기술 통계: 중심 극한 정리, 표준 오차

- **기술 통계(Descriptive Statistics)**: 데이터를 요약하고 설명하는 통계적 방법
- **추론 통계(Inferential Statistics)**: 표본(sample)을 사용하여 모집단(population)에 대해 추론하는 것
- 중심 극한 정리와 표준 오차는 이 두 영역을 잇는 매우 중요한 다리 역할을 하는 개념

중심 극한 정리 (Central Limit Theorem, CLT)

- 모집단의 분포 형태와 관계없이(정규분포가 아니어도), 해당 모집단에서 추출한 표본의 크기 n이 충분히 크다면(일반적으로 $n ≥ 30$), **표본 평균들의 분포(sampling distribution of the sample mean)**는 근사적으로 정규분포를 따른다는 정리입니다.
    - **핵심 내용**:
        1. 표본 평균의 분포는 정규분포에 가까워진다.
        2. 이 정규분포의 평균은 모집단의 평균($μ$)과 같다.
        3. 이 정규분포의 표준편차는 $σ / √n$ 이다. ($σ$: 모집단 표준편차, n: 표본 크기)
    - **의의**: 모집단의 분포를 모르더라도, 표본 평균의 분포를 정규분포로 가정하고 통계적 추론(가설 검정, 신뢰구간 추정 등)을 수행할 수 있는 강력한 이론적 근거를 제공합니다.

표준 오차 (Standard Error, SE)

- 통계량(statistic)의 표본 분포(sampling distribution)가 갖는 표준편차를 의미합니다. 즉, 표본을 반복해서 추출할 때마다 통계량(e.g., 표본 평균, 표본 비율)이 얼마나 변동하는지를 나타내는 척도입니다.
- **표본 평균의 표준 오차 (Standard Error of the Mean, SEM)**
    - 가장 흔하게 사용되는 표준 오차로, $SE = σ / √n$ 입니다. ($σ$: 모집단 표준편차)
- **현실적 문제**
    - 실제로는 모집단의 표준편차($σ$)를 알 수 없는 경우가 대부분입니다.
    - 따라서, 표본의 표준편차(s)를 대신 사용하여 표준 오차를 추정합니다: $SE ≈ s / √n$
- **의의**
    - 표준 오차는 표본 통계량이 모집단 모수(parameter)를 얼마나 정확하게 추정하는지를 나타냅니다.
    - 표준 오차가 작을수록 표본 통계량이 더 안정적이고 신뢰할 수 있음을 의미합니다.
    - 신뢰구간 계산과 가설 검정에서 핵심적인 역할을 합니다.

표준편차 vs 표준 오차

- **표준편차 (Standard Deviation)**: **하나의 표본 내**에서 데이터 값들이 평균으로부터 얼마나 퍼져 있는지를 나타내는 산포도의 척도입니다.
- **표준 오차 (Standard Error)**: **여러 표본들로부터 계산된 통계량(e.g., 표본 평균)들**이 모집단 모수 주변에 얼마나 퍼져 있는지를 나타내는 척도입니다.

적용 가능한 상황

- **중심 극한 정리**
    - 모집단의 분포를 알 수 없거나 정규분포가 아닐 때, 표본의 크기가 충분히 크다면($n≥30$) Z-검정이나 t-검정과 같은 모수적 통계 검정을 적용할 수 있는 근거가 됩니다.
    - 대규모 여론조사에서, 특정 후보에 대한 지지율(표본 비율)의 분포가 정규분포를 따른다고 가정하고 신뢰구간을 계산할 때.

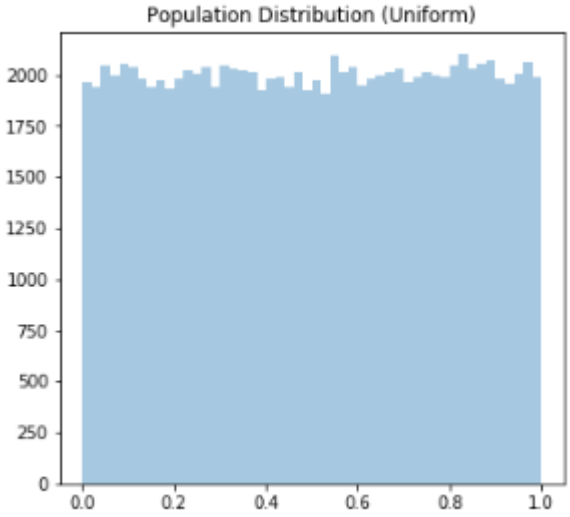
- **표준 오차**
    - **신뢰구간 추정**: '모집단 평균은 [표본 평균 - (임계값 * 표준 오차)] 와 [표본 평균 + (임계값 * 표준 오차)] 사이에 있을 것이다' 와 같이 모집단 모수의 신뢰구간을 계산할 때 사용됩니다.
    - **가설 검정**: 두 집단의 평균 차이를 검정할 때, 그 차이가 우연에 의한 변동(표준 오차)보다 통계적으로 유의미하게 큰지를 판단하는 기준(검정 통계량)을 계산하는 데 사용됩니다.

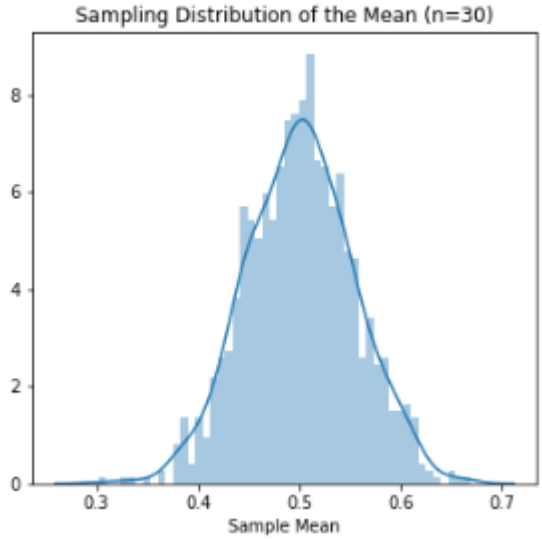
중심 극한 정리 시뮬레이션

- 균등 분포(Uniform Distribution)라는 비정규분포 모집단에서 표본을 반복적으로 추출하여, 표본 평균들의 분포가 정말로 정규분포를 따르는지 시각적으로 확인합니다.

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# 모집단 생성: 0과 1 사이의 균등 분포 (정규분포가 아님)
population = np.random.uniform(0, 1, 100000)

# 표본 크기(n)와 시뮬레이션 횟수 설정
sample_size = 30
num_simulations = 1000

# 표본 평균들을 저장할 리스트
sample_means = []

for _ in range(num_simulations):
    # 모집단에서 표본 추출
    sample = np.random.choice(population, size=sample_size)
    # 표본의 평균 계산 및 저장
    sample_means.append(sample.mean())

# 시각화
plt.figure(figsize=(12, 5))

# 모집단 분포
plt.subplot(1, 2, 1)
sns.distplot(population, bins=50, kde=False)
plt.title('Population Distribution (Uniform)')

# 표본 평균들의 분포
plt.subplot(1, 2, 2)
sns.distplot(sample_means, bins=50, kde=True)
plt.title(f'Sampling Distribution of the Mean (n={sample_size})')
plt.xlabel('Sample Mean')

plt.show()

# 왜도 및 첨도 통계량 확인 방법
from scipy import stats
# 왜도: 평균을 중심으로 좌우로 데이터가 편향되어 있는 정도
# 0보다 크면(양수) 오른쪽으로 긴 꼬리 / 0보다 작으면(음수) 왼쪽으로 긴 꼬리
```

```
print(stats.skew(sample_means))     # 0.10185051041351989
# 첨도: 뾰족함 정도
# 0은 정규분포 기준 / 0보다 크면 중앙 집중도가 정규분포보다 큼 / 0보다 작으면 중앙 집중
도가 정규분포보다 작음
print(stats.kurtosis(sample_means)) # -0.19022850947512282
```





- **결과 해석**
  - 왼쪽 그래프는 모집단의 분포로, 평평한 균등 분포를 보입니다.
  - 오른쪽 그래프는 이 모집단에서 추출한 표본들의 평균값들의 분포를 보여주며, 중심 극한 정리에 따라 종 모양의 정규분포에 가까워진 것을 확인할 수 있습니다.

표준 오차 계산

- 주어진 데이터(표본)의 표준편차를 이용하여 표본 평균의 표준 오차(SEM)를 계산합니다.

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

# 예제 데이터 (어떤 표본)
sample_data = [10, 12, 15, 13, 18, 11, 16, 14]

# 1. 수동 계산
n = len(sample_data)
sample_std = np.std(sample_data, ddof=1) # ddof=1은 표본 표준편차를 의미
standard_error_manual = sample_std / np.sqrt(n)

# 2. scipy.stats.sem 사용
standard_error_scipy = sem(sample_data)

print(f"표본 크기 (n): {n}")                                    # 8
print(f"표본 표준편차 (s): {sample_std:.4f}")                   # 2.6693
print(f"표준 오차 (수동 계산): {standard_error_manual:.4f}")    # 0.9437
print(f"표준 오차 (scipy.sem): {standard_error_scipy:.4f}")     # 0.9437
```

- **결과 해석**
  - `scipy.stats.sem` 함수를 사용하면 표본 평균의 표준 오차를 간편하게 계산할 수 있습니다.
  - 이 값은 이 표본의 평균이 모집단 평균을 얼마나 정확하게 추정하는지에 대한 불확실성의 정도를 나타냅니다.
  - 표본 크기 n이 커질수록 표준 오차는 작아지며, 이는 더 큰 표본이 더 정확한 추정치를 제공함을 의미합니다.