그로스해킹 101

6-3. A/B 테스트

A/B 테스트

A/B 테스트란?

- 집단 간 대조 실험 (2-표본 가설 검정)
- 통계적 가설 검정
- 변수 사이의 인과관계를 확인하기 위한 사회과학 실험 방법론

들어가기 전에

- 누구나 잘 알고 있다고 생각하고 그거 빨간색 버튼이랑 파란색 버튼 보여주고 뭐 많이 클릭하는지 보는거잖아?
- 누구나 잘 할 수 있다고 생각하지만 빨간색 버튼을 더 많이 눌렀으면, 그게 더 좋은거니깐 서비스 반영하면 되잖아?
- A/B 테스트를 잘 진행하려면 생각보다 고려해야 할 것들이 많다;;;

A/B 테스트 설계

- 가설 – 실험을 통해 무엇을 확인하고 싶은지. 구체적으로는 독립변수와 종속변수 식별 + 종속변수의 목표수준
- 실험집단 / 통제집단 – 실험군을 어떤 기준으로 구분하며, 어떤 비율로 할당할 것인지
- 독립변수 – 종속변수에 영향을 줄 거라고 기대되는 변수 + 각 케이스별 variation에 대한 정의
- 종속변수 – 실험의 성과를 측정할 때 사용하는 변수 + 어떻게 측정할 것인지에 대한 operational definition
- 통제변수 – 실험 결과에 영향을 미칠 수 있기 때문에, 실험집단/통제집단 모두에서 동등한 조건을 가져야 하는 변수

- 종속변수의 현재 수준과 목표 수준 – 현재 어떤 수치이고, 어느정도의 성과를 기대하는지
- Sample Size – 가설 검증에 필요한 실험 참가자의 숫자 (실험 전에 미리 정해야 함)
- 실험 기간 – Sample Size를 고려했을 때, 가설 검증을 위한 데이터를 수집하는데 필요한 기간

A/B 테스트 설계 시 고려사항

설계의 성패는 통제변수 관리와 실험집단/통제집단 샘플링

- 순차 테스트는 A/B 테스트가 아니다
  - 순차 테스트의 문제는 샘플링 오류를 발생시킬 수 있다는 것 + 기대하지 못한 외부 효과
  - A/B 테스트를 안하는 것보다는 낫다는 의견도 있으나… 음… 이 경우 샘플링 오류에서 정말 자유로운가를 굉장히 꼼꼼하게 검증 (A-B-A 테스트)
  - 반대로 말하면, 동시 테스트라고 하더라도 샘플링을 충분히 고민하지 않으면 잘못된 실험 설계가 될 수 있음

- 샘플링은 홀/짝 구분이 진리?
  - 랜덤 추출(random sampling)과 편의 추출(convenient sampling)
  - 랜덤 추출 - 통제변수가 잘 관리된 상태에서의 무작위 추출
  - 실험 전, 후로 A/A 테스트를 진행하는 것도 좋은 방법

- 테스트 유형에 따른 분석방법 구분
  - 종속변수가 범주형 (ex. 클릭여부, 가입여부) – 로지스틱 회귀, 카이제곱 검정
  - 종속변수가 이산형 (ex. 클릭횟수, 결제금액) – T검증, 분산분석

A/B 테스트 분석

- 가설을 검증하려면 어느 정도의 숫자가 필요한가 (sample size) – 샘플사이즈 계산기
  - http://www.evanmiller.org/ab-testing/sample-size.html
  - https://docs.adobe.com/content/target-microsite/testcalculator.html

- 효과를 어떻게 판단할 것인가?
  - 기본적으로는 분포와 신뢰구간, 효과 크기(effect size)를 기준으로 판단해야 함
    - P value
    - 분포
    - 신뢰구간
    - 효과크기
  - P값만 보고 단편적으로 판단하면 안되지만, 그렇다고 P값을 무시해서도 안됨

A/B 테스트 분석

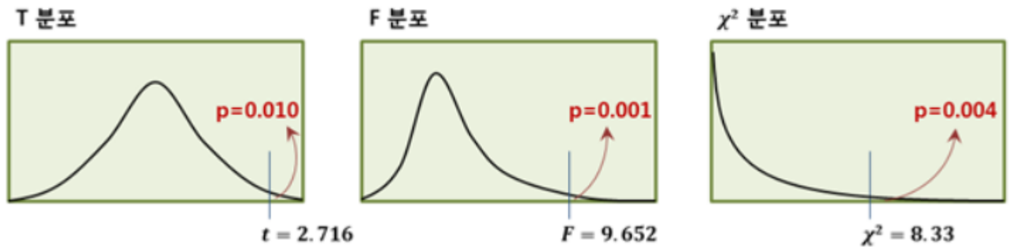
- 통계적으로 유의미하다… 는 말의 의미가 뭘까?
- 95% 신뢰수준에서 A의 클릭율이 B의 클릭율보다 유의미하게 높다?!
  - A의 클릭율이 B의 클릭율보다 높을 확률이 95%이다 (X – 이렇게 해석하면 안됨)

- 통계학에서 가설을 검증하는 방식 (빈도주의 사례)
  - 우리가 검증하고 싶은 건, A의 클릭율이 B의 클릭율보다 높다! 혹은 낮다!
  - 하지만, 통계학에서는 이런 식의 검증이 불가능함
  - A와 B의 클릭율 차이가 없다 → 차이가 없는데, 이렇게 극단적인 값이 관찰될 확률은 5% 미만 → 그럼 차이가 있는 거네(!)

귀무가설(영가설)

검정통계량
(T통계량, F통계량…)

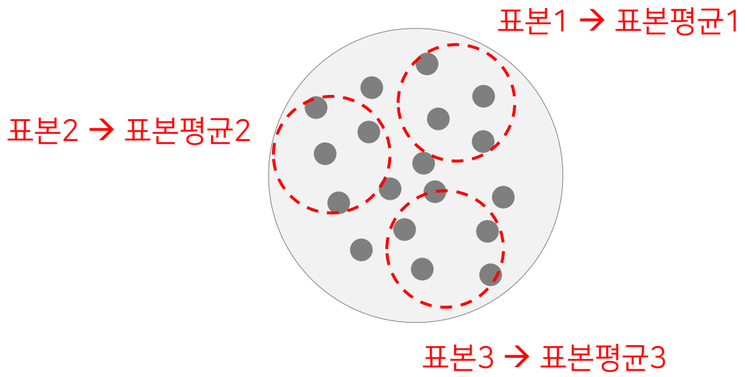
P < 0.05

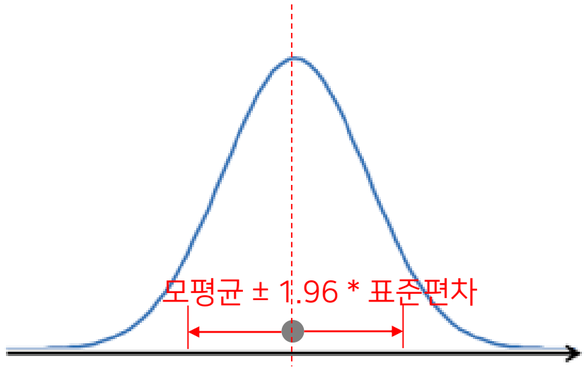


https://dermabae.tistory.com/145

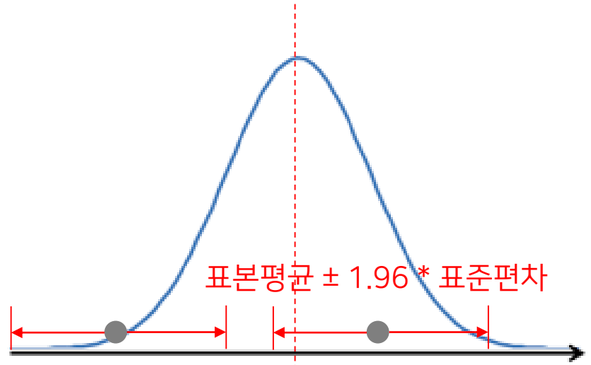
A/B 테스트 분석

- 신뢰구간

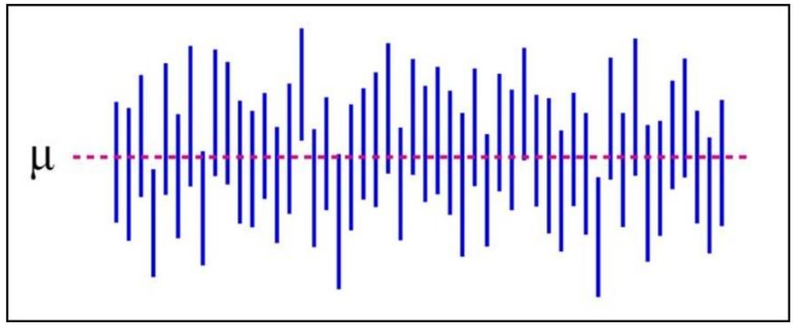




표본 평균들의 분포
(N이 충분히 크면 정규분포 → 중심극한정리)



모평균을 알 수 없으므로
거꾸로, 표본평균을 통해 모평균을 추정

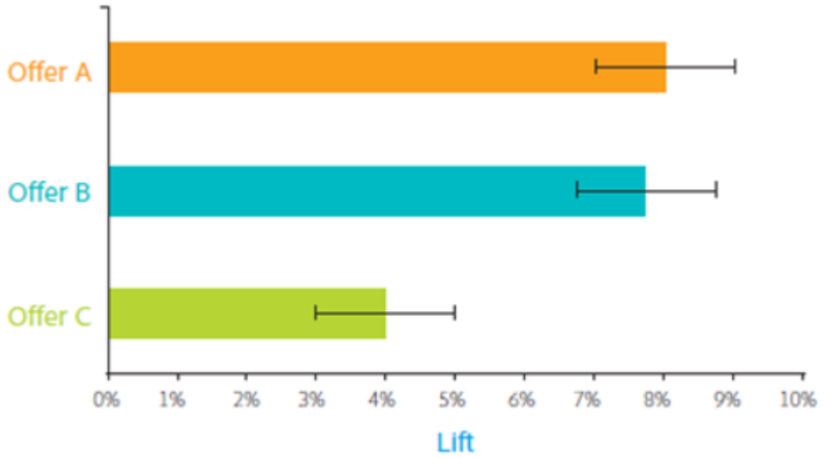


신뢰구간
- 모수가 어느 범위 안에 있는지를 확률적으로 보여주는 방법
- 95% 신뢰구간의 개념

→ 반복적으로 표본 추출을 100회 했을 때 모평균을 포함한 신뢰구간이 95개 나올 수 있다

A/B 테스트 분석

- 신뢰구간

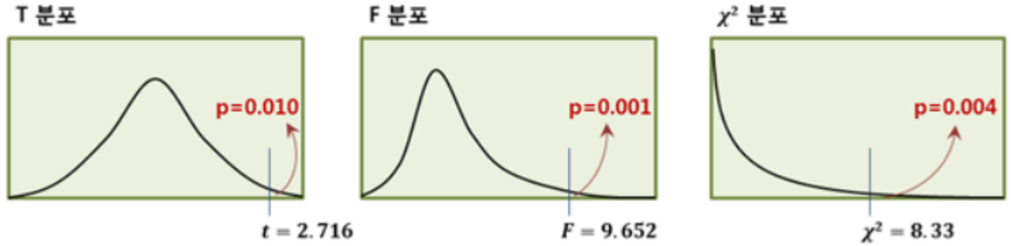


Source: Adobe

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}} \qquad F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/n-k-1} \qquad \chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

차이가 없는데, 이렇게 극단적인 값이 관찰될 확률은 5% 미만

검정통계량
(T통계량, F통계량…)



→ 그럼 차이가 있는 거네(!)

A/B 테스트 분석

- A/B test Calculator
  - 종속변수가 범주형 (ex. 클릭여부, 가입여부) – 로지스틱 회귀, 카이제곱 검정
    - https://www.evanmiller.org/ab-testing/chi-squared.html
    - http://www.abtestcalculator.com/
  - 종속변수가 이산형 (ex. 클릭횟수, 결제금액) – T검증, 분산분석
    - https://www.evanmiller.org/ab-testing/t-test.html
    - https://mathcracker.com/t-test-for-two-means

A/B 테스트 분석

- 효과크기
  - 조건 A에서 접속→구매 전환율 0.1% 상승
  - 조건 B에서 접속→구매 전환율 0.15% 상승
  - p < .01 즉, 99% 수준에서 통계적으로 유의미
  - 이 실험의 가치는??

- DAU **1000명**, ARPPU 10000원

→ B의 경우 **5만원/일** 의 추가 매출 발생

- DAU **100만명**, ARPPU 10000원

→ B의 경우 **5000만원/일** 의 추가 매출 발생

- 배너 1은 구매전환율 10%
- 배너 2는 구매전환율 20%
- p< .01 즉, 99% 수준에서 통계적으로 유의미
- 이 실험의 가치는?

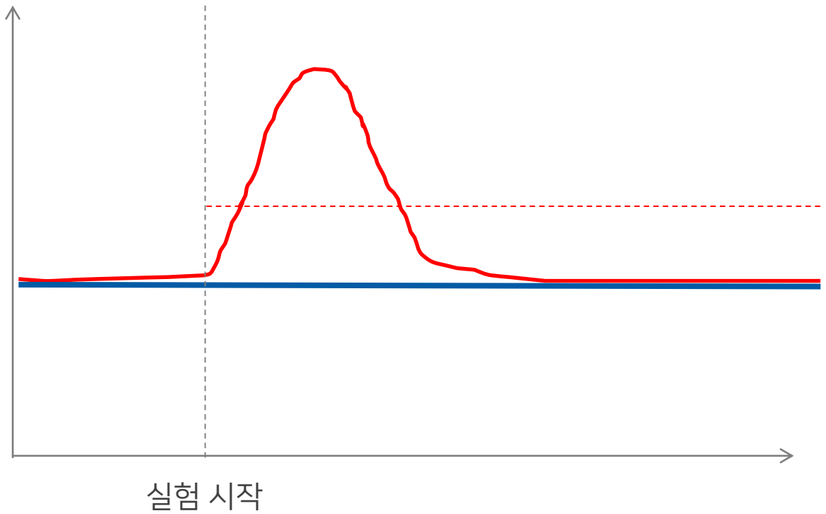
- 배너 1 상품의 profit은 10000원
- 배너 2 상품의 profit은 1000원 이라면?

A/B 테스트 참고사항

- 대표적으로 하는 A/B 테스트의 실수
  - 무가설
  - 통제변수 관리 실패
  - 단순 평균 비교
  - 엿보기 + 조기 중지
  - Delayed conversion 무시
  - A/B 테스트의 결과가 비즈니스 목표와 align 되지 않는 것

A/B 테스트 참고사항

- 시간의 흐름에 따른 차이를 살펴봐야 함
  - A/B 테스트 결과는 시간에 따라 변화하는 일이 자주 발생함
  - 새로운 기능이 나오면, 새 기능을 일단 써보는 유저가 있어서 전환율과 p-value에 영향을 줌
  - 시간의 흐름에 따른 추이 변화, 혹은 특정 브라우저 버그 / 기능 오류 등 외부 요인이 없었는지 재차 확인 필요



A/B 테스트 참고사항

- A/B 테스트의 결과는 언제까지 유효한가?
  - 잘 설계되어서 의미있는 결과가 나왔다고 해도 '앞으로도 계속 그 결과가 유효할 것이다' 라고 보장할 수 없다.
  - 계절 변화, 시장상황 변화, 사용자층 변화, 취향 변화 등 시간의 흐름에 따라 달라질 수 있다.

- 국지적 최적화의 함정
  - A/B 테스트는 주어진 조건에서의 최선의 찾는 문제
  - 애초에 조건 자체가 최선이 아니었다면, 결과로 찾은 안의 임팩트도 크지 않다

A/B 테스트 참고사항

- 실험 기간동안 보는 손해는 어쩌지?
  - A조건이 B조건보다 30% 더 좋은 성과가 난다면… B 조건에 할당한 사용자들은 낭비되는 비용?
  - 크리스마스 시즌 할인이벤트를 A/B테스트 하고 싶은데, A/B테스트 끝나면… 크리스마스도 끝나는데?

- Multi-Armed Bandit