립성 검정 귀무가설 (H_0)

관측도수 (Observed Frequency)

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

기대도수 (Expected Frequency)

실제 관측값

분할표의 각 칸의 도수

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

양 변에 n 곱함

귀무가설 하에

각 칸의 도수에 대한 기대값

이므로

귀무가설 하에서 **관측도수 = 기대도수**

각 칸의 결합확률 * n

전체 표본 n * 행과 열의 주변확률

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

위 가설에서 n 만 곱한 것일 뿐 **같은 의미**를 지님!

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

독립성 검정의 종류

모든 기대도수가 5 이상인 것을 의미

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 이상인 것을 의미



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)

대표본 + 명목형 자료 독립성 검정

피어슨 카이제곱 검정

검정통계량

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

기각역

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

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

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

검정 과정

관측도수와 기대도수의 차이가 큼! → 검정통계량의 값이 큼 → 귀무가설 기각

→ 두 변수는 독립이 아님 → 변수 간의 연관성 존재

대표본 + 명목형 자료 독립성 검정

피어슨 카이제곱 검정

검정통계량

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2_{(i-1)(j-1)}$$

기각역

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

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

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

검정 과정

관측도수와 기대도수의 차이가 큼! → 검정통계량의 값이 큼 → 귀무가설 기각

→ 두 변수는 독립이 아님 → 변수 간의 연관성 존재

대표본 + 명목형 자료 독립성 검정

피어슨 카이제곱 검정

검정통계량

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2_{(i-1)(j-1)}$$

기각역

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

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

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

검정 과정

관측도수와 기대도수의 **차이가 큼!** → 검정통계량의 값이 큼 → **귀무가설 기각**

→ 두 변수는 **독립이 아님** → 변수 간의 **연관성** 존재

대표본 + 순서형 자료 독립성 검정

순서형 자료

각 변수의 수준에 차등적인 점수 할당

행점수 $\mu_1 \leq \mu_2 \leq \dots \leq \mu_I$

열점수 $v_1 \leq v_2 \leq \dots \leq v_J$

수준 간 점수 차이는 동등할 필요 없음
목적과 데이터의 특성에 따라 할당 가능



MH 검정

검정통계량

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

기각역

$$M^2 \geq \chi_{\alpha,1}^2$$

대표본 + 순서형 자료 독립성 검정

두 변수의 추세 연관성을 확인하기 위해
피어슨 교차적률 상관계수 사용



MH 검정

검정통계량

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

기각역

$$M^2 \geq \chi_{\alpha,1}^2$$



대표본 + 순서형 자료 독립성 검정

피어슨 교차적률 상관계수

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

두 변수의 추세 연관성을 확인하기 위해

피어슨 교차적률 상관계수 사용

공분산을 두 표준편차의 곱으로 나눈

상관계수와 같은 형태



피어슨 교차적률 상관계수의 범위는 **-1에서 1** 사이

상관계수가 **0**일 때 두 변수는 **독립**

상관계수가 **-1** 혹은 **1**에 가까울 수록 두 변수간의 큰 **연관성** 존재

MH 검정

검정통계량

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

기각역

$$M^2 \geq \chi_{\alpha,1}^2$$

대표본 + 순서형 자료 독립성 검정

순서형 자료

각 변수의 수준에 차등적인 점수 할당

$$\text{행점수 } \mu_1 \leq \mu_2 \leq \dots \leq \mu_I$$

$$\text{열점수 } v_1 \leq v_2 \leq \dots \leq v_J$$

수준 간 점수 차이는 동등할 필요 없음

목적과 데이터의 특성에 따라 할당 가능



MH 검정

검정통계량

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

기각역

$$M^2 \geq \chi_{\alpha,1}^2$$

검정 과정

상관계수가 크다 → 검정통계량이 크다

→ 귀무가설(두 변수는 독립) 기각! → 변수 간의 연관성이 존재!

3

독립성 검정

대표본 + 순서형 자료 독립성 검정



순서형 자료

MH 검정

각 변수의 수준에 차등적인 점수 할당

검정통계량

독립성 검정은 두 범주형 변수의 연관성 유무만 판단하기 때문에 $r^2 \sim \chi^2_1$

구체적으로 어떻게 연관이 있는지는 파악할 수 없음

기각역

수준 간 점수 차이는 동등할 필요 없음

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

연관성 측도를 통해 변수 간 연관성의 성질을 파악!

검정 과정

상관계수가 크다 → 검정통계량이 크다

→ 귀무가설(두 변수는 독립) 기각! → 변수 간의 연관성이 존재!

4

연관성 측도

비율의 비교 척도

비율: 각 행에 따른 조건부 확률

비율의 비교 척도		
비율의 차이	상대 위험도	오즈비

두 범주형 변수가 모두 2가지 수준만을 갖는 이항변수일 때,
세 종류의 척도들을 통해 변수 간 **연관성** 파악 가능

비율의 차이

각 행의 조건부 확률 간 차이: $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

여성이 연인이 있을 조건부 확률

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

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

비율의 차이

각 행의 조건부 확률 간 차이: $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

남성이 연인이 있을 조건부 확률

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

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

비율의 차이

각 행의 조건부 확률 간 차이: $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

비율의 차이

$$\pi_1 - \pi_2 = 0.814 - 0.793 = 0.021$$



여성이 연인이 있을 확률이
남성일 때 보다 약 **0.021** 높음

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

비율의 차이

각 행의 조건부 확률 간 차이: $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

비율의 차이

$$\pi_1 - \pi_2 = 0.4 - 0.4 = 0$$



성별이 연인 유무에 영향을 미치지 못함
두 변수가 독립일 때 비율의 차이는 0

성별	연인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

상대 위험도

조건부 확률의 비 $\frac{\pi_1}{\pi_2}$

0보다 크거나 같은 값을 가짐



상대 위험도 해석

상대위험도가 1에서 멀어질수록
두 변수간 **연관성**이 크다고 파악

상대 위험도

조건부 확률의 비 $\frac{\pi_1}{\pi_2}$

0보다 크거나 같은 값을 가짐

연인이 있을 경우의 상대 위험도

$$= \frac{\pi_1}{\pi_2} = \frac{0.814}{0.793} = 1.027$$



여성일 경우 연인이 있을 확률이
약 **1.027배** 높음

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

상대 위험도

조건부 확률의 비 $\frac{\pi_1}{\pi_2}$

0보다 크거나 같은 값을 가짐

연인이 있을 경우의 상대 위험도

$$= \frac{\pi_1}{\pi_2} = \frac{0.4}{0.4} = 1$$



성별이 연인 유무에 영향을 미치지 못함

두 변수가 독립일 때 상대위험도는 1

성별	연인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

4

연관성 측도

비율의 차이 vs 상대위험도

성별	연인 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

비율의 차이 : $0.02 - 0.01 = 0.01$

상대 위험도: $0.02 / 0.01 = 2$

성별	연인 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

비율의 차이 : $0.92 - 0.91 = 0.01$

상대 위험도: $0.92 / 0.91 = 1.01$

4

연관성 측도

비율의 차이 vs 상대위험도

성별	연인 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

성별	연인 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

비율의 차이 : $0.02 - 0.01 = 0.01$

비율의 차이 : $0.92 - 0.91 = 0.01$



비율의 차이는 0.01로 같음

4

연관성 측도

비율의 차이 vs 상대위험도

성별	연인 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

성별	연인 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

상대 위험도: $0.02 / 0.01 = 2$

상대 위험도: $0.92 / 0.91 = 1.01$



상대위험도는 큰 차이가 나타남

4

연관성 측도



비율의 차이 vs 상대위험도

상대위험도에 따르면 큰 연관성을 띄는 두 변수가 연인 유무

비율의 차이의 경우 연관성이 미미하다고 잘못 판단될 수 있음

성별	연인 유무		성별	연인 유무	
	있음	없음		있음	없음
여성	0.02	0.98	여성	0.92	0.08
남성	0.01	0.99	남성	0.91	0.09



조건부 확률이 0 혹은 1의 값에 가까울 때

상대 위험도의 절대적인 차이만을 이용해 연관성을 판단하는 것은 매우 위험! $1 = 1.01$



상대위험도는 큰 차이가 나타남

4

연관성 측도

비율의 차이 vs 상대위험도



성별	연인 유무		성별	연인 유무	
	있음	없음		있음	없음
여성	0.02	0.98	여성	0.92	0.08
남성	0.01	0.99	남성	0.91	0.09

비율의 차이와 상대위험도는

후향적 연구와 같이 반응변수(Y)를 고정시킨 연구에서는
사용할 수 없다는 한계가 존재!

후향적 연구란 이미 나온 결과를 바탕으로

상대 위험도: $0.02 / 0.01 = 2$ 과거 기록을 관찰하는 연구를 의미
상대 위험도: $0.92 / 0.91 = 1.01$



상대위험도는 큰 차이가 나타남

비율의 차이와 상대위험도의 한계

후향적 연구같이 Y 변수를 고정시켰을 경우 사용할 수 없음

	위암 발병 (Y=1)	위암 발병X (Y=0)	합계
알코올 중독 O (X=1)	4	2	6
알코올 중독 X (X=0)	46	98	144
합계	50	100	150

사례군과 대조군의 비율을 연구자가 사전에 비율을 1:2로 추출
대조군의 합 변경 시 비율의 차이와 상대위험도도 달라지기 때문에

비율의 차이와 상대위험도 사용 불가

오즈비(Odds Ratio)



오즈

성공확률 / 실패확률

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



오즈 해석

성공확률이 실패확률의 몇 배인지를 의미!

오즈비(Odds Ratio)

오즈

성공확률 / 실패확률

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

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

여성이 연인이 있을 오즈
: $0.814 / 0.186 = 4.388$

남성이 연인이 있을 오즈
: $0.793 / 0.207 = 3.826$

오즈비(Odds Ratio)

오즈비

각 행 별로 계산한 오즈의 비

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

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

$$\text{오즈비} = 4.388 / 3.826 = 1.147$$



여성이 연인이 있을 **오즈**가
남성이 연인이 있을 **오즈**보다
약 **1.147**배 높다!

오즈비(Odds Ratio)

오즈비

각 행 별로 계산한 오즈의 비

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

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

$$\text{오즈비} = 4.388 / 3.826 = 1.147$$



여성이 연인이 있을 **오즈**가
남성이 연인이 있을 **오즈**보다
약 **1.147**배 높다!

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

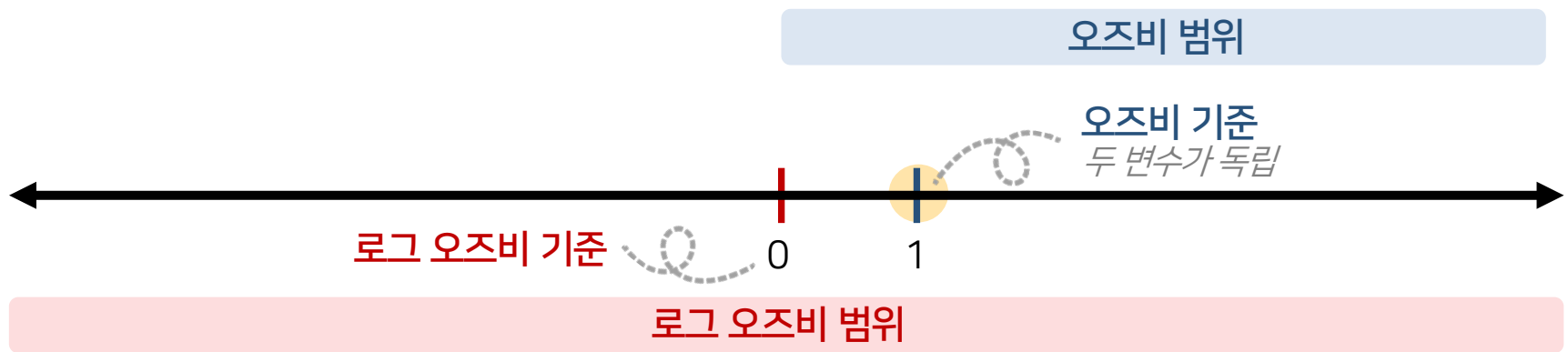


서로 역수 관계에 있는 오즈비는
방향만 반대이고 두 변수간 연관성의 정도는 같음

4

연관성 측도

로그 오즈비 (Log Odds Ratio)



로그 오즈비

오즈비에 로그(log)를 씌운 형태

로그 오즈비 (Log Odds Ratio)

	오즈비		로그 오즈비	
기준	$\theta = 1$		$\theta = 0$	
범위	$(0, \infty)$		$(-\infty, \infty)$	
기준에 따른 범위	$(0, 1)$	$(1, \infty)$	$(-\infty, 0)$	$(0, \infty)$



오즈비의 범위는 $\theta = 1$ 을 기준으로 서로 **비대칭**



로그 오즈비 (Log Odds Ratio)

	오즈비		로그 오즈비	
기준	$\theta = 1$		$\theta = 0$	
범위	$(0, \infty)$		$(-\infty, \infty)$	
기준에 따른 범위	$(0, 1)$	$(1, \infty)$	$(-\infty, 0)$	$(0, \infty)$



기준의 비대칭적인 오즈비의 범위 교정



오즈비의 장점 ①



한 변수가 고정되어 있는 경우에도 사용 가능

후향적 연구

알코올 중독	위암 발병 여부		합
	위암 환자	건강한 사람	
0	4	2	6
X	46	98	144
합	50	100	150

알코올 중독	위암 발병 여부		합
	위암 환자	건강한 사람	
0	4	6	10
X	46	294	340
합	50	300	350

오즈비의 장점 ①



한 변수가 고정되어 있는 경우에도 사용 가능

후향적 연구

	왼쪽 분할표	오른쪽 분할표	변화
비율의 차이 ($\pi_1 - \pi_2$)	0.347	0.265	있음
상대 위험도 (π_1 / π_2)	2.087	2.956	있음
오즈비 (odds1/odds2)	4.26	4.26	없음



오즈비의 장점 ①



한 변수가 고정되어 있는 경우에도 사용 가능

오즈비 값은

대조군의 크기에 관계없이 동일한 값을 가짐

	왼쪽 분할표	오른쪽 분할표	변화
비율의 차이 ($\pi_1 - \pi_2$)	0.347	0.265	이름
상대 위험도 (π_1 / π_2)	2.087	2.956	이름
오즈비 (odds1/odds2)	4.26	4.26	이름

후향적 연구에서는 오즈비만 사용 가능



오즈비의 장점 ②



행과 열의 순서가 바뀌어도 값이 동일

알코올 중독	위암 발병 여부		합
	위암 환자	건강한 사람	
0	4	2	6
X	46	98	144
합	50	100	150

위암 발병 여부	알코올 중독		합
	0	X	
위암 환자	4	46	50
건강한 사람	2	98	100
합	6	144	150

오즈비의 장점 ②

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



알코올 중독	위암 발병 여부		합
	위암 환자	건강한 사람	
0	4	2	6
X	46	98	144
합	50	100	150

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



위암 발병 여부	알코올 중독		합
	0	X	
위암 환자	4	46	50
건강한 사람	2	98	100
합	6	144	150

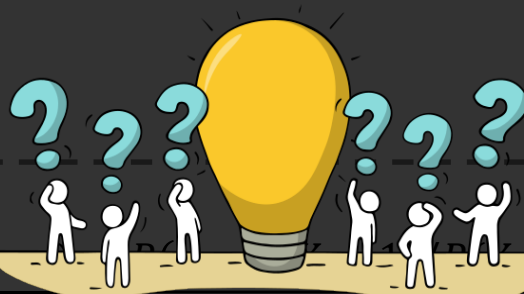
오즈비의 장점 ②

$$\begin{aligned}
 \frac{odds1}{odds2} &= \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=2)/P(Y=0|X=2)} \\
 &= \frac{\frac{P(X=1|Y=1) \times P(Y=1)}{P(X=1)}}{\frac{P(X=2|Y=1) \times P(Y=1)}{P(X=2)}} \bigg/ \frac{\frac{P(X=1|Y=0) \times P(Y=0)}{P(X=1)}}{\frac{P(X=2|Y=0) \times P(Y=0)}{P(X=2)}} \\
 &= \frac{P(X=1|Y=1)/P(X=1|Y=0)}{P(X=2|Y=1)/P(X=2|Y=0)}
 \end{aligned}$$



P(Y|X), P(X|Y) 모두 동일한 값을 가지므로
반응변수 구분 불필요

오즈비의 장점 ②



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

오즈비는 왜 이런 장점을 지닐까?

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

오즈비가 **교차적비(cross-product ratio)**이기 때문!



$P(Y|X), P(X|Y)$ 모두 동일한 값을 가지므로

반응변수 구분 불필요

교차적비 (cross-product ratio)

교차적비

분할표에서 대각선에 위치한 값끼리 곱한 수 간의 비율

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

대각성분의 곱과
비대각성분의 곱



한 변수가 고정된 상태에서 대조군의 크기가 변하거나
분할표에서 행과 열의 위치가 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		비율
	있음	없음	
여성	4 (0.02)	196 (0.98)	0.0204
남성	3 (0.01)	297 (0.99)	0.0101



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

$$p_1, p_2 \cong 0$$

실패 확률 $\cong 1$



$$\text{오즈비} \cong \text{상대위험도} \times 1$$

오즈비와 상대위험도



성공확률이 0에 가까우면

오즈비와 상대위험도가 근사한 값을 가짐

복잡한 오즈를 통한 해석이 아닌
상대위험도(확률)로 해석 가능



해석 용이

상대위험도를 계산할 수 없는 자료에
오즈비 추정



상대위험도 근사에 오즈비 사용

3차원 분할표에서의 오즈비

부분분할표에서의 연관성

조건부 오즈비

제어변수 Z가 고정되어 있을 때 X와 Y 간의 연관성 파악 가능

조건부 연관성

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

제어변수 Z의 각 수준별
교차적비 계산

3차원 분할표에서의 오즈비

조건부 오즈비

동질 연관성 (Homogeneous Association)

제어변수의 각 수준별 조건부 오즈비가 모두 같을 경우

$$\begin{aligned}\theta_{XZ(1)} &= \theta_{XZ(2)} = \dots = \theta_{XZ(J)}, \\ \theta_{YZ(1)} &= \theta_{YZ(2)} = \dots = \theta_{YZ(I)}\end{aligned}$$

대칭적 성질

조건부 독립성 (Conditional Independence)

제어변수의 각 수준별 조건부 오즈비가 모두 **1**로 같을 경우

$$\theta_{XY(1)} = \dots = \theta_{XY(K)} = 1$$

X와 Y가 서로 독립

제어변수에 관계없이 X와 Y에 대한 오즈비가 **1**

3차원 분할표에서의 오즈비

주변분할표에서의 연관성

주변 오즈비

제어변수 Z의 모든 수준을 합친 주변분할표에서의 연관성 파악 가능

주변 연관성

주변분할표			
성별 (X)	통학 여부 (Y)		주변 오즈비
	0	X	
남자	11 + 16 + 14 = 41	25 + 4 + 5 = 34	θ_{XY+} = 1.515
여자	10 + 22 + 7 = 39	27 + 10 + 12 = 49	

$$\theta_{XY+} = 1$$

주변 오즈비가 1일 때,

주변 독립성을 가짐



3차원 분할표에서의 오즈비

주변분할표에서의 연관성

주변 오즈비

분할표에서 조건부 독립성이 성립하더라도

주변 독립성이 성립하지 않을 수 있음

제어변수 Z의 모든 수준을 합친 주변분할표에서의 연관성 파악 가능

주변 연관성

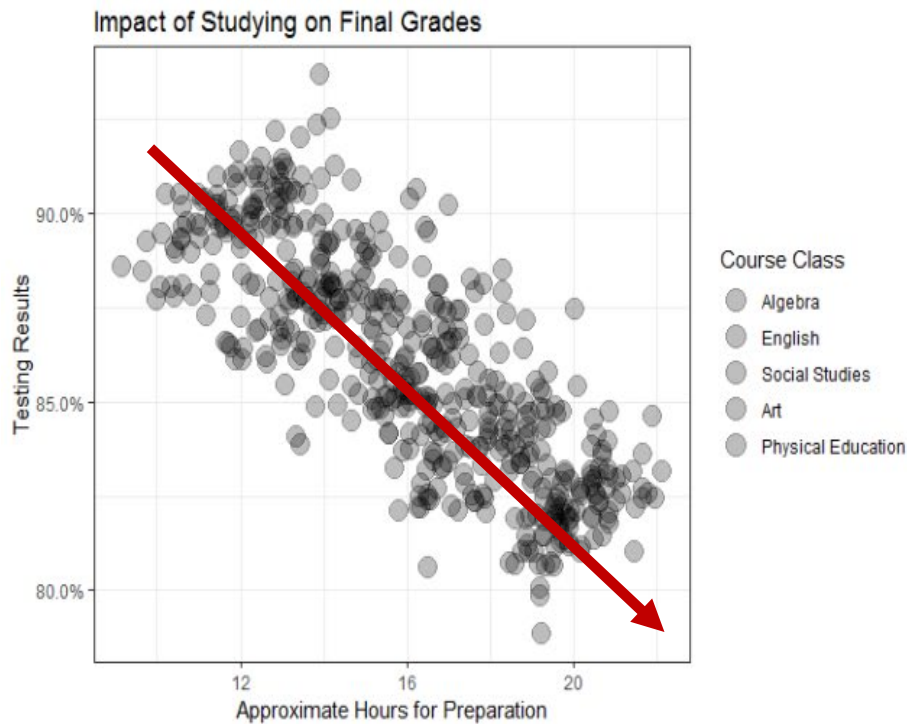


심슨의 역설 (Simpson's Paradox)

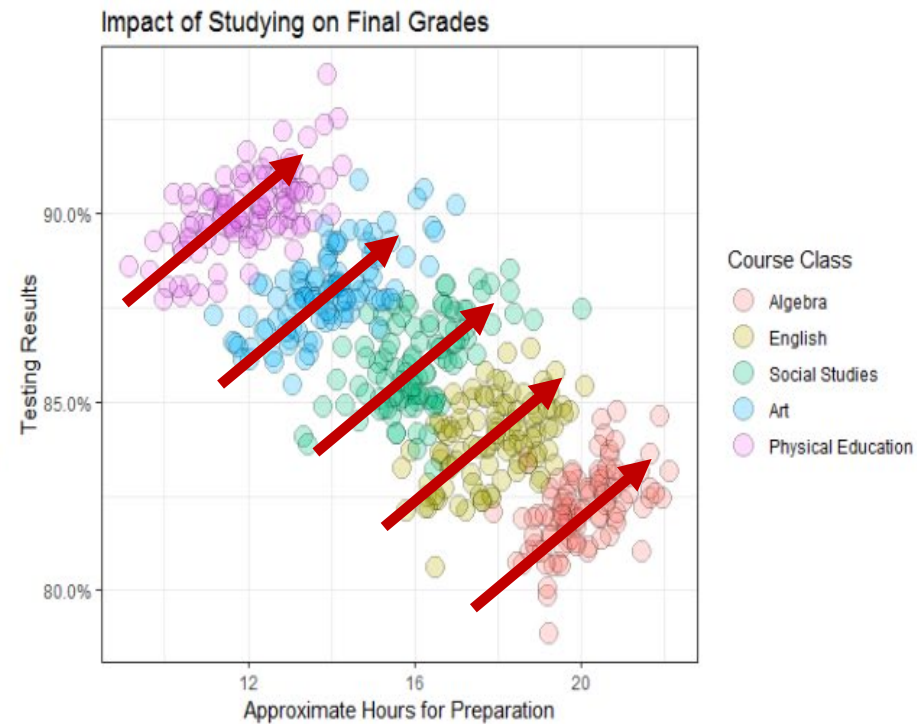
조건부 오즈비와 주변 오즈비의 $\theta_{yy_1} = 1$ 방향성이 항상 같지는 않기 때문 !

주변 오즈비가 1일 때, 주변 독립성을 가짐

심슨의 역설 (Simpson's Paradox)



전체적으로 **우하향**하는 추세선

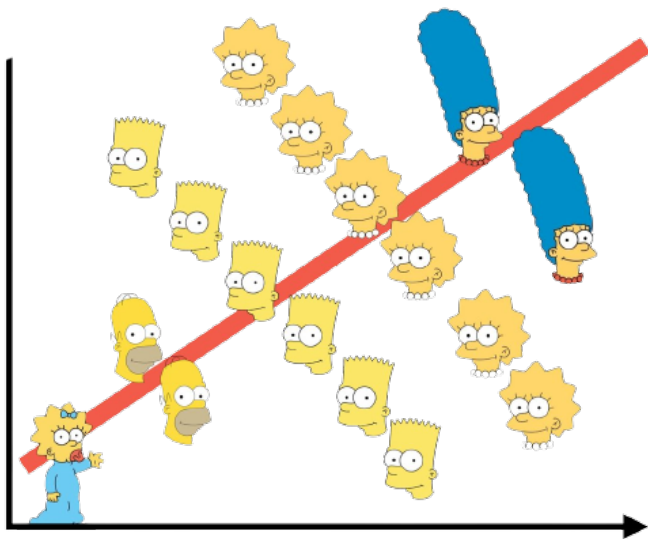


각 그룹별로 **우상향**하는 추세선

심슨의 역설 (Simpson's Paradox)

심슨의 역설

전반적인 추세는 경향성이 존재하는 것처럼 보이지만
세부 그룹으로 나뉘서 살펴볼 경우 앞선 경향성이 **사라지거나 반대로 해석**되는 경우

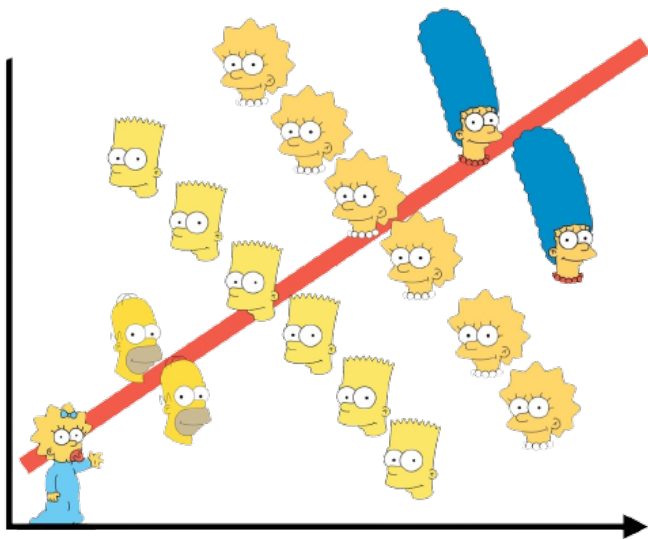


조건부 오즈비와 주변 오즈비의
연관성 방향이 다른 경우

심슨의 역설 (Simpson's Paradox)

심슨의 역설

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



전체의 추세뿐만 아니라
각 변수들의 **경향성** 확인 필요

심슨의 역설 (Simpson's Paradox)

조건부 오즈비와 주변 오즈비는 **정반대의 연관성**을 보이고 있음

오즈비의 기준은 1


부분분할표				
학과 (Z)	성별 (X)	통학 여부 (Y)		조건부 오즈비
		0	X	
경영	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
	0	X	
남자	55	10	$\theta_{XY+} = 0.90$
여자	135	22	

심슨의 역설 (Simpson's Paradox)

조건부 오즈비와 주변 오즈비는 **정반대의 연관성**을 보이고 있음

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

주변분할표				
성별 (X)	통학 여부 (Y)		조건부 오즈비	
	0	X		
남자	10	10	$\theta_{XY+} = 0.90$	제어변수인 학과 (Z)에 따라 전체 도수의 차이 가 크게 남 
여자	135	22		

심슨의 역설 (Simpson's Paradox)



제어변수 Z가 연관성을 해석하는 데 큰 영향을 끼치는 변수로 작용

부분분할표				
학과 (Z)	성별 (X)	통학 여부 (Y)		조건부 오즈비
		0	X	
경영	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
	0	X	
남자	55	10	$\theta_{XY+} = 0.90$
여자	135	22	

심슨의 역설 (Simpson's Paradox)

도수의 크기에 따른 영향력 차이로 인해

심슨의 역설 발생



제어변수 Z가 연관성을 해석하는 데 큰 영향을 끼치는 변수로 작용



부분분할표

주변분할표

학과 (Z)	성별 (X)	통학 여부 (Y)		조건부 오즈비	성별 (X)	통학 여부 (Y)		주변 오즈비
경영	남자	40	5	$\theta_{XY(1)} = 1.23$	남자	55	10	$\theta_{YY+} = 0.90$
	여자	130	20		여자	135	22	
통계	남자	15	5	$\theta_{XY(2)} = 1.2$	남자	55	10	$\theta_{YY+} = 0.90$
	여자	5	2		여자	135	22	

조건부 오즈비와 주변 오즈비의 방향성 차이에
유의하여 분석

다음주 예고

1. GLM

2. 유의성 검정

3. 로지스틱 회귀 모형

4. 다범주 로짓 모형

5. 포아송 회귀 모형



THANK YOU

