Kruskal-Wallis H Test (크루스칼-왈리스 H 검정)

- 세 개 이상의 독립적인 표본(집단)들의 중앙값이 동일한지를 검정하는 비모수적인 방법
- 일원배치 분산분석(One-way ANOVA)의 비모수 버전으로, 데이터가 정규성 가정을 만족하지 않을 때 사용
- 검정은 각 집단의 데이터를 순위로 변환한 다음, 순위의 평균을 비교하여 집단 간에 통계적으로 유의미한 차이가 있는지를 판단
- **귀무가설 (H0)**: 모든 집단의 중앙값은 동일하다.
    - 모든 집단의 분포는 동일하다.
- **대립가설 (H1)**: 적어도 한 집단의 중앙값은 다른 집단과 다르다.
    - 적어도 한 집단의 분포는 다른 집단과 다르다.

적용 가능한 상황

- 세 개 이상의 독립적인 그룹을 비교할 때
- 데이터가 정규성 가정을 충족하지 못할 때
- 데이터가 순서형 척도(ordinal scale)일 때
- 등분산성 가정이 충족되지 않을 때 (ANOVA보다 덜 민감함)

용도

`scipy.stats.kruskal` 함수는 세 개 이상의 독립적인 샘플에 대한 크루스칼-왈리스 H 검정을 수행하는 데 사용됩니다.

주의사항 (가정)

- **독립성**: 각 집단은 서로 독립적이어야 합니다.
- **분포 형태**: 각 집단의 분포 형태는 유사해야 합니다. 분포의 위치(중앙값)만 다를 뿐, 분산이나 모양은 비슷하다고 가정합니다. 만약 분포 형태가 크게 다르면 검정 결과의 해석에 주의가 필요합니다.
- **연속형 변수**: 이론적으로는 연속형 변수를 가정하지만, 순서형 변수에도 적용 가능합니다. 동점 순위(tie)가 많을 경우 p-value가 보정됩니다.

`scipy.stats.kruskal(*samples, nan_policy='propagate')`

하이퍼파라미터 (인자) 설명

- `*samples`: array_like. 비교하고자 하는 각 집단의 데이터를 개별 인자로 전달합니다. 예를 들어, 3개 집단을 비교한다면 `kruskal(group1, group2, group3)`와 같이 사용합니다.
- `nan_policy`: `{'propagate', 'raise', 'omit'}` (기본값: 'propagate'). 결측치(NaN) 처리 방법을 정의합니다.
    - `'propagate'`: 결측치가 있으면 `NaN`을 반환합니다.
    - `'raise'`: 결측치가 있으면 에러를 발생시킵니다.
    - `'omit'`: 결측치를 무시하고 계산을 수행합니다.

```python
import numpy as np
from scipy.stats import kruskal

# 데이터 생성 (세 개의 독립적인 그룹)
# 각 그룹이 정규분포를 따르지 않는다고 가정
group1 = [7, 8, 8, 9, 10, 11, 12]
group2 = [9, 9, 10, 11, 12, 13, 14]
group3 = [12, 13, 14, 14, 15, 16, 17]

# 크루스칼-왈리스 H 검정 수행
statistic, p_value = kruskal(group1, group2, group3)

print(f"Kruskal-Wallis H statistic: {statistic:.4f}") # 12.3866
print(f"P-value: {p_value:.4f}") # 0.0020

# 결과 해석: "귀무가설 기각: 적어도 한 집단의 중앙값은 다른 집단과 통계적으로 유의미한 차이가 있습니다."
alpha = 0.05
if p_value < alpha:
    print("귀무가설 기각: 적어도 한 집단의 중앙값은 다른 집단과 통계적으로 유의미한 차이가 있습니다.")
else:
    print("귀무가설 채택: 집단 간 중앙값의 차이가 통계적으로 유의미하지 않습니다.")

# 만약 귀무가설이 기각되었다면, 사후 분석(Post-hoc test)을 통해 어느 집단 간에 차이가 있는지 확인해야 합니다.
# 비모수 검정의 사후 분석으로는 Dunn's test, Conover's test 등이 사용될 수 있습니다.
# Dunn's test는 scikit-posthocs 패키지를 활용할 수 있습니다.

# !pip install scikit-posthocs
import scikit_posthocs as sp
import pandas as pd

data = pd.DataFrame({
    'value': group1 + group2 + group3,
    'group': ['group1'] * len(group1) + ['group2'] * len(group2) + ['group3'] * len(group3)
})

# Dunn's test 수행 (p-value 조정 방법: Bonferroni)
posthoc_df = sp.posthoc_dunn(data, val_col='value', group_col='group', p_adjust='bonferroni')

print("Dunn's Post-hoc test (Bonferroni correction):")
print(posthoc_df)
'''
Dunn's Post-hoc test (Bonferroni correction):
          group1    group2    group3
group1  1.000000  0.559898  0.001473
group2  0.559898  1.000000  0.091180
group3  0.001473  0.091180  1.000000
'''
```

결과 해석 방법

- **H statistic**: 검정 통계량입니다. 이 값이 클수록 집단 간의 순위 평균 차이가 크다는 것을 의미합니다.
- **P-value**: 귀무가설이 사실일 때, 현재와 같은 검정 통계량 또는 더 극단적인 값이 관찰될 확률입니다.
  - `p-value < 유의수준(alpha)`: 귀무가설을 기각합니다. 즉, 적어도 하나의 집단은 다른 집단들과 중앙값이 다르다고 결론 내릴 수 있습니다.
  - `p-value >= 유의수준(alpha)`: 귀무가설을 기각하지 못합니다. 즉, 집단 간에 중앙값 차이가 통계적으로 유의미하다고 말할 수 없습니다.

사후 분석 결과 해석

Dunn's test 결과로 나온 행렬은 각 집단 쌍 간의 p-value를 보여줍니다. 이 p-value를 유의수준과 비교하여 특정 두 집단 간에 유의미한 차이가 있는지를 판단할 수 있습니다. 예를 들어, `group1`과 `group3` 사이의 p-value가 0.05보다 작다면, 두 집단의 중앙값은 통계적으로 유의미하게 다르다고 해석할 수 있습니다.

장단점 및 대안

장점

- **정규성 불필요**: 데이터가 정규 분포를 따르지 않아도 사용할 수 있어 활용도가 높습니다.
- **이상치에 강건함**: 순위를 사용하므로 극단적인 값(이상치)에 덜 민감합니다.
- **적용 범위**: 순서형 데이터에도 적용할 수 있습니다.

단점

- **민감도**: 데이터가 실제로 정규성 및 등분산성 가정을 만족하는 경우, 일원배치 분산분석(ANOVA)에 비해 검정력이 낮을 수 있습니다.
- **구체성 부족**: 검정 결과가 유의미하더라도, 어느 집단이 다른 집단과 다른지는 알려주지 않습니다. 이를 확인하려면 별도의 사후 분석(Post-hoc test)이 필요합니다.
- **가정**: 각 집단의 분포 모양이 비슷해야 한다는 가정이 있습니다. 만약 분포 모양이 크게 다르면 결과 해석에 오류가 있을 수 있습니다.

대안

- **일원배치 분산분석 (One-way ANOVA)**: 데이터가 정규성, 등분산성, 독립성 가정을 모두 만족하는 경우 사용합니다. 크루스칼-왈리스 검정보다 검정력이 높습니다.
- **Welch's ANOVA**: 등분산성 가정을 만족하지 못할 때 사용하는 ANOVA의 변형입니다.
- **Mood's Median Test**: 집단 간 중앙값이 같은지를 검정하는 또 다른 비모수 방법입니다. 크루스칼-왈리스 검정보다 이상치에 더 강건하지만, 일반적으로 검정력은 더 낮습니다.