3. 자료의 중심과 산포 측도

김 덕 기

toby123@cbnu.ac.kr

중심을 나타내는 수치적 측도

- 수치로 자료의 특성을 나타내는 방법 :

➔ 중심위치의 측도, 산포의 측도, 비대칭의 측도(왜도) 등.

➔ 중심위치의 측도 : 자료의 중심부가 어디에 위치해 있는가를 나타내 주는 것.

(1) 산술평균, (2) 절사평균, (3) 중앙값, (4) 최빈값이 대표적인 중심측도.

① n개의 자료가 $x_1, x_2, \cdots, x_n$으로 주어져 있을 경우 :

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

② 자료가 도수분포 형태로 주어져 있을 경우 :

변수(x_i)	x_1	x_2	$\cdots$	x_k	계
도수(f_i)	f_1	f_2	$\cdots$	f_k	$\sum_{i=1}^{k} f_i = n$

$$\bar{x} = \sum_{i=1}^{k} f_i x_i / n$$

중심측도 : 절사평균, 가중평균

- 절사평균(trimmed mean) : 자료를 순서대로 나열된 자료 중 양쪽p%를 버린 후 가운데 100(1-2p)% 자료의 평균을 구한 것.

- 특징 : 산술평균은 이상치에 영향이 많은 단점이 있는데 이러한 단점을 보완하여 고안 됨.

- 가중평균(weighted mean) : 여러 개의 평균이 각각 다른 도수를 가지고 있을 경우 평균에 가중치를 부여하여 전체 평균을 산출한다.

(ex) 탁구공 5개, 3개, 2개의 평균무게가 각각 4.2gram, 4.05gram, 4.1gram이었다. 전체의 평균무게는 얼마인가?

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

$$\bar{x}_w = \frac{5 \times 4.2 + 3 \times 4.05 + 2 \times 4.1}{5+3+2} = 4.135\ gram$$

중심측도 : 중앙값(median), 최빈값(mode)

- n개의 자료를 크기 순으로 배열하였을 때 중앙의 위치에 놓인 자료의 값을 중위수 M_d 라 정의한다. (특징 : 이상치에 둔감)

$$x_1, x_2, x_3, \cdots, x_{n-1}, x_n \xrightarrow{\text{order}} x_{[1]}, x_{[2]}, x_{[3]}, \cdots, x_{[n-1]}, x_{[n]}$$

n이 홀수인 경우: $x_{[(n+1)/2]}$, n이 짝수인 경우: $\dfrac{x_{[n/2]} + x_{[n/2+1]}}{2}$

- 최빈수(mode)는 도수가 가장 많은 측정치를 말한다.
- 최빈수는 2개 이상일 수 있다.

1, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 6, 7, 8, 9, 9, 9, 9, 9, 10

무엇을 대표-값으로 사용할 것인가 ?

Ex) 어느 중소기업의 각 직위별 월급 현황. 회사의 전체 직원에 대한 월급의 대푯값(중심측도)은 어떤 값을 사용해야 하는가 ?

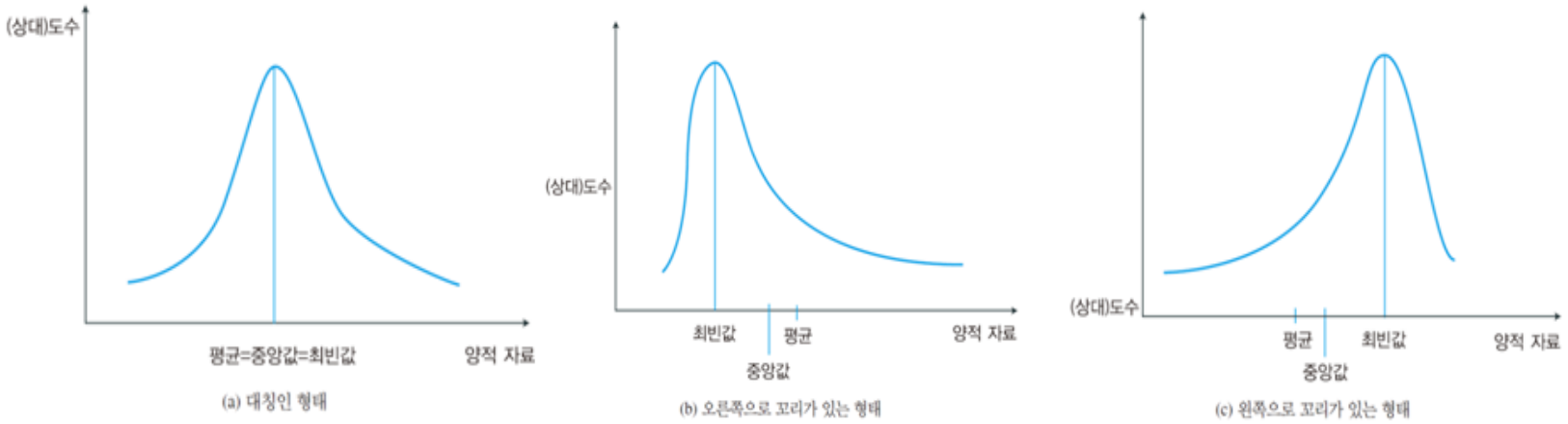
직책	인원(단위 : 명)	월급(단위 : 만원)
사장	1	4,500
전무	1	1,500
이사	2	1,000
실장	1	570
부장	3	500
과장	4	370
대리	1	300
사원	12	200
총원	25	

산술평균=570

중앙값=300

최빈값=200

자료의 중심측도간의 관계



비대칭도(skewness:왜도) :분포의 모양이 한쪽으로 치우쳐 지는 정도

$$\text{비대칭도} = S_k = \frac{3(\bar{x} - M_d)}{s}$$

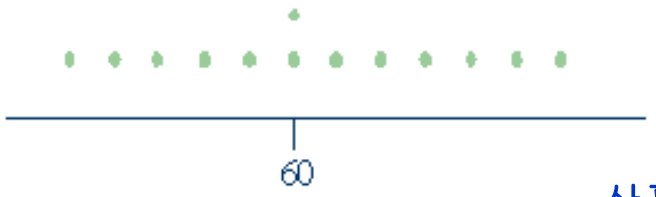
$S_k = 0$(대칭), $S_k < 0$(왼쪽 비대칭), $S_k > 0$(오른쪽 비대칭)

자료의 산포를 나타내는 측도

- 평균이 동일한 자료라 하더라도 평균을 중심으로 자료의 흩어진 정도가 다를 수 있기 때문에 대표값만으로 자료의 특성을 요약하는 것은 충분하지 않다. 따라서 대표값과 더불어 흩어진 정도를 나타내는 측도.

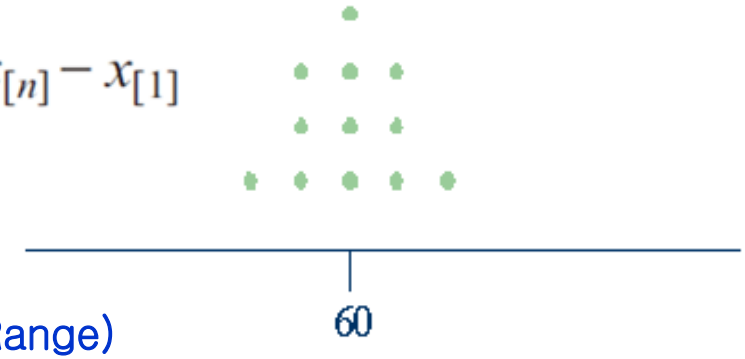
표본자료1	10	20	30	40	50	60	60	70	80	90	100	110
표본자료2	40	50	50	50	60	60	60	60	70	70	70	80

$$R = x_{\max} - x_{\min} = x_{[n]} - x_{[1]}$$



자료 1의 분포
평균=중위수=60

산포 : 범위(Range)



자료 2의 분포
평균=중위수=60

산포 측도 : 분산, 표준편차, 평균편차, 범위

모분산: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ ➡ 표본분산: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2\right] \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)
\end{aligned}
$$

표본표준편차: $s = \sqrt{s^2}$

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

➔ 표본분산을 구할 때 왜 (n-1)로 나누는가 ?

산포 측도 : 백분위수(percentile)

- 백분위수(percentile)

  전체자료를 100등분하는 값.

  ($n > 30$일 때) n개의 자료를 크기 순으로 나열하였을 때,

  제 p 백분위수는 $np/100$번째에 해당하는 자료값.

  $\Rightarrow$ 자료값 중 p%가 그 값보다 작거나 같고 $(1-p)$%가 크거나 같은 값.

제 p 백분위수 구하는 방법

① 관측값들을 크기순으로 정렬한다.

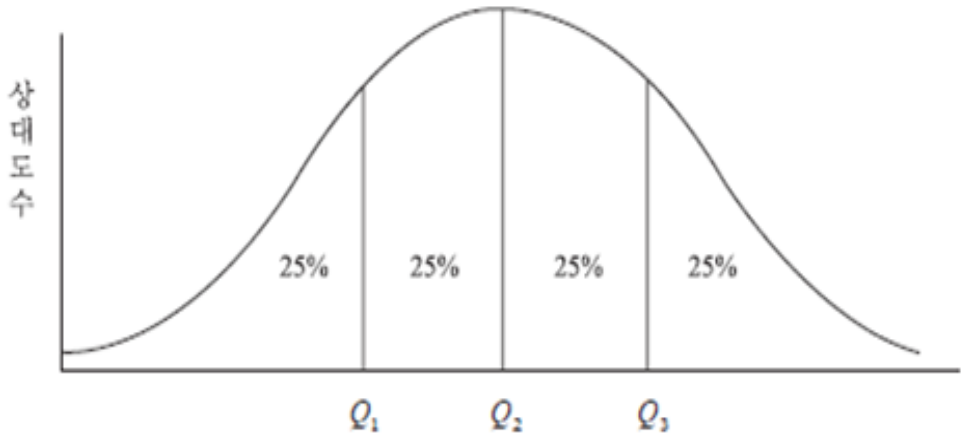
② 관측값의 개수 n에 $p/100$을 곱한다.

$$\text{제 } p \text{ 백분위수} = \begin{cases} \frac{np}{100} \text{ 가 정수인 경우} & = \left(\frac{np}{100}\right)\text{번째와 } \left(\frac{np}{100}+1\right)\text{번째의 평균} \\ \frac{np}{100} \text{ 가 정수가 아닌 경우} & = \left(\frac{np}{100}\text{의 정수 부분} + 1\right)\text{번째 값} \end{cases}$$

산포 측도 : 사분위수(quartile)

▶사분위수, 백분위수

- 사분위수(quartile)
  - ① 제 1 사분위수 Q_1은 전체자료에서 중위수보다 작은 자료값들의 중앙값
  - ② 제 2 사분위수 Q_2은 전체자료에서의 중위수
  - ③ 제 3 사분위수 Q_3은 전체자료에서 중위수보다 큰 자료값들의 중앙값



사분위수범위

$$IQR = Q_3 - Q_1$$

사분위수 : example

예제 : 어느 학과 학생들 25명의 시험 총점이 다음과 같이 크기순으로 나열되었다. 이 자료의 사분위수 Q_1, Q_2, Q_3을 구하여라.

35	37	38	41	47	50	50	53	55	58
61	65	69	70	71	72	74	79	83	85
90	93	95	97	99					

풀이

① $n = 25$이므로 $Q_2 = \frac{(25+1)}{2} = 13$번째 값 $\Rightarrow Q_2 = 69$

② Q_2보다 작은 자료값들이 12개이므로 $Q_1 = 6, 7$번째 평균 $= \frac{(50+50)}{2} = 50$

③ Q2 보다 큰 자료값들이 12개이므로 Q3 = 19, 20번째 평균 $= \frac{(83+85)}{2} = 84$

사분위구하는 방법 : 백분위수방법, Tukey 방법

변동계수 – CV(coefficient of variation)

- 변동계수(CV)란 두 조사자료의 단위가 다르거나(m, km), 단위는 같지만 평균의 차이가 너무 클 때 산포도를 비교하는데 사용.

$$\text{변동계수}(CV) = \frac{s}{\bar{x}}$$

날짜	1	2	3	4	5	6
A 회사주식	76,300	77,400	77,900	77,200	76,900	78,800
B 회사주식	7,400	7,000	7,400	6,900	7,300	7,600

A회사 – 평균 : 77417, 표준편차 : 861, 변동계수 : $861/77471 = 0.01112$

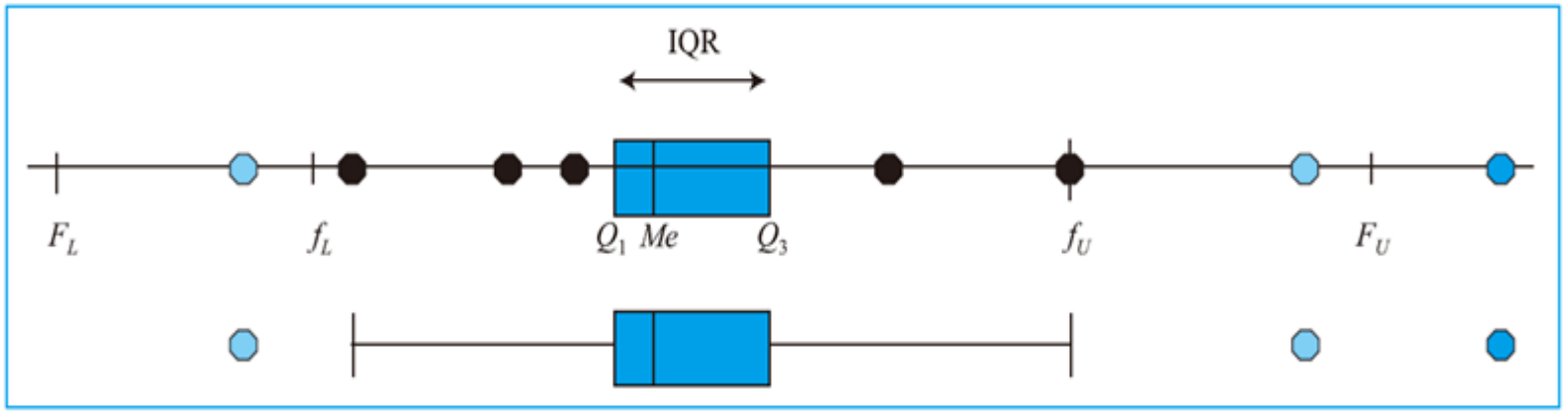
B회사 – 평균 : 7100, 표준편차 : 429, 변동계수 : $429/7100 = 0.06042$

A, B 주식의 변동성(산포)은 어느 주식이 크며, 어디에 투자하겠는가 ?

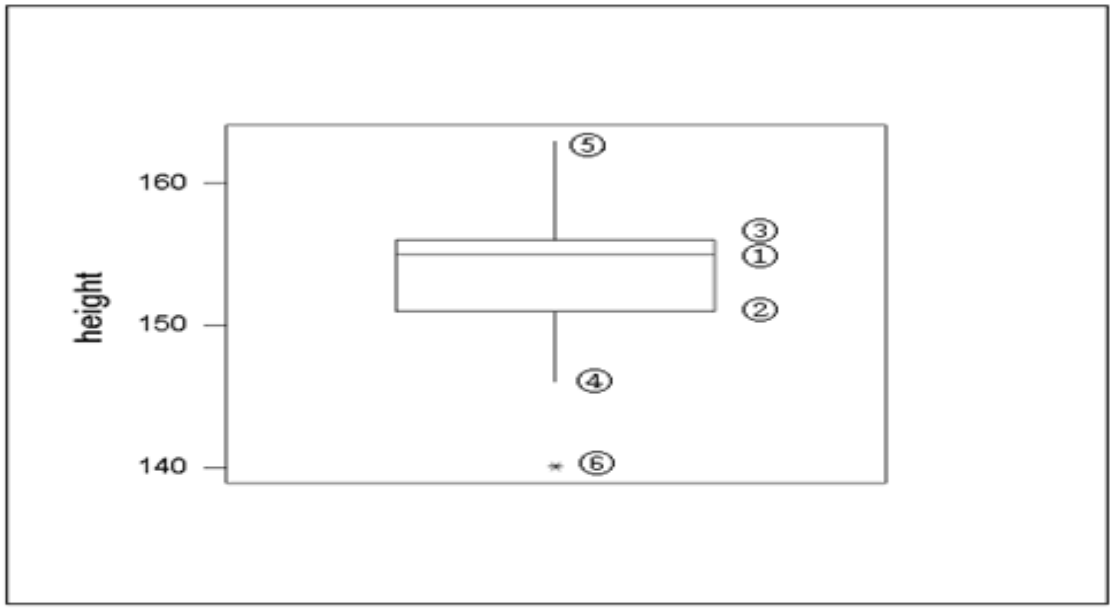
데이터 시각화 – 상자그림(box-plot)

상자그림

상자그림이란 자료들의 중심의 위치와 산포를 요약한 그림으로서, 제1사분위수, 중앙값, 제3사분위수로 상자를 그리고, 상자로부터 아래쪽 인접값과 위쪽 인접값까지 수염을 그린다. 그리고 울타리를 벗어나는 값들을 특이값으로 표시하여 상자그림을 완성한다.



상자도표-해석방법



① 중위수(M_e) - 자료를 크기순으로 나열했을 때 중앙에 해당되는 값

② 제1사분위수(Q_1) - 자료를 크기순으로 나열했을 때 25%째 해당되는 값

③ 제3사분위수(Q_3) - 자료를 크기순으로 나열했을 때 75%째 해당되는 값

④ 하한 값 - 사분위 범위(Q3-Q1)*1.5

⑤ 상한 값 - 사분위 범위(Q3-Q1)*1.5

⑥ 이상점 - 자료들 중 극단적으로 크거나 작은 즉, 특이하다고 판단되는 값

줄기-입 그림, 상자도표 – example

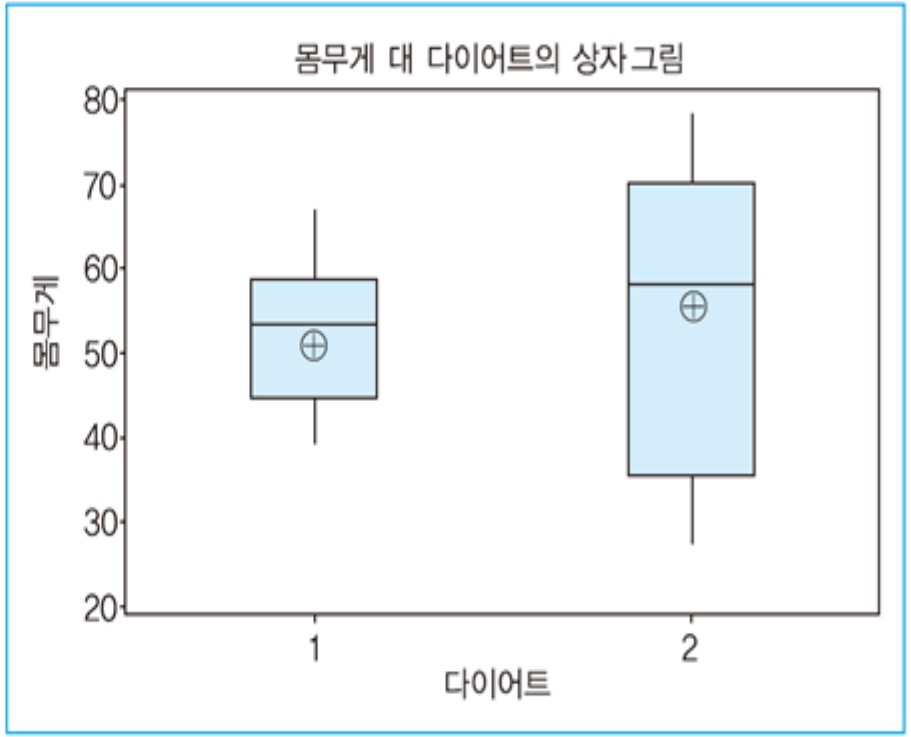
```
줄기-입 그림: weight  diet = 1    N  = 16
입 단위 = 1.0

 1  3  9
 4  4  024
 6  4  78
 8  5  02
 8  5  56789
 3  6  12
 1  6  7

줄기-입 그림: weight  diet = 2    N  = 13
입 단위 = 1.0

 1   2  7
 3   3  34
 4   3  7
 5   4  0
 5   4
 5   5
(2)  5  58
 6   6
 6   6  568
 3   7  2
 2   7  58
```

R-분석 결과의 예 : stem, boxplot



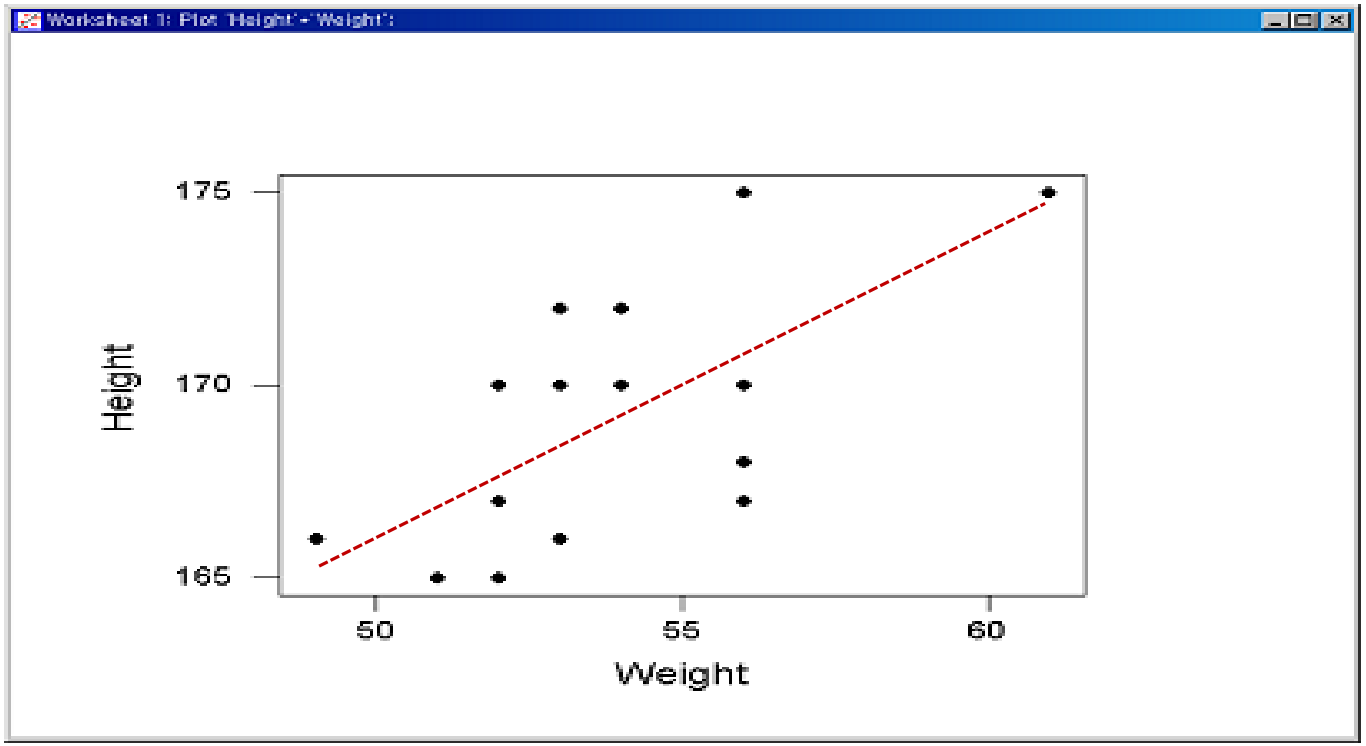
두 그룹 자료로의 확장

- 관련이 있는 두 그룹의 자료가 관측된 경우.

미스코리아 17명의 신장과 몸무게

신장	몸무게	신장	몸무게	신장	몸무게	신장	몸무게
165	51	168	56	170	56	175	56
170	54	167	56	166	49	172	53
167	52	170	52	172	54		
166	53	167	52	170	53		
165	52	175	61	172	54		

두 그룹 자료로의 확장(산점도)



신장과 몸무게의 산 점 도(scatter plot)

두 그룹 자료로의 확장(표본상관계수)

앞의 두 변수(몸무게, 신장)에 대한 자료를 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$라 하면 표본상관계수(sample correlation coefficient) r_{xy}는 다음과 같이 정의된다.

공분산(Cov(x,y))

상관계수(Corr(x,y)) ➡

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

|r| > 2/3 : 강한 상관성

|r| ~ 0 : 무상관

|r| ~ 1/2 : 보통의 상관성

|r| ~ 1/3 : 약한 상관성

두 그룹 자료로의 확장(표본상관계수)

Example) 다음과 같이 10명의 대학생들의 신장(x)과 몸무게(y)에 대한 자료를 얻었다. 이 자료를 바탕으로 두 변수에 대한 상관계수 r_{xy}를 구하여라.

신장(x)	162	163	166	168	169	171	173	174	175	179
몸무게(y)	54	56	56	64	62	64	82	67	71	74

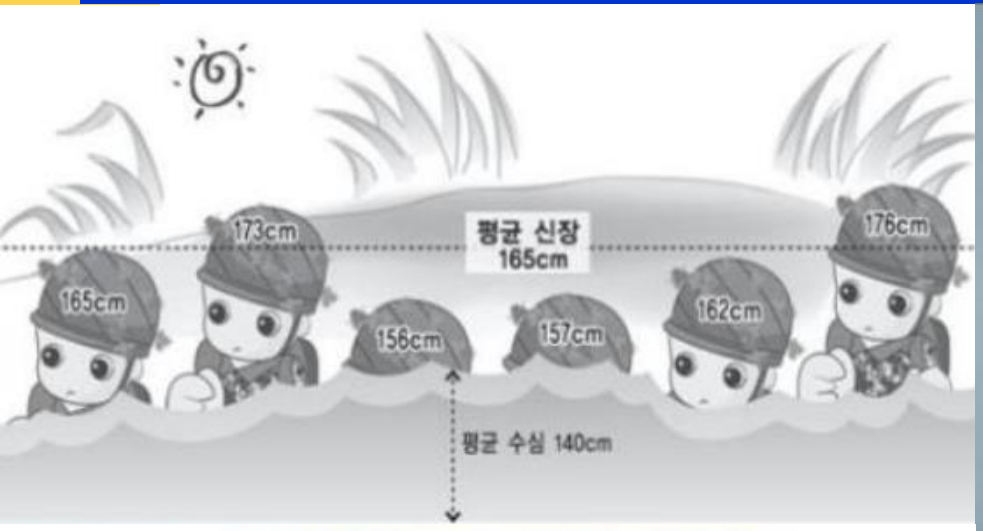
$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 170 \qquad \bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 65$$

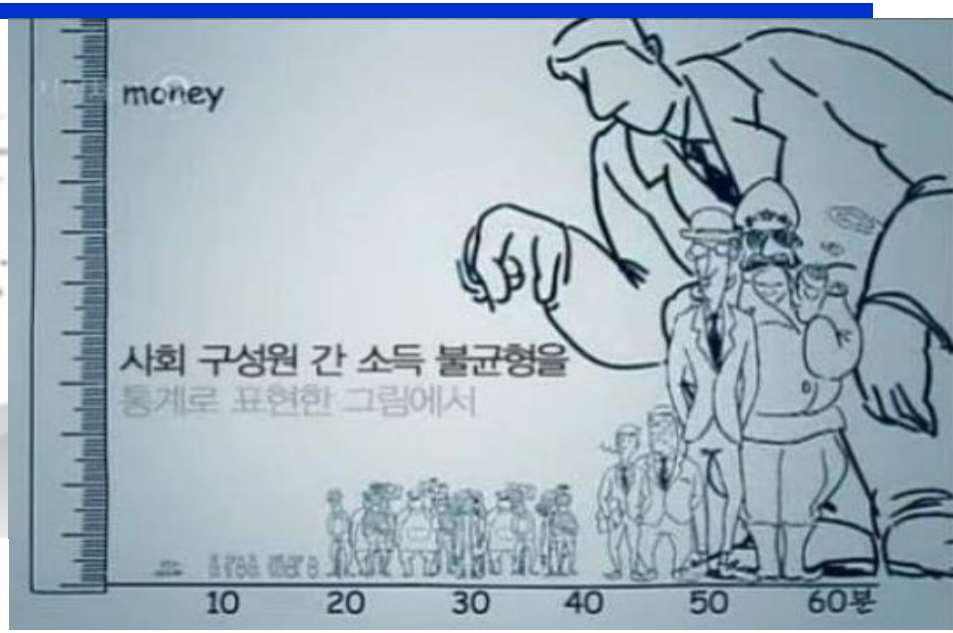
$$\sum_{i=1}^{10}(x_i - \bar{x})^2 = 266\,, \quad \sum_{i=1}^{10}(y_i - \bar{y})^2 = 704\,, \quad \sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 361$$

$$r_{xy} = \frac{361}{\sqrt{266}\,\sqrt{704}} = 0.8342$$

⇒ 강한 양의 상관이 있다.

다음 3가지 경우의 적절한 수치적요약





➢ 3 가지 그림 모두 평균으로 해석한다면 어떤 문제점이 있을까 ?