사회통계연습

변수의 조작과 자료 결합

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 새 변수 만들기
2. 결측치
3. 자료의 결합

새 변수 만들기

새 변수 만들기

두 개 이상의 변수가 주어졌을 때 새로운 변수를 만들 수 있다.

- Jamovi에서는 [데이터]-[계산]을 통해 새로운 변수를 계산할 수 있다.
- SPSS에서는 [변환]-[변수 계산]을 통해 새로운 변수를 계산할 수 있다.
- 적정한 공식이 있으면 이걸 활용할 수도 있다는 점에서, 예전에 배운 [변환]이나 [다른 변수로 코딩변경]과는 다르다!

새 변수 만들기

연습 1. COLLEGE.SAV 자료를 사용하여 전임교원 1인당 재학생수를 나타내는 변수를 만드시오. 이 비율이 지역에 따라 어떻게 다른지 확인해보시오. 또 전체교원 중 전임교원의 백분율(%)을 나타내는 변수를 계산하시오.

새 변수 만들기

연습 2. COLLEGE.SAV 자료를 사용하여 전체교원 중 전임교원의 백분율(%)을 나타내는 변수를 계산하시오. 이 비율이 지역에 따라 어떻게 다른지 확인해보시오.

새 변수 만들기

변수와 관측치가 많은 자료를 분석하다보면 혼란스러울 수 있다.

- SPSS에서 변수의 순서를 바꾸거나 삭제하는 방법을 익히자. 변수의 이름을 바꾸는 법도 연습하자.
- 아쉽게도 Jamovi에선 이런 관리가 다소 불편하다.
- 새 변수를 만든 다음에는 즉각 레이블(label)을 달아두지 않으면 잊어버린다.
- 그러나 원본은 항상 건드리지 않고 놔둔다(적어도 백업을 마련해 둔다!).

결측치

결측치

자료를 처음 살펴볼 때 빠진 부분은 없나 주의해야 한다.

- 이 과정은 종종 기술통계(descriptive statistics)를 살펴보면서 함께 이루어진다.
- 변수별로 하나하나 기술통계를 살펴보면 이상한 숫자가 실수로 입력되어 있거나 한 경우를 쉽사리 발견할 수 있다.
- 1차 자료(primary data)인 경우에는 필요에 따라 수집된 설문조사 종이 등을 재검토하고 재입력해야 할 수도 있다.
- 이 단계가 끝나기 전까지 종이 자료 등을 성급히 파쇄해서는 안된다(물론 요즘엔 종이 자료가 없는 경우가 더 많다).

결측치

- 2차 자료(secondary data)인 경우에도 이 과정을 생략해서는 안된다.
- (아무래도 입력상의 실수는 보통 없겠지만) 이 과정을 거쳐야만 자료와 친숙해질 수 있기 때문이다.
- 이 과정을 거친 뒤에 각 변수의 레이블과 자료유형(data type) 등을 정리해놓은 문서인 코드북(codebook)을 만들어 두기도 한다.

결측치

무엇보다 결측치 문제에 주목해야 한다.

- 결측치(missing values)는 아예 참여를 거부하지는 않았으나, 특정 문항에 대한 응답만큼은 거절한 경우이다.
- 결측치는 소득, 범죄경력 여부, 성관계 횟수 등 민감한 질문에 흔히 발생한다.
- 그러나 민감함 여부와는 상관없이, 질문이 잘못 설계되어 해당사항이 없는 경우에도 결측치가 발생할 수 있다(e.g., 모든 근로자에게 대학 전공-직업 일치도를 물어보는 경우).
- 빈칸으로 내버려두기도 하지만 -1, -999 같은 값으로 플래그(flag)를 세워두기도 한다(Why?).
- 이렇게 플래그 값을 준 경우 이것이 결측치를 뜻한다고 반드시 지정해야 한다.

결측치

- 만일 결측치 발생이 순수하게 임의(random)로 발생한다면 단지 추정치(estimates)의 표준오차(standard error)를 크게 할 뿐이다.
- 그러므로 이런 경우라면 결측치를 그냥 무시하고 평범하게 분석해도 사실 큰 문제가 되지 않는다(Why?).
- 그러나 응답자의 어떤 특성에 따라 결측치가 발생한다면(가령 범죄를 실제로 저질러 보았다면 범죄경력 여부에 무응답할 수도 있다), 결측치를 무시하는 것은 현명하지 못하다.
- 문제는 응답자의 어떤 특성에 따라 결측치가 발생할 가능성이 매우 높다는 점이다.

결측치

- 결측치를 무시하고 분석하면 결국 행단위 삭제(listwise deletion)를 수행하게 된다.
- 단 하나라도 사용하려는 변수에 결측치가 있다면 그 행(row)은 통째로 삭제하므로 상당히 극단적으로 표본 크기가 감소할 수 있다.
- 그러므로 분석 결과표에서 최종적으로 분석에 사용된 표본 크기를 세심하게 살펴보아야 한다.

id	age	politics	economy
1	20	1	3
2	21	1	.
3	.	4	2
4	23	2	.
5	23	.	3
6	.	2	4
7	22	3	1

결측치

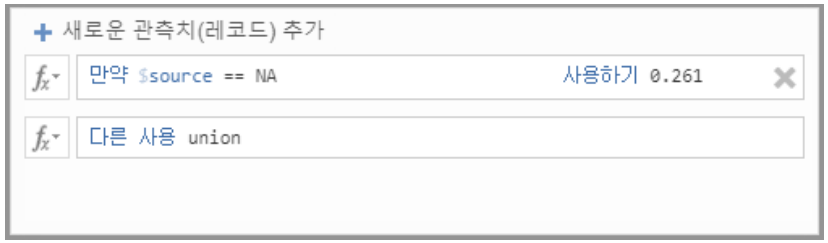
- 단순하게 말해, 결측치 문제가 너무 심각한 변수는 원칙적으로 사용하지 않는 편이 바람직하다.
- 그러나 꼭 사용해야 한다면 결측치는 기술적으로 메꿔볼 수 있다. 가장 단순한 방법은 평균 대체(mean imputation)와 중앙값 대체(median imputation) 등이다.
- 결측치를 내포하고 있는 변수를 먼저 기술통계로 살펴보아 (결측치를 제외하고) 평균이나 중앙값을 계산한 다음, 결측치는 평균 또는 중앙값으로 대체할 수 있다.

결측치

연습 3. NLSWORK_IND.SAV 자료에서 노조가입 여부(union)의 결측치 비중을 파악한 다음, 결측치를 적절히 대체하시오.

결측치

- 해당 변수의 기술통계를 살펴보고 평균이나 중앙값을 기억해 둔다.
- Jamovi라면 [데이터]-[변환]에서 새로운 변수를 만들고 평균으로 대체할 수 있다.



- SPSS에서 [분석]-[결측값 분석]을 수행하면 얼마나 결측값 비중이 높은지 파악할 수 있다.
- 아까 배웠던 [변환]-[변수 계산]에서 결측치가 담긴 변수를 하나 새로 복사한다.
- 다시 한 번 [변환]-[변수 계산]에서 평균값이나 중앙값으로 대체하되, "조건"에서 MISSING(·) 함수를 사용한다(Why?).

결측치

평균대체같은 간단한 방법들은 여러모로 결함을 가지고 있다.

- 결측치 대체 전후로 이 변수의 히스토그램을 그려보자.
- 무엇보다 결측치가 평균에 준할 것이라는 근거는 어디에도 없다(Why?).
- 성별 등 질적변수(qualitative variable)에서 결측치가 발생한 경우 그 평균이나 중앙값은 몹시 어울리지 않는다(최빈값은 그나마 좀 낫다).
- 그러나 일단은 가장 단순하기 때문에 의외로 그동안 널리 사용되어 왔다.

결측치

물론 기술적으로 훨씬 세련된 결측치 대체법도 있다.

- 전통적인 센서스(census)에서는 핫덱 대체(hot-deck imputation) 등을 사용했었다.
- 핫덱 대체는 비슷한 속성을 가진 다른 응답자의 값을 “기증받아” 무응답을 대체하는 기법이다.
- 일반적으로 무응답과 특성이 비슷한 개체를 찾기 위해서 여러 관련 변수들을 가지고 대체군을 형성한 다음에, 그 대체군 내에서 기증자를 무작위로 선택하여 기증자의 응답값으로 대체를 실시한다.
- (평균대체에 비한다면) 자료의 분포를 잘 유지해 주고 어떤 형태의 변수에도(e.g., 명목변수) 적용할 수 있다는 장점을 가지고 있다.
- 최근에는 훨씬 더 정교하고 세련된 결측치 대체법이 개발되어 사용되고 있다(학부 수준에서는 다루지 않는다).

자료의 결합

자료의 결합

때때로 두 개 이상의 자료를 결합하여 분석하기도 한다.

- 예를 들어 (한국은행에서 발표하는) 지역별 실업률이 (중앙선거관리위원회에서 발표하는) 지역별 지방선거 여당득표율에 부정적인 영향을 미친다는 연구가설을 살펴본다고 하자.
- 이 경우 (1) 살펴보려는 지방선거의 연도(가령 2022년)의 여당득표율을 구하고, (2) 그에 대응하거나 살짝 이른 시기의 지역별 실업률을 각각 따로 구해야 한다.
- 물론 두 데이터를 나중에 결합(merging)해야 하는 것은 물론이다!

자료의 결합

- 자료를 결합하려면 두 자료의 공통 식별자(common identifiers)를 기준으로 합쳐야 한다.

ID	var1	var2	var3
588	2	d	1
654	1	y	1
527	1	o	0
955	2	c	0
954	1	t	0

+

ID	var1	var2	var3
588	290	Apples	Breakfast
654	81	Bananas	Snack
527	63	Apples	Snack
955	6	Pears	Snack
954	146	Pears	Breakfast

ID	var1	var2	var3	var4	var5	var6
588	2	d	1	225	Apples	Breakfast
654	1	y	1	56	Bananas	Snack
527	1	o	0	245	Apples	Snack
955	2	c	0	46	Pears	Snack
954	1	t	0	121	Pears	Breakfast

자료의 결합

- 자료 결합에는 기본적으로 세 가지 방식이 있다: (1) 일대일(one-to-one), (2) 일대다(one-to-many), (3) 다대다(many-to-many).
- 다대다 결합은 좀처럼 사용되지 않으므로 우리 수업에서 다루지 않는다.
- 위의 예제(지역별 실업률 + 지역별 지방선거 여당득표율)는 오로지 1:1만 성립하는 상황이다(Why?).
- 주민등록대장 자료에 납세자명부 자료를 결합한다면 어떤 결합이 필요할까?
- 지난 5년간 응급환자명부의 자료에 주민등록대장 자료를 결합한다면 어떤 결합이 필요할까?
- 아파트 소유자 통계에 재산세 납입대장을 결합한다면 어떤 결합이 필요할까?

자료의 결합

- 추가(appending)로 불리우는 다른 형태의 자료 결합도 있다.
- 물론 전혀 다른 설문조사 자료를 함부로 이렇게 합쳐선 안된다(Why?).

ID	var1	var2	var3
588	2	d	1
654	1	y	1
527	1	o	0
955	2	c	0
954	1	t	0

ID	var1	var2	var3
1280	1	p	1
1917	2	t	0
1854	2	x	1
1701	2	e	0
1928	1	q	1

ID	var1	var2	var3
588	2	d	1
654	1	y	1
527	1	o	0
955	2	c	0
954	1	t	0
1280	1	p	1
1917	2	t	0
1854	2	x	1
1701	2	e	0
1928	1	q	1

자료의 결합

- Jamovi에서는 jReshape 모듈을 미리 설치하고 [Data]-[Merge Columns]를 선택한다(기능에 다소 제약이 있다).
- SPSS에서는 [데이터]-[파일 합치기] 안에서 [변수 추가] 또는 [케이스 추가]를 선택하여 수행한다. 각각 결합(merging)과 추가(appending)가 된다.
- 결합을 수행할 때는 두번째 1:1과 세번째 1:m 사이에서 신중히 선택해야 한다.
- 첫번째 '파일 순서를 기반으로 하는 일대일 합치기'는 보통 사용하지 않으니 무시해도 좋다.
- 결합을 수행할 때, 공통 식별자의 이름을 꼭 통일시켜야 한다(Why?)!

자료의 결합

연습 4. DATA1.CSV 자료에 DATA2.CSV 자료를 결합하고자 한다. 어떤 공통 식별자를 사용하는 것이 적합한지 그리고 어떤 결합 유형이 적합한지 판단하고 이를 직접 수행하시오.

자료의 결합

연습 5. INFO_MODELS.SAV 자료에 INFO_COMPANIES.SAV 자료를 결합하여 수입차 여부(foreign)와 차량무게(weight)의 연관성을 살펴보고자 한다. 어떤 공통 식별자를 사용하는 것이 적합한지, 그리고 어떤 결합 유형이 적합한지 판단하고 이를 직접 수행하시오.