계량분석
Examining Data Archives

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 27, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 통계 데이터(Statistical Data)
2. 데이터 아카이브(Data Archive)

통계 데이터(Statistical Data)

통계 데이터(Statistical Data)

통계 자료(statistical data)는 직접 수집한 1차 자료와 남이 수집한 2차 자료(secondary data)로 구분할 수 있다.

- 편의표본(convenient sample)에 근거하여 자기가 직접 1차 자료를 수집한 뒤, 이를 분석하여 논문을 쓰는 경우도 제법 많다.
- 증거 자료의 체계적인 수집과 정리, 관리도 하나의 전문적인 경험과학자의 역할이라고 언급하였다. 이에 요구되는 전문적인 능력, 자본, 시간 없이 수집된 1차 자료는 현실적으로 기준 미달인 경우가 많다.
- 이러한 데이터에 의존한 발견과 결론은 일반화(generalization) 될 수 없고 그만큼 가치가 떨어진다.

그러나 1차 자료의 수집(과 공개) 자체가 훌륭한 연구(의 일부)로 인정받는 경우도 있다.

- 내용분석(content analysis)을 수행하는 경우에는 대부분 직접 1차 자료를 수집해야 한다.
- 설문조사(survey)의 경우에도 2차 자료로는 존재하지 않는 숨겨진 모집단(hidden population)을 연구한다면 직접 1차 자료를 수집해야 한다.
- 다만 데이터를 직접 수집하다보면 석박사 과정의 기간이 (아무리 적게 잡아도) 최소한 1년은 늘어난다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

통계 데이터는 수집 절차에 따라 다음과 같이 나뉜다.

- 실험(experimental) 데이터
- 비실험(non-experimental; observational) 데이터
- 모의실험(simulation) 데이터

많은 사회과학 연구는 주로 비실험 데이터를 다루어왔다.

- 그러나 이것은 실험 데이터와 모의실험 데이터를 고려할 필요가 없다는 의미는 아니다.
- 다만 모의실험 데이터는 수리사회학과 수리통계학의 영역이므로 기회가 닿는다면 다른 수업에서 다루기로 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

실험 데이터와 비실험 데이터는 다르다.

- 실험(experiment)은 상대적으로 소수의 참가자들(subjects)을 선발한다. 비실험 데이터는 상대적으로 다수를 담고 있다.
- 실험 참가자들은 무작위(random)로 처방집단(treatment group)과 통제집단(control group)으로 나뉜다. 비실험 데이터에서도 처방집단과 통제집단이 나뉘지만 여기에는 무작위라는 보장이 없다.
- 실험 결과, 실험집단과 통제집단 사이에서 나타나는 차이는 처방의 인과적 효과(treatment effect)로 귀속된다. 비실험 데이터의 경우 그렇지 않다.
- 실험의 결과(results)가 실험 참가자들 이외의 사람들에게 일반화(generalization)될 수 있는가는 다소 애매하다. 그렇기 때문에 메타분석(meta analysis)이 필요하다. 무작위 표본(random sample)을 사용한 비실험 데이터는 적어도 상관관계의 일반화는 잘 된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

실험 데이터에서 무작위화(randomization)와 비실험 데이터에서 무작위 표본(random sample)은 분명히 다르다.

- 실험 데이터의 randomization는 어떤 subject가 treatment group과 control group 중 "오로지 우연에 의해서" 어느 쪽에 속할지 결정된다는 의미이다.
- Randomization 덕택에 실험 연구(experiment)에서 처방(treatment)과 효과(effect) 사이에 인과관계(causal relations)가 인정된다.
- 비실험 데이터에서 무작위 표본(random sample)이란 모집단(population)의 구성원이 "오로지 우연에 의해' 서' 선발된 표본(sample)을 뜻한다. 단순 무작위 표본(simple random sample)에서 population의 구성원이 표본으로 뽑힐 확률은 동일하다.
- Random sample 덕택에 표본 안에서의 분석 결과는 모집단 전체로 일반화 될 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

통계 데이터는 수집 방식에 따라 다음과 같이 나뉜다.

- 횡단면(cross-sectional) 데이터
- 시계열(time-series) 데이터
- 패널(panel) 데이터
- 반복된 횡단면(repeated cross-sectional) 데이터

많은 사회과학 연구는 주로 횡단면 데이터를 다루어왔다.

- 그러나 이것은 시계열, 반복된 횡단면, 그리고 패널 데이터를 고려할 필요가 없다는 의미는 아니다.
- 다만 그것들은 난이도가 높으므로 기회가 닿는다면 다른 수업에서 다루기로 한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

횡단면(cross-sectional) 데이터는 "주어진 시간대"에 여러 분석대상(subjects)에 관해 조사한 자료이다.

- 분석대상은 사람, 고양이, 암세포 밀도, 단어의 수 등 연구 목적에 따라 다양하다.
- 이 데이터가 수집되는 동안 시간 변화에 따른 차이는 고려되지 않는다.

시계열(time-series) 데이터는 "시간이 경과함에 따라" 하나의 분석대상의 변화하는 특성을 조사한 자료이다.

- 대표적인 예로 일별 주가지수나 연간 강수량 등을 생각해 볼 수 있다.
- 이 데이터에서 핵심은 시간이 흐르는 동안의 변화가 계속 기록된다는 점이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

패널(panel) 데이터는 여러 분석대상에 관해 “시간이 경과함에 따라” 추적하여 반복 조사한 자료이다.

- 여기서 중요한 것은 추적하였다는 점이다. 다시 말해, “여러 분석대상”은 시간 경과에 따라 재방문하여 재조사되었다.
- 패널 데이터에서 “분석대상”이 반드시 사람이여야 하는 것은 아니다. 동네(town)나 국가(country)와 같은 장소일 수도 있고 사건(event)일 수도 있다.

반복된 횡단면(repeated cross-sectional) 데이터는 여러 분석대상에 관해 “시간이 경과함에 따라” 추적하지는 않고 반복 조사한 자료이다.

- 추적이 이루어지지 않았으므로, “여러 분석대상”은 조사 시점에 따라 매번 다르다 (우연히 같은 사람이 걸렸을 수는 있다).

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

통계 데이터(Statistical Data)

통계 데이터는 여러 포멧(formats)을 취할 수 있다.

- 분석에 투입되기 이전의 순간 데이터는 반드시 숫자의 나열(수열) 형태를 갖추어야 한다.
- 하지만 아직 수열로 전환되기 이전의 자료도 일종의 데이터로 볼 수 있다. 가령 사진, 그림, 비디오, 오디오, 문서/일기/신문, 파일, 움직임에 관한 기록, 질적 자료 (qualitative data), 심지어 통계 데이터에 관한 설명 그 자체도 데이터가 될 수 있다.
- 우리는 pre-processing 이라는 절차를 통해 위와 같은 데이터를 수열의 꼴로 바꿀 수 있다. 물론 pre-processing 절차 그 자체도 배워야 한다.
- 데이터의 정의는 오로지 자신의 상상력의 한계에 의해서만 제약받는다.

이 수업에서는 pre-processing을 생략하기 위해 수열의 꼴로 주어진 데이터를 바로 분석한다.

이렇게 분석에 ready-to-go 상태에 있는 데이터를 데이터셋(dataset) 이라고도 부른다. 현실적으로 연구자들 사이에서 data와 dataset을 그리 엄격하게 구분하지는 않는다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

데이터 아카이브(Data Archive)

경험적 사회과학이 성장하고 가정용 컴퓨터(PC)가 보급된 1990년대 이후의 시점에서 데이터 아카이브(data archive)가 탄생했다.

- 자신이 수집한 자료를 한 번 쓰고 버리기보다 남들도 연구와 교육을 위해 쓸 수 있도록 배려한다.
- 경험과학의 성격상 자료 수집 자체에 전문화된 사람들이 다양한 주제로 데이터를 수집하여 공개하거나 판매하였다.
- 데이터 아카이브의 출현은 계량적 사회과학(quantitative social sciences)의 출현과 발달에 가장 중요한 인프라였다.
- 어떤 조직이 데이터 아카이브를 운영하고 통제하는 것은 말하자면 그 조직이 계량적 사회과학 계에서 기축통화를 가졌다는 말과 같다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

종합적인 주제를 모두 커버하는 데이터 아카이브와 특수한 주제만을 커버하는 데이터 아카이브로 나뉜다.

- 전세계에서 명실상부 가장 대표적인 종합형 데이터 아카이브는 Inter-university Consortium for Political and Social Research로 이른바 ICPSR (https://www.icpsr.umich.edu)라고 불리운다.
- Pew Research Center (https://www.pewresearch.org)
- Roper Center (https://ropercenter.cornell.edu)
- Harvard Dataverse (https://dataverse.harvard.edu)
- 우리나라에서 대표적인 종합형 데이터 아카이브는 2021년 현 시점에서 아마도 한국사회과학자료원(https://kossda.snu.ac.kr)과 한국사회과학데이터센터 (https://www.ksdc.re.kr)인 것 같다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

- 특수주제형 데이터 아카이브는 수가 무척 많고 여기저기 흩어져 있어서 자기 전공 분야만 잘 아는 경우가 많다.
- 경제 데이터의 경우 National Bureau of Economic Research (NBER)가 유명하다 (https://www.nber.org/research/data).
- 종교 데이터의 경우 The Association of Religion Data Archives (ARDA)가 유명하다(https://thearda.com).
- 경영/금융 데이터의 경우 Wharton Research Data Services (WRDS)가 유명하지만 기관 라이선스가 필요하다.
- 인구 데이터의 경우 Social Explorer가 제법 유명하고 편리하지만 기관 라이선스가 필요하다.

시간을 많이 들여서 관심에 부합하는 데이터 아카이브를 발굴하고 자주 살펴보는 습관이 필요하다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

여기서는 여러 데이터 아카이브 가운데 한국사회과학데이터센터(KSDC)를 방문해 보기로 하자.

- 현재 충북대학교 도서관은 홈페이지 관리 부실로 KSDC 링크가 깨진 상태를 내버려두고 있다.
- 교내라면 https://ksdcdb.kr를 입력해 직접 들어갈 수 있다. 교외의 경우 교외 접속을 하고 직접 입력해 따로 들어가야 한다.

여기서 "[2699] 마스크 및 사회적 거리두기에 대한 대국민 인식조사"를 다운받자. 원자료 뿐 아니라 코드북(codebook)이나 설문지(questionnaire)도 함께 다운받아야 한다(**[Stata 코드]** 참고).

- 먼저 설문지를 쭉 살펴보면서 어떤 문항들이 있는지 살펴보자. 무슨 변수들이 독립변수(independent variable)로, 또 종속변수(dependent variable)로 어울릴지 상상해보자.
- 무슨 이론이나 가설이나 두 변수 사이의 관계에 대해 어떤 식으로 시사하나?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

- Stata가 구버전이라서 **import spss** 명령어를 사용할 수 없으면 SPSS를 기동시켜 "새로 저장하기" 기능을 활용해 Stata 파일로 저장하자.
- "문5. 귀하는 아래의 대상 및 목적으로 마스크를 수출하는 것에 대해 어떻게 생각하십니까?"를 살펴보자. 이 안의 다섯개 문항은 마스크 수출 반대라는 태도/의견의 여러 차원들(dimensions)을 측정하고 있다. 각각의 변수에 대해 빈도분포표(frequency distribution table)를 살펴보고 해석해보자.
- 마스크 수출 반대라는 개념의 차원을 이론적으로 제대로 측정하고 있는가는 지금 당장 고민하지 말자. 그저 연습삼아 이들 변수를 모두 더해 하나의 변수를 만들어보고 그것의 빈도분포표를 살펴보자. 이 새로운 변수에서 큰 값을 무엇을 의미하나?
- 새로운 변수(**oppose**)의 히스토그램(histogram)을 그려보자.
- **oppose**과 또다른 변수인 응답자의 연령 간의 연관성을 산포도(scatterplot)로 간단히 살펴보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

데이터 아카이브(Data Archive)

이 데이터에서 “문4. 귀하는 마스크 5부제에 대해 어떻게 평가하십니까?”를 보자(**[Stata 코드]** 참고).

- 개별 변수들의 frequency distribution table을 그려보자. 해석해보자.
- 이들 변수들을 역코딩(reverse coding)하여 새로운 변수를 만들어보자. 이제 이 변수들을 해석해보자.
- 연습삼아 역코딩한 변수들을 모두 더하여 하나의 변수로 만들어보자. 이 값이 크다는 것은 무엇을 의미하는지 해석해보자.
- 이 변수의 histogram을 그려보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY