계량분석
Cross-tabulation

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 26, 2024

진행 순서

1. 교차표의 이해
2. 이원 교차표
3. table의 사용

교차표의 이해

교차표의 이해

교차표는 둘 이상의 변수 간 관계를 분석하는데 있어 탄력적인 도구이다.

- 교차표(cross-table) 또는 분할표(contingency table)는 질적 변수들의 연관성을 보여준다.
- 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.
- 수많은 교차표들을 일단 만들어 놓고 꼼꼼히 살펴보면 이론적 영감을 얻을 때가 많다.

교차표의 이해

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.

교차표의 이해

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.

교차표의 이해

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.

교차표의 이해

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.

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

이원 교차표

이원 교차표

일반적으로 이원 교차표가 가장 유용하다.

- 이원(twoway) 교차표는 두 변수의 관계를 보여준다.
- Stata에서 한국일반사회조사(KGSS) 자료를 활용하여 행복감과 건강상태 간의 연관성을 분석해보자.
- 두 변수는 각각 어떤 척도로 측정되었나 살펴보자. 두 변수 모두 질적변수일 때 교차표를 사용하게 된다!
- 개별적으로 변수를 검토하고 결측치(missing values) 문제에 유의하면서 적절하게 재부호화(recoding)하자.
- 원래 값(raw values)와 레이블(label) 사이에 매칭을 보기 불편하다면 numlabel 명령어를 사용하면 편리하다.
- 어느 쪽이 행/열에 와야 해석에 편리한지 고민해보자.

이원 교차표

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 “건강 상태가 매우 좋다”라고 응답한 사람 중에 몇 퍼센트나 “매우 행복하다”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

이원 교차표

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.

이원 교차표

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 교차표를 만들 때부터 독립변수와 종속변수를 생각하고 만들어야 한다.

이원 교차표

행 합계, 열 합계, 총 합계로 표준화한 뒤 해석해보자.

- 행복감과 건강상태 변수 중 어느 쪽이 독립변수와 종속변수로 어울릴지 판단하자.
- 옵션으로 row, column, cell은 표준화 방식을 결정한다.
- nofreq 옵션은 빈도(frequency) 표시를 아예 빼버린다.

이원 교차표

결과물을 엑셀에 옮겨보자.

- 결과를 복사하여 엑셀에 붙여넣고 보기 좋게 꾸미자(때때로 유용하다).
- 표를 엑셀로 옮길 때 putexcel 명령어를 사용할 수 있다(잘 쓰면 매우 강력하다).
- 하지만 그보다 간단한 사용자 작성 명령어가 몇 가지 있으므로(e.g., estout, asdoc, tab2xl, tabout 등) 필요에 따라 각자 선택하고 연습하자.
- 자신에게 흥미있는 자료 또는 변수를 골라 교차표를 직접 만들어보자.

table의 사용

table의 사용

보다 깔끔하게 표를 만드는 또다른 명령어를 살펴보자.

- Stata에서 tabulate과 table는 종종 혼동되지만 상이한 명령어임에 주의하자.
- 우리가 흔히 tab을 사용하면 얻는 결과는 tabulate이다.
- 삼원(three-way) 교차표를 만들거나 평균을 요약하는 표를 만드는 등의 목적으로 table이 활용될 수 있다.

table의 사용

때때로 삼원 교차표가 필요한 경우도 있다.

- 삼원 교차표는 보통 지나치게 복잡하므로 늘 추천되는 것은 아니다.
- 삼원 교차표는 심슨의 역설(Simpson's Paradox)을 초래하는 관찰가능한 이질성(observable heterogeneity)를 찾아내는데 사용될 수 있다.
- 아래 교차표는 화재 피해와 출동한 소방관의 수 패러독스(?)를 보여준다. 이는 사실 화재의 규모를 고려하지 않았기 때문에 생기는 허구적 인과성(spurious causality)에 불과하다.

	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%)

- 즉 마음 속에서는 독립변수와 종속변수 사이에 인과관계가 있다고 생각하여 이원 교차표를 제시하지만 이것만으로는 인과관계를 보여주지 못한다.
- 어떻게 하면 좋을지 고민해보자.

table의 사용

Stata에서 삼원 교차표를 만들어보자.

- 성별로 나누어 행복감과 주관적 건강상태의 교차표를 만들어보자.
- bysort와 tabulate을 섞어 이원 교차표를 두 개 만들수 있다.
- 그러나 table 명령어를 사용하여 한번에 삼원 교차표를 만들수도 있다.

table의 사용

Durhkeim이 〈자살론〉에서 이용한 수많은 표들은 조금 다르다.

- 평균을 요약하는 표를 만들 때도 이 명령어가 보다 편리하다.

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

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

table의 사용

- table 명령어도 사용하여 표준화를 시도해보자.
- 엑셀로 옮겨 표를 꾸며보자.
- 몇 개 이상의 셀(cell)을 적절하게 해석해보자.