01. Population and Sample

a. Population

연구하고자 하는 전체 그룹을 의미한다. 예를 들어, 어떤 국가의 성인 남성의 평균 키를 연구할 경우 인구는 해당 국가의 모든 성인 남성이 된다.

b. Sample

인구 중 실제 연구를 위해 선택된 부분집합이다. 모든 인구로부터 데이터를 수집하는 것이 비현실적이거나 비용이 많이 들거나 물리적으로 불가능하기 때문에 표본추출을 사용합니다.

02. Types of Sampling(표본추출 방법)

a. 왜 샘플링이 필요한가?

Feasibility

많은 경우, 대상 집단의 크기가 너무 커서 전체를 연구하는 것이 불가능하다. 예를 들어, 한 나라의 모든 성인 남성의 키를 측정하는 것은 비현실적이다.

Cost and Time Efficiency

표본 추출을 통해 연구자들은 전체 인구를 연구하는 것에 비해 더 빠르게 결과를 얻고 비용을 줄일 수 있다.

Data Quality

일부 경우에서, 표본에 초점을 맞추면 더 정확한 데이터 수집 방법을 사용할 수 있다. 왜냐하면 자원을 더 작은 그룹에 집중할 수 있기 때문이다.

b. 샘플링 방법

Random Sampling

- 정의: 모든 대상 집단의 구성원이 표본에 포함될 동일한 기회를 가진 표본 추출 방법이다. **- 장점**: 단순하고 직관적이며, 표본이 대상 집단을 대표하도록 보장한다. **- 단점**: 매우 큰 대상 집단에서 실행하는 것이 어려울 수 있다.

Stratified Sampling

- 정의: 대상 집단을 비슷한 특성을 공유하는 작은 그룹(계층)으로 나누고, 이러한 각 계층에서 독립적으로 표본을 추출한다. **- 장점**: 특정 하위 그룹이 표본에서 공정하게 대표되도록 보장한다. **- 단점**: 계층화의 기준을 결정하는 것이 어려울 수 있습니다. 인구를 비슷한 특성을 공유하는 작은 그룹(층)으로 나누고, 각 층에서 독립적으로 표본을 추출한다.

Systematic Sampling

- 정의: 첫 번째 간격 내에서 무작위로 시작점을 선택한 후 고정된 주기적인 간격(예: 10번째 구성원)을 기반으로 표본을 선택하는 방법이다. **- 장점**: 실행하기 쉽고 무작위 표본 추출보다 덜 복잡하다. **- 단점**: 주기적인 선택 과정이 표본 추출 간격과 일치하는 패턴이 있다면 편향을 도입할 수 있다.

Cluster Sampling

- 정의: 대상 집단을 군집(그룹)으로 나누고, 이 군집 중 일부를 무작위로 선택한다. 그런 다음 선택된 군집 내의 모든 구성원이 표본화된다. **- 장점**: 큰 대상 집단에 대해 비용 효율적이고 시간을 절약할 수 있다. **- 단점**: 계층화 표본 추출에 비해 표본의 대표성을 감소시킬 수 있다.

03. 기술통계 (Descriptive Statistics)

- **중심 경향의 측정**: 데이터 세트의 빈도 분포 중앙 위치를 설명하는 측정이다. (Average, Median, Mode(최빈값)).
- **분포의 측정 (변동성 측정)**: 데이터 포인트가 얼마나 퍼져 있는지를 보여주는 지표입니다 (범위, 사분위수 범위(IQR), 분산, 표준편차).
- **분포의 형태 측정**: 데이터 분포의 형태를 설명합니다 (Skewness(왜도)와 Kurtosis(첨도)).

사분위수 범위 (IQR)

- 중앙값이 분포를 반으로 나누듯, 사분위수는 분포를 4등분합한다. 사분위수 범위(IQR)는 데이터 값의 중간 50%의 범위를 정의하며, Q3(3사분위수)와 Q1(1사분위수)의 차이로 계산된다. $$IQR = Q3 - Q1$$

왜도 (Skewness)

- 왜도는 분포의 비대칭성을 측정한다. 양의 왜도는 오른쪽 꼬리가 길다는 것을, 음의 왜도는 왼쪽 꼬리가 길다는 것을 나타낸다.

첨도 (Kurtosis)

- 첨도는 분포의 "꼬리무게" 또는 분포의 정점의 뾰족함을 측정한다. Mesokurtic(Kurtosis ≈ 3), Leptokurtic(Kurtosis > 3), Platykurtic(Kurtosis < 3)으로 분류된다.

04. 데이터 스케일링 (Data Scaling)

- 데이터 스케일링은 다양한 스케일로 측정된 특성(또는 변수)을 다룰 때 매우 중요한 단계이다. 스케일링 없이는 큰 수치 범위를 가진 특성이 분석에서 작은 범위를 가진 특성을 지배할 수 있으며, 이는 편향된 결과를 초래한다. 데이터 스케일링에는 Min-max scaling, Z-score standardization 등의 방법이 있다.

Min-max scaling

Min-Max Scaling은 데이터를 0과 1 사이의 범위로 변환하는 방법이다. $X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$

여기서 X_{norm}은 정규화된 값을, X는 원래 값을, X_{min}과 X_{max}는 각각 데이터의 최소값과 최대값을 나타낸다.

Z-score standardization

Z-Score Standardization은 각 데이터 포인트를 평균이 0이고 표준편차가 1인 분포로 변환하는 방법이다.

$X_{std} = \frac{X - \mu}{\sigma}$

여기서 X_{std}은 표준화된 값을, X는 원래 값을, μ는 데이터의 평균값을, σ는 데이터의 표준편차를 나타낸다.