2. 퍼진 정도의 측도

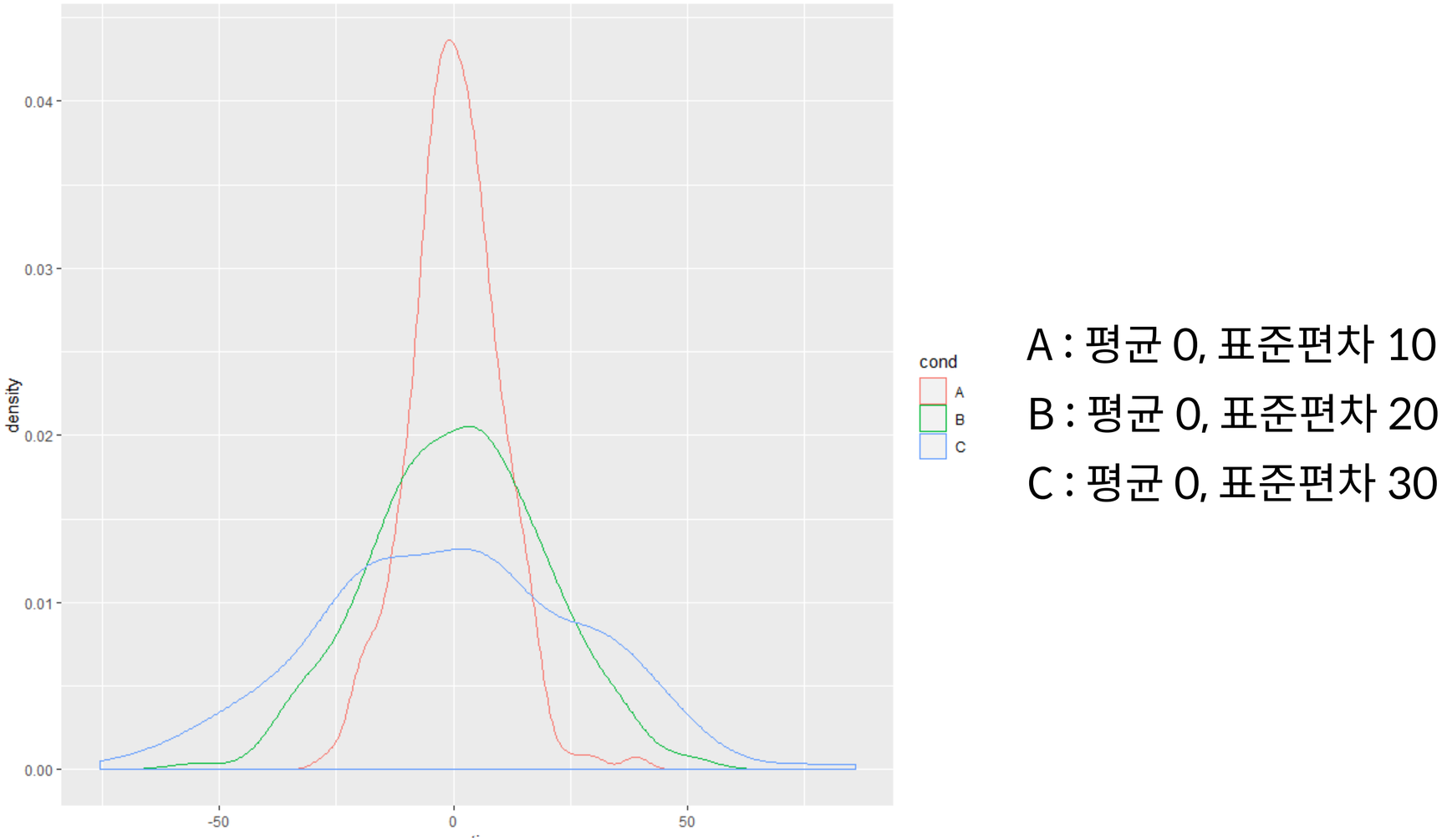
퍼진 정도의 측도

중심위치만으로 분포를 파악하기에 부족

중심위치 측도 외에 분포가 퍼진 정도를 측도할 수치가 필요

분산, 표준편차, 범위, 사분위수 등을
퍼진 정도의 측도로 사용

퍼진 정도의 측도



A : 평균 0, 표준편차 10

B : 평균 0, 표준편차 20

C : 평균 0, 표준편차 30

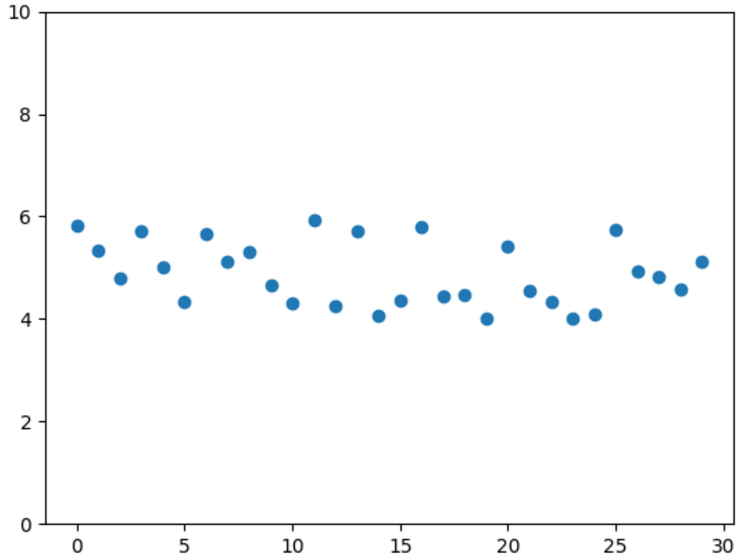
분산

```
variance()
```

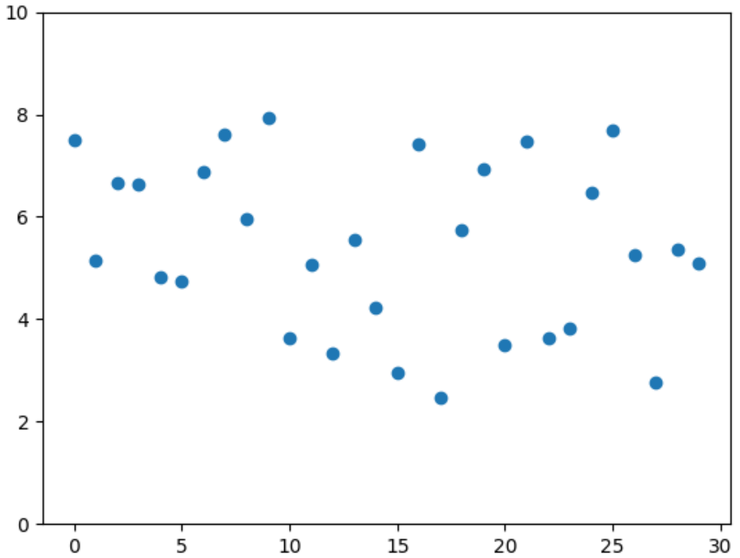
자료가 얼마나 흩어졌는지 숫자로 표현

각 관측값이 자료의 평균으로부터 떨어진 정도

분산



분산이 작다



분산이 크다

분산

관측값이 $x_1, x_2, \ldots x_n$ 이고 평균이 $\bar{x}$ 일 때,

관측값에 대한 편차 = (관측값- 평균) = $(x_i - \bar{x})$

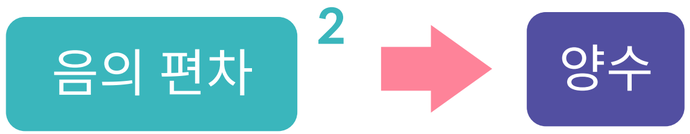
편차의 합은 항상 0

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

퍼진 정도의 측정으로 적절한 것은 편차의 평균
그렇지만 편차들의 합은 항상 0이므로 평균도 항상 0이 되어
편차의 평균은 퍼진 정도의 측도로 적합하지 않음

분산

음의 편차를 제곱하여 양수로 바꿀 수 있다



분산

편차의 제곱의 평균으로 퍼진 정도를 측정할 수 있다

$\frac{\text{편차의 제곱합}}{n}$ → 퍼진 정도를 측정

분산, s^2로 표기

관측값이 $x_1, x_2, \ldots x_n$ 이고 평균이 $\bar{x}$ 일 때,

$$\text{분산 } s^2 = \frac{(\text{편차의 제곱합})}{n} = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

표준편차

`stdev()`

분산의 단위 = 관측값의 단위의 제곱

관측값의 단위와 불일치

분산의 양의 제곱근은 관측값과 단위가 일치

분산의 양의 제곱근을 표준편차라 하고 s로 표기

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

범위(Range)

```
np.max()-np.min()
```

관측값에서 가장 큰 값과 가장 작은 값의 차이

장점
간편하게 구할 수 있고
해석이 용이함

단점
- 중간에 위치한 값은 고려되지 않음
- 극단값의 영향이 클 수 있음

백분위수

```
np.percentile()
```

중앙값을 확장한 개념

자료를 순서대로 정렬했을 때 백분율로 특정 위치의 값을 표현

백분위수

제 100×p 백분위수를 구하는 방법

1. 관측값을 오름차순으로 배열

2. 관측값의 개수(n) 에 p를 곱셈

3-1. n×p가 정수인 경우

n×p 번째로 작은 관측값과
n×p+1 번째로 작은 관측값의 평균

3-2. n×p가 정수가 아닌 경우

n×p에서 정수 부분에 1을 더한 값 m을 구한 후
m번째로 작은 관측값

사분위수

`np.percentile(25)` `np.percentile(50)` `np.percentile(75)`

백분위수의 일종으로 전체를 사등분하는 값

제1, 2, 3 분위수를 각각 Q_1, Q_2, Q_3 으로 표시

- 제 1 사분위수 : Q_1= 제 25백분위수
- 제 2 사분위수 : Q_2= 제 50백분위수
- 제 3 사분위수 : Q_3= 제 75백분위수

중앙값은 전체의 1/2에 위치하는 값이므로 제 2사분위수 및 제 50백분위수

사분위수 범위

제 3사분위수와 1사분위수 사이의 거리

사분위수 범위 IQR = 제 3사분위수 – 제 1사분위수 = $Q_3 - Q_1$

범위	사분위수 범위
전체 관측값이 퍼진 정도	관측값의 중간 50%에 대한 범위

표준편차, 범위, 사분위수 범위의 비교

평균의 특징 = 표준편차의 특징

중앙값의 특징 = 사분위수 범위의 특징

표준편차	사분위수 범위	범위
전체 관측값의 퍼진 정도를 골고루 반영 단점 : 극단적인 관측값에 의해 영향을 받음	극단값의 영향없이 퍼진 정도를 확인 가능 단점 : 제1사분위수와 제3사분위수 사이의 관측값에 대한 분포를 반영하지 않음	퍼진 정도를 나타냄 단점 : 표준편차의 단점과 사분위수 범위의 단점을 모두 가지고 있음

변동계수

퍼진 정도를 상대적으로 나타내는 수치를 사용

변동계수는 평균에 대한 상대적인 퍼진 정도를 백분율(%) 로 나타냄

$$\text{변동계수 } CV = \frac{\text{표준편차}}{\text{평균}} \times 100$$

비교 대상의 단위가 다른 경우, 단위가 없는 변동계수를 통해 퍼진 정도 비교 가능