# 예제를 통한 A/B 실험의 과정 이해

가짜연구소 인과추론팀
온라인 통제 실험 연구자로 거듭나기

2024-03-26

방태모


# 얻어갈 수 있는 것

Causal-Lab

## 실제 예제를 통해 A/B 실험에 기반한 의사결정 과정 이해하기

- 가설 설정
- 실험 설계
- 실험 실행 및 데이터 수집
- 통계 분석 결과 해석
- 의사 결정

# 1 실험 배경 이해하기

# A/B 테스트 배경

Causal-Lab

위젯을 판매하는 온라인 상점에서 실험을 하고자 함

- 마케팅 부서의 요청
  - 위젯의 할인 쿠폰 코드가 포함된 프로모션 이메일을 보내서 판매를 늘리고 싶어요!

# A/B 테스트 배경

Causal-Lab

위젯을 판매하는 온라인 상점에서 실험을 하고자 함

- 마케팅 부서의 요청
  - 위젯의 할인 쿠폰 코드가 포함된 프로모션 이메일을 보내서 판매를 늘리고 싶어요!

외부 자료에 기반한 사전 도메인 지식

- 체크아웃 도메인에 쿠폰 코드 필드를 추가하는 것만으로 수익이 저하됨

# A/B 테스트 배경

Causal-Lab

즉, 사용자가 쿠폰 코드 필드를 본다.

➔ 구매 전환 속도가 느려질 수 있고

➔ 쿠폰 코드를 검색하게 하여

➔ 심지어 사용자를 떠나게 할 수 있다.

# A/B 테스트 배경

Causal-Lab

즉, 사용자가 쿠폰 코드 필드를 본다.

➔ 구매 전환 속도가 느려질 수 있고

➔ 쿠폰 코드를 검색하게 하여

➔ 심지어 사용자를 떠나게 할 수 있다.

위 사항이 우리 프로덕에서도 발생하는지 검증해보자.

➔ A/B 테스트로!

# A/B 테스트가 갖는 2가지 가치

Causal-Lab

1. 실험적 피쳐의 인과적 효과 검증

# A/B 테스트가 갖는 2가지 가치

Causal-Lab

1. 실험적 피쳐의 인과적 효과 검증

2. 출시 전 최소 기능 구현으로 최소 비용으로 아이디어 검증
    - 본 예제가 이 사례를 잘 보여줄 것

# 본 예제의 실험군 구현 방식

Causal–Lab

실제 쿠폰 코드 시스템은 구현하지 않고,

- 체크아웃 페이지에 쿠폰 코드 필드 (UI)를 더하는 간단한 추가 변경만 수행

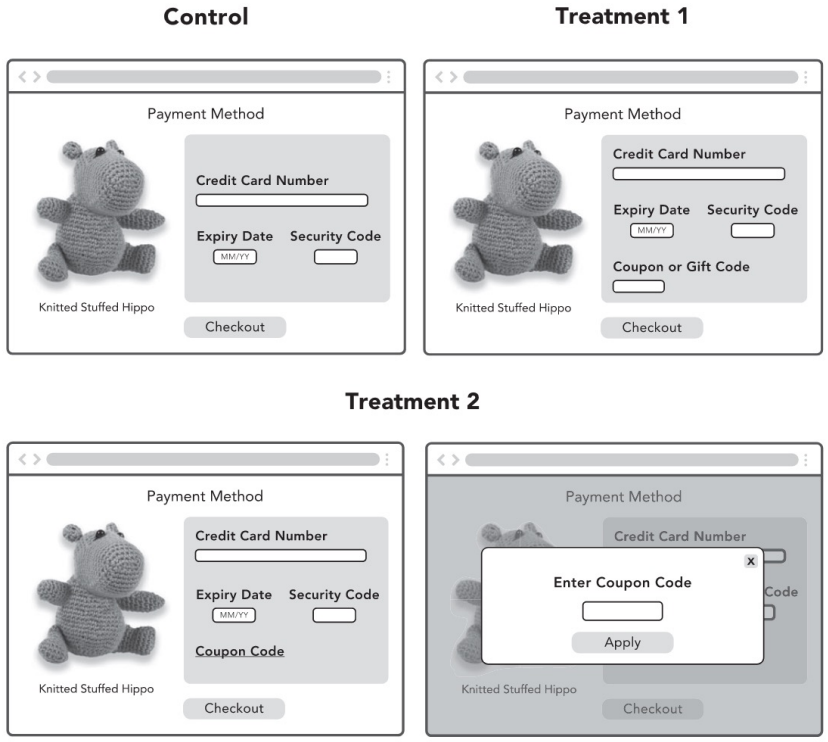


Figure 2.2 (1) Control: the old checkout page. (2) Treatment one: coupon or gift code field below credit card information (3) Treatment two: coupon or gift code as a popup

Source: Kohavi, Tang, and Xu 2020

# 2 가설 설정

# 가설 초안 설정

Causal-Lab

"**쿠폰 코드 필드**를 **체크아웃 페이지**에 더하는 것은 **매출**을 저하할 것이다."

# 가설 초안 설정

Causal-Lab

"**쿠폰 코드 필드**를 **체크아웃 페이지**에 더하는 것은 **매출**을 저하할 것이다."

가설을 설정 시 고려하면 좋은 3가지 요소

- Feature
- Domain
- Metric (OEC, Success Metric)
  - 표본 크기에 대해 표준화된 지표 사용 권장
  - **사용자 당 매출**은 좋은 OEC(Overall Evaluation Criterion)

# 가설 초안 설정

Causal-Lab

사용자 당 매출 지표의 분모로 어떤 사용자들을

고려할 것인가?

# 가설 초안 설정

Causal-Lab

**사용자 당 매출 지표**의 **분모**로 어떤 사용자들을 고려할 것인가?

- 퍼널 관점에서 고객의 쇼핑 프로세스 떠올려 보자!
  - 사이트를 방문한 모든 사용자

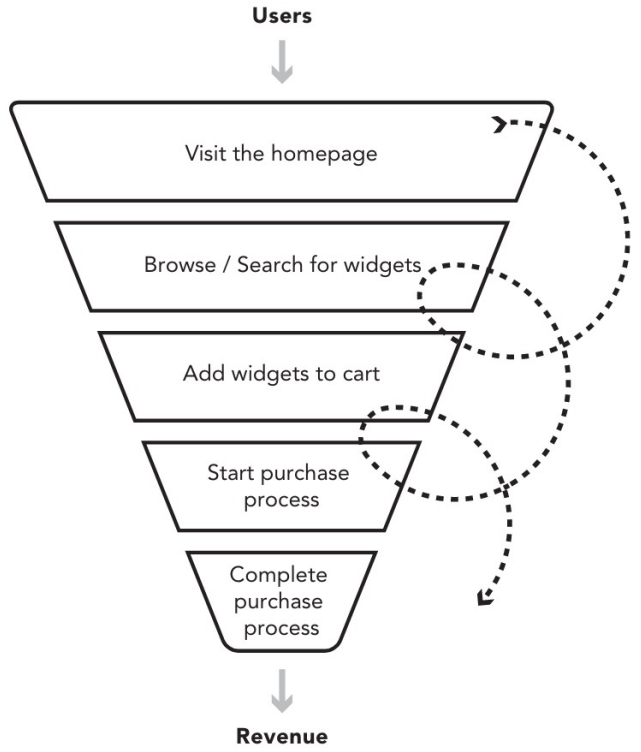


Figure 2.1 A user online shopping funnel. Users may not progress linearly through a funnel, but instead skip, repeat or go back-and-forth between steps

Source: Kohavi, Tang, and Xu 2020

# 가설 초안 설정

Causal-Lab

**사용자 당 매출 지표**의 **분모**로 어떤 사용자들을 고려할 것인가?

- 퍼널 관점에서 고객의 쇼핑 프로세스 떠올려 보자!
  - 사이트를 방문한 모든 사용자
  - 구매 프로세스를 완료한 사용자

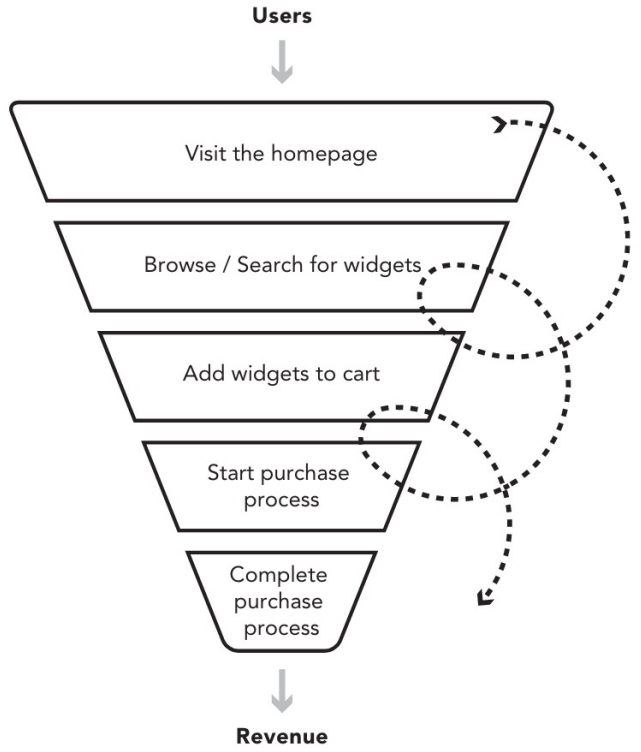


Figure 2.1 A user online shopping funnel. Users may not progress linearly through a funnel, but instead skip, repeat or go back-and-forth between steps

Source: Kohavi, Tang, and Xu 2020

# 가설 초안 설정

Causal-Lab

**사용자 당 매출 지표**의 **분모**로 어떤 사용자들을 고려할 것인가?

- 퍼널 관점에서 고객의 쇼핑 프로세스 떠올려 보자!
  - 사이트를 방문한 모든 사용자
  - 구매 프로세스를 완료한 사용자
  - **구매 프로세스를 시작한 사용자 (최적의 선택)**

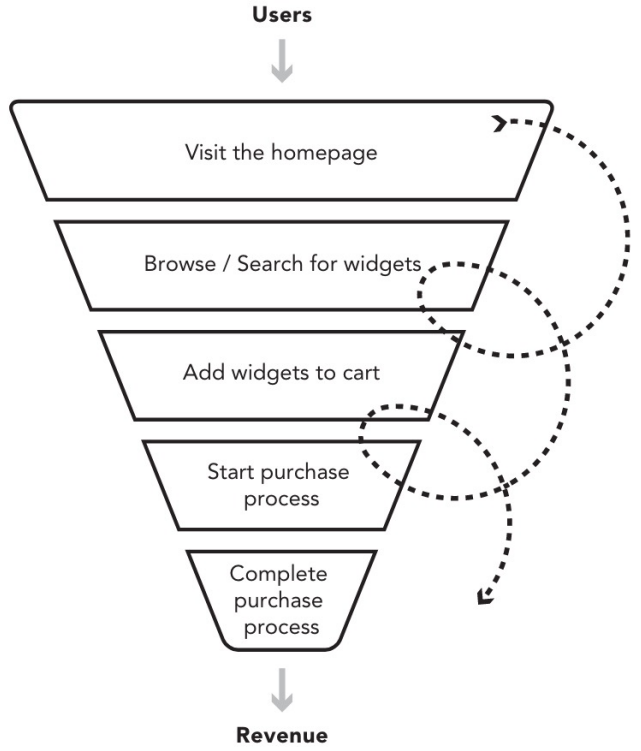


Figure 2.1 A user online shopping funnel. Users may not progress linearly through a funnel, but instead skip, repeat or go back-and-forth between steps

Source: Kohavi, Tang, and Xu 2020

# 가설 정교화

Causal-Lab

"**쿠폰 코드 필드**를 **체크아웃 페이지**에 더하면,

**구매 프로세스를 시작하는 사용자**에 대한 **사용자 당 수입**이 저하한다."

# 가설 정교화

Causal-Lab

"**쿠폰 코드 필드**를 **체크아웃 페이지**에 더하면,

**구매 프로세스를 시작하는 사용자**에 대한 **사용자 당 수입**이 저하한다."

좋은 가설이 갖추고 있는 3가지 요소

- **Feature**: 쿠폰 코드 필드
- **Domain**: 체크아웃 페이지
- **Metric** (OEC, Success Metric): 사용자 당 수입
  - 지표의 분모: 구매 프로세스를 시작하는 사용자
  - 즉, 새로운 Feature를 경험한 사용자 전원

# 3 실험 설계, 실행, 분석 전 통계적 가설 검정 개념 짚기

# 통계적 검정력

Causal-Lab

변형군 간 **실제로 차이가 있을 때**, 이 차이를 유의미하다고 **올바르게** 판별할 확률

- 검정력 분석을 통해 80%의 검정력을 달성하기 위한 표본 크기를 알아낼 수 있음
- 이에 따라 실험의 트래픽과 기간이 결정된다.
  - 지난 스터디 때 검정력과 검정력 분석에 대해 자세히 이야기 함
  - 슬라이드 보러가기

# 통계적 유의도

## P값에 의한 의사결정 = 신뢰구간에 의한 의사결정

- 즉, 관측된 검정 통계량으로 계산된 p-value가 0.05보다 작은 경우
- 95% 신뢰구간에 0이 포함되지 않을 것

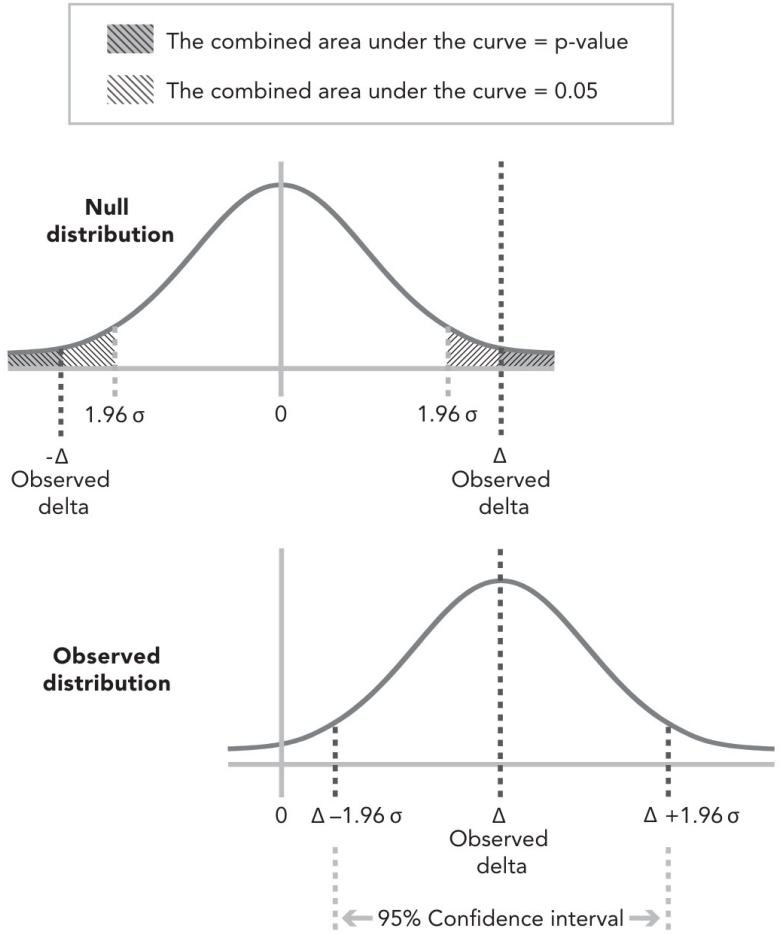


Figure 2.3 Top: Using p-value to assess whether the observed delta is statistically significant. If p-value is less than 0.05, we declare that the difference is statistically significant. Bottom: The equivalent view of using 95% confidence interval $[\Delta - 1.96\,\sigma, \Delta + 1.96\,\sigma]$ to assess statistical significance. If zero lies outside of the confidence interval, we declare significance

Source: Kohavi, Tang, and Xu 2020

# 4 실험 설계

# 실험 설계시 고려해야 할 4가지 요소

Causal-Lab

1. 무작위 추출 단위

일반적으로 **사용자**

# 실험 설계시 고려해야 할 4가지 요소

Causal-Lab

2. 무작위 추출 단위의 모집단 대상 (타겟팅)

**지리적 지역**, **플랫폼 및 장치 유형**, **앱 버전** 등

# 실험 설계시 고려해야 할 4가지 요소

Causal-Lab

3. 트래픽
   - 대규모 변경의 경우 적은 비율의 사용자로 실험 시작할 것
4. 실험 기간

위 두 항목은 **결과의 정밀도에 직접적 영향**을 미침

- **검정력 분석**과 연관
  - 얼마나 정밀한 효과를 감지할 것인가? (MDE)
  - 얼마나 많은 트래픽이 허가되어 있는가?

$$n = \frac{2\sigma^2\left(Z_{1-\beta} + Z_{1-\frac{\alpha}{2}}\right)^2}{\delta^2}$$

Source: Kohavi et al., 2022
실제 예제는 지난 스터디 슬라이드 p.71-72 참고

# 실험 설계시 고려해야 할 4가지 요소

Causal-Lab

실험 기간과 관련하여 고려해야할 사항 4가지

- **실험을 오래할수록 더 많은 사용자들이 실험에 참가**
  - 단, 사용자 당 세션 수와 분산이 증가하는 경우 예외가 발생
  - 동일한 사용자가 돌아온다는 점을 고려할 때 사용자 누적률은 저선형(sub-linear)이 될 수 있음
  - 즉, 실험 기간을 오래 가져가 표본 크기를 늘리는 것에는 한계가 존재

# 실험 설계시 고려해야 할 4가지 요소

Causal-Lab

실험 기간과 관련하여 고려해야할 사항 4가지

- **실험을 오래할수록 더 많은 사용자들이 실험에 참가**
  - 단, 사용자 당 세션 수와 분산이 증가하는 경우 예외가 발생
  - 동일한 사용자가 돌아온다는 점을 고려할 때 사용자 누적률은 저선형(sub-linear)이 될 수 있음
  - 즉, 실험 기간을 오래 가져가 표본 크기를 늘리는 것에는 한계가 존재
- **주간 효과: 평일과 주말의 사용자 분포는 다름 (∴최소 1주 권장)**

# 실험 설계시 고려해야 할 4가지 요소

Causal-Lab

실험 기간과 관련하여 고려해야할 사항 4가지

- **실험을 오래할수록 더 많은 사용자들이 실험에 참가**
  - 단, 사용자 당 세션 수와 분산이 증가하는 경우 예외가 발생
  - 동일한 사용자가 돌아온다는 점을 고려할 때 사용자 누적률은 저선형(sub-linear)이 될 수 있음
  - 즉, 실험 기간을 오래 가져가 표본 크기를 늘리는 것에는 한계가 존재
- **주간 효과: 평일과 주말의 사용자 분포는 다름 (∴최소 1주 권장)**
- **계절성: 외적 타당성(external validity)과 관련**
  - 크리스마스 시즌과 나머지 일반적 시즌의 기프트 카드 판매량은 다를 수 있음

# 실험 설계시 고려해야 할 4가지 요소

Causal-Lab

실험 기간과 관련하여 고려해야할 사항 4가지

- **실험을 오래할수록 더 많은 사용자들이 실험에 참가**
  - 단, 사용자 당 세션 수와 분산이 증가하는 경우 예외가 발생
  - 동일한 사용자가 돌아온다는 점을 고려할 때 사용자 누적률은 저선형(sub-linear)이 될 수 있음
  - 즉, 실험 기간을 오래 가져가 표본 크기를 늘리는 것에는 한계가 존재
- **주간 효과: 평일과 주말의 사용자 분포는 다름 (∴최소 1주 권장)**
- **계절성: 외적 타당성(external validity)과 관련**
  - 크리스마스 시즌과 나머지 일반적 시즌의 기프트 카드 판매량은 다를 수 있음
- **초두 효과와 신기 효과**
  - 초두 효과(primacy effect): 실험 초기 효과가 정상보다 작은 것
  - 신기 효과(novelty effect): 실험 초기 효과가 정상보다 큰 것

# 실험 설계시 고려해야 할 4가지 요소

Causal-Lab

**트래픽, 실험 기간**이 **고정**되어 있다는 가정하에 **결과의 정밀도**를 올릴 수 있는 방법

- **지표 변경 및 변환**(값 제한, 이진화, 로그 변환)을 통한 분산 축소
  ➔사용자 당 수익
  ➔~원 이상 구매한 사용자의 여부를 나타내는 지표 (교재 18장 p.278 넷플릭스 예제)
  ➔구매한 사용자의 여부를 나타내는 지표 구매여부
- CUPED를 활용한 분산 축소
  - 지난 스터디 슬라이드 p.89-90 참고

# 실험 설계시 고려해야 할 4가지 요소 - 요약

Causal-Lab

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

# 실험 설계시 고려해야 할 4가지 요소 - 본 예제에 대입

Causal-Lab

## 본 예제 실험의 설정

- 무작위 추출 단위
  - **사용자**
- 타겟팅
  - **모든 사용자** (단, 체크아웃 페이지를 방문하는 사용자만 분석에 포함)
- Relative MDE = 1%
  - 검정력 80% 하에 사용자 당 수익의 **1% 이상 변화를 감지**하고자 함
- 검정력 분석에 의해 결정된 트래픽과 실험기간
  - 트래픽
    - 100%
    - 대조군/실험군1/실험군2 34/33/33% **균등 분할 (∵검정력 최대)**
  - 실험 기간
    - 4일이지만, 주간효과를 고려하여 **7일**간 실험 진행

Causal–Lab

# 5 실험 실행 및 데이터 수집

# 실험의 실행을 위해 필요한 3가지

Causal-Lab

- 로그 데이터를 얻기 위한 계측(instrumentation) (교재 13장 참조)
  - 실험 키
  - 실험 버전
  - 변형군 할당 정보
  - …
- 실험 인프라 (교재 4장 참조)
  - **최소 필요 인프라**
    - 실험 구성(트래픽, 실험기간, 타게팅 등) 및 변형군 할당을 수행할 수 있는 실험 인프라

# 실험의 실행을 위해 필요한 3가지

Causal-Lab

- 로그 데이터를 얻기 위한 계측(instrumentation) (교재 13장 참조)
  - 실험 키
  - 실험 버전
  - 변형군 할당 정보
  - …
- 실험 인프라 (교재 4장 참조)
  - **최소 필요 인프라**
    - 실험 구성(트래픽, 실험기간, 타게팅 등) 및 변형군 할당을 수행할 수 있는 실험 인프라
  - **완성형 실험 인프라**
    - Experimentation Platform
    - Metric Store
    - Stats engine
    - Log system + Design System
    - …


# 6 결과 해석

# 적절성 검사

Causal-Lab

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

# 적절성 검사

Causal-Lab

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

적절성 검사에서 **가드레일 지표 검토**

- 반드시 개선할 필요는 없지만 성과 저하는 원하지 않는 측면 측정
- **신뢰 관련 가드레일 지표** (데이터 품질 지표)
  - SRM(Sample ratio mismatch): 변형군 할당 크기가 실험 설정을 잘 따르는가?
  - 캐시 적중률: 서버 캐시와 관련한 이야기
  - …
- **조직 관련 가드레일 지표**
  - 페이지 로드 시간: 새로운 피쳐에서 속도 저하가 있는가?
  - …

# 적절성 검사

Causal-Lab

적절성 검사가 실패하였는가?

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

# 결과 해석

Causal-Lab

## 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%] |

Source: Kohavi, Tang, and Xu 2020

# 결과 해석

Causal-Lab

## 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%] |

Source: Kohavi, Tang, and Xu 2020

- **쿠폰 코드를 추가하면 수익이 감소한다**는 패턴을 확인
  - 구매 과정을 완료하는 사용자 수가 더 적어져 수익이 감소
- 쿠폰 코드 발송 마케팅 이메일이 회수해야 하는 2가지 비용
  - 쿠폰 처리와 유지보수를 추가하는데 드는 **구현 비용**
  - **쿠폰 코드를 처음 추가할 때 발생하는 부정적 영향의 비용** (실험 결과에 따른)
- 프로모션 코드를 도입하는 아이디어는 폐기하기로 결정
- **출시 전 최소 테스트**로 A/B 테스트를 수행함으로써 많은 노력 절감!


# 7 의사결정

Causal-Lab

**올바른 의사결정**을 내릴 수 있도록,

**결과가 반복 가능**하고 **신뢰할 수 있게 하는데 많은 노력이 필요**

- 여러 지표간 트레이드오프의 고려 필요성 검토
  - 사용자 참여가 증가하는데, 수익이 감소할 경우 출시할 것인가?
  - …
- 서비스 출시 비용이 얼마나 큰가?
  - 출시 전 구현 비용
  - 유지보수 비용

# 통계적 유의도와 실무적 유의도

Causal-Lab

## 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

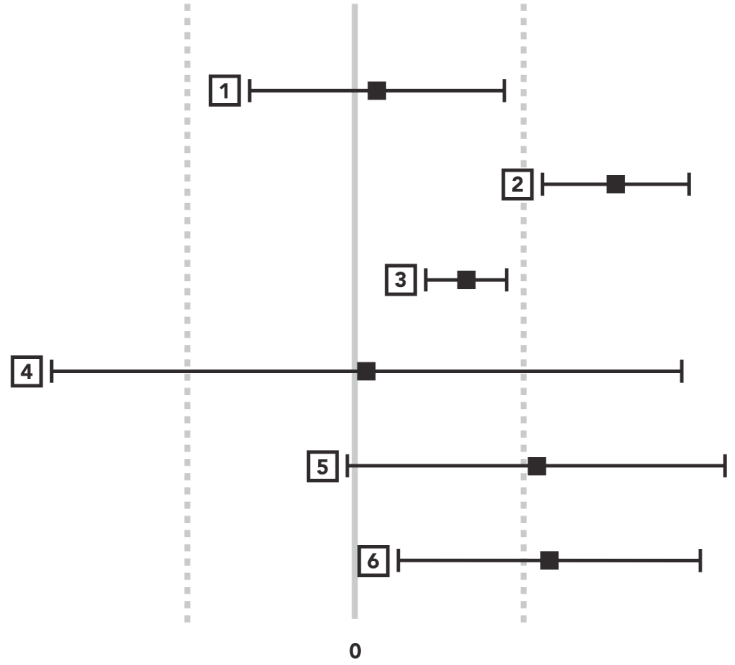


Figure 2.4 Examples for understanding statistical and practical significance when making launch decisions. The practical significance boundary is drawn as two dashed lines. The estimated difference for each example result is the black box, together with its confidence interval

Source: Kohavi, Tang, and Xu 2020

Causal-Lab

# 의사결정 간 기억해야 할 것

- 명확한 답이 나오지 않을 수 있지만 결정을 내려할때가 존재
- 이러한 상황에서는 어떤 요인을 고려하고 있는지,
- 특히 이 요인들이 실무적 및 통계적 유의도 경계에서 어떻게 반영되는지 명확하게 할 필요가 있음