

06 데이터 분석 심화 - 확증적 데이터 분석(CDA)

AI 에이전트 개발

데이터 분석 심화

원티드랩

- 소개
 - [가설 검정\(Hypothesis Testing\)](#)
- 1. 정규성 검정
 - [1\) 가설 및 판단 기준](#)
 - [2\) 검정 방법](#)
- 2. 평균 검정
 - [1\) 1개 집단 평균 검정](#)
 - [2\) 독립표본 평균 검정](#)
 - [3\) 대응표본 평균 검정](#)
 - [4\) ANOVA \(세 집단 이상 평균 비교\)](#)
- 3. 독립성 검정

소개

CDA(Confirmatory Data Analysis)는 이미 수집된 데이터를 바탕으로 사전에 세운 가설을 검증하는 분석이다. 지금까지 EDA를 통해 데이터의 특징을 발견하는 과정이었다면, CDA는 검증을 위한 과정인 것이다.

- A/B 테스트로 광고 문구 효과 검증
- 기능 추가 전/후 이탈률 비교
- 수업 방식의 점수 차이 검증
- 신약의 효과가 기존보다 우수한지에 대한 검증

가설 검정(Hypothesis Testing)

가설검정은 어떤 주장(가설)이 통계적으로 유의미한지를 확인하는 과정이다.

- 귀무가설(H_0): "차이가 없다", "효과가 없다", "변화가 없다"는 기본 가정
- 대립가설(H_1): "차이가 있다", "효과가 있다", "변화가 있다"는 주장

우리는 항상 귀무가설을 기각할 수 있는지를 검정한다.

① 검정통계량(Test Statistics)

데이터를 기반으로 계산된 검정통계량은 귀무가설이 맞는 상황에서 어느 정도의 차이가 있는지를 수치화한 값
예: 평균 차이를 검정할 때 사용하는 t값, 집단 간 분산 차이를 보는 F값 등

② 유의확률(p-value)

- 귀무가설이 참일 때, 현재 관측된 통계량과 같거나 더 극단적인 결과가 나올 확률
- 일반적으로 $p < 0.05$ 이면, 유의수준 5% 기준으로 귀무가설을 기각 → "차이가 있다"고 판단한다.

1. 정규성 검정

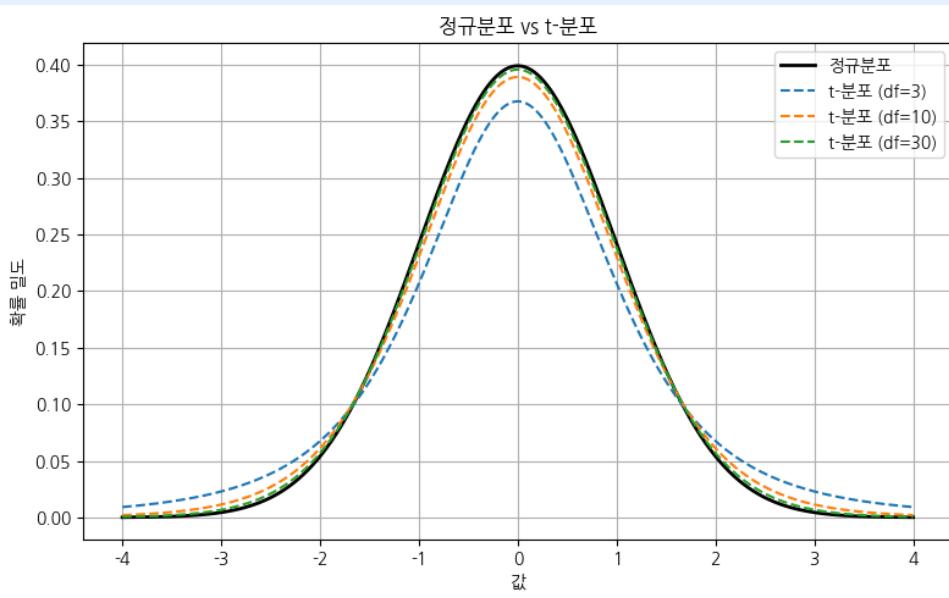
통계 분석의 많은 기법들은 데이터가 정규분포를 따른다는 전제 하에 사용된다.

따라서 분석 전에 변수가 정규분포를 따르는지 확인하는 과정이 필요하다.

① 분포(Distribution)란?

데이터가 어떻게 퍼져 있는지를 보여주는 형태이다.

정규분포는 종 모양의 대칭 형태로, 평균을 중심으로 고르게 퍼져있다.



1) 가설 및 판단 기준

- 귀무가설(H_0): 이 데이터는 정규분포를 따른다.
- 대립가설(H_1): 이 데이터는 정규분포를 따르지 않는다.
- $p\text{-value} < 0.05 \rightarrow$ 귀무가설 기각 \rightarrow 정규분포가 아니다

2) 검정 방법

- Shapiro-Wilk 검정: 소규모 데이터에서 사용
- Kolmogorov-Smirnov 검정: 대규모 데이터에서 사용

2. 평균 검정

하나 이상의 집단 간에 평균의 차이가 유의미한지 검정한다.

1) 1개 집단 평균 검정

가정

- 연속성 데이터야 한다.
- 데이터는 정규분포를 따른다.

가설 및 판단 기준

- 귀무가설(H_0): 모집단의 평균은 특정 값과 같다.
- 대립가설(H_1): 모집단의 평균은 특정 값과 다르다.
- 예시: “우리 반 수학 점수의 평균은 70점이다”
- $p\text{-value} < 0.05 \rightarrow$ 귀무가설 기각 \rightarrow 모집단의 평균은 특정 값과 다르다

2) 독립표본 평균 검정

가정

- 두 집단이 서로 독립이어야 한다.
- 각 집단은 정규분포를 따르는 연속형 데이터이다.
- 등분산성의 여부에 따라 분석 방법이 다르다.

가설 및 판단 기준

- 귀무가설(H_0): 두 집단의 평균은 같다.
- 대립가설(H_1): 두 집단의 평균은 다르다.
- 예시: “A반과 B반의 시험 평균은 차이가 없다” → 이 가정이 맞는지 검정
- $p\text{-value} < 0.05$ → 귀무가설 기각 → 두 집단의 평균은 다르다

3) 대응표본 평균 검정

가정

- 동일한 대상을 두 번 측정한 데이터이다.
- 두 값의 차이가 정규분포를 따라야 한다.

가설 및 판단 기준

- 귀무가설(H_0): 전과 후의 평균은 같다.
- 대립가설(H_1): 전과 후의 평균은 다르다.
- 예시: “운동 전과 후의 체중 변화”가 유의한지 검정
- $p\text{-value} < 0.05$ → 귀무가설 기각 → 전과 후의 평균이 다르다

4) ANOVA (세 집단 이상 평균 비교)

가정

- 각 집단은 서로 독립이어야 한다.
- 각 집단은 정규분포를 따르는 연속형 데이터이다.
- 집단 간 분산이 서로 같아야 한다(등분산성).

가설 및 판단 기준

- 귀무가설(H_0): 모든 집단의 평균은 같다.
- 대립가설(H_1): 적어도 하나의 집단 평균은 다르다.
- $p\text{-value} < 0.05$ → 귀무가설 기각 → 적어도 하나의 집단 평균은 다르다

3. 독립성 검정

가정

- 두 집단은 범주형 데이터이다.
- 각 셀 기대도수(예상 빈도수)가 충분히 커야 한다.

가설 및 판단 기준

- 귀무가설(H_0): 두 변수는 서로 독립이다 (\rightarrow 관련 없다).
- 대립가설(H_1): 두 변수는 서로 독립이 아니다 (\rightarrow 관련 있다).

- 예시: “성별과 구매 여부는 관련이 없다” → 이 가정이 맞는지 검정
 - p-value < 0.05 → 귀무가설 기각 → 두 변수는 관련이 있다