CHAPTER 3 통계적 실험과 유의성 검정 - 김지호

▼ 목차

- 실험설계
  - 통계 분석의 토대
  - 어떤 가설을 확인하거나 기각하기 위한 목표를 가짐

- 전형적인 통계적 추론이라는 **파이프 라인** 속에 있음.
  1. 가설을 세운다.
  2. 실험을 설계 한다.
  3. 데이터를 수집한다.
  4. 추론 및 결과를 도출한다.

3.1 A/B 검정

- A/B 검정은 실험군을 두 그룹으로 나누어 어느 쪽이 다른 쪽보다 더 우월한지 입증 하는 실험
- 두가지 처리법 중 하나는 기준이 되는 기존 방법이나 아무런 처리도 하지 않는다. 이를 대조군이라 함
- 주로 웹디자인이나 마케팅에 사용
- 이 실험의 핵심은 피험자가 어떤 특정 처리에 노출되는 것
- 측정 지표가 연속형변수, 횟수를 나타내는 변수에 따라 결과가 다르게 표시될 수 있음.

3.1.1 대조군은 왜 필요할까?

- 대조군이 없다면 다른 것들은 동일하다는 보장이 없다. / 어떤 차이가 어떤 처리 때문인지 확신할 수 없다.
- 대상은 일반적으로 웹 방문자이며, 측정하고자 하는 결과는 클릭 수, 구매 수, 방문 기간, 방문한 페이지 수, 특정 페이지 방문여부 등
- A/B 검정 실험에서는 미리 하나의 측정지표를 결정해야 함.

3.2 가설 검정

- 가설 검정, 유의성 검정은 전통적인 통계분석 방법.
- 목적 : 관찰된 효과가 우연에 의한 것인지 여부를 알아냄.
- 통계적 가설 검정은 연구자가 랜덤하게 우연히 일어난 일에 속지 않도록 보호하기 위한 방법

3.2.1 귀무가설

- 우연 때문이라는 가설. 실제로 우연히 일어난 일이지만 흔하지 않다.
- 그룹간의 차이는 우연에 의한 결과 이다. 즉, 원래는 차이가 없다.
- 귀무가설이 틀렸다는 것을 증명함.

3.2.2 대립가설

- 귀무가설과의 대조 (증명하고자 하는 가설)
- 귀무 가설과 대립가설은 모든 가설에 대해 설명
    - 예시) 귀무가설 : A ≤ B / 대립가설 : A > B

3.2.3 일원/이원 가설 검정

- 일원검정(one-way test) : 한 방향으로만 우연히 일어날 확률을 계산하는 가설검정
    - 우연에 의한 극단적인 결과에 대해 한 방향만을 고려하여 p값 계산
    - B는 A보다 낫다.
- 이원검정(two-way test) : 양방향으로 우연히 일어날 확률을 계산하는 가설검정
    - A는 B와 다르며 더 크거나 작을 수 있음.

3.3 재표본추출

- 재표본 추출
    - 목표 : 랜덤한 변동성을 알아보자
    - 의미
        - **관찰된 데이터의 값에서 표본을 반복적으로 추출하는 것**
        - 또한, 일부 머신러닝 모델의 정확성을 평가하고 향상시키는데에도 적용
        - 부트스트랩 데이터 집합을 기반으로 하는 각각의 의사 결정 트리 모델로 부터 나온 예측들로 부터 배깅이라는 절차를 통해 평균 예측값을 구할수 있다.
    - 유형
        - 부트스트랩 : 추정의 신뢰성 평가
        - 순열 검정 : 두개 이상의 그룹과 관련된 가설을 검증

3.3.1 순열 검정

- 두 개 이상의 표본을 함께 결합하여 관측값들을 무작위로 재표본으로 추출하는 과정.
    - 통상적으로 A/B 또는 기타 가설검정을 위해 사용되는 그룹들

- 순열 검정의 절차

1. 여러 그룹의 결과를 단일 데이터 집합으로 결합
2. 결합된 데이터를 잘 섞은 후, 그룹 A와 동일한 크기의 표본을 무작위로 (비복원) 추출함.
3. 나머지 데이터에서 그룹 B와 동일한 크기의 샘플을 무작위로 (비복원) 추출
4. C,D 등의 그룹 등에서도 동일한 작업 수행
5. 원래 샘플의 통계량(또는 추정치)과 지금 추출한 재표본에 대한 다시 계산하고 기록.
6. 1~5단계 R 번 반복하여 검정통계량의 순열 분포를 얻음

- 그룹간의 차이점을 관찰
    - 관찰된 차이가 순열로 보이는 차이의 집합에 들어가 있다면 우연히 일어날수 있는 범위 안에 있는 것

      → 어떠한 것도 증명할 수 없음
    - 관찰된 차이가 순열 밖에 있다면, **통계적으로 유의미**하다.(우연히 일어날 수 없다.)

3.3.3 전체 및 부트스트랩 순열 검정

- 전체 순열검정
    - 데이터를 무작위로 섞고 나누는 대신 실제로 나눌 수 있는 모든 가능한 조합을 찾는다.
    - 데이터 수가 적을 때 유리함
- 부트스트랩 순열검정
    - 순열 검정의 비복원 추출 과정을 복원 추출로 진행
    - 리샘플링 과정에서 모집단에서 개체를 선택할 때 임의성을 보장한다.

3.4 통계적 유의성과 p 값

- 통계적 유의성이랑 결과가 우연히 일어난 것인지, 우연히 일어날 수 없는 극단적인 것인지를 판단하는 방법.
- 우연의 변동성 바깥에 존재하면 통계적으로 유의하다

3.4.1 P 값

- p 값 : 통계적 유의성을 정확히 측정하기 위한 지표.
- 확률모형이 관측된 결과보다 더 극단적인 값을 생성할 빈도
- 관찰된 차이와 같거다 더 큰 차이를 보이는 경우의 비율로 p값을 추정할 수 있다.
    - p 가 **0.308**이라면 우연히 얻은 결과의 30% 정도가 관찰한 것만큼 극단적이거나 그 이상의 극단적인 결과를 얻은 것으로 기대됨.

3.4.2 유의수준

- p 값의 의미
    - 너무 많은 연구자가 어렴풋이 아는 p 값 개념으로 유의미한 p 값이 나올 때까지 온갖 가설검정을 수행.
    - 실제 p 값이 나타내는 것 **: 랜덤 모델이 주어졌을 때, 그 결과가 관찰된 결과보다 더 극단적일 확률**
    - 통계적으로 유의미하다는 근거를 가지기엔 약하다.

3.4.3 제1종과 제2종 오류

- 1종 오류 : 참을 거짓으로 판단
    - 보통은 1종 오류를 최소화하도록 가설을 설계한다.
- 2종 오류 : 거짓을 참으로 판단
    - 표본의 크기가 너무 작아서 효과를 알아낼 수 없다고 판단하는 것과 같다.

      = 아직 효과가 입증되지 않았다.

3.4.4 데이터 과학과 p 값

- p 값은 만능이 아니다.
- p값 또는 통계적 유의성은 효과의 크기나 결과의 중요성을 의미하지는 않는다.
- p값 그 자체는 모델이나 가설에 대한 증거를 측정하기 위한 좋은 지표가 아니다.

- 실험에서 의사결정을 좌우하는 도구로서 사용되선 안된다.
- 결정에 관련된 정보일 뿐.

3.5 t 검정

- *두 집단간의 평균이 통계적으로 유의미한 차이를 보이고 있는지의 여부를 검증할 때 사용되는 분석방법.*
- 데이터가 횟수나 측정값을 포함하는지, 표본이 얼마나 큰지, 측정 대상이 무엇인지에 따라 다양한 유형의 유의성 검정 방법 중 가장 많이 사용되는 것.
- 유의성 검정 방법 중 가장 자주 사용되는 t 검정(t-test)
    - 유의성 검정 : 관심있는 효과를 측정하기 위한 검정통계량을 지정, 관찰된 효과가 정상적인 랜덤 변이의 범위 내에 있는지 여부를 판단하는데 도움을 줌

데이터가 수치형인 아주 일반적인 2표본 비교(A/B 검정)에 주로 사용한다.

- 검정통계량(test statistic) : 관심의 차이 또는 효과에 대한 측정지표
    - *현실적으로 모집단 전체를 조사하기 힘들다. 해서 표본을 뽑아 표본통계량으로 계산하는데, 이 표본통계량을 가설검정에선 검정통계량이라고 한다.*
- t 통계량(t-statistic) : 표준화된 형태의 검정통계량
- t 분포(t-distribution) : 관측된 t 통계량을 비교할 수 있는, 기준분포

(정리)

1. 컴퓨터가 널리 보급되기 전에, 재표본 검정은 실용적이지 않았으며, 대신 통계학자들은 표준적인 분포를 참고했다.
2. 이렇게 하면 검정통계량이 표준화되어 참고할 분포와 비교할 수 있다.
3. 널리 사용되는 표준화된 통계량 중 하나가 t 통계량이다.

3.6 다중검정

- 통계학에서는 다양한 관점으로 데이터를 보고 충분한 질문을 던지다 보면 거의 항상 통계적을 유의한 결과가 나옴.

- 하지만 변수가 많아지거나 다양한 모델을 사용하다보면 **우연에 의한** 유의미한 결과가 나타날 확률이 높아짐.

  → *제1종 오류 : 어떤 효과가 통계적으로 유의미하다고 잘못된 결론을 내리게 됨*

- 지도 학습에서는 이런 위험을 낮추기 위해, 홀드아웃 세트를 사용함

  → 이전에 보지 못했던 데이터를 통해 모델을 평가할 수 있음

  - 교차검증방법은 여러가지 있음 ex) Holdout method, k-fold cross validation, Leave-p-out cross validation 등

- 다중검정에선 우연에 속을 기회는 더 증가한다. 단일 검정에선 A에 대해 1or0으로 가설을 세웠다면, 다중검정에선

  1번, A와 B가 서로 다른가?

  2번, A와 C가 서로 다른가?

  3번, B와 C가 서로 다른가?

  같이 다양한 질문이 생긴다.

  즉, 단일검정을 할 때보다 통계적 유의성에 대한 기준을 더 엄격하게 설정하게 된다. 더 작은 알파를.

  - *알파 : 유의수준. 임계값을 미리 설정해두어서 우연인 사건을 방지함. 많이 사용되는 값은 5%나 1%이다.*

- 다중검정에서 많은 것을 발견할 수 있는 기회:

  1. 여러 그룹 간의 쌍별 차이를 조사
  2. 여러 부분군에서의 결과를 알아보는 것(ex. 전체연령이 아니라 20대(부분군)에서 발견)
  3. 여러 가지 통계 모형을 적용
  4. 모델에서 많은 변수들을 사용하는 것
  5. 수많은 서로 다른 질문들을 묻는 것

- 하지만, **중복도**같은 일반적인 문제를 포함하여 여러 가지 이유로 더 많은 연구가 반드시 나은 연구를 의미하는 것은 아님.

3.7 자유도 d.f.(degrees of freedom)

- 자유도 : 표본 데이터에서 계산된 통계량에 적용되며 변화가 가능한 값들의 개수를 나타낸다.
    - ex. 9개의 값, 평균을 알고 있다면 10번째 값도 예상가능

- 표본크기 n : 해당 데이터에서 관측값의 개수(행 혹은 기록값의 개수와 같은 의미)
- 자유도가 중요한 이유?
    - 표본을 통해 모집단의 분산을 추정하고자 할 때 분모에 n을 사용하면 추정치가 살짝 아래쪽으로 편향될 것.
    - n-1로 하여 편향이 발생하지 않는다
    - *표본의 분산을 모집단의 분산에 근사해지게 하는 비율을 찾았는데 그것이 바로 n/(n-1).*
    - *이를 표본의 분산에 n/(n-1)만큼 곱하면 모집단의 분산에 근사하게 된다.*
- 자유도는 표준화된 데이터가 그에 적합한 기준 분포(t분포, F분포 등)에 맞도록 하기 위한 표준화 계산의 일부.
- 하지만 데이터 과학에서는 유의성 검정 측면에서 중요하지 않다.
    - cf) 완전히 불필요한 예측변수들이 있는 경우 회귀 알고리즘 사용하기 어렵다.
        - ex) 일주일에 7일이 있지만 요일을 지정할때 자유도는 6일이다.

3.8 분산분석

- 분산 분석 (analysis of variance, ANOVA) : 여러 그룹간의 통계적 유의미한 차이를 검정하는 통계적 절차.
    - 여러그룹(ex : A-B-C-D)의 수치를 서로 비교

      ex. 4개의 페이지로 이루어진 웹페이지에 5명의 사용자가 방문한 페이지
    - 한 쌍씩 비교하게 되면 우연히 일어난 일에 속을 가능성이 커짐
    - ''원래 4개 페이지에 할당된 세션시간이 무작위로 할당된 것인가?"라는 질문을 다루는 **총괄검정** 필요

- 쌍별비교(pairwise comparison) : 여러 그룹 중 두 그룹 간의 가설검정
- 총괄검정(omnibus test) : 여러 그룹 평균들의 전체 분산에 관한 단일 가설검정
- 분산분해(decomposition of variance) : 구성 요소 분리. 예를 들면 전체 평균, 처리 평균, 잔차 오차로부터 개별값들에 대한 기여를 뜻함
- SS(sum of squares) : 어떤 평균으로부터의 편차들의 제곱합

- ANOVA 기반의 재추출 과정
  1. 모든 데이터를 한 상자에 담아 놓음.
  2. 5개의 값을 갖는 4개의 재표본을 섞어서 추출
  3. 각 그룹의 평균을 기록
  4. 네그룹 평균 사이의 분산을 기록
  5. 2~4단계 여러번 반복

3.8.1 F 통계량

- F 통계량(F-statistic) : 그룹 평균 간의 차이가 랜덤 모델에서 예상되는 것보다 벗어나는 정도를 측정하는 표준화된 통계량
  - 비율이 높을수록 통계적으로 **유의미**
  - 잔차 오차로 인한 분산과 그룹 평균(처리 효과)의 분산에 대한 비율을 기초로함.
- F통계량을기반으로 한 ANOVA 통계 검정도 있음

3.8.2 이원 분산분석

- 위의 사례 A-B-C-D 검정은 변하는 요소(그룹)가 하나인 **일원ANOVA** 이다.
  - ex) A vs B, C vs D
- 두가지 요소를 고려하여 분석하기 위해선 **이원 ANOVA** 가 필요함
  - A(주말-토요일, 일요일) vs B(평일-월,화,수,목,금)

정리

1. ANOVA는 여러 그룹의 실험 결과를 분석하기 위한 통계적 절차
2. A/B 검정과 비슷한 절차를 확장하여 그룹 간 전체적인 편차가 우연히 발생할 수 있는 범위 내에 있는지를 평하기 위해 사용한다.
3. ANOVA의 결과 중 유용한 점 중 하나는 그룹 처리, 상호작용 효과, 오차와 관련된 분산의 구성 요소들을 구분하는 데 있다.

3.9 카이제곱검정

- 카이제곱 검정(chi-square test) : 횟수 관련 데이터에 주로 사용, 예상되는 분포에 얼마나 잘 맞는지 검정.
    - 단순 A/B 검정을 넘어 동시에 여러 가지 처리를 한 번에 테스트할 필요가 있다.
    - ex. 웹 테스트시
- 일반적으로 변수 간 독립성에 대한 귀무가설이 타당한지 평가하기 위해 r*c 분할표 함께 사용
    - r과 c는 각각 행과 열의 수 의미

3.9.1 카이제곱검정: 재표본추출 방법

- 피어슨 잔차, R : 실제 횟수와 기대한 횟수 사이의 차이를 나타냄.
- 카이제곱통계량 : 피어슨 잔차들의 제곱합

3.9.2 카이제곱검정: 통계적 이론

- 점근적 통계 이론은 카이제곱통계량의 분포가 카이제곱분포로 근사화될 수 있음을 보여줌.
- 적절한 표준 카이제곱분포는 **자유도**에 의해 결정
    - 자유도 = (r-1) * (c-1)
- 카이제곱분포는 일반적으로 한쪽으로 기울어져 있고, 오른쪽 긴 꼬리가 있다.

3.9.3 피셔의 정확검정

- 대부분의 통계 소프트웨어는 발생할 수 있는 모든 조합을 실제로 열거하고, 빈도를 집계하고, 관찰된 결과가 얼마나 극단적으로 발생할 수 있는지를 결정하는 절차를 제공한다.

  이를 피셔의 정확검정이라한다.

3.9.4 데이터 과학과의 관련성

- 카이제곱검정이나 피셔의 정확검정은 통계적 유의성을 조사하는 것으로 데이터과학과의 직접적인 연관성을 찾기가 어렵다. 따라서 최적의 처리 방법을 찾는 멀티암드 밴딧 방법이 더 정확한 해결책이라 할 수 있겠다.
- 데이터과학 응용 분야에선, **카이제곱 검정**이나 재표본추출 시뮬레이션을 필터로 사용.

  → 즉, 어떤 효과나 특징에 대해 기본적인 유의성 검정을 넘어 더 심층적인 분석이 필요할지 여부를 결정한다.
- 머신러닝에서는 자동으로 특징을 선택하기 위해 사용한다.

3.10 멀티암드 밴딧 알고리즘

- 멀티암드 밴딧(multi-armed bandit) : 고객이 선택할 수 있는 손잡이가 여러 개인 가상의 슬롯머신을 말하며, 각 손잡이는 각기 다른 수익을 가져다준다. 다중 처리 실험에 대한 비유라고 생각할 수 있다.
  - 손잡이(arm) : 실험에서 어떤 하나의 처리를 말한다.
  - 상금(수익) : 슬롯머신으로 딴 상금에 대한 실험적 비유

- 전통적인 통계적 접근 방식보다 명시적인 최적화와 좀 더 빠른 의사 결정이 목표
  - 특히 웹테스트에 사용
- 밴딧 알고리즘은 하이브리드 접근 방식을 취한다.

  ▼ 예시

  손잡이 A: 50번 중 10번 승리

  손잡이 B: 50번 중 2번 승리

  손잡이 C: 50번 중 4번 승리

A를 더 자주 잡아당기는 걸로 시작하지만 B와 C를 포기하지 않는다. A의 성과가 꾸준히 우수하다면 A에 기회를 더 많이 주겠지만, 만일 C가 더 좋아진다면 C의 기회를 더 늘리는 식으로 바꾼다.

→ A의 우위를 활용하기 위해 검증하고 나머지 B, C도 포기하지 않는다.

- 적용 알고리즘
  - 앱실론-그리디 알고리즘
    - 앱실론 : 알고리즘을 제어하는 단일 파라미터
      - 앱실론이 1이면 표준 A/B 검정
      - 앱실론이 0이면 탐욕 알고리즘
  - 톰슨의 샘플링
    - 베이지언 방식 사용
    - 베타분포(사전 정보)를 사용하여 수익의 일부 사전 부포를 가정함.

- 전통적 A/B 검정은 임의표집 과정을 기본으로 하기 때문에 수익이 낮은 것을 너무 많이 시도하게 된다.
- 이와 달리 MAB는 실험 도중에 얻은 정보를 통합하고 수익이 낮은 것의 빈도르 줄이는 쪽으로 표본 추출과정을 변경한다.
- 또한 두 가지 이상의 처리를 효과적으로 다룰 수 있다.
- 추출 확률을 수익이 낮은 처리에서 수익이 높으리라 추정되는 쪽으로 이동시키기 위한 다양한 알고리즘이 존재한다.

3.11 검정력과 표본 크기

- 실험 진행시 표본크기에 대한 고려가 중요하다.
- 표본 크기에 대한 고려는 실제로 A와 B의 차이를 밝혀낼 수 있을지에 대한 질문과 연결된다.
  - p값(가설검정의 결과)은 A와 B의 차이에 따라 달라진다.
- A,B 의 차이가 작을 수록 더 많은 데이터가 필요하다.

- 검정력 : 특정 표본 조건(크기와 변이)에서 특정한 효과크기를 알아낼 수 있을 확률을 의미

3.11.1 표본크기

- 검정력 계산의 주된 용도는 표본크기가 어느 정도가 필요한가를 추정하는 것
- 작은 차이에도 관심이 있다면 훨씬 큰 표본이 필요함.
  - **효과크기**가 표본크기를 좌우함.
  - 효과크기(effect size) : 통계 검정을 통해 판단할 수 있는 효과의 최소

    ex. 3할 3푼 타자vs 2할 타자라면 0.33-0.2=0.13이 효과크기

- **검정력 혹은 표본크기**의 계산과 관련된 다음 4가지 중요한 요소
  - 표본크기
  - 탐지하고자 하는 효과크기
  - 가설검정을 위한 유의수준
  - 검정력
- 위의 3가지를 정하면 나머지 하나를 알 수 있다.

정리

1. 통계 검정을 수행하기 앞서, 어느정도의 표본크기가 필요한지 미리 생각할 필요가 있다.
2. 알아내고자 하는 효과의 최소 크기를 지정해야 한다.
3. 또한 효과크기를 알아내기 위해 요구되는 확률(검정력)을 지정해야 한다.
4. 마지막으로, 수행할 가설검정에 필요한 유의수준을 정해야 한다.

▼ 3.10 멀티암드 밴딧 알고리즘

💡 전통적인 통계적 접근 방식보다 **최적화와 좀 더 빠른 의사 결정**을 가능하게 하며, 여러 테스트, 특히 **웹 테스트**를 위해 사용된다.

용어 정리

- **멀티암드 밴딧(MAB)** multi-armed bandit : 고객이 선택할 수 있는 손잡이가 여러 개인 가상의 슬롯머신을 말하며, 각 손잡이는 각기 다른 수익을 가져다준다. 다중 처리 실험에 대한 비유라고 생각할 수 있다.
- **손잡이** arm : 실험에서 어떤 하나의 처리를 말한다(예를 들면 '웹 테스트에서 헤드라인 A').
- **상금(수익)** win : 슬롯머신으로 딴 상금에 대한 실험적 비유(예를 들면 '고객들의 링크 클릭 수')

- 전통적인 A/B 검정의 단점:
    - 실험 결과를 통해 효과가 있다는 것을 '유추' 할 수 있지만 '입증'할 만한 증거가 없을 수 있다.
    - 실험이 끝나기 전에 이미 얻은 결과들을 이용하기 시작할 수도 있다.
    - 실험의 목적이 변할 수 있다.
    - 비즈니스 전반에서는 통계적 유의성보다는 비용과 결과를 최적화하는데 더 관심이 있다.

▼ 예시: 슬롯머신

- 손잡이 A: 50번 중 10번 승리
- 손잡이 B: 50번 중 2번 승리
- 손잡이 C: 50번 중 4번 승리

해석

- 결과를 단순히 보면 손잡이 A가 최고로 보인다.
    - A가 정말 우월하면 초기에 이익을 얻을 수 있다.
    - 아니라면 다른 사실을 발견할 기회를 놓친다.
- 다른 극단적인 접근법: `'모두가 무작위이니 모두 똑같이 잡아당기자'`
    - B와 C를 포기하지 않지만 수익이 낮을 것으로 예상되는 행위를 자주 취해야한다.
- 밴딧 알고리즘은 하이브리드 접근 방식
    - A를 더 자주 잡아당기지만 B와 C를 당길 기회를 A에게 더 부여한다.

▼ 알고리즘을 위한 파라미터: 엡실론-그리디 알고리즘

1. 0부터 1 사이의 균등분포의 난수 생성
2. 숫자가 0과 엡실론(0과 1 사이의 값) 사이에 존재하면, 50/50의 확률로 동전 뒤집기 시행
    a. 그 결과 동전이 앞면이면 제안 A 표시
    b. 동전이 뒷면이면 제안 B 표시
3. 숫자가 엡실론보다 크면, 지금까지 가장 좋은 결과를 보인 제안 표시

▼ **입실론-그리디 알고리즘, 톰슨 샘플링** 에 대해서 참고하면 좋은 링크 입니다.

1. **입실론-그리디 알고리즘 :** https://brunch.co.kr/@chris-song/62

    포스팅 중반부에 입실론-그리디 알고리즘 관련 설명이 나와있습니다!

    처음부터 읽으시면 이해가 더 쉽습니다.

2. **톰슨 샘플링** : https://brunch.co.kr/@chris-song/66