

이커머스 보고서

서론

1. A/B 테스트 개념 및 분석 목적

디지털 환경에서 사용자 경험을 개선하고 비즈니스 성과를 높이기 위해서는 직관이나 추측이 아닌 실제 데이터에 기반한 의사 결정이 필수적이다. A/B 테스트는 두 그룹으로 나누어 각기 다른 버전을 부여하고 더 나은 성과를 도출할 수 있는지 비교하는 작업이다. 통계적 검증을 거쳐 실제로 효과가 있는지 객관적으로 확인하여 리스크를 최소화하고 최적의 선택을 할 수 있다.

본 분석에서는 Control 그룹과 Treatment 그룹 간의 전환율 차이가 통계적으로 유의한지 검증하고 새로운 페이지 (new_page) 도입 여부를 결정하고자 한다.

2. 데이터 설명

kaggle에서 제공하는 A/B 테스트 샘플 데이터를 활용하여 A/B 테스트를 진행하였다. (출처: <https://www.kaggle.com/datasets/zhanguyuan/ab-testing>)

데이터는 user_id, timestamp, group, landing_page, converted 총 5개의 컬럼으로 구성되어 있으며 각 컬럼의 의미는 다음과 같다.

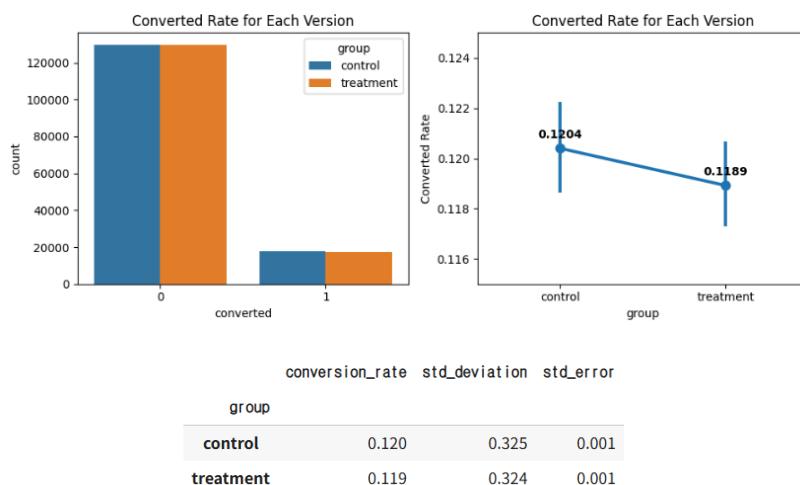
컬럼	설명
user_id	각 사용자를 구분하는 고유 식별자
timestamp	사용자가 웹사이트를 방문한 시간
group	실험 그룹 구분 (control: 대조군, treatment: 실험군)
landing_page	사용자에게 노출된 페이지 버전 (old_page/new_page)
converted	전환 성공 여부 (1: 성공, 0: 실패)

본론

1. 데이터 EDA

1-1. 데이터 분포 확인

- 웹페이지 변경 전후 전환율 파악



- 전환율

- group별 전환율은 큰 차이가 없었으며 오히려 변경 후 전환율이 1% 감소하였다.

- 표준 오차

- 샘플 크기가 매우 크고 표준 오차가 0.001 수준으로 작게 나타나기 때문에 표본 통계량의 추정치가 매우 안정적임을 알 수 있다.

1-2. 결측치 확인

- 없음

1-3. 데이터 정합성 확인

- 변수별 데이터 확인

	0
user_id	290584
timestamp	294478
group	2
landing_page	2
converted	2

dtype: int64

- group과 landing_page가 올바르게 매칭되지 않은 행 제거

landing_page new_page old_page		
group		
control	1928	145274
treatment	145311	1965

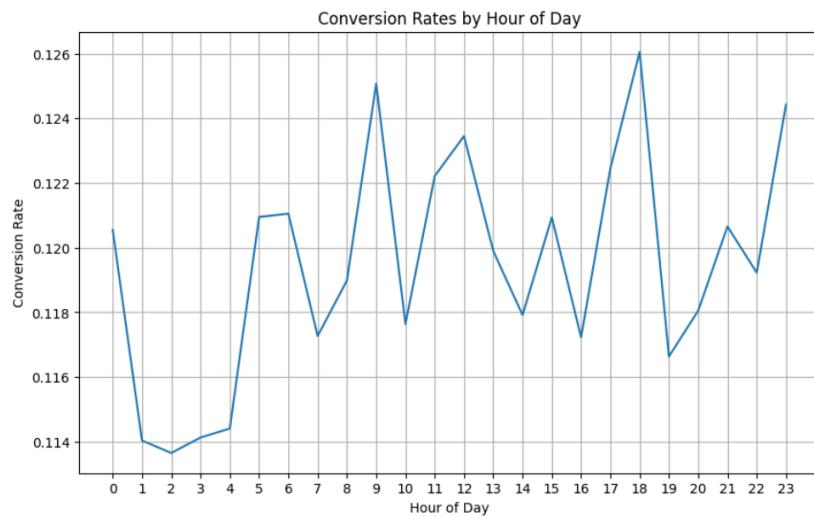
landing_page new_page old_page		
group		
control	0	143293
treatment	143397	0

사전 경험이 실험에 영향을 줄 수 있다고 판단하여 user_id 기준으로 세션이 2회 이상 존재하는 사용자(홈페이지를 한번 이상 추가로 방문한 사용자)를 제거하고 1회 방문 사용자만을 대상으로 실험을 진행하였다.

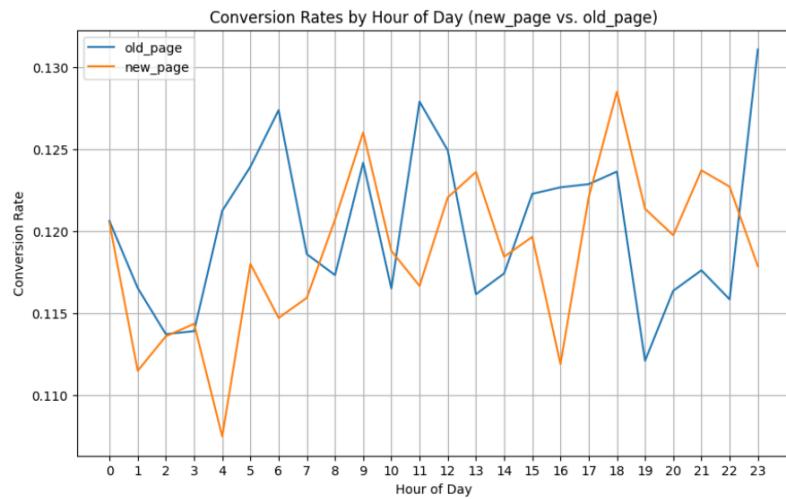
1-4. 시간대 기반 분석

시간대별 평균 전환율 변화 패턴을 분석하여 전환율이 상대적으로 낮거나 높은 시간대를 식별하여 개선이 필요한 구간을 확인했다.

- 시간대별 전체 전환율



- 웹페이지별, 시간대별 전환율



- 결과 해석

- 전체 시간대별 전환율 패턴
 - 전환율은 심야 시간대인 오전 1시부터 4시까지 낮은 수준을 보였으며, 출근 시간대인 오전 9시, 퇴근 시간대인 오후 6시, 그리고 취침 전 시간대인 밤 11시에 상대적으로 높은 수준이 나타났다.
- 버전별 시간대 전환율 특성
 - 신규 페이지(new_page)는 새벽 시간대인 오전 4시부터 7시 사이에 낮은 전환율을 보였으나, 저녁 시간대인 오후 6시부터 10시 사이에는 높은 전환율을 기록하였다.
 - 반면, 기존 페이지(old_page)는 오전 4시부터 7시 사이에 상대적으로 높은 전환율을 보였으며, 오후 7시부터 10시까지는 낮은 전환율을 나타냈다.
 - 이는 두 버전이 시간대별로 상반된 성과 패턴을 보임을 시사하며 특히 저녁 시간대 (오후6~10시)에서는 신규 페이지의 우위가 새벽, 오전 시간대 (오전 4~7시)에서는 기존 페이지의 우위가 관찰된다.

2. 실험 설계

2-1. Z-test 검증

2-1-1. 가설 설정

- 귀무 가설 : 수정된 웹페이지는 고객 전환율에 의미있는 영향을 끼치지 못한다.
- 대립 가설 : 수정된 웹페이지는 고객 전환율에 의미있는 영향을 준다.

2-1-2. 검정 방법

- z-test 사용 이유
 - 전환율과 같이 성공(1) 또는 실패(0)의 이진 데이터는 이항분포를 따르고 샘플이 충분히 크면 정규분포로 근사되기 때문에 z-test 검정 방법을 선택하였다.
- 양측 검정
 - 방향성에 대한 가설이 없고 단순히 A와 B가 다른지 확인하고 싶으므로 양측 검정을 사용하고자 한다.

2-1-3. 샘플 사이즈 결정

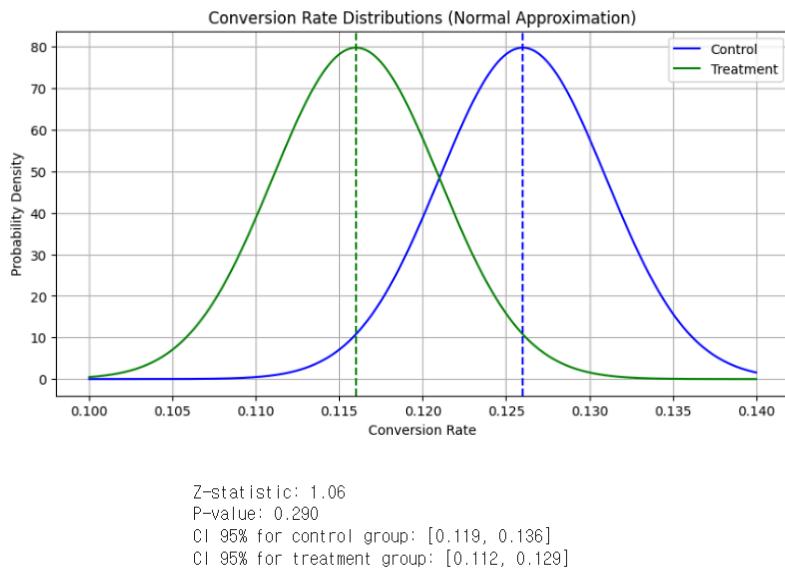
- 통계적 가정 및 파라미터 설정
 - 효과 크기(Effect Size)를 기반으로 검정력 분석(Power Analysis)을 통해 필요한 표본 크기를 계산했으며, 파라미터는 다음과 같다.
 - 기본 전환율 : 12.0%
 - 기대 전환율 : 14.0%
 - 유의수준 : 0.05
 - 검정력(Power) : 0.90
 - 0.8 : 일반적인 A/B 테스트에서 가장 많이 사용한다.
 - 0.9 : 신뢰도를 더 확보하고 싶을 때 사용한다.
 - 설정된 통계적 파라미터를 기반으로 각 그룹이 최소한으로 확보해야 할 샘플 크기(n)은 다음과 같다.
 - 각 그룹당 필요한 최소 샘플 수 (n) : 5,934
 - 총 실험에 필요한 최소 샘플 수 (Total N) : 11,868
- 그룹 별 5934개의 데이터 확인

group	conversion_rate	std_deviation	std_error
control	0.127	0.333	0.004
treatment	0.121	0.326	0.004

- 전환율 (Converstion Rate)
 - control 그룹이 treatment 그룹의 전환율보다 0.6% 높다. 이는 새로운 버전이 기존보다 전환율 상승 측면에서 성능이 떨어진다는 것을 시사한다.
- 표준 편차 (Standard Deviation)
 - 두 그룹 모두 비슷한 수준으로 낮다. 이를 통해 두 그룹의 산포 정도가 유사하다는 것을 확인할 수 있다.
- 표준 오차 (Standard Error)
 - control 그룹과 treatment 그룹의 표준 오차 값이 동일하게 나타났으며 이는 그룹에서 추정된 전환율의 불확실성 수준이 비슷함을 의미한다.

2-1-4. 결과

- 결과 요약



- z 통계량 = 1.06
 - 두 그룹의 전환율 차이를 해당 차이의 표준 오차로 나눈 값이며 1.96 (95% 신뢰수준) 이상이어야 통계적으로 유의미하다.
- p-value = 0.290
 - 만약 실제로 두 그룹 전환율에 차이가 없다는 귀무가설 하에서, 현재 관측된 전환율 차이 이상으로 극단적인 결과가 우연히 관측될 확률을 의미한다.
- 신뢰 구간(CI)
 - Control 그룹 : [11.9%, 13.6%] / Treatment 그룹 : [11.2%, 12.9%]
 - 두 그룹의 신뢰구간이 상당 부분 겹치며 이는 전환율 차이가 크지 않을 가능성이 있다.
- 결과 해석
 - z 통계량은 1.06으로 임계값 1.96보다 작으며 p-value는 0.290으로 0.05보다 크다. 이는 귀무가설을 기각할 수 없으며 수정된 웹페이지는 고객 전환율에 유의미한 영향을 미쳤다고 판단할 통계적 근거가 없음을 의미한다. 또한 두 그룹의 신뢰 구간이 상당 부분 겹치고 treatment 그룹의 구간이 control 그룹보다 낮은 값을 포함하고 있어 전환율 차이가 존재한다고 보기 어렵다.

2-2. 카이제곱 검정

2-2-1. 가설 설정

- 귀무가설(H0): group과 converted는 독립적이다 (그룹이 전환에 영향을 주지 않는다)
- 대립가설(H1): group과 converted는 종속적이다 (그룹이 전환에 영향을 준다)

2-2-2. 검정 방법

- 독립성 검정 사용 이유
 - 독립성 검정
 - 두 범주형 변수가 서로 독립적인지 연관되어 있는지 확인한다.
 - 적합도 검정
 - 관측된 데이터 분포가 기대하는 이론적 분포와 일치하는지 확인한다.

본 분석에서는 group과 converted 간 연관성을 확인하기 위해 카이제곱 독립성 검정을 수행하여 두 그룹 간 전환율 차이가 통계적으로 유의미한지 평가하였다.

2-2-3. 샘플 사이즈 결정

- 샘플 크기에 따라 검정 방법을 선택한다.

- $n < 1,000$: Fisher's Exact Test 사용 (기대 빈도가 작을 때 정확)
 - $n \geq 1,000$: Chi-Square Test 사용 (계산 효율적이고 정확)
- 본 데이터는 충분히 큰 샘플이므로 Chi-Square Test를 적용한다.

2-2-4. 결과

- 결과 요약

```
[Detailed Chi-Square Info]
Degrees of Freedom = 1
Expected Table:
[[126176.71137814 126268.28862186]
 [ 17116.28862186 17128.71137814]]
Statistic=1.413, Critical=3.841, p=0.2345
Decision: Fail to Reject H0 (Independent)
```

Test	P_value	AB Hypothesis	Comment
0 Chi Square Test	0.234547	Fail to Reject H0	A/B groups are similar!

- 임계값 : 3.841 (자유도 = 1)
- 검정 통계량 : 1.413
- p-value : 0.2345

- 결과 해석

검정 통계량은 1.446로 임계값 3.841보다 작고, p-value는 0.2345로 유의수준 0.05보다 크기 때문에 통계적으로 유의미한 차이를 발견하지 못했다. 따라서 귀무가설을 기각할 수 없으며 이는 두 변수 간 통계적으로 유의한 관련성이 없다는 것을 의미한다.

2-3. 전환율 차이 존재 여부에 대한 베이지안 A/B 테스트

2-3-1. 분석 목적

본 분석의 목적은 베이지안 A/B 테스트를 통해 새로운 웹페이지(treatment)의 전환율이 기존 웹페이지(control)와 차이가 있는지를 확률적으로 평가하는 것이다. 이를 위해 두 그룹 전환율의 사후분포를 추정하고 전환율 차이가 0보다 크거나 작을 확률을 분석한다.

2-3-2. 검정 방법

- 베이지안 검정 사용 이유
 - 사전 정보 활용
 - 과거 실험 결과나 도메인 지식을 사전 확률로 반영 가능하다.
 - 적은 데이터로도 의미 있는 추론이 가능하다.
 - 불확실성의 정량화
 - 단순히 “유의하다/아니다”가 아닌 확률 분포로 불확실성을 표현할 수 있다.
- 베이지안 검정의 주요 가정 및 목표
 - 사전 분포
 - 데이터를 관찰하기 이전에 갖고 있던 믿음이나 지식을 확률 분포로 표현한다.
 - 약한 사전 분포 : Beta(1,1)
 - 0부터 1까지의 모든 확률 값에 대해 균일한 확률 부여한다.
 - 제프리 사전 분포 (Jeffreys Prior) : Beta(0.5, 0.5)
 - 통계적 이론에 기반하여 유도된 무정보적 사전 분포이다.
 - 특정 값에 대한 선호가 거의 없으며 데이터에 판단을 맡기겠다는 의미이다. (0~1 어디든 가능하다.)

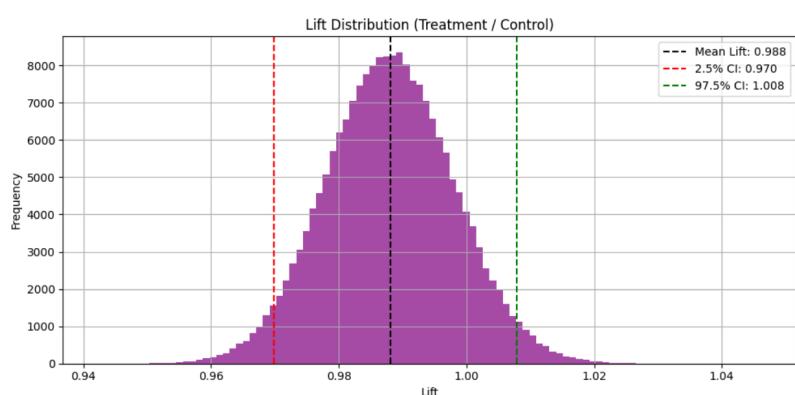
- 0이나 1과 같은 극단값 근처일 때 더 강한 믿음을 부여하며 데이터가 부족할 때 발생할 수 있는 오류를 보정하는 역할이다.
- 목표 (GO_ABS_DIFF)
 - Treatment 그룹의 전환율(pB)이 Control 그룹의 전환율(pA)보다 2% 이상 높을 것이다.
- 판단 기준 (THRESH)
 - 목표를 달성할 확률이 95% 이상일 때 ($P(pB - pA) \geq 0.95$), treatment를 승자로 판단한다.
- 사후 분포
 - 데이터를 관찰 전의 사전 분포와 관찰된 데이터의 가능도를 결합한 결과(Posterior x Prior)이다.

2-3-3. 샘플 사이즈 결정

- 베이지안 검정에서 샘플 사이즈의 의미
 - 관측 데이터가 많을수록 사후 분포의 분산이 감소하고 추정 전환율과 Lift가 안정적이다.
 - 빈도주의 검정과 달리 고정된 표본 수 n 에 대한 유의확률 계산에 의존하지 않고 실제 수집된 데이터에 따라 사후분포를 순차적으로 갱신한다.

2-3-4. 결과

- 결과 요약
 - 관측 전환율
 - control : 12.018%
 - treatment : 11.873%
 - $P(\text{treatment} > \text{control}) = 11.733\%$
 - treatment 그룹의 전환율이 control 그룹보다 높을 확률이 약 11.7%에 불과함을 의미한다.
 - 평균 Lift : 0.988
 - treatment 전환율이 control 전환율의 약 98.8% 수준이다.
 - Lift 분포는 왼쪽으로 치우쳐져 있고 오른쪽 꼬리가 짧아서 treatment가 개선 효과를 낼 가능성이 낮다.
 - 95% 신뢰구간 : [0.969, 1.008]



- treatment 그룹의 전환율은 control 그룹 대비 약 3.1% 낮은 수준에서 최대 0.8% 수준까지의 범위를 포함한다.

- 결과 해석

표면적으로는 control이 소폭 높아 보이나 차이는 0.145%p로 크지 않다. treatment가 control보다 실제로 더 높은 전환율을 가질 확률은 매우 낮으며 베이지안 기준(95%)에 현저히 미달한다. 지금 시점에서는 기존 페이지를 유지하는 것이 더 타당하다.

2-4. 목표 개선판 달성 여부에 대한 베이지안 A/B 테스트

2-4-1. 분석 목적

본 분석에서는 베이지안 A/B 테스트를 통해 새로운 웹페이지(treatment)의 전환율이 기존 웹페이지(control)보다 높을 확률과 그 차이의 크기를 추정하고, 사전에 정의한 최소 개선판 기준을 충족하는지를 평가하고자 한다.

2-4-2. 검정 방법

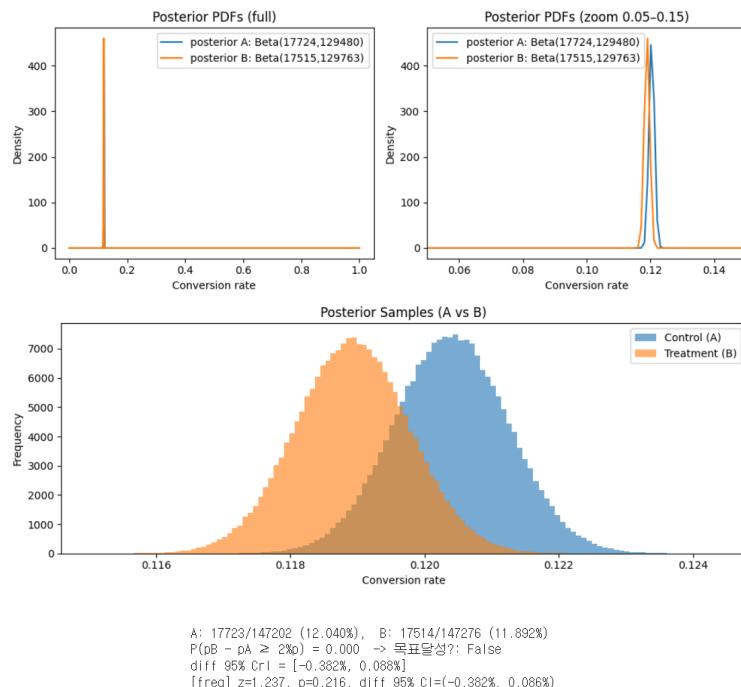
- 베이지안
 - 위와 동일하게 베이지안 검정을 사용하되 가설 목표 설정을 다르게 하여 검정한다.

2-4-3. 샘플 사이즈 결정

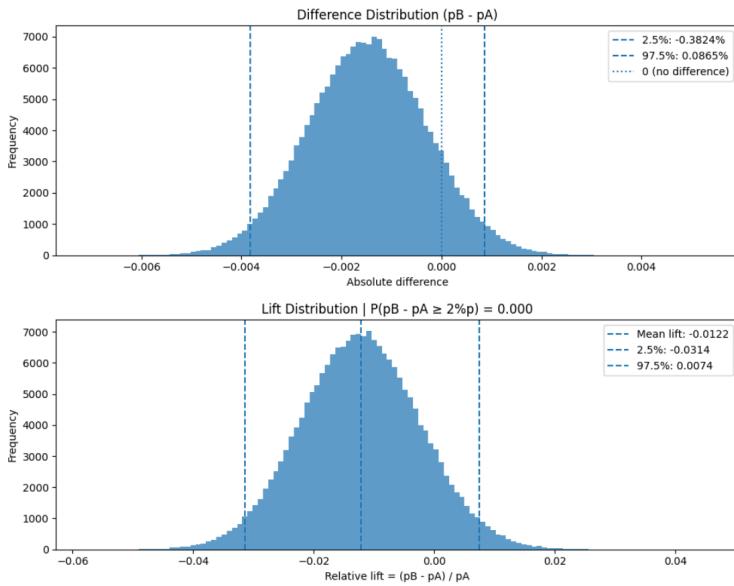
- 하이브리드 종료 조건 선택
 - 각 그룹의 최소 1000명 이상일 경우
 - 베이지안 사후확률 기준으로, treatment 전환율이 control 전환율보다 최소 2%p 이상 높을 확률이 95% 이상일 경우

2-4-4. 결과

- 결과 요약
 - 관측 전환율



- A(control) 그룹 = 12.040%
- B(treatment) 그룹 = 11.892%
- 차이($p_B - p_A$) = -0.148%p
- 95% 신뢰구간 : -0.382%p ~ +0.088%p



- treatment 그룹의 전환율은 control 그룹 대비 약 0.382%p 낮은 수준에서 최대 0.086%p 수준까지의 범위를 포함한다.
- 구간에 0이 포함되어 있어 실질적 차이가 없을 가능성이 높다.
- 빈도주의 검정

$z\text{-statistic} = -1.2369$
 $p\text{-value (two-sided)} = 0.2161$

- $z\text{-통계량} = -1.237$
- $p\text{-value} = 0.216$
 - $p\text{-value}$ 는 0.05보다 크기 때문에 통계적으로 유의하지 않다.

• 결과 해석

두 개의 조건이 충족될 경우 실험을 종료하고 treatment 그룹을 승자로 선언한다. 하지만 본 검정에서는 treatment가 control보다 전환율이 최소 2%p 더 높을 확률이 0.95에 현저히 미치지 못해 목표 수준의 확신을 얻지 못했다.

두 버전 간 전환율 차이는 통계적으로 유의미하다고 판단할 수 없으며 treatment 버전의 목표(control 대비 2%p 개선)를 달성할 가능성은 매우 낮다. 오히려 관측 전환율은 control이 treatment가 소폭 높았으나, 해당 차이 역시 통계적으로 유의하지 않았다. 따라서, 현 시점에서는 기존 버전을 유지하는 것이 합리적이다.

결론

1. 통계적 유의성 부재

세 가지 통계 검정 방법 모두에서 일관되게 통계적 유의성을 발견하지 못한다.

- Z-test : 통계량 1.06, $p\text{-value}$ 0.290
- 카이제곱 검정 : 검정 통계량 1.413, $p\text{-value}$ 0.235
- 베이지안 분석 : 95% 신뢰구간이 0을 포함 [-0.382% ~ +0.088%p]

모든 $p\text{-value}$ 가 유의수준 0.05보다 크게 상회하며 관측된 차이가 우연에 의한 것일 가능성이 높다.

2. 실질적 차이 미미

- Control 그룹 : 12.040%
- Treatment 그룹 : 11.892%
- 차이 : 단 0.148%p (실질적으로 거의 동일)

3. 신뢰 구간 중첩

- Control : [11.9%, 13.6%]
- Treatment : [11.2%, 12.9%]

두 그룹의 신뢰 구간의 상당 부분이 겹치며 treatment 구간이 오히려 낮은 쪽에 치우쳐 있어 개선 효과를 입증하기 어렵다.

4. 목표 미달성

베이지안 분석 결과, treatment 그룹이 control 그룹 대비 2%p 이상 개선될 확률은 사실상 0%로, 설정된 개선 목표를 달성하지 못했다.

5. 최종 결론

새로운 버전(Treatment)은 기존 버전(Control) 대비 통계적으로나 실질적으로 유의미한 개선을 보여주지 못한다. 따라서, 기존 버전을 유지하는 것이 최선의 선택이다.