Population : 연구 대상 집합 전체 **Sample** : population의 부분집합 **Sampling의 목적** : 실현 가능성↑ / 비용적, 시간적 효율↑ / 데이터 퀄리티↑

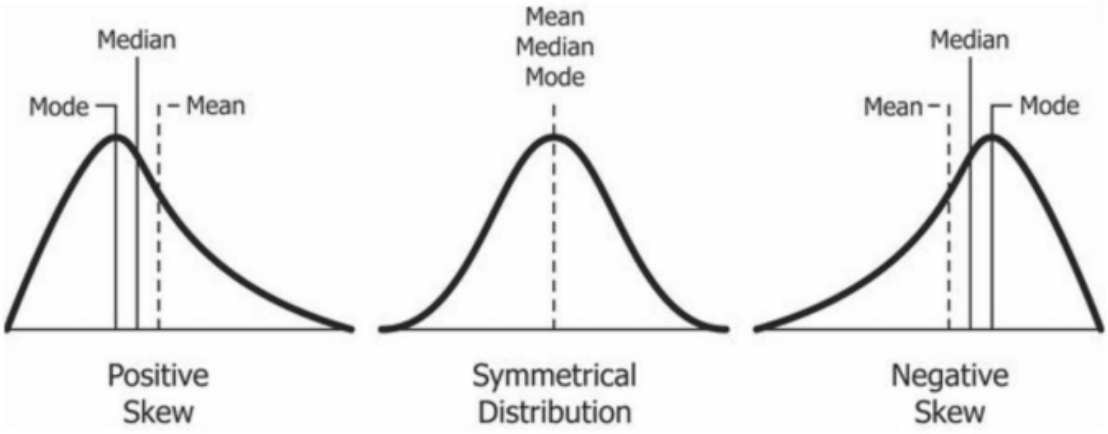
Sampling의 종류

	정의	장점	단점
Random Sampling	랜덤으로 샘플링	간단함	집합이 커지면 까다로움
Systematic Sampling	고정된 간격마다 샘플링	random sampling 보다도 간단함	샘플링 간격에 따른 패턴이 있으면 결과가 편향될 수 있음
Stratified Sampling	그룹을 부분집합으로 나누고 각각의 부분집합에서 샘플링	각각의 부분집합이 공평하게 반영됨	집합을 나누는 기준을 정하기 까다로움
Cluster Sampling	그룹을 부분집합으로 나누고 부분집합 몇 개를 통째로 샘플링	큰 집단에 대해서 비용&시간 절약됨	대표성이 떨어짐

Descriptive Statistics(설명적 통계)

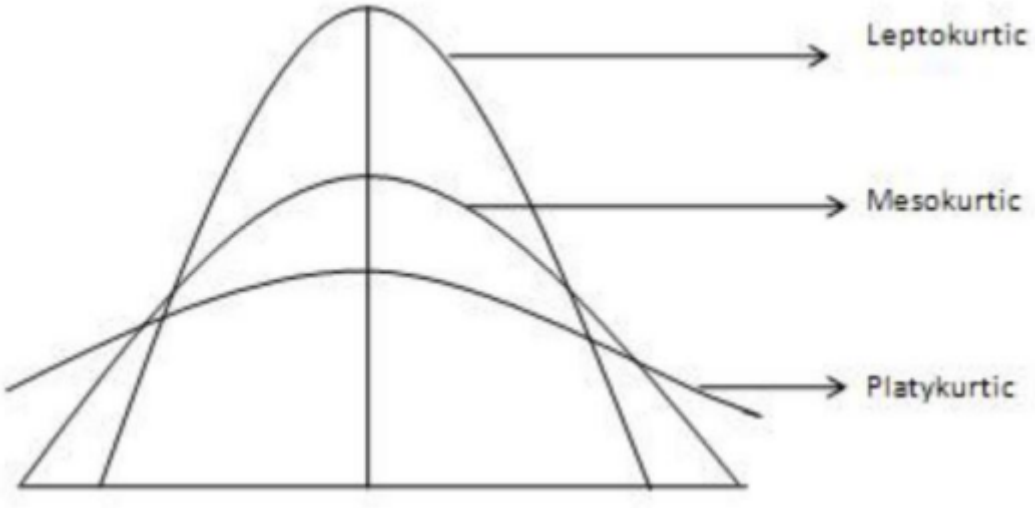
- Measures of Central Tendency
    - Mean(평균)
    - Median(중앙값)
    - Mode(최빈값)
- Measures of Spread
    - Range : 최대값-최소값
    - IQR(Interquartile Range) : 데이터값들의 중간부분 50%를 차지하는 범위. Q3(3사분위, 하위 75%)-Q1(1사분위, 하위 25%)으로 계산
    - Variance(분산) : (편차)^2의 평균
    - Standard Deviation(표준편차) : √분산
- Measures of Shape
    - Skewness(왜도) : 평균값을 기준으로 한 확률분포의 비대칭성
        - Positively Skewed(Skewness>0): 데이터가 평균값보다 왼쪽에 치우침
        - Symmetric(Skewness=0): 데이터의 왼쪽과 오른쪽이 대칭(Bell-shaped), 정규분포의 왜도와 비슷

- Negatively Skewed(Skewness<0): 데이터가 평균값보다 오른쪽에 치우침



- Kurtosis(첨도) : 확률분포의 꼬리 늘어짐 정도

  - Leptokurtic(Kurtosis>3): 두꺼운 꼬리, 뾰족한 정점
  - Mesokurtic(Kurtosis≈3): 적당한 꼬리, 정규분포의 첨도와 비슷
  - Platykurtic(Kurtosis<3): 얇은 꼬리와 평평한 정점

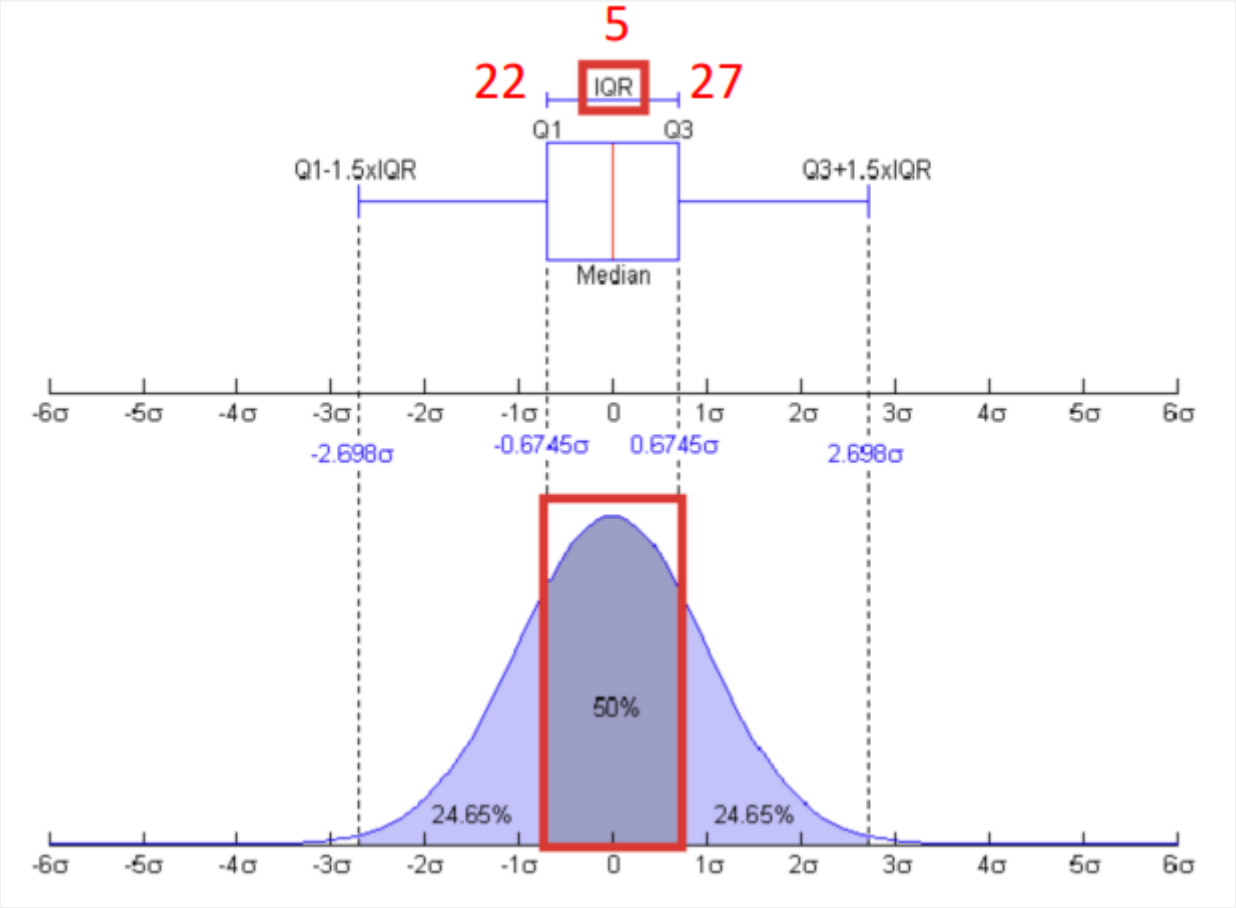


IQR

quartile(사분위수): 통계 변량을 도수 분포로 정리하였을 때 적은 것으로부터 25%, 50%, 75% 자리의 변량 값 Q1(1사분위): 하위 25% Q2(2사분위): 하위 50%(중앙값) Q3(3사분위): 하위 75% IQR(InterQuartile Range): Q3 - Q1

> **IQR로 outlier 찾기** dataset : 10, 19, 5, 21, 21, 22, 15, 22, 23, 7, 23, 23, 23, 24, 24, 24, 24, 25, 23
>
> (1) 데이터 작은 순으로 정렬 5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25
> (2) Q1, Q3값 구하기 Q1: 19 Q3: 24
> (3) IQR값 구하기 IQR=24-19=5
> (4) 경계 찾기 Lower bound(coefficient=1.5) = 19(Q1)- 5×1.5 = 11.5 Upper bound(coefficient=1.5) = 24(Q3)+ 5×1.5 = 31.5

(5) outlier 찾기 outliers below the lower bound = 3 outliers over the upper bound = 0



Data Scaling

다른 범위에 대해 측정된 특성들을 다룰 때 큰 범위의 측정값이 작은 범위의 측정값보다 더 많은 영량력을 행사하는 것을 방지

- Min-Max scaling: 각 특성값이 0에서 1사이의 범위로 scaled 됨

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where:

- X_{scaled} is the scaled value.
- X is the original value.
- $X_{\min}$ and $X_{\max}$ are the minimum and maximum values of the feature, respectively.

Ex1) 나이에 대한 데이터 [20, 22, 25, 27, 29] Min-Max scaling하기 min=20, max=29, max-min=9 [(20-20)/9=0, (22-20)/9=2/9, (25-20)/9=5/9, (27-20)/9=7/9, (29-20)/9=1]
> **Ex2) GPA데이터와 TOEIC점수 데이터 Min-Max scaling하기**

	Steve	**Tony**
GPA	4.4	2.0
TOEIC	910	920

> Steve의 GPA, TOEIC 평균: 457.2 Tony의 GPA, TOEIC 평균: 461.0
> 이것만 보면 Tony가 더 우수한 학생으로 보이지만, 이는 GPA의 scale은 0~4.5, TOEIC의 scale은 0~990임을 고려하지 않은 결과이다.
> Min-Max scaling하면
>
> | | **Steve** | **Tony** |
> |---|---|---|
> | GPA | (4.4-0)/(4.5-0)=0.98 | (2.0-0)/(4.5-0)=0.40 |
> | TOEIC | (910-0)/(990-0)=0.92 | (920-0)/(990-0)=0.93 |
>
> Steve의 GPA, TOEIC 평균: 0.95 Tony의 GPA, TOEIC 평균: 0.67 => Steve가 훨씬 우수한 학생인 것을 알 수 있다.

- Z-score standardization: 데이터가 평균값 0과 표준편차 1을 가지도록 scaled 됨

$$Z = \frac{X-\mu}{\sigma}$$

where:

- Z is the standardized value.
- X is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

> Ex) 나이에 대한 데이터 [20, 22, 25, 27, 29] Z-score standardization 하기 μ(뮤, 평균)= (20+22+25+27+29)/5=24.6 σ(시그마, 표준편차)=3.26 [(20-24.6)/3.26, (22-24.6)/3.26, (25-24.6)/3.26, (27-24.6)/3.26, (29-24.6)/3.26] > [-1.41, -0.80, 0.12, 0.74, 1.35]