범주형자료분석팀

2팀

김찬영
이혜인
김서윤
심은주
진수정

INDEX

- 범주형 자료 분석은?

X 변수

: 독립 변수 / 설명 변수 / 예측 변수 / 위험인자 /공변량 [연속형] / 요인 [범주형]

Y 변수

: 종속 변수 / 반응 변수 / 결과 변수 / 표적 변수

Y변수가 범주형 자료 일 때 '범주형 자료분석'

범주형 자료

: 범주형 자료의 가장 큰 특징은 분할표를 작성할 수 있다는 것!

명목형 자료

: 순서척도가 없는 범주형 변수

Ex) 성별(F/M), 성공여부(Y/N), 혈액형(A/B/O/AB)

순서형 자료

: 순서척도가 있는 범주형 변수

Ex) 증상 정도(괜찮음/보통/심각), 순위(1등/2등/3등)

분할표

: 범주형 변수의 결과의 도수들을 각 칸에 넣어 표로 정리한 것

- 2차원 분할표 (I*J)

  :두 개의 변수만을 분류한 분할표

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

설명 변수 : X
반응 변수 : Y

- 3차원 분할표 (I*J*K)

  :세 개의 변수를 분류한 분할표

<부분분할표>

제어변수 Z의 각 수준에서 X와 Y를 분류한 표

제어변수 Z	설명변수 X	반응변수 Y: 합격	반응변수 Y: 불합격
학과	성별	학회 합격 여부	
통계	남자	11	25
	여자	10	27
경영	남자	16	4
	여자	22	10
경제	남자	14	5
	여자	7	12

학과(변수 Z) → 합쳐짐

<주변분할표>

부분분할표를 모두 결합해서 얻은 2차원분할표

성별	학회 합격 여부	
	합격	불합격
남자	11 + 16 + 14	25 + 4 + 5
여자	10 +22 + 7	27 + 10 + 12

변수 Z를 통제하는 것이 아니라 무시함.

Z 통제할 때 Y에 대한 X의 효과를 알 수 있음.

- 분할표 사용 목적

1 예측 검정력에 대한 요약

예를 들어, 2 * 2 형태의 2차원 분할표에서 민감도와 특이도, Accuracy 등을 찾을 수 있음

2 독립성 검정 실시

제시된 변수끼리의 연관성 파악

"독립성 검정"

: 변수 간에 독립성 유무를 검정하는데 많이 사용되는 가설검정

귀무가설: $\mu_{ij} = n\pi_{ij}$ ➡ 변수들이 서로 독립 O!

대립가설: $\mu_{ij} \neq n\pi_{ij}$ ➡ 변수들이 서로 독립 X!

변수들이 서로 독립 ➡ 두 변수가 연관성 X ➡ 분석 가치 X

- 2차원 분할표 독립성 검정

대표본
- 명목형
  - 피어슨 카이제곱 검정
  - 가능도비 검정
- 순서형
  - 멘탈 스코어 검정

소표본
- 피셔의 정확 검정

- 3차원 분할표 독립성 검정

로그 선형 모형 비교

3차원 이상의 고차원 모형은
모형으로 다루는 것이 효과적!

"피어슨 카이제곱 검정"

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

- 모든 n_{ij}가 μ_{ij}와 같을 때, 최소값 0을 가짐
- n_{ij}와 μ_{ij} 사이의 차이가 커지면 X^2가 커져서 귀무가설을 기각하는 증거가 강해짐
- $\mu\_ij \geq 5$ 정도(대표본)이라면 카이제곱 분포를 따름

"가능도비 검정" : 자료가 대표본 & 명목형일 때

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

관측도수(n_{ij})와 기대도수(μ_{ij})의 차이가 크다

→ G^2 증가 귀무가설 기각

→ 변수 간의 연관성이 있다

"멘탈스코어 검정" : 자료가 대표본 & 순서형일 때

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

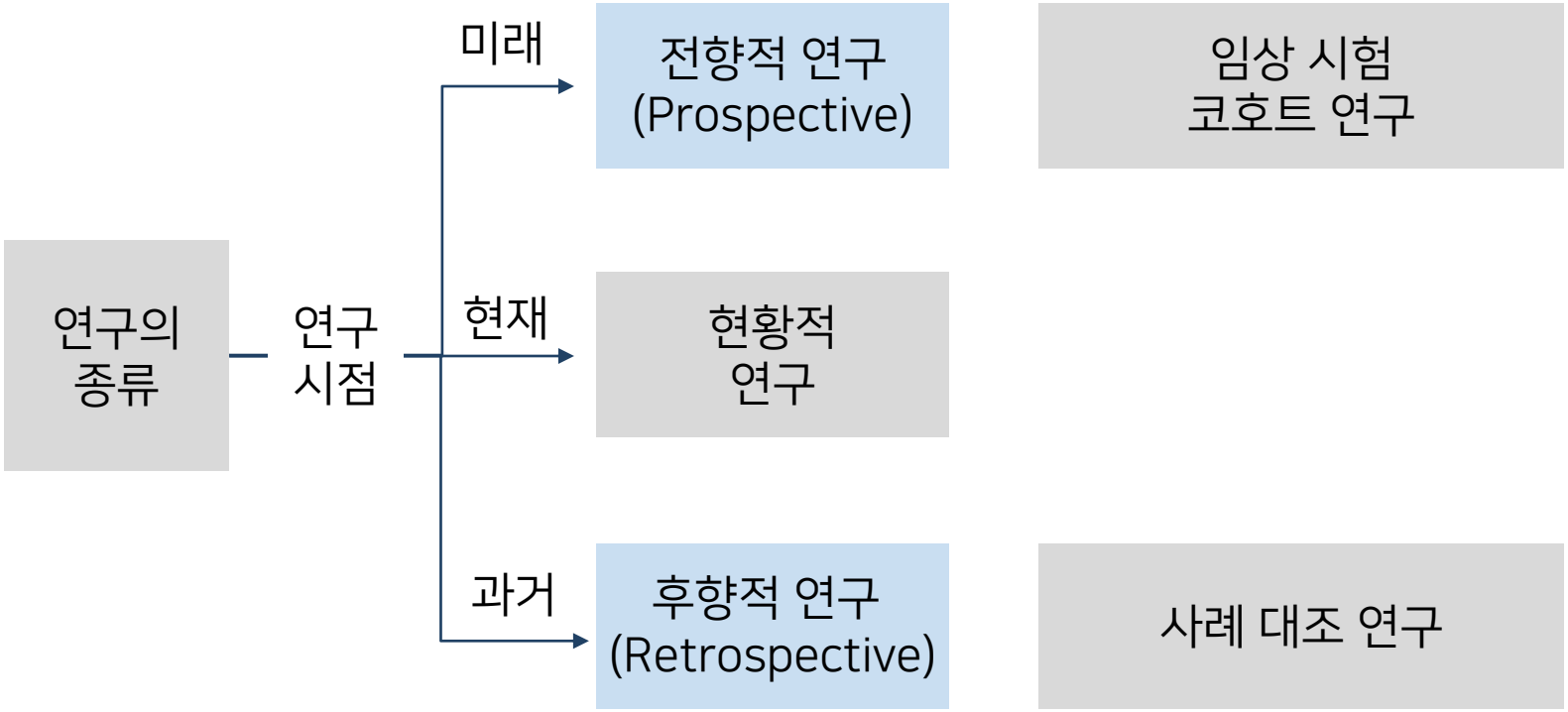
n과
r(피어슨 교차적률 상관계수)이 크다

→ M^2 증가
귀무가설 기각

→ 변수 간의
연관성이 있다

* r(피어슨 교차적률 상관계수): 변수 간의 추세 연관성 파악 가능. $-1 \le r \le 1$ ($r = 0$일때 독립)

4 연구의 종류


연구의 종류
연구 시점
미래
현재
과거
전향적 연구 (Prospective)
현황적 연구
후향적 연구 (Retrospective)
임상 시험 코호트 연구
사례 대조 연구

"오즈비" : 여러 모형에서 기초가 되는 모수

각 행의 오즈끼리의 비 : Odds ratio(θ) = $\frac{_1\text{행의 오즈}}{_2\text{행의 오즈}} = \frac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)}$

성별	연인 여부		Odds
	예	아니오	
여성	0.8144	0.1856	4.3879
남성	0.7928	0.2072	3.8262

<오즈비>

$$\frac{4.3879}{3.8262} = 1.1468$$

"여성이 연인이 있을 오즈가 남성이 연인이 있을 오즈보다 약 1.15배 높다"

- 3차원 분할표에서 오즈비

"조건부 오즈비"

- 동질연관성

$$\theta_{XY(1)} = \ \ldots = \theta_{XY(\boldsymbol{K})}$$

동질연관성은 대칭적이다!

→ XY에 동질연관성 존재하면, YZ, XZ도 동질연관성이 존재한다!

- 조건부독립성

$$\theta_{XY(1)} = \ \ldots = \theta_{XY(\boldsymbol{K})} = \mathbf{1}$$

동질연관성이 특별한 경우!

"주변 오즈비"

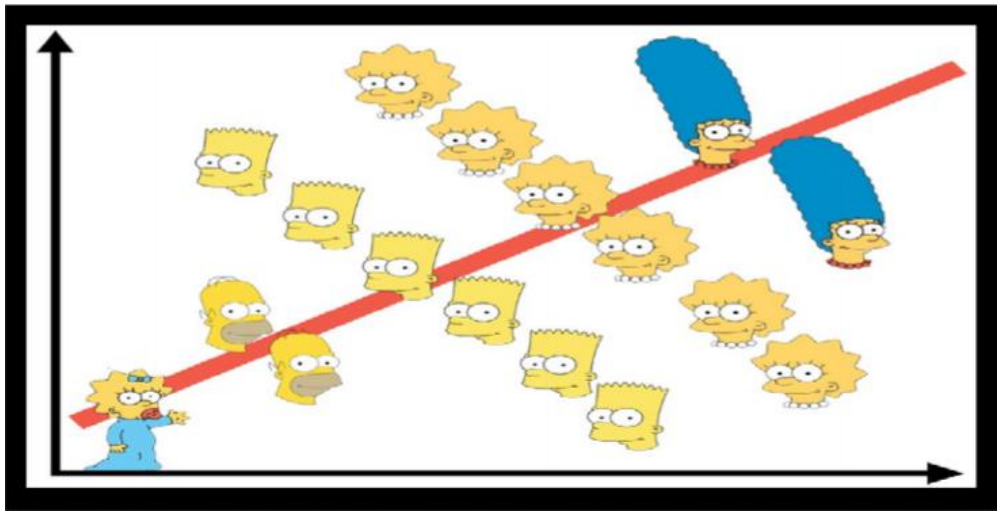
- 주변독립성

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

- 3차원 분할표에서 오즈비

"Simpson의 역설"

- 조건부오즈비와 주변오즈비의 연관성 방향이 다른 경우를 뜻함
- 도수의 크기에 따른 영향력 차이로 인해 나타남



즉, 조건부연관성과 주변연관성이 다를 수 있다는 것!