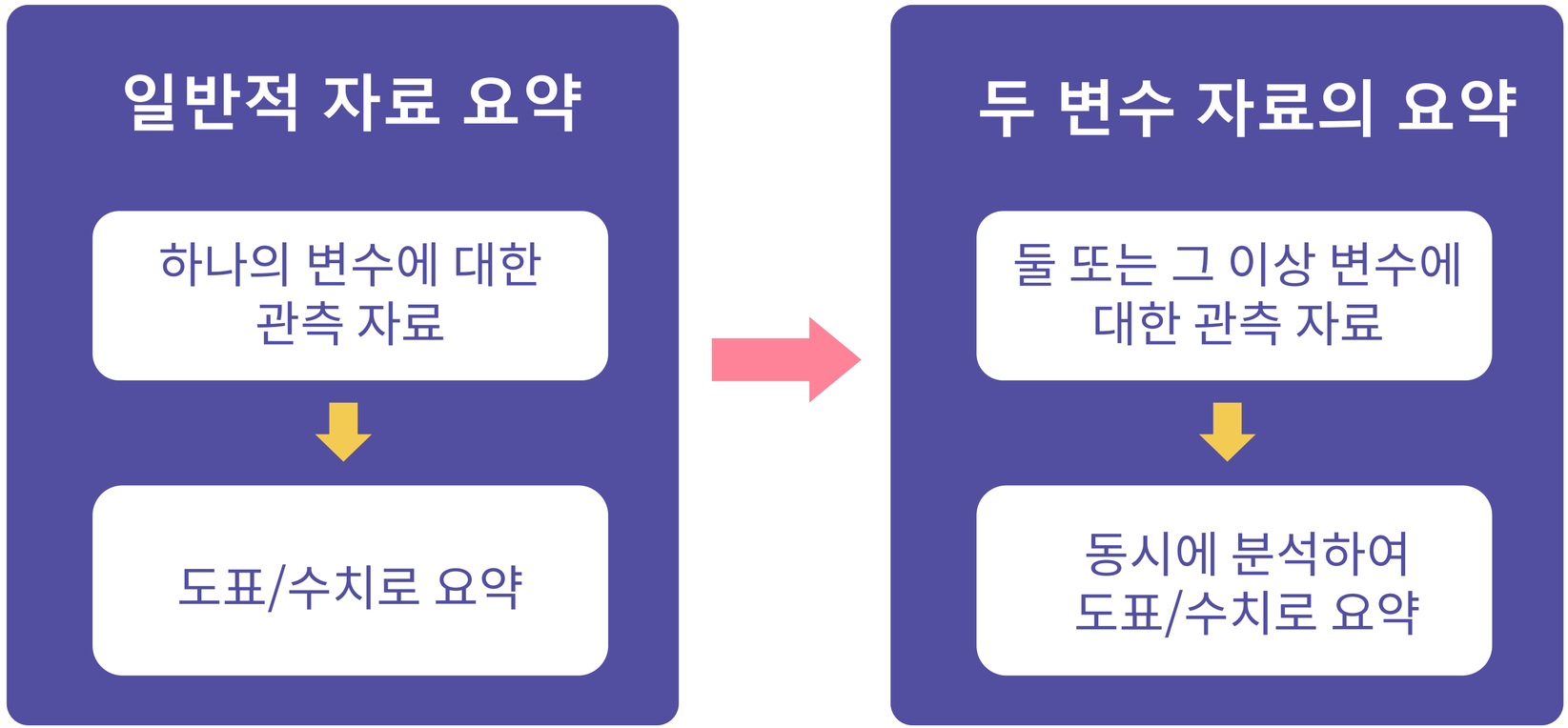
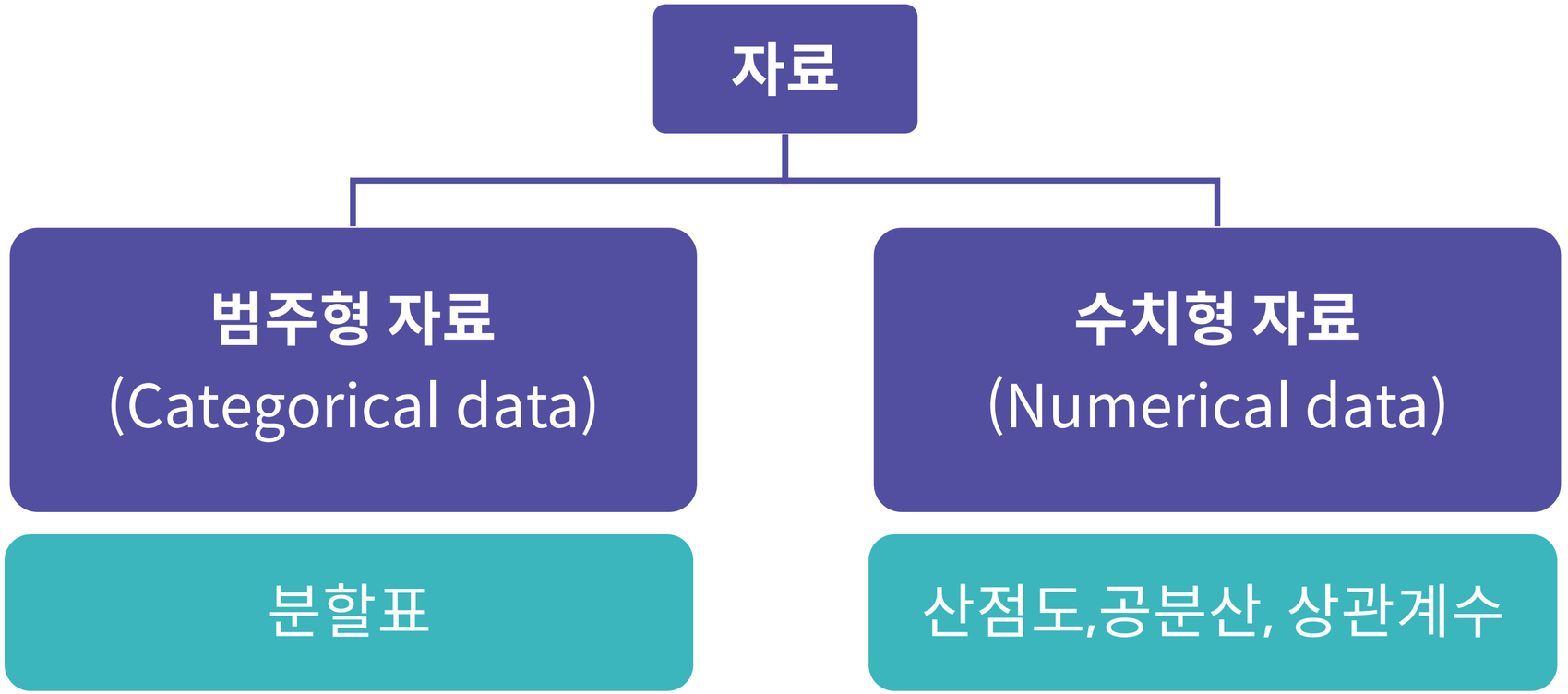
4. 두 변수 자료의 요약

두 변수 자료의 요약


일반적 자료 요약
하나의 변수에 대한 관측 자료
도표/수치로 요약
두 변수 자료의 요약
둘 또는 그 이상 변수에 대한 관측 자료
동시에 분석하여 도표/수치로 요약

두 변수 자료의 요약


자료
범주형 자료
(Categorical data)
수치형 자료
(Numerical data)
분할표
산점도,공분산, 상관계수

분할표

범주형

변수 1 변수 2

▶ 도수분포표를 2차원으로 확장한 형태로 요약

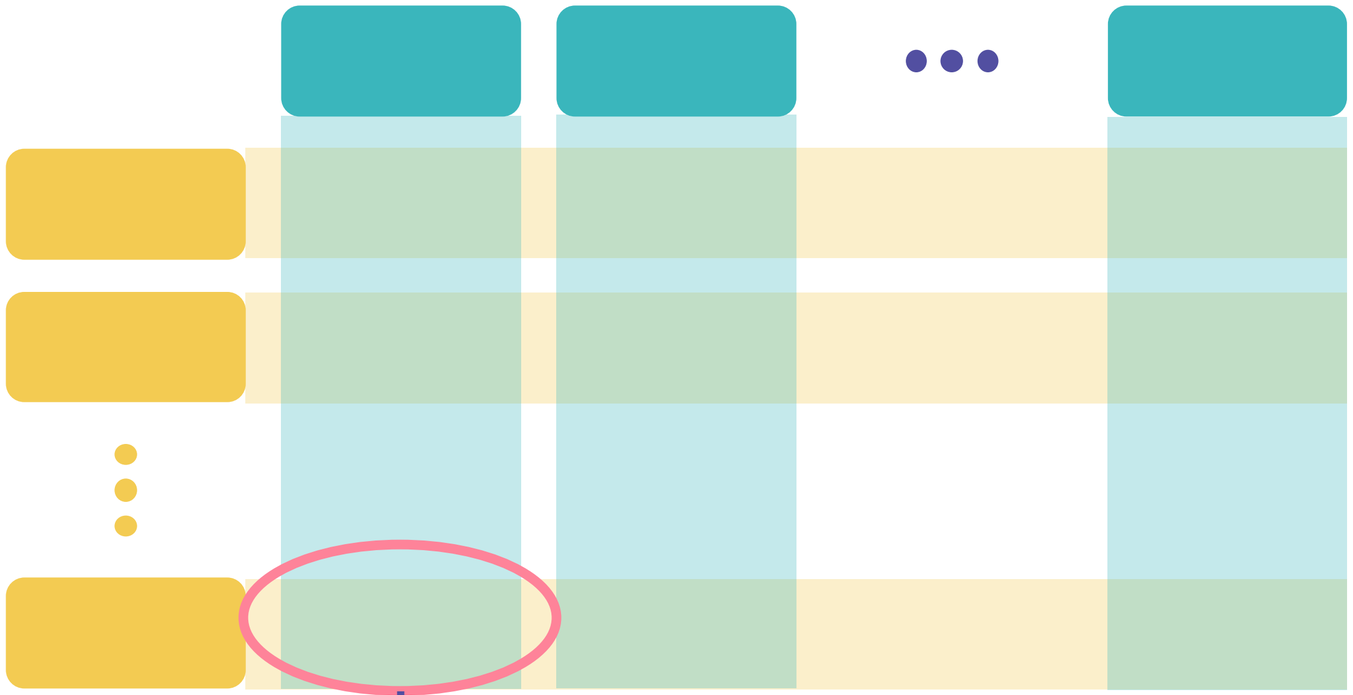
	변수 2의 범주 1	변수 2의 범주 2	변수 2의 범주 3
변수 1의 범주 1			
변수 1의 범주 2			
변수 1의 범주 3			

두 변수의 범주들이 동시에 갖는 관측값의 수

(ex : **변수 1의 범주 2**와 **변수 2의 범주 3** 사이의 상관관계)

분할표 `pd.crosstab(변수1, 변수2)`

두 범주형 변수의 요약 : 분할표



교차하는 부분에 여러 가지 값 표시 가능

예) 상대도수 -> 두 변수 사이 관련 분포 상태를 명확히 표현

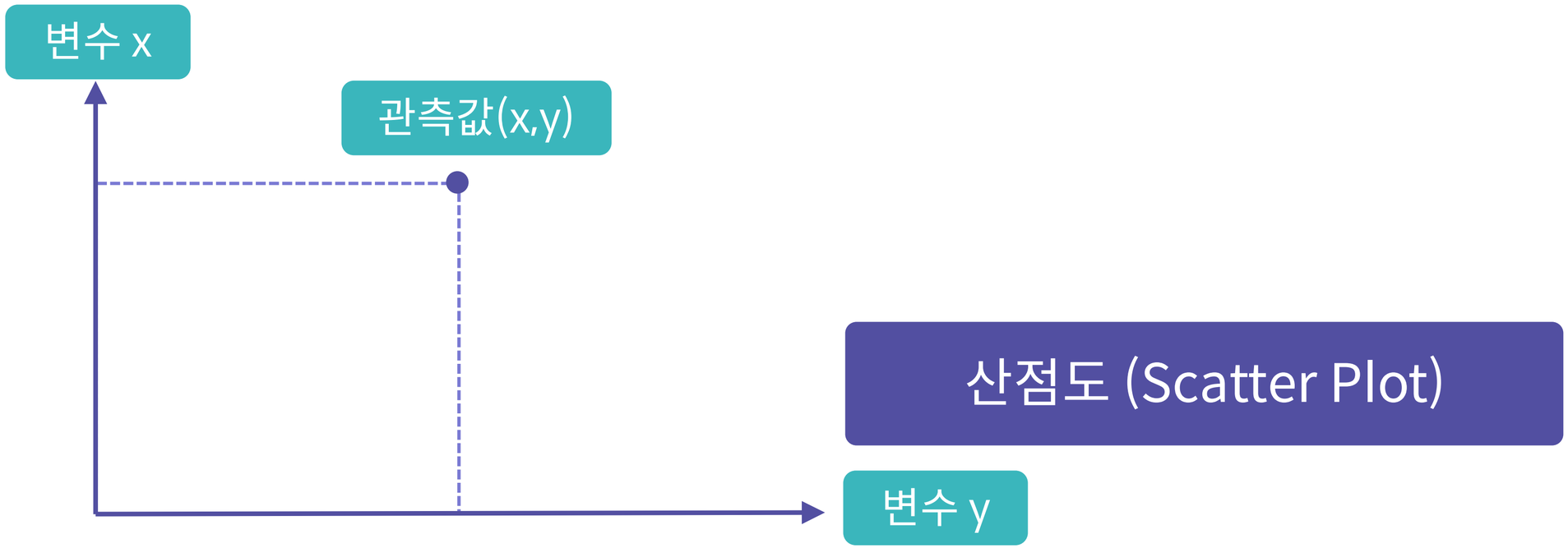
그림을 통한 두 연속형 변수의 요약 : 산점도

연속형

변수 x

변수 y

각 변수의 관측값을 좌표평면에 표시함



그림을 통한 두 연속형 변수의 요약 : 산점도

```
plt.scatter(변수1, 변수2)
```

두 변수 사이의 관계를 시각적으로 파악

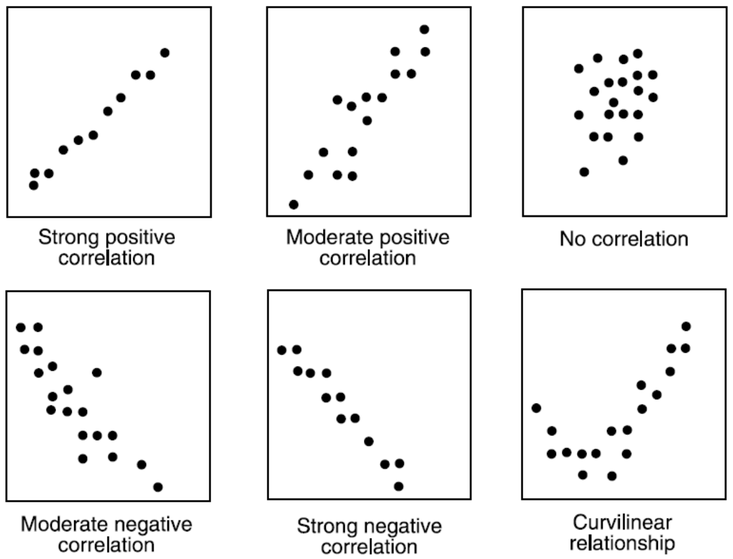
관측값이 많은 경우 점들이 띠를 형성

그림을 통한 두 연속형 변수의 요약 : 산점도

변수 x

변수 y

변수 x, 변수 y 각각에 대해 관심이 있다면
앞서 배운 기법들로 분석 가능



변수 x

변수 y

두 변수가 서로 어떤 관계인지
확인하기 위해 산점도를 사용

산점도 위의 점들의 경향

▼

곡선 등 여러 가지 형태가 가능

공분산

변수가 포함된 자료.cov()

두 변수 (x, y) 에 대하여 서로 어떤 관계를 가지는지 나타냄

- x 값과 y값이 같은 방향으로 변화할 때, 공분산 값은 양수
- x 값과 y값이 반대 방향으로 변화할 때, 공분산 값은 음수

Cov(x,y)로 표현

공분산

$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$ 로 계산

- 여기서 $\bar{x}, \bar{y}$는 평균값, x_i, y_i는 각각의 관측값

두 변수의 편차를 곱하여 더한 후자료의 개수(N) 으로 나누어줌

자료가 평균값으로부터 얼마나 멀리 떨어져 있는지 나타냄

상관계수

연속형

변수 x　변수 y

산점도의 점들이 직선에 가까운
정도를 수치로 나타내어 관계를 파악

◀ 피어슨

- 피어슨에 의해 제안되었기 때문에
  피어슨의 상관계수라고도 불림
- 상관계수는 보통 r 로 표시

상관계수

두 변수 (x, y) 에 대하여 관측값 n 개의 짝
$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$이 주어질 때 다음과 같이 계산

$$\text{상관계수 } r = \frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}}$$

$$\text{단, } \bar{x} = \frac{1}{n}\sum x_i, \bar{y} = \frac{1}{n}\sum y_i$$

$$S_{xx} = \Sigma(x_i - \bar{x})^2$$

$$S_{yy} = \Sigma(y_i - \bar{y})^2$$

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

상관계수

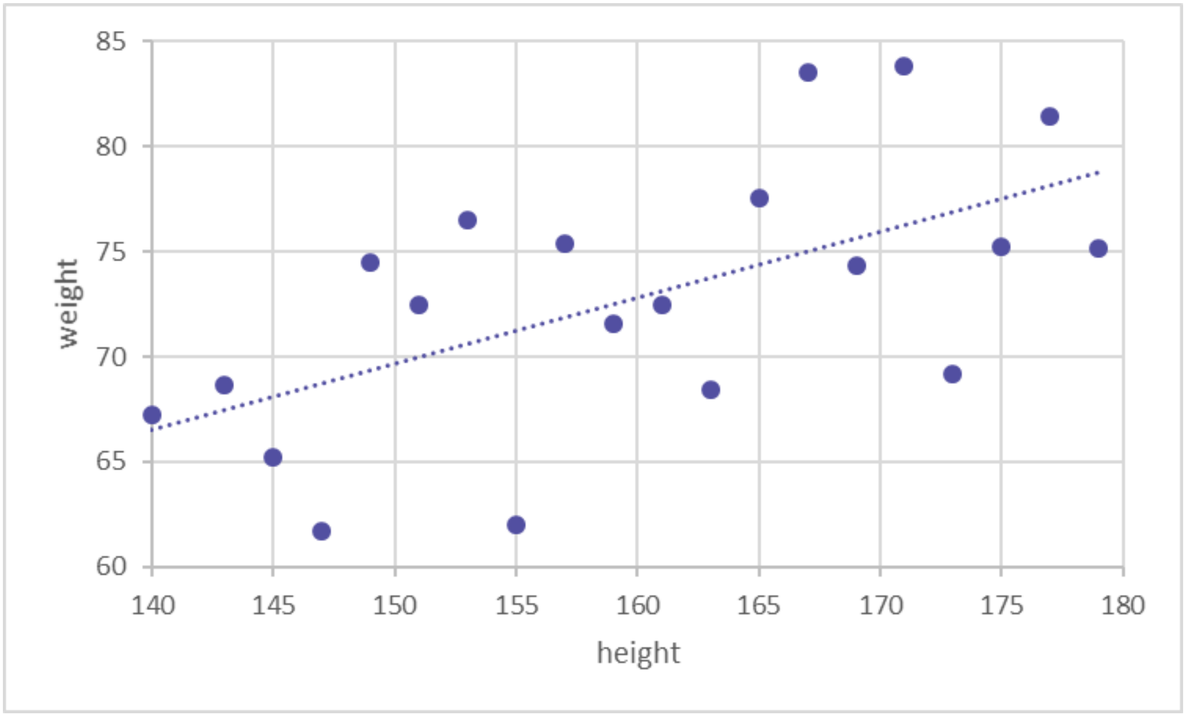
```
변수가 포함된 자료.corr()
```

표본상관계수 r은 항상 -1과 1사이에 있음

절댓값의 크기는 직선관계에 가까운 정도를 나타냄

부호는 직선관계의 방향을 나타냄

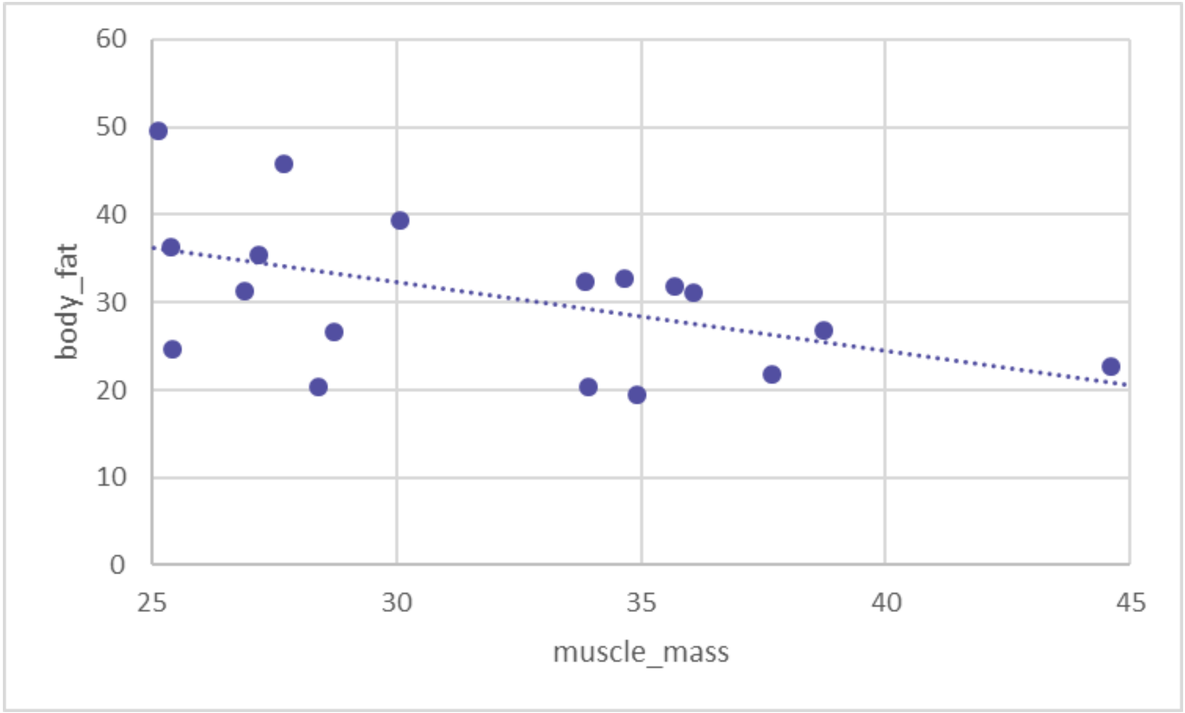
상관계수



r > 0

- 점들이 좌하에서 우상방향으로 띠를 형성
- 두 변수의 값이 **비례** 관계를 나타냄
- 이 경향 직선의 기울기는 **양수**

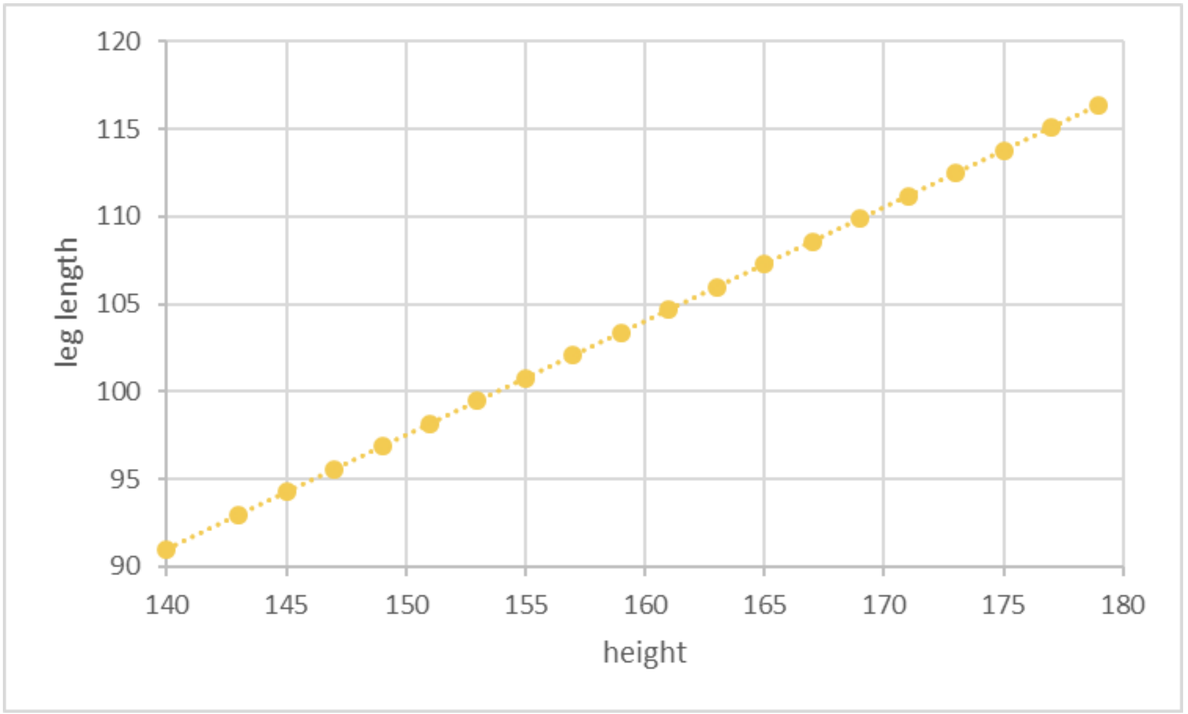
상관계수



r < 0

- 점들이 좌상에서 우하방향으로 띠를 형성
- 두 변수의 값이 **반비례** 관계를 나타냄
- 이 경향 직선의 기울기는 **음수**

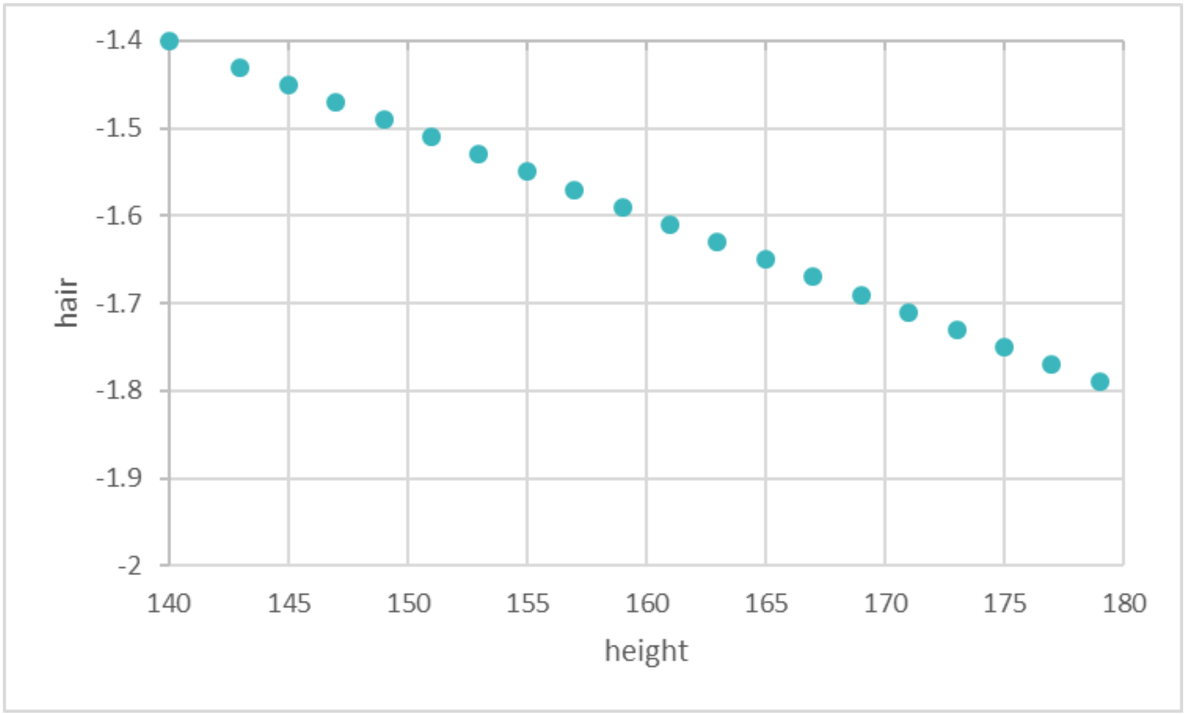
상관계수



r = +1

- 모든 점이 정확히 기울기가 양수인 직선에 위치

상관계수



r = -1

- 모든 점이 정확히 기울기가 음수인 직선에 위치

상관계수의 특징

상관계수는 단위가 없음

- 변수 x,y의 단위는 분모, 분자에서 상쇄
- 이를 이용하여 단위가 다른 변수에서 직선관계 정도를 비교가능

상관계수만으로 판단 시, 잘못된 해석 가능성

- 상관계수는 직선 관계 나타내므로 직선이 아닐 때 부적합
- 상관계수를 구하기 전 산점도를 보고 전체의 경향을 파악한 후 상관계수 계산

상관계수와 인과관계

인과관계

x가 y의 원인이 되고 있다고 믿어지는 관계

자료분석 시, 주의해야할 점

큰 상관계수값이 항상 두 변수 사이의
어떠한 인과관계를 의미하지 않는다는 사실!

상관계수와 인과관계

상어에 물린 사고 횟수가 늘어날 때

아이스크림 판매량도 같이 늘어난다

→ 상어에 물린 사고 횟수와
아이스크림 판매량은 상관 관계가 있다

→ 상어에 많이 물릴 수록 **아이스크림이 많이 팔린다**?

상관계수와 인과관계

상어 사고가 많다 → 해수욕이 많은 **여름철**이기 때문

아이스크림이 많이 팔린다 → 더운 **여름철**이기 때문

직접적인 인과관계는 상어와 아이스크림이 아니라,

여름와 상어, **여름**과 아이스크림에 있다

상관계수와 인과관계

상관 계수가 높다 ≠ 인과관계이다