사회통계
두 질적변수 사이의 관계

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 두 질적변수 사이의 연관성
2. 질적변수와 양적변수 사이의 연관성

두 질적변수 사이의 연관성

두 질적변수 사이의 연관성

두 개의 질적변수 사이의 관계를 분석할 때는 우선 교차표를 살펴본다.

- 교차표(cross-tabulation)는 여러가지로 별칭을 가지고 있고 분할표(contingency table)라고도 종종 불린다.
- 나의 지도교수(John D. McCarthy)는 이렇게 말했다.
- "어떤 사회 현상을 두고 머신러닝이나 베이즈 통계로 분석해야 한다는 등의 말이 많지만, 교차표로 일단 설명할 수 없으면 다 소용없다."
- "사회 조사는 교차표에서 시작해서 교차표로 끝난다."
- 교차표 분석만으로도 사회학 고전의 반열에 오른 연구들이 제법 많다. 가장 위대한 예는 Emile Durkheim의 〈자살(Suicide)〉이다.

두 질적변수 사이의 연관성

교차표는 엑셀에서 피벗 테이블 기능을 사용해서 쉽게 만들 수 있다.

- 피벗 테이블(pivot table)은 엑셀 뿐 아니라 대부분의 데이터베이스에서 기본적으로 내장하고 있는 중요한 기능이기 때문에 반드시 명확하게 알아야 한다.
- preference.csv를 엑셀로 불러오자. 이 자료는 10명의 학생들이 영어와 수학을 각각 얼마나 좋아하는지에 대해 리커트 5점 척도(1=매우 싫어함; · · · ; 5=매우 좋아함)에 따라 보고하고 있다.
- 데이터를 훑어보고 무엇을 시사하고 있는지 한 번 생각해 보자.

두 질적변수 사이의 연관성

- 이제 피벗 테이블 기능을 사용해 엑셀로 교차표 만들어보자!
- 전체 자료를 하이라이트한 다음, [삽입]-[피벗 테이블]을 선택한 뒤, [확인]을 클릭하자
- 우측의 [피벗 테이블 필드]가 나타나면 [열], [행], [값] 부분에 적절한 변수들을 마우스로 드래그하여 드롭한다.
- 우측의 [Σ 값]에서 [값 필드 설정]을 고르고 [합계: ...]가 아니라 [개수: ...]를 선택한다.

두 질적변수 사이의 연관성

- 표의 오른쪽 끄트머리와 아랫쪽 끄트머리에는 “총합계”가 있는데 이것들을 주변(marginal)이라고 한다. 나중에 다루겠지만 매우 중요한 정보를 제공한다.
- 피벗 테이블 위에서 우클릭 하고 [피벗 테이블 옵션]을 누르면 빈칸을 0으로 채워넣을 수 있다.
- 이제 본격적으로 꾸미기에 앞서 전체 선택(Ctrl-A) 후 복사(Ctrl-C)하여 새로운 위치에 [값(V)]으로 붙여넣는 것이 좋다(Why?).

두 질적변수 사이의 연관성

- 표에는 '영어'나 '수학' 같은 변수 레이블(variable labels)을 제시해야 한다. 이를 위해 [홈] 메뉴의 [병합하고 가운데 맞춤] 기능이 좀 유용할 수도 있다.
- 속성(attribute labels)도 제시해야 한다. 여기서는 리커트 척도 1에서 5까지 의미에 대응한다.
- 필요에 따라 행이나 열의 순서를 바꾸어야 한다(Why?).
- 표를 완성하였다면 다시 복사해서 다른 워드프로세서(Word 또는 한글) 안으로 붙여넣자.

두 질적변수 사이의 연관성

교차표 내용이 원점수 그대로인 상태에서는 해석이 다소 모호하다.

- 이 표에서는 여전히 두 변수의 관계를 해석하기에 불편하다.
- 예를 들어 "수학을 다소 좋아하지 않는다"라고 응답한 사람이 40명이라면 그게 무슨 의미일까?
- "표본 중 10%가 수학을 다소 좋아하지 않는다"라면 그 의미를 명확히 파악할 수 있다.
- 다시 말해, 교차표에서 원점수(raw score)보다 비율(proportions) 또는 백분율(percentage)를 보고해야 해석이 쉽다.

두 질적변수 사이의 연관성

그런데 백분율을 구할 때 세 가지 방법이 있고 그 해석법도 달라진다.

- (1) 행 합계(row total)로 표준화하는 방법.
- 이때는 개별 셀의 값들을 그에 대응하는 "행 합계"로 나누어준다.
- (2) 열 합계(column total)로 표준화하는 방법.
- 이때는 개별 셀의 값들을 그에 대응하는 "열 합계"로 나누어준다.
- (3) 총 합계(grand total)로 표준화하는 방법.
- 이때는 개별 셀의 값들을 "총 합계"로 나누어준다.

두 질적변수 사이의 연관성

예제 4. 어떤 조사업체는 경쟁관계에 있는 휴대폰 브랜드 S와 i에 대한 선호도 조사를 A지역과 B지역에서 실시하였다. 총 200명의 조사 결과를 교차표로 정리하여 다음과 같은 결과를 얻었다. 세 가지 방식에 따라 백분율로 표준화하고 모든 셀을 해석하시오.

	S 브랜드	I 브랜드	행 합계
지역 A	56	40	96
지역 B	50	54	104
열 합계	106	94	200

두 질적변수 사이의 연관성

(1) 행 합계로 표준화하는 방법

	S 브랜드	I 브랜드	행 합계
지역 A	56/96	40/96	96/96
지역 B	50/104	54/104	104/104
열 합계	106/200	94/200	200/200

	S 브랜드	I 브랜드	행 합계
지역 A	0.58	0.42	1
지역 B	0.48	0.52	1
열 합계	0.53	0.47	1

두 질적변수 사이의 연관성

행 합계로 표준화하였다면 그에 적절한 해석이 필요하다.

- "A 지역주민 중 S 브랜드를 선호하는 비율은 58%이고, I 브랜드를 선호하는 비율은 42%이다."
- "B 지역주민 중 S 브랜드를 선호하는 비율은 48%이고, I 브랜드를 선호하는 비율은 52%이다."
- "전체 조사응답자 중 S 브랜드를 선호하는 비율은 53%이고, I 브랜드를 선호하는 비율은 47%이다."

두 질적변수 사이의 연관성

(2) 열 합계로 표준화하는 방법

	S 브랜드	I 브랜드	행 합계
지역 A	56/106	40/94	96/200
지역 B	50/106	54/94	104/200
열 합계	106/106	94/94	200/200

	S 브랜드	I 브랜드	행 합계
지역 A	0.53	0.43	0.48
지역 B	0.47	0.57	0.52
열 합계	1	1	1

두 질적변수 사이의 연관성

열 합계로 표준화하였다면 그에 적절한 해석이 필요하다.

- "S 브랜드를 선호하는 사람 중 A 지역주민은 53%, B 지역주민은 47%이다."
- "I 브랜드를 선호하는 사람 중 A 지역주민은 43%, B 지역주민은 57%이다."
- "전체 조사응답자 중 A 지역주민은 48%, B 지역주민은 52%이다."

두 질적변수 사이의 연관성

(3) 총 합계로 표준화하는 방법

	S 브랜드	I 브랜드	행 합계
지역 A	56/200	40/200	96/200
지역 B	50/200	54/200	104/200
열 합계	106/200	94/200	200/200

	S 브랜드	I 브랜드	행 합계
지역 A	0.28	0.20	0.48
지역 B	0.25	0.27	0.52
열 합계	0.53	0.47	1

두 질적변수 사이의 연관성

총 합계로 표준화하였다면 그에 적절한 해석이 필요하다.

- “전체 조사대상자 200명 가운데 A 주민이고 S 브랜드를 선호하는 사람은 28%,”
- “A 주민이고 I 브랜드를 선호하는 사람은 20%,”
- “B 주민이고 S 브랜드를 선호하는 사람은 25%,”
- “B 주민이고 I 브랜드를 선호하는 사람은 27%이다.”
- “전체 조사응답자 중 A 주민은 48%, B 주민은 52%이고,”
- “S를 선호하는 비율은 53%, I 브랜드를 선호하는 비율은 47%이다.”

두 질적변수 사이의 연관성

해석하려는 기준이 무엇인가에 따라 표준화 방법도 달라져야 한다.

- 물론 셋 다 연습해야 한다.
- 그런데 우리는 관습에 따라 종종 독립변수(X)에 해당하는 부분을 행(row)에 놓고, 종속변수(Y)에 해당하는 부분을 열(column)에 놓는 경향이 있다.
- 이 경우 행 합계(row total)로 표준화하는 편이 해석에 편리하다(Why?)
- 엑셀에서 피벗 테이블을 만들기 전부터 독립변수와 종속변수를 생각하고 만드는 편이 좋다.
- 경우에 따라 어떤 식으로 표준화 했는지 알려주지 않고 냅다 표만 던져주는 경우도 있다. 그래도 표를 쓱 보면 어떻게 표준화 했는지 알 수 있다(Why?).

두 질적변수 사이의 연관성

행에는 독립변수를, 열에는 종속변수를 두고 행 합계로 표준화하자.

- 표를 쉽게 만드려면 결국 셀 참조(cell reference)를 고정시키는 달러 싸인($)을 능숙하게 사용해야 한다.
- 만약 [A1] 셀을 참조할 때 A$1 을 입력하면 1행은 고정되고 A열은 변화한다.
- 만약 [A1] 셀을 참조할 때 $A1 을 입력하면 A열은 고정되고 1행은 변화한다.
- 만약 [A1] 셀을 참조할 때 A1 을 입력하면 A열과 1행은 모두 고정된다.
- 이것도 연습을 제법 필요로 한다.

두 질적변수 사이의 연관성

일단 피벗 테이블을 만든 뒤 나중에 X, Y를 뒤집고 싶을 수도 있다.

- 구태여 다시 만들지 않아도 엑셀에서 표를 뒤집을 수 있다.
- 데이터는 행렬이다(첫 주 참고). 행렬에서 X와 Y를 뒤집는 것을 행렬의 전치(transposition)라고 한다.
- 수학적으로 표현하면 알기 쉬울 수도 있는데 $(X_{ij})^T = X_{ji}$ 이다.

두 질적변수 사이의 연관성

예제 5. KGSS_congl2030.csv는 2030세대 1,817명이 한국의 대기업(그룹)이 국민경제에 지금까지 얼마나 기여했고(BIGECO), 앞으로 얼마나 기여할 것인지(BIGECOFU)에 관해 조사한 자료이다. 변수는 리커트 척도에 따라 코딩되었다(1=전혀 기여하지 못했다/못할 것이다; 2=별로 기여하지 못했다/못할 것이다; 3=다소 기여했다/할 것이다; 4=크게 기여했다/할 것이다). 두 변수 사이의 연관성을 살펴보기에 적절한 분석을 수행하고, 그 결과를 해석하시오.

두 질적변수 사이의 연관성

교차표 해석은 “가장 중요한” 기초통계학적 스킬이다.

- 이론의 생명은 우아한 단순화에 있다. 2×2 도식의 가치를 이해해야 한다!
- 교차표의 통계학적으로 의미와 해석에 관해서는 우리 수업 중반부와 후반부에 더욱 깊게 배운다.
- 지금 당장은 (1) 주어진 자료를 엑셀로 교차표 만들기, (2) 세 가지 표준화 방법과 그 해석법에 집중하자.

질적변수와 양적변수 사이의 연관성

질적변수와 양적변수 사이의 연관성

양적변수와 질적변수의 연관성도 쉽게 분석할 수도 있다.

- 첫번째 방법은 정보의 손실(loss of information)을 각오하고 양적변수를 질적변수로 재부호화(recoding)하는 것이다.
- 정보의 손실이 발생하더라도 해석이 보다 직관적이 된다면 이 방법이 가장 괜찮을 수도 있다.
- 즉 교차표는 두 질적변수 뿐 아니라 두 양적변수에 대해서도 사용할 수 있다.

질적변수와 양적변수 사이의 연관성

- 두번째 방법은 양적변수의 요약통계량을 다른 질적변수의 범주에 따라 살펴보는 것이다.
- 그러므로 질적변수의 속성별(e.g., 0=남성; 1=여성)로 양적변수의 평균과 표준편차를 따로따로 구하여 비교할 수 있다.

	평균	표준편차
남자	315.7	336.5
여자	178.2	222.6
합계	255.8	300.2

- 실제 보고서에서는 시각화하여 그 차이를 한 눈에 드러내는 것이 바람직하다(나중에 배우기로 한다).

질적변수와 양적변수 사이의 연관성

예제 6. KGSS_class.csv를 활용하여 주관적 계층인식과 월소득 사이에 연관성이 존재하는지 살펴보고자 한다. 이때 RXTINC0는 월소득을 만원 단위로 나타내고, CLASS는 주관적 계층을 나타낸다(1=하의 하; 2=하의 상; 3=중의 하; 4=중의 상; 5=상의 하; 6=상의 상). 두 변수의 관계를 살펴보기 위해 적절한 분석 기법을 수행하고, 그 결과를 제시하고 해석하시오.

질적변수와 양적변수 사이의 연관성

- 먼저 첫번째 방식(양적변수를 질적변수로 변환)으로 접근해보자.
- 엑셀에서는 일단 양적변수를 그대로 피벗 테이블로 만든 뒤, 나중에 그룹화(grouping)를 통해 질적변수로 바꾸는 쪽이 편리하다.
- 이제 월소득과 주관적 계층소득의 피벗 테이블을 만들어보자. 이때 CLASS를 행으로, RXTINC0를 열로 처리하자. [값]은 [개수: ...]로 선택해야 한다(Why?).
- 이제 RXTINC0 위에서 우클릭하고 [그룹(G)]을 통해 적절히 질적변수로 바꿀 수 있다.

질적변수와 양적변수 사이의 연관성

- 다음으로 두번째 방식(양적변수를 질적변수 범주에 따라 요약)을 살펴보자.
- 아까 만든 월소득과 주관적 계층소득의 피벗 테이블을 그대로 활용하자. [값]을 [평균: RXTINC0]를 선택하자.
- 이제 [총합계] 부분에 주관적 계층(CLASS)별로 월소득 평균이 나타난다.
- 만일 [값]을 [표준편차: RXTINC0]로 하면 주관적 계층별로 월소득 표준편차도 확인할 수 있다.
- 이렇게 구한 평균과 표준편차로 적절히 표를 꾸며보자.