표본 크기 산출

- 통계적 조사를 실시할 때, 주로 진행하는 2가지 방식

    1. **전수조사**: 전체(모집단)을 조사
        - **장점** : 정확, 오차 `0`
        - **단점** : 비용, 시간 과다
    2. **표본조사**: 일부 추출 후 조사 (많이 추출 후 조사, 무작위 추출 후 조사)
        - **장점** : 적절한 비용과 시간
        - **단점** : 오차가 존재

- Sampling(표집, 표본을 뽑는 행위) 목적 → 모집단 추정

    - 표본 평균을 통해 모 평균을 추정
    - 표준오차 : 추정치에 존재하는 오차

- **너무 작은 표본**: 표본이 모집단을 잘 대표하지 못하여 추정의 정확도가 떨어지고, 통계적 검정력이 낮아져 실제로 존재하는 차이나 효과를 발견하지 못할 수 있습니다.

- **너무 큰 표본**: 필요 이상의 시간과 비용을 낭비하게 됩니다.

표본 크기는 주로 **신뢰구간(Confidence Interval)**의 폭을 결정하는 방식으로 산출됩니다.
즉, "모집단 모수(e.g., 평균, 비율)를 특정 신뢰수준(e.g., 95%)에서 특정 오차한계(e.g., ±3%) 이내로 추정하고 싶을 때, 몇 명의 표본이 필요한가?" 라는 질문에 답하는 과정입니다.

95% 신뢰구간 (Confidence Interval)

- 표준오차(X)를 바탕으로 95% 확률 구간을 구할 수 있다!
- `95% 신뢰구간 = X - 1.96 * SE ~ X + 1.96 * SE`
    - `1.96`: 95% 신뢰 수준에서의 Z-score(표본 평균에서 표준편차의 몇 배만큼 떨어져 있는지)
- 의미 : 신뢰구간 안에 모평균이 포함될 확률 95% (쉬운 설명)
  = 표본을 100번 정도 뽑으면 95번 정도는 95% 신뢰구간 안에 모평균을 포함함
  = 100번 샘플링 하고 각각 평균과 95% 신뢰구간을 구했을 때,
  신뢰구간들(100개)이 모집단 평균을 포함하는 횟수가 95번, 포함하지 않는 횟수가 5번

표본 크기 결정에 영향을 미치는 주요 요소

1. **오차 한계 (Margin of Error, E)**: 표본 통계량과 모집단 모수 간에 허용되는 최대 오차의 크기. 작게 설정할수록 더 많은 표본이 필요합니다. (e.g., ±3%)
2. **신뢰 수준 (Confidence Level)**: 추정된 신뢰구간이 실제 모집단 모수를 포함할 것이라고 확신하는 정도. 보통 95% 또는 99%를 사용합니다. 신뢰수준이 높을수록 더 많은 표본이 필요합니다.
3. **모집단의 표준편차 (σ) 또는 비율 (p)**: 모집단의 데이터가 얼마나 퍼져있는지(분산)를 나타냅니다. 분산이 클수록 불확실성이 크므로, 더 많은 표본이 필요합니다. 실제로는 모집단의 표준편차나 비율을 모르므로, 과거 연구나 파일럿 테스트(pilot test)를 통해 추정하거나, 가장 보수적인 값(비율의 경우 p=0.5)을 사용합니다.

공식 및 적용

1. 모집단 평균 추정을 위한 표본 크기

- **공식**: `n = (Z * σ / E)²`
    - `n`: 필요한 표본 크기
    - `Z`: 신뢰수준에 해당하는 Z-score (e.g., 95% 신뢰수준 -> Z=1.96, 99% 신뢰수준 -> Z=2.576)
    - `σ`: 모집단의 표준편차 (추정치 사용)
    - `E`: 최대 허용 오차 (오차 한계)
- **적용 가능한 상황**: 특정 제품의 평균 수명, 특정 지역의 평균 소득 등을 특정 오차 범위 내에서 추정하고 싶을 때.

2. 모집단 비율 추정을 위한 표본 크기

- **공식**: `n = (Z² * p * (1-p)) / E²`
    - `n`: 필요한 표본 크기
    - `Z`: 신뢰수준에 해당하는 Z-score
    - `p`: 모집단 비율 (추정치 사용). 모를 경우, 분산을 최대화하여 가장 많은 표본 크기를 요구하는 `p=0.5`를 사용.
    - `E`: 최대 허용 오차 (오차 한계)
- **적용 가능한 상황**: 대통령 선거 여론조사에서 특정 후보의 지지율, 특정 제품의 시장 점유율 등을 특정 오차 범위 내에서 추정하고 싶을 때.

1. 모집단 평균 추정을 위한 표본 크기 계산

- **문제**: 어떤 공장에서 생산되는 전구의 수명(시간)을 95% 신뢰수준에서 평균 ±50시간의 오차 한계로 추정하고 싶다. 과거 데이터에 따르면 전구 수명의 표준편차는 약 300시간으로 알려져 있다. 필요한 최소 표본 크기는?

```python
import numpy as np
from scipy.stats import norm

# 1. 파라미터 설정
confidence_level = 0.95
margin_of_error = 50  # E
population_std = 300  # σ

# 2. 신뢰수준에 따른 Z-score 계산
# 양측 검정이므로 (1 - 신뢰수준) / 2 의 확률에 해당하는 Z값을 찾고 절대값을 취하거나,
# (1 + 신뢰수준) / 2 의 확률에 해당하는 Z값을 찾음.
alpha = 1 - confidence_level
z_score = norm.ppf(1 - alpha / 2)
# 또는 z_score = norm.ppf(confidence_level + alpha / 2)

# 3. 표본 크기 공식 적용
sample_size = (z_score * population_std / margin_of_error)**2

print(f"신뢰수준: {confidence_level * 100}%")                    # 95.0%
print(f"Z-score: {z_score:.4f}")                                 # 1.9600
```

```
print(f"필요한 표본 크기 (n): {sample_size:.4f}")                # 138.2925
print(f"최소 표본 크기 (올림 처리): {np.ceil(sample_size)}")    # 139.0
```

- **결과 해석**: 약 138.29가 계산되므로, 최소 139개의 전구 표본을 검사해야 95% 신뢰수준에서 ±50시간의 오차 한계를 만족하는 평균 수명을 추정할 수 있습니다.

2. 모집단 비율 추정을 위한 표본 크기 계산

- **문제**: 전국 유권자를 대상으로 특정 후보의 지지율을 99% 신뢰수준에서 ±2%p의 오차 한계로 추정하고 싶다. 지지율에 대한 정보가 전혀 없을 때, 필요한 최소 표본 크기는?

```
# 1. 파라미터 설정
confidence_level = 0.99
margin_of_error = 0.02  # E
# 모집단 비율 p를 모르므로 가장 보수적인 0.5 사용
population_proportion = 0.5 # p

# 2. 신뢰수준에 따른 Z-score 계산
alpha = 1 - confidence_level
z_score = norm.ppf(1 - alpha / 2)

# 3. 표본 크기 공식 적용
sample_size = (z_score**2 * population_proportion * (1 - population_proportion)) /
(margin_of_error**2)

print(f"신뢰수준: {confidence_level * 100}%")                    # 99.0%
print(f"Z-score: {z_score:.4f}")                                 # 2.5758
print(f"필요한 표본 크기 (n): {sample_size:.4f}")               # 4146.8104
print(f"최소 표본 크기 (올림 처리): {np.ceil(sample_size)}")    # 4147.0
```

- **결과 해석**: 약 4147.36이 계산되므로, 최소 4148명의 유권자를 대상으로 설문조사를 해야 99% 신뢰수준에서 ±2%p의 오차 한계를 만족하는 지지율을 추정할 수 있습니다.

주의사항 및 고려사항

- **표본 추출 방법**: 위 공식들은 모집단에서 표본을 무작위로 추출(Random Sampling)한다는 가정을 전제로 합니다. 표본 추출 방법에 편향(bias)이 있다면, 표본 크기를 아무리 늘려도 신뢰할 수 있는 결과를 얻을 수 없습니다.
- **모집단 표준편차(σ) 및 비율(p)의 추정**: 표본 크기 산출 공식은 모집단의 특성(σ 또는 p)을 알아야 한다는 모순적인 상황에 놓입니다. 따라서 이 값들을 어떻게 추정하는지가 중요합니다.
    - **과거 연구/데이터 활용**: 가장 좋은 방법입니다.
    - **파일럿 연구 (Pilot Study)**: 소규모 예비 조사를 통해 σ나 p를 대략적으로 추정합니다.
    - **보수적 접근**: 비율(p)의 경우, `p=0.5`를 사용하면 필요한 표본 크기가 최대가 되므로 가장 안전합니다. 표준편차(σ)의 경우, 가능한 값의 범위(range)를 알고 있다면 `range / 4` 또는 `range / 6`을 대략적인 표준편차로 사용하기도 합니다.

- **유한 모집단 수정 (Finite Population Correction)**: 만약 표본의 크기(n)가 모집단의 크기(N)에 비해 무시할 수 없을 정도로 크다면(보통 n/N > 0.05), 계산된 표본 크기를 보정하여 줄여줄 수 있습니다. 하지만 대부분의 경우 모집단이 매우 크므로 이 보정은 생략됩니다.

- **검정력 (Power)**: 위 내용은 주로 '추정'의 정확도에 초점을 맞춘 것입니다. 만약 '가설 검정'을 목적으로 한다면, **검정력(Power)**, 즉 대립가설이 사실일 때 이를 올바르게 탐지할 확률(1-β)을 고려하여 표본 크기를 계산해야 합니다. 이는 더 복잡한 계산을 요구하며, `statsmodels.stats.power` 모듈 등을 사용할 수 있습니다.