사회통계연습

More on Association of Two or More Variables

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 17, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 둘 이상의 범주형 변수 사이의 관계
2. 둘 이상의 범주형과 숫자형 변수 사이의 관계

둘 이상의 범주형 변수 사이의 관계

둘 이상의 범주형 변수 사이의 관계

둘 이상의 범주형(categorical) 변수 사이의 관계를 볼 때는 일차적으로 교차표(cross-tabulation)를 살펴본다.

- 교차표는 여러가지 이름을 가지고 있지만, 그 중에서도 크로스탭(crosstab)이나 분할표(contingency table)가 널리 쓰인다.

"사회 조사는 crosstab에서 시작해서 crosstab으로 끝난다."

- 나의 지도교수(John D. McCarthy)도 똑같이 이야기했다: "어떤 사회 현상을 두고 머신러닝이나 베이즈 통계로 분석해야 한다는 등의 말이 많지만, crosstab으로 일단 설명할 수 없으면 다 소용없다."
- 사회통계학을 극단적으로 경멸하는 C. W. Mills도 이것(cross-classification)의 가치만은 인정함.
- 사회학의 저명한 고전들을 보면 crosstab만으로도 경험적 사회학 연구를 크게 발전시켰다. 가장 위대한 예제는 Emile Durkheim의 〈자살론(Suicide)〉이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

Crosstab은 엑셀에서 피벗 테이블(pivot table) 기능을 사용해서 매우 쉽게 만들 수 있다.

- eCampus에서 preference.csv를 다운받아 열어보자. 10명의 학생들이 독서와 수학을 각각 얼마나 좋아하는지에 대해 리커트 5점 척도(1=매우 싫어함; 5=매우 좋아함)로 보고하고 있다.
- 데이터를 훑어보고 무엇을 시사하고 있는지 한 번 생각해 보자.
- 전체 자료를 하이라이트한 다음, [삽입]-[피벗 테이블]을 선택한 뒤, [확인]을 클릭한다.
- 우측의 [피벗 테이블 필드]가 나타나면 [열], [행], [값] 부분에 적절한 변수들을 마우스로 drag-and-drop 한다.
- 이제 crosstab을 훑어보고 무엇을 시사하고 있는지 한 번 생각해 보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

Crosstab 꾸미기!

- 행이나 열 위에서 우클릭을 하고 [이동]을 통해 특정 행이나 열의 위치를 앞뒤/위아래로 이동할 수 있다.
- Crosstab 위에서 우클릭을 하고 [피벗 테이블 옵션]을 누르면 빈칸을 0으로 채워넣을 수 있다.
- 일단 crosstab 내용이 채워진 뒤에는 복사(Ctrl-C)하여 새로운 탭에 우클릭을 한 다음, 값(V)을 선택하여 붙여넣는 것이 좋다.
- 표를 예쁘게 꾸미거나 다시 복사해서 다른 워드프로세서(Word 또는 한글) 안으로 붙여넣어 넣는 것이 좋다.
- "독서"나 "수학" 같은 레이블(label)은 [홈] 메뉴의 [병합하고 가운데 맞춤] 기능이 좀 유용할 수도 있다.
- 행이나 열의 순서를 바꿀 수도 있다. 필요할 때는 반드시 바꾸어야 한다!
- 표의 오른쪽 끄트머리와 아랫쪽 끄트머리에는 "총합계"가 있는데 이것들을 주변(marginal)이라고 한다. 나중에 다루겠지만 매우 중요한 정보이므로 절대 삭제해서는 안된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

Crosstab은 범주형(categorical) 데이터를 다룬다고 했다.

- 하지만 첫 주에 언급했듯 숫자형(numerical) 자료는 범주형 자료로 변환(recoding)할 수 있다.
- 일단 숫자형 자료 그대로 crosstab을 만든 뒤, 나중에 그룹화(grouping)를 통해 범주형 척도로 바꾸어 표를 정리할 수 있다.

eCampus에서 2020년 9월.xlsx 파일을 다운받아 열자. 이것은 저번 주에도 사용한 바 있던 대기오염 데이터의 일부이다.

- 필터링 기능을 통해 강원도 태백시로 제한하자. 선별된 자료를 전체 하이라이트(Ctrl-A)한 뒤, 복사하여(Ctrl-C) 새 탭에 붙여넣자(Ctrl-V).
- 미세먼지(PM10)와 초미세먼지(PM2.5)의 crosstab을 만들어보자. 관상이 어떠한가?
- 이제 행(rows)과 열(column) 위에서 우클릭을 해서 "그룹화(grouping)"를 통해 범주형 척도로 바꾸어 표를 정리할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

eCampus에서 KGSS_public_v3.xlsx을 열자. 이것은 〈한국일반사회조사〉 2018년 자료에서 35세 이하의 남녀만을 따로 선별한 것이다.

- 행복감과 건강상태 변수를 보라. 두 변수는 각각 어떤 척도로 측정되었나?
- 두 변수의 관계를 보기 위해서는 어떤 분석 기법이 적절하다고 생각되나?
- 그 분석 기법을 수행하고 두 변수 사이의 관계를 보여주는 표를 만들어라(결측치는 표에서 삭제할 것).
- 필요최소한도로 그래프를 보기 좋게 꾸미자.

표를 만들어보고 두 변수의 관계를 해석해보라.

- 이 표를 보고 왜 "관계"를 해석하기에 다소 불편한가? 무엇이 불편했나?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

원점수(raw score) 그대로인 상태에서는 해석이 다소 모호할 때가 있다.

- 차라리 원점수보다 상대적인 비율(relative proportions)을 보면 편리하다.
- 여기서 상대적인 비율은 각 카테고리 별로 특정 셀에 보고된 응답의 비율을 의미한다. 예컨대, “전혀 행복하지 않다”라고 응답한 사람 중에 몇 퍼센트나 “매우 건강이 나쁘다” 라고 응답했는지를 보는 것이다.

그런데 상대비율을 구할 때는 세 가지 방법을 상상해 볼 수 있다!

- (1) Row total로 표준화하는 방법
- (2) Column total로 표준화하는 방법
- (3) Grand total로 표준화하는 방법

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

(1) Row total로 표준화하는 방법

- Row total (행 합계)는 각 행(row)의 합계를 나타내기 위해 추가적인 열(column) 안에 넣어놓은 숫자다.
- 표준화(standardize)한다는 말의 의미는 개별 셀(cell)을 해당 row total로 나눠주었다는 의미다.

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	총합계
매우 나쁘다	0	1	0	0	1
다소 나쁘다	1	10	14	2	27
좋지도 나쁘지도 않다	1	17	41	2	62
다소 좋다	1	13	74	10	98
매우 좋다	1	3	47	13	64
총합계	4	44	176	27	252

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	총합계
매우 나쁘다	0/1	1/1	0/1	0/1	1
다소 나쁘다	1/27	10/27	14/27	2/27	27
좋지도 나쁘지도 않다	1/62	17/62	41/62	2/62	62
다소 좋다	1/98	13/98	74/98	10/98	98
매우 좋다	1/64	3/64	47/64	13/64	64
총합계	4/252	44/252	176/252	27/252	252

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다
매우 나쁘다	0.00	1.00	0.00	0.00
다소 나쁘다	0.04	0.37	0.52	0.07
좋지도 나쁘지도 않다	0.02	0.27	0.66	0.03
다소 좋다	0.01	0.13	0.76	0.10
매우 좋다	0.02	0.05	0.73	0.20
총합계	0.02	0.17	0.70	0.11

- “건강이 매우 좋다고 응답한 사람의 73%는 다소 행복하다고 응답하였다.”

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

(2) Column total로 표준화하는 방법

- Column total (열 합계)는 alert각 열(column)의 합계를 나타내기 위해 추가적인 행(row) 안에 넣어놓은 숫자다.
- 표준화(standardize)한다는 말의 의미는 개별 셀(cell)을 해당 column total로 나눠주었다는 의미다.

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	총합계
매우 나쁘다	0	1	0	0	1
다소 나쁘다	1	10	14	2	27
좋지도 나쁘지도 않다	1	17	41	2	62
다소 좋다	1	13	74	10	98
매우 좋다	1	3	47	13	64
총합계	4	44	176	27	252

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	총합계
매우 나쁘다	0/4	1/44	0/176	0/27	1/252
다소 나쁘다	1/4	10/44	14/176	2/27	27/252
좋지도 나쁘지도 않다	1/4	17/44	41/176	2/27	62/252
다소 좋다	1/4	13/44	74/176	10/27	98/252
매우 좋다	1/4	3/44	47/176	13/27	64/252
총합계	4	44	176	27	252

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	총합계
매우 나쁘다	0.00	0.02	0.00	0.00	0.00
다소 나쁘다	0.25	0.23	0.08	0.07	0.11
좋지도 나쁘지도 않다	0.25	0.39	0.23	0.07	0.25
다소 좋다	0.25	0.30	0.42	0.37	0.39
매우 좋다	0.25	0.07	0.27	0.48	0.25

- “전혀 행복하지 않다고 응답한 사람의 25%는 건강이 매우 좋다고 응답하였다.”

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

(3) Grand total로 표준화하는 방법

- Grand total (총 합계)는 alert모든 셀(column)의 합계를 나타내기 위해 우측 하단 안에 넣어놓은 숫자다.
- 표준화(standardize)한다는 말의 의미는 개별 셀(cell)을 grand total로 나눠주었다는 의미다.

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	총합계
매우 나쁘다	0	1	0	0	1
다소 나쁘다	1	10	14	2	27
좋지도 나쁘지도 않다	1	17	41	2	62
다소 좋다	1	13	74	10	98
매우 좋다	1	3	47	13	64
총합계	4	44	176	27	252

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	총합계
매우 나쁘다	0/252	1/252	0/252	0/252	1/252
다소 나쁘다	1/252	10/252	14/252	2/252	27/252
좋지도 나쁘지도 않다	1/252	17/252	41/252	2/252	62/252
다소 좋다	1/252	13/252	74/252	10/252	98/252
매우 좋다	1/252	3/252	47/252	13/252	64/252
총합계	4/252	44/252	176/252	27/252	

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	총합계
매우 나쁘다	0.00	0.00	0.00	0.00	0.00
다소 나쁘다	0.00	0.04	0.06	0.01	0.11
좋지도 나쁘지도 않다	0.00	0.07	0.16	0.01	0.25
다소 좋다	0.00	0.05	0.29	0.04	0.39
매우 좋다	0.00	0.01	0.19	0.05	0.25
총합계	0.02	0.17	0.70	0.11	

- “전체 응답자 중 4%는 다소 건강이 나쁘고 별로 행복하지 않다고 응답하였다.”

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

각각의 표준화 방법에 따라 계산되는 상대비율은 해석방법이 달라진다!

- 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 row total을 가지고 표준화하는 편이 해석에 편리하다(Why?)
- 다시 말해, 엑셀에서 피벗 테이블을 만들 때부터 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(How?)

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

이제 엑셀에서 행(row)에는 독립변수를, 열(column)에는 종속변수를 두고 row total에 맞추어 표준화하자.

- 표를 쉽게 만드려면 결국 셀 참조(cell reference)를 고정시키는 달러 싸인($)을 능숙하게 사용해야 한다.
- [A1] 셀을 참조할 때, A$1 를 입력하면 1행을 고정시키고 A열은 움직임에 따라 변화한다.
- [A1] 셀을 참조할 때, $A1 를 입력하면 A열을 고정시키고 1행은 움직임에 따라 변화한다.
- [A1] 셀을 참조할 때, A1 를 입력하면 A열과 1행이 모두 고정된다.

다시 표를 보고 의미를 해석해보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

일단 엑셀에서 피벗 테이블을 만든 뒤, 표를 가져다 쓰다가 나중에 X, Y를 뒤집고 싶을 수도 있다.

- 구태여 다시 만들지 않아도 X와 Y를 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.
- 엑셀로도 물론 쉽게 할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

다시 KGSS 데이터로 돌아가자.

- 다음 두 개의 변수를 선택하자: (1) "다수에의 동조1: 본인과 생각이 다르더라도 대다수의 사람들의 의견을 따른다." 그리고 (2) "내집단 편향1: 사기업에서 사람을 뽑을 때 친척/친구에게 기회를 준다."
- 두 변수는 각각 어떤 척도로 측정되었나?
- 대체로 보아 각각 어떤 변수를 독립변수(X)로 그리고 종속변수(Y)로 보아야 할 것인가?
- 두 변수의 관계를 보기 위해서는 어떤 분석 기법이 적절하다고 생각되나?
- 그 분석 기법을 수행하고 두 변수 사이의 관계를 보여주는 표를 만들어라(결측치는 표에서 삭제할 것).
- 필요최소한도로 그래프를 보기 좋게 꾸며라.
- 표를 만들고 row total에 대해 표준화 한 뒤, 두 변수의 관계를 해석해보라.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형 변수 사이의 관계

Crosstab 해석상 주의사항!

- 독립변수와 종속변수라는 표현에도 불구하고 X와 Y의 관계는 결코 인과관계가 아니다.
- 교차표를 해석할 때 아직 "X와 Y 사이에 통계적으로 의미있는(statistically significant) 관계가 있다"고 말해서도 안된다.
- 이에 관해서는 우리 수업 나중에 배운다. 참고로 여러분이 저번 학기에 카이자승 검정(chi-square test)에서 배운 부분이다.
- 지금 당장은 세 가지 표준화 방법에 따라 상이한 해석법에 일단 익숙해지는 것이 중요하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

둘 이상의 범주형과 숫자형 변수 사이의 관계

둘 이상의 범주형과 숫자형 변수 사이의 관계

첫번째 추천하는 방법은 연속형 변수를 범주형으로 바꾸어 관계를 분석하는 것이다.

- 일정 정도 정보의 손실(loss of information)이 발생하지만 해석이 직관적이라면 괜찮을 수도 있다.

두번째 추천하는 방법은 연속형 변수의 요약통계량이 범주형 변수에 따라 어떻게 변화하는가를 관찰하는 것이다.

- "매우 건강하다"고 보고한 사람들의 연봉 평균 및 표준편차
- "다소 건강하다"고 보고한 사람들의 연봉 평균 및 표준편차
- "다소 건강하지 못하다"고 보고한 사람들의 연봉 평균 및 표준편차
- "매우 건강하지 못하다"고 보고한 사람들의 연봉 평균 및 표준편차

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

끝!

와~ 추석이다! 숙제나 하자~

충북대학교
CHUNGBUK NATIONAL UNIVERSITY