14. 데이터분석 기초(2)

14.1 파이썬을 이용한 기술 통계

기술통계란?

주어진 자료를 수집 및 정리, 요약해서 정보를 획득하는 통계학

자료는 그래프, 통계량으로 표현할 수 있다.

14.1.1 통계량 구하기

통계량 계산하기(1차원 데이터)

```
list1 = [
82,39,90,40,20,89,69,79,80,98,
79,88,38,58,68,80,78,73,89,93,
52,74,77,78,87,99,85,79,77,100,
85,49,68,74,80,92,88,49,80,64
]
```

위와 같은 데이터가 있다. 이 데이터의 통계량을 계산해본다.

```
# 1. 표본 개수 세기
import numpy as np

data = np.array(list1)
print(len(data)) # 40
```

```
# 2. 표본 평균 구하기
print(np.mean(data)) #74.175
```

```
# 3. 분산과 표본 분산 구하기
# ddof는 자유도이다. n개의 데이터면 n-1로 나눈다.
print(np.var(data,ddof = 1)) #338.0455128205128
print(np.std(data,ddof = 1)) #18.386014054724118
```

```
# 4. 중앙값 구하기
print(np.median(data)) #79.0

print(np.median(data[0:38]))
```

짝수이던 홀수이던 float로 나온다.

```
# 5. 백분위수와 사분위수 구하기
# 데이터를 작은 수부터 오름차순으로 정렬했을 때, 아래에서부터 해당하는 수
print(np.percentile(data, 25)) # 1분위수, 68.0
print(np.percentile(data, 50)) # 2분위수, 중앙값, 79.0
print(np.percentile(data, 75)) # 3분위수, 87.25
```

```
# 6. 최대값과 최소값
print(np.max(data)) #100
print(np.min(data)) #20
```

통계량 구하기 (2차원 데이터)

```
import pandas as pd
url = "https://raw.githubusercontent.com/sesillim/ai/main/New_Fish.csv"
data = pd.read_csv(url)
```

url에서 불러온 pandas dataframe의 통계량을 구해본다.

```
data.describe()

'''
    Weight  Length1 Length2 Length3 Height  Width
count   159.000000  159.000000  159.000000  159.000000  159.000000  159.000000
mean    398.326415  26.247170   28.415723   31.227044   8.970994    4.417486
std 357.978317  9.996441    10.716328   11.610246   4.286208    1.685804
min 0.000000    7.500000    8.400000    8.800000    1.728400    1.047600
25% 120.000000  19.050000   21.000000   23.150000   5.944800    3.385650
50% 273.000000  25.200000   27.300000   29.400000   7.786000    4.248500
75% 650.000000  32.700000   35.500000   39.650000   12.365900   5.584500
max 1650.000000 59.000000   63.400000   68.000000   18.957000   8.142000
'''
```

data.describe()를 하면 각 칼럼들의 count, mean, std, min, 1분위수, 2분위수, 3분위수, 최대값을 알 수 있다.

두 변수의 공분산 : 두 변수의 관련성을 나타내는 척도

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

분산-공분산 행렬 : 대각선은 해당되는 변수의 분산(같은 칼럼끼리의 공분산은 분산), 나머지는 공분산이다.

```
data.cov(ddof=1)

'''
    Weight  Length1 Length2 Length3 Height  Width
Weight  128148.475121   3276.882797 3524.013253 3836.368648 1111.413300
534.990098
Length1 3276.882797 99.928837   107.073431  115.136248  26.795457   14.611556
Length2 3524.013253 107.073431  114.839688  123.685458  29.416988   15.781169
Length3 3836.368648 115.136248  123.685458  134.797808  35.004389   17.194921
Height  1111.413300 26.795457   29.416988   35.004389   18.371576   5.729125
Width   534.990098  14.611556   15.781169   17.194921   5.729125    2.841935
'''
```

상관계수 행렬 구하기

피어슨 상관계수 : 공분산 값을 최대가 1, 최소가 -1이 되도록 표준화한 값
$\mu_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

상관계수 행렬 : 대각선은 1, 나머지는 상관계수

```
data.corr(method = 'pearson')

'''
    Weight  Length1 Length2 Length3 Height  Width
Weight  1.000000    0.915712    0.918618    0.923044    0.724345    0.886507
Length1 0.915712    1.000000    0.999517    0.992031    0.625378    0.867050
Length2 0.918618    0.999517    1.000000    0.994103    0.640441    0.873547
Length3 0.923044    0.992031    0.994103    1.000000    0.703409    0.878520
Height  0.724345    0.625378    0.640441    0.703409    1.000000    0.792881
Width   0.886507    0.867050    0.873547    0.878520    0.792881    1.000000
'''
```

14.1.2 그래프 그리기

히스토그램

히스토그램은 자료를 몇 개의 계급으로 나누어 각 계급에 속하는 빈도수를 막대 모양의 그래프로 나타낸 것.

1차원 배열 만들기

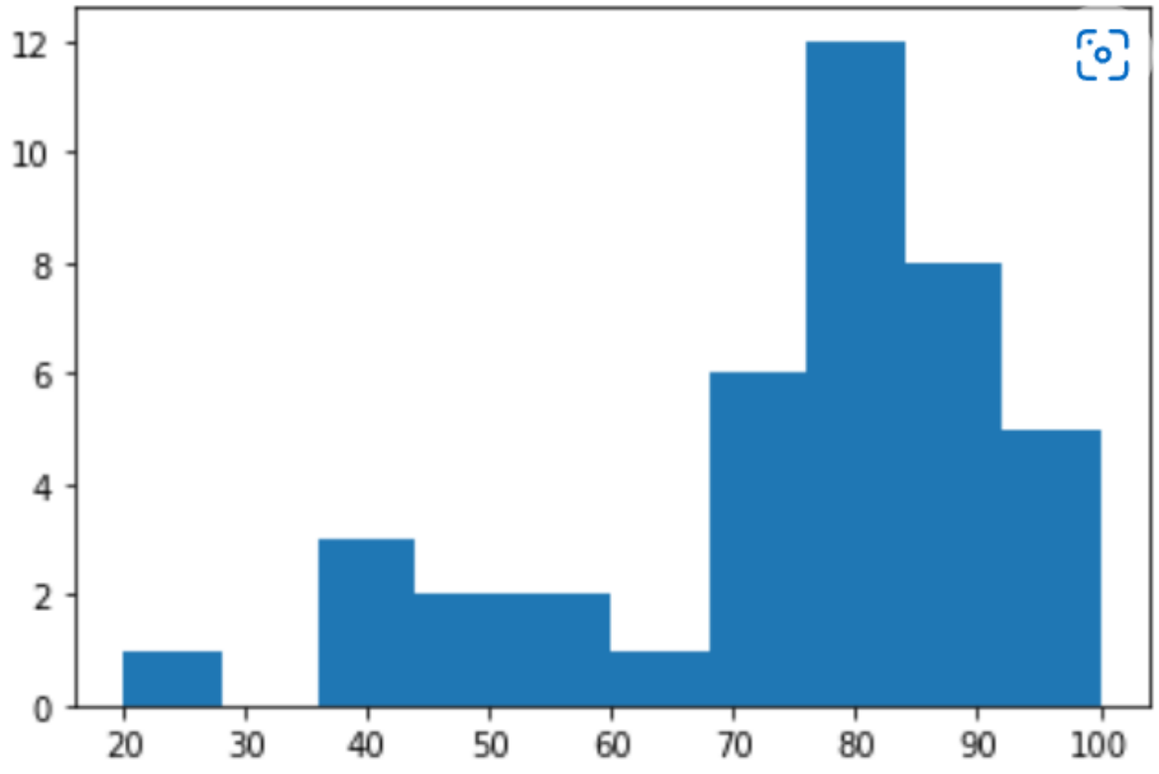
```
import matplotlib.pyplot as plt
%matplotlib inline

list1 = [
82,39,90,40,20,89,69,79,80,98,
79,88,38,58,68,80,78,73,89,93,
52,74,77,78,87,99,85,79,77,100,
85,49,68,74,80,92,88,49,80,64
]
```

기본 히스토그램 만들기

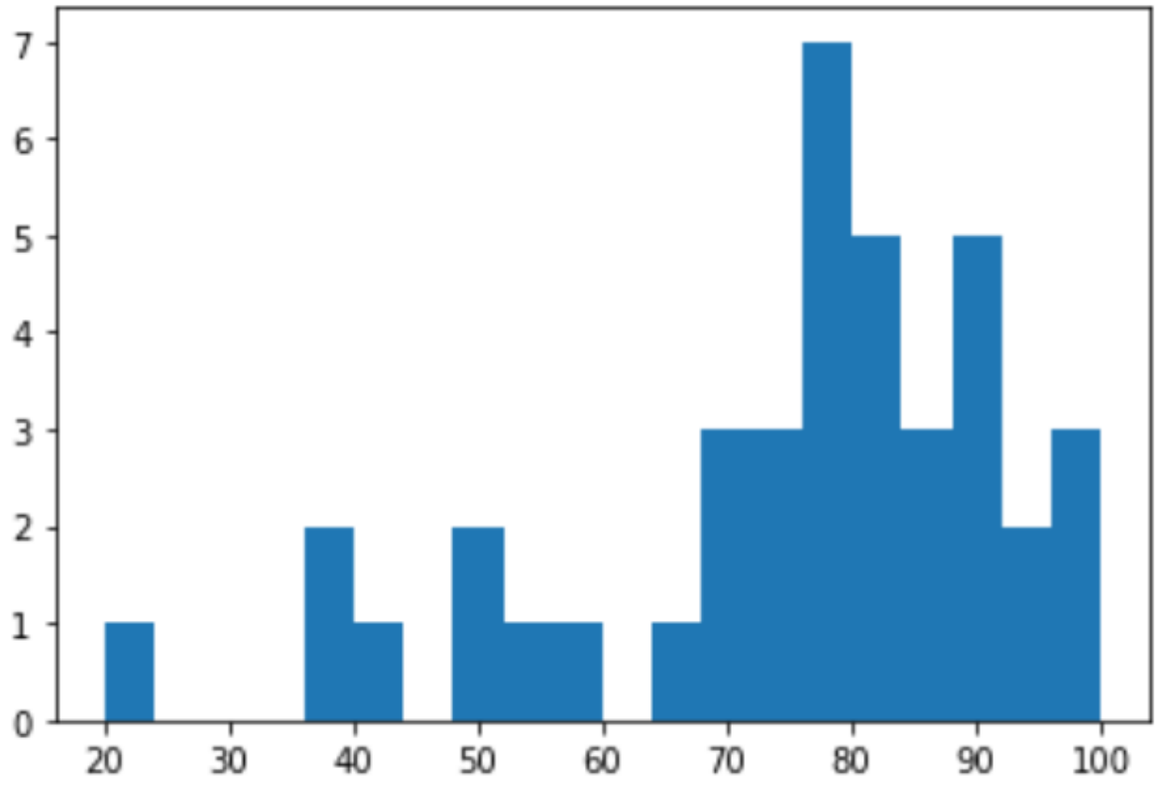
```
data = np.array(list1)

plt.hist(data)
plt.show()
```

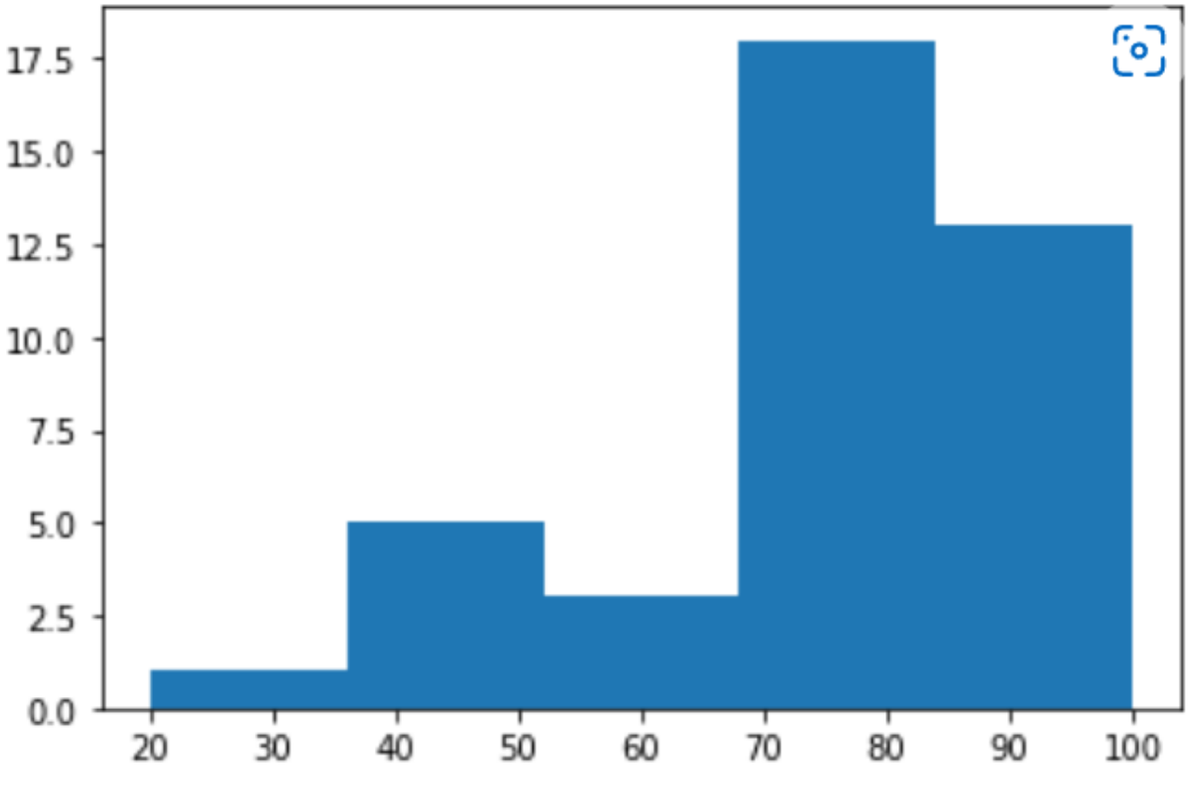


bins로 히스토그램의 계급 수 조절하기

```
plt.hist(data, bins = 20)
plt.show()
```

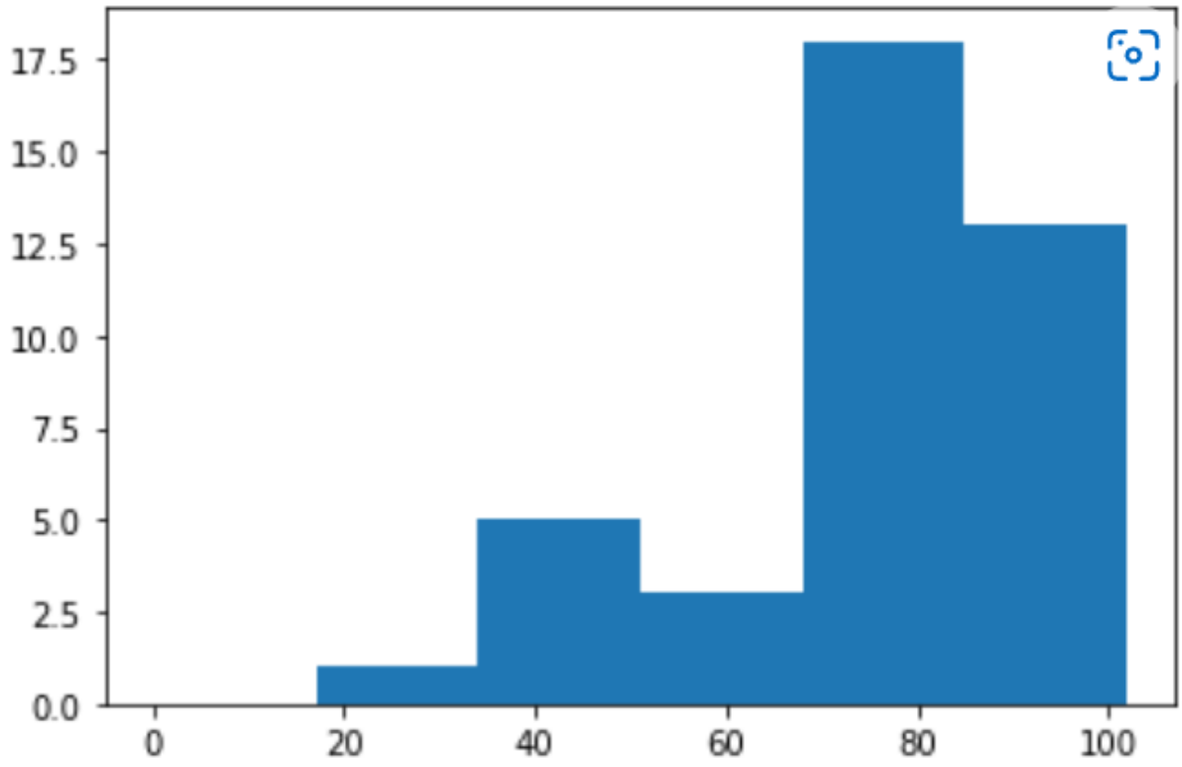



```
plt.hist(data, bins = 5)
plt.show()
```



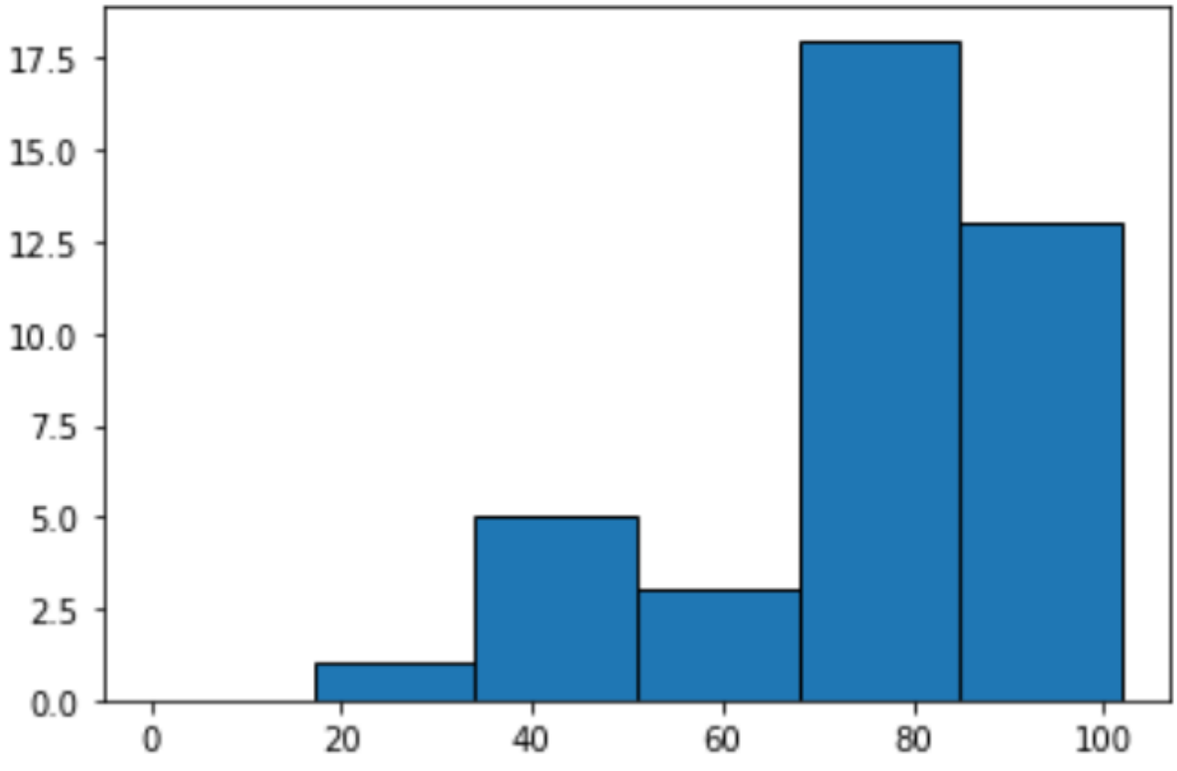
bins로 계급 구간 직접 설정하기

```
plt.hist(data, bins = [0,17,34,51,68,85,102])
plt.show()
```

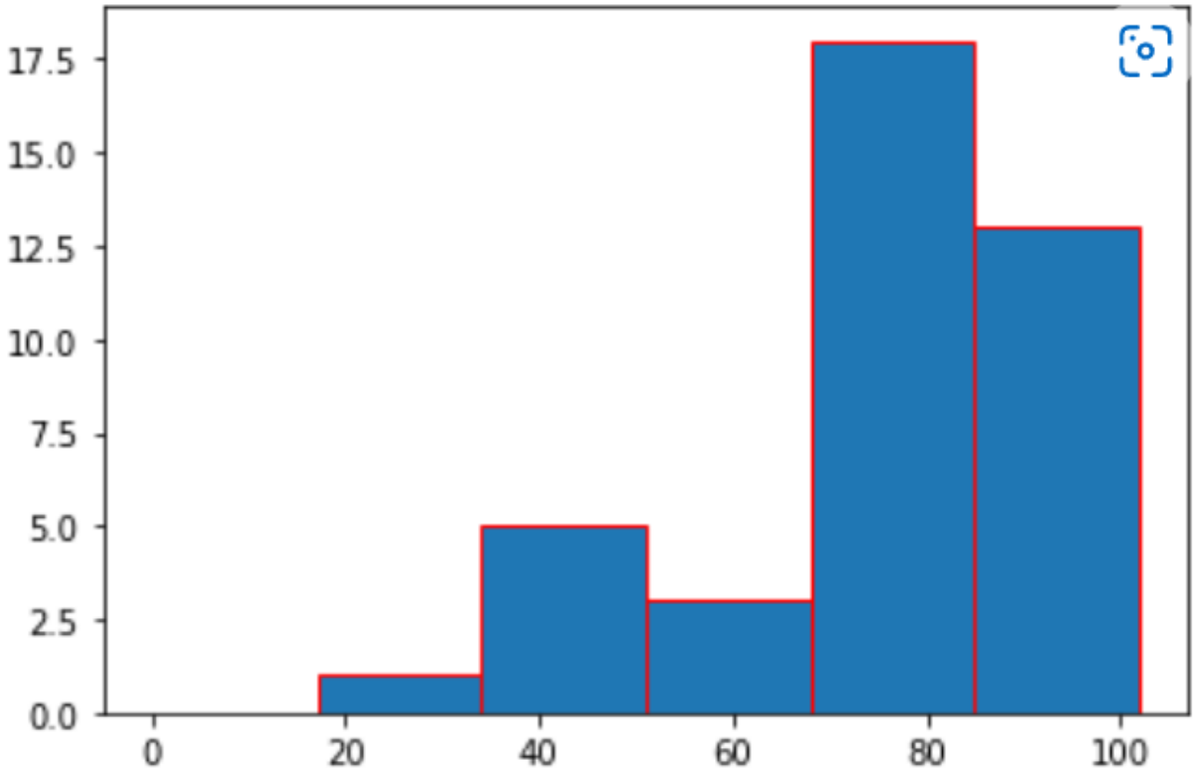


edgecolor로 구분짓는 선의 색 지정

```
plt.hist(data, bins = [0,17,34,51,68,85,102], edgecolor = 'black')
plt.show()
```




```
plt.hist(data, bins = [0,17,34,51,68,85,102], edgecolor = 'red')
plt.show()
```

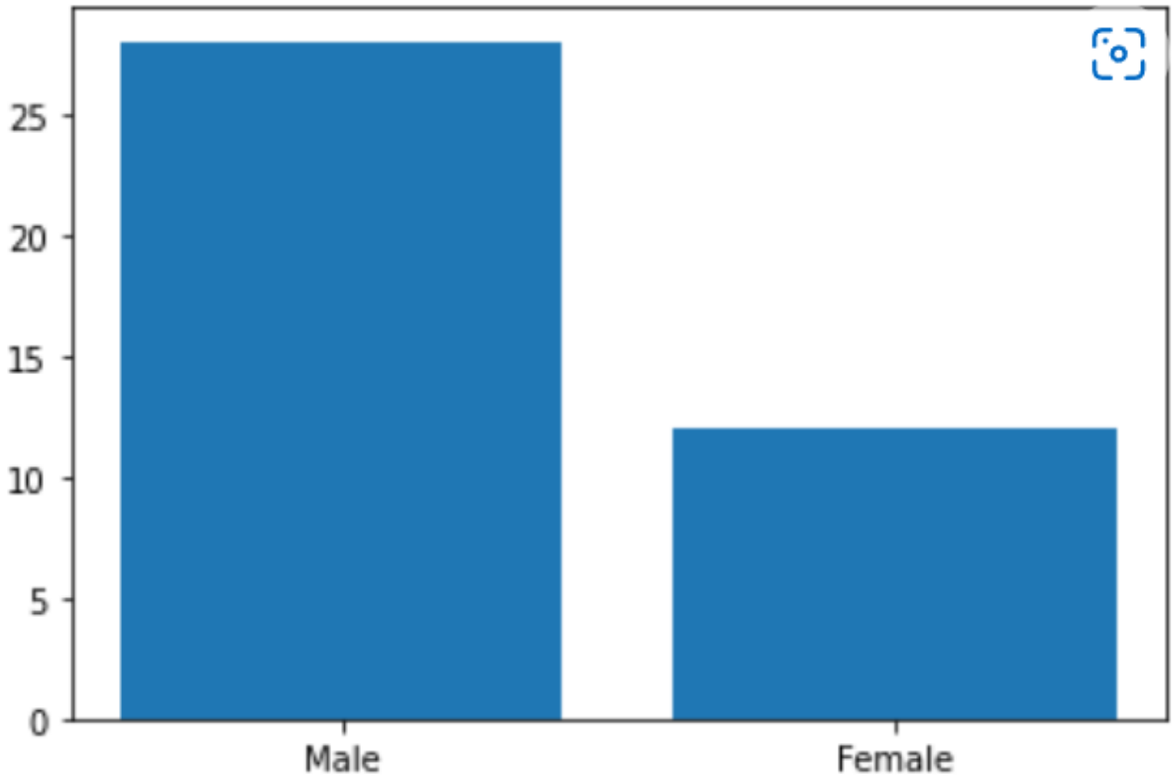


막대그래프 그리기

막대그래프는 이산형 자료에서 빈도수를 세어 표현할 때 사용한다.

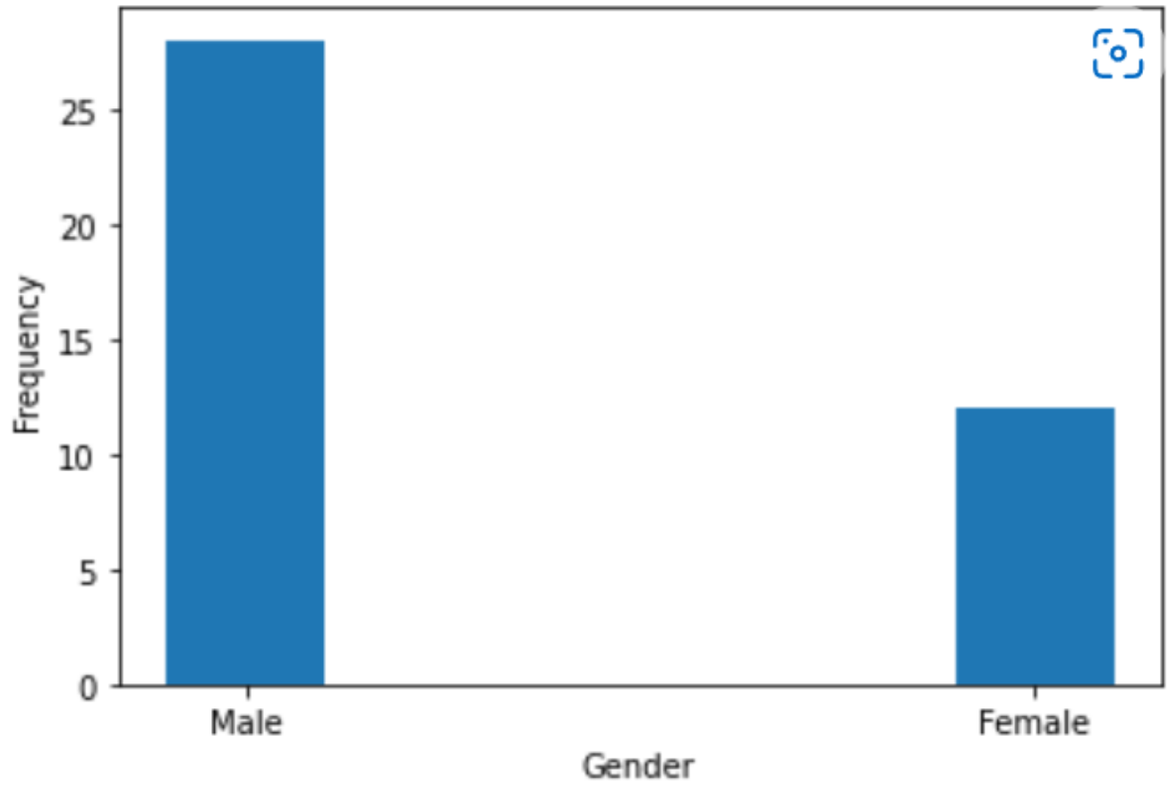
성별	빈도수
남성	28
여성	12

```
gender = ["Male", "Female"]
frequency = [28,12]
# 둘다 리스트로 저장
plt.bar(gender, frequency)
plt.show()
```

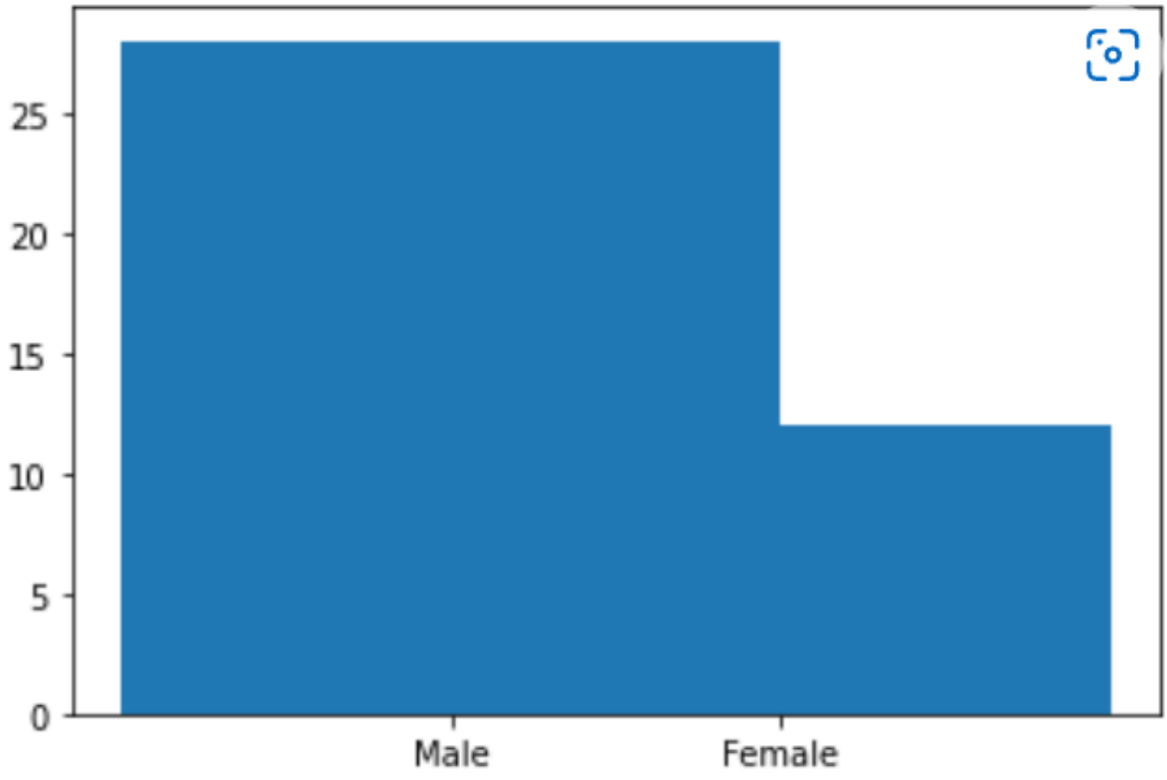


width를 써서 막대그래프 넓이 조절하기

```
plt.bar(gender, frequency, width = 0.2)
plt.xlabel("Gender")
plt.ylabel("Frequency")
```




```
plt.bar(gender, frequency, width = 2.0)
plt.show()
```



파이차트 그리기

파이차트는 전체 데이터 중 특정 데이터의 비율을 보기 쉽게 표현할 때 사용

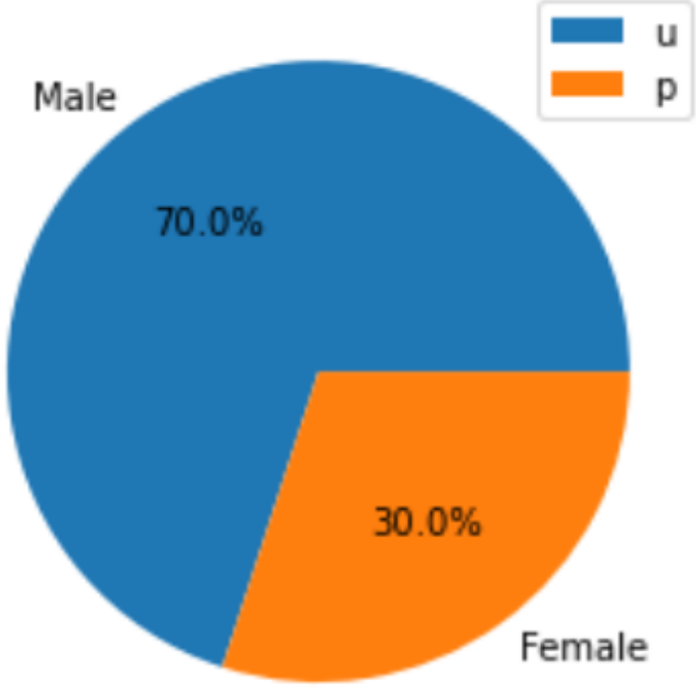
성별	빈도수
남성	28
여성	12

자료의 입력 및 기본적인 파이차트 그리기

```
data = [28,12]
plt.pie(data)
plt.show()
```

labels로 레이블 추가,autopct로 비율 추가, legend()로 범례 추가

```
data=[28,12]
label = ['Male','Female']
plt.pie(data, labels = label, autopct = '%.1f%%')
plt.legend('upper right')
plt.show()
```

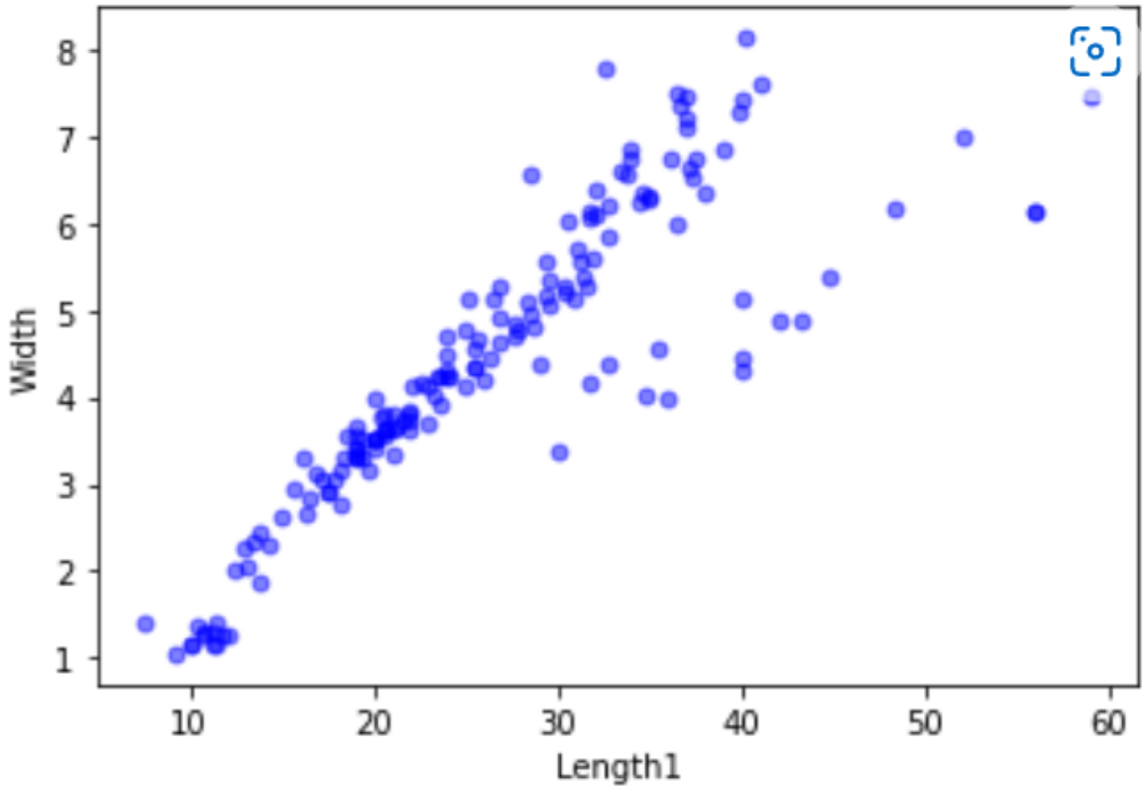


산점도

산점도는 두 변수 간의 관계를 도표 상의 점으로 표시한다.

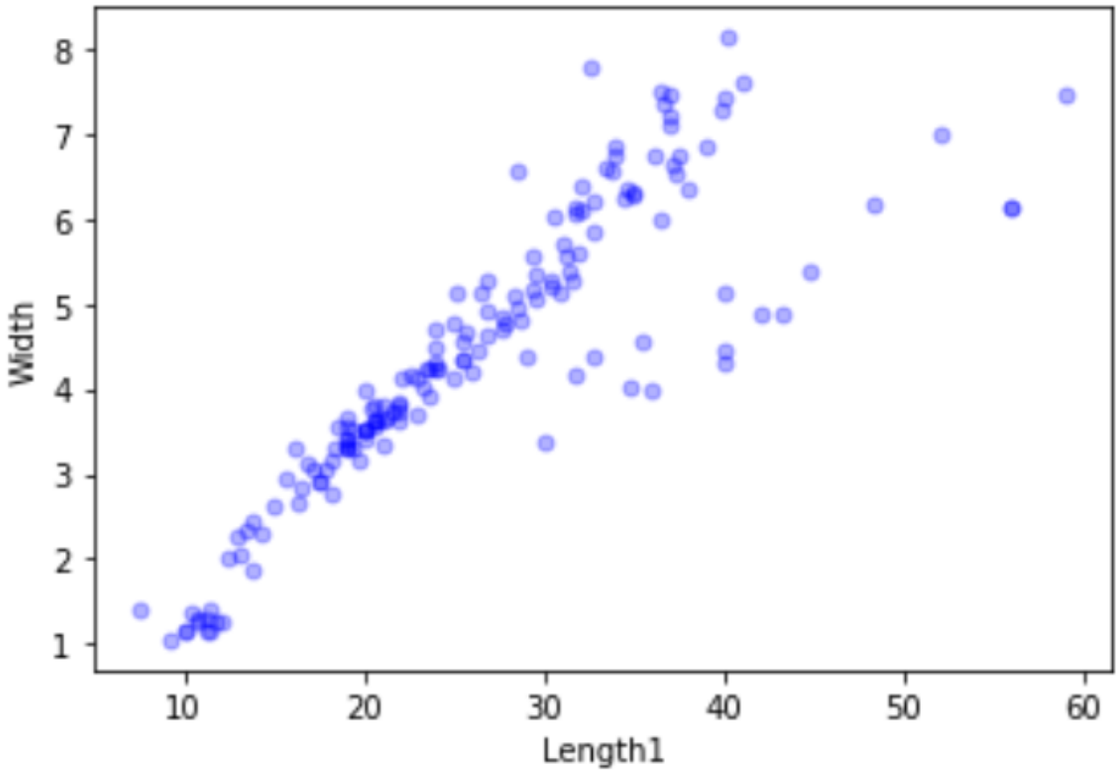
matplotlib으로 산점도 그리기

```
url = "https://raw.githubusercontent.com/sesillim/ai/main/New_Fish.csv"
data = pd.read_csv(url)
plt.plot('Length1', 'Width', data = data, ls = 'none', marker = 'o', markersize
= 5, color = 'blue', alpha = 0.5)
plt.xlabel('Length1')
plt.ylabel('Width')
plt.show()
```



pandas로 산점도 그리기

```
url = "https://raw.githubusercontent.com/sesillim/ai/main/New_Fish.csv"
data = pd.read_csv(url)
data.plot.scatter(x = 'Length1', y = 'Width', s = 25, c = 'blue', alpha = 0.3)
plt.xlabel('Length1')
plt.ylabel('Width')
plt.show()
```



14.2 파이썬 데이터 분석 패키지

14.2.1 통계적 가설 검정

표본 자료에서 얻은 표본 통계량을 바탕으로 모집단의 특성을 추측한 주장 및 가설을 기각할지 말지를 판정하는 행위

- 가설을 참이라고 가정했을 때, 표본의 검정통계량이 기각역에 들어온다면 가설을 기각함.
- p값이 0.05 이하로 매우 작으면 가설을 기각함

파이썬 scipy패키지에는 다양한 검정통계량과 관련된 가설을 검정할 수 있는 함수를 제공

- 이항검정, 카이제곱검정, 단일표본 z검정, 단일표본 t검정, 독립표본 t검정