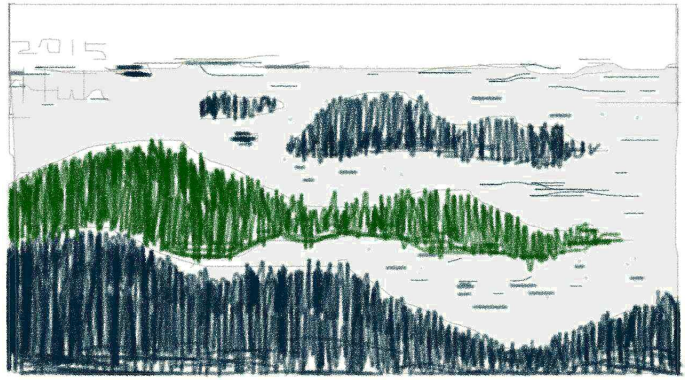
데이터 분석의 철학과 과학성

허 명 회 (고려대 교수, 통계학)

stat420@korea.ac.kr



2015.06.26. 세종대학교

키워드

- EDA vs CDA
- 빅데이터
- 데이터 분석의 철학
- 과학성의 저해 요소
  - Disguised EDA
  - 교락성(confounding)의 위험
  - 재현성 평가의 부재: 유의성 검정의 p-값과 신뢰구간
  - 예측과 설명의 분리
  - 성급한 기대감

데이터 분석의 유형

EDA, exploratory data analysis (탐색적 데이터 분석)

- 데이터의 특징과 구조에 대한 탐구 exploration
- 인사이트의 생성, 가설과 모형의 도출
- 先데이터, 後분석

CDA, confirmatory data analysis (확증적 데이터 분석)

- 인사이트, 가설, 모형의 타당성, 일반성, 재현성 평가
- 모형 적합도, 가설검정, 신뢰구간
- 계획 → 데이터 확보 → 분석

Reference: Tukey, J. (1962). "The future of data analysis". Annals of Mathematical Statistics, 33. 1-67.

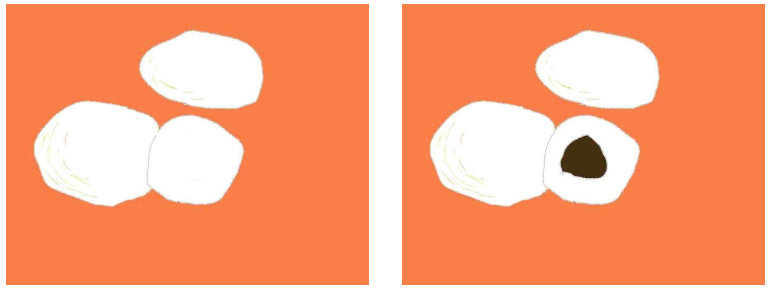
EDA vs CDA

보기 1. 감기에 걸리는 사람들과 걸리지 않는 사람들 간에 어떤 차이가 있는가를 수 십 가지 측면에서 살펴보았다. 그 결과, 비타민 C를 복용하는 사람들이 감기에 잘 걸리지 않음을 알게 되었다. 그러면, 비타민 C를 복용하면 감기에 덜 걸리게 된다고 말할 수 있는가? [EDA]
EDA로 이에 대한 답을 하긴 어렵다. 비교실험을 설계하여 새로 자료를 수집해 가설을 확인해볼 필요가 있다 [CDA].

보기 2. 대형마켓에서 고객들의 구매 내역 자료를 분석한 결과, 일부 고객들은 다른 고객들에 비해 유기농 식재료 비중이 크게 나타났다. 그들이 어떤 생각을 하는 사람들인가? 이에 대해 몇 개의 추측이 생성되었다. [EDA]
추측이 맞는가? 이를 확인하기 위하여 전체 고객의 일부를 선택하여 몇 가지 인구사회적 속성과 연소득, 그리고 소비와 삶에 대한 태도를 조사하여 구매 내역과 연결해 확인하였다 [CDA].

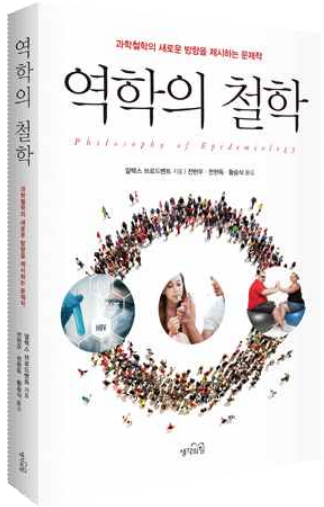
빅데이터 vs 스몰데이터

- Volume, Velocity, Variety, Veracity, ...
- 빅데이터는 EDA의 대상
- CDA가 없는 EDA는?



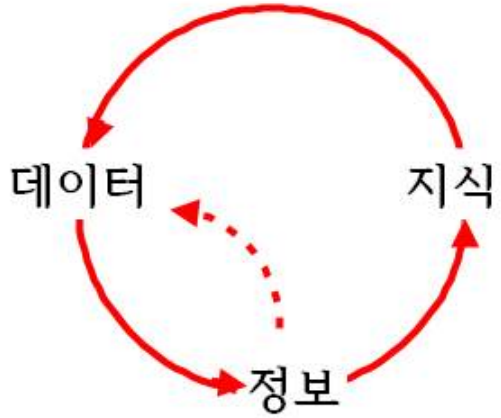
사례 1: 산욕열 puerperal fever

Ignaz Semmelweis, 1844 at Vienna General Hospital.



- Ward 1: Mortality rate 16%
  Ward 2: Mortality rate 2%
- Catholic Priest?
- Autopsy Class?

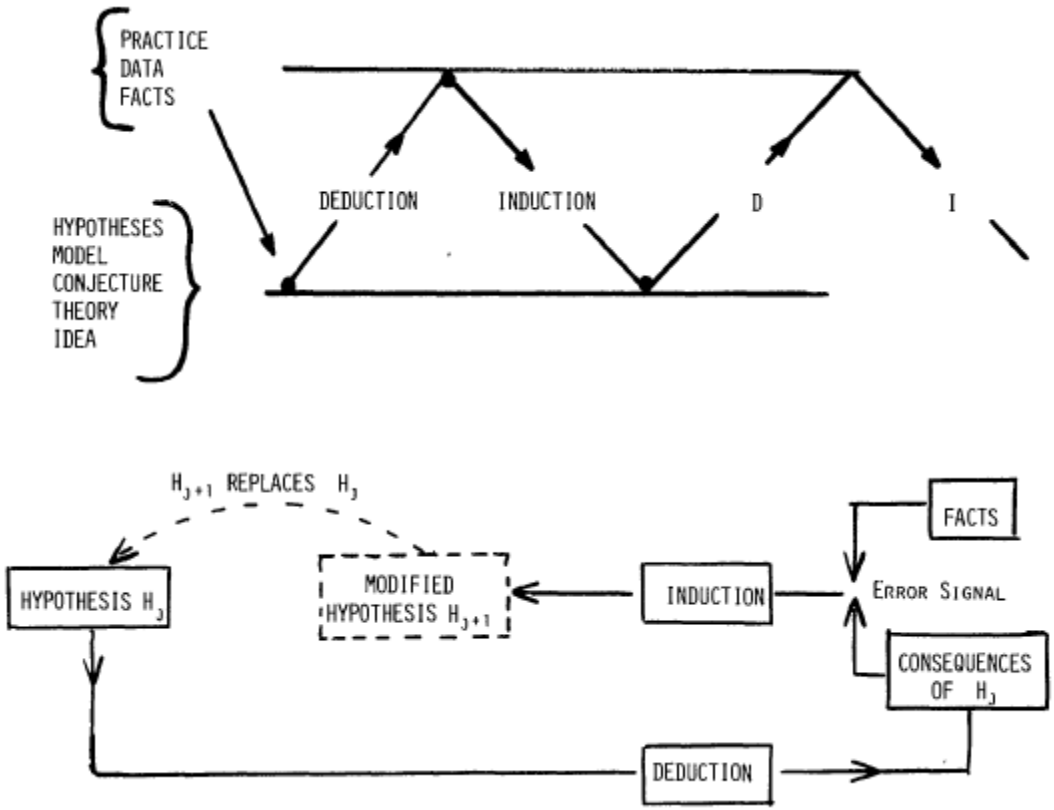
데이터 분석과 활용의 철학



- 증거주의: 데이터로 확인이 되는가?
- 재현성: 거듭 확인이 되는가?
- **實事求是**의 철학

"All models are wrong, but some are useful," George Box (1919-2013).

Advancement of Learning



Reference: Box, G. (1976). "Science and statistics". Journal of the American Statistical Association, 71, 791-799.

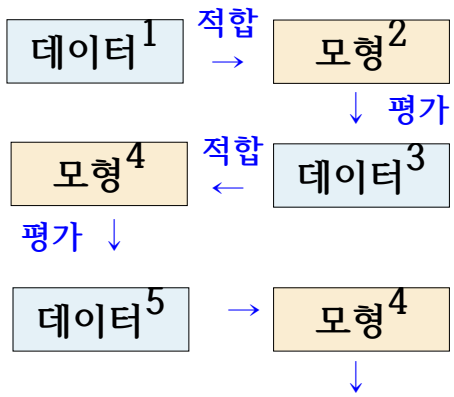
데이터와 모형의 진화

복잡성(複雜性, 현실)의 이해: 현실 세계는 많은 요인이 서로 얽혀져있다.
그렇다고 절대적으로 불가지(不可知)한 것은 아니다.

데이터: 우리가 포착한 복잡계의 한 단면이다.

모형(模型, model): 모형은 우리가 현실을 이해하는 '틀'이다. 진짜는 아니다.

데이터-모형의 진화 (evolution):



과학성의 저해 요소

- Disguised EDA
- 교락성(confounding)의 위험
- 재현성 평가의 부재
- 예측과 설명의 분리
- 성급한 기대감

Disguised EDA

- 분포의 탐색. $x_1, \cdots, x_n \sim f_\theta(x)$, θ is a location parameter.

  [정규성 검정] $H_0 : x_1, \cdots, x_n \sim N(\theta, \sigma)$ vs $H :$ *not* H_0.

  If H_0 is not rejected, assume $x_1, \cdots, x_n \sim N(\theta, \sigma)$
  and apply parametric methods.

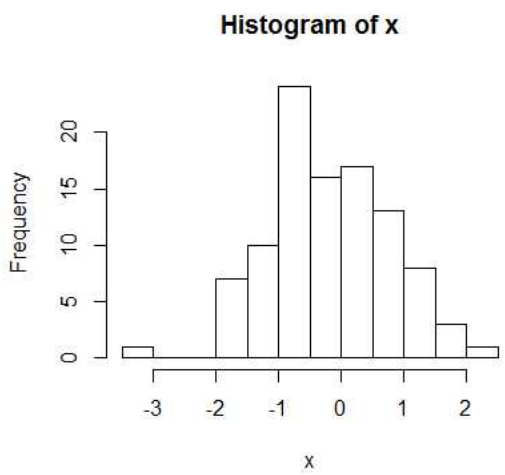
  If H_0 is rejected, apply non-parametric methods.

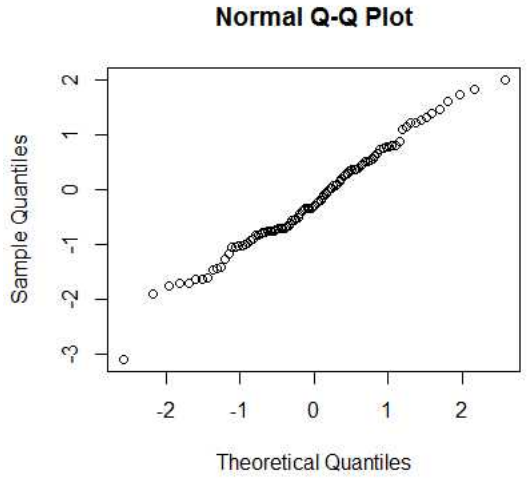
  - 통계적 검정은 H_1을 보이는데 목적이 있다.
    H_0를 정당화하기 위해 통계적 검정을 써서는 안 된다.
  - 통계적 검정은 최종 단계에서만 써야 한다.

Disguised EDA

- 분포의 탐색. $x_1, \cdots, x_n \sim f_\theta(x)$, θ is a location parameter.

  [EDA]



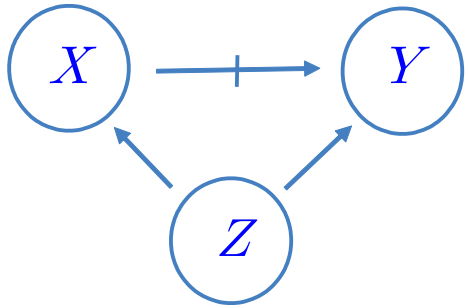


  - EDA에는 예술적 측면이 있다. 객관적이어야 한다는 부담감이 지나치다.
  - 그림을 그려야 한다. 숫자에만 의존하지 말라.

교락성(交絡性, confounding)

- 연관성이 인과성을 의미하지 않는다. Correlation does not mean causation.

  특히 비실험 자료에서는 그렇다. 그러나 지나친 엄격성은 곤란하다.



  - 다양한 가능성을 검토하고 확인해야 한다.

    예: Berkeley Admissions Data

재현성(reproducibility)

- 재현성 평가의 부재는 실험연구에서도 문제이다. 관측연구에서 재현성은 더욱 작다. 수많은 "차이의 발견"이 시도된 결과로 얻은 차이는 일과적일 가능성이 충분히 있다.
- 독립적인 Validation이 필요하다.

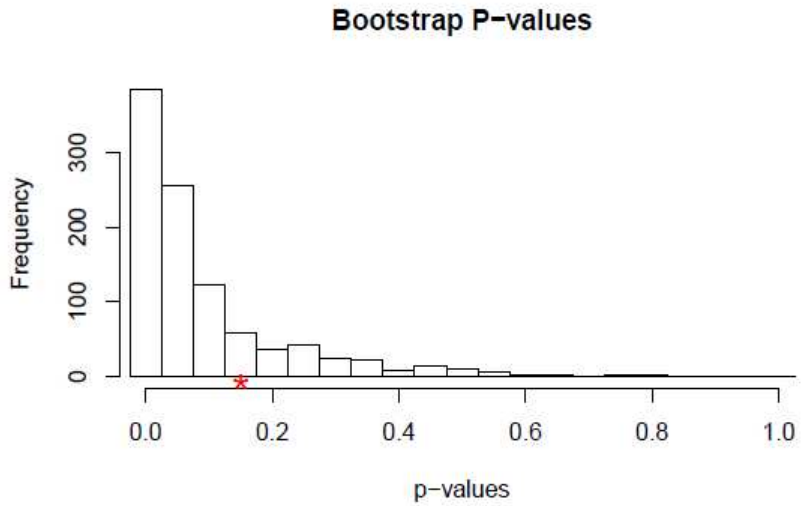
재현성(reproducibility)

- p-값이란? 가설검정에서

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta > 0$$

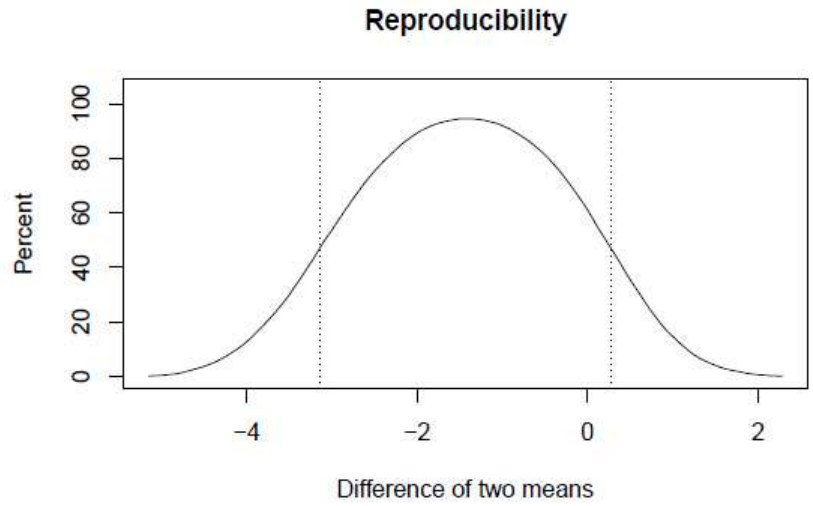
  - p-값이 0.05인 경우, 같은 크기의 재현적 연구에서 이것이 유의하게 나타날 확률은 50%이다. 실제로는 여러 이유에 의해 이보다 작다.

  - p-값도 확률변수이다. 아래는 p-value =0.048인 한 예.



재현성(reproducibility)

- 신뢰구간도 확률변수이다. 아래는 신뢰구간이 $(-3.14, 0.28)$ 인 한 사례.



- 신뢰구간 끝 값의 재현율은 50%이다. 신뢰구간 중앙점의 재현율은 95%.

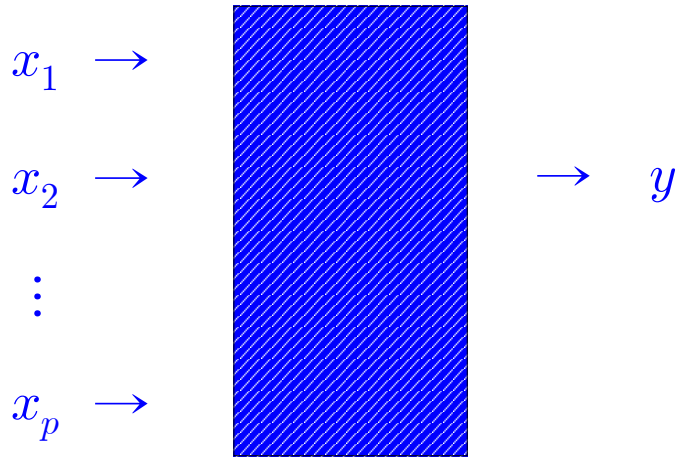
Reference: 허명회 (2014). "가설검정과 신뢰구간의 재현성", 응용통계연구, 24(4), 645-653.

예측과 설명의 분리

- 기계학습(machine learning) vs. 일반화선형모형

  - 출력모형: $\hat{y} = f(x_1, \cdots, x_p)$

  - 기계학습은 설명이 쉽지 않다. 예측변수 x_j가 어떤 역할을 하는지?



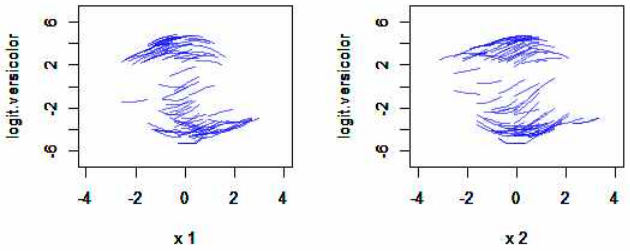
  - 설명 없는 예측이 과학(science)인가?

Reference: Shmueli (2010). "To explain or to predict", Statistical Science 25(3), 289-310.

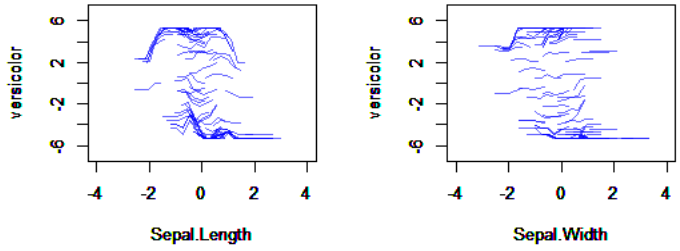
예측과 설명의 분리

- 다변수 함수 $\hat{y} = f(x_1, \cdots, x_p)$의 시각화
  - 사례: iris *Versicolor vs Virginica*

support vector machine



random forests



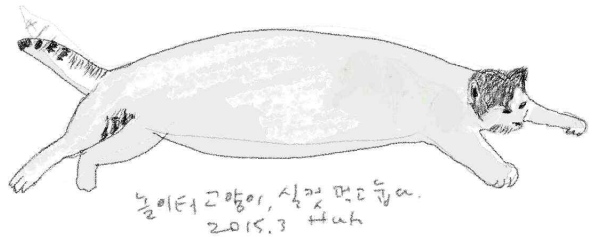
Reference: 허명회 (2014). "예측함수의 시각화", <응용데이터분석> 20장.

조급증(躁急症)

- 성급한 기대감

→ 선순환 cycle의 붕괴

→ 빅데이터에 대한 회의감



맺음말

- Data Analytics의 발전
  - □□과 □□□의 기반
  - 개인역량의 계발
  - Best Practice의 공유, 생태계 조성