계량분석
Descriptive Statistics

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 12, 2024

진행 순서

1. 파일 관리와 변수 탐색
2. 기술통계 확인하기

파일 관리와 변수 탐색

파일 관리와 변수 탐색

많은 외부 자료들은 Stata 데이터 파일이 아니다.

- 일반적인 Stata 데이터 파일이라면 use 명령어로 불러올 수 있다.
- 현재 작업 중인 데이터가 이미 있다면 , clear 옵션으로 메모리를 먼저 치우고 시작해야 한다.
- Stata 데이터 파일 포멧이 아닌 경우라면 import 명령어를 사용해서 불러와야 한다!
- (SPSS나 SAS 등) 해당 소프트웨어 Stata 포멧으로 저장하는 것도 대안이 된다.

파일 관리와 변수 탐색

- 특히 CSV (comma-separated values)가 널리 사용되며, 이것들도 Stata로 불러와 고유의 데이터 포멧으로 바꾸어 저장할 수 있다.
- import delimited 명령어는 생각보다 연구와 실무 상황에서 무척 중요하다. help import delimited를 통해 여러가지 옵션에 대한 이해도를 충분히 높여야 한다!
- 특히 텍스트 속에 콤마(,)가 들어있고 이것이 delimiter/separator로 인식되면서 데이터 전체가 잘못 인식되는 경우가 있다. 외부파일을 불러온 다음에는 꼼꼼히 살펴보아야 한다.
- SPSS 데이터 파일도 자주 사용되고 import spss로 문제없이 불러와 다른 형식으로 바꾸어 저장할 수 있다(Stata 버전이 너무 낮으면 안됨).

파일 관리와 변수 탐색

어떤 데이터는 워낙 방대하여 원하는 변수를 골라내는게 쉽지 않다.

- describe 명령어는 데이터 속에 담긴 모든 변수의 이름과 레이블(label)을 보여준다.
- 와일드카드(* 또는 ?)에 대해서는 몇 번 연습해서 감각을 얻자.
- 변수 이름에 대한 단서가 부족할 때는 lookfor가 유용하다!
- Stata에서 변수명은 대소문자에 민감하므로(case-sensitive) 주의해야 한다. 예컨대 Happy와 HAPPY와 happy는 모두 다른 변수명으로 간주된다.
- 변수 이름 바꾸기(rename)도 중요하다.
- 변수 이름을 소문자로 바꾸는 편리한 옵션은 , lower이다.

파일 관리와 변수 탐색

분석과 무관한 변수나 관측치가 너무 많으면 거추장스럽다.

- keep과 drop같은 명령어로 하위표본(subsample)을 추출해 낼 수 있다.
- 40세 이상인 관찰값은 모조리 삭제하자.
- happy, marital, sex, incom0 이외의 변수도 모조리 삭제하자.
- order 명령어는 데이터 안에서 변수의 순서(order)를 바꾼다.

기술통계 확인하기

기술통계 확인하기

자료가 주어지면 일단 그 성격을 이해하고 변수들을 요약해야 한다.

- 학부 사회통계 시간에 배운 요약통계(summary statistics) 또는 기술통계(descriptive statistics)를 쉽게 살펴볼 수 있다.
- 행복(happy) 변수의 평균(mean)과 표준편차(standard deviation), 분산(variance), 최소값(minimum), 최대값(maximum)은 각각 얼마인가?
- 1사분위수(1st quartile), 2사분위수(2nd quartile) 또는 중위값(median), 3사분위수(3rd quartile)는 각각 얼마인가?

기술통계 확인하기

명목변수나 서열변수에 대해서는 다른 접근이 필요하다.

- tabulate 명령어는 변수 안에 들어있는 값들의 빈도분포표(frequency distribution table)를 보여준다
- 만일 변수 안의 값들에 레이블이 지정되어 있으면 이것을 보여주므로 , nolabel과 같은 옵션도 유용하다.
- 여기서 나오는 DK (“Don’t Know”)는 원래 값이 -8이므로 결측치(missing values)임을 시사한다.

기술통계 확인하기

기술통계표를 제대로 다시 만들기 위해 재부호화해보자.

- 행복(happy) 변수의 역부호화(reverse coding)을 해보자.
- replace 명령어가 하나하나의 값들에 대응해 재부호화(recoding)을 하는 반면, recode는 한 번에 원하는 값들의 대응관계를 설정할 수 있다.
- 만일 tabulate 뒤에 변수 두 개를 지정한다면 두 변수 간의 교차표 (cross-tabulation)를 보여준다.

기술통계 확인하기

요약통계는 그 자체로도 유용하지만 조건별로 보면 더 흥미롭다.

- 필요에 따라 bysort와 같은 접두어(prefix)를 사용하여 상이한 조건에 국한시켜 요약통계를 살펴볼 수도 있다.
- "혼인 상태에 따라 행복도가 서로 다를 것이다"라는 가설 아래 요약통계량을 살펴보자.
- 요약통계가 시사하는 바를 해석해보자.

기술통계 확인하기

- summarize 명령어는 if 조건문(conditional statement)과 결합하여 더 흥미로운 요약통계를 제시할 수 있다.
- 한발 더 나아가, 성별에 따라 혼인 상태와 행복도의 관계에 어떤 추가적인 차이가 나타나는지 살펴보자.
- 요약통계가 시사하는 바를 해석해보자.

기술통계 확인하기

- 혼인 상태는 부호화 기준(coding scheme)이 약간 복잡하므로 “함께인가 따로인가”로만 단순화 해보자.
- 새로운 변수의 이름은 together로 하자.
- 새로운 변수에 대해 레이블을 주기 위해 label 명령어를 적절히 활용하자.
- 새로 만들 레이블의 이름은 newmar로 하자.
- tabulate 옵션으로 , miss는 결측치(missing values)를 빼지 않고 보고해준다.

기술통계 확인하기

좀 더 깔끔한 기술통계표를 만들어보자.

- dtable 명령어가 매우 편리하므로 익숙해질 때까지 연습하자.
- 명목변수나 서열변수인 경우에는 변수 이름 앞에 i. 라는 접두어를 붙이고(e.g., i.marital), 등간변수나 비율변수는 그대로 입력한다.
- 이 표를 엑셀로 옮겨 적당히 꾸며보자.
- 그 밖에도 tabstat 등 여러 다양한 대안이 있으니 필요에 따라 사용하자.

기술통계 확인하기

- tabstat도 중요하다.
- sysuse auto.dta
- tabstat mpg, statistics(mean median sd skewness kurtosis cv iqr min max) by(foreign) columns(statistics)