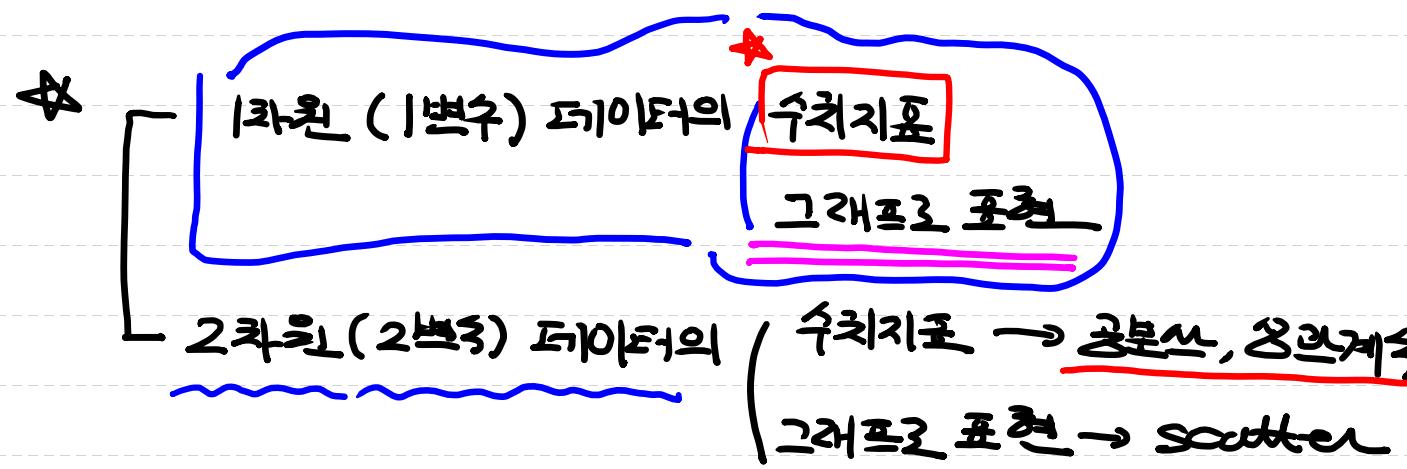


• 03/23

## "기초통계"



[1, 2, 3, 4, 5, 10000] ↑  
이상치

• 1차원 (1변수) 데이터의 수치 지표.

\* "평균값" → 데이터를 하나의 값으로 요약한 지표

→ 평균을 많이 이용!! → 단점) 이상치에 악영향

(mean) → 코드로 구해보아요~~

평균을  
평균값으로  
사용하기가  
곤란

제거평균  
(Trimmed mean)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + x_2 + x_3 + \dots + x_n)$$

① “중위값” (median)

▼ [ 1, 2, 3. 4. 5, 10, 12 ]  
    ↓  
    4가 중위값                                  → code3 구해보아요~

[ 1. 2. 3. 4. 5, 6 ]  
    ↓  
    ~~평균으로~~ 중위값 계산

② “최빈값” (mode) → 가장 많이 등장한 값

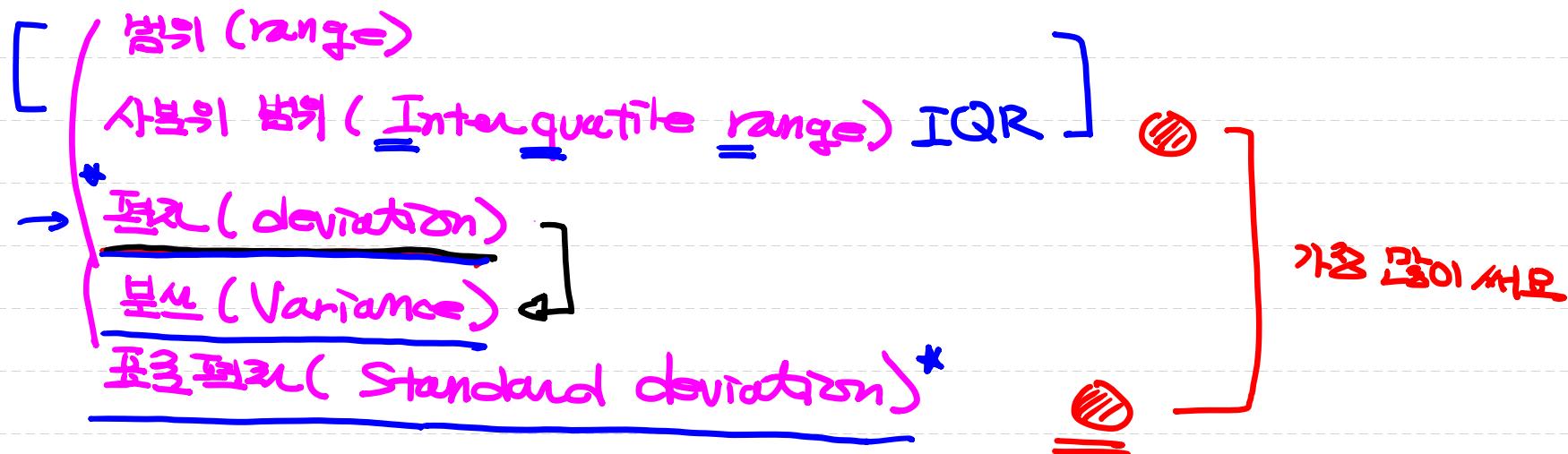
예) [ 1, 1, 1. 2, 2. 3 ] → ①  
    ↳ Numpy는 mode(> 찾고자 하는 값)이고, 대신 Pandas는 가지고 있어요

“최대, 최소” → 대푯값으로 사용하기에도 무리가 없어요~~

- 데이터가 얼마나, 어떻게 퍼져 있는지에 관심

\* 산포도 (Dispersion)

- 데이터가 흩어져 있는 정도 (변동성)을 수치로 표현하고 싶어요



편차 (deviation) : 각 데이터가 평균으로부터 어느정도 흩어져 있는지에 대한 지표

↳ 영향력을 가지고 deviation은 구해보아요 ~

↳ 구해보는건 값이 여러개예요!! → 평균이기 때문에 → 하나의 값으로 만들려고 해요

\* 분산 (Variance) ← 평균을 약간이나마 그대로 방법을 이용해내 해결 ← 그래서 평균을 구해보는걸 ⇒ ○이네!!

- 분산 (Variance) → 평균의 제곱의 평균  
 ↳ 코드로 계산해 보아요 ~

\* 표본분산 ( $\text{ddof} = 0$ )  
 불편분산 ( $\text{ddof} = 1$ )

### 수식표현

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

↓  
deviation(평균)

분산을 이용해서 데이터의 훨씬집 척도(偏差量)를 하나의 숫자로 표현할 수 있어요!!

↳ 하지만 제곱을 사용하기 때문에 실제 범위보다 더 큰값을 가지고 계산!!



그래서 원래 크기로 돌아오려면 "분산의 제곱근"을 구하면 돼요.

\* ↳ 표준 편차 (Standard deviation)



- Range (범위)

$$R_g = \underline{x_{\max}} - \underline{x_{\min}}$$

Range = 값이 크면  
 → 스프드가 커요.  
 Range = 값이 작으면  
 → 스프드가 작아요!

- 간단한 계산이 편집  
 “이상치가 아주 많음”

- Interquartile range (사분위 범위) → IQR

이상치를 빼는 때 사용

드레이터를 하위 25%, 50%, 75%.

\*  $\text{IQR} = \underline{\underline{Q_3 - Q_1}}$

Q1      ↘      Median (Q2)      ↗      Q3

사분위      2사분위      3사분위

↓      ↓      ↓

코드로 구해보아요 ~

- 1차원(1변수) 데이터의 시각화 → 데이터의 분포상태를 그래프로 표현  
[표로 표현]

데이터 분포상태를 도표로 확인

↳ 데이터가 봄하는 값으로 몇 개의 구간으로 나누고  
각 구간이 몇 개의 데이터가 놀아가 있는지를 세어도 표로 만들어요

frequency distribution Table

⇒ 도수분포표

ex)  
→ 시험성적(점수)을 이용해 도수분포표를 만들어요.

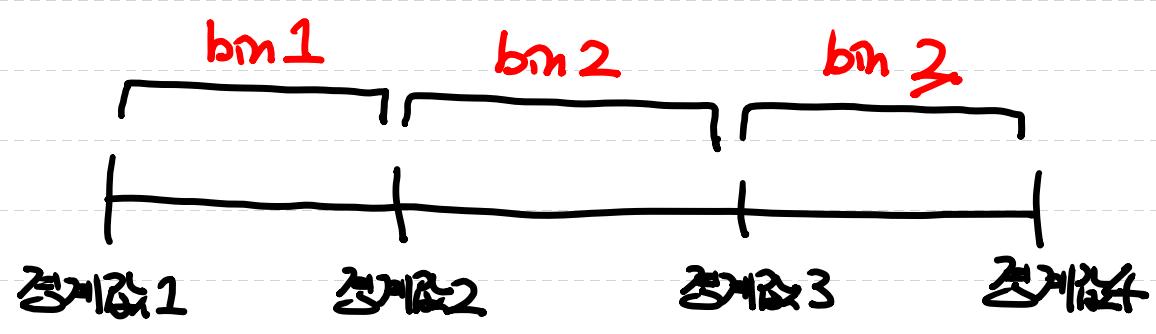
(개의) 구간 →	0점 ~ 10점 : x 명
	11점 ~ 20점 : x 명
	21점 ~ 30점 : x 명
	:
	91 ~ 100점 : x 명

\* 놓아가 놔요!!!

- 0점 ~ 10점 ⇒ 구간 class (계급)
- 각 구간(class)에 속한 흔적수: frequency (도수)
- 구간의 폭 (10점) : class Interval (계급의 크기)

★ ★ DataFrame으로  
만들어요!!

- 계급수 (class의 개수) 10개



## • Frequency Distribution Table (도수분포표)

→ class (계급)

frequency (도수)

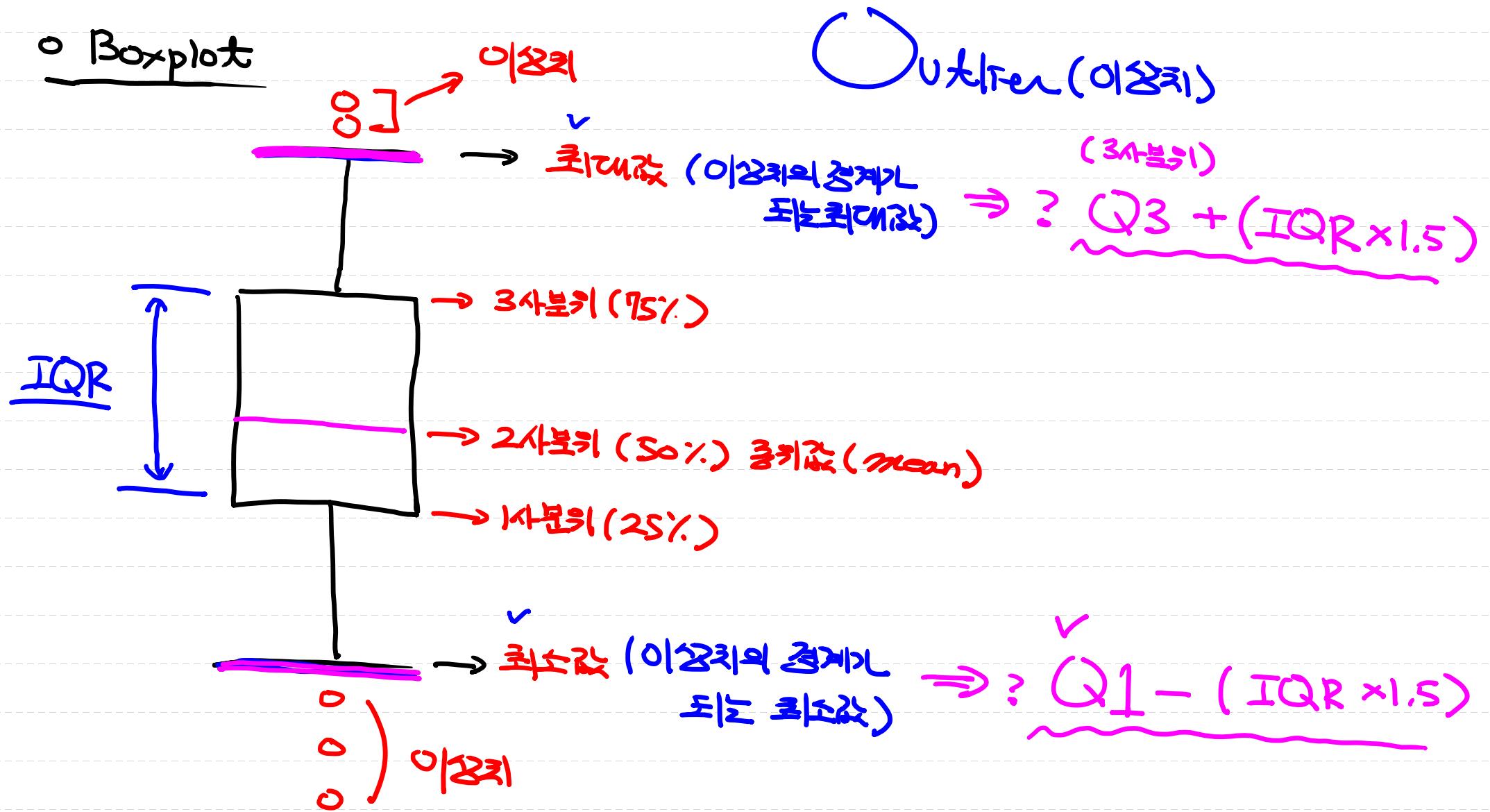
- ① Class Mark (계급값) → Class를 값으로 표현 → 중위값 이용  
(구간) (median)
- ② relative frequency (상대도수) → 조체크이터에 대해 해당 class의 도수가 얼마 만큼의 비율을 차지하고 있는가
- ③ cumulative relative frequency (누적상대도수)  
→ 해당 class까지 상대도수의 누적.

“  
포트폴리오 코드로 알아보아요”

\*  
코드로  
알아보아요~

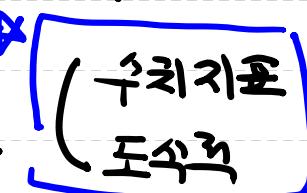
- Histogram → | 치환 데이터의 봉고를 흐름에 모아보기 | 카트고 (그래프)
- Boxplot → “ “ “ “ ”

## Boxplot



→ 우리가 가지고 있는 영어성적을 가지고 Boxplot을 그려보자요~

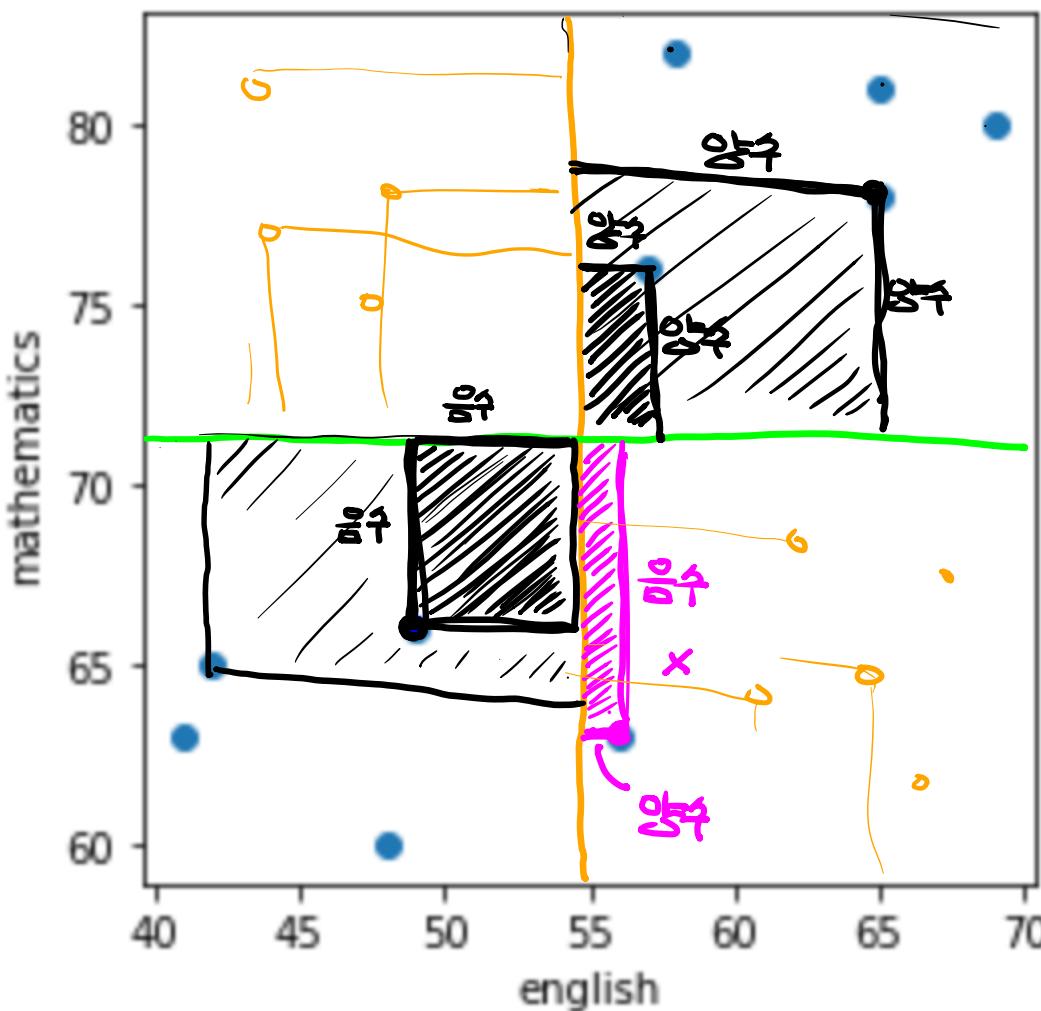
⇒ 여기까지가 1차원(1변수) 데이터에 대한

\* 
 주차자료  
도수분포

- 2차원(2변수) 데이터의 수치적표 [ 정분식 상한계 ] → 고도로 알아보아요!!  
그래프.

→ 사용할 Data를 준비!!

→ scatter를 이용해 그림으로 그려보아요!!



양수 양수

↑  
수학영어 양수

⇒ 양수 양수  
수학영어 ]

"양의 상관관계"

## • Covariance (공분산)

\*  $S_{xy} = \frac{1}{n} \sum_{i=1}^n (\bar{x} - x_i) \times (\bar{y} - y_i)$

↓      ↓  
 영어      수학  
[ ]      [ ]  
 영어표현      deviation

공분산값이 양수  $\Rightarrow$  2차원 데이터(2번쨰 데이터)가

“양의 상관관계”를 가져요 !!

\* [ ]

결국 공분산값은 숫자로 계산되며  
그 숫자크기로 2차원데이터가 얼마나 모종의  
관계가 있는지를 알 수 있습니다!!

\* 공분산이 0에 가까울 때 두 데이터는  
서로 상관관계가 없이고 모종의 있나요 !!