

# 파이썬 확률 통계

시각화를 통한 자료의 요약



/\* elice \*/

# 커리큘럼

## 1 ○ 시각화를 통한 자료의 요약

자료의 형태  
범주형 자료의 요약  
수치형 자료의 요약  
분포의 모양

## 2 ○ 논리적인 자료의 요약

중심위치의 측도  
퍼진 정도의 측도  
상자그림  
두 변수 자료의 요약

# 커리큘럼

3 ○

## 확률

사건과 확률의 개념  
순열과 조합  
독립과 조건부 확률  
확률분포

4 ○

## 가설검정 및 추론

여러 가지 확률분포  
통계적 추론  
통계적 가설검정  
검정의 종류와 과정

# 수강 대상



머신러닝, 빅데이터 공부를 위해  
통계학 기초가 필요하신 분



파이썬으로 통계를 해보고 싶은 분



논문 작성 등을 위해 통계분석이 필요하신 분

# 수강 목표

기초 통계학과 파이썬을 함께 배우며  
원리부터 응용까지 시너지를 최대화합니다.

파이썬 기초 지식만으로도  
자유자재로 **통계 분석**을 할 수 있습니다.

통계가 가진 힘을 알고,  
통계에 대한 두려움을 떨칠 수 있습니다.

# 목차

1. 자료의 형태
2. 범주형 자료의 요약
3. 수치형 자료의 요약

# 자료의 형태

# 자료의 형태

## 자료

### 수치형 자료 (Numerical data)

= 양적 자료(Quantitative data)

수치로 측정이 가능한 자료

예) 키, 몸무게, 시험 점수, 나이 등

→ 선형 회귀 분석 등 사용

### 범주형 자료 (Categorical data)

= 질적 자료(Qualitative data)

수치로 측정이 불가능한 자료

예) 성별, 지역, 혈액형 등

→ 로지스틱 회귀 분석 등 사용



# 수치형 자료

## 수치형 자료

### 연속형 자료 (Continuous data)

연속적인 관측값을 가짐

예) 원주율( $3.1415923878\dots$ ),  
시간( $09:12:23.21\dots$ ) 등

### 이산형 자료 (Discrete data)

셀 수 있는 관측값을 가짐

예) 동영상 조회수

# 자료의 형태 구분 시, 주의점

범주형 자료와  
수치 자료의 구분



자료의 숫자 표현 가능 여부

범주형 자료가  
숫자로 표현되는 경우

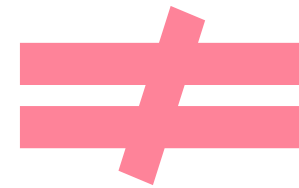
남녀 성별 구분 시, 남자를 1, 여자를 0으로  
표현하는 경우, 숫자로 표현 되었으나  
범주형 자료

수치형 자료를 범주형 자료로  
변환하는 경우

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

# 수치형 자료 구분

연속형 자료



이산형 자료

**연속형 자료는 연속적인 관측**

예 ) 시간 측정

어떤 순간은 09:12:23.21... 처럼 연속되고 있는 상태를 관측한 연속형 자료이나  
평소에는 09시 12분으로 반올림하여 표현하여 이산형 자료로 사용

# 범주형 자료

## 범주형 자료

### 순위형 자료 (Ordinal data)

범주 사이의 순서에 의미가 있음

예) 학점 (A+, A, A-)

### 명목형 자료 (Nominal data)

범주 사이의 순서에 의미가 없음

예) 혈액형 (A, B, O, AB)

# [퀴즈]

## 자료의 형태



# 범주형 자료의 요약

# 범주형 자료 요약

다수의 범주가  
반복해서 관측

관측값의 크기보다  
포함되는 범주에 관심

범주형 자료  
요약 필요

각 범주에 속하는  
관측값의 개수를 측정

전체에서 차지하는  
각 범주의 비율 파악

효율적으로 범주 간의  
차이점을 비교 가능

# 도수분포표

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



# 도수분포표 예시

강의 만족도 설문 (100명 조사)

범주	도수	상대도수
매우 만족	30	0.3
만족	10	0.1
보통	30	0.3
불만족	15	0.15
매우 불만족	15	0.15

# 도수

```
df[범주].value_counts()
```

value\_counts() 함수를 이용해 도수를 구할 수 있습니다.

# 상대도수

```
df[범주].value_counts(normalize=True)
```

value\_counts() 함수에 normalize=True 옵션을 주면  
상대도수를 구할 수 있습니다.

# 도수분포표

몇 개의 범주를 기준으로 둘 것인지에 따라  
**다양한 도수분포표**를 만들 수 있습니다.

# 한 가지 범주의 도수분포표

```
pd.crosstab(index = 범주, columns = "count")
```

index로 설정한 범주에 해당하는 도수를 계산하여  
도수분포표를 제작

# 두 가지 범주의 도수분포표

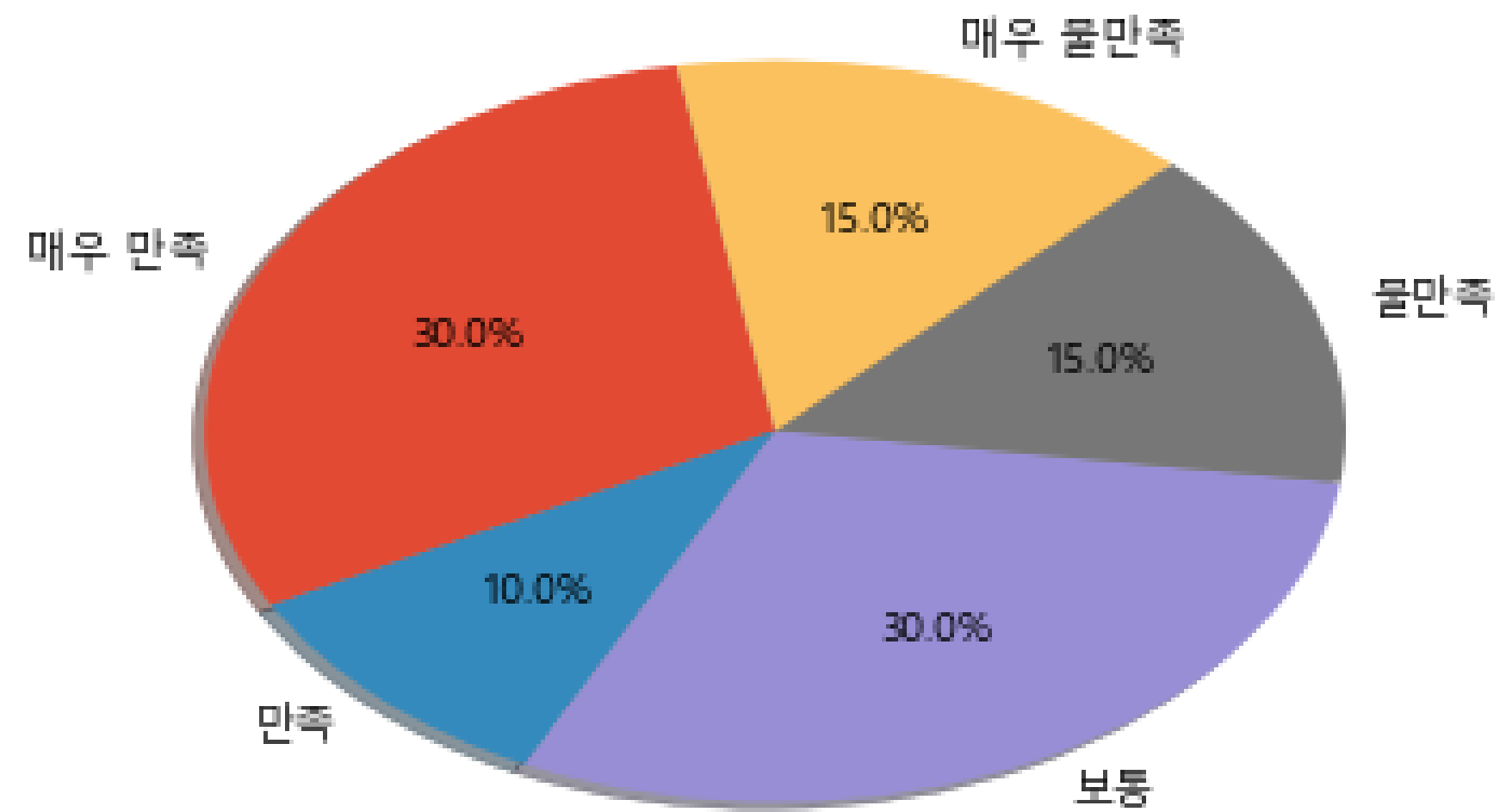
```
pd.crosstab(index = 범주, columns = 또 다른 범주)
```

index로 설정한 범주와 columns로 설정한 범주를 모두 만족하는 도수를 계산하여 **도수분포표**를 제작

# 범주형 자료의 요약: 그래프

# 원형그래프(Pie Chart)

```
plt.pie(수치, labels = 라벨)
```





# 원형그래프(Pie Chart)

**숫자의 나열보다 전체적인 분포를 이해하기 쉬운 그래프**

원을 각 범주가 차지하는 비율로 중심각을 나눠 피자처럼 조각을 나눈 형태의 그림

## 장점

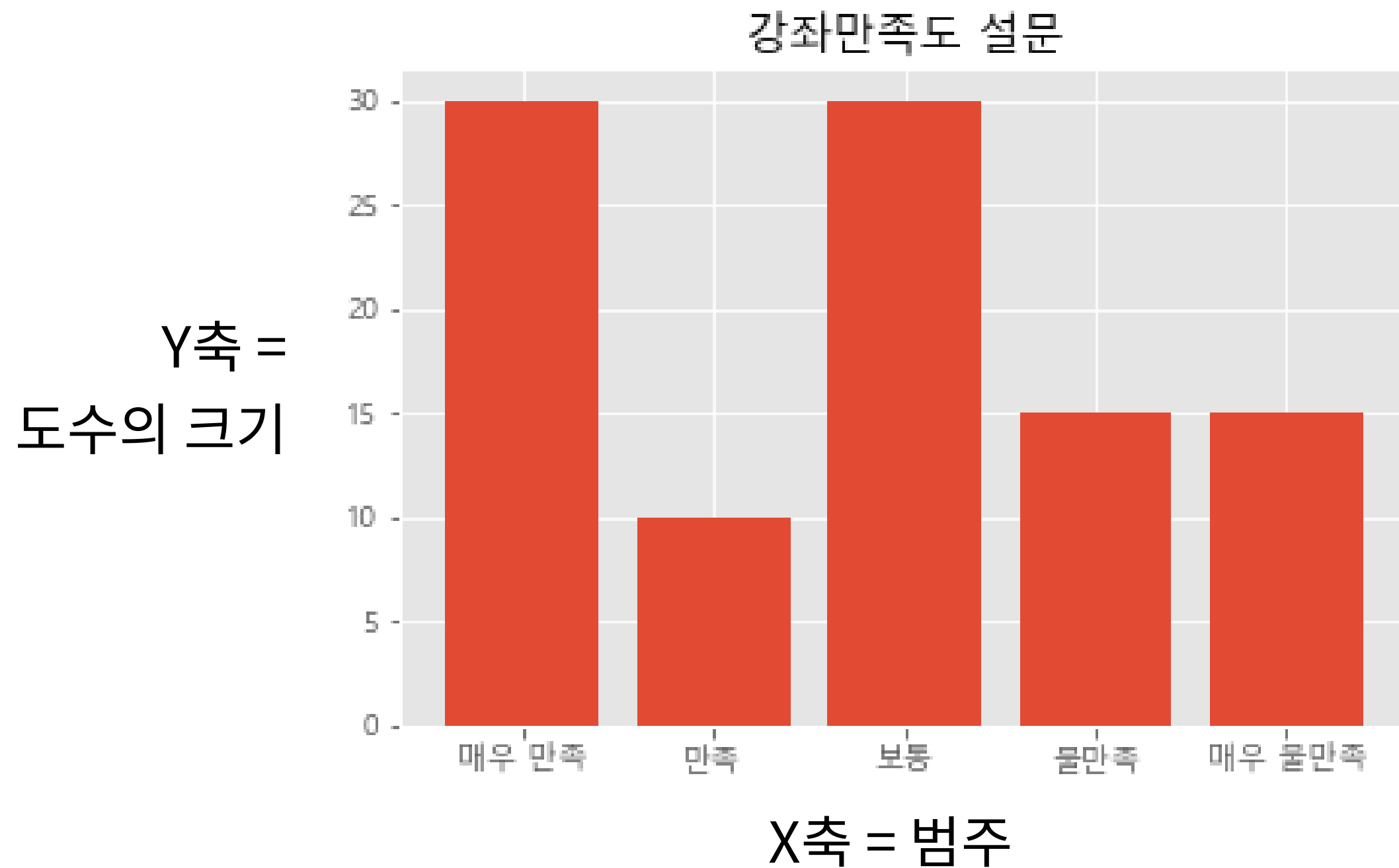
전체에서 범주가 차지하는  
비율을 파악하기 쉬움

## 단점

범주 간 도수 비교 및 도수  
크기 차이 파악이 어려움

# 막대그래프 (Bar Chart)

```
plt.bar(x = 라벨, height = 수치)
```



# 막대그래프 (Bar Chart)

**각 범주에서 도수의 크기를 막대로 그림**

그래프의 Y축 : 도수에 대한 눈금

그래프의 X축 : 범주를 나열

장점

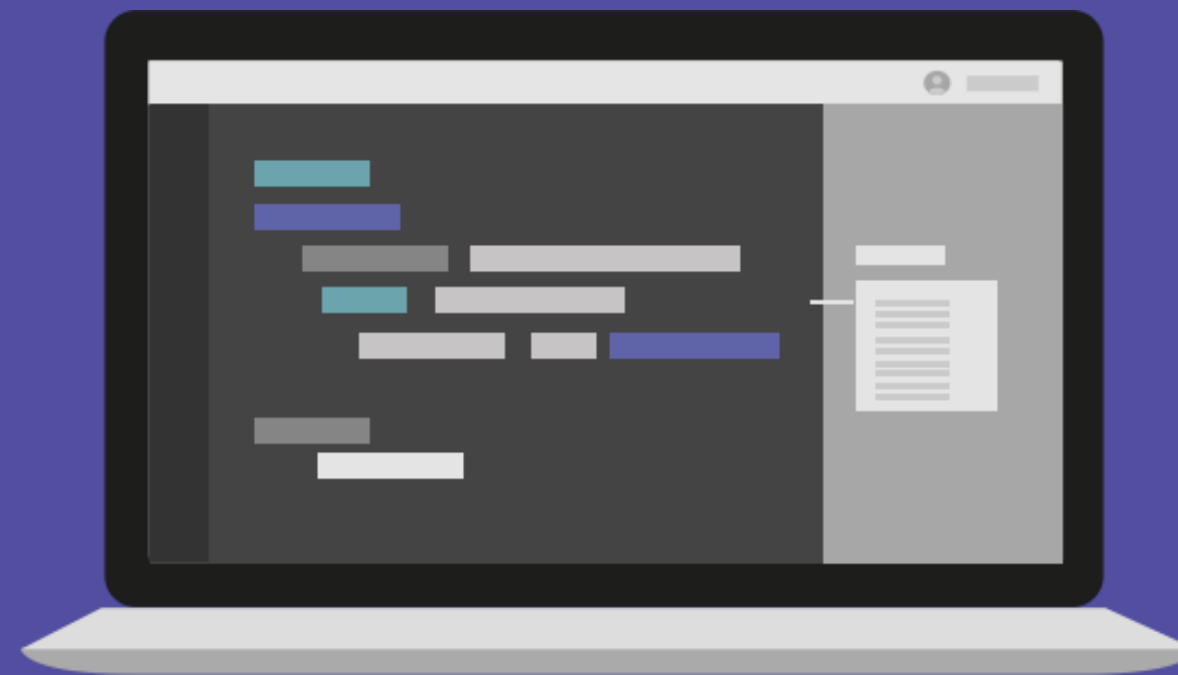
각 범주가 가지는 도수의  
크기 차이를 비교하기 쉬움

단점

각 범주가 차지하는 비율의  
비교는 어려움

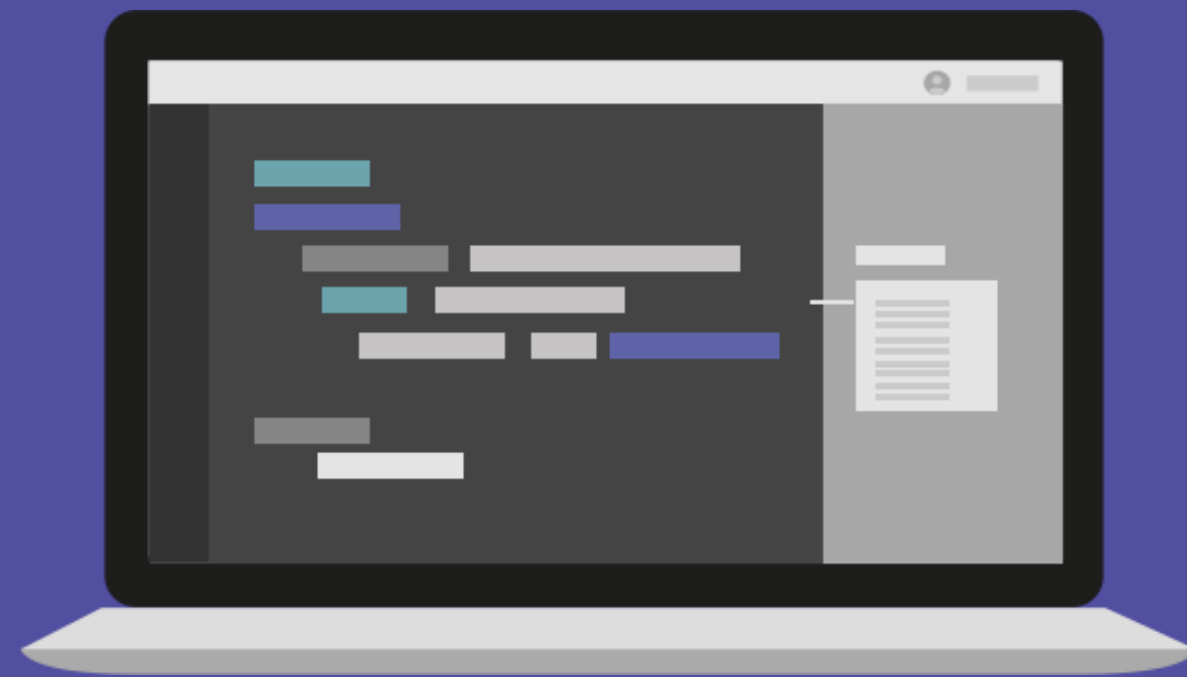
[실습]

범주형 자료 요약 : 도수와 상대도수,  
도수분포표



# [실습]

범주형 자료 요약 : 원형 그래프, 막대 그래프



# 수치형 자료의 요약

# 이산형 자료 요약

관측된 수치 자료가 셀 수 있는 경우 ▶ 이산형 자료 요약

관측값의 종류 수



범주형 자료 요약 기법



연속형 자료 요약 기법

# 연속형 자료 요약

수치 자료가 연속적으로 관측 ▶ 연속형 자료 요약

관측값의 종류 수

↑  
많음

연속형 자료 요약 기법

↓  
적음

점도표, 도수분포표, 히스토그램,  
상대도수다각형, 줄기-잎 그림



# 점도표 (dot diagram)

관측값의 개수가 상대적으로 적은 경우(20 또는 25이하) 사용

- 자료 전체의 개요를 파악 가능
- 모든 자료를 나타낼 수 있도록 줄 위에 각 관측값에 해당되는 점을 찍어 표시

연속형 자료의 경우 **중복된**  
**정보**를 판단하기 어려움



자료를 크기에 따라 묶어서  
분석하는 것이 **효율적**

# 도수분포표(Frequency Table)

- 각 관측값에 대한 도수를 측정하여 도수분포표 작성
- 연속형 자료의 경우 다수의 구간(계급)으로 나누고  
각 구간마다 관측값의 개수(도수)로 작성

계급(Class)	계급구간	계급구간의 폭
모든 연속형 자료의 범위를 나눈 구간	각 계급에 포함되는 값의 범위	계급구간의 크기

# 도수분포표 작성 순서

## 1. 자료의 범위

자료에서 최댓값과 최솟값을 찾아 자료의 범위를 구함

## 2. 계급의 폭

계급의 개수를 분포의 경향이 잘 드러날 수 있도록 정함

## 3. 계급구간

모든 관측 값을 포함하도록 계급구간의 경계점을 구함

## 4. 도수

각 계급구간에 속하는 관측값의 개수를 세어 계급의 도수를 더함

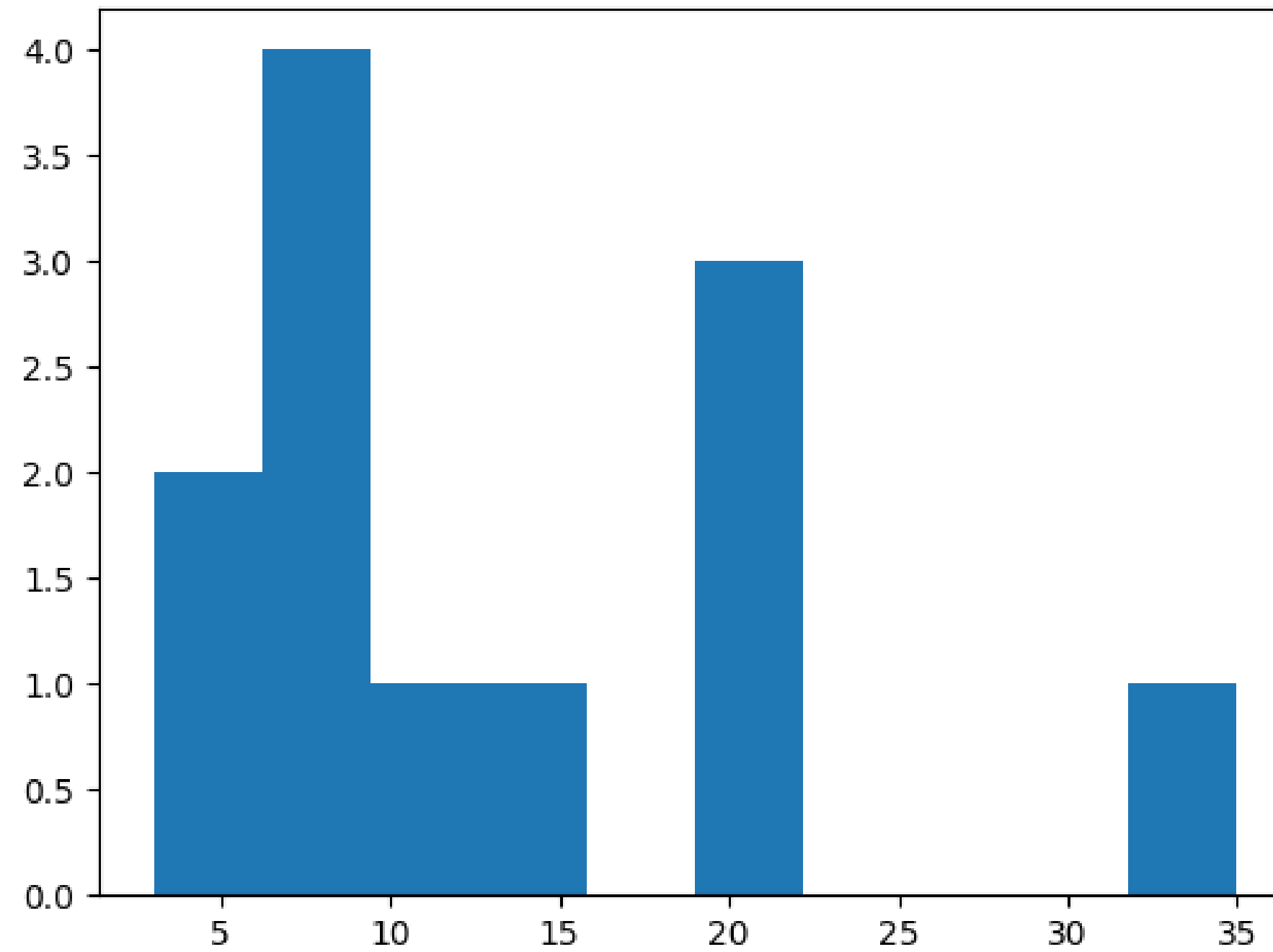
## 5. 상대도수

각 계급의 도수를 전체 관측값의 개수로 나눠 계급의 상대도수를 구함

# 수치형 자료의 요약: 그래프

# 히스토그램(Histogram)

Y 축 : 빈도



X 축 : 계급

# 히스토그램(Histogram)

```
plt.hist()
```

연속형 자료의 도수분포표를 기반으로 각 계급을 범주처럼 사용  
범주형 자료의 **막대그래프**와 같은 방식으로 그림

도수 비교



범주

막대그래프

연속

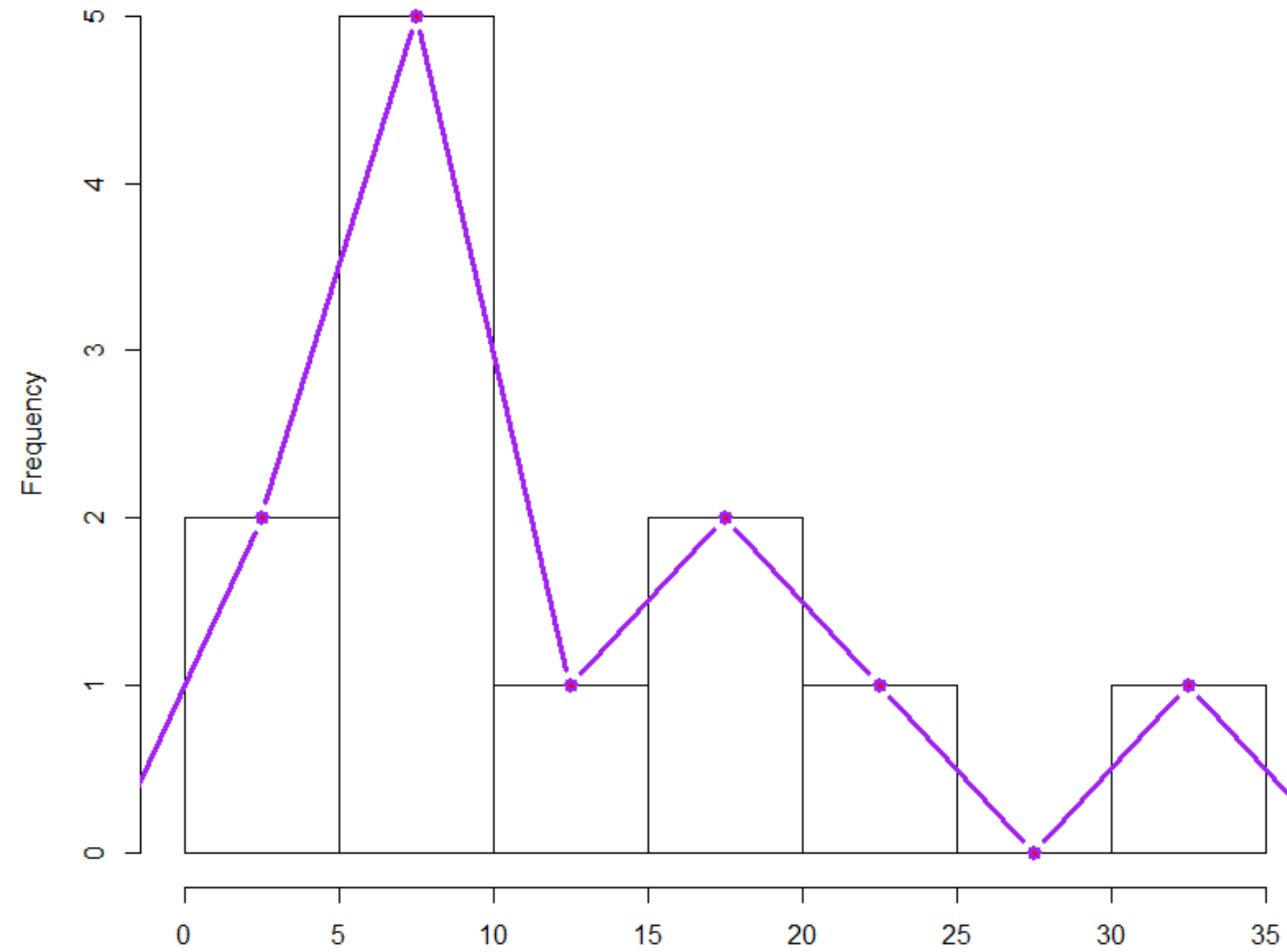
히스토그램

# 히스토그램의 특징

- 자료의 분포를 알 수 있음
- 계급구간과 막대의 높이로 그림
- 모든 계급구간의 폭이 같으면  
도수, 상대도수를 막대 높이로 사용

# 도수다각형

Y 축 : 빈도



X 축 : 구간



# 도수다각형의 특징

- 각 계급구간의 중앙에 점을 찍어 직선으로 연결함
- 관측값의 집중된 위치, 정도, 치우친 정도, 꼬리의 두터움 등 분포의 상태를 쉽게 파악
- 관측값의 변화에 따라 도수 또는 상대도수의 변화를 잘 나타냄

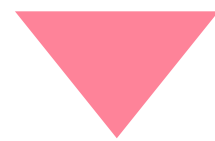
# 도수다각형과 히스토그램

## 히스토그램

옆으로 나열하여  
자료 비교

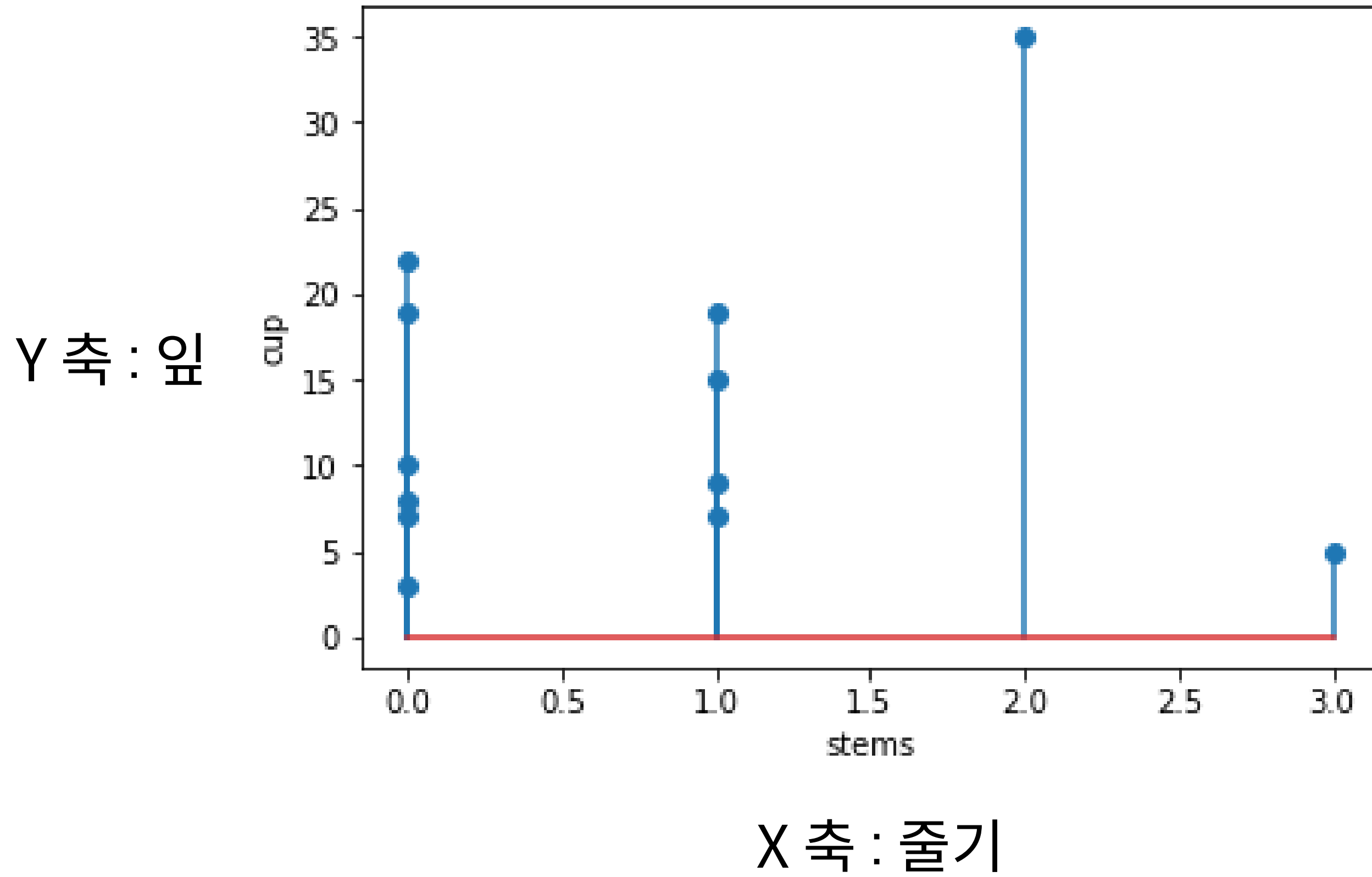
## 도수다각형

꺾은선으로 표시하여  
자료 비교



여러 자료를 비교하기 위해서는  
히스토그램보다 도수다각형이 알맞음

# 줄기-잎 그림



# 줄기-잎 그림 자료 정리 방법

예) 관측값의 자릿수가 최대 2자리 일때



1. 관측값을 보고 앞 단위와 뒷 단위를 정함
2. 앞 단위를 줄기로 하여 세로로 배열하고 수직선을 그림
3. 뒷 단위를 잎으로 하여 관측값을 앞 단위 오른쪽에 오름차순 기입

# 줄기-잎 그림 함수

```
plt.stem(줄기, 관측 값)
```

자료의 분포를 시각적으로 쉽게 파악

각 관측값도 유지 가능

함수 사용 시에 줄기 값을 따로 지정해줘야 함  
: 줄기를 데이터마다 다르게 설정할 수 있기 때문

# 줄기-잎 그림 장단점

## 장점

- 관측값을 보여주므로 최댓값, 최솟값 등의 위치 파악 쉬움
- 순서대로 배열된 관측값의 장점과 히스토그램의 장점을 모두 가지고 있음
- 그리기 쉬움

## 단점

- 관측값의 개수가 많은 경우 제한된 공간에 그리기 불가능
- 관측값이 지나치게 흩어져 있으면 부적절

[실습]

연속형 자료 요약 : 도수분포표,  
줄기-잎 그림, 히스토그램

