범주형자료분석팀

2팀
정희철
김민서
이주형
심수현
이준석

INDEX

1

범주형 자료분석

1 범주형 자료분석

범주형 자료분석의 개념

범주형 자료분석

반응변수가 **범주형**인 자료에 대한 분석

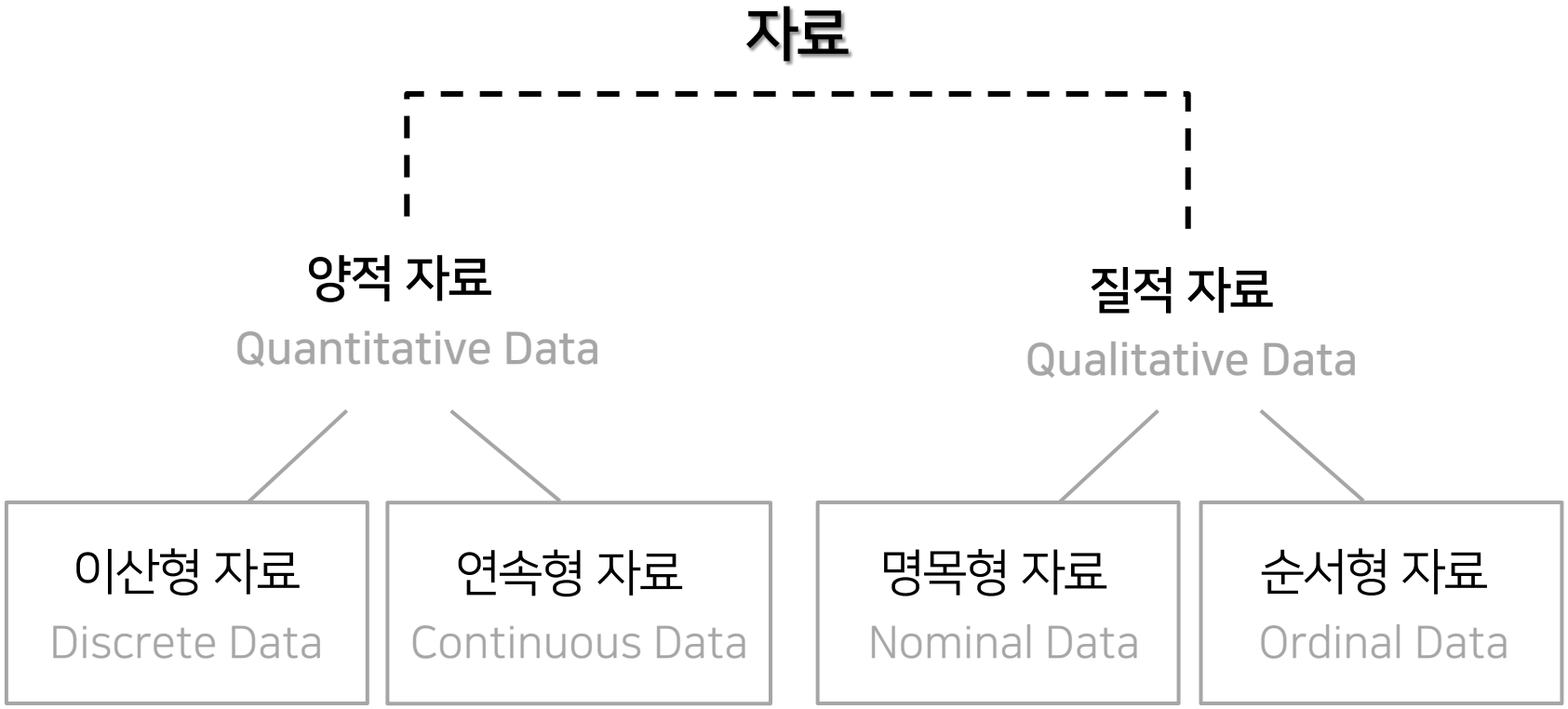
Y 변수	종속변수, **반응변수**, 결과변수, 표적변수
X 변수	독립변수, **설명변수**, 예측변수, 위험인자, 공변량 (연속형) , 요인 (범주형)

1 범주형 자료분석

자료의 형태

1 범주형 자료분석

자료의 형태



양적 자료

양적 자료 (수치형 자료)

특정 대상에 대한 **이산적** 또는 **연속적** 측정치

이산형 자료
Discrete Data

셀 수 있는 값의 형태를 취하는 자료로, **정수** 형태
ex) 개수, 3반 학생 수

연속형 자료
Continuous Data

연속인 어떤 구간에서 값을 취하는 자료로, **실수** 형태
ex) 몸무게, 키, BMI 지수

양적 자료

양적 자료 (수치형 자료)

특정 대상에 대한 **이산적** 또는 **연속적** 측정치

이산형 자료
Discrete Data

셀 수 있는 값의 형태를 취하는 자료로, **정수** 형태
ex) 개수, 3반 학생 수

연속형 자료
Continuous Data

연속인 어떤 구간에서 값을 취하는 자료로, **실수** 형태
ex) 몸무게, 키, BMI 지수

1 범주형 자료분석

양적 자료

양적 자료 (수치형 자료)

특정 대상에 대한 **이산적** 또는 **연속적** 측정치

이산형 자료
Discrete Data

셀 수 있는 값의 형태를 취하는 자료로, 정수 형태
ex) 개수, 3반 학생 수

연속형 자료
Continuous Data

연속인 어떤 구간에서 값을 취하는 자료로, 실수 형태
ex) 몸무게, 키, BMI 지수

양적 자료

양적 자료 (수치형 자료)

특정 대상에 대한 **이산적** 또는 **연속적** 측정치

양적자료는

수치적인 의미

→

① **수리적인 계산** 가능

② 일반 **회귀분석** 가능

(오차의 분포가 정규분포라고 가정했을 때)

회귀팀 클린업 참고!

1 범주형 자료분석

질적 자료

질적 자료 (범주형 자료)

2개 이상의 **범주들의 집합**으로 구성된 자료로,

범주 간 **순서의 존재유무**에 따라 **명목형**과 **순서형**으로 구분

명목형 자료
Nominal Data

범주의 **순서가 의미없이** 그 자체로만 분류된 자료

ex) 성별(남/여), 정치성향(진보/보수), 선호하는 커피

ex)

선호하는 커피

아메리카노	카페모카	카페라떼	콜드브루

1 범주형 자료분석

질적 자료

질적 자료 (범주형 자료)

2개 이상의 **범주들의 집합**으로 구성된 자료로,

범주 간 **순서의 존재유무**에 따라 **명목형**과 **순서형**으로 구분

순서형 자료
Ordinal Data

범주의 나열된 **순서가 의미 있는** 자료

ex) 만족도(나쁨/보통/좋음), 지지도

ex)

지지도

매우 반대	반대	중립	지지	매우 지지

1 범주형 자료분석

질적 자료

질적 자료 (범주형 자료)

2개 이상의 **범주들의 집합**으로 구성된 자료로,

범주 간 **순서의 존재유무**에 따라 **명목형**과 **순서형**으로 구분

특징

① **순서형 자료**에 **명목형 자료** 분석방법 적용가능

명목형 자료 → 순서형 자료 적용은 불가능!

1 범주형 자료분석

질적 자료

질적 자료(범주형 자료)

2개 이상의 범주들의 집합으로 구성된 자료로,

범주 간 순서의 존재유무에 따라 명목형과 순서형으로 구분

순서형 자료의 명목형 자료 분석방법 적용

분석과정에서 순서에 대한 정보 무시로 인해

검정력에 심각한 손실 가져올 수 있음!

특징

① 순서형 자료에 명목형 자료 분석방법 적용가능

명목형 자료 → 순서형 자료 적용은 불가능!

1 범주형 자료분석

질적 자료

질적 자료 (범주형 자료)

2개 이상의 **범주들의 집합**으로 구성된 자료로,

범주 간 **순서의 존재유무**에 따라 **명목형**과 **순서형**으로 구분

특징

② **분할표** 작성 가능

③ 각 범주에 특정 점수를 할당해 **양적자료로 활용 가능**

3주차 클린업에서 배울 예정!

2

분할표

분할표

분할표

2개 이상의 범주형 변수들을 표로 나타내는 방식

각 범주가 속하는 결과의 도수들을 각 칸에 넣어서 정리한 표

연속형 자료 — **중심, 산포도** 등의 **기술통계** 중심 자료 요약

범주형 자료 — **분할표**를 통한 자료 요약

분할표

분할표

2개 이상의 범주형 변수들을 표로 나타내는 방식

각 범주의 수준에 속하는 결과의 도수들을 각 칸에 넣어서 정리한 표

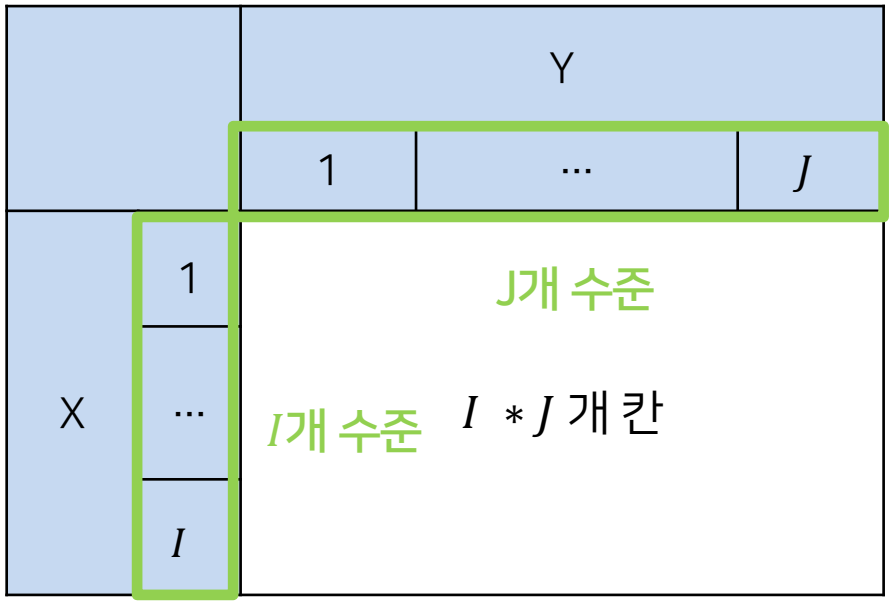
		Y		
		1	…	J
X	1	$I * J$ 개 칸		
	…			
	I			

분할표

분할표

2개 이상의 범주형 변수들을 표로 나타내는 방식

각 범주의 수준에 속하는 결과의 도수들을 각 칸에 넣어서 정리한 표



수준(level)

각 범주형 변수가 취하는 값

Ex) 성별: 여성/남성 2개의 수준

분할표

분할표

2개 이상의 범주형 변수들을 표로 나타내는 방식

각 범주의 수준에 속하는 결과의 도수들을 각 칸에 넣어서 정리한 표

분할표 사용

① 예측 **검정력에 대한 요약** 가능

② **독립성 검정** 실시 가능

3주차 분류평가지표에서 다뤄질 예정!

2 분할표

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

n_{ij} : 각 칸의 도수

n_{i+}, n_{+j} : 각 열과 행의 주변 도수

n_{++} : 총계

일반적으로 **X를 행, Y를 열로** 설정함

2 분할표

여러 차원의 분할표

3차원 분할표

세 가지의 범주형 변수를 분류한 분할표로,

기존 X와 Y에서 **제어변수(Control Variable) Z**가 추가됨

		Y		합계
Z	X1	n_{111}	n_{121}	n_{1+1}
		n_{211}	n_{221}	n_{2+1}
	합계	n_{+11}	n_{+21}	n_{++1}
	X2	n_{112}	n_{122}	n_{1+2}
		n_{212}	n_{222}	n_{2+2}
	합계	n_{+12}	n_{+22}	n_{++2}

2 분할표

여러 차원의 분할표

3차원 분할표

세 가지의 범주형 변수를 분류한 분할표로
가운 X와 Y에서 제어변수(Control Variable) Z가 추가됨

분할표는 범주형 변수의 개수에 따라

무한대로 확장 가능!

		Y		합계
Z	X1	n_{111}	n_{121}	n_{1+1}
		n_{211}	n_{221}	n_{2+1}
	합계	n_{+11}	n_{+21}	n_{++1}
	X2	n_{112}	n_{122}	n_{1+2}
		n_{212}	n_{222}	n_{2+2}
	합계	n_{+12}	n_{+22}	n_{++2}

BUT 클린업에서는 2차원과 3차원 분할표를
중점으로 다룰 계획

2 분할표

부분분할표

학과(Z)	성별(X)	자취 여부(Y)		합계
		O	X	
통계	남자	11	25	36
	여자	10	27	37
	합계	21	52	73
경제	남자	16	4	20
	여자	22	10	32
	합계	38	14	52

부분분할표

제어변수 Z의 수준에 따라

나머지 변수 X,Y를 분류한 표

↓

고정된 제어변수의 한 수준에서

반응변수에 대한 설명변수의 효과

파악 가능!

2 분할표

부분분할표

학과(Z)	성별(X)	자취 여부(Y)		합계
		O	X	
통계	남자	11	25	36
	여자	10	27	37
	합계	21	52	73
경제	남자	16	4	20
	여자	22	10	32
	합계	38	14	52

부분분할표

제어변수 Z의 수준에 따라

나머지 변수 X,Y를 분류한 표

↓

고정된 제어변수의 한 수준에서

반응변수에 대한 설명변수의 효과

파악 가능!

2 분할표

부분분할표

학과(Z)	성별(X)	자취 여부(Y)		합계
		O	X	
통계	남자	11	25	36
	여자	10	27	37
	합계	21	52	73
경제	남자	16	4	20
	여자	22	10	32
	합계	38	14	52

부분분할표

제어변수 Z의 수준에 따라

'학과'가 제어변수로 설정

학과마다 자취 여부에 대한

성별의 효과 파악 가능

고정된 제어변수의 한 수준에서

반응변수에 대한 설명변수의 효과

파악 가능!

2 분할표

주변분할표

성별 (X)	자취 여부(Y)		
	O	X	합계
남자	27	29	56
여자	32	37	69
합계	59	66	125

주변분할표

모든 제어변수의 수준을

결합해 얻은 2차원 분할표

제어변수 통제 X

↓

일반적인 **2차원 분할표**와

형태는 동일하지만, **의미가 다름!**

2 분할표

주변분할표

성별 (X)	자취 여부(Y)		
	O	X	합계
남자	27	29	56
여자	32	37	69
합계	59	66	125

주변분할표

모든 제어변수의 수준을

결합해 얻은 2차원 분할표

제어변수(학과)가 결합해

학과에 대한 정보 사라짐

일반적인 2차원 분할표와

형태는 동일하지만, 의미가 다름!

비율에 대한 분할표

비율에 대한 분할표

분할표 각 칸에 도수 대신 **비율**을 넣은 표

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

→

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

2 분할표

분할표에서의 확률 분포

	Y			합계
X	π_{11}	…	π_{1J}	π_{1+}
	…	…	…	…
	π_{I1}	…	π_{IJ}	π_{I+}
합계	π_{+1}	…	π_{+J}	1

결합 확률 (Joint Probability)

표본이 X의 i번째 수준에 속하면서

Y의 j번째 수준에 속할 확률

주변 확률 (Marginal Probability)

결합 확률의 행 또는 열의 합

2 분할표

분할표에서의 확률 분포

	Y			합계
X	π_{11}	…	π_{1J}	π_{1+}
	…	…	…	…
	π_{I1}	…	π_{IJ}	π_{I+}
합계	π_{+1}	…	π_{+J}	1

결합 확률 (Joint Probability)

표본이 X의 i번째 수준에 속하면서

Y의 j번째 수준에 속할 확률

주변 확률 (Marginal Probability)

결합 확률의 행 또는 열의 합

2 분할표

분할표에서의 확률 분포

	Y			합계
X	π_{11}	…	π_{1J}	π_{1+}
	…	…	…	…
	π_{I1}	…	π_{IJ}	π_{I+}
합계	π_{+1}	…	π_{+J}	1

π_{1+} → 행의 주변 확률

π_{+1}, π_{+J} → 열의 주변 확률

결합 확률 (Joint Probability)

표본이 X의 i번째 수준에 속하면서 Y의 j번째 수준에 속할 확률

주변 확률 (Marginal Probability)

결합 확률의 행 또는 열의 합

2 분할표

분할표에서의 확률 분포

	Y			합계
X	π_{11}	…	π_{1J}	π_{1+}
	…	…	…	…
	π_{I1}	…	π_{IJ}	π_{I+}
합계	π_{+1}	…	π_{+J}	1

조건부 확률

(Conditional Probability)

X의 각 수준에서 **Y에 대한 확률**

$$= P(Y = j \mid X = i)$$

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

2 분할표

분할표에서의 확률 분포

$$P(Y=1|X=1)=\frac{P(Y=1,X=1)}{P(X=1)}=\frac{\pi_{11}}{\pi_{1+}}$$

	Y			합계
X	π_{11}	…	π_{1J}	π_{1+}
	…	…	…	…
	π_{I1}	…	π_{IJ}	π_{I+}
합계	π_{+1}	…	π_{+J}	1

조건부 확률

(Conditional Probability)

X의 각 수준에서 **Y에 대한 확률**

$$= P(Y=j \mid X=i)$$

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

2 분할표

분할표에서의 확률 분포

	Y			합계
X	π_{11}	…	π_{1J}	π_{1+}
	…	…	…	…
	π_{I1}	…	π_{IJ}	π_{I+}
합계	π_{+1}	…	π_{+J}	1

결합 확률 (Joint Probability)

표본이 X의 i번째 수준에 속하면서

Y의 j번째 수준에 속할 확률

주변 확률 (Marginal Probability)

결합 확률의 행 또는 열의 합

모든 결합 확률의 합은 1 !

2 분할표

분할표에서의 확률 분포

	게임	쇼핑	합계
남성	0.5	0.1	0.6
여성	0.1	0.3	0.4
합계	0.6	0.4	1

게임을 가장 선호하는 남성일

결합확률 : 0.5

성별에 상관없이 게임을 가장 좋아하는

사람의 **주변확률** : 0.6

여성이라는 가정 하에 쇼핑을

가장 좋아하는 사람의 **조건부 확률** :

$$\frac{0.3}{0.4} = 0.75$$

2 분할표

분할표에서의 확률 분포

	게임	쇼핑	합계
남성	0.5	0.1	0.6
여성	0.1	0.3	0.4
합계	0.6	0.4	1

게임을 가장 선호하는 남성일 결합확률 : 0.5

성별에 상관없이 게임을 가장 좋아하는 사람의 **주변확률** : 0.6

여성이라는 가정 하에 쇼핑을 가장 좋아하는 사람의 **조건부 확률** :

$$\frac{0.3}{0.4} = 0.75$$

2 분할표

분할표에서의 확률 분포

	게임	쇼핑	합계
남성	0.5	0.1	0.6
여성	0.1	0.3	0.4
합계	0.6	0.4	1

게임을 가장 선호하는 남성일
결합확률 : 0.5

성별에 상관없이 게임을 가장 좋아하는
사람의 **주변확률** : 0.6

여성이라는 가정 하에 쇼핑을
가장 좋아하는 사람의 **조건부 확률** :

$$\frac{0.3}{0.4} = 0.75$$

3

독립성 검정

독립성 검정

독립성 검정

두 범주형 변수가 연관성이 있는지를 검정하는 방법

독립성 검정을 통해..

1) 변수 간의 연관성이 있는지 없는지 판단

2) 분석 가치 판단

3 독립성 검정

독립성 검정의 목적

독립성 검정

두 범주형 변수가 연관성이 있는지를 검정하는 방법

독립성 검정을 통해..

1) **변수 간의 연관성**이 있는지 없는지 판단

2) **분석 가치** 판단

3 독립성 검정

독립성 검정의 목적

독립성 검정

두 범주형 변수가 연관성이 있는지를 검정하는 방법

독립성 검정 결과,

두 변수가 **독립**, 즉 **연관성이 없다**고 판단

=> 관계없는 변수들이므로 이 이상의 **분석 가치가 사라짐!**

3 독립성 검정

독립성 검정의 가설

통계적 독립성

모든 **결합확률**이 **주변확률의 곱**과 동일하다.

H_0 : 두 범주형 변수는 **독립**이다. $\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$

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

관측도수와 기대도수

관측도수 (Observed Frequency) [n_{ij}]

표본의 도수, 즉 **실제 관측 값**

비율에 대한 분할표에서는 $\boldsymbol{n \times \pi_{ij}}$

기대도수 (Expected Frequency) [μ_{ij}]

귀무가설 하에 각 칸의 도수에 대한 **기대값**

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

기대도수와 관측도수

독립성 검정의 가설을 다시 표현하면?

관측도수 (Observed Frequency) $[n_{ij}]$

H_0 : 두 범주형 변수는 독립이다. $\Leftrightarrow \mu_{ij} = n\pi_{ij}$

비율에 대한 분할표에서는 $n \times \pi_{ij}$

→ 귀무가설 하에서 주변확률의 곱 = 결합확률

기대도수 (Expected Frequency) $[\mu_{ij}]$

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

귀무가설 하에 각 칸의 도수에 대한 기대값

이는 앞의 가설에서 n이 곱해졌는가의 차이일 뿐 동일한 정보량!

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

3 독립성 검정

독립성 검정의 종류

2차원 분할표 독립성 검정

<table>
<tr><td rowspan="3">대표본</td><td rowspan="2">명목형</td><td>피어슨 카이제곱 검정</td></tr>
<tr><td>가능도비 검정</td></tr>
<tr><td>순서형</td><td>MH 검정</td></tr>
<tr><td colspan="2">소표본</td><td>피셔의 정확검정</td></tr>
</table>

3 독립성 검정

명목형 자료의 독립성 검정 (대표본)

피어슨 카이제곱 검정

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

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

가능도비 검정

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

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

검정 Flow

관측도수와 기대도수의 차이가 **크다**

↓

검정통계량 (χ^2, G^2) 이 **크다**

↓

P-value 값이 **작다**

↓

귀무가설 **기각**

↓

변수간 **연관성 존재**

3 독립성 검정

순서형 자료의 독립성 검정 (대표본)

MH 검정

두 범주형 변수가 **모두 순서형**인 경우 사용

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

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

피어슨 교차적률 상관계수

검정 Flow

상관 계수 r이 **크다**

↓

검정통계량 (M^2) 이 **크다**

↓

P-value 값이 **작다**

↓

귀무가설 **기각**

↓

변수간 **연관성 존재**

3 독립성 검정

순서형 자료의 독립성 검정 (대표본)

피어슨 교차적률 상관계수 $(=r)$

MH검정 : 범주형 변수의 수준에 점수를 할당하여 변수 간 선형 추세 측정

↓

변수에 행 점수 : $u_1 \leq u_2 \leq \cdots \leq u_i$ / 열 점수 : $v_1 \leq v_2 \leq \cdots \leq v_i$ 할당 후,

두 변수간 추세 연관성을 파악하기 위해 피어슨 교차적률 상관계수를 사용 !

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

$$r = \frac{\sum(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{[\sum(u_i - \bar{u})^2 p_{i+}][\sum(v_j - \bar{v})^2 p_{+j}]}}$$

피어슨 교차적률 상관계수

일반 상관계수와 마찬가지로 $-1 \leq r \leq 1$, $r = 0$이면 독립

검정 Flow

상관계수 r이 大

↓

검정통계량 (M^2), 이 大

↓

P-value 값이 小

↓

귀무가설 기각

↓

변수간 연관성 존재

3 독립성 검정

독립성 검정의 한계

검정 통계량 값이 크다
≠ 변수간 연관성이 크다

→

범주형 변수의 **연관성 유무만** 판단
하지만 얼마나 연관이 있는지 파악 불가

변수 간 연관성의 성질을 파악하기 위해서는
연관성 측도를 알아야 함!

3 독립성 검정

독립성 검정의 한계

검정 통계량 값이 크다
≠ 변수간 연관성이 크다

→

범주형 변수의 **연관성 유무만** 판단
하지만 얼마나 연관이 있는지 파악 불가

변수 간 연관성의 성질을 파악하기 위해서는
연관성 측도를 알아야 함!

4

연관성 측도

4 연관성 측도

비율의 비교 척도

비율 : 각 행을 기준으로 두고 계산한 조건부 확률

비율의 비교 척도		
비율의 차이	상대 위험도	오즈비

두 범주형 변수가 모두 **2가지 수준만을 갖는 이항변수**일 때,
세 척도들을 통해 두 변수간 **연관성의 성질**을 파악할 수 있음

비율의 차이 (Difference of Proportions)

비율의 차이

π_i : i번째 행의 조건부 확률

조건부 확률의 차이 $= \boldsymbol{\pi_1 - \pi_2}$

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

여성이 자취경험이 있을 조건부 확률

$$= \pi_1 = \frac{509}{509 + 116} = 0.814$$

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

비율의 차이 (Difference of Proportions)

비율의 차이

π_i : i번째 행의 조건부 확률

조건부 확률의 차이 $= \boldsymbol{\pi_1 - \pi_2}$

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

여성이 자취경험이 있을 조건부 확률

$$= \pi_1 = \frac{509}{509 + 116} = 0.814$$

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

비율의 차이 (Difference of Proportions)

비율의 차이

π_i : i번째 행의 조건부 확률

조건부 확률의 차이 $= \boldsymbol{\pi_1 - \pi_2}$

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

남성이 자취경험이 있을 조건부 확률

$$= \pi_2 = \frac{398}{398 + 104} = 0.793$$

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

4 연관성 측도

비율의 차이 (Difference of Proportions)

비율의 차이

π_i : i번째 행의 조건부 확률

조건부 확률의 차이 $= \boldsymbol{\pi_1 - \pi_2}$

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

비율의 차이 $= \pi_1 - \pi_2 = 0.0216$

→ 여성일 때 자취경험이 있을 확률이 남성일 때보다 0.0216 높음 !

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

비율의 차이 (Difference of Proportions)

비율의 차이

π_i : i번째 행의 조건부 확률

조건부 확률의 차이 $= \boldsymbol{\pi_1 - \pi_2}$

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

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

→ **성별이 자취 여부에 영향을 끼치지 않음**

(= 성별과 자취여부는 서로 **독립**이다)

성별	자취 유무	
	있음	없음
여성	80 (0.4)	120 (0.6)
남성	40 (0.4)	60 (0.6)

상대위험도

상대위험도

$$\text{조건부 확률의 비} = \frac{\pi_1}{\pi_2}$$

0보다 크거나 같은 값을 가짐

상대위험도가 **1에서 멀어질수록** 두 변수간 **연관성이 크다**고 판단

자취경험이 있을 경우 상대 위험도

$$= \frac{\pi_1}{\pi_2} = \frac{0.814}{0.793} = 1.03$$

→ 여성일 경우 자취경험이 있을 확률이 약 1.03배 높다.

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

상대위험도

상대위험도

$$\text{조건부 확률의 비} = \frac{\pi_1}{\pi_2}$$

0보다 크거나 같은 값을 가짐

상대위험도가 **1에서 멀어질수록** 두 변수간 **연관성이 크다**고 판단

자취경험이 있을 경우 상대 위험도

$$= \frac{\pi_1}{\pi_2} = \frac{0.814}{0.793} = 1.03$$

→ 여성일 경우 자취경험이 있을 확률이 약 **1.03배** 높다.

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

상대위험도

상대위험도

$$\text{조건부 확률의 비} = \frac{\pi_1}{\pi_2}$$

0보다 크거나 같은 값을 가짐

상대위험도가 **1에서 멀어질수록** 두 변수간 **연관성이 크다**고 판단

자취경험이 있을 경우 상대 위험도

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

→ 두 변수가 **독립**일 때 상대위험도는 **1**

성별	자취 유무	
	있음	없음
여성	80 (0.4)	120 (0.6)
남성	40 (0.4)	60 (0.6)

4 연관성 측도

비율의 차이 vs 상대위험도

조건부 확률이 0에 가까울수록 반응변수에 대한 두 집단의 **영향력 차이가 큼**

성별	자취 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

비율의 차이 : 0.01

상대위험도 : 0.02/0.01=2

성별	자취 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

비율의 차이 : 0.01

상대위험도 : 0.92/0.91=1.01

4 연관성 측도

비율의 차이 vs 상대위험도

조건부 확률이 0에 가까울수록 반응변수에 대한 두 집단의 **영향력 차이가 큼**

성별	자취 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

비율의 차이 : 0.01

성별	자취 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

비율의 차이 : 0.01

상대위험도 : 0.02/0.0 **비율의 차이는 서로 같음** 험도 : 0.92/0.91=1.01

4 연관성 측도

비율의 차이 vs 상대위험도

조건부 확률이 0에 가까울수록 반응변수에 대한 두 집단의 **영향력 차이가 큼**

성별	자취 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

성별	자취 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

상대 위험도는 큰 차이를 보임

상대위험도 : 0.02/0.01=2

상대위험도 : 0.92/0.91=1.01

비율의 차이와 상대위험도의 한계

후향적 연구처럼 **한 변수를 고정시킨 조사**에서는 **사용 불가**

	위암환자(Y=1)	건강한 사람(Y=0)	합
알코올 중독O(X=1)	4	2	6
알코올 중독X(X=0)	46	98	144
합	50	100	150

위암환자의 비율을 1/3으로 고정 → 비율의 차이, 상대위험도 사용 불가

연구자가 환자의 비율을 어떻게 설정하는지에 따라 값이 달라지기 때문

4 연관성 측도

비율의 차이와 상대위험도의 한계

후향적 연구처럼 **한 변수를 고정시킨 조사**에서는 **사용 불가**

연구 대상의 독립변수가 이미 발생한 후에 나타난
종속변수를 대상으로 한 연구 방법

	위암환자(Y=1)	건강한 사람(Y=0)	합
알코올 중독O(X=1)	4	2	6
알코올 중독X(X=0)	46	98	144
합	50	100	150

위암환자의 비율을 1/3으로 고정 → 비율의 차이, 상대위험도 사용 불가

연구자가 환자의 비율을 어떻게 설정하는지에 따라 값이 달라지기 때문

4 연관성 측도

비율의 차이와 상대위험도의 한계

후향적 연구처럼 **한 변수를 고정시킨 조사**에서는 **사용 불가**

	위암환자(Y=1)	건강한 사람(Y=0)	합
알코올 중독O(X=1)	4	2	6
알코올 중독X(X=0)	46	98	144
합	50	100	150

위암환자의 비율을 1/3으로 고정 → 비율의 차이, 상대위험도 사용 불가

연구자가 환자의 비율을 어떻게 설정하는지에 따라 값이 달라지기 때문

4 연관성 측도

오즈비

오즈 (Odds)

$$\text{성공확률} / \text{실패확률}$$

π : 어떤 사건의 성공확률 → $odds = \frac{\pi}{1-\pi}$, $\pi = \frac{odds}{1+odds}$

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

여성일 경우 자취경험이 있을 오즈

= 0.814/0.186 = **4.388**

남성일 경우 자취경험이 있을 오즈

= 0.793/0.207 = **3.826**

4 연관성 측도

오즈비

오즈 (Odds)

$$\text{성공확률} / \text{실패확률}$$

π : 어떤 사건의 성공확률 → $odds = \frac{\pi}{1-\pi}$, $\pi = \frac{odds}{1+odds}$

성별	자취 유무	
	있음	없음
여성	309 (0.814)	116 (0.186)
남성	398 (0.793)	104 (0.207)

여성의 입장에서 자취경험이 있을 오즈

= 0.814/0.186 = 4.388

남성의 입장에서 자취경험이 있을 오즈

= 0.793/0.207 = 3.826

✓ 오즈는 실패에 비해 성공이 어느 정도의 배수로 있는지 알려줌 !

오즈비

오즈비 (Odds ratio)

각 오즈의 비를 의미함

0보다 크거나 같은 값을 가짐

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

여성의 입장에서 자취경험이 있을 오즈 = 약 4.388

남성의 입장에서 자취경험이 있을 오즈 = 약 3.826

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

여성이 자취경험이 있을 오즈가

남성이 자취경험이 있을 오즈보다 약 **1.147배** 높다

4 연관성 측도

오즈비

오즈비 (θ) 값에 따른 의미

$\boldsymbol{\theta = 1}$: 두 행의 성공 오즈가 같음, 두 변수간 연관이 없음 (= **독립**)

$\boldsymbol{\theta > 1}$: **분자**의 성공의 오즈가 더 큼

$\boldsymbol{0 < \theta < 1}$: **분모**의 성공의 오즈가 더 큼

서로 **역수관계**에 있는 오즈비

→ **방향만 반대**일 뿐 두 변수간 **동일한 크기의 연관성**을 의미함

오즈비

로그오즈비 (Log Odds Ratio)

오즈비에 **로그**를 씌운 것

오즈비		
기준	1	
전체 범위	0~∞	
기준에 따른 범위	0~1	1~∞

로그 오즈비		
기준	0	
전체 범위	-∞~∞	
기준에 따른 범위	-∞~0	0~∞

4 연관성 측도

오즈비

로그오즈비 (Log Odds Ratio)

오즈비에 **로그**를 씌운 것

오즈비		
기준	1	
전체 범위	0~∞	
기준에 따른 범위	0~1	1~∞

비대칭적인 모습을 보임

↓

로그오즈비는 **비대칭적인**

오즈비의 범위를 교정한 측도 !

오즈비

로그오즈비 (Log Odds Ratio)

오즈비에 **로그**를 씌운 것

오즈비		
기준	1	
전체 범위	0~∞	
기준에 따른 범위	0~1	1~∞

오즈비에 로그를 씌운 결과

대칭적인 두 범위로 나뉨

로그 오즈비		
기준	0	
전체 범위	-∞~∞	
기준에 따른 범위	-∞~0	0~∞

4 연관성 측도

오즈비의 장점

장점

① 한 변수가 고정되어 있을 때도 사용 가능

② 행과 열의 위치가 바뀌어도 같은 값을 가짐

4 연관성 측도

오즈비의 장점

① 한 변수가 **고정되어 있을** 때도 사용 가능

알코올 중독	위암 유무		합
	위암 환자	건강한 사람	
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	위암 유무		합
	위암 환자	건강한 사람	
O	4 (4/10)	6 (6/10)	10
	4/6		
X	46 (46/340)	294 (294/340)	340
	46/294		
합	50	300	350

→ 오즈비는 **대조군의 크기가 달라져도** 동일한 값을 가짐

4 연관성 측도

오즈비의 장점

① 한 변수가 **고정되어 있을** 때도 사용 가능

알코올 중독	위암 유무		합
	위암 환자	건강한 사람	
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	위암 유무		합
	위암 환자	건강한 사람	
O	4 (4/10)	6 (6/10)	10
	4/6		
X	46 (46/340)	294 (294/340)	340
	46/294		
합	50	300	350

→ 오즈비는 **대조군의 크기가 달라져도** 동일한 값을 가짐

4 연관성 측도

오즈비의 장점

① 한 변수가 고정되어 있을 때도 사용 가능

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

알코올 중독	위암 유무		합
	위암 환자	건강한 사람	
O	4 (4/6)	2 (2/6)	6
X	46 (46/144)	98 (98/144)	144
합	50	100	150

알코올 중독	위암 유무		합
	위암 환자	건강한 사람	
O	4 (4/10)	6 (6/10)	10
X	46 (46/340)	294 (294/340)	340
합	50	300	350

비율의 차이, 상대위험도: 대조군의 크기 변함에 따라 달라짐

오즈비: 동일한 값을 가짐

→ 오즈비는 대조군의 크기가 달라져도 동일한 값을 가짐

오즈비의 장점

② 행과 열의 위치가 바뀌어도 같은 값을 가짐

알코올 중독	위암 유무		합
	위암 환자	건강한 사람	
O	4 (4/6)	2 (2/6)	6
	4/2		
X	46 (46/144)	98 (98/144)	144
	46/98		
합	50	100	150

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

위암 유무	알코올 중독		합
	O	X	
위암 환자	4 (4/50)	46 (46/50)	50
	4/46		
건강한 사람	2 (2/100)	98 (98/100)	100
	2/98		
합	6	144	150

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

오즈비의 장점

② 행과 열의 위치가 바뀌어도 같은 값을 가짐

오즈비는 $P(Y|X)$, $P(X|Y)$ 두 조건부 확률 중

어느 것을 사용하여 정의하든 **동일한 값**을 지니기 때문!

* 베이즈 정리를 이용한 증명

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

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

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

오즈비의 장점

② 행과 열의 위치가 바뀌어도 같은 값을 가짐

오즈비는 $P(Y|X)$, $P(X|Y)$ 두 조건부 확률 중
어느 것을 사용하여 정의하든 **동일한 값**을 지니기 때문!

* 베이즈 정리를 이용한 증명

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

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

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

4 연관성 측도

오즈비의 장점

장점

① 한 변수가 고정되어 있을 때도 사용 가능

② 행과 열의 위치가 바뀌어도 같은 값을 가짐

오즈비는 교차적비(cross-product ratio)이기 때문!

오즈비의 장점

교차적비 Cross-product ratio

분할표 상에서 대각선 반대편에 있는 칸의 확률들의 곱의 비

② 행과 열의 위치가 바뀌어도 같은 값을 가짐

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

오즈비는 교차적비(cross-product ratio)이기 때문!

4 연관성 측도

부분분할표에서의 연관성

조건부 연관성

Z의 값이 고정되어 있다는 조건 하에 X와 Y의 연관성

조건부 오즈비(Conditional Odds Ratio)를 통해 파악

학과(Z)	성별(X)	자취 여부(Y)		조건부 오즈비
		O	X	
통계	남자	11	25	$\theta_{XY(1)} = 1.188$
	여자	10	27	
경제	남자	16	4	$\theta_{XY(2)} = 1.818$
	여자	22	10	
경영	남자	14	5	$\theta_{XY(3)} = 4.8$
	여자	7	12	

경영학과 한정 남자가 자취할 오즈가 여자가 자취할 오즈보다 4.8배 높음

부분분할표에서의 연관성

조건부 연관성

Z의 값이 고정되어 있다는 조건 하에 X와 Y의 연관성

조건부 오즈비(Conditional Odds Ratio)를 통해 파악

학과(Z)	성별(X)	자취 여부(Y)		조건부 오즈비
		O	X	
통계	남자	11	25	$\theta_{XY(1)} = 1.188$
	여자	10	27	
경제	남자	16	4	$\theta_{XY(2)} = 1.818$
	여자	22	10	
경영	남자	14	5	$\theta_{XY(3)} = 4.8$
	여자	7	12	

경영학과 한정 남자가 자취할 오즈가 여자가 자취할 오즈보다 **4.8배** 높음

4 연관성 측도

부분분할표에서의 연관성

동질 연관성
(Homogeneous Association)

조건부 오즈비가 **모두 같은 경우**

대칭적: XY에 **동질 연관성** 존재 → YZ, XZ도 **동질 연관성** 존재

Special case

조건부 독립성
(Conditional Independence)

조건부 오즈비가 **모두 1로 같은 경우**

주변분할표에서의 연관성

주변 오즈비

제어변수를 합쳐버린 주변분할표에서의 오즈비

성별(X)	자취 여부(Y)		주변 오즈비
	O	X	
남자	11+16+15=41	25+4+5=34	$\theta_{XY+} = 0.148$
여자	10+22+7=39	27+10+12=49	

If 주변 오즈비$(\theta_{XY+}) = 1$

주변독립성을 가짐

주변 오즈비, 주변 독립성 = 2차원 분할표의 오즈비, 독립성

(부분분할표에서 파생되었다는 점에서 용어 구분을 함)

4 연관성 측도

주변분할표에서의 연관성

주변 오즈비

제어변수를 합쳐버린 주변분할표에서의 오즈비

성별(X)	자취 여부(Y)		주변 오즈비
	O	X	
남자	11+16+15=41	25+4+5=34	$\theta_{XY+} = 0.148$
여자	10+22+7=39	27+10+12=49	

If 주변오즈비$(\theta_{XY+}) = 1$

주변독립성을 가짐

주변 오즈비, 주변 독립성 = 2차원 분할표의 오즈비, 독립성

(부분분할표에서 파생되었다는 점에서 용어 구분을 함)

4 연관성 측도

주변분할표에서의 연관성

독립성 성립 ≠ 주변 독립성 성립

성별(X)	자취 여부(Y)		주변 오즈비
	O	X	
남자	11+16+15=41	25+4+5=34	$\theta_{XY+} = 0.148$
여자	10+22+7=39	27+10+12=49	

Why?

조건부 오즈비와 주변 오즈비의 **방향성**이 항상 같지는 않음

$\theta_{XY+}) = 1$

주변독립성을 가짐

주변 오즈비, 주변 독립성 = 2차원 분할표의 오즈비, 독립성

(부분분할표에서 파생되었다는 점에서 용어 구분을 함)

+ 곧 나올 심슨의 역설..

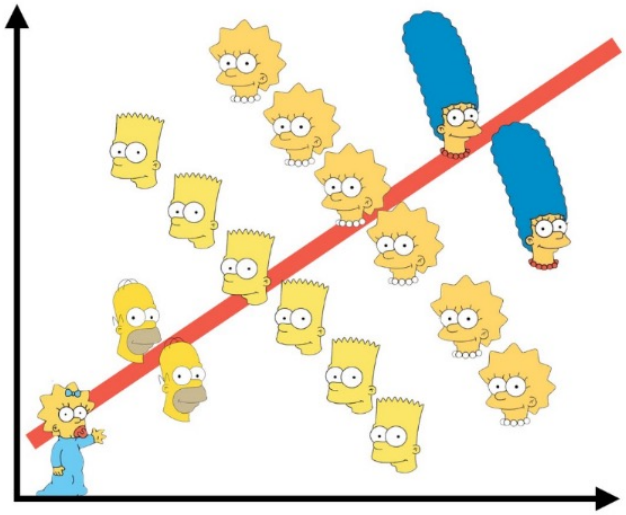
4 연관성 측도

심슨의 역설

심슨의 역설

전반적인 데이터들의 추세가 경향성이 존재하는 것처럼 보이지만

세부 그룹별로 나눠서 보면 **경향성이 사라지거나 반대로 해석되는 경우**



조건부 오즈비와 주변 오즈비가 의미하는

연관성의 방향이 서로 다르게 나타나는 경우

Ex) 심슨 가족 전체는 **우상향**, 각각의 가족 구성원은 **우하향**

4 연관성 측도

심슨의 역설

조건부 오즈비

$0 < \theta_{XY(1)}, \theta_{XY(2)} < 1$

주변 오즈비

$\theta_{XY+} > 1$

학과 (Z)	성별 (X)	자취 여부(Y)		조건부 오즈비
		O	X	
통계	남자	40	140	$\theta_{XY(1)} = 0.492$
	여자	9	64	
경제	남자	2	53	$\theta_{XY(2)} = 0.377$
	여자	1	10	

성별(X)	자취 여부(Y)		주변 오즈비
	O	X	
남자	41	193	$\theta_{XY+} = 1.61$
여자	10	74	

오즈비는 1을 기준으로 변수 간 연관성의 방향이 정해짐

위의 경우 조건부 오즈비와 주변 오즈비가 서로 반대 방향

→ 심슨의 역설 발생!

4 연관성 측도

심슨의 역설

조건부 오즈비 $0 < \theta_{XY(1)}, \theta_{XY(2)} < 1$

학과 (Z)	성별 (X)	자취 여부(Y)		조건부 오즈비
		O	X	
통계	남자	40	140	$\theta_{XY(1)} = 0.492$
	여자	9	64	
경제	남자	2	53	$\theta_{XY(2)} = 0.377$
	여자	1	10	

주변 오즈비 $\theta_{XY+} > 1$

성별(X)	자취 여부(Y)		주변 오즈비
	O	X	
남자	41	193	$\theta_{XY+} = 1.61$
여자	10	74	

오즈비는 **1을 기준으로** 변수 간 연관성의 방향이 정해짐

위의 경우 조건부 오즈비와 주변 오즈비가 서로 **반대 방향**

→ **심슨의 역설** 발생!

다음 주 예고

일반화 선형 모형(GLM)

유의성 검정

로지스틱 회귀모형

다범주 로짓 모형

포아송 회귀 모형