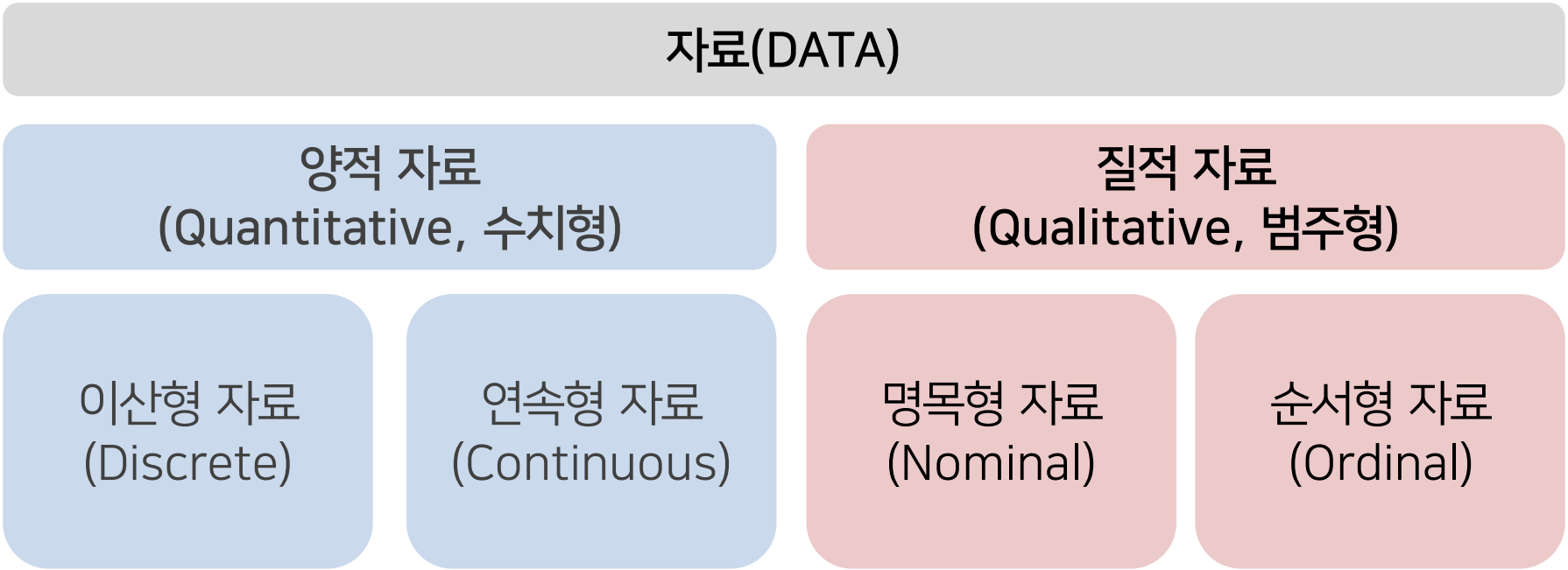
범주형자료분석팀

2팀
이지연
심예진
조장희
조혜현
진효주

INDEX

범주형 자료분석

자료의 형태



범주형 자료분석

자료의 형태

질적 자료 (Qualitative)

명목형 자료

순서형 자료 (Ordinal)

순서 척도가 있는 범주형 변수

1~5 별점으로 나타내는 영화 평점				
싫어함 (1)	좋아하지 않음 (2)	좋아함 (3)	아주 좋아함 (4)	사랑함 (5)

변수 간의 순서 O
순서형 자료 분석 방법 가능

분할표

분할표란? 「범주형 자료 변수」에 대해서만 만들 수 있음

분할표

여러 개의 범주형 변수를 기준으로 관측치를 기록하는 표

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

Ex) 2차원 분할표의 형태

분할표

여러 차원의 분할표

3차원 분할표(I*J*K)

X(설명변수), Y(종속변수), **Z(제어변수)**

고정된 Z의 한 수준에 대해서
XY의 관계를 보여줌
Z를 통제했을 때
Y에 대한 X의 효과를 알 수 있음

		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}

「부분분할표」

: Z의 각 수준에서 X와 Y를 분류한 표

독립성 검정

가설

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

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

분할표의 각 칸의 발생확률(π_{ij}) = 각 교차표의 주변확률(π_{i+} , π_{+j})의 곱

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

즉, π_{11} = π_{1+} X π_{+1} 이면

두 변수는 독립이라고 할 수 있다!

독립성 검정

기대도수와 관측도수

기대 도수와 관측 도수를 이용해 앞의 가설을 이렇게 바꿔 쓸 수 있다!

귀무가설 H_0 : 두 범주형 변수는 독립이다 ($\mu_{ij} = n \cdot \pi_{ij}$)

대립가설 H_1 : 두 범주형 변수는 독립이 아니다 ($\mu_{ij} \neq n \cdot \pi_{ij}$)

즉, 기대 도수와 관측 도수의 차이 ($\mu_{ij} - n \cdot \pi_{ij}$) 가

유의미하게 크다면, 귀무 가설을 기각할 가능성이 커지게 됨!

독립성 검정

명목형 자료 검정

피어슨 카이제곱 검정(Pearson's chi-squared test)

- 검정통계량 : $\mathrm{X}^2 = \sum \frac{(\mathrm{n_{ij}} - \mu_{\mathrm{ij}})^2}{\mu_{\mathrm{ij}}} \sim \chi^2_{(\mathrm{I}-1)(\mathrm{J}-1)}$
- 기각역 : $\mathrm{X}^2 \geq \chi^2_{\alpha,(\mathrm{I}-1)(\mathrm{J}-1)}$

가능도비 검정 (Likelihood-ratio test)

- 검정통계량 : $\mathrm{G}^2 = 2\sum \mathrm{n_{ij}} \log\left(\frac{\mathrm{n_{ij}}}{\mu_{\mathrm{ij}}}\right) \sim \chi^2_{(\mathrm{I}-1)(\mathrm{J}-1)}$
- 기각역 : $\mathrm{G}^2 \geq \chi^2_{\alpha,(\mathrm{I}-1)(\mathrm{J}-1)}$

독립성 검정

순서형 자료 검정

순서형 자료에 명목형 독립성 검정을 할 수는 있지만, 순서 정보의 손실이 일어나기 때문에 비추!

MH 검정 (Mantel-Haenszel test)

MH 검정통계량

$$M^2 = (n-1)r^2$$

- 조건 : 두 범주형 변수 모두 순서형일 때 사용
- 원리 : 범주의 각 level에 점수를 할당하여 변수 간의 선형 추세 측정!
- 추세 연관성을 파악하기 위해 피어슨 교차적률 상관계수 사용! 없는데가 없는 갓-피어슨…
- 기각역 : $M^2 \geq \chi^2_{\alpha,1}$
- n과 r이 커지면 검정통계량 M^2도 커진다. -> 귀무가설 기각 -> 변수 간 연관성!

3

독립성 검정의

검정 통계량의 값

연속형 자료에서 공

즉, 독립성 검정

연관이 있다고

변수 간 연관성

연관성 성질은 혼자 알아보기엔 위험하단다!
이 아이들 중 하나를 데려가렴.

비율의 차이 (Difference of Proportions)	상대 위험도 (Relative Risk)	오즈비 (Odds Ratio)

연관성 측도

오즈비 (Odds Ratio, OR)

오즈 (Odds) 란?

어떤 일이 일어날 승산 (공산) , 또는 가능성

성공확률 / 실패확률을 의미

$$\text{odds} = \frac{\pi}{1-\pi}$$

π = 성공확률

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…	

첫 번째 행의 오즈는 4.388, 두 번째 행의 오즈는 3.826

조건부 독립성과 주변 독립성

조건부 연관성 (Conditional Association)

부분분할표에서의 연관성

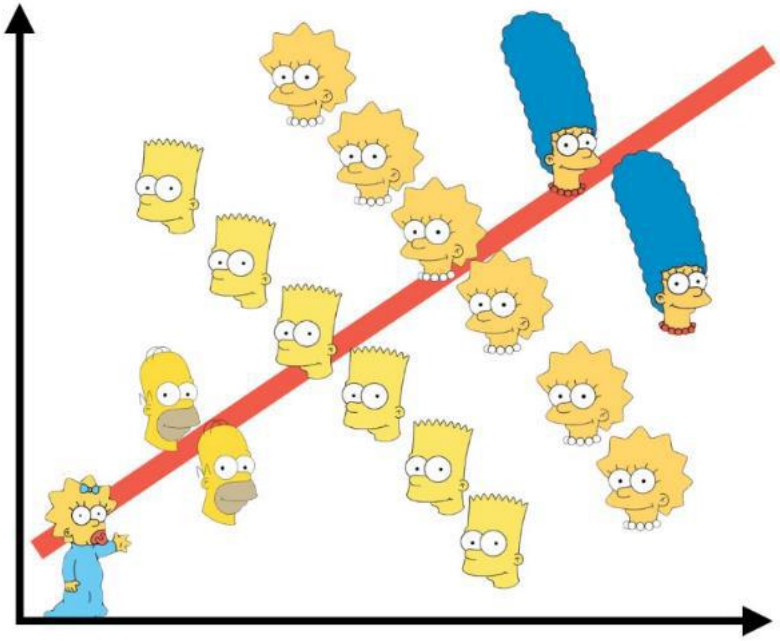
제어변수 Z의 값이 어떤 수준에서 고정되어 있다는 조건 하에서 X와 Y의 연관성

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

심슨의 역설 (Simpsons's Paradox)

전반적인 추세가 경향성이 존재하는 것으로 보이지만

그룹으로 나누어 보면 경향성이 사라지거나 해석이 반대로 되는 경우



조건부 오즈비와 주변 오즈비의

연관성 방향이 다르게 나타나는 경우!

THANK YOU