사회통계

교차표 복습과 카이제곱분석

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 교차표 톺아보기
2. χ^2 독립성 검정

교차표 톺아보기

교차표 톺아보기

최초의 사회학적 분석은 교차표로 이루어졌다.

- Emile Durkheim의 〈자살론(Suicide)〉은 교차표를 잘 활용한 위대한 고전사회학적 연구이다.
- 그는 여러 사회에 걸쳐 수많은 인구학적 특성별로 자살률을 비교하는 표를 제시하였다.
- 여러 사회의 집단적 성격에 따라 자살률이 같거나 상이함을 비교하여, 자살에 영향을 미치고 또 미치지 않는 사회적 사실(social facts)을 드러내는 독특한 비교방법론인 공변법(method of concomitant variations)을 활용하였다.
- 다만 주의할 점은 (비슷하게 생기긴 했지만) Durkheim이 사용한 수많은 표들이 일반적인 교차표와 좀 다르다는 점이다. 이에 대해서는 나중에 다시 이야기한다.

교차표 톺아보기

교차표는 사회학적 상상력의 원천이다.

- (사회통계학을 극단적으로 경멸하는) C. W. Mills도 교차표의 가치만큼은 인정하였다(특히 〈사회학적 상상력〉 장인기질론 부록을 볼 것).
- "이제까지 고립되어 있던 항목들을 서로 관련시켜 예기치 않았던 관계를 발견해냄으로써 상상력이 성공적으로 구현된다(Mills 1959[2004]: 246)."
- "상관분류 기법은 물론 양적인 자료에만 국한되는 것이 아니다. 이 기법은 실제로 옛 유형을 비판하고 명료히 하는 것뿐만 아니라 '새로운' 유형을 상상하고 파악하는 가장 좋은 방법이다(Mills 1959[2004]: 416)."
- Mills (1959[2004]: 246)도 함께 볼 것.

Mills, C. Wright. 1959. The Sociological Imagination, New York, NY: Oxford University Press. [밀즈, C. 라이트. 2004. 『사회학적 상상력』. 돌베개.]

교차표 톺아보기

교차표는 둘 이상의 변수 간 관계를 분석하는데 있어 탄력적인 도구이다.

- 원칙적으로 교차표는 두 개의 범주형 변수(categorical variables) 사이의 관계를 분석하는데 사용한다.
- 하지만 일정한 정보의 손실(information loss)을 감수한다면 숫자형 변수를 범주형 변수로 얼마든지 변환할 수 있다(e.g., 숫자형 연령에서 범주형 연령으로).
- 그러므로 교차표는 (1) 두 개의 질적변수, (2) 하나의 질적변수와 하나의 양적변수, (3) 두 개의 양적변수 등 자료유형과 무관하게 다 사용할 수 있다.

교차표 톺아보기

교차표 만들기와 해석에는 몇 가지 중요한 규칙이 있다!

- 우리는 관습에 따라 독립변수 X에 해당하는 변수를 행(row)에 놓고, 종속변수 Y에 해당하는 변수를 열(column)에 놓는다.
- 교차표는 각 셀(cell)에 빈도(frequency)를 보고하는 것에서 종종 출발한다. 하지만 사실 비율(percentage)을 보고하는 쪽이 "해석을 위해서" 훨씬 편리하다.
- 비율을 구할 때는 적어도 세 가지의 표준화(standardization) 방식이 있으며, 표준화를 어떻게 했는가에 따라 해석 또한 달라진다.
- 올바른 해석을 위해서는 조건부확률(conditional probability)과 결합확률(joint probability) 개념에 대한 이해를 요구한다.

교차표 톺아보기

이제 확률 개념을 다시 되짚어보자.

- 조건부확률이란 "다른 사건 B가 이미 일어났다는 전제 아래, 한 사건 A가 일어날 확률"을 의미한다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- 결합확률이란 "두 사건 A와 B가 동시에 일어날 확률"을 의미한다.

$$P(A \cap B) = P(A|B) \cdot P(B)$$

- 이때, $P(A \cap B) = P(B \cap A)$ 이지만, $P(A|B) \neq P(B|A)$ 이다.

교차표 톺아보기

확률에는 몇 가지 기본법칙들이 성립한다.

- 덧셈법칙(addition rule): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- 여기서 $P(A \cap B)$는 아까 설명한 결합확률이다.
- 여사건법칙(complement rule): $P(A^C) = 1 - P(A)$
- 곱셈법칙(multiplication rule): $P(A \cap B) = P(A|B) \cdot P(B)$
- 여기서 $P(A|B)$ 는 아까 설명한 조건부확률이다.

교차표 톺아보기

확률의 기본법칙을 이해하면 독립 사건과 종속 사건을 이해할 수 있다.

- 두 개의 사건 A와 B가 있을 때, $P(A|B) = P(A)$ 이거나 $P(B|A) = P(B)$ 이면 두 사건은 독립(independent)이다(Why?).
- (두 사건은 독립이 아니면 종속이므로) 위 식이 성립하지 않으면 종속(dependent)이다.
- A와 B가 독립적인 사건들이라면 위 법칙들은 더욱 단순해진다.

$$P(A \cup B) = P(A) + P(B)$$
$$P(A \cap B) = P(A) \cdot P(B)$$

교차표 톺아보기

예제 1. 당신은 성별과 종교인 여부에는 모종의 관계가 있을 것이라고 추측한다. 당신이 521명의 임의표본을 추출한 자료는 religious.csv에서 확인할 수 있다. 여성의 몇 퍼센트가 종교인인가? 종교인의 몇 퍼센트가 여성인가? 종교인이자 여성인 사람은 모두 몇 퍼센트인가?

교차표 톺아보기

- 일단 religious.csv를 엑셀로 불러오자.
- 여기서는 단지 “성별과 종교인 여부에는 어떠한 관계가 있을까”를 묻고 있을 뿐, 구체적인 가설을 제시하지 않았다.
- 자료가 일단 주어졌다면 그것들의 자료유형을 판독하는 것이 선행되어야 한다. 성별(sex)과 종교인 여부(religious)는 각각 어떠한 척도로 측정되어 있나?
- 두 변수 사이의 관계를 보기 위해 일단 교차표를 만들어보자. [삽입]-[피벗 테이블] 그리고 “Σ 값” 부분에는 반드시 “값 필드 설정”에서 “개수”가 선택되도록 유의한다.
- 표를 복사하여 아랫쪽에 “값으로” 붙여넣는다. 레이블은 적절히 복사하여 붙여넣을 수 있다. 불필요한 행과 열은 제거하자.

교차표 톺아보기

원점수 그대로인 상태에서는 해석하기 어렵다.

- 차라리 원점수(raw score)보다 상대적인 비율(proportions)을 보면 편리하다.
- 여기서 비율은 각 카테고리 별로 특정 셀에 보고된 응답의 비율을 의미한다.
- 예컨대 “나는 여성이다”라고 응답한 사람 중에 몇 퍼센트나 “나는 종교적이다” 라고 응답했는지 살펴보아야 한다.
- 그런데 상대비율을 구할 때는 세 가지 방법을 상상해 볼 수 있다!
  (1) 행 합계(row total)로 표준화하는 방법
  (2) 열 합계(column total)로 표준화하는 방법
  (3) 총 합계(grand total)로 표준화하는 방법

교차표 톺아보기

세 가지 표준화하는 방법을 혼동해서는 안된다.

- 행 합계(row total)로 표준화할 때는 개별 셀(cell)을 해당 행 합계로 나눠준다. 이때, 행 합계는 각 행(row)의 합계를 나타내기 위해 추가적인 열 안에 넣어놓은 숫자다.
- 열 합계(column total)로 표준화할 때는 개별 셀을 해당 열 합계로 나눠준다. 이때, 열 합계는 각 열(column)의 합계를 나타내기 위해 추가적인 행 안에 넣어놓은 숫자다.
- 총 합계(grand total)로 표준화할 때는 개별 셀을 총 합계로 나눠준다. 이때, 총 합계는 모든 셀(cell)의 합계를 나타내기 위해 추가적으로 우측 하단 안에 넣어놓은 숫자다.

교차표 톺아보기

표준화 방법에 따라 계산된 비율은 해석방법이 달라진다.

- 물론 셋 다 연습해야 한다. 분석 목적에 따라 다른 표준화 방식이 적용되어야 하기 때문이다.
- 그런데 우리는 관습에 따라 종종 독립변수 X에 해당하는 부분을 행(row)에 놓고, 종속변수 Y에 해당하는 부분을 열(column)에 놓는 경향이 있다.
- 이 경우 행 합계로 표준화하는 편이 해석에 편리하다(Why?).
- 이점을 고려하여 엑셀에서 피벗 테이블을 만들 때부터 독립변수와 종속변수를 생각하고 만들어야 한다.

교차표 톺아보기

원자료에서 왼쪽 표를 만든 뒤, 행 합계로 표준화하여 오른쪽 표를 만들자.

- 표는 다음과 같다.

	비종교인	종교인	합계
남자	232	68	300
여자	124	97	221
합계	356	165	521

	비종교인	종교인	합계
남자	0.77	0.23	1
여자	0.56	0.44	1
합계	0.68	0.32	1

- 이 해석은 물론 조건부확률 개념을 사용하고 있다.

$$P(\text{비종교인}|\text{남자}) = 0.77 \qquad P(\text{종교인}|\text{남자}) = 0.23$$
$$P(\text{비종교인}|\text{여자}) = 0.56 \qquad P(\text{종교인}|\text{여자}) = 0.44$$

교차표 톺아보기

이번엔 열 합계로 표준화하여 오른쪽 표를 만들자.

- 표는 다음과 같다.

	비종교인	종교인	합계
남자	232	68	300
여자	124	97	221
합계	356	165	521

	비종교인	종교인	합계
남자	0.65	0.41	0.58
여자	0.35	0.59	0.42
합계	1	1	1

- 만약 열 합계로 표준화할 경우, P(여자|종교인), P(남자|종교인), P(여자|비종교인), P(남자|비종교인)을 계산할 수 있다. 이 해석들도 조건부확률 개념을 사용하고 있다.

교차표 톺아보기

마지막으로 총 합계로 표준화하여 오른쪽 표를 만들자.

- 표는 다음과 같다.

	비종교인	종교인	합계
남자	232	68	300
여자	124	97	221
합계	356	165	521

	비종교인	종교인	합계
남자	0.45	0.13	0.58
여자	0.24	0.19	0.42
합계	0.68	0.32	1

- 이 해석은 결합확률을 사용하고 있다.

$$P(\text{비종교인} \cap \text{남자}) = 0.45 \qquad P(\text{종교인} \cap \text{남자}) = 0.13$$
$$P(\text{비종교인} \cap \text{여자}) = 0.24 \qquad P(\text{종교인} \cap \text{여자}) = 0.19$$

교차표 톺아보기

- 지금까지 배운 교차표, 결합확률, 조건부확률 개념을 동시에 생각하여 주변확률(marginal probability)을 파악할 수 있다.

$$\begin{aligned} P(\text{종교인}) &= P(\text{종교인} \cap \text{여자}) + P(\text{종교인} \cap \text{남자}) \\ &= P(\text{종교인}|\text{여자}) \cdot P(\text{여자}) + P(\text{종교인}|\text{남자}) \cdot P(\text{남자}) \end{aligned}$$

- 마찬가지로 $P(\text{비종교인})$, $P(\text{여자})$, $P(\text{남자})$도 스스로 구해보자(3주차 강의안 참고).

χ^2 독립성 검정

χ^2 독립성 검정

χ^2 분석은 이른바 범주형 변수를 분석하는 대표적 기법 중 하나이다.

- 우리는 이미 χ^2 분포에 관해 모분산에 대한 추정과 가설검정 파트에서 배웠다.
- 이제 Karl Pearson이 개발한 χ^2 독립성 검정(chi-square test of independence) 또는 χ^2 분석(chi-square analysis)이라고 불리우는 기법을 하나 더 배운다.
- Pearson은 불세출의 수학 천재, 사회진화론자(social Darwinist), 우생학자, 무신론자(freethinker), 사회주의자였다. 그가 카이제곱검정 뿐 아니라 상관분석과 회귀분석 등에도 크게 기여했다.

χ^2 독립성 검정

- χ^2 분포를 따르는 확률밀도함수를 정의하기 위해서는 독립성 가정(independence assumption)이 필요했음을 떠올리자.
- Karl Pearson의 χ^2 분석은 이 가정을 절묘하게 이용해 (교차표에 주어진) 두 변수가 서로 독립적(independent)인가 가설검정을 수행한다.
- 즉, 독립성을 가정하고 이론적으로 계산한 기대빈도(expected frequency) E와 실제 교차표의 관찰빈도(observed frequency) O를 비교하여 너무 큰 차이가 나는지 살펴보는 것이다.

χ^2 독립성 검정

기대빈도 E는 χ^2 분포의 독립성 가정에 따라 계산된다.

- 표준화된 자료가 아니라 원자료로 χ^2 독립성 검정을 시작한다.

	비종교인	종교인	합계
남자	232	68	300
여자	124	97	221
합계	356	165	521

- 성별과 종교 두 변수가 독립적이라면(=다른 모집단에서 나온 표본이라면) 곱셈법칙이 성립한다.

$$P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$$

- e.g., $P(\text{여자} \cap \text{종교인}) = P(\text{여자}) \cdot P(\text{종교인}) = (221/521) \cdot (165/521) \approx 0.13434$
- $n = 521$ 이므로 "여자 $\cap$ 종교인" 사건의 기대빈도 E는 $521 \cdot 0.13434 = 69.99$이다.

χ^2 독립성 검정

엑셀을 사용해 관찰빈도 O와 기대빈도 E를 직접 계산해보자.

- 수식을 계산할 때는 늘 괄호 사용에 주의할 것.
- 관찰빈도 O 행렬의 주변확률 정보를 이용해 옆의 기대빈도 E 행렬을 채워넣자.

	비종교인	종교인	합계
남자	232	68	300
여자	124	97	221
합계	356	165	521

	비종교인	종교인
남자		
여자		69.99

- 정답은 뒷 페이지에 있으니 먼저 엑셀로 표를 만들어 풀어보고 확인하자.

χ^2 독립성 검정

- 이제 채워진 관찰빈도 O와 기대빈도 E를 곰곰히 비교해보자!

	비종교인	종교인	합계
남자	232	68	300
여자	124	97	221
합계	356	165	521

	비종교인	종교인
남자	204.99	95.01
여자	151.01	69.99

- 기대빈도 E는 두 변수가 독립적이라는 가정에 입각했을 때 이론적으로 기대된 빈도였다.
- 다시 말해, 기대빈도 E가 관찰빈도 O와 크게 다르지 않다면 독립성 가정(=두 변수는 서로 독립적이다)는 결론에 도달한다.
- 반면, 기대빈도 E와 관찰빈도 O가 크게 다르다면 (애초에 세운) 독립성 가정이 틀렸다는 결론에 도달한다. 즉 두 변수는 연관되어 있었던 것이다!

χ^2 독립성 검정

지금까지 논의를 통해 χ^2 분석의 가설구조를 유추할 수 있다.

- χ^2 독립성 검정의 가설은 다음과 같다.

$$H_0 : \text{성별과 종교인 여부는 서로 독립적이다.}$$
$$H_a : \text{성별과 종교인 여부는 서로 독립적이지 않다.}$$

- 만일 귀무가설을 기각할 수 있었다고 할지라도 여성이 더 종교적인지 여부 등은 독립성 검정을 통해서는 알 수 없다.
- 단지 "성별과 종교성이 서로 독립적"이라는 귀무가설을 통계적으로 유의하게 기각할 수 있을 뿐이다.
- 대립가설 이상의 해석을 χ^2 독립성 검정에 멋대로 덧붙이지 않도록 꼭 주의해야 한다!

χ^2 독립성 검정

이제 (가설검정을 위한) 검정통계량을 계산한다.

- 검정통계량인 χ^2 값은 다음과 같이 계산된다.

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}$$

- 여기서 J와 K는 각각 교차표의 행과 열의 수, O_{jk}와 E_{jk}는 각각 j번째 행, k번째 열의 관찰빈도 O와 기대빈도 E를 의미한다.
- 엑셀에서 직접 검정통계량 χ^2 값을 계산해 보자.

χ^2 독립성 검정

관찰빈도 O와 기대빈도 E가 다름을 확인하는 유의성 검정을 수행한다.

- (유의성 검정에서 사용하는) 이론적 확률분포는 확률변수 χ^2의 분포이다.

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(O_{jk}-E_{jk})^2}{E_{jk}} \sim \chi^2_{(J-1)(K-1)}$$

- 만일 귀무가설대로 $O=E$라면 χ^2는 0이다. 그러나 표본에서 계산된 χ^2 값이 크다면 (그려진) χ^2 분포 위의 아주 구석진 곳에 놓일 것이다.
- 앞의 예에서는 2×2 교차표였으므로 $J=K=2$다. 자유도는 $(J-1)(K-1)=1$이다. 1의 자유도를 가진 χ^2 분포를 그린다.
- 유의수준에 따른 판정은 (이미 배운대로) 엑셀 함수 1-CHISQ.DIST$(\cdot)$를 활용한다.

χ^2 독립성 검정

자유도 계산에서 실수하지 않도록 주의하자.

- 사실 Pearson은 χ^2 독립성 검정 원리를 발표할 때 자유도를 잘못 계산했었다.
- (한참 어린) Ronald Fisher가 올바른 자유도의 계산원리를 제시하자 그는 길길이 날뛰면서 Fisher를 왕립통계학회로부터 제명하는 만행을 저질렀다.
- (조금 있다가 다시 이야기하겠지만) χ^2 독립성 검정은 원자료를 필요로 하지 않고 단지 교차표만 주어지면 곧바로 계산할 수 있다.
- 자유도가 이런 형태로 주어지는 것은 그 때문이다.

χ^2 독립성 검정

본래 χ^2 분석에는 독립성 검정 이외에 다른 유형도 있다.

- 기본적으로 독립성 검정(independence test)이 압도적으로 많이 쓰인다.
- 그러나 동질성 검정(homogeneity test)과 적합성 검정(goodness-of-fit test)도 분석 목적에 따라 나름 유용하다.
- 독립성 검정과 동질성 검정은 두 변수 사이의 독립성에 주목하는 반면, 적합성 검정에서는 단일한 변수에 대한 이론적 분포와 (실제로) 관찰된 분포 사이의 일치성 여부를 살펴본다.
- 적합성 검정은 고급통계학이나 수리사회학(mathematical sociology) 분야에서는 매우 중요한 분석도구가 되지만, 기초사회통계에서는 다루지 않는다.