계량분석
Two-Way Cross-Tabulations

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 4, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 교차표의 기초 지식
2. 두 범주형 변수 사이의 관계

교차표의 기초 지식

교차표의 기초 지식

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 머신러닝이니 혹은 베이즈 통계학이 요구된다는 등의 말이 많지만, crosstab으로 일단 설명할 수 없으면 다 소용없다."
- 분명히 사회학의 저명한 고전들을 보면 crosstab만으로도 경험적 사회학 연구를 크게 발전시켰다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

사회학적 상상력의 원천으로써 crosstab

- 사회통계학을 극단적으로 경멸하는 C. W. Mills도 이것(cross-classification)의 가치만은 인정하였다(특히 장인기질론 부록을 볼 것).
- “이제까지 고립되어 있던 항목들을 서로 관련시켜 예기치 않았던 관계를 발견해냄으로써 상상력이 성공적으로 구현된다(번역서 p. 246).”
- “상관분류 기법은 물론 양적인 자료에만 국한되는 것이 아니다. 이 기법은 실제로 옛 유형을 비판하고 명료히 하는 것뿐만 아니라 ‘새로운’ 유형을 상상하고 파악하는 가장 좋은 방법이다(번역서 p. 416).”
- 번역서 pp. 254-6도 볼 것.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

E. Durkheim의 〈자살론(Suicide)〉 역시 "이론화의 도구로써 표(tables)"를 잘 활용한 위대한 고전이다.

- Durkheim은 여러 사회에 걸쳐 수많은 인구학적 특성별로 자살률을 비교하는 표를 제시하였다.
- 여러 사회의 집단적 성격에 따라 자살률이 같거나 상이함을 비교하여, 자살에 영향을 미치고 또 미치지 않는 사회적 사실(social facts)을 드러내는 비교방법론인 Method of Concomitant Variations을 활용였다.
- 다만 주의할 점은 (비슷하게 생기긴 했지만) Durkheim이 사용한 수많은 표들이 사실 일반적인 crosstab과는 좀 다르다는 점이다. 이에 대해서는 다시 이야기한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

Crosstab은 둘 이상의 변수 사이의 관계를 분석하는데 있어 매우 탄력적이고 강력한 도구이다.

- 원칙적으로 crosstab은 두 개의 범주형(categorical) 변수들 사이의 관계를 분석하는데 사용한다.
- 하지만 일정한 정보의 손실을 감수한다면 숫자형(numerical) 변수를 범주형(categorical) 변수로 변환할 수 있다(e.g., 연령/숫자에서 연령/카테고리로).
- 그러므로 crosstab은 (1) 두 개의 범주형 변수건, (2) 하나의 범주형과 하나의 숫자형 변수건, (3) 두 개의 숫자형 변수건 어떤 상황이건 무관하게 다 사용할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

Crosstab 만들기와 해석에는 몇 가지 중요한 규칙이 있다!

- 우리는 관습에 따라 독립변수(X)에 해당하는 변수를 행(row)에 놓고, 종속변수(Y)에 해당하는 변수를 열(column)에 놓는다.
- Crosstab은 각 셀(cell)에 빈도(frequency)를 보고하는 것에서 출발하지만, 사실 비율(percentage)을 보고하는 쪽이 해석을 위해서는 편리하다.
- 비율을 구할 때는 적어도 세 가지의 표준화(standardization) 방식이 있으며, 표준화를 어떻게 했는가에 따라 해석이 달라진다.
- 상이한 해석을 위해서는 조건부 확률(conditional probability)과 결합 확률(joint probability) 개념에 대한 이해를 요구한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

오늘 우리 수업은 다음의 순서를 따른다.

1. 측정의 척도(scale of measurement)를 복습한다.
2. 2×2 행렬(matrix)이 주어졌을 때 행합계(row total), 열합계(column total), 총합계(grand total)의 세 가지 기준에 따라 표준화하는 방법을 배운다.
3. 조건부 확률(conditional probability)과 결합 확률(joint probability) 개념을 복습한다.
4. 상이한 표준화 방식에 따라 달라지는 crosstab 해석을 연습한다.
5. Three-way crosstab 만들기와 해석을 연습한다.
6. 유의성 검정(significance test)을 위해 비모수통계학의 아이디어와 χ^2 검정(chi-square test)를 공부한다.
7. Crosstab과 관련된 몇 가지 방법론적 이슈를 검토한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

측정의 척도(scale of measurement) 분류 기준을 다시 떠올리자.

- 척도(scale)는 일단 숫자형(numerical)과 범주형(categorical)로 일차적으로 구분된다. 이때 교과서에 따라서는 숫자형을 양적(quantitative) 척도, 범주형을 질적(qualitative) 척도로 다르게 부르기도 한다.
- 숫자형 척도는 다시 연속형(continuous)과 이산형(discrete)으로 구분된다.
- 솔직히 말해, 전통적 척도(명목, 서열, 등간, 비율)보다 이 분류법이 훨씬 더 중요하다!

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

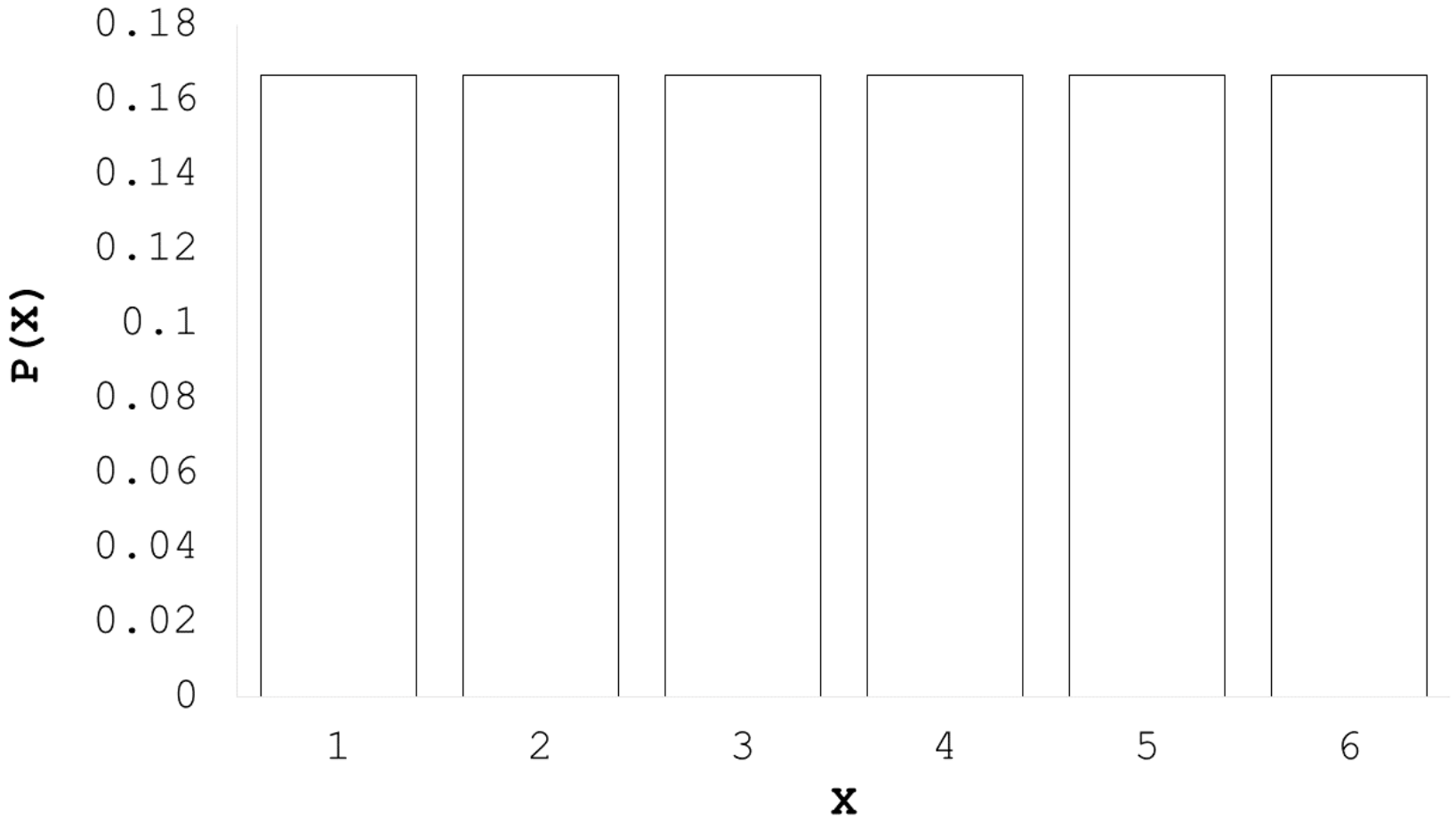
교차표의 기초 지식

연속형과 이산형에는 미묘한 차이가 있다.

- 동전 던지기나 주사위 던지기처럼 사건이 H, T 또는 1, 2, 3, 4, 5, 6로 값이 딱딱 떨어져서 셀 수 있는 경우가 있다.
- 이런 자료유형을 특별히 이산형(discrete)이라고 부른다(이산가족의 離散이다).
- 또다른 예는 분기별 목표를 달성한 사회복지사의 수, 회사를 떠나는 직원 수, 여성별로 출산한 아이의 수, 특정 달에 파산을 신청한 기업의 수 등이 있다.
- 반면에 딱딱 떨어지지 않아서 하나 둘 이렇게 셀 수 없는 경우를 연속형(continuous)이라고 부른다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

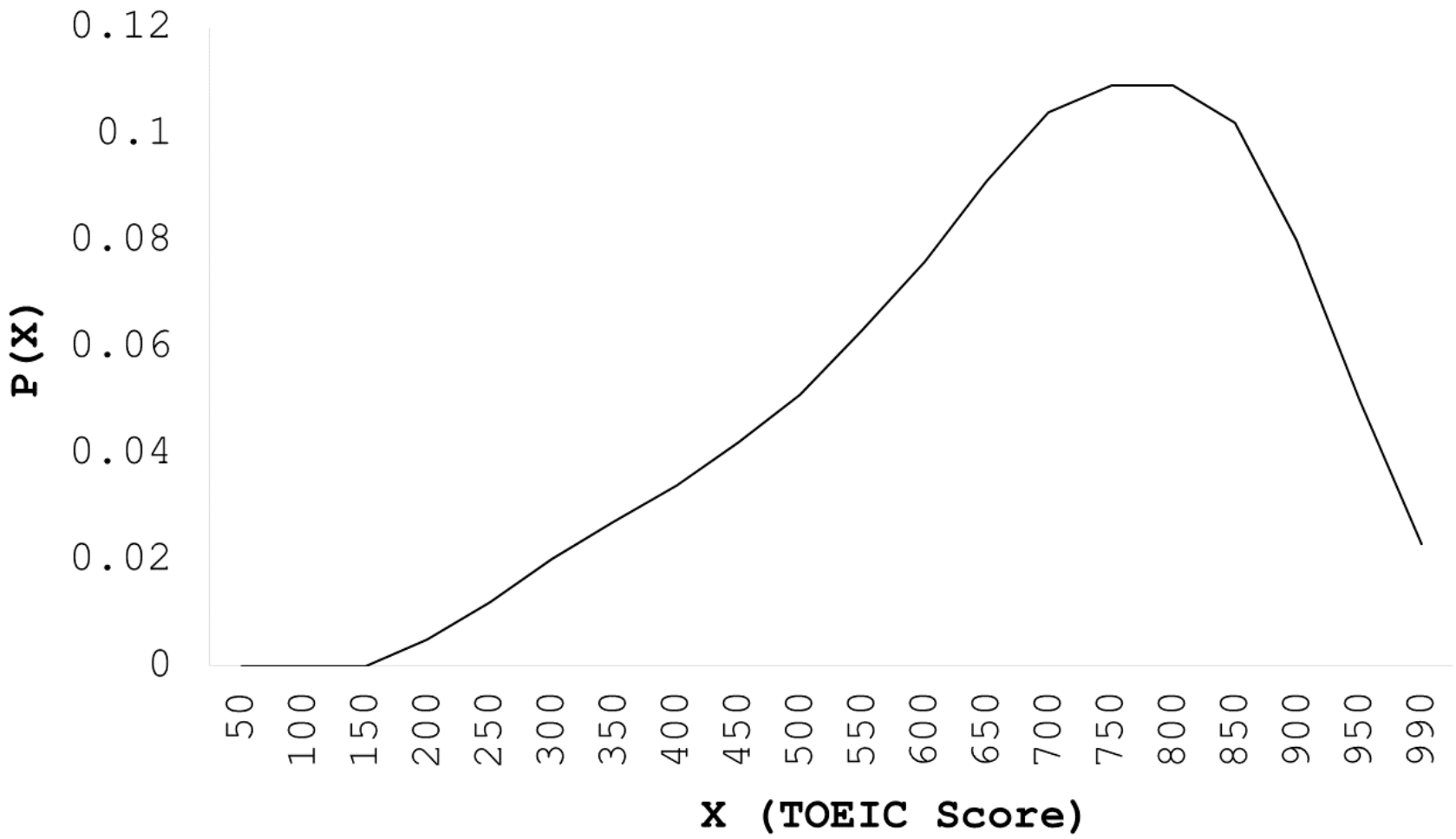
교차표의 기초 지식



이산형 그래프의 예(주사위 던지기)

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식



연속형 그래프의 예(직장인 토익 점수)

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

연속형(continuous) 척도는 어떤 예가 있을 수 있나?

- 가령 사람의 키나 체중, 소득액/세액, 펀드의 수익률, 지역별 태어난 아이의 수, 특정 작업을 완료하기까지 걸리는 시간 등은 연속확률변수가 될 수 있다.
- 왜 "여성별로 출산한 아이의 수"는 이산형(discrete)인데, "지역별 태어난 아이의 수"는 연속형(continuous)인가?
- 사람은 0.5421... 명을 낳을 수 없지만, 지역별로는 평균 따위를 계산하다보면 그런 숫자가 나올 수 있기 때문이다.

이산확률변수와 연속확률변수에서 확률의 표현이 조금씩 다르다.

- 만일 여성별로 출산한 아이의 수가 X라면, P(X=1)=0.4와 같은 표현이 가능하다.
- 만일 청주시 출산율이 X라면, $P(X=1) \approx 0$이다. 정확히 1이라는 숫자로 떨어질 확률은 무한히 작기 때문이다.
- 대신 청주시 출산율에 대해서는 다음의 표현이 가능하다:
  $P(0.8 < X < 1) = P(0.8 \leq X < 1) = P(0.8 < X \leq 1) = P(0.8 \leq X \leq 1) = 0.4$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

다음으로 확률 개념을 다시 되짚어보자.

- 조건부확률(conditional probability)이란 "다른 사건(B)이 이미 있어났다는 전제 아래, 한 사건(A)이 일어날 확률"을 의미한다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- 그 다음으로 결합확률(joint probability)이란 "두 사건이 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.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

교차표의 기초 지식

확률에는 몇 가지 기본법칙들이 성립한다.

- 덧셈 법칙(addition rule): $P(A \cup B) = P(A) + (B) - P(A \cap B)$
- 여기서 $P(A \cap B)$는 아까 설명한 결합확률(joint probability)이다.
- 여사건 법칙(complement rule): $P(A^C) = 1 - P(A)$
- 곱셈 법칙(multiplication rule): $P(A \cap B) = P(A|B) \cdot P(B)$
- 여기서 $P(A|B)$ 는 아까 설명한 조건부확률(conditional probability)이다.

확률의 기본법칙을 이해하면 독립 사건(independent event)과 종속 사건(dependent event)을 이해할 수 있다.

- 두 개의 사건 A와 B가 있을 때, $P(A|B) = P(A)$ 이거나 $P(B|A) = P(B)$ 이면 두 사건은 독립이다(Why?).
- (두 사건은 독립이 아니면 종속이므로) 위 식이 성립하지 않으면 종속이다.
- A와 B가 독립적(independent)인 사건들이라면 위 법칙들은 더욱 단순해진다: $P(A \cup B) = P(A) + P(B)$ 그리고 $P(A \cap B) = P(A) \cdot P(B)$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

두 범주형 변수 사이의 관계

eCampus에서 KGSS_under35.dta를 다운받고 Stata에서 열자. 이것은 〈한국일반사회조사〉 2018년 자료에서 35세 이하의 남녀만을 따로 선별한 것이다(**[Stata 코드]** 참고).

- 행복감과 건강상태 변수를 보라. 두 변수는 각각 어떤 척도로 측정되었나?
- 개별적으로 변수를 살펴보자. 어떻게 하면 될까? 결측치(missing values)는?
- 코딩은 제대로 되었나? 역코딩(reverse coding)하자.
- 두 변수의 관계를 보기 위해 crosstab을 만들어보자.
- 결과를 복사하여 엑셀에 붙여넣고 필요최소한도로 그래프를 보기 좋게 꾸미자.
- 레이블(label)이 없어져서 불편하지 않나? 대안이 있지 않을까? User-written commands를 찾아보자.

표를 만들어보고 두 변수의 관계를 해석해보라.

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

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 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

두 범주형 변수 사이의 관계

(1) Row total로 표준화하는 방법([Stata 코드] 참고).

- 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	행합계
매우 나쁘다	0	1	0	0	1
다소 나쁘다	1	10	14	2	27
좋지도 나쁘지도 않다	1	17	41	2	61
다소 좋다	1	13	74	10	98
매우 좋다	1	3	47	13	64
열합계	4	44	176	27	251

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다
매우 나쁘다	0.00	100.00	0.00	0.00
다소 나쁘다	3.70	37.04	51.85	7.41
좋지도 나쁘지도 않다	1.64	27.87	67.21	3.28
다소 좋다	1.02	13.27	75.51	10.20
매우 좋다	1.56	4.69	73.44	20.31
열합계	1.59	17.53	70.12	10.76

- e.g., "건강이 매우 좋다고 응답한 사람의 20%는 매우 행복하다고 응답하였다."
- e.g., "건강이 다소 또는 매우 좋다고 응답한 사람의 30.5%는 매우 행복하다고 응답하였다."
- e.g., "건강이 다소 좋다도 응답한 사람의 89%는 별로 행복하지 않거나 다소 행복하다고 응답하였다."
- e.g., "건강이 매우 나쁘다고 응답한 사람은 너무 적어 결과를 신뢰할 수 없다."

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

(2) Column total로 표준화하는 방법([Stata 코드] 참고).

- Column total (열 합계)는 각 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	행합계
매우 나쁘다	0	1	0	0	1
다소 나쁘다	1	10	14	2	27
좋지도 나쁘지도 않다	1	17	41	2	61
다소 좋다	1	13	74	10	98
매우 좋다	1	3	47	13	64
열합계	4	44	176	27	251

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	행합계
매우 나쁘다	0.00	2.27	0.00	0.00	0.40
다소 나쁘다	25.00	22.73	7.95	7.41	10.76
좋지도 나쁘지도 않다	25.00	38.64	23.30	7.41	24.30
다소 좋다	25.00	29.55	42.05	37.04	39.04
매우 좋다	25.00	6.82	26.70	48.15	25.50

- "매우 행복하다고 응답한 사람의 48%는 건강이 매우 좋다고 응답하였다."
- "다소 또는 매우 행복하고 않다고 응답한 사람의 79%는 건강이 다소 좋다고 응답하였다."
- "별로 행복하지 않다고 응답한 사람의 67%는 건강이 좋지 나쁘지도 않거나 다소 좋다고 응답하였다."
- "전혀 행복하지 않다고 응답한 사람은 너무 적어 결과를 신뢰할 수 없다."

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

(3) Grand total로 표준화하는 방법([Stata 코드] 참고).

- Grand total (총 합계)는 모든 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	행합계
매우 나쁘다	0	1	0	0	1
다소 나쁘다	1	10	14	2	27
좋지도 나쁘지도 않다	1	17	41	2	61
다소 좋다	1	13	74	10	98
매우 좋다	1	3	47	13	64
열합계	4	44	176	27	251

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

	전혀 행복하지 않다	별로 행복하지 않다	다소 행복하다	매우 행복하다	행합계
매우 나쁘다	0.00	0.40	0.00	0.00	0.40
다소 나쁘다	0.40	3.98	5.58	0.80	10.76
좋지도 나쁘지도 않다	0.40	6.77	16.33	0.80	24.30
다소 좋다	0.40	5.18	29.48	3.98	39.04
매우 좋다	0.40	1.20	18.73	5.18	25.50
열합계	1.59	17.53	70.12	10.76	

- “전체 응답자 중 29.5%는 건강이 다소 좋고 다소 행복하다고 응답하였다.”
- “전체 응답자 중 18.7%는 건강이 매우 좋고 다소 행복하다고 응답하였다.”
- “전체 응답자 중 16%는 건강이 좋지도 나쁘지도 않고 다소 행복하다고 응답하였다.”
- “전체 응답자 중 전혀 행복하지 않다고 응답한 이와 건강이 매우 나쁘다고 응답한 이는 너무 적어 결과를 신뢰할 수 없다.”

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

각각의 표준화 방법에 따라 계산되는 상대비율은 해석방법이 달라진다!

- 물론 셋 다 연습해야 한다. 분석 목적에 따라 다르기 때문이다.
- 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, Stata에서 crosstab을 만들 때부터 독립변수와 종속변수를 생각하고 만드는 편이 좋다.
- 그러나 grand total을 가지고 표준화하면 독립변수나 종속변수를 생각하지 않아도 된다(Why?)

경우에 따라 어떤 식으로 표준화 했는지 알려주지 않고 냅다 표만 던져주는 경우도 있다.

- 그래도 표를 쓱 보면 어떻게 표준화 했는지 알 수 있다(Why?)

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

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

	종교인	비종교인
여자	43.89	56.11
남자	22.67	77.33
합계	31.67	68.33

데이터로부터 왼쪽 표(crosstab)를 만든 뒤, row total을 기준으로 표준화하여 오른쪽 표를 만들었다.

- 이 표를 들여다보면 P(종교인|여자) $= 0.44$, P(비종교인|여자) $= 0.56$, P(종교인|남자) $= 0.23$, P(비종교인|남자) $= 0.77$ 임을 알 수 있다.
- 이것이 바로 조건부확률(conditional probability)이다.

만약 column total을 기준으로 표준화한 경우 다른 해석을 할 수 있다.

- 그 경우에는 P(여자|종교인), P(남자|종교인), P(여자|비종교인), P(남자|비종교인)을 구할 수 있다.
- 이것들도 물론 조건부확률(conditional probability)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

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

	종교인	비종교인	합계
여자	0.19	0.24	0.42
남자	0.13	0.45	0.58
합계	0.32	0.68	

데이터로부터 왼쪽 표(crosstab)를 만든 뒤, grand total을 기준으로 표준화를 하여 오른쪽 표를 만들었다.

- 이 표를 들여다보면 $\mathrm{P}(\text{종교인} \cap \text{여자}) = 0.19$, $\mathrm{P}(\text{비종교인} \cap \text{여자}) = 0.24$, $\mathrm{P}(\text{종교인} \cap \text{남자}) = 0.13$, $\mathrm{P}(\text{비종교인} \cap \text{남자}) = 0.45$ 임을 알 수 있다.
- 이것이 바로 결합확률(joint probability)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 범주형 변수 사이의 관계

아까 작업했던 KGSS_under35.dta를 다시 열고 다음에 답하자([Stata 코드] 참고).

- 행복감과 건강상태 변수 중 어느 쪽이 독립변수와 종속변수로 어울릴지 판단하자.
- **tabulate** 명령어 뒤에 변수 두 개를 나열하여 crosstab을 만들자.
- 옵션으로 **row**, **column**, **cell**은 표준화 방식을 결정한다.
- **nofreq** 옵션은 빈도(frequency) 표시를 아예 빼버린다.
- 해석하는 연습을 해보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY