탐색적 데이터 분석 (EDA)

- 탐색적 데이터 분석 (Explanatory Data Analysis, EDA)

  -> 데이터를 가지고 여러 방향으로 데이터를 탐색하고, 데이터의 특징과 구조로부터 얻은 정보를 바탕으로 통계모형을 만드는 분석 방법

탐색적 데이터 분석에서 가장 중요한 것

- 데이터로부터 인사이트를 파악해볼 수 있는 단서들을 도출해 내는 것!
- 이를 위해 여러 가지 방식으로 통계량도 내어보고, 시각화도 해 보는 과정이 중요함
- 내가 탐정이 되었다고 생각하자.

"어떤 비밀이 있을까..."

기술통계량 (Descriptive statistics)

- 데이터 샘플의 특징을 정량적으로 설명해 주는 통계량
- 크게 위치 모수, 척도 모수, 기타 척도 등으로 나눌 수 있음
- 탐색적 데이터 분석 뿐만 아니라 확증적 데이터 분석에도 쓰임

위치 모수 (Location parameter)	척도 모수 (Scale parameter)	기타 척도
데이터의 갯수 (count)	분산 (variation)	왜도 (skewness)
평균 (mean, average)	표준편차 (SD, standard deviation)	첨도 (kurtosis)
중앙값 (median)	변동계수 (CV, coefficient of variation)	상관계수 (Correlation coefficient))
최빈값 (mode)	범위 (range)	
최소값 (minimum)	분위수 (quantile)와 사분위수(quartile), IQR	
최대값 (maximum)	백분위수 (percentile)	

위치 모수 (Location parameter)

- 데이터 샘플의 산포를 나타내는 대푯값

데이터셋 D의 수가 n개, 데이터를 각각 $x_1, x_2, \dots, x_n$이라 한다면

갯수 : n

평균 : $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{\sum x_n}{n}$

중앙값 : $m_e = x_{\frac{n+1}{2}}$ (n이 홀수일 때)

$m_e = \frac{x_{\frac{n}{2}}+x_{\frac{n+1}{2}}}{2}$ (n이 짝수일 때)

최빈값 : x_{freq}

최소값 : x_{min}

최대값 : x_{max}

척도 모수 (Scale parameter)

- 데이터 샘플의 크기를 나타내는 대푯값

데이터셋 D의 수가 n개, 데이터를 각각 $x_1, x_2, \dots, x_n$이라 한다면

모분산 : $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

표본분산 : $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

모표준편차 : $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

표본표준편차 : $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

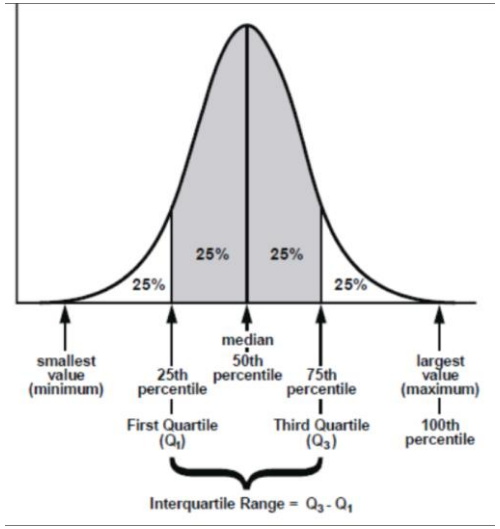
변동계수 : $CV = \frac{s}{\bar{x}}$

범위 : $x_{max} - x_{min}$

분위수, 사분위수, 백분위수, IQR

척도 모수 - 분위수, 사분위수, 백분위수, IQR

- Quantile (분위수) : 주어진 데이터를 동등한 크기로 분할하는 지점
- Quartile (사분위수) : 크기 순서로 정렬한 데이터를 4분할하는 관측값
- Percentile (백분위수) : 크기 순서로 정렬한 데이터를 100등분 했을 시 x%인 관측값
- IQR (InterQuartile Range) : $Q_3 - Q_1$

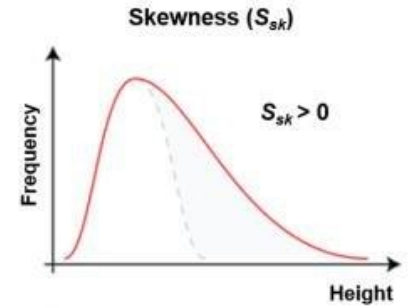


Quartiles = 4-Quantiles

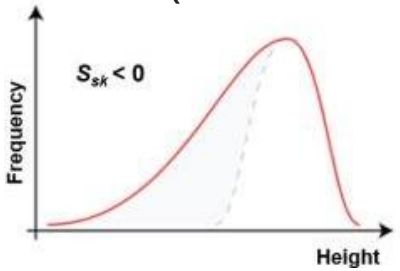
- 1사분위수 = First Quartile = 0.25 Quantile = Q_1 = 25th Percentile
- 2사분위수 = Second Quartile = 0.5 Quantile = m_e = Q_2 = 50th Percentile
- 3사분위수 = Third Quartile = 0.75 Quantile = Q_3 = 75th Percentile

기타 척도

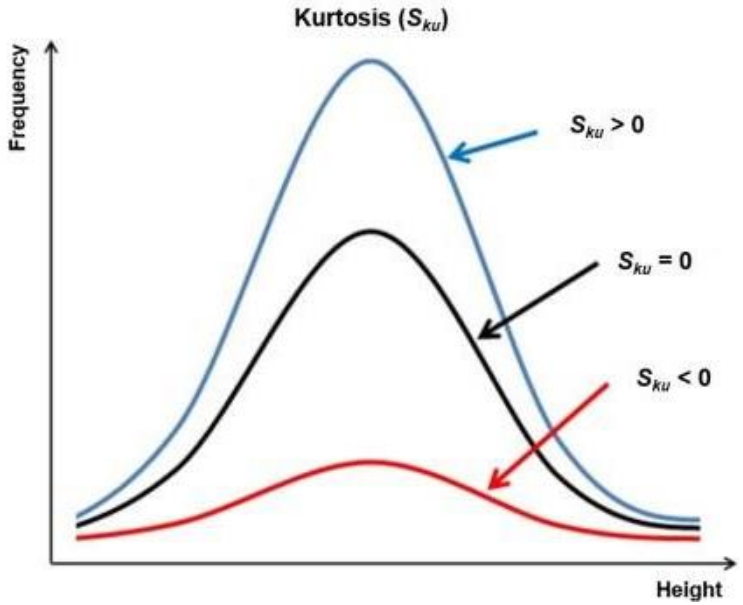
- 왜도 (Skewness) : 분포의 비대칭도를 나타내는 통계량
- 첨도 (kurtosis) : 분포가 얼마나 평균 근처에 몰려 있는지를 나타내는 통계량.



Right-skewed (Positive skewness)



Left-skewed (negative skewness)



* 정규분포의 첨도는 3

기타 척도

- 상관계수 : 두 변수 사이의 통계적 상관관계의 정도를 수치적으로 나타낸 계수.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}}$$

두 변수 (x, y)에 대하여 관측값 n개의 짝 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 이 주어질 때 상관계수 r은 다음과 같이 계산한다.

$(-1 \le r \le 1)$

$$\bar{x} = \frac{1}{n}\sum x_i \,, \bar{y} = \frac{1}{n}\sum y_i$$

$$S_{xx} = \sum (x_i - \bar{x})^2$$

$$S_{yy} = \sum (y_i - \bar{y})^2$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

시각화 (Visualization)

- 데이터의 생김새를 파악하기 위해 이미지, 다이어그램 등으로 데이터를 표시하는 방법
- 데이터 사이의 관계를 식별하거나, 숨겨진 패턴 등을 추론하는 데에 유용함

파이썬 시각화 패키지

matplotlib

seaborn