계량분석
Theory-Based Data Analysis

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

September 13, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 사회과학철학의 맥락
2. 경험적 사회과학 연구의 단계
3. 측정오류
4. 모집단과 표본, 그리고 분석단위
5. 상관관계와 인과관계

사회과학철학의 맥락

사회과학철학의 맥락

무엇이 과학적 지식인가?

- 감각적 지각을 활용하여 체계적인 관찰에서 얻은 증거에 의해 지지되는 사실관계
- 논리적으로 검증가능한 명제: "오각형의 내각의 합은 삼각형 세 개의 내각의 합과 같다."
- 경험적으로 검증가능한 명제: "지구는 달보다 크다."

무엇이 과학적 지식이 아닌가?

- 초자연적이고 형이상학적 명제들
- 신학적 명제: "신은 존재한다."
- 윤리적 명제: "사람은 약속을 지켜야 한다."
- 미적 명제: "나의 외모는 아름답다."
- 물론 이와 같은 비경험적 명제를 경험적 명제로 뒤바꾸어 경험과학적 연구 주제로 삼을 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

사회과학철학의 맥락

경험과학이란 무엇인가?

- 경험주의(empiricism)에 기초하고 있는 과학
- 경험주의는 경험을 지식의 유일한 원천으로 보는 이데올로기다.
- 경험과학자는 경험적으로 검증가능한 명제를 생산하는 것을 과업으로 삼는다.
- "경험에 의한 검증가능성(verifiability)"이 생명이다(어원: Verus/참)

경험과학에서는 귀납추론(inductive reasoning)이 보다 중요하다.

- 개별 사실에 대한 경험을 토대로 증거를 수집하여 이(=data)를 관통하는 (가능한 최대한) 일반화된 법칙을 정립하는 방향으로(구체에서 추상으로) 움직인다.
- 반면 연역추론(deductive reasoning)은 논리적으로 흠없는 명제를 세우고 이것을 통해 개별 사실을 설명하는 방향으로(추상에서 구체로) 움직인다.
- 합리주의는 선험적(a priori) 이성을 통해 진리를 발견할 수 있다고 보지만 경험과학은 그렇지 않다. 합리주의는 이성에 의한 진리(truths of reasons)와 사실에 의한 진리(truths of facts)를 구별한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

사회과학철학의 맥락

경험과학은 자료(data)와 측정(measurement)를 토대로 한다.

- 주장에는 증거(evidence)가 필요하고, 증거는 결국 자료를 토대로 한다.
- 너무나도 당연한 말이지만, 자료(data)가 축적되기 위해서는 측정도구를 필요로 한다.
- 측정이란 어떤 현상을 관찰한 후 일정한 규칙에 따라 수치를 부여하는 것이다.
- 이런 행위에는 사회현상을 계량화(quantify)할 수 있다는 철학적 세계관을 내포하고 있다. 어떤 사람은 이 관념을 배격한다.

경험과학으로서의 사회학

- 사회학의 맥락에서 경험과학이란 "사회 현상을 체계적으로 관찰하여 얻어낸 자료(data)를 토대로 분석하여 법칙이나 원리 따위를 개발하는 학문"이다.
- 넓게 보면 이렇게 "증거를 수집하고 정리하는 체계적인 절차"를 개발하는 것도 경험과학의 일부로 볼 수 있다.
- 이렇게 "증거를 수집 정리하는 체계적인 절차"를 흔히 방법(method)이라고 하며, 방법론(methodology)이란 방법(method)에 대한 학문(-logy)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

사회과학철학의 맥락

경험과학은 논쟁적이고 역사적인 개념이다.

- 이 관념은 아무런 논쟁없이 뚝 떨어져 이견없이 받아들여진 것이 아니라 수 백년 이상의 논쟁을 거쳤고 지금도 가다듬어지며 변형되고 있다.
- 합리주의가 말하는 경험주의의 한계: 〈플라톤〉 동굴의 비유와 Bertrand Russell의 칠면조 이야기
- Karl Popper. "검증가능성(verifiability)은 지나치게 까다롭거나 무용지물이다. 그보다는 반증가능성(falsifiability)이 핵심이다."
- Thomas Kuhn. "정상과학(normal science)은 이상(anomaly)을 축적해 나가다 마침내 패러다임 이행(paradigm shift)을 겪고 새로운 정상과학에 길을 양보한다."
- Imre Lakatos. "하나의 연구프로그램(research program)은 쉽사리 반증되지 않으며 여전히 발견적 힘(heuristic power)을 통해 생명력을 유지한다."
- 그 밖에도 경험과학과 (사회)과학철학에 대한 깊이있는 비판적 논의들이 현재 진행중이다. 몇몇 사회학자(이기홍@강원대, 김명희@경상국립대)는 특히 비판적 실재론(critical realism)을 논의한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

경험적 사회과학 연구의 단계

경험적 사회과학 연구의 단계

이론(theory)

- 이론은 경험적으로 검증가능한, 상호연관된 일련의 명제들(propositions)이다.
- 명제란 둘 이상의 개념들 사이의 관계에 대한 진술이다(e.g., "X가 증가하면 y는 감소한다").
- 명제는 수행하는 역할에 따라 공리(axiom), 정리(theorem), 가설(hypothesis), 발견(finding) 등 다양한 형태를 가질 수 있다.

개념(concept)

- 추상적인 현상을 경계지어 개념으로 부르지만, 이는 경험적 연구를 위해 조작적 정의(operational definition)를 거쳐 변수(variables)로 측정(measure)되어야 한다.

가설(hypothesis)과 발견(finding)

- 가설이란 (연구를 수행하기에 앞서 제시된) 둘 이상의 개념들 사이의 관계에 대한 "잠정적" 진술이다.
- 발견이란 (연구를 통해 어느 정도 입증된) 둘 이상의 개념들 사이의 관계에 대한 진술(=명제)이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

경험적 사회과학 연구의 단계

통계학의 정의는 "자료(data)를 수집, 요약, 분석하는 이론과 방법"이다.

- "A body of techniques and procedures dealing with the collection, organization, analysis, interpretation, and presentation of information that can be stated numerically."

  Kuzma, J. W. 1984. Basic Statistics for the Health Sciences. Palo Alto, CA: Mayfield.

사회통계학의 역할

- (경험과학으로서) 사회학에서는 통계학을 통해 (1) 데이터를 수집하고 그 성격을 규명하고 (2) 이를 분석하여 가설을 테스트한다.
- 연구설계를 야심차면서도 조심스럽게 수행하고 이를 통해 수집된 데이터를 꼼꼼하게 살펴보아 그 통계학적 기법과 데이터의 성격을 규명하는 것도 사회통계학자(예컨대 Paul Lazarsfeld)의 역할이다.
- 좀 더 실질적인(substantial) 사회학적 주제에 관해 세워진 경험적 가설을 테스트하는 것도 사회통계학자의 역할이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

경험적 사회과학 연구의 단계

일단 교과서대로라면 경험적 사회과학 연구는 아래와 같은 순서대로 진행된다.

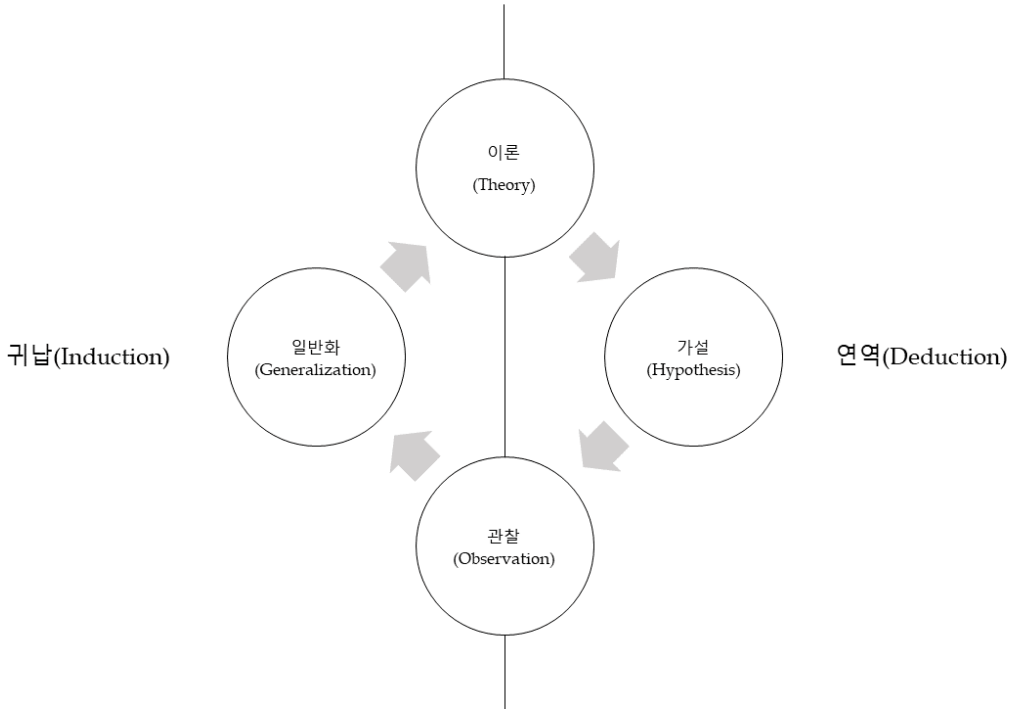
1. 문제인식: 엄격한 정의, 왜 연구가 필요한가, 기존 연구 속에 위치
2. 가설설정: 잠정적 진술의 꼴
3. 연구설계: 누구/무엇을, 언제, 어떻게 연구?
4. 자료수집: 설문조사? 실험?
5. 자료분석: 이른바 통계적 분석(statistical analysis)
6. 가설검정 및 일반화: 가설의 입증 또는 기각
7. 결과발표: 논문, 포스터 또는 학회 발표

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

경험적 사회과학 연구의 단계

물론 현실은 종종 다르다.

1. 몇몇 단계는 건너뛰거나 아예 주어진 것으로 받아들인다. 특정 단계들을 수 차례 반복하기도 한다.
2. 연역추론과 귀납추론 사이에서도 실용주의적 통합(또는 현실주의적 타협)이 일어난다.



연역과 귀납의 통합

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

경험적 사회과학 연구의 단계

그러나 타협 어쩌구하기 이전에 경계해야 할 것들도 있다.

- 이론없이 통계 분석은 사회학의 발전에 아무런 지적인 기여도 하지 못한다.
- 그것은 C. Wright Mills가 abstracted empiricism이라고 부르며 경멸했던 것 중 하나다.
- Mills는 경멸했지만 적어도 나는 사회통계학을 위한 사회통계학 연구는 의미가 있다고 본다. (Mills도 인정했듯) 그 나름대로 잡마켓에서 값어치를 할지도 모른다.
- Fortitudine vincimus, 어쩌면 특히 요즘 같은 때 젊은 세대에게는...

좋은 이론/가설과 좋은 통계적 기법 적용 사이의 관계는 몹시 미묘하다.

- 좋은 이론과 설명, 시사점에 관한 사색과 토론이 필요하다. 이 단계를 게을리하면 기법에는 흠잡을데가 없지만 내용은 시시한 논문이 나온다.
- 좋은 이론을 만들어도 그에 적합한 연구설계, 가설 수립, 통계적 기법의 활용, 결과 해석을 못하면 결국 논문을 완성하지 못한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

경험적 사회과학 연구의 단계

오늘날 경험과학의 중요한 덕목 가운데 하나는 재현가능성(replicability)다.

- 저자와 독자 사이에는 심각한 정보비대칭(information asymmetry)가 있기 마련이다.
- 가능한 범위 안에서 소스 코드와 데이터 공개해야 한다는 요청이 높아지고 있다.
- 공개된 소스와 데이터는 과학자 커뮤니티 사이에 지식의 공유를 촉진한다!
- 재현가능한 논문들은 그렇지 않은 논문보다 훨씬 피인용수가 높다는 분석도 있다.
- 현실적으로는 기밀 데이터 사용, 연구노하우 비밀유지, 오류지적의 두려움, 특허 인센티브 등으로 인해 어려움이 많다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

측정오류

측정오류

이론적 구성물: 연구자가 실재(the real)로부터 캐내고 컴퓨터 안에 집어넣어 "보이게 한" 무언가

- 경험적 사회학 연구에서는 사회 현상을 관찰하여 이를 관찰되고(observed) 측정된(measured) 데이터로 표현한다. 이 과정에서 이론적 구성물(theoretical construct)이라는 관념이 등장한다.
- 이론적 구성물이란 "an explanatory concept that is not itself directly observable but that can be inferred from observed or measured data. In psychology, many hypothesized internal processes are of this kind, being presumed to underlie specific overt behaviors. For example, a personality dimension, such as neuroticism, might be described as a theoretical construct measurable by means of a questionnaire." (APA 사전)

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

측정오류

오차(errors)

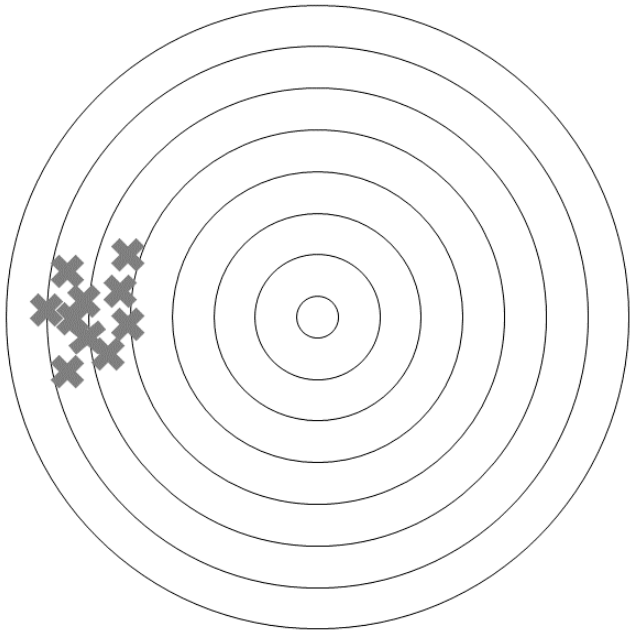
- (정치나 종교 등과는 달리) 경험과학적 연구는 틀릴 수 있다는 가능성에서 출발하여 오류가 일어나는 조건과 크기에 주목한다.
- 다양한 원인과 맥락에서 연구자가 측정하고자 하는 참 값(true score)과 측정된 값(observed score) 사이에 괴리가 발생한다.
- 인간이 수행하는 모든 측정에는 어느 정도 오차가 뒤따르기 마련이다.
- 오차로 말미암아 측정도구의 타당도(validity)와 신뢰도(reliability)가 훼손된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

측정오류

타당도는 재고자 한 개념을 얼마나 충실하게/제대로 측정했는가의 문제다.

- “the degree to which it measures what it is supposed to measure”
- e.g., 고래가 디지털 체중계에 올라섰는데 40kg이 나왔면 그 측정은 타당도가 낮다.



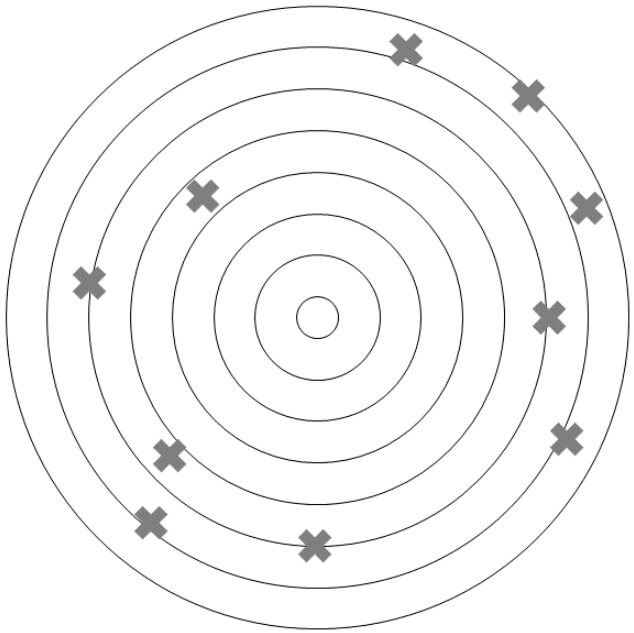
타당도와 과녁판의 비유

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

측정오류

신뢰도는 재고자 한 개념을 일관성있게(consistent) 측정했는가의 문제다.

- e.g, 고래가 디지털 체중계에 3번 측정해서 모두 다르게 나온다면 그 측정은 신뢰도가 낮다.



신뢰도와 과녁판의 비유

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

측정오류

타당도와 신뢰도 사이의 관계

- 타당도와 신뢰도는 둘 다 높을수록 좋다.
- 하지만 현실적으로 양자는 trade-off 관계에 놓여 있다(이른바 bias-variance trade-off).
- 데이터 과학(data science)에서 보면 양자 간에 균형을 맞추는 것이 타당하지만, 경제학이나 정책과학 등 사회과학적 관점에서는 종종 bias를 극소화하기 위해 신뢰도를 희생시키는 경우도 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

측정오류

측정오차(measurement error)는 여러가지 맥락에서 기인한다.

- 조사자 또는 검사도구의 맥락(e.g., 설문 내용의 모호함 등)
- 응답자 맥락(e.g., 설문에 대한 귀찮음, 무응답, 설문 문항에 대해 주관적인 해석 차이 등)
- 검사 과정/환경의 맥락(e.g., 장소가 지나치게 산만함 등)
- 조사자-응답자 상호작용(e.g., 인터뷰어-인터뷰이 권력거리/성차 등)

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

측정오류

측정오차=체계적 오차+무선 오차

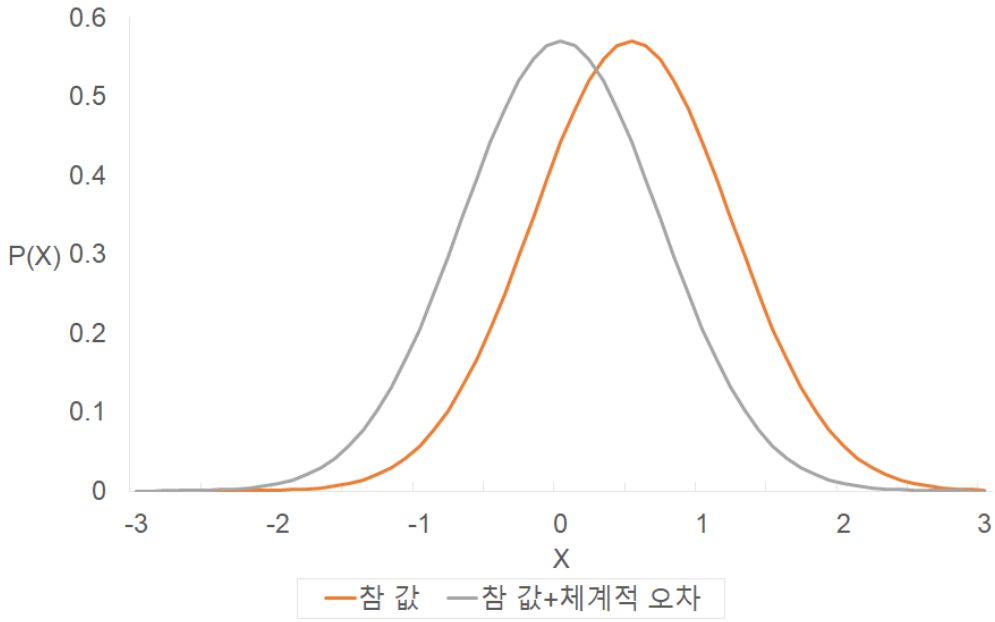
- 측정오차는 체계적 오차(systematic error)와 무선 오차(random error)로 분해된다.
- 측정된 값(X)=참 값(T)+측정오차(e)
- 오차는 사회조사의 질(quality)에 따라 심하게 영향을 받는다. 사회조사 단계에서 pilot test, feedback, 조사원 훈련 등 엄청난 노력을 기울여야 한다. 사회조사방법론과 서베이방법론에서 이를 다룬다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

측정오류

체계적 오차(systematic error)

- 체계적 오차는 “일정한 방향으로” “치우친(bias)” 오차다. 일정한 방향이라는 점에서 설명될 수도 있다.
- 측정도구(instrument)가 일관성있게 잘못 작동하여 수집된 자료가 참 자료와 일정하게 어긋나 있는 경우다(e.g., 디지털 체중계의 영점 조정이 안되어 일정하게 5kg씩 높게 측정함).
- 체계적 오차로 인해 측정된 값과 참 값 사이에는 평균(expected value)이 다르게 된다.
- 측정의 타당도(validity)의 문제로 이어진다.

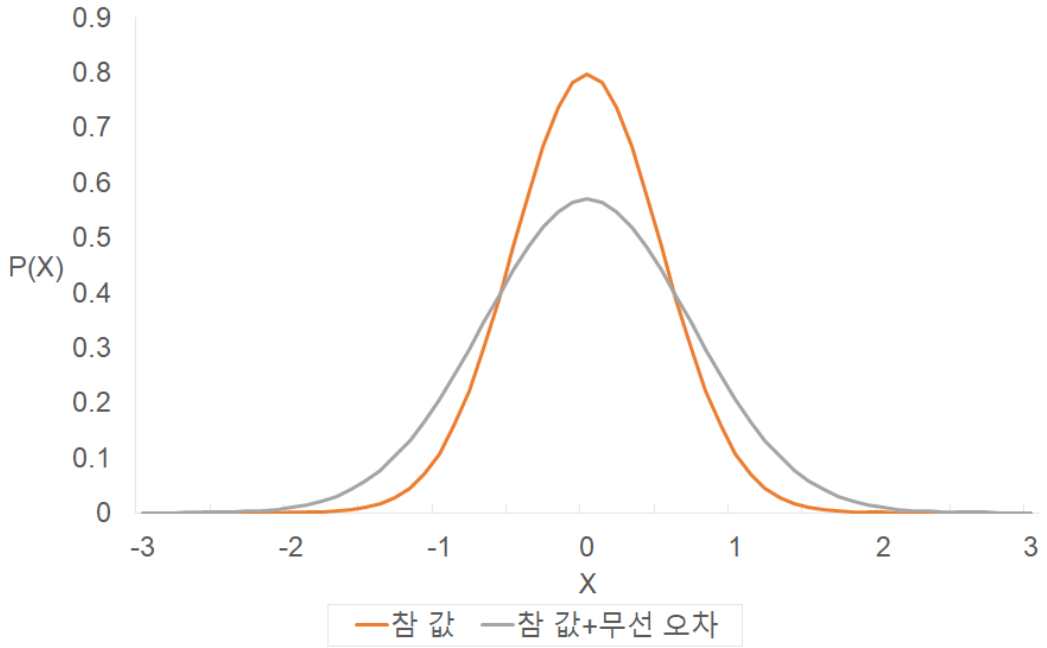


충북대학교
CHUNGBUK NATIONAL UNIVERSITY

측정오류

무선 오차(random error)

- 무선(無線)이란 선이 없다는 의미에서 만들어진 일본어다(그런데 wireless는 아니다).
- 무선 오차는 “임의(random)의 방향으로” “흩어진(variance)” 오차다. 임의적이므로 설명될 수 없다.
- 전통적인 사회과학적 맥락에서는 비교적 무해하고 자연스러운 것으로 여겨졌다(e.g., 디지털 체중계의 전기적 오류로 ± 1kg씩 다르게 측정함).
- 무선 오차로 인해 측정된 값과 참 값 사이에는 분산(variance)이 다르게 된다.
- 측정의 신뢰도(reliability)의 문제로 이어진다.



충북대학교 CHUNGBUK NATIONAL UNIVERSITY

측정오류

참 값이란 대체로 관찰할 수 없다.

- 그걸 알면 구태여 "측정된 값"을 구할 필요조차 없기 때문이다.
- 그러므로 많은 경우 측정된 값(X)-참 값(T)을 계산할 수 없어 측정오차(e)가 얼마나 큰지 알 수 없고, 측정오차 안에서 체계적 오차와 무선 오차가 각각 얼마나 큰지도 알 수도 없다.
- 특수한 연구설계를 통해 이를 드러내고 줄이려는 시도가 있지만 우리 수업에서는 다루지 않는다.

측정이론(measurement theory)에서는 측정도구의 타당도와 신뢰도를 평가하는 여러가지 통계 기술적인 방식을 발전시켜 왔다.

- 우리는 이 수업 말미에 그러한 기법들을 약간 배운다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

모집단과 표본, 그리고 분석단위

일엽지추(一葉知秋)

- “뜰 안에 잎이 하나 떨어지는 것을 보아 온 천하에 가을이 왔음을 미루어 안다.” ≪회남자(淮南子)≫ 〈설산훈편(說山訓篇)〉
- 무언가를 알기 위해 (설령 전체를 모두 살펴보지 않아도) 부분을 통해 미루어 짐작할 수 있다.
- 지식에도 어느 정도 경제적 논리가 작동한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

모집단(population)과 표본(sample)

- 당신은 72 만명이 네오청주에서 승차공유 스타트업을 운영하고 있기 때문에 도시에서 매일 소요되는 통근 시간의 평균을 알고 싶어 한다.
- 이때, 네오청주에서 출근하는 모든 사람은 모집단(population)이 된다.
- 그러나 모든 네오청주 통근자(=모집단)를 조사하기엔 비용을 감당할 수 없다.
- 따라서 당신은 네오청주 통근자 중 일부만을 골라 표본(sample)을 추출할 수 있다.

모수(parameter)와 통계량(statistic)

- 모든 네오청주 통근자의 통근 시간의 평균은 모수(parameter)라고 부른다(전에 배운 모수와는 의미가 다르다).
- 표본으로 고른 네오청주 통근자의 통근 시간의 평균은 통계량(statistic)이라고 부른다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

추청값(estimate)과 추정량(estimator)

- 당신은 네오청주에서 통근하려던 100명을 붙잡아 표본(sample)을 추출하였다.
- 그 100명의 통근 시간을 조사하여 표본(sample)의 평균(mean)을 계산해 보았더니 35분이라는 결론을 얻었다.
- 이 35분을 추정값(estimate)이라고 부른다.
- 한편, 35분은 말하자면 그래프 상에 점을 찍듯 하나의 숫자로 구한 값이므로 점추정값(point estimate)이다. 그와는 "달리 25분에서 45분 사이"라고 추정했다면, 이는 구간추정값(interval estimate)이 된다.
- 추청치를 구하는 방법을 추정량(estimator)이라고 부른다. 이 예제에서는 표본평균(sample mean)이 모평균(population mean)의 추정량인 셈이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

통계량(statistic)은 모수(parameter)의 추정량(estimator)이 된다.

- 우리는 표본(sample)에서 얻은 통계량(statistic)을 가지고 모집단(population)의 모수(parameter)를 추정하고자 하기 때문이다!
- 말할 필요도 없이, 우리가 통계량을 통해 모수를 추정하고자 할 때도 오차의 문제는 발생한다. 이에 관해서는 나중에 훨씬 심도있게 다룬다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

분석단위(unit of analysis)의 문제

- 분석단위에 따라 데이터로 입력된 한 줄(row)이 무엇을 지칭하는지 결정된다.
- 많은 사회학도들은 거의 척수반사처럼 분석단위를 개인(individuals)이라고 전제한다. 꼭 그래야 할 이유는 없다.
- 분석단위는 분석 목적과 전략에 따라 얼마든지 달라질 수 있다: 발화(act of speech), 조직(organization), 프로그램(program), 국가(country), 개월(month), 국가-연도(country-year) 등.
- 창의적인 분석단위의 선택은 경험적 사회학 연구의 훌륭한 출발점이다!

관찰단위(unit of observation)와 분석단위는 다를 수도 있다.

- 필요에 따라 관찰은 훨씬 더 작은 단위로 한 뒤, 이것을 뭉뚱그려서(aggregate) 좀 더 큰 분석단위를 만들 수도 있다.
- e.g., 3개 국가의 개인들을 관찰단위로 삼아 조사한 뒤, 이를 국가별로 뭉뚱그려 평균을 분석단위로 삼아 국가간 비교를 수행할 수도 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

원칙적으로 특정 분석단위로부터 얻은 발견(finding)은 해당 분석단위에만 국한된다.

- 생태학적 오류(ecological fallacy)는 국가나 집단과 같이 거시적인(macro) 분석단위를 통해 얻은 발견을 개인과 같이 미시적인(micro) 단위로 확장할 때 범하는 오류다(Robinson 1950).

  W. S. Robinson. 1950. "Ecological Correlations and the Behavior of Individuals." American Sociological Review 15(3): 351–357.

- 흡연과 기대수명 및 각종 발병률 사이의 관계: 국가 수준과 개인 수준

그러나 생태학적 오류를 무슨 "전가의 보도"처럼 휘둘러선 안된다.

- 미시적 분석단위가 거시적 분석단위보다 우월하다는 의미로 해석되어선 안된다. 게다가 때때로 개인 레벨 서베이는 아예 불가능하다(e.g., 비교역사사회학).
- 특히 사회학 연구에서는 거시적 발견과 해석이 그 자체로 의미있는 경우가 많다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

모집단, 표본, 관찰단위, 분석단위, 추정값: 하나의 예제

- 당신은 전세계 국가를 모집단(population)으로 삼을 수 있다.
- 전세계 모든 국가를 대상으로 사회조사를 실시하여 국가별로 랜덤하게 100명 씩 조사하면 표본(sample)은 각 국가별 100명이 된다.
- 데이터로 최초로 입력되는 관찰단위(unit of observations)는 개인이다.
- 그 뒤, 국가별로 평균 흡연량과 심장병 유병률을 계산하여 새로운 데이터를 입력하면 이제 분석단위(unit of analysis)는 국가가 된다.
- 국가별 심장병 유병률과 평균 흡연량 사이의 상관관계를 분석하여 추정값(estimate)을 얻는다. 이제 이 추정값에 근거해 모집단의 모수(parameter)를 추정한다(국가 수준에서의 일반화).
- 이 추정은 개인 수준에서 일반화 될 수 없다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

심슨의 역설(Simpson's Paradox)

- 가장 유명한 예제는 1973년 가을학기 UC Berkeley 대학원 입학 성차별 문제다 (Bickel, Hammel, and O'Connell 1975).

  Bickel, P.J., E.A. Hammel, and J.W. O'Connell. 1975. "Sex Bias in Graduate Admissions: Data From Berkeley." Science 187(4175): 398–404.

- "당시 성별 입학률(admission rate) 통계자료에 따르면 남자의 입학률이 여자의 입학률보다 훨씬 높았다. 이것은 성차별의 결과인가?"

Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12{,}763$, $\chi^2 = 110.8$, d.f. $= 1$, $P = 0$ (*18*).

Applicants	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
Men	3738	4704	3460.7	4981.3	277.3	− 277.3
Women	1494	2827	1771.3	2549.7	− 277.3	277.3

Bickel et al. (1975: 399)

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

모집단과 표본, 그리고 분석단위

- Bickel et al. (1975) 은 아닐수도 있다고 지적. 이런 학교 수준의 집계자료 (aggregate data)는 종종 학과(department)의 이질성을 감추기 때문이다.
- 학과 계열 별로 나누어 보았을 때, 여학생은 인문/사회과학 계열 지원률(application rate)이 높았고, 남학생은 수학/공학 계열 지원률이 높을 수 있다.
- 그런데 수학/공학 계열은 입학 경쟁이 상대적으로 덜한 반면, 인문/사회과학 계열은 입학 경쟁이 상대적으로 치열한 편이다. Bickel et al. (1975)은 이런 경쟁 차이의 이유로 수학 선수강 등을 지목했다.
- 그 결과, 여학생의 입학률이 낮을 수도 있다는 것. 그는 가상적 자료를 통해 학과 수준에서는 성별 입학률이 똑같지만, 집계자료에서는 달라질 수 있음을 보임.

Table 2. Admissions data by sex of applicant for two hypothetical departments. For total, $\chi^2 = 5.71$, d.f. $= 1$, $P = 0.19$ (one-tailed).

Applicants	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
	Department of machismatics					
Men	200	200	200	200	0	0
Women	100	100	100	100	0	0
	Department of social warfare					
Men	50	100	50	100	0	0
Women	150	300	150	300	0	0
	Totals					
Men	250	300	229.2	320.8	20.8	− 20.8
Women	250	400	270.8	379.2	− 20.8	20.8

Bickel et al. (1975: 400)

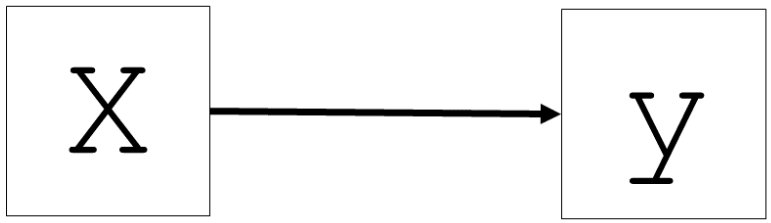
충북대학교 CHUNGBUK NATIONAL UNIVERSITY

상관관계와 인과관계

상관관계와 인과관계

종속변수와 독립변수

- 종속변수(dependent variable)란 다른 무언가에 종속되어 설명되어지는 변수다.
- 독립변수(independent variable)란 다른 무언가로부터 독립되어 설명하는 변수다.



독립변수 →종속변수

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

상관관계와 인과관계

이제는 비전공자에게도 상식이라지만...

- "상관관계는 인과관계를 뜻하지 않는다(Correlation does not imply causation)."
- 현실에서는 심지어 전공자 사이에서조차 혼동해서 쓰이고 있음.
- "상관분석은 상관관계를, 회귀분석은 인과관계를 분석하는 기법이다" →완전틀림!
- (사회)과학철학의 맥락에서 인과관계라는 관념과 방법은 (사회)과학철학에서 역사적으로도 그리고 지금 이 순간에도 치열한 논쟁의 대상이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

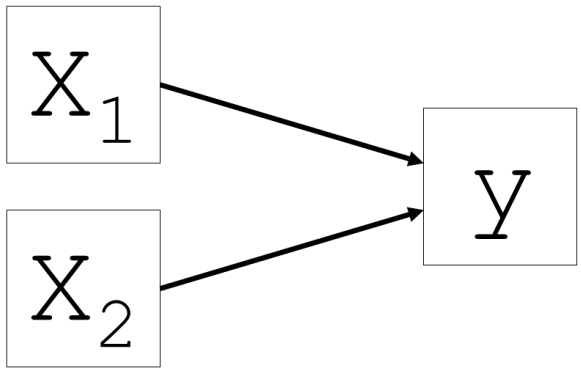
상관관계와 인과관계

John Stuart Mill이 말하는 인과관계의 세 가지 조건

1. 원인은 결과보다 시간상 선행한다.
2. 원인과 결과에는 상관관계가 있다.
3. 결과에 대해 다른 그럴듯한 대안적 설명이 없다.

많은 사회과학 연구는 (2)의 조건만을 밝힐 뿐 실제로는 인과관계의 분석이 아닌 경우가 많다.

- 우리가 동시간 대에서 수집한 횡단면자료(cross-sectional data)를 통한 분석에는 이미 (1)에 위배되기 쉽다.
- (3)의 조건은 흔히 통제변수(control variable)를 이용하여 충족된다고 생각하지만 사실은 그렇게 간단한 문제가 아니다.

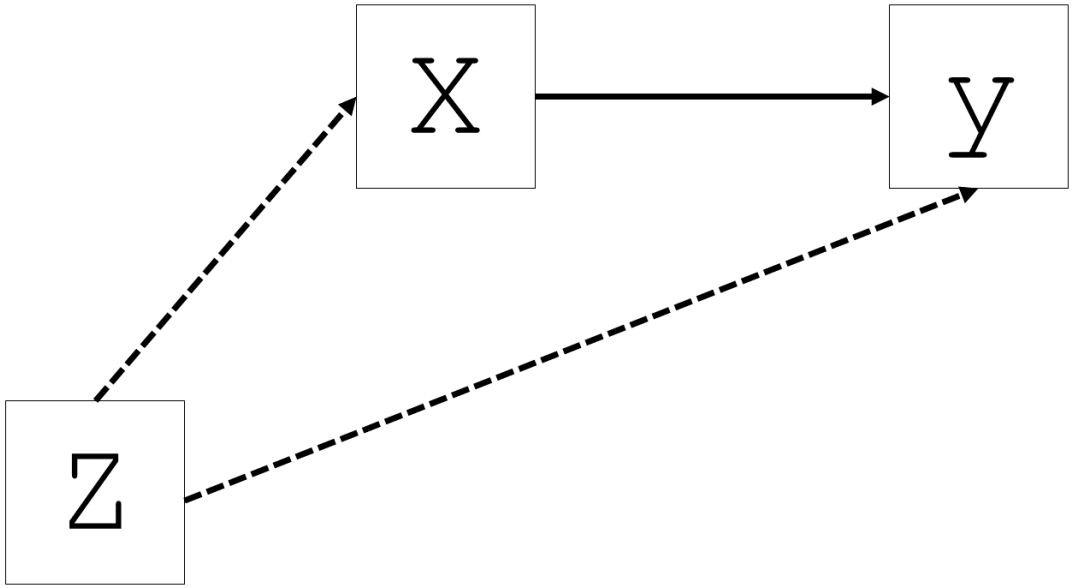


충북대학교
CHUNGBUK NATIONAL UNIVERSITY

상관관계와 인과관계

교란변수(Confounding variable)

- 교란변수란 연구상 설정된 독립변수와 종속변수에 동시에 영향을 미치는 제3의 변수다.
- 교란변수가 관찰가능(observable)하면 그나마 다행이지만 관찰불가능(unobservable)한 경우가 많다.
- 실험설계(experimental design)는 무작위화(randomization)라는 마법을 통해 교란변수의 영향력을 근본적으로 차단하는 가장 이상적인 세팅이다.
- 경험적 사회학 연구의 대부분은 비실험적(non-experimental) 데이터를 사용하므로 교란변수의 영향력을 배제하기 어렵다.



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

상관관계와 인과관계

예제 1. "높은 아이스크림 섭취량은 학업성취도를 낮춘다."

- 아이스크림은 머리를 나쁘게 한다? 이른바 허위적 관계(spurious relations).

예제 2. "교육이 여성의 임금에 미치는 영향은 어떠한가?"

- James Heckman. "여성은 유보임금(reservation wage)보다 큰 임금을 제의받아야 취업한다. 다시 말해, 취업은 randomized되지 않기 때문에 취업여성만을 표본(sample)로 삼을때, 이 추정은 표본선택편의(sample selection bias)에 취약하다."

예제 3. 심슨의 역설(Simpson's Paradox)

- 테크니컬하게 말하자면, 심슨의 역설은 교란변수가 극단적으로 강력하게 작용하여 (1) 전체 자료 수준에서 확인했을 때와 (2) 세부적인 층(stratum) 별로 확인했을 때의 두 변수의 관계가 정반대로 나타나는 현상이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY