/* elice */

파이썬으로 배우는 기초 통계

논리적인 자료의 요약

목차

1. 중심위치의 측도

수치를 통한 연속형 자료 요약

그림이나 도표에 의한 분석의 단점

- 작성자의 주관적 판단에 따라 달라지므로 일관성 및 객관성이 부족
- 시각적 자료로는 이론적 근거 제시가 쉽지 않음

많은 양의 자료를 의미 있는 수치로 요약하여
대략적인 분포상태를 파악 가능하므로 단점 보완 가능

수치를 통한 연속형 자료 요약

1) 중심위치의 측도
(measure of center)

자료의 중심위치를 나타냄

2) 퍼진 정도의 측도
(measure of dispersion)

자료가 각 중심위치로부터
얼마나 흩어져 있는지 나타냄

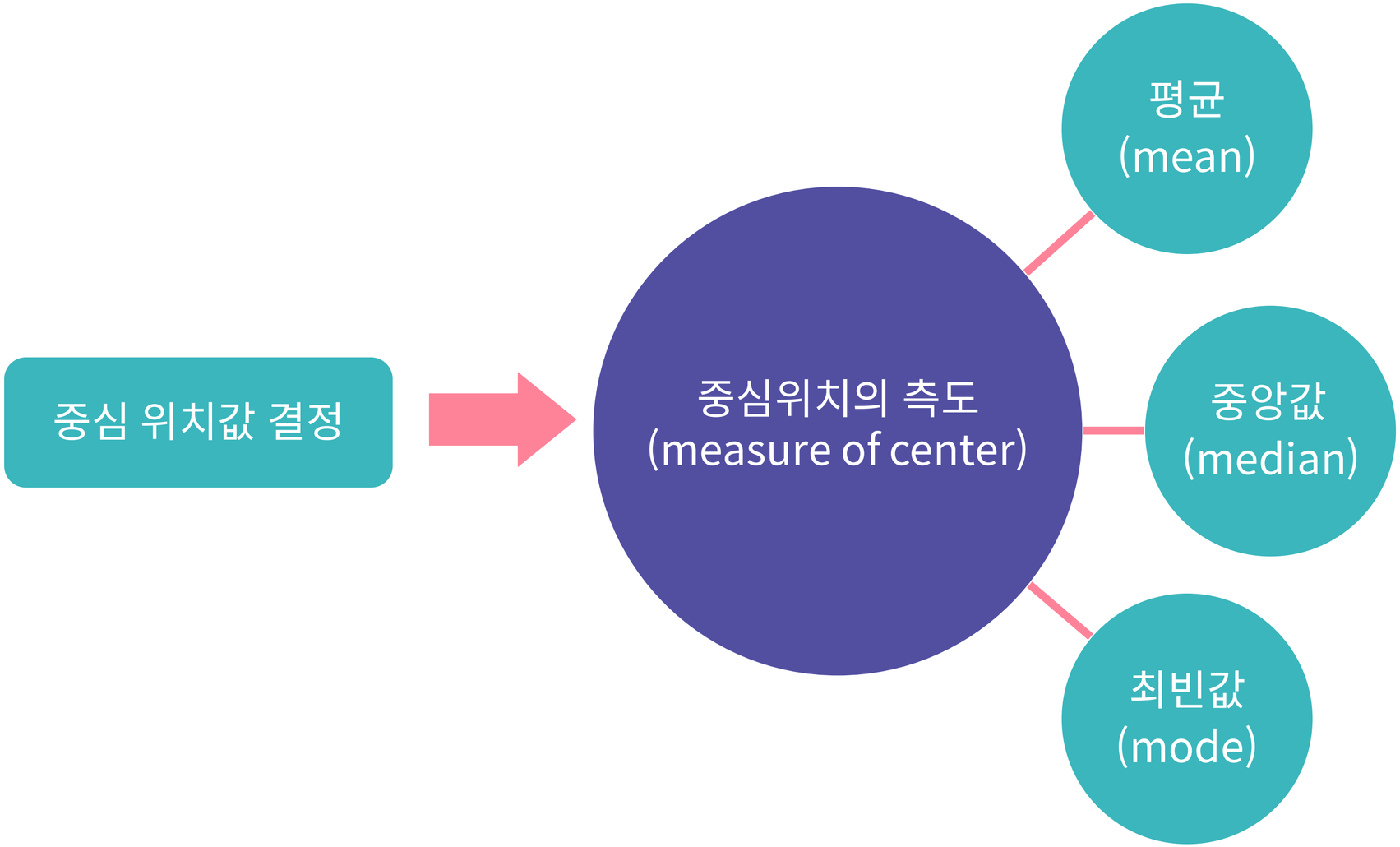
3) 도수분포표에서의 자료의 요약

자료가 이미 그룹화된 경우의 수치 요약 방법

4) 상자 그림

사분위수, 최소값, 최대값 등을 이용한 요약 방법

중심위치의 측도


중심 위치값 결정
중심위치의 측도
(measure of center)
평균
(mean)
중앙값
(median)
최빈값
(mode)

평균(Mean)

```
np.mean()
```

중심위치의 측도 중에서 가장 많이 사용되는 방법

모든 관측값의 합을 자료의 개수로 나눈 것

관측값들의 무게중심

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

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

평균의 특징

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

중앙값(Median)

```
np.median()
```

전체 관측값을 정렬했을 때 가운데에 위치하는 값

자료의 개수(n)가 홀수인 경우

$\frac{(n+1)}{2}$ 번째 관측값

자료의 개수(n)가 짝수인 경우

$\frac{n}{2}$ 번째 관측값과 $\frac{n}{2}+1$ 번째 관측값의 평균

중앙값의 특징

- 관측값을 크기 순서대로 배열할 때 중앙에 위치
- 가운데에 위치한 값 이외의 값의 크기는 중요하지 않음
- 관측값의 변화에 민감하지 않고, 극단값의 영향을 받지 않음

최빈값(Mode)

```
stats.mode()
```

관측값 중 가장 자주 나오는 값

이산형/범주형 자료에서 많이 사용

최빈값의 특징

- 연속형 자료에서 같은 값이 나오는 경우는 흔치 않으므로 최빈값을 사용하기 부적절
- 단봉형 분포를 갖는 자료에서만 유용

평균, 중앙값, 최빈값의 비교

실제 사용 빈도

평균 > 중앙값 > 최빈값

평균

- 이해하기 쉽고 통계적으로 가장 많이 사용
- 관측값이 골고루 반영
- 극단값으로 인한 영향을 많이 받음

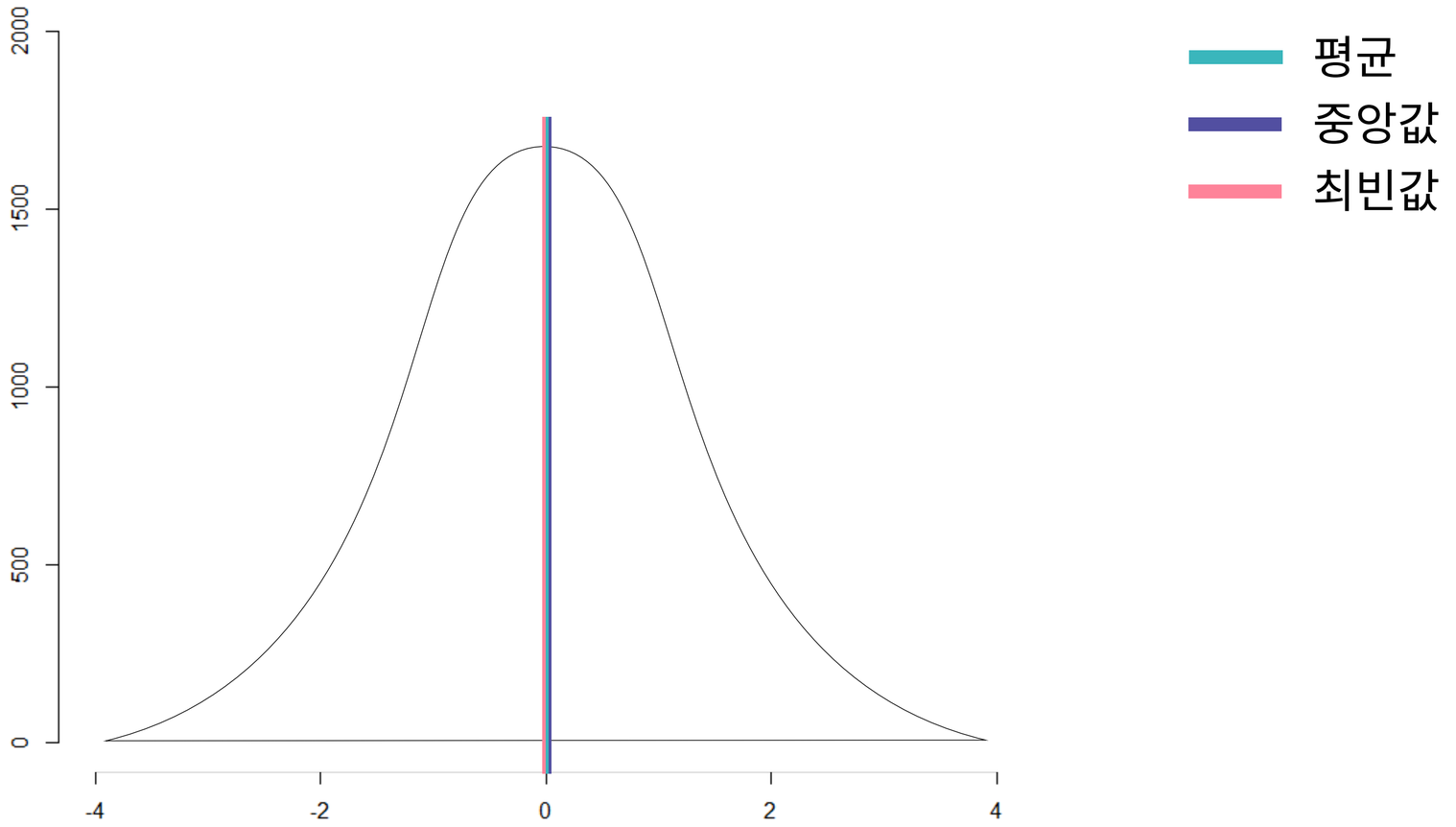
중앙값

- 중앙 부분 외 관측값의 변화에 민감하지 않음
- 극단값으로 인한 영향을 받지 않음

극단값이 있는 경우

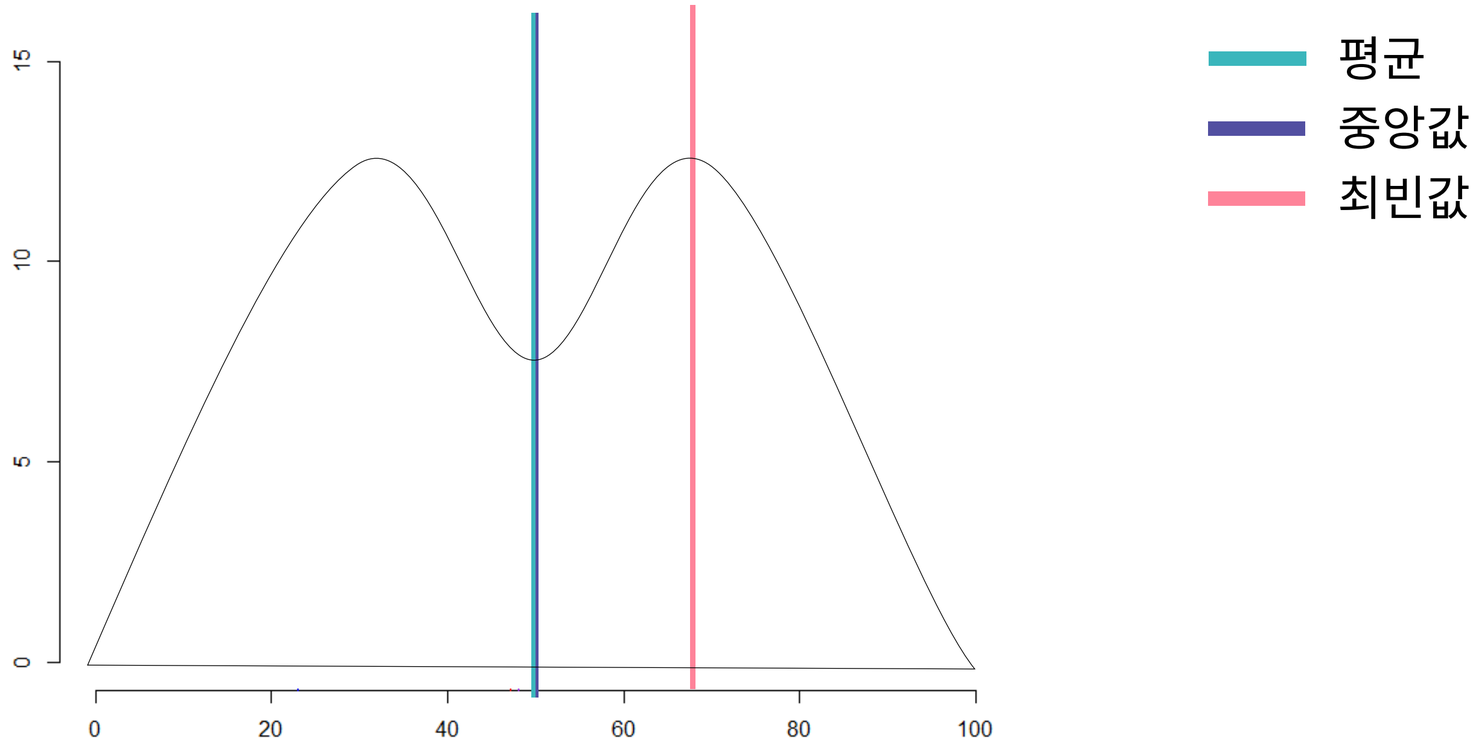
- 극단값의 영향을 배제하고 싶으면 중앙값 사용
- 전체 관측값을 모두 포함하고 싶으면 평균 사용

평균, 중앙값, 최빈값의 비교: 단봉형 대칭



평균 = 중앙값 = 최빈값

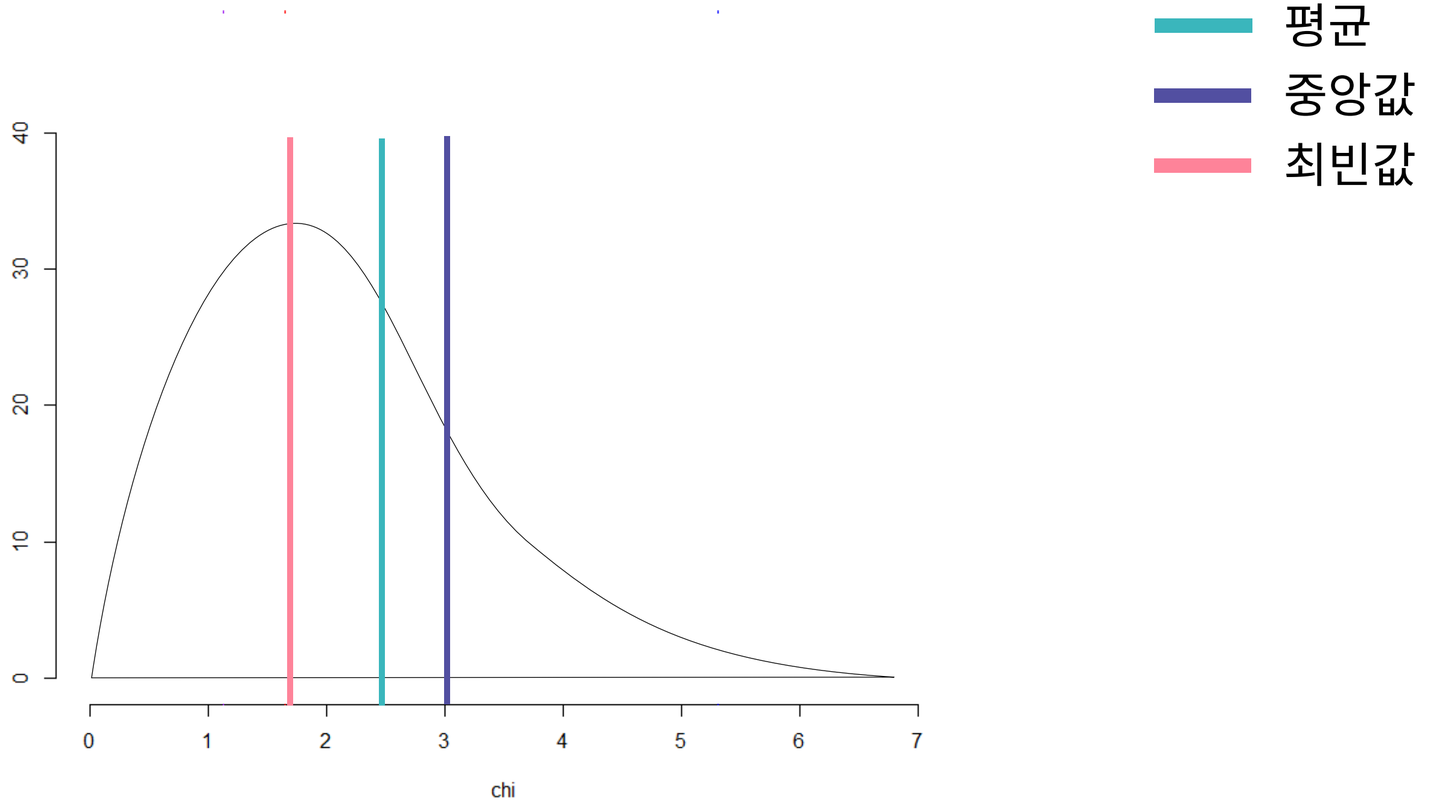
평균, 중앙값, 최빈값의 비교: 이봉형 대칭



평균 = 중앙값 ≠ 최빈값

※다봉형 분포에서 최빈값은 중심위치의 측도로 부적합

평균, 중앙값, 최빈값의 비교: 비대칭 분포



평균 ≠ 중앙값 ≠ 최빈값

비대칭 분포에서 평균과 중앙값