확률및통계

제3주 수치를 통한 연속형 자료의 요약

hylee@silla.ac.kr

수치를 통한 연속형 자료의 요약

- 그래프가 자료를 이해하기 위한 유용한 도구이기는 하나 주관적이고 일관성이 없는 결과를 도출할 가능성이 있다.
- 객관적이고 일관성 있는 결과를 위해서 수치가 사용된다.
- 자료 요약을 위한 측도의 종류로는 중심 위치의 측도, 퍼진 정도의 측도 등이 있다.
  - 중심 위치의 측도 : 평균, 중앙값, 최빈값
  - 퍼진 정도의 측도 : 분산, 표준편차, 범위, 사분위수범위

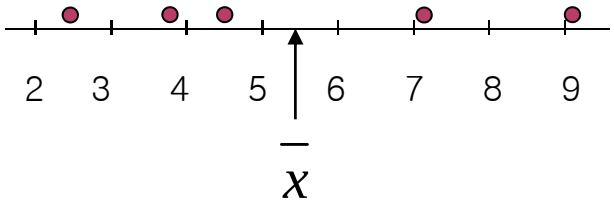
평균

- 평균 (Mean) : 가장 많이 쓰이는 중심위치의 측도
  - 자료 : $x_1, \ldots, x_n$
  - 표본평균

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

  - 예: 4.5 2.4 3.8 9.1 7.2

$$\sum x_i = 27 \qquad \bar{x} = \frac{1}{n}\sum x_i = 5.4$$



무게중심

중앙값 - 1

- 중앙값 (Median) : 자료를 순서대로 배열했을 때, 가운데 위치하는 값.
  - 자료 : $x_1, \ldots, x_n$
  - 크기 순으로 배열 : $x_{(1)}, \ldots, x_{(n)}$
  - 앞의 예에서 2.4 3.8 4.5 7.2 9.1
    - 자료의 개수가 홀수인 경우
      - 중앙값은 $\frac{n+1}{2}$ 번째 관측값 = $x_{((n+1)/2)}$
      - 예에서 자료의 개수는 5개이므로 3번째 작은 값 $x_{(3)} = 4.5$ 이 중앙값이다.

중앙값- 2

- 자료의 개수가 짝수인 경우
  - 중앙값은 $\frac{n}{2}$ 번째와 $\frac{n}{2}+1$ 번째 자료값의 평균이다.

$$\frac{x_{(n/2)} + x_{((n+1)/2)}}{2}$$

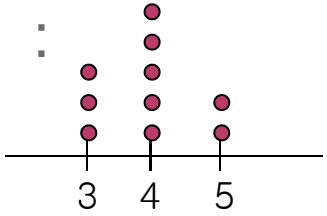
  - 예로써 5.0 의 값이 추가되었다면 자료의 개수가 6개이므로

2.4 3.8 4.5 5.0 7.2 9.1

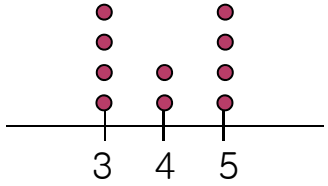
중앙값 : $\frac{x_{(3)} + x_{(4)}}{2} = \frac{4.5+5.0}{2} = 4.75$

최빈값

- 최빈값 (Mode) : 가장 자주 나오는 자료값.
  - 질적 자료나 그룹화한 수치 자료에서 자주 쓰임
  - 예 :



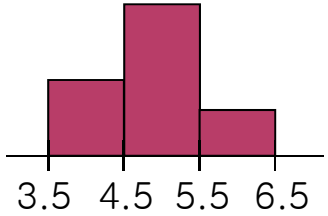
최빈값 = 4



최빈값 = 3,5

여러 개의 값이

될 수 있다.



최빈값= (4.5+5.5)/2= 5

표본평균, 중앙값, 최빈값의 특징

- 자료가 선형 변환될 때 평균과 중앙값도 같은 선형 변환으로 표현된다.
- 평균은 중앙값에 비해 아주 크거나 작은 값에 영향을 잘 받는다.
- 분포에 따라서 대표값들의 상대적인 위치가 다르다.

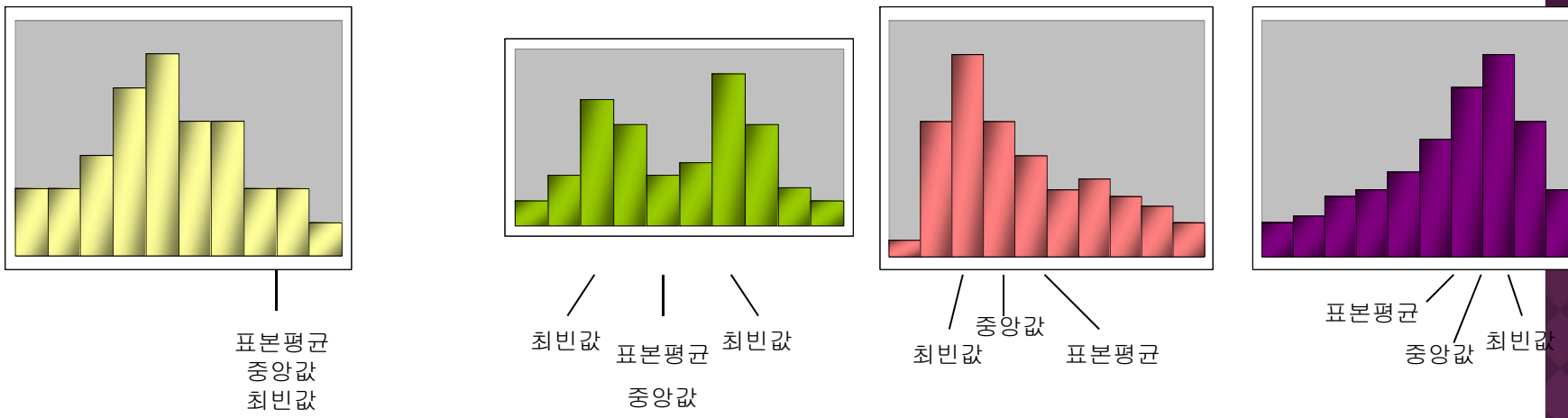


그림 : 분포의 모양에 따른 표본평균, 중앙값, 최빈값의 위치

분산과 표준편차

- 분산과 표준편차 (Variance, Standard deviation) : 자료가 얼마나 퍼져 있는가를 숫자로 표현한 것.
  - 편차 : 관측값 – 평균 = $x_i - \bar{x}$
    - 특징 : $\sum_{i=1}^{n}(x_i - \bar{x}) = 0, \quad \sum_{i=1}^{n} x_i - n\bar{x} = 0$
  - 표본분산

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

    - 계산식

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

  - 표본표준편차 :

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

선형 변환시 분산과 표준편차의 변화

$$y = ax + b \qquad s_y^{\,2} = a^2 s_x^{\,2} \qquad s_y = |a| s_x$$

$$s_y^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(ax_i - a\bar{x})^2 = \frac{a^2}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = a^2 s_x^{\,2}$$

$$\bar{y} = 59 \qquad y_i = 10x_i + 5$$

$$s_y^{\,2} = \frac{1}{4}(30^2 + 16^2 + 9^2 + 18^2 + 37^2) = 732.5 = 100 \times 7.325$$

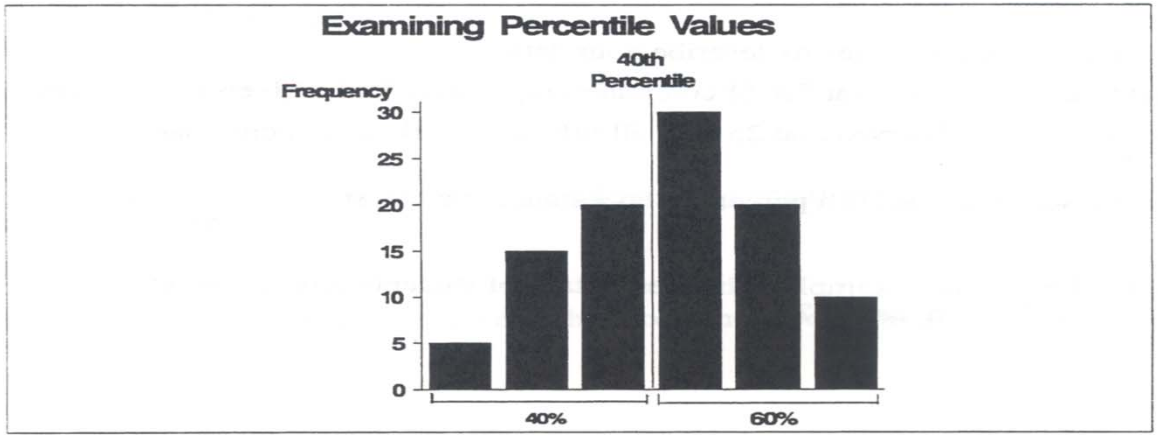
$$\sqrt{s_y} = \sqrt{732.5} = 27.065 = 10 \times 2.7065$$

◉ 예

- 자료 : 29 43 50 77 96

백분위수

- 제 $100p$-번째 백분위수 : 전체의 관측값을 $100(1-p)\%$ 와 $100p\%$ 로 나누는 값
- 즉, 관측값의 개수가 그 값 이하에 $n(1-p)$ 개 이상, 그 값 이상에 np 개 이상 있어야 한다.
- 구하는 방법
  - 자료를 크기 순으로 재배열한다.
  - np 가 정수가 아닌 경우 : 제 $100p$ 백분위수 = $x_{(\lfloor np \rfloor+1)}$ $\lfloor t \rfloor : t$의 정수부분
  - np 가 정수인 경우 : 제 $100p$ 백분위수 = $\dfrac{x_{(np)}+x_{(np+1)}}{2}$



사분위수와 사분위수범위

- 사분위수 (Quartile)
  - 제 1 사분위수 = 제 25 백분위수
  - 제 2 사분위수 = 제 50 백분위수 = 중앙값
  - 제 3 사분위수 = 제 75 백분위수

- 사분위수 범위 (IQR : Interquartile range)
  - IQR = 제 3 사분위수 – 제 1 사분위수
  - 자료가 29 43 50 77 96인 경우
    - 제 1 사분위수 Q_1 : $n=9 \quad p=0.25 \quad np=2.25 \quad Q_1 = x_{(3)} = 1.9$
      $n=9 \quad p=0.75 \quad np=6.75 \quad Q_3 = x_{(7)} = 3.4$
    - 제 3 사분위수 Q_3 : $IQR = Q_3 - Q_1 = 3.4 - 1.9 = 1.5$

표준편차와 사분위수 범위의 비교

- 표준편차는 평균의 장단점을, 사분위수 범위는 중앙값의 장단점을 갖는다.
- 즉 표준편차는 관측값의 퍼진 정도를 골고루 반영하나 극단적인 값에 크게 영향을 받는다.
- 사분위수 범위는 제**1**,제**3** 사분위수 바깥의 분포의 퍼져있는 정도는 반영이 안되고, 따라서 극단적인 값들에 영향을 받지 않는다.

기타 퍼진 정도의 측도

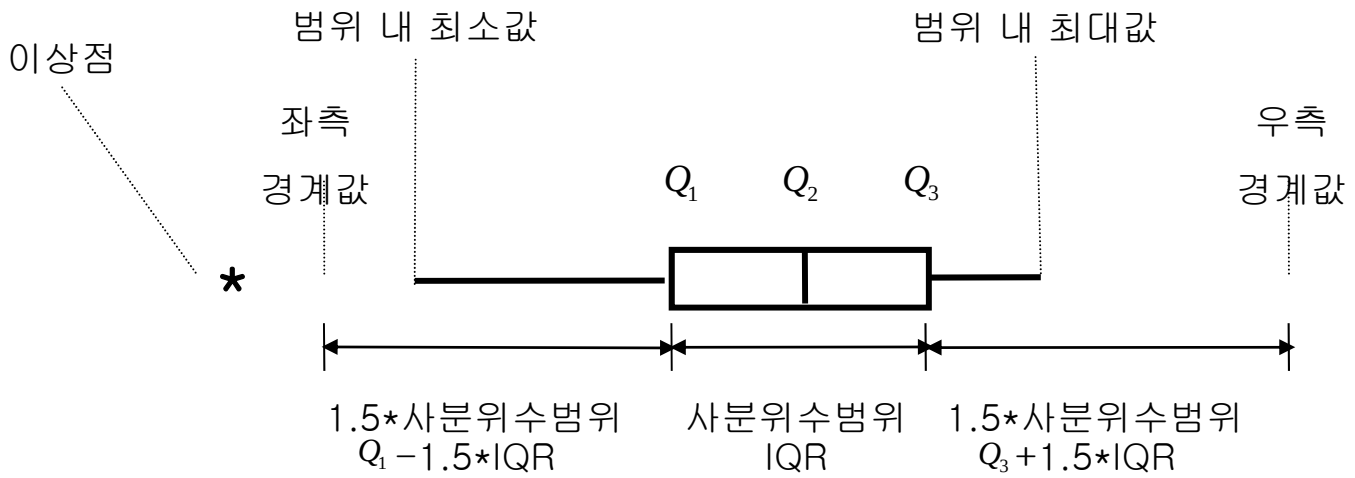
- 변동계수 (CV : Coefficient of Variation)

$$CV = \frac{s}{\bar{y}} \times 100$$

- 범위 (Range)
  - 범위 : 최대값 – 최소값
  - 표준편차와 사분위수 범위의 단점을 모두 가지고 있다. 퍼진 정도의 측도로 적절치 않다.

상자그림 (BOX PLOT)

- 중앙값, 사분위수 등을 이용하여 자료에 대한 정보를 하나의 그림으로 함축 시켜 그리는 요약 방법
- 방법
  - 최소사분위수 Q_1, 최대값을 구한다.
  - Q_3 , 를 상자로 값 연결하고 중앙값을 수직으로 표현
  - 상자의 양끝으로부터 1.5 * IQR 내에 있는 최소값과 최대값까지 선으로 연결
  - 양 경계를 벗어나는 자료 값들을 *로 표시

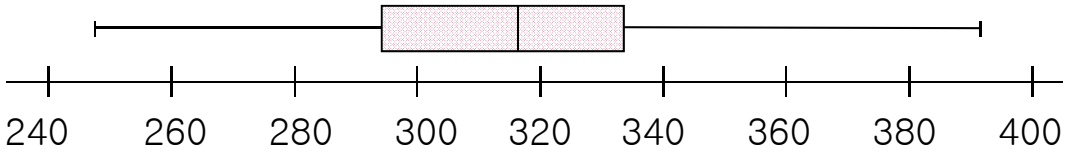


상자그림 예 - 1

$n = 20$ 248 260 270 274 295 301 308 310 315 315
320 325 332 333 334 334 356 368 370 388

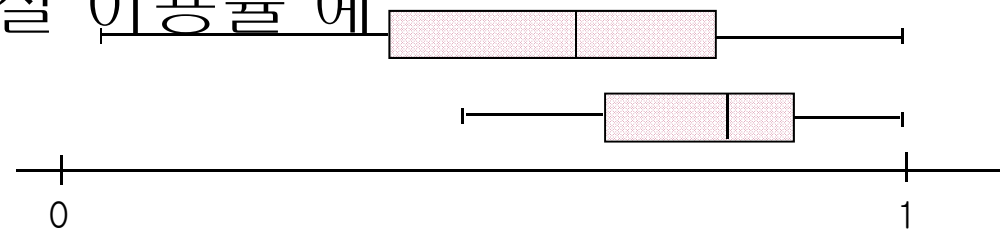
최소값=248, $Q_1 = \frac{295+301}{2} = 298,$ 중앙값= $= \frac{315+320}{2} = 317.5,$ $Q_3 = 334,$ 최대값=388

$IQR = Q_3 - Q_1 = 334 - 298 = 36,$ $1.5 \times IQR = 54,$ $298 - 54 = 244,$ $334 + 54 = 388$



상자그림 예 - 2

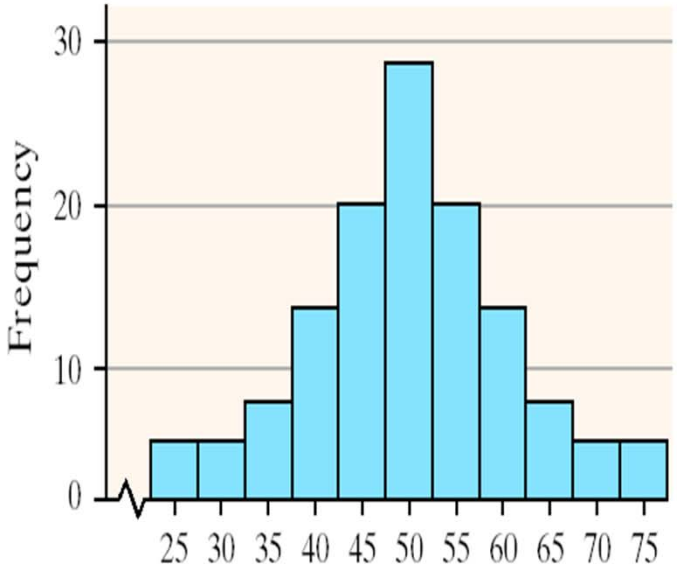
- 호텔 객실 이용률 예

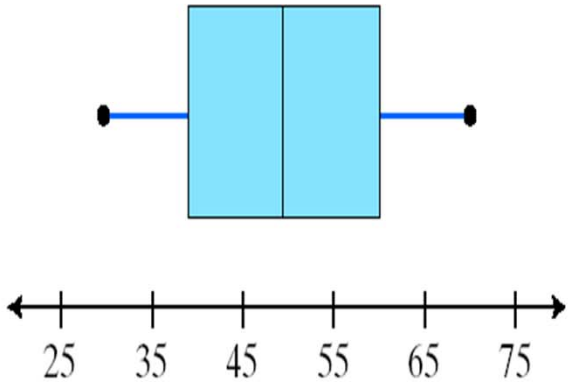


- 호텔 A의 이용률은 변화가 심하다.
- 호텔 B의 이용률이 A보다 상대적으로 높다.

히스토그램과 BOX PLOT 비교 - 1

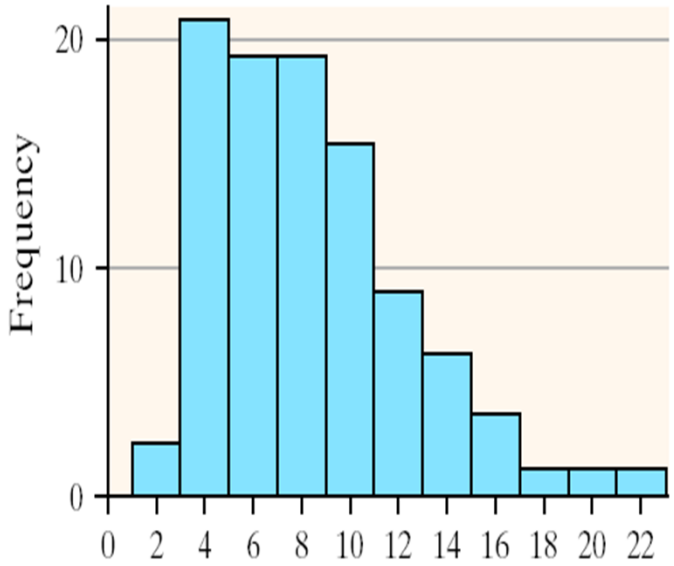
- 대칭

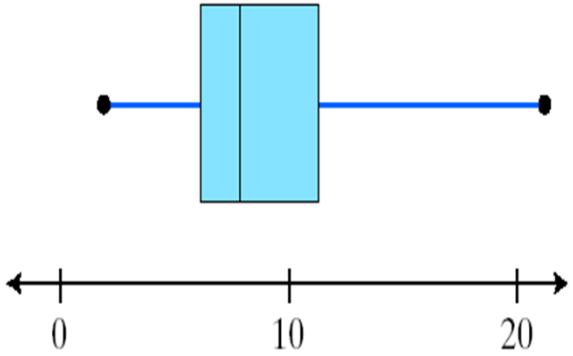




히스토그램과 BOX PLOT 비교 - 2

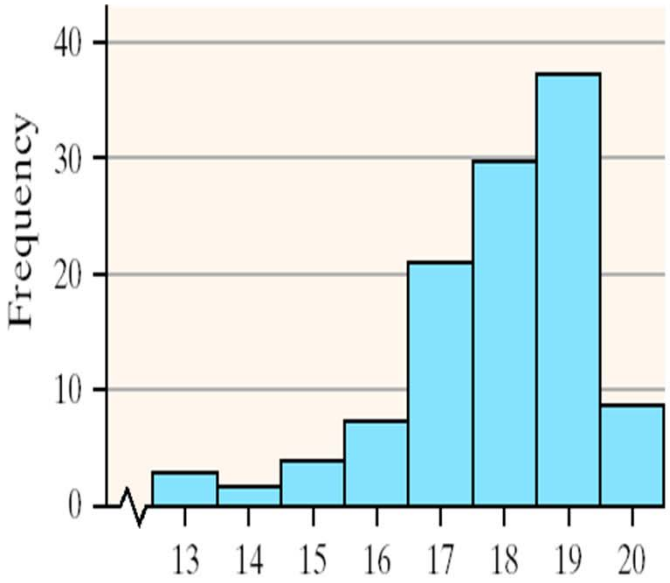
- 왼쪽으로 쏠림

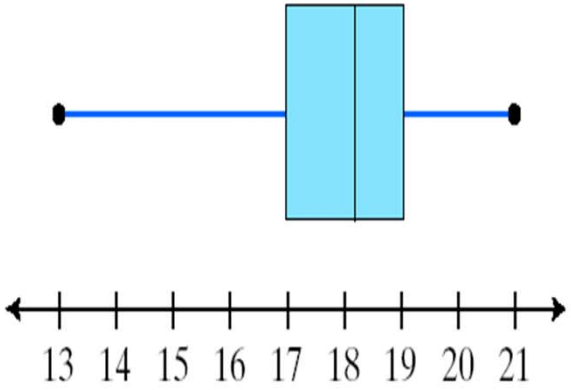




히스토그램과 BOX PLOT 비교 - 3

- 오른쪽으로 쏠림





그룹화된 자료의 요약 - 1

- 그룹화된 자료는 원 자료의 정보를 많이 상실하기 때문에 자료의 정확한 요약이 힘들다. 그러나 자료가 방대한 경우, 그룹화된 자료를 이용하여 비교적 간편하게 자료를 요약할 수 있다.

	계급값	돗수
$l_1 \sim u_1$	m_1	f_1
$l_2 \sim u_2$	m_2	f_2
$\vdots$	$\vdots$	$\vdots$
$l_k \sim u_k$	m_k	f_k
		n

계급값 : 각 계급의 중간값

$$m_i = \frac{l_i + u_i}{2}$$

- 각 계급에 포함되는 자료 값이 모두 계급 값을 l_1 u_1 갖는 것으로 간주하여 계산한다. 즉, 과 m_1 사이의 자료는 모두 이라는 값을 갖는 것으로 간주한다.

그룹화된 자료의 요약 - 2

- 평균 ($\overline{x_g}$)

$$\overline{x_g} = \frac{1}{n}\sum_{i=1}^{n} m_i f_i = \frac{1}{n}(m_1 f_1 + \cdots + m_k f_k)$$

- 분산 ($s_g^{\,2}$)

$$s_g^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(m_i - \overline{x_g})^2 f_i$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} m_i^{\,2} f_i - 2\overline{x_g}\sum_{i=1}^{n} m_i f_i + \overline{x_g}^{\,2}\sum_{i=1}^{n} f_i\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} m_i^{\,2} f_i - n\overline{x_g}^{\,2}\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} m_i^{\,2} f_i - \frac{\left(\sum_{i=1}^{n} m_i f_i\right)^2}{n}\right)$$

$$\sum f_i = n$$
$$\sum m_i f_i = n\overline{x_g}$$

그룹화된 자료의 요약 예 - 1

- 줄기 잎 그림:

9	**6 4 2 9 9 9**
10	3 8 6 9 5 4 2 6 3 0 1
11	1 3 1 0 9 1
12	1 2 8 3 8 1
13	0
14	0 7

최소값 : 92 최대값 : 147 범위 : 55 계급구간의 폭 : 10

그룹화된 자료의 요약 예 - 2

계급구간	계급값(m_i)	돗수(f_i)	$m_i f_i$
89.5 - 99.5	94.5	6	567
99.5-109.5	104.5	11	1149.5
109.5-119.5	114.5	6	687
119.5-129.5	124.5	6	747
129.5-139.5	134.5	1	134.5
139.5-149.5	144.5	2	289
		32	3574

$$\sum m_i^2 f_i = 405218$$

$$\overline{x_g} = \frac{\sum m_i f_i}{32} = 111.6875$$

$$s_g^2 = \frac{1}{31}\left(\sum m_i^2 f_i - \frac{3574^2}{32}\right) = 195.060$$

$$s_g = 13.97$$

$$\sum x_i = 3561$$

$$\overline{x} = 111.281$$

$$\sum x_i^2 = 401765$$

$$s^2 = 177.18$$

$$s = 13.31$$

그룹화된 자료의 요약 예 - 3

계급구간	계급값(m_i)	돗수(f_i)	$m_i f_i$	$m_i^2 f_i$
1-5	3	4	12	36
5-9	7	8	56	392
9-13	11	10	110	1210
13-17	15	8	120	1800
		30	298	3438

$$\overline{x}_g = \frac{298}{30} = 9.93$$

$$s_g^{\,2} = \frac{1}{29}(3438 - \frac{298^2}{30}) = 16.48$$

$$s_g = 4.06$$

종 모양의 대칭분포의 성질

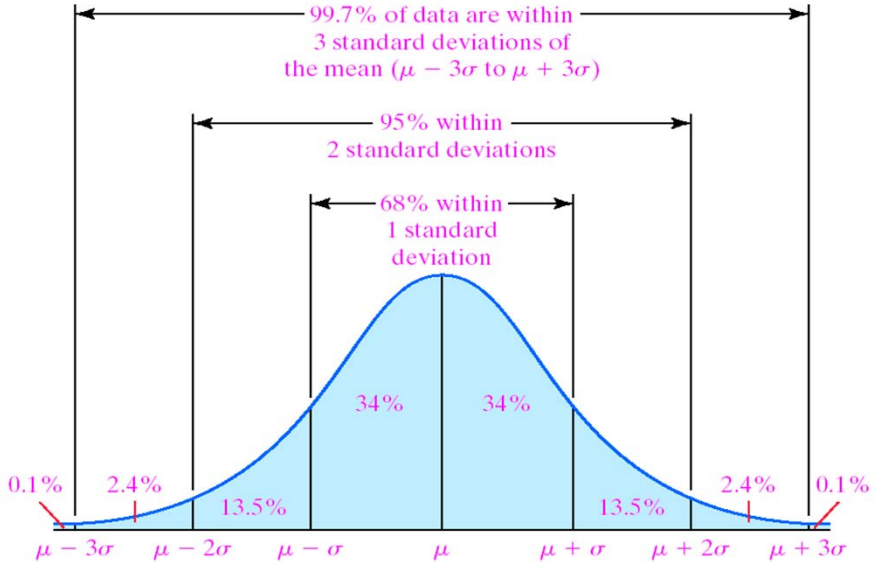
- 종 모양의 대칭분포 (Symmetric bell-shaped distribution)의 성질

$\bar{x}, s$ 가 주어져 있을 때,

$\bar{x} \pm s$ 구간에 **68%**

$\bar{x} \pm 2s$ 구간에 **96%**

$\bar{x} \pm 3s$ 구간에 **99.7%**의 자료가 포함된다.



관측값의 표준화

- z - 값 : 평균으로부터 떨어져 있는 상대적 위치의 측도

- x의 z -값 = $\dfrac{x-\bar{x}}{s}$

Thank You!