사회통계연습
자료변환과 기술통계

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 자료변환과 응용
2. 기술통계

자료변환과 응용

자료변환과 응용

척도 간에는 변환이 가능하지만 오로지 일방향으로만 가능하다!

- 자료변환(data transformation)을 특히 재부호화(recoding)라고 부른다(기억이 안나면 복습하자).
- 정보량에 차이가 있기 때문에 변환하는 순간 "사라진 정보"는 복원 불가능하게 된다.
- 즉 정보가 많은 쪽에서 정보가 적은 쪽으로만 변환 가능하다(Why?).
- 예를 들면 비율척도로 측정된 월평균 소득은 다음과 같이 서열척도로 변환할 수 있지만, 그 역은 성립하지 않는다.

비율척도	서열척도
100만원 미만	1
100만원 이상 ~ 200만원 미만	2
200만원 이상 ~ 400만원 미만	3
400만원 이상	4

자료변환과 응용

재부호화는 정보상실로 이어지지만 해석상의 편의를 위해 유용할 수 있다.

- 가령 우리는 "그 사람 월급은 481만원이야" 표현 대신 "그 사람 돈 매우 많이 벌어"라는 표현을 사용한다(Why?).
- 언어적 표현(oral presentation)은 수학 특유의 엄밀성(rigorousness)을 상실하지만, 인간 사회에서의 의사소통에 유용하다.

비율척도	서열척도	언어적 표현
100만원 미만	1	"돈을 매우 못 벌어"
100만원 이상~200만원 미만	2	"돈을 다소 못 벌어"
200만원 이상~400만원 미만	3	"돈을 다소 많이 벌어"
400만원 이상	4	"돈을 매우 많이 벌어"

자료변환과 응용

연습 1. INCOME.SAV 자료의 가구소득 변수 INCOM0을 앞서 제시된 기준대로 (비율척도를 서열척도로) 재부호화하시오.

자료변환과 응용

재부호화의 기준은 잘 생각해야 한다.

- 언어적 표현은 그 자체로 상당한 설득력을 갖게 된다(Why?). 그러나 기준점(cutpoint) 또는 임계값(threshold)에 따라 사실을 왜곡할 수 있다.
- 위의 재부호화 기준(recoding scheme)이 타당한지는 맥락에 달려있다.
- 사회통계를 연습한다는 것은 결국 통계적 결과물에 대한 언어적 표현을 연습하는 것이기도 하다(Why?).

자료변환과 응용

어떤 재부호화하는 조금 다른 상황에서 이루어진다.

- 많은 사회조사에서는 성별 변수를 {남성=1, 여성=2}으로 부호화(encoding)한다.
- 이것은 분석에 그대로 사용될 경우 해석이 다소 불편하다(Why?).
- 그러므로 성별 변수는 주로 {남성=0, 여성=1} 또는 {여성=0, 남성=1}으로 재부호화된다.

자료변환과 응용

연습 2. INCOME.SAV 자료에서 성별(SEX)을 재부호화하시오. 적절히 변수 및 입력값 레이블(label) 또한 부여하시오.

자료변환과 응용

사회통계학에서는 설문조사로 수집된 자료를 분석하는 경우가 많다.

- 최근 동향을 보면 데이터 수집방식이 훨씬 더 다원화되었지만 여전히 설문조사는 중요한 자료의 원천이다.
- 사회현상에 관해 개인의 가치와 태도에 대해 설문할 때 리커트 척도(Likert scale)가 압도적으로 많이 사용된다.
- 한편 많은 사회조사에서는 리커트 척도의 초기 부호화 기준을 종종 거꾸로 되어있다. 다시 말해, 문항에 가장 적극적인 가치/태도를 보일 경우 가장 작은 값이 부여된다.
- 하지만 문항에 가장 적극적인 가치/태도를 보일 경우 가장 큰 값이 부여되는 편이 보다 직관적이므로 역부호화가 필요하다(Why?).

자료변환과 응용

※ 이번에는 귀하의 전반적인 가치관에 대해서 여쭈어 보겠습니다.

HAPPY
66. 귀하의 요즘 생활을 고려할 때 전반적으로 얼마나 행복 또는 불행하다고 생각하십니까?

____ ① 매우 행복하다	____ ③ 별로 행복하지 않다	____ (8) 선택할 수 없음
____ ② 다소 행복하다	____ ④ 전혀 행복하지 않다	____

FAMSATIS
67. 모든 것을 고려해 봤을 때, 가족과의 관계에 얼마나 만족하십니까?

____ ① 전적으로 만족한다	____ ④ 만족하지도 불만족하지도 않는다	____ ⑦ 전적으로 불만족한다
____ ② 매우 만족한다	____ ⑤ 다소 불만족한다	____ (8) 선택할 수 없음
____ ③ 다소 만족한다	____ ⑥ 매우 불만족한다	

68. 귀하는 다음의 각 상황에 대해서 옳다고 생각하십니까, 아니면 옳지 않다고 생각하십니까?

		전적으로 옳지 않다	대부분 옳지 않다	때에 따라 옳지 않다	전혀 잘못되지 않았다	선택할 수 없음
SEXATT1	1) 남녀가 결혼 전에 성관계를 갖는 것	①	②	③	④	(8)
SEXATT2	2) 결혼한 사람이 배우자가 아닌 사람과 성관계를 갖는 것	①	②	③	④	(8)
SEXATT3	3) 동성의 성인끼리 성관계를 갖는 것(동성애)	①	②	③	④	(8)

한국종합사회조사(KGSS)에서 사용된 리커트 척도의 예제

자료변환과 응용

연습 3. VALUE.SAV 자료에서 HAPPY와 FAMSATIS 변수를 적절히 부호화하시오. 또한 SEXATT1, SEXATT2, SEXATT3의 세 변수를 그대로 사용할 수 있는지 판단하고, 만약 필요하다면 적절히 재부호화하시오.

기술통계

기술통계

우리는 기술통계를 통해 자료를 요약한다.

- 기술통계(descriptive statistics)는 자료를 요약하는 목적을 가지고 있으므로 요약통계(summary statistics)라고도 부를 수 있다.
- "자료를 요약하라" 라는 말은 곧 기술통계를 제시하고 이를 언어적으로 표현하라는 말과도 같다.
- 자료를 요약하라고 했는데 자료를 그대로 복사해서 붙여넣으면 안된다(Why?).

기술통계

무엇보다, 가장 먼저, 자료의 척도를 옳게 식별해야 한다!

- 비율척도에 근접할수록 보다 다양한 기술통계를 활용할 수 있다(교재 24페이지).

변수유형	척도	추가되는 연산	요약 통계량
질적변수	명목척도	셈	빈도(frequency)와 구성비(percentage) 최빈값(mode)
	서열척도	순위 측정	중앙값(median) 범위(range)
양적변수	등간척도	덧셈/뺄셈	산술평균(arithmetic mean) 분산(variance)과 표준편차(standard deviation)
	비율척도	곱셈/나눗셈	기하평균(geometric mean) 조화평균(harmonic mean)

기술통계

기술통계는 주로 중심성향과 산포성향으로 나뉘어 보고된다.

- 중심성향(central tendency)는 자료를 대표하는(representative) 값이 자료의 가운데 어딘가에 위치해 있다라는 아이디어에 기반한다.
- 주로 세 가지 통계가 중심성향을 파악하기 위해 사용된다:

  (1) 평균(mean)
  (2) 중앙값(median)
  (3) 최빈값(mode)

- 셋 중 어느 것이 가장 좋은가? 평균, 중위값, 최빈값이 모두 같거나 비슷하면 뭘 써도 상관없다.
- 한편 평균은 극단값(outliers; extremes)에 민감하므로 극단값이 자료 속에 끼어있는 경우 평균이 커진다.

기술통계

중심성향으로만 충분히 자료를 잘 요약할 수 있을까? 사실은 그렇지 않다.

- 다음의 데이터가 주어졌다: {-10, 0, 0, 10}
- 평균, 중앙값, 최빈치는 각각 얼마인가?
- 다음의 데이터가 주어졌다: {-100, 0, 0, 100}
- 평균, 중앙값, 최빈치는 각각 얼마인가?
- 두 데이터는 정말로 같은가? 중심성향에 근거하여 두 데이터가 잘 요약되었나?

기술통계

그러므로 관찰값들이 얼마나 흩어져 있는가를 측정하는 통계가 필요하다.

- 산포성향(dispersion tendency)는 자료를 대표하는(representative) 값이 자료의 흩어진 정도에서 나타난다는 아이디어에 기반한다.
- 주로 세 가지 통계가 산포성향을 파악하기 위해 사용된다:

  (1) 범위(range)
  (2) 사분위수간 범위(interquartile range; IQR)
  (3) 분산(variance) 또는 표준편차(standard deviation)

기술통계

연습 4. INCOME.SAV 자료를 월평균 가구소득(INCOM0)의 기술통계를 보고하시오.