제 1 장　통계학과 자료

1. 왜 통계학을 배워야 하나
2. 자료
3. 실험연구
4. 경험적 연구
5. 통계학을 대하는 자세

1. 왜 통계학을 배워야 하나

통계학의 유용성

경제학, 경영학, 사회과학, 인문과학, 법학, 의학, 공학, 자연과학 등 분야를 막론하고 통계학을 쓴다. 우리는 통계에 묻혀 산다.

- 10% 담뱃값 인상이 청소년층의 흡연을 어느 정도 줄이나?
- 직업훈련은 재 취업률을 높이는가? 임금률에 미치는 효과는?
- 외국인 직접투자가 늘면 경제성장률이 제고되는가?
- 특정 광고가 매출증가 효과를 가져왔는가?
- 노동시장에서 여성에 대한 차별이 존재하는가?

1. 왜 통계학을 배워야 하나

월별 출생률

Births are not uniformly distributed over a year.

- The typical United States pattern for births is an April-May trough.
- April-May births correspond to July-August conceptions, hottest months in the States.
- This trend is more pronounced in the southern hot states of Georgia and Louisiana than in New York.
- Less visible April-May trough as air-conditioning spreads.

1. 왜 통계학을 배워야 하나

통제된 실험에서 두 비율의 비교

'인간의 합리성' 가정을 검증하기 위한 실험

의사집단 1: 서식A를 제공. 의사집단 2: 서식B를 제공

[서식 A] 수술 환자 100명 중에서 10명은 수술 도중에 죽고, 32명은 1년 이내에 죽으며, 66명은 5년 이내에 죽는다. 방사선 치료를 받는 100명의 환자 중에서는 아무도 치료 도중에 죽지 않고, 23명이 1년 이내에 죽으며, 78명이 5년 이내에 죽는다.

[서식 B] 수술 환자 100명 중에서 90명은 살아서 수술 기간을 넘기고, 68명은 일년 이상 살아남으며, 34명은 5년 이상 살아남는다. 방사선 치료를 받는 100명의 환자 중에서는 모두 다 살아서 치료기간을 넘기고, 77명이 1년 이상 살며, 22명이 5년 이상 산다.

서식을 읽은 후 의사들은 각자 폐암 환자에게 추천할 치료법을 제시

- 의사집단 1: 80명 중 40명(50%)이 수술을 추천
- 의사집단 2: 87명 중 73명(84%)이 수술을 추천

1. 왜 통계학을 배워야 하나

엉터리 통계의 예

"딸만 둘 낳았으니 이제 아들 낳을 차례다." – 시어머니 말씀 중 –

"주식은 저점에서 사고, 고점에서 팔아야..." – 투자 전문가 조언 중 –

"원칙을 지키며 펀더멘털에 따라 투자한 결과 3개월 만에 800%가 넘는 높은 수익률을 올렸습니다" – 대학생 모의주식투자 대회 1등 팀에 대한 전문가 평가 중 –

"수능 성적과 입학 후 학점간에는 상관관계가 낮습니다. 수능 성적은 입학 후 성취를 예측하는데 별로 도움이 되지 못합니다." - 모 대학 교무처 발표 중 -

1. 왜 통계학을 배워야 하나

엉터리 통계의 예

각 타석은 독립입니다.

1. 왜 통계학을 배워야 하나

엉터리 통계의 예

아동학대, 친부모 '80%' 계부모 '10%미만'(조선일보, 2003.05.05)

- 보건복지부가 전국 17개 시,도 아동학대 예방센터와 신고전화 '1391'을 통해 접수한 아동학대 사례는 모두 4,111건이며, 이 중 2,478건이 실제 아동학대
- 복지부에 신고된 학대 사례들도 사망 등 극한 상황으로 가는 특수한 경우가 아니라, 자기 집에서 친부모에게 거의 매일 학대를 당한 사례가 주종을 이뤘다. 가해자의 80%는 친부모이며, 계부모와 양부모는 각각 10% 미만이었다.

잘못의 지적: 친부모가 계부모보다 자식을 8배 더 학대하는 경향이 있는 게 아니라 대부분의 아이가 친부모 밑에서 자라고 있음

올바른 분석: 친부모 슬하의 자녀들 집단, 계부모 슬하의 자녀들 집단 등 두 집단간 매맞는 아이의 비율을 비교하는 이른바 two-sample analysis해야 타당함

1. 왜 통계학을 배워야 하나

엉터리 통계의 예

'속도의 유혹', "곧은 길서 사고 더 많아요" (중앙일보, 2006.10.9)

- 곧게 뻗은 직선도로에서 속도의 유혹을 이기지 못하는 바람에 발생하는 교통사고가 전체 교통사고 10건 중 9건이나 된다고 합니다. 건설교통부와 경찰청이 지난해 발생한 교통사고 21만4000여 건을 분석한 결과입니다.
- 쭉 뻗은 직선 도로에 올라서면 가슴이 뻥 뚫리는 듯한 느낌을 받으신 적이 있을 겁니다. 차도 별로 많지 않다면 자연스레 액셀러레이터를 밟는 발에 힘이 들어가게 됩니다. 악마의 유혹. 사고의 위험성이 순간적으로 높아지는 순간입니다.

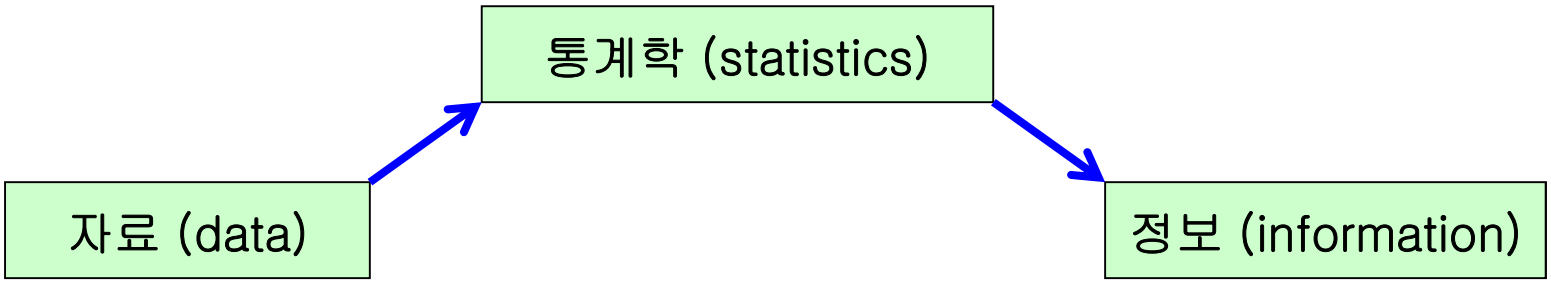
잘못의 지적: 전체 도로 중 대부분이 곧은 도로이고 굽은 도로나 오르막/내리막 길은 상대적으로 적음

올바른 분석: 쭉 뻗은 직선 도로 주행 100,000km 당 사고 발생률과 그렇지 않은 도로 주행 100,000km 당 사고발생률을 상호 비교해 보아야 함

1. 왜 통계학을 배워야 하나

통계학이란 무엇인가?

세상은 자료들로 가득 차 있다.



통계학은 자료를 정리/분석해 유용한 정보를 얻기 위한 언어이자 도구임

1. 왜 통계학을 배워야 하나

통계학의 분류

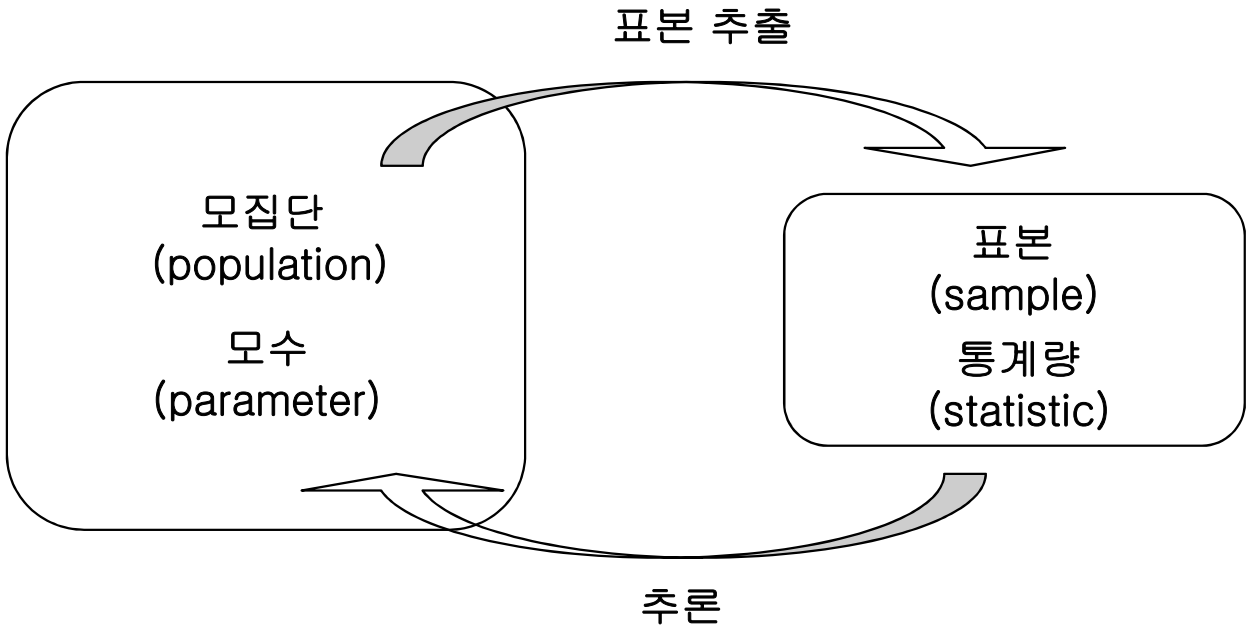
통계학은 크게 두 가지로 분류할 수 있다.

- 기술통계학 (descriptive statistics)
  - 자료를 변수 별로 따로따로 또는 관계되는 변수끼리 묶어서 요약

- 추론통계학(inferential statistics)
  - 정리된 자료에 담긴 의미를 해석하여 미지의 세계에 대해 추론

2. 자료

모집단과 표본 사이의 관계



통계학은 표본의 자료를 수집, 정리, 요약하고 나아가 요약된 자료를 토대로 그 자료의 모태가 되는 모집단에 대해 짐작, 추측해 보는 작업을 포함

2. 자료

자료의 종류

횡단면 자료(cross-sectional data)

- 한 시점에서 여러 개체를 관측한 자료

시계열 자료(time-series data)

- 한 개체를 여러 시점에 걸쳐 관측한 자료

패널 자료 (panel data) 또는 종적 자료 (longitudinal data)

- 횡단면과 시계열의 특성을 결합하여 여러 개체를 여러 시점에 걸쳐 관측한 자료

2. 자료

국내외 패널자료 예시 1

한국노동패널

PSID, NLS, NLSY

BHPS

JPSC: Japanese panel survey on consumers

Compustat, CRSP

Scanner data (마케팅 분야)

2. 자료

국내외 패널자료 예시 2

신용카드 보유자의 월별 업종별 지출액 패널 자료

통신회사(전화, 인터넷)가 보유한 개인별 통신 사용자료

전기 등 각종 공과금의 월별 지출액 자료

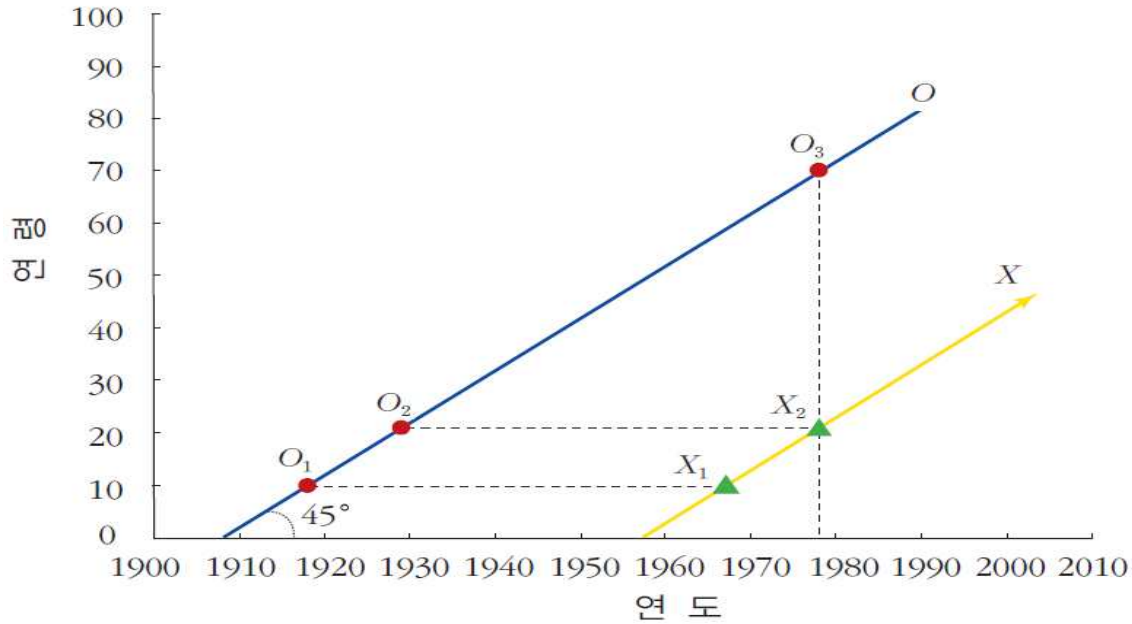
은행의 기업/가계별 대출 및 사후 관리 자료

기업의 신용등급 변화 자료

결혼정보회사의 개인별 matching history 자료

2. 자료

렉시스 도표



주: O_2와 X_2는 같은 나이를, O_3와 X_2는 같은 시대를 나타낸다.

종적 분석은 여러 개체를 여러 시점에 걸쳐 다양하게 비교함으로써 보다 풍부한 분석을 가능하게 해줌

2. 자료: 대표성 문제

유행가, 소설 속 자료의 대표성 문제

(i) "거리엔 괜찮은 사람들이 많은데 소개를 받으러 나온 자리엔 어디서 이런 사람만 나오는 거야" (유행가 구절)

(ii) "서울이라고 못난이가 없을 리야 없겠지만 대처에서는 못난이들이 거리에 나와 행세를 하지 못하고, 시골에선 아무리 못난이라도 마음 놓고 나와 다니는 때문인지, 못난이는 시골에만 있는 것처럼 흔히 시골에서 잘 눈에 뜨인다." (이태준 소설 '달밤' 중)

2. 자료: 대표성 문제

표본추출편의(**sample selection bias**)

"무작위 추출"의 개념을 악용하는 아들!

2. 자료: 대표성 문제

상기하라! 돈 잃은 수많은 사람들의 소리 없는 아우성

돈을 딴 사람만 목소리가 큽니다.

2. 자료: 대표성 문제

여론조사: 다이제스트사 vs. 갤럽

1936년 미국 대통령 선거 결과의 예측

	루즈벨트의 득표율(%)
실제선거결과	62
다이제스트사의 예측	43
다이제스트사의 예측에 대한 갤럽의 예측	44
갤럽의 예측	56

→ 실제 루즈벨트 승리

→ 다이제스트사는 랜던 승리 예측

→ 갤럽은 다이제스트사가 랜던 승리 예측할 것으로 예측

→ 갤럽은 루즈벨트 승리 예측

갤럽의 일방적 승리

- 다이제스트사의 잘못된 예측까지도 갤럽이 예측 (지피지기면 백전백승!)
- 다이제스트사의 표본추출방식에는 심각한 문제가 있었다.

2. 자료: 대표성 문제

1993년의 LA 시장 선거: 여론조사와 실제 결과의 비교

LA Times Poll on LA Mayoral Election, LA Times, May 12, 1993

- among all registered voters: 민주당 Michael Woo 후보가 공화당 Richard Riordan 후보에 6% 포인트 앞서는 것으로 나타남
- among the likely voters: 공화당 Richard Riordan 후보가 민주당 Michael Woo 후보에 7% 포인트 앞서는 것으로 나타남
- 1993년 6월 8일의 실제 선거결과: 공화당 Richard Riordan 후보가 민주당 Michael Woo 후보에 8% 포인트 앞선 승리
- 전체 등록된 유권자의 44%만이 투표. Likely voters가 all registered voters보다 모집단을 더 잘 대표함

2. 자료: 대표성 문제

생존편의(survivorship bias)

- When estimating an overall rate of return from stock investment for the past 20 years:
- If you took a random sample of 100 stocks currently trading, this would be incorrect and overstate returns due to "survivorship bias".
- You would better sample 100 random stocks that were trading 20 years ago, as some might go bankrupt or merge under adverse circumstances.
- Ryu & Yoon (2013), "Relative performance of chaebol vs. non-chaebol in Korea over the past three decades: a value path approach" : 재벌성과를 측정하는데 있어 생존편의를 고려하기 위해 value path approach 이용

2. 자료: “Final Four, Five Years Later”

종적 자료와 인생사

- 1987년 미국 NCAA 대학농구 토너먼트 Final Four 팀

  -Indiana (감독 B. Knight 우승 74 대 73)

  -Syracuse (감독 J. Boeheim 준우승)

  -UNLV (감독 J. Tarkanian)

  -Providence (감독 R. Pitino)

2. 자료: “Final Four, Five Years Later”

종적 자료와 인생사

- 1989년 11월 10일 미국 신시내티의 한 모텔 110호: 미연방 마약수사대가 5명 체포. 그 중 한 명이 Derek Brower (1987년 Syracuse 농구선수)

2. 자료: “Final Four, Five Years Later”

종적 자료와 인생사

- 1992년 4월: NY Times 기자 Ira Berkow는 5년 전 Final Four에 진출했던 4개 대학 53명 선수들의 1987년 이후 life 추적
- The Final Four, Five Years Later -- A special report: Players Find Glory is Replaced by Reality- NY Times, April 03, 1992 –
- Life after 1992, updated by Keunkwan Ryu

2. 자료: “Final Four, Five Years Later”

Indiana 우승팀

• 가드 Steve Alford, 1987년 NBA 신인드래프트서 제2라운드에 Dallas Mavericks에 지명. 1992년 Manchester College의 헤드코치. 2013년 현재 UCLA 대학 헤드코치. 아들 Bryce는 UCLA 농구선수

2. 자료: “Final Four, Five Years Later”

Indiana 우승팀

• 가드 Keith Smart, 1987년 토너먼트 결승전서 역전 2점 슛 성공시킨 Final Four MVP. 1988년 NBA 신인 드래프트서 제2라운드에 Golden State Warriors에 지명. 4게임 뛰고 방출. 미국 마이너 리그, 필리핀 등지서 선수 생활. 2010년 GS Worriors 헤드코치로 부임한 뒤 다음 해인 2011년 해임. 2012년 Sacramento Kings 헤드코치로 부임한 뒤 다음 해인 2013년 해임

2. 자료: “Final Four, Five Years Later”

Indiana 우승팀

• 센터 Dean Garrett, 1988년 NBA 신인 드래프트에서 제2라운드에 Phoenix Suns에 지명. NBA 방출 이후 1989년 유럽 진출 이후 1996년까지 이태리, 그리스 리그 등에서 선수 생활하다 1996년 NBA 복귀. 2002년 NBA 은퇴 후 2013년 현재 미네소타에서 음식점, 나이트클럽 등 개인 사업

2. 자료: “Final Four, Five Years Later”

Indiana 우승팀

- 포워드 Daryl Thomas, 1987년 NBA 신인 드래프트에서 제6라운드에 지명됨. 1992년 당시 유럽서 선수생활
- 가드 Richard Calloway, 1990년 미프로농구 마이너리그 CBA에 드래프트. 1990-91 한 시즌은 NBA Sacramento Kings 선수로도 뜀. 마이너리그, Argentina, Poland 등에서 선수 생활

2. 자료: “Final Four, Five Years Later”

Indiana 우승팀

- 센터 Todd Jadlow, 졸업 후 인디애나 감독인 Bobby Knight 밑에서 잠시 선수 스카우터로 일한 경력
- 가드 Joe Hillman, academic all-American, 1992년 당시 regional supervisor for Xerox

2. 자료: “Final Four, Five Years Later”

Indiana 우승팀

- 가드 Todd Meier, 1992년 당시 United Parcel 근무
- 포워드 Brian Sloan, 대만에서 한 시즌 선수 생활. 이후 인디애나 대학으로 복학하여 의학 전공하여 1997년 졸업, 2000년 레지던트 수료. 2000가을에 법과대학원 진학

2. 자료: “Final Four, Five Years Later”

Syracuse 준우승팀

• 포워드 Derrick Coleman, 1990년 NBA 신인드래프트서 전체 1순위로 New Jersey Nets에 지명. 1992년 NJ Nets와 5년간 1,500만 달러 계약. 2010년 3월 파산 선언

2. 자료: “Final Four, Five Years Later”

Syracuse 준우승팀

- 센터 Rony Seikaly, 1988년 NBA 신인드래프트에서 전체 9순위로 1라운드에 Miami Heats에 지명. 1999년 NBA 은퇴. 현 수백만 달러 부동산 회사 소유 및 운영. 2013년 현재 라디오 방송 및 나이트클럽 Music DJ

2. 자료: “Final Four, Five Years Later”

Syracuse 준우승팀

- 가드 Steve Thompson, NBA Sacramento Kings에서 잠시 선수생활
- 가드 Sherman Douglas, Syracuse guard, 대학 재학 중 팀 내 최고득점 선수(게임당 평균 17.2점 득점 및 7.6개 어시스트). NBA Boston Celtics에서 잠시 선수생활
- 가드 Greg Monroe, 졸업 후 1년간 미프로농구 마이너리그 선수 생활

2. 자료: “Final Four, Five Years Later”

Syracuse 준우승팀

- 포워드 Howard Triche, 1992년 당시 주류회사인 Anheuser Busch 근무
- 참고: 2013년 NCAA대학농구 16강전서 Syracuse 대학과 Indiana 대학이 맞붙었는데 Syracuse는 Howard의 조카인 Brandon의 14득점 활약에 힙 입어 Indiana를 61 대 50으로 격파하고 8강에 진출 (26년 전 삼촌의 패배 설욕: 삼촌은 1987년 결승서 20여초 남기고 역전 2점슛을 성공시킨 Indiana의 Keith Smart의 역전 슛을 막지 못함)

2. 자료: “Final Four, Five Years Later”

Syracuse 준우승팀

• 포워드 Derek Brower, 1989년 마약 거래 관련 체포. 1990년에 보석 상태로 Syracuse 대학 복귀하여 심리학 학사 마치고 이후 6개월 복역 후 출소

• 참고: NCAA 농구의 의도적 파울(intentional foul) 규칙 도입은 Derek Brower와 관련됨. 1987년 Western Kentucky 대학 선수들은 Syracuse 선수가 볼을 인바운드 하기도 전에 의도적으로 Brower에게 파울. Brower는 자유투 6개 던져 모두 실패. 다음 시즌부터 NCAA는 의도적 파울의 경우 상대방에게 자유투 2개 및 볼 소유권 유지시켜주는 의도적 파울 규칙 도입

2. 자료: “Final Four, Five Years Later”

Providence

• 가드 Billy Donovan, NBA NY Knicks, Utah Jazz 등에서 잠시 선수생활. 이후 마이너리그 거쳐 선수생활 접고 월스트리트서 7개월간 investment banker로 일하다 1992년 당시 켄터키대학 헤드코치로 자리를 옮기는 이전 스승 R. Pitino 감독 밑에서 assistant coach. 2013년 현재 Florida 대학의 헤드코치. 2006년과 2007년 두 해 연속 Florida 대학을 NCAA 토너먼트서 우승시킴

2. 자료: “Final Four, Five Years Later”

Providence

- 가드 Delray Brooks, 1992년 당시 마이너리그 선수로 뛰면서 플로리다의 drug store 체인에 근무
- 포워드 David Kipfer, 1992년 당시 Providence에 위치한 한 종이제조회사의 판매 담당

2. 자료: “Final Four, Five Years Later”

UNLV

• 포워드 Armen Gilliam, 1987년 NBA 신인드래프트서 전체 2순위로 Philadelphia 76ers에 지명. 1992년 당시 NBA Phoenix Suns 선수. 총 13년간 NBA 선수생활 후 은퇴. 2011년 사망

• 가드 Fred Banks, 1987 NCAA 토너먼트 준결승전서 인디애나와 맞붙었을 때 3점 슛 10개 성공시킴. 팀은 93-97 패배. NBA 신인 드래프트서 제2라운드에 지명. 1992년 당시 그리스서 선수생활. 2013년 현재 미국 라스베가스의 Canyon Springs 고등학교 농구감독

2. 자료: “Final Four, Five Years Later”

UNLV

• 포워드 Gerald Paddio, 1988년 NBA 신인 드래프트에서 제3라운드에 Boston Celtics에 지명됨. 1992년 당시 미프로농구 마이너리그, 유럽 리그 등서 선수생활. 미프로농구 마이너리그, 유럽, NBA 등을 오가며 선수생활 지속하다 2004년 은퇴

• 포워드 Jarvis Basnight, 1992년 당시 NBA 진출 꿈꾸었으나 결국 진출 못한 채 미국 마이너리그, 해외 리그 등에서 10년간 선수 생활하다 은퇴. 2010년 현재 가족 위탁 보호 사업에서 일하고 있음

2. 자료: “Final Four, Five Years Later”

UNLV

- Richard Robinson: UNLV 센터. 1992년 당시 라스 베가스 Clark County 소년법원에서 교도관으로 근무하다 UNLV로 복학
- 포워드 Leon Symanski, UNLV의 12명 선수 중 11번째 선수. 주전 포워드인 Armon Gilliam 훈련시키는 역할 담당의 “연습용” 선수. UNLV 대학서 호텔경영학 공부 마친 뒤 1992년 당시 라스베가스의 한 호텔서 근무

2. 자료: “Final Four, Five Years Later”

UNLV

- R가드 Gary Graham, 근무하던 은행 합병당한 뒤 해고됨, 1992년 당시 Clark County 소년원 청소년 프로그램 운영하며 대학원 재학 중

- 가드 Mark Wade, 1987-88 한 시즌만 NBA GS Warriors에서 선수 생활. 1989-90 시즌에는 단 한 게임만 NBA Dallas Mavericks에서 뛰고 선수생활 마감. 이후 2002년부터 2007년까지 여러 고등학교, 대학교 등에서 assistant coach 생활. 2007년 횡령 등 혐의로 체포, 2008년 횡령 혐의 인정하고 150일간 수감생활

2. 자료: 변수의 종류

변수의 종류

양적(quantitative) 변수: 나이, 가족의 수, 가구소득

질적(qualitative) 변수: 혼인상태, 취업여부

- 일반적으로 질적 변수도 통계처리 목적상 수치로 코딩하여 사용함

이산변수(discrete): 가족의 수처럼 2,3,4,… 등의 이산적인 값만을 취함

연속변수(continuous): 나이, 가구소득처럼 연속인 값을 취함

- 컴퓨터를 통해 숫자를 표현하면 이론상 이는 언제나 이산적일 수밖에 없음
- 현실적으로는 어떠한 연속변수도 이산적으로 근사 시켜 표현할 수밖에 없음
- 이때 그 근사의 정확도를 얼마로 할 것인가가 문제의 본질임

2. 자료: 척도의 종류

척도의 종류

명목척도 (nominal scale) – 척도의 명칭만 의미 있음

- (예) 결혼 상태에 대한 코드: { 미혼=1, 기혼=2, 이혼=3, 사별=4}

순서척도 (ordinal scale) – 명칭 및 순서가 의미를 지님

- (예) 성적 등급 - {poor=1 , fair=2 , good=3 , very good=4 , excellent=5}

간격척도 (interval scale) – 명칭, 순서 및 간격이 의미를 지님

- (예) 온도

비율척도 (ratio scale) – 명칭, 순서, 간격 및 배율 모두 의미를 지님

- 이들 척도의 경우 이른바 "절대적 원점(absolute zero point)"이 정의됨
- (예) 키, 몸무게, 재산 등

3. 실험연구

실험 연구 대 경험적 연구

실험 연구와 경험적 연구는 다르다.

많은 경우 연구자는 특정 처리(예컨대, 대학 교육, 백신 투여 등)의 효과를 처리 집단과 통제집단간 반응(예컨대, 소득, 소아마비 발병률 등)을 비교함으로써 파악하고자 한다.

여기서 처리를 가한 집단을 처리집단(treatment group), 처리를 가하지 않은 집단을 통제집단(control group)이라고 부른다.

3. 실험연구

실험 연구의 집단 배정 원리

무작위 배정 (randomized control)

- 처리집단(treatment)과 통제집단(control)으로 구분
- 확률에 의존한 무작위 배정(예컨대, 동전 던지기에 의한 배정)

이중 눈가림(double blindness)

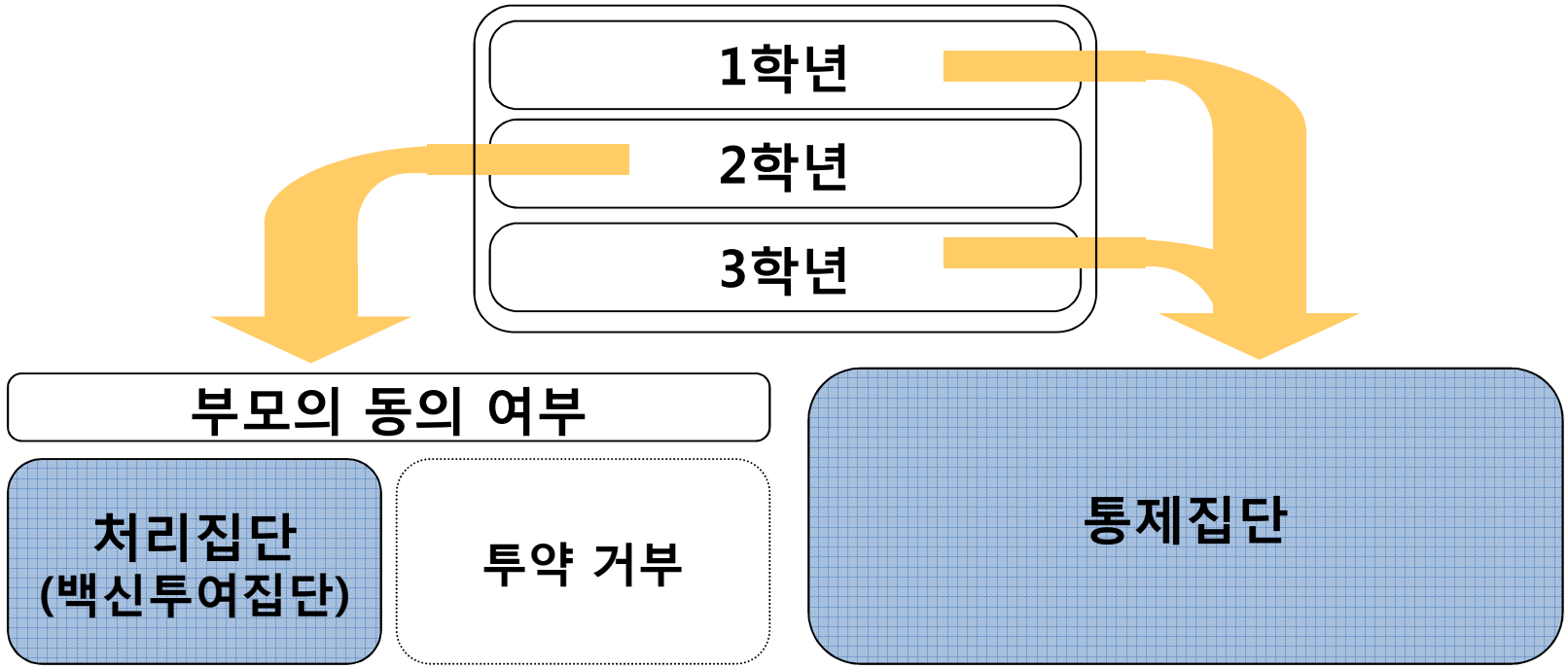
- 피험자가 본인이 처리를 받았는지 안 받았는지 모르게 조치하여 피험자의 심리적 효과 내지 위약효과(placebo effect)를 통제
- 실험자가 피험자의 소속집단을 모르게 조치하여 실험자가 피험자의 반응을 해석할 때 자의성이 개입되지 않도록 함

이상적인 실험

- 무작위로 통제된 이중 눈가림 실험

3. 실험연구

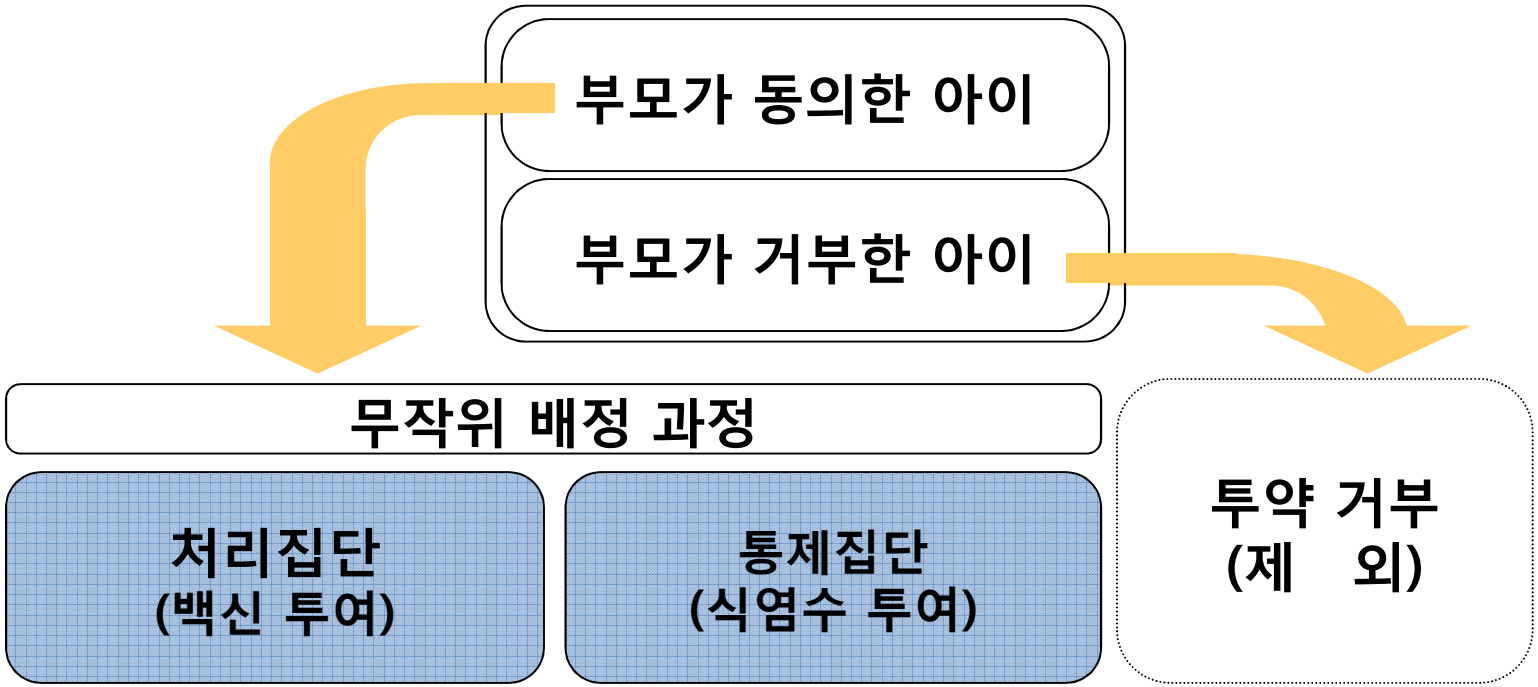
소아마비 재단 실험: 잘못된 통제



처리집단과 통제집단이 같은 표본에서 무작위로 추출되었는가? 아니다

3. 실험연구

무작위로 통제된 이중눈가림 실험: 잘 된 통제



부모가 동의한 아이들 중에서 처리집단과 통제집단에 무작위로 배정

학년 구분 제거하고 식염수가 든 가짜약 이용. 이중 눈가림 실험

3. 실험연구

잘못된 통제와 잘 된 통제의 비교

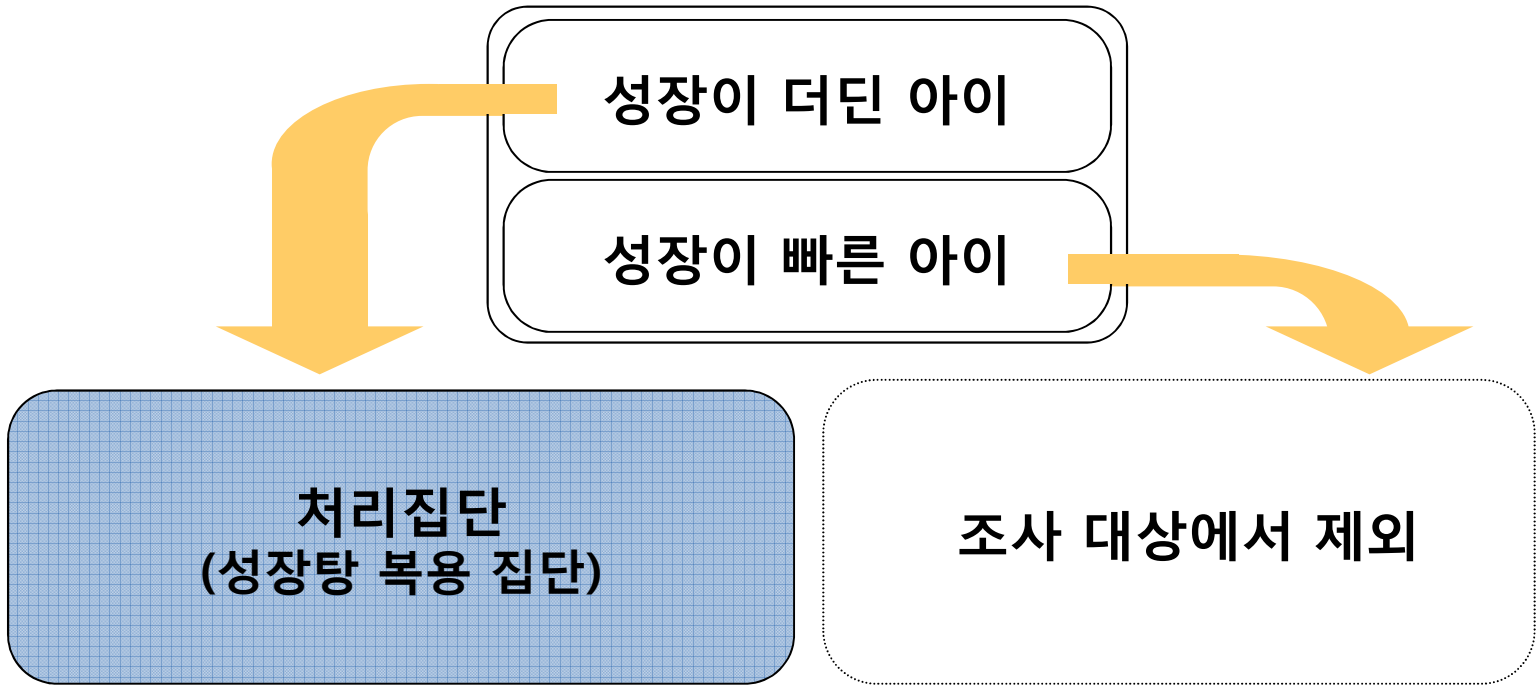
무작위 통제가 안 된 실험			무작위 통제된 이중눈가림 실험		
	표본 크기	발병률		표본 크기	발병률
처리집단	221,998	**25**	처리집단	200,745	**28**
통제집단	725,173	**54**	통제집단	201,229	**71**
투약거부집단	123,605	44	투약거부집단	338,778	46

주: 발병률은 10만명당 발병환자수를 의미함

출처: J.M.Tanner, et al., *Statistics* , 3rd ed., Wadworth & Brooks, p.12, Table1.

3. 실험 연구

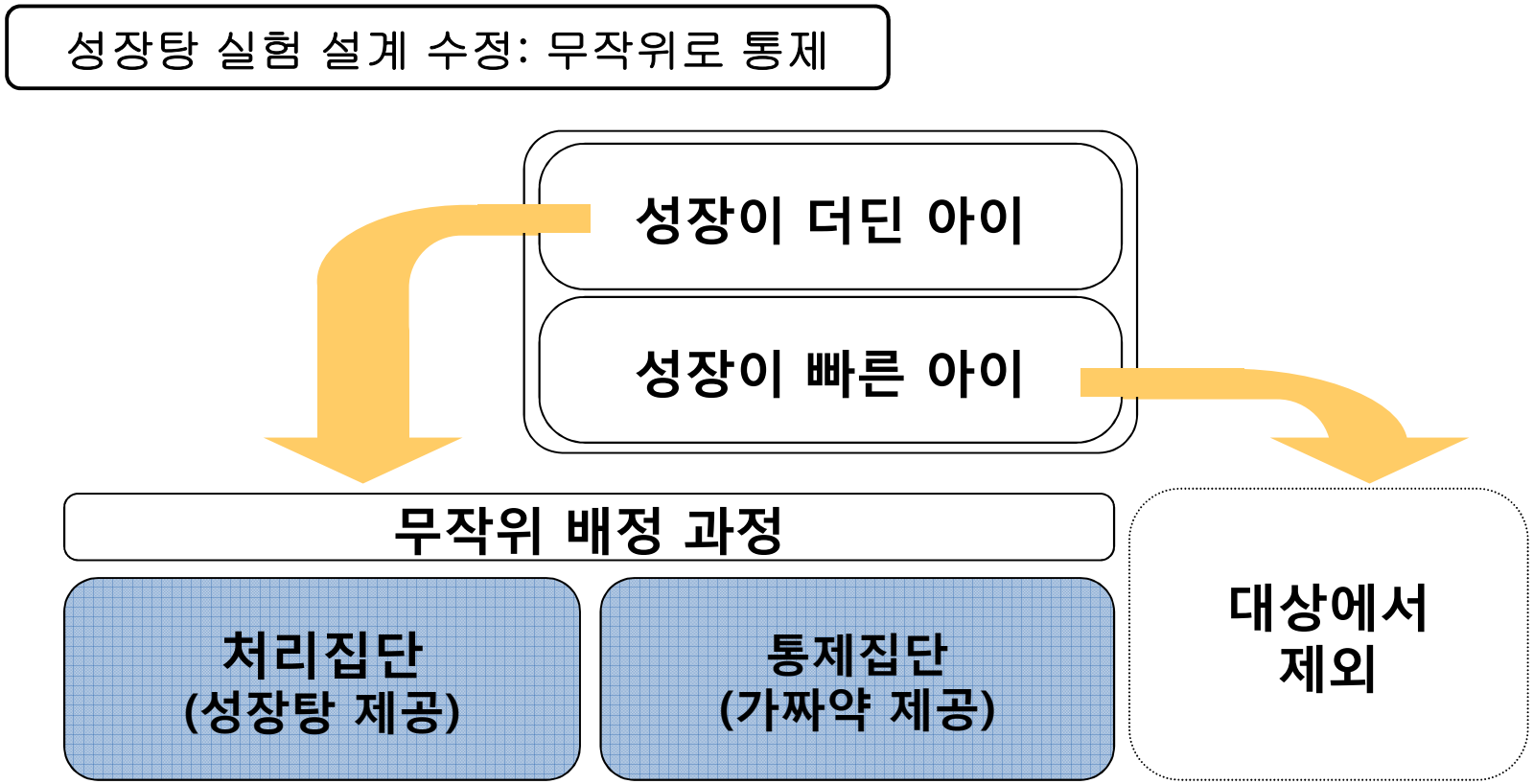
성장탕의 효과: 동시적 통제가 이루어지지 않음



성장탕 복용 '이전'과 '이후'가 비교 가능한가?

과거 성장이 더뎠던 아이들이 과연 '성장탕'의 영향으로 성장한 것인가?

3. 실험 연구



성장이 더딘 아이들을 처리집단과 통제집단에 무작위로 배정

4. 경험적 연구

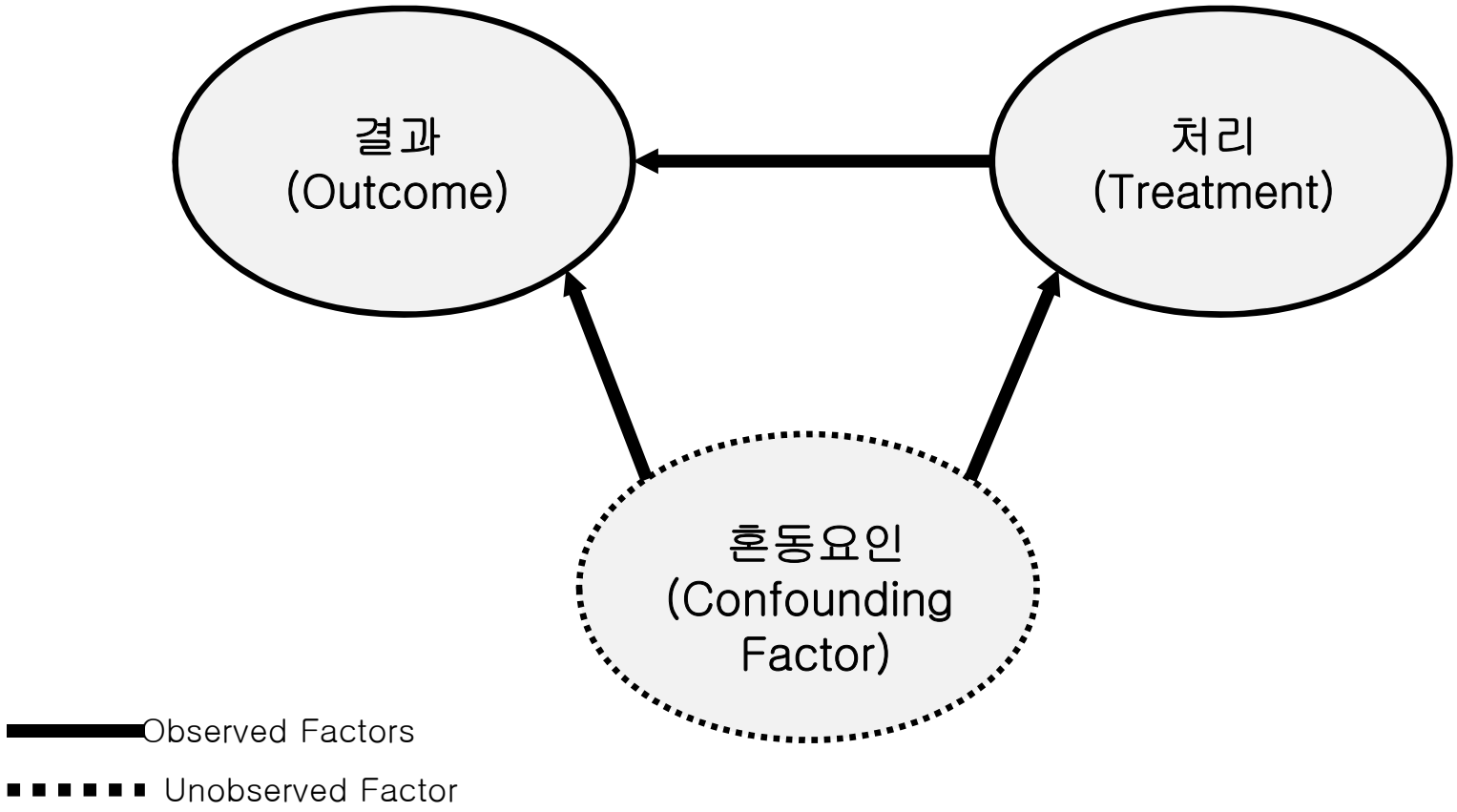
경험적 연구 대 실험 연구

경험적 연구는 통제된 실험과 달리 관측자료에 의존한다. 연구자는 개체의 행동이 가져오는 결과를 지켜볼 뿐이다.

- 흡연의 효과 연구 시 흡연자와 비흡연자의 행동이 가져오는 결과를 지켜볼 뿐
- 어느 누구도 연구자를 기쁘게 하기 위해 억지로 담배를 피거나 끊지 않는다.
- 실업자 재교육의 경우 실업자 스스로 직업훈련을 받을지, 안 받을지 결정한다.
- 최근에는 직업 훈련의 효과를 파악하기 위해 사회적 실험이 이루어지기도 한다. ("randomization out"의 개념 이용한 집단 배정)

4. 경험적 연구

경험적 연구에서 인과관계 추론 시의 문제점



4. 경험적 연구

혼동요인

통제되지 않은 제 3의 요인이 처리 여부와 관련이 있으면서 동시에 처리집단과 통제집단의 반응에 차별적인 영향을 주는 경우, 이러한 제 3의 요인을 혼동요인이라 한다.

- 예: 태아 때 초음파에 노출되면 출생 시 저체중이 초래되는가?
  - 태아에 문제가 있다고 느낄 때 초음파 검사를 하는 경향
  - 이는 역인과관계(reverse causality)

- 예: 처방을 잘 따르는 순응자(adherer)가 비순응자(non-adherer)보다 사망률이 낮은 것을 근거로 처방이 효과가 있다고 판단할 수 있는가?
  - 순응자와 비순응자는 건강에 대한 태도 및 생에 대한 애착 정도가 다르다.
  - 건강에 훨씬 더 관심이 있고 자신을 더 잘 보살피는 사람이 스스로 순응자가 된다.

4. 경험적 연구

심슨의 역설(Simpson's paradox)

하위집단에서 관찰된 관계는 하위집단들이 결합되었을 때 그 관계가 바뀌어 나타날 수 있다. 이를 **심슨의 역설**이라고 부른다. 심슨의 역설은 혼동요인을 통제할 필요성을 일깨워 준다

4. 경험적 연구

심슨의 역설(Simpson's paradox) 사례 1

예: 어느 한 대학원에 8,442명의 남성과 4,321명의 여성이 지원

- 남성 지원자의 약 44%가 합격
- 여성 지원자의 약 35%가 합격

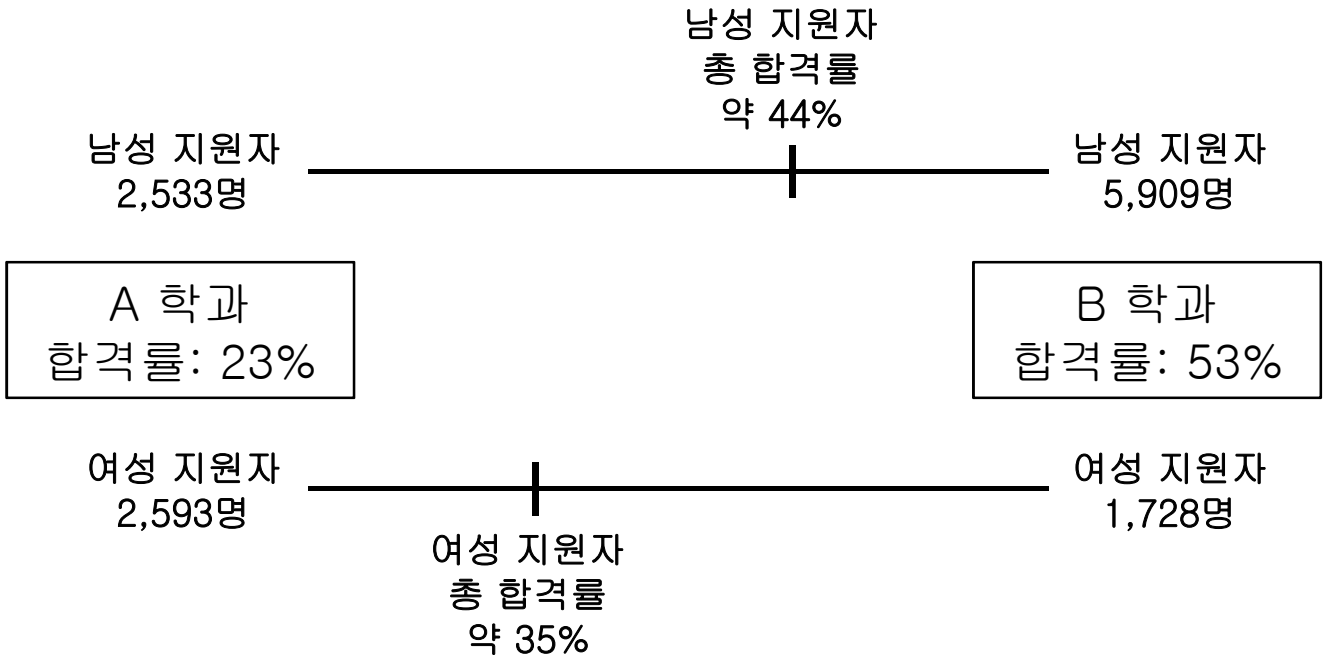
지원한 남성과 여성간 능력의 차이가 없다고 가정할 때

질문) 이 학교는 남성을 여성보다 44 대 35로 선호한다고 볼 수 있는가?

- 만약 여학생은 합격률이 낮은 학과에 몰리고 남학생은 상대적으로 들어가기 쉬운 학과에 몰렸으면 여학생의 전반적 합격률은 남학생에 비해 낮게 됨
- 학과별로 남녀 차별이 없었음에도 불구하고 전체적으로 볼 때는 남녀 합격률에 차이가 날 수 있음

4. 경험적 연구

심슨의 역설(Simpson's paradox) 사례 1



4. 경험적 연구

심슨의 역설(Simpson's paradox) 사례 2

예: 아파트 단지별로 따로따로 비교해 보면 새 아파트가 오래된 아파트보다 비쌈. 하지만 단지를 통합하여보면 이전에 지어진 아파트가 최근에 지어진 아파트보다 더 비싼 것으로 드러남. 이는 이전에 지어진 아파트가 대규모 단지에 위차해 있어 편의시설의 혜택을 보기 때문임

하위집단에서 관찰된 관계는 하위집단들이 결합되었을 그 관계가 바뀌어 나타날 수 있는 데 이를 심슨의 역설이라고 부름

심슨의 역설은 경험적 연구에서 적절한 통제의 중요성을 강조함

4. 경험적 연구

혼동요인의 통제

보다 동질적인 하위집단을 따로따로 비교함으로써 혼동요인에 대해 통제

- 예: 흡연자와 비흡연자간 사망률 단순 비교의 문제점 : 흡연자 중에는 남성이 많은데 남성은 여성에 비해 심장질환에 걸릴 가능성이 더 높다.

분리해서 비교	남성 흡연자와 남성 비흡연자 여성 흡연자와 여성 비흡연자

- 예: 나이에 따라 흡연습관이 다를 뿐 아니라, 폐암 발병률도 다르다.

더욱 세분해서 비교	나이든 남성 흡연자와 나이든 남성 비흡연자 나이든 여성 흡연자와 나이든 여성 비흡연자

4. 경험적 연구

자연실험 사례 I: 동경대학 프리미엄은 존재하는가?

- 동경대학 출신들과 타대학 출신들을 단순 비교하는 것은 부적절
- 자연실험: 일본의 동경대학은 1969년 학내사태로 인해 신입생을 선발하지 않음 (즉 동경대학 69학번은 없음)
- 69학번인 "다른 좋은 대학" 출신들은 학내사태가 없었으면 동경대학을 갈 수 있었음. 이는 마치 자연이 우리를 대신해 이들을 강제로 타대학으로 배정하는 실험을 해준 셈임. 자연실험
- 이들을 동경대의 인접 학번인 68 내지 70학번들과 비교
  - (i) 공무원 사회에서의 승진율
  - (ii) 민간부문에서의 승진율
- **동경대학 프리미엄은 다소 존재하고 민간 부문보다 공공 부문에서 강함**

4. 경험적 연구

자연실험 사례 II: 쌍둥이 연구

- 전쟁이나 병원에서의 "실수" 등으로 헤어진 뒤 뒤늦게 만난 쌍둥이를 비교
- 성장환경이 개인의 교육, 성격형성, 성공 등에 미치는 효과 분석

4. 경험적 연구

자연실험: 이중차분법, 회귀불연속 기법

- 이중차분법 (difference in difference): 비교의 비교, 즉 차이의 차이 이용하여 treatment effect 존재하는지 분석하는 기법

  -담합으로 인한 가격 상승률 추정

  -blind audition 채택으로 인한 오케스트라 여성 단원 채용 증가

- 회귀불연속 기법 (regression discontinuity): "아주 작은 차이=>처리집단과 통제집단의 구분=>두 집단간 통계적으로 의미 있는 결과의 차이 존재하는지" 분석

  -미국 하원의원 선거에서 현역 정당의 프리미엄이 존재하는가? 지난 선거에서 아주 미미한 득표율 차이로 당선이 결정된 지역만 대상으로 분석하여 현역 정당 소속 후보가 유의미하게 유리한지 분석

- 이중차분법, 회귀불연속 기법에 대해서는 별도 부록에서 좀 더 상술함

5. 통계학을 대하는 자세

통계학을 대하는 자세

잘못된 모형을 정확하게 푸는 것보다 올바른 문제를 근사적으로 푸는 게 낫다."

(Better to solve right problems approximately than to solve wrong ones exactly.)

"통계학은 미지의 세계에 대한 안내자이다." (Statistics is a guide to the unknown.)

"통계학은 스포츠와 같다. 토론하는 것보다 실천하는 게 낫다."

(Statistics is like sports. Better to practice than to discuss.)

5. 통계학을 대하는 자세

통계학을 대하는 자세

"모든 모형은 틀렸다. 다만, 몇몇 모형은 유용하다."

(All models are wrong, some are useful.)

"Everything should be made as simple as possible, but not simpler."

-Albert Einstein, 독일, 1879~1955-

"'Obvious' is the most dangerous word in mathematics."

-E. Bell, 스코트랜드, 1883~1960-