범주형자료분석팀

2팀
박지성
박지민
서희나
윤경선
이지윤

INDEX

1 범주형 자료분석

자료의 형태 | 질적 자료 (범주형 자료)

명목형 (Nominal) 자료

- 순서의 척도 없이 단순 분류된 자료
- 각 항목을 전환해도 성질의 변화가 없음
- 순서형 자료 분석 방법 사용 불가

피겨 단체전 종목			
남자 싱글	여자 싱글	남자 페어	아이스 댄스

1 범주형 자료분석

자료의 형태 | 질적 자료 (범주형 자료)

순서형 (Ordinal) 자료

- 순서의 척도가 있는 자료
- 명목형 자료 분석방법을 적용할 수 있지만 한계가 있다

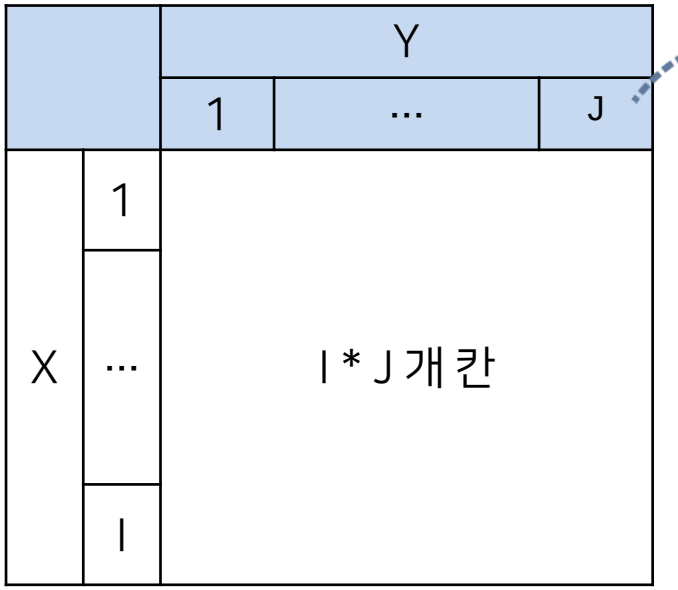
→ 순서에 대한 정보가 무시되기 때문

제품의 크기 정도				
초소형	소형	중형	대형	초대형

2 분할표

분할표 (Contingency Table)

각 범주형 변수에 속하는 결과의 도수들을 각 칸에 넣어 정리한 표



I : 변수 X의 **수준**의 개수

J : 변수 Y의 **수준**의 개수

수준 (Level)

- 범주형 변수가 취하는 값

2 분할표

여러 차원의 분할표 | 부분분할표

부분분할표				
학과(Z)	성별(X)	자취여부(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

제어변수(Z)의 각 수준에서 나머지 두 변수를 분류한 표

고정된 제어변수의 한 수준에서 X에 대한 Y의 효과 확인

2 분할표

비율에 대한 분할표 | 조건부 확률

설명변수(X)의 각 수준에서 반응변수(Y)에 대한 확률

비율에 대한 분할표			
성별(X)	자취여부(Y)		합계
	O	X	
남자	0.4	0.1	0.5
여자	0.2	0.3	0.5
합계	0.6	0.4	1

Ex) 남자인 경우에 자취를 할 **조건부확률** : 0.4/0.5 = 0.8

3 독립성 검정

독립성 검정의 가설

독립성 검정의 가설을
관측 도수와 기대 도수로 표현 가능

$$H_0 : \boldsymbol{n_{ij}}(= n_{++} \times \pi_{ij}) = \boldsymbol{\mu_{ij}}\,(= n_{++} \times \pi_{i+} \times \pi_{+j})$$

양변의 n_{++}를 지우면 결합확률 = 주변확률의 곱에 관한 식

독립성 검정은 관측 도수와 기대 도수의
차이를 비교하는 과정으로 이루어짐

3 독립성 검정

대표본 & 명목형 자료의 독립성 검정

피어슨 카이제곱 검정

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

가능도비 검정

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

검정 flow

관측 도수와 기대 도수의 차이 ↑ > 검정통계량 ↑

> P-value ↓ > 귀무가설 기각

변수 간 연관성 有

연관성 측도

두 범주형 변수가 **이항변수**일 때,
연관성을 나타내는 측도 3가지

비율의 비교 척도		
비율의 차이	상대 위험도	**오즈비**

4 연관성 측도

유의할 점 (비율의 차이, 상대위험도)

비율의 차이와 상대 위험도는 직관적인 척도

But! 후향적 연구처럼 **한 변수의 수를 고정**시킨 조사에서는 **사용 불가**

→ 대신 오즈비를 사용

후향적 연구 : 이미 나온 결과를 바탕으로 과거 기록을 관찰하는 연구

위암 환자 (사례군)	건강한 사람(대조군)	합
50	100	150

관측치를 랜덤하게 선택하지 않고

전체 표본에서 사례군의 비율을 1/3으로 고정

오즈비 (Odds Ratio)

오즈비(θ)의 값에 따른 의미

$\boldsymbol{\theta = 1}$: 두 행의 성공의 오즈가 같음, **독립**

(= 두 변수간 연관이 없음)

$\boldsymbol{\theta > 1}$: 분자의 성공의 오즈가 더 큼

$\boldsymbol{0 < \theta < 1}$: 분모의 성공의 오즈가 더 큼

서로 역수관계에 있는 오즈비

방향만 반대일 뿐, 두 변수간 동일한 크기의 연관성을 의미

4 연관성 측도

오즈비 (Odds Ratio) 장점

오즈비가 앞선 장점들을 가지는 이유 ➡ 오즈비는 교차적비

: 대각선 반대편에 있는 칸의 확률들의 곱의 비

위암 유무	알코올 중독 O	알코올 중독 X	합
위암 환자	[illegible]	[illegible]	[illegible]
건강한 사람	2	98	100
[illegible]	[illegible]	[illegible]	[illegible]

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

대각성분이 분자로, 비대각 성분이 분모로 간다

오즈비 (Odds Ratio) | 3차원 분할표

부분분할표에서의 연관성

동질 연관성(homogeneous association)

조건부 오즈비가 모두 같은 값을 가지는 경우 ($\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$)

- **대칭적** : XY에 동질 연관성 존재 → YZ와 XZ간에도 동질연관성이 존재

조건부 연관성(conditional association)

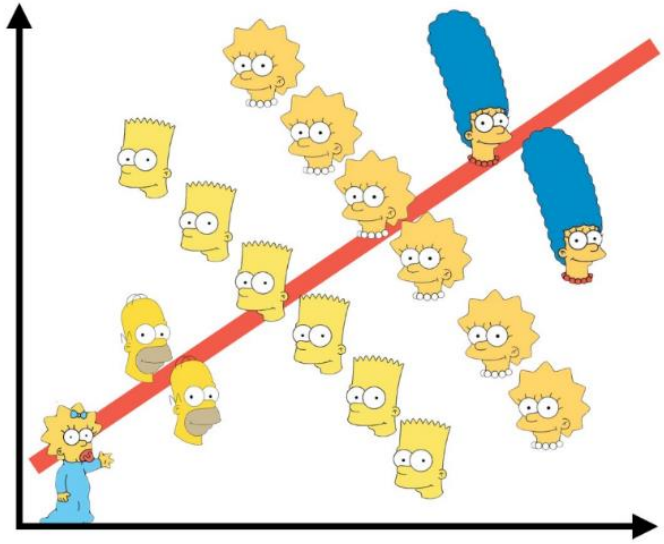
조건부 오즈비가 모두 1로 동일한 경우($\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)} = 1$)

- 어떠한 제어변수에서도 두 변수 X와 Y에 대한 오즈비가 1이라는 의미 → **독립**

오즈비 (Odds Ratio)

심슨의 역설 (Simpson's Paradox)

:전반적인 데이터의 추세에 경향성이 존재하는 것처럼 보이지만,
세부 그룹별로 나눠서 보면 **앞선 경향성이 사라지거나 반대로 해석**되는 경우



조건부 오즈비와 주변 오즈비가
의미하는 **연관성의 방향**이
서로 **다르게** 나타나는 경우