

6) 전에 배운 평균 :

- 데이터 징향이 갖고 있는 전형적인 값을 알려준다.
- 하지만 데이터 징향의 모든 정보를 제공하지 못함
- 예를 들어, 데이터들의 별위와 변화량에 대한
추가 분석이 요구된다.
- 앞서 배운 평균값, 중앙값, 최빈값 전부를
활용 하더라도 한계가 발생한다.
다음 예제 참조.

P. 125

시| 명의 농구 선수 기록 비교

선수 1

제일당 절수	7	9	10	11	13
도수	1	2	4	2	1

선수 2

제일당 절수	7	8	9	10	11	12	13
도수	1	1	2	2	2	1	1

선수 3

제일당 절수	3	6	7	10	11	13	30
도수	2	1	2	3	1	1	1

시| 선수 기록의 평균값, 중앙값, 최빈값 모두 10점인!

⇒ 하지만 각 선수의 기록은 분명히
다른 특성을 보인다.

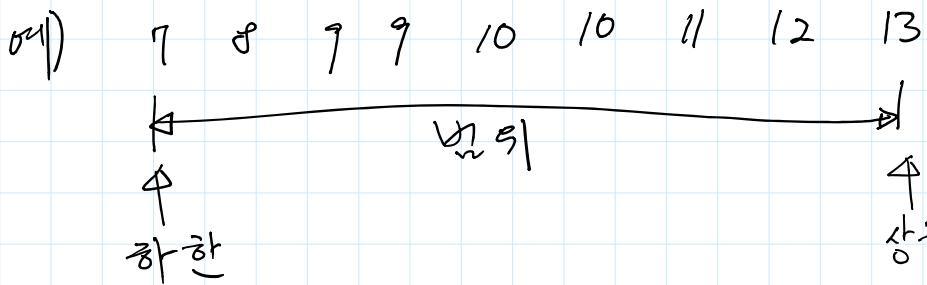
예를 들어, 기록의 일관성이 다르다.

* 기록의 일관성을 평가하는 방법 필요!

P. 126

| 일관성 평가 방법 1 : 범위 사용 |

설명: 데이터에 포함된 숫자들이 띄져있는 정도를
가장 간단하게 측정하는 방법



$$\begin{aligned} \text{범위} &= \text{상한} - \text{하한} \\ &= 13 - 7 \\ &= 6 \end{aligned}$$

P. 129

별도의

이상치

결론: 별도의 이상치의 문제를 전혀 해결하지 못함.

이유: 별도의 이상치에 대해 매우
민감하게 반응한다.

(예 1)

1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 4 5 5 5
 ↓
 하한

↑
 상한

(예 2)

1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 4 5 5 5
 ↓
 하한

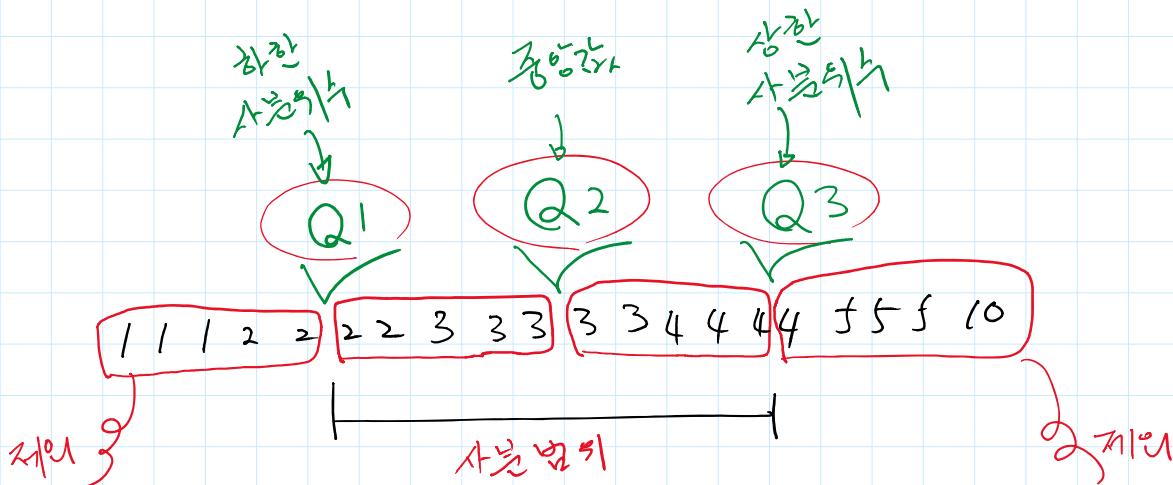
~
 10
 ↑
 상한

[P. 132] ~ [P. 133]

이상치 문제 해결방법: 사분위수

↓

데이터를 동일한 크기의 네 조각으로
나누는 방법.



⇒ 사분위수 이외의 값을 제외한 범위를,

대푯값으로 하면 이상치 문제 해결 가능!

⇒ 사용법에는 이상치를 포함하지 않는다.

책에 있는 예제

피지스케이팅 점수 매기기

6명의 심판이 아래와 같이 점수를 매겼을 경우

5.0 5.5 5.5 5.7 5.8 6.0

이제 최고점 6.0과 최하점 5.0을 제외한 나머지 점수들을 합산한다. 즉,

$$\cancel{5.0} \quad 5.5 + 5.5 + 5.7 + 5.8 \quad \cancel{6.0}$$

\downarrow

그런데 $n=6$ 인 경우,

$$\textcircled{1} \quad Q_1 = 2\text{번째 } \frac{\text{값}}{2} = 5.5$$

$$(\Leftarrow \frac{6}{4} = 1.5)$$

$$\textcircled{2} \quad Q_3 = 5\text{번째 } \frac{\text{값}}{2} = 5.8.$$

$$(\Leftarrow \frac{3*6}{4} = 4.5)$$

즉, 정확히 사분위에 있는 수들만 대상으로 점수를 매긴다.

P. 134

사분위수 찾기

1) 하한 사분위수 찾기

① $m \div 4$ 계산

② $m \div 4$ 가 정수

\Rightarrow 사분위수 위치 = 이 정수 위치의 값과

그 다음에 오는 값 사이

정한 사분위수 = 수 위치에 있는 값들의
평균값

③ $m \div 4$ 가 정수가 아님

\Rightarrow 값을 옮긴한 값의 위치.

2) 상한 사분위수 찾기

① $3m \div 4$ 계산

② $3m \div 4$ 가 정수

\Rightarrow 이 값과 다음에 위치한 값 사이

③ $3m \div 4$ 가 정수 아님

\Rightarrow 옮긴한 위치 찾는 값의 위치

상한 사분위수 = 수 위치에 있는 값들의
평균값

(P. 135)

8/10

시행당 전수	3	6	7	10	11	13	30
도수	2	1	2	3	1	1	1

① 평균 $= 30 - 3 = 27$

② 상한 사용도수

- $m = 11$

3 3 6 7 7 10 10 11
↓ Q1

30 6 10 11
↑ Q3

13 30

$\Rightarrow 3m \div 4 = 8.25$

$\Rightarrow 3 \text{ 번째 } ?_{\text{값}} = 11$

하한 사용도수

$\Rightarrow m \div 4 = 2.75$

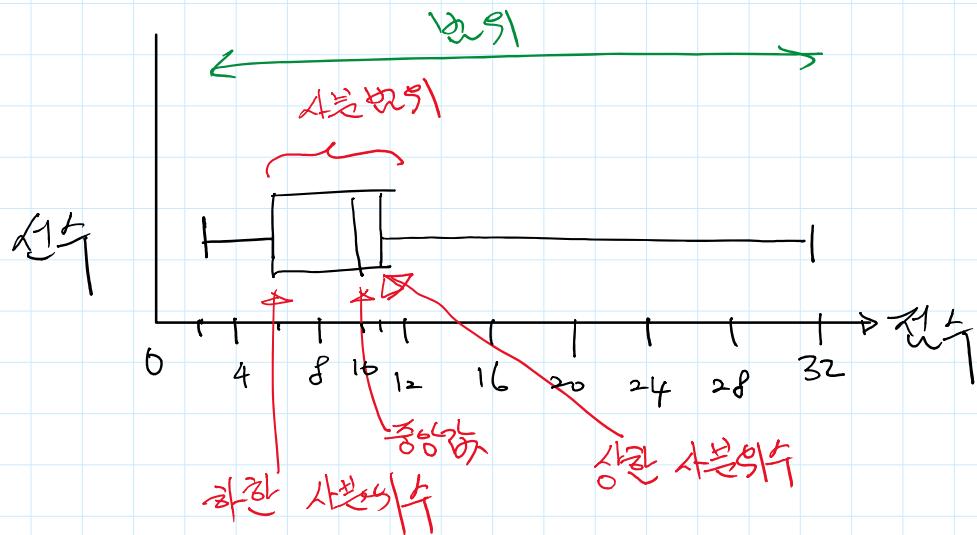
$\Rightarrow 3 \text{ 번째 } ?_{\text{값}} = 6$

③ 사용범위 $= 11 - 6 = 5$

(P. 140)

상자수법 사이어그램

앞서 자주 배운 선수의 전수 범위를
상자수로 사이어그램을 이용하여 시각화할 수 있다.



[P. 141]

예) 상자수로 사이어그램을 활용한
농구선수 기록 비교



\Rightarrow 중앙값과 사분위수로 기준으로 판단했을 때
선수 A 선택!

Φ. 143

사물의학의 경계 : 사물의학 안에 위치한 전수들이
얼마나 자주 발생하는지
또는 얼마나 자주 중앙값이
가장은 전수를 내는지
알려주지 않음.

예) 처음에 언급했던 세 명의 경우 선수들의
경우, 선수₁과 선수₂의 경우는
사물의학 조차 동일하다.
따라서 두 선수의 상자수로 차이가
그럼이 완전히 동일하다
하지만 두 선수의 기록은
흔명히 다른 특징을 나타낸다.

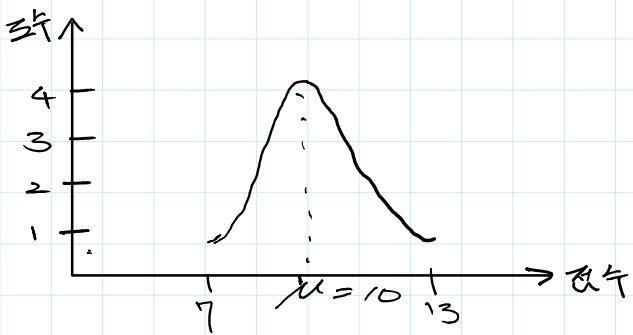
⇒ 일관성을 측정하는 방식으로
사물의학의 하위의 다른 기준이 요구됨

P. 144

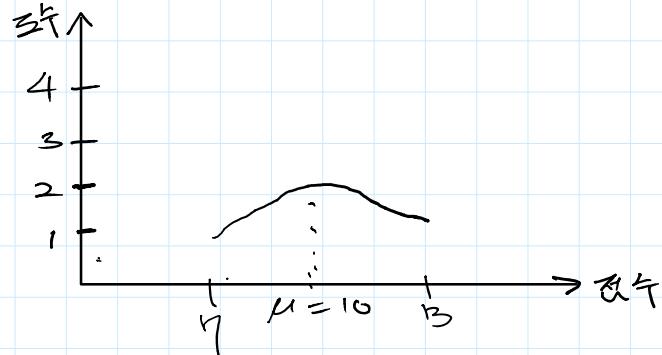
변이란?

단위히 전수들의 분포 양상이 아니라,
데이터의 안정성까지 포함하는 개념.

변이 측정
방법 예제 : 데이터 각각이 평균값으로부터
떨어져 있는 정도!



선수 1



선수 2

$\Rightarrow \left\{ \begin{array}{l} \text{선수 2의 기록이 선수 1의 기록보다} \\ \text{성과에 집중되어 있음!} \end{array} \right.$

P. 145

주의!

변수가 단순히 평균으로부터 벗어진 정도의 평균을 예측하지 않음

△ 61% : 평균거리의 장상이 있.

6) 궁구선수 전수.

시행당 천수	3	6	7	10	11	13	30
도수	2	1	2	3	1	1	1

$$\text{平均数} = \frac{6+6+14+30+11+13+35}{11}$$

$$\Rightarrow \text{평균거리} = \frac{(3-10) \times 2 + (6-10) + (7-10) \times 2 + (10-10) \times 3 + (11-10) + (13-10) + (30-10)}{11}$$

$$= \frac{-14 + (-4) + (-6) + 0 + 1 + 3 + 20}{11}$$

$$= 0$$

P. 146

문장 : 빛이 측정하기 위한 주간.

|| 짐짓으로 무역의 기지의 저승들의 행운!

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

P. 147

하지만

편차가 뜻보다 직관적일

$$\begin{aligned} \text{편차 } (\sigma) &= \sqrt{\text{분산}} \\ &= \sqrt{\frac{\sum (x-\mu)^2}{n}} \end{aligned}$$

즉, $\sqrt{\text{분산}} = \text{편차}^2 = \sigma^2$

P. 153

분산을 어떻게 계산하기

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

평균과 차이를 따로
계산할 필요가 있다.

P. 154

선수 짜증 선수

선수 1 ~ 선수 3의 기록이 다음과 같다.

(선수 1)

제일당 절수	7	9	10	11	13
도수	1	2	4	2	1

(선수 2)

제일당 절수	7	8	9	10	11	12	13
도수	1	1	2	2	2	1	1

(선수 3)

제일당 절수	3	6	7	10	11	13	30
도수	2	1	2	3	1	1	1

세 선수기록의 평균과 표준편차를 계산해 보자.

$$\text{선수 } 1 \quad : \quad \mu = 10$$

$$\sigma^2 = \frac{7^2 + 9^2 * 2 + 10^2 * 4 + 11^2 * 2 + 13^2}{16} - 10^2$$

$$= 2.2$$

$$\Rightarrow \sigma = \sqrt{2.2} = 1.48$$

선수 2 : $M = 10$

$$\sigma^2 = \frac{7^2 + 8^2 + 9^2 \times 2 + 10^2 \times 2 + 11^2 \times 2 + 12^2 + 13^2}{10} - 10^2$$

$$= 3$$

$$\Rightarrow \sigma = \sqrt{3} = 1.73$$

선수 3 : $M = 10$

$$\sigma^2 = \frac{3^2 \times 2 + 6^2 + 7^2 \times 2 + 10^2 \times 3 + 11^2 + 13^2 + 30}{11} - 10^2$$

$$= 49.27$$

$$\sigma = \sqrt{49.27} = 7.02$$

결론: 선수 1과 선수 2는 표준편차가 선수 3에 비해
상대적으로 많이 작다.
선수 3의 표준편차 7.02는 선수 3이
시행할 때마다 평균적으로
 $10 - 7.02$ 에서 $10 + 7.02$, 즉
약 3회에서 17회 사이의
점수를 기록한다는 것을 의미한다.

또한 선수 1의 경우가 선수 2의 경우보다
조금이나마 더 안정적이다라고
할 수 있다.

따라서 같은 이라면 선수 1을 선택하게 될 것이다.

P. 155

연습문제 : 평균에서 각각 있던 월급 인상 관련 예제 계산.

문제 : 연봉 인상 방식

- ① 일정하게 2,000 를씩 인상
- ② 각자의 연봉을 10%씩 인상.

a) ①의 경우 표준 편차는 어떻게 변할까?

원래 각각의 연봉을 X_i 라 하고
원래의 연봉 평균값을 μ 라 하자.

그러면 각각의 연봉은 $X_i + 2000$ 이 되고
평균값은 $\mu + 2000$ 이 된다.

따라서 새로운 표준 편차는

$$\begin{aligned}\sigma_{\text{new}}^2 &= \frac{\sum ((x+2000) - (\mu+2000))^2}{n} \\ &= \frac{\sum (x-\mu)^2}{n} \\ &= \sigma_{\text{old}}^2\end{aligned}$$

즉, 표준 편차는 변하지 않는다.

b) ②의 경우 표준편차

각 직원의 세연봉 = $1.1 * x_i$

연봉의 평균값 = $1.1 * \mu$

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

$$\Rightarrow \sigma_{\text{new}} = 1.1 * \sigma_{\text{old}}$$

즉, 원래 표준편차의 1.1 배.

즉, 연봉의 편차가 커진다.



기준에 높은 연봉을 '얕은' 경우가
처리하다.

¶. 15d

앞서 3주 선수 1과 선수 2의 경우 평균이

동일하지만 표준편차가 달랐다.

그래서 표준편수가 불과 작은 선수 1의
기록이 보다 안정적이라고 평가할 수 있다.

하지만 평균이 다르거나 다른 조건을

. . . →

갖는 테이터를 상호 비교하고자 할 때는
간단히 숫자의 크기만으로 평가할 수
없는 경우가 발생한다.

예) 두 명의 농구 선수의 슛 성공률

$$\text{선수 1} : \mu = 70\% \\ \sigma = 20\%$$

$$\text{선수 2} : \mu = 40\% \\ \sigma = 10\%$$

최종 사학의 결과

- 선수 1의 슛 성공률 75%
- 선수 2의 슛 성공률 55%

질문: 평상시 기록에 비추어 누가 더
잘한 것일까요?

가능한 답변: 선수 1의 성공률이 훨씬 높다.
따라서 선수 1이 더 잘했다.

질문: 정말 그러한가? 즉,
단순히 퍼센트값만으로 전체 그림을
파악할 수 있나?

⇒ 아니다!

선수 2의 경우 평상시 실력보다 훨씬

좋은 기록을 잘 성취했다.

새로운 질문: 누가 더 평상시보다 좋은 기록을
성취하였는지 평가하는 기준을
어떻게 설정하나?

⇒ 방법: 표준점수 (또는 표준점수를 블립)
활용!

[p. 159]

표준점수 : 두 개의 데이터 집합을 비교할 때
기준 역할을 수행한다.

표준점수 계산법

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

↳ μ 기존 데이터
 σ 기존 평균
 σ 기존 표준편차

기존 데이터의 표준 점수.

⇒ 이렇게 구한 표준 점수들은 항상
평균 (μ)이 0이고

표준편차 (σ)는 1이 된다.

$$\frac{\sum z}{n} = 0 \quad \text{이고}$$

표준 점수들의 평균

$$\frac{\sum (z-0)^2}{n} = 1 \quad \text{이다.}$$

표준 점수들의 표준편차.

⇒ 따라서 기준에 서로 다른 조건에서 얻어진 데이터를 소위 "표준화"하여 두 데이터를 비교할 수 있게 되었다.

예) 선수 1의 최근 놛 성공률의 표준점수

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

- 선수 2의 최근 경기 놛 성공률의 표준점수

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

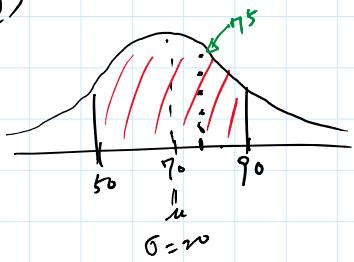
⇒ 선수 2의 표준점수가 선수 1의 표준점수 보다 높다.

(P. 160)

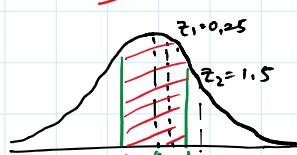
표준 점수 해석하기.

앞서 언급한 두 선수의 그래프 비교

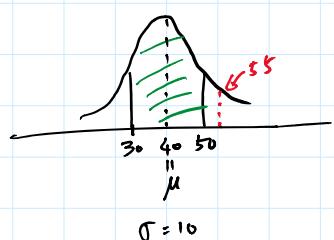
(선수 1)



$$\begin{cases} \mu = 0 \\ \sigma = 1 \end{cases}$$



(선수 2)



수평 축
평균값보다 높아
평균값보다 낮아

결론: 결론 전수 결과에 의해,
전수 2가 전수 1이 비례 정상시운라
활선 좋은 조성률을 보인 것을
다시 한번 확인할 수 있다.