

## 3장 범이와 브로 측정하기

### 1. 2장에서 배운 평균의 특성

- 데이터에 포함된 값들의 대푯에 대한 정보를 알려준다.
- 하지만 데이터에 대한 모든 정보를 제공하지는 못한다.
- 예를 들어, 데이터의 범위와 변화량에 대한 추가 특성이 요구된다.
- 평균(평균값, 중앙값, 최빈값)의 활용만으로는 한계가 있을 때 다음 예제 참조

P.125

### 2. 세 명의 나이 선수 기록 비교

선수 1

개별적 청수	7	9	10	11	13
도수	1	2	4	2	1

선수 2

개별적 청수	7	8	9	10	11	12	13
도수	1	1	2	2	2	1	1

선수 3

개별적 청수	3	6	7	10	11	13	30
도수	2	1	2	3	1	1	1

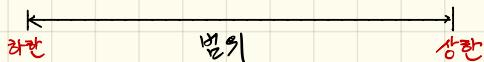
- ⇒ 세 선수의 기록 모두 동일한 평균 가짐  
 $\text{평균값} = \text{중앙값} = \text{최빈값} = 10$
- 따라서 평균을 이용하여 세 선수의 특성을 구분하지 못한다.
  - 하지만 세 선수의 기록은 불평등 다른 특성을 가짐
  - 예를 들어, 기록의 일관성이 다를 때.

p.126

### 3. 일관성 평가 방법 1 : 범위

- 범위 : 데이터에 포함된 숫자들이 퍼져있는 정도

예) 7 8 9 9 10 10 11 12 13



$$\text{범위} = \text{상한} - \text{하한} = 13 - 7 = 6$$

① 선수 1 데이터의 범위 :  $13 - 7 = 6$

② 선수 2 " " :  $13 - 7 = 6$

③ 선수 3 " " :  $30 - 3 = 27$

⇒ 선수 1과 선수 2의 일관성이 선수 3의 일관성보다 높다고 할 수 있음.

p.129

- 범위 사용의 한계

- 이상치가 있는 경우 문제 발생

주의) 제에서 다른 예제와 다른.

예) 2장 컴퓨터교실 참가자 나이 데이터

나이	19	20	21	145	149
도수	3	6	3	1	1

(단)

$$\text{범위} : 149 - 19 = 128$$

사용 범위

활용 가능

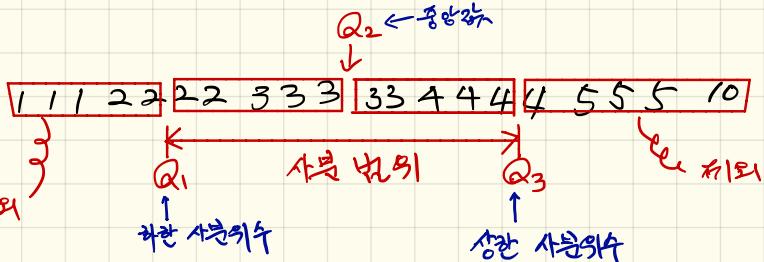
↑ { 이상치를 제외시키는  
방법 요구됨 }

{ 이상치 데이터 때문에  
범위가 너무 커졌음.  
따라서 주어진 데이터의 특성을  
범위가 제대로 반영 못함. }

• 이상치 문제 해결법 :

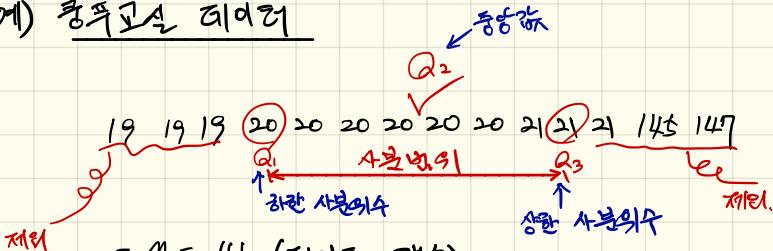
- ① 먼저 데이터를 오름차순으로 일렬로 나열할 것.
- ② 나열된 데이터를 동일한 크기로 = 것.

예)



- $n = 20$  (데이터 개수)
- $Q_1 : 20 \cdot \frac{1}{4} = 5$  번째와 6 번째의 평균값
- $Q_2 :$  중앙값
- $Q_3 : 20 \cdot \frac{3}{4} = 15$  번째와 16 번째의 평균값
- 사분위수 크기 :  $4 - 2 = 2$

예) 풍속고속 데이터



- $n = 14$  (데이터 개수)
- $Q_1 : 14 \cdot \frac{1}{4} = 3.5$  를 올림한 값에 위치한 값
- $Q_3 : 14 \cdot \frac{3}{4} = 10.5$  를 "
- $Q_2 :$  중앙값
- 사분위수 크기 :  $Q_3 - Q_1 = 21 - 20 = 1$

## • 사분위와 이상치

$$- \text{사분위} = Q_3 - Q_1$$

\* 이상치를 포함하지 않으면서 정상값을 중심으로 해서 전체 데이터의 50%에 해당하는 값들의 범위를 측정함.

$$* \text{중위수상의 사분위} = 21 - 20 = 1$$

$\Rightarrow$  정상값 ( $>0$ )을 중심으로 ± 1의 범위에서 전체 데이터의 50% 이상이 위치함.

- 사분위 이외의 값을 제외하면 이상치와 연관된 문제들을 피할 수 있음.

p.134

## • 사분위수 찾기

### - 하한 사분위수 ( $Q_1$ )

- ①  $(n * \frac{1}{4})$ 이 정수인 경우  $Q_1$ :  
이 정수값의 위치에 있는  
값과 그 다음에 있는 값의  
평균값

- ②  $(n * \frac{1}{4})$ 이 정수가 아닌 경우의  
 $Q_1$ : 나눗셈 값을 옮길한  
자리에 위치한 값

### - 상한 사분위수 ( $Q_3$ )

- ①  $(n * \frac{3}{4})$ 이 정수인 경우  $Q_3$ :  
이 정수값의 위치에 있는  
값과 그 다음에 있는 값의  
평균값

- ②  $(n * \frac{3}{4})$ 이 정수가 아닌 경우의  
 $Q_3$ : 나눗셈 값을 옮길한  
자리에 위치한 값

## 예) 증구 7수3의 데이타

제일작 첨수	3	6	7	10	11	13	30
3수	2	1	2	3	1	1	1

- 벌목 :  $30 - 3 = 27$

- 사분영수 :  $11 - 6 = 5$

$$\left\{ \begin{array}{l} Q_1 : 3 \text{ 번째 값인 } 6 \\ \text{이후} : 11 * \frac{1}{4} = 2.75 \end{array} \right.$$

$$Q_3 : 9 \text{ 번째 값인 } 11$$

$$\text{이후} : 11 * \frac{3}{4} = 8.25$$

- 책에 있는 예제 : 피자 스케이팅 점수 매기기

설정은 6명의 선수가 아래와 같이 점수를 주었을 때

5.0 5.5 5.5 5.7 5.8 6.0

- $\Rightarrow$  최고점수 6.0과 최하점수 5.0을 제외한 나머지 점수들을 합산함

~~5.0~~ 5.5 5.5 5.7 5.8 ~~6.0~~

그로테  $n=6$ 인 경우  $Q_1$ 은 2번재 위치한 값인 5.5이고

$Q_3$ 은 5번재 위치한 값인 5.8이다.

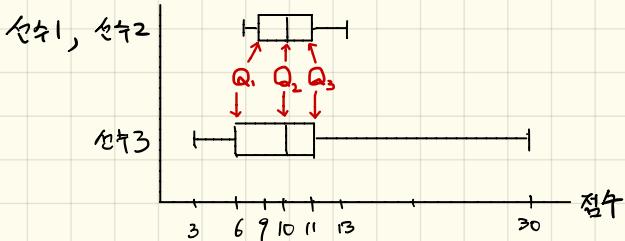
즉, 정확히 사분영수에 포함된 점수만 이용함.

p.140

#### 4. 상자수열 다이어그램

- 데이터의 범위를 한 칸에 담아둘 수 있도록 시작화하는 방법
- 포함되는 요소 :
  - 사분위수 ( $Q_1, Q_2, Q_3$ )
  - 범위
  - 사분 범위

예) 농구 선수 1~3의 상자수열 다이어그램



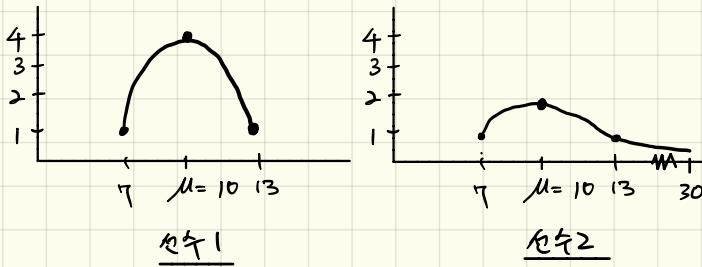
p.143

#### 5. 사분 범위의 한계

- 사분 범위 안에 위치한 데이터들이 얼마나 자주 발생하는지 또는 얼마나 자주 중앙값이 가까운 위치를 내는지 알려주지 않음.
- 예를 들어, 농구 선수 1과 선수2는 사분 범위를 이용하여도 점수 데이터를 구분하지 못함.
- 일관성을 측정하는 다른 기준이 요구됨.
- 범위(변동) 개념을 활용할 수 있음

## 6. 벤이 (변동)

예) 농구 선수1과 선수2의 템포를 선그래프로 나타내면  
기록의 안정성 측면에서 은명히 다른 특성을 보인다.



⇒ 선수1의 기록이 선수2의 기록보다 평균에  
집중되어 있을음. 즉, 데이터 각각이 평균값( $\mu$ )으로부터  
떨어져 있는 정도가 아름음.

⇒ 1번이를 사용하여 데이터 각각이 평균값으로부터  
떨어져 있는 정도를 측정할 수 있음.

- 벤이 : 단순히 절차들의 흐름 양상이 아니라, 데이터의  
안정성까지 포함하는 개념
  - 안정성 : 데이터 각각이 평균값으로부터 떨어져 있는 정도
  - 안정성을 측정하는 방법 필요.  
⇒ 분산과 표준편차 사용!

### • 주의사항 1 : 평균거리 활용

- 데이터가 단순히 평균값(10)과의 차이들의 평균값인 평균거리를 의미하기 때문.

#### 예) 농구 선수 3의 기록의 평균거리

개별형	3	6	7	10	11	13	30
수수	2	1	2	3	1	1	1
시작위 차이	-7	-4	-3	0	1	3	20

$$\begin{aligned}
 \text{평균거리} &= \frac{(3-10)*2 + (6-10) + (7-10)*2}{11} \\
 &\quad + (10-10)*3 + (11-10) + (13-10) + (30-10) \\
 &= 0
 \end{aligned}$$

- 이유 : 평균거리는 항상 0임!

### • 주의사항 2 : 평균편차 활용

- 평균편차 : 각 데이터가 평균으로 떨어진 정도의 평균값
- 농구 선수 3의 기록의 평균편차

$$\begin{aligned}
 \text{평균편차} &= \frac{|3-10|*2 + |6-10| + |7-10|*2}{11} \\
 &\quad + (10-10)*3 + (11-10) + (13-10) + (30-10) \\
 &= \frac{14 + 4 + 6 + 1 + 3 + 20}{11} = \frac{48}{11} = 4.36 \times
 \end{aligned}$$

- 하지만 평균편차는 활용성이 부족하여 사용하지 않음.
- 대신에 분산과 표준편차를 사용함.

P.146 ~147

## 7. 분산과 표준편차

- 분산 : 평균으로부터의 거리의 제곱들의 평균값

$$\text{분산 } (\sigma^2) = \frac{\sum (x - \mu)^2}{n}$$

- 표준편차 : 분산의 제곱근이며 분산보다 적관적인 정보 제공

$$\text{표준편차 } (\sigma) = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

- 분산은 숫자가 매우 큼. 따라서 분산의 제곱근인 표준편차를 보다 더 실용적으로 활용할.

P.154

### 예) 총수 선수3의 분산과 표준편차

$$\begin{aligned}\sigma^2 &= \frac{(3-10)^2 \cdot 2 + (6-10)^2 + (7-10)^2 \cdot 2}{11} \\ &\quad + (10-10)^2 \cdot 3 + (11-10)^2 + (13-10)^2 + (30-10)^2 \\ &= \frac{542}{11} = 49.27 \times \times\end{aligned}$$

$$\sigma = \sqrt{49.27} = 7.01 \times \times$$

- 이제 선수3의 표준편차를 선수1, 선수2의 표준편차와 비교해보자.

## - 선수 1의 음산과 풍진포차

개별당점수	7	9	10	11	13
도수	1	2	4	2	1

$$\bar{G}^2 = \frac{7+2+2+9}{10} = 2.2$$

$$G = 1.48xx$$

## - 선수 2의 음산과 풍진포차

개별당점수	7	8	9	10	11	12	13
도수	1	1	2	2	2	1	1

$$\bar{G}^2 = \frac{9+4+2+2+4+9}{10} = \frac{30}{10} = 3$$

$$G = 1.72xx$$

### 결론

- 풍진포차를 활용한 결과 역시 선수 3의 일관성이 선수 1과 선수 2에 비해 많이 떨어진다는 사실을 보여준다.
- 또한 풍진포차를 활용하여 선수 1과 선수 2의 기록의 일관성에 미세한 차이가 있음을 알 수 있음.
- 선수 1이 선수 2보다 기록의 일관성이 좀 더 높으며 따라서 좀 더 안정적임.

연습문제

회사 직원들의 연봉을 알려주고자 함

- 방식 ①: 모든 직원의 연봉을 동일하게 그 백분위씩 옮기기.

- 방식 ②: 각자의 연봉의 10%씩 옮기기

- 전제: 기존 연봉의 평균값 ( $M_{old}$ ) 와 표준편차 ( $\sigma_{old}$ )가 알려졌다고 가정한.

질문: 표준편차 어떻게 보면하는가?

- 방식 ①을 사용할 경우

• 직원 각자의 새 연봉:  $x + 200$

• 직원 연봉의 새 평균값:  $M_{old} + 200$

$$\Rightarrow \sigma_{new} = \sqrt{\frac{\sum ((x + 200) - (M_{old} + 200))^2}{n}}$$

$$= \sqrt{\frac{\sum (x - M_{old})^2}{n}} = \sigma_{old}$$

• 결론: 소득격차는 변하지 않음.

- 방식 ②를 사용할 경우

• 직원 각자의 새 연봉:  $x * 1.1$

• 직원 연봉의 새 평균값:  $M_{old} * 1.1$

$$\Rightarrow \sigma_{new} = \sqrt{\frac{\sum ((x * 1.1) - (M_{old} * 1.1))^2}{n}}$$

$$= 1.1 * \sqrt{\frac{\sum (x - M_{old})^2}{n}} = 1.1 * \sigma_{old}$$

• 결론: 소득격자 또한 10% 정도 더 벌어진다.

- 분산을 구하는 식

$$\begin{aligned}
 \sigma^2 &= \frac{\sum (x - \mu)^2}{n} \\
 &= \frac{\sum (x^2 - 2\mu x + \mu^2)}{n} \\
 &= \dots \quad (\text{처지면 중요하지 않음}) \\
 &= \frac{\sum x^2}{n} - \mu^2
 \end{aligned}$$

↑ 계산이 오래 간단느낌.

- 분산(표준편차)의 특성

- 개개의 데이터에 대한 정보를 반영
- 계산 쉬운 .
- 분산(표준편차)이 클수록 데이터들이 평균값으로부터 크게 흐트러진 형태를 띠있.

## f. 표준偏差

- 앞서 다른 농구 선수와 선수2의 경우 평균값이 동일하지만 표준偏差가 달랐음.

그래서 표준偏差가 보다 작은 선수의 기록이 보다 안정적이다라고 평가했음.

- 그런데 평균이 다르거나 다른 조건을 갖는 데이터를 서로 비교할 때 표준偏差가 나타내는 수자가 아닌 다른 기준으로 평가해야 하는 경우 있음.

9.15f

### 예) 두 명의 농구선수의 $\frac{1}{2}$ 성공률

선수1	$\therefore \begin{cases} M = 70\% \\ S = 20\% \end{cases}$	선수2	$\therefore \begin{cases} M = 40\% \\ S = 10\% \end{cases}$
-----	---	-----	---

### 최근 시장의 결과

- 선수1의  $\frac{1}{2}$  성공률 : 75%

- 선수2의 " : 55%

$\Rightarrow$  선수1의 성공률이 훨씬 높지만

선수1이 보다 잘했다고 말하기는 어렵다.

이유 : 평균값이 다르기 때문에 단순히 표준偏差만으로 전체 그림을 파악할 수 없음.

질문: 누가 더 평상시보다 둘은 기록을 달성하였는가를  
평가하는 기준을 어떻게 설정할 것인가?

$\Rightarrow$  표준점수를 활용하여 선수가와 선수그를 비교한 후  
있음.

**표준점수**: 두 개의 데이터 집합을 비교할 때  
기준역할을 수행함.

표준점수 계산법:

$$z = \frac{x - \mu}{\sigma}$$

$\uparrow$   
 $x$ 의 표준점수

$\Rightarrow$

선수1의 표준점수

$$z_1 = \frac{75 - 70}{20} = \frac{5}{20} = 0.25$$

선수2의 표준점수

$$z_2 = \frac{55 - 40}{10} = 1.5$$

결론: 선수2의 표준점수가 예상대로 선수1의 표준점수보다  
높다. 이는 선수2의 최근 기록이 자신의  
평상시 기록을 월등히 넘어선다는 성적을  
그대로 반영한다.

• 표준편수의 특성

- 기존 데이터들을 모두 표준화하라 시키면 평균값은 0, 표준편자는 1이 된다.

증명

- 설명

- 표준화 이전 평균값과 표준편자:  $\mu$ ,  $\sigma$
- 표준화 후 평균값과 표준편자

$$\begin{aligned}\mu_z &= \frac{\sum z}{n} \\ &= \frac{\sum (\frac{x-\mu}{\sigma})}{n} \\ &= \dots \\ &= 0\end{aligned}$$

$$\begin{aligned}\sigma_z &= \frac{\sum (z - \mu_z)^2}{n} \\ &= \dots \\ &= 1\end{aligned}$$

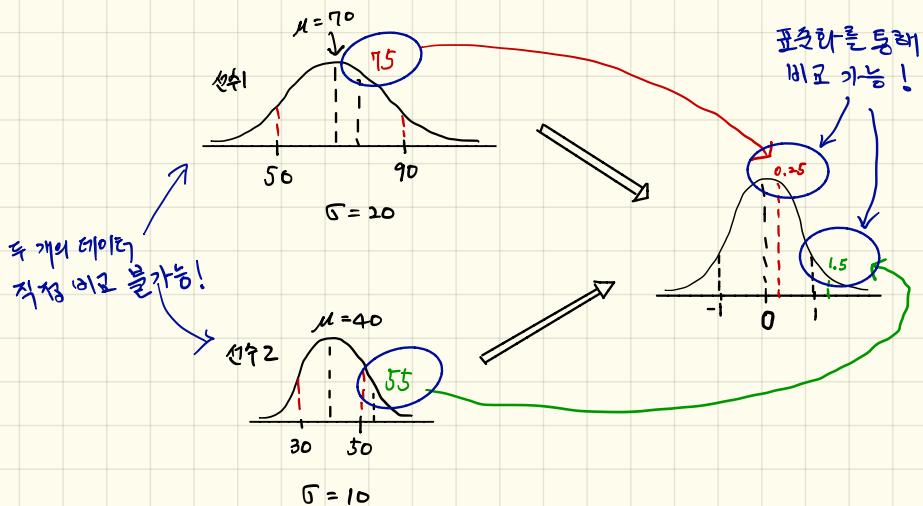
(생각의 부분은 크게 중요하지 않음)

- 즉, 표준화를 통해 모든 데이터를 아래 그림이 보여주는 것처럼 표준화 시킨다.



P. 160

### 예) 농구 선수1과 선수2의 표적골률 데이타 표준점수 의미



P. 161

### • 평균값으로부터의 표준편차

- "특정 값이 평균값으로부터 1 표준편차 이내에"

존재한다"의 의미?

$\Rightarrow$  해당 값의 표준점수가  $-1$ 과  $1$  사이에  
존재함을 의미함.

- 즉, 표준점수는 평균값으로부터의  
표준편차의 수를 의미함.