## 1.1 정형화된 데이터의 요소

정형 데이터의 종류

* 수치형(numeric) 데이터
  + 연속형(continuous) 데이터 – 온도, 풍속.
  + 이산(discrete) 데이터 – 횟수, 빈도
* 범주형(categorical) 데이터
  + 이진(binary) 데이터 – 0 or 1
  + 순서형(ordinal) 데이터 – 평점

데이터를 범주로 구분하는 이유:

* 데이터 분석을 할 때 데이터 종류가 결정적인 역할을 함. 종류마다 사용할 수 있는 기법(시각화, 통계 개념, 해석)이 다름.

## 1.2 테이블 데이터

테이블형 데이터

* 행(레코드, 사건)과 열(피쳐, 변수)로 구성된 2차원적인 데이터.

데이터 프레임과 인덱스

* Python: **pandas**
  + 기본적으로 숫자 인덱싱 있음. 다중/계층적 인덱싱 가능(pivot, reshape).
* R: **data.frame**
  + 마찬가지로 숫자 인덱싱 존재. 하지만 다중/계층적 인덱싱 불가.
    - 보완을 위해 dplyr, data.table 사용됨.

테이블형 이외의 데이터 구조

* 시계열 데이터, 공간 데이터, 그래프 데이터, …
  + 이 책에서는 다뤄지지 않음.

## 1.3 위치 추정

평균

* 가장 기본적인 위치 추정 방법. 모든 숫자들의 합을 숫자 개수로 나눔.
* 변형으로 절사평균, 가중평균 존재.
  + 절사평균: 값들을 크기 순서로 정렬한 후 양 끝 일정 개수의 값들을 제외하고 나머지 숫자들로 평균.
    - 극단값의 영향을 제거한다.
  + 가중평균: 숫자들에 가중치를 붙여 평균.
    - 특이 수치들이 존재할 때,
    - 데이터 그룹의 크기가 다를 때

중간값과 로버스트 추정

* 중간값: 숫자를 일렬로 정렬했을 때 중간에 위치하는 값.
  + 데이터 개수가 짝수라면 가운데 두 값의 평균.
  + 중간값이 많은 경우 데이터에 민감한 평균보다 위치 추정에 유리.
    - 특이값에 로버스트 하다.
* 가중 평균과 같은 목적으로 가중 중간값도 존재함.
  + 구글링 해보니 딱히 많이 사용되지는 않는듯.

특이값

* 중간값 = 특이값 영향이 없는 로버스트한 위치 추정 방법.
* 절사평균도 특이값 영향을 줄이기 위해 사용.
  + 절사평균은 중간값과 평균의 절충안
    - 특이값에 로버스트 하면서도 많은 데이터 사용.

## 1.4 변위 추정

위치는 데이터의 특징을 파악하는데 사용되는 요소 중 하나. 다른 하나가 **변이**. 변이는 산포도를 나타낸다.

표준편차와 관련 추정값들

* 가장 기본이 되는 변이 추정들은 **편차**를 기본으로 한다.
  + 편차는 관측값과 위치 추정값의 차이. 데이터가 중앙에서 얼마나 퍼져있는지.
* 변이 측정법 중 하나는 편차의 대푯값을 추정하는 것.
  + 평균은 대푯값이 될 수 없음. 편차의 합은 0.
  + 편차를 절대값 씌우고 데이터 개수 n으로 나누면 **평균절대편차.**
* 가장 유명한 추정 방법은 **분산**과 **표준편차**.
  + 분산 = 편차의 제곱을 n-1로 나눈 값.
  + 표준편차 = 분산을 제곱근한 값
* 분산, 표준편차, 평균절대편차 모두 특이값에 로버스트하지 않다.
  + 특히 분산과 표준편차는 편차를 제곱하기 때문에 더 민감하다.
* 로버스트한 변위 추정값으로는 중위절대편차(MAD)가 있다.
  + 중위절대편차: 중간값과의 편차의 절대값의 중간값.
  + 중간값의 특성과 동일하게 특이값에 로버스트하다.
  + 절사평균처럼 절사표준편차도 존재한다.

백분위수에 기초한 추정

* 변이를 추정하는 또 다른 방법: 정렬된 데이터가 얼마나 퍼져 있는지.
  + 순서통계량: 정렬(순위) 데이터를 나타내는 통계량.
  + 최대, 최솟값의 차인 범위는 특이값 분석에 유용하지만 특이값에 민감해 변이 추정에는 유용하지는 않다.
* 백분위수의 차를 사용하면 특이값의 영향을 덜 받고 변이 추정을 사용할 수 있다.
* 변이를 측정하는 대표적인 방법은 사분위범위(IQR). 25번째 백분위수와 75번째 백분위수의 차이.
  + 작은 데이터셋에서는 사분위범위가 유용할 수 있지만, 큰 데이터셋에서는 순서통계량 수집에 많은 연산이 필요로 한다.
    - 그래서 머신러닝과 통계 소프트웨어에서는 백분위수의 근삿값을 사용한다.