programmers

A/B Test 관련 통계 살펴보기

기본 통계 리뷰

A/B 테스트 트래픽 비교

실습

Grepp Inc

한기용

keeyonghan@hotmail.com

programmers

Contents

programmers

질문 리뷰

몇 개 질문들을 리뷰해보자

질문 리뷰 (1)

- A/B 테스트에서 70프로의 비용이 사용된다고 강의를 통해서 들었습니다. 이 70프로 비용중에서는 단순 SaaS에 대한 비용일까요? 인력에 대한 비용도 포함일까요?, 또한 SaaS를 직접구현 한다고 했을시 해당 비용은 몇 프로로 줄어 들까요?
- A/B를 나눌때 LEFT를 하면 버려지는 부분이 나눌때 버려지는 수가 핵심 숫자로 생각한다면 중복이 될것 같은데 이는 50:50의 영향에 적기때문에 무시되는걸까요?
- 팀에서 데이터 문해력을 높이기 위해 실행했던 프로세스, 문화 등 활동이 궁금합니다!

질문 리뷰 (2)

- ab test 결과를 얻기 위해서 초반에는 1% 유저들만 ab test 프로세스를 태우고 이후 점차 테스트하는 유저 수를 5%까지도 늘린다고 들었습니다. 이 경우 5% 에서 유의미한 결과가 나오면 테스트를 멈추고 해당 개선방안으로 의사결정을 하는지, 그렇지 않다면 5% -> 30% -> 100% 까지 점차 늘리는 방향으로 ab test에 전체 유저들을 태우려고 하는 지 궁금합니다. (물론 후자는 안할 것 같지만. 정확하게 알고 싶어서 질문드립니다!)
- 또한, ab test를 진행한 유저들 3%에서는 유의한 결과차이가 보였지만 5%으로 늘렸을 때 다시 그 유의한 차이가 줄었다면 이후 기간을 늘려서 더 테스트를 진행하는 지, 데이터 양을 늘려서 다시 한번 확인하는 지 등이 궁금합니다!
- A, B 테스트에서 A 유저와 B 유저의 비율을 비슷하게 맞추는 것이 중요하다고 하셨습니다. 여기서 궁금한 것이 만일 A 집단의 유저 수가 B 집단의 유저 수보다 더 많다면 A집단을 가장 잘 대표할 수 있는 샘플링 방법을 써서 B 집단과 비율을 맞출 수도 있을 것 같은데 혹시 써도 되는 것인지, 불가능하다면 정확도 때문일까요?

질문 리뷰 (3)

- A/B 테스트는 대고객 상대로 하는 분석 기법으로 이해했는데요, 대고객 서비스를 맡은 조직이 아니라 사내 서비스를 맡은 조직 또는 순수 데이터엔지니어 조직은 A/B 테스트를 할 일이 없는 것인가요?
- 가설과 일정 사이즈의 데이터가 있어야 한다는 점 때문에 완전 초창기 스타트업 같은 회사에서는 A/B 테스트 도입 자체가 힘든 구조가 될것 같고, 데이터팀이 있는 스타트업 정도 부터 할 수 있을것 같습니다. 제대로 이해한게 맞을까요?
- A/B 테스트 미팅에서 괜찮을 수 있는 아이디어가 기각되는 경우도 많을 것 같습니다. A/B 테스트를 시도해볼만한 아이디어인지 아닌지 판단하는 기준이 있을까요?
- 유데미에서 마케터 중심의 규칙 기반 추천을 머신러닝 방식으로 변경했다는 사례를 소개해주셨는데요, 머신러닝 방식으로 추천 방식을 변경한 뒤에 마케터들은 어떻게 됐나요??

programmers

지난 강의 리뷰

지난 강의 내용 일부를 리뷰해보자

좀더 실전적인 사용자 버킷팅 코드 (1)

- 이전 코드

```
def split_userid(id, num_of_variants=2):
   h = hashlib.md5(str(id).encode())
   return int(h.hexdigest(), 16) % num_of_variants
```

- 문제점
  - 모든 사용자들을 대상으로 50:50으로 나눔
    - 점진적인 커버리지를 늘려가며 사용자 버킷팅이 필요 (1% -> 5% …)
  - AB 테스트에 관계없이 사용자들은 항상 A에 들어가거나 B에 들어감
    - 본의 아니게 bias가 생김

좀더 실전적인 사용자 버킷팅 코드 (2)

- 개선된 코드
  - 점진적인 커버리지 증대 지원: size_of_test 파라미터 추가 (1, 5, 10, 50, 100)
    - AB 테스트에 포함되지 않는 경우 -1을 리턴
  - md5으로 새로운 숫자를 만들어낼 때 abtest_id도 추가

```
def split_userid(abtest_id, user_id, size_of_test, num_of_variants=2):
    id = user_id + abtest_id
    h = hashlib.md5(str(id).encode())
    if (int(h.hexdigest(), 16) % 100) < size_of_test:
        return int(h.hexdigest(), 16) % num_of_variants
    else
        -1
```

AB 테스트들간의 Interaction이란?

- 예를 들어 4명의 사용자를 대상으로 2개의 테스트를 동시에 진행한다고 가정
  - 아래 예처럼 지정된다면 A1-A2와 B1-B2 조합만 테스트됨
    - A1-B2, B1-A2 조합은 테스트가 되지 않음
  - 즉 두 테스트들이 독립적으로 테스트되는 것이 아님
    - 조합에 따라 결과가 달라짐
    - 후처리를 통해 영향을 살펴보는 것이 필요 (아니면 버킷팅을 미리 하는 형태로 가야함)

User ID	Test 1 (A1, B1)	Test 2 (A2, B2)
1	A1	A2
2	A1	A2
3	B1	B2
4	B1	B2

A/B 테스트 Variant 크기 비교

- A/B 테스트에서 기본은 먼저 Variant간의 트래픽 크기가 **동일한지** 살펴보는 것
  - 여기서 동일하다는 의미는 통계적으로 차이가 오차 범위안에 있는지를 의미
    - 보통 95% 신뢰구간을 사용
- 이는 binomial distribution(이항분포)에 해당 (A 혹은 B)
  - 동전 던지기에서 앞면과 뒷면의 분포와 동일

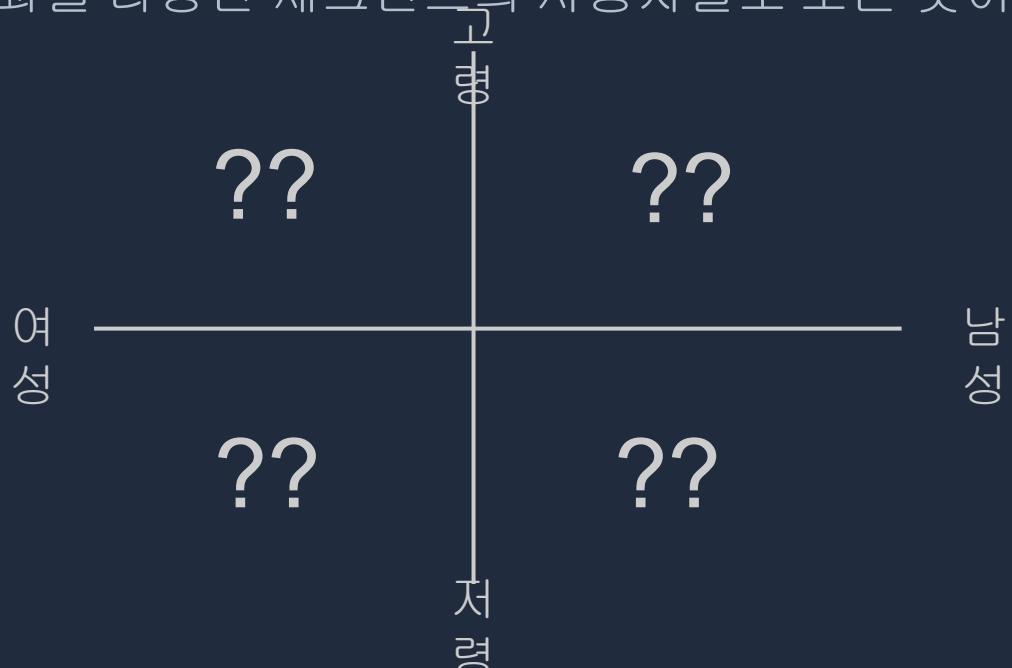
	A (Control)	B (Test)
User Size(*)	50	53
Impressions(**)	120	115
Clicks	10	15
Converted	3	1
Revenue	1	3

A/B 테스트 QA의 중요성

- 많은 A/B 테스트는 개발자 혹은 디자이너의 손을 타고 구현이 됨
- 즉 그 과정에서 A/B 테스트 제안자의 생각이 제대로 반영되지 않을 수 있음
  - 또한 코딩이 필요하다면 항상 버그의 존재 가능성이 있음
  - 이를 제대로 QA하는 것은 A/B 테스트 제안자의 책임이기도 함
- 특히 A/B 테스트 시스템 자체가 새로 만들어졌거나 변경이 생긴 경우 QA는 더 중요해짐
- 주기적으로 A/A 테스트를 수행해보는 것이 중요함

새로운 기능의 점진적 커버리지 확대

- 만일 A/B 테스트 결과가 전체적으로는 안 좋지만 특정 사용자층에 대해서만 좋다면?
- A/B 테스트 결과를 다양한 세그먼트의 사용자별로 보는 것이 중요



A/B 테스트 성공의 경우 지표는 항상 좋아야 하나?

- 테스트에 따라서는 지표가 개선되지 않아도 성공으로 간주할 수 있음
  - 예를 들어 운영비용이 적은 방식이라던지 레거시 방식을 새로운 방식으로 교체하는 경우 지표가 같은 경우도 성공으로 간주 가능
- 다시 한번 이런 부분이 가설에 분명하게 미리 명시되어 있어야함

모바일 앱 기반 A/B 테스트의 난점

- 앱단의 시간을 그대로 믿고 사용하면 낭패
  - 이는 특히 앱이 온라인이 아닌 상황에서 동작가능하는 경우 더 큰 문제가 됨
  - 앱에서 따라서는 사용자의 오프라인 이벤트를 사용자가 온라인이 되었을 때 전송받음
  - 이 경우 사용자의 과거 데이터가 AB 테스트 시작 이후에 전송될 수 있음
- 앱 업데이트가 필요하다면 테스트에 충분한 수의 앱 업데이트를 하는데 시간이 걸림

안 좋은 AB 테스트 가설 예는 무엇이 있나요?

- 잘못된 가설을 바탕으로 하는 경우
  - CS팀 문의자들의 Churn rate을 낮추겠다
- 굉장히 임팩트가 작은 기능 구현에 사용하는 경우
  - 즉 가설이 중요한 문제를 다루지 않는 경우
- B2C 환경에서 다른 가격대를 테스트하는 경우
  - 컨트롤하기가 쉽지 않음 (유데미)
- 검색엔진 UI와 검색엔진 알고리즘을 동시에 테스트하는 경우
  - 한번에 하나만 테스트 (야후 차이나)

A/B 테스트 얼마나 기다려야 하나? (1)

- 기본적으로는 적어도 일주일은 테스트를 돌려야함
  - 많은 테스트들이 처음에는 통계적 유의미한 차이를 초반에 보이다가 시간이 지나면서 없어짐
    - Data peeking problem
  - 웹 스케일에서는 z-test등에 전통 통계학에서 사용하는 최소 샘플의 크기가 아무 의미가 없음 (시대적 변화)

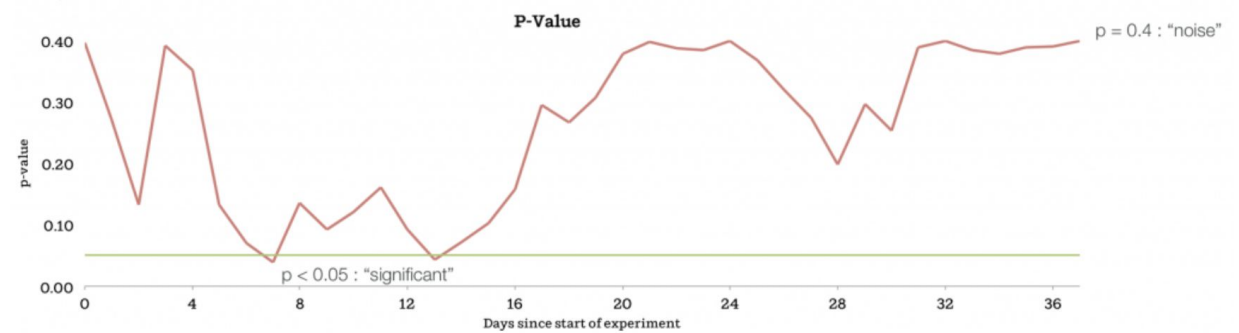


Figure 5 — Result of the price filter experiment over time.

A/B 테스트 얼마나 기다려야 하나? (2)

- 이는 최소 샘플의 크기와도 연관됨
  - 가설의 일부로 원하는 지표의 성공실패 기준에 따라 최소 트래픽 기준을 미리 설정하는 것이 좋음
  - https://www.evanmiller.org/ab-testing/sample-size.html
- 사용자들이 익숙한 UI/UX의 변경은 최소 2-4주를 실행해야함

programmers

기본 통계 리뷰와 실습

A/B 테스트 분석에 사용되는 기본 통계를 학습해보자

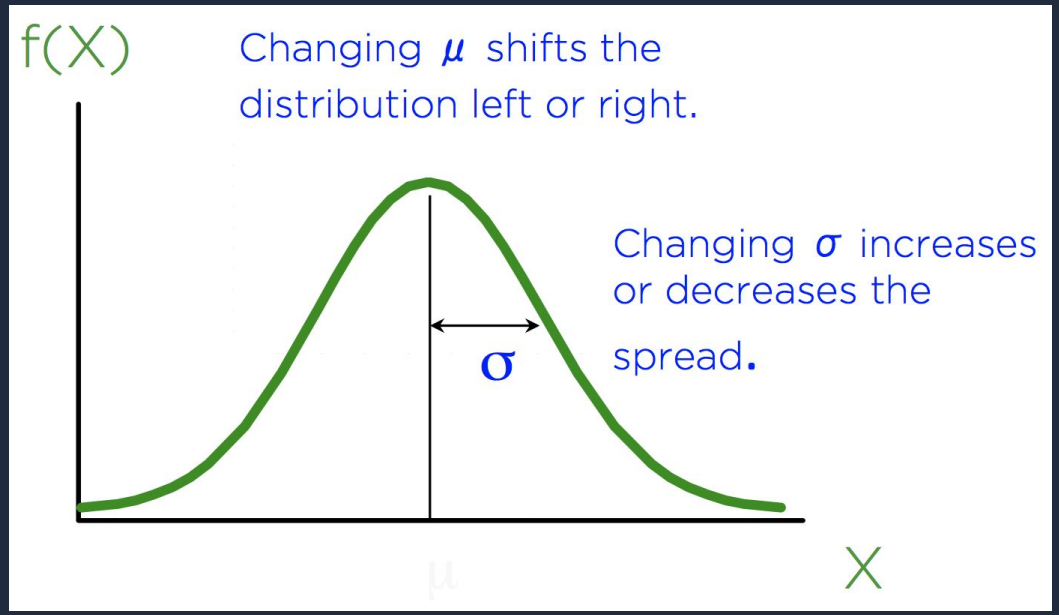
귀무가설, 정규분포, 중심극한정리, Z-test, T-test, ...

Which Side is Better in A/B Test? (1)

- Basically NULL Hypothesis (귀무 가설)
  - Assumes A and B are the same
    - 예) A와 B의 구매율은 동일하다
- CLT를 사용하면 데이터 분포를 정규분포로 변환 가능
- B-A를 계산하여 동일한지 아닌지 여부를 판단
  - t-test를 사용하여 P value 혹은 Z score를 계산하여 판단
  - 얼마나 발생하기 힘든 일인지를 보는 것! 정규분포라면 양끝단이 됨

Normal Distribution (정규 분포) - Bell Curve

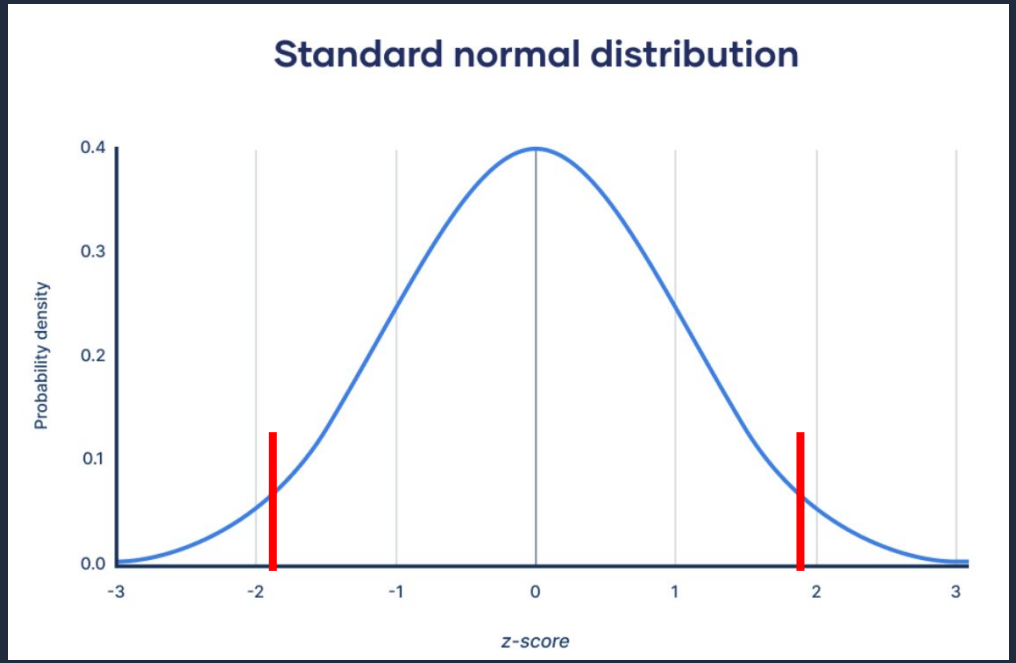
- Mean (μ) and Standard Deviation (σ) determines the shape



Standard Normal Distribution

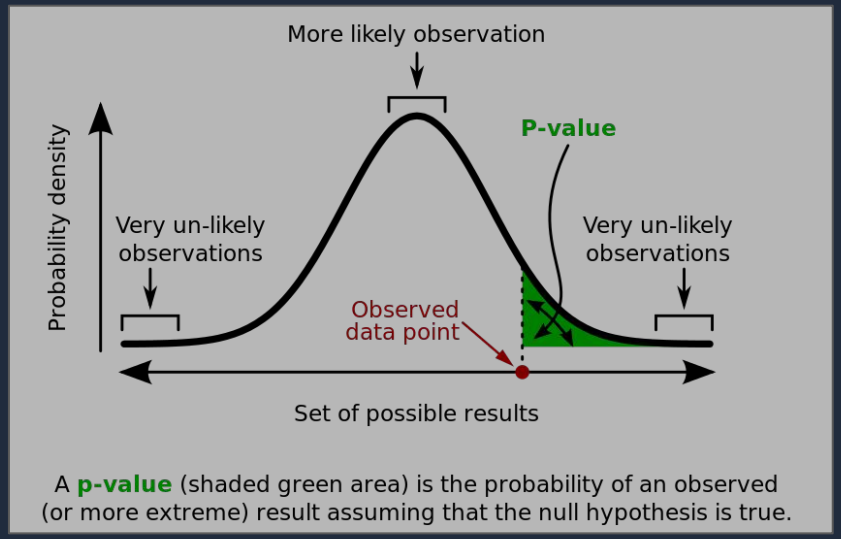
- 평균 0, 표준편차 1, 그래프 면적 1
- Z 분포라 부르기도함

- 신뢰구간과 z-score (x 축)
  - 90% -> 1.645
  - 95% -> 1.96
  - 99% -> 2.575
- 양측검정 vs. 단측검정
  - Two-sided vs. one-sided
  - 95% 단측검정이라면
    - 커져야하는 경우 1.645
    - 작아져야하는 경우 -1.645



Which Side is Better in A/B Test?

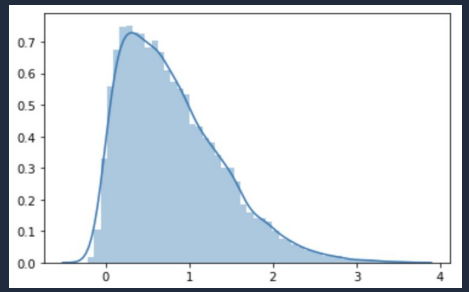
- Computation of P-value and Statistical Significance (통계적 유의성)
  - More sample and/or more difference will give you more significance

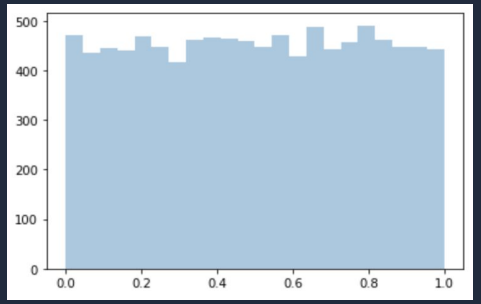


귀무 가설 기반의 AB 테스트 분석은 비교대상이 되는 데이터가 **정규분포**를 따름을 가정함

모든 데이터가 정규분포를 따르지 않음

- **Central Limit Theorem to the Rescue**
  - 중심극한 정리를 사용





Central Limit Theorem (중심 극한 정리)

- 샘플의 평균은 정규 분포를 따른다
  - Arithmetic mean (n > 30) of any distribution approximated as Normal Distribution
  - Standard deviation decreases as the data sample increases
  - 이는 모집단의 분포와 상관없음
- 샘플의 크기가 클수록 샘플을 뽑는 수가 클수록 더 정규 분포에 가까워짐
  - If we have a population and we take sufficiently large samples from it, then **the sample means** (the average of each sample) will be approximately normally distributed

Central Limit Theorem 살펴보기

- 모집단에서 샘플을 반복해서 수집 (샘플의 크기는 30 이상)
- 각 샘플의 평균을 계산
- 평균들의 평균이 모집단 평균과 유사해짐 (샘플 수가 클수록 샘플 수집을 많이 할수록)
- 샘플 평균의 히스토그램은 정규분포를 따르게 됨
- 수집된 샘플 평균들간의 **차이**나 **합**도 **정규분포**를 따름

$$u_{\bar{x}} = u$$

$u_{\bar{x}}$ = Mean of the sample means

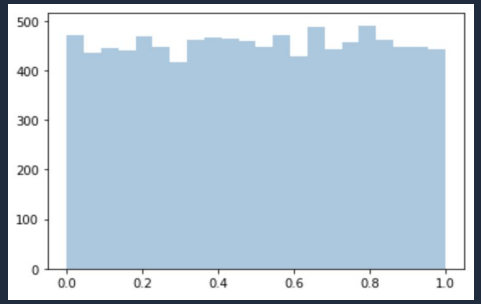
u = Population Mean

$$\text{Standard Deviation} = \frac{\sigma}{\sqrt{n}}$$

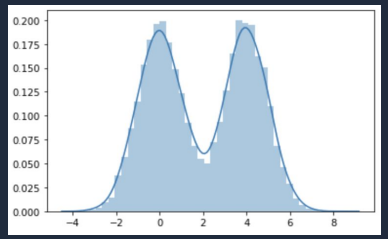
σ = the population standard deviation

n = the sample size

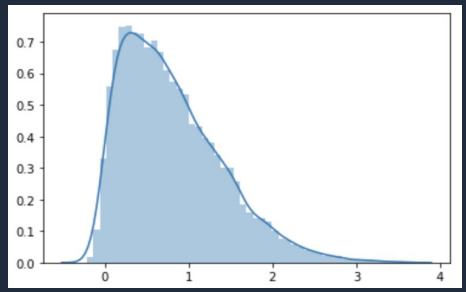
Central Limit Theorem 예제: 구글 Colab 링크



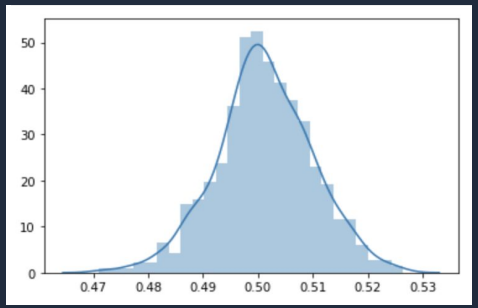
Uniform Distribution

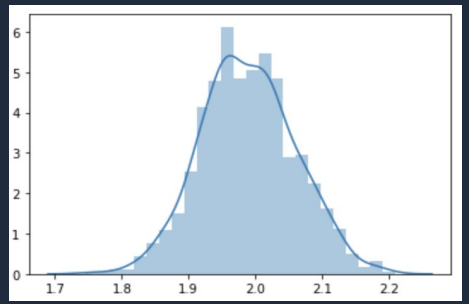


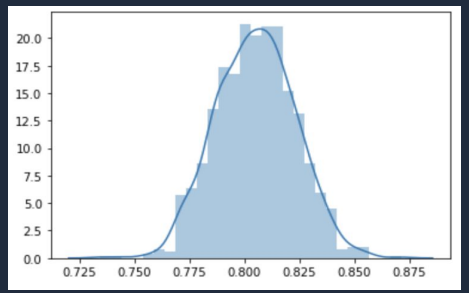
Modal Distribution



Skewed Distribution







programmers

A/B 테스트 트래픽 크기 비교

A와 B의 트래픽 크기를 통계적으로 비교해보자

트래픽 크기 비교 가설은?

AB 테스트 성공실패 지표를 비교하기 전에 제일 먼저 해야하는 일은 트래픽이 양쪽에 우리가 원하는 형태로 나눠졌는지부터 점검하는 것!

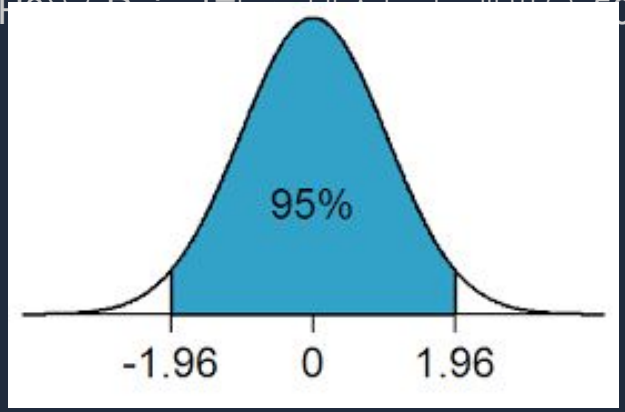
- AB 테스트 사용자 크기를 통계적으로 비교해보자
- 50:50으로 나눈 테스트라면 이는 P(A) = P(B) 혹은 P(B) = 0.5가 귀무가설(H0)이 됨
- 어떻게 비교하나?
  - CLT에 따르면 P(B)-0.5가 정규 분포를 따르게 됨
  - proportion z-test (혹은 one-sample t-test)로 유의수준(p-value)을 계산

비율 비교: Proportion z-test 공식

- 하나의 모집단에서 N개의 샘플을 통해서 나온 특정 이벤트의 확률의 평균이 P인 경우
  - 이게 P`라는 확률과 통계 $\sqrt{\frac{p(1-p)}{N}}$ 이야기할 때 다른지 아니면 같은지 z-score를 계산하는 공식
    - Z-score = (P-P`)/
  - Proportion z-test의 계산결과는 결국 z-score

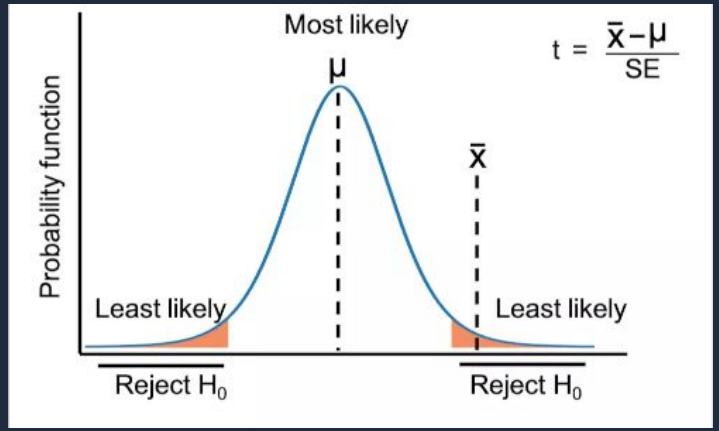
Bucket 크기 비교에 z-test 적용해보기

- 예를 들어 P가 테스트 사용자의 비율이고 N이 테스트에 속한 전체 사용자의 수 (A와 B 포함)라고 하자
  - z-score = (P - 0.5)/ $\sqrt{\frac{p(1-p)}{N}}$
- (95% 신뢰도라면) z-score가 1.96보다 크거나 -1.96보다 작으면 P는 95% 신뢰도로 봤을 때 50%가 아니라고 할 수 있음 (발생하기 힘든 일이 발생했다고 할 수 있음)
  - 이 경우 귀무가설(H0)을 P ≠ 0.5인 대안 가설을 50%가 아님 혹은 컨트롤 트래픽과 다름



t-test란?

- A t-test is a statistic method used to determine if there is a **significant** difference between **the means of two groups** based on **a sample of data**
  - T-test is when the variance (standard deviation) is unknown
- One-sample t-test와 Two-sample t-test가 존재
  - Bucket 크기 비교는 전자 (혹은 z-test)
  - Impression/click/purchase/amount는 후자에 속함



Two Sample T-Test

- https://www.statsdirect.co.uk/help/parametric_methods/utt.htm

Assuming equal variances, the test statistic is calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

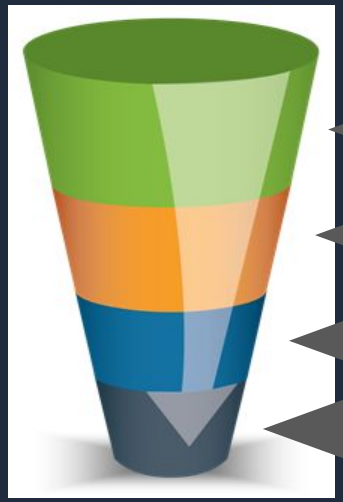
- where x bar 1 and x bar 2 are the sample means, s^2 is the pooled sample variance, n_1 and n_2 are the sample sizes

- 4강에서 자세히 살펴볼 예정

A/A Test를 통한 검증 방식 (1)

- 기본적으로는 A/B 테스트 분석과 동일
- 하지만 차이점은
  - 기존 서비스 방문 트래픽을 랜덤하게 추출(보통 날짜 기간 기반)
  - 앞서 구현한 Bucketing 로직을 적용해서 트래픽을 A와 A`로 분리
  - 그리고 기타 비교 지표들을 계산하고 그 값들이 동일함을 컨펌

A/A Test를 통한 검증 방식 (2)



Impressions

Clicks

Enrollment (Revenue)

Consumption and NPS

통계적으로 무의미한 차이가 나야함 (95% 신뢰도)

A	A`
10,000	10,500
500	480
15	16
11	12

programmers

실습: A/B 테스트 트래픽 크기 비교

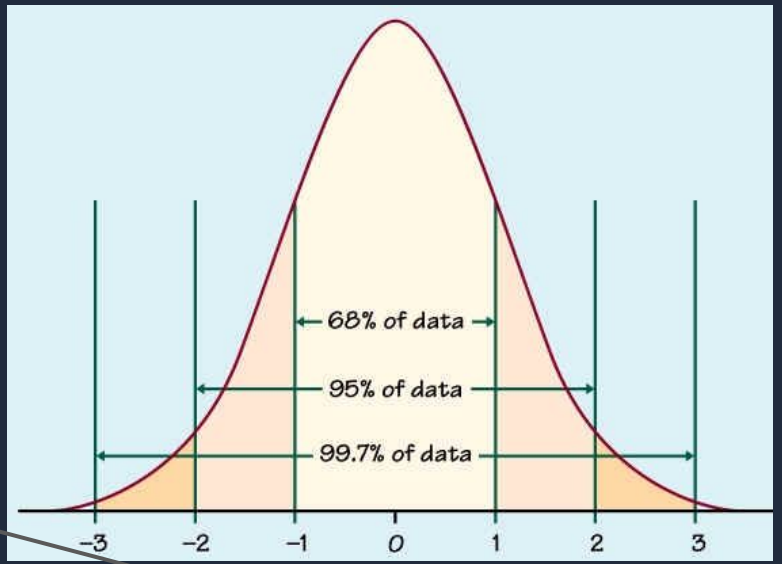
A와 B의 트래픽 크기를 통계적으로 비교해보는 실습을 해보자

실습 1

- 앞서 aa_example 테이블의 user_id를 가지고 다음을 수행
  - 앞서 설명한 파이썬 함수로 A/B로 나눴을 때 A에 속한 사용자의 수와 B에 속한 사용자의 수
  - 앞서 설명한 SQL 함수로 A/B로 나눴을 때 A에 속한 사용자의 수와 B에 속한 사용자의 수
- Colab 링크

AB 테스트 Variant 크기 검증

- A and B are supposed to be 50:50 in our case
  - $P_A = P_B = 0.5$
- H_0 (NULL Hypothesis - 귀무가설):
  - $P_B = 0.5$
- With Total N users
  - $\mu_A = N*P_A$
  - $\mu_B = N*P_B$
  - $\sigma_B = \text{sqrt}\ (P_B(1-P_B)/N)$
- **Z-Score** = $(\mu_B-\mu_A)/\sigma_B$



Z-score Distribution

실습 2. z-score를 계산해보기 (1)

- A 사용자 수는 2070
- B 사용자 수는 2056
- N = 2070 + 2056 = 4126
- 비교대상 확률 = 0.5
- P(B) = 2056/4126 = 0.498
- z-score = (0.498-0.5) / sqrt(0.498*(1-0.498)/4126)
  = -0.217

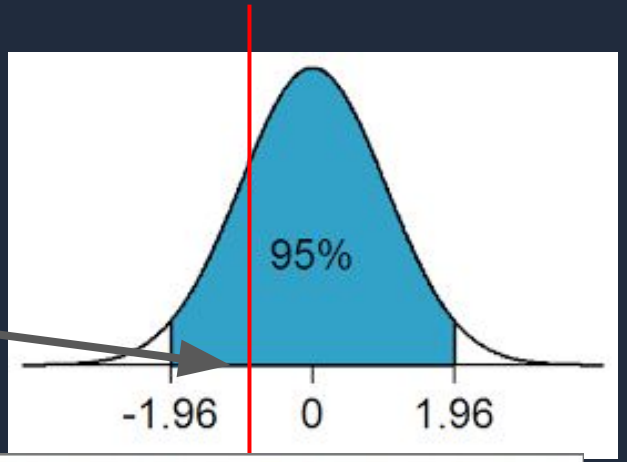
$$\text{Z-score} = (P-0.5)/\sqrt{\frac{p(1-p)}{N}}$$

귀무가설이 채택됨. 즉 B 사용자 비율은 50%라고 할 수 있음

실습 2. z-score를 계산해보기 (2)

- z-score = -0.217

귀무가설이 채택됨. 즉 B 사용자 비율은 50%라고 할 수



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

실습 2: Z-score 계산

- (A/B 테스트) 3강 리뷰
  - 파이썬으로 직접 계산
  - Scipy 모듈 사용해서 계산

관련 기타 링크/책

- Stats 101: An Intuitive Guide to A/B Testing by Michael Sussman
- Hypothesis Testing: The Basics
- 그림으로 설명하는 개념 쏙쏙 통계학
- 신입탐정의 데이터 분석 입문

programmers

3강 QA

3강 관련 질문들에 대해 이야기해보자

One sample t-test z-score 계산하는 코드 작성하기

- Python에서 A와 B의 트래픽 크기(사용자수가 될 수도 있고 세션수가 될 수도 있음)를 인자로 받아 앞서 슬라이드 15장을 바탕으로 z-score를 계산하는 코드를 만들어보기
  - Redshift에서도 SQL로 계산 가능. 아니면 Python UDF를 작성하면 분명 가능함
  - Spark에서도 Python UDF로 계산할 수도 있고 최종 정보를 다운로드받아서 Driver에서 계산해도 됨
  - 2주차 숙제에 이어서 해보거나 별도 Colab을 만들어서 구현
- 자신의 숙제에서 A,B 세션 수를 입력으로 z-score를 계산해보기 (One-sample t-test)

```
import math

def compute_zscore(n_test, n_ctrl):
    n = n_test + n_ctrl
    …
    return z_score

print(compute_zscore(2070, 2056))
```

programmers

Archived

불필요한 슬라이드들

One Sample T-Test (혹은 Proportion Z-Test) 공식

- 하나의 모집단에서 N개의 샘플을 통해서 나온 특정 이벤트의 확률의 평균이 P인 경우
  - 이게 P`라는 확률과 통계 $\sqrt{\frac{p(1-p)}{N}}$ 이야기할 때 다른지 아니면 같은지 z-score를 계산하는 공식
    - Z-score = (P-P`)/
- One sample t-test
  - 위의 공식과 동일하지만 분모가 degree of freedom으로 인해 N-1이 됨.
  - The 1-sample t-test estimates only one parameter: the population mean. The sample size of n constitutes n pieces of information for estimating the population mean and its variability. One degree of freedom is spent estimating the mean, and the remaining n-1 degrees of freedom estimate variability.

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

Things to Read Through

- https://ux.stackexchange.com/questions/148282/a-b-testing-client-wanted-the-test-run-70-30
- AB 테스트 결과를 보기 위해 필요한 트래픽의 크기를 어떻게 파악 가능한가?
  - Optimizely's A/B test sample size calculator
  - Evan Miller's Sample Size Calculator

t-test와 z-test의 차이점은 무엇인가요?

- z-distribution의 경우, 평균은 0이고 표준편차가 1이 됨
- 보통 온라인 AB 테스트에서는 최소 1000명의 사용자가 필요하며 이 정도 규모에서는 t-score가 z-score와 같다고 가정함