Categorical Data Analysis

[1주차 클린업 교안]

안녕하세요 여러분! 범주형자료분석팀에 오신 것을 진심으로 환영합니다! 👏👏👏

3주간 진행될 클린업 동안 즐겁고 유익한 시간이 되도록 많이 노력할게요!

언제든 궁금하거나 의구심이 드는 부분이 있다면 주저하지 않고 찾아주세요

저는 언제나 여러분들을 기다린답니다…

최고의 범주피플들이 되어 세상에 나가는 그 날을 기다리며~

그럼 신나는 클린업 1주차를 시작해볼까요~?😇

임지훈 010-3499-4655

목차

임지훈 010-3499-4655

Ⅰ. 범주형 자료분석(Categorical Data Analysis, CDA)

범주형 자료분석이란 반응변수가 범주형인 자료에 대한 분석을 뜻한다. 우리는 이 문장을 이해하기 위하여 **변수**와 **자료**의 의미에 대해 정확히 확인해보아야 한다.

자료(Data)란 변수와 관측치의 집합이다. 이 때, 분석하고자 하는 모집단의 특징을 변수, 변수를 측정한 값을 관측치라고 한다. 자료는 주로 각 변수를 열로 삼고 변수 별 관측치를 행으로 나열한 행렬 형태를 띈다.

예를 들어, P-SAT 학회원들을 모집단으로 삼았을 때, 나이, 전공, 성별, 거주지 등이 변수가 될 것이며, 각 학회원들의 측정값이 관측치가 되어 행에 순차적으로 나열되어 자료의 형태를 띌 것이다.

1. 변수의 구분

변수는 다음과 같이 크게 두 가지로 구분할 수 있다.

1)Y 변수 : 종속변수 / 반응변수 / 결과변수 / 표적변수

2)X 변수 : 독립변수 / 설명변수 / 예측변수 / 위험인자 / 요인 (범주형)

이처럼 같은 역할을 하는 변수도 다양한 이름으로 바꿔 부를 수 있다. 따라서 앞서 설명한 범주형 자료분석의 정의를 다시 살펴보면, 범주형 자료가 반응변수, 즉 Y 변수인 자료를 분석한다는 뜻이다.

2. 자료의 형태

다음으로 범주형 자료에 대해 알아보기 위해 자료의 형태가 어떻게 구분되는지 살펴보자. 자료가 어떤 형태를 띄는지에 따라 분석 방법이 달라질 수 있으므로 자료를 정확하게 파악하는 것은 매우 중요하다.

<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>

자료는 크게 양적 자료 (=수치형 자료)와 질적 자료 (=범주형 자료)로 나뉜다.

1) 양적 자료 (수치형 자료)

양적 자료는 관측된 값이 수치로 측정되는 자료를 말한다. 양적 자료는 관측되는 값의 성질에 따라 다시

이산형 자료와 연속형 자료로 구분된다.

① 이산형 자료 : 값을 셀 수 있는 자료

예) 나이, 금년도 성균관대학교 신입생 수

② 연속형 자료 : 연속인 어떤 구간에서 값을 취하는 자료

예) 키, 몸무게 (키를 170cm라고 말하지만 실제론 170.041…과 같이 이산적으로 표현할 수 없다.)

양적 자료는 공분산과 상관계수 등의 수치적 공식 사용이 가능하며, 정규분포를 통해 일반회귀분석이나 ANOVA (분산분석)가 가능하다. 이 부분은 회귀분석팀 클린업을 통해 공부할 수 있으므로, 우리는 질적 자료, 즉 범주형 자료에 초점을 맞춰보자!

2) 질적 자료 (범주형 자료)

질적 자료는 관측 결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료를 말한다. 질적 자료는 범주들의 성질에 따라 다시 명목형 자료와 순서형 자료로 구분된다.

① 명목형 자료 : 범주간에 순서의 의미가 없는 자료

예) 혈액형, MBTI

범주팀 회식 장소 (명목형 자료 예시)			
철문집	홍곱창	명쭈삼	두찜

② 순서형 자료 : 범주간에 순서의 의미가 있는 자료

예) 선호도 ('매우 좋다 / 좋다 / 보통 / 싫다 / 매우 싫다)

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

범주형 자료는 다음과 같은 특징을 가지고 있다.

1) 순서형 자료에 명목형 자료 분석방법을 적용할 수는 있으나, 분석하는 과정에서 순서에 대한 정보가 무시되기 때문에 검정력에 심각한 손실을 가져온다.

(반대로 명목형 자료에는 순서에 관한 정보가 없으므로 순서형 자료에 대한 분석법을 명목형 자료에 적용할 수 없다!)

2) 분할표를 작성할 수 있다. (바로 다음 챕터에서 만나요!)

3) 각 범주에 특정 점수를 할당하여 양적자료로 활용할 수 있다. (3주차에서 자세히!)

수치형 자료처럼 표현되어 있는 범주형 자료를 잘 구분해야 한다. 자료가 숫자로 표현되어 있다고 반드시 수치형 자료인 것이 아니다.

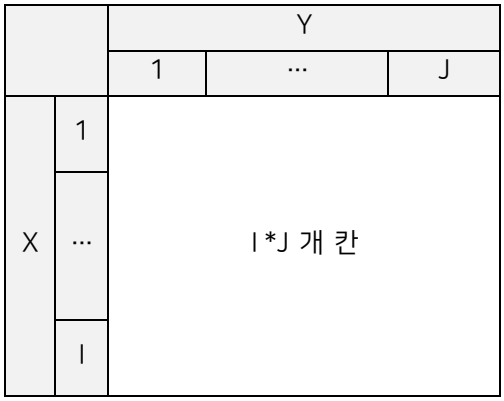
예를 들어 서울을 1, 서울 외의 지역을 0으로 표현한 자료가 있다고 하자. 여기서 0과 1 숫자 간에 수치적 정보가 포함되어 있다고 해석하는 것은 옳지 않다. 이처럼 숫자 형식의 자료가 수치형 자료인지, 범주형 자료인지 꼼꼼히 확인해보도록 하자.

II. 분할표 (Contingency Table)

분할표란 범주형에 속하는 변수들에 대한 관측값들이 도표로 요약된 자료를 일컫는다. 이는 중심 (평균, 중간값) 과 산포도 (분산, 표준편차) 등의 기술통계를 통해 진행하는 수치형 변수 분석과는 차이가 있다.

1. 분할표

아래 그림을 통해 분할표의 구성 요소에 대해 자세히 알아보자.



X 변수와 Y 변수, 총 2개의 범주형 변수가 주어졌다고 했을 때 각 변수의 카테고리 개수를 **수준(Level)** 이라고 부른다. X 변수가 성별이라면 수준은 총 2개 (남, 여) 가 될 것이다. 변수의 수준에 따라 분할표의 크기가 정해진다. 위 분할표에서 X 변수는 I개의 수준, Y 변수는 J개의 수준을 갖고 있기 때문에 분할표는 I * J 크기의 행렬의 모습을 띄고 있다.

위와 같이 범주형 자료를 분할표로 표현함으로써

1) 예측 검정력에 대한 요약이 가능해지고, (3주차 클린업에서 만나요!)

2) 독립성 검정을 실시할 수 있다. (다음 챕터에서 만나요!)

2. 여러 차원의 분할표

분할표는 2개 이상의 범주형 변수가 주어졌을 때 차원과 수준에 따라 무한가지 경우의 형태로 만들 수 있다. 하지만 복수의 범주형 변수가 주어졌을 때 분할표보다는 모델링 등의 방식을 통한 분석이 더 큰 편의성을 갖고 있으므로 본 클린업에서는 2차원 분할표와 3차원 분할표에 관하여 중점적으로 다뤄보자.

1) 2차원 분할표 (I * J)

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

위 도표는 두 개의 범주형 변수를 분류한 분할표이다. 만약 두 개의 변수가 서로 설명변수와 반응변수의 관계라면 일반적으로 설명변수를 행에, 반응변수를 열에 위치시킨다.

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

2) 3차원 분할표 (I * J * K)

세 가지 범주형 변수를 분류한 분할표로, 기존의 설명변수와 반응변수에 K개의 수준을 가진 제어변수(제한변수, Control Variable) Z가 추가된 형태를 일컫는다.

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 도표가 3차원 분할표의 형태이다. 만약 Z 변수 각 수준에서의 도수를 합쳐버리면 오른쪽 도표와 같이 2차원 분할표로 변환된다. 왼쪽의 도표를 '**부분분할표**,' 오른쪽 도표를 '**주변분할표**' 라고 부른다. 아래 도표는 부분분할표와 주변분할표의 예시자료이다.

부분분할표				
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

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

① 부분분할표

부분분할표란 X 변수와 Y 변수가 Z 변수 (제어변수)의 수준에 따라 분류된 분할표이다. 부분분할표를 통해 Z 변수의 각 수준에서의 X 변수와 Y 변수 간의 관계를 확인할 수 있다.

위 예시에서 제어변수(Z)인 '학과' 별 학회 합격 여부(Y)에 성별(X)이 미치는 영향을 확인할 수 있다.

② 주변분할표

주변분할표란 Z 변수 (제어변수)의 수준을 결합하여 만든 2차원 분할표로 X 변수와 Y 변수 간의 관계에서 Z 변수의 영향력을 제거시킨 형태이다.

위의 예시에서는 학과와 무관하게 학회 합격 여부(Y)에 성별(X)이 미치는 영향만을 확인할 수 있다.

부분분할표와 주변분할표를 통해 변수 간 연관성을 파악할 수 있는데, 이는 **오즈비** 파트에서 자세히 다뤄보는 걸로 하자!

3. 비율에 대한 분할표

비율에 대한 분할표는 각 칸에 도수 대신 비율이 들어간 분할표이다. 이 때 비율은 각 칸의 도수인 n_{ij}를 전체 도수 n_{++}으로 나누어 주면 된다.

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

π_{ij} : 전체 대비 각 칸의 비율 (즉, 확률) / π_{++} : 분할표 내 모든 칸의 확률의 합 (즉 1!)

1) 분할표에서의 확률분포

① 결합 확률 (Joint Probability) : π_{ij}

결합 확률이란 모집단에서 추출된 표본이 X 변수의 I 번째 수준과 Y 변수의 J 번째 수준을 동시에 만족할 확률로, 위 비율에 대한 분할표에서 각 칸의 확률을 말한다. 전체 확률의 합은 항상 1 이므로, 분할표 내 결합 확률의 합 역시 1이다. ($\sum \pi_{Ij} = 1$)

② 주변 확률 (Marginal Probability) : π_{i+} , π_{+j}

주변 확률이란 X 변수의 I 번째 수준이 전부 일어날 확률 (π_{i+}) 또는 Y 변수의 J 번째 수준이 전부 일어날 확률 (π_{+j}) 을 말한다. 주변 확률 역시 분포함수이기 때문에 해당 범위 내의 확률들의 합은 1 이다. ($\sum_I \pi_{i+} = \sum_J \pi_{+j} = 1$)

③ 조건부 확률 (Conditional Probability)

앞서 설명했듯, 대부분의 분할표에서 행에 위치한 X 변수는 설명변수, 열에 위치한 Y 변수는 반응변수의 역할을 한다. 이 때 X 변수의 각 수준에서의 Y 변수의 값을 조건부 확률이라고 한다. 식으로 표현하면 $\frac{\pi_{ij}}{\pi_{i+}}$ 로 표현 가능하다.

2) 확률분포 예시

아래의 도표는 연령대에 따른 선호 업종의 관계를 나타낸 분할표이다. 이를 통해 앞서 결합 확률과 주변

확률, 조건부 확률을 계산해보도록 하자.

연령대에 따른 희망 직종				
	의사 (Y=1)	회계사 (Y=2)	엔지니어 (Y=3)	합계
10대 (X=1)	78 (0.31)	23 (0.09)	29 (0.12)	130
20대 (X=2)	41 (0.16)	42 (0.17)	37 (0.15)	120
합계	119	65	66	250

위 도표에서 20대이면서 회계사를 희망할 **결합 확률**은 $\pi_{22} = \frac{42}{250}$ =약 0.17이다.

그리고 연령대에 무관하게 엔지니어를 희망할 **주변 확률**은 $\pi_{+3} = \frac{66}{250} =$ 0.264이다.

마지막으로 20대라는 가정 하에 의사를 희망할 **조건부 확률**은 $\frac{\pi_{21}}{\pi_{2+}} = \frac{41}{120}$ = 약 0.34이다.

III. 독립성 검정 (Test of Independence)

통계학원론에서 배웠듯이, 분할표가 주어졌을 때 우리는 적합도, 동질성, 독립성 검정을 실시할 수 있다. 적합도 검정이란 실제로 얻어진 관측치들의 분포가 예상한 이론의 분포와 같은 지 검정하는 것이다. 동질성 검정이란 서로 다른 모집단에서 표본추출 했을 때, 각 그룹의 확률분포가 같은 지 검정하는 것이다. 마지막으로 이번 장에서 설명하려는 독립성 검정이란 <u>두 범주형 변수 사이의 관계를 확인하기 위한 검정</u>이다.

1. 독립성 검정의 목적

독립성 검정을 통해 우리는 1) 두 변수 간 연관성 유무와 2) 분석 가치를 판단할 수 있다. 만약 독립성 검정을 통해 두 변수 간 연관성이 없다 (=독립이다) 는 결과가 도출된다면 주어진 두 변수간 관계는 더 이상 분석의 가치가 없다. 설명변수가 반응변수에 어떠한 영향도 미치지 못한다는 뜻이기 때문이다.

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 귀무가설과 대립가설은 위와 같이 설정한다. 통계학에서 X와 Y가 서로 독립이라는 것은 $P(Y|X)=P(Y)$, 즉 $P(X \cap Y)=P(X)*P(Y)$가 성립함을 의미한다. 마찬가지로 분할표 상에서 두 변수가

독립이라는 것은 모든 결합 확률이 행과 열 주변 확률의 곱과 같다($\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$)는 것이 귀무가설의 내용이다.

3. 기대도수와 관측도수

독립성 검정을 진행하기 위해 기대도수와 관측도수에 대한 이해가 필요하다.

1) 관측도수 [n_{ij}] ("O" 로도 표현)

실제 관측값을 의미하며, 분할표 내에 표시된 각 칸의 도수와 같다. 비율에 대한 분할표가 제시된 경우, 각 칸의 비율에 전체 표본 n을 곱하면 관측도수를 얻을 수 있다. ($n_{ij} = n * \pi_{ij}$)

2) 기대도수 [μ_{ij}] ("E" 로도 표현)

귀무가설 하에 계산된 각 칸에 해당하는 기댓값을 의미한다. 즉, 두 변수가 서로 독립이라는 가정 하에, 결합 확률 대신 행과 열 주변확률의 곱을 전체 표본 n에 곱하여 얻을 수 있다. ($\mu_{ij} = n \times \pi_{i+} \times \pi_{+j}$)

앞서 제시된 귀무가설과 대립가설을 기대도수와 관측도수를 이용하여 표현하면 다음과 같다.

$$\text{귀무가설 } H_0 : \mu_{ij} = n_{ij}$$

$$\text{대립가설 } H_1 : \mu_{ij} \neq n_{ij}$$

귀무가설 하에, 관측도수 $n_{ij} = n * \pi_{ij}$와 기대도수 $\mu_{ij} = n \times \pi_{i+} \times \pi_{+j}$가 같다는 것은 $\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$과 같은 의미이므로 결국 두 가설들은 의미가 같다. 본 가설 검정을 위해 기대도수와 관측도수 간의 차이를 검정통계량과 비교해볼 것이다. 이 때, 표본의 크기나 범주형 변수의 특징에 따라 적합한 검정 방법을 채택한다.

4. 독립성 검정의 종류

1. 2차원 분할표 독립성 검정

<table>
<tr><td rowspan="3">대표본</td><td rowspan="2">명목형</td><td>피어슨 카이제곱 검정 (Pearson's chi-squared test)</td></tr>
<tr><td>가능도비 검정 (Likelihood-ratio test)</td></tr>
<tr><td>순서형</td><td>MH 검정 (Mantel-Haenszel test)</td></tr>
<tr><td colspan="2">소표본</td><td>피셔의 정확검정 (Fisher's Exact test)</td></tr>
</table>

2차원 분할표 독립성 검정은 표본의 크기에 따라 일차적으로 구분된다. 여기서 대표본이란 모든 기대도수

(μ_{ij})가 ≥ 5인 것을 의미한다. (과거에는 도수 n 값을 기준으로 모든 도수(n_{ij})가 30 이상일 때 대표본이라고 정의했다. 이는 범주형 자료분석에서 활용하는 t-분포가 n 값이 커질수록 중심극한정리(Central Limit Theorem)에 의해 표준정규분포에 근사하기 때문이다. 실제로는 t-분포를 살펴보면 n 값이 100 이상일 때 표준정규분포에 수렴하는 것으로 관찰된다.)

또한 대표본인 경우 범주형 변수가 명목형 혹은 순서형 인지에 따라 검정 방법이 구분된다. 이 중 클린업을 통해 우리는 2차원 분할표 중 대표본인 경우의 3가지 검정 방법에 대해 알아보자!

✓ 3차원 분할표 독립성 검정

3차원 분할표 독립성 검정은 대표적으로 Breslow-Day test (BD test)와 Cochran-Mantel-Haenszel test (CMH test)가 있다. 하지만 고차원(3차원 이상)에서의 독립성 검정은 모형을 통한 검정이 가장 효과적이며, 로그 선형 모형을 주로 사용한다. 이 부분은 교안 마지막 부록 부분 참고!

1) 대표본 + 명목형 자료 독립성 검정

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

독립성을 가정한 귀무가설 하에서 관측도수와 기대도수 간의 차이가 없으므로($n_{ij} = \mu_{ij}$), 검정통계량은 0이 된다. 하지만 기대도수와 관측도수 간의 차이가 커질수록 검정통계량 값이 커져 귀무가설이 기각될 (=변수들 간의 독립성이 부정될) 확률이 높아진다.

✓ 검정통계량은 어떻게 카이제곱분포를 따르는가?

도수자료는 포아송 분포를 따른다는 기본 가정 하에, 대표본이기 때문에 포아송 분포의 정규근사가 가능해진다. 따라서 다음과 같이 표현 가능하다.

$$\text{Poisson}(\mu) \sim N(\mu, \mu) \ \cdots \text{(포아송 분포는 평균과 분산이 같다.)}$$

즉, 표본의 크기가 크다는 피어슨 카이제곱 검정의 조건 하에 도수자료는 정규분포를 따르므로, 각 도수(n_{ij})를 표준화시키면, 표준정규분포를 따르는 형태로 변환 가능하다.

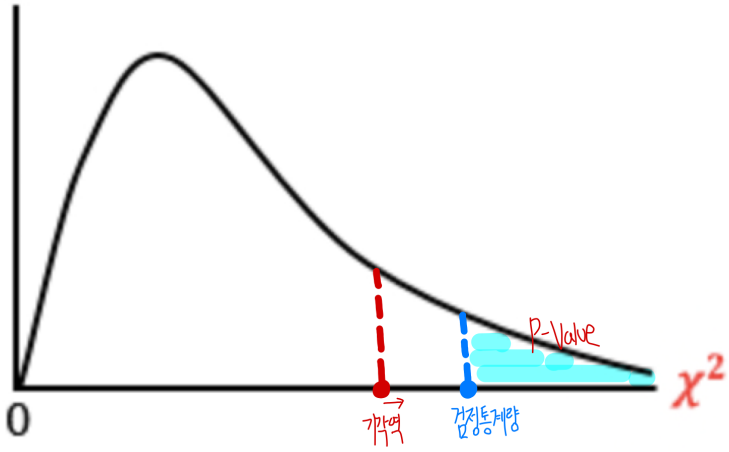
$$\frac{n_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}}} \sim Z(0,1)$$

여기서 표준정규분포의 제곱들의 합이 카이제곱분포를 따른다는 성질을 이용하면,

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

즉, 검정통계량이 자유도가 $(I-1)(J-1)$인 카이제곱분포를 따른다는 것을 알 수 있다!

따라서 정리해보자면, 아래와 같은 그림의 과정을 통해 귀무가설을 기각 혹은 채택함으로써 변수들 간의 관계를 확인할 수 있게 된다.



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

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

가능도비 검정은 일반화 가능도비 검정을 분할표에 적용한 방법으로, 피어슨 카이제곱 검정과 마찬가지로 기대도수와 관측도수 간의 차이를 통해 독립성을 확인한다. $\log\left(\frac{n_{ij}}{\mu_{ij}}\right)$는 곧 $\log n_{ij} - \log \mu_{ij}$ 이므로 둘 간의 차이가 커질수록 검정통계량이 커져 귀무가설을 기각할 확률이 높아진다. 이 때 G^2는 X^2과 근사적으로 같아지므로 피어슨 카이제곱 검정과 같은 분석 방법을 적용할 수 있다.

관측 도수와 기대 도수의 차이가 크다! -> 검정통계량 G^2가 크다! -> p-value가 작겠군! -> 귀무가설(두 변수는 독립) 기각! -> 변수 간의 연관성이 존재하겠구나!

2) 대표본 + 순서형 자료 독립성 검정

① MH 검정 (Mantel-Haenszel test)

두 변수가 모두 순서형 자료일 때 사용 가능하다. 혹은 두 변수 중 하나가 (예, 아니요) 와 같이 두 개의 수준만을 가진 명목형 변수일 때도 사용 가능하다. (그 이상은 불가능)

각 변수의 수준(Level)에 차등적인 점수를 할당함으로써 선형 추세를 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 각 수준 간 점수의 차이는 반드시 동등할 필요가 없으며, 부여 방식 역시 목적에 따라 다양하다.

- 검정통계량 : $\mathrm{M}^2 = (\mathrm{n}-1)\mathrm{r}^2 \sim \mathrm{x}_1^2$
- 기각역 : $\mathrm{M}^2 \geq \mathrm{x}_{\alpha,1}^2$

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

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

위 식은 복잡해 보이지만, 결국 공분산(분자)을 두 표준편차의 곱(분모)로 나눈다는 점에서 결국 우리가 알고 있는 상관계수와 같은 형태임을 알 수 있다. 이 때, 상관계수 r의 범위는 $-1 \leq r \leq 1$로 설정되며, r = 0일 때 두 변수는 독립이며 -1과 1에 가까워질수록 두 변수 간의 연관성이 커지는 것이다.

따라서 표본의 수인 n과 피어슨 교차적률 상관계수 $|r|$의 값이 커질수록 검정통계량 M^2의 값이 커져 귀무가설을 기각할 확률이 높아진다. 검정의 흐름을 정리해보면 다음과 같다.

상관계수 |r|이 크다! -> 검정통계량 M^2가 크다! -> p-value가 작겠군! -> 귀무가설(두 변수는 독립) 기각! -> 변수 간의 연관성이 존재하겠구나!

5. 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.

IV. 연관성 측도 (Test of Independence)

두 범주형 변수가 모두 2가지 수준만을 갖는 이항변수일 때, 우리는 세 종류의 척도들을 통해 변수 간 연관성을 파악할 수 있다.

비율의 비교 척도		
비율의 차이	상대 위험도	오즈비

여기서 비율은 각 행에 따른 조건부 확률을 의미한다. 각 척도들의 특징을 비교해 보고, 특히 **오즈비**에 대해 자세히 알아보도록 하자.

1. 비율의 차이 (Difference of Proportions)

비율의 차이는 각 행의 조건부 확률 간 차이($\pi_1 - \pi_2$)를 의미한다. 이 때 π_i는 분할표에서 i번째 행의 조건부 확률을 나타낸다. 이 때 확률은 항상 0~1 범위 안에 존재하므로 비율의 차이는 -1과 1 사이의 값이 될 것이다. 아래 예시를 통해 더 자세히 알아보도록 하자.

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

위의 분할표에 표시된 확률은 행(성별)에 따른 연인 유무의 조건부 확률을 의미한다. 이 때 연인이 있을 경우의 비율의 차이를 알고 싶다면 두 확률을 빼면 된다. 따라서 0.814-0.793= 약 0.021만큼 여성일 때 연인이 있을 확률이 높다고 할 수 있다.

성별	연인 유무	
	있음	없음
여성	0.4	0.6
남성	0.4	0.6

반면, 위의 도표의 경우, 연인이 있을 경우의 비율의 차이를 계산해 보았을 때, 0.4-0.4= 0이라는 것을 알

수 있다. 이는 성별에 무관하게 연인이 있을 확률이 같다는 뜻으로, 성별이 연인 유무에 영향을 미치지 못한다고 해석할 수 있다. 따라서 비율의 차이에 따르면 $\pi_1 - \pi_2 = 0$일 때 두 변수가 서로 독립이다.

하지만 조건부 확률이 0에 가까워질수록 반응변수에 대한 두 집단의 영향력의 차이가 커지지만, 비율의 차이는 이를 정확히 해석하는데 한계가 있다. 자세한 내용은 다음 상대위험도를 통해 살펴보자

2. 상대위험도 (Relative Risk)

상대위험도는 조건부 확률의 비($\frac{\pi_1}{\pi_2}$)를 뜻한다. 비율의 차이는 두 조건부 확률의 **차이**를 통해 연관성을 파악했다면, 상대위험도는 두 조건부 확률의 **비율**을 통해 연관성을 확인한다. 상대위험도가 클수록 두 변수의 연관성이 크다고 간주한다. 아래의 분할표를 통해 더 자세히 알아보도록 하자.

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

비율의 차이와 같은 예시의 분할표이지만, 상대위험도는 행(성별)에 따른 연인이 있을 조건부 확률의 비율을 통해 연관성을 파악한다. 즉, 0.814/0.793 = 1.027… 이므로, 여성일 경우 연인이 있을 확률이 약 1.027배 높다고 해석할 수 있는 것이다.

상대위험도($\frac{\pi_1}{\pi_2}$) 값은 ≥ 0이 될 것이며, $\frac{\pi_1}{\pi_2} = 1$ **일 때 두 변수가 서로 연관성이 없는 독립**의 관계라고 말할 수 있다.

✓ 조건부 확률이 0에 가까운 경우 (비율의 차이와 비교)

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

위의 두 분할표와 같이 조건부 확률이 0과 1에 가까운 경우에 성별에 따른 연인이 있을 확률을 비율의 차이와 상대위험도로 각각 계산해 보았을 때,

- 비율의 차이 : 0.02 – 0.01 = **0.01** , 0.92 - 0.91 = **0.01**
- 상대위험도 : 0.02/0.01 = **2** , 0.92/0.91 = **1.01**

위의 결과에서 알 수 있듯이 비율의 차이와 비교했을 때 상대위험도의 값 사이의 차가 훨씬 더 큰 것을 알 수 있다. 즉, 상대위험도에 따르면 매우 높은 연관성을 띄는 두 변수가 비율의 차이에 따르면 연관성이 매우 작은 것으로 잘못 해석될 수 있는 것이다. 따라서 조건부 확률이 0 혹은 1에 가까울 때 비율의 차이만으로 연관성을 판단하는 것은 매우 위험하다!

✓ 후향적 연구 (비율의 차이와 비교)

비율의 차이와 상대위험도는 두 변수간 연관성을 파악하는 직관적인 척도이지만, 후향적 연구처럼 한 변수의 수를 고정시킨 조사에서는 사용할 수 없다는 한계를 지닌다. 후향적 연구란, 이미 나온 결과를 바탕으로 과거 기록을 관찰하는 연구를 뜻한다.

	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

보통 후향적 연구는 사례군(Y=1)에 해당하는 사례가 많이 없는 보건 분야에 사용되기에, 랜덤으로 표본을 뽑지 않고 연구자가 정한 비율이나 숫자에 따라 열의 분포가 정해진다. 이런 후향적 연구의 경우 비율의 차이나 상대위험도를 사용하는 것이 불가능하다. 전체 표본 중 심장질환 환자의 비율이 1/3 인 것은 우리가 이미 그렇게 정해서 추출했기 때문이지, 실제 3 명 중 1 명이 심장질환을 가졌다는 의미가 아니기 때문이다. 그러므로 연구자가 대조군 (Y=0)의 합을 변경하게 된다면 그에 따라 비율의 차이와 상대 위험도도 달라지기 때문에 비율의 차이와 상대위험도는 후향적 연구에선 아무런 소용이 없게 된다.

가령 위의 경우 조건부확률을 따져보면, $\pi_1 = 0.66$, $\pi_2 = 0.32$인데, 만약 표본을 뽑을 때 건강한 사람 (Y=0)을 300 명으로 바꾼다면, 조건부 확률 값이 달라지게 돼서 비율의 차이나 상대위험도 값도 달라지게 된다. 이런 비율의 차이와 상대위험도의 한계점을 극복한 개념이 바로 오즈비 (Odds Ratio) 이다.

3. 오즈비 (Odds Ratio)

1) 오즈 (Odds)

오즈란 사전적으로는 어떤 일이 일어날 가능성이라는 의미이다. 달리 표현하면 "**성공확률 / 실패확률**", 즉 성공확률을 실패확률로 나눈 값이다. 반응변수에 두 가지 수준이 있을 때 이 중 한 가지를 대상으로 삼아 해당 경우가 발생하는 상황을 '성공' 이라고 표현하여, '성공'이 발생할 확률을 '성공확률'이라 부르고, 자연스럽게 다른 한 수준이 발생할 확률을 '실패확률'이라고 부른다. 수식으로 표현하면 다음과 같다.

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

여기서 π란 앞서 설명한 "성공확률"을 의미한다. 아래 예시를 통해 자세히 알아보자.

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…	

연인이 있을 확률을 π라고 할 때, 여성이 연인이 있을 오즈는 약 4.388이 될 것이며, 남성이 연인이 있을 오즈는 약 3.826이 될 것이다. 즉, 오즈는 성공확률이 실패확률의 몇 배인지 나타낸다.

2) 오즈비 (Odds Ratio)

오즈비는 **각 행 별로 계산한 오즈의 비**를 의미하여, 0보다 크거나 같은 값을 가진다.

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

앞선 예시를 통해 확인해보자. 여성이 연인이 있을 오즈 $(odds1)$는 약 4.388 이었고, 남성이 연인이 있을 오즈 $(odds2)$는 약 3.826이었다. 이 때 오즈비는 4.388/3.826 = 약 1.147이 된다. 즉, 여성이 연인이 있을 오즈는 남성이 연인이 있을 오즈의 약 1.147배가 높다고 할 수 있다.

① $\theta = 1$: 두 행에서 성공의 오즈가 같다, 즉 **독립**이다.

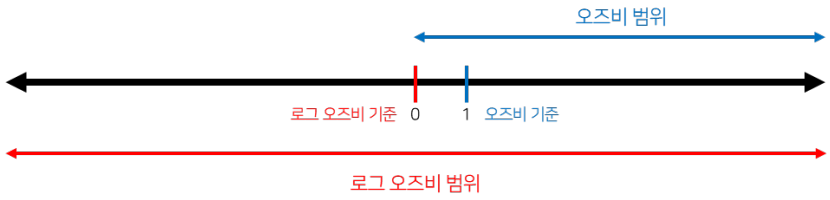
② $\theta > 1$: 첫번째 행에서의 성공의 오즈가 두번째 행보다 높다.

③ $0 < \theta < 1$: 첫번째 행에서의 성공의 오즈가 두번째 행보다 낮다.

④ 서로 역수관계에 있는 오즈비는 방향만 반대이고 연관성의 정도가 같다.

✓ 로그 오즈비 (Log Odds Ratio)

로그 오즈비는 오즈비에 Log를 씌운 형태이다. Log를 씌운 이유는 무엇일까? 그림을 통해 확인해보자.



기존 오즈비의 범위를 살펴보았을 때, 두 변수가 서로 독립인 $\theta = 1$을 기준으로 분자의 오즈가 더 큰 경우의 범위 (1~∞) 와 분모의 오즈가 더 큰 경우의 범위 (0~1) 의 크기가 서로 비대칭적인 모습을 띄고 있다.

반면 오즈비에 Log를 씌운 로그 오즈비의 경우 두 변수가 서로 독립인 0을 기준으로 분모의 값이 더 큰 경우의 범위 (−∞ ~ 0)과 분자의 값이 더 큰 경우의 범위 (0 ~ ∞)의 크기가 서로 대칭적인 모습을 띄게 된다. 즉 로그 오즈비는 기존 오즈비의 비대칭적인 범위를 교정한 측도의 역할을 한다.

3) 오즈비의 장점

① 오즈비는 상대위험도나 비율의 차이와 달리 후향적 연구와 같이 한 변수가 고정되어 있는 경우에도 사용 가능하다. 앞선 심장질환의 예시와 대조군(Y=0)의 크기를 변화시킨 분할표를 통해 확인해보자.

알코올 중독	심장질환 유무		합
	심장질환 환자	건강한 사람	
0	4 (4/6)	2 (2/6)	6
	4/2		
X	46 (46/144)	98 (98/144)	144
	46/98		
합	50	100	150

알코올 중독	심장질환 유무		합
	심장질환 환자	건강한 사람	
0	4 (4/10)	6 (6/10)	10
	4/6		
X	46 (46/340)	294(294/340)	340
	46/294		
합	50	300	350

비교군의 크기를 고정시킨 상태에서 왼쪽 분할표는 비교군과 대조군의 비율이 1:2인 경우, 오른쪽 분할표는 비교군과 대조군의 비율이 1:6인 경우이다. 이 때 비율의 차이, 상대위험도 그리고 오즈비의 크기

가 어떻게 변화하는지 아래 도표를 통해 계산해보자.

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

오즈비 값은 **대조군의 크기가 변해도 동일**한 값을 갖는 것을 확인할 수 있다.

② 오즈비는 행과 열의 위치가 바뀌어도 같은 값을 가진다.

알코올 중독	위암 유무		합
	위암 환자	건강한 사람	
0	4 (4/6)	2 (2/6)	6
	4/2		
X	46 (46/144)	98 (98/144)	144
	46/98		
합	50	100	150

위암 유무	알코올 중독		합
	0	X	
위암 환자	4 (4/50)	46 (46/50)	50
	4/46		
건강한 사람	2 (2/100)	98 (98/100)	100
	2/98		
합	6	144	150

위의 두 분할표와 같이 두 변수의 위치를 바꾸었을 때 오즈비 값이 어떻게 변하는지 계산해보자

왼쪽 : $\frac{odds1}{odds2} = \frac{4/2}{46/98} = 4.26$ 오른쪽 : $\frac{odds1}{odds2} = \frac{4/46}{2/98} = 4.26$

오즈비는 **행과 열이 서로 바뀌어도 동일**한 값이 도출된다. 어떻게 오즈비의 값이 유지될 수 있는 것일까? 이는 오즈비 값이 $P(Y|X)$를 사용하여 정의하나 $P(X|Y)$로 정의하나 서로 동일한 값을 갖기 때문이다. 이는 베이즈 정리를 통해 쉽게 증명할 수 있다.

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

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

즉, 행을 기준으로 조건부 확률을 구하여 계산한 오즈비와 열을 기준으로 조건부 확률을 구하여 계산한 오즈비의 값이 서로 같은 것이다.

✓ 교차적비 (cross-product ratio)

오즈비가 위의 두 장점 (①,②)을 가질 수 있는 이유는 오즈비가 **교차적비**이기 때문이다. 교차적비는 분할표의 대각선에 위치한 값끼리 곱한 수 간의 비율을 통해 정의되는데, 오즈비 역시 대각성분끼리의 곱과 비대각성분끼리의 곱의 비율 형태로 표현할 수 있는 것이다. 아래 수식을 통해 확인해보자.

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

교차적비의 성질에 따라 오즈비는 한 변수가 고정된 상태에서 대조군의 크기가 변하거나, 두 변수의 위치가 서로 바뀌더라도 같은 값이 유지되는 것이다. (정말 신기하고 유용하지 않나요..? ^^)

4) 3차원 분할표에서의 오즈비

지금까지 두 개의 변수 (설명변수, 반응변수)가 각각 2개의 수준을 갖고 있는 2차원 분할표에서의 오즈비에 대해 알아보았다. 다음으로 제어변수 Z가 주어진 상황에서의 오즈비를 통해 변수들 간 연관성에 대해 공부해보자.

① 부분분할표에서의 연관성

제어변수 Z가 고정되어 있을 때, X와 Y 간의 연관성을 "**조건부 연관성(Conditional Association)**" 이라고 한다. 조건부 연관성은 조건부 오즈비를 통해 파악할 수 있는데, 계산 방법은 2차원 분할표의 오즈비 계산과 매우 흡사하다. 아래 부분분할표 예시를 통해 알아보자.

부분분할표				
노트북(Z)	성별(X)	아이폰 사용 여부(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	
LG	남자	14	5	$\theta_{XY(3)}$= 4.8
	여자	7	12	

조건부 오즈비는 제어변수 Z의 각 수준별로 교차적비를 구하면 된다. (매우 간단하죠?) 조건부 오즈비를 통해 조건부 연관성을 해석해 보면, 노트북이 애플일 때 남자가 아이폰을 사용할 오즈가 여자가 아이폰을 사용할 오즈보다 약 1.188배 높다고 할 수 있다 ($\theta_{XY(1)}$= 1.188).

- 동질 연관성 (Homogeneous Association)

각 제어변수의 수준별 조건부 오즈비가 모두 같을 때 "동질 연관성이 있다"고 말한다. ($\theta_{XY(1)} = \theta_{XY(2)}$=⋯ = $\theta_{XY(K)}$) 동질연관성은 대칭적이기 때문에 X와 Y 간에 동질 연관성이 성립할 경우, Y와 Z, X와 Z 간에도 동질 연관성이 성립한다.

($\theta_{XZ(1)} = \theta_{XZ(2)}$=⋯ = $\theta_{XZ(J)}$, $\theta_{YZ(1)} = \theta_{YZ(2)}$=⋯ = $\theta_{YZ(I)}$)

- 조건부 독립성 (Conditional Independence)

각 제어변수의 수준별 조건부 오즈비가 모두 1로 같을 때 "조건부 독립성이 성립한다"고 말한다. ($\theta_{XY(1)}$= ⋯ = $\theta_{XY(K)}$= 1) 즉, 제어변수에 무관하게 X변수와 Y변수가 서로 독립인 상태이다. 조건부 독립성은 동질 연관성의 일종으로 더 엄격한 성립조건을 갖는다.

② 주변분할표에서의 연관성

앞선 예시에서는 제어변수의 제한 하의 X와 Y 간의 연관성을 확인했다면, 이번에는 Z의 각 수준의 합을 합친 주변분할표에서의 연관성에 대해 알아보도록 하자. 주변 분할표에서는 주변 오즈비를 통해 연관성을 파악할 수 있다.

주변분할표			
성별(X)	아이폰 사용 여부(Y)		주변 오즈비
	진학	비진학	
남자	11+16+14 = 41	25+4+5 = 34	$\boldsymbol{\theta_{XY+}}$**= 1.515**
여자	10+22+7 = 39	27+10+12 =49	

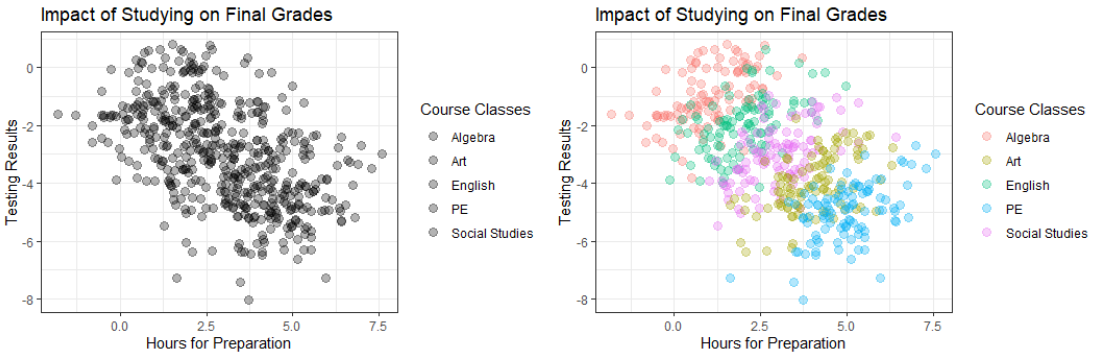
주변 오즈비는 결국 2차원 분할표에서 오즈비를 구하는 것과 같다. 하지만 주변분할표가 3차원 분할표에서 파생되었다는 것을 구분하기 위해 "주변 오즈비"라는 표현을 사용한다.

- 주변 독립성 (Marginal Independence)

주변 독립성은 주변 오즈비가 1일 때를 일컫는다. (θ_{XY+}= 1)

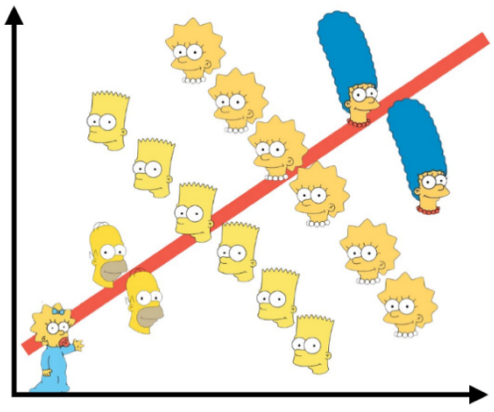
이 때 주의할 점은 분할표에서 **조건부 독립성이 성립한다고 반드시 주변 독립성도 성립하는 것은 아니라는 것**이다. 즉, 조건부 오즈비와 주변 오즈비의 방향이 항상 같지는 않다. 아래 심슨의 역설을 통해 그 이유를 알아보자.

③ 심슨의 역설 (Simpson's Paradox)



왼쪽의 그림에서 전체 플랏의 추세를 대표하는 추세선을 그려보자. 오른쪽 그림에서 색깔로 구분된 각 클래스 별 추세를 나타내는 추세선들을 그려보자. 왼쪽 그림의 추세선과 오른쪽 그림의 추세선들의 방향을 비교해보자.

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.



위 그림을 통해 심슨의 역할을 한 눈에 파악할 수 있다. 즉 우리에게 3차원 분할표나 플랏이 주어졌을 때, 조건부 독립성과 주변 독립성이 서로 다를 수 있음을 항상 유의하며 분석해야 한다. 아래 예시를 통해 마무리해보자!

부분분할표

학과 (Z)	성별 (X)	자취 여부 (Y)		조건부 오즈비
		O	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	

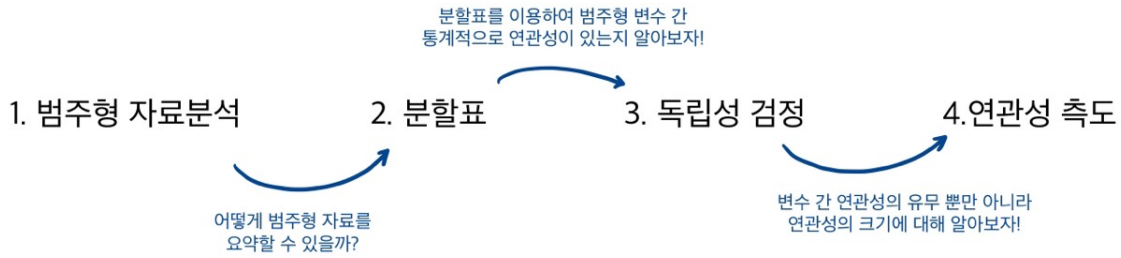
주변분할표

성별 (X)	자취 여부(Y)		주변 오즈비
	O	X	
남자	55	10	θ_{XY+}= 0.90
여자	135	22	

부분분할표에서 계산한 조건부 오즈비는 $\theta_{XY(1)} = 1.23$, $\theta_{XY(2)} = 1.2$이고 주변분할표에서 계산한 주변 오즈비는 $\theta_{XY+} = 0.90$이다. 부분분할표의 두 오즈비와 주변 오즈비는 기준값인 1을 경계로 서로 정반대의 연관성의 방향을 나타내고 있다.

심슨의 역설이 발생한 위의 예시에는 큰 특징이 존재한다. 제어변수인 학과 (Z)에 따라 전체 도수의 차이가 크게 차이가 난다. 경영학과의 전체 학생 수는 195명인 반면 통계학과 전체 학생 수는 27명 뿐이다. 즉, **학과 (Z)가 연관성을 해석하는 데 큰 영향을 미치는 변수로 작용**하였기 때문에 영향성을 반영한 조건부 오즈비와, 반영하지 않은 주변 오즈비 간에 서로 다른 결과가 도출된 것이다.

[1주차 흐름 정리]



[실습과제] (필수 아님!)

오늘 클린업에서 배웠던 내용들을 직접 코드로 구현해볼게요!

Kaggle Competition에서 유명한 Titanic 데이터를 변형하여 새롭게 데이터를 구성해보았습니다!

특히 주어진 데이터의 특징에 따라 변수 간의 독립성을 검정하는,

다양한 검정 방법을 적용해 보는 방향에 집중하였습니다!

각 문제별로 힌트들을 제시해 두었으니 참고해서 진행해 주세요!

모르는 부분이 생기면 언제든 도움 드릴테니 전혀 부담 없이 연습의 기회로 삼아보시길 바라요.

여러분 매주 세미나 준비와 패키지 등으로 바쁘신 것을 알고 있기에,

제출은 여러분들의 자유에 맡기고 제출해 주신 분들에게 피드백 드리는 방향으로 진행할게요.

실습과제에 대한 답안은 목요일 저녁 6시에 보내드릴게요~

[2주차 예고]

2주차 클린업에서는 회귀분석에서 활용하는 일반선형회귀 모형을 확장시킨 일반화 선형 모형(GLM)에 대해 배워보도록 할게요. 2주차 내용은 다소 어려울 수 있지만 여러분의 능력은 최고이기 때문에 전혀 걱정되지 않습니다~ :)

1주차 클린업 참여하시느라 고생 많으셨고, PPT와 패키지도 화이팅 해봐요~~~

임지훈 010-3499-4655

[부록]

여기서 설명되는 부분은 정말x100 Only For 참고… 클린업 내용보다 심화되는 내용이지만 궁금하실 멋진 범주러들을 위해!

I. 파이계수, 크래머의 V

앞선 클린업 교안 III. 독립성 검정에서 우리는 독립성 검정 방법을 통해 두 범주형 변수 간 연관성을 파악하는 방법에 대해 알아보았다. 하지만 **파이계수, 크래머 V**등의 통계량 계산을 통해서도 두 범주형 변수 간 연관성을 확인할 수 있다.

1. 파이계수

파이계수는 두 개의 수준을 가진 범주형 변수들 간의 연관성의 강도를 측정한다. 이 때 결과값은 0~1 사이의 값으로, 1에 가까울수록 높은 상관관계를 의미한다. 파이계수를 계산하는 방법으로 두 가지가 있다.

① 검정통계량을 이용한 계산

피어슨 카이제곱 검정의 검정통계량 X^2은 $\sum \frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}}$을 통해 구할 수 있었다. 이 검정통계량을 아래의 식에 대입하여 구한 값이 파이계수이다.

$$\varphi = \sqrt{\frac{X^2}{n}}$$

이 때, n은 전체 도수의 합이다.

② 빈도를 이용한 계산

두 이항변수로 이루어진 분할표가 다음과 같을 때, 주어진 식을 통해서도 파이계수를 구할 수 있다.

	Y=0	Y=1
X=0	A	B
X=1	C	D

$$\varphi = \frac{AD - BC}{\sqrt{(A+B)(A+C)(B+D)(C+D)}}$$

예제를 통해 두 계산을 통해 같은 파이계수 값이 도출되는지 확인해보자!

		Y		총계
		1.00	2.00	
X	1.00	A 3 (2)	B 2 (3)	5
	2.00	C 1 (2)	D 4 (3)	5
총계		4	6	10

- 검정통계량을 이용한 계산 :

$$X^2 = \frac{(3-2)^2}{2} + \frac{(2-3)^2}{3} + \frac{(1-2)^2}{2} + \frac{(4-3)^2}{3} = 1.66\,, \qquad \varphi = \sqrt{\frac{1.66}{10}} = 약\ 0.4074$$

- 빈도를 이용한 계산 :

$$\varphi = \frac{3*4-2*1}{\sqrt{(3+2)(1+4)(2+4)(3+1)}} = 약\ 0.4082$$

두 계산방법이 정확히 같은 값을 도출해내지는 않지만 거의 유사한 값을 보인다. 결국 두 방법 모두 파이 계수 계산 방법으로 활용하기에 무리가 없다.

2. 크래머의 V

크래머의 V는 두 범주형 변수가 **2개 이상의 수준**을 가진 경우에 연관성의 강도를 측정한다. 이 때 결괏값은 0~1 사이로, 1에 가까울수록 강한 연관성을 띈다. 아래 식을 통해 계산 가능하다.

$$V = \sqrt{\frac{\chi^2}{n(\min(I,J)-1)}}$$

n : 전체 도수의 합 / χ^2 : $\Sigma\frac{(O-E)^2}{E}$ / I,J : 각 변수들 수준의 개수

크래머의 V는 파이계수 계산에서 수준의 개수가 2개 이상으로 늘어났을 뿐, 큰 차이가 없다. 따라서 예시는 생략하자!

II. 피셔의 정확검정 (Fisher's Exact test)

클린업 교안에서 대표본인 경우에 대해 알아보았다면, 피셔의 정확검정 (Fisher's Exact test)는 소표본인 경우의 독립성을 검정할 때 사용한다. 카이제곱 검정법에서는 공식으로 구한 값이 근사적으로 카이제곱 분포를 따른다고 가정한 뒤 p-value를 "근사적"으로 구하지만, 피셔의 정확검정은 p-value를 정확하게 구한다. (그래서 이름이 exact test이다!)

	RH-	RH+	합계
여성	1	481	482
남성	5	513	518
합계	6	994	1000

피셔의 정확검정에서 가장 중요한 특징은 모든 주변합들이 이미 정해져 있다는 것이다. 그렇다면 우리에게 필요한 정보는 "여성이면서 RH-"인 도수(n_{11})이다. 네 개의 도수 중 단 한 칸만 정해져도 나머지 도수를 알 수 있기 때문이다. 위의 도수에서는 해당 관측 도수가 1로 기록 되어있다. 우리는 주어진 정보를 통해 초기하분포를 이용하여 여성이면서 RH- 일 확률을 구할 수 있다.

$$P(n_{11}=1)=\frac{\binom{482}{1}\binom{518}{5}}{\binom{1000}{6}}=\text{약 } 0.1074$$

- H_0 : 두 변수는 독립이다.
- H_1 : 두 변수는 연관성이 있다. (양측 검정)

✓ 이 때 단측 검정을 통한 가설도 세울 수 있으나, 어느 방향으로든 연관성이 있는지 확인하기 위한 일반적 가설인 양측 검정만 설명하겠다.

그렇다면 우리가 구하고자 하는 p-value의 의미는 다음과 같다:

"성별과 RH 혈액형 사이에는 관련성이 없다는 것(H_0)이 진실인데, 여성일수록 RH-일 확률이 높아지거나 낮아지는 관련성이 있다(H_1)고 결론을 내릴 확률"

두 변수 사이에 연관성이 있다면 해당하는 도수는 극도로 작아지거나 커질 것이다. 따라서 초기하분포 확률이 1인 경우 (0.1074) 보다 더 작은 (일어나기 힘든) 경우를 모두 더해 p-value로 삼고 검정을 진행한다.

(일반적인 독립성 검정에서는 p-value는 연속적인 범위에 대한 확률을 구하는 거라면, 피셔의 정확검정은 각 경우의 이산적인 확률을 다 더한 것이 곧 p-value가 된다.)

여성, RH-인 수	초기하 분포
0	0.0191
1	0.1074
2	0.2512
3	0.3122
4	0.2174
5	0.0804
6	0.0123

0.1074보다 작거나 같은 확률을 모두 더하면 0.2192가 나온다. 만약 유의성 기준이 0.3 이었다면 p-value 값이 더 작으므로 귀무가설을 기각하고 두 변수 간에 연관성이 있다는 결론이 도출된다.

정리하자면, 피셔의 정확검정은 다음과 같은 흐름을 따른다.

"RH-인 사람 중 여성이 1 명밖에 안되네?"
→"1 명보다 적거나 6 명에 가깝다는 것은 여성과 RH-의 연관성이 있다는 것 아닐까?"
→[귀무가설] "성별이랑 RH 혈액형 간에 아무런 관계가 없다고 가정해보자."
→"그때 RH-중 1 명이 여성일 확률보다 일어나기 힘든 경우의 확률의 합이 얼마라고? 0.2192 라고."
→"이게 현실적으로 일어나기는 어려운 일 아니야?" (유의성 기준 0.3 보다 낮은 수치)
→"그러면 여성일수록 RH-일 확률이 높아지거나 낮아질 수 있다는 것 (두 변수 간 연관성이 있다는 것)이 진실이고 그런 현실 속에서 발생한 일이라고 보는 게 낫겠다."

피셔의 정확검정은 대표본의 경우에 사용이 불가한 것이 아니다. 사실 피셔의 정확검정을 통해 독립성을 검정하는 것이 가장 정확한 방법이다. 하지만 대표본의 경우, 독립성 검정을 위해 초기하분포를 이용해 모든 p-value 값을 계산하여 검정을 진행하는 것이 소모적이고 비효율적이므로, 이를 소표본 검정에서만 진행한다고 말하는 것이다.

III. 3차원 분할표 독립성 검정 (BD test, CMH test)

3차원 분할표에서는 제어변수 Z의 영향에 따라 제어변수의 각 수준별 X 변수와 Y 변수 간의 연관성의 방향이 다르게 나타날 수 있다. 따라서 각 제어변수 Z 별 연관성 간에 큰 차이가 나지 않는지 동질성 검정을 먼저 진행해야 하는데, 이때 대표적인 동질성 검정 방법이 BD test (Breslow-Day test)이다.

1. BD test (Breslow-Day test)

BD test는 **오즈비의 동질성 검정**을 위해 고안된 카이제곱 검정법이다. 각 제어변수의 수준별 오즈비가 동질성을 갖는다는 가설 하에 검정을 진행한다.

- H_0 : 제어변수의 수준별 오즈비는 동질성을 갖는다. ($\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(J)}$)

① 귀무가설이 기각되지 않은 경우

제어변수의 수준별 오즈비가 동질성을 갖는다는 뜻으로, 각각의 수준을 구분하여 연관성을 확인하는 대신 CMH test를 통해 공통적인 연관성을 구하여 해석하는 것이 무리가 없다.

② 귀무가설이 기각된 경우

제어변수의 수준별 오즈비가 서로 다른 방향으로 나타나, CMH test를 통해 도출한 공통 오즈비가 해석의 가치를 잃는다. 이 경우에는 각 수준별로 구분하여 연관성을 확인해야 한다. (CMH test 사용 불가)
이 때 공통 오즈비를 통해 해석하려고 한다면, 심슨의 역설에 의해 왜곡된 결론이 도출될 위험이 있다.

- BD test의 검정을 위한 검정통계량과 검정 과정은 매우 복잡하므로, 수식만 첨부하는 정도로 마무리하겠다… (그래도 궁금하신 분들은 웰컴!)

$$\chi^2_{BD} = \sum_{k=1}^{r} \frac{[\, n_{k11} - E(n_{k11}; \hat{\theta}_{MH}) \,]^2}{Var(n_{k11}; \hat{\theta}_{MH})}$$

이 때 검정통계량은 자유도가 k-1 (k는 제어변수 Z의 수준의 개수)인 카이제곱 분포 $\mathrm{x}^2_{\mathrm{k}-1}$를 따른다.

2. CMH test (Cochran-Mantel-Haenszel test)

BD test에 따라 제어변수 수준 별 연관성들 간의 동질성이 지켜졌다는 가정 하에 사용할 수 있는 독립성 검정 방법이 CMH test다. CMH test는 제어변수의 모든 수준에서 조건부 독립이 성립하는지를 검정한다. 만약 조건부 독립이 성립하지 않는다면, 즉 제어변수 Z의 수준에 따라 X와 Y가 유의미한 연관성을 갖는다면, 제어변수의 모든 수준을 총합한 X와 Y 간의 공통 오즈비를 계산해볼 수 있다.

① CMH 검정을 통해 해석

앞서 설명했듯이 BD test를 통해 동질성이 성립하는 경우에만 CMH 검정을 진행한다. CMH 검정을 통해 조건부 독립이 성립한다는 결론이 도출된 경우에는 두 변수 X와 Y 간의 분석이 무의미해진다. 반대로 조건부 독립이 성립하지 않는다면, 공통 오즈비를 계산하여 이를 통해 X와 Y간의 관계를 해석하는 것이 타당성을 갖는다. 우선 귀무가설은 다음과 같다.

- H_0 : X변수와 Y변수는 제어변수의 모든 수준에서 조건부 독립성을 갖는다 (오즈비가 1이다).

우리는 아래의 검정통계량 계산식을 통해 검정통계량 값의 p-value 값을 통해 검정을 진행한다.

✓ 검정통계량

$$X^2_{CMH} = \frac{[\sum_{k=1}^{r} a_k - \sum_{k=1}^{r} E(a_k)\,]^2}{\frac{\sum_{k=1}^{r}(a_k+b_k)(a_k+c_k)(b_k+c_k)(c_k+d_k)}{(n^3-n^2)}}$$

✓ 기각역 : $X^2 \geq x^2_{\alpha,1}$

위의 검정통계량 계산식은 복잡해 보이지만 결국 **관측도수 n_{11k}와 기대도수 간 차이의 제곱을 분산으로 나눠준 값**이다. 이 때, 위 계산식에서 a_k는 2 X 2의 분할표 중 어떤 도수로 지정해도 무관하다. 즉, 변수 별 위치가 고정되어 있는 것이 아니다.

즉, 제어변수 Z가 총 r개의 수준을 가지고 있다고 가정한 후, 각 제어변수별로 관측도수와 기대도수의 차이의 제곱 (분자)를 각 제어변수별 분산식(분모)으로 나누어 합한 값이 CMH 검정통계량이 된다. 이

값을 바탕으로 검정을 진행한다. p-value 값이 작아 귀무가설이 기각되는 경우에는 변수 간의 연관성이 있다고 해석할 수 있게 된다.

② 공통 오즈비 계산

앞선 CMH 검정에서 변수 간 연관성이 있다고 해석된 경우 공통 오즈비를 계산함으로써 제어변수 Z에 따른 변수 X와 Y 간의 연관성을 요약해 볼 수 있다. 아래의 예시가 CMH 검정을 통해 연관성이 있음을 확인하였다고 가정하자. 이 때 우리는 제어변수의 수준별 비만(X)과 당뇨(Y) 간의 연관성을 종합한 하나의 공통 오즈비를 통해 X와 Y 간의 연관성을 해석하고자 한다.

부분분할표				
연령대(Z)	비만(X)	당뇨(Y)		조건부 오즈비
		O	X	
50대 이하	O	10	90	$\boldsymbol{\theta_{XY(1)}}$= 1.476
	X	35	465	
50대 이상	O	36	164	$\boldsymbol{\theta_{XY(2)}}$= 1.53
	X	25	175	

주변분할표			
비만(X)	당뇨(Y)		주변 오즈비
	O	X	
O	10	90	$\boldsymbol{\theta_{XY+}}$= 1.93
X	35	465	

위 예시에서 주변 오즈비의 값은 1.93 이지만 조건부 오즈비는 1.476 &1.53으로, 주변 오즈비가 과장되어있음을 알 수 있다. 따라서 단순하게 주변 오즈비를 조건부 오즈비들을 종합한 공통 오즈비로 여기는 것은 위험하다. 이를 감안하여 두 오즈비를 합하기 위한 방법이 CMH 추정량이다. 이를 위해 아래의 식이 필요하다.

$$\widehat{OR_{CMH}} = \frac{\sum \frac{a_i d_i}{n_i}}{\sum \frac{b_i c_i}{n_i}}$$

(a_i , b_i , c_i , d_i 는 2 X 2 분할표의 도수)

위의 식에 대입한 결과 공통 오즈비는 1.52임을 알 수 있다. 따라서, 연령대의 영향을 반영하여 해석해보면 당뇨 환자가 비만 환자일 오즈가 약 1.52배 높다고 추론 가능하다.

주의할 점은 CMH 검정은 주로 3차원 분할표가 2*2*K의 형태, 즉 X와 Y의 수준이 각각 2개인 분할표가 K개의 제어변수를 가지고 있을 때 사용한다는 점이다.

정리해보자면, BD test는 제어변수의 수준별 동질 연관성을($\theta_{XY(1)}$= … = $\theta_{XY(K)}$= n), CMH test는 조건부 독립성을($\theta_{XY(1)}$= … = $\theta_{XY(K)}$= 1) 확인하기 위한 검정법이다. BD test부터 CMH test로 이르는 flow를 정리해보면 아래와 같다.

BD test의 귀무가설(X,Y는 동질 연관성 갖는다) 수용 → CMH test → CMH test 귀무가설(X,Y는 조건부 독립) 기각 → 공통 오즈비 계산

✓ 참고 :

https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Biological_Statistics_(McDonald)/02%3A_Tests_for_Nominal_Variables/2.10%3A_Cochran-Mantel-Haenszel_Test

https://www.slideserve.com/arleen/introduction-to-mantel-haenszel-estimate