표집
Sampling

CE730 통계와 금융

조남운

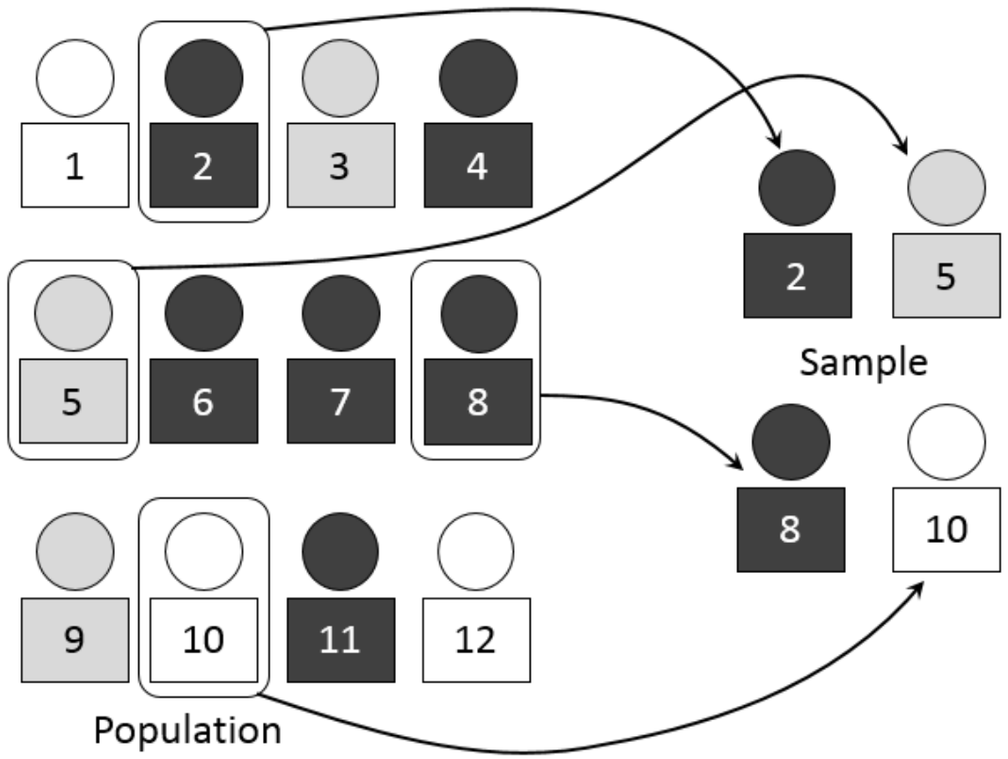
Namun Cho/ mailto:namun@snu.ac.kr

주제

- 표집의 개념 및 용어
- 확률표집
- 비확률표집
- 편의 (bias)

표집
Sampling

- 모집단의 일부인 표본(sample)을 선정하여 측정/조사하는 행동
- 모집단을 잘 대표할 수 있어야 함



https://en.wikipedia.org/wiki/Sampling_(statistics)#/media/File:Simple_random_sampling.PNG

Case Study: 리터러리 다이제스트 vs. 갤럽

- 1936년 미국 대통령선거
  - 공화당 후보: 알프레드 랜던
  - 민주당 후보: 프랭클린 루즈벨트
- 두 집단이 설문 조사 실시함
  - 리터러리 다이제스트
  - 조지 갤럽

https://www.youtube.com/watch?v=kGbKZ2ihYVk

리터러리 다이제스트의 표집 방법

- 미국의 유권자 1천만명에게 우편으로 설문지를 발송
  - 240만명에게 응답을 받음
  - 역사적으로 유례없이 압도적인 표본수였음
- 분석 결과 알프레드 랜던 후보의 당선을 예측

https://en.wikipedia.org/wiki/The_Literary_Digest#/media/File:LiteraryDigest-19210219.jpg

Namun Cho/ mailto:namun@snu.ac.kr

갤럽의 표집방법

- 1500명을 대상으로 면접조사를 실시함
- 프랭클린 루스벨트가 56%의 지지율로 당선될 것이라 예측

https://en.wikipedia.org/wiki/George_Gallup#/media/File:George_Gallup.png

결과

- 루스벨트가 62%로 당선됨
- 질문1: 왜 압도적으로 많은 표본을 대상으로 조사한 리터러리 다이제스트는 예측에 실패했는가?
- 질문2: 왜 1500명에 불과한 표본을 대상으로 조사한 조지 갤럽은 예측에 성공했는가?


THE BALTIMORE NEWS
The Largest Daily Circulation in the Entire South
COMPLETE MARKETS
Hoover Loses New York State By 500,000
ROOSEVELT ELECTED IN DEMOCRATIC LANDSLIDE
ROOSEVELT CARRIES CITY BY 2 TO 1
President and Vice President-Elect
NEW U. S. EXECUTIVE WINS NEW YORK BY 500,000; SOUTH SOLID AGAIN
Final Returns May Show Victory of Unprecedented Proportions, Carries Chicago By Some

Namun Cho/ mailto:namun@snu.ac.kr

관련용어

- 전수조사 Complete Enumeration Survey
- 표본조사 Sample Survey
- 표집틀 Sampling Frame
- 표본모집단 Sampled Population
- 목표모집단 Target Population

전수조사
Complete Enumeration Survey

- 모집단 전체를 조사
- 장점: 모수 그 자체를 조사한 것
- 단점: 고비용
- 예
  - 통계청의 사업체 기초통계조사
  - 인구센서스
    - http://www.census.go.kr/mainView.do
  - Twitter Crawling

표본조사
Sample Survey

- 모집단의 일부를 조사
- 장점: 조사의 제약이 적음
- 단점: 모수와의 차이 (오차) 발생
- 대부분의 조사가 여기에 해당
  - 선물시장통계조사
  - 금융기관 대출행태조사
  - 기업자금사정실태조사

표집틀
Sampling Frame

- 표본을 구하기 위한 모집단의 목록
- 리터러리 다이제스트의 표집틀
  - 1천만명의 주소 수집
  - 잡지의 정기구독자
  - 전화번호부
  - 자동차 등록명부
  - 사교클럽 인명부

표본모집단, 목표모집단

- 표본모집단 (Sampled Population)
  - 표본이 구해지는 모집단
- 목표모집단 (Target Population)
  - 정보를 얻고자 하는 모집단
- 이상적인 상황이라면 이 두 모집단은 같아야 함
- 현실에서는 격차가 발생

표본모집단≠목표모집단

- 표집을 위해서는 표집틀이 필요
- 표집틀 자체를 항상 구할 수 있는 것이 아니라는 것이 문제
- 예: 통계청 기업체 업종별실태조사 (2003)
  - 2003년 현재 전체 기업체 리스트가 존재하지 않음 ⇒ 2001년 사업체기초통계조사(전수조사)를 표집틀로 사용하여 표집
    - 표본모집단: 2001년 전체 사업체
    - 목표모집단: 2003년 전체 사업체

표집의 종류

- 확률표집 Probability Sampling
  - 모집단의 각 개체가 표본이 될 확률이 동일
- 비확률표집 Non-probability Sampling
  - 모집단의 각 개체가 표본이 될 확률이 다름

확률표집의 종류
Probability Sampling

- 단순임의표집 Simple Random Sampling
- 층화임의표집 Stratified Random Sampling
- 계통임의표집 Systematic Random Sampling
- 집락임의표집 Cluster Random Sampling
- 다단계 표집 Multi-stage Sampling

Notations

기호	의미
N	모집단의 크기
n	표본의 수
f	표집 비율

단순임의표집
Simple Random Sampling

- 임의 표집 방법 중 가장 간단
- 기본적 표집방법
- 절차
  - 표집틀에 색인번호(index)를 부여: 1 ... N
  - 1~N 중에서 난수 (random number) n 개를 생성
  - 난수에 해당하는 대상이 표본이 됨

층화임의표집
Stratified Random Sampling

- 모집단을 몇 개의 그룹으로 나눈 뒤 각 그룹 안에서 단순임의표집을 수행
- 그룹 선정의 기준
  - 그룹내에서는 가능한한 동질적일 것
  - 그룹 사이에서는 이질적일 것
- 예:
  - 지역 그룹 층화 표집
  - 소득 수준 층화 표집

계통임의표집
Systematic Random Sampling

- 단순임의표집은 표본 수 (n) 만큼의 난수가 필요함
- 계통임의표집은 하나의 난수로 표본을 추출
- 절차
  - 표본추출간격 K := N/n 계산
  - 1-K 중 하나의 숫자를 임의로 선택 (K0)
  - K0을 시작점으로 K번째 표본을 선택
- **표집틀 순서에 패턴이 있을 경우 사용하면 안됨**

집락임의표집
Cluster Random Sampling

- 조사비용 절감을 위한 표집 전략
- 절차
  - 모집단을 집락(Cluster)으로 분할
    - 통상적으로 행정구역 등으로 분할
  - 집락을 단순임의표집
  - 표본 집락에 대해 전수조사

Namun Cho/ mailto:namun@snu.ac.kr

다단계표집
Multi-stage Sampling

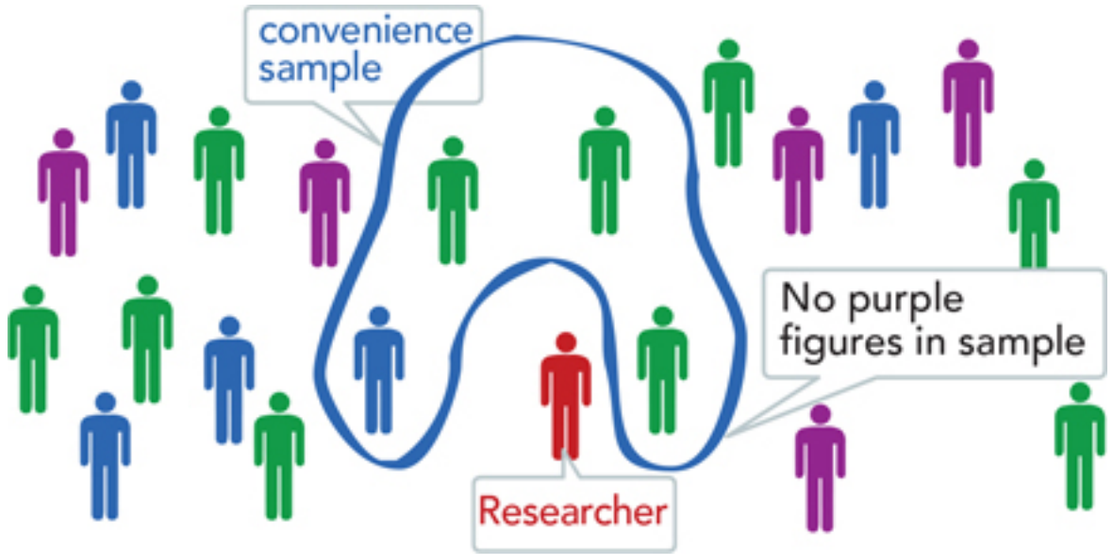
- 단순임의, 층화임의, 계통임의, 집락임의표집을 복합적으로 사용
- 대부분의 조사들이 채택하는 방식
- 예: 서울지역조사 (2단계표집)
  - 1단계: 집락임의표집: 행정동을 추출
  - 2단계: 층화임의표집: 추출한 행정동 내에서 층화임의표집 (연령)

비확률표집
Non-probability Sampling

- 추출된 표본이 모집단을 잘 대표한다는 보장이 없음
- 표집틀을 만들 수 없는 경우 차선책으로 사용
- 편의표집 Convenience Sampling
- 목적표집 Purposive Sampling
- 할당표집 Quota Sampling
- 눈덩이표집 Snowball Sampling

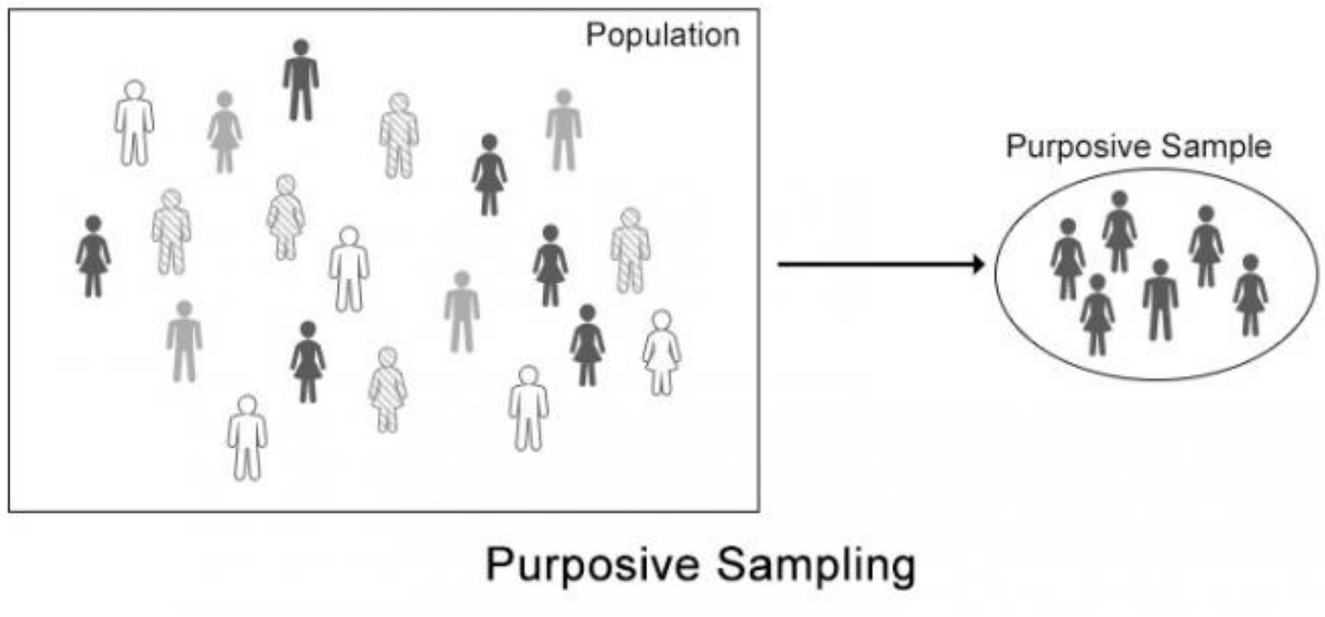
편의표집
Convenience Sampling

- bias의 편의가 아님
- 거리에서 눈에 띄는대로 인터뷰 (조사) 를 하는 표집
- 장점: 저비용
- 단점: 데이터의 신뢰성 부족



목적표집
Purposive Sampling

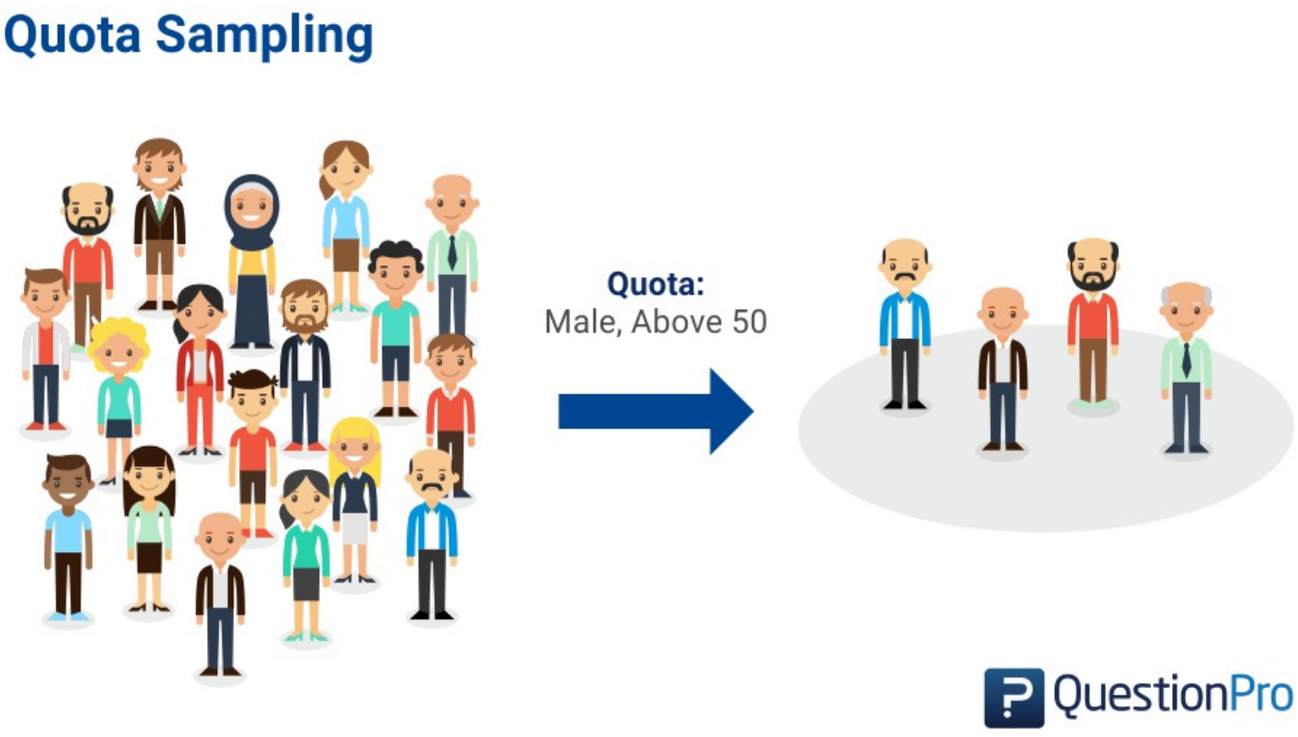
- 특정 특성을 가지는 모집단의 조사를 할 때 사용
- 예: 흡연자 대상 설문
  - 거리에서 일반인에게 흡연자인지 질문
    - YES: 설문시작
    - NO; 다른 사람에게 감



https://research-methodology.net/sampling-in-primary-data-collection/purposive-sampling/

할당표집
Quota Sampling

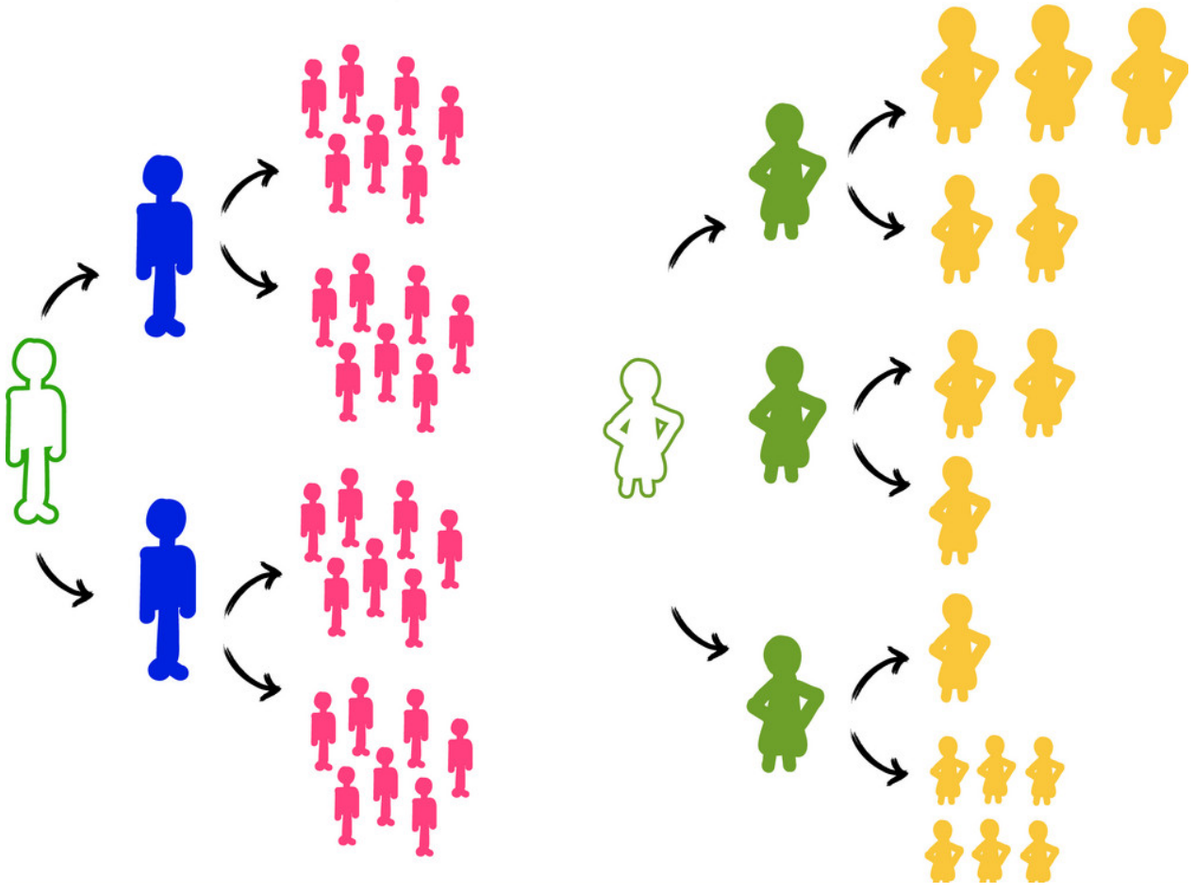
- 층화임의표집과 유사
- 그룹별로 표본수를 할당한다는 측면에서는 동일
- 그룹 내 표본을 편의표집 등 비확률 표집으로 추출한다는 측면에서는 다름



https://www.questionpro.com/blog/quota-sampling/

눈덩이표집
Snowball Sampling

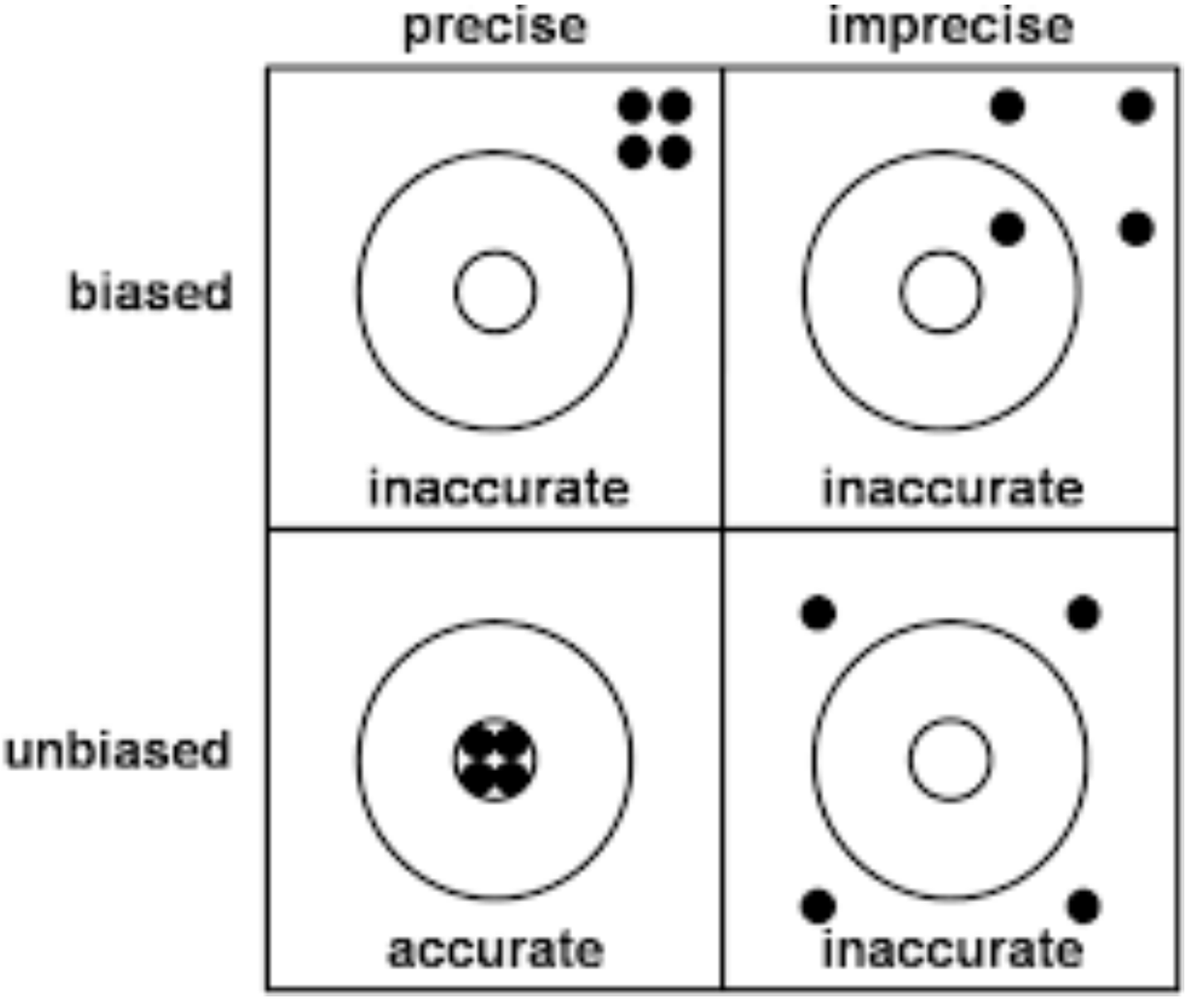
- 특정 성질을 가지는 사람을 조사 후 동일 특성을 가지는 사람을 소개받는 방법
- 예: 페이스북 유저 샘플링
  - 페이스북 유저 조사후,
  - 해당 유저의 팔로워를 조사



https://www.vectorstock.com/royalty-free-vector/snowball-sampling-the-sampling-methods-vector-13352210

편의
Bias

- 모수와 표본에 의한 모수추정치간의 차이
- 가능한한 제거되어야 할 것



편의의 종류

- 데이터스누핑 편의 Data-snooping Bias
- 데이터마이닝 편의 Data Mining Bias
- 표본선택 편의 Sample Selection Bias
- 시간 편의 Time-period Bias

데이터스누핑 편의 Data-snooping Bias

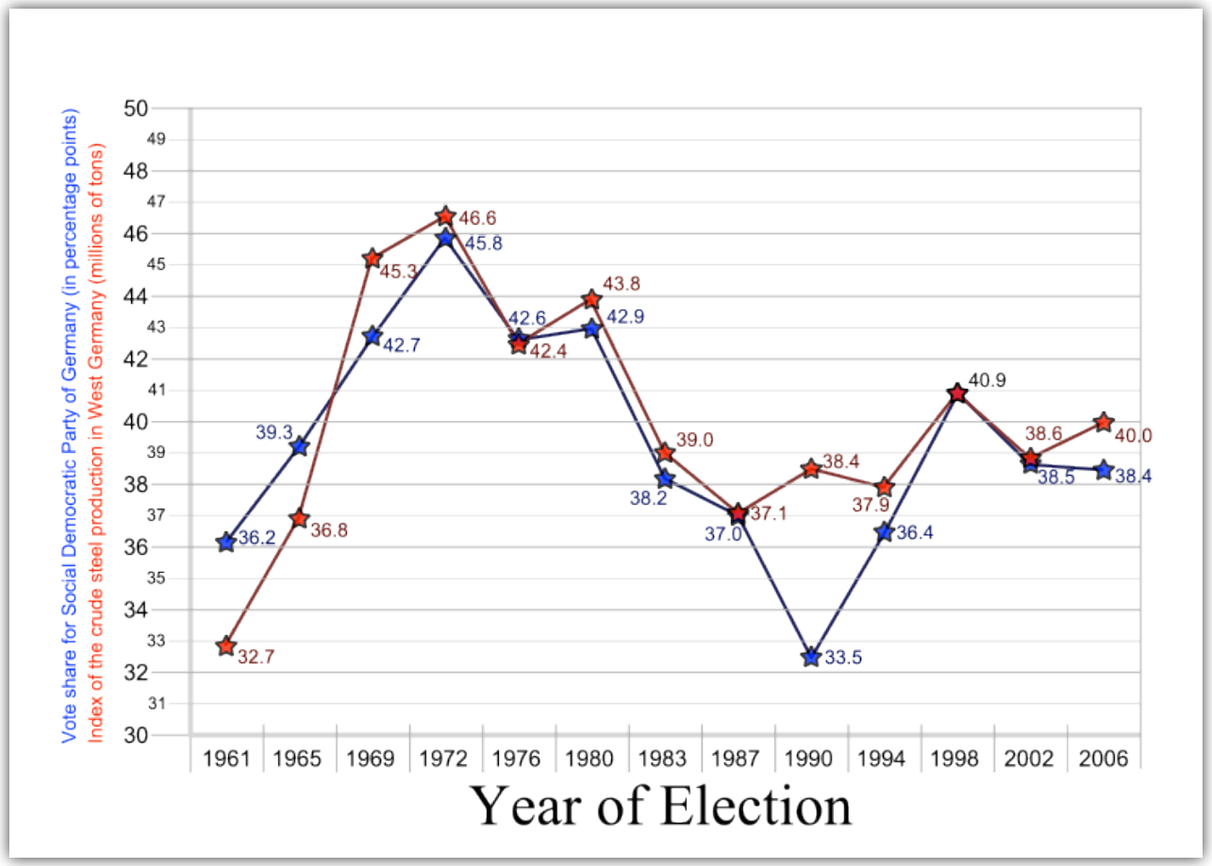
- 주어진 표본을 추론, 모델 선택 등에 반복적으로 이용하는 과정에서 실제 모수와 차이가 있는 결과를 얻게 되는 상황
- 예
  - 약품의 효과를 추정하는 과정에서 별다른 유의한 관계가 관찰되지 않았음. 하지만 이 과정에서 성별에 따른 효과 검토결과 여성그룹에서 유의한 관계가 관찰됨 $\Rightarrow$ 가설을 수정
  - 이러한 수정된 가설은 표본편의에 의한 오류로 만들어진 가설이 될 가능성이 존재함

데이터마이닝 편의
Data Mining Bias

- 조사연구에 대한 가설을 표본의 특성으로부터 추출할 때 발생할 수 있는 편의
- 데이터스누핑 편의와 밀접한 연관
- 예
  - 미어샤이트의 법칙

미어샤이트의 법칙

- 독일 사민당의 득표율과 해당 년도의 조강(crude steel) 생산량간의 매우 강력한 양의 상관관계 존재
- 인과관계가 없는 상관관계의 사례



Namun Cho/ mailto:namun@snu.ac.kr

표본선택 편의
Sample Selection Bias

- 특정 성격을 가진 모집단의 구성요소가 표본 선정 방식의 결함으로 포함되지 않는 경우 발생
- 예: 리터러리 다이제스트의 1936년 표본표집
  - 주로 전화번호부로부터 표본틀을 설정
  - 이러한 경우 전화가 없는 가정 (1936년에 전화기는 가난한 집에 없는 경우가 많았음)은 표본에서 제외됨 - 저소득층의 표본선정되지 못함

금융 통계에서의 예

- 성장주와 가치주의 비교
  - 성장주: 수익신장률이 높은 기업의 주식. 앞으로 고성장할 것으로 기대되는 기업의 주식. 주당 순이익에 비해 높은 가격에 거래됨
  - 가치주: 주당 순이익에 비해 낮은 가격에 거래되는 주식 (저평가주)
- 방법: 10년전 성장주 50개와 가치주 50개의 현재 성과를 비교 $\Longrightarrow$ 10년간 파산합병된 기업이 배제되므로 표본선택 편의 발생 가능성이 높음

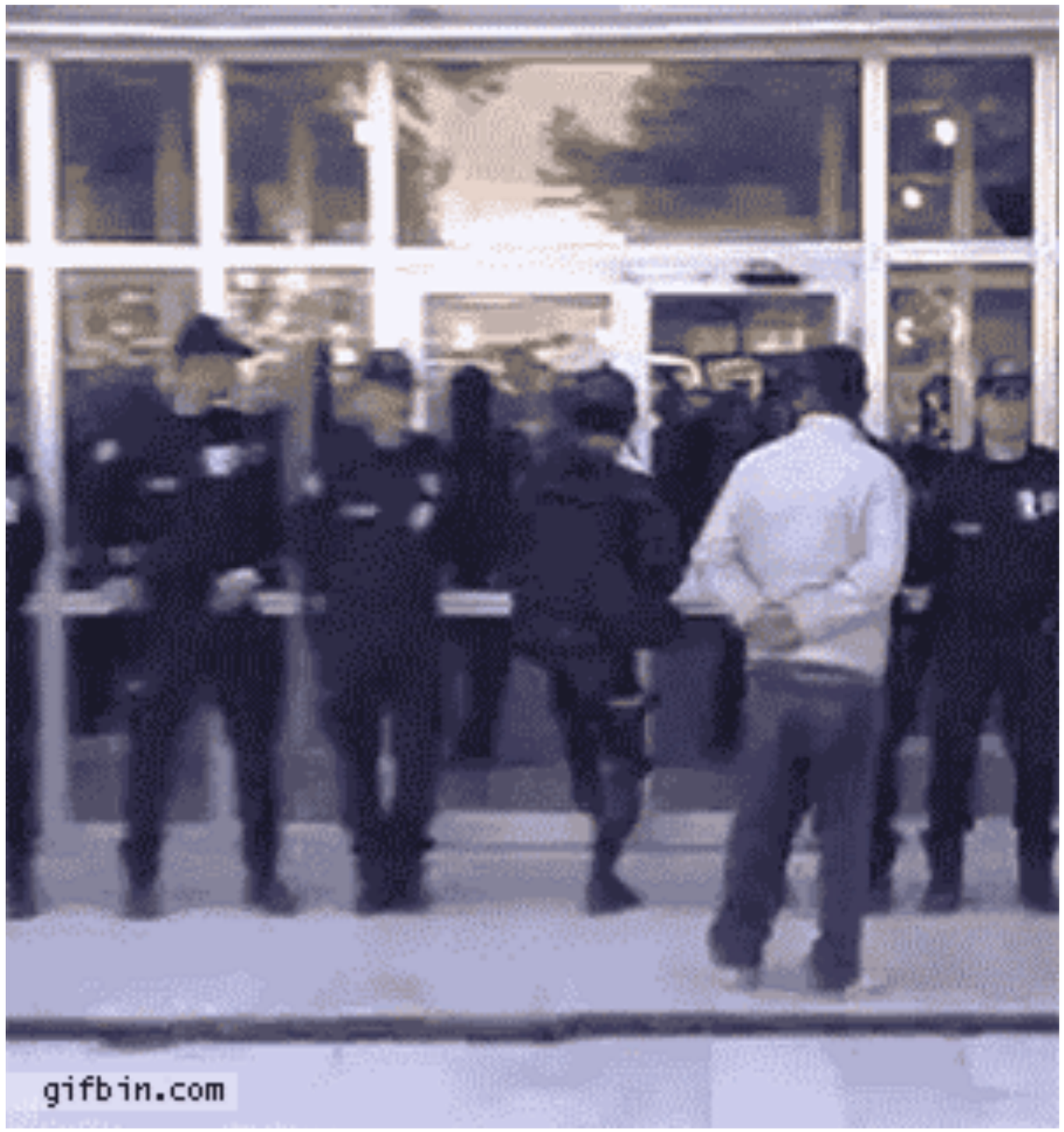
시간편의
Time-period Bias

- 자료수집기간동안 외부의 개입 (충격 shock)이 있을 때 발생하는 편의
- 자료 수집 기간 동안 시간의 흐름에 따라 모집단에 발생한 근본적 변화가 편의의 원인이 됨

Next Topic

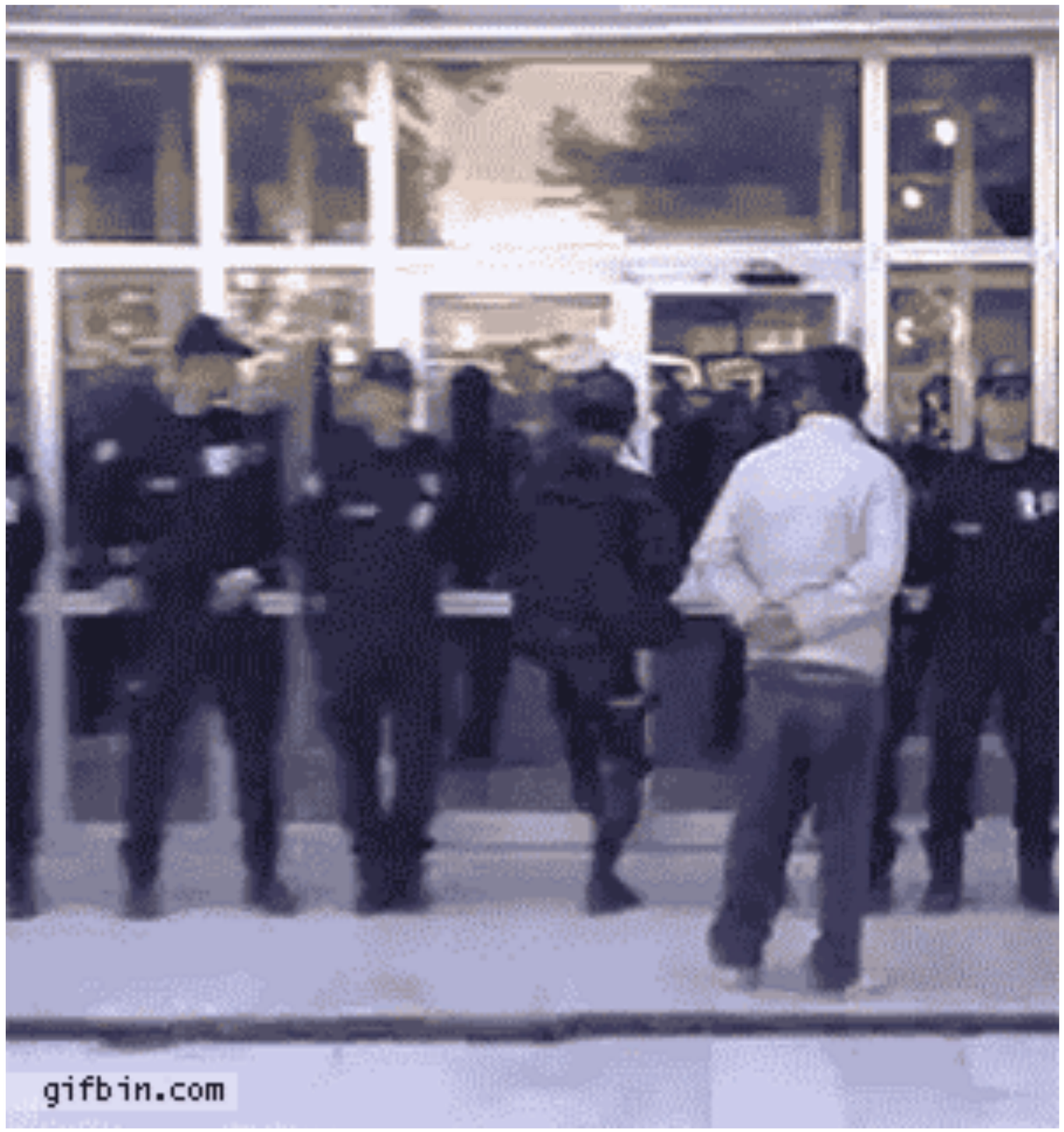
- 기술통계

수고하셨습니다!



Namun Cho/ mailto:namun@snu.ac.kr

수고하셨습니다!



Namun Cho/ mailto:namun@snu.ac.kr