사회통계

카이제곱분석의 연습과 방법론적 이슈들

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. χ^2 분석의 연습
2. 교차표와 관련된 방법론적 이슈

χ^2 분석의 연습

χ^2 분석의 연습

χ^2 분석은 상당히 중요하며 제법 연습이 필요하다.

- 기본적으로 두 개의 질적변수(즉 범주형 변수)가 주어졌다면 교차표를 만들고 이를 표준화한 뒤 해석을 시도해야 한다.
- 교차표를 만들었다면 곧장 χ^2 분석을 고려해볼 만 하다.
- 셀(cell) 안의 숫자를 모두 다 해석할 필요는 없고, 두 변수 간 연관성이 있는지 없는지 밝혀야 한다.
- χ^2 분석 결과의 과잉해석에는 주의해야 한다.

χ^2 분석의 연습

예제 2. 사회계층에서 결혼은 계층적 단절을 지속시키는 중요한 매커니즘 가운데 하나로 알려져 있다. 비슷한 교육수준을 갖춘 사람들 사이에서 결혼관계, 즉 동질혼(assortative marriage)가 한국에서도 나타나지 않는지 가설을 검증하고자 한다. 〈한국일반사회조사〉 자료의 일부인 homogamy.csv를 엑셀로 불러와 이 아이디어를 테스트하시오.

χ^2 분석의 연습

- 이 연구질문(research question)은 다음과 같다: "응답자의 최종학력과 배우자의 최종학력에는 어떠한 관계가 있을까?"
- 두 변수 모두 질적변수이므로 이 연구질문에 대해 적절한 기법 중 하나는 χ^2 독립성 검정이다. 연구질문을 보다 구체화시켜 다음의 귀무가설과 대립가설을 세울 수 있다.

  H_0: 응답자의 최종학력과 배우자의 최종학력은 서로 독립적이다.
  H_a: 응답자의 최종학력과 배우자의 최종학력은 서로 연관되어 있다.
- 두 변수 사이의 관계를 보기 위해 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.

χ^2 분석의 연습

- χ^2 독립성 검정을 위해 기대빈도 E 표를 원래의 관찰빈도 O 표 밑에 만든다.
- 검정통계량 χ^2 값은 다음과 같이 계산한다.

$$\chi^2_{(j-1)(k-1)} = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} = 14082.99$$

- 엑셀에서 1-CHISQ.DIST(14082.99, 64, TRUE)로 유의확률을 계산한다.
- 0.01 유의수준에서 귀무가설을 기각할 수 있으므로 "통계적으로 유의하게" 두 변수는 서로 연관되어 있다고 결론내릴 수 있다. 그러나 두 변수 사이에 구체적으로 어떤 관계가 있는지는 χ^2 독립성 검정을 통해 말할 수 없다.
- 이제 χ^2 독립성 검정과는 별개로 이 표를 행 합계로 표준화(row standardization)한 뒤, 교차표를 해석해보자. 해석할 때는 각 행(row)에서 가장 큰 값과 두번째로 큰 값을 기준으로 판단하면 편리하다.

χ^2 분석의 연습

예제 3. 당신은 나흘간 단 한숨도 자지 않고 사회통계학 숙제를 완성하였다. 최종 제출을 위해 학교로 후다닥 달려가던 도중, 걸어나오던 친구와 부딪쳐 그만 커피를 그 위에 왕창 쏟고 말았다. 참고로 당신이 작성한 이 표는 지난 100년 간 브라질에서 나비의 개체수와 텍사스에서 관찰된 비바람의 횟수 사이 관계를 묘사한 것이다. 이 표를 복원하시오.

	약한 비바람	큰 비바람	Total
적게 펄럭임	20	30	50
많이 펄럭임	30	40	70
Total	50	70	120

χ^2 분석의 연습

- χ^2 독립성 검정의 원리를 일단 이해했다면 좀 더 빠르게 계산을 수행할 수 있다.

$$E_{11} = \frac{50 \cdot 50}{120} = 20.8$$
$$E_{12} = \frac{70 \cdot 50}{120} = 29.2$$
$$E_{21} = \frac{50 \cdot 70}{120} = 29.2$$
$$E_{22} = \frac{70 \cdot 70}{120} = 40.8$$

- 어떻게 이렇게 완벽에 가깝게 복원할 수 있었을까? 그건 두 변수가 서로 아무 상관도 없다는 것이 자명하므로 $E = O$이기 때문이다.
- 다시 한 번 기억해야 할 것은 "χ^2 독립성 검정이 기본적으로 두 변수가 독립적이라는 가정 아래에서 기대빈도 E를 계산한다"는 점이다.

χ^2 분석의 연습

예제 4. 자타가 공인하는 "매의 눈"인 당신은 자신이 알바하는 가게를 찾아온 (서로 마주 앉은) 커플들이 주문한 안주 메뉴에 관해 최근 30일 동안 임의표본을 수집하였다. 이 교차표에 근거하였을 때, 두 커플이 주문하는 음식이 서로 독립적인지 여부를 검정하라.

	군고구마	찐감자	Total
군고구마	20	30	50
찐감자	30	20	50
Total	50	50	100

χ^2 분석의 연습

- (이제 슬슬 눈치를 챘겠지만) χ^2 독립성 검정은 사실 교차표만 주어져도 그냥 수행할 수 있다.
- 애시당초 χ^2 값을 계산하기 위해 원자료(raw data) 전체가 꼭 필요한 것은 아니다. 교차표만 있어도 관찰빈도 O, 기대빈도 E, 자유도 J와 K를 알기엔 충분하다.
- 특별히 원자료가 주어지지 않았는데도 χ^2 통계량을 쉽게 계산할 수 있다는 것이 큰 장점이다. 이 장점을 활용하기 위해 많은 연구자들은 교차표를 제시할 때 (별 이유가 없어도) 거의 기계적으로 χ^2 독립성 검정을 수행하고 그 결과를 제시한다.
- 표 언저리에 간단히 $\alpha\%$ 유의수준에서 두 변수 사이의 연관성 유무를 보고하여 교차표를 좀 더 유용하게 만들 수 있기 때문이다.

	군고구마	찐감자	Total
군고구마	20	30	50
찐감자	30	20	50
Total	50	50	100

(참고: $\chi^2 = 4$; $p < 0.05$.)

χ^2 분석의 연습

- 사업장 규모에 따라 각종 안전보건관리 조직 및 구성이 상이할까?

<표 2> 사업장 규모에 따른 안전보건조직 및 구성의 차이

사업장 안전보건관리 조직 및 구성	실태	사업장 규모(상시근로자 수)				전체	$\chi 2$
		-49인	50-99인	100-299인	300인-		
안전보건관리 조직 구성	미구성	40 (11.7)	227 (20.3)	166 (14.3)	29 (8.3)	462 (15.6)	38.156***
	구성	303 (88.3)	893 (79.7)	992 (85.7)	320 (91.7)	2508 (84.4)	
사업장내 안전관리자 선임형태	비전담 및 대행	107 (31.2)	833 (74.4)	689 (59.5)	122 (35.0)	1751 (59.0)	302.489***
	전담	236 (68.8)	287 (25.6)	469 (40.5)	227 (65.0)	1219 (41.0)	
안전보건 관리시스템 또는 대처방안 마련	그렇지 않다	56 (16.3)	159 (14.2)	94 (8.1)	36 (10.3)	345 (11.6)	29.085***
	그렇다	287 (83.7)	961 (85.8)	1064 (91.9)	313 (89.7)	2625 (88.4)	

*** $p<0.001$

최서연. 2019. "사업장의 안전보건활동이 안전문화 정착에 미치는 영향: 사업장 규모를 중심으로." 인문사회 21 10(4): 1105-1118.

교차표와 관련된 방법론적 이슈

교차표와 관련된 방법론적 이슈

질적변수가 아니라 양적변수의 경우에도 교차표를 그릴 수 있을까?

- 물론이다. 정보의 손실(information loss)을 감수한다면 양적변수를 질적변수로 얼마든지 재부호화(recoding)할 수 있다. 그러나 그 역은 불가능하다.
- 엑셀에서는 "그룹" 기능으로 쉽게 범주화할 수 있다.
- 다만 범주화의 범위를 고려하여 많은 셀에서 0이 나오지 않도록 주의를 기울여야 한다!

교차표와 관련된 방법론적 이슈

표본 크기가 작은데 셀에 아주 작은 숫자만 들어간다면 문제가 된다.

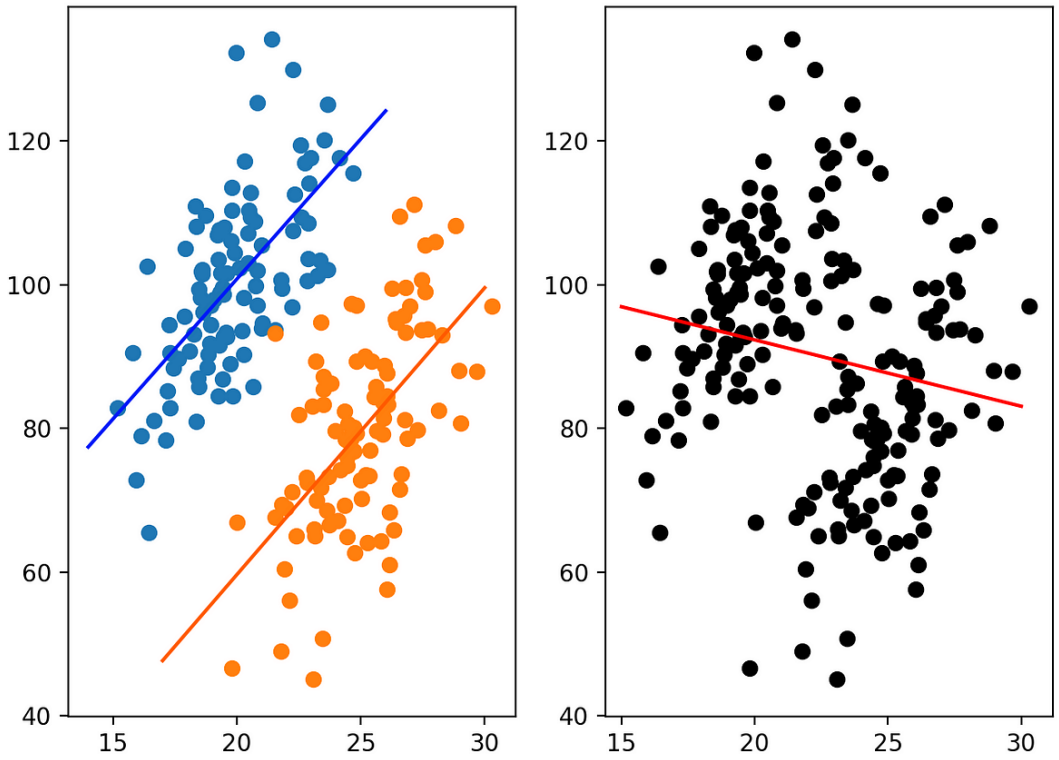
- 먼저 특정 셀에 아주 작은 숫자(특히 5보다 작은 수)가 들어간다는 것의 의미를 고민해야 한다.
- 이 문제는 아주 작은 표본($n < 40$)에 기반하여 2×2 교차표를 만들 때 제기되는 문제이다. 사실 오늘날 사회과학에서는 좀처럼 이런 상황이 없다.
- (2×2보다) 큰 교차표에서는 최소한 80% 이상의 셀이 5보다 큰 값을 가져야 한다.
- 이런 경우 다음의 세 가지 대응방법이 있다.

  (1) 범주를 다시 만든다. 2×2 교차표에서는 이 방법이 유효하지 않다.
  (2) (χ^2 독립성 검정에서) Yates의 연속성 보정(continuity correction)을 한다.
  (3) (χ^2 독립성 검정 대신) Fisher의 정확 검정(exact test)을 사용한다.

교차표와 관련된 방법론적 이슈

종종 두 개의 변수를 넘어 제3의 변수를 고려해야 한다.

- 주어진 자료 '전체'를 사용하다보면 내부의 심각한 이질성(heterogeneity)을 무시하고 연관성을 파악하기 쉽다.
- 하지만 이론적으로 더 깊은 고민을 통해 독립변수와 종속변수에 "동시에 영향을 끼치는" 이질성 요인을 반드시 찾아내야 할 때가 있다.



교차표와 관련된 방법론적 이슈

연관성을 살펴볼 때는 조건별로 주의깊게 뜯어보아야 한다.

- 화재 피해와 출동한 소방관의 수에 관한 패러독스에 관해 살펴보자.

	피해 적음	피해 큼	합계
소규모 출동	97 (69.8%)	49 (30.2%)	146 (100%)
대규모 출동	42 (32.2%)	103 (67.8%)	145 (100%)
합계	139 (50.2%)	152 (49.8%)	291 (100%)

- 이것은 패러독스도 뭣도 아니다. 화재의 규모라는 제3의 변수를 고려하지 않았기 때문에 생기는 허구적 인과성(spurious causality)에 불과하다.
- 독립변수와 종속변수 사이에 인과관계가 있다고 믿고 χ^2 검정을 수행하지만 (단순한) 교차표는 이를 증명해주지 않는다. 다만 독립성 여부를 보여줄 뿐이다.
- 심슨의 역설(Simpson's Paradox)을 언제나 주의해야 한다.

교차표와 관련된 방법론적 이슈

예제 5. 앞서 우리나라에서도 동질혼이 나타남을 확인하였지만, 심슨의 역설에 주의하여 새로운 가설을 검증해보자. 남녀를 구분해서 살펴보면 동질혼이 아니라 이질혼(heterogamy), 즉 강혼(hypogamy)과 승혼(hypergamy)의 규칙성이 발견될 것으로 예상된다. 이에 따라 응답자의 최종학력과 배우자의 최종학력 사이의 관계를 남녀별로 각각 나누어 살펴보시오. 적절하게 표준화하여 해석하시오. χ^2 독립성 검정을 통해 각각의 표에서 두 변수가 독립적인지 여부도 검정하시오.

교차표와 관련된 방법론적 이슈

Durhkeim의 〈자살론〉 교차표는 살짝 다른 종류의 것이다.

- Durkheim (1951[1897] 208)의 표를 잠깐 살펴보자.

	Suicides per million inhabitants	
	Urban population	*Rural population*
1866–69	202	104
1870–72	161	110

- 엑셀에서는 "개수" 대신 "합"이나 "평균"을 선택하면 매우 쉽게 〈자살론〉에서 쓰인 교차표를 흉내낼 수 있다.
- 이러한 교차표에 대해서는 보통 χ^2 검정을 수행하지 않는다.

교차표와 관련된 방법론적 이슈

예제 6. census.csv를 엑셀로 불러오시오. 미국의 권역(region) 별로 도시화 수준(urbanrate)에 따라 이혼 건수(divorce)의 평균이 어떻게 다른지 나타내는 교차표를 작성하시오. 이렇게 만든 표가 지금까지 배운 교차표와 구체적으로 어떻게 다른지도 설명하시오.

교차표와 관련된 방법론적 이슈

- 클릭 몇 번 하면 뚝딱 만들수 있다.

	A	B	C	D	E	F
1						
2						
3	**평균 : divorce**	**열 레이블**				
4	**행 레이블**	**1**	**2**	**3**	**4**	**총합계**
5	N Cntrl	2477	17438	43817	50997	24336
6	NE	4694		24205	27812	19304
7	South	16396	28827	31790	71579	27763
8	West	5768	3760	18943	33017	21076
9	**총합계**	**9919**	**18955**	**30434**	**35994**	**23679**
10						
11