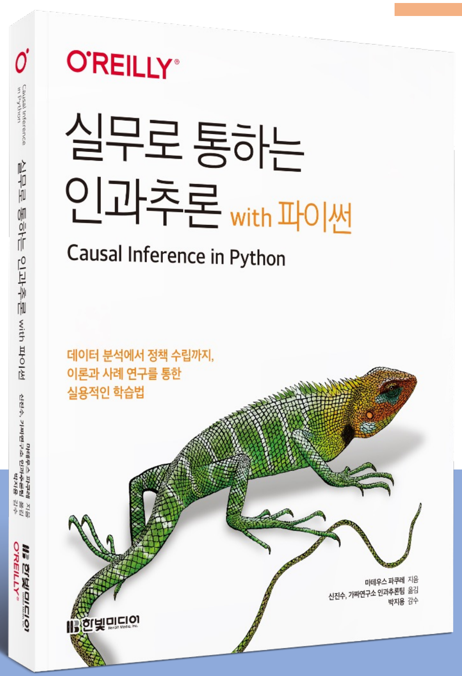
가짜연구소

Causal-Lab

『실무로 통하는 인과추론 with 파이썬』 특강

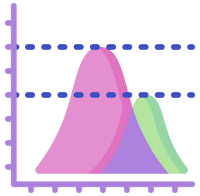


온라인 통제 실험 소개 및 통계 분석 기초

방태모 | G마켓

본인 소개

통계학 전공



Data Scientist

가천대 길병원 GABC

Data Scientist

Gmarket

글쓰기

[블로그](#)
[요즘IT](#)

온라인 통제 실험

가짜연구소

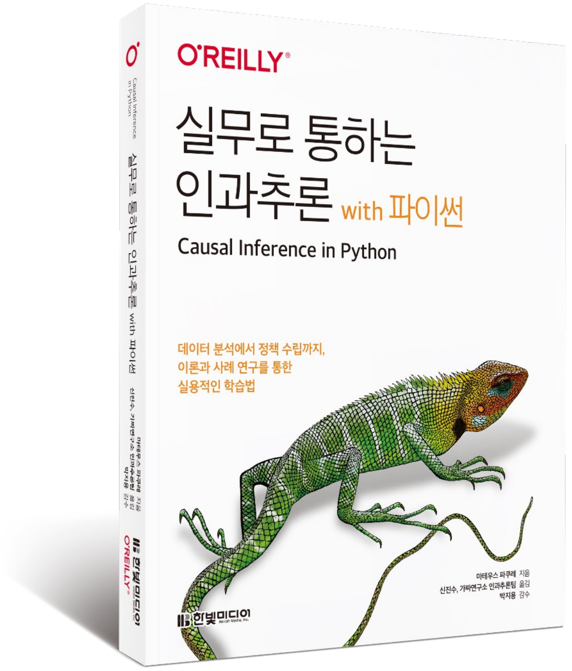
Causal-Lab

taemobang@gmail.com

[링크드인](#)

오늘 할 이야기

- 온라인 통제 실험 소개
- 무작위 배정 방법
- 기초적인 통계 분석 이론 소개



1 온라인 통제 실험 소개

온라인 통제 실험이란?

온라인 통제 실험 (Online Controlled Experiments, OCE)

사용자 단위 무작위 할당

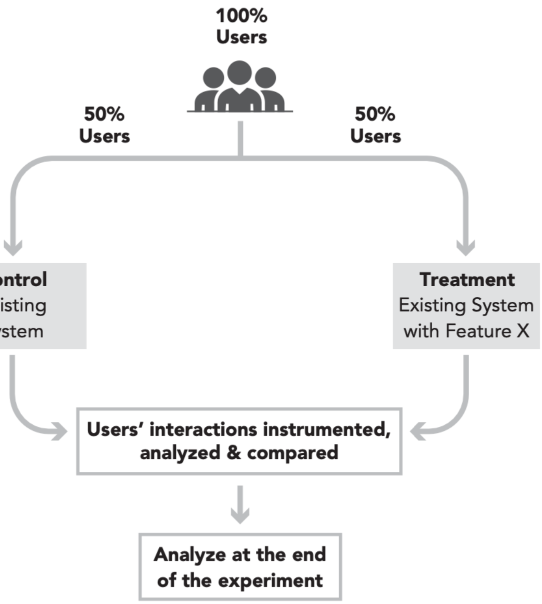


Figure 1.2 A simple controlled experiment: An A/B Test

그림 1. 온라인 통제 실험 (Source: Kohavi, Tang, and Xu 2020)

온라인 통제 실험이란?

온라인 통제 실험 (Online Controlled Experiments, OCE)

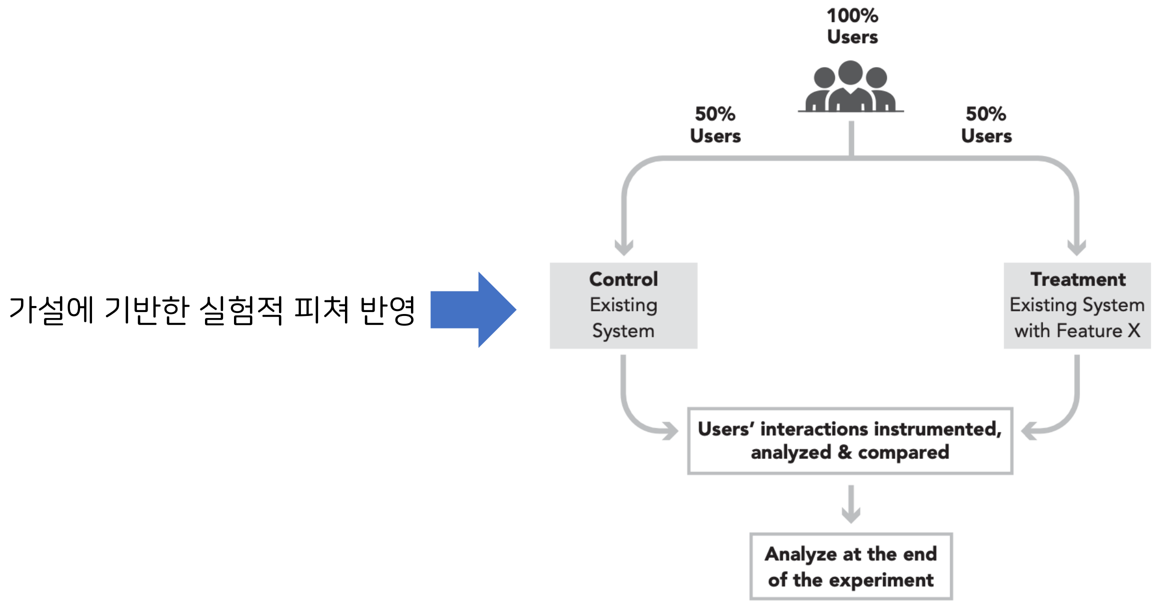


Figure 1.2 A simple controlled experiment: An A/B Test

그림 1. 온라인 통제 실험 (Source: [Kohavi, Tang, and Xu 2020](#))

온라인 통제 실험이란?

온라인 통제 실험 (Online Controlled Experiments, OCE)

정해진 실험 기간 동안 실험 데이터 수집

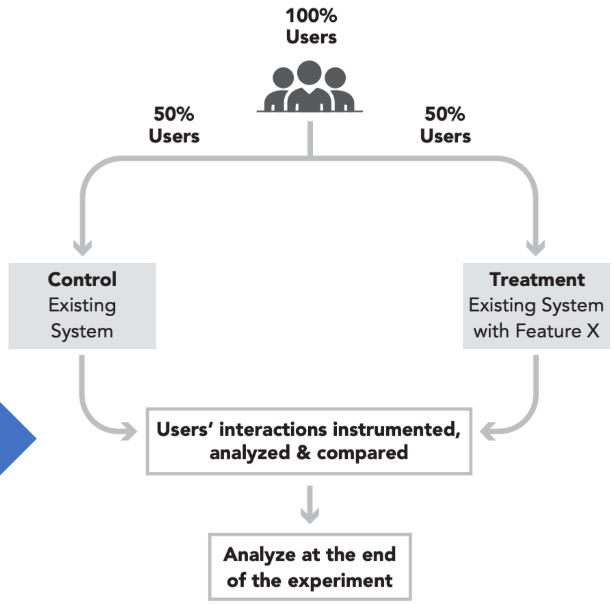


Figure 1.2 A simple controlled experiment: An A/B Test

그림 1. 온라인 통제 실험 (Source: [Kohavi, Tang, and Xu 2020])

온라인 통제 실험이란?

온라인 통제 실험 (Online Controlled Experiments, OCE)

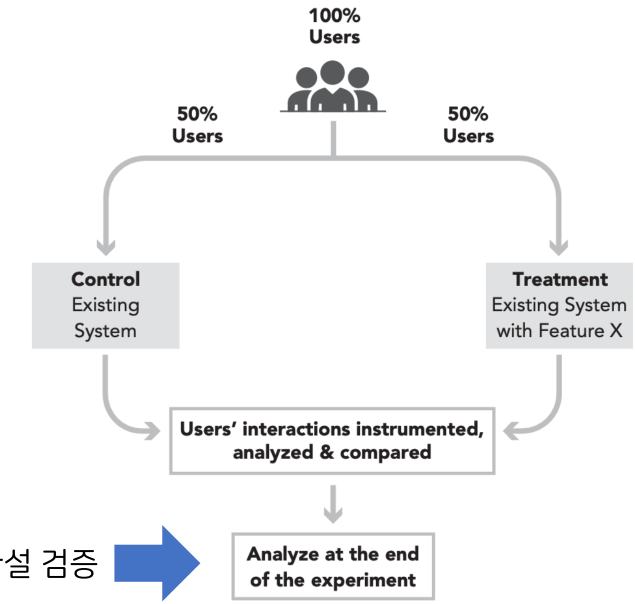


Figure 1.2 A simple controlled experiment: An A/B Test

실험 기간 종료 후, 실험 분석을 통해 가설 검증

그림 1. 온라인 통제 실험 (Source: Kohavi, Tang, and Xu 2020)

온라인 통제 실험의 등장 배경

Causal-Lab

온라인 통제 실험, 그 이전에는 RCT(Randomized Controlled Trial)

- 20세기 초 통계학의 대가 R.A. Fisher 경에 의해 발전된 방법론
- 1920년대 농업, 1940년대 의학 분야, 그 이후 제조업에서까지 번영을 이뤄냄
- 2000년대 이르러 world-wide-web, 웹서비스의 성장에 따라 온라인으로까지 확장

온라인 통제 실험의 등장 배경

Causal-Lab

왜 온라인 영역까지 확장되었는가?

- 온라인 서비스를 운영하는 기업의 최대 관심

온라인 통제 실험의 등장 배경

Causal-Lab

왜 온라인 영역까지 확장되었는가?

- 온라인 서비스를 운영하는 기업의 최대 관심: 고객이 진정 원하는 것은 무엇인가?

온라인 통제 실험의 등장 배경

Causal-Lab

왜 온라인 영역까지 확장되었는가?

- 온라인 서비스를 운영하는 기업의 최대 관심: 고객이 진정 원하는 것은 무엇인가?

우리 서비스, 프로덕트에 준 변화에 따른 대규모 고객 반응을 지표화하고 분석하면,

고객이 진정 원하는 것을 찾아낼 수 있지 않을까? (온라인 실험의 목적은 Confirm이 아닌 Learn)

온라인 통제 실험의 등장 배경

단, 여기서 고객 반응의 변화는 오로지 우리가 준 변화에 의해 발생한 부분만 발라내서 봐야한다!

- 온라인 서비스를 운영하는 기업에서 실험을 도입하게 되는 가장 큰 모티베이션

온라인 통제 실험의 등장 배경

어떻게 우리가 준 변화의 부분만 오로지 발라내서 볼 수 있는가?

➔ 인과추론(Causal Inference)

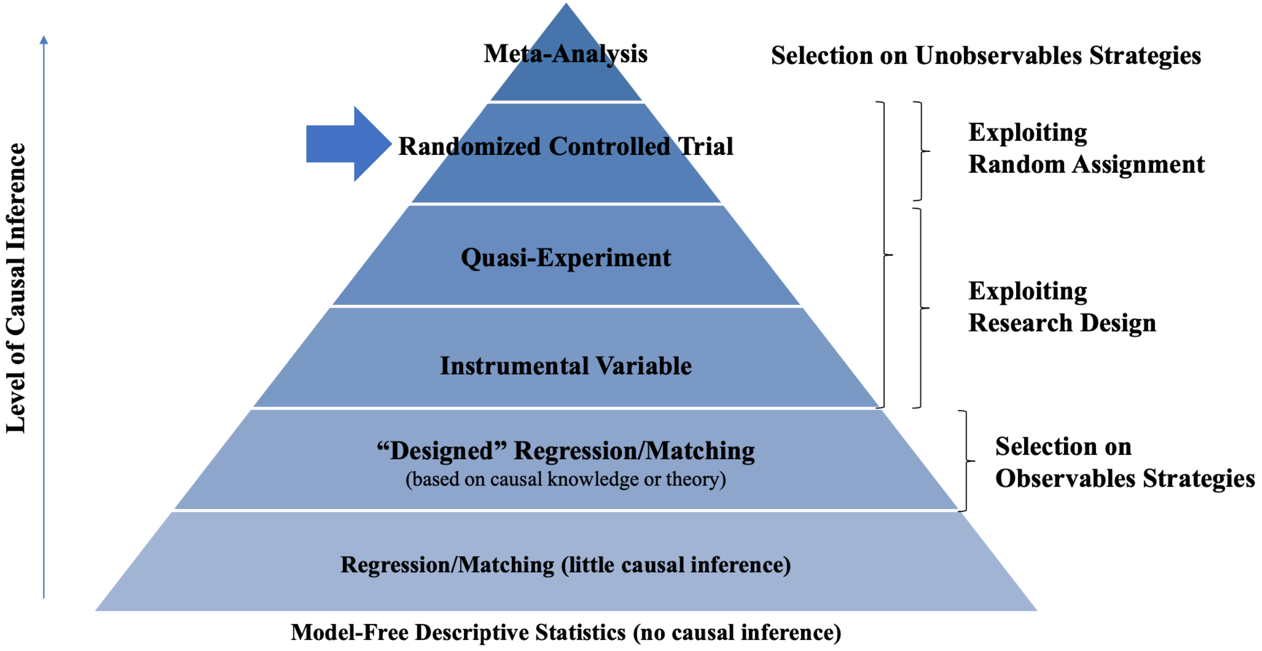


그림2. 인과추론 방법론 피라미드 (Source: Korea Summer Workshop on Causal Inference 2022)

2 무작위 배정

온라인 실험에서의 무작위 배정

"기본적으로 랜덤화는 처치와 잠재적 결과를 독립적으로 만듭니다." – Chapter 2 p.68

➔ ATE(Average Treatment Effect, 평균처치효과) 식별

온라인 실험에서의 무작위 배정

Causal-Lab

"기본적으로 랜덤화는 처치와 잠재적 결과를 독립적으로 만듭니다." – Chapter 2 p.68

➔ ATE(Average Treatment Effect, 평균처치효과) 식별

온라인 통제 실험에서 무작위 배정은 어떻게 이루어지는가?

해시 함수

- 온라인 통제 실험에서 무작위 배정은 해시 함수를 통해 이루어짐
  - Murmur Hash (주로 사용)
  - MD-5 Hash
  - …

해시 함수

- 온라인 실험에서 무작위 배정은 해시 함수를 통해 이루어짐
  - Murmur Hash (주로 사용)
  - MD-5 Hash
  - …

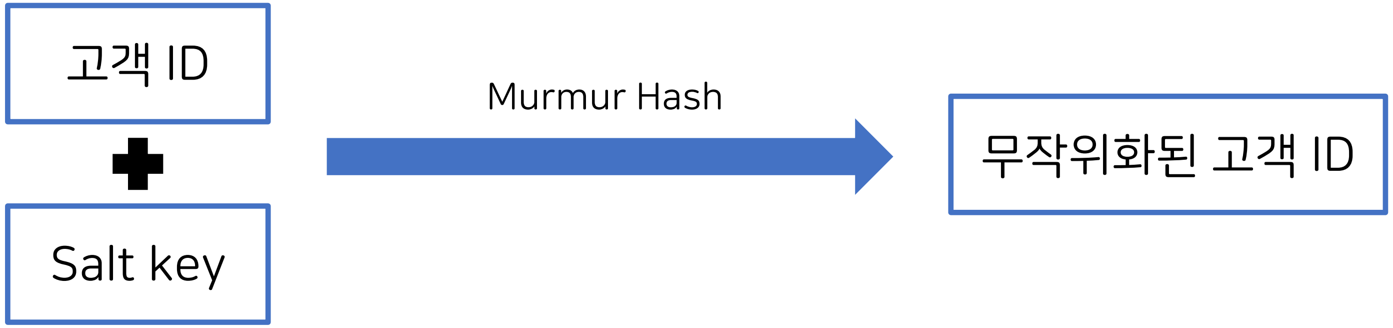


그림 3. 해시 알고리즘을 통한 고객 ID 무작위화

무작위 배정 방법 자세히 이해하기

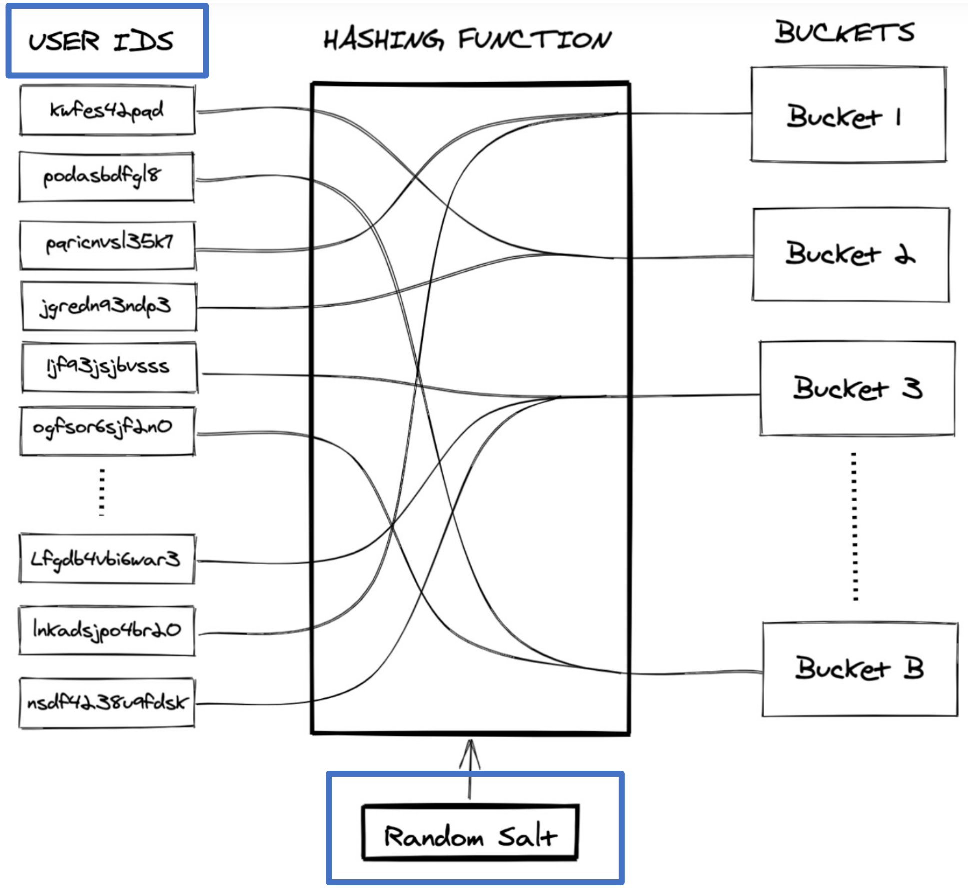


그림 4. 해시 알고리즘을 이용한 무작위화 및 고객 버켓팅 (Source: Spotify, Schultzberg, Kjellin, and Rydberg 2021)

무작위 배정 방법 자세히 이해하기

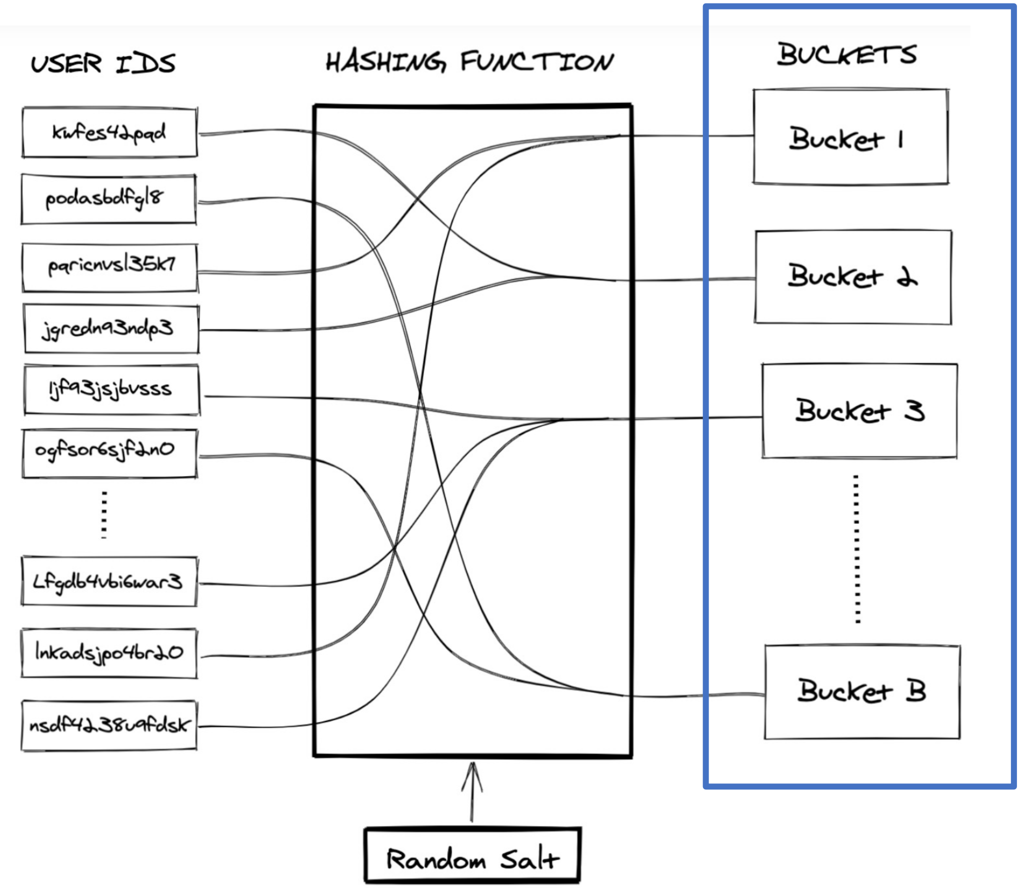


그림 4. 해시 알고리즘을 이용한 무작위화 및 고객 버켓팅 (Source: Spotify, [Schultzberg, Kjellin, and Rydberg 2021](#))

무작위 배정 방법 자세히 이해하기

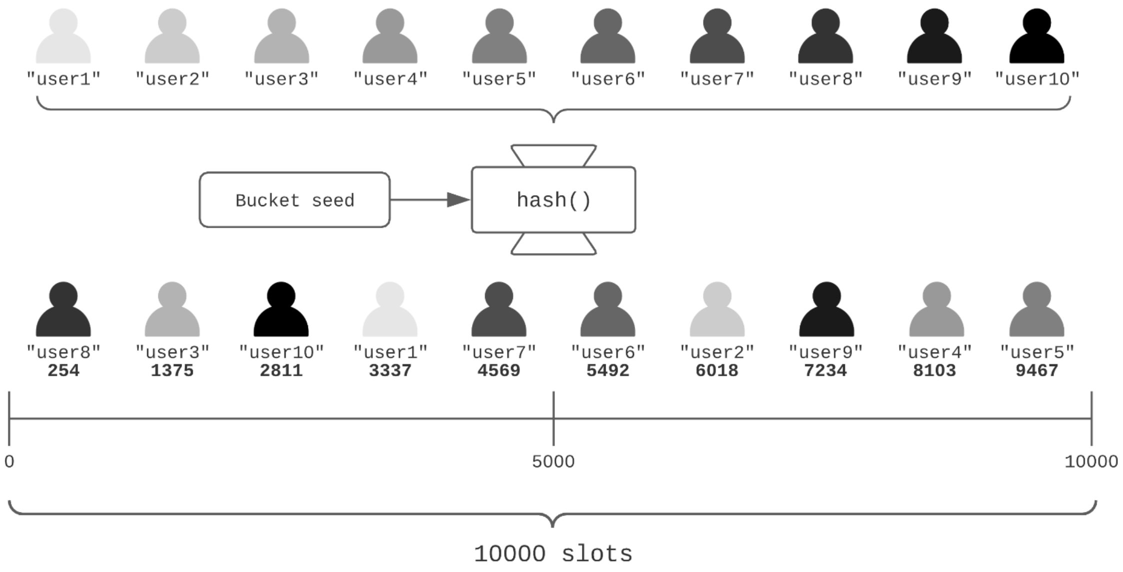


그림 5. 해시 알고리즘을 이용한 무작위화 및 고객 버켓팅 (Source: [Hackle Docs])

무작위 배정 방법 자세히 이해하기

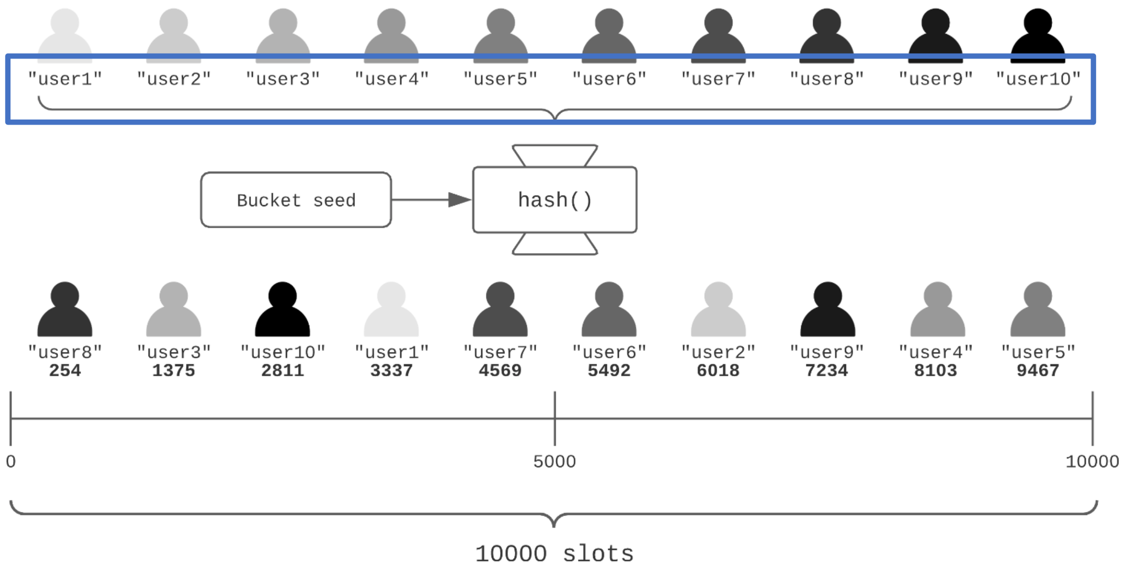


그림 5. 해시 알고리즘을 이용한 무작위화 및 고객 버켓팅 (Source: Hackle Docs)

무작위 배정 방법 자세히 이해하기

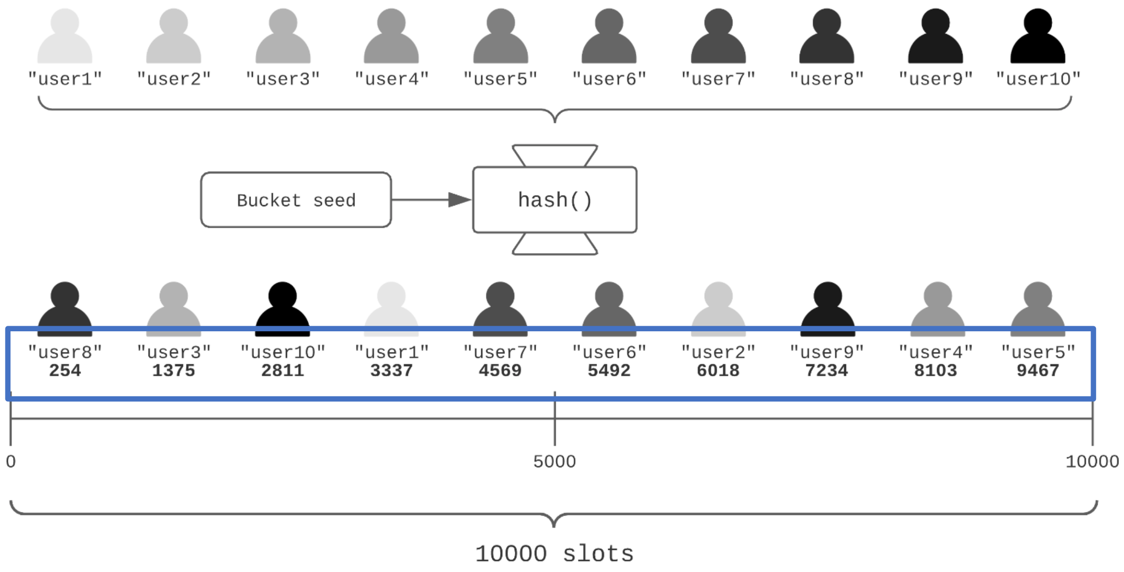


그림 5. 해시 알고리즘을 이용한 무작위화 및 고객 버켓팅 (Source: [Hackle Docs])

무작위 배정 방법 자세히 이해하기

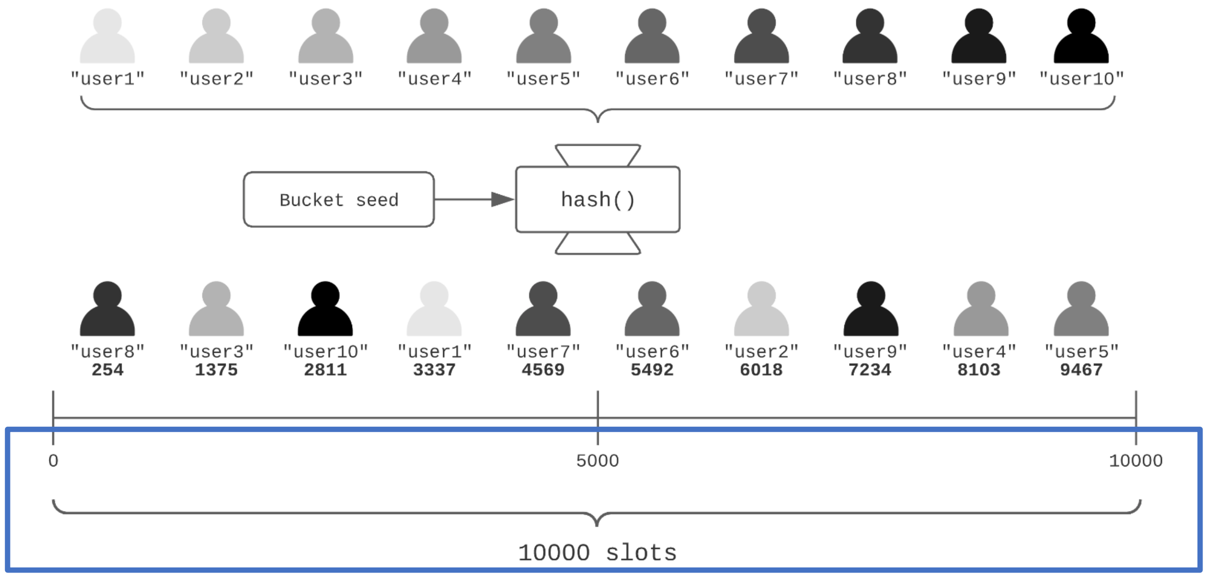
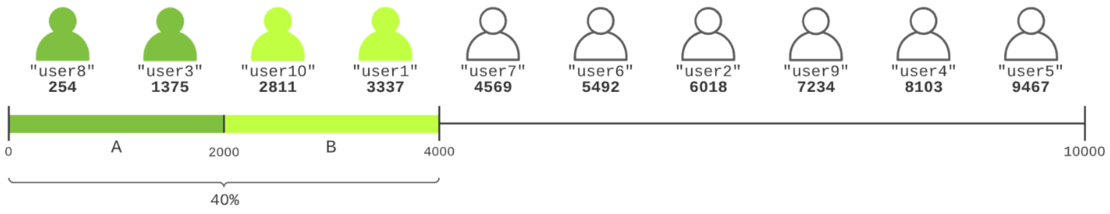


그림 5. 해시 알고리즘을 이용한 무작위화 및 고객 버켓팅 (Source: Hackle Docs)

무작위 배정 방법 자세히 이해하기

테스트 그룹 분배



트래픽 할당 40% / 그룹분배 A(50) : B(50)

그림 6. 버켓팅에 기반한 실험 변형군 분배 (Source: [Hackle Docs])

3 통계 분석 기초 I. 표준오차

표준오차

표준편차와 표준오차는 다르다.

표준편차(standard deviation, SD)

- 개별 관측값(e.g. Y)의 퍼진 정도

표준오차(standard error, SE)

- 통계량(e.g. $\bar{Y}$)의 퍼진 정도

표준오차

$$\bar{Y} = \frac{1}{n}\sum Y_i$$

$\rightarrow$ 표본평균

$$Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$$

$$Var(\bar{Y}) = Var\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

표준오차

Causal-Lab

$$\bar{Y} = \frac{1}{n}\sum Y_i$$ → 표본평균

$$Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$$ → 개별 관측값의 분산 = 표준편차의 제곱

$$Var(\bar{Y}) = Var\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$

표준오차

$$\bar{Y} = \frac{1}{n}\sum Y_i$$ → 표본평균

$$Var(Y) = \hat{\sigma}^2 = \frac{1}{n-1}\sum (Y_i - \bar{Y})^2$$ → 개별 관측값의 분산 = 표준편차의 제곱

$$Var(\bar{Y}) = Var\left(\frac{1}{n}\sum Y_i\right) = \frac{1}{n^2} * n * \hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$$ → 표본평균의 분산 = 표준오차의 제곱

표준오차

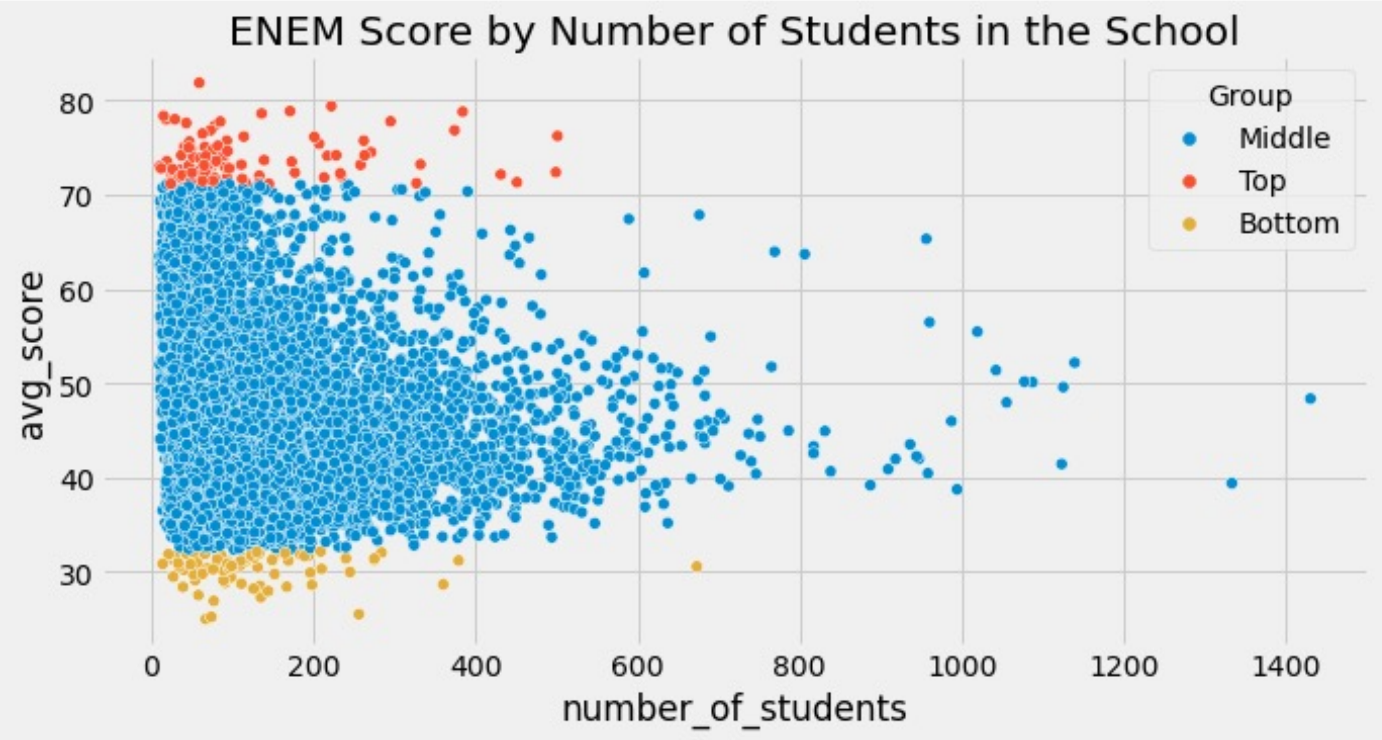


그림 7. 표본 수에 따른 학생점수(표본평균)의 분포 (Source: 교재 Chapter 2. 그림 2-1)

표준오차

관심 지표(i.e. 통계량 = 표본평균)의 중심 위치와 함께 퍼진 정도를 고려하는 것은 굉장히 중요

- 퍼진 정도는 불확실성을 이야기해줌
- 불확실성 = 우리가 관측한 실험 데이터가 우연히 관측된 것인가?

4 통계 분석 기초 II. 신뢰구간

확률변수

통계학을 이해하기 위한 첫 걸음

- 확률변수(random variables)는 분포를 갖는다.

확률변수

공정한 6면 주사위를 던지는 확률실험에서 표본공간 S를 정의하면:

$$S = \{1, 2, 3, 4, 5, 6\} = \{w_1, w_2, \cdots, w_5, w_6\}.$$

이때 표본공간에서 정의되는 함수 X를

$$X = \{\text{원소 } w\text{가 짝수인 경우}\}, w \in S$$

으로 정의하면,

X의 영역은 $A = \{2, 4, 6\}$이 되며, 함수 X는 실수값을 갖는 함수가 되므로 확률변수가 됩니다.

확률변수는 분포를 갖는다.

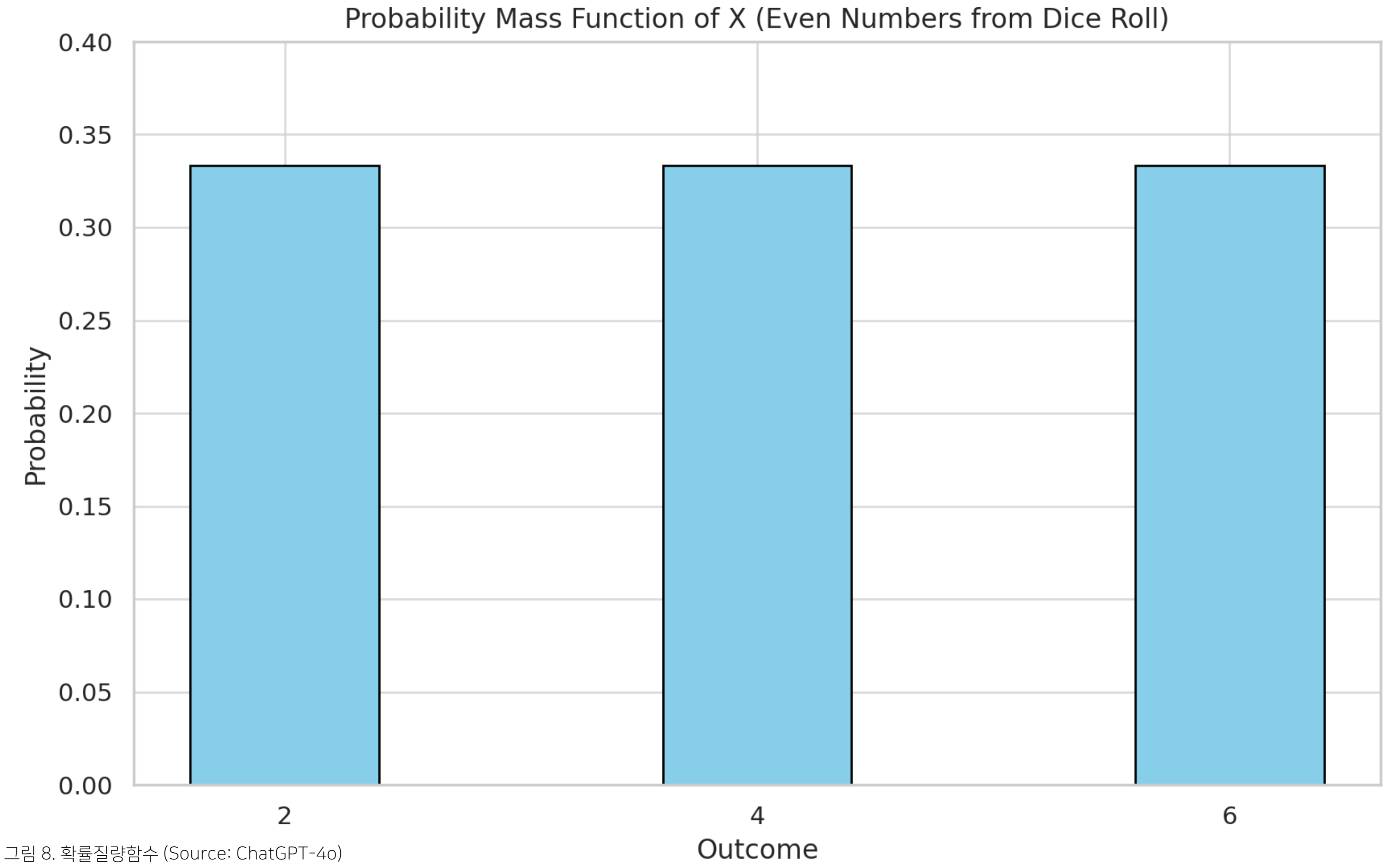


그림 8. 확률질량함수 (Source: ChatGPT-4o)

통계량과 추정량

통계량 (statistic)

- 확률변수들의 함수
  - $\bar{Y} = \frac{1}{n}\sum Y_i$
  - $\sum Y_i$
  - …

추정량(estimator)

- 모수(parameter) 추정을 위해 사용되는 통계량

- 모평균 추정에 사용하는 추정량
  - 표본평균 $\bar{Y} = \frac{1}{n}\sum Y_i$

표본평균은 분포를 갖는다.

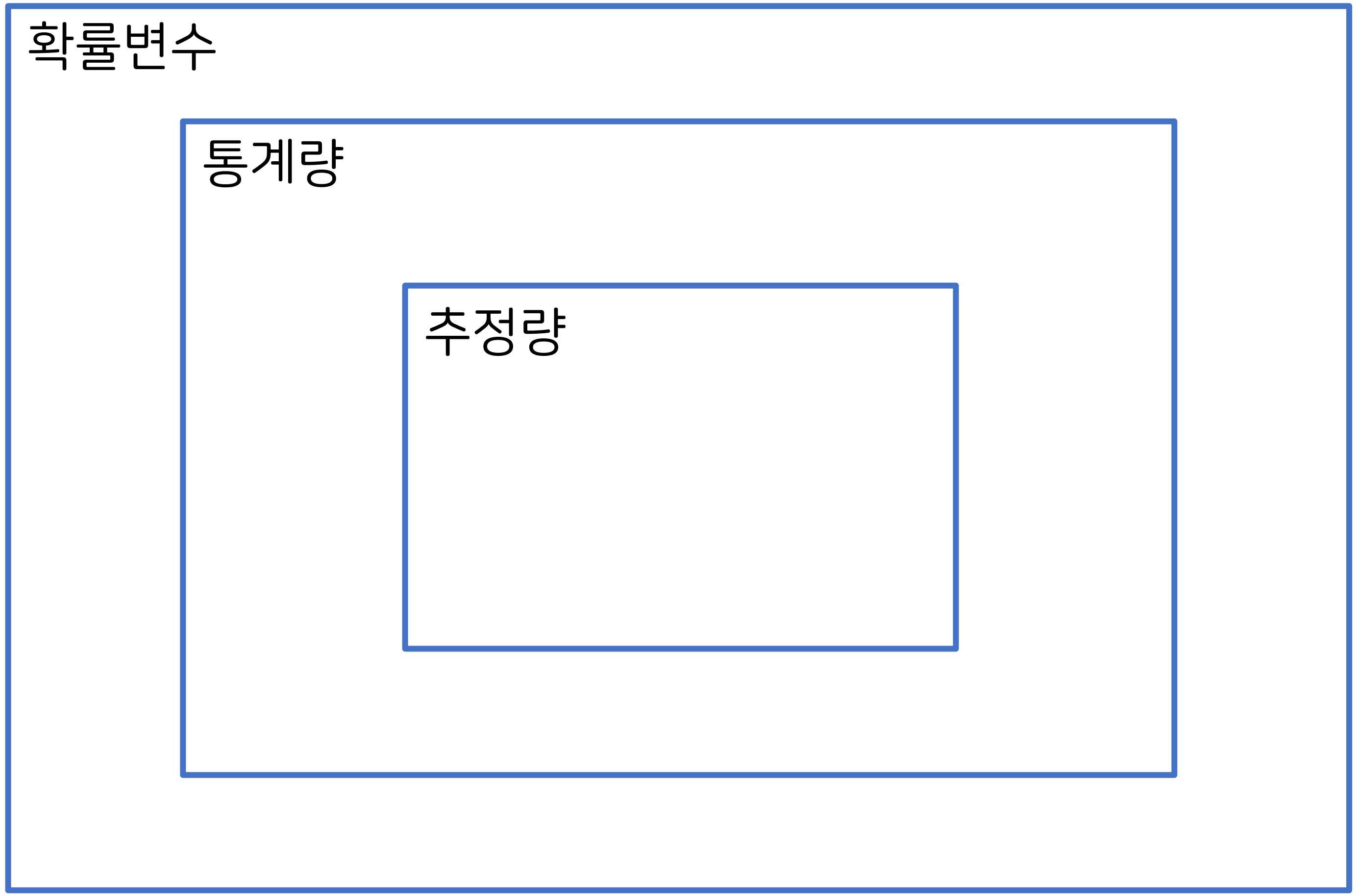


그림 10. 확률변수, 통계량, 추정량의 포함관계

- 즉 표본평균은 추정량이자, 통계량이자, 확률변수
- 따라서, 표본평균은 분포를 갖는다.

표본평균의 분포

$X \sim N(74, 2^2)$ 분포를 갖는 시험 점수로부터

500명의 학생들을 샘플링하여 시험 점수(i.e. 표본평균)를 계산하는 것을 10,000번 반복

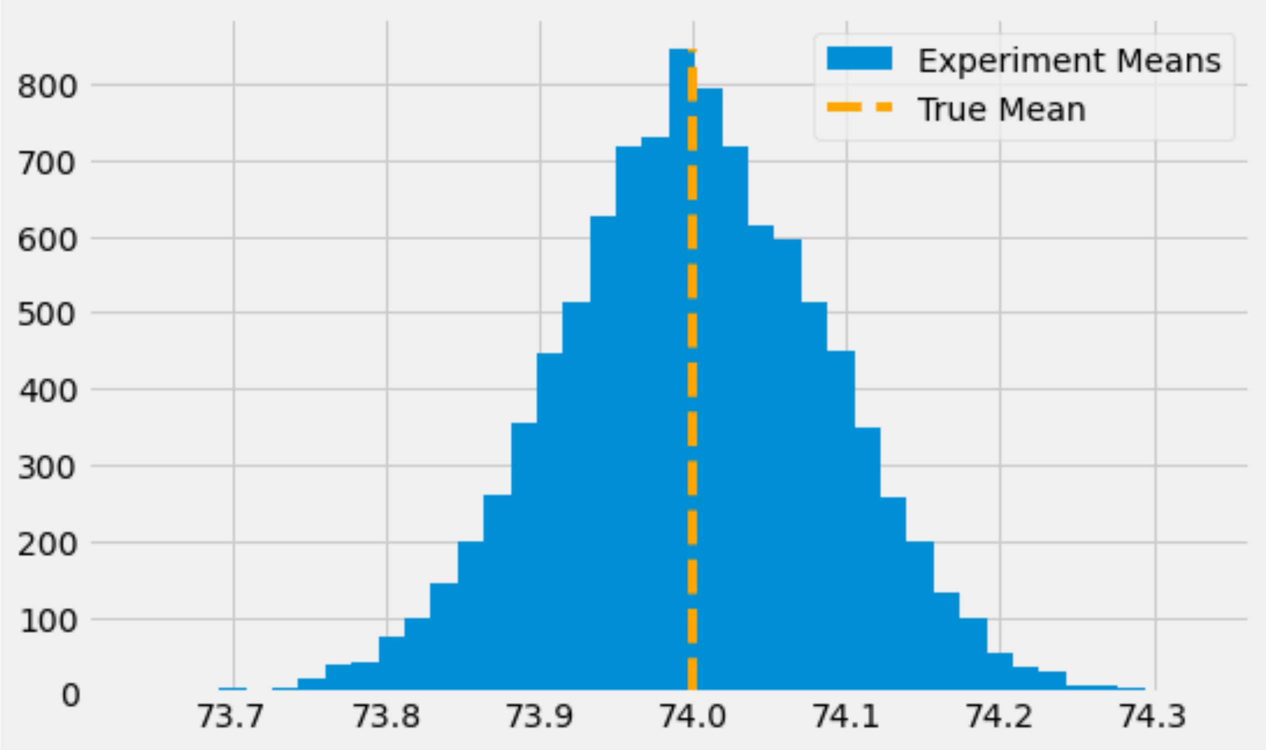


그림 9. 시뮬레이션 스터디를 통해 확인한 표본평균의 분포 (Source: 교재 웹북 Chapter 03)

표본평균의 분포

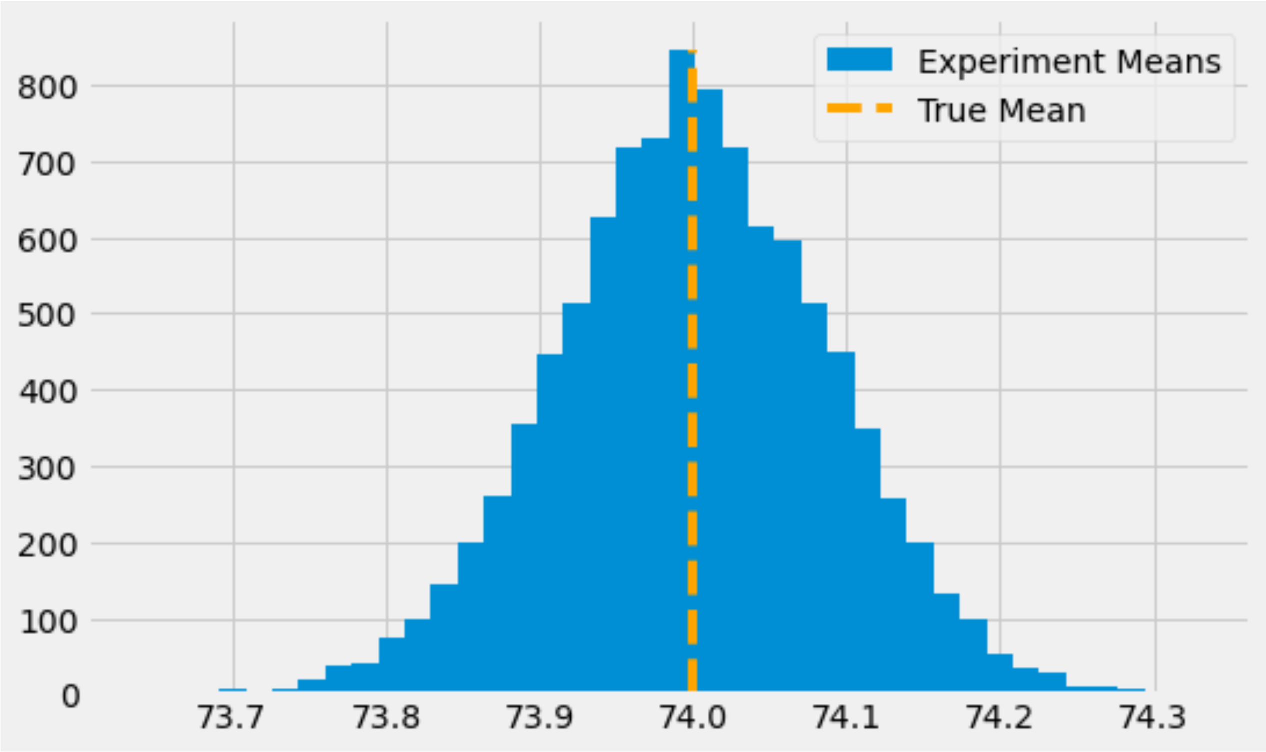


그림 9. 시뮬레이션 스터디를 통해 확인한 표본평균의 분포 (Source: 교재 웹북 Chapter 03)

중심극한정리(CLT, Central Limit Theorem)

- 표본 수가 충분히 크면 모집단의 분포에 관계없이, 표본평균의 분포는 정규분포로 근사한다.
- 실험 관심 지표에 관한 통계적 검정이 굉장히 쉬워진다.

신뢰구간

관심 지표의 분포가 보장되니 신뢰구간을 추정할 수 있음

- 95% 신뢰구간

$$\bar{x} \pm Z_{1-\alpha/2} * SE = \bar{x} \pm Z_{0.975} * \frac{\sigma}{\sqrt{n}} = \bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

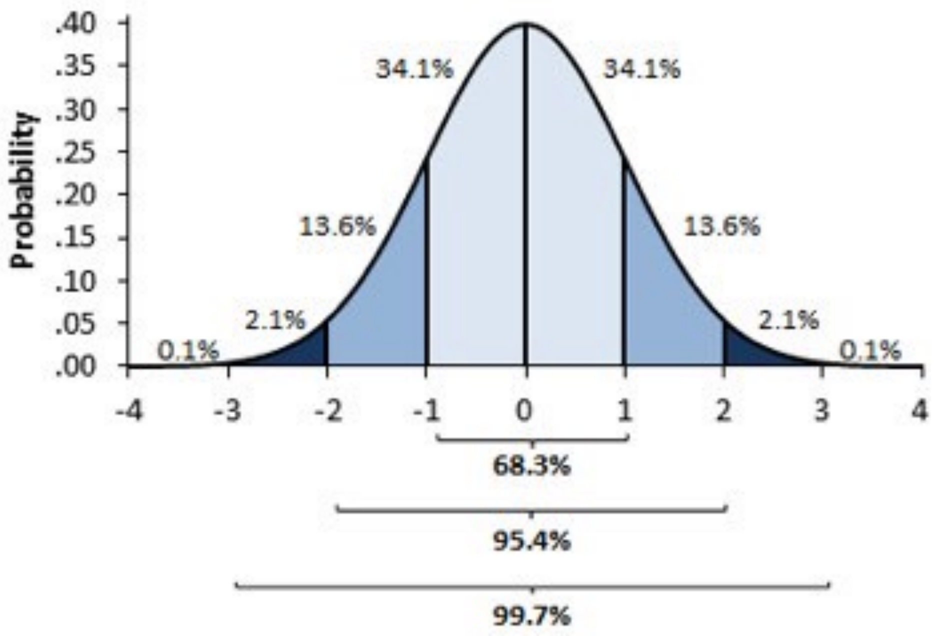


그림 10. 표준정규분포의 면적과 분위수 값 (Source: 교재 웹북 Chapter 03)

신뢰구간의 해석 방법

신뢰구간 (confidence interval)

- 빈도주의 관점
- 95% 신뢰구간에 관한 해석
  - 100번 실험했을 때, 관심 지표가 해당 구간에 95번 포함된다.
- 빈도주의 관점에서 추정된 신뢰구간은 왜 이러한 해석이 나오는가?
  - 확률표본 (i.i.d.를 충족하는 표본, random sample)
  - 관심 모수 (e.g. 모평균, 모분산, …)는 Fixed value

$$X_1, X_2, \cdots, X_{n-1}, X_n \sim i.i.d.\ N(\mu, \sigma^2)$$

5 통계 분석 기초 III. P값에 기반한 가설검정

2 표본 t 검정

2표본 t 검정(two-sample t -test)

- 모평균에 관한 검정 이론으로 실험군과 대조군 차이가 실제로 존재하는지, 우연에 의한 것인지 결정하는 통계적 가설 검정 방법

2 표본 t 검정

2표본 t 검정이 필요한 상황 예시

1. 관심 지표 $\bar{Y}$: 사용자 당 주문 건수

2. 가설 세우기
   - 대립가설(H_1, 주장하고 싶은 가설): 실험군과 대조군 간 사용자 당 주문 건수에 차이가 있다
   - 귀무가설(H_0, 내 주장에 반하는 가설): 실험군과 대조군 간 사용자당 주문 건수에 차이가 없다.

$$H_0: \mu_{trt} = \mu_{con} \ \ vs \ \ H_1: \mu_{trt} \neq \mu_{con}$$

2 표본 t 검정

3. 검정력 분석에 의해 정해진 트래픽, 실험 기간 동안 실험 데이터를 모으기

4. 실험 종료 후, 관측한 실험 데이터를 바탕으로 검정통계량인 t 통계량 계산

$$T = \frac{\Delta}{\sqrt{Var(\Delta)}}, where\ \Delta = \bar{Y}_{trt} - \bar{Y}_{con}$$

2 표본 t 검정

5. p값 계산

p값은 관측된 t 통계량보다 더욱 극단적인 값이 나올 확률을 의미함

$$\text{P−value=Prob}(T \geq |t|) = \text{Prob}(T \geq t) * 2$$

유의수준 5% 하에 계산된 p-value가 0.05보다 작으면 우리 주장에 반하는 귀무가설을 기각시킬 수 있게됨

2 표본 t 검정

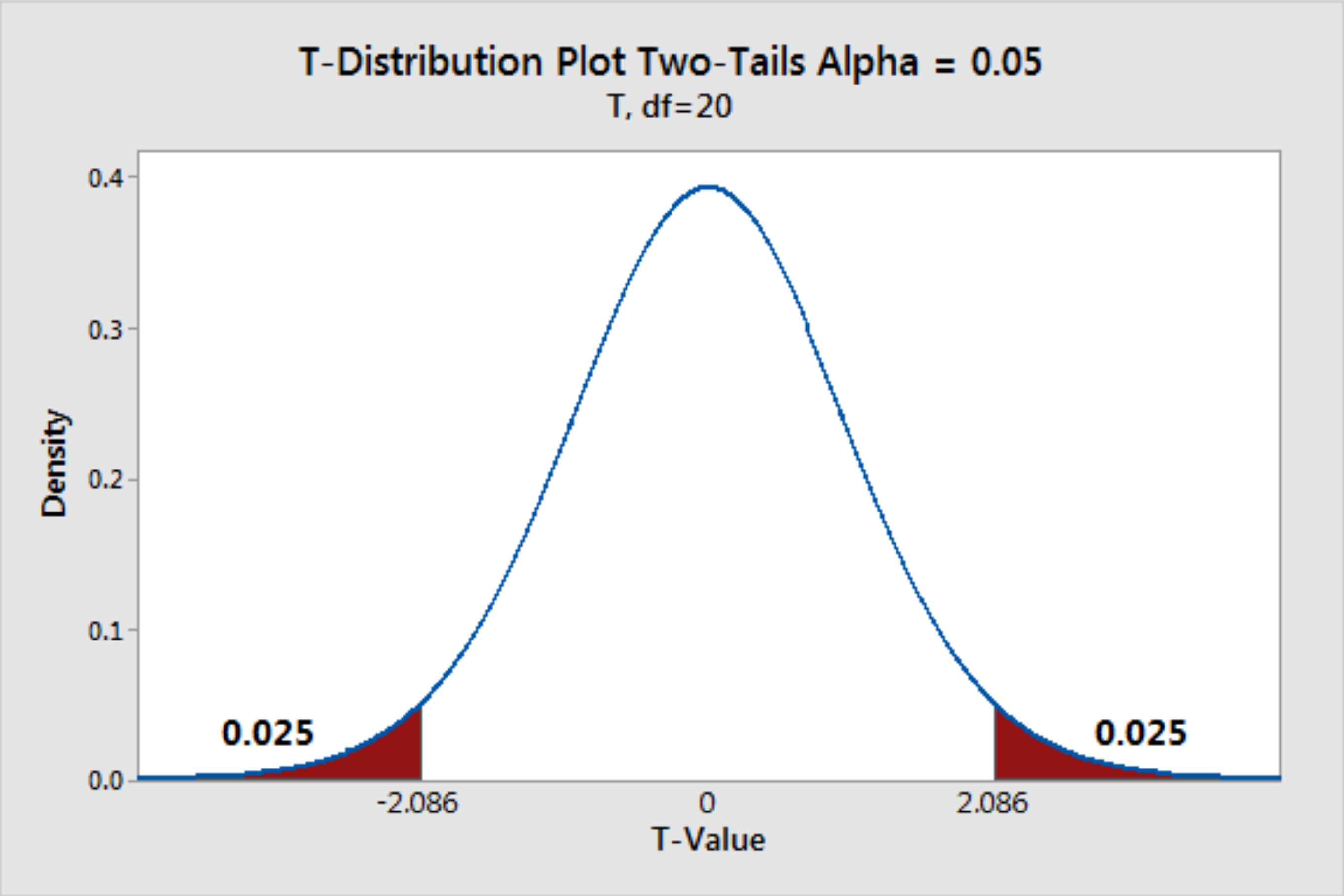


그림 11. 양측 t 검정 (Source: One-Tailed and Two-Tailed Hypothesis Tests Explained - Statistics By Jim)

2 표본 t 검정

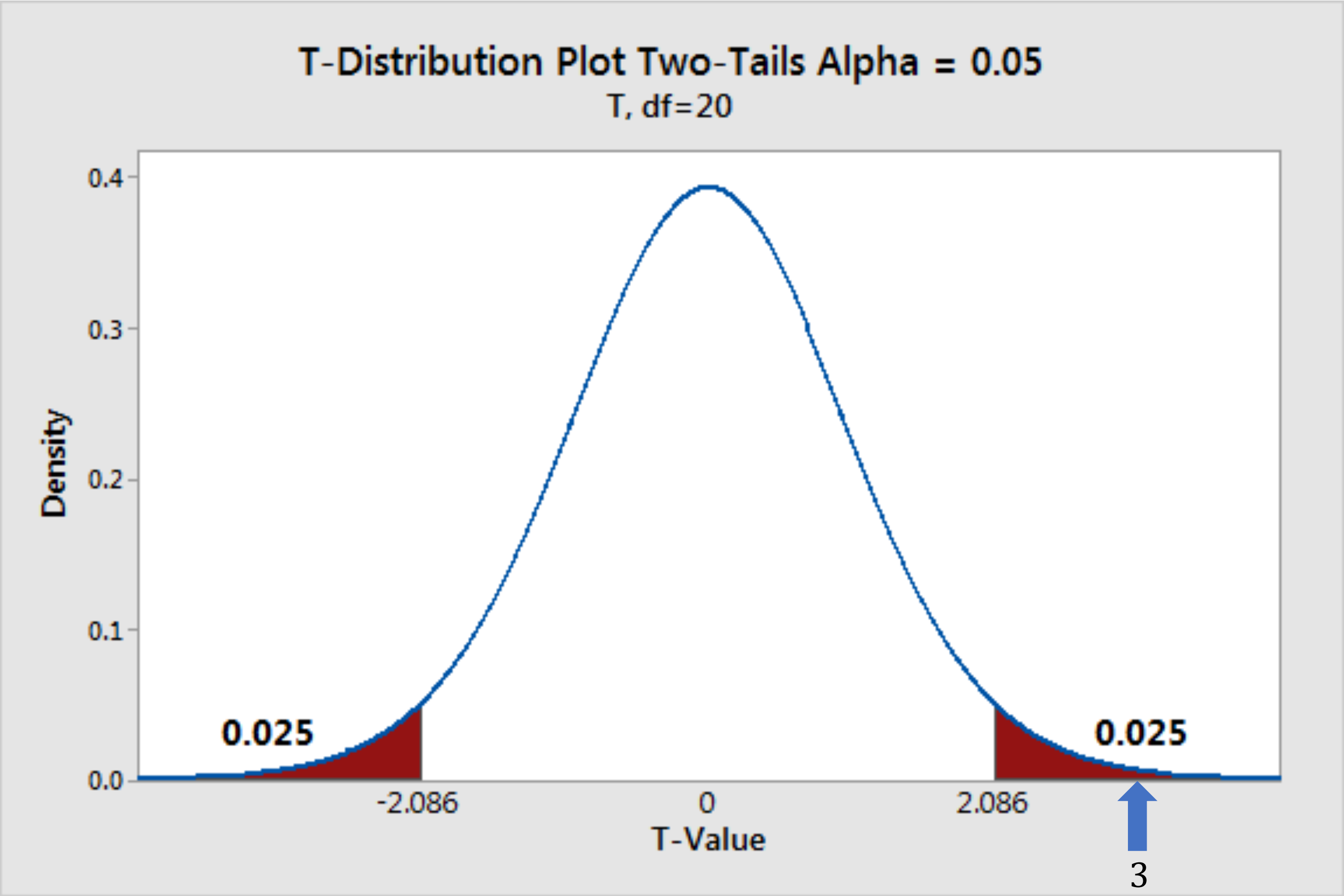


그림 11. 양측 t 검정 (Source: One-Tailed and Two-Tailed Hypothesis Tests Explained - Statistics By Jim)

6 통계 분석 기초 IV. 검정력과 검정력 분석

검정력이란?

Causal-Lab

검정력(Power, $1-\beta$)

- 실험 변형군 간 의미있는 차이가 있음을 올바르게 탐지할 확률
    - (학술적 표현) 귀무가설(H_0)을 올바르게 기각시킬 확률

검정력이란?

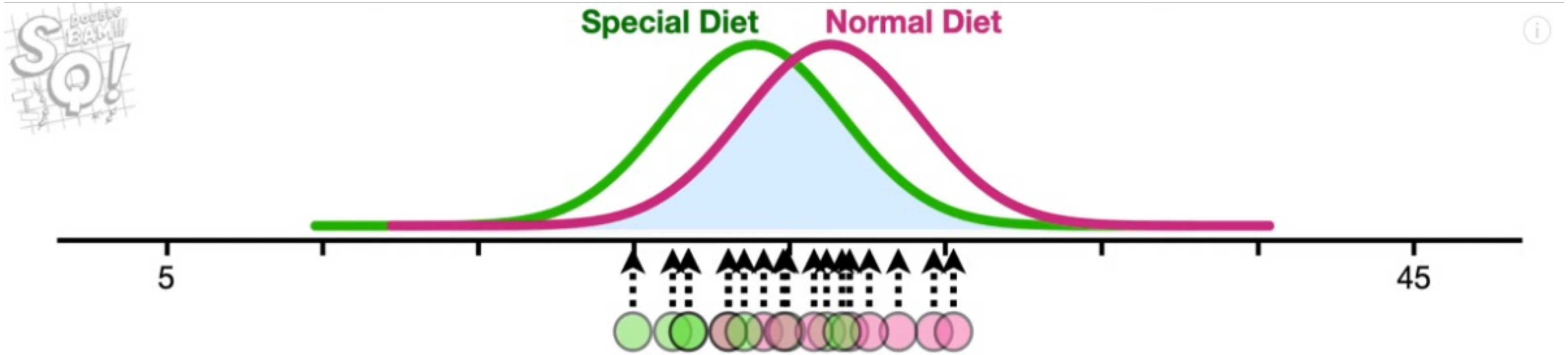


그림 14. 특별 식단과 일반 식단을 각각 섭취한 쥐 그룹의 무게 분포 (Source: Statquest, Statistical Power, Cleary Explained!)

검정력 분석이란?

특정 수준의 검정력을 갖기 위해 얼마나 많은 표본을 수집해야하는지 알려주는 분석 방법론

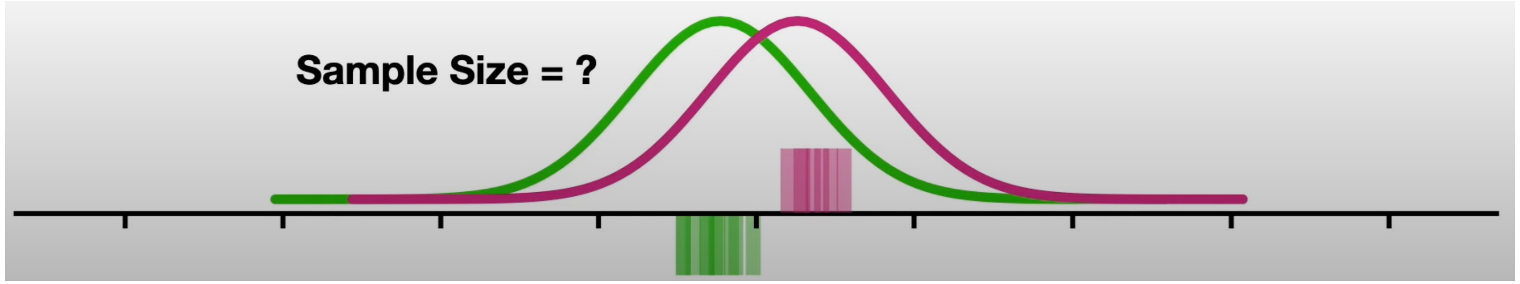


그림 15. 검정력 분석을 통해 얻고자 하는 것 (Source: Statquest, Power Analysis, Cleary Explained!)

산업 표준의 검정력 분석

산업 표준인 검정력$(1-\beta)$ 80%, 유의수준(α) 5% 하에 검정력 분석 식

$$n = \frac{16\sigma^2}{\delta^2}$$

자세한 유도 과정은 [Kohavi et al., 2022] 참고

산업 표준의 검정력 분석

$$n = \frac{16\sigma^2}{\delta^2}$$

여기서 포인트는

- 검출해내고 싶은 효과(δ)가 얼마인지에 따라 필요한 표본 크기는 달라진다는 것
- 과거 데이터에 기반한 관심 지표의 표준편차(σ)에 따라 필요한 표본 크기는 달라진다는 것

검정력 분석 계산 예시

$$n = \frac{16\sigma^2}{\delta^2}$$

다음의 상황을 가정

- 구매 전환율에 관한 10%의 상대적 변화를 한 번의 실험에서 검출하고 싶음
- 이때 과거 데이터(i.e. 대조군)의 구매 전환율은 3.7%였음

검정력 분석 계산 예시

$$n = \frac{16\sigma^2}{\delta^2}$$

다음의 상황을 가정

- 구매 전환율에 관한 10%의 상대적 변화를 한 번의 실험에서 검출하고 싶음
- 이때 과거 데이터(i.e. 대조군)의 구매 전환율은 3.7%였음

$$\sigma^2 = p * (1 - p) = 3.7\% \ * (1 - 3.7\%) = 3.563\%$$

$$\delta = 3.7\% \ * 10\% = 0.37\%$$

검정력 분석 계산 예시

$$n = \frac{16\sigma^2}{\delta^2}$$

다음의 상황을 가정

- 구매 전환율에 관한 10%의 상대적 변화를 한 번의 실험에서 검출하고 싶음
- 이때 과거 데이터(i.e. 대조군)의 구매 전환율은 3.7%였음

$$\sigma^2 = p * (1 - p) = 3.7\% * (1 - 3.7\%) = 3.563\%$$

$$\delta = 3.7\% * 10\% = 0.37\%$$

이에 따라 산업 표준 하의 검정력 분석에 기반해 계산된 두 그룹 각각에 필요한 표본 크기는:

$$n = \frac{16\sigma^2}{\delta^2} = \frac{16 * 3.563\%}{(0.37\%)^2} = 41{,}642$$

온라인 실험에서의 검정력 분석

이렇듯 검정력 분석에서는 관심지표의 표준편차와 MDE에 의해 표본 크기가 결정됨

$$n = \frac{16\sigma^2}{\delta^2}$$

퀴즈) 온라인 실험에서의 표본 크기는 2가지에 의존하는데, 이 2가지는 무엇일까요

온라인 실험에서의 검정력 분석

Causal-Lab

이렇듯 검정력 분석에서는 관심지표의 표준편차와 MDE에 의해 표본 크기가 결정됨

$$n = \frac{16\sigma^2}{\delta^2}$$

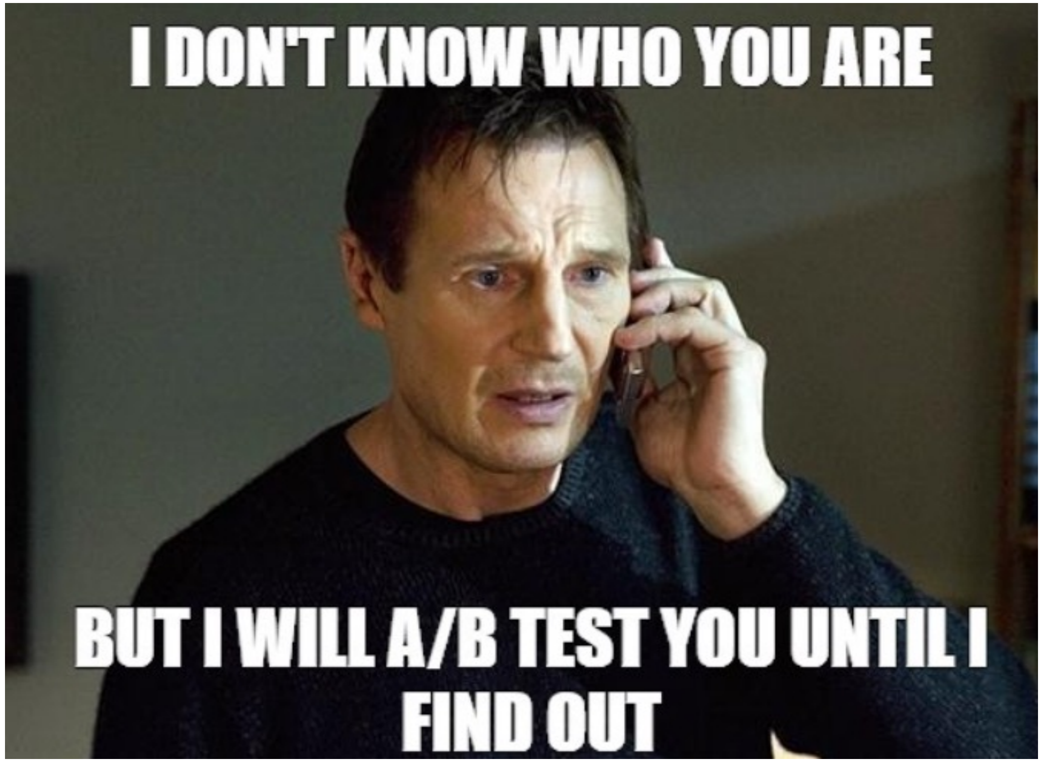
퀴즈) 온라인 실험에서의 표본 크기는 2가지에 의존하는데, 이 2가지는 무엇일까요?

1. 트래픽
2. 실험 기간

오늘 한 이야기 요약

요약

첫째, 온라인 통제 실험 (a.k.a A/B 테스트)은 고객이 진정 원하는 것을 찾기 위한 최고의 도구



Source: Meme-arsenal.com

요약

둘째, 무작위 배정은 곧 인과추론을 가능하게끔 하며 온라인 통제 실험에서는 이에 해시 함수가 이용된다.

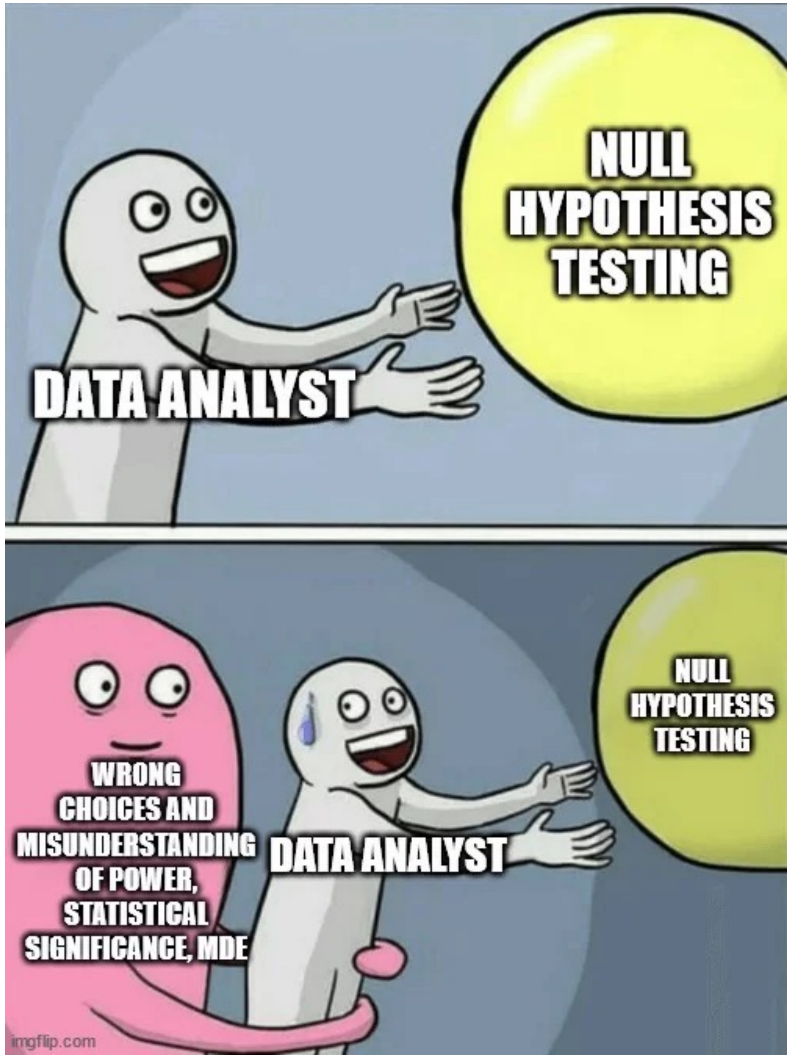
Source: Meme-arsenal.com

요약

셋째, 오늘 소개한 기초적인 통계 분석 방법론들은

실험 결과의 기본적 신뢰도를 보장하기 위해 굉장히 중요

- 검정력 80%와 유의수준 5%를 보장하기 위해 필요한 최소 표본 크기를 알려주는 검정력 분석
- 관측한 관심지표의 불확실성을 측정하는 표준오차
- 실험군과 대조군 간 관심지표의 차와 함께 불확실성까지 고려하여 통계적 의사결정을 도와주는 신뢰구간과 p-value



Source: aurimas