평균

① 산술평균

$\overline{x} = (x_1 + x_2 + \cdots + x_{n-1} + x_n) \div n$ where x is variable, n is number of data

② 기하평균 Geometric Average

$$\overline{X_G} = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_{n-1} \cdot x_n}$$

- 연 평균 성장률이나 전년대비 수익평균 구할때 사용

③ 조화평균 Harmonic Average

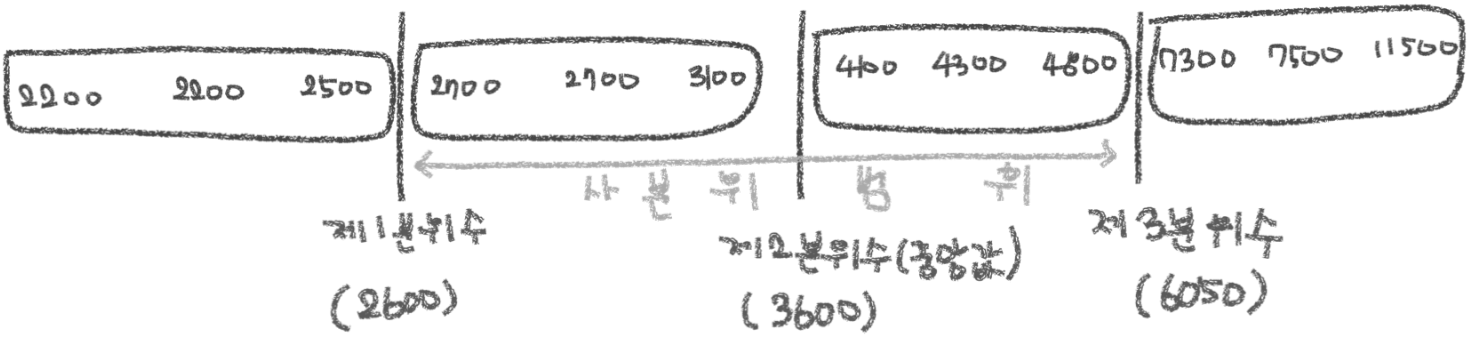
$$\overline{X_H} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_{n-1}} + \frac{1}{x_n}}$$

- 일정거리 이동시 평균 속도 구할때 사용

최댓값, 최솟값, 분위수, 사분위 범위, 분산 (표준편차) 등의 지표를 이용하여 데이터의 흩어진 정도 파악

분위수

n개의 데이터를 작은 수부터 큰 수의 순으로 늘어놓고 그것을 K등분했을때 경계된 수치를 분위수라 한다



사분위 범위 = 제3분위수 − 제1분위수

데이터가 중앙값 주위에 집중할수록 사분위 범위는 작아진다.

편차

편차 (d_i) = 데이터 값 (x_i) − 평균값 $(\overline{x})$

편차가 큰 데이터가 많으면 분산 크기가 큰 데이터 세트라고 할수 있다.

분산

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$ 여기서 S는 표준편차.

변동계수 Coefficient of Variation

$$CV = S / \overline{X}$$

- 두개의 데이터가 흩어진 정도를 비교하는 경우에 사용

상관계수 Coefficient of Correlation

두 변수가 한쪽이 증가할때 증가(감소)하는 직선적인 관계를 가질때 '상관'이라 한다.

피어슨의 적률 상관계수

$$\text{상관계수 } r = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=0}^{n}(y_i - \overline{y})^2}}$$

- $-1 <$ 상관계수 $r < 1$
- $|r| \to 1$ 이면 상관이 강해진다
- $\begin{bmatrix} r > 0 \\ r < 0 \end{bmatrix}$ 이면 $\begin{bmatrix} \text{양의 상관} \\ \text{음의 상관} \end{bmatrix}$

- 세 소비자의 순위 데이터에 대해 '순위가 일치'하는 경우에 ○을, '순위가 불일치'하는 경우에 ×를 할당한다.

소비자	x의 순위	y의 순위	소비자 1	소비자 2	소비자 3
1	1	2			
2	2	1	×		
3	3	4	○	○	
4	4	3	○	○	×

	소비자 1	소비자 2	소비자 3	계
○의 수	2	2	0	4
×의 수	1	0	1	2

- 켄달의 순위상관계수는 A=○의 수, B=×의 수, n=데이터 쌍의 수(예에서는 4)로 했을 때, 다음 식으로 구할 수 있다. 같은 순위가 있는 경우는 계산식이 달라진다.

켄달의 순위상관계수

타우

$$\tau = \frac{(A-B)}{(n\text{개에서 }2\text{개를 골라내는 조합의 수})}$$

$$= \frac{4-2}{\frac{1}{2}\cdot 4\cdot (4-1)} = 0.33$$

[통계학도감 참조]