정규성 검정: Shapiro-Wilk, Kolmogorov-Smirnov

정규성 검정(Normality Test): 주어진 데이터 샘플이 정규분포(Normal Distribution)를 따르는 모집단에서 추출되었는지를 통계적으로 확인하는 절차

- 많은 모수적 통계 분석 방법은 데이터가 정규분포를 따른다는 가정을 전제로 함
  - 분석 수행 전, 이 가정이 만족되는지를 검증하는 것이 매우 중요
- 모수적 통계 분석 방법 : t-test, ANOVA, 선형 회귀 등

정규성 검정의 가설 설정

- **귀무가설 (H_0)**: 데이터는 정규분포를 따른다.
- **대립가설 (H_1)**: 데이터는 정규분포를 따르지 않는다.

검정 결과 `p-value`가 유의수준(e.g., 0.05)보다 **작으면** 귀무가설을 기각하고, "데이터가 정규분포를 따르지 않는다"고 결론 내립니다.
반대로 `p-value`가 유의수준보다 **크면** 귀무가설을 기각할 수 없으므로, "데이터가 정규분포를 따른다고 볼 수 있다"고 판단합니다.

Shapiro-Wilk Test

- 정규성 검정 방법 중 검정력(power)이 가장 뛰어나다고 알려짐
- 표본 크기가 비교적 작을 때(n < 50, 또는 수천 개까지) 가장 널리 권장되는 방법

Kolmogorov-Smirnov (K-S) Test

- 데이터의 누적분포함수(CDF)와 정규분포의 누적분포함수 간의 최대 차이를 검정 통계량으로 사용
- `scipy.stats.kstest`에서는 평균과 표준편차를 직접 지정해야 하는 등 사용이 다소 번거로움
- `Shapiro-Wilk` 테스트보다 검정력이 낮은 경향이 존재
- `K-S 테스트`는 특정 분포(정규분포 등)를 따르는지 검정하는 것뿐만 아니라, 두 샘플이 동일한 분포에서 추출되었는지를 검정하는 데에도 사용 가능

시각적 방법 (Q-Q Plot)

- `Quantile-Quantile 플롯`: 데이터의 분위수와 정규분포의 분위수를 2차원 평면에 점으로 나타낸 그래프
- 데이터가 정규분포를 따른다면, 점들이 거의 직선 형태를 띰
- 통계적 검정과 함께 시각적으로 정규성을 판단하는 매우 효과적인 방법

적용 가능한 상황

- **모수 통계 분석 전제 조건 확인**: t-test, ANOVA, 선형 회귀 분석 등을 수행하기 전에 데이터(또는 회귀 분석의 경우 잔차)가 정규성 가정을 만족하는지 확인할 때.
- **데이터 변환 결정**: 데이터가 정규성을 따르지 않을 때, 로그 변환이나 Box-Cox 변환 등 데이터 변환의 필요성을 판단하기 위해.

예제 데이터 생성

```
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

# 1. 정규분포를 따르는 데이터 생성
np.random.seed(0)
normal_data = np.random.normal(loc=10, scale=2, size=100)

# 2. 정규분포를 따르지 않는 데이터 생성 (균등 분포)
uniform_data = np.random.uniform(low=0, high=20, size=100)
```

1. Shapiro-Wilk Test

- 표본 데이터의 정규성 검정 (가장 권장되는 방법)
- **주의사항**
  - 표본 크기가 너무 크면(수천 개 이상), 아주 작은 편차에도 귀무가설을 기각하여 정규분포가 아니라고 판단할 수 있음
  - Q-Q 플롯과 히스토그램을 함께 보고 실질적인 분포 형태를 판단하는 것이 중요

```
# 정규분포 데이터에 대한 검정
shapiro_stat_normal, shapiro_p_normal = stats.shapiro(normal_data)
print(f"--- Shapiro-Wilk Test on Normal Data ---")
print(f"Statistic: {shapiro_stat_normal:.4f}, p-value: {shapiro_p_normal:.4f}")
if shapiro_p_normal > 0.05:
    print("귀무가설 기각 실패: 데이터는 정규분포를 따릅니다.")
else:
    print("귀무가설 기각: 데이터는 정규분포를 따르지 않습니다.")
'''
--- Shapiro-Wilk Test on Normal Data ---
Statistic: 0.9927, p-value: 0.8689
귀무가설 기각 실패: 데이터는 정규분포를 따릅니다.
'''

# 균등분포 데이터에 대한 검정
shapiro_stat_uniform, shapiro_p_uniform = stats.shapiro(uniform_data)
print(f"\n--- Shapiro-Wilk Test on Uniform Data ---")
print(f"Statistic: {shapiro_stat_uniform:.4f}, p-value: {shapiro_p_uniform:.4f}")
if shapiro_p_uniform > 0.05:
    print("귀무가설 기각 실패: 데이터는 정규분포를 따릅니다.")
else:
    print("귀무가설 기각: 데이터는 정규분포를 따르지 않습니다.")
'''
--- Shapiro-Wilk Test on Uniform Data ---
Statistic: 0.9574, p-value: 0.0026
귀무가설 기각: 데이터는 정규분포를 따르지 않습니다.
'''
```

- **결과 해석**:

    - **Normal Data**: p-value(0.8689)가 0.05보다 크므로, 귀무가설을 기각하지 못합니다. 즉, 데이터가 정규분포를 따른다고 볼 수 있습니다.
    - **Uniform Data**: p-value(0.0026)가 0.05보다 작으므로, 귀무가설을 기각합니다. 즉, 데이터가 정규분포를 따르지 않는다고 결론 내립니다.

2. Kolmogorov-Smirnov (K-S) Test

- 표본이 특정 이론적 분포(e.g., 정규분포, 균등분포 등)를 따르는지 검정
- **주의사항**
    - `kstest`에서 정규성을 검정하려면, 검정 대상 분포를 `'norm'`으로 지정하고, `args` 파라미터에 표본의 평균과 표준편차를 직접 전달해야 함
        - 그렇지 않으면 표준정규분포(평균=0, 표준편차=1)와 비교하게 되어 잘못된 결과가 도출됨
    - `args` 파라미터 미 지정 시, 데이터가 특정 분포를 따르는지 확인 가능
        - 아래 예시는 기본 모수 `loc=0, scale=1`를 따른다는 가정
        - `norm`: 정규분포
        - `uniform`: 균등분포
        - `expon`: 지수분포
    - `stats.ks_2samp(normal_data, uniform_data)`를 사용하면 두 데이터의 분포를 비교 가능

```
# 정규분포 데이터에 대한 검정
ks_stat_normal, ks_p_normal = stats.kstest(normal_data, 'norm', args=
(normal_data.mean(), normal_data.std(ddof=1)))
print(f"--- K-S Test on Normal Data ---")
print(f"Statistic: {ks_stat_normal:.4f}, p-value: {ks_p_normal:.4f}")
# Statistic: 0.0643, p-value: 0.7778

# 균등분포 데이터에 대한 검정
ks_stat_uniform, ks_p_uniform = stats.kstest(uniform_data, 'norm', args=
(uniform_data.mean(), uniform_data.std(ddof=1)))
print(f"\n--- K-S Test on Uniform Data ---")
print(f"Statistic: {ks_stat_uniform:.4f}, p-value: {ks_p_uniform:.4f}")
# Statistic: 0.0830, p-value: 0.4712

# 두 데이터 비교
ks_stat, ks_p = stats.kstest(normal_data, uniform_data)
print(f"\n--- K-S Test on Each Two Data ---")
print(f"Statistic: {ks_stat:.4f}, p-value: {ks_p:.4f}")
# Statistic: 0.3200, p-value: 0.0001
```

- **결과 해석**
    - Normal Data의 p-value(0.7778)는 0.05보다 크다.
    - Uniform Data의 p-value(0.4712)도 0.05보다 크다. → 검정력이 Shapiro-Wilk 검정보다 약하다.

Q-Q Plot

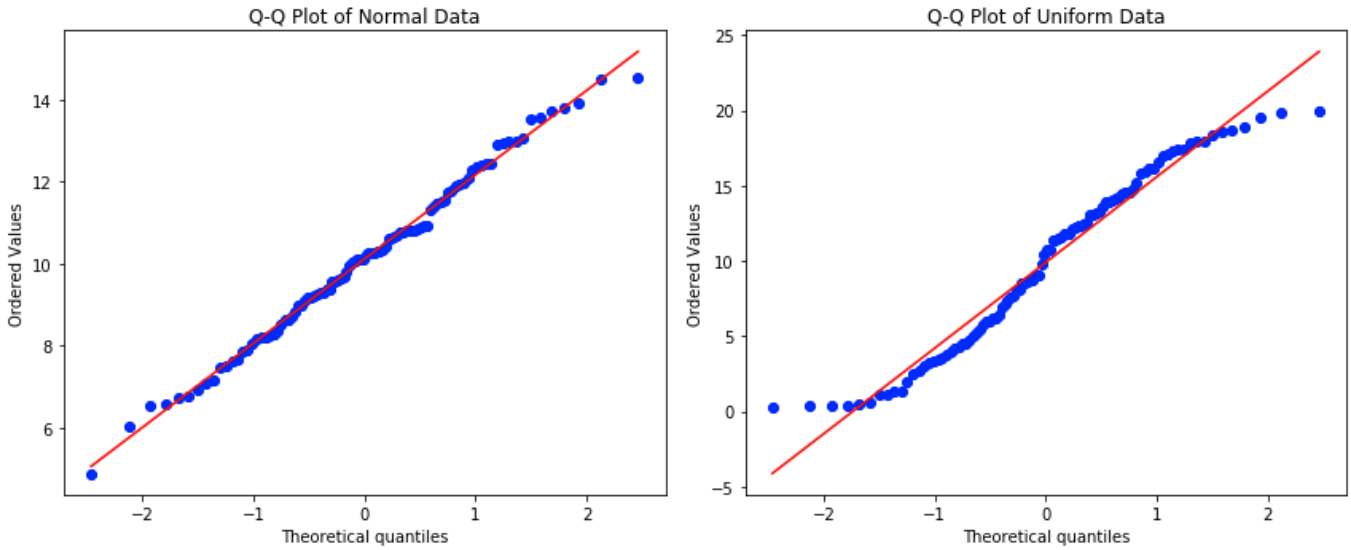
- 데이터의 분위수와 이론적 정규분포의 분위수를 비교하여 정규성을 시각적으로 확인

```
plt.figure(figsize=(12, 5))

# 정규분포 데이터의 Q-Q Plot
plt.subplot(1, 2, 1)
stats.probplot(normal_data, dist="norm", plot=plt)
plt.title("Q-Q Plot of Normal Data")

# 균등분포 데이터의 Q-Q Plot
plt.subplot(1, 2, 2)
stats.probplot(uniform_data, dist="norm", plot=plt)
plt.title("Q-Q Plot of Uniform Data")

plt.tight_layout()
plt.show()
```



- **결과 해석**
  - 왼쪽의 정규분포 데이터는 점들이 빨간색 대각선 위에 거의 일직선으로 놓여 있어 정규성을 만족함을 시각적으로 보여준다.
  - 반면, 오른쪽의 균등분포 데이터는 점들이 S자 곡선 형태로 정규분포를 따르지 않음을 보여준다.
    - 연구자에 따라 정규분포를 거의 만족하는 것으로 볼 수도 있어 경험적 판단에 맡기는 것이 중요

검정 방법 선택 가이드

- **일반적인 경우**: **Shapiro-Wilk Test**를 우선적으로 사용합니다. 검정력이 가장 우수하다고 알려져 있습니다.
- **대용량 데이터**: 표본 크기가 매우 클 경우(e.g., n > 5000), 정규성 검정은 사소한 차이에도 매우 민감하게 반응하여 p-value가 작게 나오는 경향이 있습니다. 이 경우, 통계적 검정 결과에만 의존하기보다는 **Q-Q Plot과 히스토그램을 함께 보고 실질적으로 데이터가 정규분포에 '충분히 가까운지'**를 판단하는 것이 더 중요합니다. 중심 극한 정리에 따라, 표본 크기가 매우 크면 평균의 분포는 정규분포로 근사하므로, 약간의 비정규성은 후속 분석(t-test 등)에 큰 영향을 미치지 않을 수 있습니다.