신뢰성 관련 가드레일 지표

가짜연구소 인과추론팀
온라인 통제 실험 연구자로 거듭나기

2024-05-21

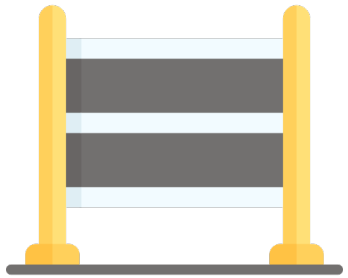
문정하

가드레일 지표

Causal-Lab

: 가정이 위반될 경우에 실험자에게 경고하기 위해 설계된 중요한 지표

- 비즈니스를 보호하는 지표 (6,7장)
  - 예) 사용 편의성이 목표일 때 보안이 가드레일 지표
- 실험 결과에 대한 신뢰성 관련 지표
  - SRM(Sample Ratio Mismatch)



SRM 발생 예시 – 시나리오1

Causal–Lab

- 설계 1 : 1
- 대조군 : 사용자 821,588명 vs. 실험군 : 사용자 815,482명

This spreadsheet is available at https://bit.ly/srmCheck
Probability that two sample counts are equal (no SRM, or Sample Ratio Mismatch)
By Ronny Kohavi (note to self: @cs version)
Fill in the yellow part

		Expected
Expected ratio (control/(control+treatment))	0.5	
Count in Control	821,588	818,535
Count in Treatment	815,482	818,535
C/pop	0.5019	
Z	4.7722	
P-value (t-dist, no plug-in estimate)	**1.82E-06**	Delta to best estimate
Chi-squared	0.000002	-1.57E-10
Z (plug-in estimate)	4.7723	
Plug-in estimates below		
P-value (normal)	1.82149E-06	-4.57E-10
P-value (t-dist)	1.82165E-06	########

If the p-value is very low (we use <0.001), then you likely have a sample-ratio mismatch and the experiment results are not trustworthy

SRM 발생 예시 – 시나리오2

	Treatment	Control	Delta	Delta %	P-Value	P-Move
▼ Metadata						
ScorecardId	96699772					
Sample Ratio [by user]	0.9938 = 959,716 (T) / 965,679 (C)					P=2e-5
Sample Ratio [by page]	0.9914 = 6,906,537 (T) / 6,966,740 (C)					-
Trigger Rate [by user]						-
Trigger Rate [by page]						-
▼ Main Metrics						
▼ Success Metrics						
Sessions/UU				+0.54%	0.0094	12.8%
				+0.20%	7e-11	>99.9%
				+0.49%	2e-10	>99.9%
				-0.46%	4e-5	99.5%
				+0.24%	0.0001	99.0%

- P-value : 2e-5 < 0.05 → 결과 신뢰할 수 없음

표 출처 :A/B 테스트 교재 21장 P.309

SRM 발생 예시 – 시나리오2

Causal–Lab

	Treatment	Control	Delta	Delta %	P-Value	P-Move	Treatment	Control	Delta	Delta %	P-Value	P-Move
▼Metadata												
ScorecardId	96699772						96762547					
Sample Ratio [by user]	0.9938 = 959,716 (T) / 965,679 (C)					P=2e-5	0.9993 = 924,240 (T) / 924,842 (C)					P=0.6580
Sample Ratio [by page]	0.9914 = 6,906,537 (T) / 6,966,740 (C)					-	0.9955 = 6,652,169 (T) / 6,682,151 (C)					-
Trigger Rate [by user]						-	0.9604 = 1,849,082 (T+C) / 1,925,395 (T+C)					-
Trigger Rate [by page]						-	0.9612 = 13,334,320 (T+C) / 13,873,277 (T+C)					-
▼Main Metrics												
▼Success Metrics												
Sessions/UU				+0.54%	0.0094	12.8%				+0.19%	0.3754	0.2%
				+0.20%	7e-11	>99.9%				+0.04%	0.1671	10.1%
				+0.49%	2e-10	>99.9%				+0.13%	0.0727	24.6%
				-0.46%	4e-5	99.5%				-0.12%	0.2877	7.4%
				+0.24%	0.0001	99.0%				+0.01%	0.8275	0.7%

- SRM 발생 원인 : bot filtering issue, outdated Chrome browser
- P-value : 0.6580 > 0.05 → 결과 버리지 않아도 됨

표 출처 :A/B 테스트 교재 21장 P.309

SRM 발생 원인

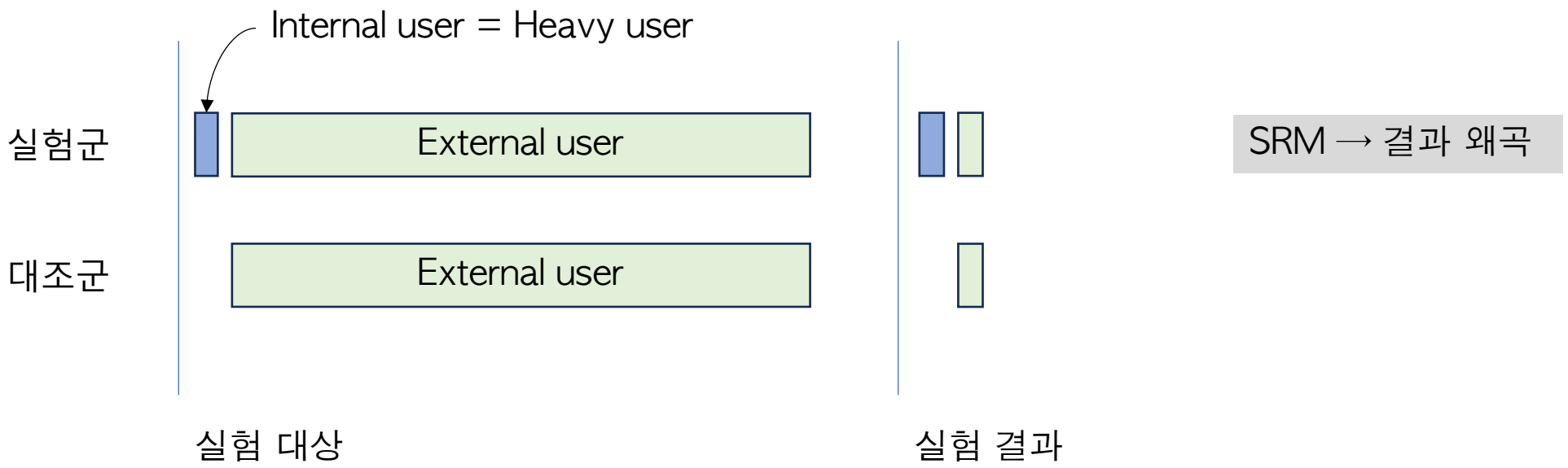
Causal-Lab

- 사용자 랜덤화의 버그
- 데이터 파이프라인 문제
- 잔여 효과
- 잘못된 트리거 조건
- 실험의 영향을 받은 속성을 기반으로 하는 트리거링

SRM 발생 원인

- **사용자 랜덤화의 버그**
  - ➢ 램프업 절차, 다른 실험에서 사용된 유저 제외, 실험 전 공변량 균형잡기 시도, 등으로 인해 현실에서 랜덤화에 어려움이 존재

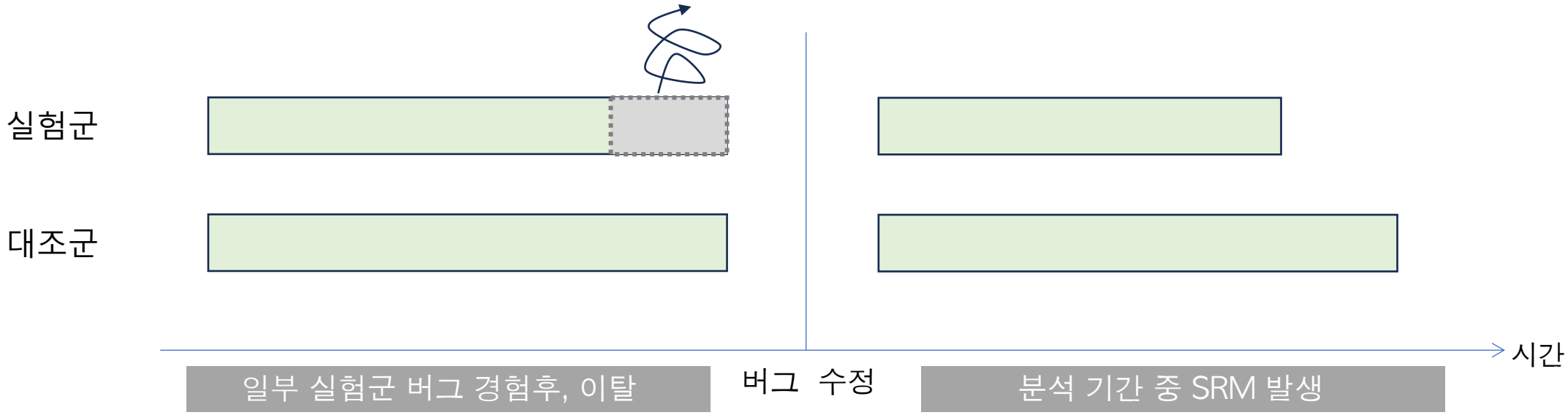
[마이크로소프트 예시]



SRM 발생 원인

- **데이터 파이프라인 문제**
  - 봇 필터링 문제 발생

- **잔여 효과**



SRM 발생 원인

Causal-Lab

- **잘못된 트리거 조건**
  - 페이지 리디렉션 후, 실험 페이지에 도달한 사용자 (☹️)
  - (개인적인 추측) 페이지 리디렉션 대상 모든 사용자 (😃)

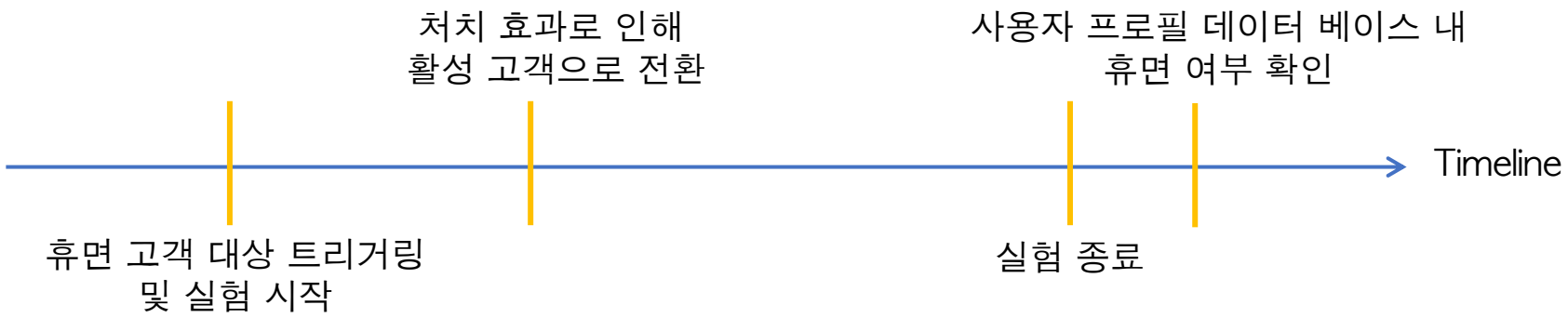
➢ 영향을 받을 수 있는 모든 사용자를 포함해야 하며, 리디렉션으로 인한 일부 손실 가능성 때문

➢ (개인적인 추측) 리디렉션된 대상을 제외하고 계산할 경우, 실험 효과를 과대 추정할 수 있음

SRM 발생 원인

- **잘못된 트리거 조건**
  - 페이지 리디렉션 후, 실험 페이지에 도달한 사용자 (☹)
  - (개인적인 추측) 페이지 리디렉션 대상 모든 사용자 (😃)
  - ➢ 영향을 받을 수 있는 모든 사용자를 포함해야 하며, 리디렉션으로 인한 일부 손실 가능성 때문
  - ➢ (개인적인 추측) 리디렉션된 대상을 제외하고 계산할 경우, 실험 효과를 과대 추정할 수 있음

- **실험의 영향을 받은 속성을 기반으로 하는 트리거링**



SRM 디버깅

Causal-Lab

- **랜덤화 시점 또는 트리거 시점의 이전 단계에 차이가 없는지 검증**
  - 예) 결제 시점이 트리거 조건일 경우, 그 이전 단계에 대한 검증 필요

- **실험군 및 대조군 할당이 올바른지 검증**
  - 다수의 실험의 실험군으로 포함되지는 않는지
  - 예) 실험1 : 검정색 글씨 vs. 파란색 글씨 / 실험2 : 검은색 글씨 유저 중, 배경색 1 vs. 배경색2

- **데이터 처리 파이프라인의 단계를 따라 SRM의 원인이 있는지 확인**
  - 예) 봇 필터링 : 최다 사용자들이 경험적인 사용량 임계값을 넘어 봇으로 잘못 분류되는 경우

SRM 디버깅

Causal-Lab

- **실험 시작 후 초기 기간 제외**
  - 여러 실험에서 대조군 공유로 인한 앱푸시, 등의 지연

- **세그먼트의 샘플 비율 확인**
  - 매일 개별적으로 살펴보기
  - 실험군의 실험 비율 변화, 다른 실험에 의한 실험 트래픽 변화, 유저 특성 변화, 등

- **다른 실험과의 공통점 체크**
  - 실험군과 대조군의 할당 비율이 다른 실험과 유사해야 함 (왜 그래야할까요…? 👀)

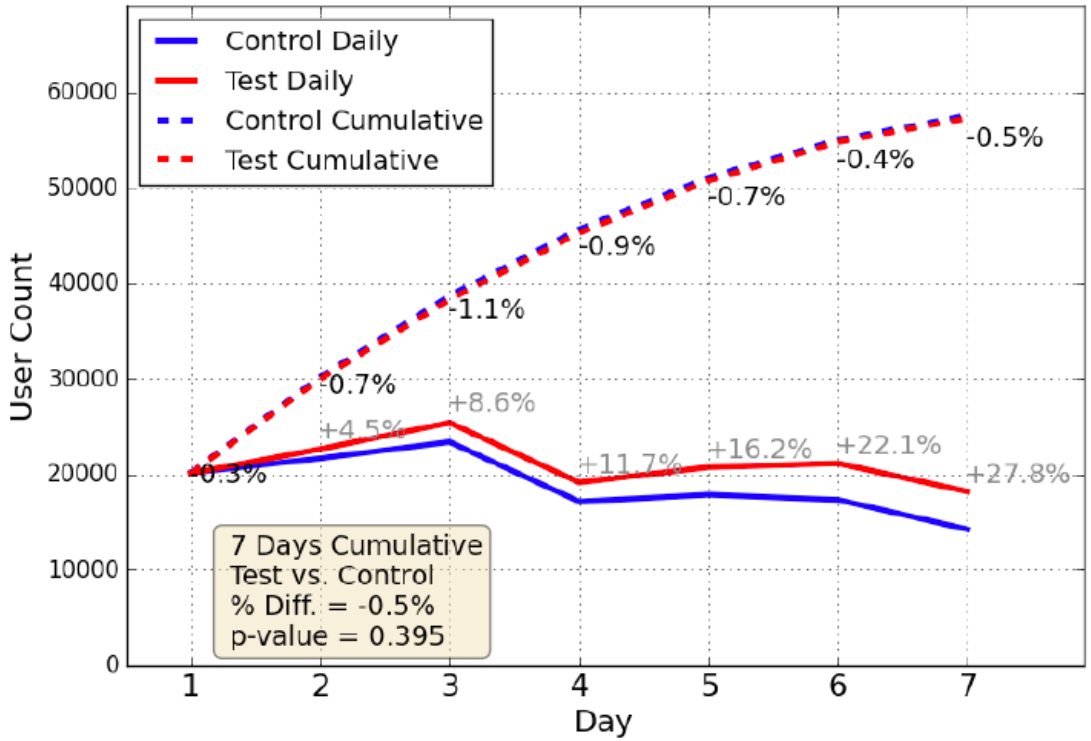
기타 신뢰성 관련 가드레일 지표

Causal-Lab

- **원격측정 정확도**
  - 처치가 클릭 추적의 손실률에 영향을 미치는 경우, 이중 로깅을 사용하는 클릭 혹은 웹사이트에 대한 내부 referrer을 통해 결함을 평가할 수 있는 지표를 통한 정확도 문제 발견 가능
- **캐시 적중률**
  - 공유 리소스에 대한 지표를 통해 예상치 못한 요인을 식별하는 데 도움이 될 수 있음
- **쿠키 쓰기 속도**
  - 실험군 및 대조군에서 영구적(비세션) 쿠키를 쓰는 속도
- **빠른 쿼리**
  - 동일한 사용자로부터 1초 이내에 검색 엔진에 도착하는 둘 이상의 검색 쿼리의 비율

Cumulative user count

- “Although there is a growing difference in daily unique user counts between test and control, the cumul ative user count remains balanced. In theory, increased user retention (by treatment effects) leads to i ncreased daily unique users, but will not affect the cumulative user count. Overall, the 7 day cumulativ e user count is balanced and passes a Chi-square test.”



Zhao, Zhenyu, Miao Chen, Don Matheson, and Maria Stone. 2016. “Online Experimentation Diagnosis and Troubleshooting Beyond AA Validation.” DSAA 2016: IEEE International Conference on Data Science and Advanced Analytics. IEEE. 498–507. doi:https://ieeexplore.ieee.org/document/7796936.

SRM as an indicator of positive treatment behavior

Causal–Lab

4.6 SRMs Can Have a Positive Cause

While most SRMs occur due to issues in one or more of the experiment stages, some SRM can be an indicator of positive treatment behavior. For example, when the load speed of the treatment variant is improved, the likelihood of the load event being logged increases. Similarly, if the treatment variant increases engagement of the users with the product (for example, users are forced to click on two items as opposed to only one in control) – the likelihood of losing both events during telemetry collection is smaller compared to only losing the single event.

*"If the **treatment is getting us to recover additional users** in the logs **it almost always** means that it made some **performance** improvements to the website. So, these are wins!"*

The findings and example data in this section are only a small subset of the knowledge that we constructed about SRMs during this study. In the next section, we provide an overview of the common SRM types we inferred from this data.

Fabijan, Aleksander, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer, and Pavel Dmitriev. 2019. "Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners." KDD '19: The 25th SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, Alaska, USA: ACM.

논의 주제

Causal-Lab

- (P.311) “문제를 디버깅하는데 도움이 되는 경우를 제외하고는 절대 다른 지표를 보지 말아야 한다.” → 무슨 의미일까요?

- 샘플 비율이 불일치한 것을 분석시 알게 되었을 경우, 버그를 경험한 유저를 제외하고 분석하는 것 외, IPW (Inverse Propensity Weighting) 방법과 같이 사후적으로 보정해주는 것은 안될까요?