ch3.통계

통계학

- 데이터 속의 차이와 패턴을 설명하여 이를 바탕으로 미래를 예측하는 학문

차이를 확인하는 데이터 요약

데이터 요약

- 데이터를 요약하면 그속에서 차이를 확인할 수 있다.
  예) 평균을 확인하면 데이터는 평균보다 큰 쪽과 평균보다 작은 쪽으로 나뉜다.

- 모델링 전에 먼저 데이터 요약을 통해 데이터의 특징을 살피고 어떤 차이가 있는지 살펴보면,
  분석의 방향을 설정하는데 도움이 된다.

변수와 관측치

- 데이터는 수많은 변수와 관측치로 이루어져 있다.
  - 변수(variable) : 열
  - 관측치(observation) : 행

- 데이터는 공간으로 표현할 수 있다.

- 데이터 분석 : 변수들이 만들어내는 공간의 특징을 설명하고 그 속에 점처럼 흩어져 있는 관측치의 패턴을 찾는 과정.

기호의 약속

- 변수의 개수 p, 관측치의 개수 n

  - 변수의 개수 : p -> 차원
  - 관측치의 개수 : n -> 점
  - 데이터의 크기 : n x p -> p차원 속에 n개의 점이 있다.

- 변수 : x, y, z

- 관측치 : (아래 첨자 알파벳) xn
  - x1 : x의 첫번째 관측치
  - xn : x의 n번째 관측치

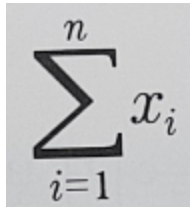
- 관측치들이 모여서 변수가 된다.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}$$

합계

합계를 뜻하는 “Summation”의 첫글자 S에 해당하는 그리스어 대문자 **Σ**로 표현

- 데이터분석을 위해 합계(SIGMA, 시그마)가 많이 등장



“x_i들(i는 1부터 n까지)을 다 더해라”

기술통계량과 변수

- 통계량(statistic) : 변수의 특징을 설명하기 위해 계산된 숫자들.

- 기술 통계량(Descriptive statistics) : 데이터의 특징을 설명하는 통계량.
  - 평균, 분산, 최솟값, 중앙값, 최댓값

- 예시) 100명의 몸무게 데이터 -> 100개의 숫자를 전부 살펴보는 것이 아니라,
  - 100개의 숫자를 모두 더하고, 100으로 나눈 값으로 몸무게의 특징을 살펴봄 -> 평균

- 변수의 종류
  1. 연속형 변수(continuous) : 몸무게, 키 -> 순서를 정하거나 값들을 더해서 통계량을 계산한다.
  2. 범주형 변수(categorical) : 성별, 나라 -> 묶어서 개수를 센다.

연속형 변수

- 정렬 : 크기에 따라 순서대로 줄 세우는 과정
  예) 0 1 2 3 4 5 6 7 8 9

- 순서 통계량 : 오름차순으로 정렬된 값들

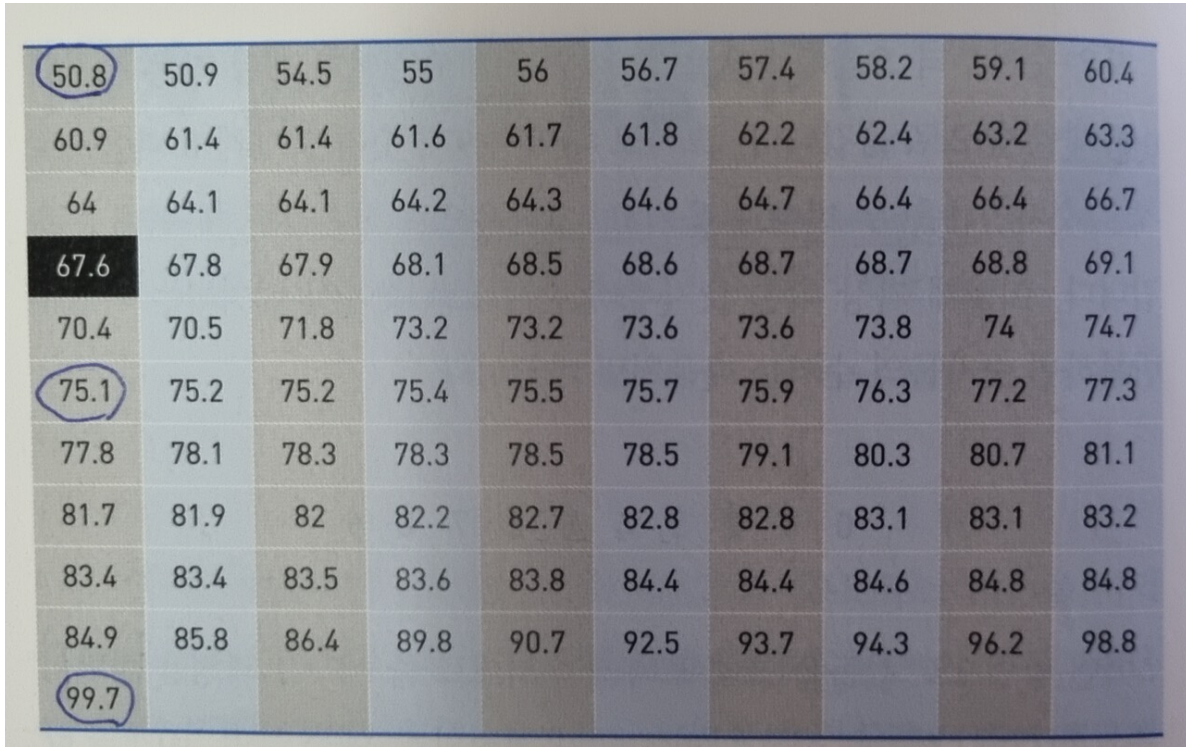
- 분위수(Quantile)
  - 백분위수(Percentile) : 100등분, 기호%를 사용.
  - 사분위수(Quartile) : 0%, 25%, 50%, 75%, 100%
  - 최소값(minimum) : 0% 수
  - 최대값(maximum) : 100% 수
  - 중앙값(median) : 50% 수, 101명 중 51번째 수

연속형 변수

- 예시) 프로듀스 101

- 최소값, 최대값
  - 최소값 : 50.8
  - 최대값 : 99.7

- 백분위 '하위 30%'가 탈락이라면?
  - 기준 : 31번째 값 67.6
  - 통계학에서는 기본이 오름차순이기 때문에,
    보통 '하위'라는 표현은 생략하고 '30%지점'이라고 표현



50.8	50.9	54.5	55	56	56.7	57.4	58.2	59.1	60.4
60.9	61.4	61.4	61.6	61.7	61.8	62.2	62.4	63.2	63.3
64	64.1	64.1	64.2	64.3	64.6	64.7	66.4	66.4	66.7
67.6	67.8	67.9	68.1	68.5	68.6	68.7	68.7	68.8	69.1
70.4	70.5	71.8	73.2	73.2	73.6	73.6	73.8	74	74.7
75.1	75.2	75.2	75.4	75.5	75.7	75.9	76.3	77.2	77.3
77.8	78.1	78.3	78.3	78.5	78.5	79.1	80.3	80.7	81.1
81.7	81.9	82	82.2	82.7	82.8	82.8	83.1	83.1	83.2
83.4	83.4	83.5	83.6	83.8	84.4	84.4	84.6	84.8	84.8
84.9	85.8	86.4	89.8	90.7	92.5	93.7	94.3	96.2	98.8
99.7									

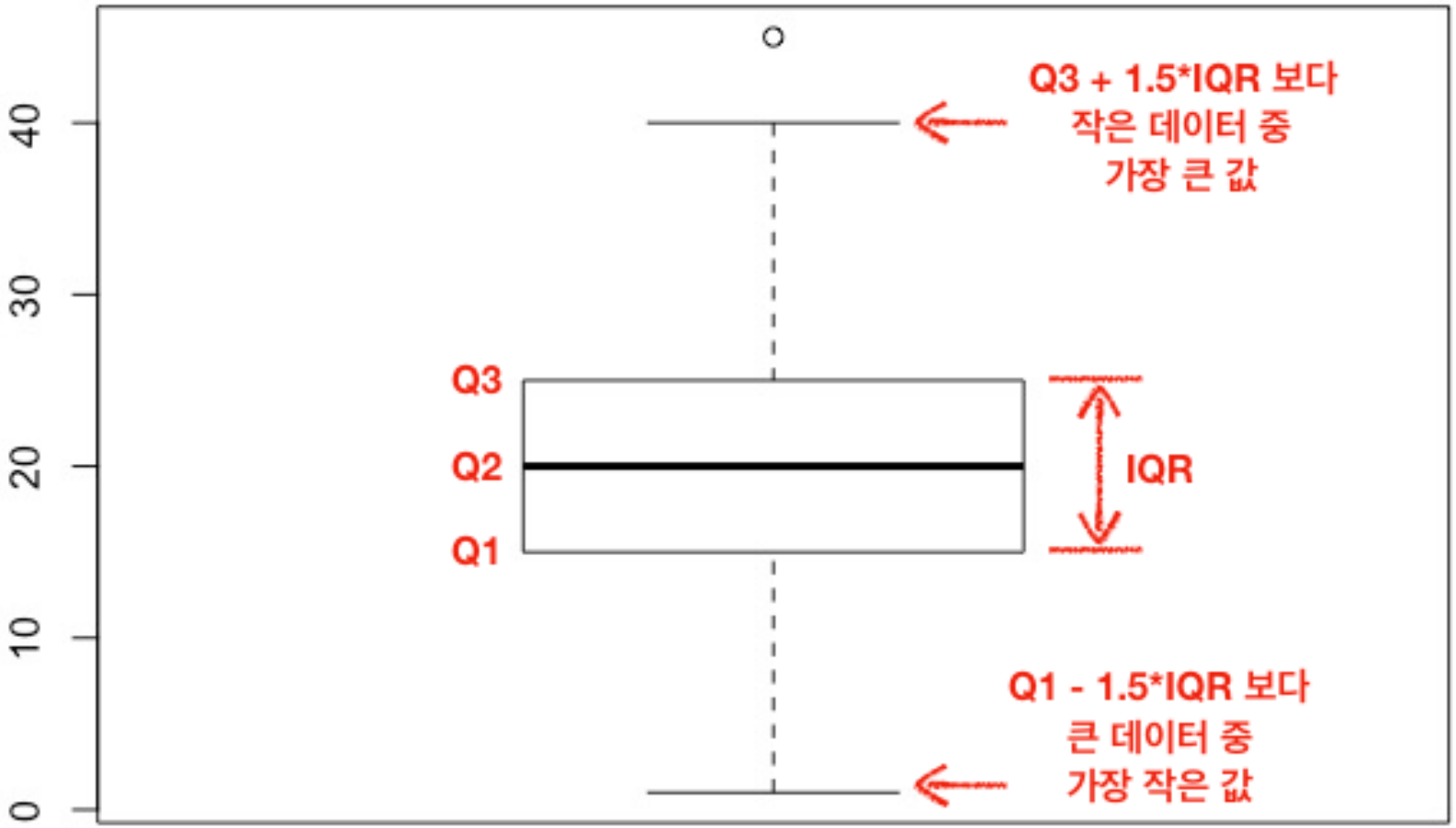
사분위수(Quartile)와 다섯 숫자 요약

하나의 연속형 변수로 사분위수 5개를 계산하고 의미를 찾는 과정을 “다섯 숫자 요약”이라고 표현함

- 최소값(minimum) : 가장 작은 값. 0%에 위치한 값
- Q1 : 25%에 위치한 값
- Q2 = 중앙값(Medium) : 50%에 위치한 값
  - 101명 중 51번째 값이 중앙값이 된다.
- Q3 : 75%에 위치한 값
- 최대값(maximum) : 100%에 위치한 값

상자그림

- 사분위수를 그림으로 표현한 것

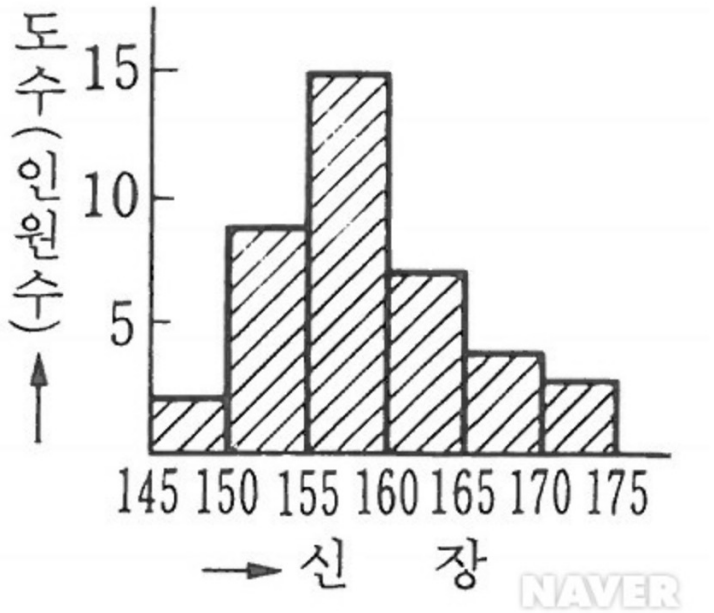


히스토그램

- 표로 되어있는 도수분포표를 그래프로 표현한 것

신장 (cm)	인원수
145이상 ~ 150미만	2
150이상 ~ 155미만	9
155이상 ~ 160미만	15
160이상 ~ 165미만	7
165이상 ~ 170미만	4
170이상 ~ 175미만	3
합계	40

신장 도수분포표



신장 히스토그램

상자그림과 히스토그램 비교

- 상자그림
  - 장점 : 어떤 연속형 변수든 5개의 값으로 표현할 수 있음
  - 한계 : 세부적인 패턴은 놓칠 수 있음
  - 사용 : 간단하고 빠른 분석을 하거나 그룹에 따른 차이를 확인할 때

- 히스토그램
  - 장점 : 구간을 잘 나누면 패턴은 얼마든지 확인 가능
  - 한계 : 5개보다 훨씬 많은 값을 확인해야 할 수도 있음
  - 사용 : 하나의 변수에 대해 좀 더 자세히 살펴볼 때

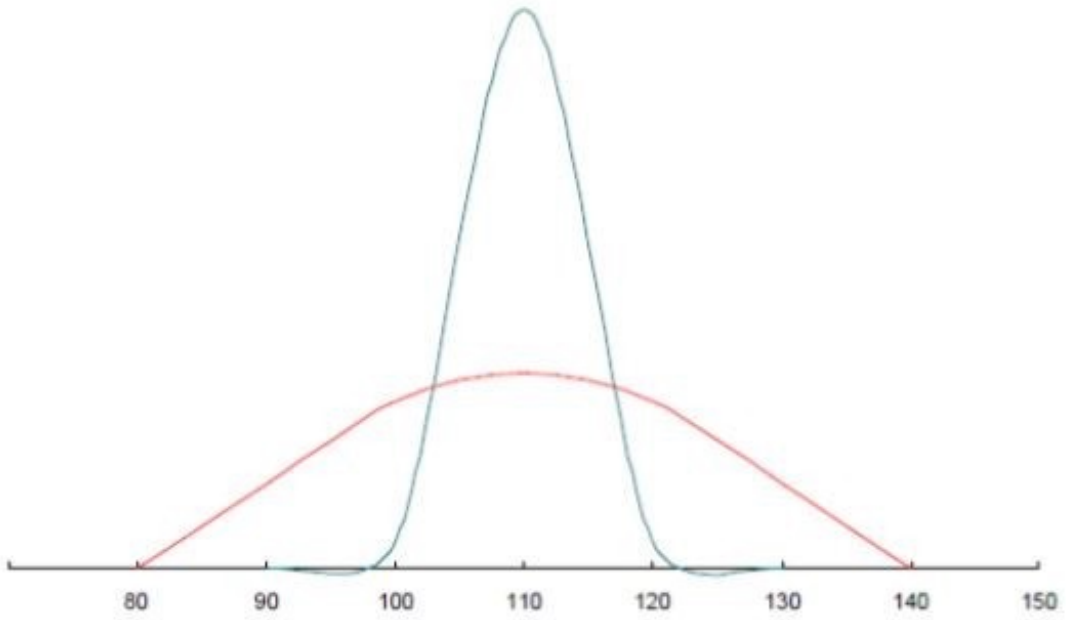
평균(mean)

- 변수의 합계가 고정되어 있을 때, 모든 관측치가 똑같이 나눠 가질 수 있는 값

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

X_i : i번째개체의측정값

$\bar{X}$: 집단의평균



평균

분산(variation)과 표준편차(standard deviation)

- 평균에서 떨어져 있는 거리

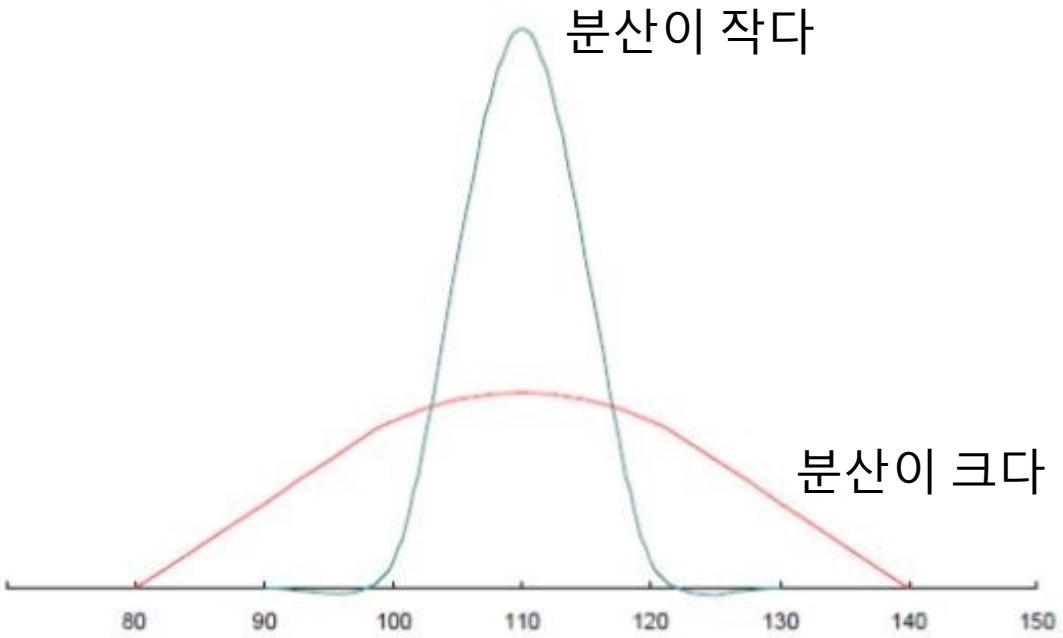
$$v = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}, \qquad \sigma = \sqrt{v}$$

X_i : i번째 개체의 측정값

$\bar{X}$: 집단의 평균

v : 집단의 분산

σ : 집단의 표준편차



- 모든 관측치가 똑같은 값을 가지면 분산은 0이 된다.

- 관측치들이 서로 다른 값을 가지면 분산은 0보다 커진다.
  관측치들이 서로 큰 차이가 있을수록 분산은 점점 커지고 분산이 클수록 관측치 간 불평등이 심하다는 것을 의미한다.

표준화(standardization)

- 중심화 : 각 관측치에서 평균을 빼는 것
  - 중심화를 하면 각 관측치가 평균보다 큰지 작은지 알 수 있다.

- 척도화 : 각 관측치를 표준편차로 나누는 것
  - 전반적인 크기를 보정할 뿐 아니라 단위를 없애준다.

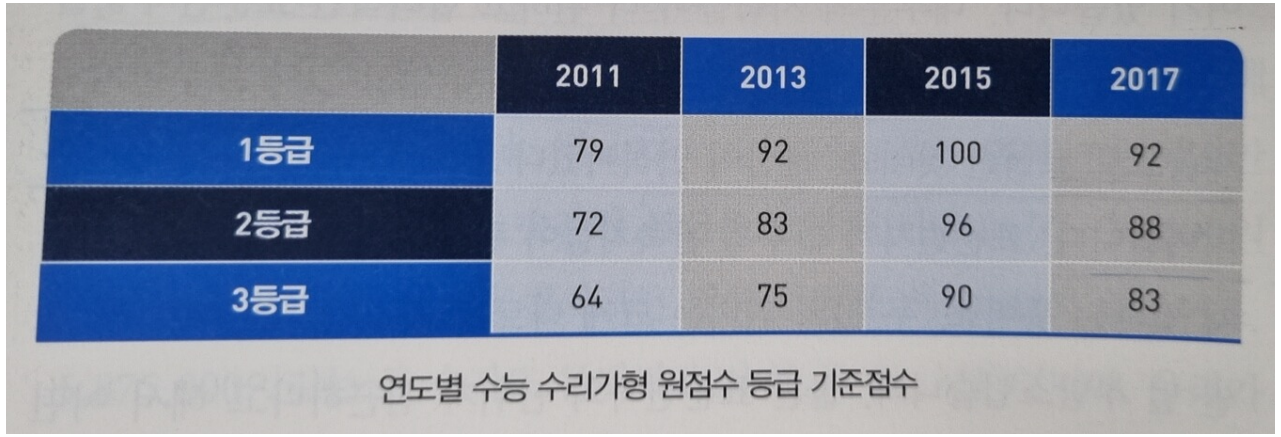
- 표준화 : 변수를 먼저 중심화하고 다음으로 척도화하는 과정
  - 표준화를 거친 변수의 평균은 0이 되고 표준편차는 1이 된다.
  - 따라서 서로 다른 변수들을 비교할 수 있다.

불수능과 물수능

- 2011년 80점을 받은 학생과 2015년 100점을 받은 학생 중 어느 쪽이 더 잘한 것일까?

  2011년 : 평균 47.8, 표준편차 19.7

  2015년 : 평균 55.4, 표준편차 28.5



	2011	2013	2015	2017
1등급	79	92	100	92
2등급	72	83	96	88
3등급	64	75	90	83

연도별 수능 수리가형 원점수 등급 기준점수

불수능과 물수능

● 2011년 80점을 받은 학생과 2015년 100점을 받은 학생 중 어느 쪽이 더 잘한 것일까?

	평균/표준편차	표준화	표준 점수
2011 80 점	평균 : 47.8 표준편차 : 19.7	$\frac{80 - 47.8}{19.7} = 1.63$	$1.63 \times 20 + 100 = 132.6$
2015 100 점	평균 : 55.4 표준편차 : 28.5	$\frac{100 - 55.4}{28.5} = 1.56$	$1.56 \times 20 + 100 = 131.2$

● 시험이 어려울수록 점수는 하향평준화된다.

-> 표준점수로 환산해보니 2011년 80점을 받은 학생이 더 잘했다.

실생활의 표준화

- IQ점수는 평균이 100인 표준화된 점수를 사용

- 옷 사이즈도 신체 치수의 평균을 기준으로 100(혹은 Medium)이라고 하고, 조금 크면 105(Large), 조금 작으면 95(Small)라고 함.

- 여성복 55 사이즈는 1979년 여성 평균 키 155cm, 평균 가슴둘레 85cm를 기준으로 잡은 것

변수 간 상관관계

상관관계(correlation)

- 공분산(covariance) : 변수 x와 y를 함께 사용해서 계산한 분산

$$\text{분산} = \sigma^2 = \frac{\sum (X-\bar{X})^2}{n}$$

$$\text{공분산} = Cov[X, Y] = E[(X-\bar{X})(Y-\bar{Y})] = \sigma_{XY}$$

$$= \frac{\sum (X_i-\bar{X})(Y_i-\bar{Y})}{n}$$

- 공분산과 상관관계
  - 공분산 값이 양수일 때 “두 변수가 양의 상관관계”
  - 공분산 값이 음수일 때 “두 변수가 음의 상관관계”

상관계수(correlation coefficient)

- 상관계수 : 표준화된 두 변수의 공분산
  - 두 변수의 상관계수가 -1일수록 음의 상관관계
  - 두 변수의 상관계수가 1일수록 양의 상관관계
  - 두 변수의 상관계수가 0일수록 관계가 없다.