

비율 검정 (Proportion Test)

- 한 개 또는 두 개의 모집단에서 특정 속성을 가진 개체의 비율(proportion)에 대한 가설을 검정하는 통계적 방법
- 표본 비율을 사용하여 모집단 비율을 추론하며, 이항 분포를 정규 분포로 근사하여 Z-검정을 수행하는 것이 일반적

단일 표본 비율 검정 (One-Sample Proportion Test)

- 하나의 모집단 비율이 특정 값과 같은지를 검정합니다.
- **귀무가설 (H0)**: 모집단 비율(p)은 특정 비율(p_0)과 같다. ($p = p_0$)
- **대립가설 (H1)**: 모집단 비율은 p_0 과 다르다 ($p \neq p_0$), p_0 보다 크다 ($p > p_0$), 또는 p_0 보다 작다 ($p < p_0$).

두 표본 비율 검정 (Two-Sample Proportion Test)

- 두 모집단의 비율이 서로 같은지를 검정합니다.
- **귀무가설 (H0)**: 두 모집단의 비율은 같다. ($p_1 = p_2$)
- **대립가설 (H1)**: 두 모집단의 비율은 다르다. ($p_1 \neq p_2$)

적용 가능한 상황

- **단일 표본 비율 검정**:
 - 특정 제품의 불량률이 알려진 기준치(예: 3%)와 동일한지 검정할 때.
 - 특정 선거 후보의 지지율이 50%를 넘는지 확인할 때.
- **두 표본 비율 검정**:
 - 두 가지 다른 광고(A/B 테스트)의 클릭률(CTR)에 차이가 있는지 비교할 때.
 - 남성과 여성의 흡연율이 동일한지 검정할 때.

구현 방법

`statsmodels.stats.proportion` 라이브러리의 `proportions_ztest` 함수를 사용하여 단일 표본 및 두 표본 비율 검정을 수행할 수 있습니다.

주의사항 (가정)

- **독립성**: 표본 내의 각 관측치는 서로 독립적이어야 합니다.
- **표본 크기**
 - 검정의 정확성을 위해 표본 크기가 충분히 커야 합니다.
 - 일반적으로 $n \cdot p \geq 10$ 이고 $n \cdot (1-p) \geq 10$ (단일 표본) 또는 $n_1 \cdot p_1$, $n_1 \cdot (1-p_1)$, $n_2 \cdot p_2$, $n_2 \cdot (1-p_2)$ 가 모두 5 또는 10 이상일 때 (두 표본) 정규 근사가 유효하다고 봅니다.
 - 이 조건이 만족되지 않으면 **이항 검정(Binomial Test)**이나 **피셔의 정확 검정(Fisher's Exact Test)**을 사용하는 것이 더 적절합니다.

1. 단일 표본 비율 검정 (One-Sample Proportion Test)

코드 예시

```
statsmodels.stats.proportion.proportions_ztest(count, nobs, value=None,
alternative='two-sided', prop_var=False)
```

하이퍼파라미터 (인자) 설명

- **count**: int or array_like. 성공 횟수 (관심 있는 사건의 발생 횟수).
- **nobs**: int or array_like. 총 시행 횟수 (표본 크기).
- **value**: float. 귀무가설에서의 모집단 비율(p_0). **None**일 경우 두 표본 비율 검정으로 간주됩니다.
- **alternative**: {'two-sided', 'smaller', 'larger'} (기본값: 'two-sided'). 대립가설의 종류를 지정합니다.
 - 'two-sided': $p \neq p_0$ (양측 검정)
 - 'smaller': $p < p_0$ (단측 검정)
 - 'larger': $p > p_0$ (단측 검정)
- **prop_var**: False or float. 비율의 분산을 계산할 때 사용될 비율 값. **False**이면 귀무가설의 비율(**value**)을 사용하고, **True**이면 표본 비율을 사용합니다.

```
import numpy as np
from statsmodels.stats.proportion import proportions_ztest

# 예시: 한 공장에서 생산된 제품 1000개 중 45개가 불량품으로 나왔다.
# 이 공장의 불량률이 3%라고 할 수 있는가? (유의수준 5%)

# 귀무가설: 공장의 불량률은 3%이다. ( $p = 0.03$ )
# 대립가설: 공장의 불량률은 3%가 아니다. ( $p \neq 0.03$ )

count = 45 # 불량품 수 (성공 횟수)
nobs = 1000 # 총 생산량 (표본 크기)
value = 0.03 # 귀무가설의 비율

# 단일 표본 비율 검정 수행
stat, p_value = proportions_ztest(count, nobs, value, alternative='two-sided')

print(f"Z-statistic: {stat:.4f}") # 2.2881
print(f"P-value: {p_value:.4f}") # 0.0221

# 결과 해석: "귀무가설 기각: 공장의 불량률은 3%와 통계적으로 유의미한 차이가 있습니다."
alpha = 0.05
if p_value < alpha:
    print("귀무가설 기각: 공장의 불량률은 3%와 통계적으로 유의미한 차이가 있습니다.")
else:
    print("귀무가설 채택: 공장의 불량률이 3%라는 주장을 기각할 수 없습니다.")

# 만약 표본 크기가 작다면 (e.g.,  $n \cdot p < 10$ ), 이항 검정을 사용
from scipy.stats import binom_test

# 예시: 20개의 제품 중 3개가 불량 ( $n \cdot p_0 = 20 \cdot 0.03 = 0.6 < 10$ )
p_val_binom = binom_test(x=3, n=20, p=0.03, alternative='two-sided')
print(f"\nBinomial test P-value (for small sample): {p_val_binom:.4f}") # 0.0210
```

결과 해석 방법

- **Z-statistic:** Z-검정 통계량. 표본 비율이 귀무가설의 비율로부터 표준오차의 몇 배만큼 떨어져 있는지를 나타냅니다.
- **P-value:** 귀무가설이 참일 때, 현재와 같은 검정 통계량 또는 더 극단적인 값이 나올 확률입니다.
 - $p\text{-value} < \text{유의수준}$: 귀무가설을 기각하고 대립가설을 채택합니다.

2. 두 표본 비율 검정 (Two-Sample Proportion Test)

코드 예시

```
from statsmodels.stats.proportion import proportions_ztest
import numpy as np

# 예시: 두 가지 다른 웹사이트 디자인(A, B)의 구매 전환율 비교
# 귀무가설: 두 디자인의 구매 전환율은 동일하다. ( $p_A = p_B$ )
# 대립가설: 두 디자인의 구매 전환율은 다르다. ( $p_A \neq p_B$ )

# A 디자인: 1500명 방문, 100명 구매
# B 디자인: 1600명 방문, 130명 구매

count = np.array([100, 130]) # 각 그룹의 성공 횟수
nobs = np.array([1500, 1600]) # 각 그룹의 표본 크기

# 두 표본 비율 검정 수행 (value=None 으로 설정)
stat, p_value = proportions_ztest(count, nobs, alternative='two-sided')

print(f"Z-statistic: {stat:.4f}") # -1.5482
print(f"P-value: {p_value:.4f}") # 0.1216

# 결과 해석: "귀무가설 채택: 두 웹사이트 디자인의 구매 전환율 차이는 유의미하지 않습니다."
alpha = 0.05
if p_value < alpha:
    print("귀무가설 기각: 두 웹사이트 디자인의 구매 전환율에는 통계적으로 유의미한 차이가 있습니다.")
else:
    print("귀무가설 채택: 두 웹사이트 디자인의 구매 전환율 차이는 유의미하지 않습니다.")

# 참고: 두 표본 비율 검정은 2x2 분할표에 대한 카이제곱 독립성 검정과 통계적으로 동일한 결과를 제공합니다.
#  $Z^2 = \chi^2$ 
from scipy.stats import chi2_contingency

# [성공, 실패] 형태로 분할표 생성
# 성공: 구매, 실패: 비구매
observed = np.array([
    [100, 1500 - 100], # A 디자인
    [130, 1600 - 130] # B 디자인
])

chi2, p_val_chi2, _, _ = chi2_contingency(observed, correction=False) #
```

correction=False로 설정해야 $Z^2 = \chi^2$ 관계 성립

```
print(f"\nChi-squared statistic: {chi2:.4f}") # 2.3970
print(f"P-value from Chi-squared test: {p_val_chi2:.4f}") # 0.1216
print(f"Z-statistic squared: {stat**2:.4f}") # 2.3970,  $Z^2$  값과  $\chi^2$  값이 동일함을 확인
```

결과 해석 방법

- **Z-statistic:** 두 표본 비율 간의 차이를 표준오차 단위로 나타낸 값입니다.
- **P-value:** 두 모집단의 비율이 같다고 가정할 때, 현재 표본에서 관찰된 것과 같거나 더 큰 차이가 나타날 확률입니다.
 - $p\text{-value} < \text{유의수준}$: 두 모집단의 비율이 다르다고 결론 내립니다.

장단점 및 대안

장점

- **직관적이고 간단함:** 비율이라는 이해하기 쉬운 척도를 사용하여 가설을 검정할 수 있습니다.
- **다양한 활용:** A/B 테스트, 여론조사 분석 등 비즈니스 및 연구 분야에서 널리 사용됩니다.

단점

- **표본 크기 제약:** 정규분포 근사를 사용하므로 표본 크기가 충분히 크지 않으면 결과의 신뢰도가 떨어집니다.

대안

- **이항 검정 (Binomial Test):** 단일 표본 비율 검정에서 표본 크기가 작을 때 사용하는 정확한 검정 방법입니다. `scipy.stats.binom_test` 또는 `scipy.stats.binomtest`로 구현할 수 있습니다.
- **피셔의 정확 검정 (Fisher's Exact Test):** 두 표본 비율 검정에서 표본 크기가 작을 때 (특히 2x2 분할표에서 기대 빈도가 5 미만인 셀이 있을 때) 사용하는 정확한 검정 방법입니다. `scipy.stats.fisher_exact`로 구현할 수 있습니다.
- **카이제곱 검정 (Chi-squared Test):** 두 표본 비율 검정은 2x2 분할표에 대한 카이제곱 독립성/동질성 검정과 동일한 결과를 제공합니다. 카이제곱 검정은 3개 이상의 비율을 비교하는 경우로 쉽게 확장될 수 있습니다.