A/B 테스트 조금 더 잘하기

with 통계학 양념 한 소금

TABLE OF CONTENTS

왜 A/B 테스트를 하는가?

현실 세계에서

한정된 자원으로

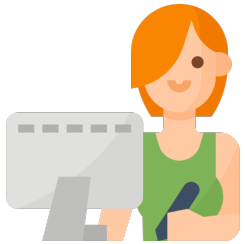
새로운 아이템의 효과를 검증하기 위한

가장 좋은 방법

Ceteris Paribus :

all things being equal, 다른 모든 조건이 동일하다면

왜 A/B 테스트를 하는가?



UX/UI 디자이너:
새로운 앱 디자인을 만들었습니다, 사장님!
이번에 이 UI를 새로 적용하면 구매 전환율이 무척 높아질 거에요!

사장님:
그렇구먼! 보기에도 좋고 아주 좋네.
바로 시행합세! (근데 실제 유저들이 싫어하면 어떻게 하지?)

왜 A/B 테스트를 하는가?

UX/UI 디자이너:
새로운 앱 디자인을 만들었습니다, 사장님!
이번에 이 UI를 새로 적용하면 구매 전환율이 무척 높아질 거에요! 가설

사장님:
그렇구먼! 보기에도 좋고 아주 좋네.
바로 시행합세! (근데 실제 유저들이 싫어하면 어떻게 하지?)

행동 리스크

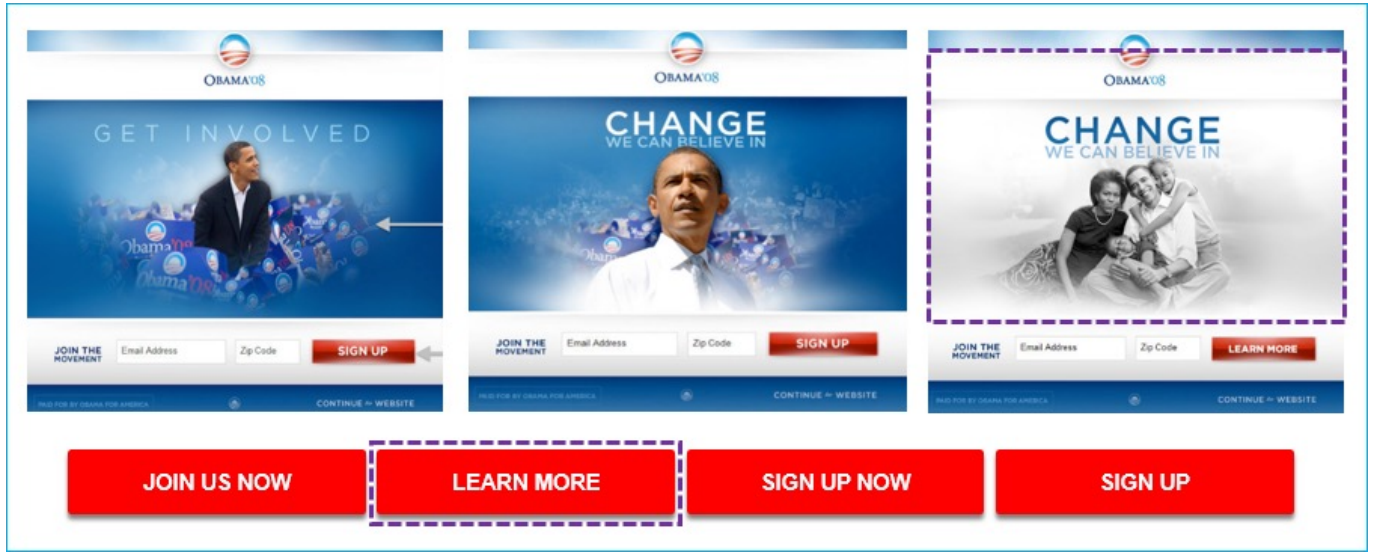
가설이 실제로 그럴 거라고 믿고, "진행시켜!"  라는 사장님의 명령에 UI를 바꿔버리면(**행동**)…

1) 가설이 맞는 경우(신규 UI => 전환율 증가)
: 문제 전혀 없고 신규 UI 그대로 쓰면 됨

2) 가설이 틀린 경우(신규 UI => 전환율 감소, **리스크**가 현실이 되는 순간)
: UI가 맘에 안 든 유저들이 이탈하고 다시는 돌아오지 않을 수도 있음 ㅠㅠ

왜 A/B 테스트를 하는가?

사례 1) 오바마 대선 캠프



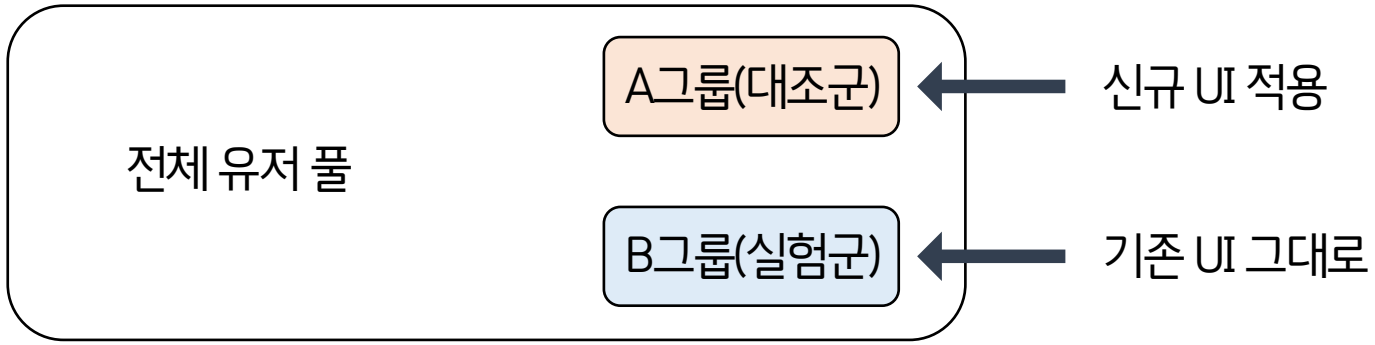
가족의 모습이 담긴 이미지와
LEARN MORE 버튼은
기존안 대비 각각 18.6%, 13.1%의
가입률 증가를 보였으며

두 아이템을 조합한 페이지의 경우
40.6% 가입률 증가

<A/B 테스트로 서비스 성장시키기>, 그로스해킹

왜 A/B 테스트를 하는가?

가설을 행동으로 바로 옮기기 전에
가설의 타당성과 리스크를 사전에 테스트해보아야 함
(마치 신약 임상실험을 하듯)



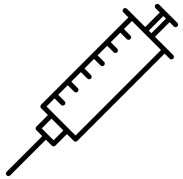
사장님:
흠…그럼 전체 유저 중에서 대략 사람들 뽑아서 A그룹과 B그룹으로 나누면 되나?

아니요. Ceteris Paribus를 확실히 해야 합니다!!!
(all things being equal, 다른 모든 조건이 동일하다면)

어떻게 해야 정확하게 할 수 있는가?

*신약 물질은 2~30대 여성에게 더 잘 반응하는 특징이 있음

A그룹(대조군)



위약 투여

바이러스
감소율 15%

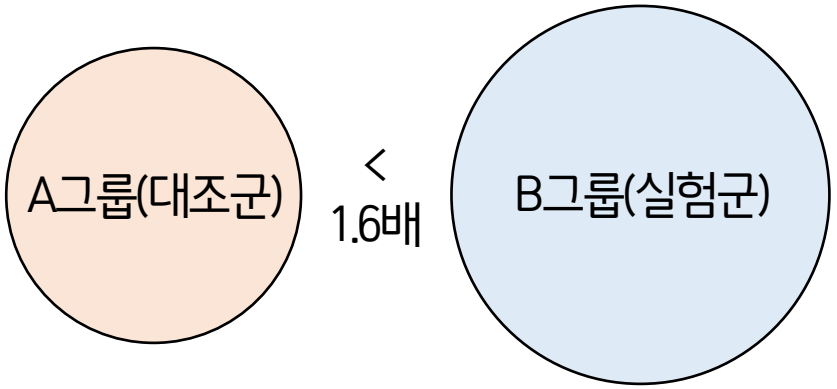
B그룹(실험군)

신약 투여

바이러스
감소율 25%

<
1.6배

그룹 별 2~30대 여성의 비율



결론 : 망한 실험
(약에 의한 효과인지
집단을 잘못 나눠서 생긴 효과인지 알 수 없음)

가급적 두 집단의 모든 조건을 동일하게 해야 한다!!!
Ceteris Paribus

어떻게 해야 정확하게 할 수 있는가?

두 집단을 최대한 동질적으로

뭘 동일하게 해야 하는가?

신약 임상실험	UI A/B 테스트
동질적인 환자가 실험/대조군을 구성하고 있는지 (연령, 성별, 지병 등)	동질적인 유저가 A/B 그룹을 구성하고 있는지 (연령, 성별, 기존 사용도 및 만족도 등)
실험이 어느 시점에 진행되는지 (동일 시점이 아니라면 병세가 악화될 수 있음)	테스트가 어떤 시점에 진행되는지 (동일 시점이 아니라면 User experience에 이미 변화가 있을 수 있음)
실험 환경은 동일한지	유저가 동일한 환경에서 UI에 노출되는지 (브라우져 등)

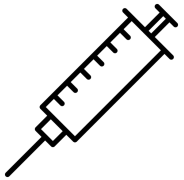
다만, A/B테스트의 경우
임상실험처럼 정확하게
A/B 집단을 나누는 것은 불가능하다

(유저들을 불러모아 가둬놓고
실험을 할 수 없으니…)

어떻게 해야 정확하게 할 수 있는가?

진짜 차이가 있는 건지 어떻게 알 수 있는가?

A그룹(대조군) B그룹(실험군)



위약 투여

신약 투여

바이러스
감소율 15%

바이러스
감소율 25%

Statistical Test 활용

T검정, 모비율 차이 검정 등

실험을 통해 얻은 A-B 그룹 간의 차이가
우연히 발생한 것은 아닌지 통계적으로 검증하는 것

진짜 차이가 있는 건지 어떻게 알 수 있는가?

실제로 차이가 있다는 것이 입증되기 위해서는…

1. 적정량의 데이터

케이스 1)
각 **3명**으로 이루어진 A/B 그룹에 대해 실험을 해 보니,
B그룹의 지표가 2% 더 높았다.

케이스 2)
각 **30명**으로 이루어진 A/B 그룹에 대해 실험을 해 보니,
B그룹의 지표가 2% 더 높았다.

진짜 차이가 있는 건지 어떻게 알 수 있는가?

실제로 차이가 있다는 것이 입증되기 위해서는…

2. 적정량의 차이

케이스 1)
A/B 그룹에 대해 실험 후 지표를 보니 각각 20%, 20.3%의 지표를 보였다.
: 0.3% 정도의 차이는 우연히 발생할 수도 있지 않을까?

케이스 2)
A/B 그룹에 대해 실험 후 지표를 보니 각각 20%, 25%의 지표를 보였다.
: 5% 정도면 **확실히 차이가 있네!**

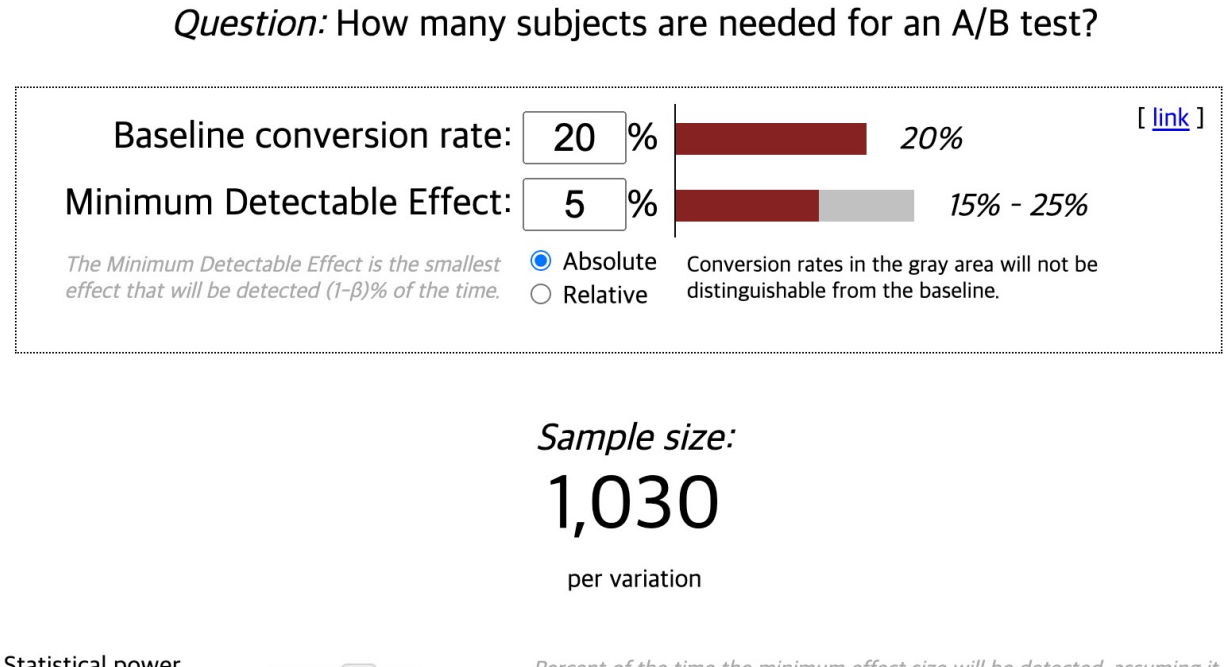
“확실히 차이가 있네!”라는 문장에 수치적인 근거를 보태는 것이 통계 검정의 위력!

어떻게 해야 정확하게 할 수 있는가?

진짜 차이가 있는 건지 어떻게 알 수 있는가?

적정량의 데이터가 얼마인지 어떻게 아는가?

https://www.evanmiller.org/ab-testing/sample-size.html



1) Baseline:
기존 지표를 통해 파악한
현재의 통상적인 지표 수준

2) Minimum Detectable Effect:
"이만큼은 차이가 나야
차이가 있다고 판단하겠다"고
연구자가 설정한 값

3) Statistical Power : (1 – 2종 오류)
1 –
실제로는 A/B 그룹에 차이가 있는데도
차이가 없다고 할 확률

4) Significance Level: 유의수준, 1종 오류
실제로는 차이가 없는데도
차이가 있다고 할 확률

어떻게 해야 정확하게 할 수 있는가?

진짜 차이가 있는 건지 어떻게 알 수 있는가?

적정량의 차이인지 어떻게 아는가?: p-value

p값에 의해서 실제로 차이가 있는지 알기 위해서는…

1) A/B 집단의 지표 차이가 없다고 가정한다(귀무가설, 무죄 추정의 원칙)

2) 그런데 실험을 통해 얻은 데이터를 실제로 보니
차이가 없다고 하기에는 너무 차이가 많이 난다
(예를 들면 10%의 차이라던지…) (검정통계량)

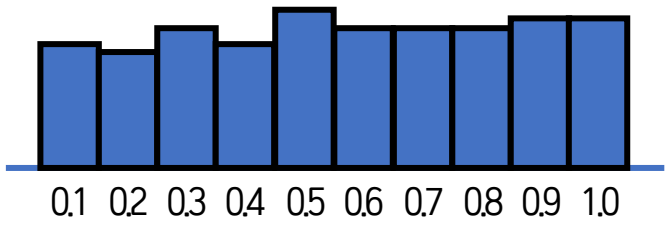
3) 그 너무 많은 10%의 차이가 우연히 발생했을 확률이 p값이다
"실제로 A/B 집단 사이의 차이는 없는데,
우연히 그런 큰 차이가 발생했을 확률"

보통 그 확률이 5% 이하면(즉 p값이 0.05 이하라면),
차이가 없다고 더 이상 가정할 수 없다고 이야가한다
(귀무가설 기각)

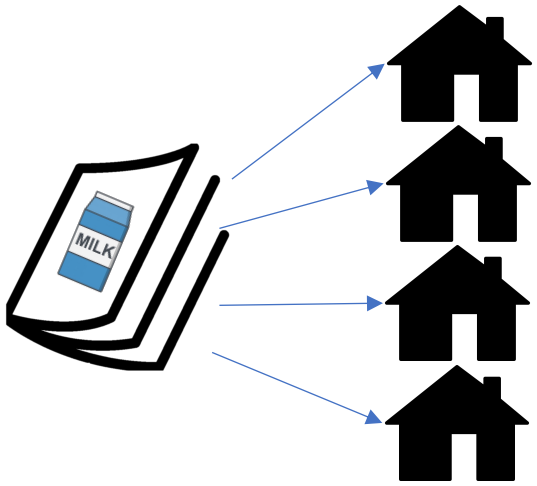
A/B 테스트를 수행하는 다양한 방법

유제품 회사 사장님:
요즘 판매량이 너무 부진한데… 홍보물 하나 제작해서 서울시 가정에 뿌려봐!

홍보물 전달 후
구매할 사전 확률



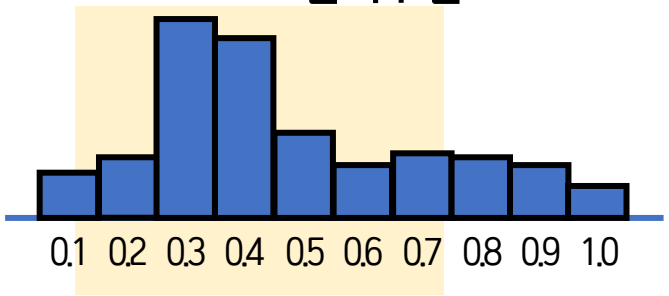
사전 확률: 믿음 혹은 기존의 정보



30 가정에 홍보물 전달 후 확인해 보니,
총 11가정이 유제품을 구매함
(11/30 ~ 0.3)

홍보물 전달 후
구매할 사후 확률

평균: 0.3,
95% 신뢰구간: 0.12 ~0.78

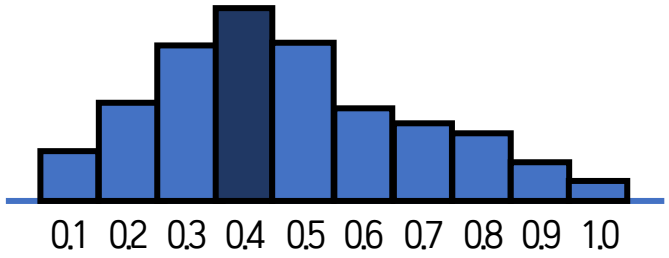


사후 확률:
실제로 데이터를 수집 후
업데이트한 확률

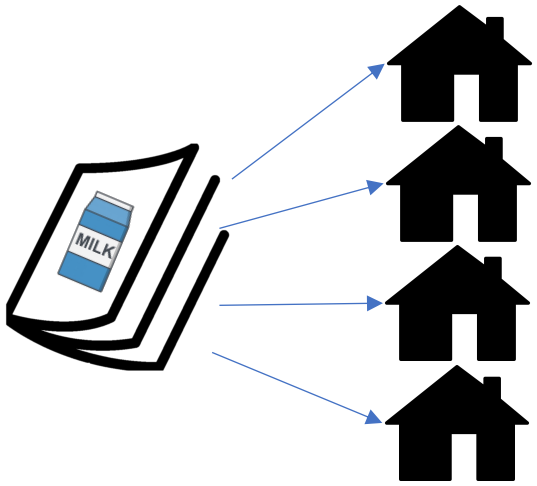
A/B 테스트를 수행하는 다양한 방법

새로 오신 에이스 상무님:
내 기존 판촉 경험에 따르면, 홍보물 전달하고 나서 보통 40% 정도는 구매하던데?

홍보물 전달 후
구매할 사전 확률



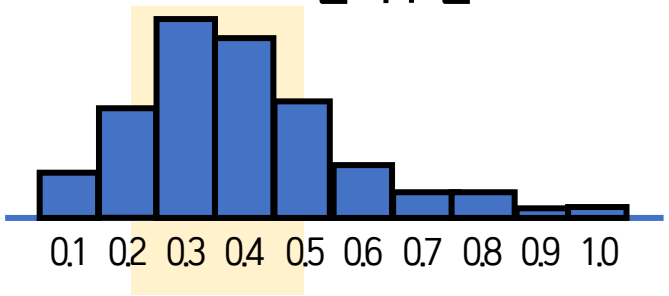
사전 확률: 믿음 혹은 기존의 정보



30 가정에 홍보물 전달 후 확인해 보니,
총 11가정이 유제품을 구매함
(11/30 ~ 0.3)

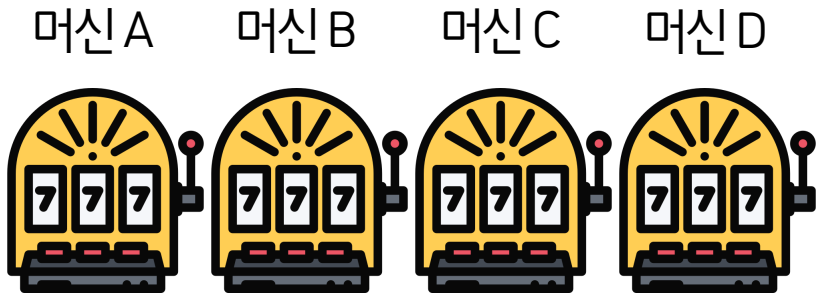
홍보물 전달 후
구매할 사후 확률

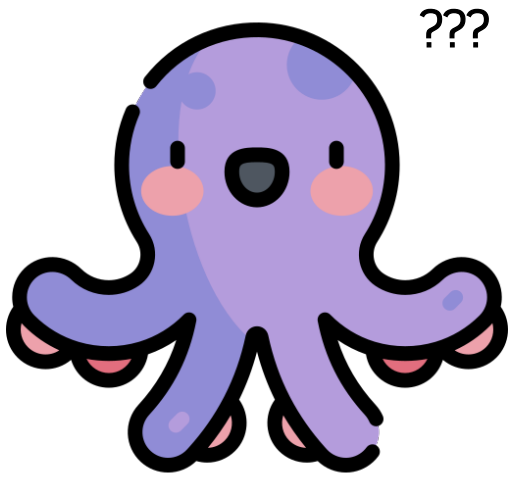
평균: 0.35,
95% 신뢰구간: 0.28~0.43



사후 확률:
실제로 데이터를 수집 후
업데이트한 확률

A/B 테스트를 수행하는 다양한 방법





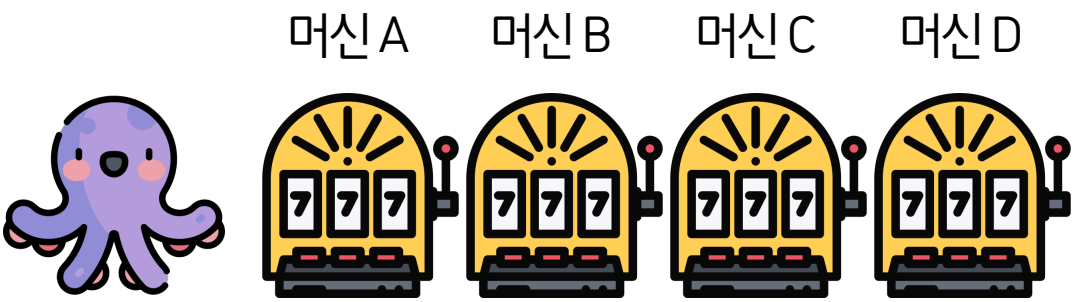
슬롯머신마다 수익률이 다르다고 가정하자.
문어씨는 슬롯머신 중 어떤 슬롯머신이 수익률이 높은지 모른다.

그래서 팔이 많은 문어씨의 장점을 활용해
모든 슬롯머신을 동시에 땡겨 보았다.

이 때 각각의 머신에서 다음과 같은 수익을 얻었다고 하자.

머신	수익
A	1000
B	50
C	0
D	500

A/B 테스트를 수행하는 다양한 방법



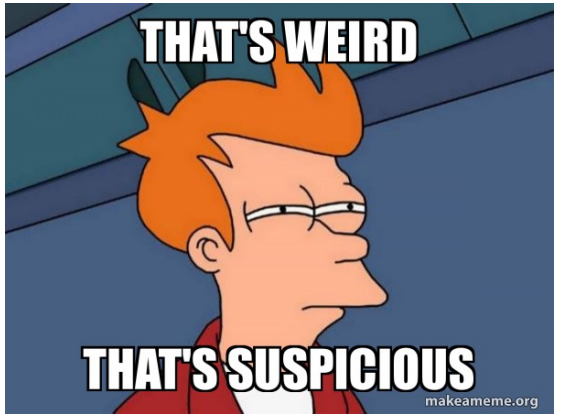
머신	수익
A	1000
B	50
C	0
D	500

Exploitation 활용
지르고 보는 타입

인생은 한방이다.
A와 D에서 많이 나왔으니
A,D에 모든 돈을 건다!

Exploration 탐색
안정적으로 가는 타입

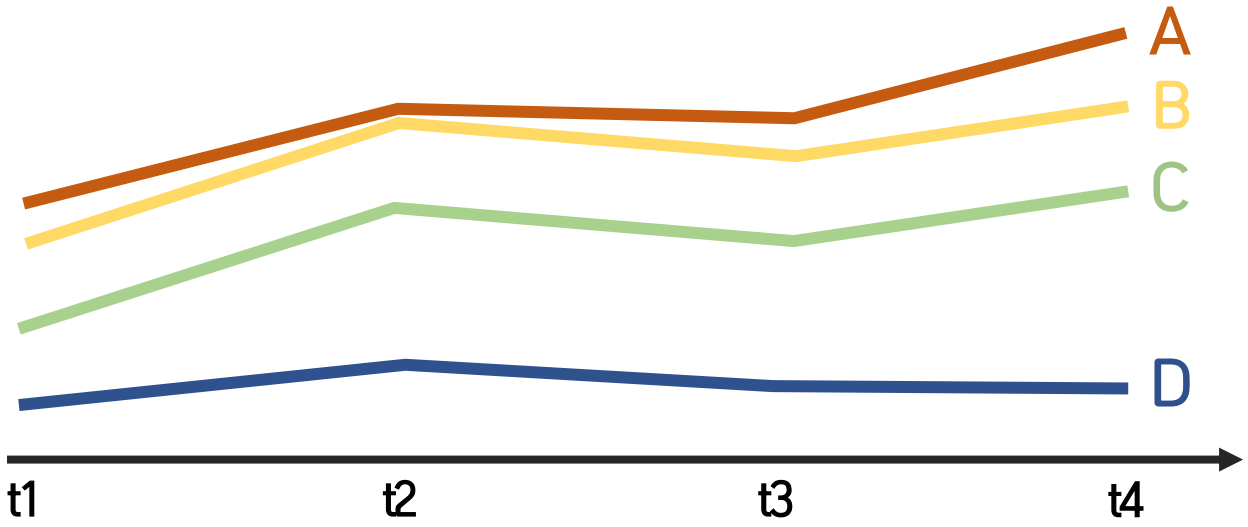
슬롯머신은 확률적인 거 아닌가?
이번에는 B,C에서
더 높은 수익이 나올 수도 있다.
무작정 A,D에 지르면 다 잃는다.



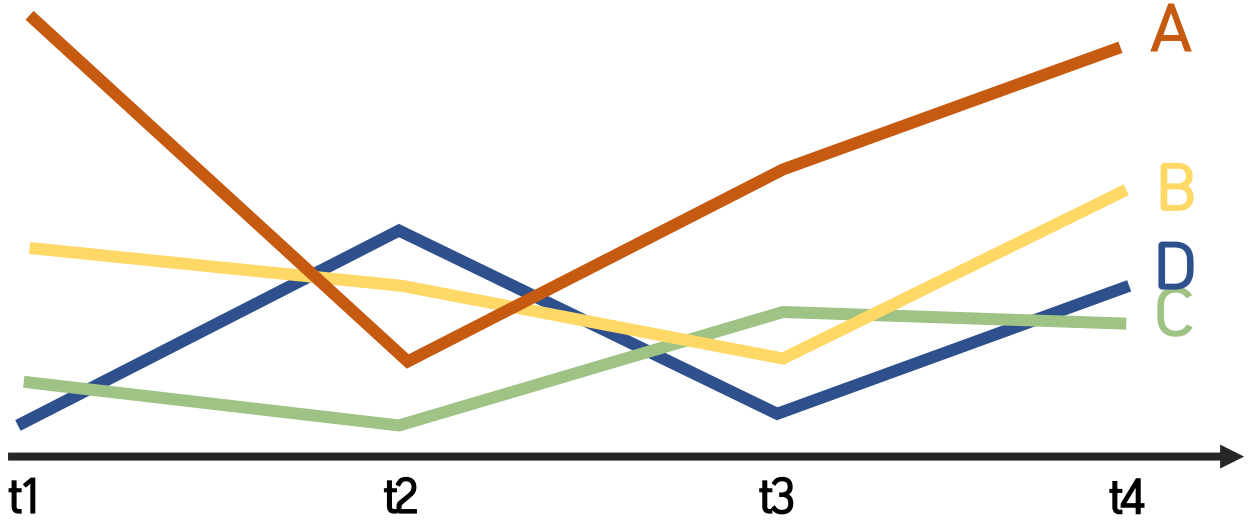
A/B 테스트를 수행하는 다양한 방법

꼭 활용과 탐색 둘 중 하나만 할 필요는 없다

활용과 탐색 중 어떤 쪽에 더 집중할지 매 라운드마다 그 확률을 조정할 수 있음



수익이 좋은 슬롯머신은 계속 좋네?
활용을 더 열심히 해야지!



수익이 슬롯머신 땡길 때마다 계속 달라지네…
일단은 활용은 좀 덜 하고 계속 탐색해 봐야겠다.

마치며

1. Ceteris Paribus를 기억하자!
   A/B그룹이 완벽히 같을 수는 없어도, 최대한 동질적으로

2. 어떠한 지표를 개선 대상으로 삼을지 처음부터 확실히!

3. 얼마나 차이가 나야 할까?
   성숙한 서비스에서 Dramatic한 차이가 나는 경우는 거의 없다
   구글에서도 1~2%의 UV 차이만 나도 굉장히 큰 차이라고 본다고 함…

4. 다양한 최신 방법들을 활용해보자