F-검정 (F-test)

- **F-검정(F-test)**
  - 두 개 이상의 집단(모집단)의 **분산(variance)**이 서로 같은지를 검정
  - 혹은 회귀 분석에서 모델 전체의 유의성을 검정하는 데 사용되는 통계적 방법
  - F-검정은 검정통계량으로 **F-분포(F-distribution)**를 사용

F-분포의 특징:

- 두 개의 카이제곱(χ^2) 분포의 비율로 정의되며, 항상 양수 값을 가집니다.
- 두 개의 자유도, 즉 분자의 자유도(df1)와 분모의 자유도(df2)에 의해 모양이 결정됩니다.
- 오른쪽으로 긴 꼬리를 갖는 비대칭적인 분포입니다.

F-검정의 주요 용도:

1. **두 집단의 등분산성 검정 (F-test for Equality of Variances)**:
   - 두 집단의 분산이 동일한지를 검정합니다. 이는 독립표본 t-검정의 전제 조건인 등분산성을 확인하는 데 사용될 수 있습니다.
   - **가설**:
     - $H_0: \sigma_1^2 = \sigma_2^2$ (두 집단의 분산은 같다)
     - $H_1: \sigma_1^2 \neq \sigma_2^2$ (두 집단의 분산은 다르다)
   - **검정통계량**: $F = s_1^2 / s_2^2$ ($s_1^2 > s_2^2$ 이도록 큰 분산을 분자에 둠)
   - **주의 사항**
     - 데이터가 **정규분포를 따른다는 가정**에 매우 민감합니다.
     - 데이터가 정규분포를 따르지 않으면 결과가 부정확해지므로, 실제로는 더 강건한(robust) **Levene 검정**이나 **Bartlett 검정**을 사용하는 것이 일반적입니다.

2. **분산분석 (Analysis of Variance, ANOVA)**:
   - **세 개 이상**의 집단 간의 **평균**이 동일한지를 검정하는 데 F-검정이 핵심적으로 사용됩니다.
   - ANOVA는 전체 분산을 집단 간 분산(between-group variance)과 집단 내 분산(within-group variance)으로 분해합니다.
   - **가설**:
     - $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ (모든 집단의 평균은 같다)
     - $H_1:$ 적어도 한 집단의 평균은 다르다
   - **검정통계량**: $F = (\text{집단 간 분산}) / (\text{집단 내 분산})$
   - **해석**: 만약 집단 간의 차이(분산)가 집단 내부의 변동(분산)에 비해 충분히 크다면, F-값이 커지고 귀무가설을 기각하게 됩니다. 즉, 집단 간 평균에 유의미한 차이가 있다고 결론 내립니다.

3. **회귀 분석 (Regression Analysis)**:
   - 회귀 모델 전체가 통계적으로 유의미한지를 검정합니다. 즉, 적어도 하나의 독립변수가 종속변수를 설명하는 데 기여하는지를 확인합니다.
   - **가설**:
     - $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ (모든 회귀계수가 0이다. 즉, 모델은 유의미하지 않다)
     - $H_1:$ 적어도 하나의 β_i는 0이 아니다 (모델은 유의미하다)

    - **해석**: F-검정의 p-value가 유의수준보다 작으면, 귀무가설을 기각하고 "모델이 통계적으로 유의미하다"고 판단합니다.

1. 두 집단의 등분산성 검정 (직접 구현)

- **문제**: 두 개의 다른 생산 라인에서 생산된 제품의 무게 데이터가 있다. 두 생산 라인의 제품 무게의 분산이 동일한가? (데이터는 정규분포를 따른다고 가정)

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

# 표본 데이터 생성
np.random.seed(0)
line_a_weights = np.random.normal(loc=100, scale=5, size=20)
line_b_weights = np.random.normal(loc=100, scale=8, size=20)

# 1. 각 그룹의 분산 계산
var_a = np.var(line_a_weights, ddof=1)
var_b = np.var(line_b_weights, ddof=1)

# 2. F-통계량 계산 (큰 분산 / 작은 분산)
if var_a > var_b:
    f_statistic = var_a / var_b
    dfn = len(line_a_weights) - 1 # 분자의 자유도
    dfd = len(line_b_weights) - 1 # 분모의 자유도
else:
    f_statistic = var_b / var_a
    dfn = len(line_b_weights) - 1
    dfd = len(line_a_weights) - 1

# 3. p-value 계산 (양측 검정이므로 꼬리 확률에 2를 곱함)
# F.cdf는 왼쪽 꼬리 확률, F.sf는 오른쪽 꼬리 확률 (1 - cdf)
p_value = f.sf(f_statistic, dfn, dfd) * 2

print("--- F-test for Equality of Variances ---")
print(f"Variance of Line A: {var_a:.4f}")
print(f"Variance of Line B: {var_b:.4f}")
print(f"F-statistic: {f_statistic:.4f}")
print(f"p-value: {p_value:.4f}")

if p_value < 0.05:
    print("귀무가설 기각: 두 집단의 분산은 유의미하게 다릅니다.")
else:
    print("귀무가설 기각 실패: 두 집단의 분산은 차이가 없습니다.")
'''
--- F-test for Equality of Variances ---
Variance of Line A: 19.0213
Variance of Line B: 95.0513
F-statistic: 4.9971
p-value: 0.0010
귀무가설 기각: 두 집단의 분산은 유의미하게 다릅니다.
'''
```

- **결과 해석**
  - p-value(0.0010)가 0.05보다 작으므로, 귀무가설을 기각합니다.
  - 즉, 두 생산 라인의 제품 무게 분산은 통계적으로 유의미하게 다르다고 결론 내릴 수 있습니다.
  - **하지만 실제 분석에서는 Levene 검정을 사용하는 것이 더 안전합니다.**

2. 분산분석(ANOVA)에서의 F-검정

- **문제**: 세 가지 다른 비료(A, B, C)를 사용하여 키운 식물들의 키 데이터가 있다. 비료의 종류에 따라 식물의 평균 키에 유의미한 차이가 있는가?
- **함수**: `stats.f_oneway(*args)`

```
from scipy.stats import f_oneway

# 표본 데이터 생성
fertilizer_a = [20, 22, 19, 21, 23]
fertilizer_b = [25, 27, 26, 24, 28]
fertilizer_c = [18, 20, 19, 17, 16]

# 일원배치 분산분석(One-way ANOVA) 수행
f_statistic, p_value = f_oneway(fertilizer_a, fertilizer_b, fertilizer_c)

print("--- F-test in One-way ANOVA ---")
print(f"F-statistic: {f_statistic:.4f}")
print(f"p-value: {p_value:.4f}")

if p_value < 0.05:
    print("귀무가설 기각: 적어도 한 비료의 효과는 다른 비료와 다릅니다.")
else:
    print("귀무가설 기각 실패: 비료 종류에 따른 평균 키 차이가 없습니다.")
'''
--- F-test in One-way ANOVA ---
F-statistic: 32.6667
p-value: 0.0000
귀무가설 기각: 적어도 한 비료의 효과는 다른 비료와 다릅니다.
'''
```

- **결과 해석**
  - p-value(0.0000)가 0.05보다 매우 작으므로, 귀무가설을 기각합니다.
  - 즉, 세 가지 비료의 종류에 따라 식물의 평균 키에는 통계적으로 유의미한 차이가 있다고 결론 내릴 수 있습니다.
  - 어떤 그룹 간에 차이가 있는지는 사후 분석(Post-hoc test)을 통해 추가로 확인해야 합니다.