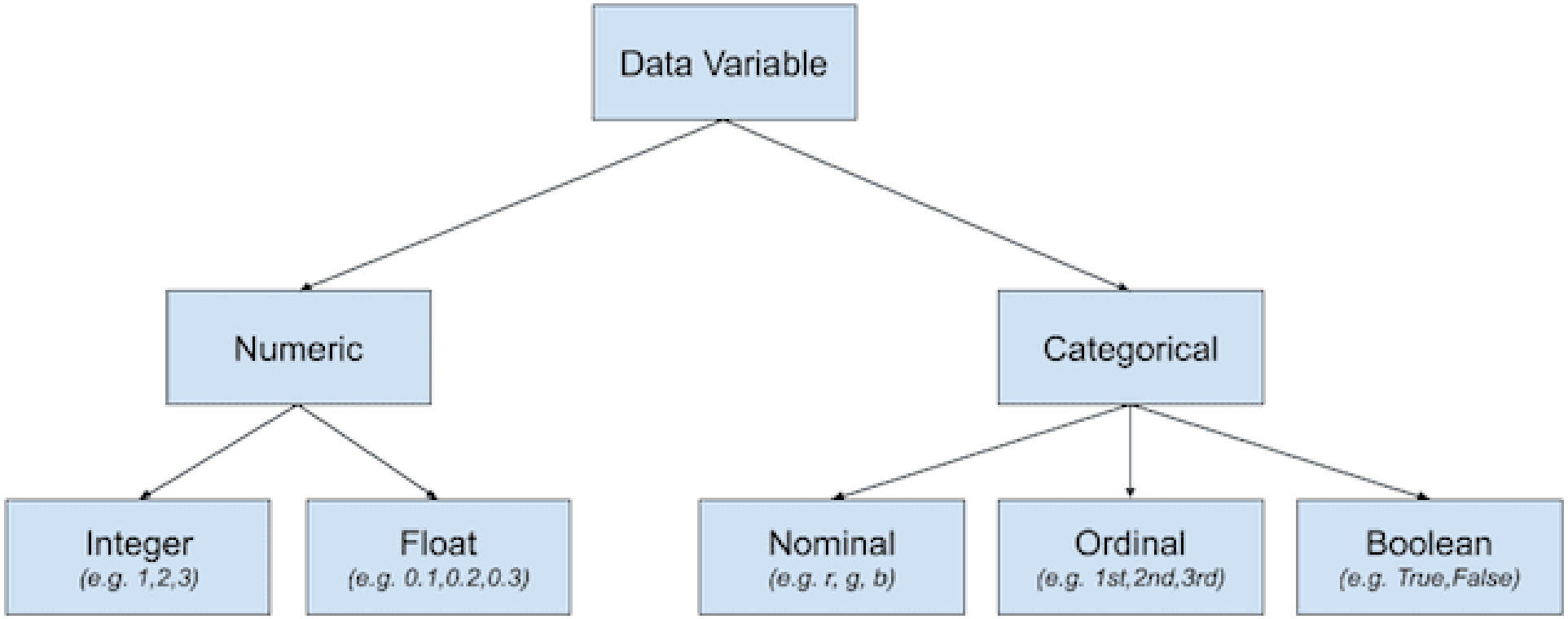
KOSA
Korea Software Industry Association

주성분 분석(PCA)

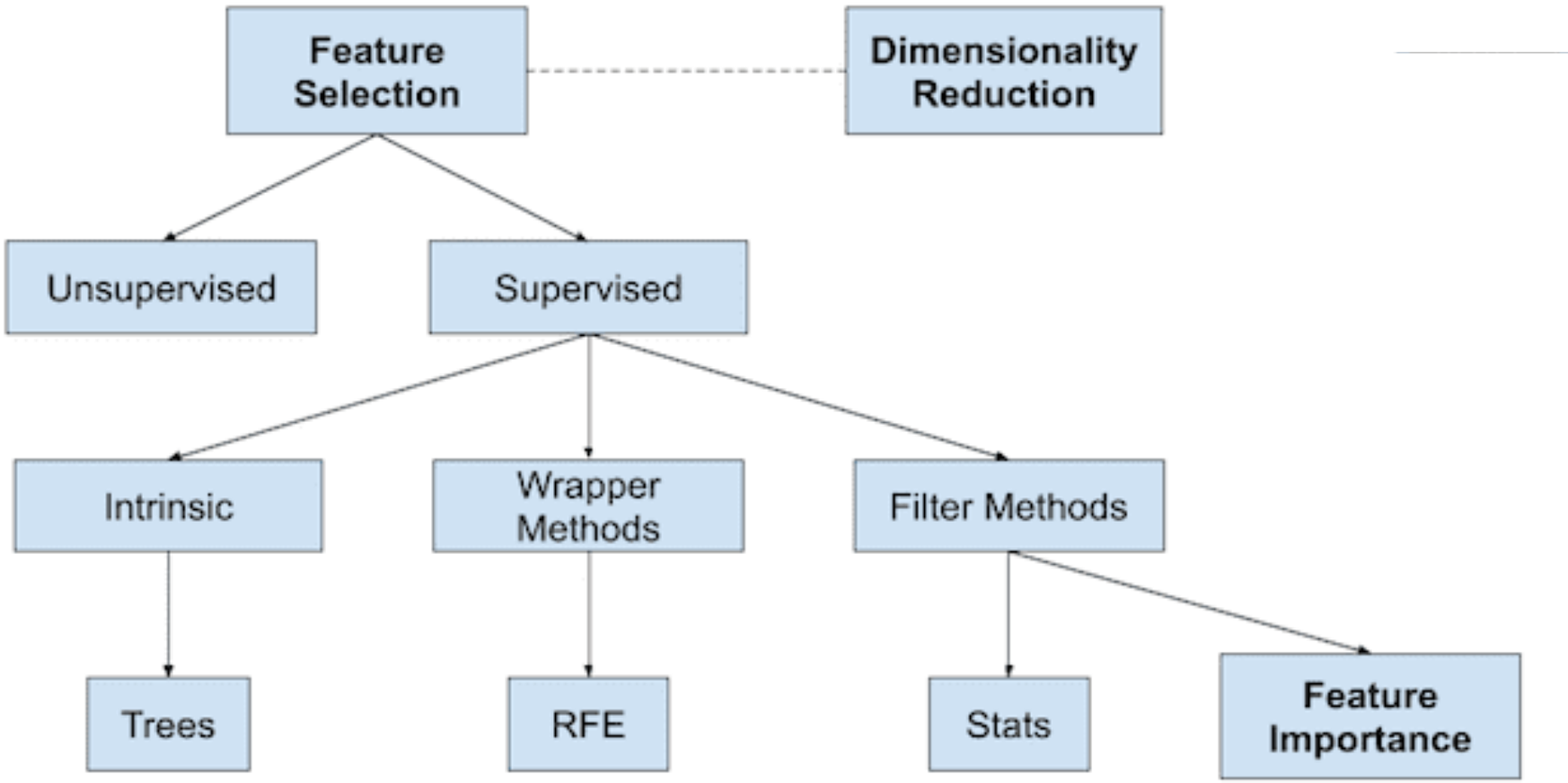
2021. 3. 23

정 준 수 Ph.D

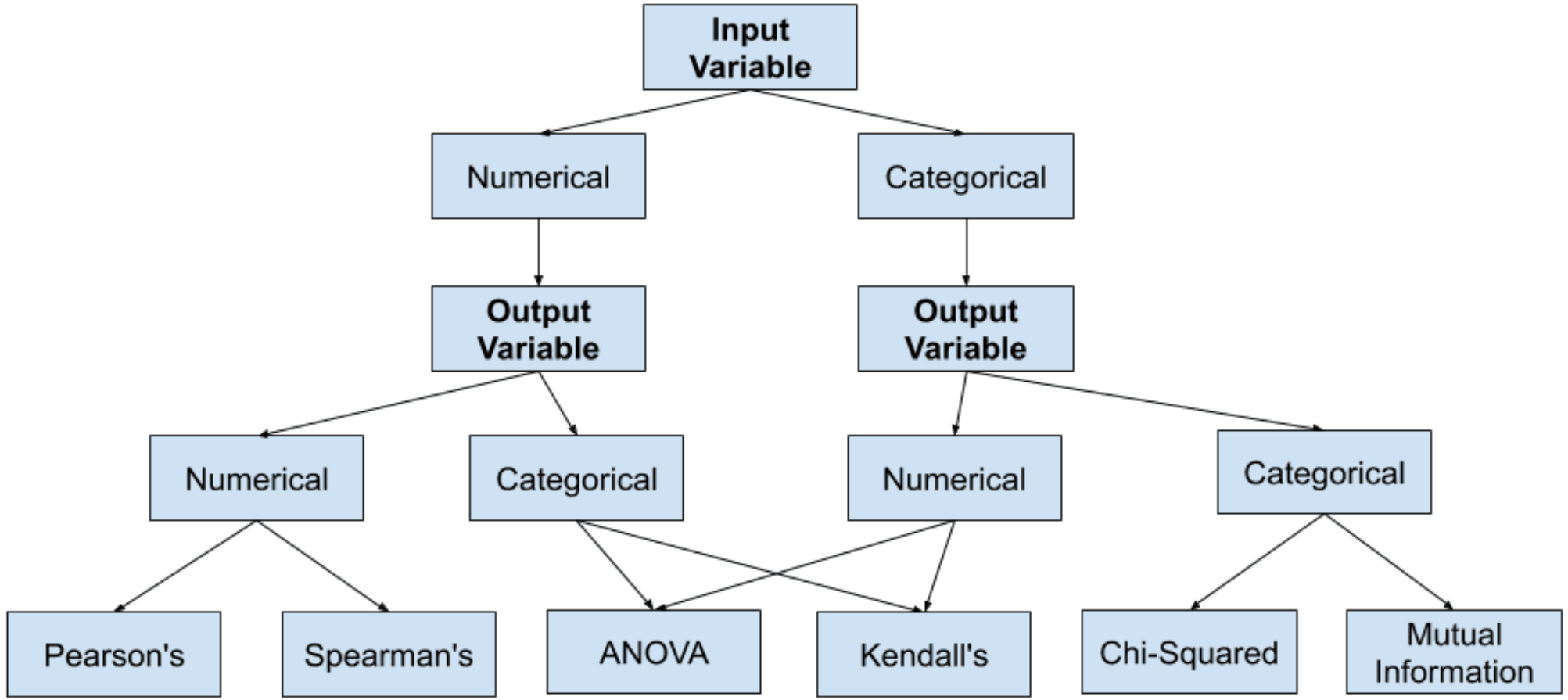
Overview of Data Variable Types



Overview of Feature Selection Techniques



How to Choose a Feature Selection Method



모집단과 샘플(표본)

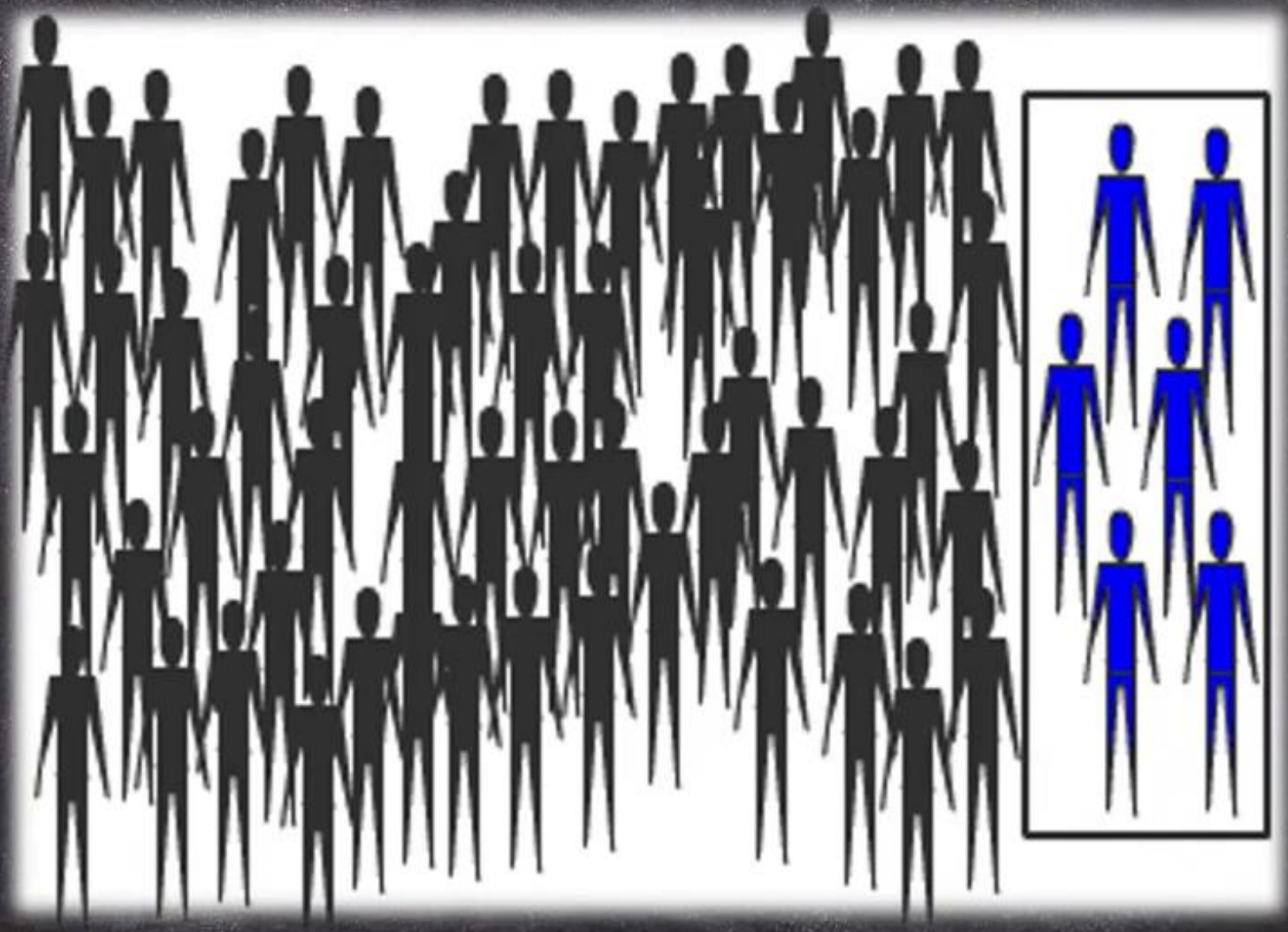
모집단

관 측 치 = N

평 균 값 = μ

분　　산 = σ^2

표준편차 = σ



표본(샘플)

관 측 치 = n

평 균 값 = $\bar{X}$

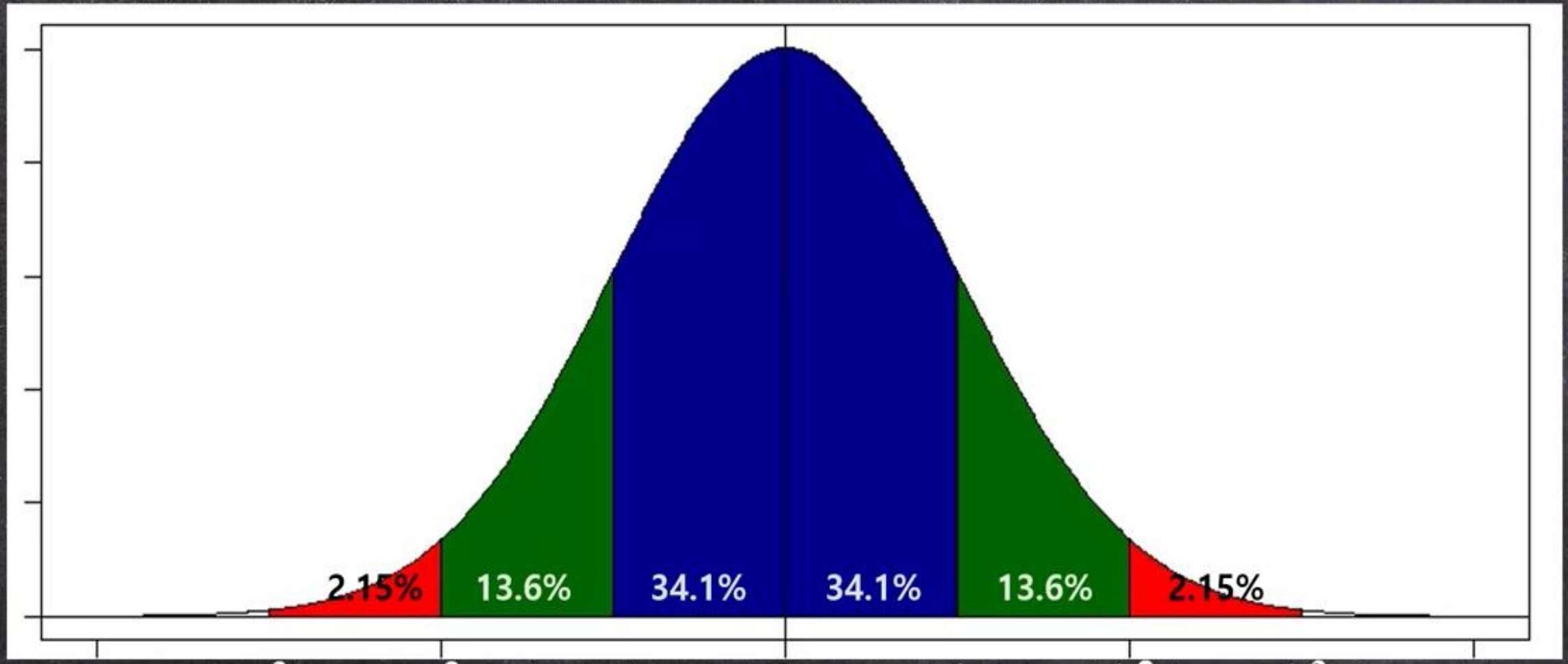
분　　산 = s^2

표준편차 = s

t-test에 대한 보다 깊은 이해

- 그렇습니다. 표준편차는 데이터에 큰 문제가 없는 한은 의미 없는 우연히 퍼져 있는 정도입니다.
- 즉, 우리의 데이터는 평균값 3을 중심으로 랜덤하게 1.58 정도씩 좌우로 퍼져 있는 것입니다.
- 그렇다면, 다시 앞의 1.4cm의 차이로 돌아가 봅시다

정규분포



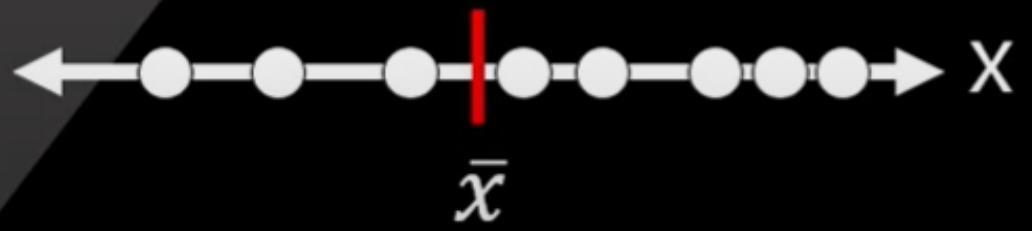
차원축소 (Dimension Reduction)

- 차원축소의 장점
  - 관측이 불가능한 보이지 않는 (Unobservable) 대상을 측정하기 위해 관측이 가능한 보이는 (Observable) 것을 이용
  - 너무 많은 변수를 줄여주어 보다 적은 변수로도 원하는 대상을 측정할 수 있도록 함
- 차원축소의 단점
  - 만능이 아님
  - 설명 불가능한 경우가 많고 연구자의 의도대로 결과를 조절 가능
  - 정보의 손실
  - 직관적인 이해가 쉽지 않음

공분산

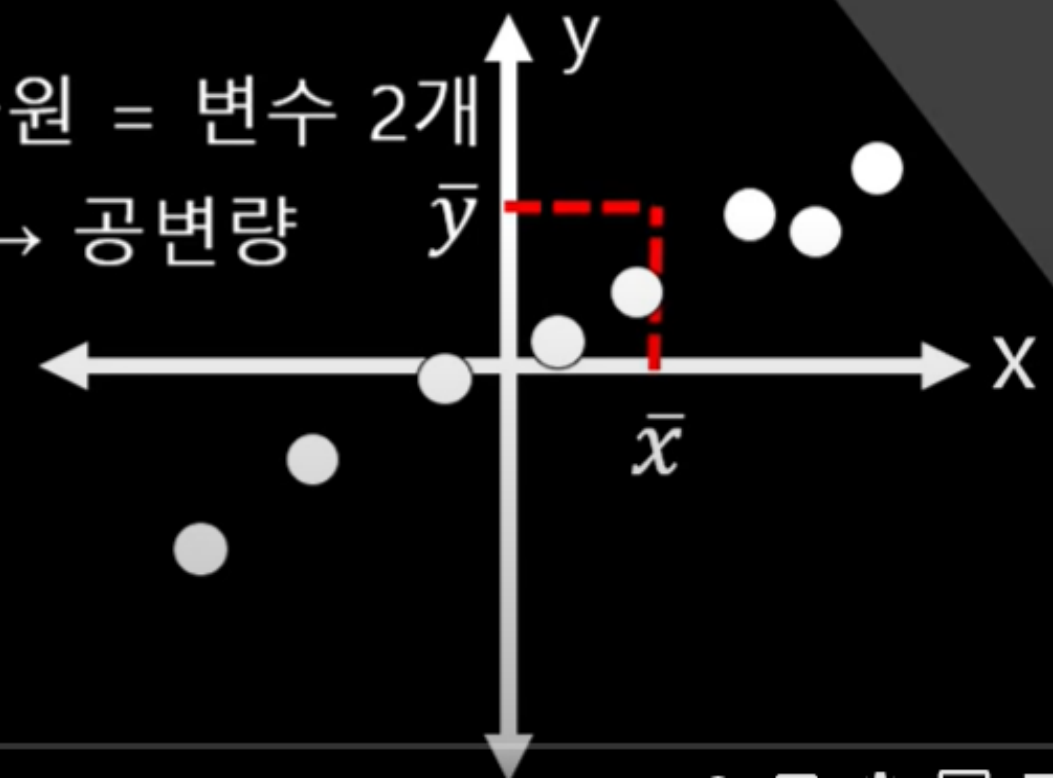
- 그러므로,
  - 분산이 한 변수의 평균값을 중심으로 퍼져 있는 평균적인 거리
  - 공분산이란 두 변수의 평균값을 중심으로 퍼져 있는 평균적인 거리(?)

1차원 = 변수 1개



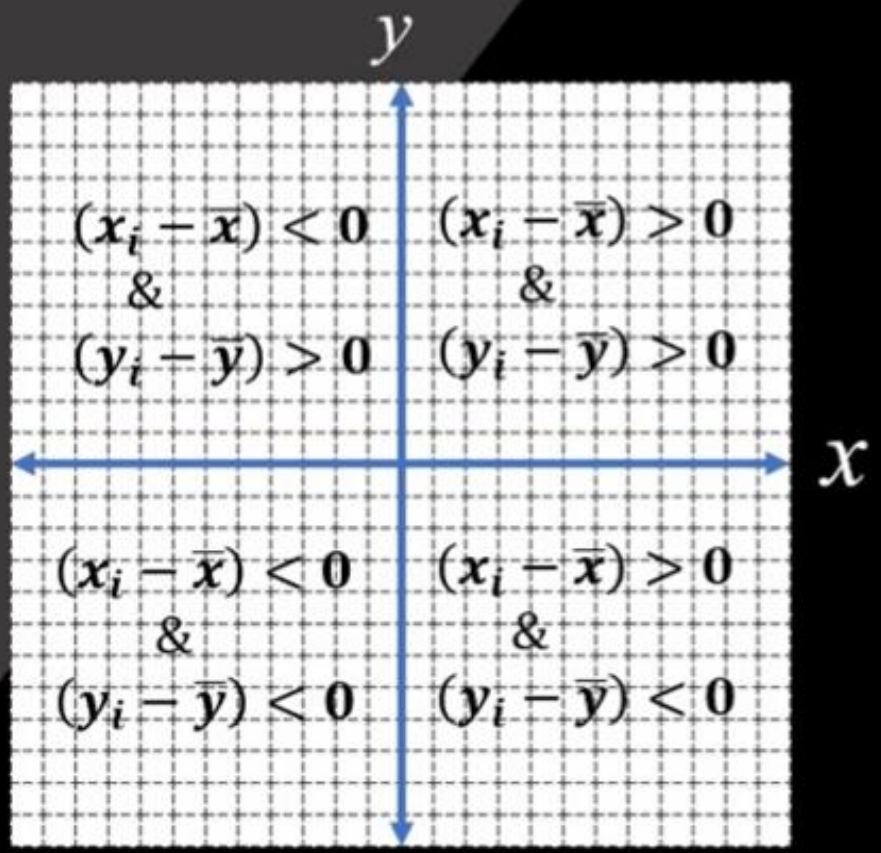
2차원 = 변수 2개

→ 공변량

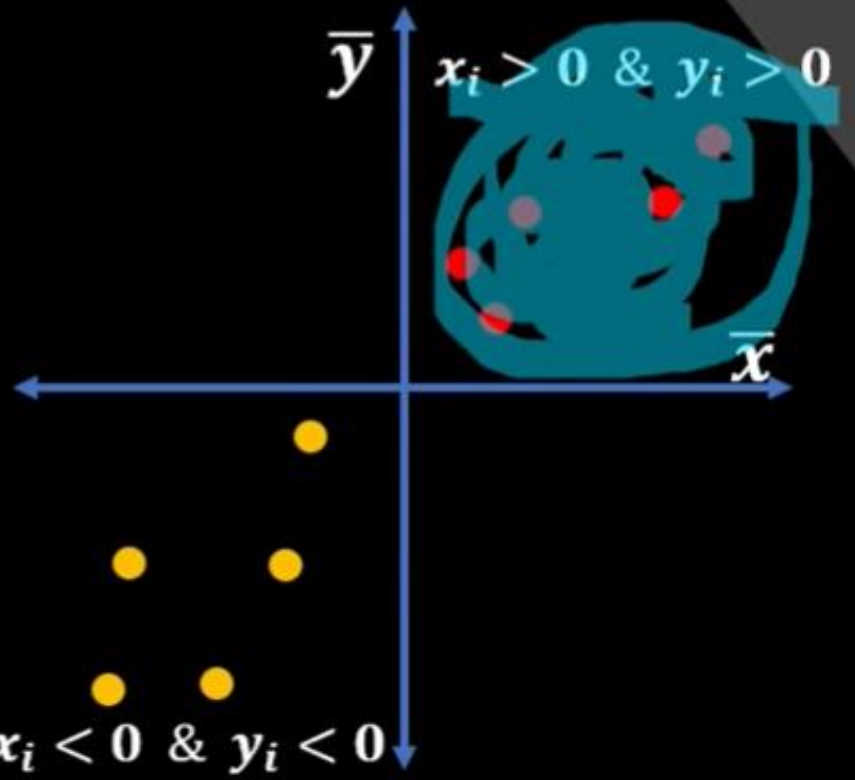


공분산

- 공분산의 종류

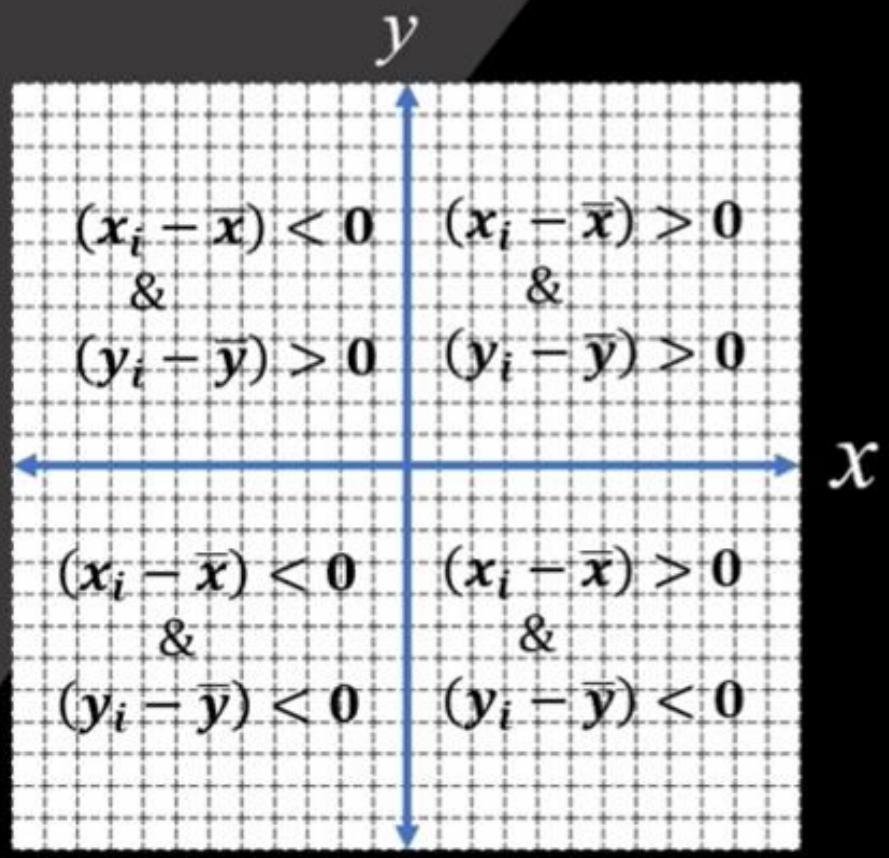


- 쉬운 이해를 위해
  - $\bar{x} = 0$ & $\bar{y} = 0$ 고정해 봅시다

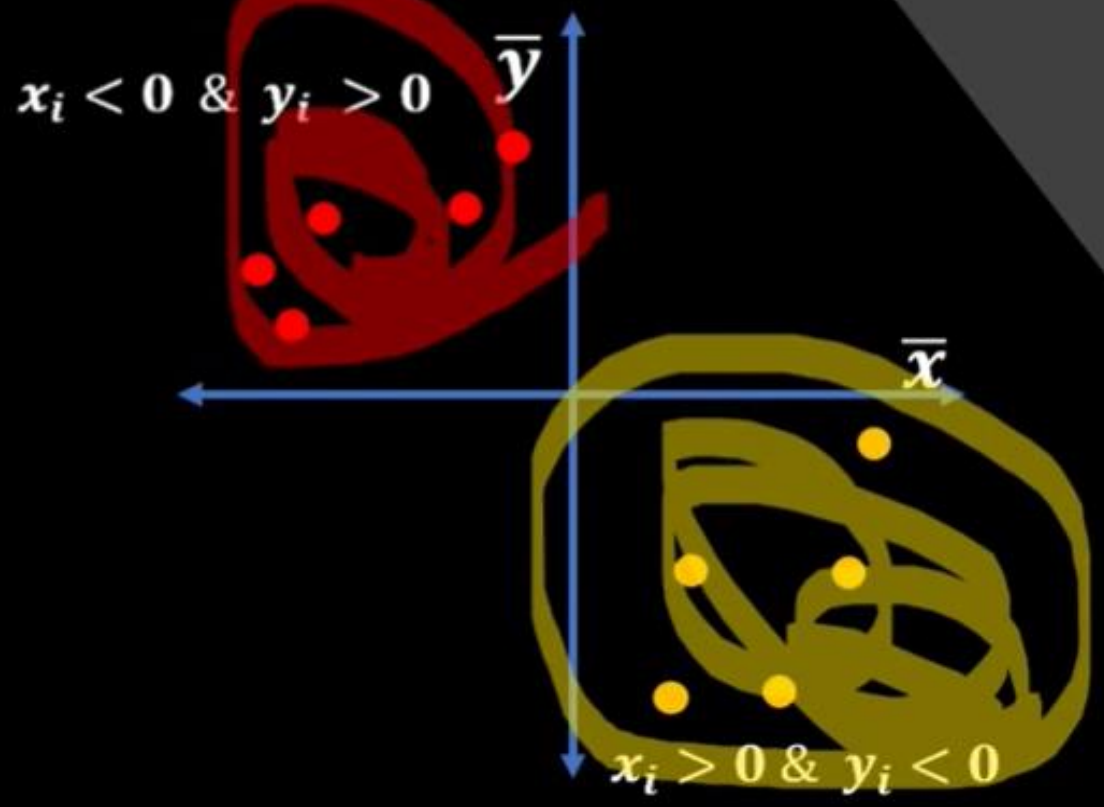


공분산

- 공분산의 종류

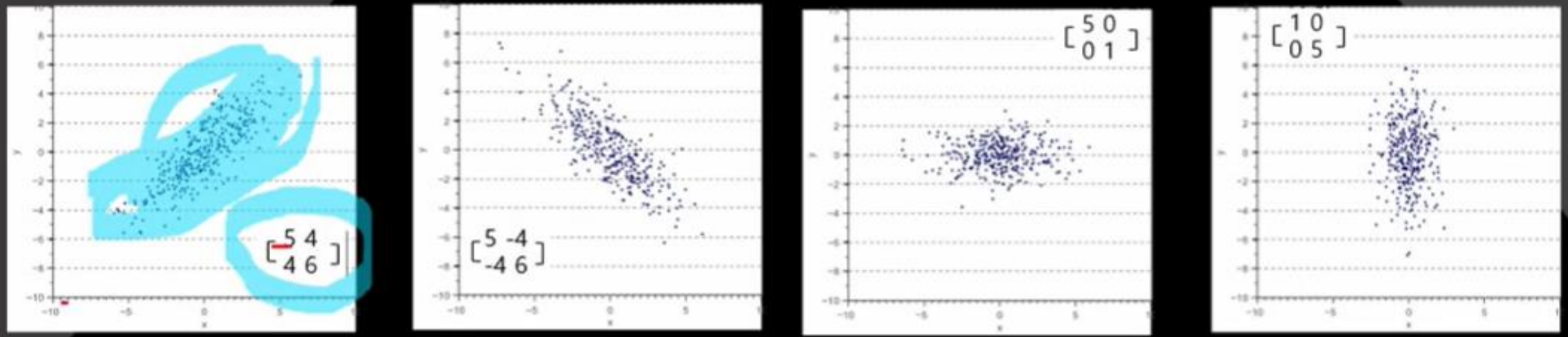


- 쉬운 이해를 위해
  - $\bar{x} = 0$ & $\bar{y} = 0$ 고정해 봅시다



공분산

- 이러한 관계를 그림과 함께 variance covariance matrix로 표현



- 어디서 많이 보지 않았나요?
  - 그렇습니다. 바로 상관관계와 아주 비슷해 보입니다!!!

공분산과 상관계수

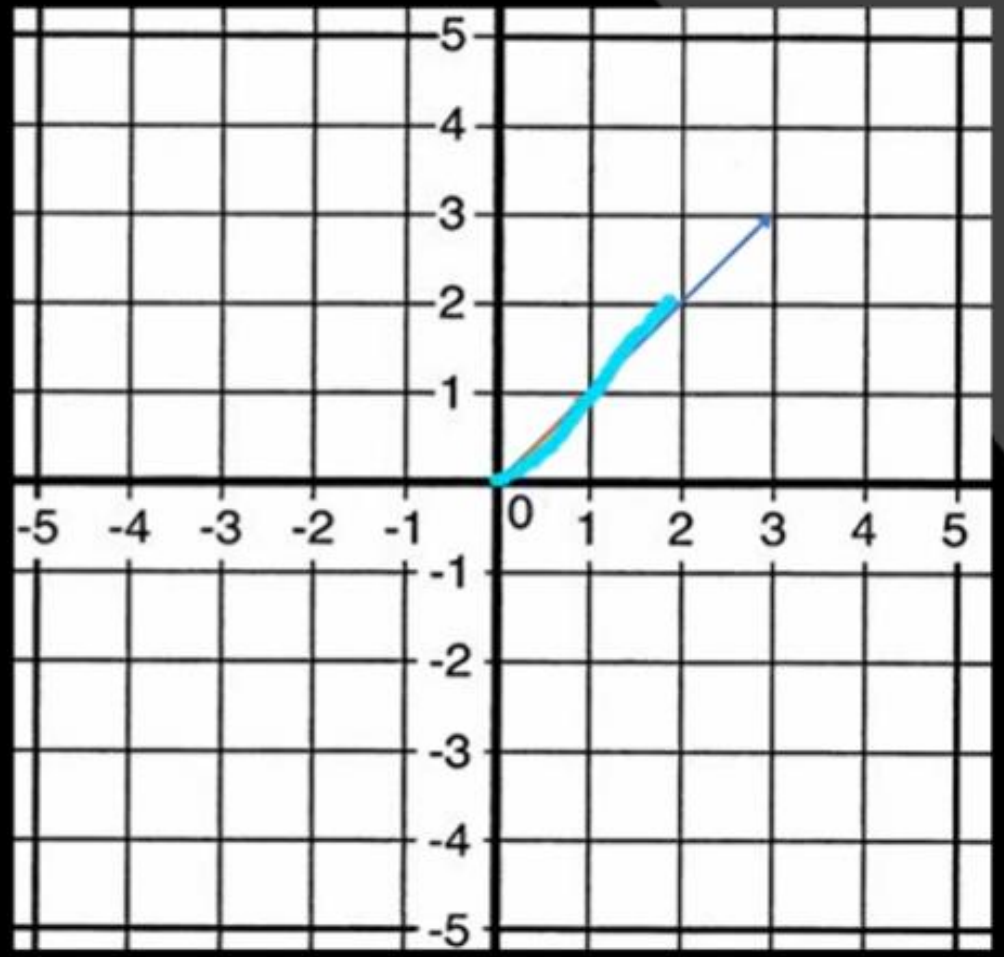
- 공분산
  - $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$
  - 공분산: X의 편차와 Y의 편차를 곱한것의 평균
    - Cov(X, Y) > 0: X ↑ Y ↑
    - Cov(X, Y) < 0: X ↑ Y ↓
    - Cov(X, Y) = 0: No linear relationship
    - 공분산은 두 변수 간에 양의 상관관계가 있는지, 음의 상관관계가 있는지 정도만 알려줌
    - 그러나, 상관관계가 얼마나 큰지는 제대로 반영하지 못함
- 상관계수
  - $\rho = \frac{Cov(x,y)}{\sqrt{Var(x) \times Var(y)}}$
  - 상관계수: 표준화된 공분산 값
    - 절대값은 1 보다 작거나 같음
    - X와 Y가 완벽한 선형 관계라면 상관계수는 1 혹은 -1
    - 상관계수가 1 또는 -1에 근접할수록 힘(power)가 커짐
    - 여기서 힘(power)이란 점들이 모인 정도를 말함

고유값과 고유벡터

- 계산을 해보면...
  - $AX = \begin{bmatrix} 2.0 & 1.0 \\ 1.0 & 2.0 \end{bmatrix} \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}$

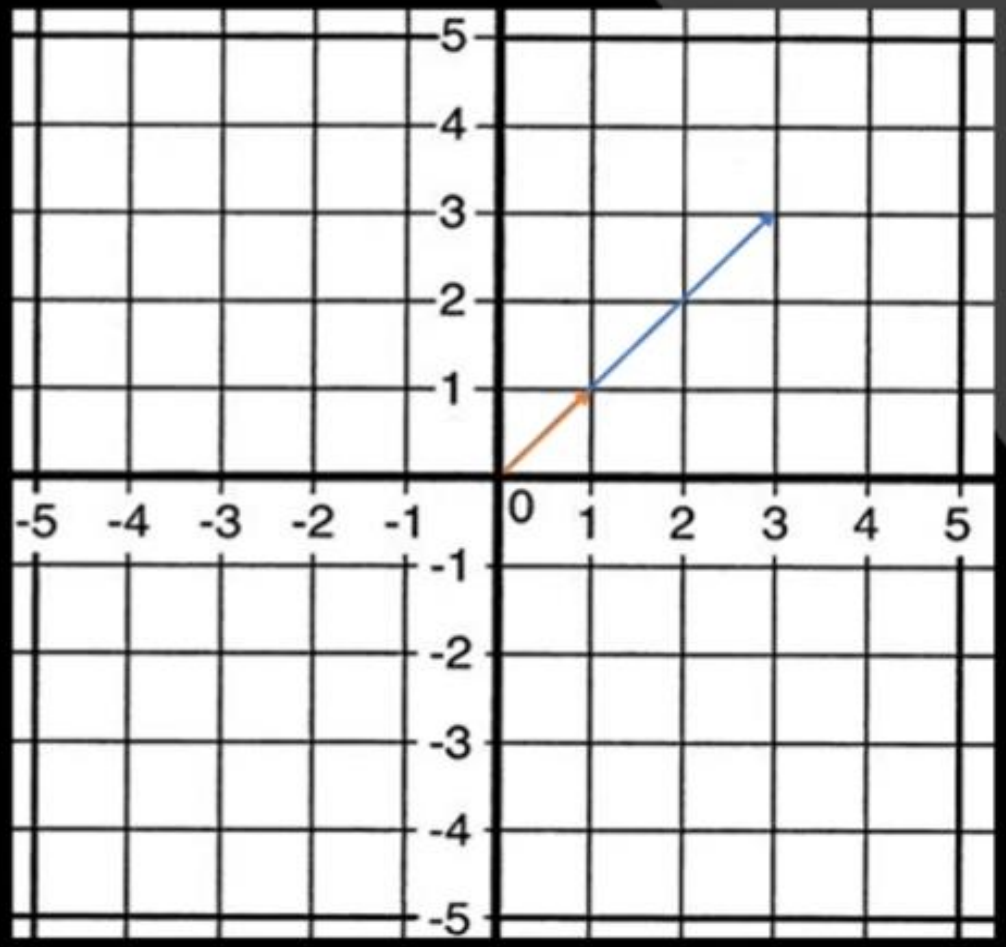
$$= \begin{bmatrix} 2.0+1.0 \\ 1.0+2.0 \end{bmatrix}$$

$$= \begin{bmatrix} 3.0 \\ 3.0 \end{bmatrix}$$



고유값과 고유벡터

- 계산을 해보면...
  - $AX = \begin{bmatrix} 3.0 \\ 3.0 \end{bmatrix}$
- 무슨 뜻일까요?
  - 어느 벡터(X)에 Matrix A를 곱하니 벡터의 방향은 그대로 이고 벡터의 길이만 변하였음
  - 이를 다르게 표현하면
  - $AX = \lambda X$
  - $(A - \lambda I)X = 0$ (I는 Identity Matrix)



고유값과 고유벡터

- 계산을 해보면...
  - $AX = \begin{bmatrix} 3.0 \\ \\ 3.0 \end{bmatrix}$
- 무슨 뜻일까요?
  - 어느 벡터(X)에 Matrix A를 곱하니 벡터의 방향은 그대로 이고 벡터의 길이만 변하였음
  - 이를 다르게 표현하면
  - $AX = \lambda X$
  - $(A - \lambda I)X = 0$ (I는 Identity Matrix)

- 이 식이 성립하려면
  - $det\,(A - \lambda I) = 0$
- 이게 무슨 뜻일까요?
  - 예를 들어 봅시다

고유값과 고유벡터

- 고유값은 3과 1이므로 이를 이용해 고유벡터를 구하면...
  - $AX = \lambda X$ 이므로, $\lambda = 3$인 경우
  - $\begin{bmatrix} 2.0 & 1.0 \\ 1.0 & 2.0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = 3 \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$
  - $2X_1 + X_2 = 3X_1$
  - $X_1 + 2X_2 = 3X_2$
  - $\therefore X_1 = X_2$
  - $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

고유값과 고유벡터

- 그래서??? 생각을 바꿔보자!!!
- 우리가 가진 데이터를 variance-covariance Matrix로 바꾼다면...
  - 그래서 그 분산-공분산 행렬이 다음과 같다면
  - $\Sigma = \begin{bmatrix} 2.0 & 1.0 \\ 1.0 & 2.0 \end{bmatrix}$
  - 이 Matrix의 고유값은 1과 3인데,
  - 또한, 고유값의 합 (1+3=4)은 두 분산의 합과 같음 (2+2=4)
  - 또한, 고유값의 곱 (1×3=3)은 $det(\Sigma) = (2\times2 - 1\times1) = 3$으로 동일

주성분 분석

- 주성분 분석의 특징
  - Unique variance가 없다
  - Total variance = Common variance

- 주성분 분석의 목적은 무엇인가?
  - 고차원의 데이터를 저차원으로 줄이는 (환원시키는) 것
  - 공통된 (상관관계가 높은) 변수들을 줄여서 주성분을 찾는 것
    - 사용된 변수의 개수 > 주성분의 개수

주 성분분석

- 주성분을 뽑아 낼 때의 원칙
  - 분산이 가장 커지는 축을 첫 번째 주성분으로 하고
  - 분산이 두번째로 커지는 축을 두 번째 주성분으로 하며
  - 이런 방식으로 주성분을 뽑아 냄
    - 1st principal component's variance > 2nd PC's variance > 3rd PC's variance > ...
  - 각 주성분 (PC)은 서로간 직교 (90도) 함
    - 공분산 행렬의 고유벡터 이므로
    - 뒤에 예를 보면서 하나씩 살펴볼 것임
  - 주성분 분석은 가장 큰 분산을 갖는 부분공간을 보존하는 최적의 선형변환으로 주성분 몇 개로 전체 분산을 설명하려는 시도

주성분 분석의 예

- 우선, 표준화를 하자!!
  - 표준화 = z-score로 변환
  - 장점
    - 모든 변수의 평균은 0, 분산은 1로 동일해 짐
    - 변수의 스케일의 차이로 인한 분산의 왜곡 방지
    - 모든 변수의 분산을 동일하게 맞춤
    - 변수의 개수 = 총분산
      - 예) 총분산 = 2
  - 방법
    - z-score = $\frac{X-\mu}{\sigma}$

ID	국어	영어
1	1.1	1.0
2	0.9	1.0
3	0.5	0.3
4	-0.2	0.2
5	-0.9	-1.0
6	0.5	0.5
7	1.1	0.8
8	-1.2	-1.3
9	-1.7	-1.8
10	-0.2	0.2
평균	**0**	**0**
분산	**1**	**1**