

범주형자료분석팀

2팀

임지훈
안은선
강세현
심현구
하희나

INDEX

1. 범주형 자료분석
2. 분할표
3. 독립성 검정
4. 연관성 측도

1

범주형 자료분석

1

범주형 자료분석

변수의 구분

범주형 자료분석

반응변수가 범주형인 자료에 대한 분석

분석 대상인 모집단의 특징

변수와 관측치의 집합

EX) P-SAT 학회원들

나이	전공	성별
25	통계학과	여자
22	통계학과	남자
23	문헌정보학과	여자
...



변수를 열로 삼고
변수 별 관측치를 행으로
나열한 행렬 형태

1

범주형 자료분석

변수의 구분

범주형 자료분석

반응변수가 범주형인 자료에 대한 분석

분석 대상인 모집단의 특징

변수와 관측치의 집합

EX) P-SAT 학회원들

나이	전공	성별
25	통계학과	여자
22	통계학과	남자
23	문헌정보학과	여자
...



변수를 열로 삼고
변수 별 관측치를 행으로
나열한 행렬 형태

1

범주형 자료분석

변수의 구분

범주형 자료분석

반응변수가 범주형인 자료에 대한 분석

Y 변수

종속변수, 반응변수
결과변수, 표적변수

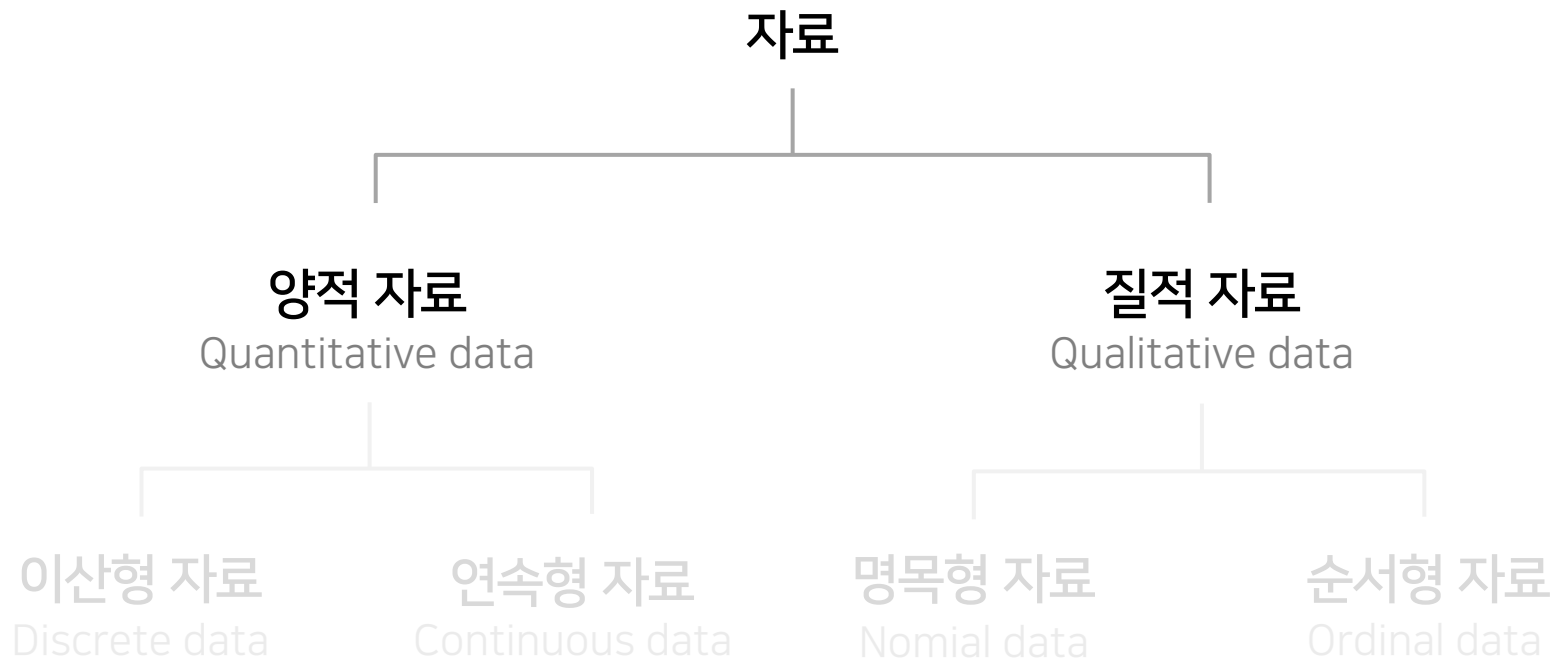
X 변수

독립변수, 설명변수
예측변수, 위험인자
요인 (범주형)



범주형 자료분석은 반응변수 Y가 범주형인 자료 분석을 의미

자료의 형태



자료의 형태



자료의 형태

양적 자료

관측값이 수치로 측정되는 자료

이산형 자료

Discrete data

값을 셀 수 있는 자료

EX) 나이, 금년도 성균관대학교 신입생 수

연속형 자료

Continuous data

연속인 어떤 구간에서 값을 취하는 자료

EX) 키, 몸무게

자료의 형태

양적 자료

관측값이 수치로 측정되는 자료

양적 자료 특징

- ✓ 공분산과 상관계수와 같은 수치적 공식 사용 가능
 - ✓ 정규분포 가정을 통해 일반회귀분석이나 ANOVA가 가능
- 회귀분석팀 클린업 참고!

자료의 형태

질적 자료

관측 결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료

명목형 자료

Nominal data

범주간에 순서의 의미가 없는 자료

EX) 혈액형, MBTI

범주형 회식 장소

철문집	홍곱창	명주삼	두짚	나누미
-----	-----	-----	----	-----

자료의 형태

질적 자료

관측 결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료

순서형 자료

Ordinal data

범주간에 순서의 의미가 있는 자료

EX) 지지도

지지도

매우 반대	반대	중립	지지	매우 지지
-------	----	----	----	-------

자료의 형태

질적 자료

관측 결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료

질적 자료 특징

✓ 순서형 자료에 명목형 자료 분석 방법 적용 가능

명목형 자료에 순서형 자료 분석 방법 적용 불가능

명목형 자료에는 순서에 관한 정보가 없기 때문!

자료의 형태

질적 자료



관측 결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료

순서형 자료에 명목형 자료 분석방법 적용 시,
분석하는 과정에서 순서에 대한 정보가 무시되기 때문에
검정력에 심각한 손실이 발생할 수 있음!

✓ 순서형 자료에 명목형 자료 분석 방법 적용 가능

명목형 자료에 순서형 자료 분석 방법 적용 불가능

명목형 자료에는 순서에 관한 정보가 없기 때문!

자료의 형태

질적 자료

관측 결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료

질적 자료 특징

- ✓ 분할표 작성 가능
 - ✓ 각 범주에 특정 점수를 할당하여 양적자료로 활용 가능
- 3주차 클린업에서 배울 예정!

2

분할표

분할표

분할표

범주형에 속하는 변수들에 대한 관측값들이 도표로 요약된 자료

수치형 자료

중심, 산포도 등의
기술통계를 통해 분석 진행

범주형 자료

분할표를 통해 분석 진행

분할표

분할표

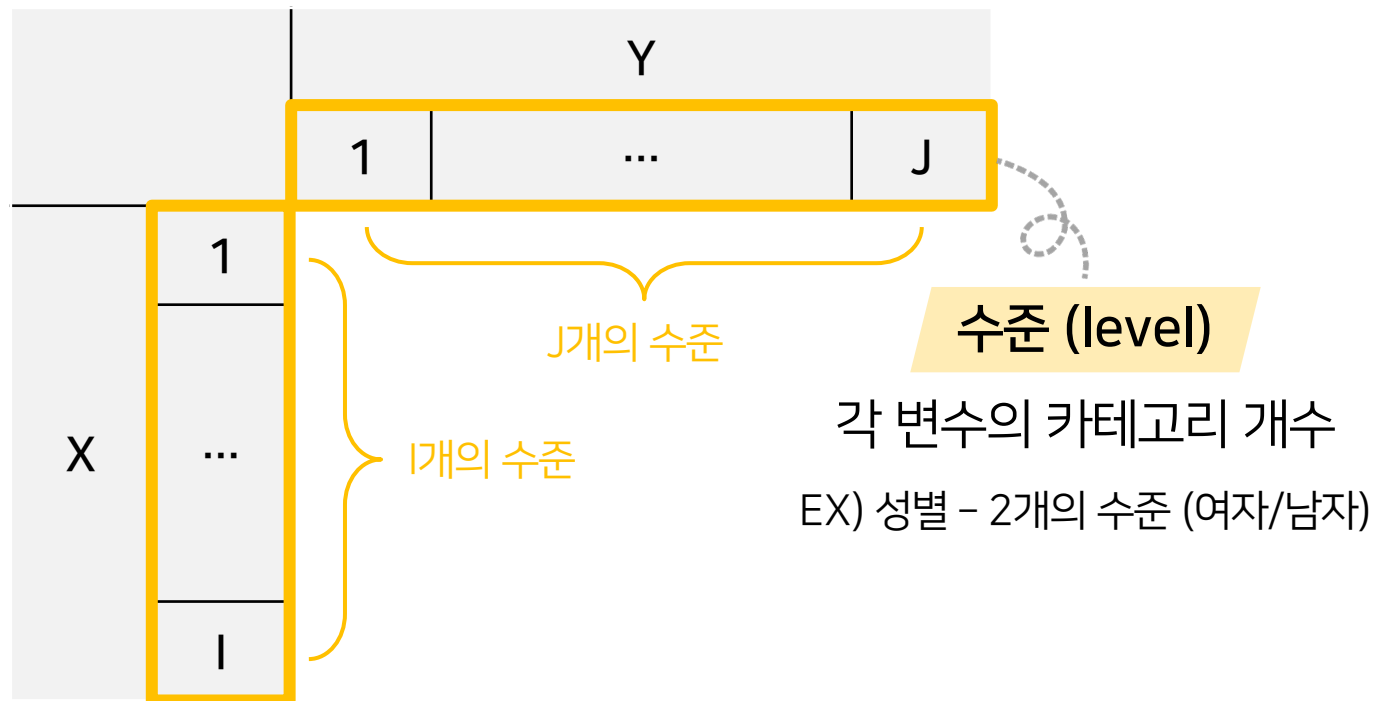
범주형에 속하는 변수들에 대한 관측값들이 도표로 요약된 자료

		Y		
		1	...	J
X	1	$I * J$ 개 칸		
	...			
	I			

분할표

분할표

범주형에 속하는 변수들에 대한 관측값들이 도표로 요약된 자료



분할표

분할표



범주형에 속하는 변수들에 대한 관측값들이 표로 요약된 자료

분할표 표현을 이용하면

예측 검정력에 대한 요약이 가능해지고

독립성 검정을 실시할 수 있음

J개의 수준

수준 (level)

X

...

I개의 수준

각 변수의 카테고리 개수

EX) 성별 - 2개의 수준 (여자/남자)

분할표

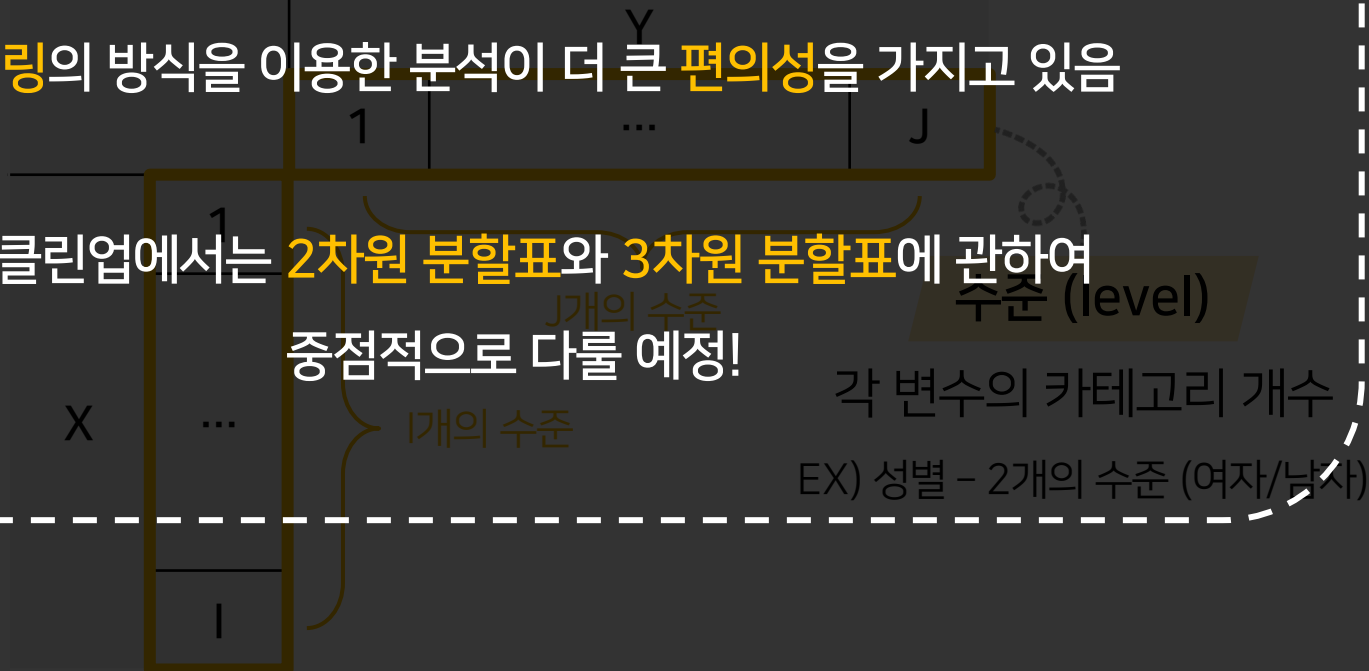


분할표

범주형에 속하는 변수들에 대한 관측값들이 도표로 요약된 자료
차원과 수준에 따라 **무한가지**의 경우의 형태로 **분할표**를 만들 수 있음

하지만 **복수의 범주형 변수**가 주어졌을 때는
모델링의 방식을 이용한 분석이 더 큰 **편의성**을 가지고 있음

본 클린업에서는 **2차원 분할표**와 **3차원 분할표**에 관하여
중점적으로 다룰 예정!



여러 차원의 분할표

2차원 분할표

두 개의 범주형 변수를 분류한 분할표

	Y			합계
X	n_{11}	...	n_{1j}	n_{1+}

	n_{i1}	...	n_{ij}	n_{i+}
합계	n_{+1}	...	n_{+j}	n_{++}



X

설명변수

Y

반응변수

주로 설명변수를 행에
반응변수를 열에 위치

여러 차원의 분할표

2차원 분할표

두 개의 범주형 변수를 분류한 분할표

	Y			합계
X	n_{11}	...	n_{1j}	n_{1+}

	n_{i1}	...	n_{ij}	n_{i+}
합계	n_{+1}	...	n_{+j}	n_{++}

 n_{ij}

각 칸의 도수

 n_{i+}

각 행의 주변 도수

 n_{+j}

각 열의 주변 도수

'+' 첨자는 해당하는 위치의
도수를 모두 더했다는 의미

여러 차원의 분할표

3차원 분할표

세 개의 범주형 변수를 분류한 분할표

기존의 설명변수와 반응변수에 K개의 수준을 가진 제어변수 Z가 추가된 형태

		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}



 설명변수
 반응변수
 제어변수

여러 차원의 분할표

학과	성별	학회 합격 여부		합계
		합격	불합격	
통계	남자	11	25	36
	여자	10	27	37
	합계	21	52	73
경영	남자	16	4	20
	여자	22	10	32
	합계	38	14	52

부분분할표

X 변수와 Y 변수가
Z 변수의 수준에 따라
분류된 분할표



제어 변수의 각 수준에서의
설명 변수와 반응 변수 간의
관계 확인 가능

여러 차원의 분할표

학과	성별	학회 합격 여부		합계
		합격	불합격	
통계	남자	11	25	36
	여자	10	27	37
	합계	21	52	73
경영	남자	16	4	20
	여자	22	10	32
	합계	38	14	52



학과를 제어변수로 설정
학과 별 학회 합격에
성별이 미치는 영향을
확인 가능!

여러 차원의 분할표

성별	학회 합격 여부		합계
	합격	불합격	
남자	11+16	25+4	56
여자	10+22	27+10	69
합계	59	66	125

주변분할표

Z 변수의 수준을
결합하여 만든 2차원 분할표



설명변수와 반응변수 간의
관계에서 제어 변수의
영향력을 제거시킨 형태

여러 차원의 분할표

성별	학회 합격 여부		합계
	합격	불합격	
남자	11+16	25+4	56
여자	10+22	27+10	69
합계	59	66	125



제어변수인 학과의 수준을 결합
학과와 무관하게 학회 합격 여부에
성별이 미치는 영향만을 확인 가능!

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

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


 π_{++}

모든 칸의 확률의 합

 π_{ij}

전체 대비 각 칸의 확률



π_{ij} 은 각 칸의 도수인 n_{ij} 를

전체 도수 n_{++} 으로

나누어서 구함

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}



결합확률

Joint Probability

- ✓ 각 칸의 확률
- ✓ 모집단에서 추출된 표본이
X 변수의 **I 번째 수준**과
Y 변수의 **J 번째 수준**을
동시에 만족하는 확률

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}



결합확률

Joint Probability

- ✓ 각 칸의 확률
- ✓ 모집단에서 추출된 표본이
X 변수의 **I 번째 수준**과
Y 변수의 **J 번째 수준**을

모든 칸의 확률의 합 = 1

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}



주변확률

Marginal Probability

- ✓ X 변수의 **I 번째 수준**이
전부 일어날 확률
- ✓ Y 변수의 **J 번째 수준**이
전부 일어날 확률

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계	
X	π_{11}	...	π_{1J}	π_{1+}	<div> <div>행의 주변확률</div> <div>Marginal Probability</div> <div> <div>✓ X 변수의 I 번째 수준이 전부 일어날 확률</div> <div>✓ Y 변수의 J 번째 수준이 전부 일어날 확률</div> </div> </div>
	
	π_{I1}	...	π_{IJ}	π_{I+}	
합계	π_{+1}	...	π_{+J}	π_{++}	열의 주변확률

비율에 대한 분할표

비율에 대한 분할표

각 칸에 도수 대신 비율이 들어간 분할표

	Y			합계
X	π_{11}	...	π_{1J}	π_{1+}

	π_{I1}	...	π_{IJ}	π_{I+}
합계	π_{+1}	...	π_{+J}	π_{++}



조건부 확률
Conditional Probability

✓ X 변수의 각 수준에서의
Y 변수의 값

✓ $\frac{\pi_{ij}}{\pi_{i+}} = P(Y = j | X = i)$

비율에 대한 분할표

	연령대에 따른 희망 직종			
	의사 (Y=1)	회계사 (Y=2)	엔지니어 (Y=3)	합계
10대 (X=1)	78	23	29	130
20대 (X=2)	41	42	37	120
합계	119	65	66	250



20대이면서 회계사를 희망할 결합 확률

$$\pi_{22} = 42/250 \approx 0.17$$

비율에 대한 분할표

	연령대에 따른 희망 직종			
	의사 (Y=1)	회계사 (Y=2)	엔지니어 (Y=3)	합계
10대 (X=1)	78	23	29	130
20대 (X=2)	41	42	37	120
합계	119	65	66	250



연령대와 상관없이 엔지니어를 희망할 주변 확률

$$\pi_{+3} = 66/250 = 0.264$$

비율에 대한 분할표

	연령대에 따른 희망 직종			
	의사 (Y=1)	회계사 (Y=2)	엔지니어 (Y=3)	합계
10대 (X=1)	78	23	29	130
20대 (X=2)	41	42	37	120
합계	119	65	66	250

20대라는 가정 하에 의사를 희망할 조건부 확률



$$\frac{\pi_{21}}{\pi_{2+}} = 41/120 \approx 0.34$$



3

독립성 검토

독립성 검정

분할표가 주어졌을 때 가능한 검정

적합도 검정

동질성 검정

독립성 검정

적합도 검정

실제로 얻어진 관측치들의 분포가 예상한 이론의 분포와 같은 지를 검정

동질성 검정

서로 다른 모집단에서 표본추출 했을 때 각 그룹의 확률분포가 같은 지를 검정

독립성 검정

두 범주형 변수 사이의 관계를 확인하기 위한 검정

독립성 검정

분할표가 주어졌을 때 가능한 검정

적합도 검정

동질성 검정

독립성 검정

적합도 검정

실제로 얻어진 관측치들의 분포가 예상한 이론의 분포와 같은 지를 검정

동질성 검정

서로 다른 모집단에서 표본추출 했을 때 각 그룹의 확률분포가 같은 지를 검정

독립성 검정

두 범주형 변수 사이의 관계를 확인하기 위한 검정

독립성 검정

분할표가 주어졌을 때 가능한 검정

적합도 검정

동질성 검정

독립성 검정

적합도 검정

실제로 얻어진 관측치들의 분포가 예상한 이론의 분포와 같은 지를 검정

동질성 검정

서로 다른 모집단에서 표본추출 했을 때 각 그룹의 확률분포가 같은 지를 검정

독립성 검정

두 범주형 변수 사이의 관계를 확인하기 위한 검정

독립성 검정

분할표가 주어졌을 때 가능한 검정

적합도 검정

동질성 검정

독립성 검정

적합도 검정

실제로 얻어진 관측치들의 분포가 예상한 이론의 분포와 같은 지를 검정

동질성 검정

서로 다른 모집단에서 표본추출 했을 때 각 그룹의 확률분포가 같은 지를 검정

독립성 검정



두 범주형 변수 사이의 관계를 확인하기 위한 검정

독립성 검정의 목적

독립성 검정

두 변수 사이의 관계를 확인하기 위한 검정

독립성 검정의 목적

- ✓ 두 변수 간 **연관성 유무** 판단 가능
- ✓ **분석 가치** 판단 가능

독립성 검정의 목적



독립성 검정

두 변수 사이의 관계를 확인하기 위한 검정

독립성 검정의 결과, 두 변수가 독립으로 도출



독립성 검정이 모전

설명변수가 반응변수에 어떠한 영향도 미치지 못한다는 의미



- ✓ 두 변수 간 연관성 없음 판단 가능
- ✓ 분석의 가치가 없다고 판단 가능!
- ✓ 분석 가치 판단 가능

독립성 검정의 가설

독립성 검정

두 변수 사이의 관계를 확인하기 위한 검정

독립성 검정의 가설

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

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

독립일 때 $P(Y|X)=P(Y)$, $P(X \cap Y)=P(X) \cdot P(Y)$ 가 성립함을 이용

두 변수가 독립이라는 것은 모든 결합 확률이 행과 열 주변 확률의 곱과 같다는 의미!

관측도수와 기대도수

관측도수 (Observed Frequency)	기대도수 (Expected Frequency)
실제 관측값 분할표 내 각 칸의 도수	귀무가설 하에 각 범주에 기대되는 도수
$n_{ij} = n \cdot \pi_{ij}$	$\mu_{ij} = n \cdot \pi_{i+} \cdot \pi_{+j}$

독립성 검정 귀무가설 (H_0)

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

양 변에 n 곱함

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

관측도수

기대도수



귀무가설 하에서
두 가설은 같은 의미를 가짐

관측도수와 기대도수

관측도수 (Observed Frequency)	기대도수 (Expected Frequency)
실제 관측값 분할표 내 각 칸의 도수	귀무가설 하에 각 범주에 기대되는 도수
$n_{ij} = n \cdot \pi_{ij}$	$\mu_{ij} = n \cdot \pi_{i+} \cdot \pi_{+j}$

독립성 검정 귀무가설 (H_0)

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

양 변에 n 곱함

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

관측도수

기대도수



귀무가설 하에서
두 가설은 같은 의미를 가짐

독립성 검정의 종류

모든 기대도수가 5 이상임을 의미

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)

독립성 검정의 종류

모든 기대도수가 5 이상임을 의미



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)

대표본 + 명목형 자료의 독립성 검정

피어슨 카이제곱 검정

검정통계량

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

기각역

$$X^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$$

가능도비 검정

검정통계량

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

기각역

$$G^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$$

검정 과정

관측도수와 기대도수의 차이가 큼 → 검정통계량의 값이 큼 → p-value 값이 작음
 → 귀무가설 기각 → 두 변수는 독립이 아님 → 변수 간의 연관성 존재

대표본 + 명목형 자료의 독립성 검정



피어슨 카이제곱 검정

검정통계량

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

기각역

$$X^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$$

가능도비 검정

검정통계량

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

기각역

$$G^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$$

검정 과정

관측도수와 기대도수의 차이가 큼 → 검정통계량의 값이 큼 → p-value 값이 작음
 → 귀무가설 기각 → 두 변수는 독립이 아님 → 변수 간의 연관성 존재

대표본 + 명목형 자료의 독립성 검정

피어슨 카이제곱 검정

검정통계량

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

기각역

$$X^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$$

가능도비 검정

검정통계량

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

기각역

$$G^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$$

검정 과정

관측도수와 기대도수의 **차이가 큼** → 검정통계량의 값이 큼 → p-value 값이 작음
 → **귀무가설 기각** → 두 변수는 **독립이 아님** → **변수 간의 연관성 존재**

대표본 + 순서형 자료 독립성 검정

순서의 정보 포함

각 변수의 수준에 점수 할당

행점수 $u_1 \leq u_2 \leq \dots \leq u_I$

열점수 $v_1 \leq v_2 \leq \dots \leq v_J$

각 수준 간 점수의 차이는 반드시 동등할 필요 X
목적에 따라 부여 방식을 다르게 할 수 있음



MH 검정

검정통계량

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

기각역

$$M^2 \geq \chi_{\alpha,1}^2$$

대표본 + 순서형 자료 독립성 검정

순서의 정보 포함

각 변수의 수준에 점수 할당

행점수 $u_1 \leq u_2 \leq \dots \leq u_I$

열점수 $v_1 \leq v_2 \leq \dots \leq v_J$

각 수준 간 점수의 차이는 반드시 동등할 필요 X
목적에 따라 부여 방식을 다르게 할 수 있음



MH 검정

검정통계량

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

기각역

$$M^2 \geq \chi_{\alpha,1}^2$$



두 변수의 추세 연관성을 확인하기 위해

피어슨 교차 적률 상관계수 사용

3

독립성 검정



대표본 + 순서형 자료 독립성 검정

순서의 정보 포함 피어슨 교차적률 상관계수

MH 검정

각 변수의 수준에 점수 할당

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

검정통계량 $M^2 = (n-1)r^2 \sim \chi^2_1$

행점수 $u_1 \leq u_2 \leq \dots$

열점수 $v_1 \leq v_2 \leq \dots$

기각역

공분산을 두 표준편차의 곱으로 나눈 $M^2 \geq \chi^2_{\alpha,1}$

상관계수와 같은 형태

r 의 범위는 -1에서 1 사이

$r = 0$ 일 때 두 변수는 독립

r 이 -1 혹은 1에 가까울 수록 두 변수 간의 큰 연관성이 존재

두 변수의 추세 연관성을 확인하기 위해
피어슨 교차 적률 상관계수 사용

3

독립성 검정

대표본 + 순서형 자료 독립성 검정

순서의 정보 포함

각 변수의 수준에 점수 할당

행점수 $u_1 \leq u_2 \leq \dots \leq u_I$

열점수 $v_1 \leq v_2 \leq \dots \leq v_J$

각 수준 간 점수의 차이는 반드시 동등할 필요 X
목적에 따라 부여 방식을 다르게 할 수 있음



MH 검정

검정통계량

$$M^2 = (n - 1)r^2 \sim \chi^2_1$$

기각역

$$M^2 \geq \chi^2_{\alpha,1}$$

검정 과정

n과 상관계수 |r|이 큼 → 검정통계량의 값이 큼 → p-value 값이 작음

→ 귀무가설 기각 → 두 변수는 **독립이 아님** → 변수 간 연관성 존재

3

독립성 검정

대표본 + 순서형 자료 독립성 검정

순서의 정보 포함



MH 검정

각 변수의 수준에 점수 할당

행점수 $u_1 \leq u_2 \leq \dots \leq u_I$

열점수 $v_1 \leq v_2 \leq \dots \leq v_J$

각 수준 간 점수의 차이는 반드시 동등할 필요 X

목적에 따라 부여 범위를 다르게 할 수 있음

검정통계량

$$M^2 = (n-1) \hat{\eta}^2 \sim \chi^2_{1}$$

기각역

$$M^2 \geq \chi^2_{\alpha,1}$$

연관성 측도를 통해 변수 간 연관성의 성질을 파악!

검정 과정

n 과 상관계수 $|r|$ 이 큼 \rightarrow 검정통계량의 값이 큼 \rightarrow p-value 값이 작음

\rightarrow 귀무가설 기각 \rightarrow 두 변수는 독립이 아님 \rightarrow 변수 간 연관성 존재

4

연관성 측도

비율의 비교 척도

비율의 비교 척도		
비율의 차이	상대 위험도	오즈비

비율 : 각 행에 따른 조건부 확률

두 범주형 변수가 모두 2가지 수준만을 갖는 이항변수일 때,
세 종류의 척도들을 통해 변수 간 **연관성 파악** 가능

비율의 차이(Difference of Proportions)

비율의 차이

π_i : i번째 행의 조건부 확률

조건부 확률 간 차이 : $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

여성이 연인이 있을 조건부 확률

$$= \pi_1 = \frac{509}{509 + 116} = 0.814$$

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

비율의 차이(Difference of Proportions)

비율의 차이

 π_i : i번째 행의 조건부 확률조건부 확률 간 차이 : $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

여성이 연인이 있을 조건부 확률

$$= \pi_1 = \frac{509}{509 + 116} = 0.814$$

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

비율의 차이(Difference of Proportions)

비율의 차이

 π_i : i번째 행의 조건부 확률조건부 확률 간 차이 : $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

남성이 연인이 있을 조건부 확률

$$= \pi_2 = \frac{398}{398 + 104} = 0.793$$

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

비율의 차이(Difference of Proportions)

비율의 차이

 π_i : i번째 행의 조건부 확률조건부 확률 간 차이 : $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

비율의 차이

$$= \pi_1 - \pi_2 = 0.814 - 0.793 = 0.021$$



여성이 연인이 있을 확률이
남성일 때보다 약 **0.021** 높음

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

비율의 차이(Difference of Proportions)

비율의 차이

π_i : i번째 행의 조건부 확률

조건부 확률 간 차이 : $\pi_1 - \pi_2$

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

비율의 차이

$$= \pi_1 - \pi_2 = 0.4 - 0.4 = 0$$



성별이 연인 유무에 영향을 미치지 못함

두 변수가 독립일 때 비율의 차이는 0

성별	연인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

상대위험도(Relative Risk)

상대위험도

π_i : i번째 행의 조건부 확률

조건부 확률의 비 : $\frac{\pi_1}{\pi_2}$

0보다 크거나 같은 값을 가짐



상대위험도 해석

상대위험도가 1에서 멀어질수록

두 변수간 연관성이 크다고 판단

상대위험도(Relative Risk)

상대위험도

 π_i : i번째 행의 조건부 확률조건부 확률의 비 : $\frac{\pi_1}{\pi_2}$

0보다 크거나 같은 값을 가짐

연인이 있을 경우의 상대위험도

$$= \frac{\pi_1}{\pi_2} = \frac{0.814}{0.793} = 1.027$$



여성일 경우 연인이 있을 확률이
약 **1.027배** 높음

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

상대위험도(Relative Risk)

상대위험도

 π_i : i번째 행의 조건부 확률조건부 확률의 비 : $\frac{\pi_1}{\pi_2}$

0보다 크거나 같은 값을 가짐

연인이 있을 경우의 상대위험도

$$= \frac{\pi_1}{\pi_2} = \frac{0.4}{0.4} = 1$$



성별이 연인 유무에 영향을 미치지 못함

두 변수가 독립일 때 상대위험도는 1

성별	연인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

비율의 차이 vs 상대위험도

성별	연인 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

비율의 차이 : 0.01

상대위험도 : $0.02/0.01 = 2$

성별	연인 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

비율의 차이 : 0.01

상대위험도 : $0.92/0.91 = 1.01$

4

연관성 측도

비율의 차이 vs 상대위험도

성별	연인 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

비율의 차이 : 0.01

성별	연인 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

비율의 차이 : 0.01

✓ 비율의 차이는 0.01로 같음

4

연관성 측도

비율의 차이 vs 상대위험도

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

✓ 상대위험도는 큰 차이가 나타남

상대위험도 : $0.02/0.01 = 2$

상대위험도 : $0.92/0.91 = 1.01$

4

연관성 측도



비율의 차이 vs 상대위험도

성별	연인 유무		성별	연인 유무	
	있음	없음		있음	없음
여성	0.02	0.98	여성	0.92	0.08
남성	0.01	0.99	남성	0.91	0.09

비율의 차이가 작더라도
상대위험도가 클 수 있음



조건부 확률이 0 혹은 1에 가까울 때
비율의 차이만으로 연관성을 판단하는 것은 위험할 수 있음!

✓ 상대위험도는 큰 차이가 나타남

상대위험도 : $0.02/0.01 = 2$

상대위험도 : $0.92/0.91 = 1.01$

4

연관성 측도

비율의 차이 vs 상대위험도



조건부 확률이 0에 가까워질수록 반응변수에 대한 두 집단의 영향력 차이가 커짐

비율의 차이와 상대위험도는

후향적 연구와 같이 Y 변수를 고정시켰을 경우

사용할 수 없다는 한계가 존재!

후향적 연구란 이미 나온 결과를 바탕으로

과거 기록을 관찰하는 연구를 의미

성별	면인 유무		성별	면인 유무	
	있음	없음		있음	없음
여성	0.02	0.08	여성	0.92	0.08
남성	0.01	0.09	남성	0.91	0.09

✓ 상대위험도는 큰 차이가 나타남

상대위험도 : $0.02/0.01 = 2$

상대위험도 : $0.92/0.91 = 1.01$

비율의 차이와 상대위험도의 한계

후향적 연구같이 **Y 변수를 고정**시켰을 경우 **사용할 수 없음**

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

심장질환 환자의 비율을 1/3으로 고정 ➡ 비율의 차이, 상대위험도 사용 불가
대조군(Y=0)의 합을 변경한다면 조건부 확률이 달라져 값이 달라지기 때문

오즈비(Odds Ratio)

π : 성공 확률

오즈(Odds)

성공확률 / 실패확률

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



오즈 해석

오즈는 성공확률이 실패확률의 몇 배인지를 의미!

오즈비(Odds Ratio)

오즈(Odds)

 π : 성공 확률

성공확률 / 실패확률

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

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

여성이 연인이 있을 오즈
: $0.814/0.186 = 4.388$

남성이 연인이 있을 오즈
: $0.793/0.208 = 3.826$

오즈비(Odds Ratio)

오즈비

각 행 별로 계산한 오즈의 비

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

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

$$\text{오즈비} = 4.388/3.826 = 1.147$$



여성이 연인이 있을 오즈가
남성이 연인이 있을 오즈보다
약 1.147배 높다

오즈비(Odds Ratio)

오즈비

각 행 별로 계산한 오즈의 비

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

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

$$\text{오즈비} = 4.388/3.826 = 1.147$$



여성이 연인이 있을 오즈가
남성이 연인이 있을 오즈보다
약 1.147배 높다

오즈비

오즈비 값에 따른 의미

$\theta = 1$: 두 행에서 성공의 오즈가 같음, 독립

$\theta > 1$: 첫 번째 행에서의 성공의 오즈가 두 번째 행보다 높음

$0 < \theta < 1$: 첫 번째 행에서의 성공의 오즈가 두 번째 행보다 낮음



서로 역수관계에 있는 오즈비

방향만 반대이고 두 변수간 연관성의 정도는 같음

오즈비

오즈비 값에 따른 의미

$\theta = 1$: 두 행에서 성공의 오즈가 같음, 독립

$\theta > 1$: 첫 번째 행에서의 성공의 오즈가 두 번째 행보다 높음

$0 < \theta < 1$: 첫 번째 행에서의 성공의 오즈가 두 번째 행보다 낮음



서로 역수관계에 있는 오즈비는
방향만 반대이고 두 변수간 연관성의 정도는 같음

로그 오즈비(Log Odds Ratio)

로그 오즈비

오즈비에 로그(log)를 씌운 형태

오즈비		
기준	$\theta = 1$ (독립을 기준으로)	
범위	$0 \sim \infty$	
기준에 다른 범위	$0 \sim 1$	$1 \sim \infty$

로그 오즈비		
기준	$\theta = 0$ (독립을 기준으로)	
범위	$-\infty \sim \infty$	
기준에 다른 범위	$-\infty \sim 0$	$0 \sim \infty$

4

연관성 측도

로그 오즈비(Log Odds Ratio)

로그 오즈비

오즈비에 로그(log)를 씌운 형태

오즈비		
기준	$\theta = 1$ (독립을 기준으로)	
범위	$0 \sim \infty$	
기준에 따른 범위	$0 \sim 1$	$1 \sim \infty$

$\theta = 1$ 을 기준으로
분모의 오즈가 더 큰 경우

$\theta = 1$ 을 기준으로
분자의 오즈가 더 큰 경우



오즈비의 범위가
비대칭적



기존 오즈비의 비대칭적인 범위를 교정

4

연관성 측도

로그 오즈비(Log Odds Ratio)

로그 오즈비

오즈비에 로그(log)를 씌운 형태

로그를 씌워
비대칭적인 범위를 교정!



대칭적인 두 범위

로그 오즈비			
기준	$\theta = 0$ (독립을 기준으로)		
범위	$-\infty \sim \infty$		
기준에 다른 범위	<table border="1"> <tr> <td>$-\infty \sim 0$</td><td>$0 \sim \infty$</td></tr> </table>	$-\infty \sim 0$	$0 \sim \infty$
$-\infty \sim 0$	$0 \sim \infty$		

$\theta = 0$ 을 기준으로
분모의 오즈가 더 큰 경우

$\theta = 0$ 을 기준으로
분자의 오즈가 더 큰 경우



오즈비의 장점

- ✓ Y 변수가 고정되어 있는 경우에도 사용 가능
- ✓ 행과 열의 위치가 바뀌어도 같은 값을 가짐

알코올 중독	심장질환 유무		합계
	심장 질환자	건강한 사람	
0	4	2	6
X	46	98	144
합계	50	100	150




알코올 중독	심장질환 유무		합계
	심장 질환자	건강한 사람	
0	4	6	10
X	46	294	340
합계	50	300	350

고정!

고정!

오즈비의 장점

- ✓ Y 변수가 고정되어 있는 경우에도 사용 가능
- ✓ 행과 열의 위치가 바뀌어도 같은 값을 가짐

	왼쪽 분할표	오른쪽 분할표	변화
비율의 차이 ($\pi_1 - \pi_2$)	$\frac{4}{6} - \frac{46}{144} = 0.347$	$\frac{4}{10} - \frac{46}{340} = 0.265$	다름
상대위험도 (π_1/π_2)	$\frac{4/6}{46/144} = 2.087$	$\frac{4/10}{46/340} = 2.956$	다름
오즈비 (odds1/odds2)	$\frac{4/2}{46/98} = 4.26$	$\frac{4/6}{46/294} = 4.26$	같음 

오즈비의 장점

- ✓ Y 변수가 고정되어 있는 경우에도 사용 가능
- ✓ 행과 열의 위치가 바뀌어도 같은 값을 가짐


알코올 중독	심장질환 유무		합계
	심장 질환자	건강한 사람	
0	4	2	6
X	46	98	144
합계	50	100	150



심장질환 유무	알코올 중독		합계
	0	X	
심장 질환자	4	46	50
건강한 사람	2	98	100
합계	6	144	150

오즈비의 장점

- ✓ Y 변수가 고정되어 있는 경우에도 사용 가능
- ✓ 행과 열의 위치가 바뀌어도 같은 값을 가짐

	왼쪽 분할표	오른쪽 분할표	변화
오즈비	$\frac{4/2}{46/98} = 4.26$	$\frac{4/46}{2/98} = 4.26$	 같음

오즈비의 장점

- ✓ Y 변수가 고정되어 있는 경우에도 사용 가능
- ✓ 행과 열의 위치가 바뀌어도 같은 값을 가짐



P(Y|X)로 정의할 때와 P(X|Y)로 정의할 때의 값이 동일하기 때문!


$$\frac{\text{odds1}}{\text{odds2}} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} = \frac{P(Y = 1|X = 1)/P(Y = 0|X = 1)}{P(Y = 1|X = 2)/P(Y = 0|X = 2)}$$

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

오즈비의 장점

- ① Y 변수가 고정되어 있는 경우에도 사용 가능
- ② 행과 열의 위치가 바뀌어도 같은 값을 가진다

왜 이런 장점들이 가능할까?

	왼쪽 분할표	오른쪽 분할표	변화
오즈비	$\frac{4/2}{46/98} = 4.26$	$\frac{4/46}{2/98} = 4.26$	<div>  같음 </div>

P(Y|X)로 정의할 때와 P(X|Y)로 정의할 때
값이 동일하기 때문!

교차적비 (Cross-Product Ratio)

교차적비

분할표의 대각선에 위치한 값끼리 곱한 수 간의 비율을 통해 정의
대조군이 바뀌거나 행과 열의 위치가 바뀌어도 같은 값을 가짐

	Y=1	Y=2
X=1	π_{11}	π_{12}
X=2	π_{21}	π_{22}

대각성분의 곱과
비대각성분의 곱의 비율의 형태

$$\theta = \frac{\text{odds1}}{\text{odds2}} = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)} = \frac{\pi_{11} / \pi_{22}}{\pi_{21} / \pi_{12}} = \frac{\pi_{11} \pi_{22}}{\pi_{12} \pi_{21}} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

3차원 분할표에서의 오즈비

3차원 분할표는 **두 개의 분할표**를 가짐

▪
▪
▪
▪

부분분할표

조건부 오즈비

동질 연관성

조건부 독립성

주변분할표

주변 오즈비

주변 독립성

부분분할표에서의 연관성

조건부 오즈비

부분분할표에서 제어변수가 고정되어 있을 때
 설명변수와 반응변수 간의 연관성을 나타내는 지표
 조건부 연관성 파악 가능!

부분분할표

노트북(Z)	성별(X)	아이폰사용여부 (Y)		조건부 오즈비
		사용	비사용	
애플	남자	11	25	$\theta_{XY(1)} = 1.188$
	여자	10	27	
삼성	남자	16	4	$\theta_{XY(2)} = 1.818$
	여자	22	10	



조건부 오즈비는
 제어변수의 각 수준별
 오즈비를 이용해 계산

부분분할표에서의 연관성

조건부 오즈비

부분분할표에서 제어변수가 고정되어 있을 때
 설명변수와 반응변수 간의 연관성을 나타내는 지표
 조건부 연관성 파악 가능!

부분분할표

노트북(Z)	성별(X)	아이폰사용여부 (Y)		조건부 오즈비
		사용	비사용	
애플	남자	11	25	$\theta_{XY(1)} = 1.188$
	여자	10	27	
삼성	남자	16	4	$\theta_{XY(2)} = 1.818$
	여자	22	10	



노트북이 애플일 때
 남자가 아이폰을 사용할 오즈가
 여자가 아이폰을 사용할 오즈보다
 약 1.188배 높음

부분분할표에서의 연관성

동질 연관성
(Homogeneous Association)

조건부 오즈비가 모두 같은 경우

$$\theta_{XY(1)} = \theta_{XY(2)} = \dots = \theta_{XY(K)}$$

X와 Y 간에 동질 연관성이 성립할 경우

Y와 Z, X와 Z 간에도 동질 연관성이 성립

동질연관성의 대칭적 성질!

조건부 독립성
(Conditional Independence)

조건부 오즈비가 1로 모두 같은 경우

부분분할표에서의 연관성

동질 연관성
(Homogeneous Association)

조건부 오즈비가 모두 같은 경우

$$\theta_{XY(1)} = \theta_{XY(2)} = \dots = \theta_{XY(K)}$$

X와 Y 간에 동질 연관성이 성립할 경우

Y와 Z, X와 Z 간에도 동질 연관성이 성립

동질연관성의 대칭적 성질!



더 엄격한 성립조건이 적용

조건부 독립성
(Conditional Independence)

조건부 오즈비가 **1**로 모두 같은 경우

주변분할표에서의 연관성

주변 오즈비

제어변수의 영향력을 제거한 주변분할표의 오즈비를 이용하여 계산
제어변수와 무관하게 설명변수와 반응변수 간의 연관성 확인

주변분할표

성별(X)	아이폰 사용 여부(Y)		주변 오즈비
	사용	미사용	
남자	11+16+14 = 41	25+4+5 = 34	θ_{XY+} = 1.515
여자	10+22+7 = 39	27+10+12 = 49	



제어변수의 각 수준에
해당하는 도수를 합침

주변분할표에서의 연관성

주변 오즈비

제어변수의 영향력을 제거한 주변분할표의 오즈비를 이용하여 계산

제어변수와 무관하게 설명변수와 반응변수 간의 연관성 확인

주변분할표

성별(X)	아이폰 사용 여부(Y)		주변 오즈비
	사용	미사용	
남자	11+16+14 = 41	25+4+5 = 34	θ_{XY+} = 1.515
여자	10+22+7 = 39	27+10+12 = 49	



주변 오즈비는
2차원 분할표의 오즈비를
계산하는 것과
동일한 원리!

주변분할표에서의 연관성

주변 오즈비

제어변수의 영향력을 제거한 **주변분할표**의 오즈비를 이용하여 계산
제어변수와 무관하게 설명변수와 반응변수 간의 **연관성** 확인

주변분할표

성별(X)	아이폰 사용 여부(Y)		주변 오즈비
	사용	미사용	
남자	11+16+14 = 41	25+4+5 = 34	θ_{XY+} = 1.515
여자	10+22+7 = 39	27+10+12 = 49	

주변 독립성은
주변 오즈비가 1일 때를 의미!



주변 오즈비가
1이 아니기 때문에
주변 독립성이
성립되지 않음

주변분할표에서의 연관성



주변 오즈비

제어변수의 영향력을 제거한 주변분할표의 오즈비를 이용하여 계산
 조건부 독립성이 성립할 때
 제어변수와 무관하게 설명변수와 반응변수 간의 연관성 확인
 반드시 주변 독립성이 성립하는 것은 아님!

주변분할표



성별(X)	아이폰 사용 여부(Y)		주변 오즈비
	사용	미사용	
남자	11+16+14 = 41	25+4+5 = 34	θ_{XY+} = 1.515
여자	10+22+7 = 39	27+10+12 = 49	

심슨의 역설 (Simpson's Paradox)

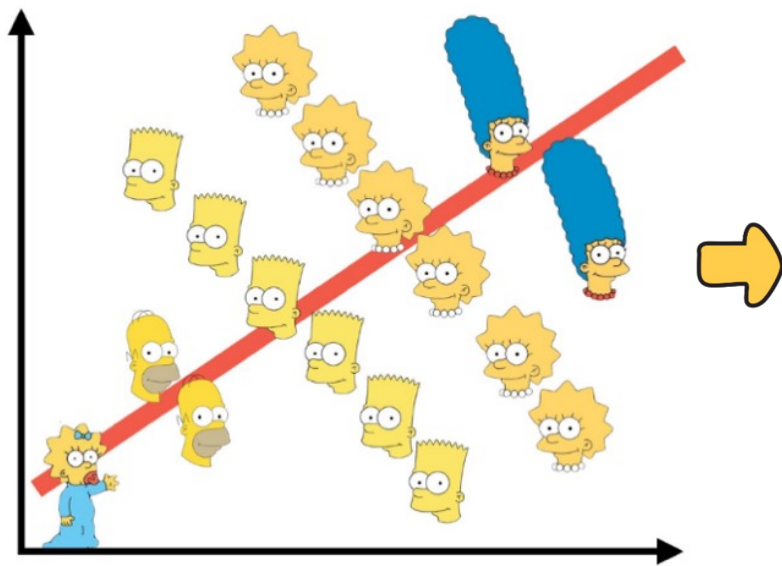


주변 오즈비가
 1이 아니기 때문에
 조건부 독립성이
 성립되지 않음

심슨의 역설

심슨의 역설

전반적인 추세가 경향성이 존재하는 것으로 보이지만
그룹으로 나누어 개별적으로 보면 경향성이 사라지거나 해석이 반대로 되는 경우



제어변수의 영향이 없는 **주변 오즈비**와
제어변수의 영향이 있는 **조건부 오즈비**의
연관성 방향이 다르게 나타남

심슨의 역설



조건부 오즈비의 방향과 주변 오즈비의 방향이 다르게 나타남

부분분할표

학과(Z)	성별(X)	자취 여부 (Y)		조건부 오즈비
		0	X	
경영	남자	40	5	$\theta_{XY(1)}$ = 1.23
	여자	130	20	
통계	남자	15	5	$\theta_{XY(2)}$ = 1.2
	여자	5	2	

주변분할표

성별	자취여부		주변 오즈비
	0	X	
남자	55	10	$\theta_{XY(1)}$ = 0.90
여자	135	22	

심슨의 역설



조건부 오즈비의 방향과 주변 오즈비의 방향이 다르게 나타남

부분분할표

학과(Z)	성별(X)	자취 여부 (Y)		조건부 오즈비
		0	X	
경영	남자	40	5	$\theta_{XY(1)}$ = 1.23
	여자	130	20	
통계	남자	15	5	$\theta_{XY(2)}$ = 1.2
	여자	5	2	



부분분할표

제어변수의 모든 수준에서
남성의 오즈 > 여성의 오즈



남자가 자취할 오즈가 여자에 비해
약 1.2배 더 높음

심슨의 역설



조건부 오즈비의 방향과 주변 오즈비의 방향이 다르게 나타남

주변분할표

남성의 오즈 < 여성의 오즈



남자가 자취할 오즈가 여자에 비해
0.9배 더 낮음



주변분할표

성별	자취여부		주변 오즈비
	0	X	
남자	55	10	$\theta_{XY(1)} = 0.90$
여자	135	22	

심슨의 역설



조건부 오즈비의 방향과 주변 오즈비의 방향이 다르게 나타남



제어변수인 학과의
수준별 도수 간의 큰 차이가 존재



학과의 연관성을 해석하는 데
큰 영향을 미치는 변수로 작용

주변분할표

성별	자취여부		주변 오즈비
	0	X	
남자	55	10	$\theta_{XY(1)} = 0.90$
여자	135	22	

심슨의 역설



조건부 오즈비의 방향과 주변 오즈비의 방향이 다르게 나타남

따라서

주변분할표

분석과정에서 3차원 분할표가 주어진다면

제어변수 조건부 독립성과 주변 독립성이 서로 다를 수 있음을

수준별 도수 간의 큰 차이가 유의하면서 분석을 진행!



학과가 연관성을 해석하는 데
큰 영향을 미치는 변수로 작용

성별	자취여부		주변 오즈비
	O	X	
남자	55	10	$\theta_{XY(1)} = 0.90$
여자	135	22	

다음주 예고

1. GLM이란?
2. 유의성 검정
3. GLM 모형의 종류
4. 로지스틱 회귀 모형
5. 다범주 로짓 모형
6. 포아송 회귀 모형

감사합니다
