■ 수치를 이용한 자료정리

- ○ 그래프 같은 시각적 기법은 자료의 특성을 파악하는데 있어 중요한 정보를 제공하지만 그것을 보는 사람에 따라 주관적으로 해석될 수 있음
- ○ 자료분석의 최종 결과는 객관적으로 그 자료의 특성을 나타내는 수치로 제시

□ 중심위치

- ○ n개의 수치형 자료: $x_1, x_2, \ldots, x_n$
  - x_i는 i번째 표본의 값
  - n을 **표본크기(sample size)**
- ○ 중심위치로 가장 많이 사용되는 통계값은 표본평균이며 대체 통계값으로 중앙값, 절사평균, 최빈값 등이 있음

① 표본평균(sample mean)

◦ 표본평균은 표본의 합을 표본크기로 나눈 값

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

- $\overline{x}$는 x bar라고 읽는데 통계학에서 bar 표시는 해당 자료의 평균을 의미

◦ 평균을 중심으로 좌우의 무게가 같은 무게중심

◉ 오름차순으로 정렬된 자료 $x_1, \ldots, x_n$의 무게중심이 $\bar{x}$이고 $\bar{x}$의 좌측에 m개의 자료가, 나머지가 우측에 있다면

$$\sum_{i=1}^{m}(\bar{x}-x_i)=\sum_{i=m+1}^{n}(x_i-\bar{x})$$

○ $\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-n\bar{x}=0$가 되어 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

○ $x_i-\bar{x}$: i번째 표본의 **편차(deviation)**이며 편차의 합은 0

◉ 통계학 관련 학과 취업률

○ 통계학 관련 42개학과의 취업률의 합은 2486.4

$$\bar{x} = \frac{55.6 + 83.3 + \cdots + 41.2 + 56.3}{42} = \frac{2486.4}{42} = 58.77$$

○ 표본비율(sample proportion)

◦ 관측값이 어떤 범주에 속하면 x_i의 값을 1, 속하지 않으면 0으로 표시

◦ 전체 표본 중에서 이 범주에 포함된 표본의 수는 $y = x_1 + \cdots + x_n$ 이며 이 범주에 포함된 표본비율은

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

◦ **표본비율 또한 일종의 표본평균**으로 이해할 수 있음

◉ 통계학전공 학생의 취업률

○ 임의로 선택된 42개 통계학 관련 학과를 2010년 8월과 2011년 2월에 졸업한 1568명(남자 715명, 여자 853명) 중 취업대상자 1371명(남자 628명, 여자 760명)의 자료

○ 건강보험DB직장가입자와 해외취업자는 791명(남자 367명, 여자 423명)인 것으로 조사

【표 2.12】 통계학과 졸업생의 취업률

취업률	취업자	취업대상자	취업률
전체	791	1371	55.70%
남자	367	628	58.44%
여자	423	760	55.66%

○ 이상점(outlier)

- ○ 대부분의 관측값들에서 멀리 떨어져 있는 일부 관측값
- ○ 표본을 수집하는 과정에서 이상점이 자료에 포함되는 경우와 아닌 경우 표본평균을 비교해 보면 값이 차이가 크게 나는 경향이 있음 ⇨ 이상점에 로버스트(robust)하지 않음

◉ 8명의 졸업생의 초임월급 실수령액(단위 만원) 자료

200, 225, 210, 205, 205, 220, 350, 205

- ○ 8명의 수령액 합은 1820만원이고 평균은 $\bar{x} = \frac{1820}{8} = 227.5$
- ○ 문제는 8명 중 7명의 수령액이 평균보다 낮아 평균이 중심위치로 적절한가에 대한 의문
- ○ 이와 같은 결과는 자료 중 350만원이라는 값이 다른 자료와 너무 동떨어져 있어 평균의 값을 크게 만들었기 때문에 발생

- 표본추출과정에서 이상점이 포함될 수도 있고 안 될 수도 있는데 포함여부에 따라 결과에 차이가 크게 발생한다면 대푯값으로써 적절하지 않음
- 표본평균은 이상점에 영향을 많이 받기 때문에 자료에 이상점이 있는 경우 안정적인 중심위치로 적절하지 않음
- 중심위치로 표본평균을 사용하려면 계산 전에 자료에 이상점이 있는지를 먼저 확인

○ 표로 정리된 자료의 표본평균

◦ 원자료를 알 수 없기 때문에 정확한 표본평균을 계산할 수 없지만 근사적인 값은 구할 수 있음

【표 2.13】 수치형자료의 도수분포표

계급	도수	상대도수	누적 상대도수	계급 중간값	밀도
$[L_1,\ U_1)$	f_1	r_1	c_1	m_1	d_1
$[L_2,\ U_2)$	f_2	r_2	c_2	m_2	d_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_k,\ U_k]$	f_k	r_k	c_k	m_k	d_k

- 계급중간값 $m_j = (L_j + U_j)/2$을 구함
- m_j는 j번째 계급에 속하는 관측값들의 대표하는 값이고 $m_j f_j$는 해당 계급의 관측값의 합에 대한 근사값
- k개의 계급이 있는 경우 표본평균은

$$\overline{x}_g = \frac{1}{n}\sum_{j=1}^{k} m_j f_j = \sum_{j=1}^{k} m_j \left(\frac{f_j}{n}\right) = \sum_{j=1}^{k} m_j r_j$$

◉ 통계학 관련학과 취업률

【표 2.14】 2011년 통계학 관련학과 취업률

취업률	도수	상대도수	누적 상대도수	계급 중간값
10%이상~40%미만	3	0.071	0.071	25
40%이상~50%미만	6	0.143	0.214	45
50%이상~60%미만	13	0.310	0.524	55
60%이상~70%미만	10	0.238	0.762	65
70%이상~80%미만	6	0.143	0.905	75
80%이상~100%	4	0.095	1.000	90

$$\overline{x_g} = \frac{1}{42}(25\times 3 + 45\times 6 + \cdots + 90\times 4)$$
$$= 25\times 0.071 + 45\times 0.143 + \cdots + 90\times 0.095 = 60.01$$

○ 원자료를 이용하여 나온 표본평균 58.77와 비슷한 것을 볼 수 있다.

② 표본중앙값(sample median)

- 자료를 크기순서대로 나열했을 때 가운데 위치에 있는 값으로 표본중위수라고도 함
- **순서통계량(order statistics)** : 표본을 오름차순으로 정렬했을 때 i 번째로 작은 값을 $x_{(i)}$ 라고 하면 $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ 가 성립
  - 예) 만약 $n=5$ 이면 3번째 순서통계량 $x_{(3)}$, $n=6$ 이면 3번째와 4번째 순서통계량 사이인 $(x_{(3)}+x_{(4)})/2$ 을 표본중앙값이라고 정의

○ 표본중앙값의 일반식

$$\tilde{x} = \begin{cases} x_{(k_1)}, & n = \text{홀수} \\ \frac{1}{2}\left(x_{(k_2)} + x_{(k_2+1)}\right), & n = \text{짝수} \end{cases}$$

- $k_1 = (n+1)/2$이고 $k_2 = n/2$이다.

◉ 통계학 관련학과 취업률

○ 오름차순으로 정렬

19.6	22.7	31.6	40.5	41.0	41.2	41.3	43.4	46.3	50.0	52.4	52.8
53.1	53.8	54.8	55.6	55.6	55.6	56.5	58.1	58.6	59.5	60.7	61.9
63.6	64.3	64.5	64.6	65.2	65.4	66.7	67.9	71.4	71.4	72.1	73.3
77.1	78.4	80.0	81.3	83.3	91.3						

○ $n = 42$ 이므로 취업률의 표본중앙값은 21번째와 22번째 순서통계값 58.6과 59.5의 평균

$$\tilde{x} = \frac{58.6 + 59.5}{2} = 59.05$$

○ 표본중앙값은 극단적인 값에 영향을 받지 않음

- 예) 취업률 자료에서 19.6이 0으로 가거나 91.3이 100으로 가도 표본중앙값의 변화는 없음

⇨ 이상점의 유무에 관계없이 안정적은 중심위치를 제공한다는 것을 의미하며 이를 이상점에 로버스트(robust)하다고 함

○ 표본중앙값을 계산하는데 있어 자료의 값들은 순서통계량을 구하는데 이용될 뿐이고 중앙에 있는 하나 또는 두 개의 관측값만 직접 사용 ⇨ 자료가 가지고 있는 정보를 다 활용하지 못함

○ 어떤 값을 중심위치로 사용해야 하는가?

- 두 통계값을 계산하여 차이가 크지 않으면 표본평균을 차이가 크면 중앙값을 사용하는 방법을 제안 ⇦ 두 값의 차이가 크다는 것은 자료 중에 이상점이 있을 가능성이 높기 때문
- 일반적으로 임금이나 소득에 관련된 자료에는 이러한 이상점들이 종종 발생하기 때문에 대푯값으로 평균을 사용하면 체감하는 것보다 높게 느껴지는 경우가 있음

◉ 미국 실질임금

- 2013년 5월 1일 The New York Times에 F. Norris가 쓴 'Can Every Group Be Worse Than Average? Yes.'
- 물가를 보정한 미국인 실질임금의 중앙값은 13년 전보다 0.9% 증가
- 고용된 사람들을 교육수준에 따라 그룹은 나누어 2000년 대비 2013년 실질임금의 중앙값을 비교하면 모두 감소

그룹	고교 중태	고교 졸업	대학 중태	대졸 이상
증가율	-7.9%	-4.7%	-7.6%	-1.2%

⇨ 심슨의 역설(Simpson's paradox)

○ 13년 동안 각 그룹에 해당되는 인원에 변동으로 발생

- 그룹 내에서는 실질소득이 줄어 듬
- 상대적으로 보수가 높은 대졸이상 인원의 채용은 늘어나고 고졸 이하의 채용은 줄어 듬
- 전체적으로 임금이 상승한 것처럼 보임

③ 표본절사평균(sample trimmed mean)

- 표본평균은 모든 자료의 정보를 사용하지만 이상점에 로버스트 하지 않은 반면 표본중앙값은 로버스트하지만 자료의 정보를 다 활용하지 못한다는 장단점
- 두 통계값이 가지고 있는 장점을 살리면서 단점을 줄여주는 통계값
- α% 표본절사평균은 순서통계량의 하위 α%에서 상위 α%까지의 자료를 이용하여 표본평균을 계산
  - α**백분위수(percentile)** : 순서통계량에서 하위 α%의 값
  - $p = \alpha/100$이면 p**분위수(quantile)** = α백분위수

- 상위 α%의 값은 (100-α)백분위수 또는 $(1-p)$분위수
- 하위 α%에 해당되는 순서를 계산하면 상위 α%는 반대방향에서 동일한 순서에 해당되는 값
- 예) 30개의 자료가 있고 하위 α%가 4번째 순서통계량이었다면 상위 α%는 위에서 4번째인 27번째 순서통계량 ⇨ α% 표본절사평균은 4번째 순서통계량부터 27번째 순서통계량까지 24개 자료의 평균

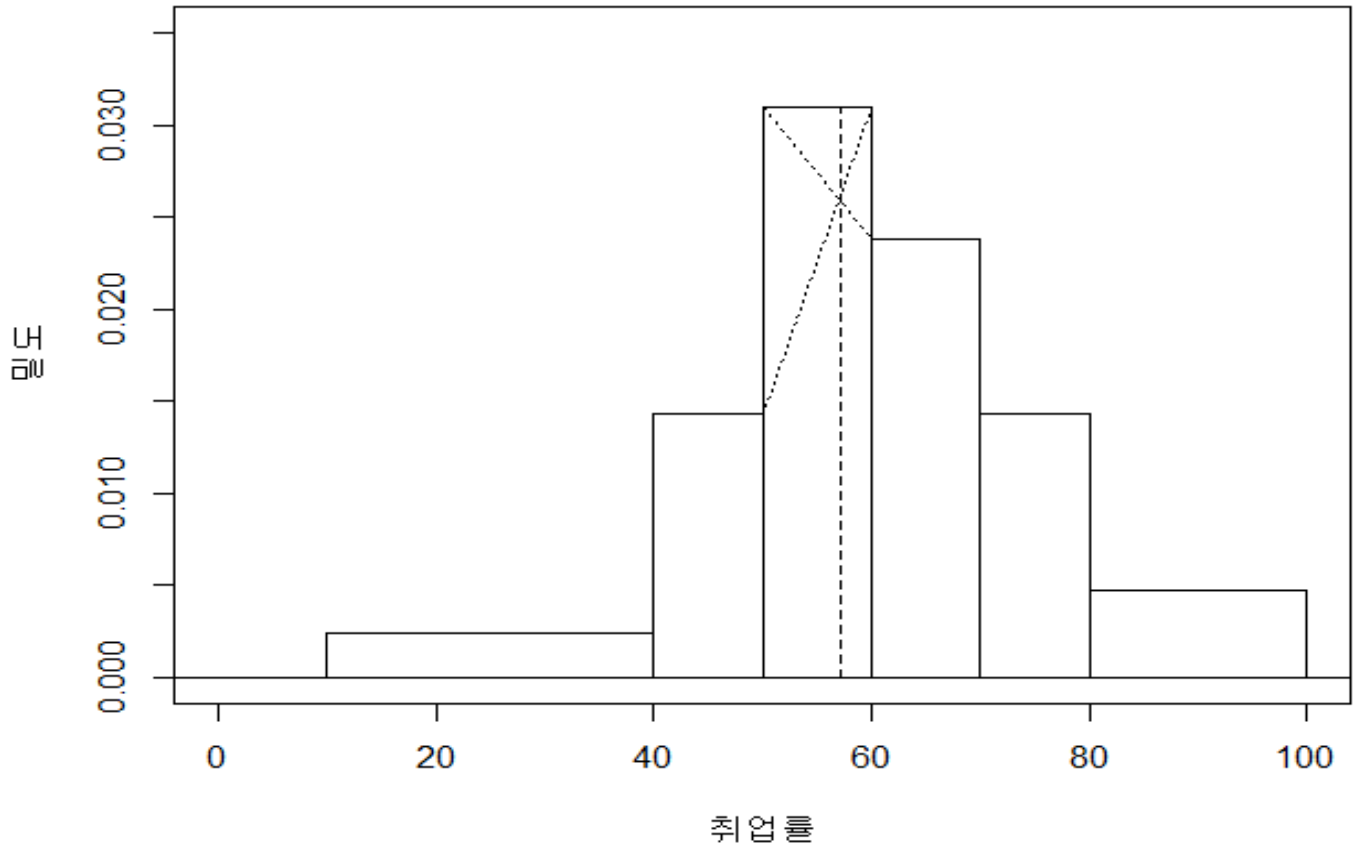
○ 적절한 크기의 α를 정하면 자료에 포함된 이상점이 제외시키면서 (100-2α)%만큼의 관측값을 사용 ⇨ 많은 자료정보를 사용하면서 로버스트한 중심위치를 제공

【표 2.15】 2013 ISU 여자피겨스케이트 점수표

선수	요소	심판									평균	절사 평균
		1	2	3	4	**5**	6	7	8	9		
김연아	Skating Skills	9.50	9.25	9.25	9.00	**8.75**	9.50	9.25	9.25	9.00	9.19	**9.21**
	Transition/Linking Footwork	9.00	9.25	8.75	8.75	**8.50**	9.25	9.00	8.75	8.75	8.89	8.89
	Performance/ Execution	10.0	10.0	9.00	9.25	**8.50**	9.50	9.75	9.00	9.00	9.33	**9.36**
	Choreography/ Composition	9.50	9.75	9.00	9.25	**8.50**	9.00	10.0	8.75	9.00	9.19	9.18
	Interpretation	10.0	10.0	9.25	9.00	**8.50**	9.25	10.0	9.00	9.00	9.33	**9.36**
아시다 마오	Skating Skills	8.50	8.75	8.50	8.25	**8.75**	8.50	8.50	8.75	9.00	8.61	8.61
	Transition/Linking Footwork	8.25	8.50	8.25	8.00	**8.50**	8.25	8.00	8.25	8.00	8.22	8.21
	Performance/ Execution	8.50	9.00	8.50	8.25	**8.75**	8.75	8.75	8.50	8.50	8.61	8.61
	Choreography/ Composition	9.00	9.00	8.50	8.50	**8.75**	8.50	8.75	8.50	8.50	**8.67**	8.64
	Interpretation	8.50	8.75	8.50	8.50	**9.00**	8.75	8.75	8.25	9.00	8.67	8.68

④ 표본최빈값(sample mode)

- 자료 중 빈도가 가장 많은 값
- 연속형 자료의 경우에는 자료의 값을 직접 사용하기보다는 그룹화하여 히스토그램을 그리고 간단하게 가장 높은 밀도를 가지는 구간의 중간값을 최빈값으로 사용하거나 내사법을 이용하여 가장 높은 밀도의 위치를 추정
- 여러 개 나올 수 있어 자주 사용하는 통계값은 아니지만 일봉 형태의 히스토그램에서는 가장 높은 밀도를 가지는 부분으로 중요한 위치



【그림 2.14】 통계학 관련전공 취업률 최빈값

□ 퍼짐의 측도

- ○ 중심위치만큼 중요한 통계값이 **산포(dispersion)**
- ○ 자료들이 얼마나 퍼져 있는가를 나타낼 뿐만 아니라 중심위치가 얼마나 안정적인지에 대한 중요한 정보를 제공
  - 자료가 조밀하게 모여 있는 경우 중심위치에 대한 정확도는 높아지지만 넓게 퍼져 있는 경우 중심위치의 변동성이 커지기 때문에 신뢰도가 떨어짐

① 범위(range)

◦ 자료 중 가장 큰 값과 작은 값의 차이

$$\text{범위} = x_{(n)} - x_{(1)}$$

- 예) 취업률 자료에서 최고 취업률은 91.3%이고 최저 취업률은 19.6%

  ⇨ 취업률 자료의 범위: 91.3% - 19.6% = 71.7%

◦ 표본은 최대값 $x_{(n)}$과 최소값 $x_{(1)}$을 계산하는데만 이용하기 때문에 많은 정보를 활용하지 못함

◦ 이상점이 있으면 전체 형태와 관계없이 범위가 클 수 있어 범위를 통해 퍼진 정도를 평가하기에는 무리가 있음

② 사분위범위(Interquartile-Range)

- **사분위수(quartile)** : 자료를 동일한 비율로 4등분 할 때의 세 위치
- 자료를 오름차순으로 정렬했을 때
  - 25% 지점: 제1사분위수(Q_1)
  - 50% 지점: 제2사분위수(Q_2) = 표본중앙값
  - 75% 지점: 제3사분위수(Q_3)
- 사분위(간)범위는 제3사분위수와 제1사분위수의 차이

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

○ 사분위수 계산 I

- $k = np$, $p = 0.25, 0.5, 0.75$ 계산
- k가 정수이면 $\left(x_{(k)} + x_{(k+1)}\right)/2$, 아니면 $x_{(k'+1)}$, k'는 k의 정수부분

◉ 취업률 자료

- $n = 42$이므로 $p = 0.25$일 때 k=42×0.25=10.5 ⇨ $Q_1 = x_{(11)} = 52.4$
- $p = 0.75$ 일 때 k=42×0.75=31.5 ⇨ $Q_3 = x_{(32)} = 67.9$
- $IQR = 67.9 - 52.4 = 15.5$

○ 사분위수 계산 II

- ◦ $k=(n-1)p+1, \quad p=0.25, 0.5, 0.75$ 계산
- ◦ k가 정수이면 $x_{(k)}$가 해당 사분위수, 아니면 비례에 의한 내사법을 적용

◉ 취업률 자료

- ◦ $n=42$이므로 Q_1에 해당되는 위치는 41×0.25+1=11.25
- ◦ 11번째와 12번째 순서통계값 사이에 있으며 비례식에 의해

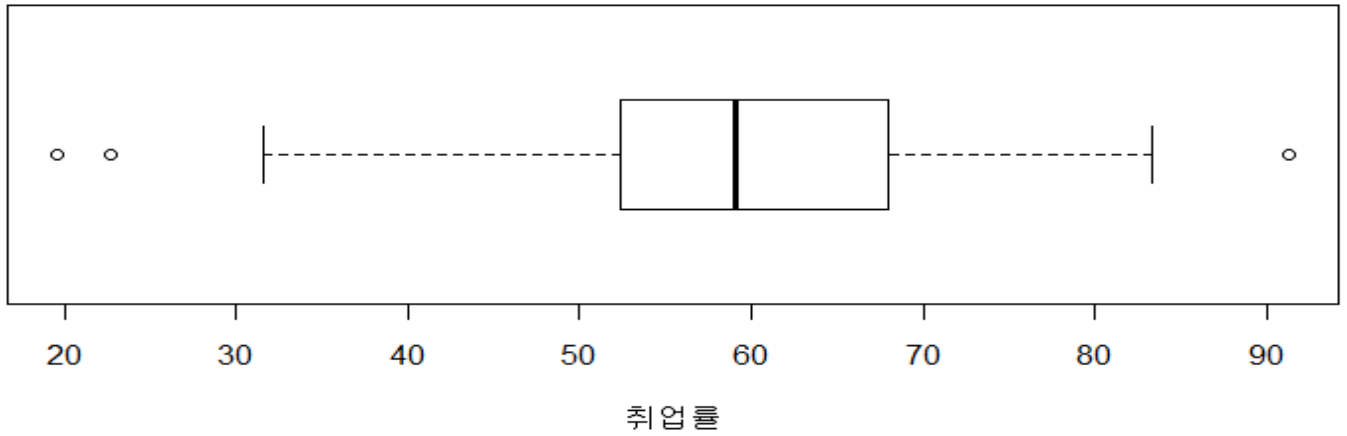
$$\begin{aligned} Q_1 &= 0.75\times x_{(11)} + 0.25\times x_{(12)} \\ &= 0.75\times 52.4 + 0.25\times 52.8 = 52.5 \end{aligned}$$

- ◦ Q_3는 31.75번째 위치로 0.25×66.7+0.75×67.9=67.6
- ◦ $IQR = 67.6 - 52.5 = 15.1$

○ 상자그림(box plot)

◦ Tukey라는 통계학자에 의해 제안된 그림

◦ 그룹 간의 비교나 이상점 검출 등에 사용되는 그림



【그림 2.15】 통계학 관련학과 취업률의 상자그림

◦ Q_1, Q_2, Q_3을 계산하여 직사각형의 상자를 표시

- $Q_1 = 52.5,\ Q_2 = 59.05,\ Q_3 = 67.6$
- Q_1과 Q_2의 거리가 Q_2와 Q_3의 거리보다 짧은 것은 Q_1과 Q_2 사이에 있는 자료가 좀 더 조밀하게 모여 있음

◦ $IQR,\ L = Q_1 - 1.5 \times IQR,\ U = Q_3 + 1.5 \times IQR$를 계산

- $L = 52.5 - 1.5 \times 15.1 = 29.85,\ U = 67.6 + 1.5 \times 15.1 = 90.25$

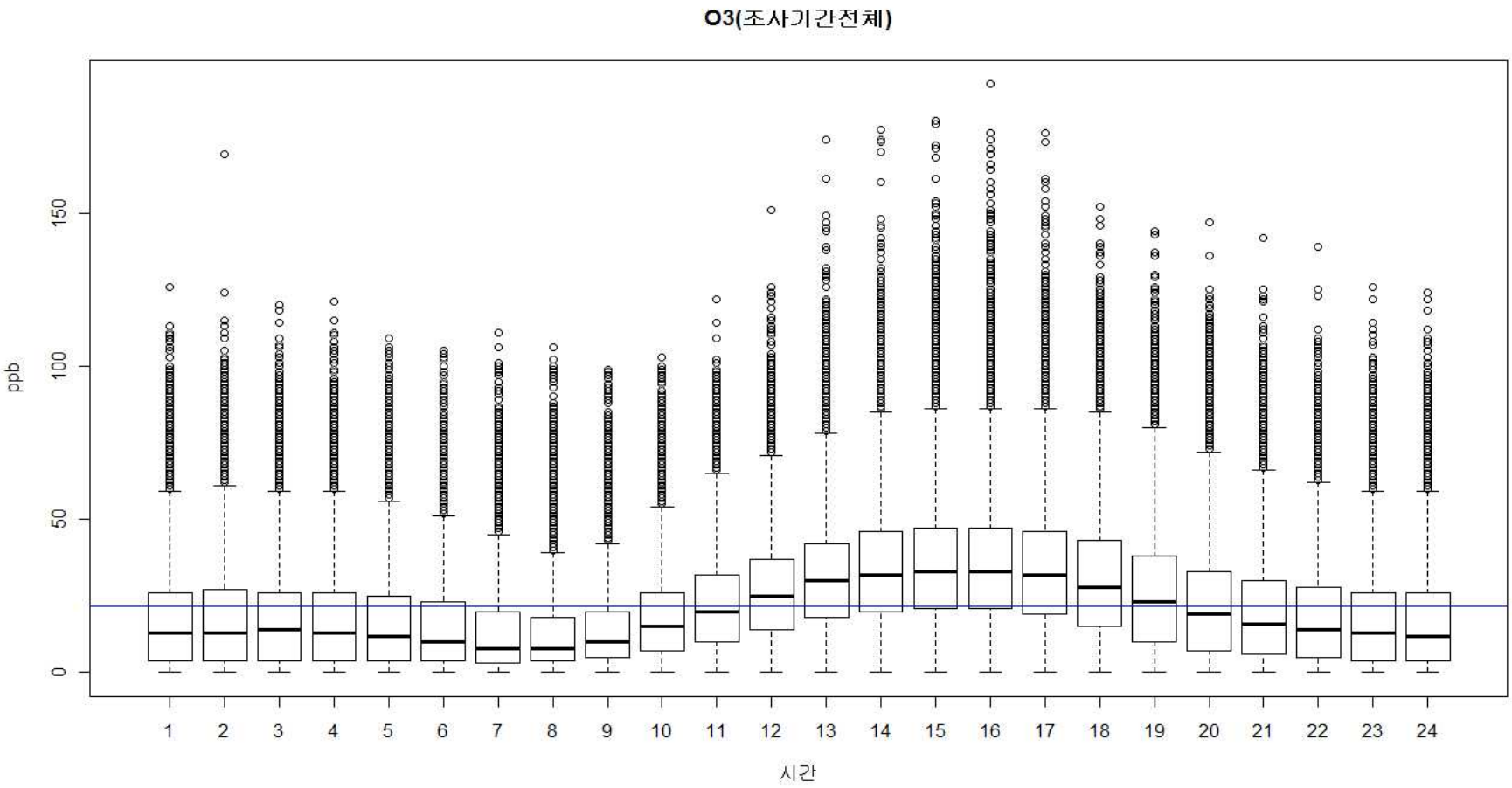
◦ L보다 큰 관측값 중 가장 작은 값, U보다 작은 관측값 중에 가장 큰 값에 직선에 직선을 표시하고 상자와 연결

- 31.6과 83.3에 직선표시

◦ 직선 밖의 관측값은 이상점으로 ○로 표시: 19.6, 22.7, 91.3

◉ 오존(O_3)자료

- 대기오염측정소의 주요 오염원에 대한 기초자료를 확보하기 위해 주요 100개 대기오염측정소를 선정
- 2005년 1월 1일 1시부터 2008년 12월 31일 24시까지 매 시간별로 측정된 오존(O_3)자료
- 모든 시간대에서 많은 이상점이 발견
- 오존의 오염도는 7시부터 9시까지 낮아졌다가 11시부터 급격히 증가하여 15시부터 17시경에 가장 높은 오염도를 가지다 서서히 감소하는 형태
- 중간의 실선은 전체 오존오염도 평균을 표시한 것



【그림 2.16】 시간대별 오존 오염도

③ 표본분산과 표본표준편차

- 범위나 사분위수범위의 경우 특정 위치의 두 값을 이용
- 모든 자료들 간의 거리의 합을 이용하는 방법은?
- **거리(distance)**: 임의의 점 a, b, c에 대해,
  - $a = b$이면 $D(a, b) = 0$이고 그 역도 성립
  - $D(a, b) = D(b, a)$
  - $D(a, b) \leq D(a, c) + D(c, b)$
  - 예) $D(a, b) = |a - b|$, $D(a, b) = (a - b)^2$

○ 모든 관측값들 간 거리의 합

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|, \qquad \sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2$$

- 자료들이 넓게 퍼져 있으면 이 합들은 커질 것이고 모여 있으면 작아짐
- n^2 개의 거리 합을 계산

○ 임의의 중심위치 a에서 자료들이 떨어져 있는 거리의 합

$$L_1(a) = \left|x_1 - a\right| + \left|x_2 - a\right| + \cdots + \left|x_n - a\right| = \sum_{i=1}^{n}\left|x_i - a\right|$$

$$L_2(a) = (x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_n - a)^2 = \sum_{i=1}^{n}(x_i - a)^2$$

○ **이 측도를 사용하기 위해서는 a값을 정해야 하는데 어떤 값으로 선택해야 할까?**

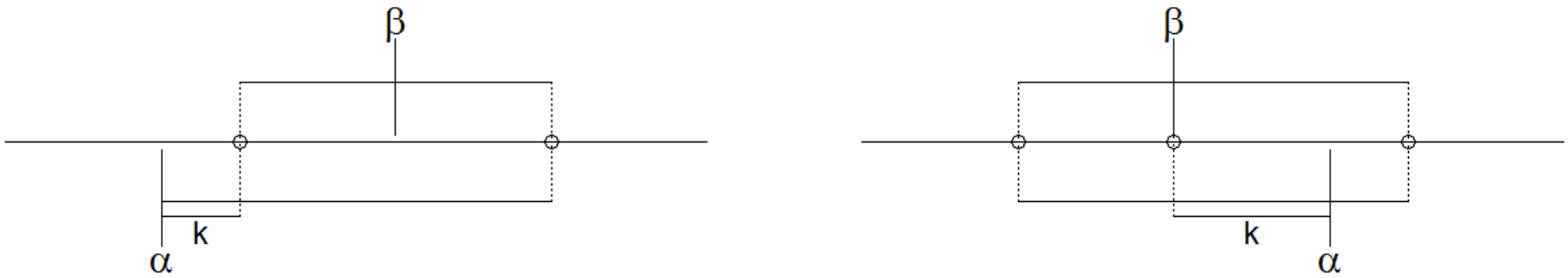
※ a가 좋은 중심위치가 되려면 자료들과의 거리가 가능한 짧아야 하며 결국 **거리의 합을 최소로 만드는 값**

○ $L_2(a)$를 a에 대해 미분한 식이 0이 되는 값

$$\frac{dL_2(a)}{da} = -2\sum_{i=1}^{n}(x_i - a) = -2\left\{\sum_{i=1}^{n}x_i - na\right\} = 0$$

$$\Rightarrow\ a = \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{x}$$

○ $L_1(a)$의 경우 a로 미분불가능



【그림 2.17】 절대편차합의 비교

○ 자료가 2개 있을 때, $L_1(\beta)$가 $L(\alpha)$보다 밑에 있는 선분의 길이 k의 두 배 작음

○ 자료가 3개인 경우에는 $L(\beta)$가 $L(\alpha)$보다 k만큼 작음

⇨ $L_1(a)$를 최소로 만드는 a는 표본중앙값

○ 퍼져있는 정도를 나타내는 통계값

- $L_1(\tilde{x}) = \sum_{i=1}^{n} \left| x_i - \tilde{x} \right|$

- $L_2(\overline{x}) = \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2$ ⇦ 편차의 제곱합

○ 편차의 합이 0 ⇦ $\frac{dL_2(a)}{da} = 0$ 를 만족하는 $a = \overline{x}$

○ 표본분산(sample variance)

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

- 표본분산은 n개의 편차를 사용하는 것 같지만

  $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$이라는 제약조건 때문에 $n-1$개의 편차 정보를 사용

- $n-1$: 자유롭게 가질 수 있는 편차의 개수라고 해 **자유도(degree of freedom)**이라고 함

○ 표본분산 간이식

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

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

○ 표본표준편차(sample standard deviation)

- $x^2+y^2=r^2 \Rightarrow \sqrt{x^2+y^2}=r$
- 표본분산은 편차의 제곱합을 이용하기 때문에 분산의 단위는 관측값 단위의 제곱
- 눈으로 이해하는 산포와 일치하기 위해서는 자료를 측정할 때의 단위로 표시

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

◉ 취업률 자료

○ 표본의 합과 제곱합

$$\sum_{i=1}^{42} x_i = 2468.4, \quad \sum_{i=1}^{42} x_i^2 = 154975.4$$

○ 편차의 제곱합

$$\sum_{i=1}^{42} (x_i - \overline{x})^2 = 154975.4 - \frac{2468.4^2}{42} = 9904.006$$

○ 표본분산과 표본표준편차

$$s^2 = \frac{9904.006}{41} = 241.56, \quad s = \sqrt{241.56} = 15.54$$

○ 표준화

◦ 수능시험은 과목별로 난이도가 다를 수 있기 때문에 원점수로 과목 간 성적을 비교 X ⇨ 표준화점수

$$z_i = \frac{x_i - \bar{x}}{s}$$

◦ $\sum_{i=1}^{n} z_i = \frac{1}{s}\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \Rightarrow \bar{z} = 0$

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

○ 표준화는 평균을 0, 표준편차를 1이 되도록 만들어 측정 단위에 영향을 받지 않게 중심위치와 척도(scale)를 조정하고 상대적 비교가능

○ 원자료 대신 도수분포표 형태로 주어진 경우

- m_i: i번째 계급의 중간값
- f_i: i번째 계급의 도수

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 \simeq \sum_{i=1}^{k}\left(m_i-\overline{x}_g\right)^2 f_i$$

- 표본분산과 표본표준편차

$$s_g^2=\frac{1}{n-1}\sum_{j=1}\left(m_j-\overline{x}_g\right)^2 f_i\,, \quad s_g=\sqrt{s_g^2}$$

◉ 취업률 자료

$$
\begin{aligned}
s_g^2 &= \frac{1}{42-1}\left\{(25-60.01)^2\,3+(45-60.1)^2\,6+\ \cdots\ +(90-60.1)^2 4\right\} \\
&= \frac{10550}{41} = 257.32 \\
s_g &= \sqrt{257.32} = 16.04
\end{aligned}
$$

④ 변동계수(coefficient of variation)

○ 표준편차가 평균에 영향을 받는 경우

- 예) 후진국의 소득분포와 선진국의 소득분포를 비교
- 예) 유아와 성인의 신장이나 체중의 분포를 비교

⇨ 비교 그룹간의 평균이 큰 차이가 있고 자료의 특성이 평균이 커지면 산포도 커지는 경향이 있기 때문

○ 실제로 0을 하한으로 가지는 많은 자료들이 이러한 성질을 가지고 있음

○ 표준편차만 이용하여 산포를 비교하는 것은 적절하지 않을 수 있어 평균으로 표준편차를 보정

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

- 100을 곱하는 이유는 표본평균에 비해 표본표준편차가 얼마나 큰지를 % 개념으로 표시하기 위한 것으로 100을 생략할 수도 있음

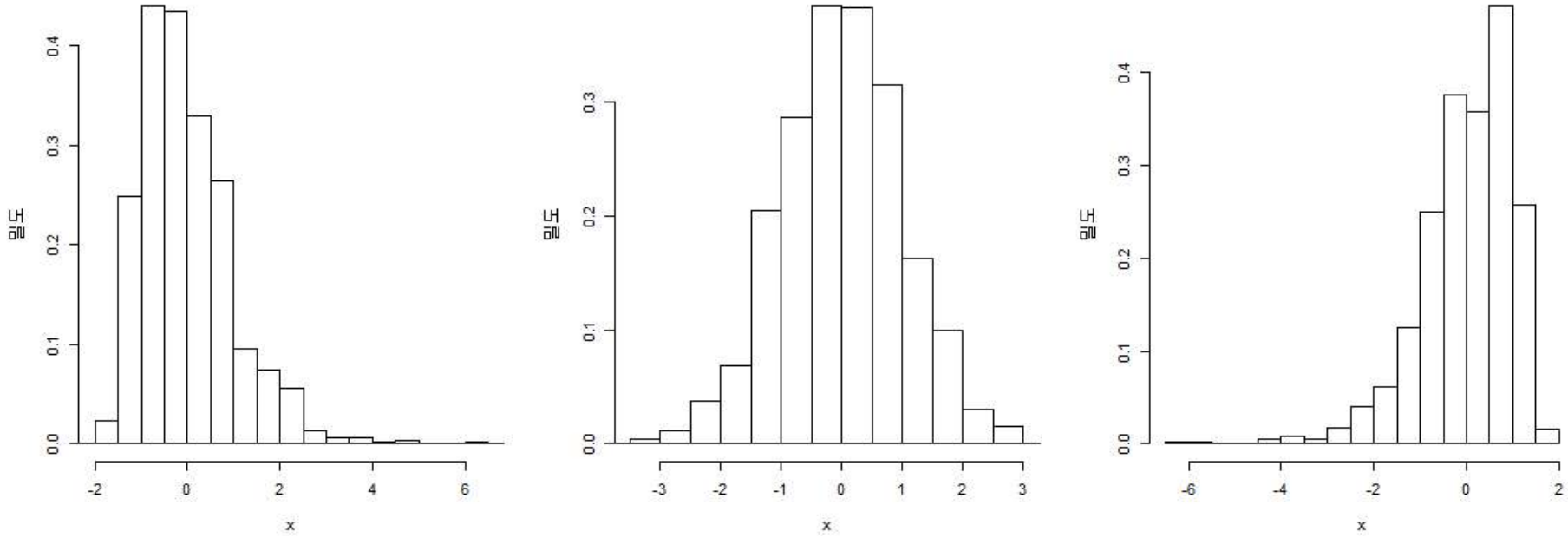
○ 신장과 체중과 같이 단위가 전혀 다른 자료들의 퍼져있는 정도를 비교할 때에도 사용

□ 분포의 형태

- ○ 수치형 자료에 대한 통계분석 방법은 대부분 모집단이 중심위치를 기준으로 좌우대칭인 형태를 가진다고 가정
- ○ 통계분석의 적절성은 분석방법에서 가정한 조건을 자료가 얼마나 만족하고 있는지에 영향을 받음
- ○ 자료의 분포 형태에 대한 측도이면서 자료가 모집단의 가정을 얼마나 만족하는지에 대한 측도로 사용

○ 왜도(skewness)

- ○ 자료가 중심위치를 기준으로 대칭적으로 분포되어 있는지는 히스토그램이나 상자그림을 통해 확인



【그림 2.18】 히스토그램

◦ 【그림 2.18】은 모두 평균이 0이고 표준편차가 1이지만 형태가 다른 자료의 히스토그램

◦ 왜도: 피어슨(Karl Pearson) 제안

$$\sqrt{b_1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

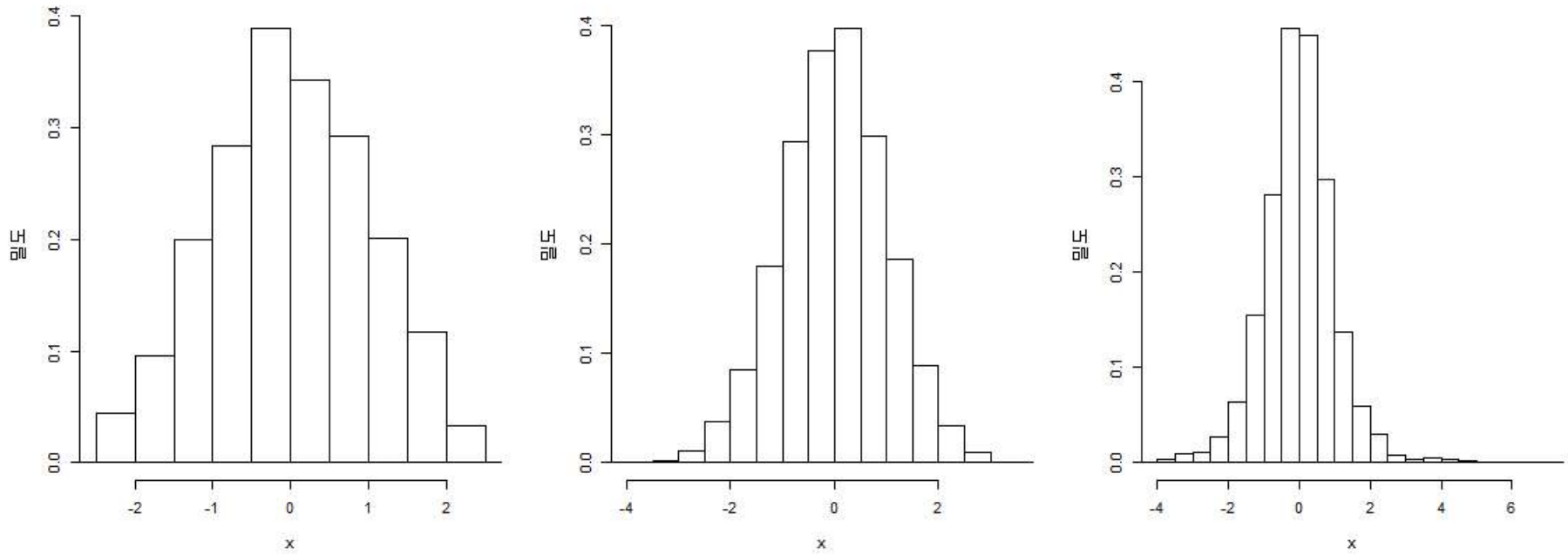
- $(x_i - \overline{x})^3$: 평균을 중심으로 왼쪽은 음수, 오른쪽은 양수
- 자료가 평균에서 멀어질수록 큰 음수나 큰 양수가 됨

◦ 좌우가 비슷한 형태를 가진다면 음수와 양수가 상쇄되어 $\sqrt{b_1}$은 0 근처 ⇨ 대칭적(symmetric)

- 왼쪽 그림: 오른쪽의 꼬리부분이 길게 부분되어 있어 큰 양수값을 가지는 자료가 있어 $\sqrt{b_1}$은 대칭일 때 보다 큰 값을 가짐 ⇨ 양의 왜도(positive skewness)를 가짐 또는 오른쪽으로 왜도(skewed to the right)됨
- 오른쪽 그림: 대칭일 때 보다 작은 값 ⇨ 음의 왜도(negative skewness)를 가짐 또는 왼쪽으로 왜도(skewed to the left)됨
- 두터운 꼬리(heavy tail) : 꼬리가 길게 분포된 것
- 수정된 왜도: $\sqrt{b_1} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3$

○ 첨도(kurtosis)

◦ 양쪽꼬리가 얼마나 두터운지를 나타내는 값



【그림 2.19】 히스토그램

○ 【그림 2.19】은 평균이 0이고 표준편차가 1인 자료의 히스토그램

○ 모두 대칭적인 형태를 가지나 꼬리가 왼쪽은 짧고 오른쪽은 길며 중간은 중간정도

○ 첨도: 피어슨(Karl Pearson) 제안

$$b_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4$$

- $(x_i - \overline{x})^4$: 평균을 중심으로 자료가 멀리 있으면 큰 값
- 항상 양수가 되며 분포의 중심보다는 꼬리부분이 얼마나 두터운지에 따라 영향을 많이 받음

○ 정규분포의 경우 이론적으로 첨도는 3

$$b_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - 3$$

○ 수정된 첨도

$$b_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

○ 심한 왜도를 가지거나 양쪽 꼬리가 두터운 경우에는 자료 중 이상점이 있을 가능성이 높아짐

○ 왜도나 첨도는 자료의 분포 형태를 나타내는 측도뿐만 아니라 분석 방법의 적절성을 확인하기 위한 측도로 사용

◉ 취업률 자료

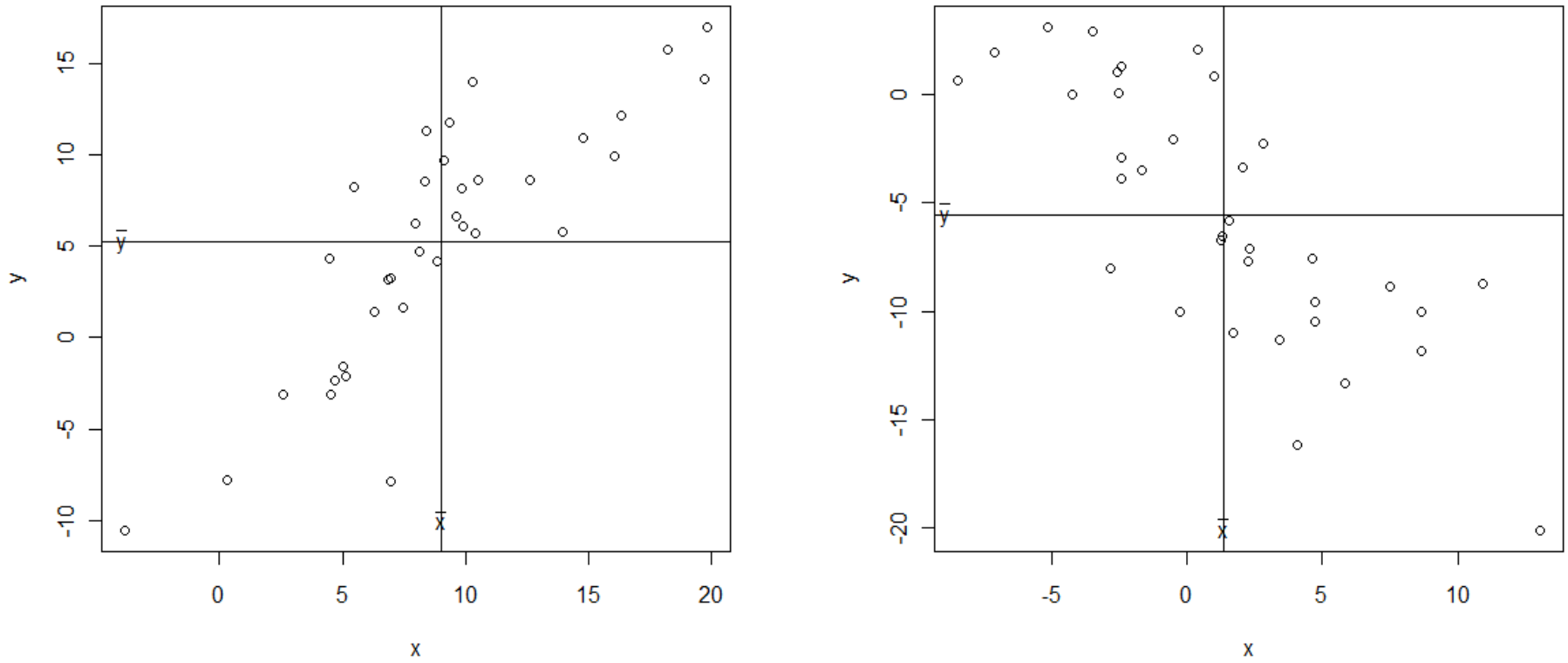
- $\sum x_i = 2468.4,\ \sum x_i^2 = 154975.4,\ \sum x_i^3 = 10211388,$
  $\sum x_i^4 = 699463185$
- 피어슨 왜도와 첨도

$$\sqrt{b_1} = \frac{1}{n-1}\sum\left(\frac{x_i - \overline{x}}{s}\right)^3 = -\frac{16.22}{41} = -0.396$$

$$b_2 = \frac{1}{n-1}\sum\left(\frac{x_i - \overline{x}}{s}\right)^4 = \frac{127.37}{41} = 3.107$$

■ 공분산과 상관계수

- 산점도를 통해 두 수치형 변수 간에 관계가 있는지를 시각적으로 확인
- 두 수치형 변수 간에 **직선관계**가 어느 정도인지를 나타내는 통계값
- 자료: $(x_1, y_1), (x_2, y_2), \ \ldots\ , (x_n, y_n)$



【그림 2.20】 양의 기울기와 음의 기울기를 가지는 산점도

- 양의 기울기를 가지는 경우 $(\overline{x}, \overline{y})$를 중심으로 1과 3사분면에 자료들이 많고 길게 분포
- 음의 기울기를 가지는 경우 대부분의 자료들이 2와 4사분면에서 길게 분포
- 자료의 직선관계를 표시하고자 할 때 $(\overline{x}, \overline{y})$을 중심으로 1과 3, 2와 4사분면의 자료가 동일한 성질을 가짐

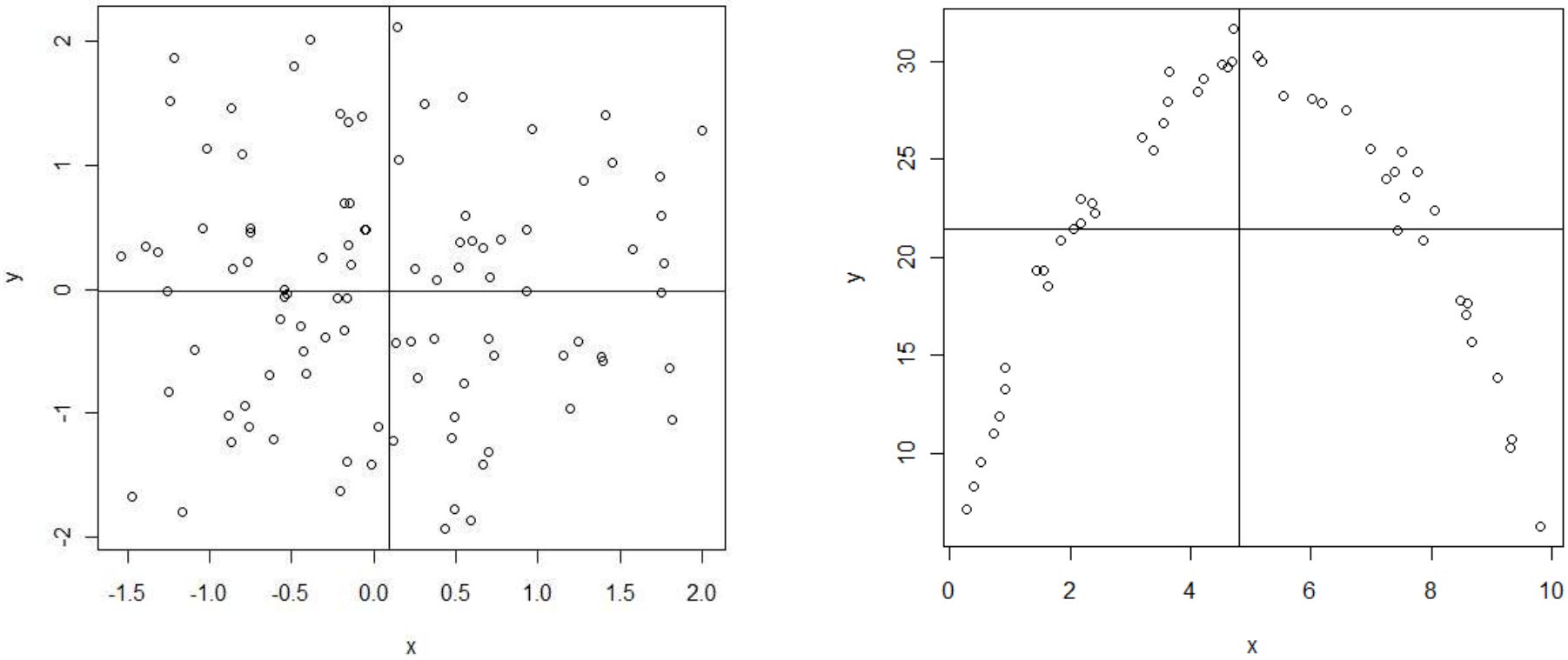
$$(x_i - \overline{x})(y_i - \overline{y})$$

⇨ 1과 3사분면은 양수, 2과 4분면의 값은 음수

○ 표본공분산(sample covariance)

$$c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

- ◦ 왼쪽 산점도와 같이 양의 기울기를 가지는 선분에 자료가 모여 있으며 c는 양의 값
- ◦ 오른쪽 산점도와 같이 음의 기울기를 가지는 선분에 모여 있으며 음의 값
- ◦ 표본분산은 $\frac{1}{n-1}\sum(x_i - \overline{x})^2 = \frac{1}{n-1}\sum(x_i - \overline{x})(x_i - \overline{x})$가 되는데 뒤에 있는 항에서 x_i를 y_i로 바꾸면 c



【그림 2.21】 직선관계가 없는 산점도의 예

○ 표본공분산의 간편식

$$
\begin{aligned}
c &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i\,y_i - n\overline{x}\,\overline{y}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i\,y_i - \frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right\}
\end{aligned}
$$

◉ 올림픽 개최 연도와 육상 100미터 우승기록

【표 2.16】 올림픽 100미터 자료 정리

	번호	x	y	x^2	y^2	xy
남자	1	1900	11	3610000	121	20900
	2	1904	11	3625216	121	20944
	⋮	⋮	⋮	⋮	⋮	⋮
	26	2012	9.63	4048144	92.7369	19375.56
	합	50924	266.95	99770832	2745.034	522514.2
	번호	x	z	x^2	z^2	xz
여자	1	1928	12.2	3717184	148.84	23521.6
	2	1932	11.9	3732624	141.61	22990.8
	⋮	⋮	⋮	⋮	⋮	⋮
	20	2012	10.75	4048144	115.5625	21629
	합	39456	223.67	77851232	2505.158	441064.2

○ 연도와 남자 우승기록의 표본공분산

$$c = \frac{1}{26-1}\left(522514.2 - \frac{1}{26}(50924)(266.95)\right)$$
$$= \frac{-338.177}{25} = -13.527$$

○ 연도와 여자 우승기록의 표본공분산

$$c = \frac{1}{20-1}\left(441064.2 - \frac{1}{20}(39456)(223.67)\right)$$
$$= \frac{-191.976}{19} = -10.104$$

○ 두 자료 모두 음의 기울기를 가지는 직선관계

○ 표본상관계수(coefficient of correlation)

◦ 공분산의 문제점은 측정 단위에 영향을 받기 때문에 그 값 자체로 선형관계의 정도를 알 수는 없음

- 예) 연도와 우승기록의 관계에서 우승기록을 초 단위가 아닌 분 단위로 표시하면 남자의 표본공분산은 –13.527/60=-0.225

◦ 피어슨의 표본상관계수: 표준화된 표본공분산

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

○ 표본상관계수 간편식

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x}\,)(y_i - \overline{y}\,) = \sum_{i=1}^{n} x_i\, y_i - n\,\overline{x}\,\overline{y}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x}\,)^2 = \sum_{i=1}^{n} x_i^2 - n\,\overline{x}^{2}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y}\,)^2 = \sum_{i=1}^{n} y_i^2 - n\,\overline{y}^{2}$$

$$\Rightarrow\ r = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}\ \sqrt{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\ \sqrt{S_{yy}}}$$

○ 표본상관계수의 성질

- $-1 \leq r \leq 1$
- 자료들이 어떤 기울기를 가지는 직선에 조밀하게 모일수록 $|r|$는 1에 근접
- 음의 기울기인 경우 r는 음수이며 음의 상관관계가 존재한다고 하고 양의 기울기인 경우 양수가 되며 양의 상관관계가 존재한다고 함
- 모든 관측값들이 직선 위에 위치하면 $|r| = 1$이 된다.
- $|r| \simeq 0$이면 상관관계가 없다고 함

◉ 올림픽 개최 연도와 우승기록

○ 남자의 상관계수

$$s_{xx} = 99770832 - \frac{50924^2}{26} = 30302.15$$

$$s_{yy} = 2475.034 - \frac{266.95^2}{26} = 4.177$$

$$r_{xy} = \frac{-338.177}{\sqrt{30302.15}\,\sqrt{4.177}} = -0.951$$

○ 여자의 상관계수

$$s_{xx} = 77851232 - \frac{39456^2}{20} = 12435.2$$

$$s_{zz} = 2505.158 - \frac{223.67^2}{20} = 3.745$$

$$r_{xz} = \frac{-191.976}{\sqrt{12435.2}\,\sqrt{3.745}} = -0.890$$

○ 두 표본상관계수 모두 –1에 가까운 값을 가지는 것으로 나타났으며 이는 연도와 우승기록 간에는 확실한 음의 상관관계가 있음

○ 표본상관계수는 두 변수 간에 직선관계가 있는지를 나타낼 뿐 인과관계를 나타내는 것은 아님

- 예) 휴대전화 보급률과 기대수명에 대한 상관계수를 구해보면 매우 높은 양의 상관관계를 가짐 ⇨ 기대수명을 늘리기 위해 휴대전화 보급을 늘려야 한다?

○ 잠복변수(lurking varaible): 두 변수에 영향을 주거나 관계가 있는 변수

○ 제3의 변수에 의해 나타나는 상관관계를 허위상관(spurious correlation) 또는 가짜상관 ⇨ 각각의 변수에서 잠복변수의 영향을 제거하고 표본상관계수를 계산하여 관련성을 파악