영등포구

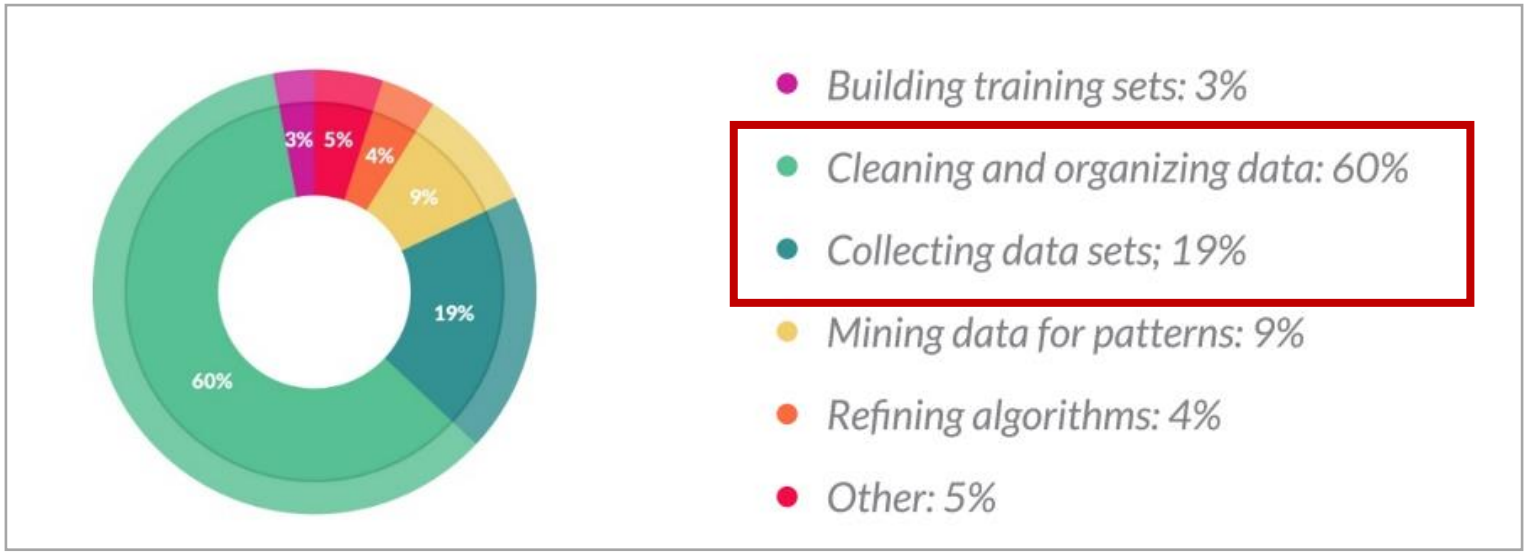
영등포구 청년을 위한

생성형AI 활용 데이터 시각화 교육

데이터 전처리

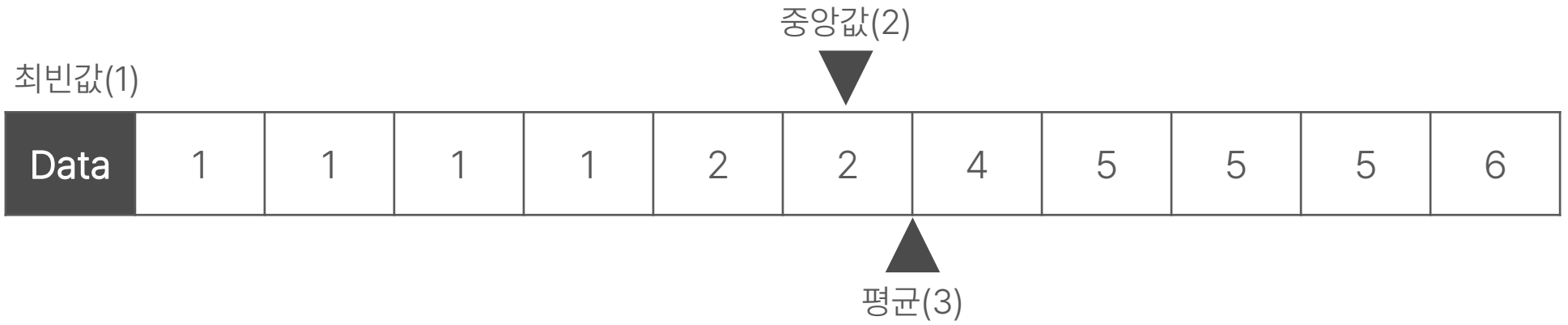
데이터 전처리

- **데이터를 탐색(EDA, Exploratory Data Analysis)하고 분석에 용이하게 변경**
  - 분석이 불가능할 정도로 지저분한 데이터를 처리하는 것(극단치, 이상치, 결측치)
  - 머신러닝에 적용할 수 있도록 가공하는 것
  - 데이터 셋을 만들고 정제하는데 약 80%



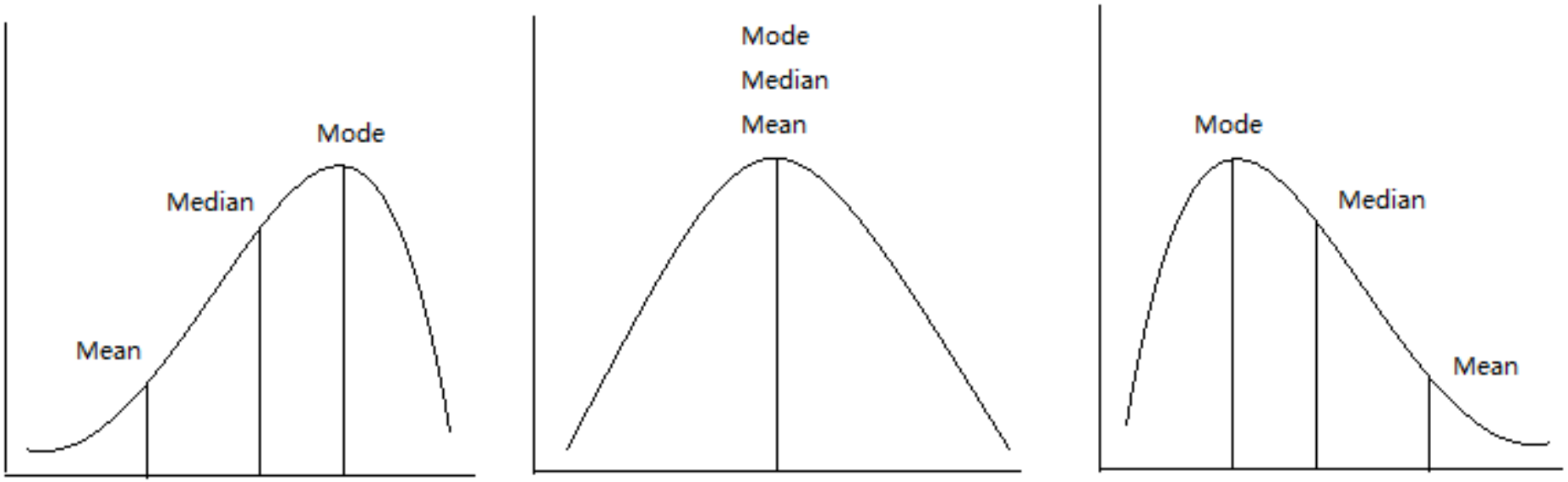
데이터 대푯값

- **하나의 데이터로 전체를 대표하는 값**
  - 평균 : 전체의 수를 더하고 이를 데이터 수로 나눈 값. 데이터 전체의 중심에 해당. 이상치(특이값)에 약함
  - 중앙값 : 데이터를 작은 값부터 순서대로 나열 했을 때 한가운데 위치하는 값. 이상치(특이값)에 영향을 받지 않음.
  - 최빈값 : 데이터 중에서 가장 많이 나타나는 값



데이터 대푯값

- 대푯값에 따라서 데이터 분포의 모양은 아래 3가지로 나타남



Left skew

평균 < 중앙값 < 최빈값

Normal Distribution

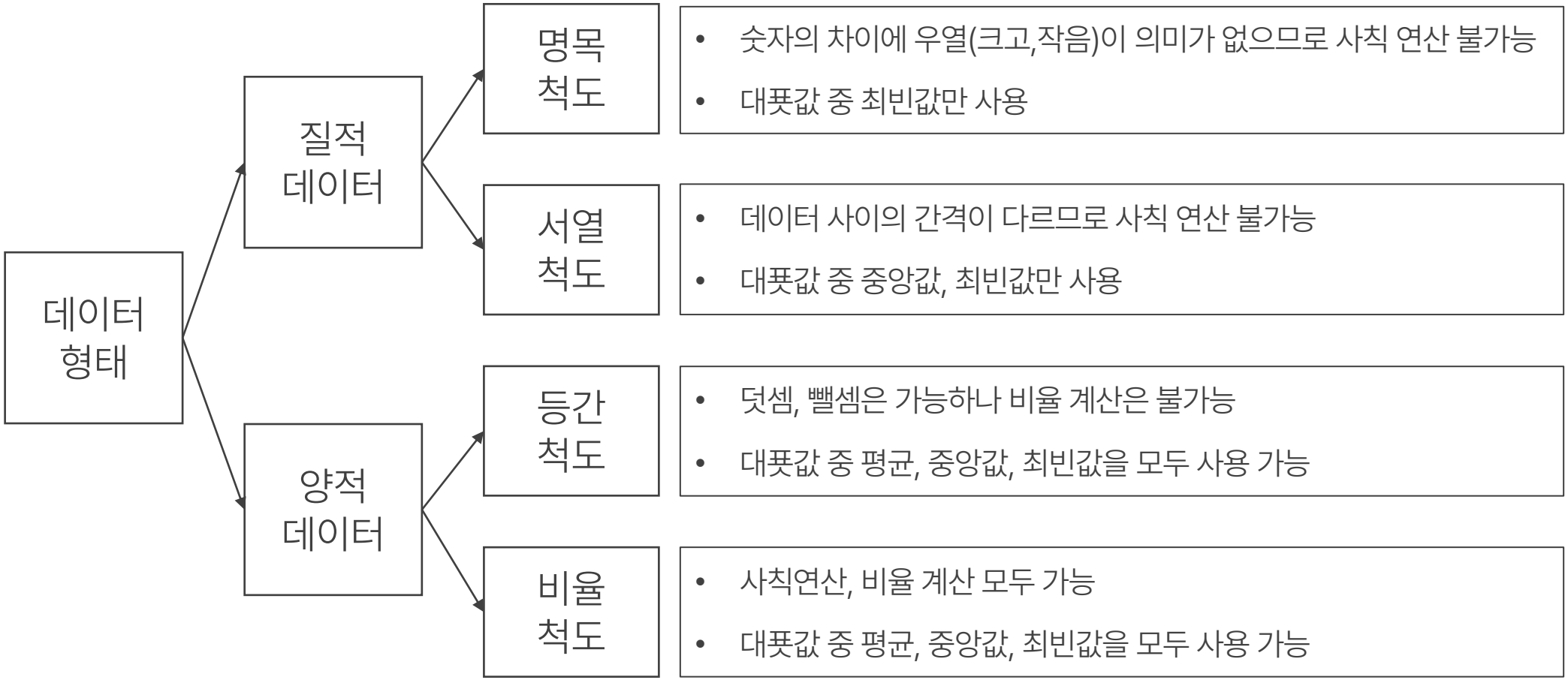
평균 ≒ 중앙값 ≒ 최빈값

Right skew

평균 > 중앙값 > 최빈값

데이터 대푯값

- **데이터 성질에 따라 사용할 수 있는 대푯값이 다름**



척도로 데이터 분류(명목척도)

- **명목척도는 원래 수치 데이터가 아니지만 수치를 부여하여 쉽게 사용**
  - 남자(1) / 여자(2) or 여자(0) / 남자(1) 등과 같이 사용
  - 주소에서 서울=1, 부산=2, 대전=3 등과 같이 사용
  - 우편번호 15231, 23561 등과 같이 사용

- **숫자의 차이에 우열(크고,작음)이 의미가 없으므로 사칙 연산 불가능**

- **대푯값 중 최빈값만 사용**

척도로 데이터 분류(서열척도)

- **서열척도는 순서에 의미가 있는 데이터**
  - 시험 성적 1등, 2등, 3등
  - 1. 매우 불만, 2. 불만, 3. 보통, 4. 만족, 5. 매우 만족
- **데이터 사이의 간격이 다르므로 사칙 연산 불가능**
- **대푯값 중 중앙값, 최빈값만 사용**

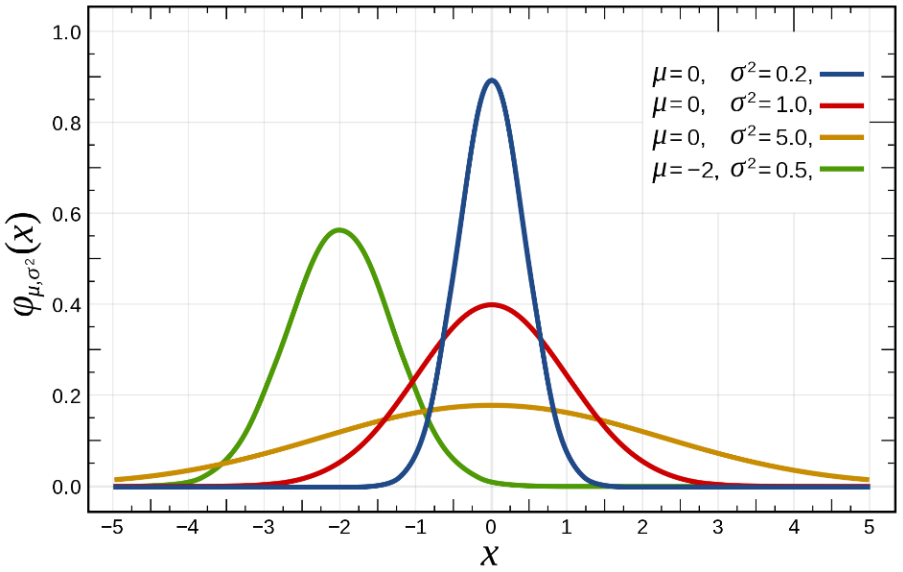
척도로 데이터 분류(등간척도, 비율척도)

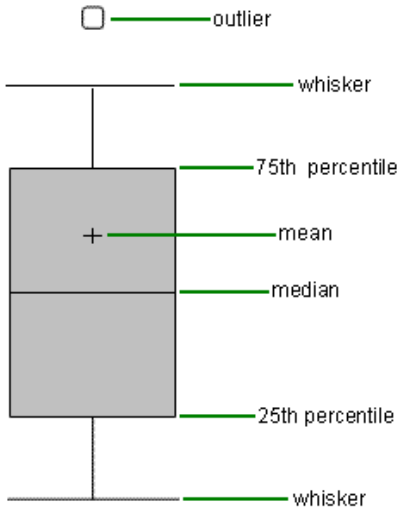
- **등간척도는 데이터간의 간격이 같은 데이터**
  - 온도(섭씨·화씨), 득점 등이 있음
  - 덧셈, 뺄셈은 가능하나 비율 계산은 불가능
  - 대푯값 중 평균, 중앙값, 최빈값을 모두 사용 가능

- **비율척도는 가장 다루기 쉬운 데이터**
  - 키, 몸무게, 시간, 돈, 절대온도 등이 있음
  - 사칙연산, 비율 계산 모두 가능
  - 대푯값 중 평균, 중앙값, 최빈값을 모두 사용 가능

데이터 산포도

- **데이터 전체가 어떻게 퍼졌는지, 흩어짐은 어느정도 인지를 나타냄**
  - 분산(표준편차) : 데이터의 흩어짐 정도를 나타내는 값. 분산과 표준편차는 같은 내용
  - 사분위범위 : 중심 근처의 데이터 흩어짐 정도를 보는 지표
  - 범위 : 데이터가 위치하는 폭(최대-최소)을 나타내는 값

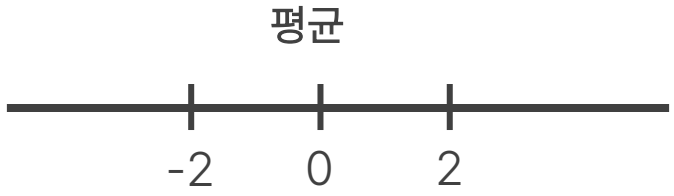




데이터 산포도

- **분산(표준편차)**
  - 데이터의 흩어짐 정도를 나타내는 값(0에 가까울수록 평균 근처에 집중)

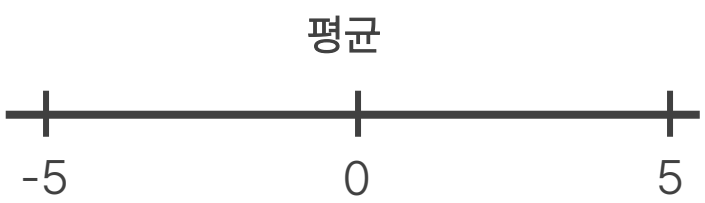
$$분산 = (표준편차)^2$$



$$분산 = \frac{(-2)^2+2^2}{2} = 4$$

표준편차 = 2

$$표준편차 = \sqrt{분산}$$

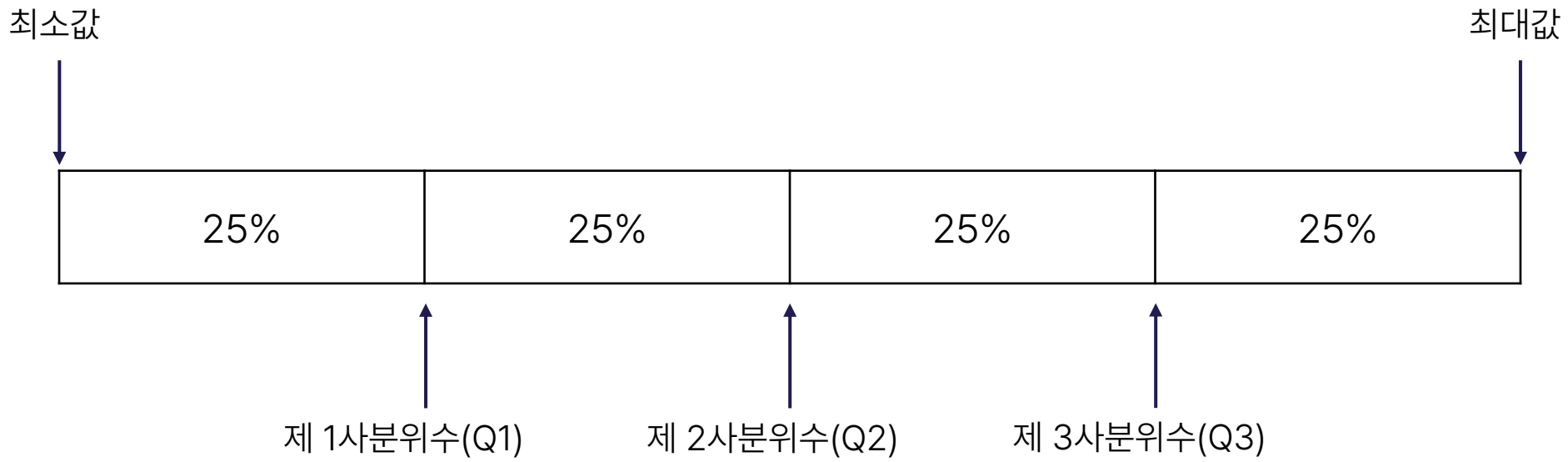


$$분산 = \frac{(-5)^2+5^2}{2} = 25$$

표준편차 = 5

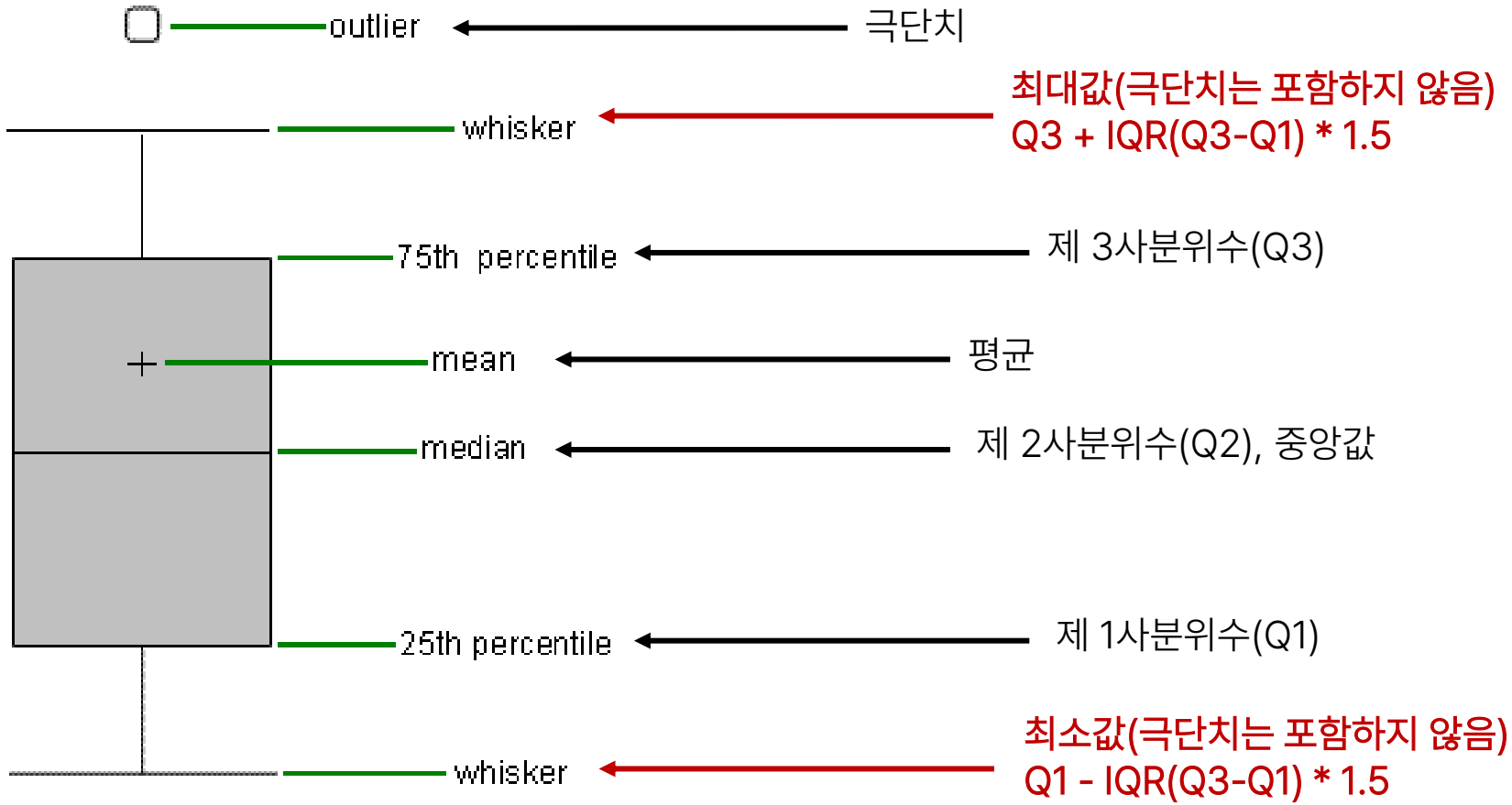
데이터 산포도

- **데이터를 4등분하는 3개의 값을 의미**
  - 제 1사분위수(Q1) : 중앙 값을 기준으로 작은 쪽 데이터의 중앙 값
  - 제 2사분위수(Q2) : 중앙 값과 동일
  - 제 3사분위수(Q3) : 중앙 값을 기준으로 큰 쪽 데이터의 중앙 값



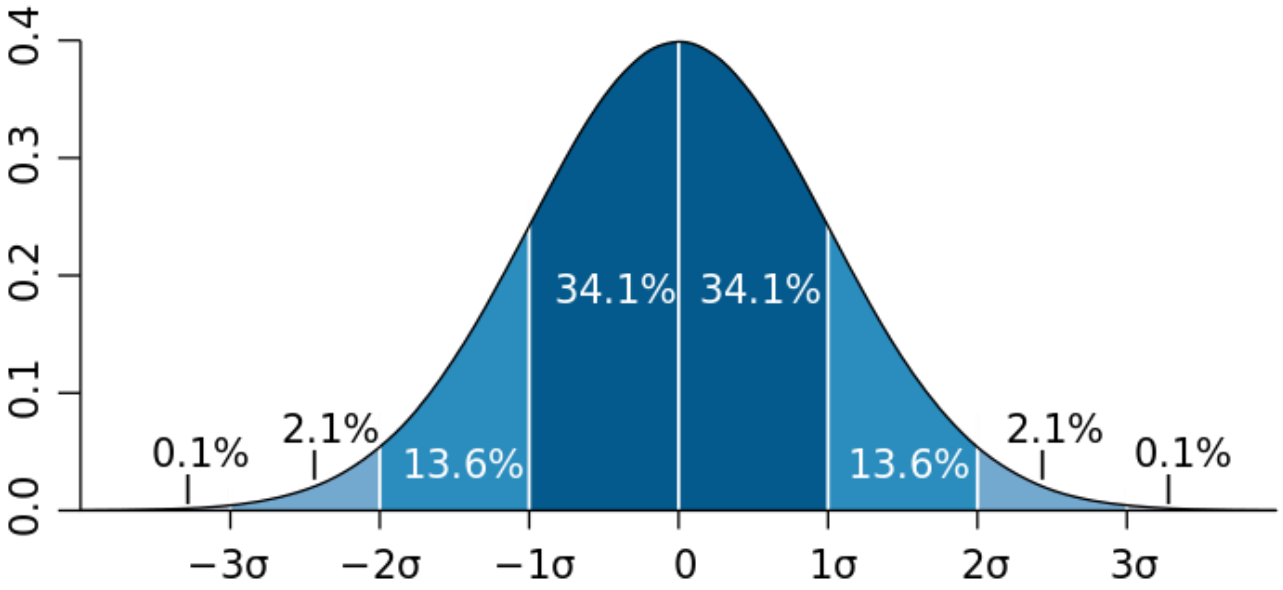
데이터 이상치 확인

- **상자 수염 그림(box and whisker plot)**
  - 데이터의 분포 상태를 알기 쉽게 표현하기 위한 그림



데이터 이상치 확인

- **68-95-99.7 규칙(3시그마 규칙)**
  - 약 68%의 값들이 평균에서 양쪽으로 1 표준편차 범위(μ±σ)에 존재
  - 약 95%의 값들이 평균에서 양쪽으로 2 표준편차 범위(μ±2σ)에 존재
  - 거의 모든 값들(실제로는 99.7%)이 평균에서 양쪽으로 3표준편차 범위(μ±3σ)에 존재



상점의 판매량 예측하기

- **데이터(판매 내역)을 이용하여 년,월,일 단위 상점 및 상품의 판매량 예측**

- **특정 상점의 판매량 예측 가능**
  - A라는 상점의 OOOO년 O월 판매량을 예측
- **특정 상품의 판매량 예측 가능**
  - B라는 상품의 OOOO년 판매량을 예측
- **특정 상점의 특정 상품 판매량 예측 가능**
  - A상점 B상품의 OOOO년 O월 O일 판매량을 예측

구분	내용	데이터 명세
date	판매일자	2013.01.01 ~ 2014.12.31
shop_id	상점 식별 번호	60개의 상점
item_id	상품 식별 번호	21807개의 상품
item_price	상품 가격	최대 : 307980 최소 : -1
item_cnt_day	일일 판매량	최대: 2169 최소: -22

감사합니다.