범주형 자료분석 1주차 교안

근본과 행복이 넘치는 범주형 자료분석팀에 오신 팀원 여러분 진심으로 환영합니다~!

한학기 동안 여러분들과 함께하게 하게 되어 정말 기쁘고 설렙니다.

범주형 자료분석은 생소할 수도 있지만, 모두 열심히 공부해서 3 주 뒤에는

범주깡패가 되어보도록 합시다ㅎㅎ

P-SAT 활동하면서 궁금하거나 어려운 점이 있을 때,

언제든지 저에게 연락주시면 최선을 다해 돕겠습니다!

사소한 것이라도 좋으니 저를 많이 이용해주세요~~

자 그럼 범주팀 1 주차 클린업 출바알~

귀여운 범주팀 로고>_< 이름은 호희에요><

(26 기 선형대수학팀 팀장 황정현 그림)

목차

1. 범주형 자료분석이란?

우리가 한 학기 동안 살펴볼 **범주형 자료분석(Categorical Data Analysis, CDA)**은 범주형 반응변수에 대한 자료 분석을 의미한다. 이 의미를 잘 이해하기 위해 범주는 무엇인지, 반응변수는 무엇인지, 또 자료는 무엇인지에 대해 살펴보자.

■ 변수와 자료

자료 수집의 대상이 되는 모집단의 특성을 **변수(variable)**, 변수의 측정치를 **관측치(observation)**라고 한다. 예를 들어, 모집단을 대학생이라고 하면 대학생의 성별, 나이, 기상 시간, 공부 시간이 변수가 되고, 각 변수에 해당하는 값인 남, 23세, 12시, 1시간은 관측치가 된다. 이 변수와 관측치를 합치면 **자료(data)**가 된다. 쉽게 말해 자료는 변수를 열로, 관측치를 행으로 해서 만들어진 행렬이다.

■ 변수의 구분

- **X변수** : 독립변수 / 설명변수 / 예측변수 / 위험인자
- **Y변수** : 종속변수 / 반응변수 / 결과변수 / 표적변수

변수는 크게 X변수와 Y변수로 나뉘고, 각각 이처럼 다양한 이름을 갖고 있다. 방금 말했듯 범주형 자료분석은 범주형 반응변수에 대한 자료분석을 의미한다. 즉, Y변수가 범주형일 때의 분석이라는 뜻이다! 그렇다면 범주형은 무슨 의미인지 알아보기 위해 자료의 형태를 살펴보자.

■ 자료의 형태

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

자료는 위의 표처럼 구분할 수 있고, 결론부터 말하자면, 범주형 자료는 질적자료를 의미한다. 즉, 종합해보자면, 범주형 자료분석은 **Y변수가 질적변수인 자료에 대한 분석**을 의미하는 것이다! 이것이 바로 우리 범주팀의 목표이자 정체성이다.

아직 질적변수가 무엇인지 다루지 않았기에 범주형 자료분석의 의미가 지금은 와 닿지 않을 수도 있다. 또한 변수의 형태에 따라 자료 분석 방법이 결정되기 때문에 적절한 분석을 위해서는 변수의 형태를 구별하는 것이 중요하다. 따라서 위 표에 나오는 각 자료가 어떤 특징을 갖고 있는지 알아볼 것이다.

1) **양적자료** (a.k.a 수치형 자료)

수량의 형태를 가진 자료로, 이산형 자료와 연속형 자료가 있다.

- **이산형 자료** : 이산적인 값을 갖는 자료

  Ex) 코로나 확진자 수, 나이 등

- **연속형 자료** : 연속적인 값을 갖는 자료

  Ex) 체온, 몸무게 등

우리가 흔히 접하는 양적 자료는 숫자들이 의미를 갖기 때문에 공분산과 상관계수 등의 수치 계산이 가능하고, 정규분포 가정하에 일반회귀분석이나 다중회귀분석이 가능하다. 이 부분은 회귀분석팀의 클린업을 통해 자세히 알아보기로 하고, 우리의 관심사인 범주형 자료에 더 집중을 해보자구요~

2) **질적자료** (a.k.a **범주형 자료**)

측정의 단위가 여러 범주들의 집합으로 구성된 자료로, 명목형 자료와 순서형 자료로 나뉜다.

Ex) 좋아하는 가수 : 김범수 / 나얼 / 박효신 / 이수

범주형 자료는 범주에 순서가 있는가 없는가를 기준으로, 없으면 명목형 자료, 있으면 순서형 자료로 구분된다.

- **명목형 자료**

  : 순서 척도 없이 단순히 분류된 자료

  Ex) 성별(남/여), 혈액형(A/B/AB/O)

혈액형 (명목형 자료 예시)			
A	B	AB	O

- **순서형 자료**

  : 순서 척도가 있는 범주형 자료

  Ex) 소득수준(상/중/하), 영화평점(1~5)

1~5 별점으로 나타내는 영화평점 (순서형 자료 예시)				
1(싫어함)	2(좋아하지 않음)	3(좋아함)	4(아주 좋아함)	5(사랑함)

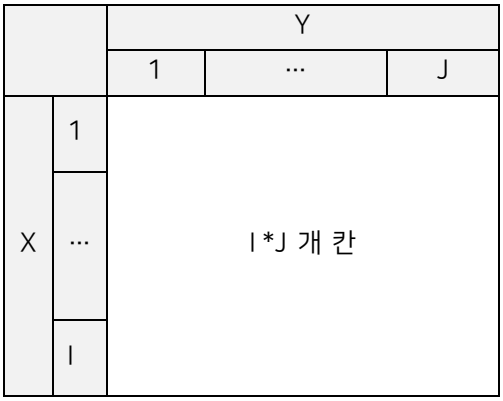
순서형 자료는 순서형 자료를 다루는 방법이 따로 존재하지만, (3주차 때 배워보아요~) 명목형 자료에 대한 분석 방법도 사용 가능하다. 이게 무슨 말이냐면, 순서 없이 단순 분류했다고 가정한 채 분석을 진행하는 것이다. 하지만 이렇게 하면 원래 자료가 갖고 있는 순서에 대한 정보가 무시되기 때문에 검정력에 심각한 손실을 가져온다.

2. 분할표

연속형 자료에 대한 기술통계 분석은 자료의 대표값(평균, 중간값 등)과 산포도(분산, 표준편차) 등에 초점을 맞춘다. 반면, 우리가 중점적으로 다룰 범주형 자료분석은 분할표를 통해 자료를 요약한다. 분할표를 사용하면 예측 검정력에 대한 요약이 가능하고(3주차 때 배워보아요~), 독립성 검정을 실시할 수 있다(이따 배워요~).

■ 분할표

앞서 말했듯, 범주형 변수의 결과를 표로 정리한 것을 **분할표(Contingency Table)**라고 한다. 분할표 안에 여러 개의 범주형 변수를 기준으로 관측치를 기록한다. 무슨 말인지 I x J 형태의 2차원 분할표 그림을 통해 살펴보자.



X변수와 Y변수가 모두 범주형 변수라고 하면 이런 2차원 I x J 분할표를 만들 수 있다. 차원은 변수의 개수를 의미하고, 행 I는 X변수의 수준을, 열 J는 Y변수의 수준을 의미한다. 여기서 **수준(level)**은 범주형 변수가 취하는 값을 의미한다. 변수가 성별이면 수준은 남,여로 2개가 된다.

■ 여러 차원의 분할표

분할표는 여러 가지 형태가 존재한다. 차원과 수준의 수에 따라 무한가지 경우의 분할표를 만들 수 있다. 그 중에서 우리는 자주 쓰이는 2차원과 3차원 분할표를 살펴볼 것이다.

1) 2차원 분할표(I x J)

두 개의 변수만으로 분류한 분할표이다.

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

n_{ij}는 각 칸의 (빈)도수를, n_{i+}, n_{+j} 는 각 열과 행의 주변(marginal) 도수를 표현한다. 쉽게 말해 중간합계라고 생각하면 될 듯하다. 그리고 n_{++}은 모든 n_{ij}의 합, 즉 총계를 의미한다. 여기서 '+'는 그 위치에 해당하는 도수를 모두 더했다는 의미의 첨자이다. 예시로 2차원 분할표를 알아보자!

최애 싹쓰리 멤버(Y)				
	유두래곤	비룡	린다.G	합계
남성	78	15	46	139
여성	49	23	37	109
합계	127	38	83	248

위 분할표는 2차원 2 x 3 분할표이다. 반응변수는 최애 싹쓰리 멤버(유행은 조금 지났지만 3인조 그룹이 생각이 안 나서,,,)이고 독립변수는 성별이다. 이 경우 n_{11}의 값은 78로 남성이면서 유두래곤이 최애인 사람이 78명이라는 것을 나타낸다. 주변도수를 보면, n_{1+}은 남성이 총 139명임을 나타내고, n_{+3}은 린다.G가 최애인 사람이 83명임을 나타낸다. 이런 식으로 2차원 분할표를 이해하고 해석하면 된다!

2) 3차원 분할표(I x J x K)

3차원 분할표는 세 변수를 분류한 분할표로, 2차원 분할표에서의 설명변수와 반응변수 외에 **제어변수(Control Variable)**가 추가된다. K는 제어변수의 수준을 의미한다.

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_{+++}

왼쪽의 분할표가 3차원 분할표이고, 여기서 제어변수 Z의 각 수준에서 분류된 것을 합쳐버리면 Z변수가 사라지면서 오른쪽의 2차원 분할표가 된다. 왼쪽을 **부분분할표**, 오른쪽을 **주변분할표**라고 한다. 2차원 분할표와 마찬가지로 n의 첨자의 의미는 동일하다.

- **부분분할표(Partial table)**

  : 제어변수 Z의 각 수준에서 X와 Y를 분류한 표로, 고정된 Z의 한 수준에 대해서 X와 Y의 관계를 보여준다. 쉽게 말해, Z를 통제했을 때 Y에 대한 X의 효과를 알 수 있다는 뜻이다.

- **주변분할표(Marginal table)**

  : 부분분할표를 모두 결합해서 얻은 2차원 분할표로, Z변수를 다 합쳐 버리기 때문에 제어변수 Z를 통제하지 않고 무시한다.

3차원보다 큰 다차원 분할표는 존재는 하지만, 사실 3차원 초과의 고차원에서는 분할표 보다는 모형으로 다루는 것이 더 효과적이다. (모형은 2주차 때 배워보아요~)

위에서 다뤘던 3차원 분할표 내용을 범주팀 대대로 내려오는 연애 예시를 통해 다시 살펴보자.

부분분할표

학과(Z)	성별(X)	연애 여부(Y)		
		O	X	합계
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
경제	남자	14	5	19
	여자	7	12	19
	합계	21	17	38

주변분할표

성별(X)	연애 여부(Y)		
	O	X	합계
남자	41	34	75
여자	39	49	88
합계	80	83	163

왼쪽은 3차원 2 x 2 x 3 분할표로, 세 변수 X : 성별(남/여), Y : 연애 여부(O/X) , Z : 학과(통계/경영/경제)가 존재한다. 이 경우에 첨자의 첫번째 글씨는 X, 두번째는 Y, 세번째는 Z를 의미한다. 즉 n_{211}의 값은 10으로 여자이고 연애를 하는 통계학과 학생이 10명임을 나타낸다. n_{+21}은 52로, 성별 상관없이 연애를 하지 않는 통계학과 학생이 52명임을 나타낸다. n_{++3}은 38로 경제학과 학생이 38명임을 의미한다.

반면, 오른쪽은 2차원 2 x 2 주변분할표로, 왼쪽 부분분할표에서 제어변수 Z(학과)를 다 합쳐서 만든 분할표이다. 가령 노란 박스로 나타낸 n_{11+}은 41로, 학과에 상관없이 연애하는 남학생이 41명임을 나타낸다. 이 41이라는 수치는 왼쪽 표에서 노란 박스로 나타낸 11(통계학과 연애남) + 16(경영학과 연애남) +14(경제학과 연애남)로 계산되었다. 또한 n_{+++}은 163으로, 총 163명의 학생을 대상으로 한 자료임을 알 수 있다.

■ 비율에 대한 분할표

지금까지는 도수에 대한 분할표를 다뤘다면, 지금부터는 비율에 대한 분할표를 알아보자. 분할표의 비율은 각 칸의 도수인 n_{ij}를 전체 도수 n_{++}로 나누어 주면 된다.

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

각 칸의 π는 전체에서의 비율을 의미하고 이는 확률이라고 할 수 있다. 확률의 합은 1이니까 모든 칸의 확률의 합을 나타내는 π_{++}은 1이 된다. 각 칸의 비율인 π_{11}은 $\frac{n_{11}}{n_{++}}$이다. 이런 방식으로 각 칸의 π값을 구하면 된다. 가끔 다른 범주 책을 보면 π가 아니라 p로 표현된 책이 있을 텐데, p는 π의 추정값, 즉 표본비율을 의미한다. π는 모비율, p는 표본비율이라고 생각하면 된다.

비율에 대한 분할표의 완벽한 이해를 위해 분할표에서의 확률분포 용어를 정리해보자.

1) **결합 확률**(Joint probability)

표본이 두 범주형 변수 X와 Y로 분류될 때, X의 i번째 수준과 Y의 j번째 수준을 동시에 만족하는 확률로,

$P(X=i, Y=j)$로 나타낸다. 쉽게 말해 i행과 j열에 속할 확률 즉, π_{ij}라고 할 수 있다. 위의 비율에 대한 분할표에서 각 칸의 확률이랑 같다!

2) **주변 확률**(Marginal probability)

결합확률의 행과 열의 합으로, $P(X=i)$ 또는 $P(Y=j)$을 의미하고, π_{i+}(행의 확률), π_{+j}(열의 확률)로 나타낸다. 이는 위의 분할표에서 각 행 혹은 각 열의 합과 같다.

3) **조건부 확률**(Conditional probability)

X의 각 수준에서 Y에 대한 확률로, $P(Y|X=i)$를 의미하고, $\frac{\pi_{ij}}{\pi_{i+}}$로 나타낸다. 이는 위의 분할표에서 행이나 열의 합을 분모로, 각 칸의 확률을 분자로 한 값과 같다.

비율에 대한 분할표와 위의 확률분포 개념의 완벽한 이해를 위해 이전에 들었던 싹쓰리 2차원 분할표 예시를 다시 살펴보자.

최애 싹쓰리 멤버(Y)				
	유두래곤 (Y=1)	비룡 (Y=2)	린다.G (Y=3)	합계
남성 (X=1)	78(0.31)	15(0.06)	46(0.19)	139(0.56)
여성 (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)

괄호 안의 값이 각 칸의 비율이 되고, 모두 합치면 1이 된다.(여기선 소수 둘째 자리 까지만 나타내는 바람에 1은 안 된다..ㅎㅎ) 비율에 대한 분할표는 도수 빼고 괄호 안의 비율만 남겨서 표현하면 된다. 이 예시에서 π_{11}은 0.31로, 전체 사람 중에서 유두래곤을 좋아하는 남성일 확률은 0.31임을 나타낸다. π_{+2}은 0.15로 성별 상관없이 비룡을 좋아할 확률은 0.15임을 의미한다.

여기서 결합확률은 0.31, 0.06과 같이 각 칸의 비율과 같다. 즉, $P(X=1, Y=1)=\pi_{11}=0.31$.

또한 주변확률은 0.5, 0.56과 같이 행이나 열의 비율의 합을 의미한다. 즉 , $P(X=1)=\pi_{1+}=0.56$.

마지막으로 조건부 확률을 살펴보자. 예를 들어 여성이라는 가정 하에 린다.G가 최애인 사람일 조건부 확률을 구해보자면 $P(Y=3|X=2)=\frac{P(X=2,Y=3)}{P(X=2)}=\frac{\pi_{23}}{\pi_{2+}}=\frac{0.15}{0.43}=0.35$가 된다.

3. 독립성 검정

범주형 자료분석을 할 때, 대부분의 범주형 자료는 분할표의 형태로 얻어진다. 이 분할표 자료에 대해서 연구자들은 보통 적합성, 독립성, 동질성의 문제에 관심이 있다. 그 중에서 우리는 변수 간의 독립성을 알아보는 독립성 검정에 초점을 두고 진행할 예정이다. (적합성과 동질성에 대해 자세히 알고 싶다면 저에게 말씀해주세요~!)

■ 독립성 검정이란?

독립성 검정은 두 범주형 변수 간에 관계가 있는지를 검정하는 방법이다. 연속형 변수인 경우 두 변수 간의 관계를 알기 위해서 상관분석, 회귀분석 등을 활용하지만, 범주형 변수는 독립성 검정이나 통계량 계산을 통해 변수들이 서로 통계적으로 유의한 관련성을 갖는지 파악한다. I x J 꼴의 2차원 분할표에서는 연관성을 나타내는 측도 통계량인 파이계수, 크래머의 V, 굿맨-크루스칼의 감마 등을 사용한다. 하지만 우리는 2 x 2 형태의 2차원 분할표에서 가설 검정을 통해 연관성을 확인하는 독립성 검정을 위주로 알아볼 예정이다. (통계량에 대해 자세히 알고 싶다면 저에게 말씀해주세요~!)

1) 독립성 검정의 목적

독립성 검정을 통해 ① 변수 간의 연관성이 있는지 없는지를 판단하고, ② 분석가치를 판단할 수 있게 된다. 만약 두 변수가 독립(=연관성이 없음)이라면, 더 이상 분석을 진행할 이유가 없음을 의미한다. 즉, 관계가 없는 두 변수에 대한 분석가치가 없다는 뜻이다.

2) 독립성 검정의 가설

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

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

귀무가설은 분할표의 결합확률(π_{ij})이 주변확률(π_{i+}, π_{+j})의 곱과 같다는 것이다. 통계학원론 확률 파트에서 $P(A \cap B) = P(A)P(B)$이면 독립임을 배웠는데 이를 떠올리면 될 것 같다.

3) 독립성 검정의 종류

- 2차원 분할표의 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)

4가지 독립성 검정이 있는데, 우리는 색칠한 부분인 대표본의 독립성 검정만 살펴볼 예정이다. 대표본 검정

의 경우 카이제곱분포 근사를 통하여 독립성 검정을 진행한다.

- 3 차원 분할표의 독립성 검정 (참고)

아래 두 검정은 2 x 2 x K 형태의 3 차원 분할표에서 사용할 수 있는 독립성 검정이다. 그렇지만 앞서 말했듯 3 차원 이상의 고차원에서는 변수 간의 관계를 모형으로 다루는 것이 효과적인데, 로그 선형 모형을 통해 가능하다.

- BD 검정 (Breslow-Day test)

  : 오즈비의 동질성 검정을 위해 고안된 카이제곱 검정

- CMH검정 (Cochran-Mantel-Haenszel test)

  : XY간의 조건부 독립성이 성립하는지 확인하는 카이제곱 검정

4) 기대 도수와 관측도수

독립성 검정의 방법을 이해하기 위해서 먼저 기대 도수와 관측 도수의 개념을 알아야 한다.

- **관측 도수** (observed frequency)

  : 표본의 도수, 즉 실제 관측값을 의미하고 n_{ij}로 표기한다. 분할표에서 각 칸의 도수와 같다. 비율로 나타낸 분할표의 경우, 전체 표본크기 n과 결합확률 π_{ij}를 곱해서 구할 수 있다. 즉 $n_{ij} = n \times \pi_{ij}$이다.

- **기대 도수** (expected frequency)

  : 귀무가설(두 범주형 변수는 독립) 하에 각 칸의 도수의 기댓값을 의미하고, $E(n_{ij})$ 혹은 μ_{ij}로 표기한다. 기대 도수 값은 전체 표본크기 n과 주변확률의 곱으로 구해진다. 즉 $\mu_{ij} = n \times \pi_{i+} \times \pi_{+j}$이다.

위에서 봤던 독립성 검정의 가설을 $\mu_{ij} = n \times \pi_{i+} \times \pi_{+j}$ 식을 통해 다음과 같이 관측 도수와 기대 도수로 표현할 수도 있다.

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

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

만약 두 변수가 독립이라면 주변확률의 곱은 결합확률과 같아지니까($\boldsymbol{\pi_{ij} = \pi_{i+} \times \pi_{+j}}$), 기대 도수의 관점에서 $\mu_{ij} = \boldsymbol{n \times \pi_{i+} \times \pi_{+j} = n\pi_{ij}}$가 된다. 즉, 위에서 봤던 가설과 지금 보는 가설이 결국 전체 도수 n만 안 곱했을 뿐, 두 가설이 같다는 소리다!

뒤이어 설명하겠지만, 독립성 검정의 방법은 관측 도수와 기대 도수의 차이를 비교하는 방식으로 이루어진다. 검정 통계량은 기대 도수와 관측 도수의 차이를 나타내는 형식으로 이루어져 있고, 기대 도수와 관측 도수의 차이가 클수록 검정 통계량의 값이 커져서 귀무가설을 기각할 확률이 높아진다. 귀무가설을 기각한다는 것은 두 변수가 독립이 아니란 소리니까 두 변수는 연관이 있다고 결론을 내릴 수 있다! 독립성 검정의 논리 흐름은 대충 이렇고(뒤에서 질리도록 이야기 할 예정..ㅎ), 그렇다면 독립성 검정을 어떻게 하는건지 자세히 알아보도록 하자. 앞서 말했듯, 우리는 2차원 분할표에서 대표본인 경우의 독립성 검정 방법만 알아볼 예정이다.

■ 명목형 자료의 독립성 검정 (대표본)

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

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

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

피어슨 카이제곱 검정 통계량 X^2은 앞서 말했듯 관측 도수와 기대 도수의 차이로 표현된다. 귀무가설 하에서 X^2값은 0이 되고, 관측 도수(n_{ij})와 기대 도수(μ_{ij})의 차이가 클수록 X^2값도 커져서 결과적으로 귀무가설을 기각할 확률이 커지게 된다. 검정의 flow를 다시 정리해보자면,

관측 도수와 기대 도수의 차이가 큼 → 검정통계량 X^2가 큼 → p-value가 작음 → 귀무가설 기각 → 변수 간의 연관성이 존재 (독립 아님)

2) **가능도비 검정** (Likelihood-ratio test)

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

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

일반화 가능도비를 이용해서 분할표 자료에 대한 검정을 실시하는 방법이다. 가능도비 검정 역시 피어슨 카이제곱 검정과 마찬가지로 관측 도수와 기대 도수의 차이로 표현된다. 로그의 성질을 떠올려보면, $\log\left(\frac{n_{ij}}{\mu_{ij}}\right) = \log n_{ij} - \log \mu_{ij}$ 이니까 G^2은 관측 도수와 기대 도수의 차이를 표현함을 알 수 있다. 또한 일반화 가능도비 검정의 근사이론에 의해(자세한 건 통추입 시간에..ㅎ) G^2는 X^2과 근사적으로 자유도가 같다. 따라서 가능도비 검정은 피어슨 카이제곱과 같은 검정 flow를 갖는다. 검정 flow를 한 번 더 설명해 보자면,

관측 도수와 기대 도수의 차이가 큼 → 검정통계량 G^2가 큼 → p-value가 작음 → 귀무가설 기각 → 변수 간의 연관성이 존재 (독립 아님)

■ 순서형 자료의 독립성 검정 (대표본)

순서형 자료에 위의 두 명목형 독립성 검정 방법을 적용할 수는 있다. 하지만, 순서 정보의 손실이 일어나기 때문에 주의해야한다.

MH 검정 (Mantel-Haenszel test)

MH검정은 두 범주형 변수가 모두 순서형인 경우에 사용하는 검정이다. MH검정은 범주형 변수의 level에 점수를 할당해서 변수 간의 선형추세를 측정하는 식으로 진행된다. 예를 들면 level이 하,중,상이면 1,2,3점을 할당하는 식이다.

- 검정통계량 : $M^2 = (n-1)r^2 \sim x^2_1$

- 기각역 : $M^2 \geq x^2_{\alpha,1}$

검정 통계량을 자세히 보면 r이 있는데 이는 피어슨 교차적률 상관계수를 의미한다.(진짜 없는데가 없는 갓피어슨..)

- 피어슨 교차적률 상관계수

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

피어슨 교차적률 상관계수 r의 식은 다음과 같고, 여기서 u_i는 분할표에서 행으로 있는 변수의 수준에 부여한 점수, v_i는 열로 있는 변수의 수준에 부여한 점수를 의미한다 ($u_1 \leq u_2 \leq \cdots \leq u_I$, $v_1 \leq v_2 \leq \cdots \leq v_j$). 식이 복잡해 보이지만 공분산을 (표본)표준편차의 곱으로 나눈다는 점에서 결국 우리가 아는 상관계수 공식과 같은 꼴이라는 것을 알 수 있다. 우리가 아는 상관계수와 마찬가지로 피어슨 교차적률 상관계수는 $-1 \leq r \leq 1$의 범위를 갖고, r값이 0 에 가까울수록 변수 간의 연관성이 적고 1 이나 -1 에 가까울수록 연관성이 크다고 할 수 있다.(사실 상관계수 값만 봐도 대충 연관성 있는지는 파악 가능하지만 자랑스러운 통계학도로서 검정을 통해 깔끔하게 결론을 도출해보자,,ㅎㅎ)

다시 MH 검정으로 돌아와서 검정통계량을 살펴보면, r값이 클수록 검정통계량 M^2값도 커진다. 즉 변수 간의 연관성이 클수록 검정통계량 값도 커져서 기각할 확률이 높아지게 된다. 또한 표본크기 n이 클수록 M^2도 커진다. MH 검정은 위의 두 명목형 변수 검정방법과 원리는 살짝 다르지만 결국은 비슷한 흐름을 갖는다. MH 검정의 flow 를 정리해보면,

상관계수 r이 큼→ 검정통계량 M^2가 큼 → p-value가 작음 → 귀무가설 기각 → 변수 간의 연관성이 존재 (독립 아님)

■ 독립성 검정의 한계

독립성 검정은 두 범주형 변수가 연관성이 있는지 없는지 여부만을 판단하기 때문에 구체적으로 얼마나 연관이 있는지는 파악할 수 없다. 검정통계량의 값이 크다고 해서 더 연관이 크다고 말할 수 없기 때문이다. 따라서 변수 간 연관성의 성질을 파악하기 위해 연관성 측도를 알아야 한다.

4. 연관성 측도

드디어 1주차 마지막 파트이다! 이 파트에서 우리는 두 범주형 변수가 모두 2가지 수준을 갖는 이항변수일 때 (ex 성별), 연관성을 나타내는 측도 세 가지를 알아볼 예정이다.

확률의 비교 척도		
비율의 차이	상대 위험도	오즈비

오즈비만 왜 빨간색으로 해 놓았을까? 정말 중요한 개념이기 때문이다. 막말로 오늘 배운 거 다 까먹고(진짜 그러면 안돼용,,ㅎㅎ) 오즈비 하나만 기억해도 1주차 클린업은 성공이라고 할 정도로 중요한 개념이다! 그러니 조금은 지치지만 기운 내서 열심히 배워보도록 하자~~

■ 비율의 차이 (Difference of Proportions)

비율의 차이는 조건부 확률의 차이($\pi_1 - \pi_2$)를 의미한다. π_i는 i번째 행의 조건부확률을 의미한다. 무슨 소리인지 범주 전통의 애인 예시를 다시 꺼내 살펴보자.

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

이 분할표에서 괄호 안의 확률은 여성일때와 남성일때의 조건부 확률을 의미한다. 애인이 있는 경우의 조건부 확률의 차이를 보고싶다면, π_1은 여성일 때 애인이 있을 조건부확률(0.814)이 되고, π_2는 남성일 때 애인이 있을 조건부확률(0.793)이 된다. 애인이 없는 경우의 확률의 차이를 보고싶으면 각각 π_1을 여성일 때 애인이 없을 조건부확률(0.186), π_2를 남성일 때 애인이 없을 조건부확률(0.207)로 놓고 차이를 계산하면 된다. (보통은 있다고 하는 경우를 비교한다.) 전자의 경우를 보면, $\pi_1 - \pi_2 = 0.8144 - 0.7928 = 0.0216$으로, 여성일 때 애인이 있을 확률이 남성일 때보다 0.0216 높다고 할 수 있다. 비율의 차이 값은 $-1 \leq \pi_1 - \pi_2 \leq 1$의 범위를 갖고, 만약 아래의 예시처럼 $\pi_1 - \pi_2$ 값이 0이면, 두 비율의 차이가 없다는 소리니까 연관이 없다고 할 수 있다.

성별	애인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

이 표의 경우 여성일 때 애인이 있을 확률과 남성일 때 애인이 있을 확률의 차이가 없다는 것이므로, 성별이 애인 여부에 영향을 끼치지 않는다는 결론을 내릴 수 있다.

■ 상대위험도 (Relative Risk)

상대위험도는 조건부 확률의 비($\frac{\pi_1}{\pi_2}$)를 의미한다. 위의 비율의 차이는 말그대로 조건부 확률의 "차이"였고 여기서는 "비율"을 확인한다. 상대위험도가 클수록 변수 간의 연관성이 크다고 할 수 있다. 다시 연인 예시를 보면,

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

위에서처럼 π_1을 여성일 때 애인이 있을 조건부확률, π_2를 남성일 때 애인이 있을 조건부확률로 놓고 상대위험도를 계산하면 $\frac{\pi_1}{\pi_2} = \frac{0.814}{0.793} = 1.027$이 된다. 즉, 여성일 때 애인이 있을 확률이 남성일 때 애인이 있을 확률 보다 1.027배 높다고 할 수 있다. 상대위험도는 $\frac{\pi_1}{\pi_2} \geq 0$의 범위를 갖고, $\frac{\pi_1}{\pi_2} = 1$이면 확률이 같다는 소리니까 변수 간에 연관이 없다고 할 수 있다. 상대위험도를 볼 때는 주의해야할 점이 있다. 바로 확률이 0이나 1에 가까울 때는 영향력 차이가 많이 난다는 점이다. 예시를 살펴보자.

성별	애인 유무	
	있음	없음
여성	0.02	0.98
남성	0.01	0.99

성별	애인 유무	
	있음	없음
여성	0.92	0.08
남성	0.91	0.09

왼쪽이나 오른쪽 분할표처럼 조건부 확률이 0이나 1에 가까울 때, 두 분할표의 (1)비율의 차이를 계산해 보면 모두 0.02 - 0.01 = 0.92 – 0.91 = 0.01로 같고 차이의 정도도 0.01로 작다. 하지만, 두 분할표의 (2)상대위험도를 보면 0.02 / 0.01 = 2와 0.92 / 0.91 = 1.01로 2배 정도 차이를 보인다. 즉, 비율의 차이가 낮아 보여도 상대위험도는 클 수 있다.

비율의 차이와 상대위험도는 직관적인 측도이긴 하지만, **후향적 연구**처럼 한 변수의 수를 고정시킨 조사에서는 사용이 불가하다는 단점이 존재한다. 후향적 연구는 이미 나온 결과를 바탕으로 과거 기록을 관찰하는 연구이다. 무슨 말인지 예시를 통해 살펴보자.

	폐암 환자 ($Y=1$)	건강한 사람 ($Y=0$)	합
과거 흡연 O ($X=1$)	4	2	6
과거 흡연 X ($X=0$)	46	98	144
합	50	100	150

이처럼 폐암과 흡연의 연관성을 보기 위해 폐암이라는 이미 나온 결과를 바탕으로 과거 흡연 기록을 관찰하는 연구가 후향적 연구다. 후향적 연구는 주로 보건 분야에서 많이 진행되고, 지금처럼 사례군(폐암O)과 대조군(폐암X)을 비교하는 사례-대조 연구의 경우가 많다. 폐암환자처럼 사례가 흔하지 않은 경우, 랜덤으로 표본을 뽑지 않고 폐암 여부에 의해 정해진 비율이나 숫자에 따라 사람을 뽑게 된다. 이 예시에서는 폐암 환자 비율을 1/3으로 고정했다. 이럴 경우 비율의 차이나 상대위험도는 사용할 수 없게 된다. 연구자가 폐암 환자 비율을 몇으로 정하는지에 따라 값이 달라지기 때문이다. 가령 위의 경우 조건부확률을 따져보면, $\pi_1 = \frac{4}{6} = 0.66$, $\pi_2 = \frac{46}{144} = 0.32$ 인데, 만약 표본을 뽑을 때 건강한 사람을 200명으로 바꾼다면, 조건부 확률 값이 달라지게 돼서 비율의 차이나 상대위험도 값도 달라지게 된다. 이처럼 변수의 수를 고정시킨 조사에서는 두 측도는 사용이 불가하다. 하지만 이런 경우 오늘의 주인공, 오즈비를 사용하면 된다!

■ 오즈비 (Odds Ratio)

1) 오즈 (Odds)

오즈란 성공확률/실패확률을 의미한다. 우리가 알고 있는 확률과는 비슷해 보이지만 다른 개념이다! π가 어떤 사건의 성공확률이라고 하면 오즈는 다음과 같이 표현할 수 있다.

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

식을 π에 대해 정리하면, 아래와 같이 나타낼 수도 있다.

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

오즈 개념의 이해를 위해 이번에도 연애 예시를 들어보자.

성별	애인 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…	

여성의 입장에서 애인이 있을 오즈는, (애인 있는 확률)/(없는 확률) = 0.814/0.186 = 4.388이라고 할 수 있고, 남성의 입장에서 애인이 있을 오즈는 (애인 있는 확률)/(없는 확률) = 0.793/0.207 = 3.826이라고 할 수 있다. 이런 식으로 성공을 실패로 나눠주면 실패에 비해 성공이 몇 배인지를 나타내는 오즈가 된다.

2) 오즈비 (Odds Ratio)

오즈비란 각 오즈의 비를 의미한다. 한마디로 오즈/오즈라고 할 수 있다.

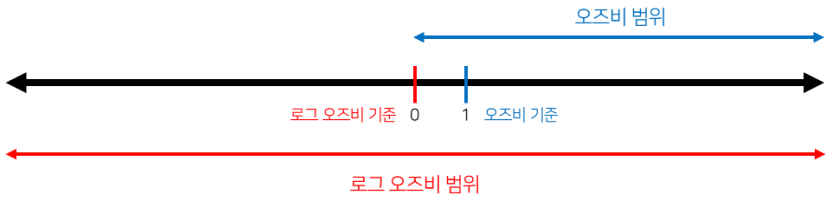
$$\theta = \frac{odds1}{odds2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

오즈비는 $\theta \geq 0$의 범위를 갖는다. 만약 두 행의 오즈가 같다면 오즈비 $\theta = 1$이 되고, 이는 두 변수 간에 연관이 없음을 의미한다. $\theta > 1$이면 분자의 오즈가 더 크다는 의미이고, $0 < \theta < 1$이면 분모의 오즈가 더 크다는 의미이다. 서로 역수관계에 있는 오즈비는 방향만 반대일 뿐 같은 연관성 정도를 갖는다. 예를 들어 오즈비가 4이거나 0.25(1/4)는 어느 오즈를 분자에 두고 분모에 두는지에 따라 값이 달라진 것일 뿐 연관성 정도는 같다.

아까 살펴본 애인 예시에서 성별에 따른 애인이 있을 오즈는 각각 4.388, 3.826이 나왔다. 따라서 오즈비를 계산해보면 $\theta = \frac{4.388}{3.826} = 1.147$이 나오고, 이를 해석하자면 여성이 애인이 있을 오즈가 남성이 애인이 있을 오즈보다 1.147배 높다는 의미이다.

로그 오즈비

로그 오즈비는 말그대로 오즈비에 log를 씌운 것이다. 복잡하게 log는 왜 씌울까? 오즈비의 범위를 생각해보면 이해가 된다.



기존 오즈비의 범위는 $\theta \geq 0$, 즉 0보다 크거나 같을 때로 한정된다. 그 안에서 $\theta = 1$을 기준으로 분자가 더 큰지 분모가 더 큰지 결정된다. 그림처럼 0~1과 1~∞로 나뉘는 셈이니까 상당히 비대칭적이다. 하지만 로그를 취해주면 $-\infty < \log\theta < \infty$의 범위를 갖게 된다. 뿐만 아니라, 0이 기준점이 돼서 $-\infty$~0과 0~∞로 범위가 대칭적으로 바뀌게 된다. 즉, 로그 오즈비는 기존의 비대칭한 오즈비의 범위를 교정한 측도이다.

- 오즈비의 장점

① 오즈비는 후향적 연구처럼 한 변수가 고정되어 있을 때도 사용이 가능하다. 앞서 말했듯 비율의 차이나 상대위험도는 값이 그때그때 바뀌기 때문에 후향적 연구에서 사용할 수 없었다. 하지만 오즈비는 대조군의 크기가 달라져도 같은 값을 갖는다. 위에서 후향적 연구라고 말했던 폐암 예시를 들어보자.

	폐암 환자 $(Y=1)$	건강한 사람 $(Y=0)$	합
과거 흡연 O $(X=1)$	4	2	6
과거 흡연 X $(X=0)$	46	98	144
합	50	100	150

분할표에서 볼 수 있듯이 사례군(폐암O)과 대조군(폐암X)을 50:100으로 고정한 후 후향적 연구를 진행했다. 여기서 대조군의 크기를 100이 아닌 200으로 달리했을 때 비율의 차이, 상대위험도, 오즈비가 각각 어떻게 되는지 살펴보고자 한다.

과거 흡연	폐암 여부		합
	O	X	
O	4(4/6)	2(2/6)	6
	4/2		
X	46(46/144)	98(98/144)	144
	46/98		
합	50	100	150

과거 흡연	폐암 여부		합
	O	X	
O	4(4/8)	4(4/8)	8
	4/4		
X	46(46/242)	196(196/242)	242
	46/196		
합	50	200	250

괄호 안은 각 칸의 행 기준 조건부 확률을 의미하고, 노란색으로 표시된 부분은 각 행에서 폐암일 오즈이다. 왼쪽은 사례군과 대조군이 50:100인 연구의 분할표, 오른쪽은 50:200인 연구의 분할표이다.

- 비율의 차이

왼쪽 : $\pi_1 - \pi_2 = \frac{4}{6} - \frac{46}{144} = 0.347$ 오른쪽 : $\pi_1 - \pi_2 = \frac{4}{8} - \frac{46}{242} = 0.309$

대조군이 달라짐에 따라 비율의 차이 값이 달라졌다.

- 상대위험도

왼쪽 : $\frac{\pi_1}{\pi_2} = \frac{4/6}{46/144} = 2.087$ 오른쪽 : $\frac{\pi_1}{\pi_2} = \frac{4/8}{46/242} = 2.63$

상대위험도 역시 대조군이 달라짐에 따라 값이 달라졌다.

- 오즈비

왼쪽 : $\frac{4/2}{46/98} = 4.26$ 오른쪽 : $\frac{4/4}{46/196} = 4.26$

대조군의 크기가 달라졌음에도 오즈비는 그대로다! (우와 신기해 나만 신기해..?) 왜 이런 신기한 결과가 나오는지는 오즈비의 다른 장점을 설명하고 난 후에 말하도록 하겠다.

② 오즈비는 행과 열의 위치가 바뀌어도 사용 가능하다.

과거 흡연	폐암 여부		합
	O	X	
O	4(4/6)	2(2/6)	6
	4/2		
X	46(46/144)	98(98/144)	144
	46/98		
합	50	100	150

폐암 여부	과거 흡연		합
	O	X	
O	4(4/50)	46(46/50)	50
	4/46		
X	2(2/100)	98(98/100)	100
	2/98		
합	6	144	150

왼쪽 분할표는 우리가 계속 쓰던 폐암 연구 분할표이고, 오른쪽은 이 분할표의 행과 열을 바꾼 분할표이다. 즉 행이 폐암여부가 되고 열이 과거 흡연이 된다. 각각의 오즈비를 계산해보면,

$$\text{왼쪽} : \frac{4/2}{46/98} = 4.26 \qquad \text{오른쪽} : \frac{4/46}{2/98} = 4.26$$

역시나 오즈가 동일하다! (신기신기,,) 비율의 차이나 상대위험도는 조건부 확률이 달라지기 때문에 당연히 값이 변한다. 왜 이런 결과가 가능할까? 이는 오즈비 값이 $P(Y|X)$를 사용하여 정의하나 $P(X|Y)$를 사용하여 정의하나 서로 동일한 값을 갖기 때문이다. 무슨 말인지 조건부확률의 정의와 베이즈 정리를 이용해서 확인해보자.

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

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

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

결국 행을 기준으로 조건부 확률을 따지나, 열을 기준으로 조건부 확률을 따지나 오즈비는 동일하다는 뜻이다!

오즈비가 장점 ①, ②를 갖는 이유는 바로 오즈비가 **교차적비**(cross-product ratio)이기 때문인데, 교차적비는 대각선에 있는 칸 확률들의 곱의 비율로 정의된다. 식으로 적자면 다음과 같다.

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

오즈비의 식을 정리해보면 분할표의 대각성분의 곱이 분자로, 비대각성분이 분모로 가서 결국에는 교차적비를 이룬다. 따라서 오즈비를 교차적비라고 부르기도 한다! 오즈비가 교차적비이기 때문에 변수가 고정된 상태에서 대조군이 바뀌더라도 그 값은 유지되고, 행과 열을 바꾸더라도 값이 변하지 않는 것이다.

3) 3차원 분할표에서의 오즈비

- 조건부 독립성과 주변 독립성

위에서 봤던 3차원 분할표를 떠올려보자. 제어변수가 추가된 부분분할표와 제어변수가 합쳐진 주변분할표를 배웠었다. 부분분할표에서의 연관성을 **조건부 연관성**(Conditional association)이라고 하는데, 이는 제어변수 Z의 값이 고정되어 있다는 조건 하에서 X와 Y의 연관성을 의미한다. 조건부 연관성은 **조건부 오즈비**(Conditional odds ratio)를 통해 알 수 있다. 위에서 봤던 학과별, 성별 연애여부 부분분할표를 다시 살펴보자.

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

제어변수가 고정된 상태에서 구한 오즈비가 바로 조건부 오즈비이다. 이 분할표에서는 제어변수인 학과를 고정한 상태에서 오즈비를 구하면 조건부 오즈비가 된다. 예를 들어 $\theta_{XY(3)}$= 4.8은 경제학과 한정 남자가 연애를 할 오즈가 여자가 연애를 할 오즈보다 4.8배 높다고 해석할 수 있다.(경제학과 남학우들 대단해,,) 조건부 오즈비가 모두 같은 경우를 **동질 연관성**(homogeneous association)이라고 정의한다. 동질 연관성은 대칭적이기 때문에 XY에 동질 연관성이 있으면 YZ, XZ도 동질 연관성이 있다. 즉, 조건부 오즈비 θ_{XY}끼리 다 같으면 θ_{YZ}나 θ_{XZ}끼리도 다 같다는 뜻이다.

만약 조건부 오즈비가 모두 같은데 심지어 그 값이 "1"로 같다면 **조건부 독립성**(Conditional Independence)이라고 한다. 즉, $\theta_{XY(1)}$= $\theta_{XY(2)}$= ⋯ = $\theta_{XY(K)}$= 1인 경우 조건부 독립성이라고 한다. 조건부 독립성은 동질 연관성의 특별한 경우라고 할 수 있겠다.

부분분할표에서의 연관성을 살펴봤다면 이번에는 주변분할표에서의 연관성을 알아보자. 예시로 썼던 위의 분할표에서 제어변수 학과를 합쳐서 주변분할표를 만들어보자.

주변분할표			
성별(X)	연애 여부(Y)		주변 오즈비
	O	X	
남자	41	34	$\boldsymbol{\theta_{XY+}=\frac{41/34}{39/49}=0.148}$
여자	39	49	

학과에 상관없이 성별과 연애 여부만 따지는 주변분할표가 완성되었다. 이렇게 만들어진 주변분할표에서는 **주변 오즈비**(Marginal odds ratio)를 통해 연관성을 알아볼 수 있다. 주변오즈비는 제어변수를 합쳐버린 주변분할표에서의 오즈비이고, 만약 주변오즈비가 1일 때 **주변 독립성**(Marginal Independence)을 갖는다고 할 수 있다. 즉, $\boldsymbol{\theta_{XY+}=1}$ 일 때 주변 독립성을 갖는다고 말한다. 사실 주변 오즈비나 주변 독립성은 그냥 2차원 분할표에서의 오즈비나 독립성과 동일하다. 단지 부분분할표에서 왔다는 것을 살리기 위해 별도로 용어를 쓸 뿐이다.

여기서 주의할 부분은 조건부 독립성이 성립한다고 해서 주변 독립성이 성립되는 것은 아니라는 점이다. 쉽게 말해 조건부 오즈비가 1이라고 해서 주변 오즈비가 1이 아닐 수 있다는 뜻이다.

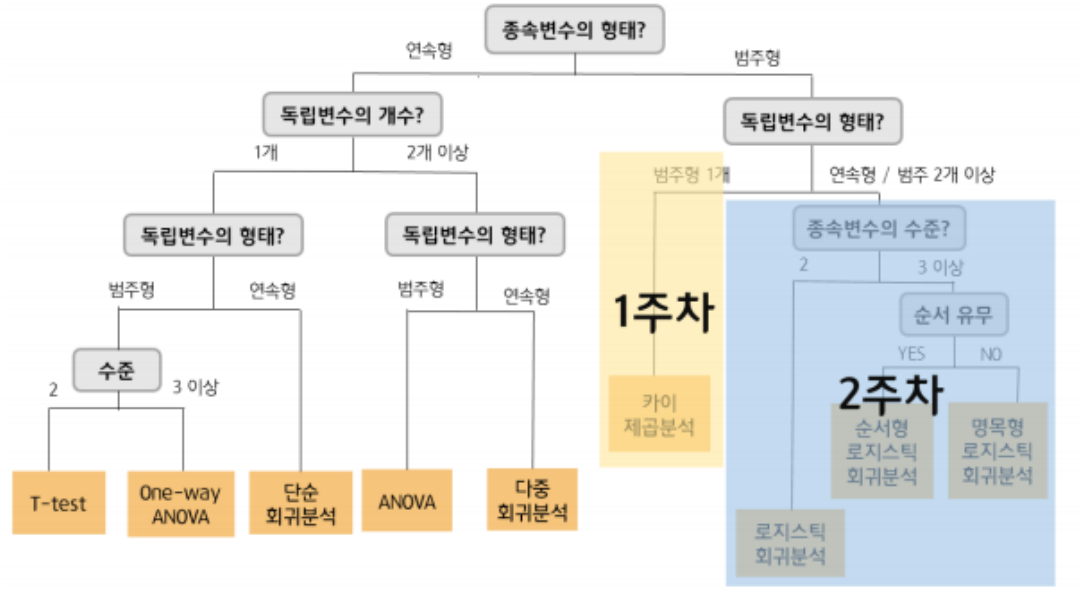
부분분할표				
학과(Z)	성별(X)	연애 여부(Y)		조건부 오즈비
		O	X	
국어 국문	남자	18	12	$\boldsymbol{\theta_{XY(1)}=\frac{18/12}{12/8}=1}$
	여자	12	8	
문헌 정보	남자	2	8	$\boldsymbol{\theta_{XY(2)}=\frac{2/8}{8/32}=1}$
	여자	8	32	

그동안 예시를 들었던 상경계열 말고 인문학부의 연애를 살펴보자. 국어국문학과나 문헌정보학과의 조건부 오즈비가 1로 같음을 확인할 수 있다. 즉, 조건부 독립성이 성립한다는 뜻이다. 하지만 이때의 주변 분할표를 보면,

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

주변 오즈비가 2임을 확인할 수 있다. 즉, 주변 오즈비가 1이 아니므로 주변 독립성은 성립하지 않는다. 인문학부 예시에서는 조건부 독립성은 성립하지만 주변 독립성이 성립하지 않는다는 점을 확인할 수 있었다.

다음주 예고!



<이번주 실습 과제!>

저번 학기 범주팀의 주제분석 데이터를 변형해서 새로 데이터를 만들어봤습니당~

이 데이터를 갖고 아래의 작업들을 수행해주세요!

0. 데이터 불러오기/살펴보기

- 아래 코드 그대로 실행하면 됩니다!

```
library(data.table)
library(tidyverse)
data = fread('1 주차실습.csv') %>% mutate_all(as.factor)
str(data)
```

- 변수 설명

  $sex : 성별 (M/F)

  $h1n1_vaccine : 독감 백신 접종 여부(Y/N)

  $income_poverty : 소득수준 (Below Poverty / <= $75,000, Above Poverty / > $75,000)

  $h1n1_knowledge : 독감에 대한 지식수준 (Low / Medium / High)

1. 원하는 변수 2 개 골라서 2 차원 분할표 만들어보기
2. 원하는 변수 3 개 골라서 3 차원 분할표 만들어보기
3. 독립성 검정 시행하기
   - 명목형 변수는 명목형 변수끼리, 순서형 변수는 순서형 변수끼리, 알맞은 검정방법으로
   - 검정결과(기각여부) 및 해석(변수끼리 독립 여부)도 쓰기!
4. sex와 h1h1_vaccine 변수로 2차원 분할표(2X2) 만들고 비율의 차이, 상대위험도, 오즈비 계산하기
   - 직접 구하는 법
   - library(epiR) 사용해서 구하는 법
   - 해석

목요일(9/9)까지 코드/마크다운을 저에게 갠톡으로 보내주세요!