• 04강 공분산과 상관계수 →

데이터 처리 및 머신러닝에서 매우매우 중요한 개념임
평균, 분산은 그 데이터의 특징을 나타내기 위한 도구들임.
그러나 더 확장해서 **두 데이터 간의 관계**를 밝히고 싶다.
실제로 머신러닝, 빅데이터 분석 등등은
전부 두 데이터 간의 관계를 분석하고 싶은 거임.
우리가 쉽게 알 수 있는 데이터로부터
우리가 알기 어려운 데이터나 미래의 데이터를
"예측" or "분석" 하는 것이 데이터 분석의 전부임
그래서 두 데이터 간의 관계를 구할 때
가장 기본적으로 사용되는 게 공분산

- 공분산 (covariance) : 두 데이터 간의 관계

$[a,b,c]$ 와 $[d,e,f]$
평균 m　　평균 n

$$cov = \frac{(a-m)(d-n)+(b-m)(e-n)+(c-m)(f-n)}{N}$$ ← 편차끼리 곱셈

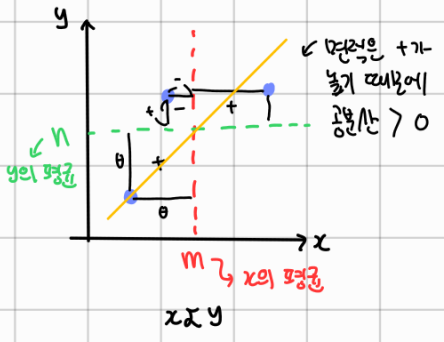
N 개수 3

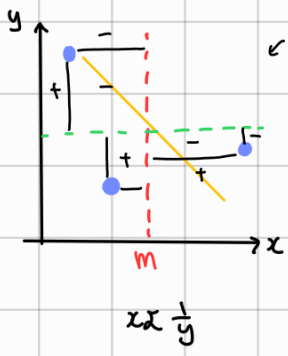
문제) $[1,2,3]$ 과 $[1,3,5]$ 의 공분산은?
$m=2$　　$n=3$

$$cov = \frac{(1-2)(1-3)+(2-2)(3-3)+(3-2)(5-3)}{3} = \frac{2+2}{3} = \frac{4}{3}$$

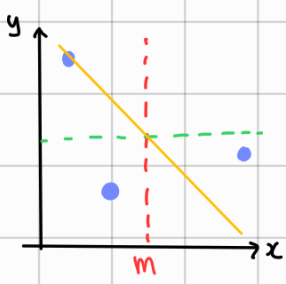
값은 구했는데 의미를 잘 모름...

- 두 데이터의 관계





공분산 < 0



- 공분산 > 0 ⇒ 비례
- 공분산 < 0 ⇒ 반비례

] 선에 붙어있다면 더 '확실히' 늘고 줄다는 말.

- 선을 그었을 때 대각선에 가까울수록 공분산 값의 절대값은 커진다.

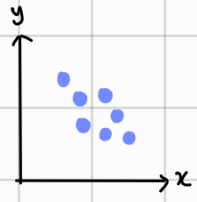
- 상관계수 : 두 데이터 간의 관계 (표준화)

$$r = \frac{Cov}{\sigma_x \sigma_y}$$

(Cov: 공분산, σ_x: x의 표준편차, σ_y: y의 표준편차)

→ r값은 $-1 \le r \le 1$

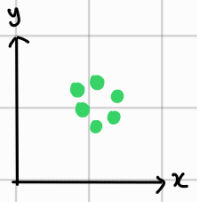
원래는 이것을 스케일링 이라고 함



$-1 \le r < 0$

음의 상관

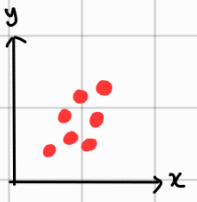
$x \propto \frac{1}{y}$



$r = 0$

독립

→ 서로 영향을 주지 않음



$0 < r \le 1$

양의 상관

$x \propto y$

- 공학에서 0.8 이상, 자연과학에서 0.6 이상이면 상관관계가 높다고 함
- 0~0.2면 상관관계 X

→ 쓸모없다는 게 아님. 두 데이터가 서로 독립적이라는 정보도 충분히 좋은 연구가 될 수 있다.

→ 대표적으로 상관관계는 없는데 높다고 인식하는 것 → 미신, 슬럼프 등등...

$$r = \frac{Cov}{\sigma_x \sigma_y} = \frac{\frac{4}{3}}{\sqrt{\frac{2}{3}}\sqrt{\frac{44}{3}}} = \frac{\frac{4}{3}}{\sqrt{\frac{16}{3}}} = \frac{\frac{4}{3}}{\frac{4}{3}} = 1$$

현실에선 잘 나오지 않는 숫자...