분산분석 (ANOVA)

분산분석(Analysis of Variance, ANOVA): **세 개 이상**의 집단 간의 **평균**을 비교하는 데 사용되는 강력한 통계적 방법

- 이름은 '분산'분석이지만, 실제 목적은 집단 간 '평균'의 차이를 검정하는 것
- T-검정을 여러 번 반복해서 사용하면 1종 오류가 증가하는 문제(다중 비교 문제)가 발생
  - ANOVA는 이를 해결하기 위해 모든 집단의 평균이 같은지를 한 번에 검정

ANOVA의 핵심 원리는 데이터의 **총 변동(분산)**을 두 가지 요소로 분해하는 것

1. **집단 간 변동 (Between-group variation)**
   - 각 집단의 평균이 전체 평균으로부터 얼마나 떨어져 있는지를 표현
   - 독립변수(요인, factor)의 효과에 의해 발생하는 변동
2. **집단 내 변동 (Within-group variation)**
   - 각 집단 내의 데이터들이 해당 집단의 평균으로부터 얼마나 퍼져 있는지를 표현
   - 무작위 오차(random error)에 의해 발생하는 변동

ANOVA는 이 두 변동의 비율, 즉 **F-통계량**을 계산하여 가설을 검정 `F = (집단 간 분산) / (집단 내 분산)`
만약 집단 간의 차이(집단 간 분산)가 우연에 의한 변동(집단 내 분산)보다 충분히 크다면,
F-값이 커지고 이는 집단 간 평균에 유의미한 차이가 있음을 시사합니다.

ANOVA의 종류

- **일원배치 분산분석 (One-way ANOVA)**
  - 하나의 독립변수(요인)에 따라 종속변수의 평균이 다른지를 검정
  - e.g. 비료 종류(A, B, C)에 따른 식물 키의 차이
- **이원배치 분산분석 (Two-way ANOVA)**
  - 두 개의 독립변수와 이들 간의 상호작용(interaction effect)이 종속변수의 평균에 미치는 영향을 검정
  - e.g. 비료 종류와 토양 종류에 따른 식물 키의 차이 및 두 요인의 상호작용 효과

ANOVA의 기본 가정

1. **독립성**: 각 집단의 표본은 서로 독립적으로 추출되어야 합니다.
2. **정규성**: 각 집단의 데이터는 정규분포를 따라야 합니다.
3. **등분산성**: 모든 집단의 분산은 동일해야 합니다.

사후 분석 (Post-hoc Analysis)

ANOVA 검정 결과 `p-value`가 유의수준보다 작아서 **귀무가설이 기각**되면, "적어도 한 집단의 평균은 다르다"는 사실을 알 수 있다.
하지만 **어떤 집단들 간에** 차이가 있는지는 알 수 없어, 확인을 위해 수행하는 추가적인 분석을 **사후 분석**이라고 합니다.

- **Tukey's HSD (Honestly Significant Difference) Test**
  - 가장 널리 사용되는 사후 분석 방법 중 하나
  - 모든 가능한 집단 쌍(pair)에 대해 평균 차이를 동시에 검정

    - 다중 비교로 인한 1종 오류의 증가를 제어
- **Bonferroni Correction**
    - 가장 간단하고 보수적인 방법
    - 유의수준(α)을 비교하는 쌍의 개수(k)로 나누어(α/k), 더 엄격한 기준으로 각 쌍을 t-검정
    - 1종 오류를 확실히 막지만, 검정력이 낮아져 실제 차이를 놓칠 수 있음 (2종 오류 증가)

1. 일원배치 분산분석 (One-way ANOVA)

- **문제**: 세 가지 다른 교육 방법(A, B, C)으로 학생들을 가르친 후, 시험 점수를 비교했다. 교육 방법에 따라 학생들의 평균 점수에 유의미한 차이가 있는가?
- **함수**: `scipy.stats.f_oneway`, `statsmodels.formula.api.ols`, `pingouin.anova`

```
import pandas as pd
from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# 데이터 생성
method_a = [85, 88, 79, 92, 84]
method_b = [75, 78, 81, 72, 79]
method_c = [90, 94, 88, 91, 95]

# 1. ANOVA 검정 (scipy)
f_statistic, p_value = f_oneway(method_a, method_b, method_c)
print("--- One-way ANOVA (scipy) ---")
print(f"F-statistic: {f_statistic:.4f}")
print(f"P-value: {p_value:.4f}")

if p_value < 0.05:
    print("귀무가설 기각: 교육 방법에 따른 평균 점수 차이가 유의미합니다.")
else:
    print("귀무가설 기각 실패: 평균 점수 차이가 유의미하지 않습니다.")
'''
--- One-way ANOVA (scipy) ---
F-statistic: 18.3175
P-value: 0.0002
귀무가설 기각: 교육 방법에 따른 평균 점수 차이가 유의미합니다.
'''

# 2. 사후 분석 (Tukey's HSD)
# 데이터를 long format으로 변환
df = pd.DataFrame({'score': method_a + method_b + method_c,
                   'group': ['A'] * 5 + ['B'] * 5 + ['C'] * 5})

tukey_result = pairwise_tukeyhsd(endog=df['score'], groups=df['group'],
alpha=0.05)
print("\n--- Tukey's HSD Post-hoc Test ---")
print(tukey_result)
'''
--- Tukey's HSD Post-hoc Test ---
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
```

```
group1 group2 meandiff p-adj   lower    upper  reject
--------------------------------------------------
     A      B     -8.6 0.0104 -15.0692 -2.1308   True
     A      C      6.0 0.0701  -0.4692 12.4692  False
     B      C     14.6 0.0002   8.1308 21.0692   True
--------------------------------------------------
'''
```

- **결과 해석**:
  - **ANOVA**
    - `p-value(0.0002)`가 0.05보다 작으므로, 세 교육 방법 간에 유의미한 평균 점수 차이가 있다고 결론 내립니다.
  - **Tukey's HSD**
    - 결과 테이블의 `reject` 열을 보면, A-B, B-C 에서 `True`, A-C에서 `False`로 나타납니다.
    - B 방법은 A, C와 통계적으로 유의미한 차이가 나는 것을 알 수 있고, A, C 사이는 통계적으로 유의미한 차이가 크지 않음을 알 수 있습니다.
    - `p-adj` 열은 다중 비교를 조정한 p-value를 나타냅니다.

2. 이원배치 분산분석 (Two-way ANOVA)

- **문제**: 비료 종류(fertilizer)와 토양 종류(soil)가 식물의 성장(growth)에 미치는 영향을 알아보고자 한다. 각 요인의 주 효과(main effect)와 두 요인 간의 상호작용 효과(interaction effect)를 검정하고 싶다.
- **함수**: `statsmodels.formula.api.ols`, `pingouin.anova`

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
import pingouin as pg

# 데이터 불러오기
df_two_way = pg.read_dataset('anova2')

# 1. 이원배치 분산분석 (statsmodels)
# C(Blend), C(Crop), 그리고 상호작용항 C(Blend):C(Crop)
model = ols('Yield ~ C(Blend) + C(Crop) + C(Blend):C(Crop)',
data=df_two_way).fit()
anova_table_sm = sm.stats.anova_lm(model, typ=2)
print("--- Two-way ANOVA (statsmodels) ---")
print(anova_table_sm)
'''
--- Two-way ANOVA (statsmodels) ---
                      sum_sq    df         F    PR(>F)
C(Blend)            2.041667   1.0  0.003768  0.951730
C(Crop)          2736.583333   2.0  2.525235  0.107978
C(Blend):C(Crop)  2360.083333   2.0  2.177813  0.142223
Residual          9753.250000  18.0       NaN       NaN
'''

# 2. 이원배치 분산분석 (pingouin)
print("\n--- Two-way ANOVA (pingouin) ---")
anova_table_pg = pg.anova(data=df_two_way, dv='Yield', between=['Blend', 'Crop'],
```

```
detailed=True)
print(anova_table_pg)
'''
--- Two-way ANOVA (pingouin) ---
         Source            SS  DF           MS         F     p-unc       np2
0         Blend      2.041667   1     2.041667  0.003768  0.951730  0.000209
1          Crop  2736.583333   2  1368.291667  2.525235  0.107978  0.219105
2  Blend * Crop  2360.083333   2  1180.041667  2.177813  0.142223  0.194834
3      Residual  9753.250000  18   541.847222       NaN       NaN       NaN
'''
```

- **결과 해석**: (`statsmodels` 기준)
  - **C(Blend)**: p-value(`PR(>F)`=0.9517)가 0.05보다 크므로, **비료 종류에 따른 수확량(Yield)의 차이는 통계적으로 유의하지 않습니다.** 즉, 비료 종류(Blend)에 따라 수확량의 평균 차이는 없다고 볼 수 있습니다.
  - **C(Crop)**: p-value(`PR(>F)`=0.1080)가 0.05보다 크므로, **토양 종류(Crop)에 따른 수확량의 차이 또한 유의하지 않습니다.**
  - **C(Blend):C(Crop)**: p-value(`PR(>F)`=0.1422)가 0.05보다 크므로, **비료와 토양 간의 상호작용 효과도 유의하지 않습니다.** 즉, 특정 비료의 효과가 토양 종류에 따라 달라진다고 보기 어렵습니다.
  - `pingouin` 라이브러리 결과 역시 동일한 결론을 보여주며, 보다 간결하고 직관적인 테이블 형식으로 확인할 수 있습니다. → **종합적으로, 비료 종류, 토양 종류, 그리고 두 요인 간의 상호작용 모두 수확량에 유의한 영향을 미치지 않는 것으로 나타났습니다.**