범주형자료분석팀

2팀
이지연
심예진
조장희
조혜현
진효주

INDEX

1

범주형 자료분석

범주형 자료분석

변수와 자료

변수

자료 수집의 대상이 되는 모집단의 특성

관측치

변수의 측정치

자료(DATA)

변수와 관측치로 이루어진 모임
열: 변수/행:측정치

범주형 자료분석

변수의 구분

X변수

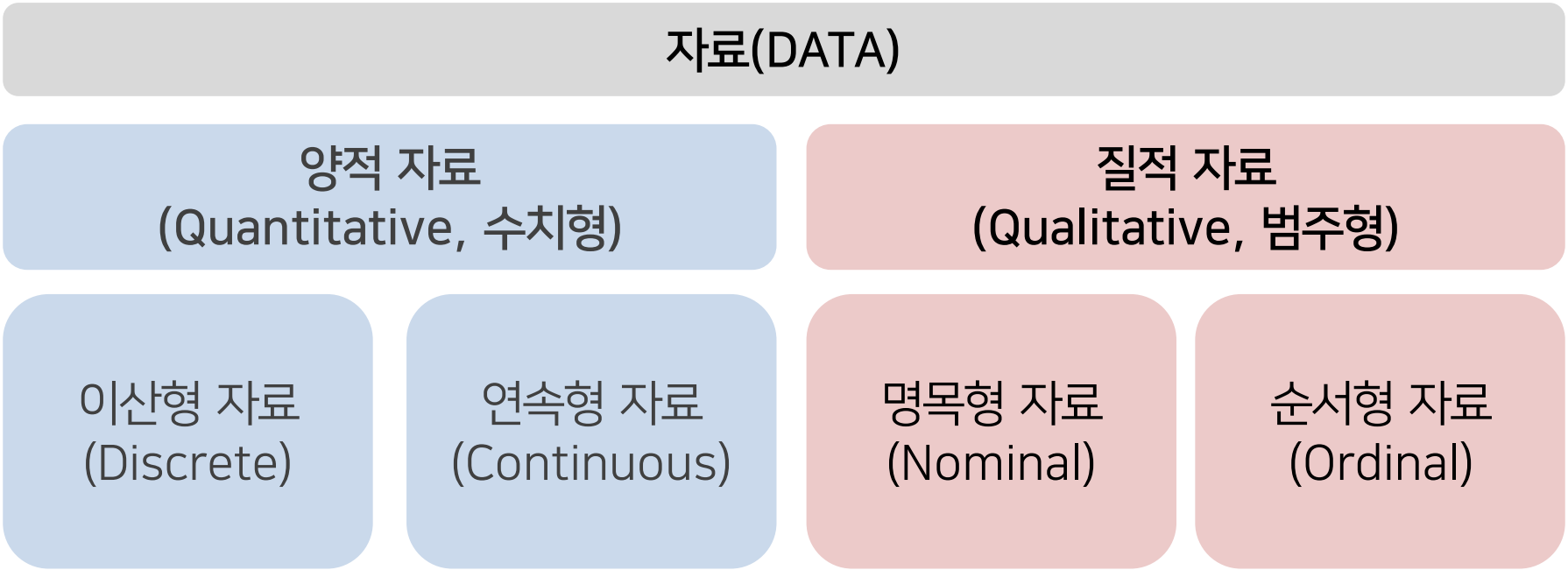
독립변수/설명변수/예측변수/위험 인자/공변량/요인

Y변수

종속변수/반응변수/결과변수/표적변수

범주형 자료분석

자료의 형태



1 범주형 자료분석

자료의 형태

자료(DATA)

양적 자료
(Quantitative)

질적 자료
(Qualitative)

이산형 자료
(Discrete)

연속형 자료
(Continuous)

명목형 자료
(Nominal)

순서형 자료
(Ordinal)

범주형 자료분석이란?

Y변수가 범주형 자료일 때의 분석을 의미

범주형 자료가 뭐였더라? 자료의 형태를 알아보자!

범주형 자료분석

자료의 형태

양적 자료
(Quantitative)

이산형 자료
(Discrete)

- 이산적인 값을 갖는 데이터

Ex) 자녀의 수, 사건 발생 수, 나이 등

연속형 자료
(Continuous)

- 연속적인 값을 갖는 데이터

Ex) 신장, 체중, 온도 등

범주형 자료분석

자료의 형태

질적 자료 (Qualitative)

명목형 자료 (Nominal)

순서형 자료

순서 척도가 없는 범주형 변수

혈액형			
A	B	AB	O

변수 간의 순서 X
순서형 자료 분석 방법 사용 불가

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

2

분할표

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

2차원 분할표(I*J)

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

$\boldsymbol{n_{ij}}$: 각 칸의 도수

$\boldsymbol{n_{i+}, n_{+j}}$: 각 열과 행의 주변(marginal) 도수

+ 첨자는 그 위치에 해당하는 도수를 모두 더했다는 의미의 첨자!

분할표

여러 차원의 분할표

3차원 분할표(I*J*K)

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

「주변분할표」

		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」

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

분할표

여러 차원의 분할표

3차원 분할표(I*J*K)

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

「주변분할표」

: 제어변수를 모두 결합해서 무시한 표

X(설명변수), Y(종속변수)

부분분할표를 결합해서 얻은

2차원 분할표로

Z를 통제하지 않고 무시함

비율에 대한 분할표

결합확률

표본이 두 범주형 반응 변수 X와 Y로 분류될 때,
X의 i번째 수준과 Y의 j번째 수준을 **동시에 만족**하는 확률

$$\sum \pi_{ij} = 1$$

	아메리카노	라떼	카푸치노	합계
남성	78(0.31)	15(0.06)	46(0.19)	139(0.56)
여성	49(0.19)	23(0.09)	37(0.15)	109(0.44)
합계	127	38	83	248

「성별에 따른 커피 취향」

비율에 대한 분할표

주변확률

결합분포의 각 행과 열의 **합**

$$\pi_{i+} = \sum_{n=1}^{j} \pi_{in} \qquad \pi_{+j} = \sum_{n=1}^{i} \pi_{nj}$$

	아메리카노	라떼	카푸치노	합계
남성	78(0.31)	15(0.06)	46(0.19)	139(0.56)
여성	49(0.19)	23(0.09)	37(0.15)	109(0.44)
합계	127(0.5)	38(0.15)	83(0.34)	248

「성별에 따른 커피 취향」

분할표

비율에 대한 분할표

조건부확률

X의 각 수준에서의 Y에 대한 확률

$$\frac{\pi_{ij}}{\pi_{i+}}$$

	아메리카노	라떼	카푸치노	합계
남성	78(0.56)	15(0.11)	46(0.33)	139(1)
여성	49(0.45)	23(0.21)	37(0.34)	109(1)
합계	127	38	83	248

「여성 중 아메리카노가 취향일 확률」

3

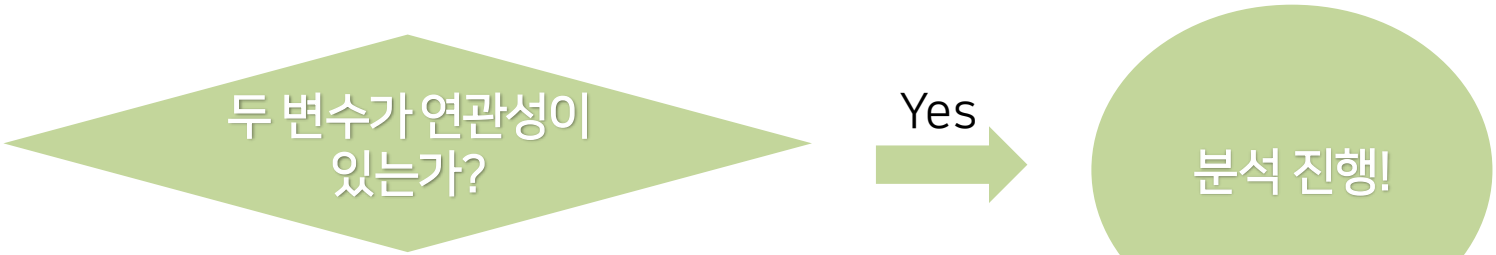
독립성 검정

독립성 검정

독립성 검정

독립성 검정

두 범주형 변수가 **독립**인지 검정하는 것!



"이런 데이터는 아예 쓸 데가 없어요"

범주형 변수는 연속형 변수의 상관계수처럼
변수 간의 상관성을 나타내기 **어려움**

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

기대 도수와 관측 도수

기대 도수 expected frequency $[\mu_{ij}]$

두 변수가 독립일 때 각 칸의 도수에 대한 **기댓값** $E(n_{ij})$

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

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

「확률 테이블」

X 전체도수 n

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

「기대도수 테이블」

독립성 검정

기대 도수와 관측 도수

관측 도수 observed frequency $[n_{ij}]$

각 칸에 해당하는 관측 수

: 전체 n개의 관찰 값 중 i행 j열에 해당 하는 값

ex) n_{11} = 분할표에서 1행 1열의 값

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

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

「확률 테이블」

X 전체도수 n

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

「관측도수 테이블」

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

독립성 검정의 종류

2차원 분할표 독립성 검정

대표본	명목형	피어슨 카이제곱 검정 (Pearson's chi-squared test)
		가능도비 검정 (Likelihood-ratio test)
	순서형	MH 검정 (Mantel-Haenszel test)
소표본		피셔의 정확검정 (Fisher's Exact test)

2차원 분할표 독립성 검정 ?

- 3차원 이상의 고차원에서는 변수 간의 관계를 모형으로 다루는 것이 효과적
- *로그 선형 모형*

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

명목형 자료 검정

피어슨 카이제곱 검정통계량

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

가능도비 검정통계량

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

피어슨 카이제곱 검정과 가능도비 검정은
둘 다 같은 flow를 갖는다!

3 독립성 검정

명목형 자료 검정

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

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

"

기대 도수와 관측 도수의 차이 가 크다는 것은,
검정통계량 X^2 나 G^2 이 크다는 것이다.
분자에 있거나 log 안에 있으니까.

검정통계량 X^2 나 G^2 이 크다는 것은,
귀무가설을 기각한다는 것이다. p-value가 작으니까.

귀무가설을 기각한다는 것은,
변수 간에 연관성이 존재한다는 것이다.
그것이 Fun cool sexy한 검정이니까.

"

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!

순서형 자료 검정

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

피어슨 교차적률 상관계수가 뭐죠?

MH 검정 (Mantel-Haenszel test)

★ 피어슨 교차적률 상관계수 ★

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

$(-1 \le r \le 1\ ,\ r = 0$ 일 때 독립)

- 조건 : 두 범주형 변
- 원리 : 범주의 각 lev
- 추세 연관성을 파악하기 위해 피어슨 교차적률 상관계수 사용! 없는데가 없는 갓-피어슨…
  복잡해 보이지만 우리가 아는 그 상관계수와 비슷하다~!
- 기각역 : $M^2 \ge \chi^2_{\alpha,1}$
- n과 r이 커지면 검정통계량 M^2도 커진다. -> 귀무가설 기각 -> 변수 간 연관성!

독립성 검정의 한계

- 검정 통계량의 값이 엄청 크다고 해서 변수 간 연관성이 더 큰 것은 아님
  연속형 자료에서 공분산이 크다고 상관관계가 큰 것은 아닌 것처럼…
- 즉, 독립성 검정은 변수 간 연관성의 유무 여부만 판단하는 한계가 있음
- 연관이 있다고 판단되는 변수들이 얼마나 연관됐는 지는 알 수 없음
- 변수 간 연관성의 성질을 파악하기 위해 연관성 측도를 알아야 함!

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)

4

연관성 측도

연관성 측도

비율의 차이

범주형 변수가 이항 변수일 때,

여러가지 방법으로 범주별 비교가 가능

[이항변수 간의 연관성을 나타내는 측도]

비율의 비교 척도		
비율의 차이	상대 위험도	오즈비

연관성 측도

비율의 차이

비율의 차이 : 조건부 확률의 차이 $(\pi_1 - \pi_2)$

(π_i = i번째 행의 조건부 확률)

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

범주팀 대대로 내려오는 연인 유무 예시를 통해 알아보자 ^^….

연관성 측도

비율의 차이

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

비율의 차이 = (위의 행의 성공확률 – 밑의 행의 성공확률)

0.814 - 0.793 = 0.021

➡ 여성일 경우, 연인이 있을 확률이 0.021만큼 더 높다고 해석

연관성 측도

비율의 차이

범위 : $-1 \leq \pi_1 - \pi_2 \leq 1$

독립일 경우 : $\pi_1 - \pi_2 = 0$

성별	연인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

비율의 차 : 0.4 - 0.4 = 0

반응변수와 설명변수가 독립

상대위험도 (Relative Risk, RR)

조건부 확률의 비 = $\frac{\pi_1}{\pi_2}$

상대위험도가 클 수록 변수 간 연관성이 큼

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

상대 위험도 = 0.814/0.793 = 1.027

➡ 여성일 경우, 연인이 있을 확률이 1.027배 높다고 해석

연관성 측도

상대위험도 (Relative Risk, RR)

범위 : $\frac{\pi_1}{\pi_2} \geq 0$

독립일 경우 : $\frac{\pi_1}{\pi_2} = 1$

성별	연인 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

성별	연인 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

비율의 차이는 0.02 - 0.01 = 0.92 - 0.91 = 0.01로 매우 작지만,

상대 위험도는 0.02/0.01 = 2, 0.92/0.91 = 1.01로 영향력 차이를 보임

연관성 측도

오즈비 (Odds Ratio, OR)

비율의 차이와 상대위험도는 직관적이지만,

한 변수의 수를 고정시킨 조사에서는 사용이 불가함

	심장질환 있음 (Y = 1)	심장질환 없음(Y = 0)	합
알코올 중독 O (X = 1)	4	2	6
알코올 중독 X (X = 0)	46	98	144
합	50	100	150

관측치를 랜덤하게 선택하지 않고,

전체 표본 중 심장질환자의 비율을 1/3로 고정해서 추출

➡ 비율의 차이와 상대위험도 대신에 **오즈비를 사용**

오즈비 (Odds Ratio, OR)

후향적 연구란?

비율의 차이와 상대위험도는 직관적이지만,
한 변수의 수를 고정시킨 조사에서는 사용이 불가함

이미 난 결과를 바탕으로 과거기록을 관찰하는 연구

Ex) 심장 질환 Ex) 알코올 중독

	심장질환 있음 (Y = 1)	심장질환 없음(Y = 0)	합
알코올 중독 O (X = 1)	4	2	6
알코올 중독 X (X = 0)	46	98	144
합	50	100	150

후향적 연구에서는 열이 고정되어

비율의 차와 상대위험도를 사용할 수 없다!

관측치를 랜덤하게 선택하지 않
전체 표본 중 심장질환자의 비율을 1/3로 고
비율의 차이와 상대위험도 대신에 오즈비

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

연관성 측도

오즈비 (Odds Ratio, OR)

오즈비 (Odds Ratio) 란?
각 오즈의 비

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

- 범위 : $\theta \geq 0$
- 독립일 때 : $\theta = 1$ (= 두 행에서 성공의 오즈가 같다.)
- $\theta > 1$ 이면 첫번째 행에서의 성공의 오즈가 두번째 행보다 높음
- $0 < \theta < 1$ 이면 첫번째 행에서의 성공의 오즈가 두번째 행보다 낮음
- 역수 관계의 오즈비는 방향만 반대이고 연관성은 같음

연관성 측도

오즈비 (Odds Ratio, OR)

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

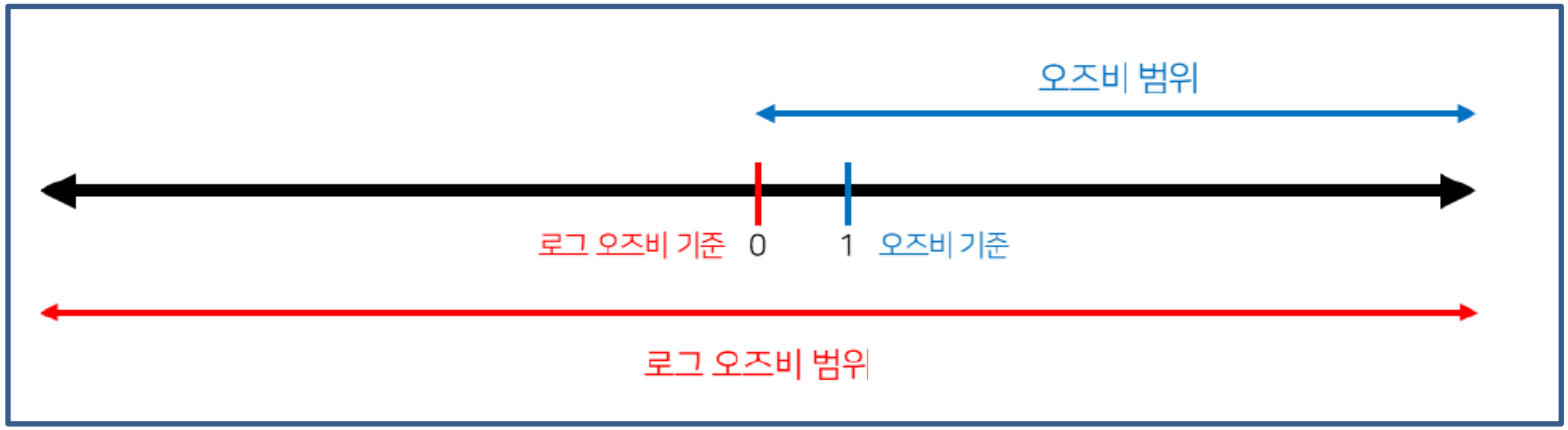
오즈비 = 4.388 / 3.826 = 1.147

➡ 여성이 연인이 있을 오즈가 남성이 연인이 있을 오즈보다 1.147배 더 높다고 해석

로그 오즈비 (Log Odds Ratio)

로그 오즈비란?

기존의 비대칭한 오즈의 범위를 교정한 측도



기존 오즈비의 기준인 1이 0이 되면서 범위가 대칭으로 교정

$$-\infty < \log\theta < \infty$$

오즈비의 장점

1. 한 변수가 고정되어 있을 때도 사용 가능
대조군의 크기가 달라져도 오즈비는 같음

성별	연인 유무		합
	있음	없음	
여성	10 (1/4)	30 (3/4)	40
	1/3		
남성	20 (1/3)	40 (2/3)	60
	1/2		
합	30	70	100

비율을 3:7로 고정한 후향적 연구를 진행

연관성 측도

오즈비의 장점

성별	연인 유무		합
	있음	없음	
여성	10 (1/4)	30 (3/4)	40
	1/3		
남성	20 (1/3)	40 (2/3)	60
	1/2		
합	30	70	100

성별	연인 유무		합
	있음	없음	
여성	10 (1/4)	300 (3/4)	310
	1/30		
남성	20 (1/3)	400 (2/3)	420
	1/20		
합	30	700	730

대조군(연인 없음)의 크기가 70에서 700으로 바뀌어

비율의 차와 상대위험도는 바뀌었지만 오즈비는 같음!

$$\frac{1/3}{1/2} = \frac{1/30}{1/20} = \frac{2}{3}$$

오즈비의 장점

오즈비의 값을 P(Y|X)를 사용해 정의하나 P(X|Y)로 정의하나 동일하기 때문!

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

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

연관성 측도

오즈비의 장점

2. 오즈비는 행과 열의 위치가 바뀌어도 같음

성별	연인 유무		합
	있음	없음	
여성	10 (1/4)	30 (3/4)	40
	1/3		
남성	20 (1/3)	40 (2/3)	60
	1/2		
합	30	70	100

$$\frac{1/4 \times 2/3}{3/4 \times 1/3} = \frac{1/6}{1/4} = \frac{2}{3}$$

연인	성별		합
	여성	남성	
있음	10 (1/3)	20 (2/3)	30
	1/2		
없음	30 (3/7)	40 (4/7)	70
	3/4		
합	40	60	100

$$\frac{1/3 \times 4/7}{2/3 \times 3/7} = \frac{4/21}{6/21} = \frac{2}{3}$$

연관성 측도

오즈비의 장점

오즈비 = 교차적비(cross-product ratio)

대각선에 있는 칸 도수 혹은 확률의 곱

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

정말 멋있는 오즈비 …

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	

연관성 측도

조건부 독립성과 주변 독립성

동질 연관성 (Homogeneous Association)

조건부 오즈비가 모두 같은 경우 [$\theta_{XY(1)} = \theta_{XY(2)} = \cdots$]

Z의 각 수준에서 XY의 연관성이 모두 같음

XY에 동질연관성이 존재하면 YZ와 XZ에도 동질연관성이 존재

조건부 독립성 (Conditional Association)

조건부 오즈비가 모두 1로 같은 경우 [$\theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1$]

동질연관성의 특별한 경우

조건부 독립성과 주변 독립성

「제어변수 Z를 합친 주변분할표」

주변분할표			
성별 (X)	대학원 진학 여부 (Y)		주변 오즈비
	진학	비진학	
남자	11 + 16 + 14 = 41	25 + 4 + 5 = 34	θ_{XY+} = 0.148
여자	10 + 22 + 7 = 39	27 + 10 + 12 = 49	

주변오즈비 : 주변분할표에서의 오즈비

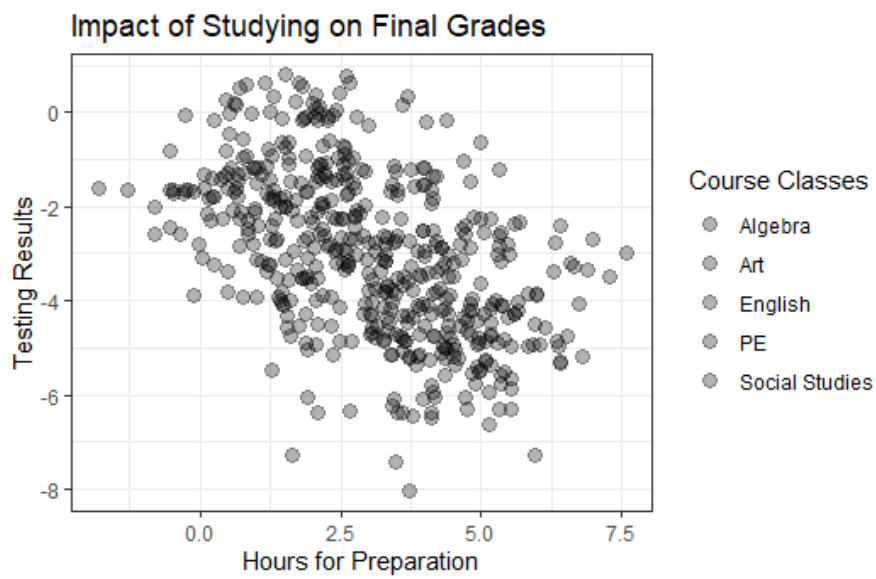
주변 오즈비가 1일 때 [$\boldsymbol{\theta_{XY+}}$ =1] 주변독립성을 가짐

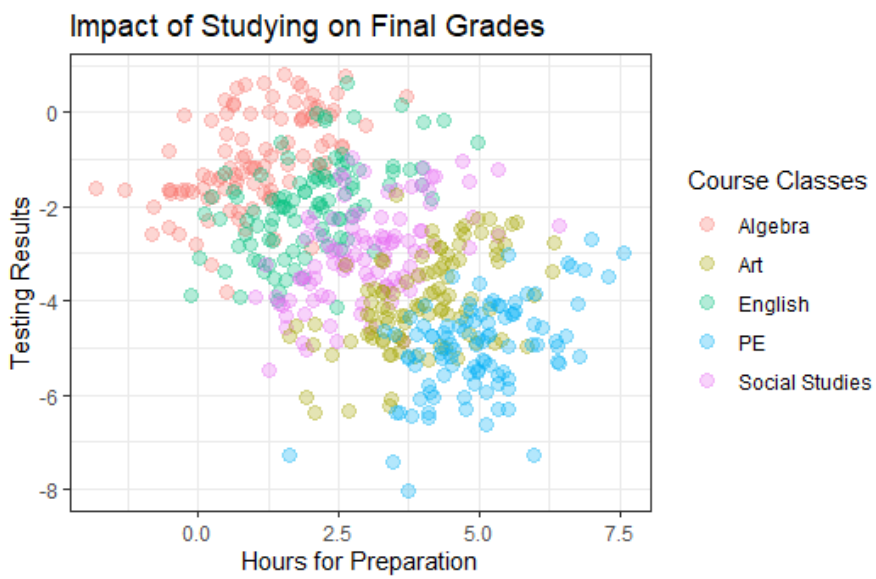
연관성 측도

조건부 독립성과 주변 독립성

조건부독립성이 성립한다고 해서
주변독립성이 성립되는 것은 아님!

제어변수 Z를 합친 주변분할표





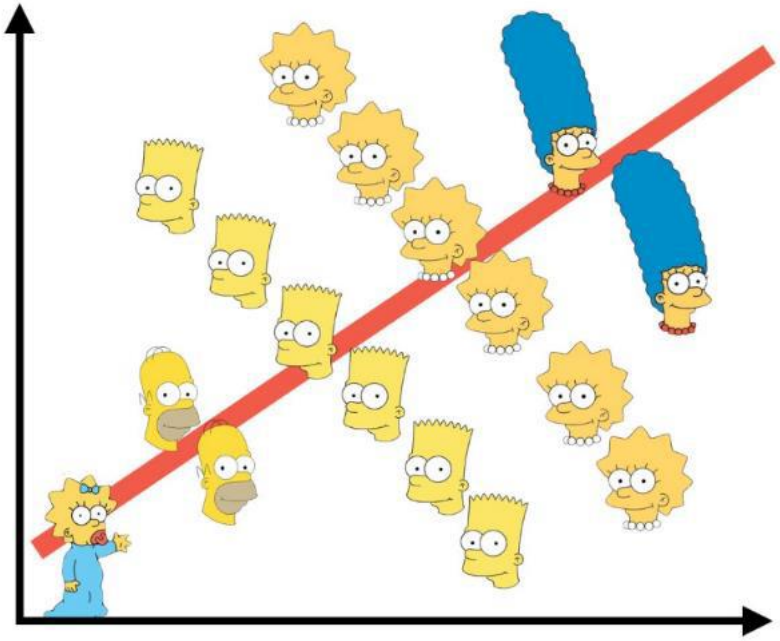
주변 오즈비가 1일 때 $[\theta_{XY+}=1]$ 주변독립성을 가짐

이 그래프를 해석해보자!

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!

심슨의 역설 (Simpsons's Paradox)

부분분할표			
학과	성별	대학원 진학 여부	
		진학	비진학
통계	남자	53	414
	여자	11	37
경영	남자	0	16
	여자	4	139

주변분할표		
성별	대학원 진학 여부	
	진학	비진학
남자	53 + 0 = 53	414 + 16 = 430
여자	11 + 4 = 15	16 + 139 = 155

조건부 오즈비

$\theta_{XY(1)} = 0.43,\ \theta_{XY(2)} = 0$

주변 오즈비

$\theta_{XY+} = 1.446$

오즈비는 1이 기준이므로 연관성 방향이 반대

→ **제어변수인 학과가 중요한 변수로 작용**

THANK YOU