## 1.1 정형화된 데이터의 요소

정형 데이터의 종류

* 수치형(numeric) 데이터
  + 연속형(continuous) 데이터 – 온도, 풍속.
  + 이산(discrete) 데이터 – 불량품 수, 사고 건수
* 범주형(categorical) 데이터
  + 이진(binary) 데이터 – 0 or 1
  + 순서형(ordinal) 데이터 – 평점, 성적표
  + 명목형(nominal) 데이터 – 성별, 색깔, 취미

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.
  + 편차의 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 제곱들의 합을 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.

# 1.5 데이터 분포 탐색하기

위치 추정은 말 그대로 데이터의 위치 추정.

변이 추정은 데이터가 밀집인지 퍼져있는지 추정.

지금까지의 추정은 하나의 수치로 알아봤는데, 전체적으로 어떻게 분포하는지 파악하는 방법을 알아보자.

백분위수와 상자그림

* 백분위수는 변이 추정에도 사용되지만, 분포 추정에도 사용된다.
* 상자그림은 백분위수로 표현한 분산을 그림으로 나타내는 것.
  + 기본적으로 수염은 사분위범위(IQR)의 1.5배 이상 멀리 나가지 않도록 한다.

도수분포표와 히스토그램

* 도수분포표는 일정한 크기로 구간을 나누고 데이터를 담아둔 표.
  + 도수분포표는 구간을 일정하게 나눈 것, 백분위수는 구간 안에 같은 수의 데이터.
* 히스토그램은 도수분포표를 시각화한 것.
  + 구간의 크기를 바꿔보는 것으로도 유용한 정보를 얻을 수 있음.
  + 막대는 서로 공간 없이 붙어있다.

밀도추정

* 밀도추정은 커널밀도추정으로 통해 히스토그램을 연속된 선으로 표현한 것.
  + 히스토그램과의 가장 큰 차이는 y축이 개수가 아닌 비율을 표현한다.

# 1.6 이진데이터와 범주 데이터 탐색하기

범주형 데이터는 비율과 퍼센트로 데이터를 표현한다.

* 범주에는 서로 다른 집합(남자와 여자), 요인의 수준(낮음, 높음), 구간 별로 나뉜 수치 등이 있다.
* 범주형 데이터를 시각화하는 방법으로는 막대도표(barplot)이 있다.
  + x축은 범주, y축은 횟수나 비율을 나타낸다.

막대도표와 히스토그램은 유사하다.

* 히스토그램은 수치를 순서를 고려한 요인들로 변환한 것.
  + 수치를 범주로 변환하면 데이터의 복잡도를 줄일 수 있다.
  + 데이터 분석 초기 단계에 feature들 사이 관계를 알아볼 수 있다.

최빈값

* 가장 자주 등장하는 값(들). 수치형 데이터에서는 자주 사용하지 않는다.

기댓값

* 기댓값 = 가중평균
  + 5%가 30만원, 15%가 5만원, 나머지 80%는 구매를 안한다면,
    - *EV* = (0.05)(300) + (0.15)(50) + (0.8)(0) = 22.5
    - 미래의 기댓값(0, 5, 30만원)과 가중치(0.05, 0.15, 0.8)의 곱을 모두 더한 값.
* 요인변수의 수준을 요약하는데 사용된다.

# 1.7 상관관계

EDA에서 feature 사이나 feature와 target 사이 상관관계 조사는 필수적이다.

* 대표적인 상관계수로는 피어슨, 스피어맨, 타우 상관계수가 있다.
  + 스피어맨이나 타우 계수는 데이터를 순위를 기초로 하기 때문에 비선형적인 데이터에도 유용하지만, 작은 양의 데이터나 특별한 가설검정에만 사용된다.
  + 피어슨 상관계수는 변수1과 변수2의 평균으로부터 편차를 곱한 값들의 평균(N-1로 나눔)을 각 변수들의 표준편차로 나눈 값이다.
* 평균과 표준편차와 같이 특잇값에 민감하다.
  + Trim으로 좀 더 로버스트하게 변환해서 사용할 수도 있다.

산점도

* 두 변수 사이 관계를 시각화하는 가장 기본적인 방법.

# 1.8 두 개 이상의 변수 탐색하기

평균과 분산같은 추정값들은 하나의 변수만 다룬다(일변량분석). 상관분석의 경우 두 변수를 비교한다(이변량분석). 물론 다변수분석도 가능하다.

육각형 구간과 등고선(수치형 vs. 수치형)

* 산점도는 데이터 개수가 적을 땐 유용하지만, 데이터 개수가 많아지면 육각형 구간(hexagonal binning)과 등고도표(contour plot)가 유리하다.
  + 육각형 구간은 데이터를 점으로 표시하는 대신 육각형 구간의 데이터 개수를 색깔로 표현한다.
  + 등고도표는 산점도 위에 등고선을 덧붙인다. 같은 등고선은 동일한 밀도를 가지고 꼭대기로 오를수록 밀도가 높다.
* 다른 방법으로 히트맵이 있다.
  + 세 방법 모두 이차원상의 밀도를 시각화하는데 사용된다.

범주형 vs. 범주형 변수

* 분할표(contingency table)는 두 범주형 변수를 분석하는데 유용하다.
  + 분할표는 해당 항목의 빈도와 백분율을 보여준다.

범주형 vs. 수치형 변수

* 상자그림(boxplot)은 범주형 변수로 분류된 수치형 변수를 시각화해 비교하는 간단한 방법이다.
* 바이올린 도표(violinplot)은 상자그림을 보완했다.
  + y축을 따라 밀도추정을 시각화했다. 상자그림에서 보이지 않는 데이터의 분포를 볼 수 있다.
  + 상자그림은 특잇값을 파악하는데 유용하다.

다변수 시각화하기

* 조건화라는 개념을 통해 두 변수를 비교하는 도표를 많은 변수 분석에 적용할 수 있다.
  + 예를 들어, 특정 지역의 주택 면적과 토지과세율의 상관관계를 파악할 때, 해당 지역을 우편번호별로 범주를 분류해 세밀하게 파악할 수 있다.