계량분석
Introductory Stata (II)

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 13, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 다른 포멧의 파일 불러오기/저장하기
2. 원하는 변수를 찾아내 살펴보기
3. 요약통계 확인하기

다른 포멧의 파일 불러오기/저장하기

다른 포멧의 파일 불러오기/저장하기

많은 외부 자료들은 Stata 데이터 파일의 포멧(format)이 아니다(**[Stata 코드]** 참고).

- eCampus에서 sociology.csv와 KGSS.zip을 연습 파일로 다운로드 받자.
- 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 delimited**를 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 spss**로 문제없이 불러와 다른 형식으로 바꾸어 저장할 수 있다(Stata 버전이 너무 낮으면 안됨).

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

원하는 변수를 찾아내 살펴보기

원하는 변수를 찾아내 살펴보기

몇몇 데이터는 수많은 변수와 관측치를 가지고 있으므로 원하는 변수를 찾는게 쉽지 않다(**[Stata 코드]** 참고).

- **describe** 명령어는 데이터 속에 담긴 모든 변수의 이름과 레이블(labels)을 보여준다.
- 와일드카드(*, ?, -)에 대해서는 몇 번 연습해서 감각을 얻자.
- 변수 이름에 대한 단서가 전혀 없을때는 lookfor가 유용할 수도 있다!
- Stata에서 변수명은 case-sensitive하므로 주의해야 한다. 예컨대 Happy와 HAPPY와 happy는 모두 다른 변수명으로 간주된다.
- 변수 이름 바꾸기(**rename**)도 중요하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

원하는 변수를 찾아내 살펴보기

때때로 분석에 직결되지 않은 변수와 관측치가 너무 많으면 거추장스럽다 (**[Stata 코드]** 참고).

- **keep**과 **drop**같은 명령어로 서브샘플(subsample)을 추출해 낼 수 있다.
- 2018년 외의 관찰값은 모조리 삭제하자. happy, marital, sex 이외의 변수도 모조리 삭제하자.
- **order** 명령어는 데이터 안에서 변수의 순서(order)를 바꾼다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

원하는 변수를 찾아내 살펴보기

tabulate 명령어는 변수 안에 들어있는 값들의 빈도분포표(frequency distribution table)를 보여준다(**[Stata 코드]** 참고).

- 만일 변수 안의 값들에 레이블(label)이 지정되어 있으면 이것을 보여주므로 , **nolabel**과 같은 옵션도 유용하다.
- 여기서 나오는 DK ("Don't Know")는 원래 값이 -8이므로 결측치(missing values)임을 시사한다.

행복(HAPPY) 변수의 역코딩(reverse coding)을 해보자.

- **replace** 명령어가 하나하나의 값들에 대응해 리코딩(recoding)을 하는 반면, **recode**는 한 번에 원하는 값들의 대응관계를 설정할 수 있다.
- 만일 **tabulate** 뒤에 변수 두 개를 지정한다면 두 변수 간의 교차표(cross-tabulation)를 보여준다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

요약통계 확인하기

요약통계 확인하기

학부 사회통계 시간에 배운 요약통계(summary statistics) 또는 기술통계(descriptive statistics)를 쉽게 살펴볼 수 있다(**[Stata 코드]** 참고).

- 앞서 역코딩(reverse coding)한 행복(**happiness**) 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?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

요약통계 확인하기

요약통계는 그 자체로도 유용하지만, 상이한 조건에 국한시켜 볼 때 더욱 흥미로운 시사점을 제공한다(**[Stata 코드]** 참고).

- 필요에 따라 **bysort**와 같은 prefix를 사용하여 상이한 조건에 국한시켜 요약통계를 살펴볼 수도 있다.
- "혼인 상태에 따라 행복도가 서로 다를 것이다"라는 가설 아래 요약통계량을 살펴보자.
- 한발 더 나아가, 성별에 따라 혼인 상태와 행복도의 관계에 어떤 추가적인 차이가 나타나는지 살펴보자. 이를 위해 **if 조건문**을 활용하는 것도 대안이 된다.
- 요약통계가 시사하는 바를 해석해보자.
- Takeaway! **summarize** 명령어는 **bysort** 같은 prefix를 더 복잡하게 하거나 **if** 와 결합하여 더 흥미로운 요약통계를 제시할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

요약통계 확인하기

혼인 상태 코딩 스킴(coding scheme)이 약간 복잡하므로 "함께인가 따로인가"로만 단순화 해보자(**[Stata 코드]** 참고).

- 새로운 변수의 이름은 **together**로 하자.
- 새로운 변수를 만들었으니 레이블(label)도 부여하자. 새로 만들 레이블의 이름은 newmar로 하자.
- **tabulate** 옵션으로 **, miss**는 결측치(missing values)를 빼지 않고 보고해준다.

이제 다시 한 번 요약통계량을 살펴보자.

- 아래 결과는 무엇을 시사하는가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY