기초통계연습

SPSS와 Jamovi 입문

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

July 4, 2023

진행 순서

안내사항

안내사항

수업에 관해 몇 가지 안내사항이 있습니다.

- 머지않아 양적 방법론을 사용해 논문을 쓰고자 하는 사람들을 위한 수업입니다.
- 펜과 노트를 준비하여 수식을 '손으로' 옮겨적는 연습을 하세요.
- 영어와 수학을 피하지 마세요.
- 강의안과 자료는 [여기]에서 무료로 다운받으세요. 공유 또는 판매하지 마세요. 오늘이 지나면 링크는 파기합니다.
- 메일로 질의응답하지 못하는 점 양해바랍니다.

안내사항

이 수업에 앞서 요구되는 기초지식이 있습니다.

- 통계학에 관한 기초 지식이 필요합니다.
- 특히 기술통계(descriptive statistics), 이론적 확률분포(theoretical probability distribution), 추정(estimation)과 가설검정(hypothesis test)의 논리를 알고 있어야 합니다.
- 이것들을 모르더라도 따라올 수 있고 “어떻게(how)” 하는지는 배울 수 있지만 “왜(why)” 하는지는 모르게 됩니다.
- 필요한 교재는 서점이나 도서관에서 자유롭게 선택할 수 있습니다.
- 통계 소프트웨어에 대한 지식과 경험은 필요하지 않습니다.

통계분석 패키지 선택

통계분석 패키지 선택

통계학은 두 개의 얼굴을 가지고 있다.

- 과학(science)과 손기술(art)!
- 손기술은 생각보다 매우 중요하며 연습해두지 않으면 연구나 실무에서 통계분석을 전혀 쓸 수 없다.
- 본격적인 분석을 위해 반복숙달이 필요하고, 편리하게 통계분석만을 위한 전용 패키지를 사용하게 된다.
- 손기술에의 숙련을 위해서는 투자한 시간이 절대적으로 많아야 한다. 머리 못지 않게 엉덩이로 공부한다!

통계분석 패키지 선택

왜 통계분석 패키지를 배워야 할까?

- 지난 주까지 기초통계학을 학습하였다면, 이 수업부터는 좀 더 본격적인 통계분석 패키지 사용법을 배운다!
- 여러분이 "뭔가 배우긴 했는데 실무나 연구에서 해보라고 하면 자신이 없는데..." 라고 느낀다면 지극히 정상이다.
- 실습에서 수학은 전혀 사용하지 않는다. 이것은 장인이 공학을 필요로 하지 않는 것과 마찬가지이다.
- 그럼 우리는 왜 지난 주에 기본 원리를 배웠을까? 스스로의 통계분석에 확신을 갖고 (나 자신과) 남에게 설명할 수 있기 위해서이다.

통계분석 패키지 선택

엑셀은 그 나름대로 중요한 강점을 가지고 있다.

- R보다 훨씬 배우기 쉽고 졸업 후 실무에서도 훨씬 다양한 분야에서 압도적으로 많이 쓰인다.
- 분석이 이루어지는 과정을 실시간으로 지켜볼 수 있다. 비슷한 장점을 가진 도구는 Matlab이 있는데 이쪽은 배우기가 R만큼이나 어렵다.
- 엑셀은 "통계 학습의 도구"로는 유용해도 "통계 분석의 도구"로는 상당히 불편하다 (불가능하지 않다).
- (억지로 또다른 단점을 생각해보면) 엑셀을 너무 잘하면 회사에서 유능한 인간으로 취급받으므로 일거리가 늘어난다.

통계분석 패키지 선택

R은 통계학과 데이터사이언스 분야에서 폭넓게 사용된다.

- 유저들의 커뮤니티가 활성화되어 있어 지금 이 순간에도 통계분석을 위한 여러 새로운 도구가 개발되어 무료로 공개되고 있다.
- 단순한 통계분석, 시각화(visualization), 데이터 사이언스(data science)에 이르기까지 다양한 분야에서 활용되는 분석툴이다.
- 오픈소스(open source) 소프트웨어이므로 그 사용이 무료다.
- 배우기 어렵고 수많은 시행착오를 요구하며 (통계학이나 데이터사이언스 자체를 떠나) R 자체를 배우는데 상당한 시간을 써야 한다.
- (약간 주관적이지만) 남이 쓴 코드를 알아보기 어렵다. 즉 코드가 지저분하다.

통계분석 패키지 선택

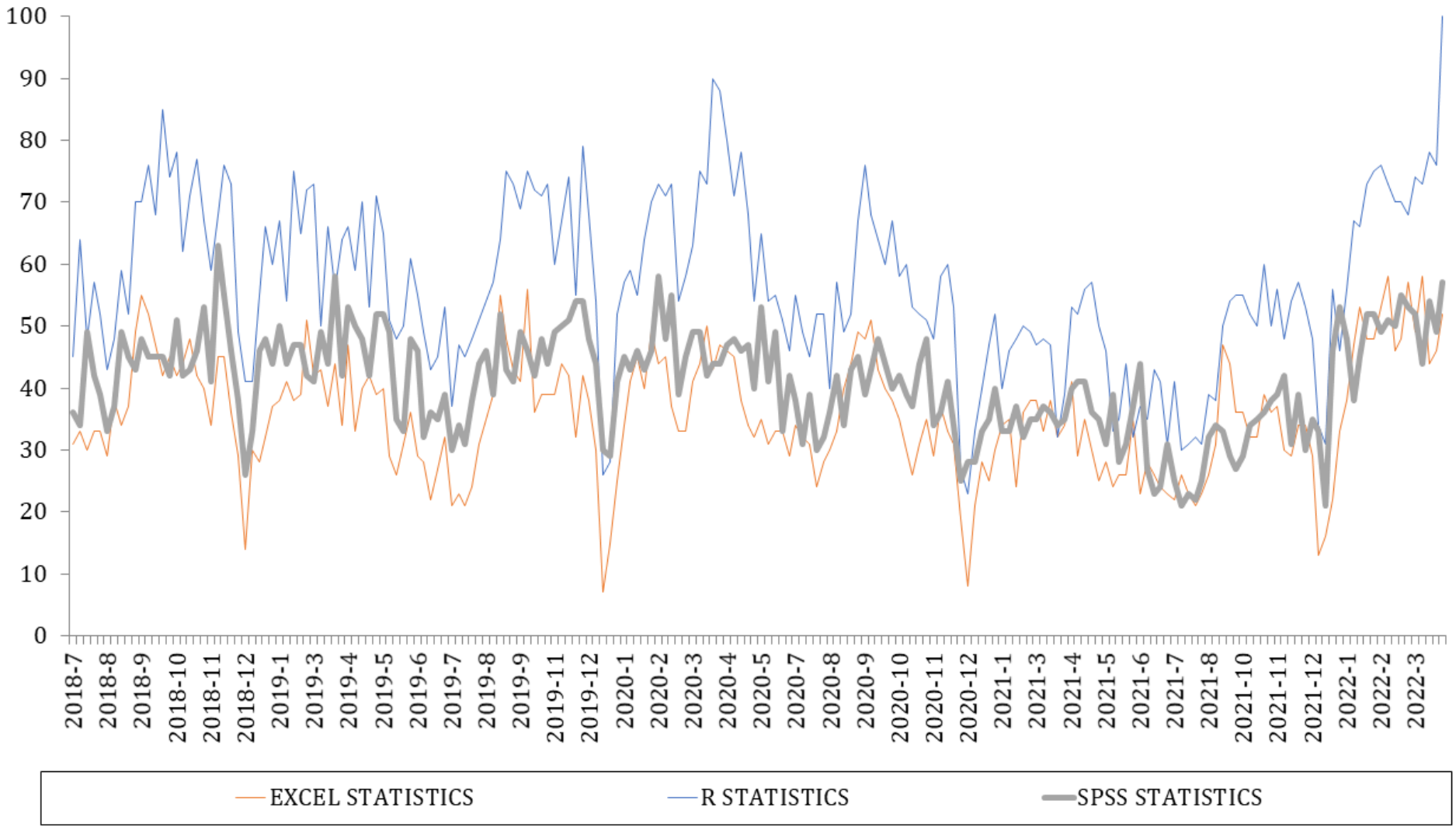
SPSS가 우리의 선택이다!

- SPSS는 배우고 사용하기 쉽다. 졸업 후 학교 밖 현업기관에서도 폭넓게 사용된다. 이용자 규모가 상당히 크기 때문에 인터넷을 통해 모르는 것을 질문하거나 자료를 찾아보기에 용이하다.
- 다만 (1) 개인이 구매하기에 너무 비싸고 (2) 고급의 통계분석에 제약이 있다.
- 요즘에는 빅데이터 붐으로 학교에서도 점차 SPSS 대신 R을 가르치자는 유행이 나타나고 있다. 취지는 알지만 학습곡선(learning curve)이 가파르기 때문에 중도포기자가 다수 나온다.
- 응용통계학을 배우기도 벅찬데 R까지 배우는 부담은 줄이고 일단 SPSS로 시작하는 것도 나름 괜찮은 선택이다.

통계분석 패키지 선택

- SPSS 라이센스는 상당한 고가이므로 개인은 물론이고 기관조차도 구매해주지 않을 수도 있다.
- 이런 경우 (SPSS보다 기능이 제약되어 있어도) 비슷한 오픈소스 소프트웨어를 사용한다.
- 특히 JASP와 PSPP 그리고 Jamovi가 널리 알려져있다.
- 이 수업에서 우리는 (SPSS 뿐 아니라) Jamovi를 함께 배우게 된다.

통계분석 패키지 선택



소스: Google Trends

자료분석의 첫 출발

자료분석의 첫 출발

자료와 관찰단위 그리고 분석단위의 개념을 되짚어보자.

- 자료(data): 분석 대상의 속성(attributes)을 기록한 수열(series).
- "자료의 관찰"은 사람을 단위로 이루어져도, 나중에 "자료의 분석"은 각 사람이 속한 지역 또는 국가 단위로 이루어질 수도 있다.
- 관찰단위(unit of observations): 자료의 관찰과 수집이 이루어지는 단위
- 분석단위(unit of analysis): 분석이 수행되고 발견이 일반화(generalization)될 수 있는 단위

자료분석의 첫 출발

열은 변수가 되고, 행은 관측치가 된다.

- 앞서 데이터는 수열이라고 말했지만, 사실 여러 변수(variables)를 갖기 때문에 각각의 변수가 하나씩 수열을 이루어서 결국 행렬(matrix)이 된다. 즉 데이터는 사각형이다.
- 행렬은 행과 열의 조합이다. 행은 가로고 관측치(observations)가 된다. 열은 세로고 변수(variables)가 된다.
- 변수는 변(vary)할 수 있는(able) 숫자이고, 관측치는 해당 변수에 실제 채워진 값들이다.
- 데이터를 남에게 설명할 때 "변수는 몇 개, 관측치는 몇 개"와 같은 식으로 표현할 수 있다.

자료분석의 첫 출발

hsb2.sav 파일을 SPSS로 불러오자.

- SPSS를 기동하기 위해서는 (1) 해당 아이콘을 더블클릭하거나 (2) SPSS 데이터파일을 더블클릭하거나 (3) Ctrl-Esc 다음 SPSS를 입력한다.
- 각 윈도우가 무엇을 의미하는지 살펴보자.
- 맨 아랫쪽 [데이터 보기(D)]와 [변수 보기(V)]를 각각 클릭해보자.
- 변수와 관측치는 각각 몇 개인가?
- '레이블'을 통해 각 변수의 의미와 부호화 방식(coding scheme)을 확인해보자.
- 이 자료의 관찰단위는 무엇인가?

자료분석의 첫 출발

각각의 변수마다 자료유형이 다를 수 있다.

- 자료유형(data type)이 다르다는 것은 보다 구체적으로 측정의 척도(scales of measurement)가 다르다는 것을 뜻한다.
- 측정의 척도란 변수가 정의(define)되고 유형화(categorize) 되는 방식을 뜻한다.
- 보통 네 종류의 척도를 거론할 수 있다: 명목(nominal), 서열(ordinal), 등간(interval), 비율(ratio).

자료분석의 첫 출발

네 종류의 척도에는 위계적인 관계가 있다.

- 명목척도(nominal scale): 범주(category)가 존재할 뿐, 그들 사이에 우열이나 대소가 없음.
- 서열척도(ordinal scale): 범주들 사이에 서열 혹은 순서(order)가 있음.
- 등간척도(interval scale): 범주들 사이에 우열이나 대소가 있고, 그 간격은 동등함 (등간; equal interval).
- 비율척도(ratio scale): 범주들 사이에 우열이나 대소가 있고, 그 간격도 동일하며, 절대영(absolute zero) 또한 의미를 가짐.

자료분석의 첫 출발

좀 더 단순하게 질적-양적 구분을 하기도 한다.

- 명목척도와 서열척도를 질적 변수(qualitative variable)로, 등간척도와 비율척도를 양적 변수(quantitative variable)로 묶을 수도 있다.
- 종종 리커트 척도(Likert scale)는 본래 성질상 서열척도이지만, 편의상 대체로 등간(approximately interval)이라고 부르며 마치 양적 변수처럼 분석하기도 한다.
- 혹은 같은 개념에 대해 여러 개의 리커트 척도를 활용하여 측정한 뒤, 이를 모두 더하거나 평균을 구한 점수를 대체로 등간척도로 보기도 한다.

자료분석의 첫 출발

연습 1. hsb2.sav 파일에서 사용된 변수의 척도를 파악해보자.

변수	척도	변수	척도
id		read	
gender		write	
race		math	
ses		science	
schtyp		socst	
prog			

자료분석의 첫 출발

SPSS에서 자료유형을 확인하고 변환해보자.

- SPSS의 한국어 번역은 통계용어규정을 제대로 반영하고 있지 않음에 주의하자.
- 측도는 measure의 번역어이고, nominal, ordinal, scale은 각각 명목형, 순서형, 척도로 번역되었다.
- 위의 척도(scale)는 결국 비율척도과 등간척도를 뭉뚱그린 것이다. 사실 실무상 이런 경우가 종종 있다(Why?).
- 자료를 보고 필요에 따라 측도를 바꾸어보자. 측도를 옳게 설정하는 것은 종종 건너뛰어도 된다(하지만 가끔씩 필수적이다).

자료분석의 첫 출발

자료유형의 판별은 기초적이지만 매우 필수적이다.

- 거의 모든 통계학이나 데이터분석 수업이 자료유형 식별로부터 출발한다.
- 여러 자료유형을 살펴보고 명목, 서열, 등간, 비율, 또는 대체로 등간인지 식별하는 연습을 계속해야 한다.

자료분석의 첫 출발

연습 2. states.sav 파일을 Jamovi로 불러와 앞서 살펴본 내용을 다시 한 번 복습해보자.

자료 변환

자료 변환

척도 간에는 변환이 가능하지만 오로지 일방향으로만 가능하다!

- 자료 변환의 가장 중요하고 흔한 유형은 재부호화(recoding)이고 이를 필요에 따라 능숙하게 수행할 수 있어야 한다.
- 정보가 많은 쪽에서 정보가 적은 쪽으로만 변환 가능하다(Why?).
- 정보량에 차이가 있기 때문에 변환하는 순간 "사라진 정보"는 복원 불가능하게 된다.
- 예를 들면 비율척도로 측정된 월평균 소득은 다음과 같이 서열척도로 변환할 수 있지만, 그 역은 성립하지 않는다.

비율척도	서열척도
100만원 미만	1
100만원 이상 ~ 300만원 미만	2
300만원 이상 ~ 500만원 미만	3
500만원 이상	4

자료 변환

재부호화는 종종 정보상실로 이어지지만 해석상 편리할 수도 있다.

- 가령 우리는 "그 사람 월급은 481만원이야" 표현 대신 "그 사람 월급 꽤 많이 벌어"라는 표현을 사용한다(Why?).
- 언어적 표현(oral presentation)은 수학 특유의 엄밀성(rigorousness)을 상실하지만 인간 사회에서의 의사소통에 유용하다는 점을 기억하자.

비율척도	서열척도	언어적 표현
100만원 미만	1	"돈을 매우 못 벌어"
100만원 이상~300만원 미만	2	"돈을 다소 못 벌어"
300만원 이상~500만원 미만	3	"돈을 다소 많이 벌어"
500만원 이상	4	"돈을 매우 많이 벌어"

자료 변환

질적 변수의 기초는 가변수에서 출발한다.

- 가변수(dummy variable)란 처방, 조건, 또는 상황 등이 존재하면(present) 1로, 그것이 부재하면(absent) 0으로 더미 코딩(dummy coding)된 변수이다.
- 예를 들어, “성별이 여성이다”에 관한 가변수라면 ‘여성이다(1)’ 또는 ‘여성이 아니다(0)’ 중 하나가 된다.
- 많은 사회조사에서는 성별 변수를 {남성=1, 여성=2}으로 부호화(coding)한다.
- 이것은 분석에 그대로 사용될 경우 해석이 다소 불편하기 때문에(Why?), 가변수로 자료 변환해야 한다.
- 성별 변수는 주로 {남성=0, 여성=1} 또는 {여성=0, 남성=1}으로 재부호화된다.

자료 변환

- SPSS에서는 [변환]-[다른 변수로 코딩변경(R)] 기능을 이용한다.
- 이 기능은 가변수 만들기 뿐 아니라 거의 모든 재부호화(recoding)에 탄력적으로 사용할 수 있다.
- 새롭게 만든 변수에는 적절한 변수 레이블(variable label)과 값 레이블(value label)을 부여한다.
- 만약 [같은 변수로 코딩변경(S)]을 수행하면 이전 변수 내용을 뒤바꾸게 되므로 주의해야 한다.

자료 변환

연습 3. income.sav 파일을 SPSS로 불러들여 소득(INCOM0)과 성별(SEX)을 앞서 제시된 기준대로 재부호화하시오. 변수 및 값 레이블을 적절히 부여하시오.

자료 변환

물론 범주형 변수를 가변수들로 변환할 수도 있다.

- 가령 5명의 사회경제적 지위(socioeconomic status; SES)를 아래와 같이 세 범주 (1=low; 2=middle; 3=high)로 입력하였다고 하자.

id	ses
1	low
2	middle
3	high
4	high
5	middle

- 이 변수를 쪼개 다음과 같이 더미 코딩할 수 있다:

  (1) "ses가 low이다"의 가변수(low)로 그렇다(1)/아니다(0).
  (2) "ses가 middle이다"의 가변수(middle)로 그렇다(1)/아니다(0).
  (3) "ses가 high이다"의 가변수(high)로 그렇다(1)/아니다(0).

자료 변환

- 사회경제적 지위(ses) 변수 하나를 3개의 가변수로 재부호화한 셈이다.

id	ses	low	middle	high
1	low	1	0	0
2	middle	0	1	0
3	high	0	0	1
4	high	0	0	1
5	middle	0	1	0

- 잘 보면 (어디든지) 한 줄은 결국 필요가 없다. 나머지 두 줄에서 얼마든지 추측이 가능하기 때문이다.
- SPSS에서 [변환]-[가변수 작성]을 활용하면 범주형 변수를 가변수로 간단히 바꿀 수 있다.

자료 변환

연습 4. hsb2.sav 자료를 SPSS에서 불러들여 인종 변수(race)의 자료유형을 파악하고 가변수로 재부호화하시오. 변수 및 값 레이블을 적절히 부여하시오.

자료 변환

응용통계학에서는 설문조사로 수집된 자료를 분석하는 경우가 많다.

- 최근 동향을 보면 데이터 수집방식이 훨씬 더 다원화되었지만 여전히 설문조사는 중요한 자료수집의 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(Why?), 역부호화(reverse-coding)가 필요하다.

자료 변환

※ 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

자료 변환

연습 5. poleff.sav 자료를 SPSS에서 불러들여 정치적 효능감(political efficacy)을 구성하는 모든 문항들을 적절히 재부호화 해보자. 재부호화한 자료는 다른 이름으로 저장하자.

자료 변환

원자료는 소중하며 때때로 복구 불가능함을 기억하자.

- 자료분석이 이루어지는 동안 새로운 변수를 추가, 삭제, 편집하는 등 원자료를 바꾸게 된다.
- 이때 만일 원자료를 조금이라도 수정했다면 덮어씌워 저장(overwrite)해서는 안된다.
- 파일 이름 뒤에 날짜를 부여하는 등 자신만의 습관을 만들고 백업(backup)을 늘 보관해야 한다.
- 필요에 따라 비망록 텍스트파일을 만들어 수정사항을 기록해 놓는 것도 필요하다.
- 직접 원자료를 수정하지 않고 구문(Syntax)을 작성·보관할 수도 있다.
- SPSS에서는 [확인]을 클릭하기 전에 [붙여넣기(P)]를 클릭하면 구문을 살펴볼 수 있다. SPSS 구문을 공부할 바에야 차라리 Stata, R, SAS 등을 배우는 편을 추천한다 (Why?).

자료 변환

Jamovi가 자료 변환에 편리한 도구는 아닌 것 같다.

- Jamovi 안에서는 [Data]에서 [Compute]와 [Transform]를 통해 자료 변환을 수행할 수는 있다. 그런데 사용자 편의성이 다소 부족한 것 같다.
- 차라리 엑셀에서 함수를 사용해 자료를 모두 정리 및 재부호화하고 따로 저장하자.
- 그리고나서 최종적인 분석만 Jamovi에서 하는 편이 시간 절약에 유리하다.

새 변수 만들기

새 변수 만들기

적절한 공식에 따라 계산하여 새로운 변수를 만들 수 있다.

- 연구 목적에 따라 두 개 이상의 변수를 사용해 새로운 변수를 만들 수도 있다. 그러한 상황을 상상해보자.
- SPSS에서는 [변환]-[변수 계산]을 통해 새로운 변수를 계산할 수 있다.
- 예전에 배운 [다른 변수로 코딩변경]으로는 할 수 없는 것들이다(Why?).
- 많은 변수와 관측치가 주어진 대형자료를 분석하다보면 무척 혼란스러우므로 변수의 순서를 바꾸거나 삭제하는 방법을 익히자.

새 변수 만들기

연습 6. college.sav 자료를 SPSS에서 불러들여 전임교원 1인당 학생수를 나타내는 변수를 만들어보자.

자료분석의 첫 출발

연습 7. poleff.sav 자료를 SPSS에서 불러들여 정치적 효능감을 측정하는 합성지수(composite index)를 구성해보자. 만일 하나의 문항만으로 정치적 효능감을 측정해야 한다면 그 문항은 어떤 것일까? 자료에서 주어진 각각의 문항은 어떤 척도로 볼 수 있을까? 모든 문항에 대해 전반적으로 긍정적으로 대답한다면 그 사람의 정치적 효능감을 어떻게 말할 수 있을까? 각 문항에 대한 응답 총점은 어떤 척도로 볼 수 있을까?

새 변수 만들기

많은 변수와 관측치가 주어진 대형자료를 분석하다보면 무척 혼란스러울 수 있다.

- 변수의 순서를 바꾸거나 삭제하는 방법을 익히자(그러나 원본은 항상 건드리지 않고 놔둔다).
- 변수의 이름을 바꾸는 법도 연습하자.
- 새 변수를 만든 다음에는 즉각 레이블(label)을 달아두지 않으면 잊어버린다.
- 측도 역시 적절히 설정해 두어야만 분석할 때 애로사항이 없다.