계량분석
Data Archives

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 19, 2024

진행 순서

1. 1차 자료와 2차 자료
2. 실험 자료와 관찰 자료
3. 횡단면 자료, 시계열 자료, 그리고 종단 자료
4. 데이터 아카이브

1차 자료와 2차 자료

1차 자료와 2차 자료

통계자료는 1차 자료와 2차 자료로 구분할 수 있다.

- 편의표본(convenient sample)에 근거하여 자기가 직접 1차 자료(primary data)를 수집한 뒤, 이를 분석하여 논문을 쓰는 경우도 제법 많다.
- 증거자료의 체계적인 수집과 정리, 관리도 하나의 전문적인 경험과학자의 역할이다.
- 그러나 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.

1차 자료와 2차 자료

- 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.
- 설문조사의 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.
- 2차 자료(secondary data)가 현실적인 대안이 된다.

1차 자료와 2차 자료

- 최근에는 데이터 사이언스의 영향력이 커지면서 연구자 스스로 온라인에서 1차 자료를 수집하는 경우가 늘어났다.
- (Python이나 R과 같은) 전문적인 코딩 언어를 학습해야 한다. 웹 스크레이핑(web scraping)을 배워야 한다.
- 연구 주제에 그에 걸맞아야 한다(e.g., 온라인 사회운동 등).
- 행정자료(administrative data)를 직접 분석하는 경우가 크게 늘어났다. 가령 소득불평등을 연구한다면 국세청이 직접 생산한 자료를 바로 분석하는 것이 가장 이상적이다(Why?).
- 행정자료는 여러모로 표본(sample)과는 성격이 크게 다르다. 어떤 의미에서 이것은 모집단(population)에 가깝다.

실험 자료와 관찰 자료

실험 자료와 관찰 자료

전통적으로 통계 자료는 연구설계에 따라 다음과 같이 나눌 수도 있다.

1. 실험 자료(experimental data)
2. 비실험 자료(non-experimental data) 또는 관찰 자료(observational data)
3. 모의실험 자료(simulation data)

실험 자료와 관찰 자료

실험 자료와 비실험 자료의 차이를 명확히 이해해야 한다.

- 실험 자료는 상대적으로 소수의 참가자들(subjects)을 선발(recruitment)한다.
- 관찰자료는 대체로 많은 사람을 조사한다. 최근에는 실험에서도 상대적으로 다수를 선발하는 추세가 있다.
- 실험 자료에서 참가자들은 광고 등의 매체를 통해 초대되는 것이 보통이다.
- 관찰자료에서는 무작위 표집(random sampling)을 강조한다. 여기서는 조사대상을 초대할 때 나타나는 선택편의(selection bias)의 문제를 엄격히 경계한다(Why?).

실험 자료와 관찰 자료

- 실험 참가자들은 무작위(random)로 처방집단(treatment group)과 통제집단(control group)으로 나뉜다.
- 비실험 자료에서도 물론 처방집단과 통제집단이 나뉘지만, 이때 두 집단의 구분이 무작위라는 보장이 없다(Why?).
- 실험의 경우, 실험집단과 통제집단 사이에서 나타나는 차이는 처방효과(treatment effect)로 귀속된다. 비실험 데이터의 경우 그렇다는 보장이 없다(Why?).

실험 자료와 관찰 자료

- 실험에서 얻은 결과(results)가 실험 참가자 이외의 사람들에 대해서도 일반화(generalization)될 수 있는가 확신할 수 없다.
- 이러한 목적을 위해서는 메타분석(meta analysis)이 필요하다(Why?).
- (무작위 표집하였다면) 관찰자료를 분석하여 얻은 결과는 조사대상자 이외의 사람들에 대해서도 일반화될 수 있다.

실험 자료와 관찰 자료

실험 자료와 관찰자료에서 무작위성 또는 임의성은 다른 의미로 통용된다.

- 실험 자료에서 무작위화(randomization) 또는 임의할당(random assignment)와 관찰자료에서 임의표본(random sample)을 확실히 구분해야 한다.
- 실험 자료에서 임의할당은 어떤 실험 참가자가 처방집단과 통제집단 중 어느 쪽에 속할지 “오로지 우연에 의해서” 결정된다는 뜻이다.
- 임의할당 덕택에 실험연구에서 두 집단 사이의 차이를 처방효과라고 볼 수 있게 된다. 즉 상관관계(correlations)를 넘어 인과관계(causal relations)를 발견할 수 있다.
- 다만 임의표집은 아니므로 모집단 전체로 일반화를 보장할 수 없다.

실험 자료와 관찰 자료

- 관찰자료에서 임의표본(random sample)이란 모집단의 구성원이 “오로지 우연에 의해서” 선발된 표본임을 뜻한다.
- 단순임의표집(simple random sampling)에서는 모집단의 구성원이 표본으로 뽑힐 확률이 같다.
- 임의표집 원리 덕택에 표본 안에서의 분석 결과는 모집단 전체로 일반화 될 수 있다.
- 다만 임의할당이 아니므로 분석 결과를 통해 얻은 연관성이 인과관계를 보장하지 못한다.

실험 자료와 관찰 자료

(심리학을 제외한) 많은 사회과학 연구에서는 대체로 관찰자료를 사용한다.

- 전통적인 사회과학은 실험이 불가능하거나 매우 어려운 주제인 경우가 많다.
- 너무 여기에 집착해서 실험 자료와 모의실험 자료를 아예 고려하지 않는 것은 위험하다.
- 실험 자료의 분석은 비교적 근래에 사회과학 분야에서 다시 한 번 크게 주목받고 있다.
- 관찰자료도 자연실험(natural experiments)의 상황에서는 실험 자료처럼 분석될 수도 있다(Why?).
- 모의실험(simulation)은 수리사회학(mathematical sociology)의 영역이고 이 분야는 (사회학 안에서는) 마이너한 분야라서 제대로 배우기 어렵다.

횡단면 자료, 시계열 자료, 그리고 종단 자료

횡단면 자료, 시계열 자료, 그리고 종단 자료

통계자료는 수집 시기에 따라 다음과 같이 나눌 수 있다.

1. 횡단면 자료(cross-sectional data)
2. 시계열 자료(time-series data)
3. 종단 자료(longitudinal data) 또는 패널 자료(panel data)
4. 반복된 횡단면 자료(repeated cross-sectional data)

횡단면 자료, 시계열 자료, 그리고 종단 자료

(경제학을 제외한) 많은 사회과학 연구는 주로 횡단면 자료를 다루어왔다.

- 횡단면 자료는 "주어진 시간대"에 "여러" 분석대상을 조사하여 얻는다.
- 관찰단위(unit of observations)는 사람, 국가, 암세포 밀도, 단어의 수 등 연구 목적에 따라 다양하다.
- 이 데이터가 수집되는 동안 시간 변화에 따른 차이는 고려되지 않는다.
- 횡단면 자료는 다른 자료에 비해 상대적으로 분석이 용이하므로 가장 먼저 배우게 된다.

횡단면 자료, 시계열 자료, 그리고 종단 자료

- 시계열 자료는 “주어진 대상’에 대하여 “여러” 시간을 조사하여 얻는다.
- 대표적인 관찰단위는 일별 주가지수나 연간 강수량 등을 생각해 볼 수 있다.
- 대상은 오로지 하나일 뿐이다. 가령 주가지수나 강수량 등은 하나의 측정대상이 된다.

횡단면 자료, 시계열 자료, 그리고 종단 자료

- 패널 자료는 “여러” 분석대상에 관해 “여러” 시간동안 추적 조사하여 얻는다.
- 패널 자료에서 관찰단위가 반드시 사람이여야 하는 것은 아니다. 동네(town)나 국가(country), 심지어 사건(event) 일 수도 있다.
- 여기서 중요한 것은 여러 분석대상이 시간 경과에 따라 각양각색으로 변화하는 모습을 “추적하여” 조사했다는 점이다. 그러므로 매 시간대 같은 분석대상이 조사된다.

횡단면 자료, 시계열 자료, 그리고 종단 자료

- 반복된 횡단면 자료는 데이터는 “여러” 분석대상에 관해 “여러” 시간동안 조사하여 얻지만, 같은 사람을 추적하지는 않은 경우에 속한다.
- 추적이 이루어지지 않았으므로 ‘여러 분석대상은 조사 시점에 따라 매번 다르다 (우연히 같은 사람이 걸렸을 수는 있다).
- 가령 사회조사가 매년 1,000명 씩 조사한다고 하자. 같은 사람을 (이사하더라도) 쫓아다니면서 조사한다면 패널자료가 되지만, 매해 다른 사람을 조사하면 반복된 횡단면 자료가 된다.

데이터 아카이브

데이터 아카이브

데이터 아카이브는 사회과학의 경험적 연구에서 보물창고와도 같다.

- 데이터 아카이브(data archive)의 출현은 계량적 사회과학(quantitative social sciences)의 출현과 발달에 가장 중요한 인프라였다.
- 경험적 사회과학이 성장하고 가정용 컴퓨터(PC)가 보급된 1990년대 이후의 시점을 고려해야 한다.
- 자신이 수집한 자료를 한 번 쓰고 버리기보다 남들도 연구와 교육을 위해 쓸 수 있도록 한다는 취지이다.
- 기존 연구의 재현(replications)을 위하여 기성자료를 공유하는 경우도 있다.
- 어떤 기관/대학이 데이터 아카이브를 운영하고 통제하는 것은 말하자면 그 곳이 계량적 사회과학 계에서 기축통화를 운영한다는 말과 같다(Why?).

데이터 아카이브

데이터 아카이브에도 종류가 다양하다.

- 데이터 아카이브에 따라 종합적인 주제를 모두 커버하기도 하고 특수한 주제만을 다루기도 한다.
- 전세계에서 명실상부 가장 대표적인 종합형 데이터 아카이브는 Inter-university Consortium for Political and Social Research (ICPSR)이다.
- Pew Research Center
- Roper Center
- 재현 자료는 Harvard Dataverse에 많다.
- 우리나라에서 대표적인 종합형 데이터 아카이브는 2024년 기준 한국사회과학자료원(KOSSDA)과 한국사회과학데이터센터(KSDC)이다.

데이터 아카이브

- 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.
- 종교 데이터의 경우 The Association of Religion Data Archives (ARDA)가 유명하다.
- 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.

데이터 아카이브

한국사회과학자료원(KOSSDA)에 가서 적당한 자료를 간단히 살펴보자.

- 교내라면 https://kossda.snu.ac.kr를 입력해 곧바로 들어갈 수 있다.
- 교외의 경우 학교 도서관에서 교외 접속을 하고 따로 링크를 따라가서 들어가야 한다.
- 여기서 '성평등한 정치대표성 확보를 위한 정치인식조사, 2020'를 다운받자.
- 원자료(raw data) 뿐 아니라 설문지(questionnaire)도 함께 다운받아야 한다.

데이터 아카이브

- 먼저 설문지를 쭉 살펴보면서 어떤 문항들이 있는지 살펴보자.
- 무슨 변수들이 독립변수로, 어떤 변수들이 종속변수(dependent variable)로 어울릴지도 상상해보자.
- 원칙대로라면 명확한 이론과 가설에 의지하여 변수를 선택해야 한다.
- 현실적으로는 2차 자료에 의존하게 되므로 변수 쇼핑(variable shopping)과 어느 정도 타협할 수 밖에 없다.

데이터 아카이브

- 인구통계학적 변수들의 기술통계표(descriptive statistics table)를 만들어보자.
- 이때 인구통계학적 변수들은 성별, 연령(대), 거주지역, 거주지역 규모, 최종학력, 월평균 가구소득 등을 포함하자.
- 관심있는 변수에 대해 빈도분포표(frequency distribution table)를 살펴보고 간단히 해석해보자.

데이터 아카이브

- 조사참여자 별로 사회적 변화를 위해 참여한 활동의 갯수를 계산해보자. 이때 오프라인 방식과 온라인 방식을 나누어 계산해보자.
- 두 변수의 빈도분포표를 살펴보고 이 새로운 변수들에서 큰 값을 무엇을 의미할지 생각해보자.
- 새로운 합성지수들의 히스토그램(histogram)을 그려보자.
- act1과 응답자의 출생연도 간의 연관성을 산점도(scatterplot)로 살펴보고, act2와 출생연도 간 산점도도 그려보자.

데이터 아카이브

- 정치적 효능감(political efficacy)를 측정하기 위해 A2_1와 A2_2의 평균을 외적(external) 효능감으로, A2_3을 내적(external) 효능감으로 사용하자.
- 필요에 따라 변수들을 역부호화(reverse coding)하여 적절히 새로운 변수를 만들어보자.
- 성별에 따라 두 정치적 효능감 변수의 차이를 간단히 비교해보자.
- 다음 기사([링크])를 읽고 이를 직접 재현해보자.