

한국자료분석학회 (충북대학교)

우리들의 학문, 그 너머의 무엇

What I learned from what I taught and studied

허명희 (고려대학교 통계학과 명예교수)

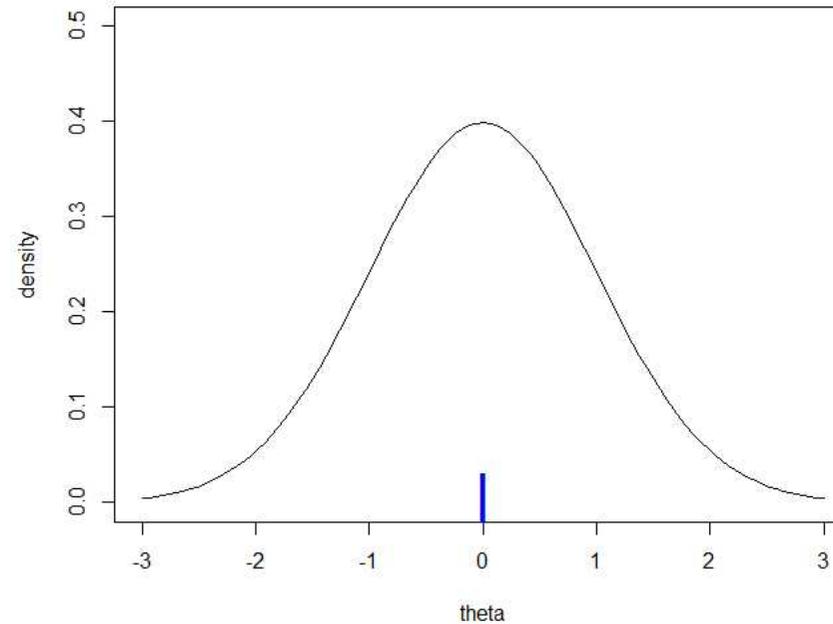
stat420@korea.ac.kr

2025년 11월 28일

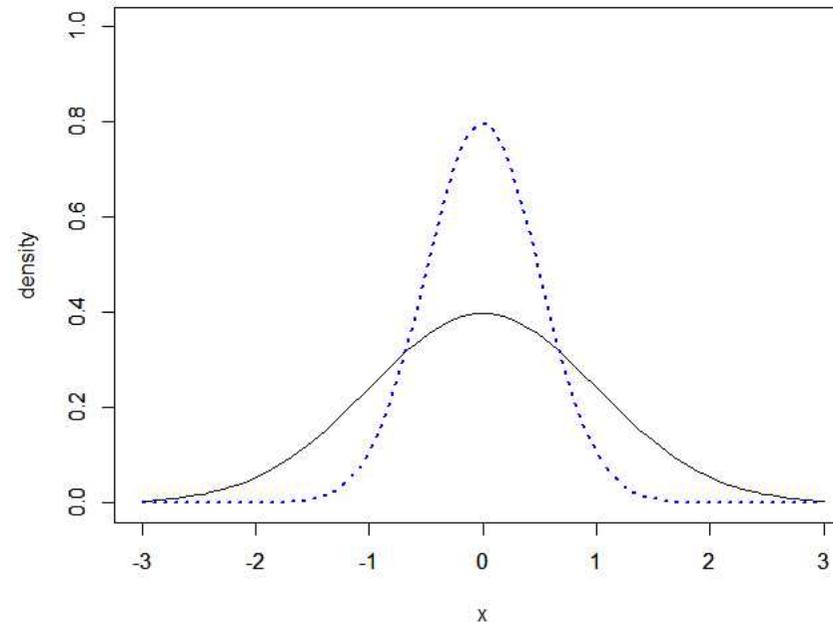
15개의 키워드

- 비편향성
- 최소분산
- 로버스트
- 도박문제+켈리법칙
- EDA (탐색적 데이터분석)
- 통계적 유의성
- 베イズ 철학
- 회귀
- 과잉학습
- 자유 vs 규제
- 다양성과 독립성
- 신경망
- 강화학습
- GPT
- 양자역학

1. 비편향성(unbiasedness): $E(T; \theta) = \eta(\theta)$ 일 때 통계량 T 가 $\eta(\theta)$ 를 비편향적으로 추정한다고 한다. 좌우의 치우침 없이 균형감을 갖고 판단한다는 의미인 '비편향성'은 통계학도로선 제1의 궁극적 목표이다.



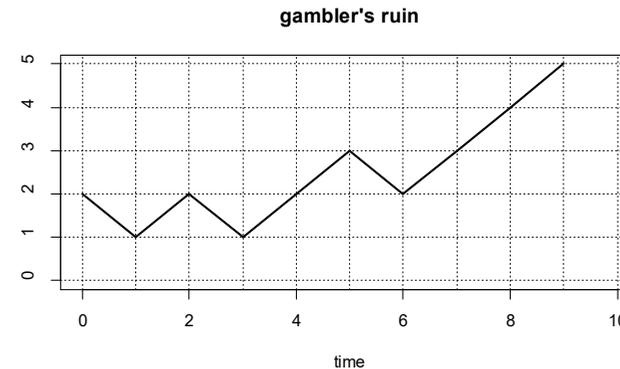
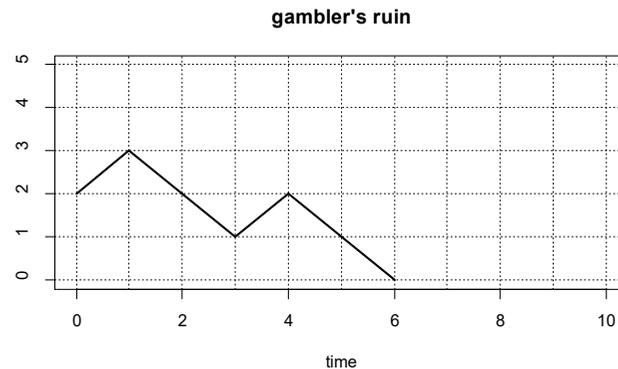
2. 최소분산(minimum variance): 분산은 산포(散布), 즉 분포의 퍼짐(spread)을 나타낸다. 산포는 내재적 잡음, 불안정, 변동스러움이다. 이것을 없앨 수는 없다. 그러나 이것이 최소로 통제된 일상을 지향한다.



3. 로버스트(robust)하다: 흔히 ‘강건하다’로 번역되지만 ‘튼실하다’이 더 낫다.
비바람을 견디어내는 느티나무가 튼실함의 표상이다. 위기에 버티는 바닥 힘이
진짜 실력이다.



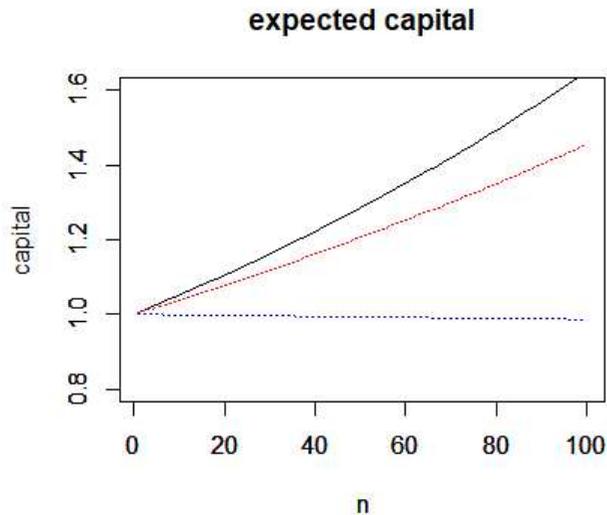
4. 도박문제(gambler's ruin): 초기자산 k 에서 시작한 gambler가 매번 1을 걸고 성공 확률 p 인 플레이를 거듭하는 과정을 생각하자. p 가 0.5인 공정 게임에서도 목표액 n 이 사전에 설정되지 않으면 gambler가 파산할 확률은 1이다. 과욕 금지!



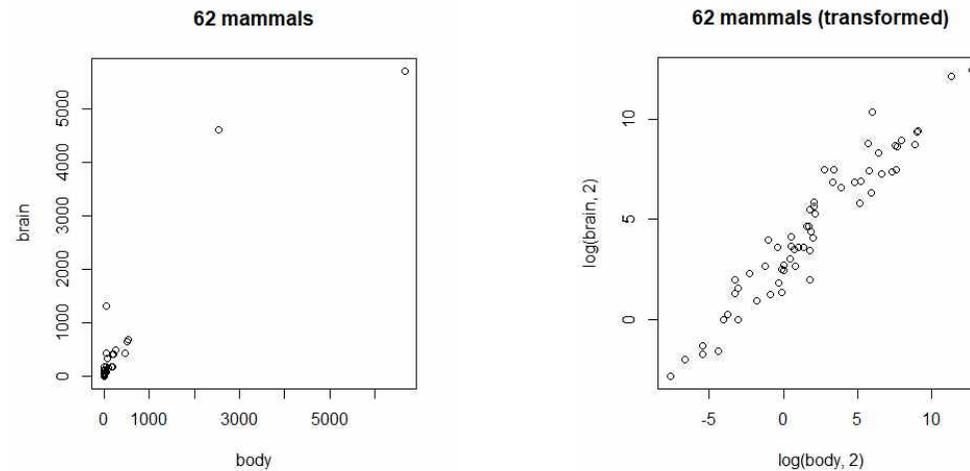
* $p = 0.5$; 파산확률 $q_k = 1 - k/n$

4.a 켈리 법칙(Kelly's rule): 매번, 자산의 일정 비율 $r < 1$ 을 걸고 성공 확률 p 인 플레이를 거듭하는 과정을 생각한다. $p \geq 0.5$ 인 경우, r 의 최적값 r^* 는 $p - q$ 이다. 프로야구에서 상위 팀의 정규시즌 승률은 0.55 근처이다. 그때의 r^* 는 $0.55 - 0.45 = 0.1$ 이다. 인생은 수십 번의 플레이로 구성된다. 과감과 소심 사이에서 최적의 베팅을 해야 한다.

* $p = 0.55$; solid line $r = 0.1$, red line $r = 0.05$, blue line $r = 0.2$



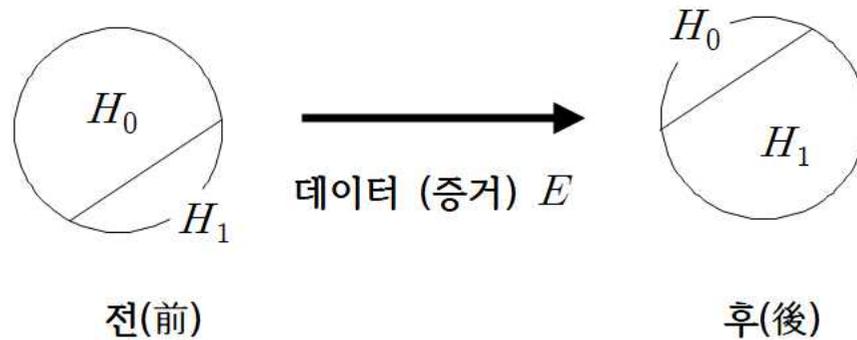
5. EDA(exploratory data analysis)의 가치: 데이터에 대한 탐색(exploration)은 빈둥거리는 것처럼 보일 수 있는 비(非) 목표지향적 작업이다. 탐색적 과정에서 바른 목표가 설정되고 명료해진다. 당장의 비효율이 장기적으로 가치가 있을 수 있다. [8:2 법칙]



mammals의 body와 brain: 변수 변환 전과 후의 산점도

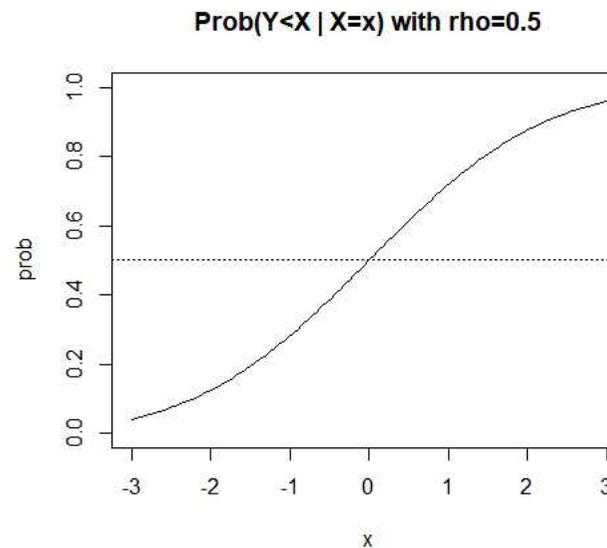
6. 통계적 유의성(statistical significance): 충분한 수의 사례에서 확인된 것만을 따라야 한다. 인간은 one-shot learning을 하는, 빨리 알아채는 존재이다. 그러나 잘못도 빈번하다. 새 주장은 통계적 유의성이 확인되는 때에 수용하는 것이 맞다.

7. 베이즈 철학(Bayesian philosophy): 경험적 사실엔 잡음이 섞여 있다. 원인과 결과는 1대 1 대응이 아니다. 그런 이유로, 관측 사실에 대한 해석에 복수의 틀(frame)을 고려할 필요가 있다. 각 틀에 대한 믿음에는 개인 간 차이가 있다. 그러나 베이즈적 업데이트 후에는 차이가 좁혀진다. [선순환 시스템]

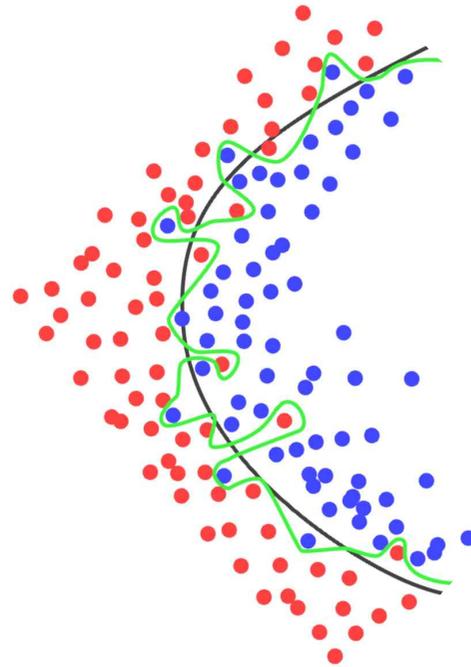


8. 회귀(regression): 모평균보다 우수한 부모의 자녀들은 부모보다 덜 우수할 확률이 크다. 부모들은 자녀가 자신의 기대에 차지 않는다고 압박해서는 안 된다. 자신들이 설정한 잣대를 자녀에 가져다대지 않아야 한다. 성장기 자녀는 각자 나름의 우수성이 있다. 그것이 찾아내고 키워지는 환경을 만들어주는 것이 부모가 할 일이다.

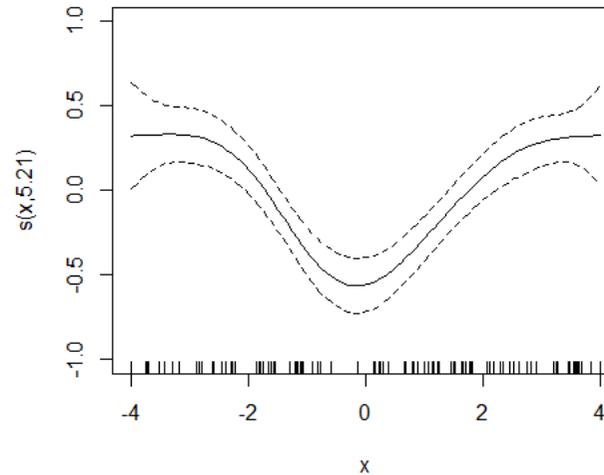
* Regression to the mean:



9. 과잉학습(over-training): 데이터의 세세한 부분까지 맞춘 모형은 정교해 보이지만 재현성이 떨어진다. 일반화에 취약하기 때문이다. 과잉학습이 단기적 목표 달성에는 효과적이지만 장기적 성장을 방해한다.



10. 자유(freedom) vs. 규제(regularization): 모형은 유연하여야 하지만 지나친 자유는 해가 된다. 적당히 규제된, 그러나 유연한 모형이 일반화에서 성능이 좋다. 자유와 규제는 튜닝(tuning)의 건(件)이다.

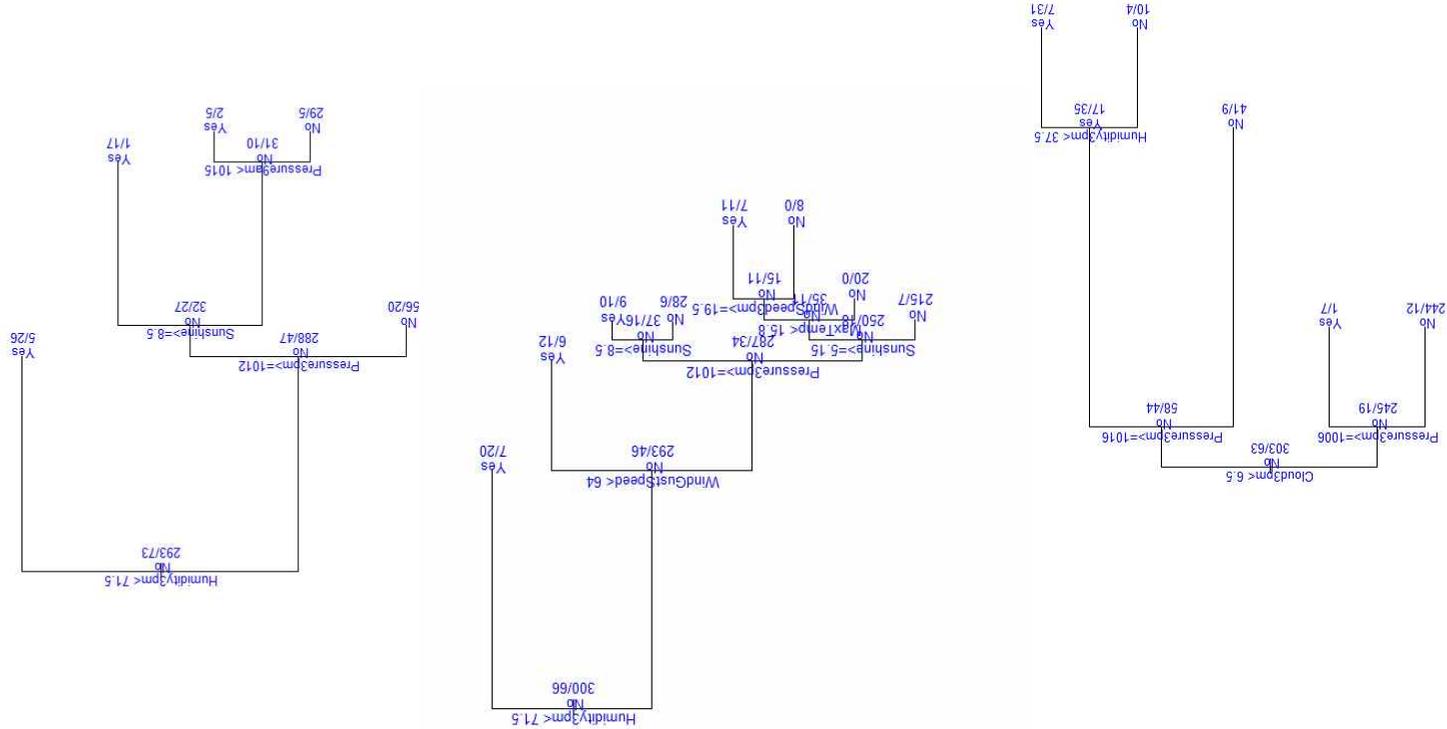


* n 개의 이변량 관측 $(x_1, y_1), \dots, (x_n, y_n)$ 에 대한 비모수적 회귀

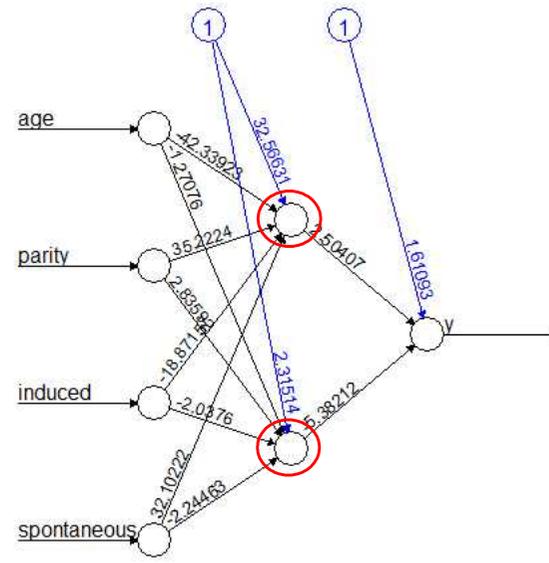
$$\text{Minimize } \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt \quad \text{w.r.t. } g.$$

여기서 $\lambda > 0$ 는 tuning parameter.

11. 다양성과 독립성: 복잡한 상황에서는 우수한 모형 하나보다 덜 우수한 여러 모형들의 총합이 더 우수하다 (모형 앙상블, model ensemble). 전제 조건은 다양성과 독립성이다.



12. 신경망: 심층 신경망의 성능은 여러 겹의 잠재 층 구조에서 나온다. 드러난 것의 내부에 그것을 결정하는 요인들이 숨어있다. 그것을 통찰해야 한다.

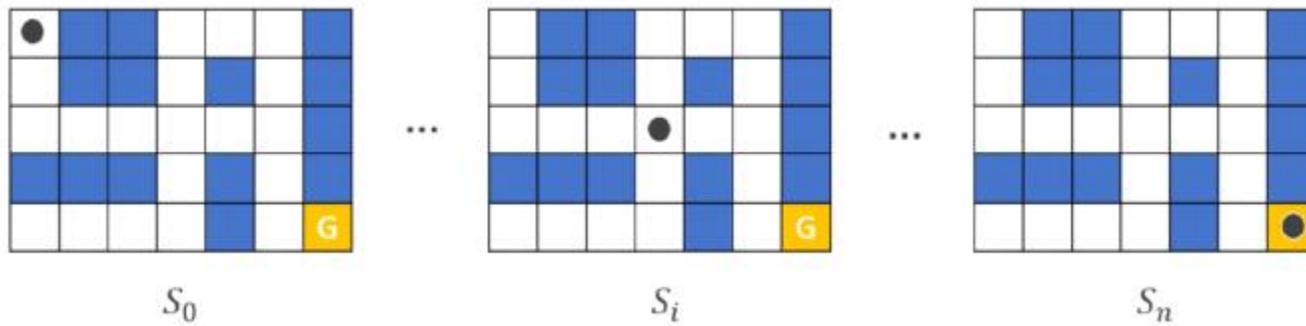


Error: 115.647259 Steps: 1527

13. 강화학습: Balance between Exploitation and Exploration

우연을 회피하면 기회도 없다. 실패로부터 배운다. 물론 실패는 쓰라리다.
균형이 중요하다.

Maze Problem



14. GPT: Generative Pre-trained Transformer

Pre-training 대 Fine-tuning의 비율은 8:2가 맞다. 목적 지향적이지 않은 '읽고 쓰고 말하기'가 기본이다.



사진 출처 <https://blog.naver.com/popopo2013/221525847133>

15. 양자역학이 우리에게 주는 교훈

측정이 상태(행태)를 바꾼다. 정량적 평가의 범람, 데이터 되어짐으로부터 자유로워야 한다.

참고: https://youtu.be/dyvOICRheVc?si=g09j9-DLN5298_QE (양자역학 세바시)



16. 경험적 모형(empirical model)

Cf. All models are wrong, but some are useful. (G.E.P. Box)

모형은 실제(reality)에 대한 근사일 뿐이다. 우리들은 불완전성을 당연히 받아들인다.

삶이 그대를 속일지라도 슬퍼하거나 노여워하지 말라!

(Aleksandr Sergeevich Pushkin)

