

# 2021학년도 2학기 언어와 컴퓨터

## 제13강 벡터, 통계, 데이터 시각화

박수지

서울대학교 인문대학 언어학과

2021년 10월 27일 수요일

## 오늘의 목표

- 1 Matplotlib로 막대그래프와 산점도를 그릴 수 있다.
- 2 (NumPy를 이용하여) 벡터와 행렬을 만들고 계산할 수 있다.
- 3 (SciPy를 이용하여) 수치형 데이터의 통계량을 구할 수 있다.
- 4 데이터를 중심 경향성, 산포도, 상관관계로 기술할 수 있다.

## 오늘 못 하면 다음 주에 할 일

- 1 Matplotlib로 막대그래프와 산점도를 그릴 수 있다.
- 2 (NumPy를 이용하여) 벡터와 행렬을 만들고 계산할 수 있다.
- 3 (SciPy를 이용하여) 수치형 데이터의 통계량을 구할 수 있다.
- 4 데이터를 중심 경향성, 산포도, 상관관계로 기술할 수 있다.

## 데이터의 유형

- 범주형
  - 명목형
  - 순서형
- 수치형
  - 이산형
  - 연속형

## 예시: 언어학 데이터

명목형 ‘밀덕’(밀리터리 덕후)을 어떻게 발음합니까?

- [밀덕], [밀떡]

순서형 이 문장이 자연스럽습니까?

- 매우 어색/어색한 편/보통/...

이산형 각 단어에 장애음이 몇 개 있는가?

- 0, 1, 2, 3, ...

연속형 어두 자음의 VOT가 몇 ms인가?

- -14.15, 3.60, 23.61, -7.42, ...

## 벡터

(주로 수치형) 데이터를 표현하는 방식

성분 표시  $v = (v_1, v_2, \dots, v_n) \leftarrow n$ -차원 벡터

- 연산
- 덧셈  $(5, 0, 1) + (7, 5, 0) = (12, 5, 1)$
  - 상수배  $2 \times (5, 0, 1) = (10, 0, 2)$

- 속성
- 내적  $(5, 0, 1) \cdot (7, 5, 0) = (5 \times 7) + (0 \times 5) + (1 \times 0) = 35$
  - 길이  $\|(5, 0, 1)\| = \sqrt{(5, 0, 1) \cdot (5, 0, 1)} = \sqrt{26}$
  - 거리  $\|(5, 0, 1) - (7, 5, 0)\| = \|(-2, -5, 1)\| = \sqrt{30}$

## 벡터 연산 시도

```
>>> [5, 0, 1] + [7, 5, 0] # 연결  
[5, 0, 1, 7, 5, 0]  
>>> 2 * [5, 0, 1] # 반복  
[5, 0, 1, 5, 0, 1]
```

## 해결 방법

- 1 정의를 따라 코드를 작성한다.
- 2 NumPy를 사용한다.

## 벡터의 덧셈의 정의

$$v + w = (v_1 + w_1, v_2 + w_2, \dots, v_n + w_n)$$

```
1 def vector_add(v, w):  
2     result = []  
3     for i in range(len(v)):  
4         result.append(v[i] + w[i])  
5     return result
```

## 개선

- 두 개 이상의 열에 대한 반복문  $\Rightarrow$  zip() 함수 사용
- 열에 대응하는 리스트  $\Rightarrow$  List comprehension 사용

## 벡터의 덧셈, 뺄셈, 상수배

```
1 def vector_add(v, w):  
2     return [v_i + w_i for v_i, w_i in zip(v, w)]  
3  
4 def vector_subtract(v, w):  
5     return [v_i - w_i for v_i, w_i in zip(v, w)]  
6  
7 def scalar_multiply(c, v):  
8     return [c * v_i for v_i in v]
```

실습 코드:

[https://github.com/insight-book/data-science-from-scratch/blob/master/scratch/linear\\_algebra.py](https://github.com/insight-book/data-science-from-scratch/blob/master/scratch/linear_algebra.py)

## 데이터를 기술하는 방법

### ■ 통계량

- 중심 경향성: 평균, 중앙값, 최빈값 – 어디에 몰려 있는가?
- 산포도: 표준편차, 사분위수 – 얼마나 흩어져 있는가?
- 상관관계

### ■ 시각화

- 히스토그램: 한 가지 데이터의 분포
- 산점도: 두 가지 데이터의 관계



# 중심 경향성

## 평균

- 계산이 간편하다.
- 데이터의 변화에 따라 변한다.  $\Rightarrow$  이상치에 민감하다.

## 중앙값

- 이상치가 포함되어도 큰 영향을 받지 않는다.
- 데이터를 크기순으로 정렬해야 한다.  $\Rightarrow$  계산량이 많아진다.

# “퍼짐 경향성”

## 편차

$(\text{편차}) = (\text{관측치}) - (\text{평균})$

- 편차의 합은 항상 0이다.

⇒ 데이터가 얼마나 퍼져 있는지를 반영할 수 없다.

⇒ 편차의 “크기”를 사용해야 한다.

## 표준편차

편차의 크기를 측정하는 방법

- 1 절댓값을 취해서 더한다. ⇒ 미분을 할 수 없다.

- 2 제곱을 해서 더한다. ⇒ 채택  
⇒ 다 더한 뒤 제곱근을 취하여 원래의 값과 같은 1차로 만든다.

# 두 변수 사이의 관계

## 공분산

두 변수가 각각의 평균에서 얼마나 떨어져 있는지를 측정하는 통계량

cf 분산: 하나의 변수가 평균에서 얼마나 떨어져 있는가?

편차를 곱해서 더한다. → 상관관계에 따라 음수가 될 수 있다.

## (비편향)공분산

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}, \quad \bar{x} = (x \text{의 평균}), \bar{y} = (y \text{의 평균})$$

편향공분산:  $(n - 1)$  대신  $n$ 으로 나눈 것

# 두 변수 사이의 관계

## 상관관계

$$\frac{(\text{x와 y의 공분산})}{(\text{x의 표준편차}) \times (\text{y의 표준편차})}$$

값의 범위 단위에 상관 없이 항상  $-1$ 에서  $1$  사이의 값을 가진다.

양의 상관관계  $x$ 가 증가할 때  $y$ 도 증가한다.

음의 상관관계  $x$ 가 증가할 때  $y$ 는 감소한다.

# 상관관계의 주의사항

## 상관관계 이외의 관계

상관관계가 0인 두 변수

$x = [-2, -1, 0, 1, 2]$

$y = [2, 1, 0, 1, 2]$

## 상관관계 $\neq$ 연관성

상관관계가 1인 두 변수

$x = [-2, -1, 0, 1, 2]$

$y = [99.98, 99.99, 100, 100.01, 100.02]$

# 상관관계의 주의사항

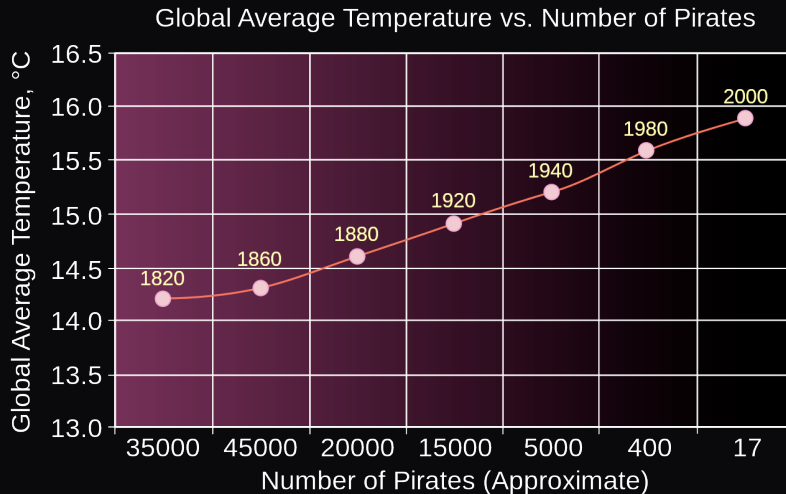
## 상관관계 $\neq$ 인과관계

접속 시간과 친구 수 사이에 상관관계가 존재하는데...

- 친구가 많기 때문에 접속 시간이 늘어났는가?
- 오래 접속하다 보니 친구가 늘어났는가?
- ...이도 저도 아닌 우연인가?

실습 코드: <https://github.com/insight-book/data-science-from-scratch/blob/master/scratch/statistics.py>

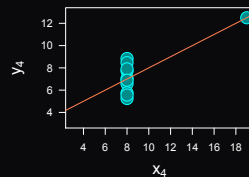
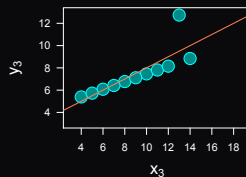
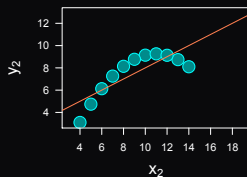
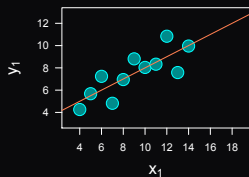
# 상관관계의 주의사항



[https://commons.wikimedia.org/wiki/File:PiratesVsTemp\(en\).svg](https://commons.wikimedia.org/wiki/File:PiratesVsTemp(en).svg)

# Anscombe's quartet

같은 평균, 같은 분산, 같은 상관관계



<https://commons.wikimedia.org/wiki/File:Anscombe.svg> 가공

$x$ 의 평균 9  
 $x$ 의 분산 11

$y$ 의 평균 7.5  
 $y$ 의 분산 4.125

상관관계 0.816



# 심슨의 역설

전체의 대소 관계와 부분의 대소 관계가 달라지는 현상

## 전체 친구 수: 서부 > 동부

지역	사용자 수	평균 친구 수
서부	101	8.2
동부	103	6.5

## 학위별 친구 수: 서부 < 동부

지역	학위	사용자 수	평균 친구 수
서부	박사	35	3.1
동부	박사	70	3.2
서부	기타	66	10.9
동부	기타	33	13.4

## 오늘 배운 것

- 벡터와 행렬의 연산
- 기술통계량
  - 평균, 중앙값, 최빈값
  - 표준편차, 4분위수
  - 상관계수

## 다음 주에 배울 것

- N-그램
  - SLP3e Ch. 3
- 조건부확률
  - 밑바닥부터... 6장