

쿠키캣 보고서

서론

1. A/B 테스트 개념 및 분석 목적

3개 이상의 동일한 색상 타일을 연결하여 보드를 정리하는 퍼즐 게임은 플레이어가 레벨을 클리어하며 점진적으로 난이도가 상승하는 구조를 가진다. 게임을 진행하는 과정에서 플레이어는 특정 시점에 '게이트'에 도달하게 되며 이 게이트는 다음 단계로 진행하기 위해 일정 시간 대기하거나 인앱 결제를 선택하도록 설계되어 있다.

이러한 게이트는 단순히 수익 창출을 위한 장치가 아니라 플레이어에게 의도적인 휴식을 제공함으로써 게임의 몰입도와 장기적인 지속성을 조절하는 중요한 요소로 작용한다. 따라서 게이트의 위치는 사용자 경험과 수익, 그리고 재방문 행동에 직접적인 영향을 미칠 수 있는 핵심 설계 변수라고 볼 수 있다.

본 실험에서는 첫 번째 게이트의 위치를 기존 30레벨에서 40레벨로 이동시켰을 때 사용자 행동에 어떠한 변화가 발생하는지를 분석하고자 한다. 이를 위해 게이트가 30레벨에 위치한 기존 버전을 Control 그룹(gate_30)으로 설정하고, 게이트를 40레벨로 이동시킨 버전을 Treatment 그룹(gate_40)으로 설정하여 A/B 테스트를 진행하였다.

분석의 주요 목적은 게이트 위치 변경이 플레이어의 재방문 행동, 즉 리텐션에 유의미한 영향을 미치는지를 검증하는 것이다. 두 그룹 간 리텐션 지표를 비교하고 통계적 검증을 통해 관측된 차이가 우연이 아닌 실제 효과인지 확인함으로써, 향후 게이트 위치 조정에 대한 합리적인 의사결정을 도출하고자 한다.

2. 데이터 설명

kaggle에서 제공하는 A/B 테스트 샘플 데이터를 활용하여 A/B 테스트를 진행하였다.

(<https://www.kaggle.com/datasets/arpitdw/cokie-cats?resource=download>)

데이터는 user_id, version, sum_gamerounds, retention_1, retention_7 총 5개의 컬럼으로 구성되어 있으며 각 컬럼의 의미는 다음과 같다.

컬럼	설명
user_id	각 사용자를 식별하는 아이디
version	게임 버전(Control 대조그룹(gate_30), Treatment 실험그룹(gate_40))
sum_gamerounds	총 플레이한 게임 라운드 합
retention_1	신규 사용자 중 첫 방문 이후 1일차에 재방문했는지 여부
retention_7	신규 사용자 중 첫 방문 이후 7일차에 재방문했는지 여부

본론

1. 데이터 EDA

1-1. 전체 데이터 분포 확인

```
##### Shape #####
(90189, 5)

##### Data Types #####
userid      int64
version     object
sum_gamerounds  int64
retention_1  bool
retention_7  bool
dtype: object

##### Null Values of Data #####
userid      0
version     0
sum_gamerounds  0
retention_1  0
retention_7  0
dtype: int64

##### Describe of the Numerical Datas #####
count      mean      std  min  25%   50%   75%   max
sum_gamerounds  90189.0  51.872457  195.050858  0.0  5.0  16.0  51.0  49854.0

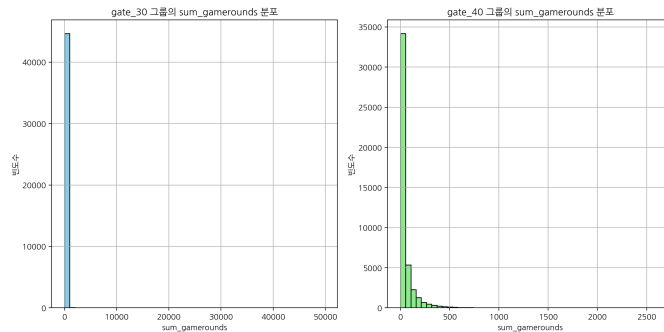
##### duplicated #####
0

##### nunique #####
90189
```

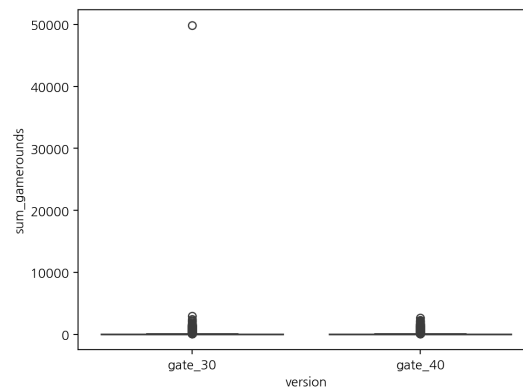
- sum_gamerounds 변수는 정규분포를 따르지 않을 가능성이 있다.
 - 중앙값 16인 반면 평균은 51.8로 평균이 중앙값보다 훨씬 큰 것을 알 수 있다.

- 대부분의 사용자는 적은 수의 게임 라운드를 플레이했지만 일부 사용자가 많은 수의 게임 라운드를 플레이하여 전체 평균을 끌어올렸을 가능성이 있다.
- 정규성 가정을 사용하는 t-test 대신 비모수 검정을 고려하는 것이 적절하다.

1-2. Control, Treatment 그룹의 sum_gamerounds 분포 확인

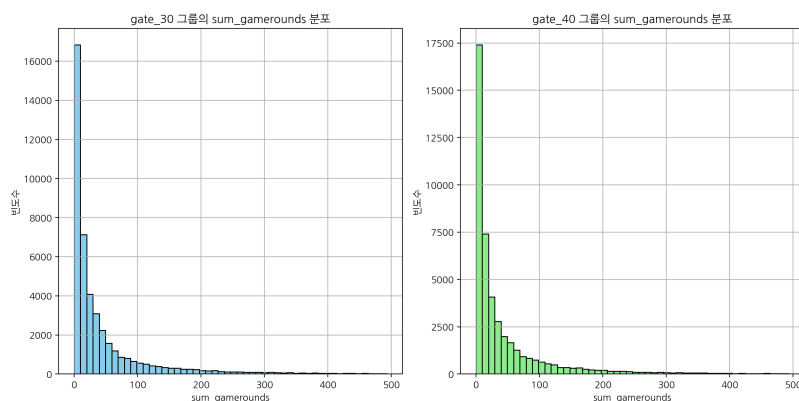


- 각 그래프의 sum_gamerounds(x축)를 확인했을 때, Control group의 sum_gamerounds는 50,000 이상이 존재하는 것을 알 수 있다.



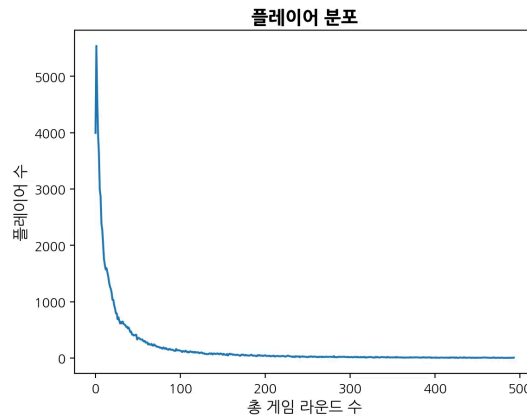
- 위의 시각화 자료를 통해 Control group의 sum_gamerounds에 극단값이 존재하는 것을 알 수 있다.
 - 눈에 띄는 이상치를 제거하고 박스플롯을 확인했으나 데이터가 0 근처에 밀집된 특성으로 인해 IQR 기준이 지나치게 좁게 설정되어 정상적인 값까지 제거되는 문제가 발생하였다. 이에 따라 IQR 기반 이상치 제거는 적절하지 않다고 판단하였다.
 - 최종적으로 Quantile Trimming 방법을 적용하여 0.01 이하 0.99 이상의 데이터들을 제거하면서도 실제 사용자 분포를 보다 안정적으로 반영할 수 있도록 하였다.

1-3. 극단값 제거 후 Control, Treatment 그룹의 sum_gamerounds 분포 확인



- Control group의 x축(sum_gamerounds축)이 50,000에서 500까지 줄어드는 것을 확인할 수 있다.
- Control, Treatment group 두 분포 모두 오른쪽으로 긴 꼬리를 가진 우측 비대칭 형태라는 것을 알 수 있다.

1-4. 유저 게임 이용 형태



- 사용자 분포는 초반에 매우 집중되는 것을 알 수 있다.
- 총 게임 라운드 수가 높아질수록 사용자 수는 급격히 감소함을 알 수 있다.
- 게임을 설치했음에도 한 번도 게임을 플레이하지 않은 유저들이 3994명 존재한다.

1-5. 게임 재방문율(리텐션율)

- 전체 리텐션율 확인

1-day retention ratio: 43.99%
7-days retention ratio: 17.83%

- 전체 플레이어 중 43.99%가 게임 설치 다음 날 접속하는 것을 알 수 있다.
- 전체 플레이어 중 17.83%가 게임 설치 다음 날 접속하는 것을 알 수 있다.
- Control, Treatment Group 나눠서 리텐션율 확인

	userid	retention_1	retention_7	sum_gamerounds
version				
gate_30	44254	0.442920	0.182537	1976494
gate_40	45037	0.436907	0.174190	1999480

- Treatment group의 리텐션율이 Control group의 리텐션율보다 더 낮을 것을 확인할 수 있다.
- 게이트의 위치 변화가 유저들에게 실제 영향을 줄 수 있다는 결론에 이른다.
- Control group을 유지하는 방향이 보다 합리적인 선택으로 판단된다.

2. 실험 설계

2-1. Mann-Whitney U Test

2-1-1. 가설 설정

- 귀무가설 : Control group과 Treatment group의 성과 지표 분포는 동일하다.
- 대립가설 : Control group과 Treatment group의 성과 지표 분포는 동일하지 않다.

2-1-2. 검정 방법

- Mann-Whitney U Test는 두 독립 집단의 데이터를 순위로 변환하여 분포 또는 중앙 위치의 차이를 검정하는 비모수적 방법으로 정규성 가정을 필요로 하지 않는다.

- 본 분석에서 사용한 성과 지표는 분포가 정규성을 만족하지 않거나 이상치의 영향을 받을 가능성이 있어, 평균 비교에 기반한 모수 검정보다 순위 기반의 비모수 검정이 적합하다고 판단하였다.
- 또한 Control group과 Treatment group은 서로 독립적인 집단으로 구성되어 있어, 두 집단 간 분포 차이를 검정하는데 Mann-Whitney U Test가 적절한 방법이라고 판단하였다.

2-1-3. 결과

- 결과 요약

Test Type	p-value	AB Hypothesis	Comment
Non-Parametric (Mann-Whitney U)	0.04789	Reject H0	A/B groups are different!

- Shapiro-Wilk 테스트 결과, 정규분포 가정이 기각되어 비모수 검정인 Mann-Whitney U 테스트를 사용했다.
- p-value = 0.04789
- 결과 해석
 - p-value가 0.05 보다 작으므로, 두 그룹의 분포가 동일하다는 귀무가설을 기각할 수 있다. 이는 게이트 위치 변경에 따라 사용자 행동 분포에 차이가 발생했음을 의미한다.

2-2. Bootstrapping

2-2-1. 가설 설정

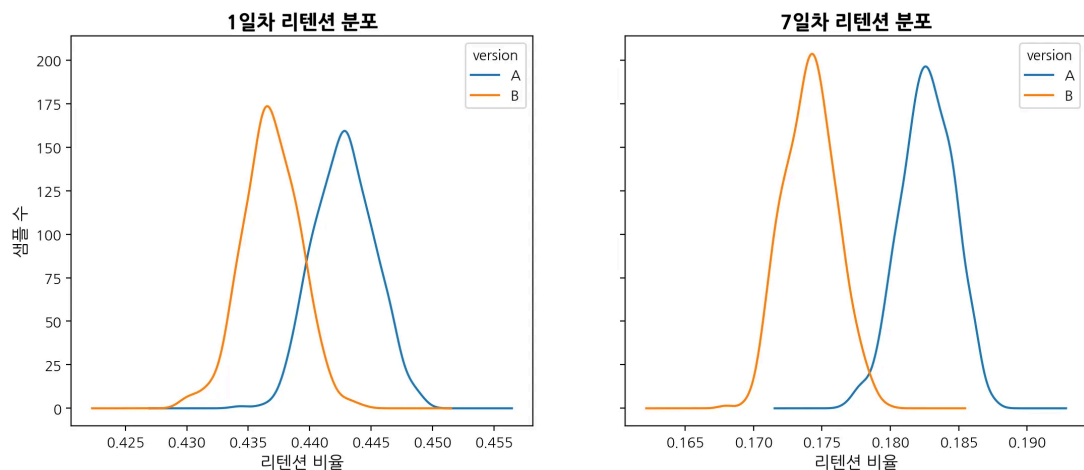
- 귀무가설 : Control group과 Treatment group의 전환율 차이는 0이다.
- 대립가설 : Control group과 Treatment group의 전환율 차이는 0이 아니다.

2-2-2. 검정 방법

- Bootstrapping 검정은 관측된 표본을 복원추출로 재추출하여 통계량의 분포를 추정하고, 이를 통해 가설 검정이나 신뢰구간을 구하는 비모수적 방법이다.
- 분포 가정이 불필요하고 표본 수가 적어도 사용 가능하며 복잡한 통계량에 적용 가능하다.
- 리텐션율 추정치의 불확실성을 부트스트래핑을 통해 평가하였다. 이를 통해 표본 데이터로부터 계산된 리텐션율이 가질 수 있는 변동 범위를 추정하였다.
- 전환율(또는 리텐션율)은 이항 분포 기반의 비율 지표로 정규성 가정을 충족하기 어렵고, 집단별 표본 수 차이 또는 분포의 비대칭성이 존재할 수 있어 모수적 검정보다 비모수적 접근이 적합하다고 판단하였다.
- 이에 따라 표본으로부터 전환율 차이의 경험적 분포를 직접 추정할 수 있는 Bootstrapping 방법을 사용하여 Control group과 Treatment group 간 전환율 차이가 0이라는 가설을 검정하였다.

2-2-3. 결과

- 결과 요약



- 두 그래프 모두 Control group의 분포가 Treatment group의 분포보다 오른쪽으로 이동해 있다.
- Control group의 retention이 더 높은 경향을 보인다.
- 결과 해석
 - 재표본화를 통해 생성된 분포에서 Control group의 리텐션 분포가 Treatment group보다 전반적으로 오른쪽에 위치하는 경향이 관측되었다.
 - 반복적인 샘플링 과정에서도 Control group이 상대적으로 높은 전환율을 보일 가능성이 크다는 것을 의미한다.
 - 따라서 두 그룹의 전환율 차이가 0이라는 귀무가설에 대해, 완전히 동일하다고 보기 어렵다는 방향의 근거를 제공한다.

2-3. 카이제곱

2-3-1. 가설 설정

- 귀무가설 : 게임 버전과 retention은 서로 독립이다. → 게임 버전과 관계 없이 retention은 거의 같다.
- 대립가설 : 게임 버전과 retention은 독립이 아니다. → 게임 버전에 따라 retention이 다르다.

2-3-2. 검정 방법

- 카이제곱 검정은 관측된 빈도와 기대되는 빈도의 차이를 비교하여 범주형 자료에서 통계적으로 유의한 차이나 관계가 있는지를 검정하는 방법이다.
- 독립성 검정은 게임 버전 변수가 retention에 영향을 주는지 검증하기 위해서 사용된다.
- 본 분석에서 retention은 유지 여부로 구분되는 범주형 변수이며, 게임 버전 또한 범주형 변수이므로 두 변수 간의 독립성 여부를 검정하기에 적합한 방법이라고 판단하였다.
- 또한 각 집단의 관측 빈도를 기반으로 게임 버전에 따라 retention 분포에 차이가 존재하는지를 확인하기 위해 카이제곱 독립성 검정을 적용하였다.

2-3-3. 결과

2-3-3-a. sum_gamerounds 0 값을 포함하는 실험

- 결과 요약

retention_1의 경우

```
카이제곱 통계량: 3.2511381280488836
p-value: 0.07137388196755491
자유도: 1
통계적으로 유의미한 차이가 없음
기대 빈도표
retention_1      0      1
gate_30      24787.215979  19466.784021
gate_40      25225.784021  19811.215979
```

- p-value: 0.07
 - 유의수준 0.05보다 크므로 게임 버전과 retention_1이 독립이라는 귀무가설을 기각할 수 없다.
- 카이제곱 통계량: 3.25
 - 자유도 1에서 고정값인 임계값 3.841보다 작아, 귀무가설을 기각할 수 없다.
 - 게임 버전에 따라 retention_1 비율에 통계적으로 유의한 연관성이 없음을 알 수 있다.

retention_7의 경우

```
카이제곱 통계량: 10.556773743285333
p-value: 0.001157630723934999
자유도: 1
통계적으로 유의미한 차이가 있음
기대 빈도표
retention_7      0      1
gate_30      36362.315037  7891.684963
gate_40      37005.684963  8031.315037
```

- p-value: 0.001
 - 유의수준 0.05보다 작으므로 게임 버전과 retention_7이 독립이라는 귀무가설을 기각한다.
- 카이제곱 통계량: 10.556
 - 자유도 1에서의 임계값 3.841보다 커 귀무가설을 기각한다.
 - 게임 버전에 따라 retention_7 비율에 통계적으로 유의한 연관성이 있음을 알 수 있다.
- 결과 해석
 - retention_1의 경우, 게임 버전과의 통계적으로 유의한 연관성이 확인되지 않았다.
 - retention_7의 경우, 게임 버전에 따라 사용자 행동 차이가 발생했음을 시사한다.

2-3-3-b. sum_gamerounds 0 값을 포함하는 실험

- 결과 요약

retention_1의 경우

```

카이제곱 통계량: 2.554644555016835
p-value: 0.10997054465977232
자유도: 1
기대 빈도표
[[22873.8127015 19443.1872985]
 [23232.1872985 19747.8127015]]
통계적으로 유의미한 차이가 없음
retention_1      0      1
gate_30      22873.812702 19443.187298
gate_40      23232.187298 19747.812702

```

- p-value: 0.110
 - 유의수준 0.05보다 크므로 게임 버전과 retention_1이 독립이라는 귀무가설을 기각할 수 없다.
- 카이제곱 통계량: 2.554
 - 자유도 1에서 임계값 3.841보다 작아 귀무가설을 기각할 수 없다.
 - 게임 버전에 따라 retention_1 비율에 통계적으로 유의한 연관성이 없음을 알 수 있다.

retention_7의 경우

```

카이제곱 통계량: 9.611011953238634
p-value: 0.001934140315170838
자유도: 1
통계적으로 유의미한 차이가 있음
기대 빈도표
retention_7      0      1
gate_30      34431.770766 7885.229234
gate_40      34971.229234 8008.770766

```

- p-value: 0.0019
 - 유의수준 0.05보다 작으므로 게임 버전과 retention_7이 독립이라는 귀무가설을 기각한다.
- 카이제곱 통계량: 9.611
 - 자유도 1에서 임계값 3.841보다 커 귀무가설을 기각한다.
 - 게임 버전에 따라 retention_7 비율에 통계적으로 유의한 연관성이 있음을 알 수 있다.
- 결과 해석
 - retention_1의 경우, 게임 버전과의 통계적으로 유의한 연관성이 확인되지 않았다.
 - retention_7의 경우, 게임 버전에 따라 사용자 행동 차이가 발생했음을 시사한다.

2-3-4. 전체 결과 해석

- sum_gamerounds의 값이 0을 포함, 불포함에 관계 없이 retention_7에서는 독립이 아니라는 결과가 나타난다.

2-4. Bootstrap을 이용한 신뢰구간

2-4-1. 가설 설정

- 귀무가설 : Control group과 Treatment group의 retention의 평균 차이는 0이다.
- 대립가설 : Control group과 Treatment group의 retention의 평균 차이는 0이 아니다.

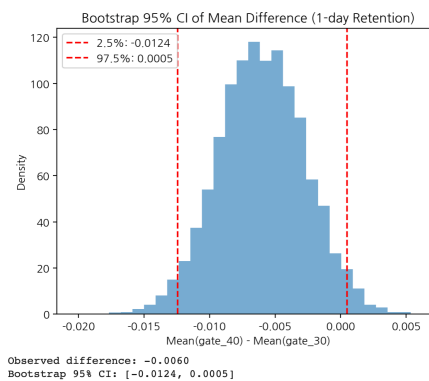
2-4-2. 검정 방법

- 본 분석에서는 다음과 같은 부트스트랩 절차를 사용하였다.
 1. 각 그룹(gate_30, gate_40)에서 원본 데이터 크기와 동일한 수의 표본을 복원추출을 수행한다.
 2. 재표본된 두 그룹의 평균 차이(gate_40 - gate_30)를 계산한다.
 3. 위 과정을 5,000회 반복하여 평균 차이의 부트스트랩 분포를 생성한다.
 4. 생성된 분포의 2.5% 및 97.5% 분위수를 이용해 평균 차이에 대한 95% 신뢰구간을 계산한다.
 - 신뢰구간에 0이 포함될 경우
 - 두 그룹 간 차이가 0일 가능성을 배제할 수 없다.
 - 귀무가설을 기각할 수 없다.
 - 신뢰구간에 0이 포함되지 않을 경우
 - 평균 차이가 0과 통계적으로 구분된다.
 - 귀무가설 기각 가능하다.
- retention 평균 차이에 대한 분포 가정을 하기 어려워, 표본 데이터를 반복 재추출하는 부트스트랩 방법을 통해 평균 차이의 신뢰구간을 추정하고 통계적 유의성을 판단하였다.

2-4-3. 결과

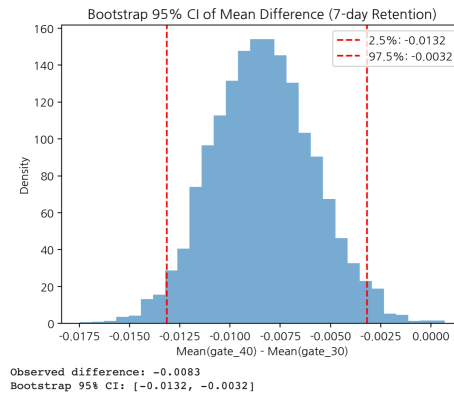
- 결과 요약

retention_1



- 재표본된 두 그룹의 평균 차이는 -0.0060이다.
- 신뢰 구간(빨간 점선) 사이에 0을 포함한다.

retention_7



- 재표본된 두 그룹의 평균 차이는 -0.0083이다.
- 신뢰 구간(빨간 점선) 사이에 0을 포함하지 않는다.
- 재표본된 두 그룹의 평균 차이(gate_40 - gate_30)를 계산한 값이 음수값(-0.0083)이기 때문에 Control group(gate_30)이 Treatment group(gate_40)보다 평균적으로 높은 retention을 가진다.
- 결과 해석
 - retention_1에서는 두 그룹의 retention 평균 차이가 같아 귀무가설을 기각하지 못한다.
 - retention_7에서는 두 그룹의 retention 평균 차이가 같지 않아 귀무가설을 기각한다.
 - retention_7에서는 Control group이 Treatment group보다 더 높은 retention을 가지고 있음을 알 수 있다.
 - Control group을 유지하는 방향이 보다 합리적인 선택으로 판단된다.

2-5. Cliff's Delta

2-5-1. 가설 설정

- 귀무가설 : Control, Treatment group의 분포가 동일하며 Cliff's Delta 값이 0이다.
- 대립가설 : Control, Treatment group의 분포가 동일하지 않아 Cliff's Delta 값이 0이 아니다.

2-5-2. 검정 방법

- Cliff's Delta는 두 독립 집단 간 차이의 크기와 방향을 비모수적으로 측정하는 효과 크기 지표이다.
- 해석 방식
 - Cliff's Delta의 값은 -1에서 1 사이의 범위를 가지며, 0에 가까울수록 두 집단 간 차이가 거의 없음을 의미한다. 양수일 경우 집단 A의 값이 집단 B보다 전반적으로 큰 경향이 있음을 뜻하고, 음수일 경우 그 반대를 의미한다. 절댓값이 1에 가까울수록 두 집단 간 차이가 매우 크다는 것을 의미한다.
- 본 분석에서는 통계적 유의성 여부뿐만 아니라 Control group과 Treatment group 간 차이가 실제로 어느 정도 크기 인지 그 방향성을 함께 파악하기 위해 Cliff's Delta를 사용하였다.
- 또한 성과 지표의 분포가 정규성을 만족하지 않을 가능성이 있어, 분포 가정에 의존하지 않고 집단 간 차이를 정량적으로 표현할 수 있는 비모수적 효과 크기 지표가 적합하다고 판단하였다.

2-5-3. 결과

- 결과 요약

Cliff's Delta: 0.0060
효과 크기 해석: negligible

- retention_1
 - Control, Treatment group에서 임의로 뽑은 두 사용자 간의 리텐션을 비교에서 Control group이 Treatment group보다 클 확률이 약 0.60% 더 높다.

Cliff's Delta: 0.0083
효과 크기 해석: negligible

- retention_7
 - Control, Treatment group에서 임의로 뽑은 두 사용자 간의 리텐션을 비교에서 Control group이 Treatment group보다 클 확률이 약 0.83% 더 높다.
- 결과 해석
 - Control group이 Treatment group 보다 좋은 결과를 보이는 것은 맞으나 Cliff's Delta 값이 0에 너무 가깝기 때문에 두 그룹의 차이는 무시 가능하다는 통계적 결과를 보인다.

2-6. two-proportion Z-test

2-6-1. 가설 설정

- 귀무가설 : Control, Treatment group의 retention 비율은 같다.
- 대립가설 : Control, Treatment group의 retention 비율은 다르다.

2-6-2. 검정 방법

- 독립 집단의 비율이 서로 같은지를 검정하는 통계적 방법이다.
- 두 집단의 모집단 비율 차이가 우연에 의한 것인지, 아니면 통계적으로 유의미한 차이인지 판단한다.
- retention은 이항 변수이며 두 집단이 독립적으로 구성되어 있으므로, 두 집단 간 비율 차이를 검정하기 위해 two-proportion Z-test를 사용하였다.

2-6-3. 결과

- 결과 요약

retention_1의 경우

Z 통계량: 1.8098335131034136
 p-value: 0.07032160914438722
 귀무가설 채택: 두 그룹의 리텐션율은 통계적으로 유사합니다.

- Z 통계량 : $1.80 < 1.96$
- p-value : $0.07 > 0.05$
 - 유의수준 0.05보다 크므로 귀무가설을 기각하지 못한다.

retention_7의 경우

Z 통계량: 3.25786180276105
 p-value: 0.001122550593303269
 귀무가설 기각: 두 그룹의 리텐션율은 유의하게 다릅니다.

- Z 통계량 : $3.25 > 1.96$
- p-value : $0.001 < 0.05$
 - 유의수준 0.05보다 작으므로 귀무가설을 기각한다.
- 결과 해석
 - retention_1의 경우 게이트 위치를 30레벨에서 40레벨로 변경하더라도 재방문 행동에는 뚜렷한 영향을 미치지 않았음을 의미한다.
 - retention_7의 경우 게이트 위치 변경이 사용자 재방문 행동에는 영향을 미쳤을 가능성이 있음을 의미한다.

2-7. 순열 검정

2-7-1. 가설 설정

- 귀무가설 : Control, Treatment group은 동일한 분포에서 추출되었으며, 관측된 통계량의 차이는 무작위적 라벨 할당으로 설명 가능하다.
- 대립가설 : Control, Treatment group은 동일한 분포에서 추출되지 않았으며, 관측된 통계량의 차이는 무작위적 라벨 할당으로 설명되지 않는다.

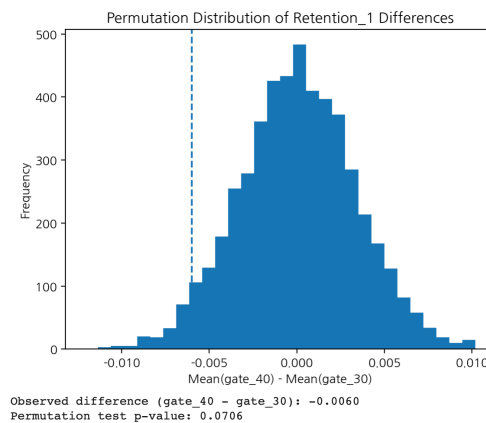
2-7-2. 검정 방법

- 두 개 이상의 표본을 결합한 후 샘플을 무작위로 resampling하여, 최종적으로 두 집단 간의 차이가 있는지 판단하는 검정 방법이다.
- 샘플 사이즈가 작은 두 그룹이 있을 때 서로 분포가 다른지를 판단할 때 사용하는 검정 방법이다.
- 본 분석에서는 Control group과 Treatment group의 표본 수가 제한적이거나 분포 가정이 어려운 상황에서, 두 집단 간 차이가 단순한 무작위 변동으로 설명 가능한지 확인하기 위해 순열 검정을 적용하였다.
- 이를 통해 관측된 통계량 차이가 무작위적 라벨 할당 하에서 얼마나 극단적인 값인지를 평가하고, 두 집단 간 분포 차이의 통계적 유의성을 판단하였다.

2-7-3. 결과

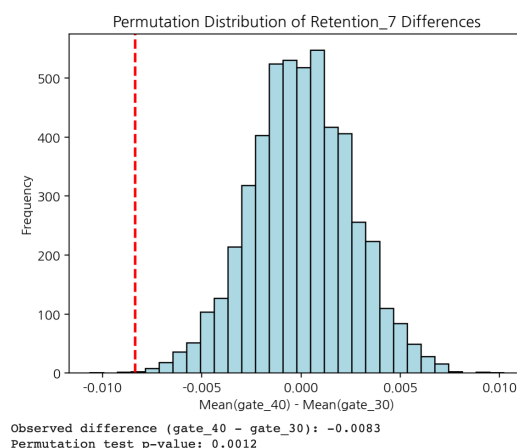
- 결과 요약

retention_1의 경우



- 관측된 평균 차이는 -0.0060이다.
- p-value = 0.0706
- 관측된 차이가 $\text{Mean}(\text{gate}_{40}) - \text{Mean}(\text{gate}_{30}) = -0.0083 \rightarrow$ Treatment group가 Control group보다 약 0.60%p 낮으므로 Control group의 retention이 높다.

retention_7의 경우



- 관측된 평균 차이는 -0.0083이다.
- p-value \approx 0.0012
- 관측된 차이가 $\text{Mean}(\text{gate}_{40}) - \text{Mean}(\text{gate}_{30}) = -0.0083 \rightarrow$ Treatment group가 Control group보다 약 0.83%p 낮으므로 Control group의 retention이 높다.

- 결과 해석
 - retention_1의 경우 p-value가 0.0706으로 유의수준 0.05를 넘어 통계적으로 유의미한 차이라고 보기는 어렵다. 해당 차이가 무작위적인 라벨 할당으로 충분히 설명 가능한 수준임을 의미한다.
 - retention_7의 경우 두 그룹이 동일한 분포에서 추출되었다고 보기 어렵다는 결론에 도달하였다. 이는 게이트 위치를 40레벨로 이동시킨 Treatment group에서 사용자 유지에 부정적인 영향이 발생했을 가능성을 시사한다.
 - Control group을 유지하는 방향이 보다 합리적인 선택으로 판단된다.

2-8. 순차 검정

2-8-1. 가설 설정

- 귀무가설 : Control, Treatment group 간 retention 차이는 0이다.
- 대립가설 : Control, Treatment group 간 retention의 차이는 0이 아니다.

2-8-2. 검정 방법

- 순차검정은 데이터를 한 번에 모두 모아 분석하는 대신, 데이터가 들어오는 과정에서 중간중간 검정을 수행하며 실험을 조기에 종료할 수 있도록 설계된 검정 방법이다.
- 표본 크기를 미리 고정하지 않고 결론이 충분히 명확해지는 시점에서 실험을 멈추는 것을 목표로 한다.
- **Alpha spending 방식**
 - 전체 유의수준 α 를 미리 정해두고, 이를 여러 번의 중간 분석에 나누어 사용하는 방식이다.
- 실험 진행 중에도 통계적 판단을 가능하게 하고, 제1종 오류를 통제하면서 효율적으로 실험을 종료하기 위해 순차검정을 사용하였다.

2-8-3. 결과

- 결과 요약

retention_1의 경우

	Look	Samples	p-value	Threshold	Stop Early
0	1	17858	0.0748	0.01	False
1	2	35716	0.0838	0.01	False
2	3	53574	0.1176	0.01	False
3	4	71432	0.0801	0.01	False
4	5	89291	0.0703	0.01	False

- 모든 Look 단계(1~5)에서 p-value가 유의수준 0.01 이상으로 조기 종료 조건이 충족되지 않았다.
- 최종 Look에서도 p-value가 0.0703으로 유의수준 0.01을 초과하여 통계적으로 유의하지 않다.

retention_7의 경우

	Look	Samples	p-value	Threshold	Stop Early
0	1	17858	0.2077	0.01	False
1	2	35716	0.0422	0.01	False
2	3	53574	0.0161	0.01	False
3	4	71432	0.0018	0.01	True
4	5	89291	0.0011	0.01	True

- Look 1~3에서는 p-value가 모두 0.01 이상으로 나타나 통계적으로 유의하지 않았으며, 데이터 수집을 계속 진행 한다.
- Look 4에서 p-value가 0.0018로 유의수준 0.01 미만으로 떨어져 조기 종료 조건이 충족된다.
- Look 5에서도 p-value가 0.0011로 더 낮아져, 유의성이 유지되며 결과의 일관성이 강화된다.

- 결과 해석
 - retention_1의 경우, 모든 중간 분석(Look 1~5)에서 관측된 p-value가 사전에 설정한 유의수준 0.01보다 높아 귀무가설을 기각할 수 없다.

- retention_7의 경우, 데이터가 추가로 누적되면서 Look 4 단계에서 p-value가 유의수준 0.01 미만으로 하락하였다. 해당 시점부터 귀무가설을 기각할 수 있다. Look 5에서도 유의성이 유지됨에 따라, 관측된 차이가 일관된 효과임을 추가적으로 확인할 수 있다.
- Control group을 유지하는 방향이 보다 합리적인 선택으로 판단된다.

2-9. 베이지안 A/B Test

2-9-1. 목표 설정

- Treatment group의 retention이 Control group보다 높을 확률을 추정한다.

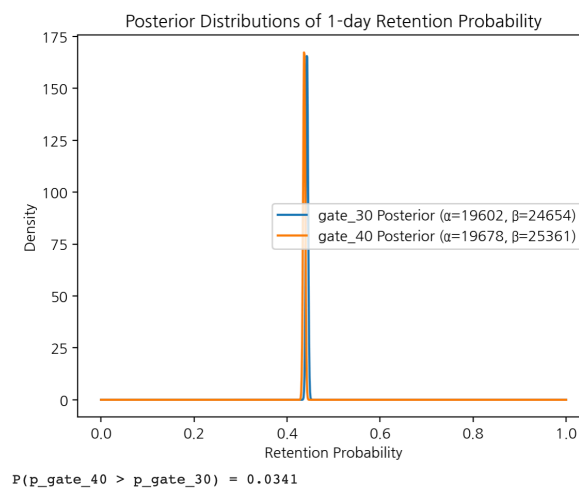
2-9-2. 검정 방법

- 사후분포에서 각 그룹의 리텐션 확률을 다수 샘플링한 뒤 다음 확률을 계산한다.
- $P(p_{\text{gate_40}} > p_{\text{gate_30}})$ 은 gate_40의 리텐션 확률이 gate_30보다 클 확률을 의미한다. 샘플링 기반 몬테카를로 방법을 사용하여 해당 확률을 근사적으로 추정한다.
- 본 분석에서는 단순한 유의성 판단을 넘어 Treatment group이 실제로 더 나은 성과를 보일 가능성을 확률적으로 해석하기 위해 베이지안 A/B Test를 적용하였다.
- 이를 통해 실험 결과를 의사결정 관점에서 직관적으로 해석하고, 두 그룹 간 리텐션 차이에 대한 불확실성을 함께 고려하고자 하였다.

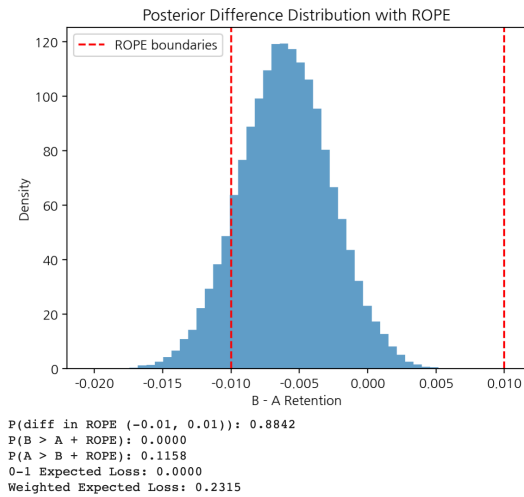
2-9-3. 결과

- 결과 요약

retention_1의 경우

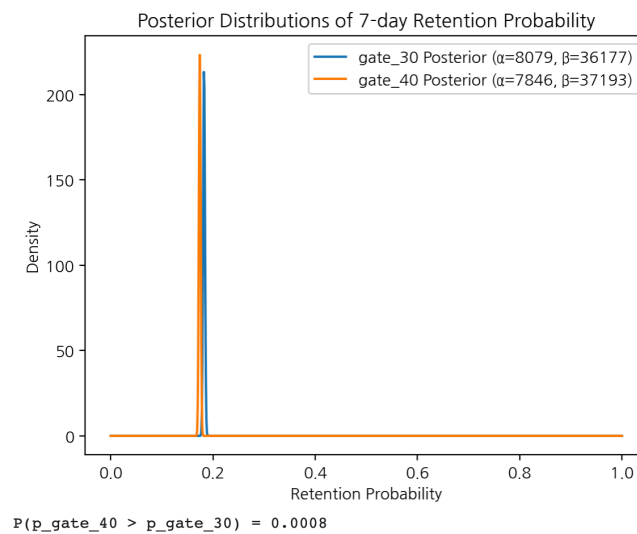


- gate_40이 더 좋을 확률 약 3.4%로 gate_30이 더 좋을 확률은 약 96.6%이다.
- 두 분포는 거의 겹치지만 중심이 약간 다르며, gate_30이 우측에 위치해있다.

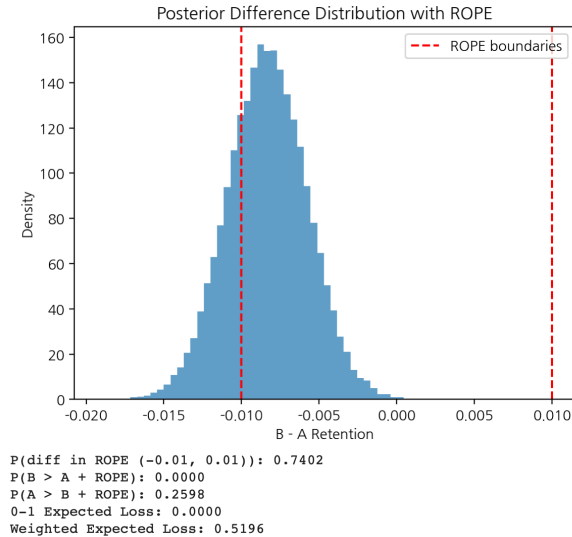


- $P(\text{diff in ROPE}) = 0.8842$
 - 두 그룹 차이가 $\pm 1\%$ 이내일 확률 약 88.4%이다.
- $P(B > A + \text{ROPE}) = 0.0000$
 - B가 A보다 의미 있게 높을 확률 0%이다.
- $P(A > B + \text{ROPE}) = 0.1158$
 - A가 B보다 의미 있게 높을 확률 약 11.5%이다.

retention_7의 경우



- gate_40이 더 좋을 확률 약 0.08%로 gate_30이 더 좋을 확률은 99.92%이다.
- 두 분포는 거의 겹치지만, gate_30이 우측에 위치해있다.



- $P(\text{diff in ROPE}) = 0.7402$
 - 두 그룹 차이가 $\pm 1\%$ 이내일 확률 약 74.0%이다.
- $P(B > A + \text{ROPE}) = 0.0000$
 - B가 A보다 의미 있게 높을 확률 0%이다.
- $P(A > B + \text{ROPE}) = 0.2598$
 - A가 B보다 의미 있게 높을 확률 약 25.9%이다.
- 결과 해석
 - retention_1의 경우, gate를 30에서 40으로 옮기면 리텐션이 약간 감소할 가능성이 있으나 실질적으로 리텐션 차이는 미미하다.
 - retention_7의 경우, gate_30이 통계적으로 우세하며 gate를 40으로 변경 시 장기 유지율이 유의하게 감소하나 전체적으로는 큰 차이는 아니다.
 - retention_1과 retention_7 분포 대부분이 ROPE 안에 포함되므로 두 그룹 리텐션 차이 크지 않는다.

결론

1. 분석 목적 요약

- 게이트 위치를 30레벨(gate_30) → 40레벨(gate_40)로 변경했을 때 리텐션(retention_1, retention_7)에 유의미한 변화가 있는지 검증한다.
- 통계적 검증을 통해 우연에 의한 차이인지 실제 효과인지 판단하여 의사결정 도출한다.

2. 핵심 결과 요약

- retention_1
 - 대부분의 검정에서 통계적으로 유의하지 않다.
 - 결론: 게이트 위치 변경이 단기 재방문에는 뚜렷한 영향이 없다.(귀무가설 기각 불가)
- retention_7
 - 다수의 검정에서 통계적으로 유의한 차이가 존재한다.
 - 방향성: gate_40의 retention_7이 gate_30보다 낮다.
 - 결론: 게이트를 40으로 옮길 경우 장기 재방문이 악화될 가능성이 크다.(귀무가설 기각)

3. 검정 방법별 결론 일관성

- 카이제곱 / two-proportion Z-test / 순열 검정 / 순차 검정에서 retention_7은 일관되게 유의, retention_1은 일관되게 비유의하다.

- 부트스트랩 신뢰구간: retention_7은 0 미포함 → 차이가 존재함을 시사한다.
- 베이지안 A/B: gate_40 우세 확률 매우 낮음, ROPE 내 비중은 크지만 방향은 gate_30 우세하다.
- 효과 크기(Cliff's Delta): 차이 규모는 크지 않을 수 있으나, 방향성과 유의성은 반복적으로 확인된다.

4. 최종 의사결정

- retention_1만 보면 큰 차이가 없어 보이나, retention_7에서 gate_40이 불리하므로 gate_30 유지가 더 합리적이라고 판단된다.