범주형자료분석팀

2팀

김찬영
이혜인
김서윤
심은주
진수정

INDEX

1

자료의 형태

- 변수 구분

X 변수

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

Y 변수

: **종속 변수 / 반응 변수 / 결과 변수** / 표적 변수

- 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'

- 자료의 형태

자료

- 양적 자료 (Quantitative)
  - 이산형 자료 (Discrete)
  - 연속형 자료 (Continuous)
- 질적 자료 (Qualitative)
  - 명목형 자료 (Nominal)
  - 순서형 자료 (Ordinal)

양적자료 정규분포

: 측정이나 셈 같은 수량의 형태를 가진 자료

이산형 자료

: 이산적인 값을 갖는 데이터

Ex) 자녀의 수, 사건 발생 수

연속형 자료

: 연속적인 값을 갖는 데이터

Ex) 신장, 체중

범주형 자료

이항분포 / 다항분포 / 포아송 분포 / 음이항 분포

: 측정의 단위가 여러 범주들의 집합으로 구성되어 있는 자료

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)

범주형 자료

이항분포 / 다항분포 / 포아송 분포 / 음이항 분포

: 측정의 단위가 여러 범주들의 집합으로 구성되어 있는 자료

명목형 자료

순서형 자료에 대한 분석방법 사용 불가능!

순서형 자료

명목형 자료에 대한 분석방법 사용 가능!

- BUT! 순서에 대한 정보 무시 -> 심각한 검정력 손실
- 순서형 자료에 일정 점수 할당해 양적자료로 다루기도 함

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)

2

분할표

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

학과	성별	학회 합격 여부	
		합격	불합격
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

- 3차원 분할표 (I*J*K)

  :세 개의 변수를 분류한 분할표

<부분분할표>

제어변수 Z의 각 수준에서 X와 Y를 분류한 표

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

Z 통제할 때 Y에 대한 X의 효과를 알 수 있음.

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.

부분분할표 vs. 주변분할표

- 3차원 분할표 (I*J*K)

  :세 개의 변수를 분류한 분할표

<부분분할표>

제어변수 Z의 각 수준에서 X와 Y를 분류한 표

학과	성별	학회 합격 여부	
		합격	불합격
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

A-HA! Z의 여부에 따라 이렇게 구분되는구나~ 이-지이지!

- 비율에 대한 분할표

  :각 칸을 전체 도수 n으로 나누어 줌

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

$\boldsymbol{\pi_{ij} = n_{ij} \div n}$ →

X \ Y	Y		합계
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
합계	π_{+1}	π_{+2}	$\pi_{++} = 1$

이를 활용하여 다양한 분할표에서의 확률 분포를 구할 수 있다!

- 분할표에서 확률 분포

결합 확률 (joint probability)

$\boldsymbol{\pi_{ij}}$: i행과 j열에 속할 확률인 X와 Y의 결합 분포

$\sum_{i,j} \boldsymbol{\pi_{ij}} = \mathbf{1}$: 모든 확률 합을 더하면 1

	Y		**합계**
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
합계	π_{+1}	π_{+2}	$\pi_{++} = 1$

π_{12}는 1행 2열에 속하는 확률인 X와 Y의 결합 분포

π_{ij}의 전체 합은 1

- 분할표에서 확률 분포

주변 확률 (marginal probability)

: 결합분포의 행과 열의 합으로 각각 정의

π_{i+} : 행의 분포
π_{+j} : 열의 분포

	Y		합계
X	π_{11}	π_{12}	π_{1+}
	π_{21}	π_{22}	π_{2+}
합계	π_{+1}	π_{+2}	$\pi_{++} = 1$

π_{1+}와 π_{2+}은 각 행에 대한 확률의 합

- 분할표에서 확률 분포

조건부 확률 (conditional probability)

: X가 주어졌을 때에 Y에 대한 조건부 확률

$\frac{\pi_{ij}}{\pi_{i+}}$: 조건부 분포

	Y_1	Y_2	합계
X_1	π_1	$1-\pi_1$	1
X_2	π_2	$1-\pi_2$	1

<예시>

학과	연인여부		합계
	예	아니오	
통계	0.8144	0.1856	1
경영	0.7928	0.2072	1

→ '통계' 중, 연인이 있을 확률 !

~~이거 맞죠..? 자료의 신뢰도가..^^~~

- 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

3

독립성검정

"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

"독립성 검정"

조건부 확률 (관측 도수)

$$\boldsymbol{n_{ij} = n * \pi_{ij}}$$

기대 도수(expected frequency)

: 주변확률의 곱으로 만들어 짐

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

→ 귀무가설이 참일 때의 기댓값 $E(\boldsymbol{n_{ij}})$

두 값이 얼마나 일치하는지 비교하는 것!

"독립성 검정"

조건부 확률 (관측 도수)

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

기대 도수(expected frequency)

: 주변확률의 곱으로 만들어 짐

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

→ 귀무가설이 참일 때의 기댓값 $E(n_{ij})$

즉, 결합 확률이 주변 확률의 곱과 일치하는 지 확인하는 것

$\pi_{ij} = \pi_{i+} * \pi_{+j}$면 독립!

변수끼리 독립이면 주변확률을 통해 결합확률을 구할 수 있음

- 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
- μ_{ij}≥5 정도(대표본)이라면 카이제곱 분포를 따름

독립성 검정의 FLOW..

관측 도수와 기대 도수 차이↑ → 검정통계량 X^2↑ → P-값↓ → 귀무가설 기-각 → 변수 간의 연관성 有

이 FLOW는 다음 주 범주와도 함께하니.. 지금 여러분들 머리속에 저-장☆

"가능도비 검정" : 자료가 대표본 & 명목형일 때

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

관측도수(n_{ij})와 기대도수(μ_{ij})의 차이가 크다

→

G^2 증가 귀무가설 기각

→

변수 간의 연관성이 있다

피어슨 카이제곱 검정 vs. 가능도비 검정

- 변수의 범주를 명목형으로 다룬다.
- 표본이 충분히 커야 한다!
- 두 검정 통계량은 표본이 클 때 χ^2로 수렴하고, 수치적으로 유사한 값을 가진다.
- 카이제곱 분할은 G^2만 가능!

```
> matrix2_2
   couple
sex   O   X
  F 185  94
  M 227  25
```

```
> matrix2_2 %>% assocstats()
                    X^2 df   P(> X^2)
Likelihood Ratio 45.536  1 1.4983e-11
Pearson          43.028  1 5.3953e-11
```

"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)

피어슨 교차적률 상관계수가 뭐죠?

$$r = \frac{\sum_{ij}(\mu_i - \bar{\mu})(v_i - \bar{v})p_{ij}}{\sqrt{\left[\sum_i(\mu_i - \bar{\mu})^2 p_{i+}\right]\left[\sum_j(v_i - \bar{v})^2 p_{+j}\right]}}$$

복잡해 보이지만, 우리가 아는 상관계수랑 같은 감성..

변수 간의 **'추세 연관성'** 파악 가능하다!

$$-1 \leq r \leq 1$$

$r = 0$일 때, 독립

"멘탈스코어 검정" *: 자료가 대표본 & 순서형일 때*

- 범주 수준에 점수를 할당하여 선형추세를 측정
- 두 변수 모두 순서형이거나, 두 변수 중 한 변수가 두 가지 범주를 갖는 명목형 변수일 때 사용 가능
- 순서형 자료에 명목형 자료의 검정 방법(X^2 or G^2)을 쓰면 정보의 손실 발생

(자료의 형태에도 다 검정의 짝이 있는 법.. 행복해라 순서형 자료!)

- 독립성 검정의 한계

변수 간의 연관성이 있는지 여부만 알 수 있을 뿐

어떻게 연관되어 있는지는 알 수 없음!

확률의 비교를 알아야 한다!

4

연구의 종류

4 연구의 종류

연구의 종류 — 연구 시점

미래 → 전향적 연구 (Prospective) — 임상 시험, 코호트 연구

현재 → 현황적 연구

과거 → 후향적 연구 (Retrospective) — 사례 대조 연구

"전향적 연구" : 연구시점이 미래일 때

: 하나의 표본을 추적해 가는 연구

<예시> 성별에 따른 맥주 취향 분석

연구자가 그룹을 선정한 뒤 결과 관찰

	맥주 취향(Y)			
성별 (X)		가벼운	보통	흑
	남성	51	56	25
	여성	39	21	8

"전향적 연구" : 연구시점이 ***미래***일 때

- 설명변수(X) 고정됨
- 행의 분포 고정
- 행의 합(n_{i+}) 고정

	맥주 취향(Y)			
성별 (X)		가벼운	보통	흑
	남성	51	56	25
	여성	39	21	8

"후향적 연구" : 연구시점이 과거일 때

: 이미 나온 결과를 바탕으로 과거 기록을 관찰하는 연구

<예시> 전공에 따른 주별 공부시간을 확인하는 연구

➡ 연구자가 결과를 선정한 뒤 과거의 원인 관찰

		주별 공부시간(Y)		
전공 (X)		0-10	11-20	20 이상
	인문	68	119	70
	사회과학	106	103	52
	경영	131	127	51
	공학	40	81	52

"후향적 연구" : 연구시점이 과거일 때

- 반응변수(Y) 고정됨
- 열의 분포 고정
- 열의 합(n_{+j}) 고정

<예시> 전공에 따른 주별 공부시간을 확인하는 연구

		주별 공부시간(Y)		
전공 (X)		0-10	11-20	20 이상
	인문	68	119	70
	사회과학	106	103	52
	경영	131	127	51
	공학	40	81	52

후향적 연구는 확률의 비교 측도에서

~~비율의 차~~ ~~상대위험도~~

오즈비만 사용 가능

위대한 오-즈비는 조금 이따 알아보도록 하자! ㄱㄷㄱㄷ

4 연구의 종류

종류 전향적 연구 후향적 연구

“후향적 연구”

- 반응변수(Y) 고정됨
- 열의 분포 고정
- 열의 합(n_{+j}) 고정

<예시> 전공에 따른 주별 공부시간을 확인하는 연구

5

확률의 비교

- 확률의 비교 척도

이항반응변수에 대하여 **두 그룹을 비교하는 척도**들을 제시

확률의 비교 척도

비율의 차이

상대 위험도

오즈비

* 확률의 비교에서는 **조건부 확률**을 사용한다!

조건부 확률의 차: $\pi_1 - \pi_2$

<도수 분할표>

X	Y		합계
	Y_1	Y_2	
X_1	n_{11}	n_{12}	n_{1+}
X_2	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

<비율에 대한 분할표>

X	Y		합계
	Y_1	Y_2	
X_1	$\boldsymbol{\pi_1}$	$1-\pi_1$	1
X_2	$\boldsymbol{\pi_2}$	$1-\pi_2$	1

<예시>

비율의 차이 = 위의 행의 성공확률 – 밑의 행의 성공확률

성별	연인 여부		합계
	예	아니오	
여성	509 (0.8144)	116 (0.1856)	625 (1)
남성	398 (0.7928)	104 (0.2072)	502 (1)
합계	907	220	1127

0.8144 – 0.7928 = 0.0216

여성일 경우 연인이 있을 확률이 0.0216만큼 높다!

- 값

  1. 범위: -1 ≤ $\pi_1 - \pi_2$ ≤ 1

  2. 독립: $\pi_1 - \pi_2 = 0$

성별	연인 여부	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

0.4 – 0.4 = 0

여성일 때 연인이 있을 확률이 남성일 때와 차이가 없다

성별이 연인 여부에 영향을 끼치지 않는다

반응변수와 설명변수는 독립!

- 단점

1. 후향적 연구에서 사용하지 못함

- 열이 고정되어 있어 $P(Y = 1|X = 1)$이 아닌 $P(X = 1|Y = 1)$만 알 수 있음.

- 추후 오즈비 파트에서 자세히 다뤄보자!

2. 0이나 1에 가까워질수록 차이를 제대로 반영하지 못함

- 다음 내용인 상대위험도와의 비교를 통해 더 자세히 알아보자!

- 정의

조건부 확률의 비: $\frac{\pi_1}{\pi_2}$ → 아까는 조건부 확률끼리의 '차이'였다면 상대 위험도는 '비율 ' 이다.

성별	연인 여부		합계
	예	아니오	
여성	0.8144	0.1856	1
남성	0.7928	0.2072	1

이렇게 되어 있을 때 상대 위험도는 0.8144 / 0.7928 = 1.027…

즉, 여성일 경우 연인이 있을 확률이 1.027배 높다는 것!

- 값

1. 범위: $\frac{\pi_1}{\pi_2} \geq 0$ 2. 독립: $\frac{\pi_1}{\pi_2} = 1$

성별	연인 여부	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

0.4 / 0.4 = 1

여성일 때 연인이 있을 확률과 남성일 때 확률의 비가 1

성별이 연인 여부에 영향을 끼치지 않는다

반응변수와 설명변수는 독립!

- 단점

1. 후향적 연구에서 사용하지 못함

- 확률의 차와 같은 이유이다..!

- 추후 오즈비 파트에서 자세히 다뤄보자!

2. 0이나 1에 가까워질수록 차이를 제대로 반영하지 못함

- 단점

1. 후향적 연구에서 사용하지 못함

- 확률의 차와 같은 이유이다..!

- 추후 오즈비 파트에서 자세히 다뤄보자!

2. 0이나 1에 가까워질수록 차이를 제대로 반영하지 못함

무슨말이야..? 다음 장에서 자세하게 알아보아보자!

다들 재밌게 듣고있죠^^?

- 단점

<0에 가까울 때>

성별	연인 여부	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

<1에 가까울 때>

성별	연인 여부	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

비율의 차이 : 0.02 – 0.01 = 0.92 – 0.91 = 0.01

상대 위험도 : 0.02 / 0.01 = 2 > 0.92 / 0.91 = 1.01

비율의 차이는 같지만, 상대위험도는 차이가 크다!

비율의 차이 vs 상대 위험도

비율의 차이

0이나 1에 가까울수록 차이를 잘 나타내지 못한다

0.02 – 0.01 = 0.01 = 0.92 – 0.91 = 0.01

비율의 차가 동일!

상대 위험도

0이나 1에 가까울수록 값의 차이가 크다

0.02 / 0.01 = 2 > 0.92 / 0.91 = 1.01

0에 가까울수록 크다!

"오즈 (Odds)" : 실패에 비해 성공이 몇 배인가

$$\text{Odds} = \frac{\text{성공}}{\text{실패}} = \frac{\pi}{1-\pi}$$

$$\pi = \frac{odds}{1+odds}$$

성별	연인 여부	
	예	아니오
여성	509 (0.8144)	116 (0.1856)
	0.8144/0.1856 = 4.3879⋯	
남성	398 (0.7928)	104 (0.2072)
	0.7928/0.2072 = 3.8262⋯	

"오즈비" : 여러 모형에서 기초가 되는 모수

각 행의 오즈끼리의 비 : Odds ratio(θ) = $\dfrac{1\text{행의 오즈}}{2\text{행의 오즈}} = \dfrac{\pi_1/(1-\pi_1)}{\pi_2(1-\pi_2)}$

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"

- 단점

1. 범위: $\theta \geq 0$

2. 독립: θ = 1 → 두 행에서의 성공의 오즈가 같다는 의미이므로..!

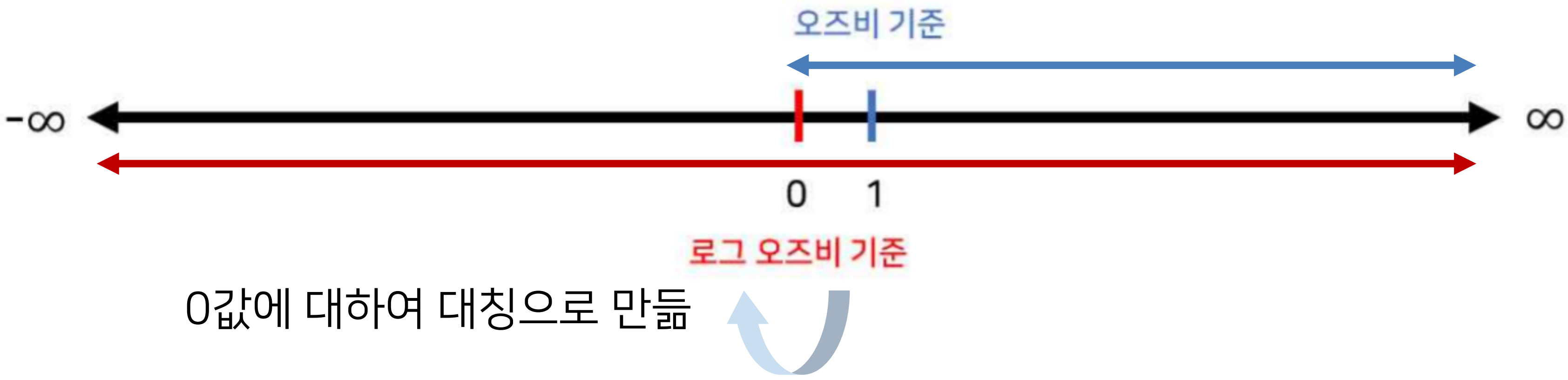
- θ > 1: 1행의 성공의 오즈 > 2행의 성공의 오즈
- 0 ≤ θ < 1: 1행의 성공의 오즈 < 2행의 성공의 오즈
- 역수 관계에 있는 오즈비: 같은 정도의 연관성을 뜻한다
  단지 방향만 반대일 뿐!

→ 2와 ½는 같은 정도의 연관성을 의미한다!

"로그 오즈 비" : 오즈비에 log를 씌운 것

→ 비대칭한 범위 교정

기존 오즈비는 1에 대해 대칭



0값에 대하여 대칭으로 만듦

log1 = 0.

- 오즈비의 특징

  1. 교차적비로 쉽게 구할 수 있음.

<비율 분할표>

성별	연인 여부	
	있음	없음
여성	0.48	0.52
남성	0.4	0.56

<도수 분할표>

성별	연인 여부	
	있음	없음
여성	24	26
남성	44	56

$$\text{오즈비} = \frac{0.48/0.52}{0.44/0.56} = \frac{0.48 * 0.52}{0.44 * 0.56} = \frac{24 * 56}{44 * 26} = 1.1748$$

- 오즈비의 특징

  1. 교차적비로 쉽게 구할 수 있음.

  2. 도수가 0일 경우, 0.5를 더해서 오즈비 계산

<도수 분할표>

성별	연인 여부	
	있음	없음
여성	0	26
남성	44	56

→

<도수 분할표>

성별	연인 여부	
	있음	없음
여성	0.5	26.5
남성	44.5	56.5

$$\text{오즈비} = \frac{0 \ * 56}{44 \ * 26} = 0$$

$$\text{오즈비} = \frac{0.5 \ * 56.5}{44.5 \ * 26.5} = 0.0239$$

- 오즈비의 특징

  1. 교차적비로 쉽게 구할 수 있음.

  2. 도수가 0일 경우, 0.5를 더해서 오즈비 계산

  3. 각 행의 조건부확률이 0에 가깝다면 오즈비와 상대 위험도가 근사

성별	연인 여부	
	예	아니오
여성	0.02	0.98
	0.02/0.98 = **0.0204**	
남성	0.01	0.99
	0.01/0.99 = **0.0101**	

<상대도수와 오즈비>

$$\frac{0.02}{0.01} = \frac{0.0204}{0.0101}$$

근사함!

상대위험도인 확률로 쉬운 해석 가능해짐

- 장점 1. 후향적 연구에서 사용 가능

오즈비 값은 P(Y|X), P(X|Y)
둘 중 어느 것을 사용해 정의해도
서로 동일한 값을 갖는다! (조건부 확률 공식)

$$P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)}$$

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

- 장점 1. 후향적 연구에서 사용 가능

후향적 연구는 열의 분포(Y의 분포)가 이미 고정되어 있기에 P(X|Y)만 의미가 생김

성별	연인 여부		합계
	예	아니오	
여성	10 (1/4)	30 (3/4)	40
	1/3		
남성	20 (1/3)	40 (2/3)	60
	1/2		
합계	30	70	100

성별	연인 여부		합계
	예	아니오	
여성	10 (1/4)	300 (3/4)	40
	1/30		
남성	20 (1/3)	400 (2/3)	60
	1/20		
합계	30	700	100

대조군의 크기를 달리 한다면 P(Y|X)는 바뀌게 됨

5 확률의 비교

비율의 차이 상대 위험도 오즈비

- 장점 1. 후향적 연구에서 사용 가능

후향적 연구는 열의 분포(Y의 분포)가 이미 고정되어 있기에 P(X|Y)만 의미가 생김

$$(1/3) / (1/2) = (1/30) / (1/20) = (2/3)$$

성별	연인 여부		합계
	[illegible]	[illegible]	
여성	[illegible] (1/4)	[illegible] (3/4)	40
	1/3		
남성	20 (1/3)	40 (2/3)	60
	1/2		
합계	30	70	100

성별	연인 여부		합계
	[illegible]	[illegible]	
여성	[illegible] (1/4)	[illegible] (3/4)	40
	1/30		
남성	20 (1/3)	400 (2/3)	60
	1/20		
합계	30	700	100

대조군의 크기를 달리 한다면 P(Y|X)는 바뀌게 됨

비율의차와 상대위험도는 대조군 크기에 따라 변하지만, **오즈비의 경우**

대조군의 크기와 상관없이 항상 똑같다!

- 장점 2. 행과 열이 바뀌어도 사용 가능

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

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

오즈비는 행과 열의 순서가 바뀌어도

(1/3) / (1/2) = (1/2) / (3/4) = 2/3로 같다!

- 3차원 분할표에서 오즈비

"조건부 오즈비"

- 동질연관성 : Z 통제 시, Z의 각 수준에서 XY의 연관성이 모두 같을 때

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

동질연관성은 대칭적이다!

XY에 동질연관성 존재하면, YZ, XZ도 동질연관성이 존재한다!

- 조건부독립성 : 조건부 오즈비가 1로 같을 때 즉, XY가 서로 독립일 때

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

"주변 오즈비"

- 주변독립성 : Z를 합쳤을 때의 오즈비

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

조건부독립성과 주변독립성

<부분 분할표>

학과	성별	학회 합격 여부(Y)	
		합격	불합격
통계	남자	11	25
	여자	10	27
경영	남자	16	4
	여자	22	10
경제	남자	14	5
	여자	7	12

학과 Z가 합쳐짐

<주변 분할표>

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

각 학과 별 성별과 합격 여부(XY)에 대한 오즈비

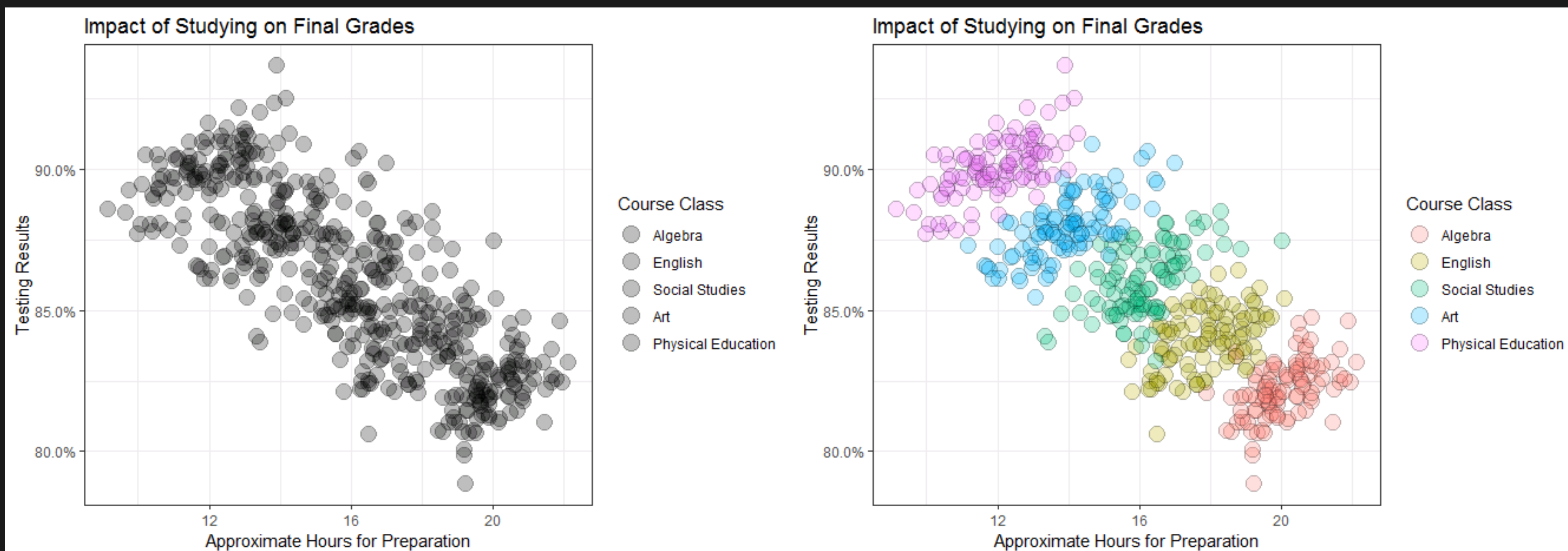
조건부 오즈비

부분분할표에서의 오즈비

주변 오즈비(=1일 때 주변독립성 가짐)

조건부독립성과 주변독립성

그러나, **조건부독립성이 성립**된다고 해서
주변독립성이 성립되는 것은 아니다!



치열했던 그 때의 기억이 떠오르는가..

몇몇 분들은 면접에서 보았을 이 구면의 plot들..

출제위원: 데마 우두머리 갓정현

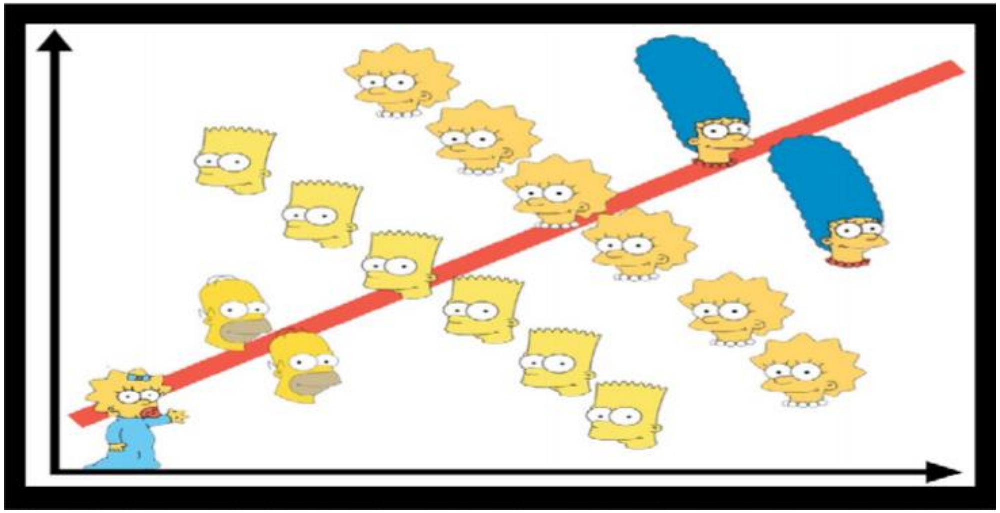
- *주변독립성*

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

- 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!

조건부연관성과 주변연관성

<부분 분할표>

학과	성별	학회 합격 여부(Y)	
		합격	불합격
통계	남자	53	414
	여자	11	37
경영	남자	0	16
	여자	4	139

학과 Z가 합쳐짐

<주변 분할표>

성별	학회 합격 여부(Y)	
	합격	불합격
남자	53+0	414+16
여자	11+4	139+37

조건부 오즈비: 0.43, 0

주변 오즈비: 1.4462

오즈비는 1이 기준이므로,
이는 연관성 방향이 서로 반대임을 알 수 있다!

도수 차이가 큰 제어 변수인 학과가 중요한 변수로 작용했기에 이를
무시하는 주변 분할표에서는 다른 결과가 나오는 것이다

2주차 예고

1. GLM
2. 유의성 검정
3. 로지스틱 회귀 모형
4. 포아송 회귀 모형
5. 로그 선형 모형

THANK YOU