온라인 통제 실험의 모범 사례

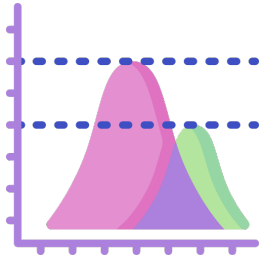
Korea Summer Workshop on Causal Inference 2024

2024-06-13

방태모

본인 소개

통계학 전공



Data Scientist

가천대 길병원 G-ABC

Data Scientist

Gmarket

글쓰기

[블로그](#)
[요즘IT](#)

Taemo Bang

Data Scientist @ Gmarket | Experimentation | Causal Inference | Time Series

온라인 통제 실험

고객이 원하는 가치를 찾기 위한 최고의 인과추론 방법론

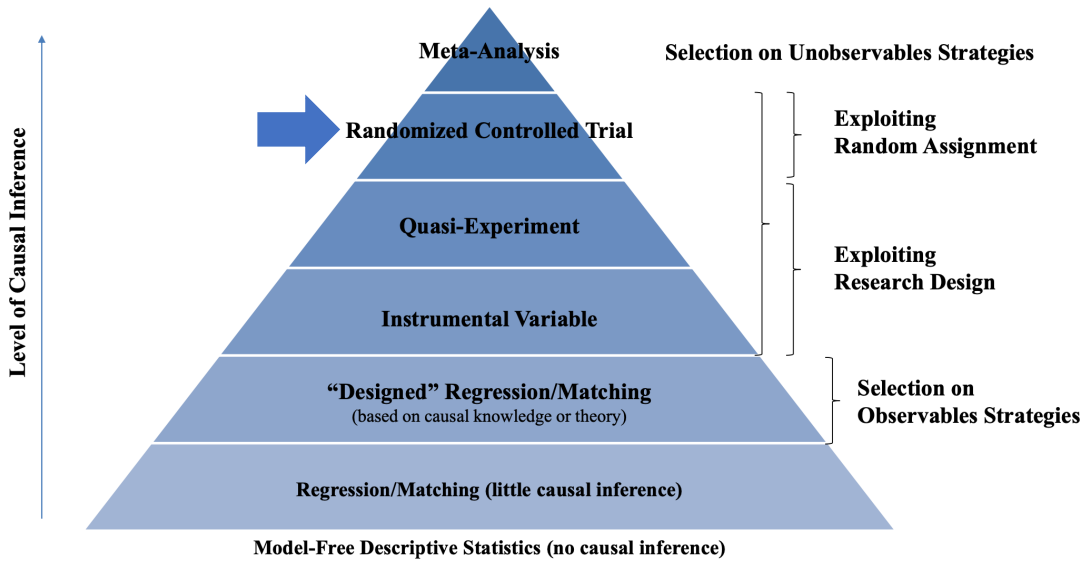
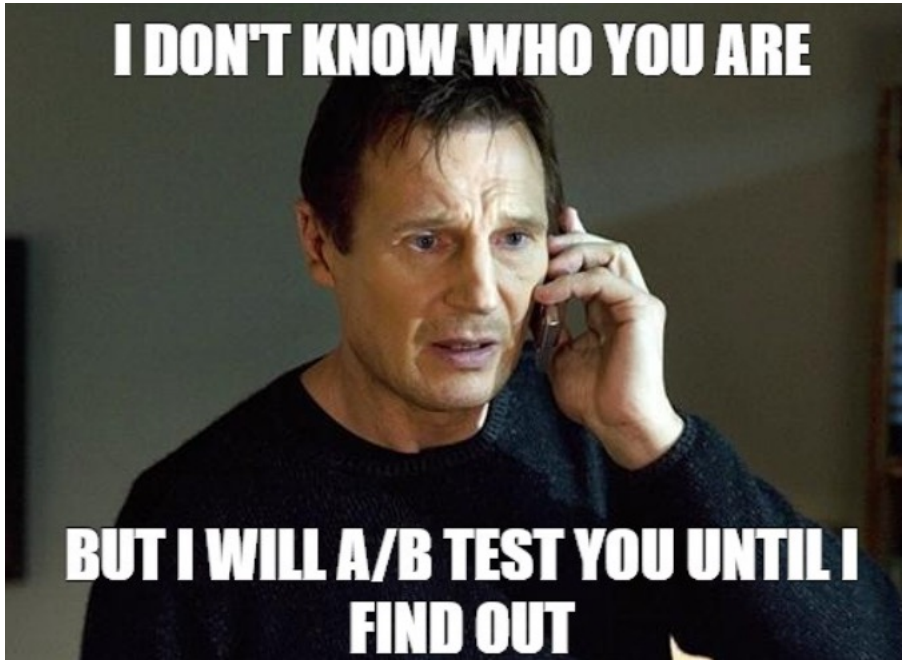


그림 1. 인과추론 방법론 피라미드 (Source: Korea Summer Workshop on Causal Inference 2022)



발표 목적

A/B 실험의 신뢰도를 보장하기 위해 라이프사이클 전반에 걸쳐 어떤 것들을 고려해야하는가?

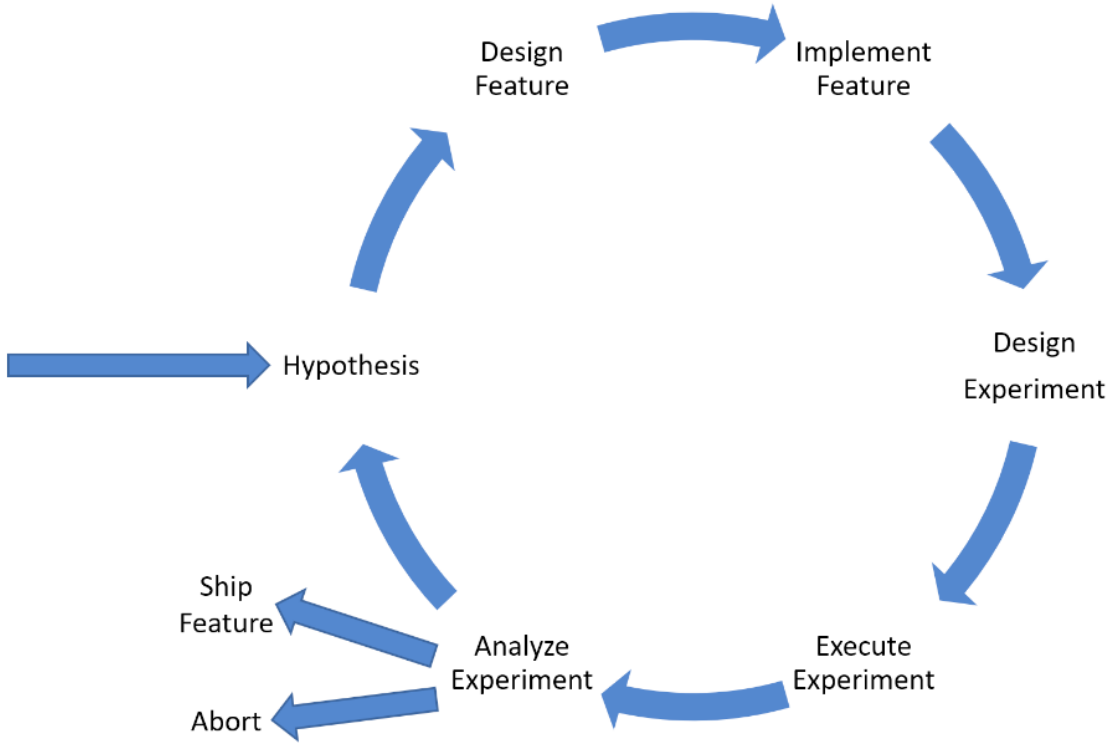


Figure 1. The experimentation lifecycle.

그림 2. 실험의 라이프사이클 (Source: Gupta et al., 2018)

A/B 실험 분석의 라이프사이클

Pre-Experiment stage - 가설 설계 및 피쳐 디자인, 실험 설계

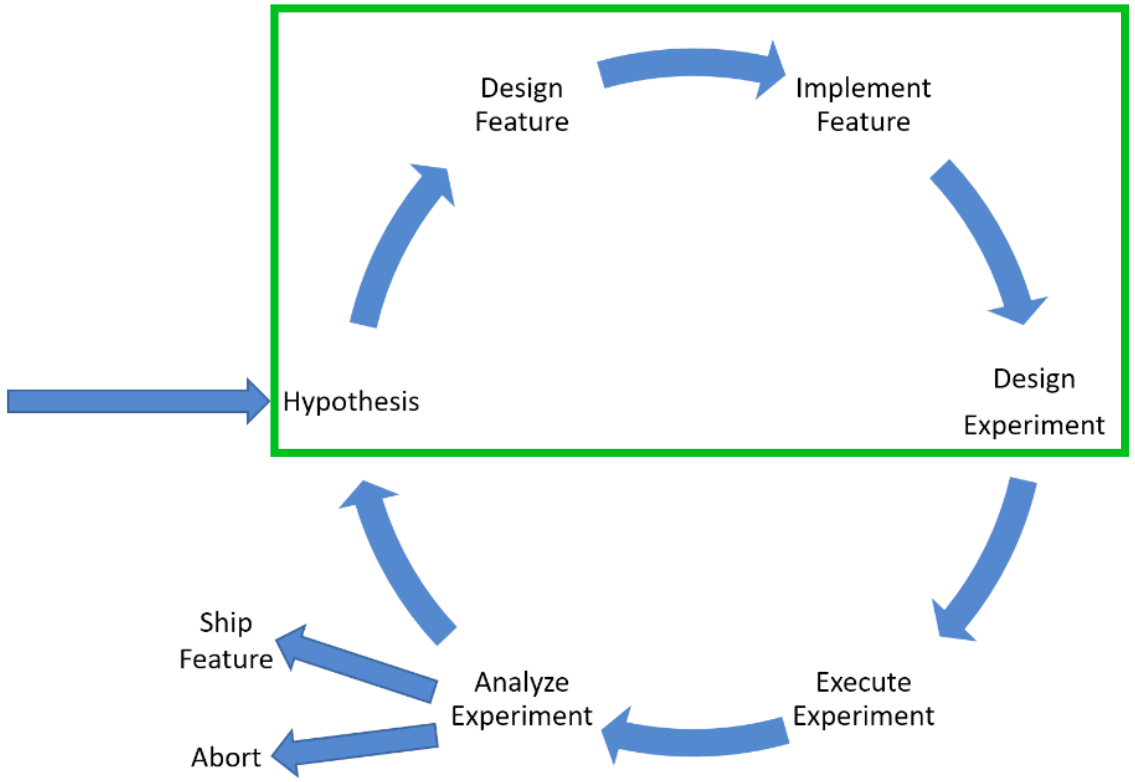


Figure 1. The experimentation lifecycle.

그림 2. 실험의 라이프사이클 (Source: Gupta et al., 2018)

A/B 실험 분석의 라이프사이클

During-Experiment stage - 모니터링

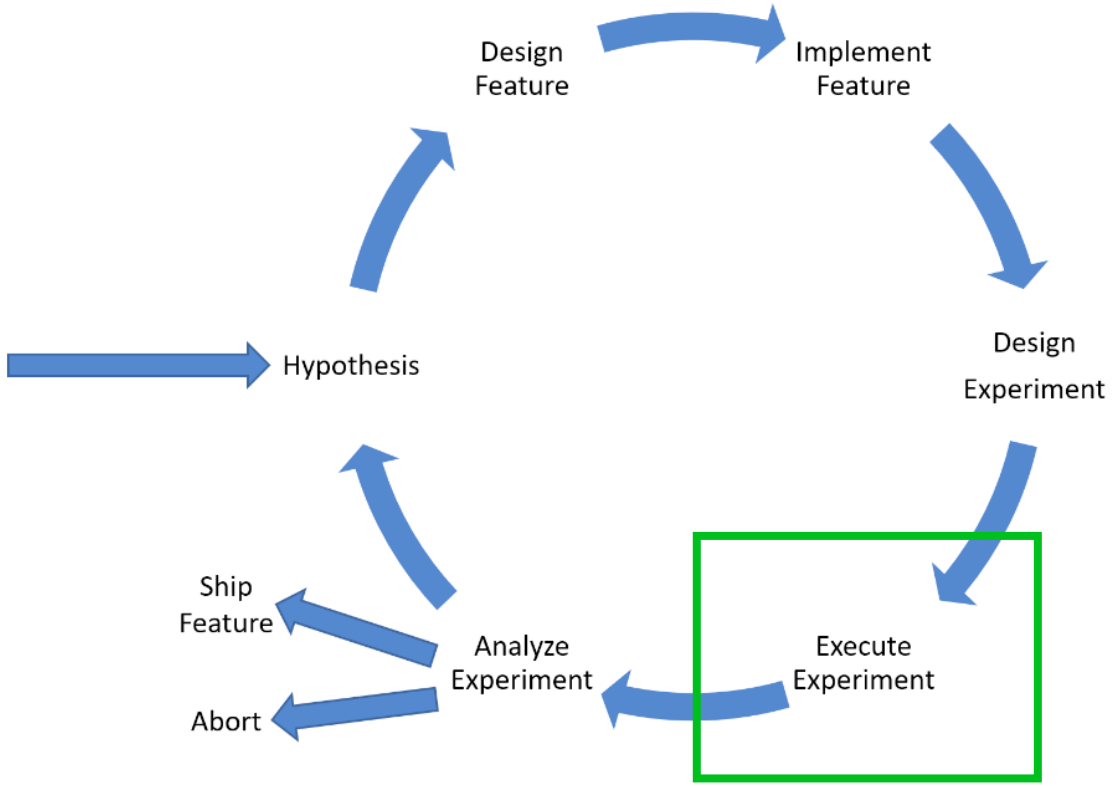


Figure 1. The experimentation lifecycle.

그림 2. 실험의 라이프사이클 (Source: Gupta et al., 2018)

A/B 실험 분석의 라이프사이클

Post-Experiment stage - 적절성 검사, 통계 분석 결과 해석, 의사 결정

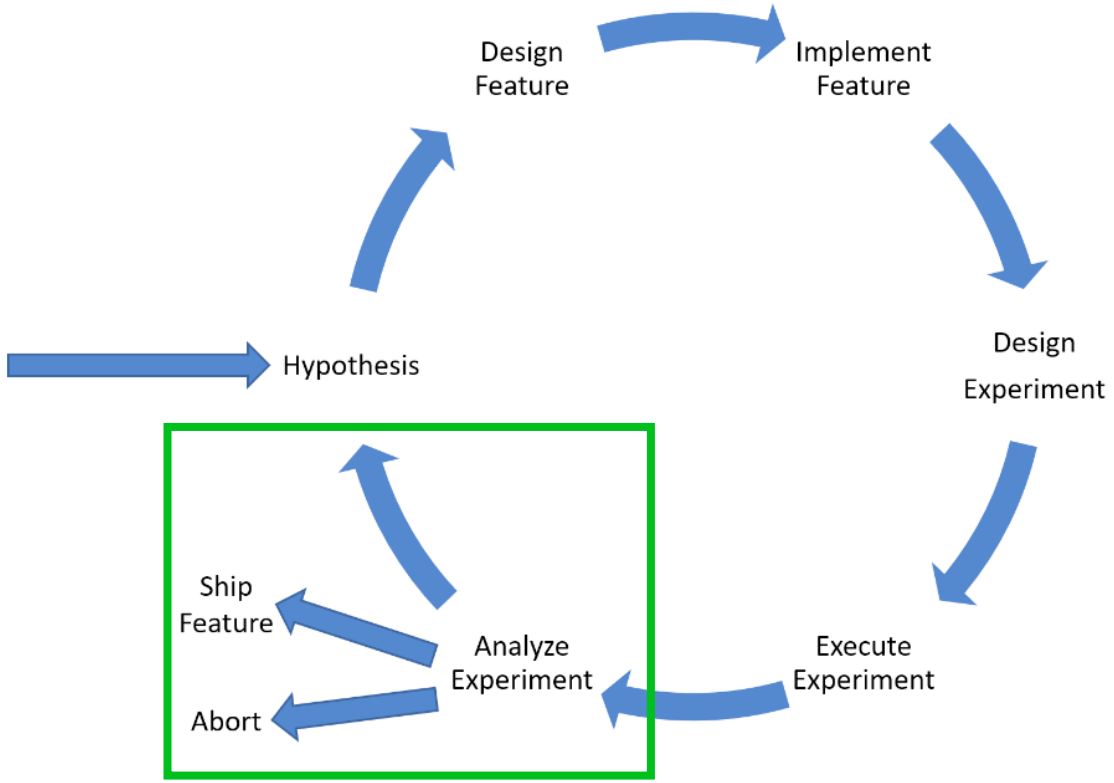


Figure 1. The experimentation lifecycle.

그림 2. 실험의 라이프사이클 (Source: Gupta et al., 2018)

신뢰도 높은 A/B 실험에 기반한 의사결정 과정을
예제를 통해 이해해봅시다.

1 Pre-Experiment stage - 가설 설계 및 피쳐 디자인

예제의 실험 배경

온라인 커머스 기업의 마케팅 부서에서 들어온 요청

“할인 쿠폰 코드가 포함된 프로모션 이메일을 보내서 판매를 늘리면 어떨까요?”

과거 연구 사례 및 동종업계 프로덕트 실험 사례 연구

- Dr.Footcare는 쿠폰 코드 필드를 추가한 뒤, 통계적으로 유의한 수준의 수익을 잃음
- GoodUI.org에서 쿠폰 코드 필드를 제거한 결과, 긍정적인 고객 반응을 보임

과거 연구 사례 및 동종업계 프로덕트 실험 사례 연구

- Dr.Footcare는 쿠폰 코드 필드를 추가한 뒤, 통계적으로 유의한 수준의 수익을 잃음
- GoodUI.org에서 쿠폰 코드 필드를 제거한 결과, 긍정적인 고객 반응을 보임

즉, 쿠폰 코드 필드를 더하는 것 자체만으로 수익이 감소될 수 있다!

고객 가치 상상해보기

사용자가 쿠폰 코드 필드를 본다.

➔ 쿠폰 코드를 검색하게 하여

➔ 구매를 망설이게 만듦

➔ 이에따라 구매 전환 속도가 느려짐

➔ 심지어 사용자를 떠나게 할 수 있음

피쳐 디자인

이에 따라 최소 비용으로 가설을 검증하기 위해 실제 쿠폰 코드 시스템은 구현하지 않고,

- **체크아웃 페이지**에 **쿠폰 코드 필드 UI**를 더하는 간단한 추가 변경만 수행

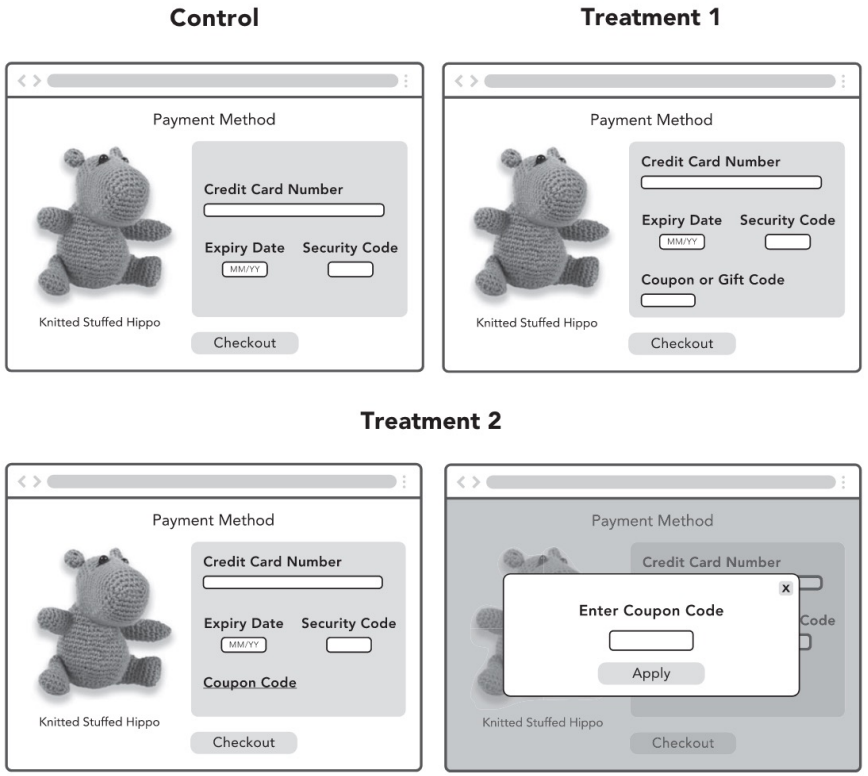


Figure 2.2 (1) Control: the old checkout page. (2) Treatment one: coupon or gift code field below credit card information (3) Treatment two: coupon or gift code as a popup

그림 3. 실험 피쳐와 변형군 상세 구성 (Source: Kohavi, Tang, and Xu 2020)

가설 초안 설계

"**쿠폰 코드 필드**를 **체크아웃 페이지**에 더하면, **구매 결정을 느리게** 만들어 **매출**을 저하할 것이다."

가설 초안 설계

"**쿠폰 코드 필드**를 **체크아웃 페이지**에 더하면, **구매 결정을 느리게** 만들어 **매출**을 저하할 것이다."

가설 초안 설계 시 고려하면 좋은 4가지 요소

- Feature
- Domain
- Value
- Primary metric
  - 표본 크기에 대해 표준화된 지표 사용 권장 -> **사용자 당 매출**

가설 구체화

사용자 당 매출 지표의 분모로 어떤 사용자들을 고려할 것인가?

가설 구체화

퍼널 관점의 고객 쇼핑 프로세스

- 사이트를 방문한 모든 사용자

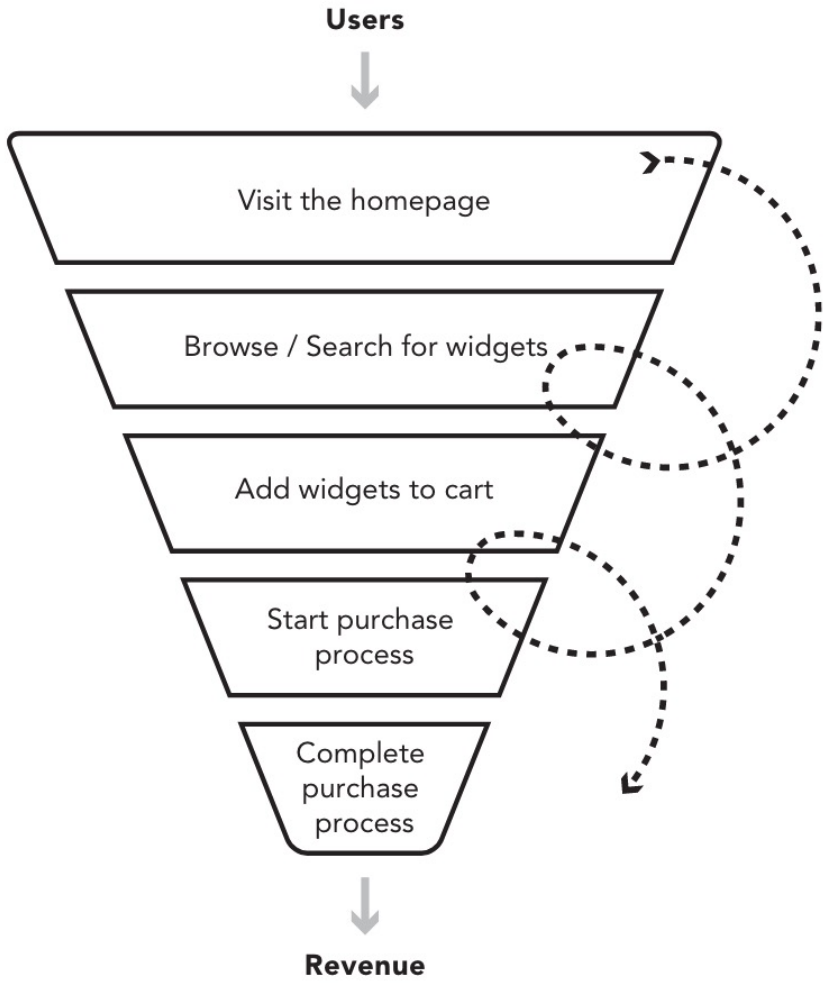


Figure 2.1 A user online shopping funnel. Users may not progress linearly through a funnel, but instead skip, repeat or go back-and-forth between steps

그림 4. 구매 완료 퍼널 (Source: Kohavi, Tang, and Xu 2020)

가설 구체화

퍼널 관점의 고객 쇼핑 프로세스

- 사이트를 방문한 모든 사용자
- 구매 프로세스를 완료한 사용자

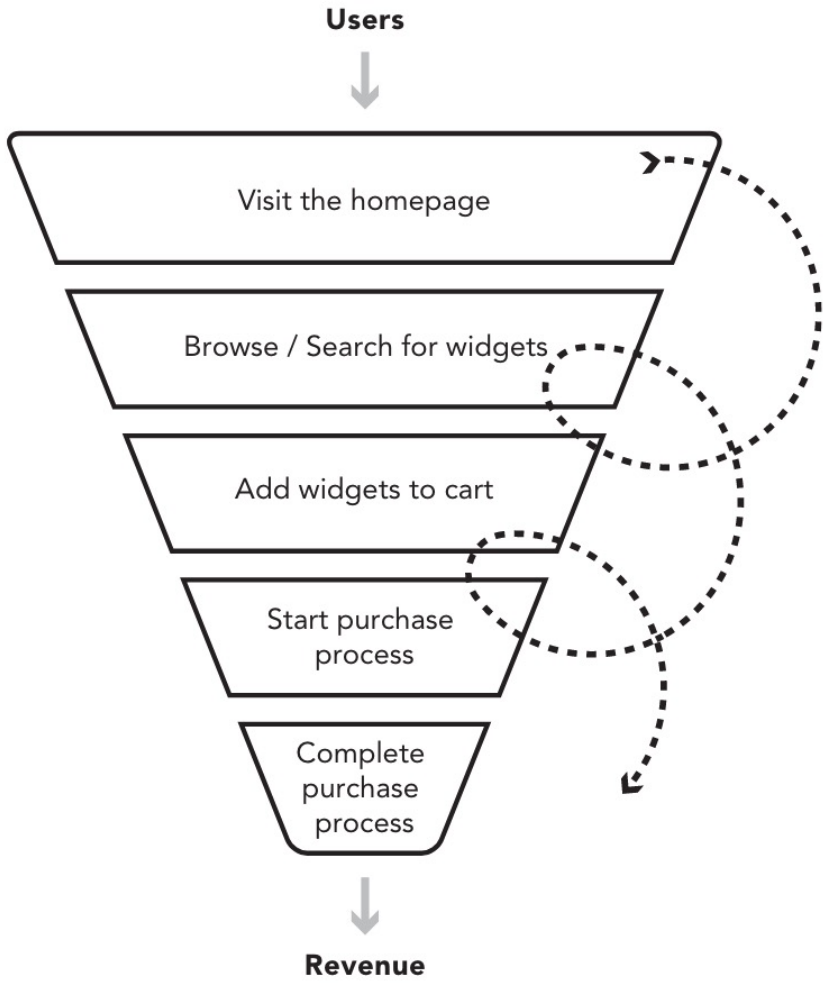


Figure 2.1 A user online shopping funnel. Users may not progress linearly through a funnel, but instead skip, repeat or go back-and-forth between steps

그림 4. 구매 완료 퍼널 (Source: Kohavi, Tang, and Xu 2020)

가설 구체화

퍼널 관점의 고객 쇼핑 프로세스

- 사이트를 방문한 모든 사용자
- 구매 프로세스를 완료한 사용자
- **구매 프로세스를 시작한 사용자 (최적의 선택)**

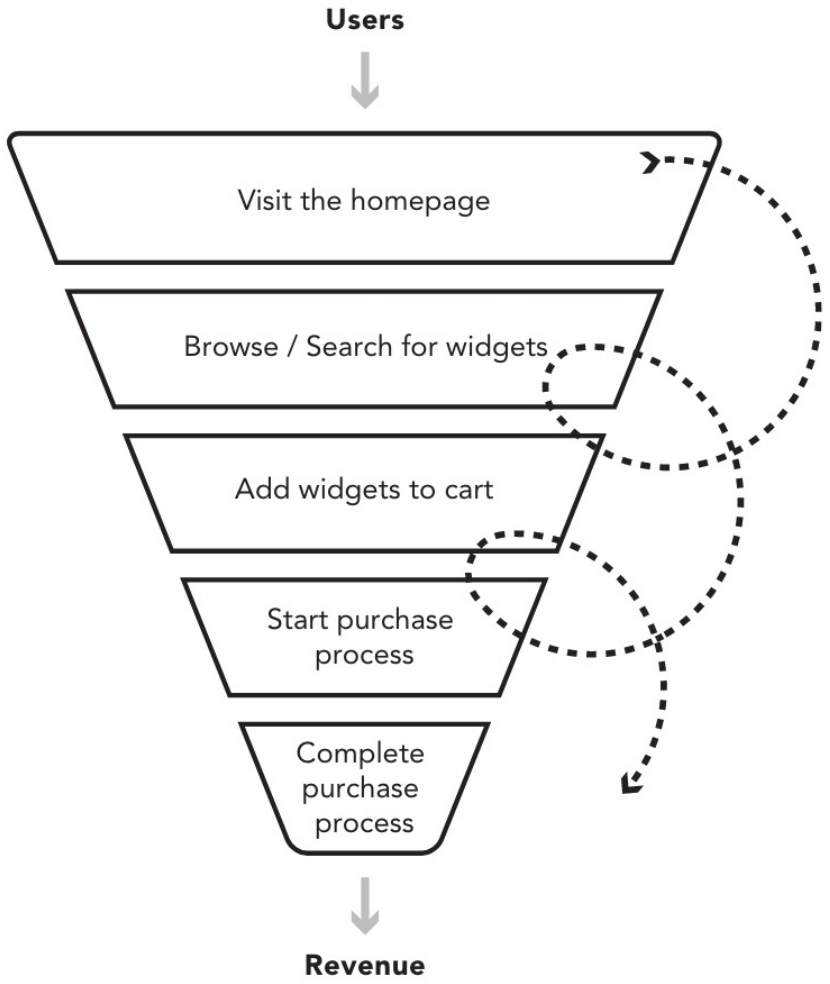


Figure 2.1 A user online shopping funnel. Users may not progress linearly through a funnel, but instead skip, repeat or go back-and-forth between steps

그림 4. 구매 완료 퍼널 (Source: Kohavi, Tang, and Xu 2020)

최종 가설

"**쿠폰 코드 필드**를 **체크아웃 페이지**에 더하면,

구매 프로세스를 시작하는 사용자들의 **구매 결정을 느리게** 만들어

사용자 당 매출이 저하한다."

좋은 가설이 갖는 5가지 요소

- Feature
- Domain
- Trigger
- Value
- Primary metric

실험 피쳐 디자인 및 정교한 가설 설계의 좋은 재료

- 과거 연구 사례 및 동종업계 프로덕트 실험 사례 연구
- 고객 가치 상상

1 Pre-Experiment stage - 실험 설계

온라인 실험 설계의 4요소

1. 무작위 추출 단위

일반적으로 **사용자**

2. 무작위 추출 단위의 모집단 대상 (타겟팅)

지리적 지역, **플랫폼 및 장치 유형**, **앱 버전** 등

3. 실험 할당량 (Allocation)
    - 대규모 변경의 경우 적은 비율의 사용자로 실험 시작할 것

4. 실험 기간

온라인 실험 설계의 4요소

1. 무작위 추출 단위

일반적으로 **사용자**

2. 무작위 추출 단위의 모집단 대상 (타겟팅)

지리적 지역, **플랫폼 및 장치 유형**, **앱 버전** 등

실험의 신뢰도 보장을 위해 검정력 분석으로 결정해야함

3. 실험 할당량 (Allocation)
    - 대규모 변경의 경우 적은 비율의 사용자로 실험 시작할 것
4. 실험 기간

실험 기간과 할당량을 결정하는 방법

검정력 분석 (Power Analysis)

- **특정 수준의 검정력과 유의수준**을 만족시키기 위해 얼마나 많은 표본을 수집해야하는가?

실험 기간과 할당량을 결정하는 방법

검정력 분석 (Power Analysis)

- **특정 수준의 검정력과 유의수준**을 만족시키기 위해 얼마나 많은 표본을 수집해야하는가?

검정력(Power, $1 - \beta$)

- 귀무가설(H_0)을 올바르게 기각시킬 확률
- 본 예제의 귀무가설
  - 쿠폰 필드를 체크아웃 페이지에 더해도, 변형군 간 사용자 당 매출에는 차이가 없다.

산업 표준의 검정력 분석

$$n = \frac{16\sigma^2}{\delta^2}$$

위 식의 조건

- 산업 표준 = 검정력$(1-\beta)$ 80%, 유의수준(α) 5%
- 양측 검정
- 변형군 균등 비율

산업 표준의 검정력 분석

$$n = \frac{16\sigma^2}{\delta^2}$$

여기서 포인트는

- **검출해내고 싶은 최소 효과(MDE, δ)**가 얼마인지에 따라 필요한 표본 크기는 달라진다는 것
- **수행할 실험의 대조군의 관심 지표의 분산(σ^2)**에 따라 필요한 표본 크기는 달라진다는 것

생각해보면 당연한 수식

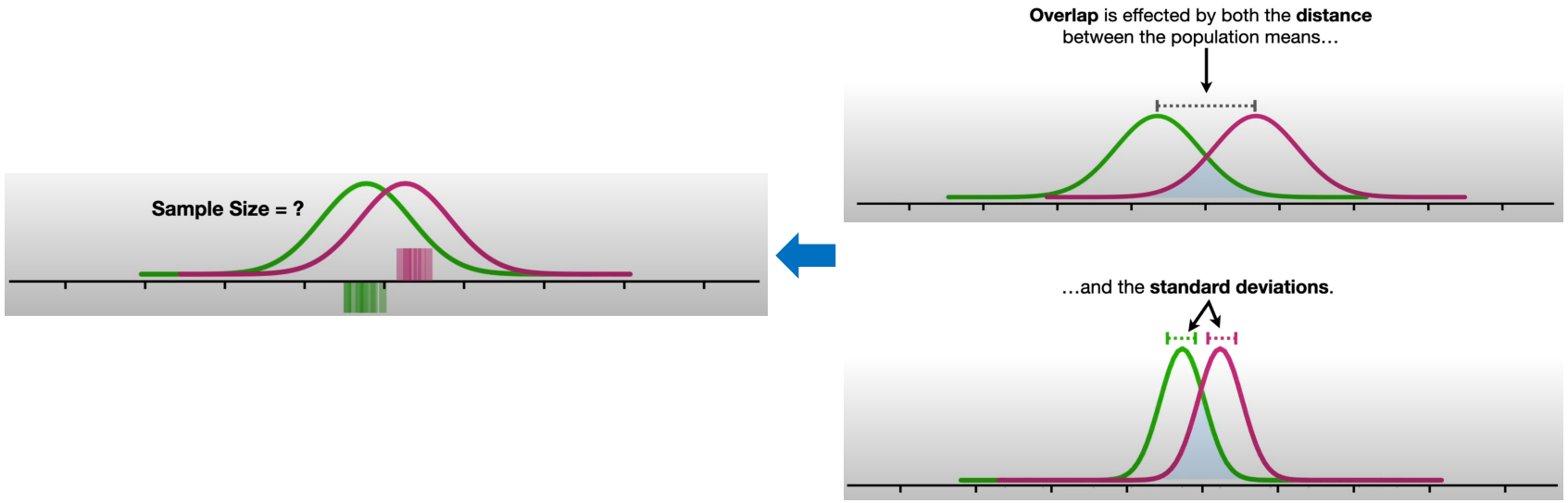


그림 5. 검정력 분석 수식의 맥락 (Source: Statquest, Power Analysis, Clearly Explained!)

생각해보면 당연한 수식

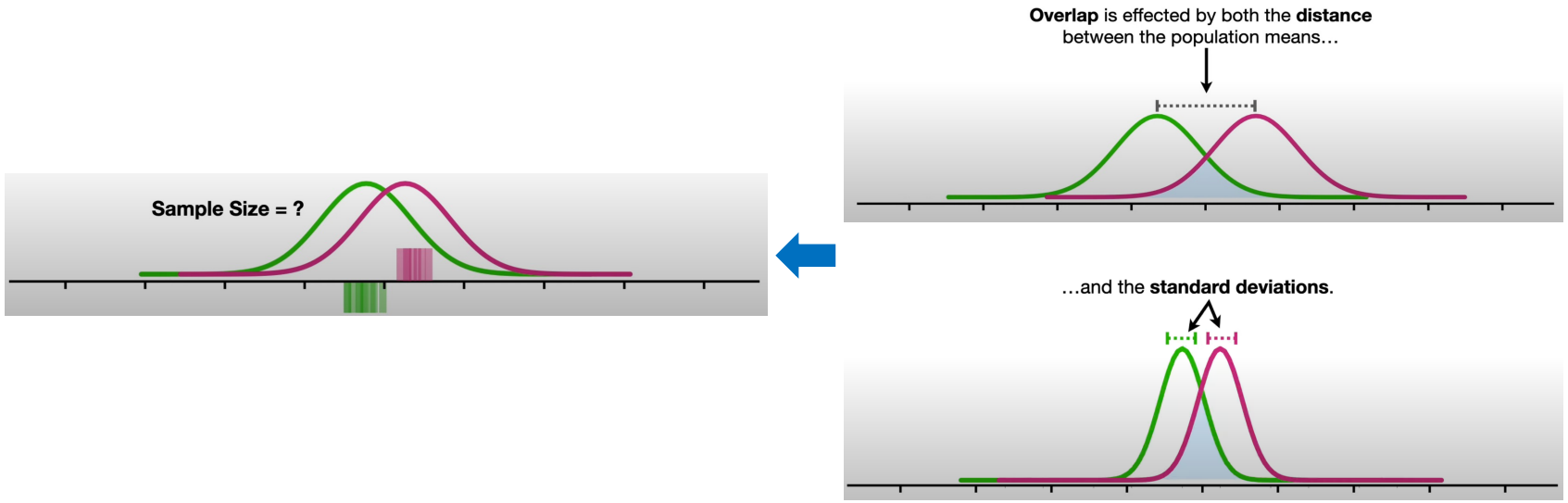


그림 5. 검정력 분석 수식의 맥락 (Source: Statquest, Power Analysis, Clearly Explained!)

온라인 실험에서 표본 크기는 실험 할당량과 실험 기간에 의존함

온라인 실험 설계의 4요소

1. 무작위 추출 단위

일반적으로 **사용자**

2. 무작위 추출 단위의 모집단 대상 (타겟팅)

지리적 지역, **플랫폼 및 장치 유형**, **앱 버전** 등

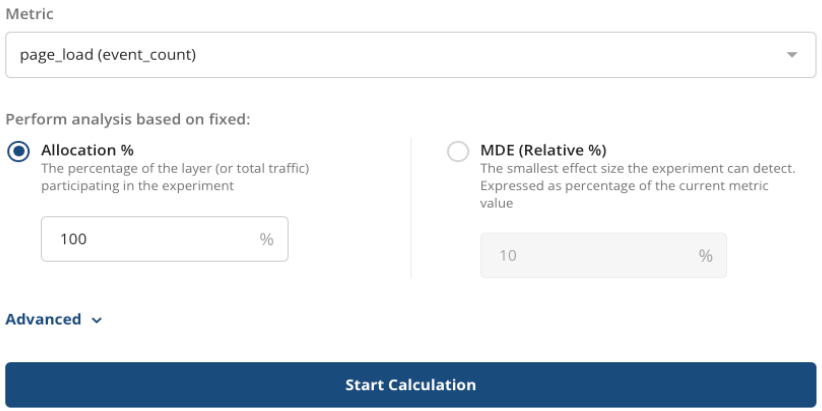
실험의 신뢰도 보장을 위해 검정력 분석으로 결정해야함

3. 실험 할당량 (Allocation)
    - 대규모 변경의 경우 적은 비율의 사용자로 실험 시작할 것
4. 실험 기간

실제 실험 플랫폼 제공 방식

실험 할당량 입력

- 실험 기간에 따른 Relative MDE 계산



Power Analysis Calculator

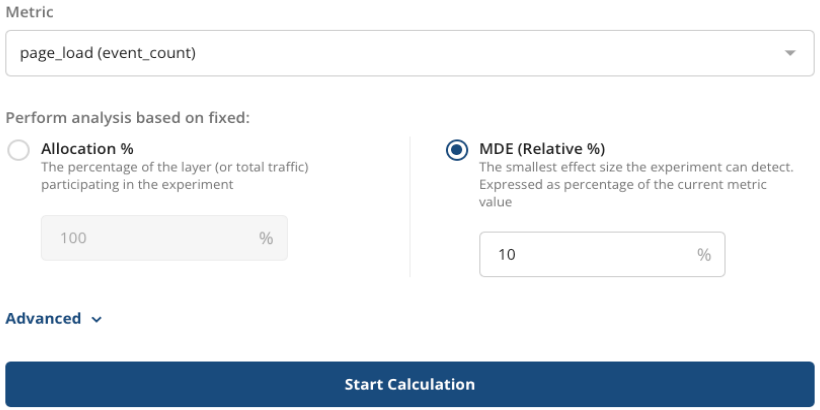
Estimates based on metric statistics calculated across all users in the project

Number of Weeks	MDE (Relative %)	Control Group Units	Test Group Units
1	21.6%	5200	5200
2	8.07%	37200	37200
3	7.49%	43200	43200
4	7.12%	47800	47800

Source: Stagsig Docs, Power Analysis

Relative MDE 입력

- 실험 기간에 따른 필요 트래픽 계산



Power Analysis Calculator

Estimates based on metric statistics calculated across all users in the project

Number of Weeks	Allocation	Control Group Units	Test Group Units
1	>100%	5190	5190
2	65.1%	24200	24200
3	56.1%	24200	24200
4	50.7%	24200	24200

Source: Stagsig Docs, Power Analysis

실험 기간 설정 시 고려해야 할 4가지 사항

- **검정력을 늘리기 위해 실험 기간을 길게 잡는 것은 한계가 있음 (∴최대 4주 허용)**
  - 실험 기간이 길어져 기존에 방문한 사용자들만 재방문을 할 경우 사용자 누적률이 저선형(sub-linear)이 됨
  - 이에 따라 사용자 당 세션 수와 분산이 증가 -> 실험 민감도 감소

실험 기간 설정 시 고려해야 할 4가지 사항

- **검정력을 늘리기 위해 실험 기간을 길게 잡는 것은 한계가 있음 (∴최대 4주 허용)**
  - 실험 기간이 길어져 기존에 방문한 사용자들만 재방문을 할 경우 사용자 누적률이 저선형(sub-linear)이 됨
  - 이에 따라 사용자 당 세션 수와 분산이 증가 -> 실험 민감도 감소
- **주간 효과: 평일과 주말의 사용자 분포는 다름 (∴최소 1주 권장)**

실험 기간 설정 시 고려해야 할 4가지 사항

- **검정력을 늘리기 위해 실험 기간을 길게 잡는 것은 한계가 있음 (∴최대 4주 허용)**
  - 실험 기간이 길어져 기존에 방문한 사용자들만 재방문을 할 경우 사용자 누적률이 저선형(sub-linear)이 됨
  - 이에 따라 사용자 당 세션 수와 분산이 증가 -> 실험 민감도 감소
- **주간 효과: 평일과 주말의 사용자 분포는 다름 (∴최소 1주 권장)**
- **계절성: 외적 타당성(external validity)과 관련 (∴ 반복 실험 권장)**
  - 크리스마스 시즌과 나머지 일반적 시즌의 기프트 카드 판매량은 다를 수 있음

실험 기간 설정 시 고려해야 할 4가지 사항

- **검정력을 늘리기 위해 실험 기간을 길게 잡는 것은 한계가 있음 (∴최대 4주 허용)**
  - 실험 기간이 길어져 기존에 방문한 사용자들만 재방문을 할 경우 사용자 누적률이 저선형(sub-linear)이 됨
  - 이에 따라 사용자 당 세션 수와 분산이 증가 -> 실험 민감도 감소
- **주간 효과: 평일과 주말의 사용자 분포는 다름 (∴최소 1주 권장)**
- **계절성: 외적 타당성(external validity)과 관련 (∴ 반복 실험 권장)**
  - 크리스마스 시즌과 나머지 일반적 시즌의 기프트 카드 판매량은 다를 수 있음
- **초두 효과와 신기 효과 (∴최소 1주 권장)**
  - 초두 효과(primacy effect): 실험 초기 효과가 정상보다 작은 것
  - 신기 효과(novelty effect): 실험 초기 효과가 정상보다 큰 것

온라인 실험 설계의 4요소 - 요약

- 무작위 추출 단위
  - **일반적으로 사용자**
- 무작위 추출 단위의 모집단 대상
  - **타겟팅** (지리적 지역, 플랫폼 및 장치 유형 등)
- 어느 정도 규모의 실험이 필요한가?
  - **실험 할당량**
- 실험을 얼마나 오래 진행할 것인가?
  - **실험 기간**

온라인 실험 설계의 4요소 - 요약

- 무작위 추출 단위
  - **일반적으로 사용자**
- 무작위 추출 단위의 모집단 대상
  - **타겟팅** (지리적 지역, 플랫폼 및 장치 유형 등)
- 어느 정도 규모의 실험이 필요한가?
  - **실험 할당량**
- 실험을 얼마나 오래 진행할 것인가?
  - **실험 기간**

실험의 신뢰도 보장을 위해 검정력 분석으로 결정해야함

본 예제 실험의 설정 요약 - 최종 가설, 피쳐 디자인, Primary metric

쿠폰 코드 필드를 **체크아웃 페이지**에 더하면, **구매 프로세스를 시작하는 사용자들**의 **구매 결정을 느리게** 만들어 **사용자 당 매출**이 저하한다.

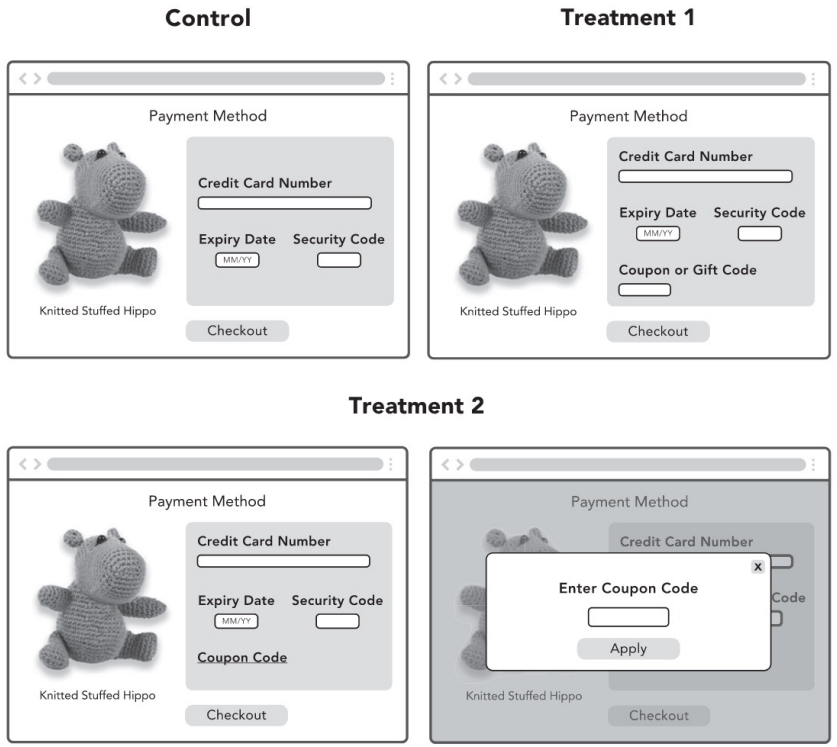


Figure 2.2 (1) Control: the old checkout page. (2) Treatment one: coupon or gift code field below credit card information (3) Treatment two: coupon or gift code as a popup

그림 3. 실험 피쳐와 변형군 상세 구성 (Source: Kohavi, Tang, and Xu 2020)

본 예제 실험의 설정 요약 - 실험 설계

- 무작위 추출 단위: **사용자**
- 타겟팅: **모든 사용자** (단, Trigger는 체크아웃 페이지를 방문하는 사용자)
- 검정력 분석에 의해 결정된 실험 할당량과 실험 기간
  - 사용자 당 수입의 **1% 이상 상대적 변화를 감지**하고자 함 (Relative MDE = 1%)
  - 할당량: **100%** (최대 검정력을 위해 변형군 균등 분할)
  - 실험 기간: 4일이면 충분하지만, **주간효과를 고려해 7일**간 실험 진행

1 Pre-Experiment stage - 실험 전 편향 교정

실험 전 편향 (Pre-experiment bias)

실험 시작 전 실험군과 대조군에 존재하는 편향

실험 시작 전에는 두 군 간 관심 지표에 차이가 존재하는 경우

➔ 순수한 실험 효과 감지가 어려움

실험 전 편향 발생의 3가지 원인

Carryover effect

- 이전 실험에 실험군으로 참여한 효과가 남아있는 것

Ordering effect

- 실험 1에 참가했냐, 실험 2에 참가했냐에 따라 실험 3에서의 효과가 달라지는 것

Random imbalance

- 고객들을 각 군에 랜덤하게 할당하더라도 우연히 편향이 발생하는 경우

실험 전 편향 극복할 수 있는 2가지 방법 - I. Seedfinder

1. Randomization과 Retrospective A/A Tests 활용 (Seedfinder)

적절한 실험군, 대조군 구성 찾기

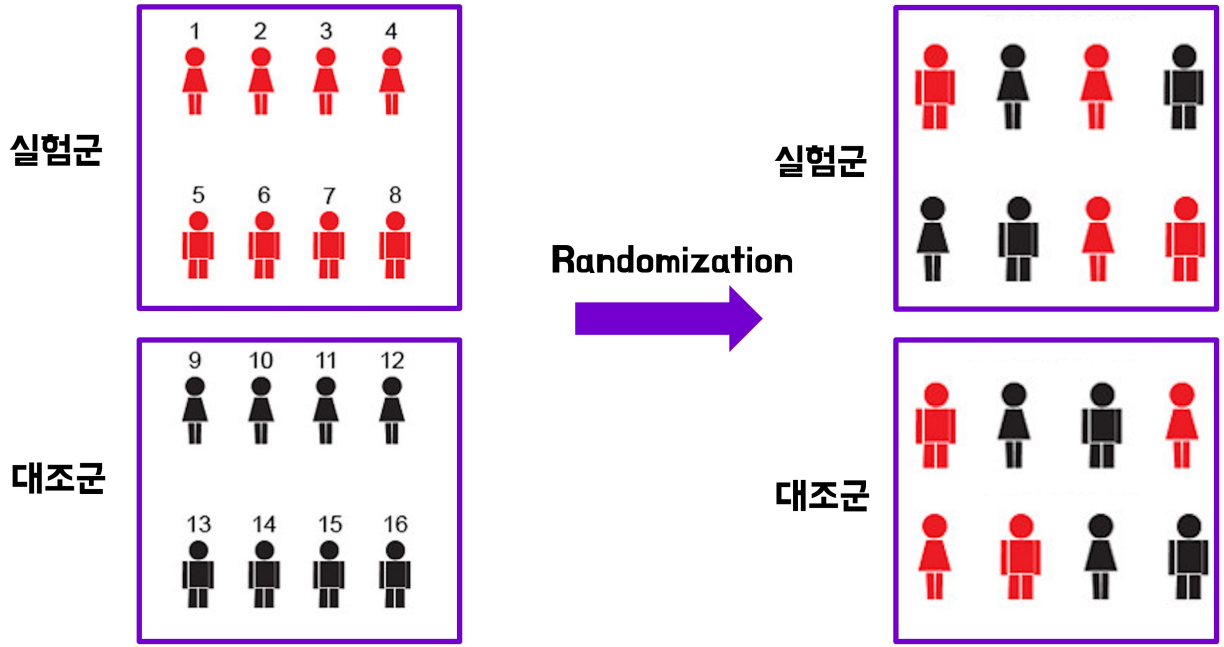


그림 5. Randomization (Source: Burger, Vaudel, and Barsnes 2021)

온라인 통제 실험의 Randomization 방식

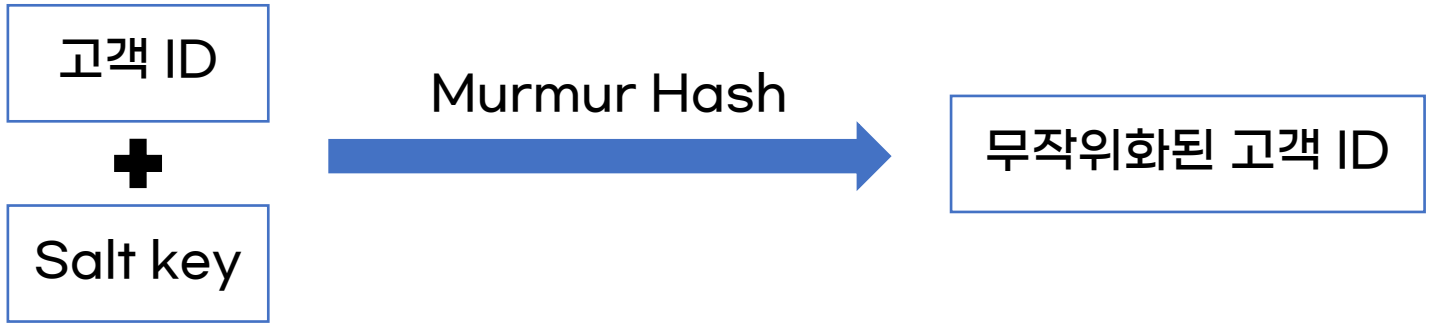


그림 6. 해시 알고리즘을 통한 고객 ID 무작위화

실험 전 편향 극복할 수 있는 2가지 방법 - I. Seedfinder

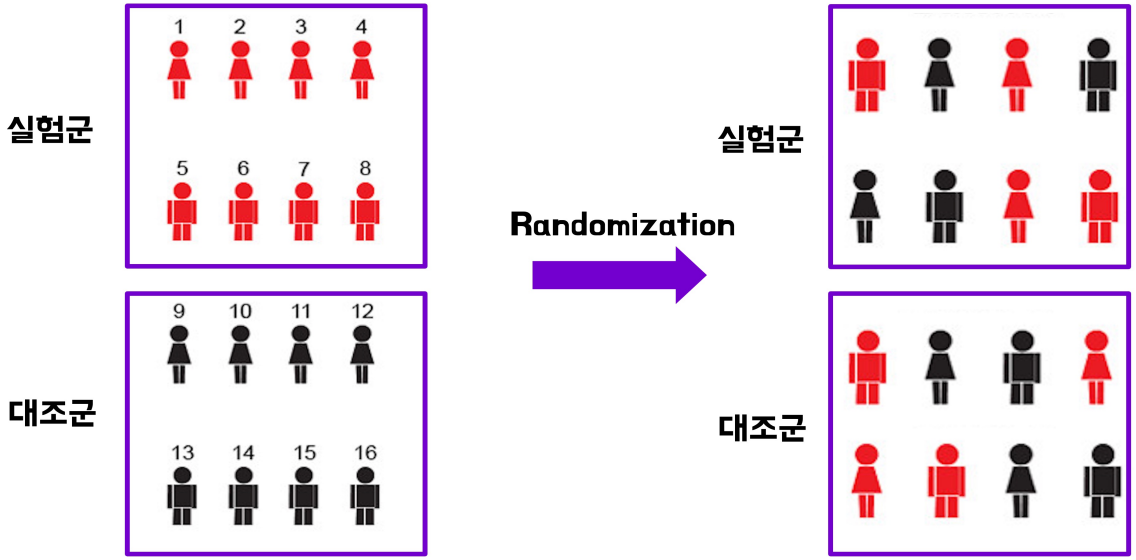


그림 5. Randomization (Source: Burger, Vaudel, and Barsnes 2021)

Salt key만 랜덤하게 바꿔주면 무한정 섞을 수 있음

실험 전 편향 극복할 수 있는 2가지 방법 - I. Seedfinder

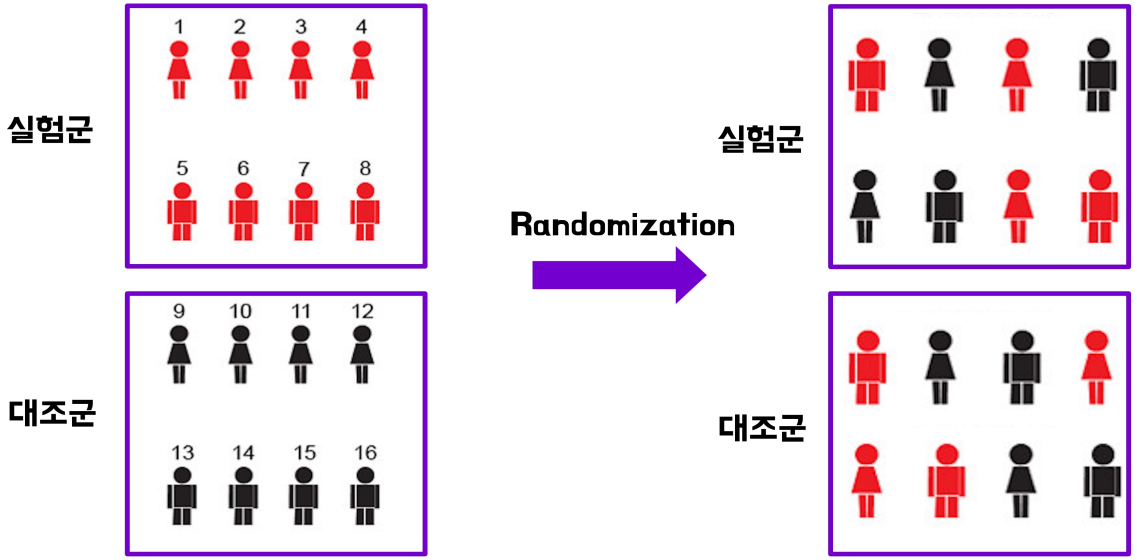


그림 5. Randomization (Source: Burger, Vaudel, and Barsnes 2021)

Salt key만 랜덤하게 바꿔주면 무한정 섞을 수 있음

- Retrospective A/A Test: 최근 1주 내지 2주 데이터로 두 군 간 관심지표 차이를 확인하여 통계적 검증
- Seedfinder: Retrospective A/A Test를 수차례 반복하여 가장 Bias가 적은 Salt key (Seed) 반환

실험 전 편향 극복할 수 있는 2가지 방법 - II. CUPED

CUPED (Controlled-experiment Using Pre-Experiment Data)

- 실험 전 데이터를 활용하여 실험 전 편향을 교정하여 실험 민감도를 올리는 방법론

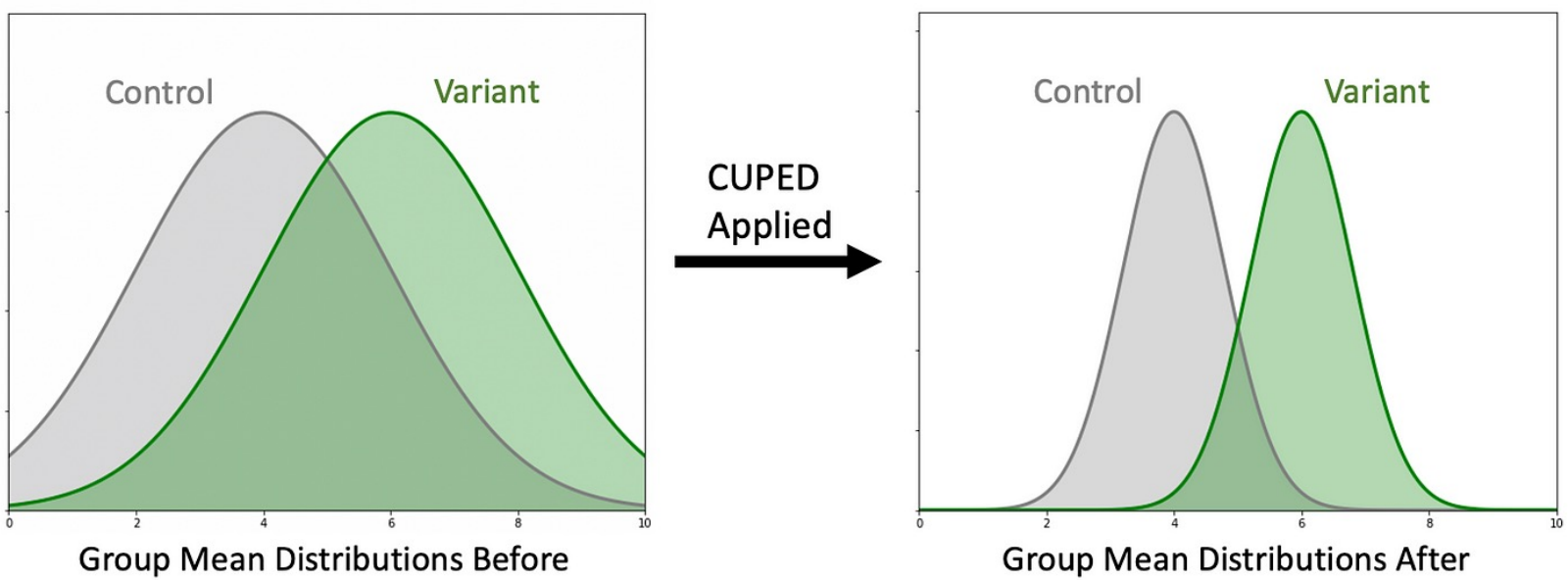


그림 7. CUPED에 의해 줄어든 관심 지표의 분산 (Source: Statsig Blog, CUPED on Statsig)

실험 전 편향 극복할 수 있는 2가지 방법 - II. CUPED

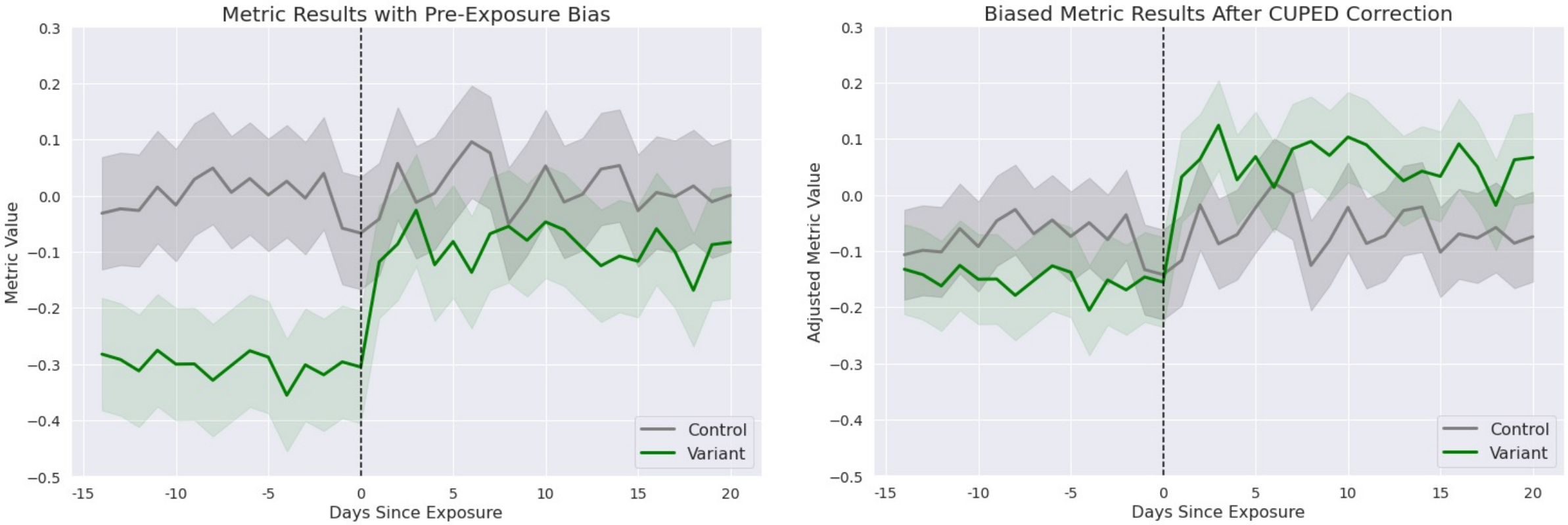


그림 8. CUPED에 의해 편향이 교정되며 실험 결과가 바뀌는 경우 (Source: Statsig Blog, CUPED on Statsig)

실제 실험 플랫폼 제공 방식

1. Seedfinder

- 구체적 제공 방법 미공개
- 각 기업 실험 인프라에 맞춰서 풀자

2. CUPED

- Primary Metrics에 대해 실험 결과 리포트에서 자동으로 적용되도록 구현

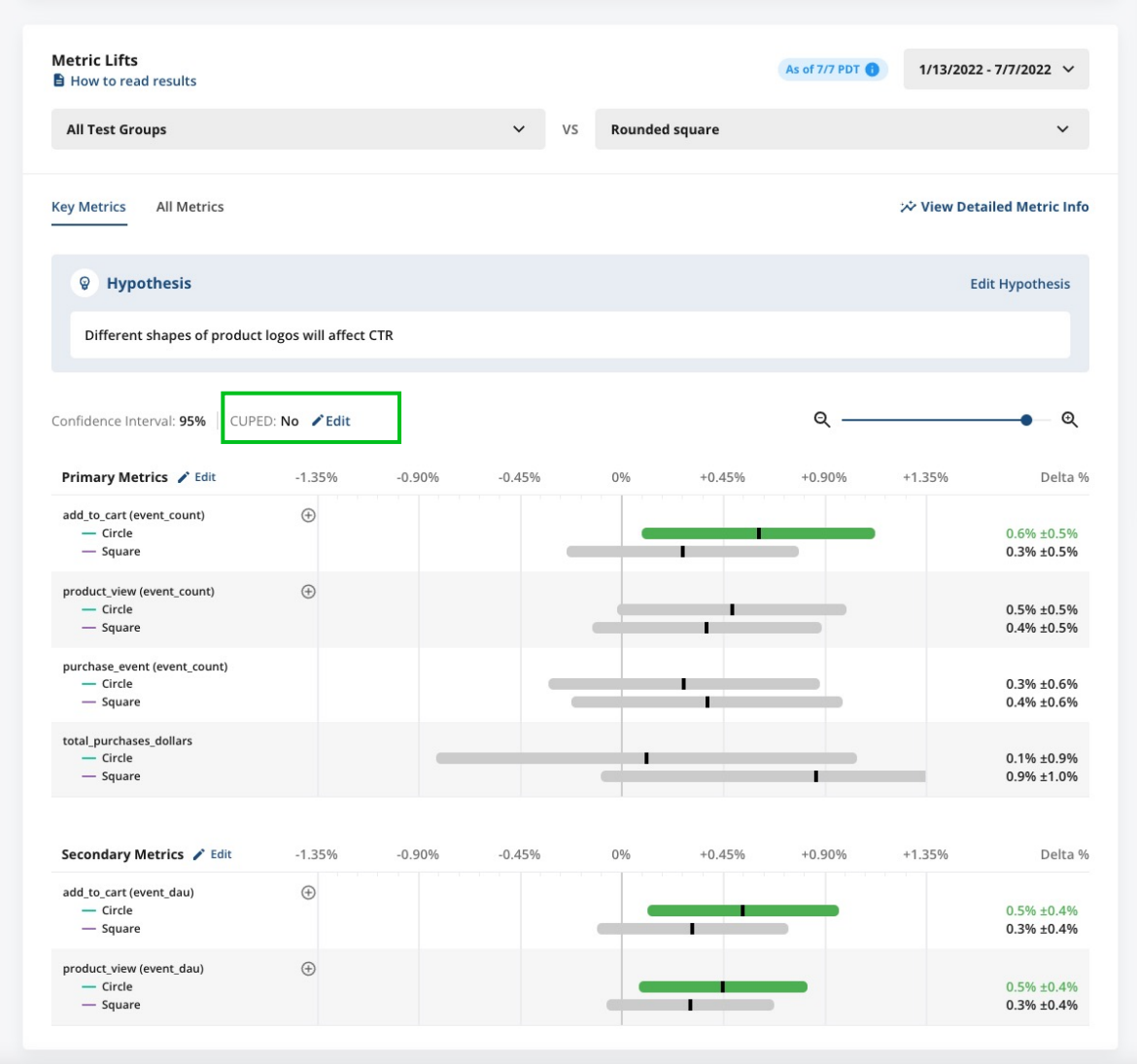


그림 9. CUPED 제공 방식 (Source: Stagsig Docs, Read Results)

2 During-Experiment stage

모니터링

예상치 못한 효과의 조기 감지를 위해 필요한 실험 지표

유형	설명	예
데이터 품질 지표	나머지 지표들을 신뢰할 수 있는지에 대해 결정하는 지표	SRM (Sample Ratio Mismatch)
Local Features 및 진단 지표	• 세부 피쳐 각각의 성과를 측정하여 Primary metric의 움직임을 해석하는데 도움을 주는 지표 • 실험이 잘 진행되고 있는지 진단할 수 있는 지표	세부 피쳐 별 CTR
가드레일 지표	반드시 개선할 필요는 없지만, 성과 저하를 원하지 않는 지표	페이지 로드 시간

표 1. 모니터링 지표 (Source: Microsoft, 2021)

3 Post-Experiment stage

적절성 검사

결과 해석 전 가장 먼저 **적절성 검사(sanity checks)** 실시

데이터 품질 지표(신뢰 관련 가드레일 지표) 검토

- SRM(Sample ratio mismatch): 변형군 할당 비율이 실험 설정을 잘 따르는가?

조직 관련 가드레일 지표 검토

- 페이지 로드 시간: 새로운 피쳐에서 속도 저하가 있는가?

적절성 검사

SRM이 발생한 경우

- **기초적 실험 설계, 인프라 또는 데이터 처리에 문제가 있다**는 뜻
- 즉, **실험 결과를 신뢰할 수 없음**
- 마이크로소프트의 사내 실험플랫폼은 **SRM이 발생 시, 실험 결과 블라인드 처리**
  - Gupta et al., 2018 p.10 4) Verifying data quality 참조

통계 분석 결과 해석

적절성 검사 통과 시 결과 해석 시작

Table 2.1 *Results on revenue-per-user from the checkout experiment.*

	Revenue-per-user, Treatment	Revenue-per-user, Control	Difference	p-value	Confidence Interval
Treatment One vs. Control	$3.12	$3.21	–$0.09 (–2.8%)	0.0003	[–4.3%, –1.3%]
Treatment Two vs. Control	$2.96	$3.21	–$0.25 (–7.8%)	1.5e-23	[–9.3%, –6.3%]

표 2. 예제 실험의 2-표본 t검정 결과 (Source: Kohavi, Tang, and Xu 2020)

통계 분석 결과 해석

적절성 검사 통과 시 결과 해석 시작

Table 2.1 *Results on revenue-per-user from the checkout experiment.*

	Revenue-per-user, Treatment	Revenue-per-user, Control	Difference	p-value	Confidence Interval
Treatment One vs. Control	$3.12	$3.21	–$0.09 (–2.8%)	0.0003	[–4.3%, –1.3%]
Treatment Two vs. Control	$2.96	$3.21	–$0.25 (–7.8%)	1.5e-23	[–9.3%, –6.3%]

표 2. 예제 실험의 2-표본 t검정 결과 (Source: Kohavi, Tang, and Xu 2020)

- **쿠폰 코드를 추가하면 수익이 감소한다**는 패턴을 확인
  - 구매 과정을 완료하는 사용자 수가 더 적어져 수익이 감소

의사결정

Table 2.1 *Results on revenue-per-user from the checkout experiment.*

	Revenue-per-user, Treatment	Revenue-per-user, Control	Difference	p-value	Confidence Interval
Treatment One vs. Control	$3.12	$3.21	–$0.09 (–2.8%)	0.0003	[–4.3%, –1.3%]
Treatment Two vs. Control	$2.96	$3.21	–$0.25 (–7.8%)	1.5e-23	[–9.3%, –6.3%]

표 2. 예제 실험의 2-표본 t검정 결과 (Source: Kohavi, Tang, and Xu 2020)

- 쿠폰 코드 발송 마케팅 이메일이 회수해야 하는 2가지 비용
  - **쿠폰 코드를 처음 추가할 때 발생하는 부정적 영향의 비용**
  - **쿠폰 처리 시스템 구현 비용** 및 **유지보수 비용**
- **(최종 결정) 프로모션 코드를 도입 아이디어 폐기**
- **출시 전 최소 테스트**로 A/B 테스트를 수행함으로써 많은 노력 절감!

실제 실험 플랫폼 제공 방식

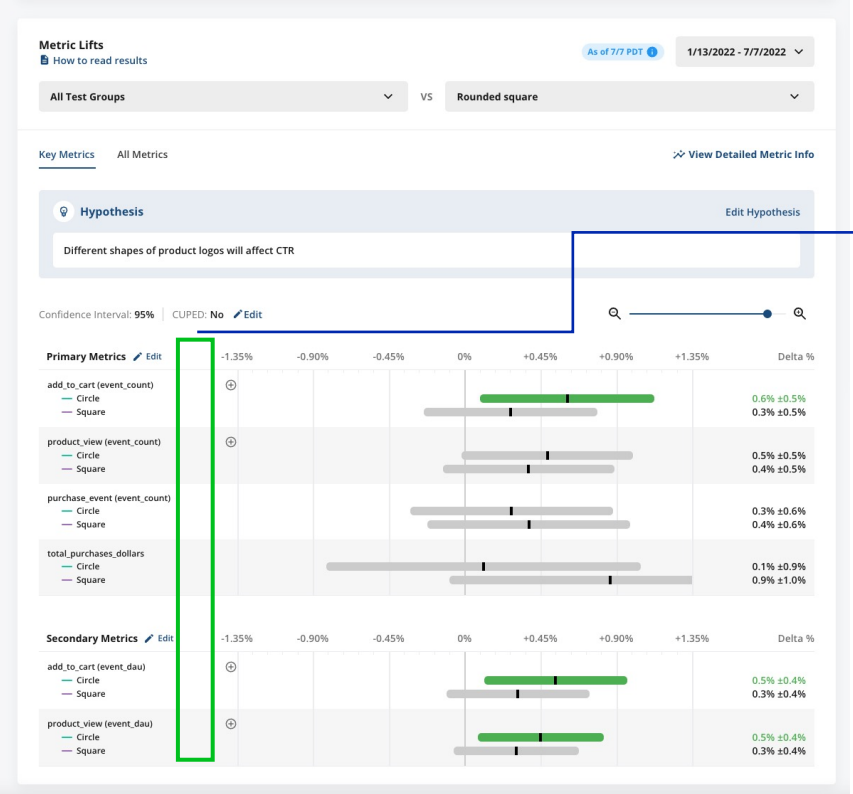


그림 10. 2-표본 t검정 결과 제공 방식 (Source: Statsig Docs, Read Results)

- P-value가 위치할 자리
- 우측에 시각화된 구간은 95% 신뢰구간을 의미함
  - 구간에 0이 포함 안되면 귀무가설을 기각 시킬 수 있음
  - 유의수준 5% 하의 p-value로 내린 결론과 같은 결정을 내리게 됨

통계적 유의도와 실무적 유의도

1. 통계적 유의도 X / 실무적 유의도 X
- 변화가 별 효과 없음
- 실험 반복 또는 아이디어 포기

2. 통계적 유의도 O / 실무적 유의도 O
- 출시

3. 통계적 유의도 O / 실무적 유의도 X
- 출시할 가치가 없을지도 모름

4. 중립 그 자체..
- 어떤 결정을 내릴만한 증거가 부족 (신뢰구간 넓이)
- 더 많은 표본을 확보하여 후속 테스트 실행 권장

5. 조금 더 긍정적 중립
- 4번과 동일한 결정
- 단, 4번 보단 조금 더 긍정적으로 권장

6. 매우 긍정적 중립
- 4번과 동일한 결정
- 단, 5번 보다 더욱 긍정적으로 후속 테스트 권장
- 만약 당장 출시 결정을 해야만 하는 상황이라면, 출시를 택하는게 합리적

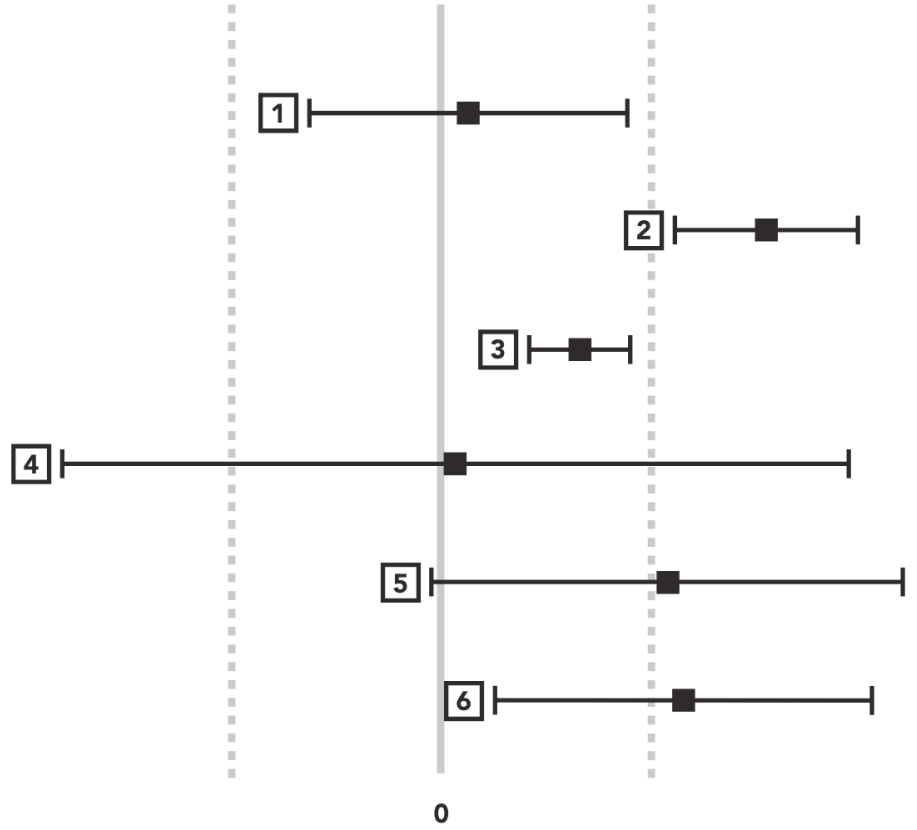


Figure 2.4 Examples for understanding statistical and practical significance when making launch decisions. The practical significance boundary is drawn as two dashed lines. The estimated difference for each example result is the black box, together with its confidence interval

그림 11. 통계적 유의도와 실무적 유의도(점선) (Source: Kohavi, Tang, and Xu 2020)

발표 내용 요약

- Pre-Experiment stage
  - 고객 가치에 기반한 가설 설계 및 피쳐 디자인
  - 정교한 실험 설계
  - 실험 전 편향 교정
- During-Experiment stage
  - 예상치 못한 효과의 조기 감지를 위한 모니터링
- Post-Experiment stage
  - 적절성 검사
  - 통계 분석 결과 해석
  - 신뢰구간과 실무적 유의도를 활용한 의사결정

A/B 실험을 통해 올바른 의사결정을 내릴 수 있도록

결과가 반복 가능(재현성)하고 신뢰할 수 있게 하는데 많은 노력 필요

References

[1] Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press. https://experimentguide.com/

[2] Gupta, S., Ulanova, L., Bhardwaj, S., Dmitriev, P., Raff, P., & Fabijan, A. (2018). The Anatomy of a Large-Scale Experimentation Platform. *2018 IEEE International Conference on Software Architecture (ICSA)*, 1-109. https://doi.org/10.1109/ICSA.2018.00009

[3] Machmouchi, W., Gupta, S., Zhang, R., & Fabijan, A. (2020, July 31). Patterns of Trustworthy Experimentation: Pre-Experiment Stage. *Microsoft Research*. https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/patterns-of-trustworthy-experimentation-pre-experiment-stage/

[4] Microsoft. (2021, January 25). Patterns of Trustworthy Experimentation: During-Experiment Stage. *Microsoft Research*. https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/patterns-of-trustworthy-experimentation-during-experiment-stage/

[5] Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., & Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 786-794. https://doi.org/10.1145/2339530.2339653

[6] Stewart, M. (2022, October 27). *Product Experimentation Best Practices*. https://www.statsig.com/blog/product-experimentation-best-practices

[7] Craig. (2022, July 7). *CUPED on Statsig*. Medium. https://blog.statsig.com/cuped-on-statsig-d57f23122d0e

[8] StatQuest with Josh Starmer (Director). (2020). *Statistical Power, Clearly Explained!!!* https://www.youtube.com/watch?v=Rsc5znwR5FA

[9] Burger, B., Vaudel, M., & Barsnes, H. (2021). Importance of Block Randomization When Designing Proteomics Experiments. *Journal of Proteome Research*, *20*(1), 122-128. https://doi.org/10.1021/acs.jproteome.0c00536

[10] Statsig Docs, https://docs.statsig.com/