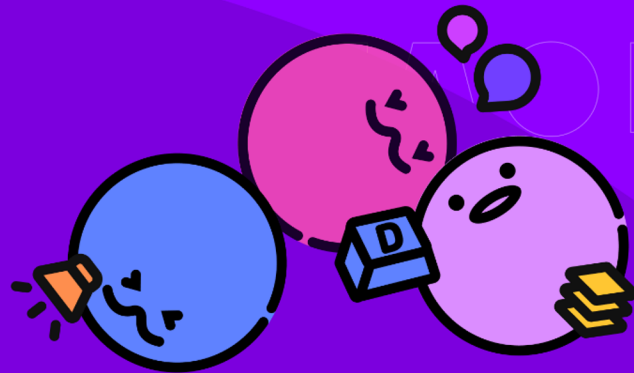


# 당신의 A/B Test를 신뢰할 수 없는 이유

방태모 | G마켓

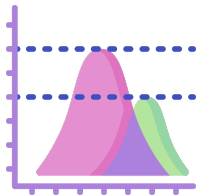


# 본인 소개

데이터야.놀자 2023



통계학 전공



Data Scientist



Data Scientist



Blog



[taemobang@gmail.com](mailto:taemobang@gmail.com)



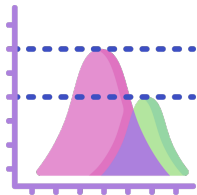
[링크드인](#)

# 본인 소개

데이터야.놀자 2023



통계학 전공



Data Scientist



Data Scientist



Blog



[taemobang@gmail.com](mailto:taemobang@gmail.com)

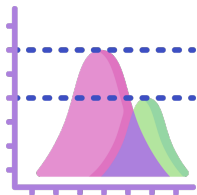


[링크드인](#)

# 본인 소개

데이터야.놀자 2023

통계학 전공



Data Scientist



Data Scientist

Gmarket

Blog



[taemobang@gmail.com](mailto:taemobang@gmail.com)

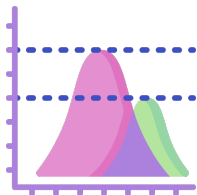


[링크드인](#)

# 본인 소개

데이터야.놀자 2023

통계학 전공



Data Scientist



Data Scientist



Blog



[taemobang@gmail.com](mailto:taemobang@gmail.com)

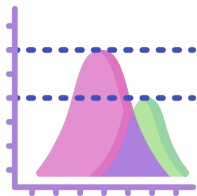


[링크드인](#)

# 본인 소개

데이터야.놀자 2023

통계학 전공



Data Scientist



Data Scientist



Blog



[taemobang@gmail.com](mailto:taemobang@gmail.com)



[링크드인](#)



**먼저,  
A/B Test에 대해 짚고 넘어가봅시다.**

# A/B Test?

데이터야. 놀자 2023

## 온라인 통제 실험 (Online Controlled Experiments, OCE)

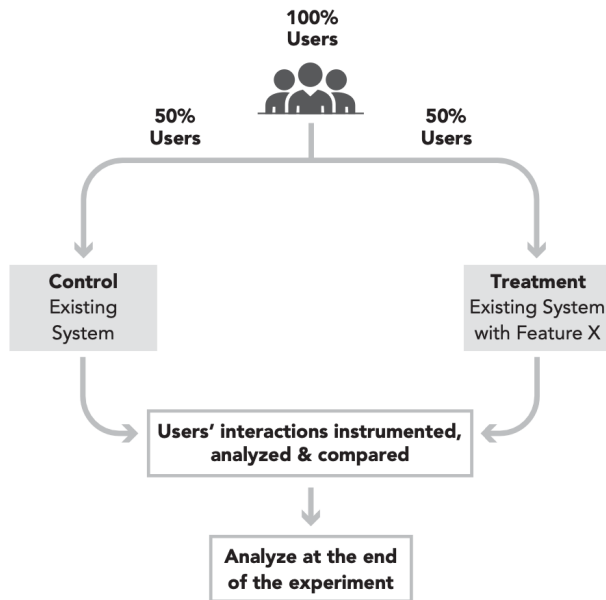


Figure 1.2 A simple controlled experiment: An A/B Test

Source: [Kohavi, Tang, and Xu 2020](#)



# A/B Test?

데이터야. 놀자 2023

## 온라인 통제 실험 (Online Controlled Experiments, OCE)

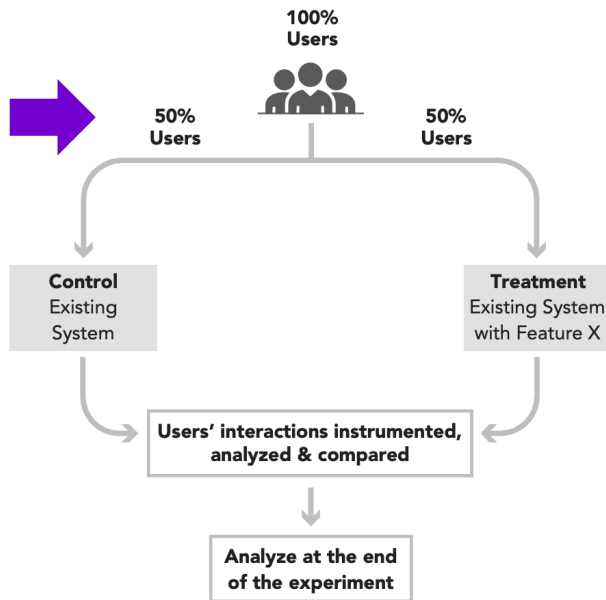


Figure 1.2 A simple controlled experiment: An A/B Test

Source: [Kohavi, Tang, and Xu 2020](#)

# A/B Test?

데이터야. 놀자 2023

## 온라인 통제 실험 (Online Controlled Experiments, OCE)

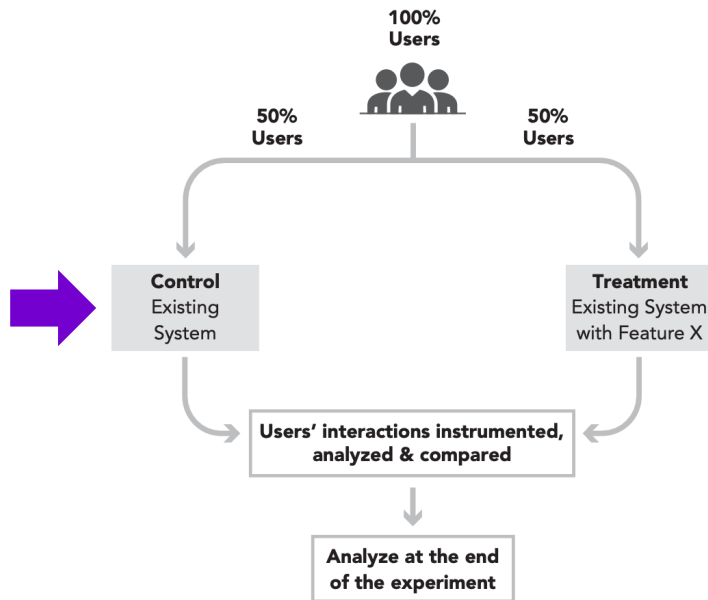


Figure 1.2 A simple controlled experiment: An A/B Test

Source: [Kohavi, Tang, and Xu 2020](#)

# A/B Test?

데이터야. 놀자 2023

## 온라인 통제 실험 (Online Controlled Experiments, OCE)

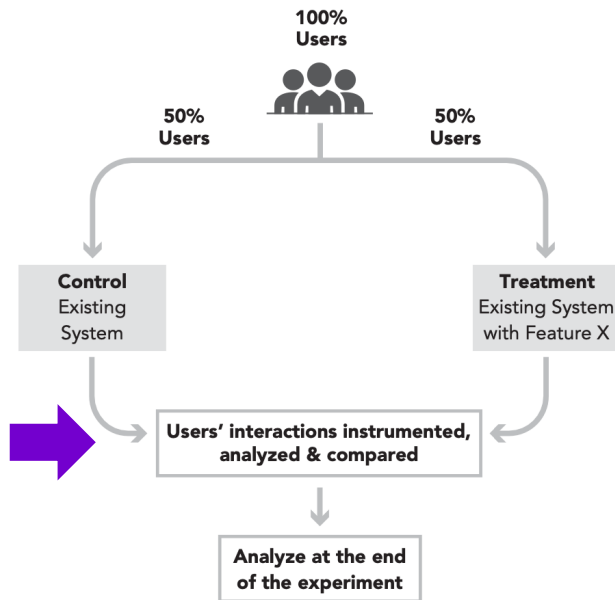


Figure 1.2 A simple controlled experiment: An A/B Test

Source: [Kohavi, Tang, and Xu 2020](#)

# A/B Test?

데이터야. 놀자 2023

## 온라인 통제 실험 (Online Controlled Experiments, OCE)

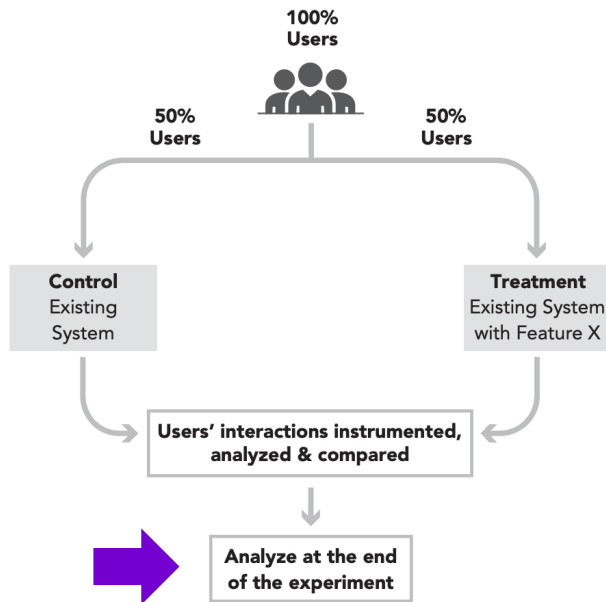


Figure 1.2 A simple controlled experiment: An A/B Test

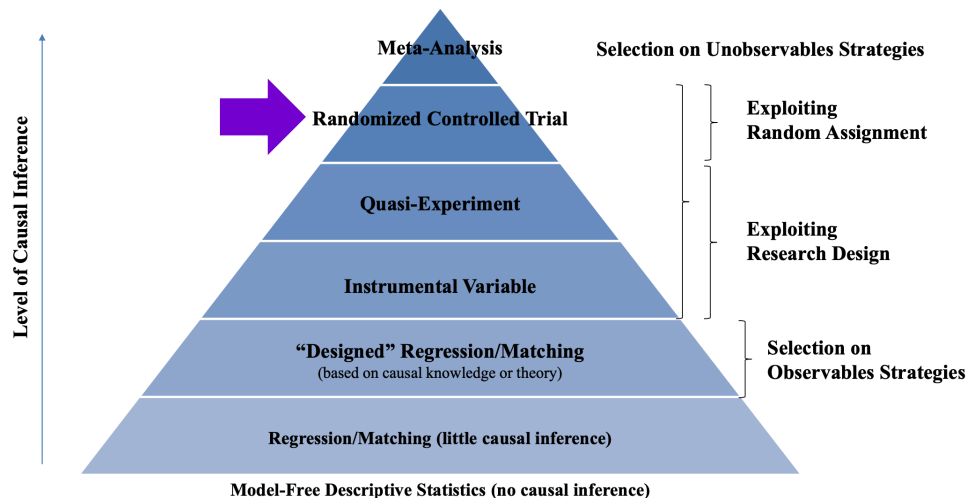
Source: [Kohavi, Tang, and Xu 2020](#)

# 왜 A/B Test인가?

데이터야. 놀자 2023

온라인 통제 실험, 그 이전에는 RCT(Randomized Controlled Trial)

-> 농업, 의학에서 먼저 번영



Source: [Korea Summer Workshop on Causal Inference 2022](#)

# 왜 A/B Test인가?

온라인 서비스에서 우리의 관심은?

# 왜 A/B Test인가?

온라인 서비스에서 우리의 관심은?

- 우리가 준 변화의 순수한 효과(Revenue, UX)
- Online 영역으로 RCT를 확장하자
- 온라인 통제 실험 (OCE, a.k.a A/B Test)

# A/B Test? 너무 쉬워보이는데?

데이터야. 놀자 2023

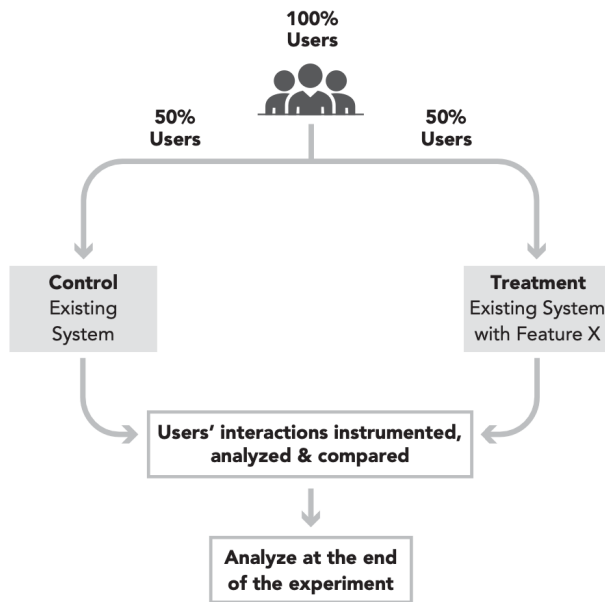


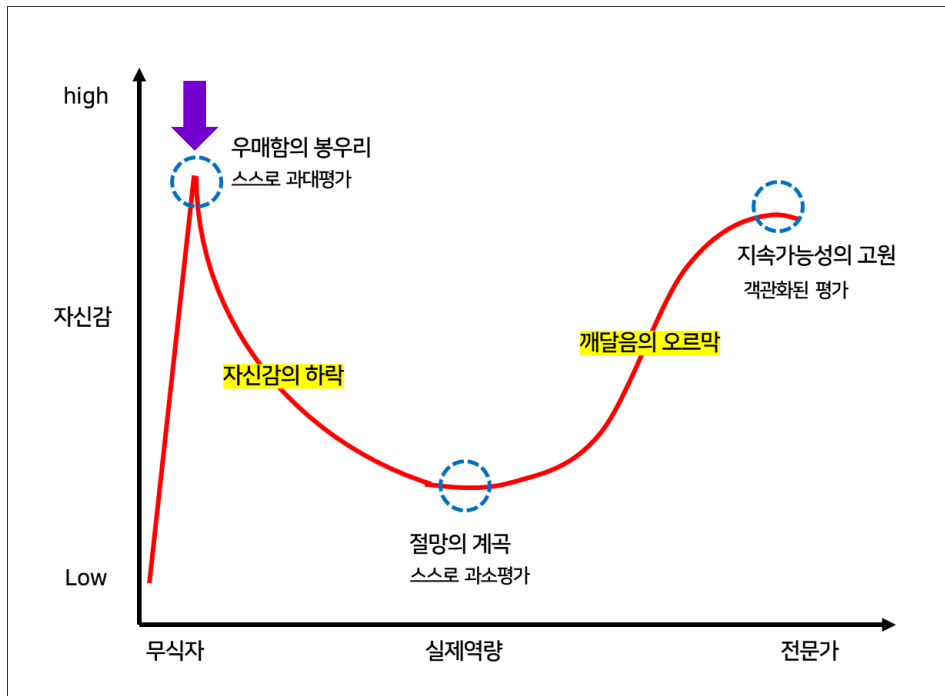
Figure 1.2 A simple controlled experiment: An A/B Test

Source: [Kohavi, Tang, and Xu 2020](#)



# A/B Test? 너무 쉬워보이는데?

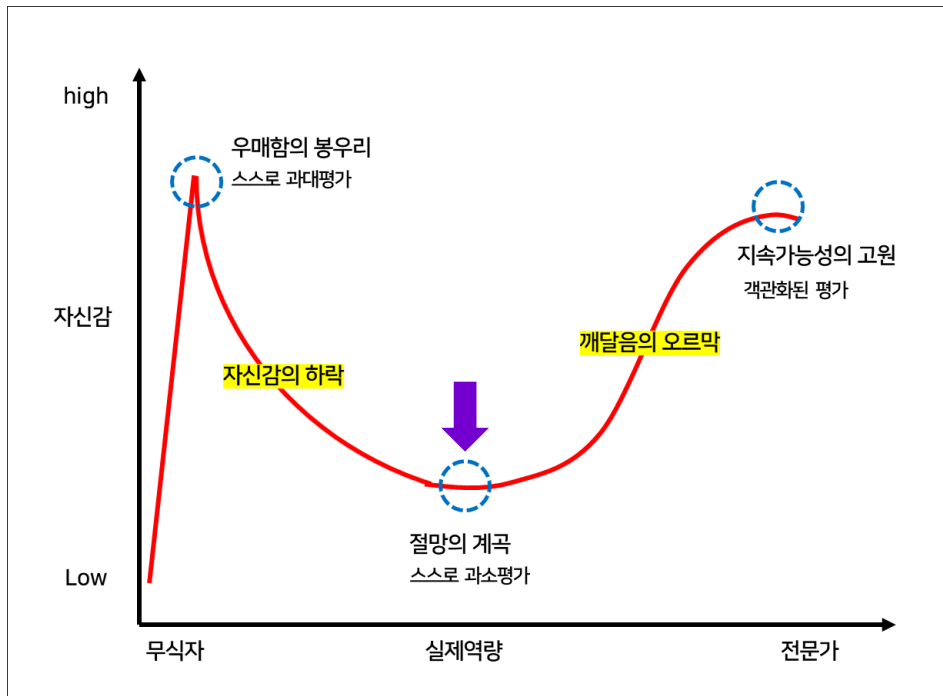
데이터야.놀자 2023



Source: <https://www.home-learn.co.kr/newsroom/news/A/1355>

# 지금 내 상태

데이터야. 놀자 2023



Source: <https://www.home-learn.co.kr/newsroom/news/A/1355>

# 왜 이런 실수를 저질렀을까..

의학 도메인

Data Scientist



온라인 서비스 도메인

Data Scientist

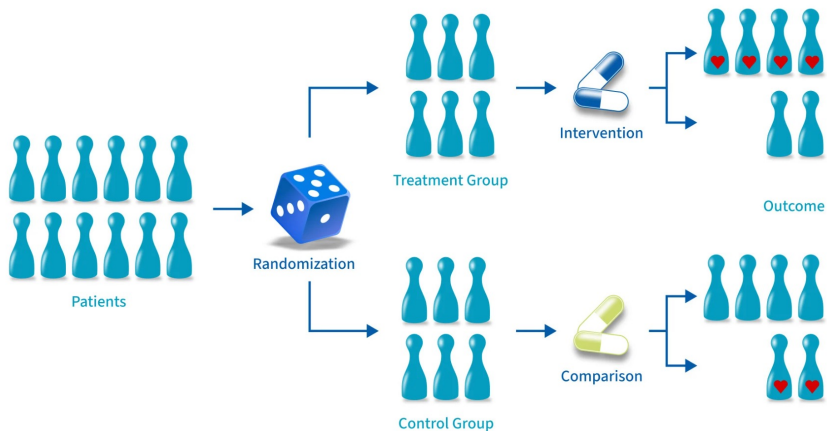


# RCT -> A/B Test

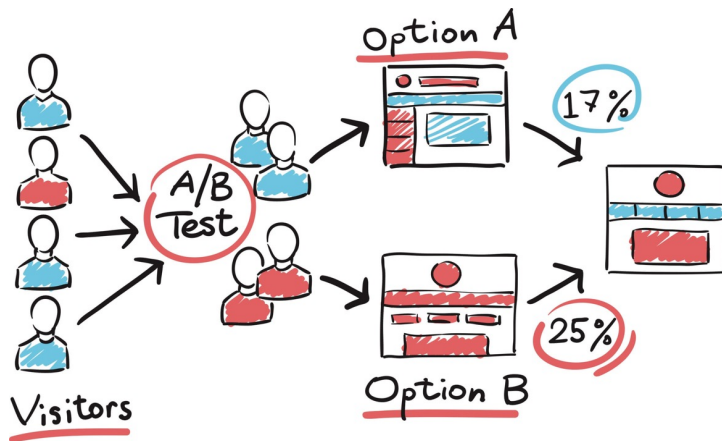
데이터야. 놀자 2023

거기서 거기 아니야?

Randomized Controlled Trial



Source: <https://www.simplypsychology.org/randomized-controlled-trial.html>



Source: <https://diggintravel.com/airline-ab-testing/>

# 둘은 어떻게 다른가?

RCT

A/B Test

# 둘은 어떻게 다른가?

## RCT

- 오프라인 환경
- 표본 확보가 어렵고 비용이 큼
- 반복 실험이 어려움
- 실험 설계가 복잡

## A/B Test

# 둘은 어떻게 다른가?

## RCT

- 오프라인 환경
- 표본 확보가 어렵고 비용이 큼
- 반복 실험이 어려움
- 실험 설계가 복잡
- 반복측정 분산분석, 요인 설계, 분할구 설계
- ...

## A/B Test

# 둘은 어떻게 다른가?

## RCT

- 오프라인 환경
- 표본 확보가 어렵고 비용이 큼
- 반복 실험이 어려움
- 실험 설계가 복잡
- 반복측정 분산분석, 요인 설계, 분할구 설계
- ...

## A/B Test

- 온라인 환경
- 표본 확보가 쉽고 비용이 작음
- 반복 실험이 용이
- 실험 설계가 매우 간단
- 자칫하면 쉽다고 생각될 수도..



# 무엇이 A/B Test를 어렵게 만드는가?

## A/B Test

- 온라인 환경
- 표본 확보가 쉽고 비용이 작음
- 반복 실험이 용이

**이제 본격적으로,  
신뢰하기 어려운 A/B Test의  
문제 사례를 제시하고자 합니다.**

# 다음의 경우에 해당하십니까?

이전 실험에 사용한 실험 모집단을 재사용하고 있음

사용자 단위의 지표를 사용하고 있지 않음

# 다음의 경우에 해당하십니까?

이전 실험에 사용한 실험 모집단을 재사용하고 있음

사용자 단위의 지표를 사용하고 있지 않음

A/A Test를 수행하고 있지 않음

**그럼,  
저는 당신의 A/B Test를  
신뢰하기 어려울 것 같습니다.**

# 그래서, 오늘 이야기 해볼 부분은..

이전 실험에 사용한 실험 모집단을 재사용하고 있음

사용자 단위의 지표를 사용하고 있지 않음

A/A Test를 수행하고 있지 않음

- 위 문제들을 인식할 수 있었던 계기
- 각 상황에 발생하는 함정과 해결책
- 그 외, 우리가 인식하고 있어야 할 것들

# 얻어가실 수 있는 것

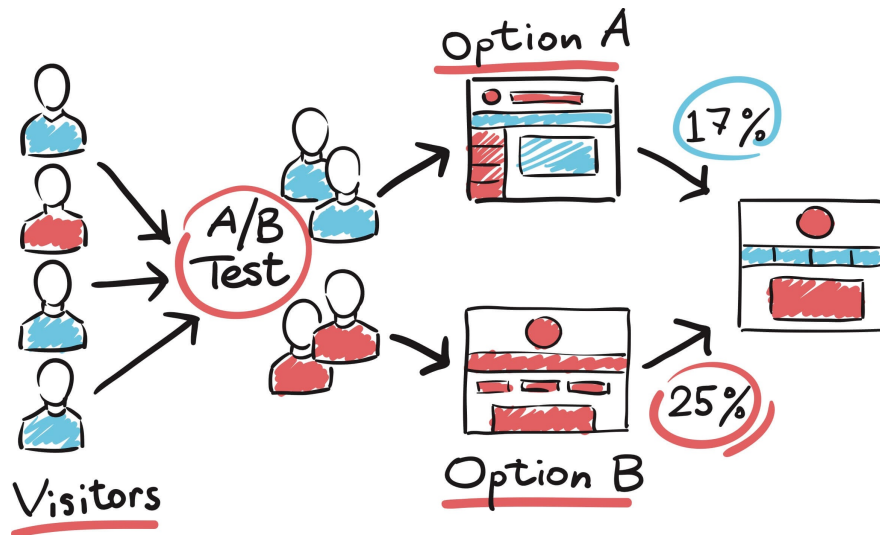
신뢰할 수 있는 A/B Test를 설계하고  
분석 결과를 정확하게 측정하는 것은  
이렇게나 어렵고, 많은 노력이 필요하구나.

**첫 번째 문제 상황,  
이전 실험에 사용한 모집단을  
재사용할 경우**



# 실험 모집단을 재사용하고 계십니까?

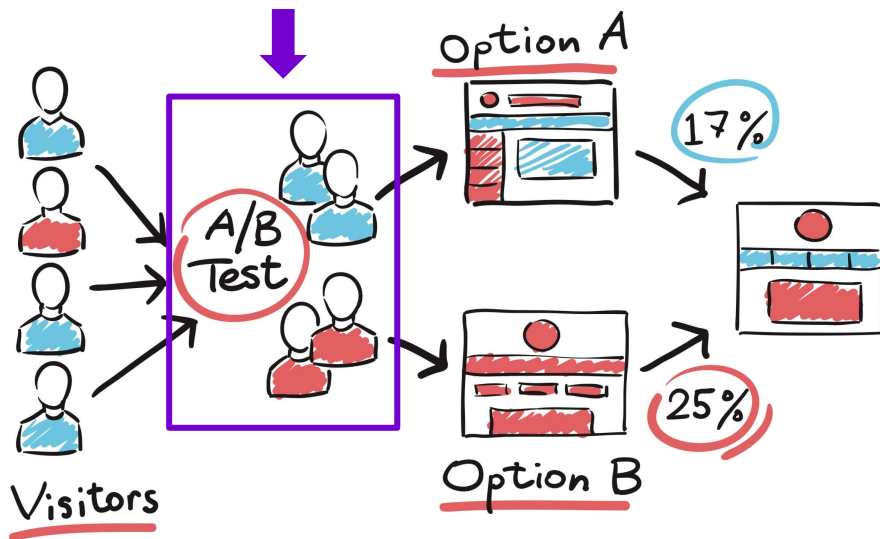
하기 세팅으로 2주간 실험 진행



Source: <https://diggintravel.com/airline-ab-testing/>

# 실험 모집단을 재사용하고 계십니까?

실험 종료 후, 해당 모집단을 곧바로/1-2주 이내 다음 실험에 투입



Source: <https://diggintravel.com/airline-ab-testing/>

# 어떤 문제가 발생하는가?

우연한 계기로 수행한 A/A Test로 발견한 문제

→ 잔류 효과 (Carryover effect)

→ 실험 종료 후에도 실험 효과가 약 3주간 잔존

# 마이크로소프트의 연구 사례

데이터야. 놀자 2023

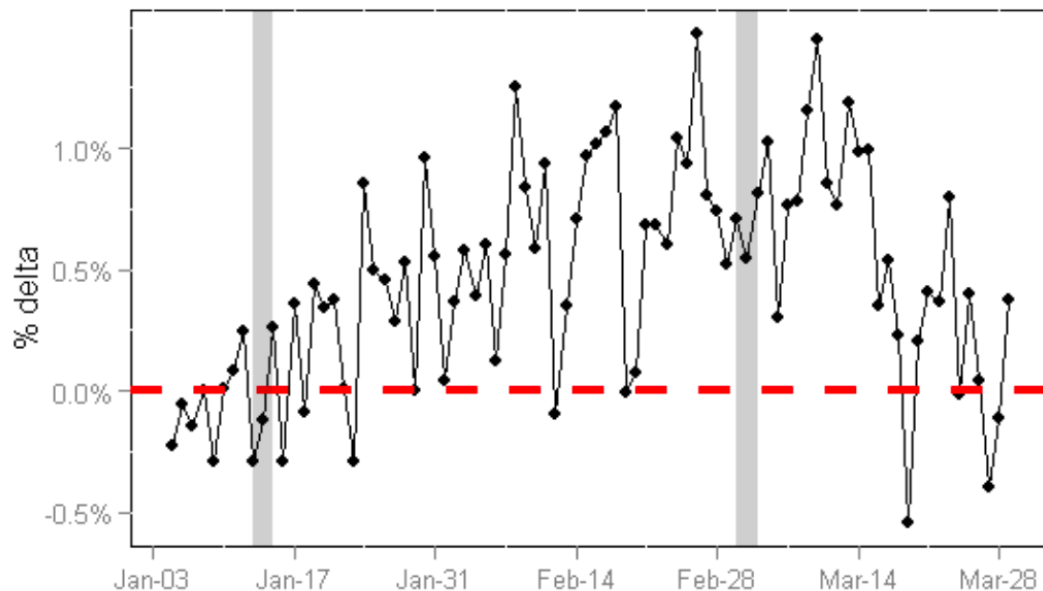


Figure 8: Carryover Effects Lasted Weeks

[Kohavi et al. 2012](#)

# 마이크로소프트의 연구 사례

데이터야. 놀자 2023

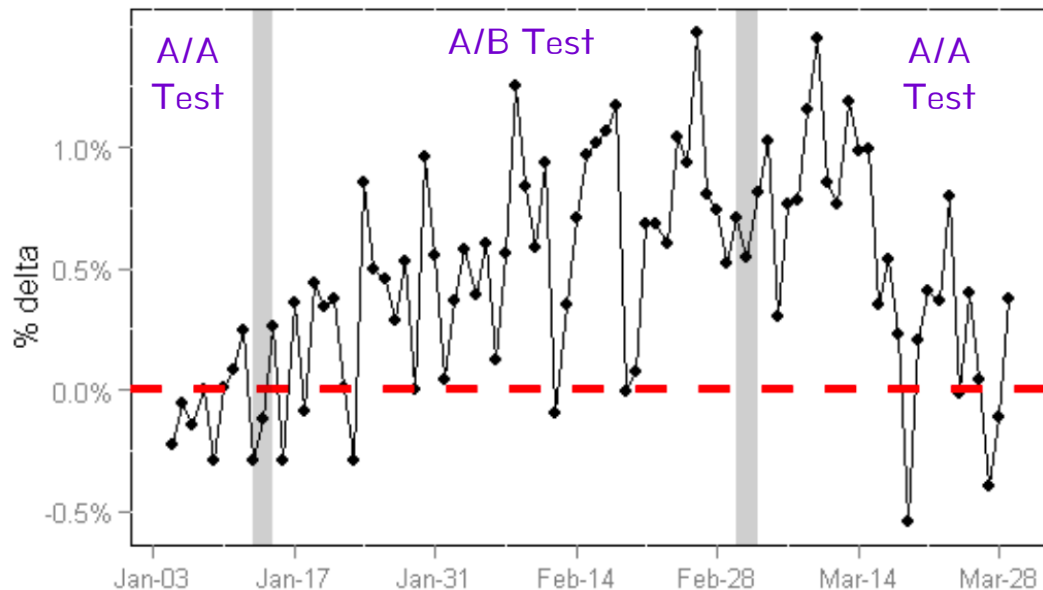


Figure 8: Carryover Effects Lasted Weeks

[Kohavi et al. 2012](#)

# 마이크로소프트의 연구 사례

데이터야. 놀자 2023

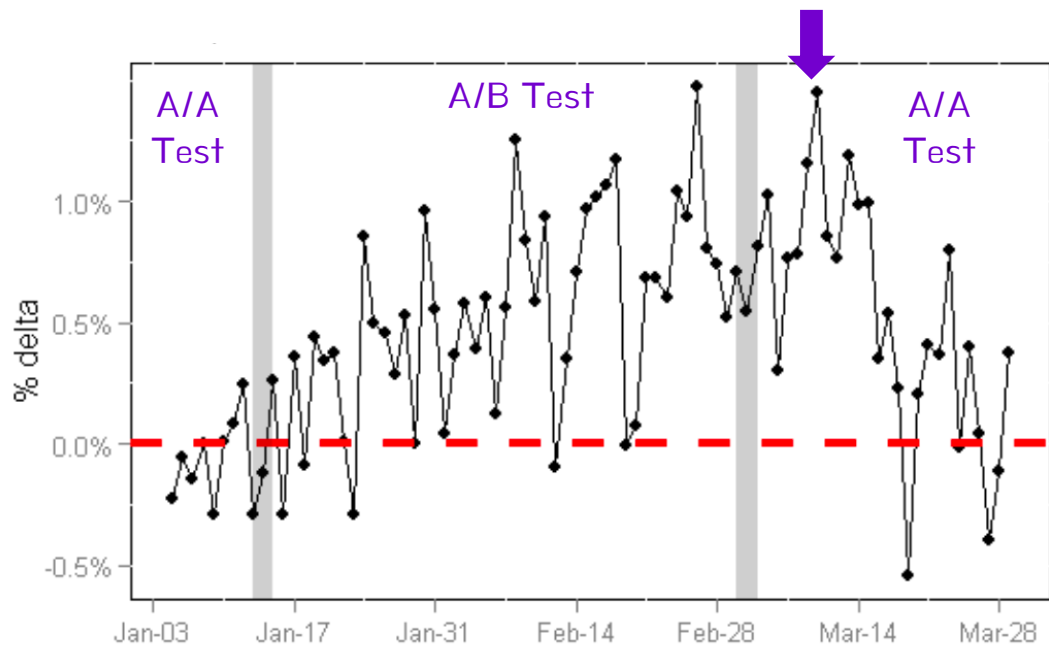


Figure 8: Carryover Effects Lasted Weeks

[Kohavi et al. 2012](#)

# 마이크로소프트의 연구 사례

데이터야. 놀자 2023

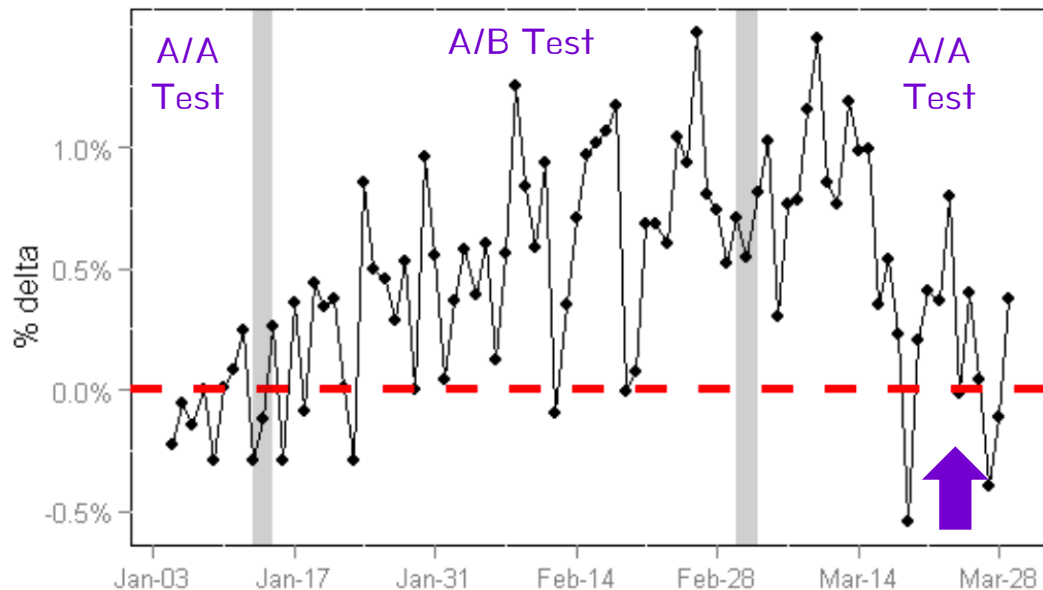


Figure 8: Carryover Effects Lasted Weeks

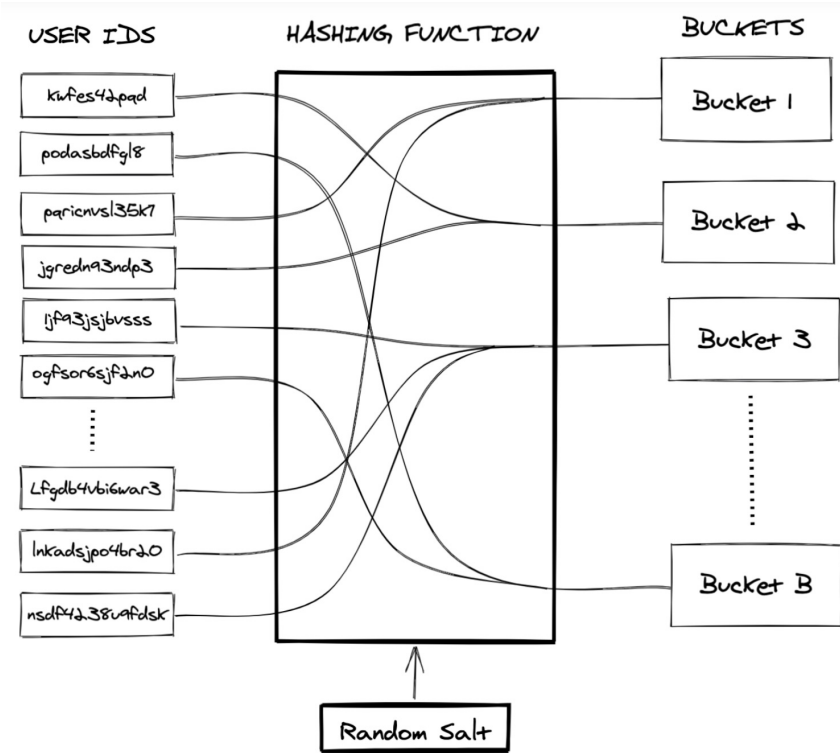
[Kohavi et al. 2012](#)

# 어떻게 해결할 수 있는가?

데이터야. 놀자 2023

실험을 진행할 때마다 Randomization

→ 즉, 실험군, 통제군 재할당(Re-randomization)



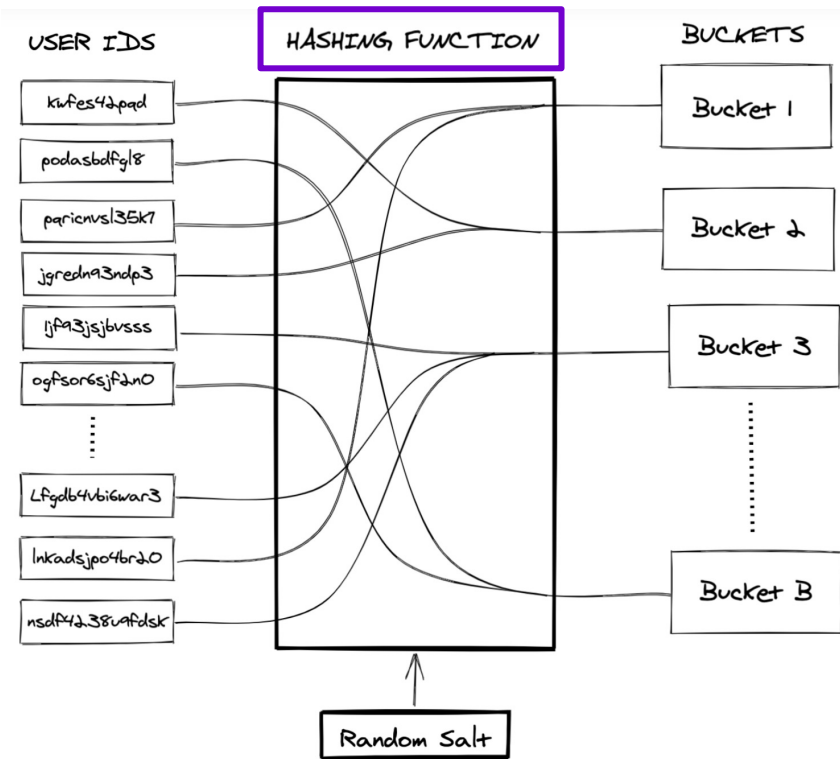


# 어떻게 해결할 수 있는가?

데이터야. 놀자 2023

실험을 진행할 때마다 Randomization

→ 즉, 실험군, 통제군 재할당(Re-randomization)

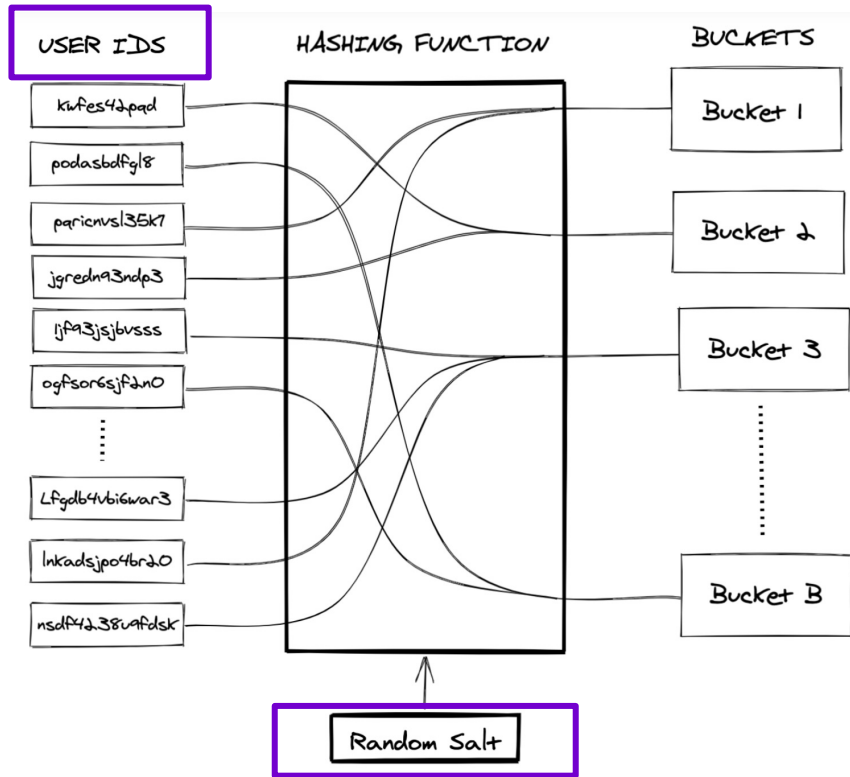


# 어떻게 해결할 수 있는가?

데이터야 놀자 2023

실험을 진행할 때마다 Randomization

→ 즉, 실험군, 통제군 재할당(Re-randomization)

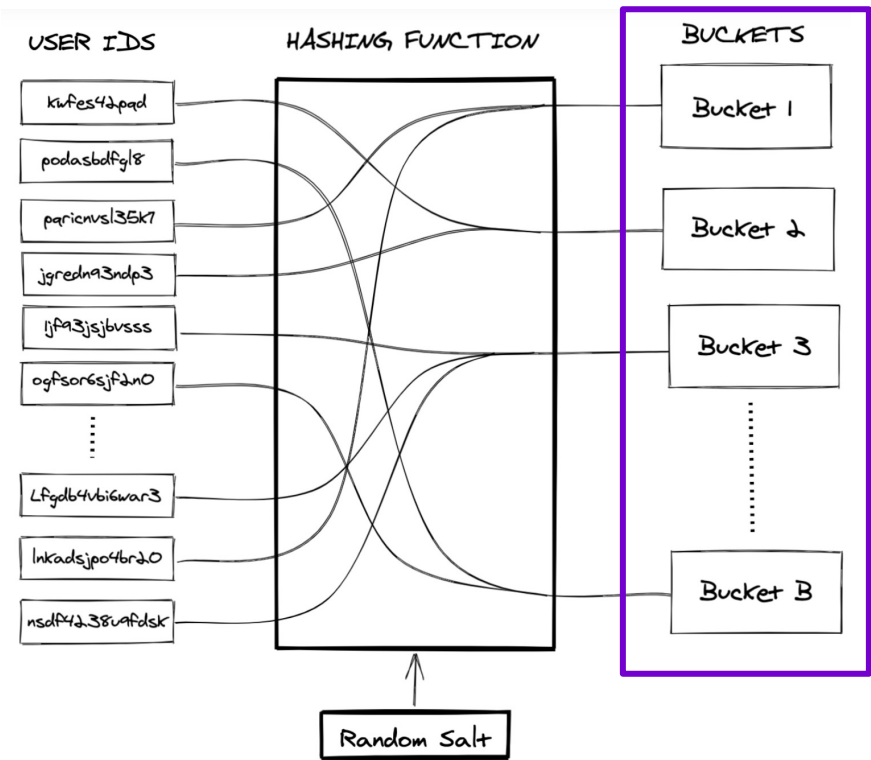


# 어떻게 해결할 수 있는가?

데이터야. 놀자 2023

실험을 진행할 때마다 Randomization

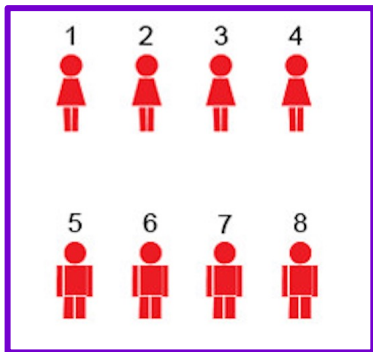
→ 즉, 실험군, 통제군 재할당(Re-randomization)



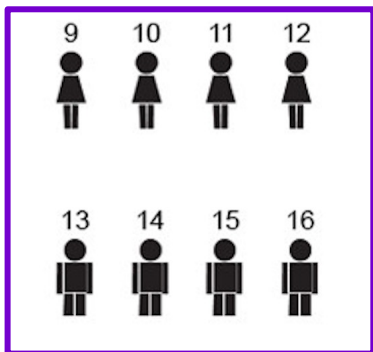
# Randomization의 효과

데이터야. 놀자 2023

실험군



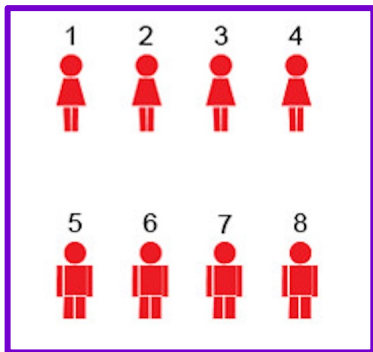
대조군



# Randomization의 효과

데이터야. 놀자 2023

실험군

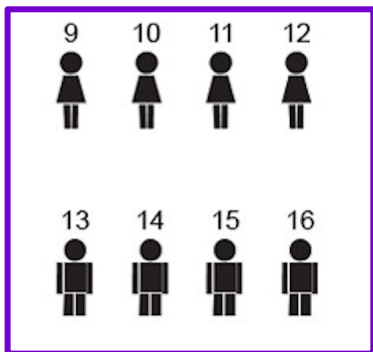


이 셋팅 그대로 재실험 하면?



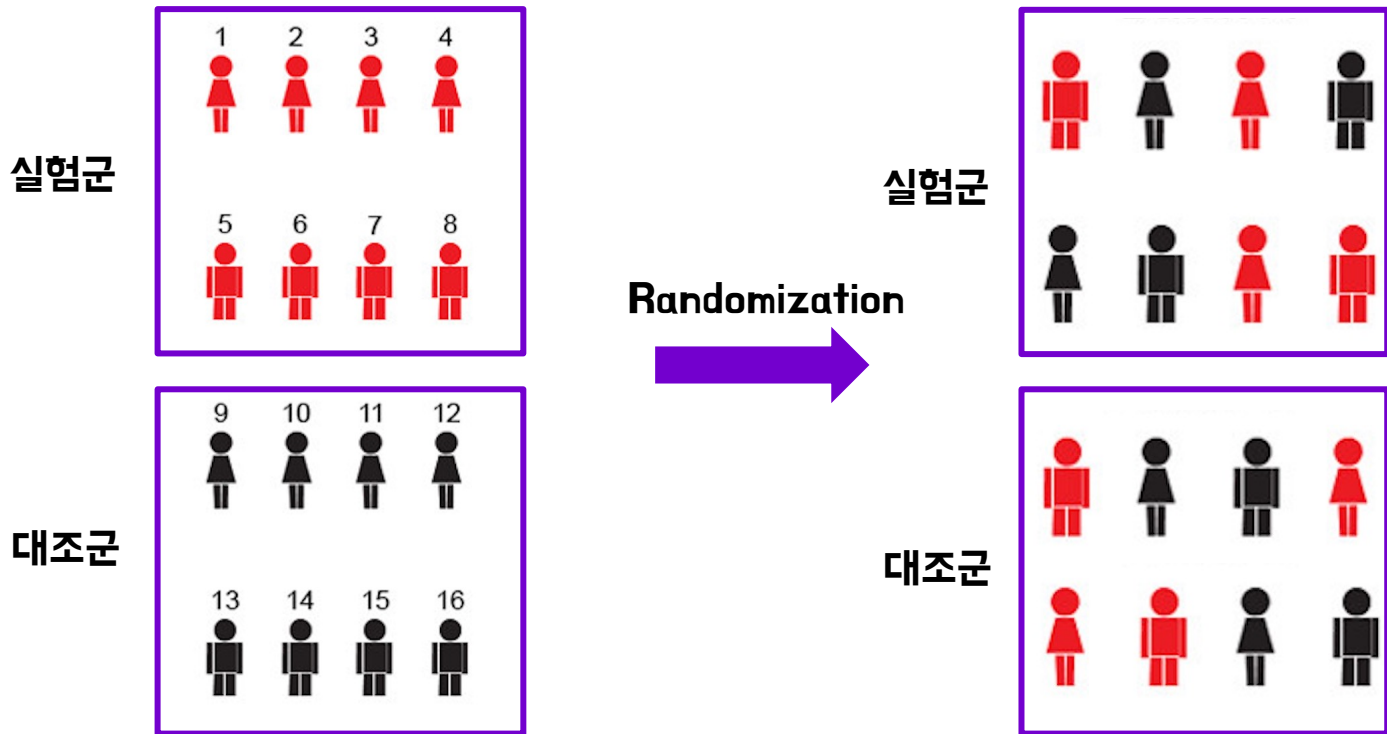
잔류 효과 발생

대조군



# Randomization의 효과

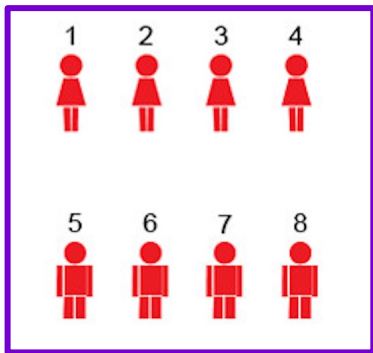
데이터야. 놀자 2023



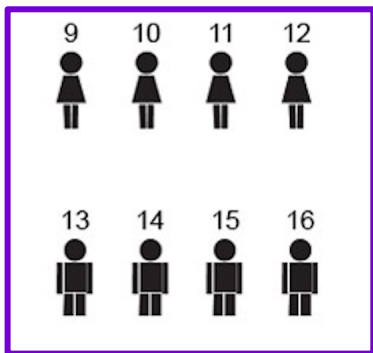
# Randomization의 효과

데이터야. 놀자 2023

실험군



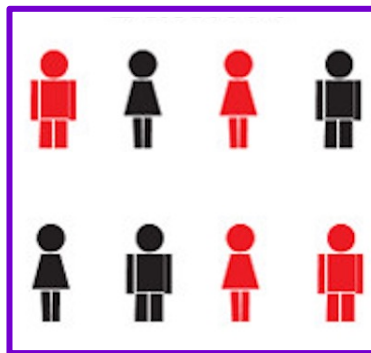
대조군



Randomization



실험군



대조군



잔류효과  
희석

**두 번째 문제 상황,  
사용자 단위 지표를  
사용하지 않는 경우**



# 사용자 단위 지표를 사용하지 않고 계십니까?

먼저, 실험 단위와 분석 단위에 대해 알아보자

# 사용자 단위 지표를 사용하지 않고 계십니까?

먼저, 실험 단위와 분석 단위에 대해 알아보자

→ (일반적인) 실험 단위: 사용자

→ 분석 단위: 분석에 사용되는 지표의 단위

# 실험 단위 $\neq$ 분석 단위

분석 단위: 노출

→ CTR

# 실험 단위 $\neq$ 분석 단위

분석 단위: 노출

→ CTR

분석 단위: 사용자

→ Click per visitor

→ GMV per visitor

→ Order per visitor

→ ...

# 실험 단위 $\neq$ 분석 단위

분석 단위: 노출

→ CTR

분석 단위: 사용자

→ Click per visitor

→ GMV per visitor

→ Order per visitor

→ ...

# 실험 단위 $\neq$ 분석 단위

분석 단위: 노출

→ CTR

→ 실험 단위  $\neq$  분석 단위

→ 문제 발생!

분석 단위: 사용자

→ Click per visitor

→ GMV per visitor

→ Order per visitor

→ ...

# 어떤 문제가 발생하는가?

지표의 분산에 Bias가 생깁니다.

# 어떤 문제가 발생하는가?

지표의 분산에 Bias가 생깁니다.

→ 우리의 관심 지표를  $\bar{Y}$ 라고 하자.



# 어떤 문제가 발생하는가?

지표의 분산에 Bias가 생깁니다.

→ 우리의 관심 지표를  $\bar{Y}$ 라고 하자.

$$\rightarrow \bar{Y} = \frac{1}{n} \sum Y_i$$

$$Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$

$$Var(\bar{Y}) = Var\left(\frac{1}{n} \sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

# 어떤 문제가 발생하는가?

지표의 분산에 Bias가 생깁니다.

→ 우리의 관심 지표를  $\bar{Y}$ 라고 하자.

$$\bar{Y} = \frac{1}{n} \sum Y_i$$

➡  $Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$

$$Var(\bar{Y}) = Var\left(\frac{1}{n} \sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

# 어떤 문제가 발생하는가?

지표의 분산에 Bias가 생깁니다.

→ 우리의 관심 지표를  $\bar{Y}$ 라고 하자.

$$\bar{Y} = \frac{1}{n} \sum Y_i$$

$$Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$

$$\rightarrow Var(\bar{Y}) = Var\left(\frac{1}{n} \sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

# 어떤 문제가 발생하는가?

지표의 분산에 Bias가 생깁니다.

→ 우리의 관심 지표를  $\bar{Y}$ 라고 하자.

$$\bar{Y} = \frac{1}{n} \sum Y_i$$

$$Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$

$$Var(\bar{Y}) = Var\left(\frac{1}{n} \sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

**Note.**


표준편차(standard deviation)과 표준오차(standard error)를 혼동하지 말자!

# 어떤 문제가 발생하는가?

지표의 분산에 Bias가 생깁니다.

→ 우리의 관심 지표를  $\bar{Y}$ 라고 하자.

$$\bar{Y} = \frac{1}{n} \sum Y_i$$


$$Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$

$$Var(\bar{Y}) = Var\left(\frac{1}{n} \sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

**Note.**

표준편차(standard deviation)과 표준오차(standard error)를 혼동하지 말자!


# 어떤 문제가 발생하는가?

지표의 분산에 Bias가 생깁니다.

→ 우리의 관심 지표를  $\bar{Y}$ 라고 하자.

$$\bar{Y} = \frac{1}{n} \sum Y_i$$

$$Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$


$$Var(\bar{Y}) = Var\left(\frac{1}{n} \sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

**Note.**

표준편차(standard deviation)과 표준오차(standard error)를 혼동하지 말자!

# 분산의 Bias가 야기하는 문제

실험군과 통제군의 지표 간 차에 관심이 있다고 하자.

$$\text{검정통계량} = \frac{\bar{Y}_{trt} - \bar{Y}_{con}}{S.E(\bar{Y}_{trt} - \bar{Y}_{con})} = \frac{\bar{Y}_{trt} - \bar{Y}_{con}}{\sqrt{Var(\bar{Y}_{trt} - \bar{Y}_{con})}}$$

\*  $S.E$  = 표준오차



# 분산의 Bias가 야기하는 문제

실험군과 통제군의 지표 간 차에 관심이 있다고 하자.

$$\text{검정통계량} = \frac{\bar{Y}_{trt} - \bar{Y}_{con}}{S.E(\bar{Y}_{trt} - \bar{Y}_{con})} = \frac{\bar{Y}_{trt} - \bar{Y}_{con}}{\sqrt{Var(\bar{Y}_{trt} - \bar{Y}_{con})}}$$

\*  $S.E$  = 표준오차



분산 과대 추정

→ False negative

분산 과소 추정

→ False positive



**실험 단위  $\neq$  분석 단위, 그런데 왜 Bias가 발생하는가?**

(일반적인) 실험 단위: 사용자

## 실험 단위 $\neq$ 분석 단위, 그런데 왜 Bias가 발생하는가?

(일반적인) 실험 단위: 사용자

→ 실험 단위와 분석 단위가 다르면?

→ i.i.d. 가정이 깨짐 (independent and identically distributed)

→ 통계학의 확률표본(random sample)에 깔려있는 기본 가정

## 실험 단위 $\neq$ 분석 단위, 그런데 왜 Bias가 발생하는가?

(일반적인) 실험 단위: 사용자

- 실험 단위와 분석 단위가 다르면?
- i.i.d. 가정이 깨짐 (independent and identically distributed)
- 통계학의 확률표본(random sample)에 깔려있는 기본 가정
- 분산 추정에 Bias 발생

**마이크로소프트 연구진들은 이를 원래부터 알고있었는가?**

**마이크로소프트 연구진들은 이를 원래부터 알고있었는가?**

**No, 그들 또한 이 부분을 완벽히 놓치고 있었다.**

**마이크로소프트 연구진들은 이를 원래부터 알고있었는가?**

**No, 그들 또한 이 부분을 완벽히 놓치고 있었다.**

**→ 모르고 지나칠 수 있는 함정은 어디서, 언제든지 발생할 수 있다**

**그렇다면 그들은 이 함정을 어떻게 인지할 수 있었는가?**

**그렇다면 그들은 이 함정을 어떻게 인지할 수 있었는가?**

**A/A Test의 병행을 통해서**



## A/A Test를 통해 어떻게 이를 발견할 수 있었는가?

관심 지표가 1,000개라고 하고, A/A Test 기간 내 관측한 지표들의 p-value를 구했다고 하자.

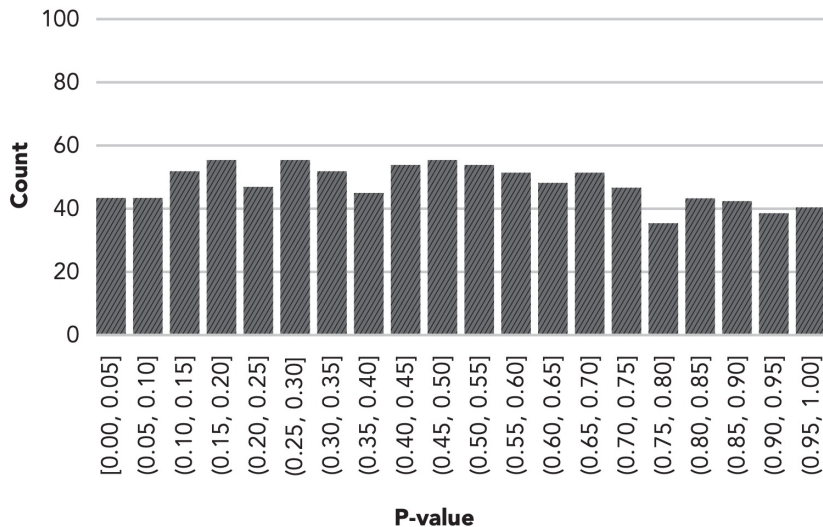
→ 이때 1,000개의 p-value에 관한 히스토그램을 그려보면

## A/A Test를 통해 어떻게 이를 발견할 수 있었는가?

관심 지표가 1,000개라고 하고, A/A Test 기간 내 관측한 지표들의 p-value를 구했다고 하자.

→ 이때 1,000개의 p-value에 관한 히스토그램을 그려보면

→ 다음과 같이 균일 분포를 띠어야함([Blocker et al., 2006](#))

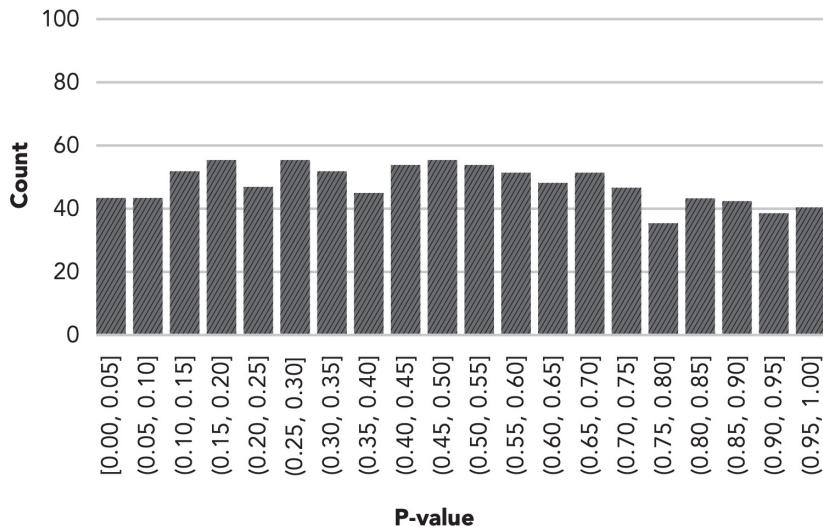


## A/A Test를 통해 어떻게 이를 발견할 수 있었는가?

관심 지표가 1,000개라고 하고, A/A Test 기간 내 관측한 지표들의 p-value를 구했다고 하자.

→ 이때 1,000개의 p-value에 관한 히스토그램을 그려보면

→ 다음과 같이 균일 분포를 띠어야함([Blocker et al., 2006](#))



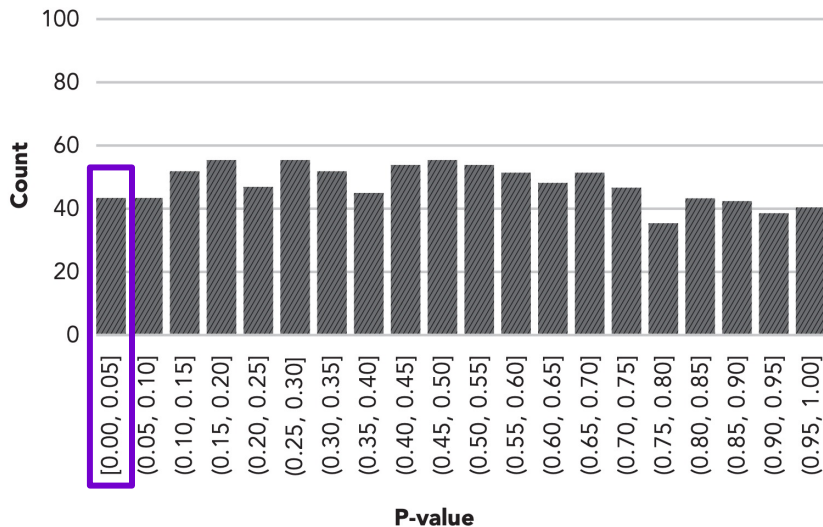
A/A Test  
성공

## A/A Test를 통해 어떻게 이를 발견할 수 있었는가?

관심 지표가 1,000개라고 하고, A/A Test 기간 내 관측한 지표들의 p-value를 구했다고 하자.

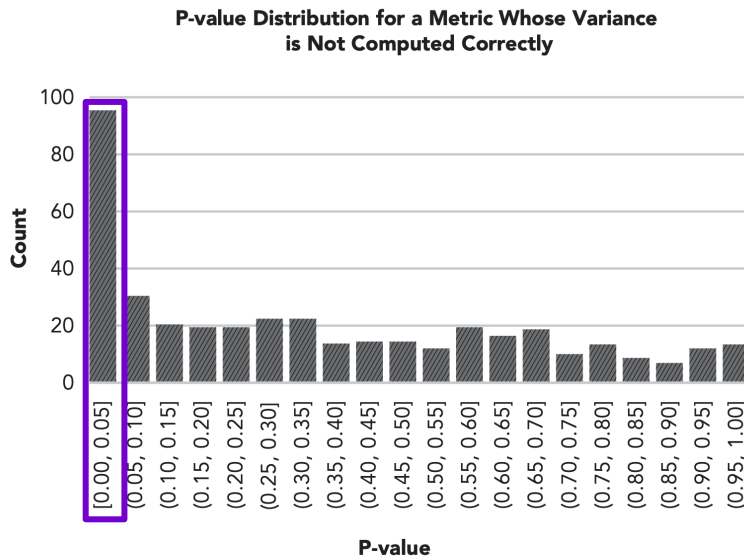
→ 이때 1,000개의 p-value에 관한 히스토그램을 그려보면

→ 다음과 같이 균일 분포를 띠어야함([Blocker et al., 2006](#))



A/A Test  
성공

그런데, 마이크로소프트에서 히스토그램을 그려본 결과..



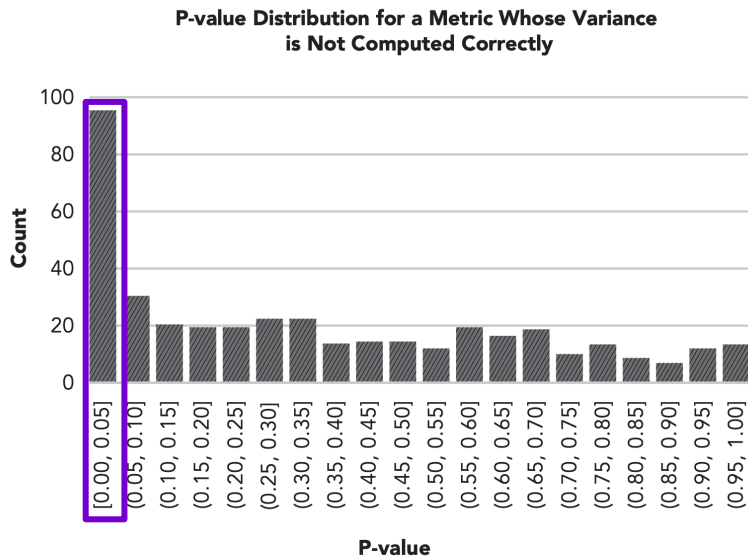
A/A Test  
실패

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

[Kohavi, Tang, and Xu 2020](#)

# 원인이 뭘까?..

데이터야. 놀자 2023



실험 단위  $\neq$  분석 단위  
i.i.d. 가정 깨짐  
분산 과소추정

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

[Kohavi, Tang, and Xu 2020](#)

## 우린 사용자 단위의 다른 지표가 꼭 필요해

e.g. CTR

→ 어떻게 해결할 수 있을까?

→ Delta Method를 통한 지표의 분산 추정 ([Deng, Lu, and Litz, 2017](#))

## 우린 사용자 단위의 다른 지표가 꼭 필요해

e.g. CTR

→ 어떻게 해결할 수 있을까?

→ Delta Method를 통한 지표의 분산 추정 ([Deng, Lu, and Litz, 2017](#))

→ Delta Method에 의해 조정된 대조군 대비 실험군의 CTR 변화율  $\Delta\% = \frac{\bar{Y}_{trt} - \bar{Y}_{con}}{\bar{Y}_{con}}$ 의 분산은?



# 우린 사용자 단위의 다른 지표가 꼭 필요해

e.g. CTR

→ 어떻게 해결할 수 있을까?

→ Delta Method를 통한 지표의 분산 추정 ([Deng, Lu, and Litz, 2017](#))

→ Delta Method에 의해 조정된 대조군 대비 실험군의 CTR 변화율  $\Delta\% = \frac{\bar{Y}_{trt} - \bar{Y}_{con}}{\bar{Y}_{con}}$ 의 분산은?

→  $Var(\Delta\%) = \frac{1}{\bar{Y}_{con}^2} Var(\bar{Y}_{trt}) + \frac{\bar{Y}_{trt}^2}{\bar{Y}_{con}^4} Var(\bar{Y}_{con})$  ([Kohavi, Tang, and Xu 2020](#))

# 우린 사용자 단위의 다른 지표가 꼭 필요해

e.g. CTR

→ 어떻게 해결할 수 있을까?

→ Delta Method를 통한 지표의 분산 추정 ([Deng, Lu, and Litz, 2017](#))

→ Delta Method에 의해 조정된 대조군 대비 실험군의 CTR 변화율  $\Delta\% = \frac{\bar{Y}_{trt} - \bar{Y}_{con}}{\bar{Y}_{con}}$ 의 분산은?

→  $Var(\Delta\%) = \frac{1}{\bar{Y}_{con}^2} Var(\bar{Y}_{trt}) + \frac{\bar{Y}_{trt}^2}{\bar{Y}_{con}^4} Var(\bar{Y}_{con})$  ([Kohavi, Tang, and Xu 2020](#))

→  $Var(\Delta\%) = Var\left(\frac{\bar{Y}_{trt} - \bar{Y}_{con}}{\bar{Y}_{con}}\right) = Var\left(\frac{\bar{Y}_{trt}}{\bar{Y}_{con}}\right)$  (Delta Method 적용 X)

# 우린 사용자 단위의 다른 지표가 꼭 필요해

e.g. CTR

→ 어떻게 해결할 수 있을까?

→ Delta Method를 통한 지표의 분산 추정 ([Deng, Lu, and Litz, 2017](#))

→ Delta Method에 의해 조정된 대조군 대비 실험군의 CTR 변화율  $\Delta\% = \frac{\bar{Y}_{trt} - \bar{Y}_{con}}{\bar{Y}_{con}}$ 의 분산은?

→  $Var(\Delta\%) = \frac{1}{\bar{Y}_{con}^2} Var(\bar{Y}_{trt}) + \frac{\bar{Y}_{trt}^2}{\bar{Y}_{con}^4} Var(\bar{Y}_{con})$  ([Kohavi, Tang, and Xu 2020](#))

→  $Var(\Delta\%) = Var\left(\frac{\bar{Y}_{trt} - \bar{Y}_{con}}{\bar{Y}_{con}}\right) = Var\left(\frac{\bar{Y}_{trt}}{\bar{Y}_{con}}\right)$  (Delta Method 적용 X)

→ Delta Method에 관한 더 자세한 사항은 [Deng, Lu, and Litz, 2017](#) 참고

# 지금까지 한 이야기들을 정리하면

실험 모집단 재사용

→ 발생하는 문제: 잔류 효과

→ 발견 계기: A/A Test

→ 해결 방법: 새로운 실험 할당 시 Re-Randomization 통한 새로운 실험 집단 구성

# 지금까지 한 이야기들을 정리하면

## 실험 모집단 재사용

→ 발생하는 문제: 잔류 효과

→ 발견 계기: A/A Test

→ 해결 방법: 새로운 실험 할당 시 Re-Randomization 통한 새로운 실험 집단 구성

사용자 단위 지표를 사용하지 않는 경우 (실험 단위  $\neq$  분석 단위)

→ 발생하는 문제: 지표의 분산 편향

→ 발견 계기: A/A Test

→ 해결 방법: Delta Method, Bootstrap sampling

# 지금까지 한 이야기들을 정리하면

## 실험 모집단 재사용

→ 발생하는 문제: 잔류 효과

→ 발견 계기: A/A Test

→ 해결 방법: 새로운 실험 할당 시 Re-Randomization 통한 새로운 실험 집단 구성

사용자 단위 지표를 사용하지 않는 경우 (실험 단위  $\neq$  분석 단위)

→ 발생하는 문제: 지표의 분산 편향

→ 발견 계기: A/A Test

→ 해결 방법: Delta Method, Bootstrap sampling

**마지막 문제 상황,  
우린 앞선 2가지 사례 모두 해당하지 않는데?  
A/A Test 필요없는거 아니야?**

**당연히 제 대답은..**

**NO.**

**→ A/A Test는 우리가 모르고 지나칠 수 있는 함정을 발견할 수 있는 유일한 무기**



# A/A Test의 중요성은 한없이 강조된다.

“온라인 통제 실험의 이론은 잘 알려져있지만 실제로 구현할 때는 여러가지 함정이 있다.

A/A Test는 실험 플랫폼에 대한 신뢰를 구축하는 데 매우 유용하다.”

- [Kohavi, Tang, and Xu 2020](#)

“우리는 분포 불일치와 플랫폼 이상을 포함한 문제를 발견하기 위해 다른 실험과 병행해서 계속해서 A/A Test를 실행하는 것을 강력하게 추천한다.”

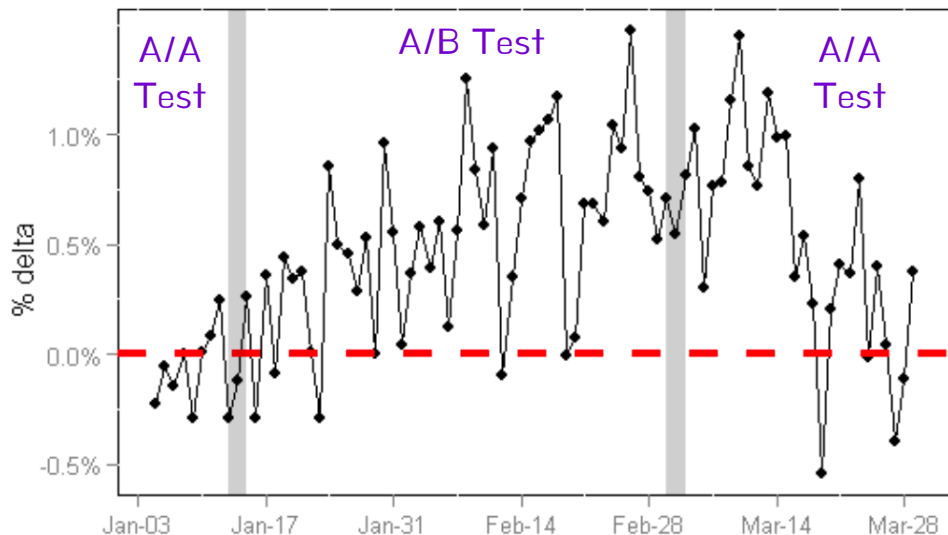
- [Kohavi, Tang, and Xu 2020](#)

# A/A Test의 중요성은 한없이 강조된다.

마이크로소프트의 ExP 팀이 약 2007년부터 쌓아온 온라인 통제실험과 관련한 수많은 함정과 레슨들은 결코 A/A Test 없이는 발견될 수 없었을 것

# A/A Test의 중요성은 한없이 강조된다.

마이크로소프트의 ExP 팀이 약 2007년부터 쌓아온 온라인 통제실험과 관련한 수많은 함정과 레슨들은 결코 A/A Test 없이는 발견될 수 없었을 것



**A/A Test를  
항상 병행함으로써  
A/B Test를 위한  
예열을 해두자!**

Figure 8: Carryover Effects Lasted Weeks

**아직 못다한 이야기..**

**우리가 추가적으로 인식하고 있어야할 것들**

# 너무나도 많습니다..

잔류 효과

지표의 분산 추정에서 발생하는 함정

A/A Test의 필요성

A/A Test를 병행하는 방법

CLT의 안정적인 보장을 위한 충분한 실험 대상의 확보 필요성

Non-exclusive 병렬 실험 환경을 갖춘 실험 플랫폼의 필요성

가드레일 지표의 필요성 (e.g. SRM, Latency)

검정력, MDE

다중검정

장기 효과의 측정 필요성

지표의 Variance reduction

...

# 너무나도 많습니다..

잔류 효과

지표의 분산 추정에서 발생하는 함정

A/A Test의 필요성

A/A Test를 병행하는 방법

CLT의 안정적인 보장을 위한 충분한 실험 대상의 확보 필요성

Non-exclusive 병렬 실험 환경을 갖춘 실험 플랫폼의 필요성

가드레일 지표의 필요성 (e.g. SRM, Latency)

검정력, MDE

다중검정

장기 효과의 측정 필요성

지표의 Variance reduction

...



관심 있으신 분들은  
편하게 말씀 주세요!

**이거 언제 다 공부하나요?...**

이 책만 보셔도 됩니다.

원서

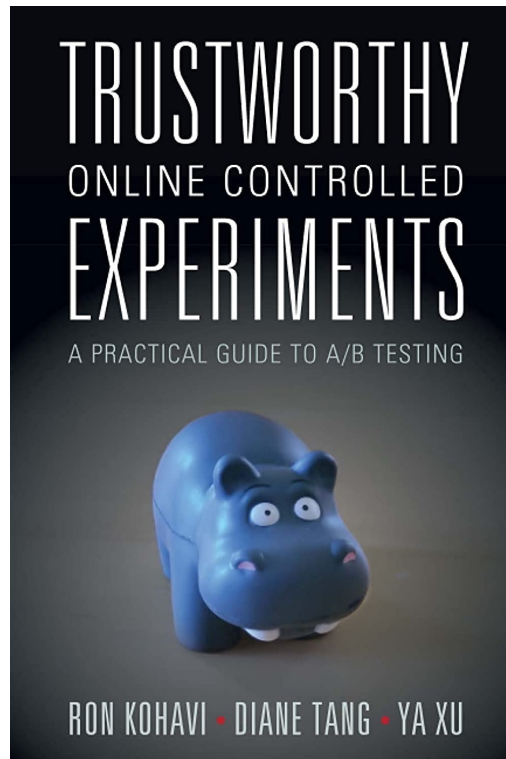


번역본



많은 행간이 생략된 어려운 책

Reference 표기가 상세함





**오늘 전달드린 것이 너무 많았죠?  
이것만 가져갑시다!**

**숫자를 얻는 것은 쉽다.**

**하지만, 신뢰할 수 있는 숫자를 얻는 것은 어렵다.**

**끊임없이 의심하자!**

**최소한 A/A Test로**