계량분석
Stata 입문

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 5, 2024

진행 순서

1. 선수과목 핵심요약
2. Stata 라이선스, 구동 및 설정
3. Stata 기초 명령어

선수과목 핵심요약

선수과목 핵심요약

사회통계(학부 기초통계)에서는 다음의 내용을 다루었다.

- 강의소개(제1주차)
- 기술통계(제2주차)
- 확률론(제3주차)
- 이산/연속확률분포(제4주차)
- 표집(제5주차)
- 통계추정(제6주차)
- 가설검정(제7주차)
- 중간시험(제8주차)
- 평균비교(제9주차)
- 분산비교(제10주차)
- 카이제곱분석(제11주차)
- 상관분석(제12주차)
- 회귀분석(제13주차)
- 회귀분석 조금 더(제14주차)
- 기말시험(제15주차)

* 충북대학교 사회학과 2학년 1학기 전공필수 사회통계 기준으로 주차를 나타냄.

선수과목 핵심요약

제1주차: 강의소개

- 강의소개
- 경험과학이란 무엇인가?
- 자료유형과 척도
- 엑셀의 설치와 기동

제2주차: 기술통계

- 기술통계란 무엇인가?
- 빈도분포표와 그래프
- 데이터의 요약(I): 중심성향
- 데이터의 요약(II): 산포성향

제3주차: 확률이론

- 확률이론의 기초
- 베이즈 정리
- 확률분포
- 누적확률분포

제4주차: 이산확률분포와 연속확률분포

- 이론적 확률분포들
- 대표적인 이산확률분포: 이항분포
- 대표적인 연속확률분포: 정규분포

선수과목 핵심요약

제5주차: 표집

- 모수를 통한 모집단의 추정
- 표본분포와 중심극한정리
- 표본평균의 표본분포
- 표본비율의 표본분포

제6주차: 통계추정

- 신뢰구간과 오차범위
- 모평균의 신뢰구간
- 모비율의 신뢰구간

제7주차: 가설검정

- 가설검정의 논리
- 모평균에 대한 가설검정
- 모비율에 대한 가설검정

제8주차: 중간시험

선수과목 핵심요약

제9주차: 평균비교

- t-분포와 자유도
- 단일표본과 2표본
- 독립표본 t-검정
- 쌍체표본 t-검정
- 등분산 가정에 관한 코멘트

제10주차: 분산비교

- 분산분석의 논리
- F-값과 t-값의 관계
- F-분포와 자유도

제11주차: 카이제곱분석

- 교차표
- 카이제곱-분포
- 기대값과 관찰값의 차이

제12주차: 상관분석

- 공분산과 상관계수
- 상관계수는 무엇이고 무엇이 아닌가
- 상관계수 그래프

선수과목 핵심요약

제13주차: 회귀분석

- 오차의 제곱의 최소화
- 회귀계수는 무엇이고 무엇이 아닌가
- 추정치의 표준오차
- 모형적합도
- 결과 보고

제14주차: 회귀분석 조금 더

- 다중회귀분석
- 제곱항과 그 해석
- 상호작용항과 그 해석
- 경로분석의 기초
- 회귀분석 가정에 관한 코멘트

제15주차: 기말시험

선수과목 핵심요약

회귀분석은 기초통계의 끝이자 중급통계의 시작이다.

- 시간이 지나면 기초통계를 잊는 것은 (엔트로피의 법칙처럼) 자연스러운 일이다.
- 복습은 이런 자연적인 힘을 인위적으로 거스르는 것이므로 몹시 괴롭다.
- 기초통계 교재를 공부하여 다시 떠올려야 한다.

Stata 라이선스, 구동 및 설정

Stata 구동 및 화면 설정

Stata에는 다양한 라이선스가 있다.

- Stata에도 다양한 버전(IC, SE, MP 등)이 있지만 일단 신경쓰지 말고 학교 것을 사용하자.
- 다행스럽게도 한국 벤더는 굉장히 유연하게 이 문제에 대응하고 있다.
- 나중에 기관에서 구매를 추진할 일이 있으면 (최소한 SE나) MP를 사도록 유도하자.

Stata 구동 및 화면 설정

Stata를 구동하고 화면을 설정하자.

- 먼저 여러 윈도우의 이름과 용도를 파악하면 두려울 것이 없다.
- 편한 위치로 윈도우를 옮겨서 자신의 취향에 맞도록 설정하자.
- 오래 바라보다 보면 눈알이 터질 것 같으므로 화면 색깔과 폰트 혹은 폰트 사이즈를 변경하자.
- 의외로 이걸 모르는 사람이 많은데 Stata는 SPSS처럼 메뉴를 사용한 분석 (point-and-click)이 가능하다.

Stata 기초 명령어

Stata 기초 명령어

대체 do 코드란 무엇이고 왜 중요한가?

- do 코드는 한 줄씩 순서대로 실행되는 일종의 배치 스크립트(batch scripts)이다.
- 메뉴에서 선택하거나 명령어 창(command line)을 사용하다보면 (그 순간에는 편리하지만) 자신의 데이터 관리나 분석 내용을 재현(replication)하기 몹시 불편하다. 게다가 타인과 협업하기란 불가능에 가까워진다.
- 반면 do 코드는 일단 한 번 짜놓으면 실행시킬 때마다 매번 같은 작업을 수행하게 된다.
- 조금 귀찮더라도 (1) 메뉴나 (2) 명령어 창보다는 (3) do 코드를 적극 활용하자.
- do 파일의 백업(backup)을 날짜별로 틈틈이 보관하며 폴더 관리도 체계적으로 해두자.

Stata 기초 명령어

최대한 보기 좋게 꾸며야 한다.

- Do-file Editor를 사용하여 초보적인 명령어를 연습해보자.
- do 코드를 짤 때는 일부러 들여쓰기(indentation) 할 것을 추천!
- 여백의 미를 충분히 활용할 것!
- 타인의 코드를 읽거나 읽혀야 할 일이 있고, 또 수 주/개월/년 뒤 자신의 코드를 되돌아볼 때도 있다(e.g, R&R 등).

Stata 기초 명령어

언제나 do 코드의 내용은 심미적으로 잘 관리하자.

- 각주달기(annotation)를 철저히! 자신을 단기기억상실증 환자인 것처럼 전제해야 한다.
- 레이블(label)이란 컴퓨터에게는 의미없지만 사람에게는 중요한 노트다.
- 한 줄의 레이블은 * 뒤에, 여러 줄의 레이블은 /* 과 */ 사이에 기록해 둔다.
- 명령어(command)나 command 뭉치마다 레이블을 달아서 자신의 의도를 전달한다.
- 자신이 그 안에 써넣은 명령어를 모두 기억할 수 있다고 믿어서는 안된다!

Stata 기초 명령어

언제나 do 코드는 실행하기 쉬운 상태로 유지하자.

- do 코드 안에서 오류가 일어나는 라인을 남겨두지 말자(Why?).
- 실행(run)의 단축키는 Ctrl-D이다. 전체를 실행시킬 수도 있지만 하이라이트한 부분만 실행할 수도 있다.
- 언제나, 언제나, 언제나 자신의 코드를 백업하자!
- 백업을 안해서 후회하는 경우가 생각보다 많다. 해서 후회하지는 않는다.
- 원자료(raw data)는 가능한한 건드리지 않는다(Why?). 모든 재부호화(recoding)은 do 코드 안에서 해결한다.
- 수업에서 일단 Stata를 공부하기로 한 이상 깔끔한 do 코드를 만드는게 무엇보다 남는 일이다.

Stata 기초 명령어

때때로 do 코드와는 별개로 분석의 결과물을 기록해 둘 필요가 있다.

- 사실 do 코드와 자료 파일을 직접 공유하는게 최선일 때가 많다.
- 그러나 보안상 이유로 데이터를 공유할 수 없지만 결과물은 보여줘야 할 때도 있다.
- 때로는 분석 자체에 너무 긴 시간이 걸려 로그(log) 파일을 보여주는 편이 빠를 때도 있다.
- 결과물 기록을 남기려면 log를 활용하자.

Stata 기초 명령어

다음의 순서대로 기초 명령어를 연습해보자.

- 먼저 do file에 나의 의도에 관한 레이블을 기록하자.
- 로그(log)를 켜기 위해 log using mylog를 입력하자.
- 내가 어느 폴더에 있는지 확인하기 위해 cd나 pwd를 입력하자. cd는 change directory의 머릿글자이고 pwd는 path of the (current) working directory의 머릿글자이다.
- 이 폴더에 있는 파일 목록을 살펴보기 위해 dir를 입력하자.
- 특정한 폴더로 이동하기 위해서 cd [folder name]를 활용하자.
- 마음에 드는 파일이 있다면 use [file name]를 통해 불러오자.

Stata 기초 명령어

당장은 데이터가 없으니 Stata에서 제공하는 웹 데이터를 살펴보자.

- webuse으로 미국의 클래식한 자동차 목록을 불러오자.
- 데이터 안의 변수 목록을 살펴보자.
- 새로운 변수를 만들어보고, 조건에 따라 이를 변경해보자. 변수를 지울 수도 있다.
- 새로운 파일로 저장해보고 로그를 종료하자.

Stata 기초 명령어

변수 조작은 매우 중요한 기법이므로 능수능란하게 수행할 수 있어야 한다.

- 다음을 간단히 연습해보자.
- (1) length와 weight을 곱한 새로운 변수를 만들자. 이름은 volume으로 하자.
- (2) price가 10,000보다 높으면 1, 그와 같거나 낮으면 0으로 하는 새로운 변수를 만들자. 이름은 자기 방식대로 하자.
- (3) price가 10,000을 넘고 외제인 차종을 나타내는 새로운 변수를 만들자. 이름은 luxury가 좋겠다.