범주형 자료분석 1주차 교안

범주형자료분석팀에 오신 여러분들을 진심으로 환영합니다. 한학기동안 여러분들과 함께하게 되어 정말 기쁘고 설렙니다. 언제든지 궁금한 것이 있을 때 질문해주시면 최선을 다해(!) 같이 궁금해보도록 하겠습니다. 그러면 1주차 클린업 시작해봅시다~

*호랑이 **범**과 주머니 **주**의 귀여운 범주팀 로고입니다 🦁*

(선대대장 황정현 팀장님 그림)

- 목차 -

Table of Contents

1. 범주형 자료분석이란?

■ 변수와 자료

자료 수집의 대상이 되는 모집단의 특성을 **변수(variable)**, 변수의 측정치를 **관측치(observation)** 라고 한다. **자료(data)**는 이러한 변수와 관측치로 이루어진 모임이다. 대학생의 아침식사 여부, 기상 시간, 점심 메뉴, 음주 횟수 등에 관심이 있다면 이들을 각각 변수라 한다. 각 변수를 열, 측정치를 행으로 만들어진 행렬을 자료, 즉 데이터라 하는 것이다.

■ 변수의 구분

- Y변수 (종속변수 / 반응변수 / 결과 변수 / 표적 변수)
- X변수 (독립변수 / 설명변수 / 예측 변수 / 위험 인자 / 공변량 [연속형 자료] / 요인 [범주형 자료])

범주형 자료분석의 의미를 명확히 이해하기 위해 변수를 구분하는 다양한 이름들을 알아보았다. 그렇다면 우리가 앞으로 배울 **'범주형 자료분석'**이란 무엇일까? 바로 **범주형 반응변수에 대한 자료 분석**을 의미한다.

■ 자료의 형태

그렇다면 범주형 자료란 무엇인지 알기 위해, 먼저 자료의 형태에 대해 알아보자. 변수의 형태에 따라 자료 분석 방법이 결정되기 때문에 적합한 분석방법을 찾기 위해선 변수의 형태를 구별할 줄 알아야 한다.

<table>
<tr><td rowspan="4">자료</td><td rowspan="2">양적 (Quantitative) 자료</td><td>이산형 (Discrete) 자료</td></tr>
<tr><td>연속형(Continuous) 자료</td></tr>
<tr><td rowspan="2">질적 (Qualitative) 자료</td><td>명목형 (Nominal) 자료</td></tr>
<tr><td>순서형 (Ordinal) 자료</td></tr>
</table>

1) 양적 자료 (a.k.a. 수치형 자료)

수량의 형태를 가진 자료로, 이산형 자료와 연속형 자료가 있다.

- 이산형 자료 : 이산적인 값을 갖는 데이터 (ex. 자녀의 수, 사건 발생 수, 나이 등)

- 연속형 자료 : 연속적인 값을 갖는 데이터 (ex. 신장, 체중, 온도 등)

양적 자료는 공분산과 상관계수 등의 수치적 공식 사용이 가능하며, 정규분포를 통해 일반회귀분석이나 ANOVA (분산분석)가 가능하다. 이부분은 회귀분석팀 클린업을 통해 더 자세히 알아보도록 하자! 우리의 메인은 범주형 자료!

2) 질적 자료 (a.k.a. 범주형 자료)

측정 단위가 **여러 범주들의 집합으로 구성된 자료**를 범주형 자료라고 한다. 정치성향(진보적/중립적/보수적),

생존여부(생존/사망), 혈액형(A/B/O/AB) 등이 이에 해당한다.

범주형 자료는 다시 명목형 자료와 순서형 자료로 구분된다. 이를 구분하는 척도는 바로 '범주에 순서가 있는가 없는가' 이다.

- 명목형 (Nominal) 자료

순서 척도가 없는 범주형 변수로, 예시로 성별(F/M), 성공여부(성공/실패), 혈액형(A/B/O/AB) 등이 있다. 명목형 자료는 순서형 자료 분석 방법 자체가 사용 불가하다. (당연한 말.. 이지만 비교를 위해 알아 두자!)

혈액형 (명목형 자료 예시)			
A	B	AB	O

- 순서형 (Ordinal) 자료

순서 척도가 있는 범주형 변수로, 예시로 학년, 순위, 소득 수준(상,중,하) 등이 있다. 순서형 자료는 명목형 자료에 대한 분석 방법이 사용 가능하지만(!), 순서에 대한 정보가 무시되기 때문에 검정력에 심각한 손실을 가져온다. (순서형 자료는 범주에 일정 점수를 할당하여 양적 자료 형태로 다룰 수도 있는데, 이는 3주차 인코딩 부분에서 다룰 예정!)

1~5 별점으로 나타내는 영화 평점 (순서형 자료 예시)				
싫어함	좋아하지 않음	좋아함	아주 좋아함	사랑함

- 범주형 자료와 관련된 분포와 모형

가. 이항분포 : 로지스틱 회귀 모형(Logistic Regression model) 범주의 꽃 로지스틱… 2주차에 만나자

나. 다항분포 : 다항 로지스틱 회귀 모형(Multinomial Regression model) 얘도 2주차에 만나자!

다. 포아송 분포 : 로그 선형 모형, 포아송 회귀 모형, 카우시 모형

라. 음이항 분포 : 음이항 회귀 모형

모형에 관련된 내용은 2주차에서 다룰 예정이므로, 지금은 범주형 자료분석에 이러한 분포들이 사용되는구나 정도로 알아 두면 좋겠다.

2. 분할표 (Contingency Table)

■ 분할표란?

연속형 자료에 대한 기술통계 분석은 자료의 중심(평균, 중간값 등)과 분산 정도(분산, 표준편차) 등에 초점을 맞추고 있다. 반면 범주형 자료분석은 분할표를 통해 자료를 쉽게 요약할 수 있다.

분할표란 **여러 개의 범주형 변수를 기준으로 관측치를 기록하는 표**이다. 표의 행과 열들이 이러한 범주형 변수에

해당한다. 즉, 범주형 자료 변수에 대해서만 만들 수 있는 표이다.

I *J 형태의 분할표 예시를 통해 분할표가 어떻게 생겼는지 살펴보자.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Y</td></tr>
<tr><td>1</td><td>…</td><td>J</td></tr>
<tr><td rowspan="3">X</td><td>1</td><td colspan="3" rowspan="3">I *J 개 칸</td></tr>
<tr><td>…</td></tr>
<tr><td>I</td></tr>
</table>

X변수는 I개의 수준, Y는 J개의 수준으로 이루어진 범주형 자료이며, 수준은 각 범주의 level이다. 위의 예시는 가장 간단한 형태인 두 개의 범주형 변수만을 구분한 것으로, 2차원 분할표의 형태이다.

■ 여러 차원의 분할표

1) 2차원 분할표 (I * J)

두 개의 변수만으로 분류한 분할표이다. (X는 설명변수, Y는 종속변수)

	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 도수를, n_{i+}, n_{+j} 는 각 열과 행의 주변(marginal) 도수를 표현한다. 여기서 '+'는 그 위치에 해당하는 도수를 모두 더했다는 의미의 첨자이다.

직관적인 예시를 들어서 성별과 커피 취향이라는 두 범주형 변수에 대해 분할표를 만들 수 있다.

성별에 따른 커피 취향				
	아메리카노	라떼	카푸치노	합계
남성	78	15	46	139
여성	49	23	37	109
합계	127	38	83	248

2) 3차원 분할표 (I * J * K)

삼원분할표는 세 변수를 분류한 분할표로, 설명변수와 반응변수 외에 제어변수(제한변수, control variable)가 추가된다.

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 분할표와 달리 **제어변수(control variable)** Z가 생겼다. 주목할 점은 왼쪽의 부분분할표에서 제어변수의 각 수준에서 분류된 걸 합쳐버리면 Z변수가 사라지면서 2차원 분할표인 주변분할표의 모양으로 바뀐다는 것이다.

부분분할표

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

주변분할표

성별	학회합격여부	
	합격	불합격
남자	11 + 16+ 14	25 + 4 + 5
여자	10 + 22 + 7	27 + 10 + 12

직관적인 예시로 세 개의 변수(X : 성별, Y : 합격 여부, Z : 학과) 를 이용하여 3차원 분할표인 부분분할표와 Z(학과)의 효과를 제거한 주변분할표를 그려봤다. 둘의 형태와 관계를 기억하면서 부분분할표와 주변분할표의 뜻을 알아보자.

- 부분분할표 (partial table)

  부분분할표는 제어변수 Z의 각 수준에서 X와 Y를 분류한 표로 고정된 Z의 한 수준에 대해서 XY의 관계를 보여준다. 즉, Z를 통제했을 때 Y에 대한 X의 효과를 알 수 있다. *(조건부연관성, 동질연관성을 알 수 있다. 연관성은 오즈비 파트에서 자세히 알아보자.)*

- 주변분할표 (marginal table)

  주변분할표는 부분분할표를 모두 결합해서 얻은 2차원 분할표로 제어변수 Z를 통제하지 않고 무시한다. *(주변연관성을 알 수 있는데, 이것도 오즈비 파트에서!)*

3차원 이상의 다차원 분할표도 존재하지만, 잘 쓰이지 않으므로 3차원까지 알아보도록 하자. *(우리는 3차원에 살고 있으니까)* 그리고 2주차에도 배우겠지만, 3차원 이상의 고차원에서는 모형으로 다루는 것이 효과적이다.

■ 비율에 대한 분할표

지금까지 도수에 대한 분할표를 다뤘다면 이번에는 비율에 대한 분할표를 알아보자. 비율은 각 도수인 n_{ij}를 전체 도수 n으로 나누어 주면 된다.

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

이전에 든 2차원 분할표를 다시 예로 들어보자.

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

괄호 안의 값이 각 칸의 비율이 되고, 이는 확률이므로 합치면 1이 된다. (확률의 합은 1이니까!)

분할표에서의 확률분포 용어를 정리해보자.

1) 결합 확률 (joint probability)

표본이 두 범주형 반응변수 X와 Y로 분류될 때, X의 i번째 수준과 Y의 j번째 수준을 동시에 만족하는 확률이다. 즉 i행과 j열에 속할 확률로 π_{ij} 로 표현한다. $(\sum \pi_{ij} = 1)$

2) 주변 확률 (marginal probability)

결합분포의 행과 열의 합이다. π_{i+}(행의 분포), π_{+j}(열의 분포)로 나타내며, +는 그 위치에 해당하는 전체 첨자들의 합을 나타낸다.

3) 조건부 확률 (conditional probability)

대부분의 분할표에서 하나의 변수는 반응변수(Y), 다른 변수는 설명변수(X)인 경우가 많다. 이때 X의 각 수준에서의 Y에 대한 확률을 조건부확률이라고 한다. 조건부확률의 분포를 조건부 분포라고 하며, 식은 $\frac{\pi_{ij}}{\pi_{i+}}$ 와 같다.

참고로 원서를 보면 π와 p가 혼재되어 나타나는데, 이 때 p는 π의 추정값, 즉 표본비율이다. 정리하면 π는 모비율, p는 표본비율, n은 도수를 표현한다.

3. 독립성 검정 (Test of Independence)

■ 독립성 검정이란?

독립성 검정은 두 범주형 변수가 독립인지 검정하기 위한 방법이다. 연속형 변수의 경우 두 변수 간의 관계를 알기 위해서 상관분석, 회귀분석 등을 활용하지만, 범주형 변수의 경우 상관계수처럼 변수 간의 상관성을 나타내기 어렵다. 따라서 독립성 검정을 통해 두 범주형 변수가 서로 통계적으로 유의한 관련성을 갖는지 파악할 수 있다. *(두 범주형 변수 간 관련성을 나타내는 측도로 파이계수, 크래머V 등의 통계량이 존재하지만, 우리는 가설 검정을 통해 연관성을 확인하는 독립성 검정에 대해서만 알아보도록 하겠다.)*

1) 독립성 검정의 목적

독립성 검정을 통해 우리는 **1) 두 변수가 연관성이 있는지 없는지 판단**할 수 있고, **2) 분석 가치를 판단**할 수 있다. 만약 두 변수가 독립(= 연관성이 없다)이라면 더 이상 분석을 진행할 이유가 없음(= 분석 가치가 없다)을 알 수 있다. 즉, 관계가 없는 두 변수에 대해 굳이 분석할 이유(분석 가치)가 없음을 독립성 검정을 통해 파악하는 것이다.

2) 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. ($\pi_{ij} \neq \pi_{i+} \cdot \pi_{+j}$)

독립성 검정의 가설은 다음과 같다. 분할표의 각 칸의 발생확률(π_{ij})이 각 교차표의 주변확률(π_{i+}, π_{+j})의 곱과 같다는 것이다. 다시 말하면 두 변수의 동시 발생 확률이 단순히 두 발생 확률의 곱으로 표현된다는 것이 귀무가설이다. *(교재나 다른 자료에서는 π 대신 표본비율 p를 쓰기도 한다.)*

3) 기대 도수와 관측 도수

독립성 검정의 방법을 이해하기 위해서 먼저 기대 도수와 관측 도수의 차이를 알아보자.

관측 도수 (observed frequency) $[n_{ij}]$ *('O'로도 표현)*

기대 도수 (expected frequency) $[\mu_{ij}]$ *('E'로도 표현)*

기대 도수는 전체 표본크기 n과 주변확률의 곱으로 구해진다. $[\mu_{ij} = n \times \pi_{i+} \times \pi_{+j}]$ 이 기대 도수는 귀무가설이 참일 때, 즉 두 변수가 독립일 때 각 칸의 도수에 대한 기댓값 $E(n_{ij})$ 을 나타낸다.

이 기대 도수와 관측 도수의 차이를 비교하는 방식으로 독립성 검정이 이루어진다. 곧 등장할 독립성 검정의 통계량들을 살펴보면 이해가 쉬울 것이다!

귀무가설 H_0 : $\mu_{ij} = n\pi_{ij}$

대립가설 H_1 : $\mu_{ij} \neq n\pi_{ij}$

독립성 검정의 가설을 기대 도수를 통해 다음과 같이 표현할 수 있다. 이는 결국 전체 도수 n만 안 곱했을 뿐 위의 가설(결합확률 = 주변확률의 곱)과 일치한다. 우리는 결국 기대 도수와 관측 도수의 차이 $[\ \mu_{ij} - n\pi_{ij}]$ 가 클수록 귀무가설을 기각(변수들은 서로 연관성이 있다) 할 수 있는 근거가 커지게 된다.

기대 도수의 추정값 : $\hat{\mu}_i = np_{i+}p_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}$

4) 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)

변수가 2개인 2차원 분할표에서 4가지 독립성을 배울 것이다. 대표본의 경우 카이제곱분포 근사를 통해 독립성 검정을 하고, 소표본의 경우 정확 검정을 통하여 독립성 검정을 한다. 우리는 대표본 독립성 검정 위주로 살펴보고자 한다.

- 3차원 분할표 독립성 검정

  - *Breslow-Day test (BD test) : 오즈비의 동질성 검정을 위해 고안된 카이제곱 검정법*
  - *Cochran-Mantel-Haenszel test (CMH test) : XY간의 조건부 독립성이 성립하는지 확인하는 카이제곱 검정법*

위 두 검정은 3차원 분할표(2 x 2 x K 분할표) 에서 사용할 수 있는 독립성 검정이다. 그렇지만 사실 3차원 이상의 고차원에서는 변수 간의 관계를 모형으로 다루는 것이 효과적인데, 로그 선형 모형을 통해 가능하다.

■ 명목형 자료의 독립성 검정

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

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

(또는 $X^2 = \sum \frac{(O-E)^2}{E}$ 로도 표현)

피어슨 카이제곱 검정 통계량 X^2은 스코어 통계량에 해당한다.

이 통계량은 모든 n_{ij}와 μ_{ij}가 같을 때, 즉 관측 도수와 기대 도수가 같을 때 최소값 0을 갖는다. 표본크기 n이 고정되어 있을 때, n_{ij}와 μ_{ij} 사이의 차이가 커지면 X^2가 커져서 결과적으로 귀무가설을 기각하는 증거가 더 강해진다.

피어슨 카이제곱 검정의 flow를 잘 알아 두자.

관측 도수와 기대 도수의 차이가 크다! -> 검정통계량 X^2가 크다! -> p-value가 작겠군! -> 귀무가설 기각! -> 변수 간의 연관성이 존재하겠구나!

$\mu_{ij} \geq 5$ 정도(대표본의 기준)라면 카이제곱 분포를 따른다.

식에 대한 유도는 각 칸의 도수가 포아송 분포를 따른다는 것을 이용한다. (도수자료가 포아송 분포를 따른다는 것은 알아 두자!)

$\frac{n_{ij}-E(n_i)}{\sqrt{Var(n_{ij})}} = \frac{n_{ij}-\mu_{ij}}{\sqrt{\mu_{ij}}} \sim N(0,1)$ *식을 카이제곱 분포 형태로 만들어주면* $X^2 = \sum \frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}}$ *이 된다.*

2) 가능도비 검정 (Likelihood-ratio test)

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

일반적인 가능도비 검정 통계량을 이차원 분할표에서의 검정 통계량으로 정리한 것이다. 위의 피어슨 카이제곱 검정과 마찬가지로 관측 도수 n_{ij}와 기대 도수 μ_{ij}의 차이가 커질수록 통계량 G^2가 커지는 원리이다.

관측 도수와 기대 도수의 차이가 크다! -> 검정통계량 G^2가 크다! -> p-value가 작겠군! -> 귀무가설 기-각! -> 변수 간의 연관성이 존재하겠구나!

G^2 역시 카이제곱분포에 근사하므로 위의 flow를 그대로 적용할 수 있다!

■ 순서형 자료의 독립성 검정

순서형 자료에 명목형 독립성 검정을 할 수는 있지만, 순서 정보의 손실이 일어나기 때문에 주의해야 한다. 생각보다 많은 범주형 자료가 순서형 자료이므로 잘 알아 두고 유용하게 사용하자.

1) MH 검정 ((Mantel-Haenszel test)

MH 검정은 두 범주형 변수가 모두 순서형일 때 쓰는 검정이다. *두 변수 중 한 범주가 오직 두 level (yes or no)를 갖는 명목형 변수일 때도 사용할 수 있지만 그 이상은 적절치 않다.*

범주의 각 level에 점수를 할당하여 변수 간의 선형추세를 측정한다.

- 행점수 : $u_1 \leq u_2 \leq \cdots \leq u_I$
- 열점수 : $v_1 \leq v_2 \leq \cdots \leq v_J$

이 때 두 변수의 추세 연관성을 파악하기 위해 **피어슨 교차적률 상관계수** r을 사용한다.

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

- $-1 \leq r \leq 1$
- $r = 0$ 일 때 독립

(공분산을 X와 Y의 표본표준편차의 곱으로 나눈 것으로 복잡해 보이지만 우리가 아는 상관계수 공식이다!)

MH 검정통계량은 다음과 같다.

$$M^2 = (n-1)r^2 \sim x_1^2$$

표본크기 n과 r이 커지면 통계량 M^2도 커지게 된다. 역시 같은 원리로 M^2이 커지면 귀무가설을 기각하고

변수 간의 연관성이 있다고 파악할 수 있다!

이 독립성 검정은 카이제곱 분포의 자유도가 1이기 때문에 위의 명목형 검정방법에 비해 자유도가 적다. 따라서 조금 표본수가 적더라도 카이제곱 분포에 더 근사한다는 장점이 있다.

■ 독립성 검정의 한계

독립성 검정은 두 범주형 변수가 연관성이 있는지 없는지 그 유무만을 판단하기 때문에 구체적으로 어떻게 연관이 있는지는 파악할 수 없다. 따라서 변수 간 연관성의 성질을 파악하기 위해 연관성 측도를 알아야 한다.

4. 연관성 측도

두 범주형 변수가 모두 2가지 범주만을 갖는 이항변수일 때 우리는 여러가지 방법으로 범주별 비교가 가능하다. 우리는 이항변수의 연관성을 나타내는 측도 세 가지를 알아볼 것이다.

비율의 비교 척도		
비율의 차이	상대 위험도	오즈비

이 중에서도 오즈비가 특히 중요하고 앞으로도 계속 사용하게 될 것이다. (로그 오즈비, 조건부 오즈비, 주변 오즈비 등등..) 그러니까 얼른 친해지도록 해보자!

■ 비율의 차이 (Difference of Proportions)

비율의 차이는 조건부 확률의 차이 $[\pi_1 - \pi_2]$ 이다.. π_i는 i번째 행의 조건부확률을 뜻하는데, 앞으로 비율의 비교 파트에서는 모두 조건부 확률을 사용한다는 것을 명심하자! 예시를 통해 알아보자.

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

이 분할표에서 비율의 차이는 위의 행의 성공확률 - 밑의 행의 성공확률을 의미한다. 0.814 - 0.793 = **0.021** 이므로, 여성일 경우 연인이 있을 확률이 0.021만큼 더 높다고 해석할 수 있다.

- 범위 : $-1 \leq \pi_1 - \pi_2 \leq 1$
- 독립일 때 $\pi_1 - \pi_2 = 0$

성별	연인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

이렇게 비율의 차가 0.4 - 0. 4 = 0이 되는 경우에 여성일 때 연인이 있을 확률이 남성일 때와 차이가 없다. 즉 성별이 연인 여부에 영향을 끼치지 않는 것으로, 반응변수와 설명변수가 독립이라는 것을 알 수 있다!

■ 상대위험도 (Relative Risk, RR)

상대위험도는 조건부 확률의 비 [$\frac{\pi_1}{\pi_2}$] 를 뜻한다. 비율의 차이가 조건부 확률끼리의 '차이' 였다면, 상대위험도는 '비율'인 것이다. 상대위험도가 클수록 변수 간에 연관성이 큰 것으로 간주한다. 같은 예시를 통해 알아보자.

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

이 때 상대위험도는 0.814/0.793 = **1.027…** 이다. 즉, 여성일 경우 연인이 있을 확률이 1.027**배** 높다고 해석하면 되는 것이다.

- 범위 : $\frac{\pi_1}{\pi_2} \geq 0$
- 독립일 때, $\frac{\pi_1}{\pi_2} = 1$
- 확률이 0이나 1에 가까울 때는 영향력 차이가 많이 난다. 비율의 차이가 낮아 보여도 상대위험도가 클 수 있으므로 주의가 필요하다.

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

이렇게 조건부 확률의 0이나 1에 가까울 때, 비율의 차이는 0.02 - 0.01 = 0.92 - 0.91 = 0.01 으로 매우 작지만, 상대 위험도는 0.02/0.01 = 2, 0.92/0.91 = 1.01 로 영향력에서 엄청난 차이를 보인다.

비율의 차이와 상대위험도는 직관적인 척도이지만 한 변수의 수를 고정시킨 조사에서는 사용이 불가하다는 단점이 있다. 반면 오즈비는 이러한 경우에도 문제 없이 사용될 수 있으며, 대칭적으로 구해지기 때문에 반응변수와 설명변수의 구별 없이 같은 값이 나오게 된다. 예시를 통해 알아보자.

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

알코올중독과 심장질환의 연관성을 보기 위해 심장질환을 가진 사람과 그렇지 않은 사람을 각각 50명, 100명씩 선정하여 알코올 중독여부와 비교한다고 했을 때, 우리는 상대위험도를 사용할 수 없다. 왜냐하면 각 관측치가 독립적으로 랜덤하게 선택된 것이 아니고 심장질환 여부에 의해 정해진 비율 혹은 숫자에 따라 선정된 집단이기 때문이다. 전체 표본 중 심장질환자의 비율이 1/3 인 것은 우리가 이미 그렇게 정해서 추출했기 때문이지, 실제로 3명 중 1명이 심장질환을 가졌다는 의미가 아니기 때문이다. 이러한 경우 추정된 확률은 의미가 없게 되고 추정된 확률의 비인 상대위험도 역시 소용이 없게 된다. 하지만 이런 경우 우리는 오즈비를 사용하면 된다! 가장 많이 사용하는 오즈비에 대해 알아보자.

위 예시처럼 이미 난 결과(심장질환)를 바탕으로 과거기록(알코올 중독 여부)을 관찰하는 연구를 후향적 연구라고 한다. **후향적 연구**는 위의 예시의 50명과 100명처럼 열이 고정되는 특징이 있다. 비율의 차와 상대위험도는 이러한 이유로 후향적 연구에서 사용할 수 없다.

■ 오즈비 (Odds Ratio, OR)

(드디어 1주차 교안의 하이-라이트 오즈비다!) 오즈비를 알기 전에 먼저 오즈란 무엇인지 알아보자.

1) 오즈 (Odds)

오즈(Odds)란 사전적 의미로는 어떤 일이 일어날 승산 (공산) 또는 가능성이라는 뜻으로, 정확한 의미는 성공확률 / 실패확률, 즉 실패 분의 성공을 의미한다.

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

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

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, 두번째 행의 오즈는 3.826이다. 이런 식으로 성공을 실패로 나눠주면 실패에 비해 성공이 몇배인지를 나타내는 오즈가 된다. 오즈는 범주형 자료분석에서 무궁무진하게 활용되므로 의미를 정확하게 파악해두자!

2) 오즈비 (Odds Ratio)

오즈비란 각 오즈의 비를 뜻한다. *(cf. 상대위험도는 두 조건부확률의 비)*

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

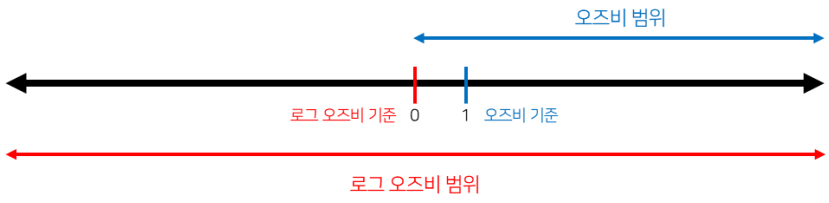
- 범위 : $\theta \geq 0$
- 독립일 때, $\theta = 1$ (= 두 행에서 성공의 오즈가 같다)
- $\theta > 1$ 이면 첫번째 행에서의 성공의 오즈가 두번째 행보다 높다. ($\pi_1 > \pi_2$ 와 같은 의미)
- $0 < \theta < 1$ 이면 첫번째 행에서의 성공의 오즈가 두번째 행보다 낮다. ($\pi_2 > \pi_1$ 와 같은 의미)
- 서로 역수관계에 있는 오즈비는 방향만 반대이고 연관성은 같은 정도이다. 예를 들어 오즈비가 4.0이거나 0.25인 사실은 단순히 어떻게 행과 열을 분류하여 표시하는가에 따라 결정되는 것이다.

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…	

이 예시에서 오즈비는 4.388/3.826 = 1.147 이 된다. 즉 여성이 연인이 있을 오즈가 남성이 연인이 있을 오즈보다 1.147배 더 높다고 할 수 있는 것이다!

- 로그 오즈비 (Log Odds Ratio)

로그 오즈비는 말그대로 오즈비에 log를 씌운 것이다. 왜 오즈비에 log를 씌울까? 오즈비의 범위를 생각해보면 쉽게 맞출 수 있다. 로그 오즈비는 기존의 비대칭한 오즈비의 범위를 교정한 측도이다.



1차원 좌표계를 통해 기존 오즈비와 로그 오즈비의 범위를 나타낸 것이다. 기존 오즈비의 기준이었던 1에 로그를 씌우면 0이 되면서 범위가 대칭으로 교정되었다. ($-\infty < \log\theta < \infty$)

- 오즈비의 장점

가. 오즈비는 상대위험도와 달리 한 변수가 고정되어 있을 때도 사용이 가능하다. (즉, 후향적 연구에서도 사용이 가능하다!) 이는 대조군의 크기가 달라져도 오즈비는 같기 때문이다.

말로만은 이해하기 어려운 설명이니 예시를 통해서 알아보자.

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

위의 분할표처럼 연인 있음(사례군)과 연인 없음(대조군)의 비율을 3:7로 고정한 후향적 연구를 진행했다고 가정해보자. 만약 아래의 분할표처럼 대조군의 크기를 달리하여 연구를 진행한다면 $P(Y|X)$가 바뀌게 된다.

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

대조군의 크기가 70에서 700으로 바뀌었을 때, 비율의 차와 상대위험도는 제멋대로 바뀌지만 오즈비는 $\frac{1/3}{1/2} = \frac{1/30}{1/20} = \frac{2}{3}$ 로 똑같다!! *(신기하죠?)*

이는 오즈비 값이 $P(Y|X)$를 사용하여 정의하나 $P(X|Y)$로 정의하나 서로 동일한 값을 갖기 때문이다. 이는 조건부확률 공식 $[P(Y|X) = \frac{P(X|Y)\times P(Y)}{P(X)}]$을 이용하면 쉽게 증명할 수 있다.

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

(Y=1이면 성공, Y=0이면 실패 / X=1이면 1그룹, X=2이면 2그룹)

나. 오즈비는 행과 열의 위치가 바뀌어도 같은 값을 갖는다.

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

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

위의 예시처럼 행과 열이 바뀌어도 오즈비는 (1/3)/(1/2) = (1/2)/(3/4) = 2/3 으로 같다! *(알면 알수록 매력적인 오즈비…)* 위와 같은 성질 덕분에 오즈비는 **교차적비 (cross-product ratio)**라고도 하는데, 이는 대각선에 있는 칸 확률들의 곱 $\pi_2/(1-\pi_2)$과 $\pi_2/(1-\pi_2)$의 비와 같기 때문이다. 교차적비는 다음과 같이 더 쉽게 구할 수 있다.

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

3) 3차원 분할표에서의 오즈비

- 조건부 독립성과 주변 독립성

위의 분할표 파트에서 3차원 분할표인 부분분할표와 제어변수를 합쳐 표현한 주변분할표를 다시 떠올려보자. 부분분할표에서의 연관성을 **조건부 연관성(conditional association)**이라고 하는데, 이는 제어변수 Z의 값이 어떤 수준에서 고정되어 있다는 조건 하에서 X와 Y의 연관성을 뜻한다.

조건부 연관성은 다음과 같이 **조건부 오즈비**를 통해 알 수 있다. *(Z의 각 수준별로 교차적비를 구하는 거라고 생각하면 된다. 어렵지 않다!)*

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	

조건부 오즈비가 모두 같다면 [$\theta_{XY(1)} = \theta_{XY(2)}$=…] , 즉 Z의 각 수준에서 XY의 연관성이 모두 같은 경우를 **동질 연관성(homogeneous association)**이라고 정의한다. 동질연관성은 대칭적이기 때문에 XY에 동질연관성이 존재하면 YZ와 XZ에도 동질연관성이 존재한다.

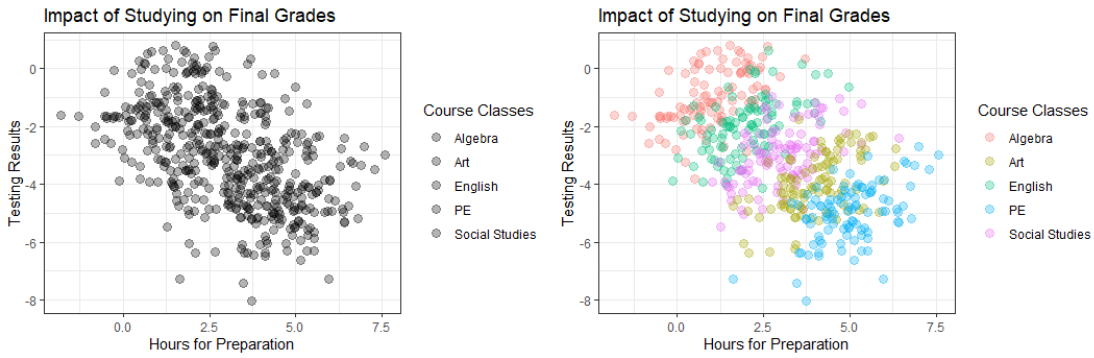
조건부 오즈비가 모두 1로 같다면 [$\theta_{XY(1)}$= … = $\theta_{XY(K)}$= 1] , 즉 XY가 서로 독립인 경우를 **조건부 독립성(conditional association)**이라고 한다. 조건부독립성은 동질연관성의 특별한 경우라고 할 수 있다. (오즈비가 모두 같으니까!)

이제 제어변수 Z를 합쳐서 주변분할표를 만들어보자.

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	

주변오즈비는 제어변수를 합쳐버린 주변분할표에서의 오즈비이다. 이 오즈비가 1일 때[θ_{XY+} = 1], **주변 독립성**을 갖는다고 할 수 있다.

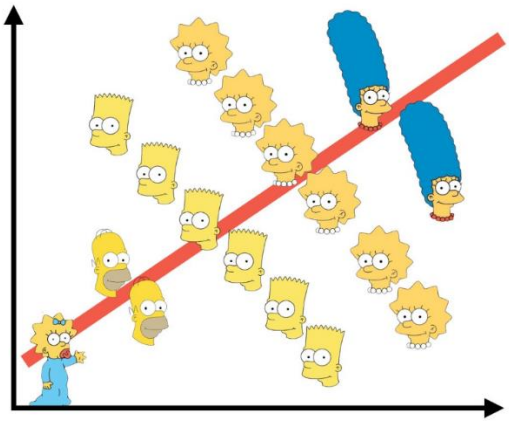
여기서 주의할 부분은 **조건부독립성이 성립한다고 해서 주변독립성이 성립되는 것은 아니다!** 조건부 오즈비와 주변 오즈비의 방향성이 항상 같은 것은 아니기 때문이다. 이게 대체 무슨 뜻인지 말로만 들으면 아리송할 수 있다. 아래의 플랏을 해석해보자.



이 플랏을 완벽히 해석할 줄 아는 멋진 범주러가 되기 위해… 1주차 클린업 마지막 파트인 심슨의 역설에 대해 알아보자.

- 심슨의 역설 (Simpson's Paradox)

심슨의 역설이란 영국의 통계학자 에드워드 심슨이 정리한 역설로, 전반적인 추세가 경향성이 존재하는 것으로 보이지만, 그룹으로 나눠서 개별적으로 보게 되면 경향성이 사라지거나 해석이 반대로 되는 경우를 말한다. 즉, 조건부 오즈비와 주변 오즈비의 연관성 방향이 다르게 나타나는 경우를 말한다.



사실 이 그림 한 장으로 이 개념은 말끔히 정리된다.

이처럼 조건부 오즈비와 주변 오즈비의 방향성이 항상 같은 것은 아니다. 즉, 조건부 연관성과 주변 연관성이 다를 수 있다는 것이다! 이렇게 연관성이 달리 나타나는 경우를 심슨의 역설이라고 하며 분할표를 해석할 때나 플랏을 해석할 때 이러한 상황이 나타날 수 있음을 유의하며 분석해야 한다.

이번엔 심슨의 역설이 나타나는 분할표의 예시를 살펴보자.

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

부분분할표에서 조건부 오즈비는 $\theta_{XY(1)} = 0.43,\ \theta_{XY(2)} = 0$ 이고, 주변분할표에서는 주변 오즈비가 $\theta_{XY+} = 1.446$이다. 오즈비는 1이 기준이므로, 이는 연관성의 방향이 서로 반대임을 알 수 있다.

두 분할표를 자세히 보면 제어변수인 학과에 따라서 도수의 차이가 상당히 크다는 것을 알 수 있다. 즉 제어변수인 학과가 중요한 변수로 작용했기 때문에 이를 무시하는 주변 분할표에서는 다른 결과가 나오는 것이다!