

계량분석

Chi-square Analysis

김현우, PhD¹

¹충북대학교 사회학과 조교수

September 26, 2024



진행 순서

- 1 χ^2 독립성 검정
- 2 교차표의 시각화
- 3 연습문제

χ^2 독립성 검정

χ^2 독립성 검정

비모수 통계학 개념을 먼저 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(e.g., 정규분포 등)을 전제로 한다.
- 하지만 이 가정은 사례 수 N 이 충분히 크면 중심극한정리(central limit theorem)에 의해 자동적으로 충족된다.
- 다시 말해 “모집단의 분포에 대한 가정이 필요없으면 사례 수 N 이 작아도 큰 문제를 일으키지 않는다”는 의미가 된다.
- 이로 인해 비모수 통계학은 꽤 많은 상황에서 유용하다.



χ^2 독립성 검정

χ^2 독립성 검정이 비모수 통계학의 대표 주자로 여겨진다.

- χ^2 분포(chi-square distribution)는 모집단의 분산에 대한 추정/가설검정에서도 등장하지만, 그보다는 χ^2 독립성 검정(chi-square test of independence)에서 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.



χ^2 독립성 검정

기대빈도 E 는 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 E 는 $521 \cdot 0.13434 = 69.99$ 이다.

χ^2 독립성 검정

Stata에서 관찰빈도 O 와 기대빈도 E 를 직접 계산해보자.

- 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 display 명령어를 사용해 풀어보고 확인하자.



χ^2 독립성 검정

- 이제 채워진 관찰빈도 O 와 기대빈도 E 를 꼼꼼히 비교해보자!

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

	비종교인	종교인
남자	204.99	95.01
여자	151.01	69.99

- 기대빈도 E 는 두 변수가 독립적이라는 가정에 입각했을 때 이론적으로 기대된 빈도였다.
- 다시 말해, 기대빈도 E 가 관찰빈도 O 와 크게 다르지 않다면 독립성 가정(=두 변수는 서로 독립적이다)는 결론에 도달한다.
- 반면, 기대빈도 E 와 관찰빈도 O 가 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 : 성별과 종교인 여부는 서로 독립적이다.

H_a : 성별과 종교인 여부는 서로 독립적이지 않다.

- 만일 귀무가설을 기각할 수 있었다고 할지라도 여성이 더 종교적인지 여부 등은 독립성 검정을 통해서 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 J 와 K 는 각각 교차표의 행과 열의 수, O_{jk} 와 E_{jk} 는 각각 j 번째 행, k 번째 열의 관찰빈도 O 와 기대빈도 E 를 의미한다.
- 엑셀에서 직접 검정통계량 χ^2 값을 계산해 보자.



χ^2 독립성 검정

관찰빈도 O 와 기대빈도 E 가 다름을 확인하는 유의성 검정을 수행한다.

- (유의성 검정에서 사용하는) 이론적 확률분포는 확률변수 χ^2 의 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^2 는 0이다. 그러나 표본에서 계산된 χ^2 값이 크다면 (그러진) 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.



χ^2 독립성 검정

Stata에서 χ^2 독립성 검정을 수행해보자.

- 옵션을 사용하면 쉽게 O_{jk} , E_{jk} , 그리고 χ^2 를 계산할 수 있다.
- 독립성 검정의 결과를 적절히 해석해보자.
- 교차표를 제시할 때는 표준화했더라도, 독립성 검정은 빈도 자료를 근거로 수행한다는 것에 주의하자.
- 앞서 골랐던 또다른 질적 변수 두 개에 관하여 χ^2 독립성 검정을 수행해보고 교차표 아래에 덧붙여보자.



χ^2 독립성 검정

χ^2 분석은 원자료가 없어도 그냥 수행할 수 있다.

- 눈치를 이미 챘을수도 있지만 생각해 보자. 기대빈도 E 와 χ^2 통계량을 계산하는데 구태여 원자료는 필요가 없다.
- 이미 만들어진 교차표에서 관찰빈도 O 만 확인할 수 있으면 충분하다.
- Stata에서도 tabi 명령어로 즉석에서 χ^2 독립성 검정을 수행할 수 있다.
- 물론 빈도를 사용해야 하고 표준화된 값을 근거로 χ^2 독립성 검정을 하지 않아야 한다.



교차표의 시각화

교차표의 시각화

이원 교차표는 막대차트로 종종 시각화된다.

- 시각화가 꼭 의미있는 것은 아닐수도 있다. 그러나 이것이 인기있는 옵션으로 활용된다.
- 단순히 범주형 변수가 주어졌을 때, 빈도(frequency)나 밀도(density)를 시각화할 때는 histogram을 사용한다.
- 그러나 이미 빈도값이나 밀도값이 직접 주어졌을 때, 즉각 시각화할 때는 graph bar를 사용한다.
- 여러분이 직접 빈도값이나 밀도값을 계산해야 할 수도 있다. 이 경우에는 collapse를 사용한다.



교차표의 시각화

- 어떻게 시각화해야 할지 모르겠으면 우선 다른 연구자가 그려놓은 자료를 보고 배워야 한다.
- 특히 y 축, x 축의 의미를 잘 확인해보고 막대 하나하나가 무엇을 지칭하는지 살펴보자.
- 머리 속에서 어떤 식으로 그림이 나와주어야 하는지 먼저 종이에 그려본 다음에 코드를 구현하자.



연습문제

연습 1. KGSS 2023년 자료를 사용하여 다음 지시를 수행하시오.

- 이 SPSS 자료를 Stata로 불러오시오.
- 원하는 변수 하나를 골라 적절히 재부호화(recoding)하시오.
- 연령대와 해당 변수 간의 교차표를 작성하고 적절한 유의성 검정을 수행하시오.
- 교차표를 시각화하시오.
- 교차표의 의미를 해석하시오.

