의료 데이터 통계검정 실습

개념 → 가설 → 예시 → 실습 → 해석

> 데이터 파일: train.csv
> Outcome: 질환 여부(0=없음, 1=있음)

환경 준비 및 패키지 설치

다음 패키지를 설치합니다.

```
pip install pandas numpy scipy matplotlib
```

> 필수: pandas, numpy, scipy, matplotlib
> 선택: statsmodels (ANOVA 사후검정 등 고급기능)

Step 0. 데이터 불러오기와 EDA

```
import pandas as pd

df = pd.read_csv("train.csv")
print(df.shape)
df.head()
```

```
df.info()
df.describe(numeric_only=True)
df['Outcome'].value_counts(dropna=False)
```

> EDA: 본격 분석 전 결측, 분포, 범위를 확인하는 과정

의료 데이터 주의: 0 값을 결측으로 처리

```
zero_cols = ["Glucose","BloodPressure","SkinThickness","Insulin","BMI"]
df[zero_cols] = df[zero_cols].replace(0, pd.NA)
df[zero_cols].isna().mean().sort_values(ascending=False)
```

결측치(NA): 미측정값을 의미. 0을 실제값으로 두면 평균·상관이 왜곡될 수 있음

Part A. t-test (독립 2집단 평균 비교)

t-test 소개와 예시

두 독립 집단의 평균을 비교할 때 사용하는 검정.

예시

- Outcome 0 vs 1의 Glucose 평균 차이
- 약 A vs 약 B의 체온 평균 차이

> 독립: 한 집단의 데이터가 다른 집단에 영향을 주지 않음

A-1) 가설 설정

가설	설명
H0(귀무가설)	두 집단의 평균에 차이가 없다
H1(대립가설)	두 집단의 평균에 차이가 있다

> 왜 t-test인가: 연속형 변수(Glucose)를 이분집단(Outcome)으로 나눈 두 집단 평균 비교이기 때문

A-2) 실습: Outcome별 Glucose 평균 차이

```
from scipy.stats import ttest_ind

df_t = df[["Outcome","Glucose"]].dropna()
g0 = df_t[df_t.Outcome==0]["Glucose"]
g1 = df_t[df_t.Outcome==1]["Glucose"]

stat, p = ttest_ind(g0, g1, equal_var=False)  # Welch t-test
print("t-stat:", stat, "p-value:", p)
g0.mean(), g1.mean()
```

> p-value: 이런 차이가 우연히 발생할 확률. 기준 $p < 0.05$이면 유의한 차이로 본다

> p-value: 1.3×10^{-35}

- 0.05보다 훨씬 작음
- 이 차이가 우연히 나올 확률 = 0에 가깝다
- 사실상 통계적으로 매우 유의하다

p-value	해석
0.05 ~ 0.01	유의함
0.01 ~ 0.001	매우 유의함
< 0.0001	극도로 유의함

A-3) 효과크기(Cohen’s d)

Cohen’s d란?

> “두 집단 평균 차이가 실제로 얼마나 큰가?”를
> 표준편차 기준으로 나타낸 값

```
import numpy as np

sd_pooled = np.sqrt((g0.var(ddof=1) + g1.var(ddof=1)) / 2)
d = (g1.mean() - g0.mean()) / sd_pooled
print("Cohen's d:", d)
```

효과크기: 실제 차이의 크기를 수치로 표현. 대략 0.2=작음, 0.5=중간, 0.8=큼

Cohen’s d	해석	실제 의미 비유
0.2	작은 효과	차이가 있지만 실생활에선 미미
0.5	중간 효과	확실히 차이를 느낄 정도
0.8	큰 효과	차이가 꽤 큰 편
1.0 이상	매우 큰 효과	두 집단이 거의 다른 종족처럼 구분됨

> d=1.15
> 당뇨 유무에 따라 혈당은 단순한 통계적 차이가 아니라 임상적으로도 큰 차이

Part B. 상관검정 (연속형과 연속형)

상관 소개와 예시

두 연속형 변수가 함께 움직이는 정도를 측정.

예시

- BMI와 BloodPressure의 동반 상승 여부

> 상관계수 범위: -1에서 1. 0에 가까울수록 관계가 약함. 상관은 인과를 의미하지 않음

B-1) 실습: Pearson와 Spearman

Pearson(피어슨) 상관계수란?

- 두 연속형 변수가 직선적으로 함께 변화하는 정도를 측정한 값
- BMI가 높아지면 혈압도 함께 올라간다?
  → 이럴 때 피어슨 상관계수를 씀

Spearman(스피어만) 상관계수란?

- 두 변수 간의 단조적 관계(Monotonic Relationship) 를 측정하는 상관계수
  즉, 순위(서열) 기반으로 관계를 평가

```python
from scipy.stats import pearsonr, spearmanr

pair = df[["BMI", "BloodPressure"]].dropna().copy()
pair = pair.apply(pd.to_numeric)

r, p = pearsonr(pair["BMI"], pair["BloodPressure"])
rho, p2 = spearmanr(pair["BMI"], pair["BloodPressure"])

print("Pearson r:", r, "p:", p)
print("Spearman rho:", rho, "p:", p2)
```

> 피어슨: 선형 관계에 민감. 스피어만: 순위 기반 단조 관계에 민감

결과

분석	상관계수	p-value	해석
Pearson r	**0.293**	2.97e-14	**선형적 양의 상관**, 통계적으로 매우 유의
Spearman rho	**0.312**	4.63e-16	**단조적 양의 상관**, 통계적으로 매우 유의

절대값	해석
0.1~0.3	약한 상관
0.3~0.5	중간 정도
> 0.5	강한 상관

> BMI가 높아질수록 혈압도 함께 증가하는 경향

Part C. 카이제곱 검정(χ^2) (범주형과 범주형)

χ^2 소개와 예시

두 범주형 변수 간 독립성 여부를 검정.

예시

- Glucose를 구간화한 고혈당 여부와 Outcome의 관련성

> 연속형 변수를 구간화하여 범주형으로 변환하면 χ^2 검정 적용 가능

C-1) 실습: Glucose 구간화와 교차표

```
cut = pd.cut(df["Glucose"], bins=[-float("inf"), 125, float("inf")],
             labels=["정상~전단계","고혈당"], include_lowest=True)

tab = pd.crosstab(cut, df["Outcome"])
print(tab)
```

C-2) 실습: χ^2 독립성 검정

```
from scipy.stats import chi2_contingency

chi2, p, dof, expected = chi2_contingency(tab.fillna(0))
print("chi2:", chi2, "p-value:", p, "dof:", dof)
```

> H0(귀무가설): 두 변수는 독립
>
> 판정: p < 0.05면 독립이 아님(관련 있음)

결과

χ^2 값	해석
~3.84	p ≈ 0.05 (유의 기준선)
~10.8	p ≈ 0.001 (상당히 유의)
> 20	매우 큼, 유의성 매우 높음
> 100	거의 완벽한 차이

> $p \approx 9 \times 10^{-26}$

이런 결과가 우연히 나올 확률이 사실상 없다

즉, 두 변수는 독립이 아니다(관련 있다)

결과 해석 체크리스트

1) p-value: 우연설을 얼마나 의심할 수 있는가
2) 효과크기: 실제 차이의 크기는 어느 정도인가
3) 표본 크기: 표본이 매우 크면 작은 차이도 유의해질 수 있음
4) 맥락: 임상적 의의, 업무적 의사결정과 연결

최종 요약

질문	변수 유형	검정법
Outcome별 Glucose 평균 비교	연속 + 이분집단	t-test
BMI그룹별 BloodPressure 비교	연속 + 3개 이상 집단	ANOVA
BMI ↔️ BloodPressure 관계	연속 + 연속	상관(피어슨/스피어만)
Glucose구간 × Outcome 연관	범주 + 범주	χ^2 독립성