계량분석

Three-Way Cross-Tabulations and Tables

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 4, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 둘 이상의 범주형 변수 사이의 관계
2. χ^2 독립성 검정
3. 교차표와 관련된 방법론적 이슈

둘 이상의 범주형 변수 사이의 관계

둘 이상의 범주형 변수 사이의 관계

eCampus에서 WVS_Cross-National_Wave_7_stata_v2_0.dta을 다운로드 받아 열자(**[Stata 코드]** 참고).

- 종속변수로서 "(이민자가 늘어나면) Increases the crime rate" 라는 문항을 찾자. 이 문항의 빈도분포를 살펴보고 결측치를 정리하자. Recoding이 필요하다면 하자. 변수 이름도 바꾸자.
- 독립변수로서 "When jobs are scarce, employers should give priority to people of this country over immigrants" 라는 문항을 찾자. 마찬가지로 정리하자.
- 행과 열은 어떻게 결정할 것인가, 표준화 방식은 어떻게 할 것인가를 결정하자.
- 이제 crosstab을 만들고 해석해 보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

삼원 교차표(three-way contingency table)을 연습하자.

- 이론적인 고민을 통해 독립변수와 종속변수의 사이에 존재하는 매개변수(mediating variables)나 조절변수(moderating variables)를 찾아낼 수 있다. 이론적으로 의미있는, 관찰가능한 이질성(observable heterogeneity)를 찾아낸다는 생각도 좋다.
- 여기서는 일단 "Losing my job or not finding a job"를 찾고, 분석에 투입하기 전 변수를 정리하자.
- **bysort** prefix로 이 변수에 따라 나누어 crosstab을 만들고 해석해 보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

χ^2 독립성 검정

비모수 통계학(non-parametric statistics)이라는 개념을 먼저 약간 이해할 필요가 있다.

- 여기서 비모수(non-parametric)이라는 단어는 추정할 모집단(population)의 분포에 대한 가정을 필요로 하지 않음을 의미한다.
- 일반적인 추리통계학(inferential statistics)에서는 모집단(population)의 분포에 대한 가정(보다 구체적으로 말해 정규분포 가정)을 전제로 한다. 하지만 이 가정은 사례 수(N)가 크면 중심극한정리(central limit theorem)에 의해 자동적으로 충족된다.
- 다시 말해, 모집단의 분포에 대한 가정이 필요없으면 사례(N)가 작아도 큰 문제를 일으키지 않는다는 의미다. 이로 인해 비모수 통계학은 꽤 많은 상황에서 유용하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

χ^2 독립성 검정(chi-square test of independence) 또는 χ^2 분석(chi-square analysis)은 비모수 통계학의 대표 주자로 여겨진다!

- χ^2 분포(chi-square distribution)는 모집단(population)의 분산(variance)에 대한 추정에서도 등장하지만, 그보다는 χ^2 분석에서 더 자주 쓰인다.
- χ^2 분석을 통해 교차표에서 주어진 두 변수 사이에 통계적으로 유의한(statistically significant) 관계가 있는지 확인할 수 있다.
- 여기서는 이론적으로 기대된 빈도(expected frequency)과 실제 관찰된 빈도(observed frequency) 사이를 비교한다.
- 이 비교의 목적은 교차표의 두 변수, 즉 두 개의 표본(sample)이 같은 모집단에서 나온 것인지를 확인하는 것이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

기대된(expected) 빈도의 계산은 "어떤 가정"에 입각해 있다.

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

- "성별"과 "종교인"이라는 두 변수가 독립된(independent) 다른 모집단에서 나온 표본이라면 다음의 곱셈 법칙이 성립한다:

$$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{A}) \cdot \mathrm{P}(\mathrm{B})$$

- e.g., $\mathrm{P}(\text{여자} \cap \text{종교인}) = \mathrm{P}(\text{여자}) \cdot \mathrm{P}(\text{종교인}) = (221/521) \cdot (165/521) \approx 0.13434$
- 이때, $\mathrm{N} = 521$ 이므로 $521 \cdot 0.13434 = 69.99$ 가 "여자 $\cap$ 종교인" 사건의 기대된(expected) 빈도이다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

Stata를 사용해 관찰된(observed) 빈도와 기대된(expected) 빈도를 직접 계산해보자(**[Stata 코드]** 참고).

- **display** 명령어를 통해 수식을 계산할 수 있다. 괄호 사용에 주의할 것!
- "관찰된 빈도(O)" 행렬의 주변확률(marginal probability) 정보를 이용해 옆의 "기대된 빈도(E)" 행렬을 채워넣자.

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

	종교인	비종교인
여자		
남자		

- 정답은 뒷 페이지에 있으니 참고하자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

이제 채워진 관찰된 빈도(O)와 기대된 빈도(E)를 곰곰히 비교해보자!

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

	종교인	비종교인
여자	69.99	151.01
남자	95.01	204.99

- 기대된 빈도(E)는 두 변수가 독립된 표본이라는 가정에 입각했을 때 이론적으로 기대된 빈도였다.
- 다시 말해, 기대된 빈도(E)가 관찰된 빈도(O)와 크게 다르지 않다면 두 변수는 서로 독립된 모집단으로부터 추출된 표본일 것이다.
- 반면, 기대된 빈도(E)와 관찰된 빈도(O)가 크게 다르다면 두 변수는 독립되지 않은 모집단으로부터 추출된 표본이라는 결론에 도달한다. 즉 두 변수는 연관되어 있는 것이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

지금까지 논의를 통해 χ^2 분석의 영가설 또는 귀무가설(null hypothesis)이 "두 변수는 독립적인 모집단에서 추출된 표본이다" 또는 더 쉽게 "두 변수는 서로 독립적이다"임을 알 수 있다.

- H_0: "성별과 종교인 여부는 서로 독립적이다."

자연스럽게 대안가설 또는 대립가설(alternative hypothesis)은 "두 변수는 독립적인 모집단에서 추출된 표본이 아니다" 또는 더 쉽게 "두 변수는 서로 연관되어 있다"가 된다.

- H_a: "성별과 종교인 여부는 서로 연관되어 있다."

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

χ^2 통계량(chi-square statistics)는 다음과 같이 계산된다([Stata 코드] 참고):

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

- 여기서 j와 k는 각각 교차표의 행과 열의 수, O_{jk}와 E_{jk}는 각각 j번째 행, k번째 열의 관찰된(observed) 빈도와 기대된(expected) 빈도를 의미한다.
- 앞의 예에서는 2×2 교차표였으므로 $j = k = 4$다.
- Stata에서 **display** 명령어를 사용해 직접 계산해 보자. 그냥 더하기, 빼기, 나누기, 제곱 수준의 계산이다.
- 사실 Stata 명령어를 사용해 쉽게 O_{jk}, E_{jk}, 그리고 χ^2를 계산할 수 있다.
- eCampus에서 religious.dta 데이터를 사용해 연습해 보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

Stata에서 **Pr=0.000** 하고 나오는 유의성(significance) 해석은 좀 더 복잡하다.

- 위와 같이 (표본에서 계산된) χ^2 통계량이 "영가설(null hypothesis)로 설정한 모집단의 성격"과 현저한(significant) 차이가 있다면 영가설을 기각(reject)하게 된다.
- 먼저 "영가설(null hypothesis)로 설정한 모집단의 성격"을 묘사한 이론적 확률분포를 그리게 된다. 여기서는 df 만큼의 자유도(degree of freedom; df)를 가진 χ^2 분포(chi-square distribution)를 뜻한다.
- 위와 같이 계산된 χ^2 통계량은 이론상 그 분포를 따른다.

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

- 자유도(df)는 행과 열의 숫자에서 1씩 빼서 곱한 값으로, 위 예제에서는 $[(2-1)(2-1)] = 1$이다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

- 그런데 표본에서 계산된 χ^2 통계량의 위치가 앞서 그려진 χ^2 분포 위에서 아주 구석진 곳에 놓여진다면, 이것은 그 표본의 성격이 "영가설(null hypothesis)로 설정한 모집단의 성격"과는 현저하게(significant) 다름을 의미한다.
- 보다 구체적으로, 연구자가 설정한 어떤 임계값(critical value)을 넘어섰을 때 표본의 성격과 영가설로 설정한 모집단의 성격 사이에 통계적으로 유의한 (statistically significant) 차이가 있다고 결론짓게 된다.
- 이 경우, 우리는 "영가설을 기각(reject)한다."
- 위 예제와 관련하여 Stata에서 보고된 p-value는 "영가설(null hypothesis)로 설정한 모집단의 성격에 비추어 보았을 때 표본에서 발견된 성격이 나타날 확률이 0.1% (=0.001)보다도 작음"을 의미한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

χ^2 독립성 검정

Crosstab을 χ^2 분석에 따라 해석할 때는 반드시 주의해야 할 점이 있다!

- 앞서 설명했듯 χ^2 분석 영가설 "두 변수가 상호 독립적이다" 였고 대립가설은 "두 변수가 상호 연관되어 있다" 일 뿐, 그 이상의 해석을 멋대로 덧붙여서는 안된다.
- 예컨대 학력 수준(고졸 이하/고졸/대졸/대학원졸 이상)과 기후관련 정보수집 태도(매우 관심없음/다소 관심없음/유보적임/다소 관심있음/매우 관심있음) 사이에서 χ^2 분석을 수행한 뒤, 단지 이에 근거하여 "교육 수준이 높아질수록 기후관련 정보수집 태도가 적극적이 된다" 식의 해석을 하면 틀린 것이다.
- 위 예제에서는 단지 "학력 수준과 기후관련 정보수집 태도는 서로 독립적인 사건 또는 현상이 아니며 통계적으로 유의한(statistically significant) 어떤 연관성을 지니고 있다"라고까지만 해석할 수 있을 뿐이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

아까 작업했던 KGSS_under35.dta에서 자신이 원하는 “범주형 변수” 두 개를 골라 다음을 연습하자.

- 두 변수 중 어느 쪽이 독립변수와 종속변수로 어울릴지 판단하자.
- 각 변수의 빈도분포는 어떻게 되어 있나? 결측치는 어떻게 되어있나? 역코딩이 필요한가?
- 행과 열은 어떻게 결정할 것인가?
- 표준화 방식은 어떻게 할 것인가?
- Crosstab을 만들고 해석해 보자.
- χ^2 독립성 검정을 수행하고 의미를 해석해 보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표와 관련된 방법론적 이슈

교차표와 관련된 방법론적 이슈

범주형 변수가 아니라 연속형 변수가 나오면 어떻게 할까?([Stata 코드] 참고).

- 정보의 손실(information loss)을 감수한다면 연속형 변수를 범주형 변수로 recode 할 수 있다. 그러나 그 역은 불가능하다(e.g., 숫자형 나이는 범주형 나이로 recoding 할 수 있지만, 반대는 불가능하다).
- Stata에서 **recode** 명령어로 연속형 변수를 범주형 변수로 바꾸어 crosstab을 만들수 있다. 연습해보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표와 관련된 방법론적 이슈

(눈치를 이미 챘을수도 있지만) χ^2 분석은 (데이터 없이) crosstab만 주어져도 그냥 수행할 수 있다(**[Stata 코드]** 참고).

- 생각해보라. 관찰빈도(O)와 기대빈도(E)를 아는데 원자료(raw data)는 필요가 없다. Crosstab만 있으면 충분하다. 물론 χ^2 통계량을 계산하는데도 원자료는 필요없다.
- Stata에서도 **tabi** 명령어로 즉석으로 crosstab을 만들어 χ^2 분석을 곧장 수행할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표와 관련된 방법론적 이슈

Durhkeim가 자살론에서 이용한 수많은 표들은 엄밀히 말해 지금껏 우리가 배운 crosstab과는 다른 것이다(**[Stata 코드]** 참고).

- Stata의 **table** 명령어를 조금 더 공부해야 한다. 이것은 crosstab을 계산할 때 쓴 **tabulate** 명령어와 다른 것이다.
- **table** 명령어로 일원(one-way), 이원(two-way), 삼원(one-way) 등 고차원의 교차표를 만들 수 있고, 익숙하게 해두면 (나중에 통계 논문을 쓸 때) 결과표를 쉽게 꾸미고 export 할 수도 있다.
- 무엇보다 Durkheim의 자살론에서 쓰인 표를 흉내내려면 이 명령어가 필수적이다.

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

Durkheim (1951[1897]: Free Press 판본 208페이지).

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

교차표와 관련된 방법론적 이슈

심슨의 역설(Simpson's Paradox)을 언제나 기억할 것!

- 화재 피해와 출동한 소방관의 수 패러독스(?)는 제법 유명하다.

	Low Damage	High Damage	Total
Few Firefighters	97 (69.8%)	49 (30.2%)	146 (100%)
Many Firefighters	42 (32.2%)	103 (67.8%)	145 (100%)
Total	139 (50.2%)	152 (49.8%)	291 (100%)

- 이것은 패러독스도 뭣도 아니다. 화재의 규모를 고려하지 않았기 때문에 생기는 허구적 인과성(spurious causality)에 불과하다. 그러나 이런 일이 입문자들의 계량적 사고방식에서는 흔하게 발견된다.
- 마음 속에서는 독립변수와 종속변수 사이에 인과관계가 있다고 생각하여 이론적으로 구상하지만, 단순한 crosstab은 이를 증명해주지 않는다. 다만 독립성 여부를 검정해줄 뿐이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY