요약 통계량

- 표, 그래프보다 더 간단
- 대표값: 간단한 수치
  - 평균 mean
  - 중앙 median
  - 모드 **mode**
- 산포도: 데이터가 대표값에서 얼마나 떨어져 있는지를 나타냄
  - 분산 variance
  - 표준편차 standard deviation
  - **사분위범위, 상자그림**
  - **변동계수**
  - **공분산 covariance**

대표값

임의의 수 n개를 더해라.

$$x_1 + x_2 + \cdots + x_n = \sum_{i=1}^{n} x_i$$

mean

임의의 수 n개에 대한 mean을 구해라.

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{x_1}{n} + \frac{x_2}{n} + \cdots + \frac{x_n}{n} = \sum_{i=1}^{n} \frac{1}{n} x_i = \bar{x}$$

sample mean

population mean μ

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$ 1, 2, **3**, 4, 5

빈도

가중평균: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$$

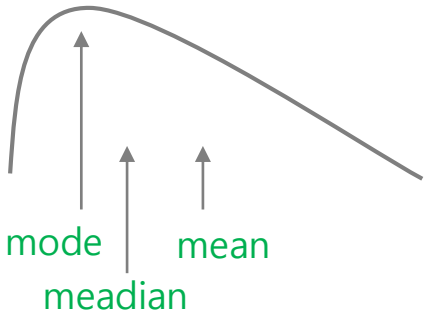
절사평균 : 큰 쪽 작은 쪽 각각 k개씩 제거한 나머지의 평균

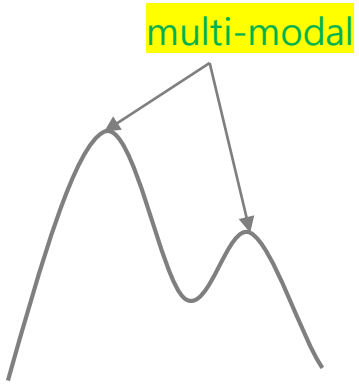
중위수; median

크기 순서대로 정렬 3, 5, 7, 7, 9 3, 5, 7, 7

최빈값

mode, multi-modal

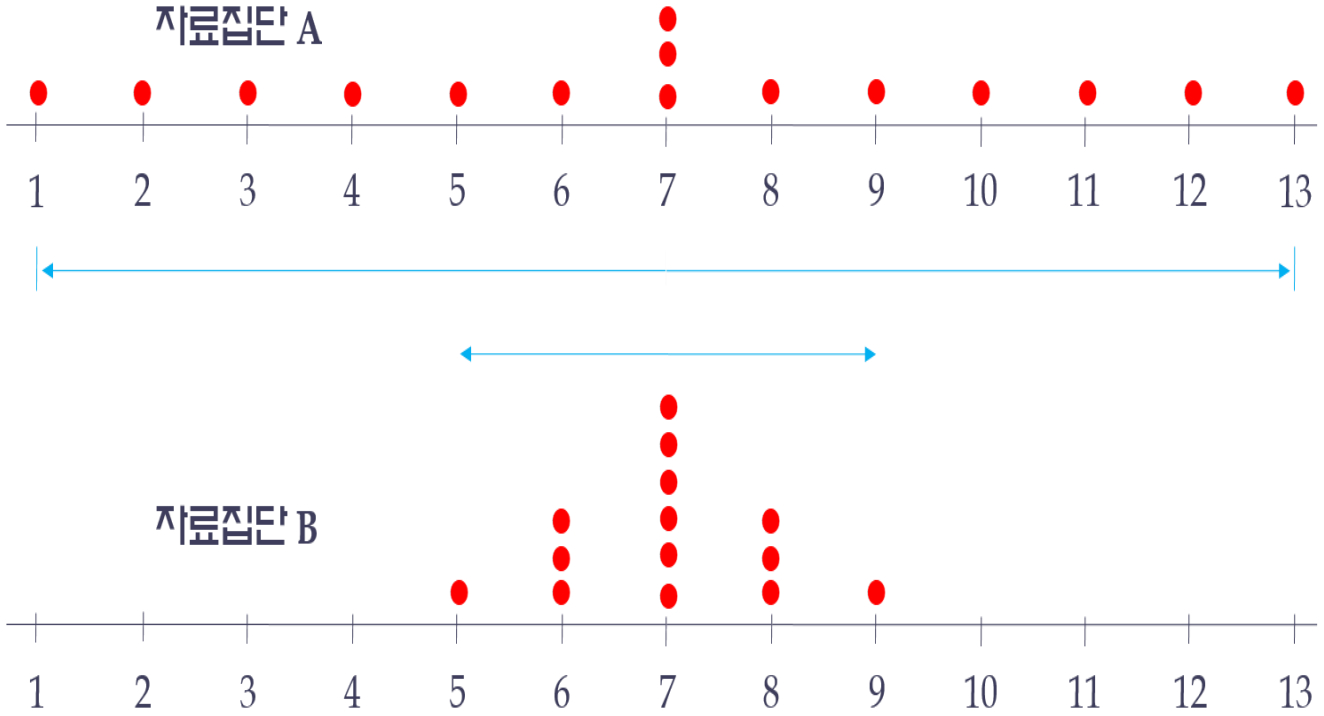




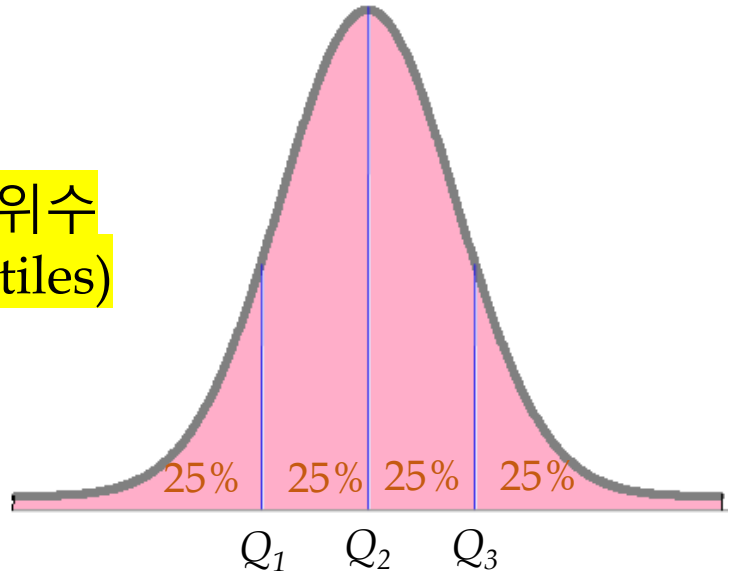
산포도

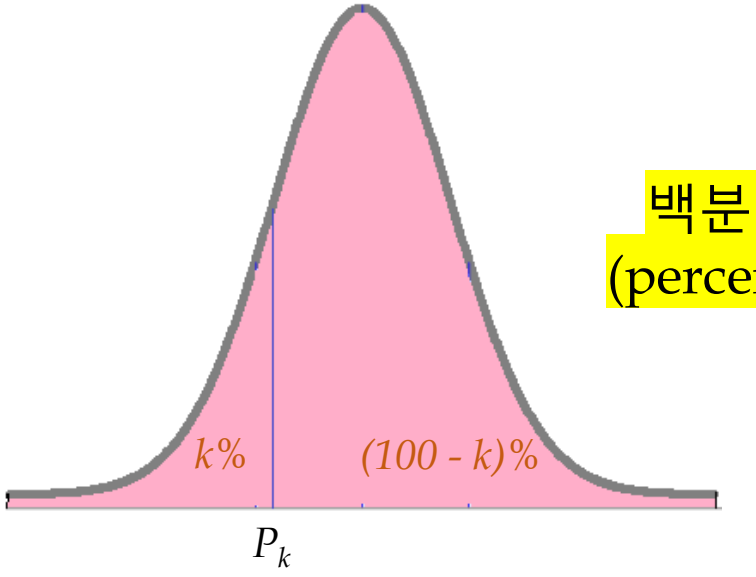
자료값이 평균을 중심으로 어떻게 분포되는지 보여준다.

자료집단 A [1, 2, 3, 4, 5, 6, 7, 7, 7, 8, 9, 10, 11, 12, 13]
자료집단 B [5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9]



사분위수
(quartiles)





백분위수
(percentiles)

$$m = \frac{k}{100} n$$

1. 자료값을 가장 작은 수부터 크기 순서로 재배열한다.
2. $m = kn/100$을 계산한다.
3. m이 정수이면 $P_k = (x_{(m)} + x_{(m+1)}) / 2$이고 m이 정수가 아니면, m보다 큰 가장 작은 정수에 해당하는 위치의 자료값이다.

제1사분위수 $Q_1 = P_{25}$ 25 백분위수
제2사분위수 $Q_2 = P_{50}$ 50 백분위수 (= 중위수)
제3사분위수 $Q_3 = P_{75}$ 75 백분위수

[예제 16]
다음 자료집단에 대한 30-백분위수와 사분위수를 구하라.

67 84 79 62 78 36 38 57 48 87 83 90 60 25 50 94 60 62 97 43

풀이

먼저 자료값을 크기 순서로 재배열한다.

25 36 38 43 48 50 57 60 60 62 62 67 78 79 83 84 87 90 94 97

$n = 20,\ k = 30 \quad m = 6 \qquad P_{30} = (x_{(6)} + x_{(7)}) / 2 = (50 + 57) / 2 = 53.5$

$n = 20,\ k = 25 \quad m = 5 \qquad Q_1 = P_{25} = (x_{(5)} + x_{(6)}) / 2 = (48 + 50) / 2 = 49$

$n = 20,\ k = 50 \quad m = 10 \qquad Q_2 = P_{50} = (x_{(10)} + x_{(11)}) / 2 = (62 + 62) / 2 = 62$

$n = 20,\ k = 75 \quad m = 15 \qquad Q_3 = P_{75} = (x_{(15)} + x_{(16)}) / 2 = (83 + 84) / 2 = 83.5$

사분위수 범위; interquartile range

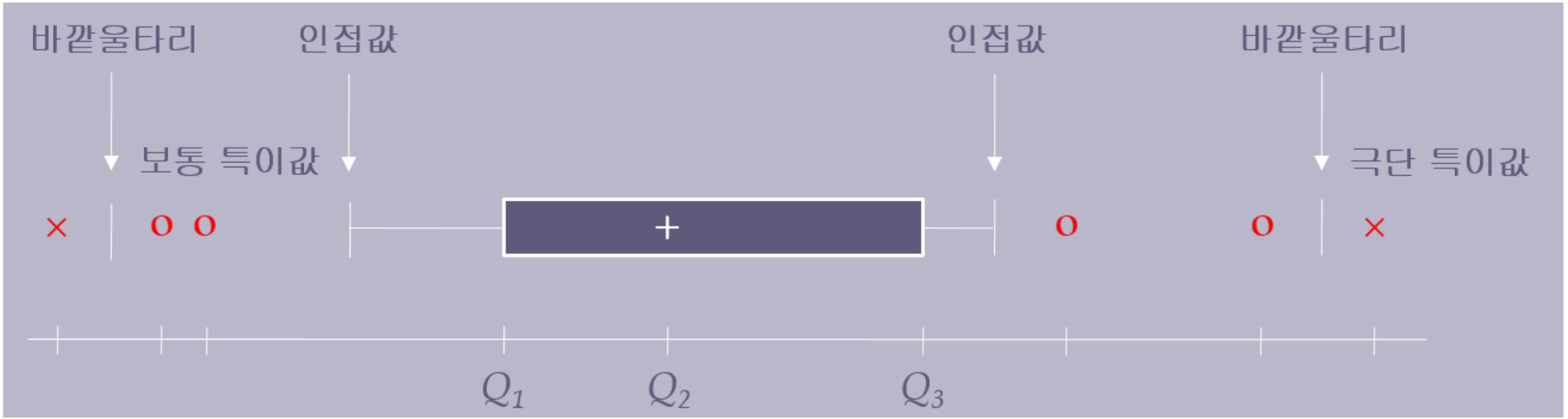
$$\text{IQR} = Q_3 - Q_1$$

사분위수범위는 중앙의 50%에 해당하는 자료에 대한 범위이며,
특이값에 대한 영향을 전혀 받지 않는다.

67 84 79 62 78 36 38 57 48 87 83 90 60 25 50 94 60 62 97 43

$Q_1 = 49$, $Q_3 = 83.5$이므로 IQR = $83.5 - 49 = 34.5$

상자 그림 (box plot)



1. 안울타리(inner fence) : 사분위수 Q_1과 Q_3에서 각각 1.5IQR 만큼 떨어져 있는 값 $f_l = Q_1 - 1.5IQR$, $f_u = Q_3 + 1.5IQR$
2. 바깥울타리(outer fence) : 사분위수 Q_1과 Q_3에서 각각 3IQR 만큼 떨어져 있는 값 $F_l = Q_1 - 3\text{IQR}$, $F_u = Q_3 + 3\text{IQR}$
3. 인접값(adjacent value) : 안울타리 안에 놓이는 가장 극단적인 자료값; 아래쪽 안울타리보다 큰 가장 작은 자료값과 위쪽 안울타리보다 작은 가장 큰 자료값
4. 보통 특이값(mild outlier) : 안울타리와 바깥울타리 사이에 놓이는 자료값
5. 극단 특이값(extreme outlier) : 바깥울타리 외부에 놓이는 자료값

[예제 18]
다음 자료에 대한 상자그림을 그려라.

49.6	50.5	49.9	51.6	49.6	48.7	49.7	49.1	48.7	51.0	50.1	48.7	50.4	50.6
51.5	49.4	51.1	49.8	49.8	49.0	47.2	50.4	49.1	50.5	50.9	49.8	49.6	49.3
50.5	50.2	52.0	50.7	50.4	48.6	50.9	51.2	50.7	48.5	50.0	51.3	47.6	49.1
51.0	51.9	49.5	49.7	48.6	49.7	48.5	48.3						

풀이

① 이 자료를 크기 순서로 재배열하여 사분위수를 구한다.

$$Q_1 = x_{(13)} = 49.1, \quad Q_2 = (x_{(25)} + x_{(26)})/2 = 49.8, \quad Q_3 = x_{(38)} = 50.7$$

따라서 사분위수범위는 IQR = *50.7 – 49.1 = 1.6*이다.

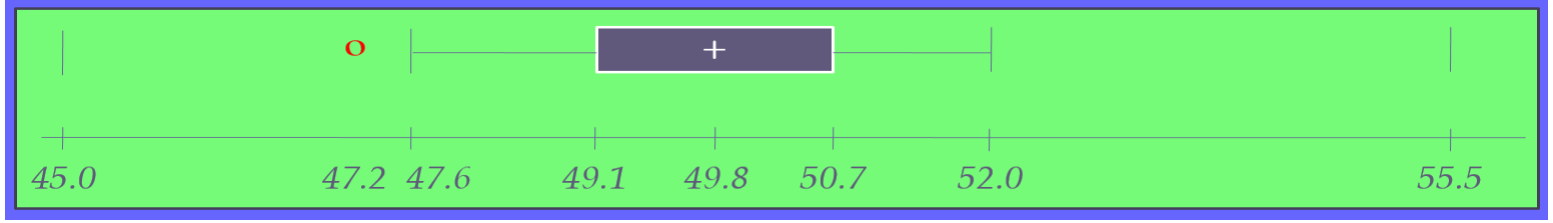
② 안울타리와 인접값을 구한다.

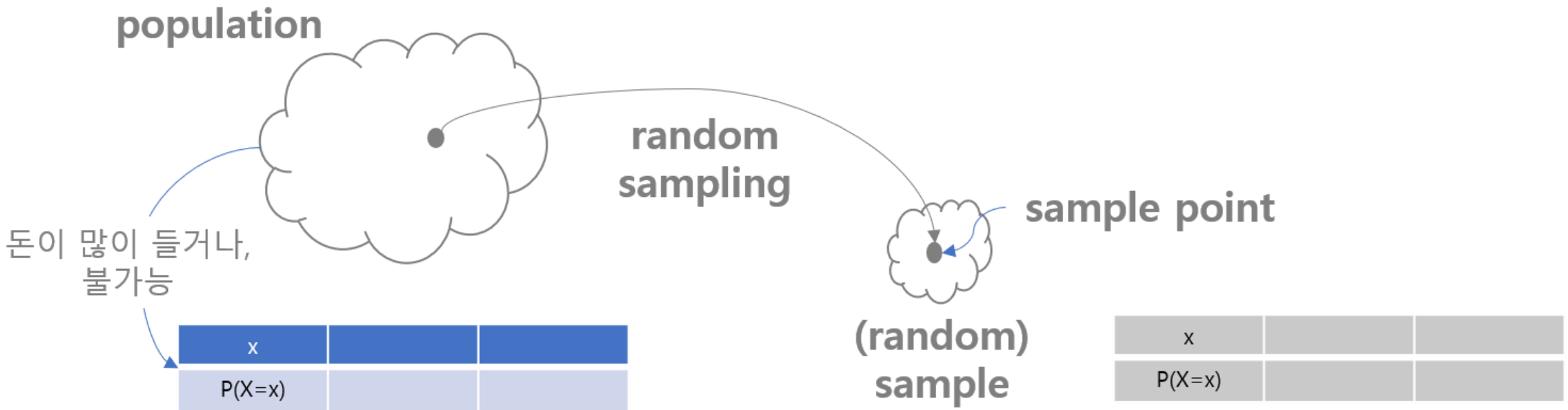
$$f_l = Q_1 - 1.5IQR = 49.8 - 2.4 = 47.4$$
$$f_u = Q_3 + 1.5IQR = 50.7 + 2.4 = 53.1$$

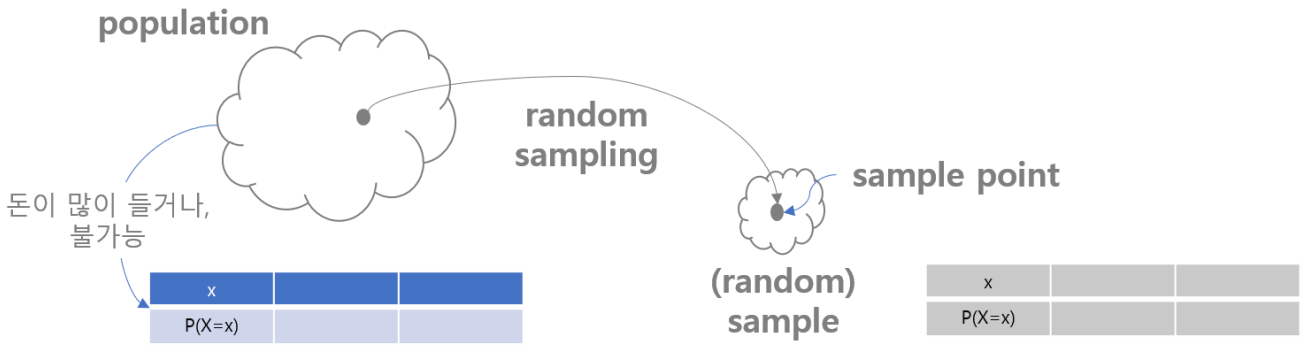
③ 바깥울타리를 구한다.

$$F_l = Q_1 - 3IQR = 49.8 - 4.8 = 45.0$$
$$F_u = Q_3 + 3IQR = 50.7 + 4.8 = 55.5$$

④ 자료값 47.2는 아래쪽 바깥울타리와 인접값 사이에 있으므로 보통 특이값이다. 그러나 최대 자료값이 위쪽 인접값이므로 중위수보다 큰 특이값은 없다.




population
random
sampling
sample point
돈이 많이 들거나,
불가능
x
P(X=x)
(random)
sample
x
P(X=x)



‘parameters’; 모수

‘모평균’; mean μ

‘statistics’; 통계량

‘표본평균’; sample mean $\bar{x}$

분산

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

분산

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

표본분산으로부터,
모분산 추정하는 것
나중에 배운다.

변동계수; coefficient of variation

두 집단의 산포도를 비교하고 싶을 때 사용

$$CV_P = \frac{\sigma}{\mu} \times 100(\%)$$

$$CV_S = \frac{s}{\bar{x}} \times 100(\%)$$

변동계수가 클수록 상대적으로 넓게 분포를 이룬다.

측정 단위가 동일하지만 평균이 큰 차이를 보이는 두 자료집단
측정단위가 서로 다른 두 자료집단에 대한 산포의 척도를 비교할 때 사용

예를 들어,
- 신생아의 몸무게와 산모의 몸무게(단위는 같으나, 평균의 차가 큰 경우)
- 키와 몸무게(단위가 서로 다른 경우)

참고자료

- 공학인증을 위한 확률과 통계, 이재원, 이욱기, 북스힐