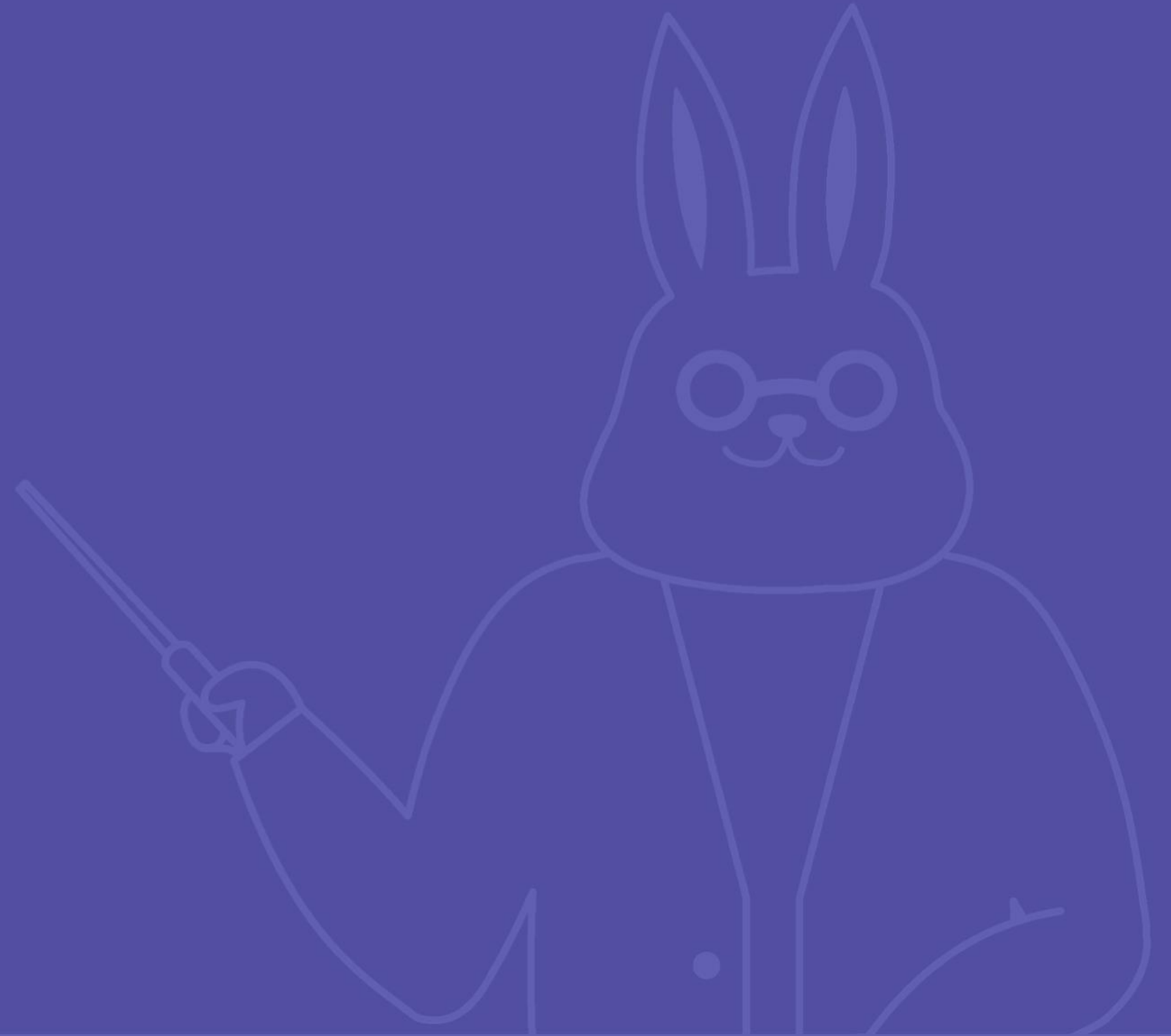




# 머신러닝 시작하기

## 01 자료 형태의 이해

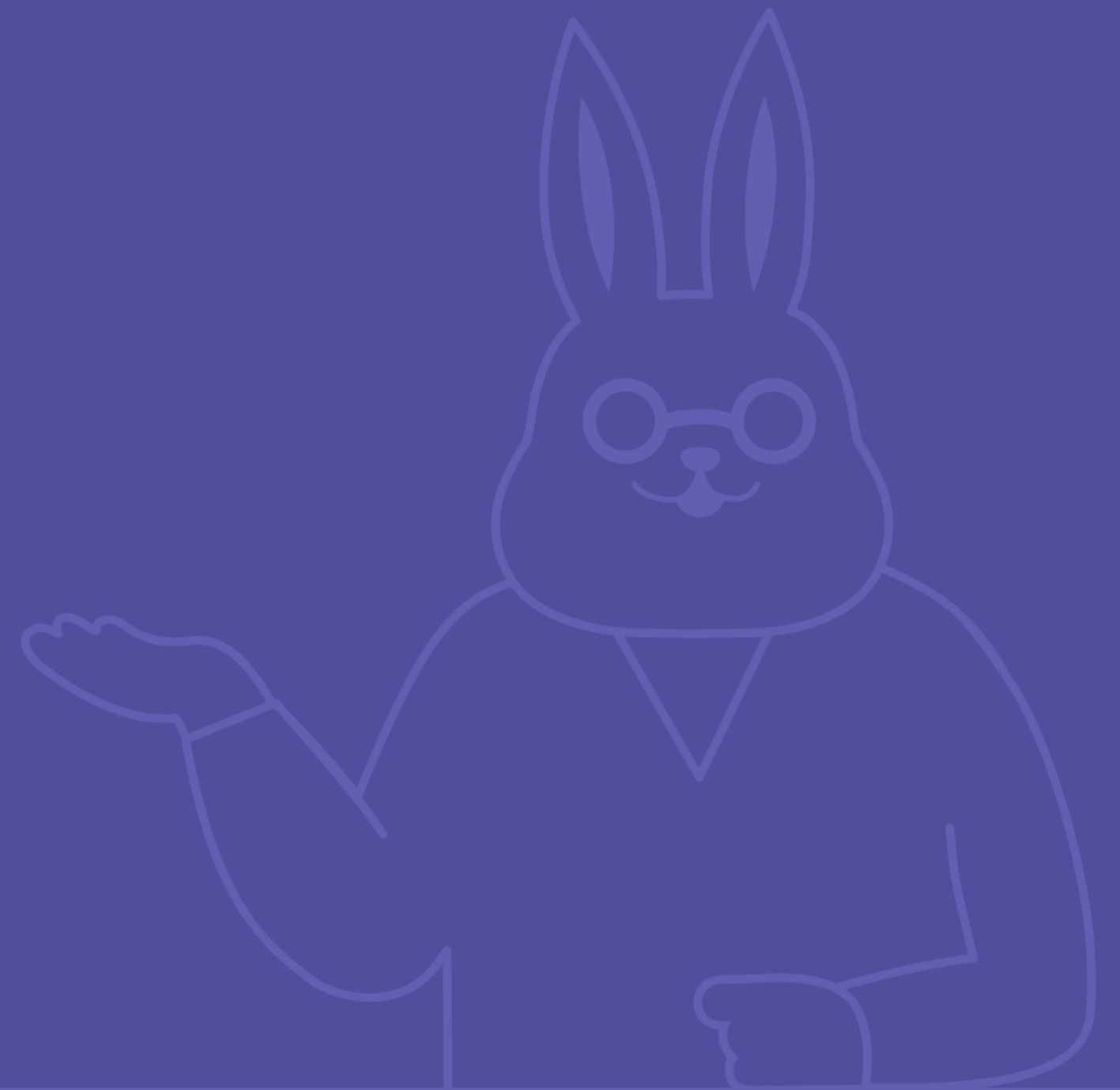


# 목차

- 01. 자료의 형태
- 02. 범주형 자료의 요약
- 03. 수치형 자료의 요약

01

# 자료의 형태

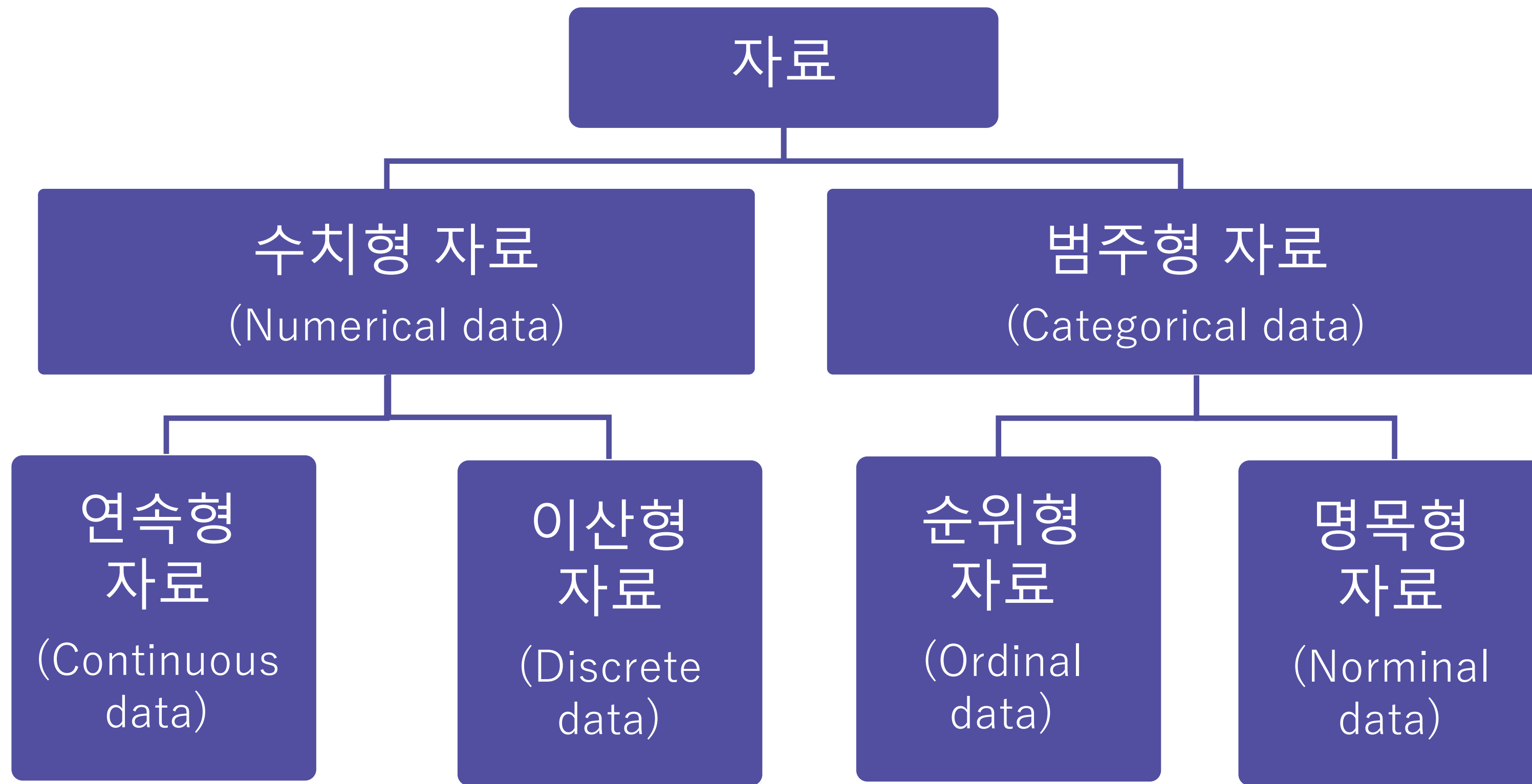


## ✓ 자료의 형태를 알아야 하는 이유

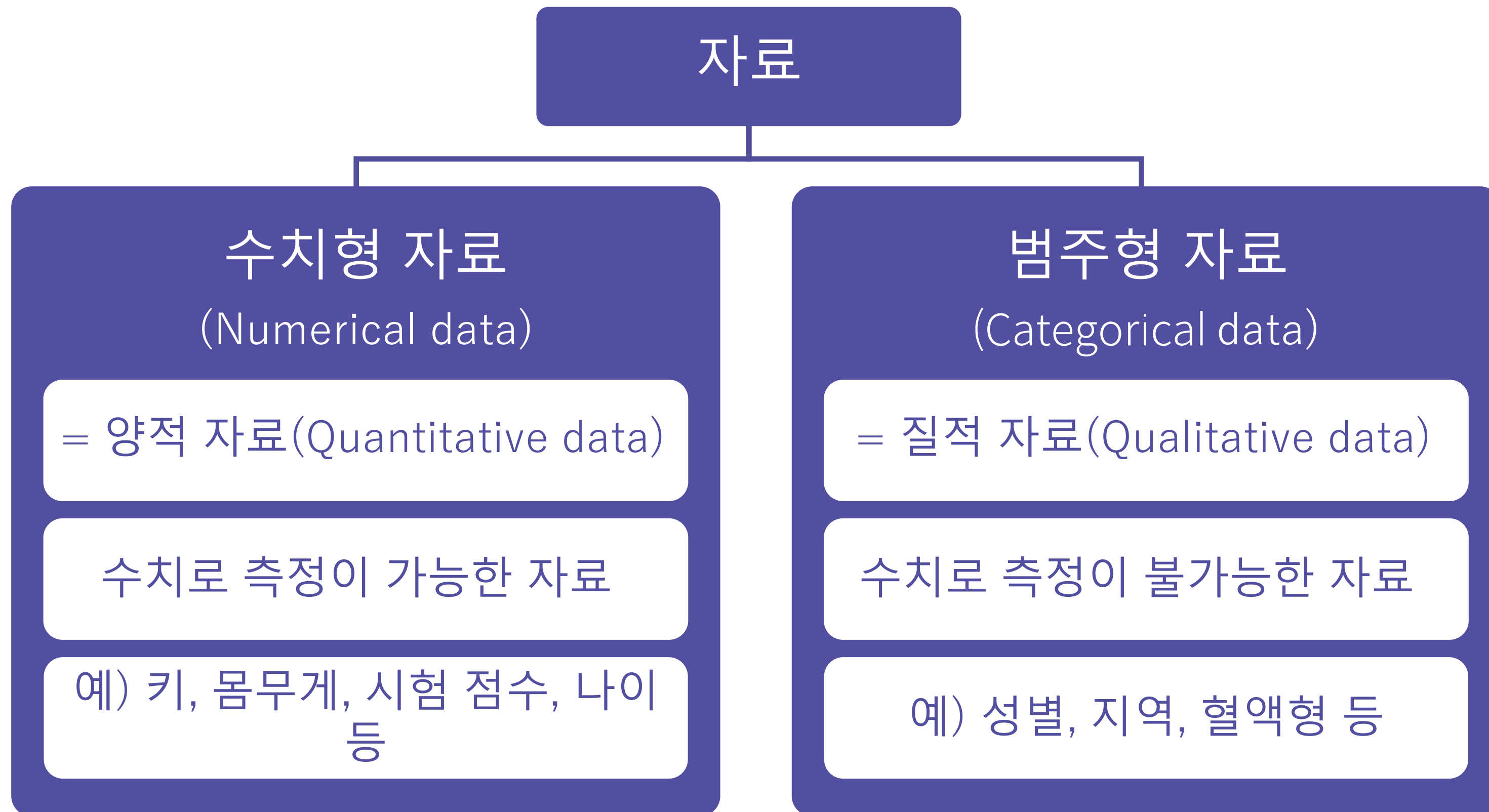
머신러닝은 데이터라는 디지털 자료를 바탕으로 수행하는 분석방식  
**자료의 형태를 파악함**은 머신러닝 사용하기 위한 **필수 과정**으로  
아래 물음의 답을 얻을 수 있음

- **데이터가 어떻게 구성되어 있을까?**
- 어떤 머신러닝 모델을 사용해야 할까?
- 데이터 전 처리를 어떻게 해야 할까?
- ...

## ✔ 자료 형태 구분

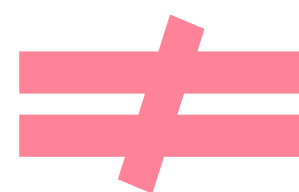


## ✔ 자료 형태 구분



## ✓ 자료의 형태 구분 시, 주의점

범주형 자료와  
수치 자료의 구분



자료의 숫자 표현 가능 여부

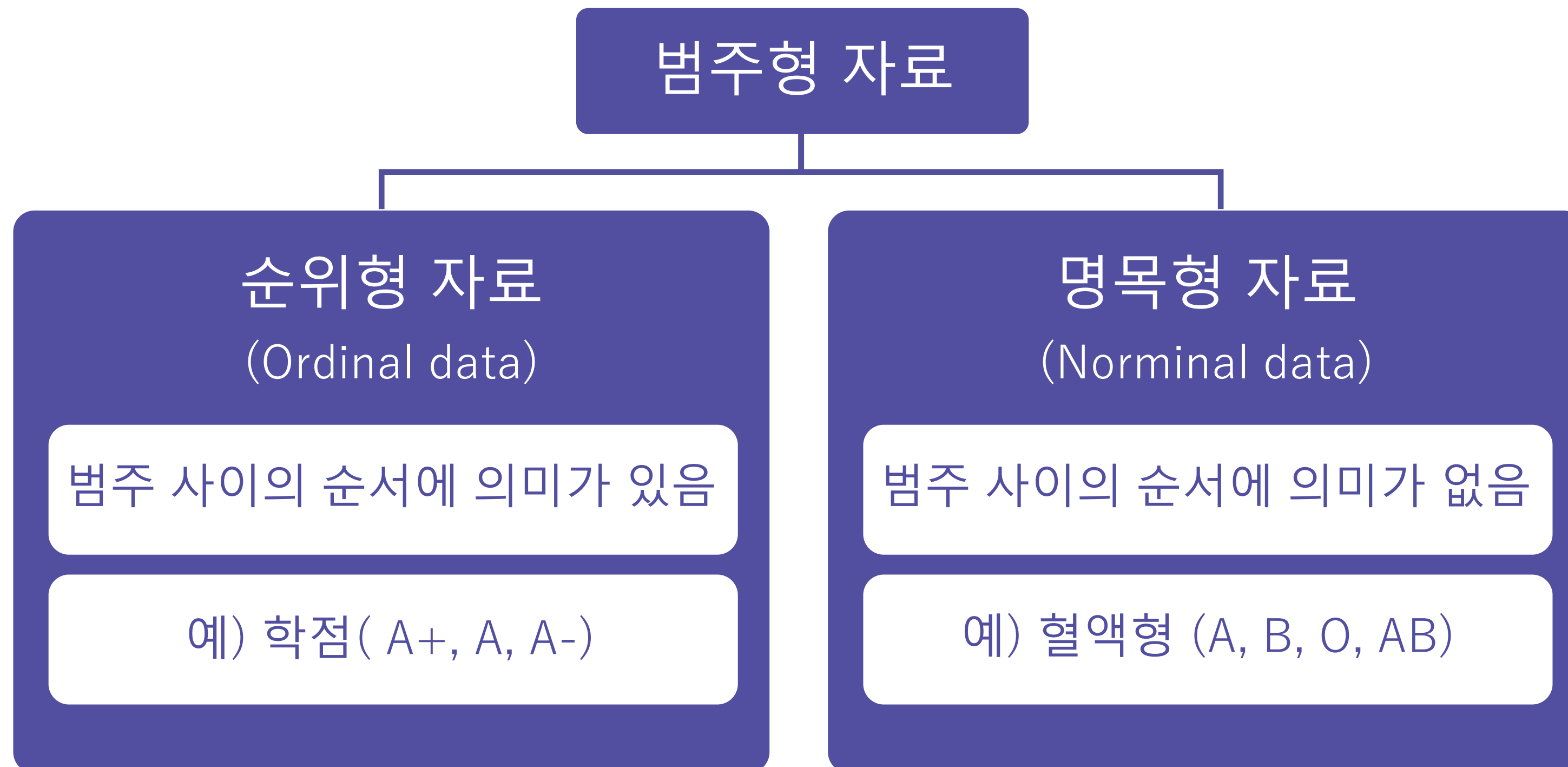
범주형 자료가  
숫자로 표현되는 경우

남녀 성별 구분 시, 남자를 1, 여자를 0  
으로 표현하는 경우, 숫자로 표현 되었  
으나 범주형 자료

수치형 자료를 범주형 자료  
로 변환하는 경우

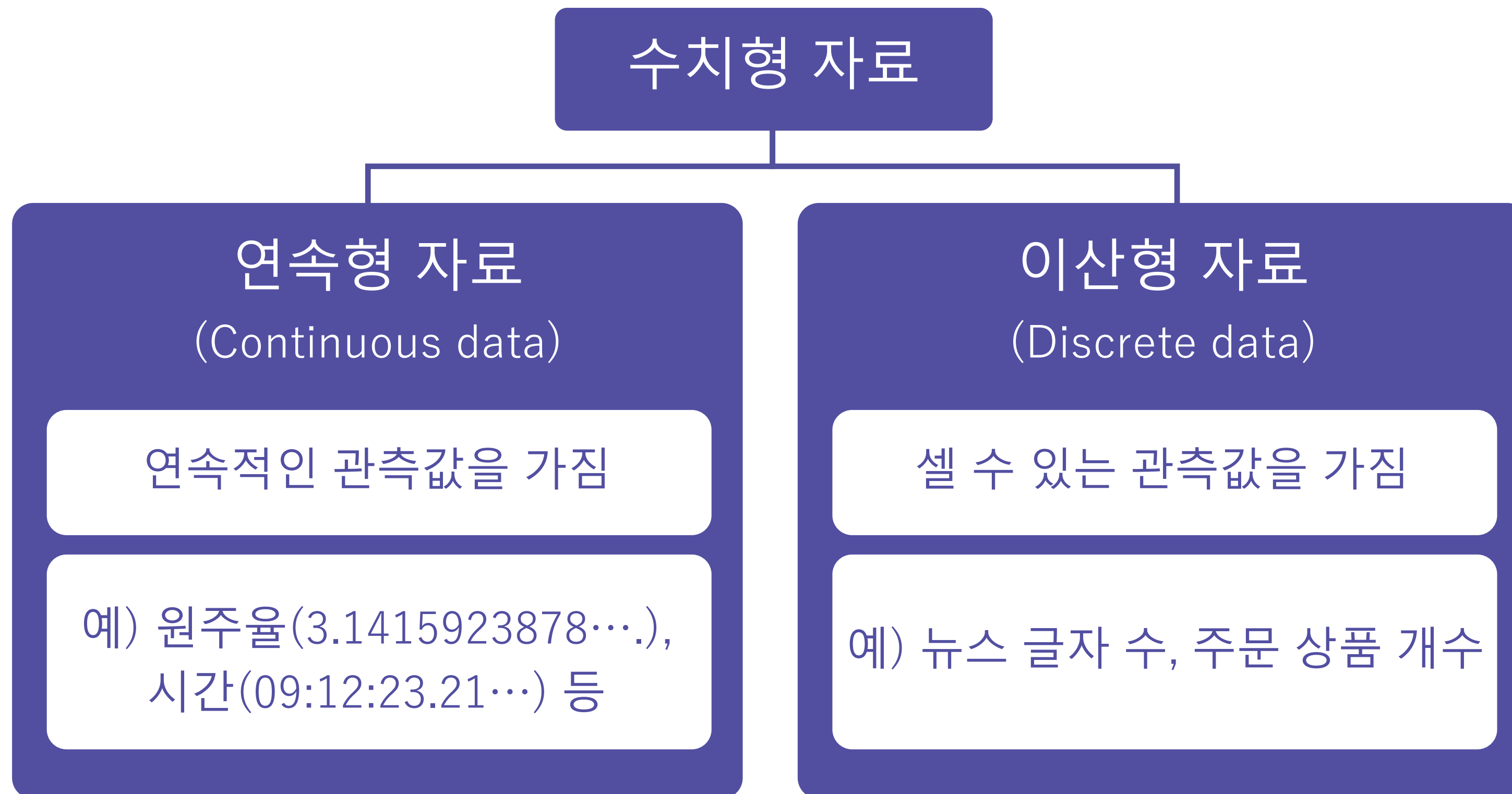
나이 구분 시, 나이 값은 수치형 자료지  
만 10 ~ 19세, 20 ~ 29세 등 나이 대에  
따라 구간화 하면 범주형 자료

## ✓ 범주형 자료 구분





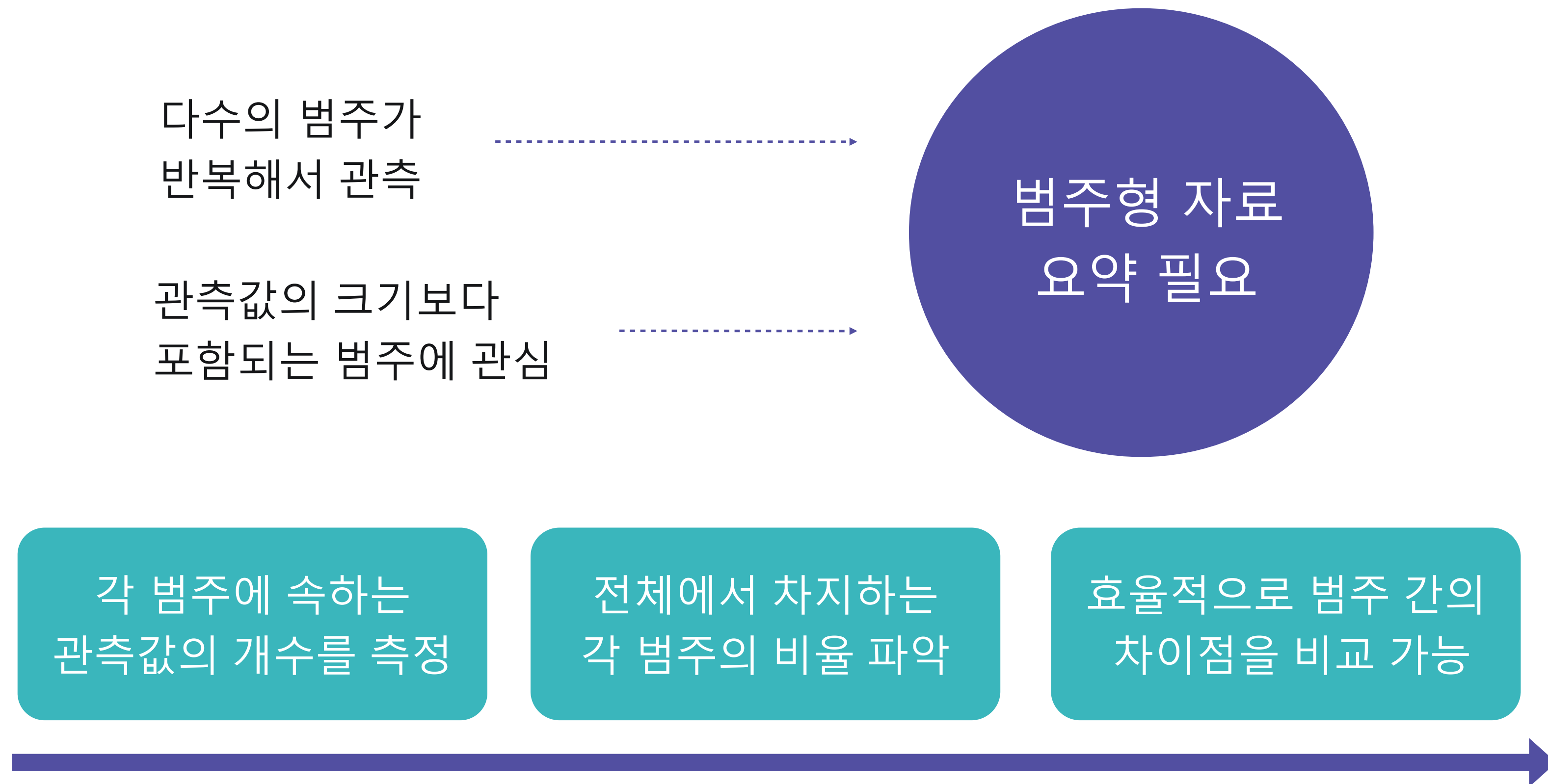
## ✔ 수치형 자료 구분



02

# 범주형 자료의 요약





✔ 도수분포표

강의 만족도 설문 (100명 조사)

| No. | ID    | 만족도   |
|-----|-------|-------|
| 1   | 23512 | 매우 만족 |
| 2   | 12351 | 보통    |
| 3   | 12532 | 만족    |
| 4   | 25432 | 불만족   |
| ... | ...   | ...   |
| 100 | 21353 | 보통    |

✔ 도수분포표

강의 만족도 설문 (100명 조사)

| 범주     | 도수 | 상대도수 | 누적 상대도수 |
|--------|----|------|---------|
| 매우 만족  | 30 | 0.3  | 0.3     |
| 만족     | 10 | 0.1  | 0.4     |
| 보통     | 30 | 0.3  | 0.7     |
| 불만족    | 15 | 0.15 | 0.85    |
| 매우 불만족 | 15 | 0.15 | 1.00    |

## ✔ 도수분포표 정의

도수  
(Frequency)

각 범주에 속하는 관측값의 **개수**

```
value_counts()
```

상대도수  
(Relative Frequency)

도수를 자료의 **전체 개수**로 나눈 **비율**

```
value_counts(normalize=True)
```

도수분포표  
(Frequency Table)

범주형 자료에서 **범주와 그 범주에 대응**하는 도수, 상대도수를 나열해 표로 만든 것

## ✓ 막대그래프 (Bar Chart)

## 강의 만족도 설문 (100명 조사)

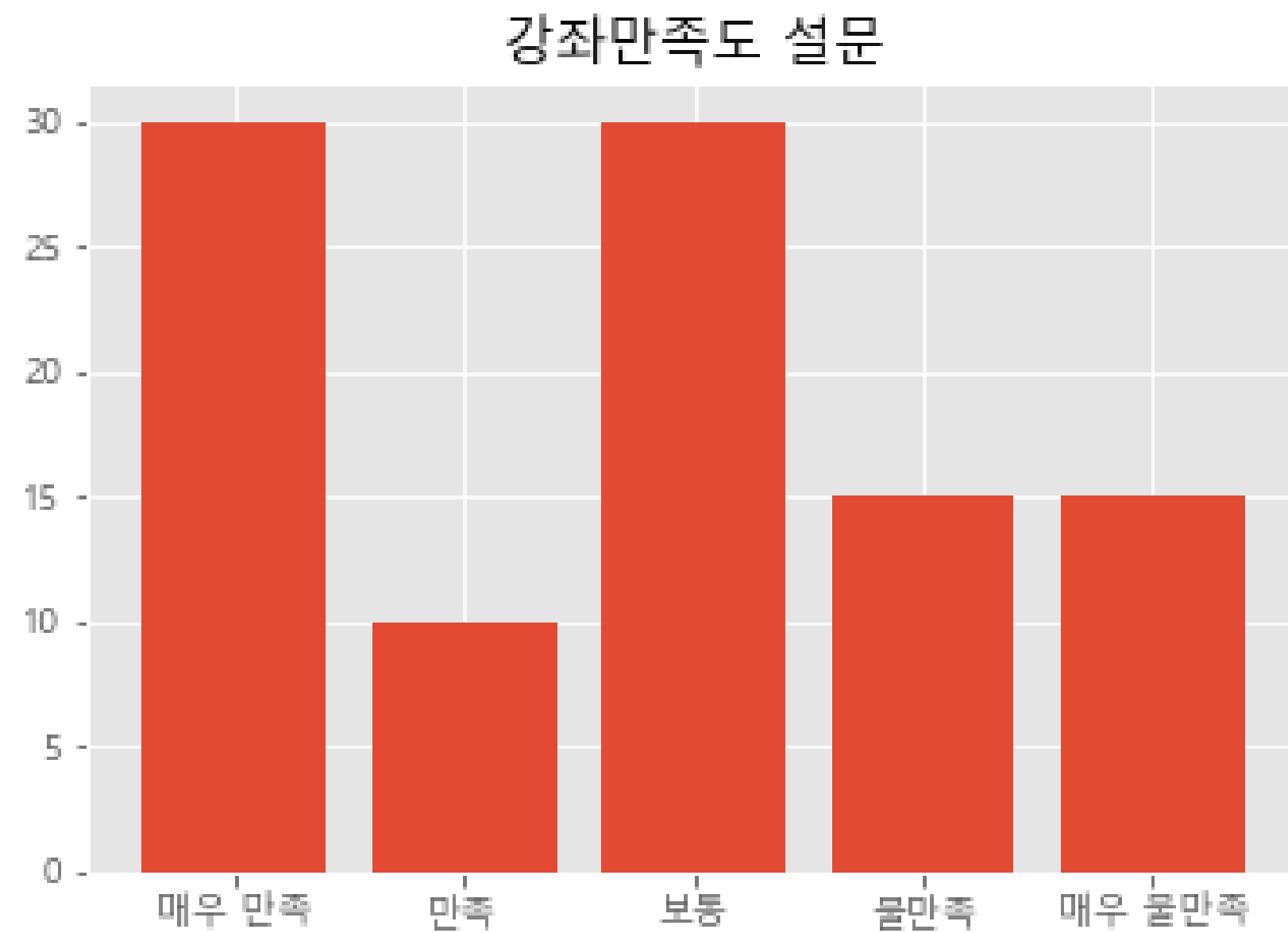
| 범주     | 도수 |
|--------|----|
| 매우 만족  | 30 |
| 만족     | 10 |
| 보통     | 30 |
| 불만족    | 15 |
| 매우 불만족 | 15 |

각 범주들의 **크기 차이**를  
잘 보여줄 수 있는  
방법이 있을까?

✓ 막대그래프 (Bar Chart)

`plt.bar()`

y축 =  
도수의 크기





## ✓ 막대그래프 (Bar Chart)

각 범주에서 도수의 크기를 막대로 그림

그래프의 y축 : 도수에 대한 눈금

그래프의 x축 : 범주를 나열

### 장점

각 범주가 가지는 도수의  
크기 차이를 비교하기 쉬움

### 단점

각 범주가 차지하는 비율의  
비교는 어려움

03

# 수치형 자료의 요약



## ✓ 수치를 통한 자료 요약

### 수치형 자료의 특징

- 범주형 자료와 달리 수치로 구성되어 있기에 통계값을 사용한 요약이 가능함
- 시각적 자료로는 이론적 근거 제시가 쉽지 않은 단점을 보완함



많은 양의 자료를 의미 있는 수치로 요약하여  
대략적인 분포상태를 파악 가능

## ✓ 평균(Mean)

```
np.mean()
```

관측값들을 대표할 수 있는 통계값

수치형 자료의 통계값 중 가장 많이 사용되는 방법

모든 관측값의 합을 자료의 개수로 나눈 것

자료  $x_1, x_2, \dots, x_n$  의 평균을  $\bar{x}$  로 표기

$$\bar{x} = \frac{\text{모든 관측값의 합계}}{\text{총 자료의 개수}} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

### ✓ 평균의 특징

- 관측값의 산술평균으로 사용
- 통계에서 기초적인 통계 수치로 가장 많이 사용
- 극단적으로 큰 값이나 작은 값의 영향을 많이 받음

✓ 퍼진 정도의 측도

평균만으로 분포를 파악하기에 부족

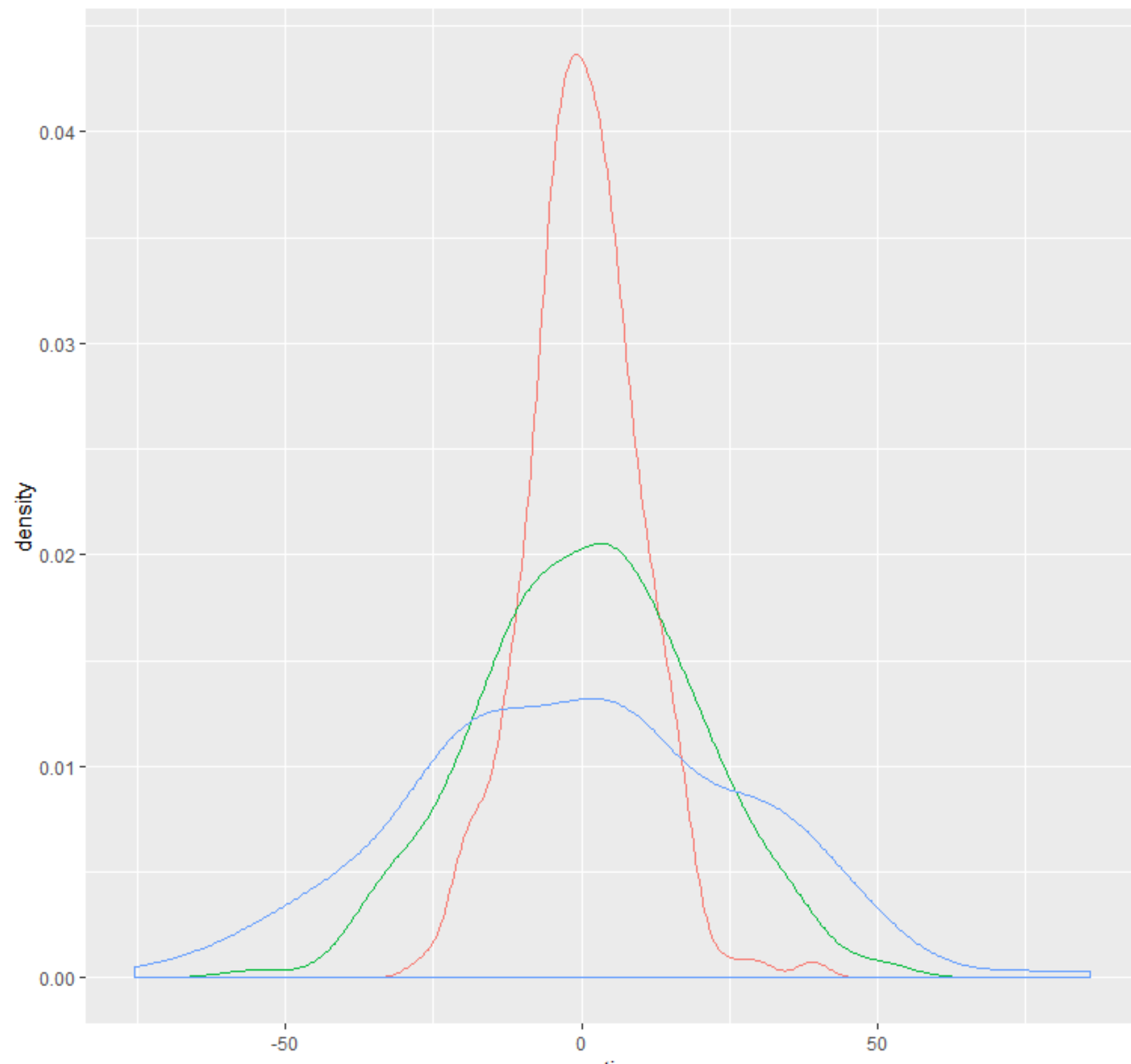


평균 외에 분포가 퍼진 정도를 측도할 수치가 필요



분산, 표준편차 등을  
퍼진 정도의 측도로 사용

✓ 퍼진 정도의 측도



A : 평균 0, 분산 10

B : 평균 0, 분산 20

C : 평균 0, 분산 30

## ✓ 분산

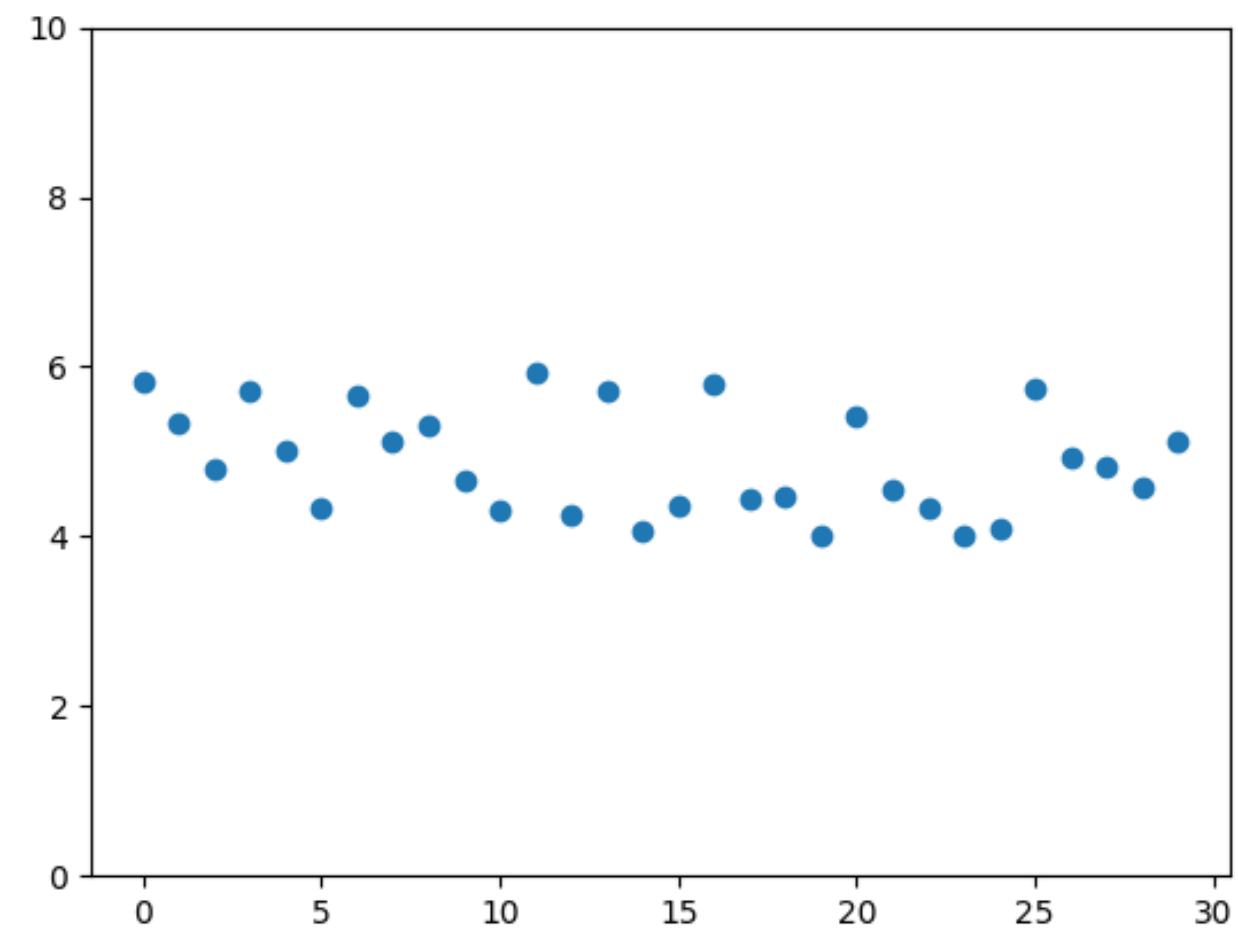
```
from statistics import variance  
variance()
```

자료가 얼마나 흩어졌는지 숫자로 표현

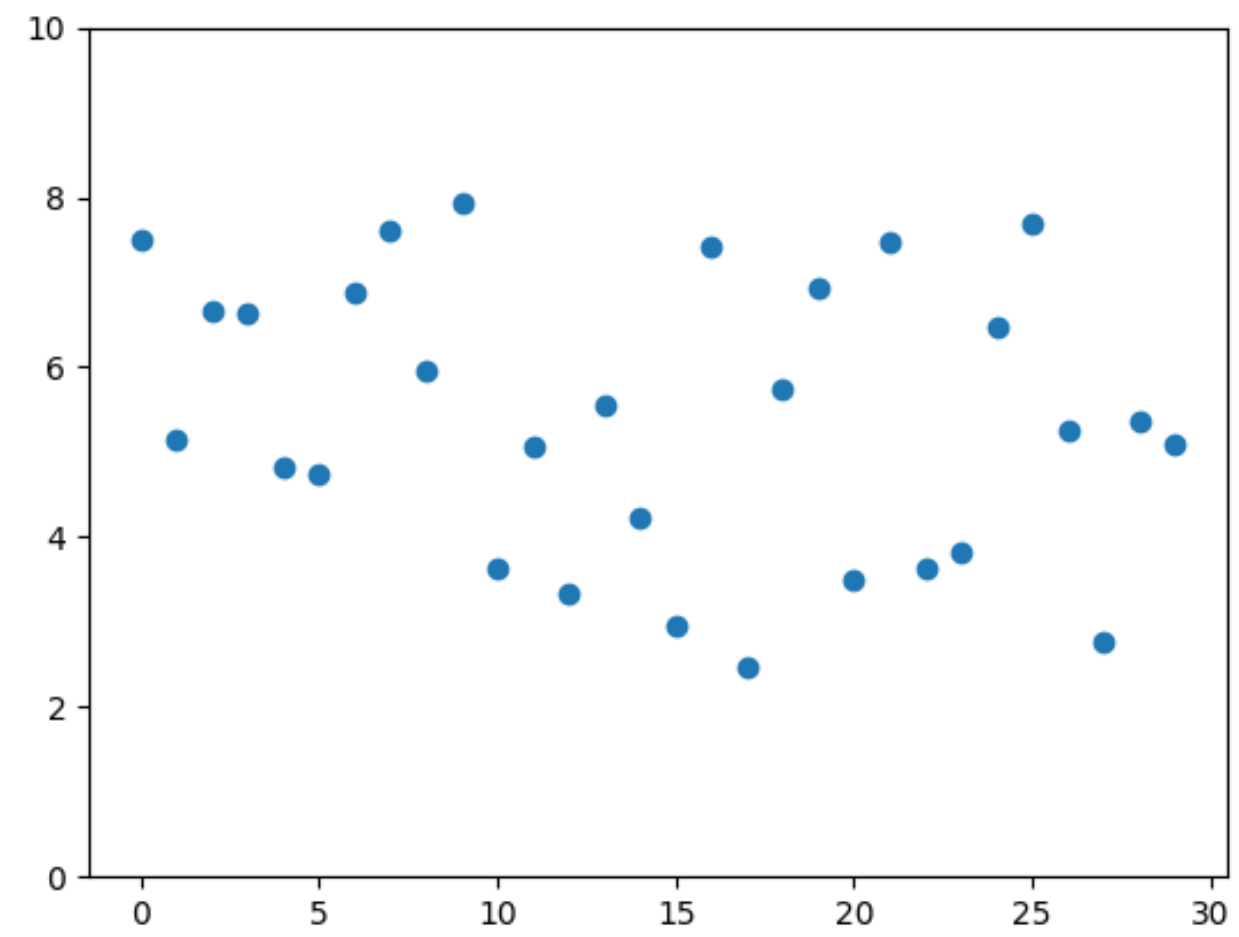
각 관측값이 자료의 평균으로부터 떨어진 정도



✓ 분산



분산이 작다



분산이 크다

## ✓ 표준편차

```
from statistics import stdev  
stdev()
```

분산의 단위 = 관측값의 단위의 제곱

관측값의 단위와 불일치

분산의 양의 제곱근은 관측값과 단위가 일치

분산의 양의 제곱근을 표준편차라 하고  $s$ 로 표기

$$s = +\sqrt{s^2}$$

## ✓ 히스토그램(Histogram)

```
plt.hist()
```

수치형 자료를 일정한 범위를 갖는 범주로 나누고  
막대그래프와 같은 방식으로 그림

도수 비교 ▶

범주

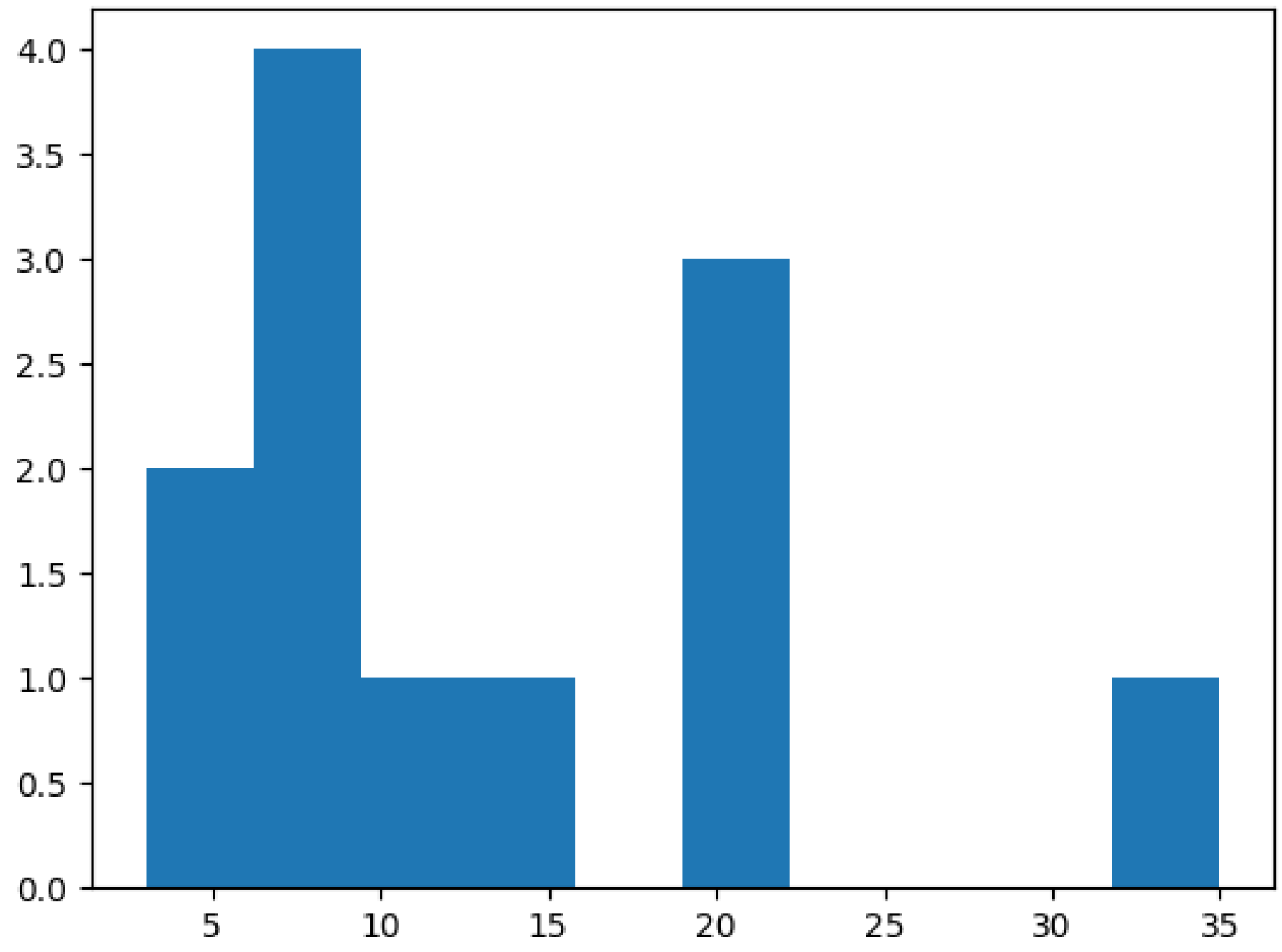
막대그래프

수치

히스토그램

✔ 히스토그램(Histogram)

Y 축 : 빈도



X 축 : 계급

### ✓ 히스토그램의 특징

- 자료의 분포를 알 수 있음
- 계급구간과 막대의 높이로 그림
- 도수, 상대도수를 막대 높이로 사용

# 크레딧

/\* elice \*/

코스 매니저

이해솔

콘텐츠 제작자

이해솔

강사

이해솔

감수자

-

디자이너

강혜정

# 연락처

TEL

070-4633-2015

WEB

<https://elice.io>

E-MAIL

[contact@elice.io](mailto:contact@elice.io)

