A/A 테스트

가짜연구소 인과추론팀
온라인 통제 실험 연구자로 거듭나기

2024-05-07

박혜지

Causal-Lab

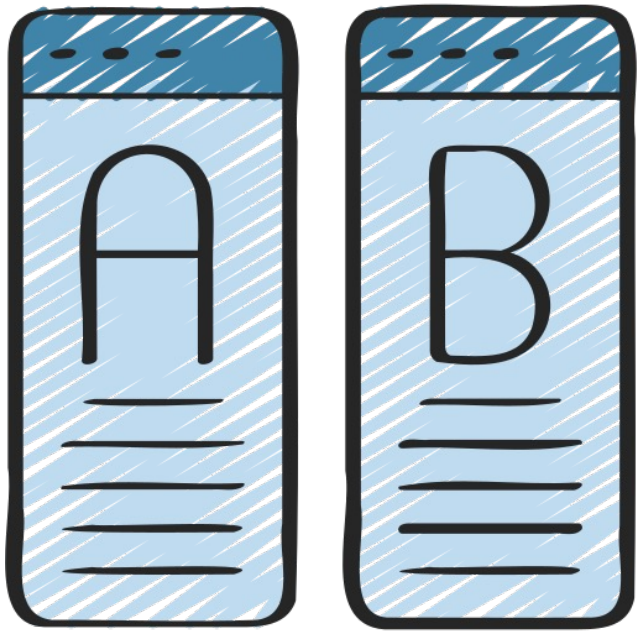
1. A/A 테스트란?
2. A/A 테스트가 필요한 이유
3. A/A 테스트 실행 방법
4. A/A 테스트가 실패한 경우
5. Discussion

A/A 테스트란?

Causal-Lab

과연 우리 조직이 쓰는 A/B 테스트 툴, 믿을 수 있을까?

A/A 테스트란?



A 그룹과 B 그룹에 동일한 실험 처치

🧐 시스템이 올바르게 작동한다면?

✓ 반복시행의 **95%는 통계적으로 유의미한 차이가 없음**
(5%의 경우 0.05 미만의 p값)

✓ t-검정을 수행할 때 반복 시행에서 얻은 p값의 분포는
균등 분포에 가까워야 함

A/A 테스트가 필요한 이유

Causal-Lab

1종 오류의 통제 여부

- 표준 분산 계산이 일부 지표에 대해 올바르지 않은 경우나 정규성 가정을 충족하는지 검증

지표의 변동성 평가

- A/A 테스트를 통해 시간에 따른 지표 분산의 변화를 확인 가능
- 더 많은 사용자가 실험에 할당됨에 따라 평균 분산의 감소 여부 예측

실험 그룹 이월효과 확인

- 이전 실험의 모집단을 재활용하는 경우 실험군과 대조군의 편향 여부 확인
- 연속 A/A 테스트를 사용해 이월 효과 식별

A/A 테스트가 필요한 이유

Causal-Lab

데이터 정합성 확인

- 주요 지표에 대해 내부 데이터 통계 수치와 실험 플랫폼 수치가 일치하는지 확인 ex/ 사용자 수, 수익, 클릭률(CTR)
- 플랫폼의 실험 그룹 할당 인원과 내부 측정 데이터 일치 여부 및 누락 인원 확인

통계적 검정력 계산을 위한 분산 추정

- A/A 테스트 지표의 분산을 통해 MDE(Minimum Detectable Effect) 검증에 필요한 실험 기간 예측

A/A 테스트가 필요한 이유

Causal-Lab

예시1. 분석 단위와 랜덤화 단위가 다르다

- 사용자별로 랜덤화했으나, 페이지별로 분석해야하는 경우가 있음

Check!

User (랜덤화 단위) != PV (분석 단위), Click (분석 단위) : 독립성 가정 위반! PV, Click은 확률 변수로서 작용, 상관관계 지표

⇢ *Delta Method 혹은 Bootstrap 방식 필요*

- CTR 계산 방법

① Total Click / Total PV

$$CTR_1 = \sum_{i=1}^{n}\sum_{j=1}^{K_i} X_{i,j} \Big/ N$$

② 사용자별 CTR의 평균

$$CTR_2 = \sum_{i=1}^{n} \frac{\sum_{j=1}^{K_i} X_{i,j}}{K_i} \Big/ n$$

A/A 테스트가 필요한 이유

Causal-Lab

예시1. 분석 단위와 랜덤화 단위가 다르다

User	PV	Click
A	1	0
B	2	2

① Total Click / Total PV

$$CTR_1 = \sum_{i=1}^{n}\sum_{j=1}^{K_i} X_{i,j} \Big/ N \quad \rightarrow \quad CTR_1 = \frac{0+2}{1+2} = \frac{2}{3}$$

분산 추정시 독립성 가정을 위반하기 때문에
Delta Method / Bootstrap 방식으로 계산

② 사용자별 CTR의 평균

$$CTR_2 = \sum_{i=1}^{n} \frac{\sum_{j=1}^{K_i} X_{i,j}}{K_i} \Big/ n \quad \rightarrow \quad CTR_2 = \frac{0}{1} + \frac{2}{2} \Big/ 2 = \frac{1}{2}$$

이상치에 대해 영향을 덜 받아 권장

A/A 테스트가 필요한 이유

Causal-Lab

예시2. Optimizely의 거짓 양성 오류

- Optimizely의 초기 버전은 **중간 과정 테스트(결과 Peeking)와 조기 종료 장려**를 통해 거짓 양성 결과 도출 오류
  - *“테스트가 통계적으로 유의한 수준에 도달하면 답을 얻을 수 있다”*
  - *충분한 실험 기간을 확보하지 않고, 실험 도중 유의미한 결과가 나오면 실험 중단*

- *A/A Test를 통해 오류를 찾아내 "옵티마이즐리의 새로운 통계 엔진" 출시*
  ** How Optimizely (Almost) Got me Fired*

예시3. 브라우저 리디렉션

- 새로운 웹사이트 구축 후 이전 버전과 새버전의 A/B Test시 변형군 B 사용자는 새 웹사이트로 리디렉션 **> B 실패 확률 높음!**

문제점

1. **리디렉션시 1~2초의 대기시간이 발생하여 성능면에서 차이 발생**
2. **리디렉션의 오류 발생 가능성 (봇의 신규 페이지 대규모 크롤링, 리디렉션 미수행, 오류 발생)**
3. **북마크와 공유 링크의 데이터 오염**

리디렉션이 없도록 구성

or

대조군, 실험군 모두 리디렉션

A/A 테스트가 필요한 이유

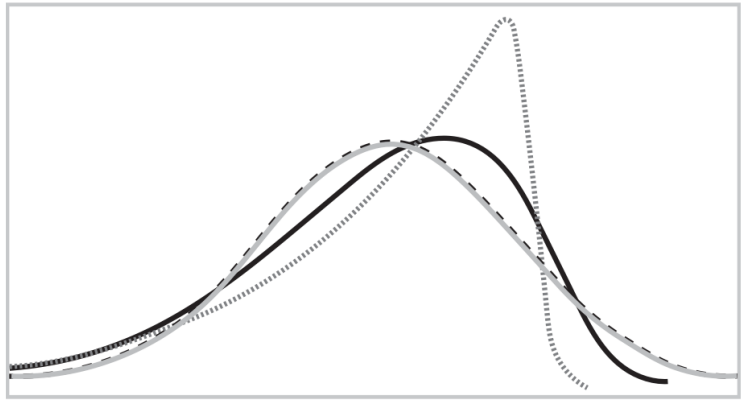
Causal-Lab

* LRU (Least Recently Used Algorithm)
: 페이지를 교체할 때 가장 오랫동안 사용되지 않은 페이지를 교체대상으로 삼는 방식
* Cache
: 연산에 필요한 데이터, 값을 미리 가져다 놓는 임시 메모리

예시4. 균등하지 않은 분할

- 균등하지 않은 분할을 할 경우(ex. 10% / 90%) 리소스의 분배가 더 큰 변형군에 유리하게 적용
  - 큰 실험군은 캐시 히트율이 높아 응답 속도 등이 좋아지고 결과가 좋아질 가능성도 높아짐
  - 실험과 일치하는 비율로 A/A 테스트 통과 권장 (90%/10%로 실험할 경우 A/A 테스트도 동일한 비율로 진행)

- 균등하지 않은 분할을 할 경우(ex. 10% / 90%) 정규 분포로의 수렴 속도가 다를 수 있음
  - 한 쪽으로 과도하게 치우친 분포 > 중심극한 정리는 백분율이 다르면 평균은 정규 분포로 수렴하지만 수렴 속도 달라짐



17장 정규성 가정/

표본 평균 $\bar{Y}$가 정규 분포를 갖기 위해 필요한

최소 표본 수에 대한 하나의 경험 법칙은 각 변수에 대해 355s^2 (s : 왜도)

A/A 테스트가 필요한 이유

Causal-Lab

예시5. 하드웨어 차이

- 페이스북은 새로운 서비스 구축 후 신형/구형 AB 테스트 진행
- 하드웨어의 차이로 인해 A/A 테스트 실패

A/A 테스트 실행 방법

Causal-Lab

실행 방법

1. A/B 테스트 시스템 사용 전 A/A 테스트 실행
2. 1,000개의 A/A 테스트를 시뮬레이션 & p값의 분포 시각화

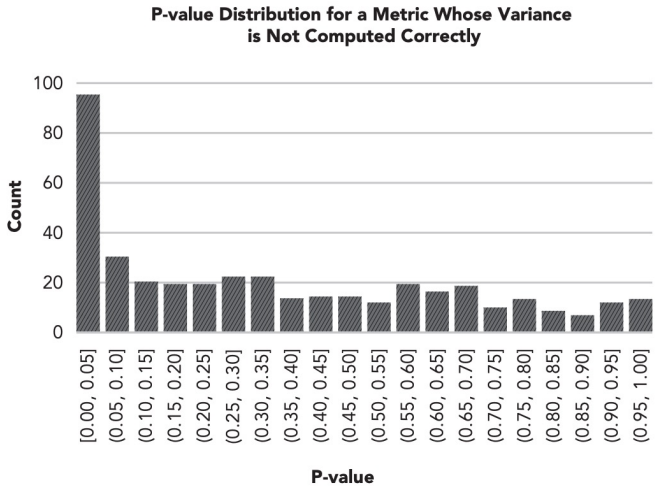


Figure 19.1 Non-uniform p-value distribution from A/A tests for a metric whose variance is not computed correctly because the analysis unit is not equal to the randomization unit

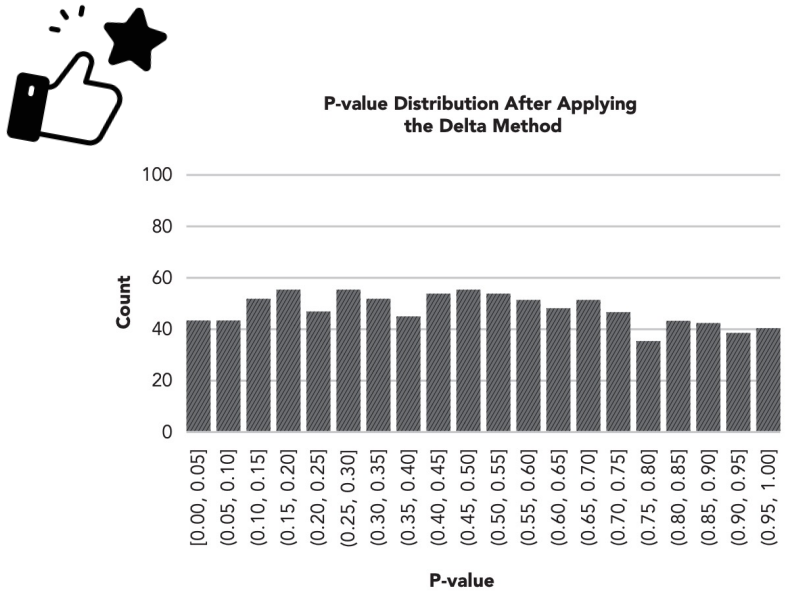


Figure 19.2 Distribution is close to uniform after applying the delta method to compute variance

올바르게 계산된 지표로 p값의 분포가 균등하다면 A/A 테스트 통과

A/A 테스트 실행 방법

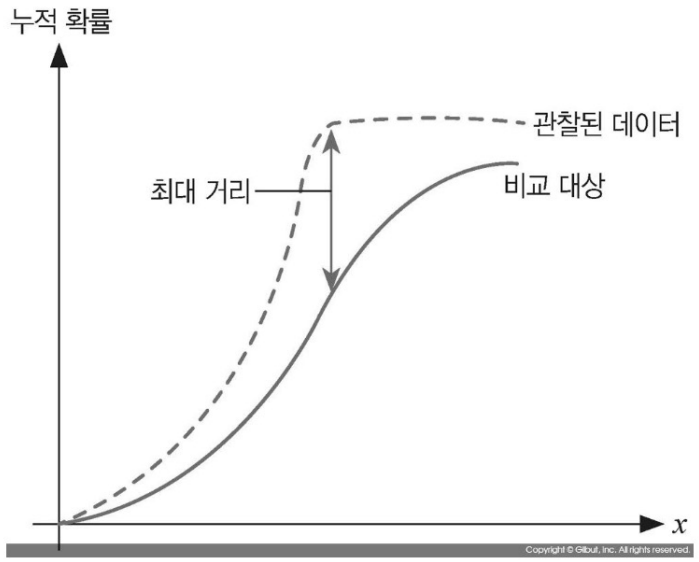
Causal-Lab

균등성 평가

1. Komogorov-Smirnov 검정 (K-S 검정)

데이터의 누적 분포함수와 비교하고자 하는 누적분포 함수간의 최대 거리를 통계량으로 사용

- H_0 (귀무가설) : 데이터는 특정 분포를 따른다.
- H_1 (대립가설) : 데이터는 특정 분포를 따르지 않는다.



A/A 테스트 실행 방법

Causal-Lab

균등성 평가

2. Anderson-Darling 검정 (AD 검정)

데이터가 특정 분포를 얼마나 잘 따르는지 검정, 일반적으로 분포가 데이터에 더 적합할 수록 AD 통계량이 작음

특정 분포의 꼬리에 K-S검정보다 가중치를 더 두어 수행 > 정규 분포 검증에 강력

- H_0 (귀무가설) : 데이터는 특정 분포를 따른다.
- H_1 (대립가설) : 데이터는 특정 분포를 따르지 않는다.

Anderson-Darling 검정 통계량(A^2)은 아래와 같이 계산됩니다.

$$A^2 = -n - S$$

여기서 n은 표본 데이터의 크기, S는 아래와 같이 계산됩니다.

$$S = \sum_{i=1}^{n} \frac{(2i-1)}{n} \left[\ln F\left(Y_i\right) + \ln\left(1 - F\left(Y_{n+1-i}\right)\right) \right]$$

A/A 테스트가 실패한 경우

Causal-Lab

1. 분포가 치우쳐있고 균등에 가깝지 않음

일반적으로 지표의 분산 추정 문제

- 랜덤화 단위와 분석 단위가 일치하는가?
  - 일치하지 않는다면 독립성 가정 위반!
  - Delta Method or Bootstrap

- 지표가 매우 치우친 분포를 갖고 있는가?
  - 샘플 수가 적은 경우 정규분포에 근사하지 않을 수 있음
  - 제한된 지표 또는 최소 표본 크기 설정 필요

17장 정규성 가정/

- **표본 평균 $\bar{Y}$가 정규 분포를 갖기 위해 필요한 최소 표본 수에 대한 하나의 경험 법칙은 각 변수에 대해 355s^2 (s : 왜도)**
- **수익과 같은 왜도가 높은 지표 > 지표 변경 or 값 제한을 통해 왜도 축소**

A/A 테스트가 실패한 경우

Causal-Lab

2. p값이 0.32 주변에 몰려있음

이상치 o의 t통계량 영향

$$T = \frac{\Delta}{\sqrt{var}}(\Delta)$$

- Delta -> o/n 혹은 음수에 수렴
- 특이값에 의해 T값이 1이나 -1에 가까워 약 0.32의 p값에 맵핑

- 이상치의 원인 조사 및 데이터 제한 필요

A/A 테스트가 실패한 경우

Causal-Lab

3. 분포가 큰 간격으로 값이 몇 개의 점에 몰려있음

데이터가 대부분 단일값 (예:0)을 가지고 몇 개의 0이 아닌 값을 가질 경우

- 0이 아닌 일부의 이산 값만 통계량으로 활용할 수 있기 때문에 p 계산에 제약
- 새로운 실험이 희소한 사건을 더 빈번하게 발생시킬 경우 실험 효과가 더 커지고 통계적으로 유의하게됨
    - 앞의 시나리오만큼 큰 문제를 야기하지 않음

Discussion

Causal-Lab

논의

1. 1000번의 A/A 테스트… 현실적으로 가능한가?

2. CTR 계산 방식 2번 (사용자의 CTR의 평균) 방식으로 사용하는 것은 정말 적절한 방식인가?

3. 실제 현업에서 A/A 테스트를 진행했던 사례

4. *모든 것이 제어되고 있는 상황이라면, 당신은 A/A 테스트를 실행하고 있다.*
   -> 실험군을 어느 정도까지 통제하는 것이 적절한가? 공휴일, 프로모션 등 모든 값을 한 번에 더해 Delta를 비교해도 괜찮을까