꽁꽁 얼어붙은..

분산 추정 및 민감도 개선 : 함정 및 해결책

분산 추정 및 민감도 개선 : 함정 및 해결책

1. 분산을 정확하게 추정하는 것 (일반적인 함정)
2. 통계적 가설 검정의 민감도를 얻기 위해 분산을 줄이는 방법

분산을 잘 추정하자! : 일반적인 함정

- **Q. 분산을 잘못 추정하면, p 값과 신뢰구간이 잘못돼 가설 검정의 결론에 오류가 발생한다.**
  - 분산을 실제보다 크게 추정 -> 거짓 음성 (제 2종 오류)
  - 분산을 실제보다 과소 추정 -> 거짓 양성 (제 1종 오류)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

T 통계량 감소 -> p 값 증가 -> 임계값 넘기 힘들어짐 -> 거짓음성

분산을 잘 추정하자! : 일반적인 함정 – 델타 vs 델타 %

- 실험 결과 보고
  - Yc 대조군 : 평균 10
  - Yt 실험군 : 평균 10.01
  - 실험군이 0.01 늘었다 (0.01 / 10) * 100% = 1% 늘었습니다? (X)
  - (0.01 / Yc의 분산) * 100% 으로 구해야 된다 (O)

  왜냐? 모든 대조군 값이 10인건 아니니까!
  10으로 나누는것은 모든 대조군 값을 10으로 가정한 것임
  분모도 분산을 넣어주세요

분산을 잘 추정하자! : 일반적인 함정 – 분석 단위가 실험 단위와 다른 경우

- 클릭률 = $\frac{\text{총 클릭 수}}{\text{총 페이지뷰}} * 100$
- 클릭 당 수익 = $\frac{\text{총 수익}}{\text{총 클릭 수}} * 100$

분석 대상 : 사용자가 아니라 페이지 뷰, 클릭수가 됨
(즉) 같은 사람이 여러 번의 페이지 뷰를 가질 수 있음

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

IID (독립항등분포) 성립
-> 표본들이 모두 독립적이다
상관관계가 없다.

분산을 잘 추정하자! : 일반적인 함정 – 분석 단위가 실험 단위와 다른 경우

비율 지표 -> 사용자 수준 지표의 평균

$$M = \frac{\bar{X}}{\bar{Y}} \tag{18.4}$$

$\bar{X}$와 $\bar{Y}$는 극한에서 이변량 결합 정규분포로 수렴하므로, 두 평균의 비율인 M도 정규 분포이다. 따라서 델타 방법으로 분산을 (Deng et al. 2017)과 같이 추정할 수 있다(식 18.5 참조).

$$var(M) = \frac{1}{\bar{Y}^2} var(\bar{X}) + \frac{\bar{X}^2}{\bar{Y}^4} var(\bar{Y}) - 2\frac{\bar{X}}{\bar{Y}^3} cov(\bar{X}, \bar{Y}) \tag{18.5}$$

Δ%의 경우, Y^t와 Y^c는 독립적이므로 식 18.6을 참조하라.

$$var(\Delta\%) = \frac{1}{\bar{Y}^{c2}} var(\bar{Y}^t) + \frac{\bar{Y}^{t2}}{\bar{Y}^{c4}} var(\bar{Y}^c) \tag{18.6}$$

$$\mathrm{Var}\left(\frac{X}{Y}\right) = \frac{\mathrm{Var}(X)}{E(Y)^2} + \frac{E(X)^2 \cdot \mathrm{Var}(Y)}{E(Y)^4} - \frac{2 \cdot E(X) \cdot \mathrm{Cov}(X,Y)}{E(Y)^3}$$

분산을 잘 추정하자! : 일반적인 함정 – 분석 단위가 실험 단위와 다른 경우

- 부트스트랩 : 연구자가 추출한 표본에서 표본을 반복 추출하면 중심 극한 정리의 논리에 따라 모집단의 분포가 정규분포에 근접

분산을 잘 추정하자! : 일반적인 함정 – 이상치

표본 평균 차이의
통계적인 지표

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

두 표본 그룹 평균의 차이

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{Var\left[\bar{X}_1 - \bar{X}_2\right]}$$

두 그룹 간 평균 차이에 대한
불확실도

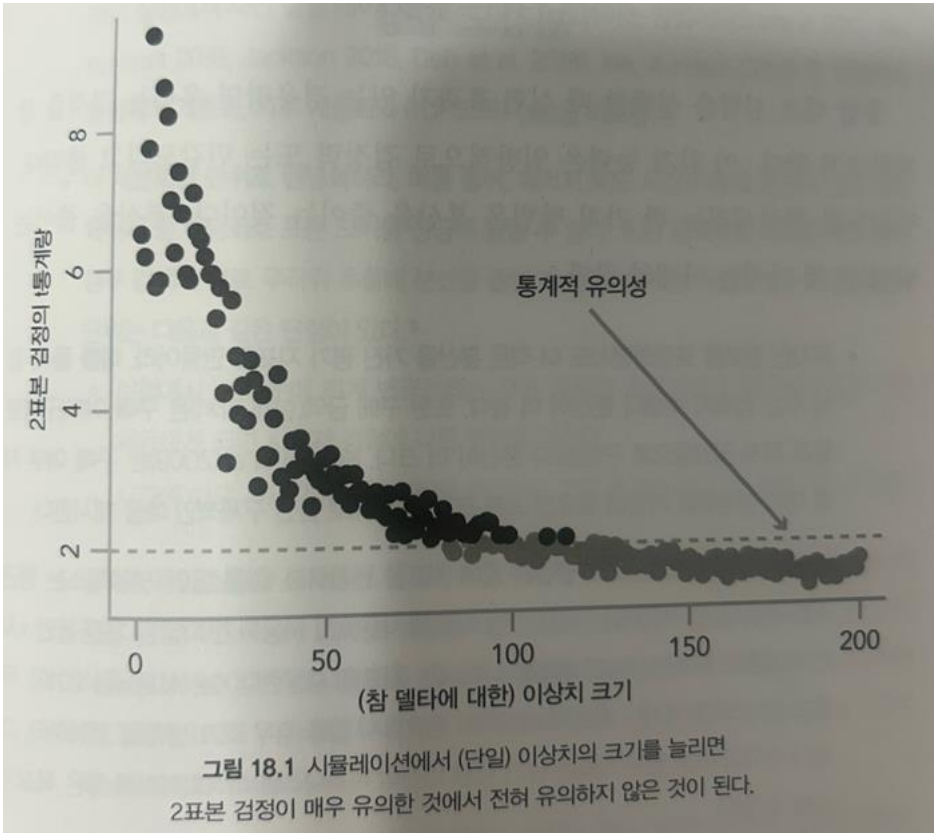


그림 18.1 시뮬레이션에서 (단일) 이상치의 크기를 늘리면 2표본 검정이 매우 유의한 것에서 전혀 유의하지 않은 것이 된다.

- 이상치 크기 -> 평균을 증가시키지만 , 분산이 더 많이 커짐 -> t통계량이 감소 -> 유의하지 않음

분산을 줄이자! : 민감도 향상

- 민감도를 향상시키는 한가지 방법 = 분산을 줄이는 것

1) 유사한 정보를 포함하지만 , 적은 분산을 가진 지표를 만들자

검색 수 -> 검색자 수
구매금액(실제 가치) -> 구매 여부 (구매o,구매x)

2) 이진화 , 로그 변환을 통해 지표를 변환

스트리밍 시간 -> 스트리밍 여부 (이진화)

3) 트리거 분석 : 명확한 기준(트리거 이벤트)을 설정
-> 관련 없는 데이터 또는 정보의 영향을 최소화 시킬 수 있음

분산을 줄이자! : 민감도 향상

4) 계층화 , 통제변수 , CUPED를 사용

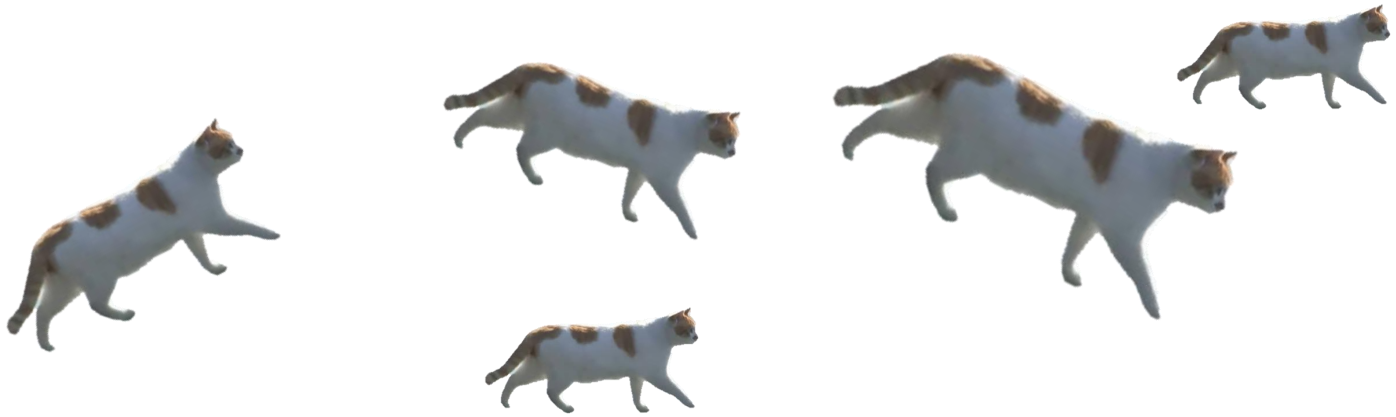
5) 더 세분화된 단위로 랜덤화

6) 쌍으로 묶인 실험을 설계하기

7)대조 집단을 통합하라

느낀점과 같이 이야기 해보고 싶은 사항

- 읽어도 어려워요



- 막상 개념을 알아도 실무에서 써보지 않은 사항들은 와닿지 않아

(실무에서 분산을 줄이고, 분산에 대한 추정의 중요성을 느낀 사례가 있으신가요?)

꽁꽁 얼어붙은..

감자합니다