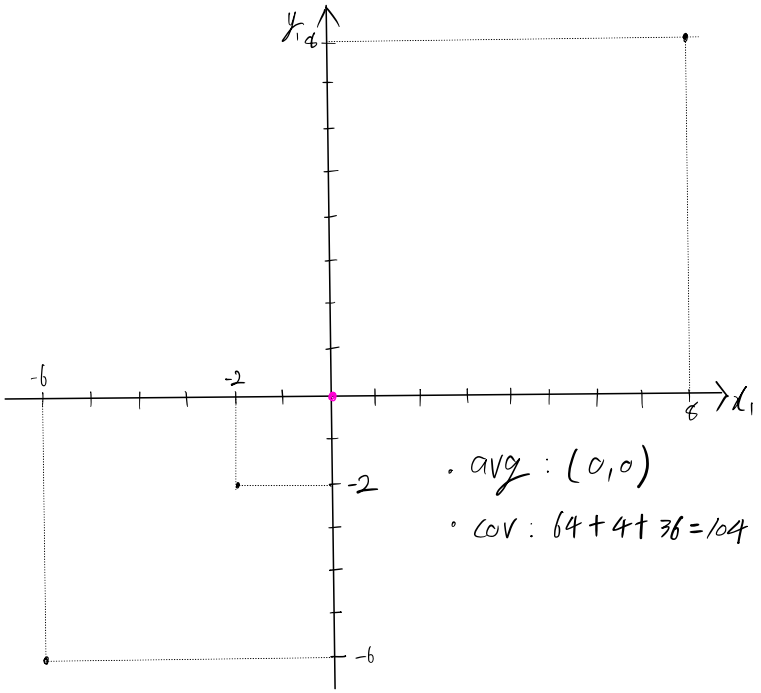
$$COV(X,Y) = \frac{(x_1-\mu_x)(y_1-\mu_y)+(x_2-\mu_x)(y_2-\mu_y)+...+(x_n-\mu_x)(y_n-\mu_y)}{n}$$

분자의 첫 항을 봅시다.

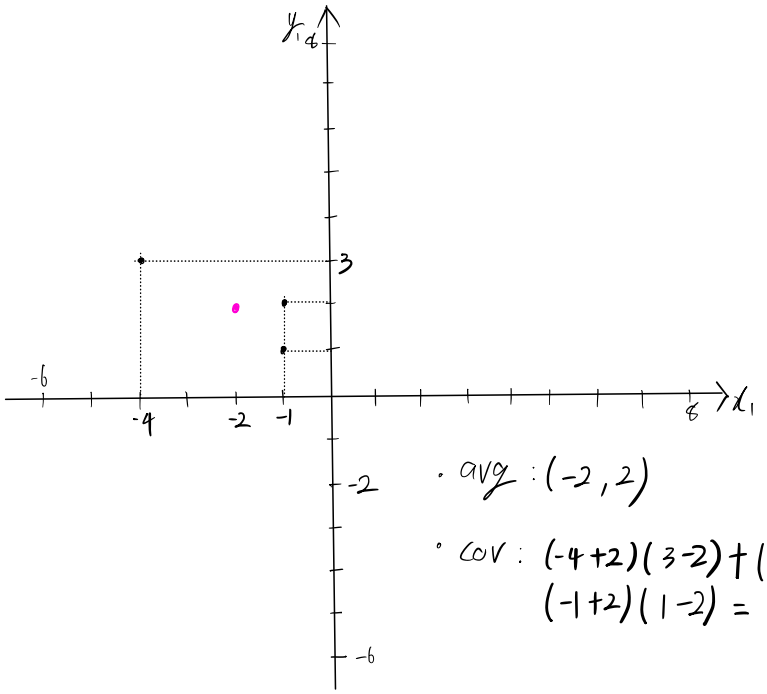
$$(x_1 - \mu_x)(y_1 - \mu_y)$$

위 식의 몇가지 성질을 쉽게 알아낼 수 있습니다.

1) x_1이 x의 평균보다 크고, y_1도 y의 평균보다 크다면 위 값은 양수가 됩니다.
2) x_1이 x의 평균보다 작고, y_1도 y의 평균보다 작다면 위 값은 양수가 됩니다.
3) x_1이 x의 평균보다 크고, y_1도 y의 평균보다 작거나 그 반대의 경우 위 값은 음수가 됩니다.
4) x_1과 y_1이 평균에서 멀 수록 위 값의 절댓값이 커집니다.



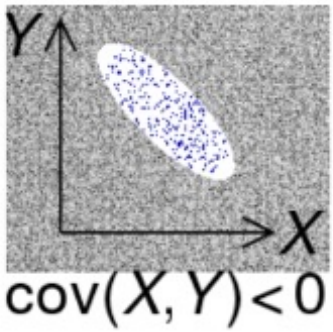
- avg : (0, 0)
- cov : 64 + 4 + 36 = 104

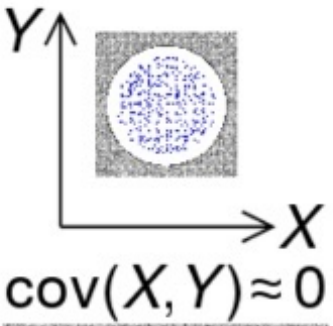


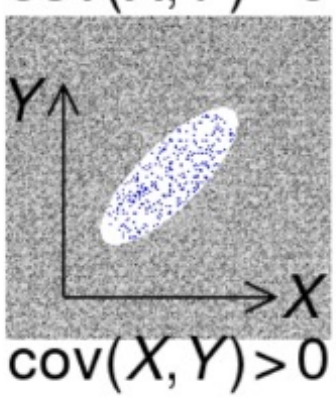
- avg : (-2, 2)
- cov : $(-4+2)(3-2)+(-1+2)(2-2)+$
  $(-1+2)(1-2) = -2+0+-1 = -3$

공분산

공분산(共分散, 영어: covariance)은 2개의 확률변수의 선형 관계를 나타내는 값이다.[1] 만약 2개의 변수중 하나의 값이 상승하는 경향을 보일 때 다른 값도 상승하는 선형 상관성이 있다면 양수의 공분산을 가진다.[2] 반대로 2개의 변수중 하나의 값이 상승하는 경향을 보일 때 다른 값이 하강하는 선형 상관성을 보인다면 공분산의 값은 음수가 된다. 이렇게 공분산은 상관관계의 상승 혹은 하강하는 경향을 이해할 수 있으나 2개 변수의 측정 단위의 크기에 따라 값이 달라지므로 상관분석을 통해 정도를 파악하기에는 부적절하다. 상관분석에서는 상관관계의 정도를 나타내는 단위로 모상관계수로는 그리스 문자 ρ를, 표본상관계수로는 알파벳 s를 사용한다.







두 개의 확률 변수 X 와 Y 의 상관성과 공분산의 부호.

공분산이라 하면 (분산과는 다르게)
하나의 변수가 아닌 두 변수 사이의 관계를 나타낸다고 생각하면 될 것이다.
체고와 도체중의 관계가 궁금하다고 가정하자.
체고가 커질수록 도체중이 증가할까?
물론 당연히 도체중이야 증가한다.
도체수율은 부의상관을 나타내지만...

각설하고, 이 두 변수간의 변동을 공분산 Cov(X, Y) 이라 한다.

공분산 값은 아래와 같이 나타낸다.

Cov(X, Y) > 0 X가 증가 할 때 Y도 증가한다.
Cov(X, Y) < 0 X가 증가 할 때 Y는 감소한다.
Cov(X, Y) = 0 공분산이 0이라면 두 변수간에는 아무런 선형관계가 없으며 두 변수는 서로 독립적인 관계에 있음을 알 수 있다.
그러나 두 변수가 독립적이라면 공분산은 0이 되지만, 공분산이 0이라고 해서 항상 독립적이라고 할 수 없다.

공분산의 개념은 우리가 흔히 사용하는 상관계수와 연관지어 생각해 보아야 한다.
공분산을 구하다 보면,
공분산 값이 항상 일정하지 않기 때문에 비교하고자 한다면 계산도 해야하며 머리가 아파온다.

-000000... <= Cov(X, Y) <= 000000...(ex)

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

그래서 이를 표준화 시켜주는 작업으로 공분산에 표준편차로 나누어 주면
값이 -1 <= Corr(x, y) <= 1 사이 범위로 좁혀지면서 우리는 쉽게 비교할수가 있어진다.
이것이 바로 상관계수 Corr(x, y)인 것이다.

X: 공분산 행렬 에러 매트릭스.

$$cov(X) = \frac{X^T \cdot X}{n}$$

cov(X): 대칭 행렬
n: or 'n-1'

- covariance matrix의 i번째 행과 j번째 열은 i번째 feature와 j번째 feature가 서로 함께 변하는 정도를 의미한다.