카이제곱 검정 (Chi-squared Test)

- 관찰된 빈도가 기대되는 빈도와 통계적으로 유의미하게 다른지를 검증하기 위해 사용되는 통계적 방법
- 주로 범주형 데이터 분석에 사용되며, 분석 목적에 따라 적합도 검정, 독립성 검정, 동질성 검정으로 나뉨
- **귀무가설 (H0)**: 관찰된 빈도와 기대 빈도 사이에 차이가 없다.
    - 두 변수는 서로 독립이다 / 각 집단의 비율은 동일하다.
- **대립가설 (H1)**: 관찰된 빈도와 기대 빈도 사이에 유의미한 차이가 있다.
    - 두 변수는 서로 독립이 아니다 / 각 집단의 비율은 동일하지 않다.
- 검정 통계량 공식: `Σ (관찰 빈도 - 기대 빈도)² / 기대 빈도`
- 검정 통계량이 카이제곱 분포를 따르는 것을 이용하여 가설을 검정

적용 가능한 상황

- **적합도 검정 (Goodness of Fit Test)**: 하나의 범주형 변수에 대해 관찰된 빈도 분포가 이론적인(기대하는) 분포와 일치하는지 검정할 때.
    - 예: 주사위를 60번 던졌을 때 각 눈이 나온 횟수가 기대치(10번)와 일치하는가?
- **독립성 검정 (Test of Independence)**: 두 개의 범주형 변수가 서로 연관성이 있는지(독립적인지) 검정할 때.
    - 예: 흡연 여부와 폐암 발병 여부 사이에 연관성이 있는가?
- **동질성 검정 (Test of Homogeneity)**: 서로 다른 모집단에서 추출된 표본들의 특정 범주형 변수에 대한 비율이 동일한지 검정할 때.
    - 예: 두 개의 다른 도시에서 특정 정당에 대한 지지율이 동일한가?
    - 독립성 검정과 계산 과정은 동일하지만, 연구 설계와 가설 설정의 관점에서 차이가 있습니다.

구현 방법

`scipy.stats.chi2_contingency` (독립성/동질성 검정)와 `scipy.stats.chisquare` (적합도 검정) 함수가 주로 사용됩니다.

주의사항 (가정)

- **빈도 데이터**: 데이터는 각 범주에 해당하는 빈도(frequency) 형태여야 합니다.
- **기대 빈도**: 각 셀(cell)의 기대 빈도는 5 이상이어야 한다는 가정이 일반적으로 통용됩니다. 만약 기대 빈도가 5 미만인 셀이 전체의 20%를 초과하면, 카이제곱 검정의 정확도가 떨어질 수 있습니다.
    - 이 경우, **피셔의 정확 검정(Fisher's Exact Test)**을 대안으로 사용하거나, 관련 범주를 통합하여 기대 빈도를 높이는 방법을 고려해야 합니다.
- **독립성**: 각 관측치는 서로 독립적이어야 합니다.

---

1. 적합도 검정 (Goodness of Fit Test)

코드 예시

`scipy.stats.chisquare(f_obs, f_exp=None, ddof=0, axis=0)`

하이퍼파라미터 (인자) 설명

- `f_obs`: array_like. 관찰된 빈도.
- `f_exp`: array_like, optional. 기대 빈도. 만약 `None`이면, 모든 범주에서 동일한 빈도를 기대한다고 가정합니다.
- `ddof`: int, optional. 자유도(Delta Degrees of Freedom). 최종 자유도는 `k-1-ddof`로 계산됩니다 (k는 범주 수). 모수 추정에 사용된 제약 조건의 수를 의미합니다.

```
import numpy as np
from scipy.stats import chisquare

# 예시: 한 해 동안 특정 상점에서 요일별 방문자 수가 균일한지 검정
# 귀무가설: 요일별 방문자 수는 차이가 없다 (모두 동일하다).
# 대립가설: 요일별 방문자 수는 차이가 있다.

# 관찰 빈도 (월, 화, 수, 목, 금, 토, 일)
observed_freq = np.array([100, 120, 110, 105, 140, 155, 150])

# 기대 빈도: 총 방문자 수를 요일 수(7)로 나눈 값
total_visitors = np.sum(observed_freq)
expected_freq = np.full_like(observed_freq, total_visitors / 7)

# 카이제곱 적합도 검정 수행
# f_exp를 명시하지 않으면 자동으로 균일 분포를 가정하여 계산합니다.
statistic, p_value = chisquare(f_obs=observed_freq)

print(f"Chi-squared statistic: {statistic:.4f}") # 24.0341
print(f"P-value: {p_value:.4f}")                 # 0.0005

# 결과 해석: "귀무가설 기각: 요일별 방문자 수에는 통계적으로 유의미한 차이가 있습니다."
alpha = 0.05
if p_value < alpha:
    print("귀무가설 기각: 요일별 방문자 수에는 통계적으로 유의미한 차이가 있습니다.")
else:
    print("귀무가설 채택: 요일별 방문자 수는 균일하다고 볼 수 있습니다.")
```

결과 해석 방법

- **Chi-squared statistic**: 검정 통계량. 관찰 빈도와 기대 빈도의 차이를 나타내는 값입니다.
- **P-value**: 귀무가설이 참일 때, 현재와 같은 검정 통계량 또는 더 극단적인 값이 나올 확률입니다.
    - `p-value < 유의수준`: 관찰된 빈도 분포가 기대 분포와 유의미하게 다르다고 결론 내립니다.

---

2. 독립성/동질성 검정 (Test of Independence/Homogeneity)

코드 예시

`scipy.stats.chi2_contingency(observed, correction=True, lambda_=None)`

하이퍼파라미터 (인자) 설명

- `observed`: array_like. 분할표(contingency table) 형태의 관찰 빈도 데이터.
- `correction`: bool, optional (기본값: `True`). **예이츠의 연속성 수정(Yates' correction for continuity)** 적용 여부. 2x2 분할표에서 카이제곱 분포의 근사 정확도를 높이기 위해 사용됩니다. 자유도가 1일 때만 적용됩니다.
- `lambda_`: float or str, optional. `Power-divergence` 통계량을 계산할 때 사용되는 파라미터.
  - `lambda_='pearson'`: 피어슨 카이제곱 통계량 (기본값, 1)
  - `lambda_='log-likelihood'`: G-test (우도비 검정) (0)

```
import pandas as pd
from scipy.stats import chi2_contingency

# 예시: 교육 수준과 소득 수준 간의 연관성(독립성) 검정
# 귀무가설: 교육 수준과 소득 수준은 서로 독립이다.
# 대립가설: 교육 수준과 소득 수준은 서로 연관되어 있다.

# 데이터 (분할표)
data = {
    '고졸': [50, 70, 30],
    '대졸': [80, 120, 50],
    '대학원졸': [40, 90, 70]
}
index = ['소득 하', '소득 중', '소득 상']
observed_table = pd.DataFrame(data, index=index)

print("관찰 빈도 분할표:")
print(observed_table)
'''
관찰 빈도 분할표:
         고졸   대졸  대학원졸
소득 하  50     80    40
소득 중  70    120    90
소득 상  30     50    70
'''

# 카이제곱 독립성 검정 수행
chi2, p_value, dof, expected = chi2_contingency(observed_table)

print(f"Chi-squared statistic: {chi2:.4f}") # 19.6261
print(f"P-value: {p_value:.4f}")            # 0.0006
print(f"Degrees of Freedom: {dof}")         # 4
print("기대 빈도 분할표:")
print(pd.DataFrame(expected, index=index, columns=data.keys()))
'''
기대 빈도 분할표:
         고졸   대졸       대학원졸
소득 하  42.5   70.833333  56.666667
소득 중  70.0  116.666667  93.333333
소득 상  37.5   62.500000  50.000000
'''
```

```
# 결과 해석: "귀무가설 기각: 교육 수준과 소득 수준은 통계적으로 유의미한 연관이 있습니다."
alpha = 0.05
if p_value < alpha:
    print("귀무가설 기각: 교육 수준과 소득 수준은 통계적으로 유의미한 연관이 있습니다.")
else:
    print("귀무가설 채택: 교육 수준과 소득 수준은 서로 독립이라고 볼 수 있습니다.")
```

결과 해석 방법

- **chi2**: 카이제곱 검정 통계량.
- **p_value**: p-value.
    - `p-value < 유의수준`: 두 변수가 서로 연관되어 있다고 결론 내립니다 (독립성 검정). 또는, 집단 간 비율이 다르다고 결론 내립니다 (동질성 검정).
- **dof**: 자유도. `(행의 수 - 1) * (열의 수 - 1)`로 계산됩니다.
- **expected**: 기대 빈도 행렬. `chi2_contingency` 함수가 반환하는 값으로, 각 셀의 기대 빈도를 확인할 수 있습니다. 이를 통해 기대 빈도가 5 미만인 셀이 있는지 확인할 수 있습니다.

장단점 및 대안

장점

- **간단하고 빠름**: 계산이 비교적 간단하고 해석이 용이합니다.
- **넓은 적용 범위**: 다양한 범주형 데이터 분석에 활용될 수 있습니다.
- **가정이 적음**: 정규성 같은 강력한 분포 가정이 필요 없습니다.

단점

- **기대 빈도 제약**: 기대 빈도가 너무 작으면 검정의 신뢰도가 떨어집니다.
- **연관성의 강도 측정 불가**: 두 변수 간의 연관성 유무는 알려주지만, 연관성의 강도나 방향을 직접적으로 알려주지는 않습니다.
    - 연관성의 강도를 측정하려면 **크래머의 V(Cramér's V)**나 **파이 계수(Phi coefficient)** 같은 추가적인 측정이 필요합니다.
- **결과의 구체성 부족**: 3개 이상의 범주를 가진 변수에서 귀무가설이 기각될 경우, 어느 범주 간에 유의미한 차이가 있는지는 알려주지 않습니다. 이를 위해서는 **사후 분석(Post-hoc analysis)** (예: 각 셀의 잔차 분석)이 필요합니다.

대안

- **피셔의 정확 검정 (Fisher's Exact Test)**: 2x2 분할표에서 기대 빈도가 5 미만인 셀이 있을 때 사용하는 정확한 확률 기반 검정 방법입니다. `scipy.stats.fisher_exact` 함수로 구현할 수 있습니다.
- **G-검정 (G-test, 우도비 검정)**: 카이제곱 검정과 유사한 목적을 가지며, 특히 표본 크기가 작을 때 더 나은 성능을 보일 수 있습니다. `chi2_contingency`에서 `lambda_='log-likelihood'`로 설정하여 수행할 수 있습니다.
- **맥니마 검정 (McNemar's Test)**: 대응되는 두 범주형 변수의 비율 변화를 검정할 때 사용합니다. (예: 특정 약물 투여 전/후의 효과 비교).