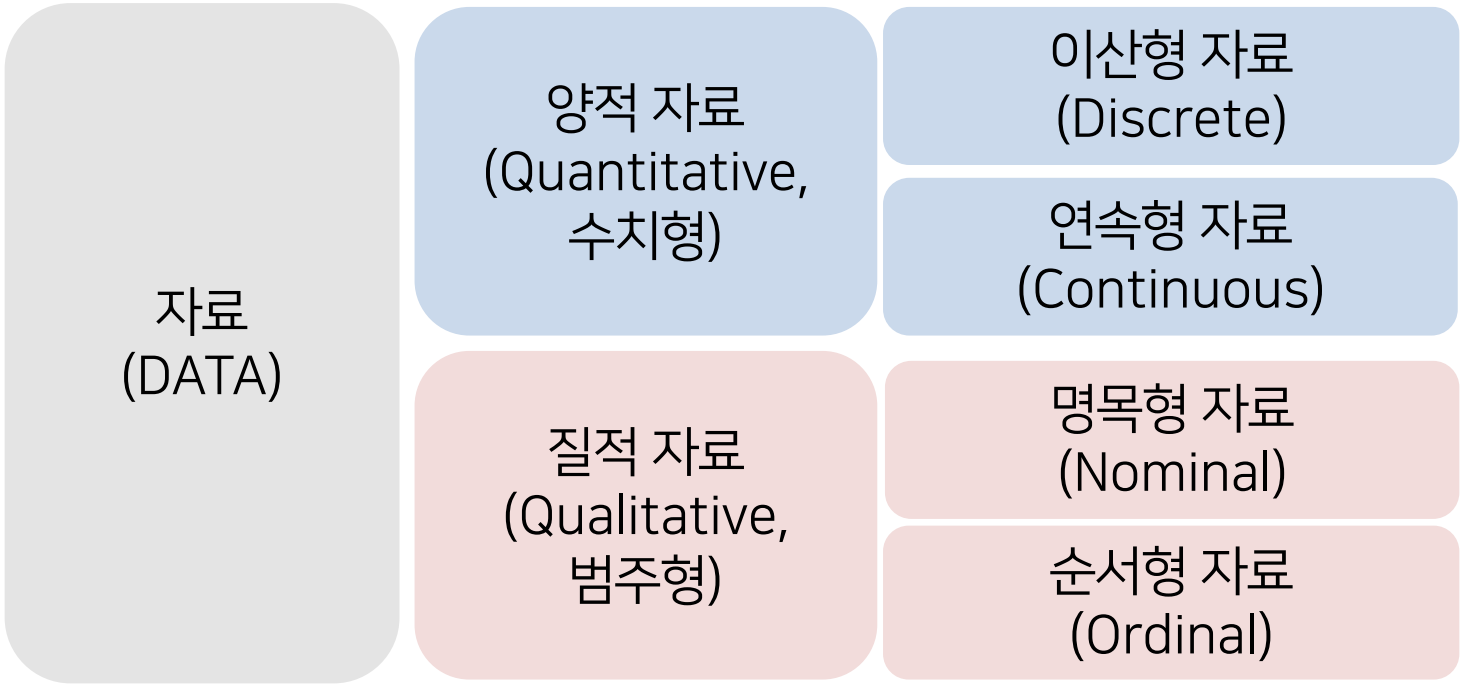
범주형자료분석팀

2팀
조장희
위재성
김지현
조수미
송지현
김민지

INDEX

범주형 자료분석

자료의 형태



분할표

분할표란?

분할표

		Y			
		1	2	…	J
X	1	I * J 개 칸			
	2				
	…				
	I				

범주형 변수의 관측치를 기록한 표

분할표

여러 차원의 분할표

3차원 분할표

부분분할표

		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}

주변분할표

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

2차원 분할표에서 제어변수 Z 추가

분할표

비율에 대한 분할표

결합 확률

X의 i번째 수준과 Y의 j번째 수준을 동시에 만족하는 확률

	BMI			합계
	정상 (Y=1)	과체중 (Y=2)	비만 (Y=3)	
남자 (X=1)	78(0.31)	15(0.06)	46(0.19)	139(0.56)
여자 (X=2)	49(0.19)	23(0.09)	37(0.15)	109(0.43)
합계	127(0.5)	38(0.15)	83(0.34)	248(1)

전체 인원 중에서

남자이고 BMI가 정상일 확률

$P(X = 1, Y = 1)$ = 0.31

즉, 결합 확률 = 각 칸의 확률

독립성 검정

관측도수 & 기대도수

관측도수

실제 관측값

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

기대도수

도수의 기댓값

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

관측 도수와 기대 도수의 차이를 비교

$H_0 : \pi_{ij} = \pi_{i+} \times \pi_{+j}$ (같은 가설)

$H_1 : \pi_{ij} \neq \pi_{i+} \times \pi_{+j}$

$H_0 : \mu_{ij} = n\pi_{ij}$

$H_1 : \mu_{ij} \neq n\pi_{ij}$

3 독립성 검정

명목형 자료

1. 피어슨 카이제곱 검정

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

2. 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

관측도수와 기대 도수의 차이 ↑ → 검정통계량 ↑ → P-value ↓

→ H_0 기각 가능성 ↑ → 변수 간의 연관성O

피어슨 교차적률 상관계수(r)

순서형 자료

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

잠깐! r이 뭐죠?

M^2 =(공분산/표준편차의 곱) 인 상관계수의 같은 형식

무슨 상관입니까?

▶공통점

- $-1 \leq r \leq 1$의 범위
- r 값이 0에 가까울수록 변수 간의 연관성 ↓

▶ 차이점

범주 수준에 점수를 할당

- $u_1 \leq u_2 \leq \cdots \leq u_I$,
- $v_1 \leq v_2 \leq \cdots \leq v_J$).

검정 flow

n과r ↑ → 검정통계량 ↑ → P-value ↓

H_0 기각 가능성 ↑ → 변수 간의 연관성O

연관성 측도

두 범주형 변수가 **이항 변수**일 때, 연관성을 나타내는 측도

(이항 변수 간의 연관성을 나타내는 측도)

비율의 차이	상대 위험도	오즈비

비율의 차이

비율의 차이 = 조건부 확률의 **차이**

$$(\pi_1 - \pi_2)$$

π_i = i 번째 행의 조건부 확률

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

여성일 때 애인이 있을 확률이

남성일 때보다 0.814 - 0.793 = 0.0216 만큼 더 높음

연관성 측도

오즈비 (Odds Ratio, OR)

오즈 (Odds) 성공확률 / 실패확률 을 의미

$$\text{odds} = \frac{\pi}{1-\pi} \quad (\pi = \text{성공확률})$$

성별	애인 유무	
	있음	없음
여성	509(0.814)	116(0.186)
	0.814/0.186 = 4.388 = 오즈	
남성	398(0.793)	104(0.207)
	0.793/0.207 = 3.826 = 오즈	

여성의 입장에서 애인이 있을 확률은 애인이 없을 확률의 4.388배

오즈비 (Odds Ratio, OR)

오즈비 각 오즈의 비

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

- 범위 : $\theta \geq 0$
- 역수관계의 오즈비는 방향만 반대이고, 연관성은 같음

  오즈비가 4인 경우 연관성 = 오즈비가 0.25(1/4)인 경우 연관성

오즈비 (Odds Ratio, OR)

성별	애인 유무	
	있음	없음
여성	509 (0.814)	116 (0.186)
	오즈 = 0.814/0.186 = 4.388	
남성	398 (0.793)	104 (0.207)
	오즈 = 0.793/0.207 = 3.826	

오즈비는 $\theta = \frac{4.388}{3.826} = 1.147$ 임

여성이 애인이 있을 오즈가 남성이 애인이 있을 오즈보다

1.147배 더 높음

오즈비

장점

오즈비 장점의 이유는 바로 교차적비

$$\theta = \frac{\pi_1/(1-\pi_{1)}}{\pi_2/(1-\pi_{2)}} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

분할표의 대각성분의 곱이 분자로, 비대각성분이 분모로 감.

변수가 고정된 상태에서 대조군이 바뀌더라도 값 유지.

행과 열을 바꾸더라도 값 유지.

오즈비

3차원 분할표에서의 오즈비

조건부 독립성 성립 ≠ 주변 독립성 성립

부분분할표				
학과(Z)	성별(X)	연애여부(Y)		조건부 오즈비
		O	X	
국어국문	남자	18	12	$\theta_{XY(1)}=\frac{18/12}{12/8}=1$
	여자	12	8	
문헌정보	남자	2	8	$\theta_{XY(2)}=\frac{2/8}{8/32}=1$
	여자	8	32	

주변분할표			
성별(X)	연애여부(Y)		주변 오즈비
	O	X	
남자	20	20	$\theta_{XY+}=\frac{20/20}{20/40}=2$
여자	20	40	