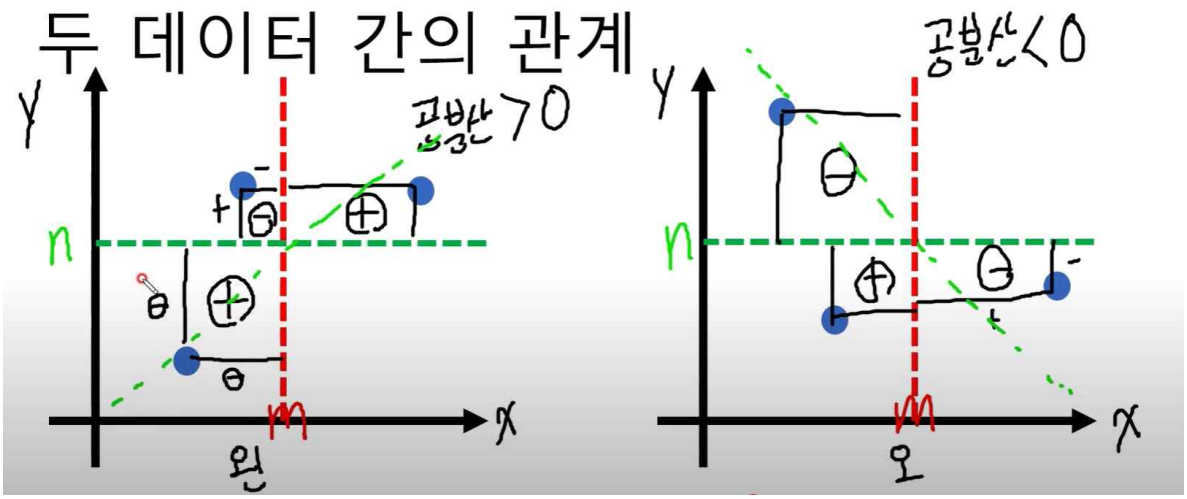
공분산과상관계수는 데이터처리및머신러닝에서 매우중요한 개념이다.
지금까지 다루었던 평균 분산은 그 데이터특징을 나타내는 도구였다.
이제는 두데이터간의관계를 분석하기위해서 두데이터의 관계를 구할때의
공분산이다. covariance
문제[1,2,3]와[1,3,5]의 공분산은?
[a,b,c] ==m와[d,e,f]==n 은 각각의 편차끼리의 곱을 더해서 n으로나눈다.
cov=(a-m)(d-n)+(b-m)(e-n)+(c-m)(f-n)/n
(1-2)(1-3)+(2-2)(3-3)+(3-2)(5-3)/3
2+2=4/3



왼쪽에서 x가오를때에 y도 같이오른다.공분산이 크다는 것은 하나가 오를때에 확실히하나가 오른다.공분산이 작다는거는 하나가 오를때에 확실히 하나가 작아진다.
그래서 얼마나 큰거냐를 알기위해 표준화를 해준다.
상관계수 두데이터의 관계(표준화)
스케일링후 -1<=r<=1
v=cov/σxσy 공학과학에서 0.8이면 상관관계가높다.
4/3/√2/3*√8/3 =1이 나온다.
1이나오면 뭔가 이상하다.