

## 계량분석

## Course Introduction

김현우, PhD<sup>1</sup>

<sup>1</sup>충북대학교 사회학과 조교수

September 6, 2021

# 진행 순서

- 1 강사 소개
- 2 사회학과 계량적 분석
- 3 대학원 계량분석의 두 얼굴: Science와 Art
- 4 수업의 구성과 당부사항
- 5 통계분석 패키지의 선택
- 6 충북대 사회학과 커리큘럼과 이 수업

## 김현우

- Pennsylvania State University 사회학 박사학위 취득(2017)
- Edna Bennett Pierce Prevention Research Center 박사후 연구원 (2017-2019)
- Edna Bennett Pierce Prevention Research Center 연구조교수 (2019-2021)
- 충북대학교 사회학과 조교수 (2021.9-)

## 사회학과 계량적 분석

오늘날 통계적 리터러시가 갖는 중요성이 폭발적으로 증대하였다.

- 제4차 산업혁명의 기반기술(빅데이터, 인공지능, IoT 등) 전문가 수요
- 급속히 증가하는 데이터의 양: 오늘날 전 세계 데이터의 90%가 최근 2년 동안 생성
- 통계학은 인공지능, 빅데이터 분석, 기계학습(machine learning) 등의 기초

최근 10년 사이 사회학 방법론 분야에도 눈부신 발전이 있었다.

- (만약) 계량적 분석을 택하여 공부한다면 이를 필수적으로 따라가야 한다.
- 여러분의 졸업 이후에도 이 발전은 (더 빠르게) 계속 될 것이며 스스로 그 지식을 업데이트할 수 있어야 한다.

## 사회학과 계량적 분석

사회학에서 (고급) 계량적 분석은 이미 핵심적 주류다.

- ASR 2021년 8월호에 실린 7편의 논문 중 6편의 논문이, AJS 2021년 7월호에 실린 5편의 논문 중 3편의 논문이 계량적 분석을 사용하였다.
- 여러분의 세부전공과 상관없이 만약 계량분석을 공부하지 않으면 3/4 (=9/12)은 읽고 비판하지 못한다.

“에이, 그래도 나는 계량분석을 전공할 건 아닌데...”

- 물론 모든 사람이 계량적 분석 자체를 전공해야 하는 것은 아니다.
- 하지만 이 수업에서 다루는 토픽은 **선형회귀분석(linear regression analysis)의 기초와 그 응용**으로, 이는 사회학에서 쓰이는 모든 계량분석의 기초적인 뼈대를 구성한다.
- 다시 말해, 이 수업이 다루는 토픽은 기초일 뿐이고 전혀 수준높은 계량분석이 아니다.

# 대학원 계량분석의 두 얼굴: Science와 Art

사회통계학은 두 개의 얼굴을 가지고 있다.

- 과학으로서의 측면(Science)과 손기술로서의 측면(Art)

Science: 설명할 수 있는 힘

- 내가 지금 하고 있는 것이 무엇인지 직관적으로 명확하게 알아야 한다.
- 남에게 그리고 논문 속에서 자신의 분석을 확실하게 설명할 수 있어야 한다.

Art: 반복속련의 힘

- 손기술/연습은 생각보다 매우 중요하며 연습해두지 않으면 연구나 실무에서 전혀 쓸 수 없다.
- 투자한 시간이 절대적으로 많아야 한다. 머리 못지 않게 엉덩이로 공부한다!

# 대학원 계량분석의 두 얼굴: Science와 Art

## 수학에 관한 코멘트

- 통계학은 본디 수학의 분과학문이다. 수학을 제대로 하지 않고 건너뛰다는 것은 사실 넌센스다.
- 강의계획서와 강의안을 만드는 동안에도 사실 몇 번이나 마음을 바꾸었다. 여러 다른 교수님들과도 상의를 해보았다.
- 최종적으로 수학을 필요최소한만 하고 모두 건너뛰기로 하였다. **수학보다 직관을 강조**하기로 한 셈이다.
- 수학을 건너뛰는 것에는 장단점이 있다. 중요한 장점은 **사회통계학에 대한 흥미를 잃지 않게 돕는다**는 것이다.
- 그러므로 지금은 일단 직관적으로 공부한 다음, 나중에 점점 더 호기심이 생겨 계량분석을 한층 깊게 공부하고 싶다면 그때 가서 수학을 하자!

## 수업의 구성과 당부사항

선수과목: 학부 사회통계 및 사회조사방법론

- 만약에 기억이 잘 안나면 학부 시절의 교재와 노트를 다시 보아야 한다.
- 지금 학부에서 사회통계연습이 비대면 강좌로 개설되고 있다. 필요하다면 가서 듣자!
- 知之爲知之, 不知爲不知, 是知也.

교재는 필수가 아니지만...

- 자습용으로 중요하므로 반드시 하나 정도는 (구판/중고라도) 구입하는 것을 강추!
- 정확히 알고 구입하는 것이 아니라면 계량경제학이나 수리통계학과 같이 지나치게 수학적인 것은 피할 것.
- 전공은 전혀 상관없다(e.g., 경영통계학, 관광통계학 등).

강의계획서는 수업이 진행하면서 조금씩 계속 업데이트된다.

- <https://github.com/hxk271/Syllabi>





## 수업의 구성과 당부사항

## 수업에서 분석할 데이터 준비

- 이 수업을 가이드로 삼아 여러분은 스스로 관심있는 프로젝트를 직접 수행하여야 한다.
- 첫 출발점은 분석할 데이터를 준비하는 것!
- 먼저 관심가는 주제(경제사회학/환경사회학/조직사회학/노동문제 등)를 먼저 특정할 것. 그리고 나서 데이터를 고른다. 반드시 한국 데이터가 아니어도 된다.

## 어디에서 데이터를 확보할까?

- ICPSR: <https://www.icpsr.umich.edu>
- KGSS: <http://kgss.skku.edu>
- GSS: <https://gss.norc.org>
- KSDC: <https://www.ksdc.re.kr>
- 사업체패널: <https://www.kli.re.kr/wps>
- 한국노동패널: <https://www.kli.re.kr/klips>

## 수업의 구성과 당부사항

## 데이터 확보는 제법 중요한 일!

- 좋은 데이터 찾기에 시간을 많이 쏟아야 한다. 좋은 데이터를 얻는 것은 좋은 논문의 출발점.
- TITO (Trash In, Trash Out).
- 지도교수님과 반드시 이런 이야기를 나누고 조언을 받을 것!

이상적으로는 이 수업에 좋은 데이터를 들고와서 석박사 논문을 위한 분석을 바로 하는 것!

- 만일 논문 주제를 정하지 못했거나 데이터를 찾지 못할 것 같다면 석박사 논문과는 다른 주제의 (관심가는) 데이터라도 상관없다.
- 아무래도 이 수업을 위한 데이터를 찾을 수 없을 것 같다면, 최대한 빨리 강사와 상담할 것.
- 아무리 늦어도 중간시험 이전까지는 자기 데이터를 확보해야 한다.

## 수업의 구성과 당부사항

## 영어에 관한 코멘트

- 나도 안다. 처음엔 좀 어색할 수도 있다.
- 그런데 통계학은 결국 영어권에서 그 핵심적인 발전이 이루어진 분야. 좋건 싫건 이것이 오리지널.
- 자꾸 우리말과 엮어서 반복할 예정! 결국엔 둘 다 기억해야 함.

필요하다면 **한국통계학회 통계용어집**을 이용할 것.

- [http://www.kss.or.kr/bbs/board.php?bo\\_table=psd\\_sec](http://www.kss.or.kr/bbs/board.php?bo_table=psd_sec)

좋은 소식도 있다!

- 영어가 훨씬 직관적이라 통계학 속에 담긴 진짜 의미나 취지를 오히려 알아듣기 쉽다.
- 게다가 우리말이라고는 해도 순우리말이 아니라 일본식 한자가 많다.

# 통계분석 패키지의 선택

다섯 개의 강력한 대안들: SPSS, SAS, Stata, R, Python

가격 측면

- SPSS, SAS는 일단 아웃. 개인은 도저히 감당할 수 없고 반드시 기관이 구입해야 한다.
- Stata는 개인이 살짝 부담이긴 한데 살 수는 있음. 물론 기관은 부담없이 구입해 줄 수 있다.
- R과 Python은 무료.

## 러닝커브(learning curve) 측면

- R과 Python은 러닝커브가 가파르다. 언어 자체를 배우는데 시간을 꽤 써야한다. 일단 습득하면 가장 많은 것들(머신러닝/웹스크래이핑/빅데이터/GIS/SNA/베이스통계 포함)을 할 수 있다!
- Stata과 SAS는 러닝커브가 살짝 있지만 한 학기 안으로 할 만하다. 요즘엔 베이스통계나 머신러닝 등을 조금 다룰 수 있지만 R과 Python만큼 강력하지는 않다.
- SPSS는 러닝커브가 거의 없다. 구문(syntax)를 배우려면 약간 시간이 걸리긴 한다. 그런데 제약이 제법 많다(예컨대 parametric survival analysis 등).



## 통계분석 패키지의 선택

## 왜 결론은 Stata인가?

- 일단 SAS는 내가 모름. 미안합니다~
- SPSS는 현업(조사기관)에서 많이 쓰이지만 그 밖에서는 잘 안쓰임. 게다가 통계학을 알면 쉽게 독학할 수 있다.
- Python과 R은 머신러닝/웹스크래이핑/빅데이터 등에서 강력하므로, 학생들을 위해서는 이쪽의 전망이 낮지만 두 개의 단점: (1) 러닝커브 때문에 포기자가 나옴; (2) 분석 노가다(!)할 때 Stata가 좀 더 편리함.
- 미국 Penn State, UPenn, Notre Dame 등에서도. 보통 1학점 짜리 랩에서 공부.



## 외부 특강과의 비교

- 많은 학생들이 통계분석 패키지를 공부하고 논문 쓰기 위해서 대학 바깥에서 특강을 듣는다.
- KOSSDA 외부 특강: <https://kossda.methods.snu.ac.kr/>
- 이 수업은 특히 “기초통계학” 및 “중급통계학”과 같은 수요에 대응한다.

## 충북대 사회학과 커리큘럼과 이 수업

충북대 사회학과 학부에서 개설할 계량분석 관련 수업들

- **사회통계**(2학년): 통계학의 논리적 기초 중심. Excel
- **사회통계연습**(2학년): 실습 중심. SPSS (비대면이면 JASP) 또는 R.
- **소셜데이터분석**(3학년): 파이썬 입문, 기계학습(machine learning), 웹 자료수집(web scraping), 커뮤니티 매핑(community mapping), 텍스트 분석(text analysis)

## 충북대 사회학과 대학원에서 개설할 계량분석 관련 수업들

- **계량분석1**(대학원): 이 수업
- **계량분석2**(대학원): 범주형 자료분석(Categorical Data Analysis). 계량분석1이 선행수강.
- **고급사회통계세미나**(대학원): 수요에 따라 종단자료 분석(Longitudinal Data Analysis), 다층모형 분석(Multilevel Modeling), 사건사분석(Event-History Analysis), 소셜네트워크분석(Social Network Analysis), 공간회귀분석(Spatial Regression Analysis) 약간, 비실험적 인과분석(Causal Inference with Observational Data) 등.

## 충북대 사회학과에서 개설되지 “않는” 계량분석 관련 수업들

- **데이터베이스**: 데이터의 보관 및 접근, 시스템(DBMS) 구축(SQL 포함), 데이터 엔지니어링 이슈(특히 클라우드 컴퓨팅)
- **서베이 방법론**: 확률표집원리, 조사 및 문항설계, 조사품질관리, 패널관리, 가중치 및 결측치 핸들링, 온라인 서베이 등