2. 자료의 정리

김 덕 기

toby123@cbnu.ac.kr

자료의 정리 및 시각화-범주형 자료

(1) 범주형 자료에 적합한 그래프

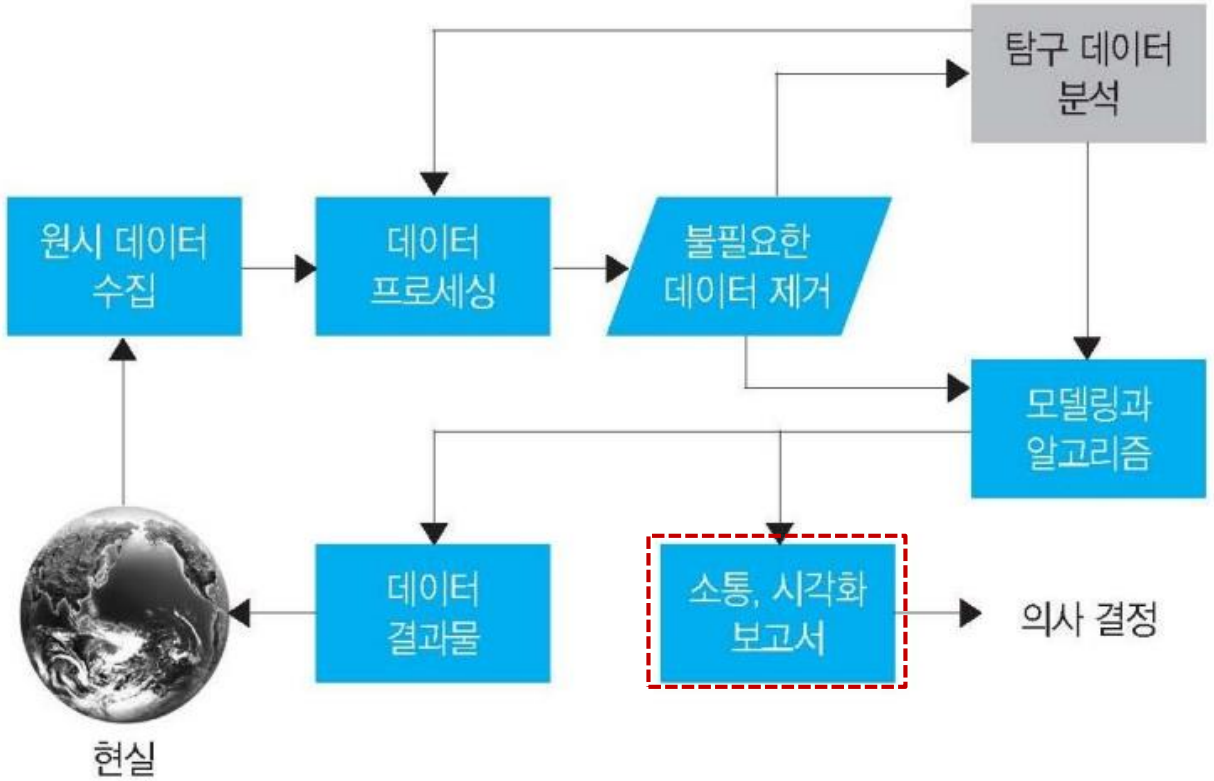
<table>
<tr><td rowspan="3">막대도표</td><td colspan="2">단순 막대도표 : 하나의 범주형 자료의 경우</td></tr>
<tr><td>수평누적 막대도표</td><td rowspan="2">둘 이상의 범주형 변수에 의해 2차원 이상의 분할표로 표현되는 자료의 경우</td></tr>
<tr><td>수직누적 막대도표</td></tr>
<tr><td>선도표</td><td colspan="2">단순 선도표 : 단순 막대도표를 선도표로 표현한 것
다중 선도표 : 누적 막대도표를 선도표로 표현한 그림
하락-선 선도표</td></tr>
<tr><td>면적도표</td><td colspan="2">단순 면적도표
수직누적 면적도표</td></tr>
<tr><td>원도표</td><td colspan="2"></td></tr>
<tr><td>히스토그램</td><td colspan="2"></td></tr>
</table>

자료의 정리 및 시각화-연속형 자료

(2) 연속형 자료에 적합한 그래프

히스토그램	범주형 자료의 형태로 변환 후 히스토그램을 그림
산점도	단순산점도 : 두 변수의 관계를 그림으로 표현 3-차원산점도 : 세 변수 사이의 관계를 나타내고자 할 때 산점도행렬 : 세 개 이상의 변수들 사이의 관계를 알아보고자 할 때 겹쳐그리기
P-P도표	귀무가설 하의 확률분포의 누적비율과 자료들의 누적비율의 산점도
Q-Q도표	귀무가설 하의 확률분포의 분위수와 자료들의 경험적분포의 분위수의 산점도
상자도표	한 범주형 변수들의 각 수준별로 상자그림을 작성
오차막대도표	한 범주형 변수들의 각 수준별로 신뢰구간, 표준편차, 표준오차 등이 수직선의 형태로 평균과 함께 출력
순차도표	시간을 X-축으로 하고 관측값을 Y-축으로 하는 산점도
시계열도표	시계열 자료의 패턴을 알아보기 위한 도표

데이터 시각화-Data Science Process



데이터 사이언스 프로세스 흐름도

데이터 시각화 – 중요성

- A Pictore is worth a thousans words.
  – Frederic R. Barnard (1927)
- Every picture tells a story.
  – Rod Stewart (1971)
- Graphs are essential to good statistical analysis.
  – F.J. Anscombe (1973)
- Visualization is critical to data analysis
  – W. Cleveland (1993)

자료의 도표화 –원자료(연속형)–도수분포표

표 2.1 분단위로 기록된 40개의 국제통화 시간

12	10	5	25	20	19	16	27	9	13
8	20	19	12	2	9	25	23	51	30
5	15	33	14	23	18	24	15	1	12
25	23	18	37	27	23	21	20	2	4

표 2.2 국제통화시간의 도수분포표

통화시간 (계급)		도수	백분율 (%)
	5분 미만	4	10.0
5분 이상	10분 미만	5	12.5
10분 이상	15분 미만	6	15.0
15분 이상	20분 미만	7	17.5
20분 이상	25분 미만	9	22.5
25분 이상	30분 미만	5	12.5
30분 이상	35분 미만	2	5.0
35분 이상	40분 미만	1	2.5
40분 이상		1	2.5
	합계	40	100.0

자료의 도표화–범주형–도수분포표+막대도표

진로 (범주)	학생 수 (도수)
공무원	12
기업체	25
대학원진학	5
해외연수	7
어학 및 자격증준비	1
합계	50

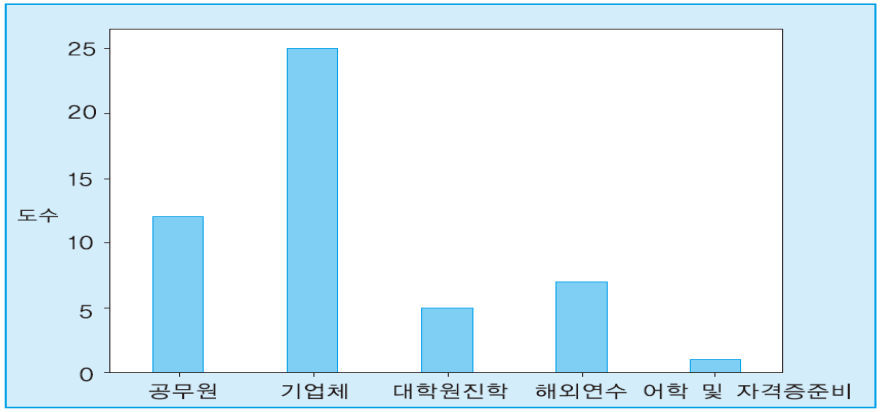


그림 2.2 졸업 후 진로에 대한 막대그래프

표 2.5 졸업 후 진로의 도수분포표

진로 (범주)	학생 수 (도수)	백분율 (%)
공무원	12	24.0
기업체	25	50.0
대학원진학	5	10.0
해외연수	7	14.0
어학 및 자격증준비	1	2.0
합 계	50	100.0

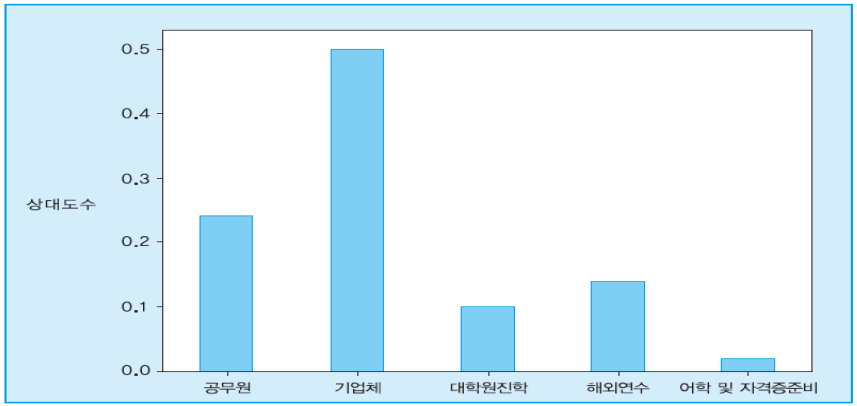


그림 2.3 졸업 후 진로에 대한 막대그래프

질적 자료의 정리 – 원그래프

표 2.5 졸업 후 진로의 도수분포표

진로 (범주)	학생 수 (도수)	백분율 (%)	원그래프 안의 범주들의 각도
공무원	12	24.0	86.4
기업체	25	50.0	180.0
대학원진학	5	10.0	36.0
해외연수	7	14.0	50.4
어학 및 자격증준비	1	2.0	7.2
합 계	50	100.0	360.0

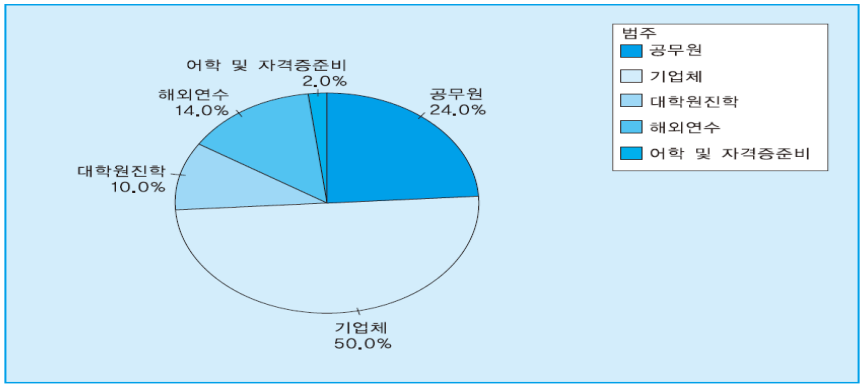


그림 2.4 졸업 후 진로에 대한 원그래프

360도 * 0.24=86.4도

꺾은선그래프-순차도표-시계열그래프

표 2.6 자동차 부품회사의 분기별 매출실적(단위: 억 원)

분기	매출
작년 1/4	12.0
2/4	10.5
3/4	16.0
4/4	13.5
금년 1/4	23.0
2/4	20.5
3/4	25.0
4/4	26.5
합 계	147.0

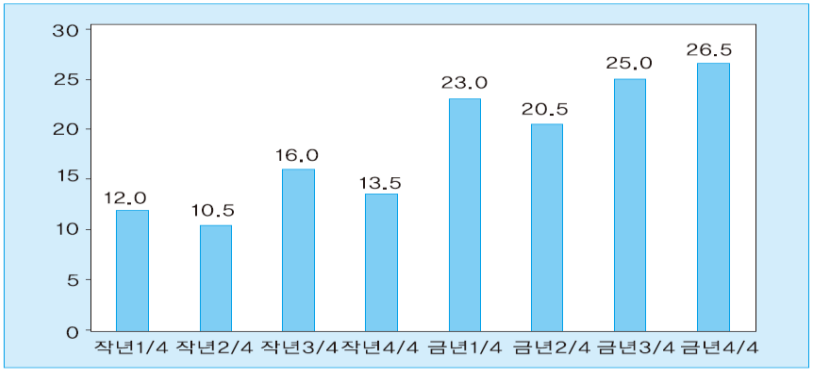


그림 2.5 자동차 부품회사의 분기별 매출실적의 막대그래프

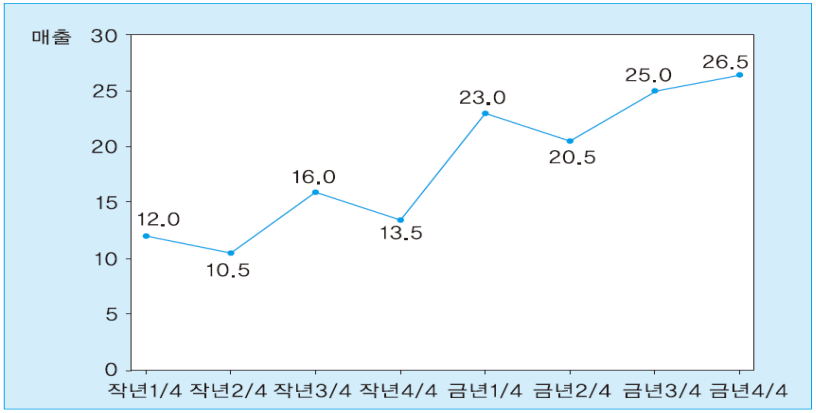


그림 2.6 자동차 부품회사의 분기별 매출실적의 꺾은선그래프

양적 자료의 정리-1

도수분포표 양 → 질

- 전체 자료를 동일한 간격을 가지는 서로 중복되지 않는 몇 개의 계급 구간으로 나누어 각 구간에 속하는 도수를 세어 나타낸 표

일반적인 작성절차

1. 자료 중 최대값과 최소값을 찾아 범위(=최대값-최소값)을 구함.
2. 자료의 크기에 따라 5~15정도의 계급의 수를 정하고, **계급의 폭 (=범위÷계급의 수)**을 정함.
3. 첫 계급구간의 **시작값(=최소값-자료값의 최소단위×½)**을 정하고, 2에서 구한 계급의 폭에 따라 나머지 계급을 설정.
4. 각 계급의 도수(상대도수, 누적상대도수 등)을 구함.

* 계급의 수: $K = 1 + \log_2 N$ [sturges formula : N(자료총수), K(계급수)]

양적 자료의 정리-2

표 2.7 남학생 50명의 체중

72	74	73	76	66	86	70	71	77	71
70	72	71	72	70	72	79	74	70	74
72	77	78	72	69	68	76	67	69	73
72	73	66	67	72	68	68	67	71	67
69	75	70	68	73	70	68	69	70	71

계급의 수가 5인 경우

계급	도수
65.5 ~ 70.5	22
70.5 ~ 75.5	21
75.5 ~ 80.5	6
80.5 ~ 85.5	0
85.5 ~ 90.5	1
합 계	50

계급의 수가 7인 경우

계급	도수
65.5 ~ 68.5	11
68.5 ~ 71.5	16
71.5 ~ 74.5	15
74.5 ~ 77.5	5
77.5 ~ 80.5	2
80.5 ~ 83.5	0
83.5 ~ 86.5	1
합 계	50

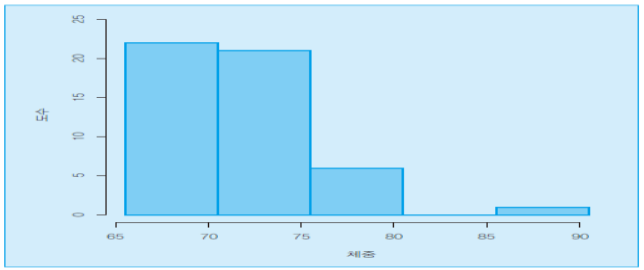
계급의 수가 9인 경우

계급	도수
63.75 ~ 66.25	2
66.25 ~ 68.75	9
68.75 ~ 71.25	16
71.25 ~ 73.75	12
73.75 ~ 76.25	6
76.25 ~ 78.75	3
78.75 ~ 81.25	1
81.25 ~ 83.75	0
83.75 ~ 86.25	1
합 계	50

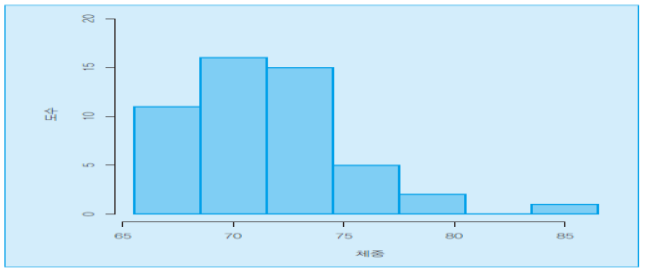
양적 자료의 정리-히스토그램

히스토그램 분포경향 이해

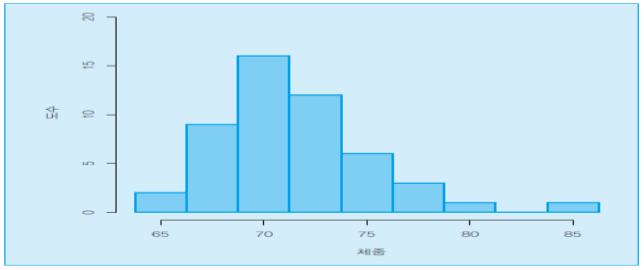
히스토그램이란 수평축에 계급구간을 표시하고, 수직축에는 각 계급구간에 해당되는 계급의 도수를 나타낸 그래프를 의미한다.



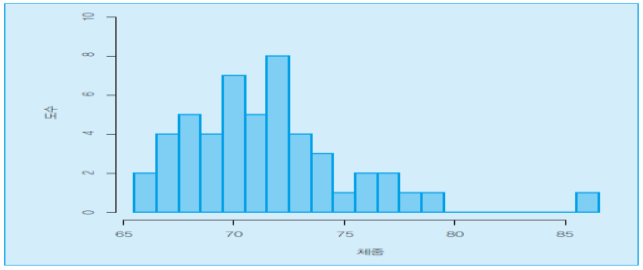
(a) 계급의 수가 5일 때



(b) 계급의 수가 7일 때



(c) 계급의 수가 9일 때



(d) 계급의 수가 21일 때

그림 2.7 각 계급의 수에 따른 히스토그램

양적 자료의 정리-각종 그래프

표 2.9 상대도수분포표

계 급	상대도수
63.75 ~ 66.25	2/50 = 0.04
66.25 ~ 68.75	9/50 = 0.18
68.75 ~ 71.25	16/50 = 0.32
71.25 ~ 73.75	12/50 = 0.24
73.75 ~ 76.25	6/50 = 0.12
76.25 ~ 78.75	3/50 = 0.06
78.75 ~ 81.25	1/50 = 0.02
81.25 ~ 83.75	0/50 = 0.00
83.75 ~ 86.25	1/50 = 0.02
합 계	1.00

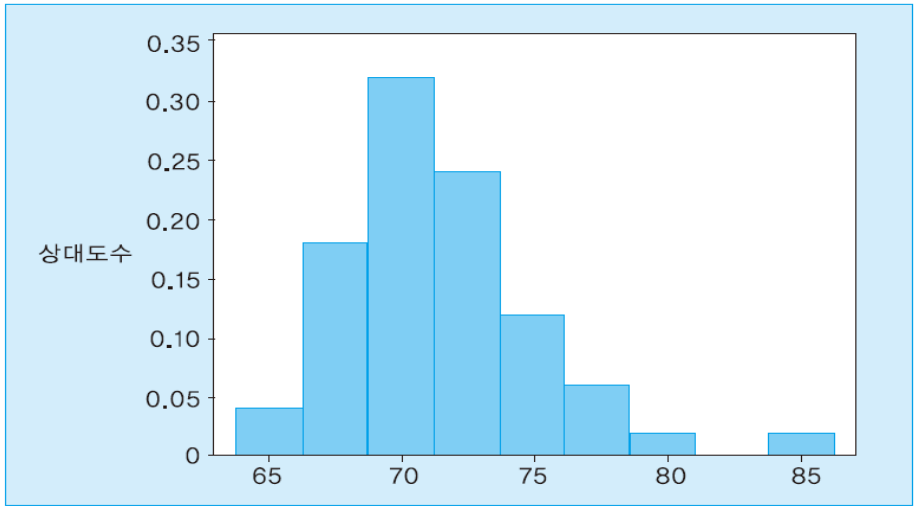


그림 2.8 상대도수분포도

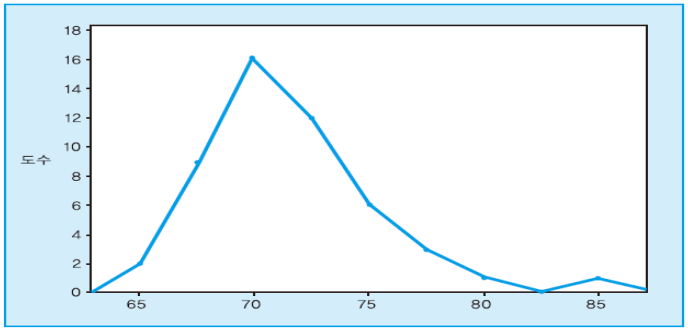


그림 2.9 도수다각형

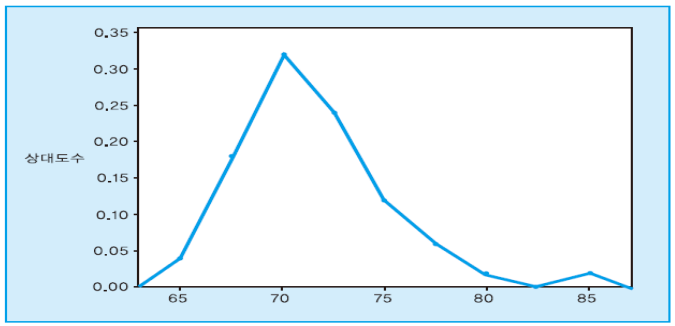


그림 2.10 상대도수다각형

누적상대도수분포표-누적상대도수분포도

표 2.10 누적상대도수분포표

계급	상대도수	누적상대도수
63.75 ~ 66.25	2/50 = 0.04	2/50 = 0.04
66.25 ~ 68.75	9/50 = 0.18	11/50 = 0.22
68.75 ~ 71.25	16/50 = 0.32	27/50 = 0.54
71.25 ~ 73.75	12/50 = 0.24	39/50 = 0.78
73.75 ~ 76.25	6/50 = 0.12	45/50 = 0.90
76.25 ~ 78.75	3/50 = 0.06	48/50 = 0.96
78.75 ~ 81.25	1/50 = 0.02	49/50 = 0.98
81.25 ~ 83.75	0/50 = 0.00	49/50 = 0.98
83.75 ~ 86.25	1/50 = 0.02	50/50 = 1.00
합 계	1.00	

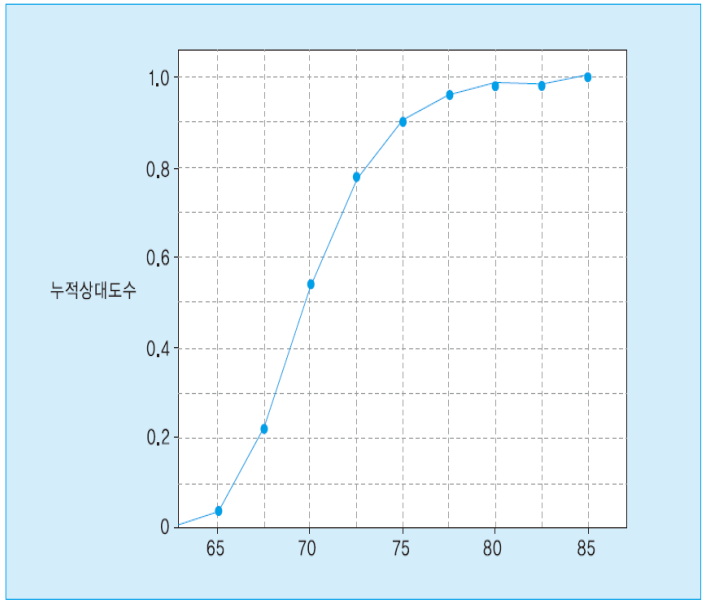
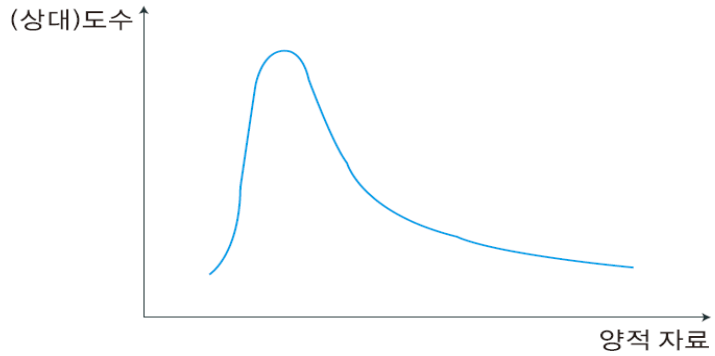
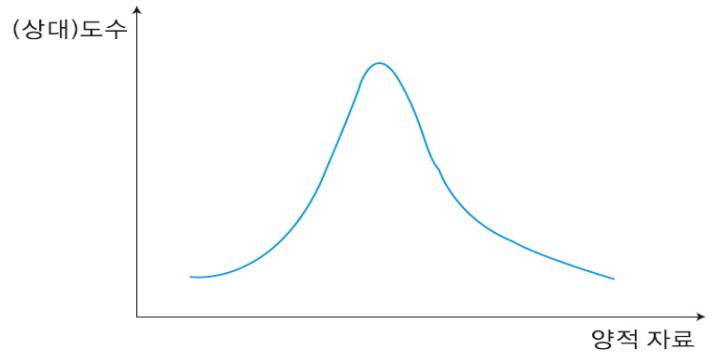


그림 2.11 누적상대도수분포도

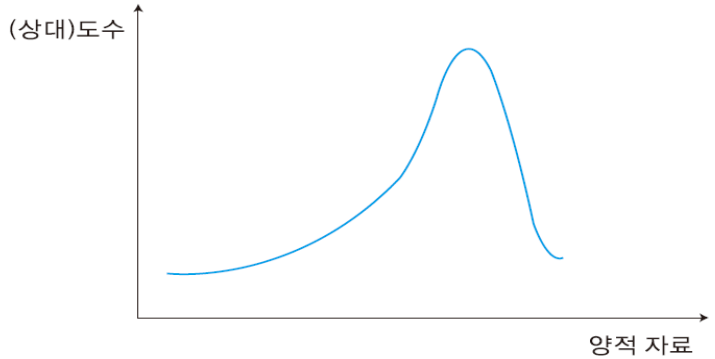
히스토그램의 형태–대칭, 비대칭, 쌍봉분포



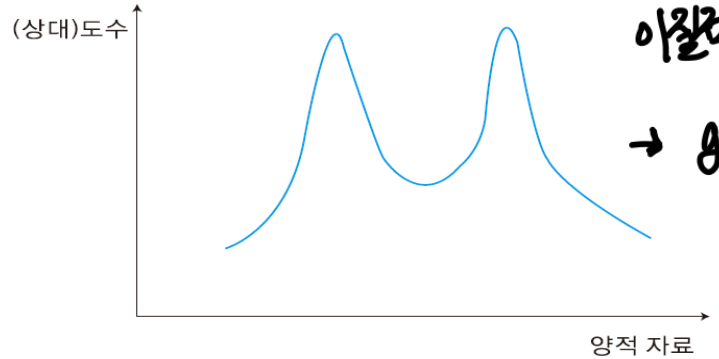
(a) 오른쪽으로 꼬리가 있는 형태



(b) 대칭인 형태



(c) 왼쪽으로 꼬리가 있는 형태



(d) 대칭이나 봉우리가 두 개인 형태

그림 2.12 히스토그램의 여러 가지 모양

원본 데이터 변형 X

줄기-잎 그림(stem and leaf plot)

줄기-잎 그림

줄기-잎 그림이란 히스토그램과 같은 공식적인 그래프로 표현하기 이전에 예비적 분석에 사용하는 그래프로서, 원자료로부터 공통적인 큰 값들을 줄기로 정하고, 각 관측값들의 구체적인 값들을 잎으로 정하여, 각 줄기에 잎을 붙인 모양을 갖춘 그래프를 의미한다.

줄기	잎
6	6 6 7 7 7 7 8 8 8 8 8 9 9 9 9
7	0 0 0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 4 4 4 5 6 6 7 7 8 9
8	6

그림 2.13 남학생 50명의 체중에 대한 줄기-잎 그림

줄기	잎
6	6 6 7 7 7 7 8 8 8 8 8 9 9 9 9
7	0 0 0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 4 4 4
7	5 6 6 7 7 8 9
8	
8	6

그림 2.14 각 줄기를 이등분하였을 때, 남학생 50명의 체중에 대한 줄기-잎 그림

줄기-잎-그림-2

줄기	잎
6	6 6 7 7 7 7
6	8 8 8 8 8 9 9 9 9
7	0 0 0 0 0 0 0 1 1 1 1 1
7	2 2 2 2 2 2 2 2 3 3 3 3
7	4 4 4 5
7	6 6 7 7
7	8 9
8	
8	
8	
8	6

그림 2.15 각 줄기를 오등분하였을 때, 남학생 50명의 체중에 대한 줄기-잎 그림

줄기-잎-그림-3

- 히스토그램과 같이 자료의 분포 모양을 파악할 수 있으며 관측값 개개의 정보도 얻을 수 있는 장점이 있다. 반면 자료의 수가 너무 많거나 흩어져 있는 경우에는 적절하지 않다.

15^-	2
15^+	88889
16^-	00000001223334
16^+	56777899
17^-	000011133344
17^+	6778889
18^-	0013

그림 13 | 통계학과 신입생의 키에 대한 두 줄기-잎 그림

줄기를 세분화한 경우

	15^-	2
	15^+	88889
33	16^-	000000012234
977	16^+	56789
44333111000	17^-	0
9888776	17^+	
3100	18^-	

그림 14 | 통계학과 남자 신입생과 여자 신입생의 키에 대한 두 줄기-잎 그림

남녀로 구분한 경우

2. 자료의 요약-수치요약

담당교수 : 김 덕 기

toby123@cbnu.ac.kr

수치적 요약-중심측도, 산포측도, 분포측도

- 수치로 자료의 특성을 나타내는 방법 :

➔ 중심위치의 측도, 산포(퍼진 정도)의 측도, 분포의 측도(왜도, 첨도) 등.

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

- 도표나 그림을 이용한 자료의 요약은 주관적이며 일관성이 부족
  - 히스토그램은 계급구간의 폭, 시작값에 따라 달라진다.

- 수치를 이용한 요약은 객관적
- 중심위치의 측도(measure of center)
  - 평균, 중앙값, 최빈값
- 퍼진 정도의 측도(measure of dispersion)
  - 자료가 중심위치로부터 얼마나 흩어져 있는가?
  - 분산, 표준편차, 범위, 사분위수범위, 변동계수

민감성

둔감성

중심 측도 : 평균(mean)

- 자료: $x_1, x_2, \ldots, x_n$

평균 = mean

- 표본평균(sample mean): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- 중심위치의 측도로 가장 많이 사용된다.
- 단점: 극단적으로 아주 크거나 작은 값의 영향을 많이 받을 수 있다. (not robust)

- 예) 관측값
  - 89, 74, 91, 88, 72, 84
  - $n = 6,\ \sum_{i=1}^{6} x_i = 498,\ \bar{x} = 83$

10% 절사평균 : $\bar{x}_t$

중심 측도 : 절사평균, 가중평균

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

이상치 다수 존재 시 영향 감소

- 특징 : 산술평균은 이상치(outlier)에 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?

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

평균

빈도 (가중치)

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

중심 측도 : 중앙값(median)

- 자료를 크기 순으로 배열할 때, 중앙에 위치하는 값
  - 자료의 수 (n)가 홀수이면 $\frac{n+1}{2}$ 번째 관측값
  - 자료의 수가 짝수이면 $\frac{n}{2}$ 번째 관측값과 $\frac{n}{2}+1$ 번째 관측값의 평균
- 관측값의 50% 이상이 중앙값 이상이고, 관측값의 50% 이상이 중앙값 이하이다.

- 예) 관측값
  - 89, 74, 91, 88, 72, 84
  - 크기 순: 72, 74, 84, 88, 89, 91
  - 중앙값 = (84+88)/2 = 86

- 평균과 달리 관측값의 변화에 민감하지 않고, 큰 관측값이나 작은 관측값에 영향을 받지 않는다. (robust)

중심 측도 : 최빈값(mode)

- 최빈수(mode)는 관측 값 중에서 빈도수가 가장 큰 값.
- 주로 이산형 자료나 범주형 자료에 대한 중심 위치로 사용.
- 연속형 자료에 대해서는 중심 위치의 측도로 적절하지 않다.
- 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 중심 값으로 사용할 것인가 ?

예제 : 어느 중소기업의 각 직위별 월급 현황. 회사의 전체 직원에 대한 월급의 대푯값(중심 측도)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	

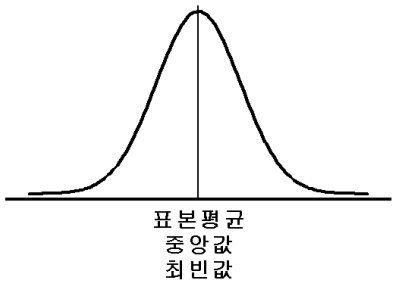
가중평균=570

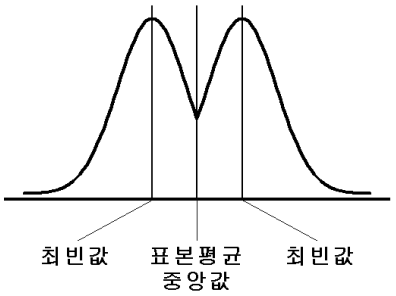
중앙값=300

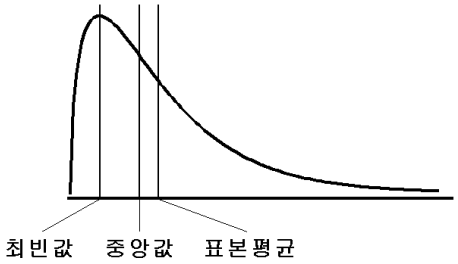
최빈값=200

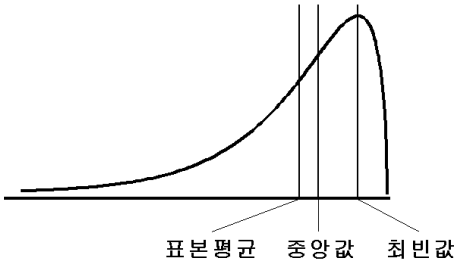
➔ 편중된 자료(skewed data)나, 이상치(outlier)가 있는 경우 중앙값이 표본평균보다 중심 위치의 측도로 적절할 수 있다

자료의 중심측도간의 관계









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

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

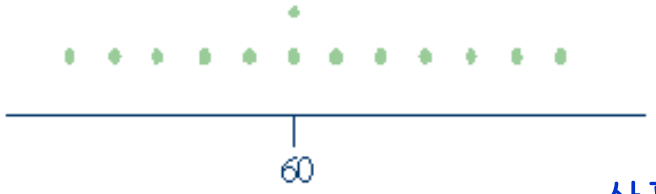
M_d = 중앙값

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

자료의 산포를 나타내는 측도

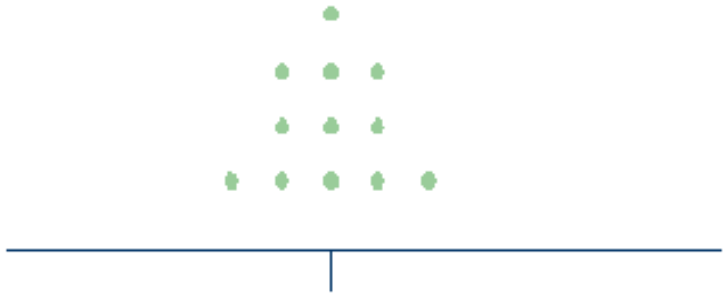
- 평균이 동일한 자료라 하더라도 평균을 중심으로 자료의 흩어진 정도가 다를 수 있기 때문에 대표값만으로 자료의 특성을 요약하는 것은 충분하지 않다. 따라서 대표값과 더불어 흩어진 정도를 나타내는 측도.

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



자료 1의 분포
평균=중위수=60

산포 : 범위(Range)



자료 2의 분포
평균=중위수=60

$\bar{x} = \frac{\sum x_i}{\text{자유도}} \rightarrow n$

자유도 : 통계량 계산할 때, 필요한 최소한의 데이터 수

산포 측도 : 분산, 표준편차, 평균편차, 범위

모분산: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

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

$\frac{\text{변동}}{\text{자유도}}$

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 - N\mu^2}{N} = \frac{\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 / N}{N}$$

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

모표준편차: $\sigma = \sqrt{\sigma^2}$

표본표준편차: $s = \sqrt{s^2}$

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

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

편차 $= \sum_{i=1}^{n}(x_i - \bar{x})$

변동 $=$ 편차^2

분산 $= \frac{\text{변동}}{\text{자유도}}$

➔ 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.

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

이상치 판별 기준 : $Q_1 - 1.5 * IQR = f_L$
$Q_3 + 1.5 * IQR = f_U$

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

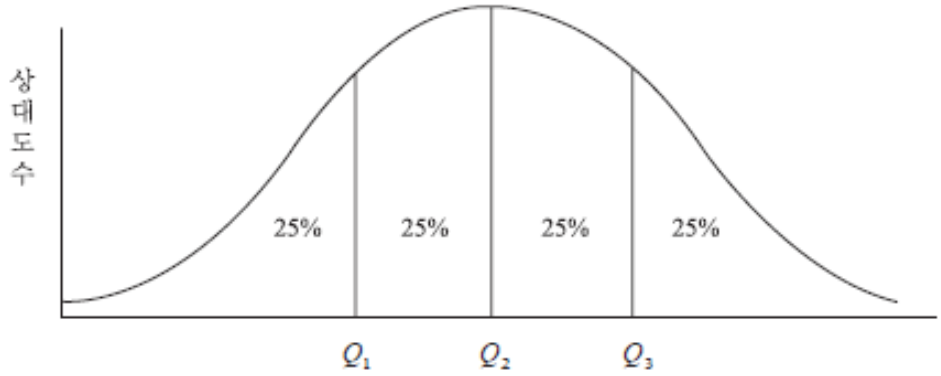


그림 1.2 사분위수와 자료구조 (무한모집단)

안정적인 범위 (25~75%)

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 $= \frac{(83+85)}{2} = 84$

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 : $861/77471 = 0.01112$

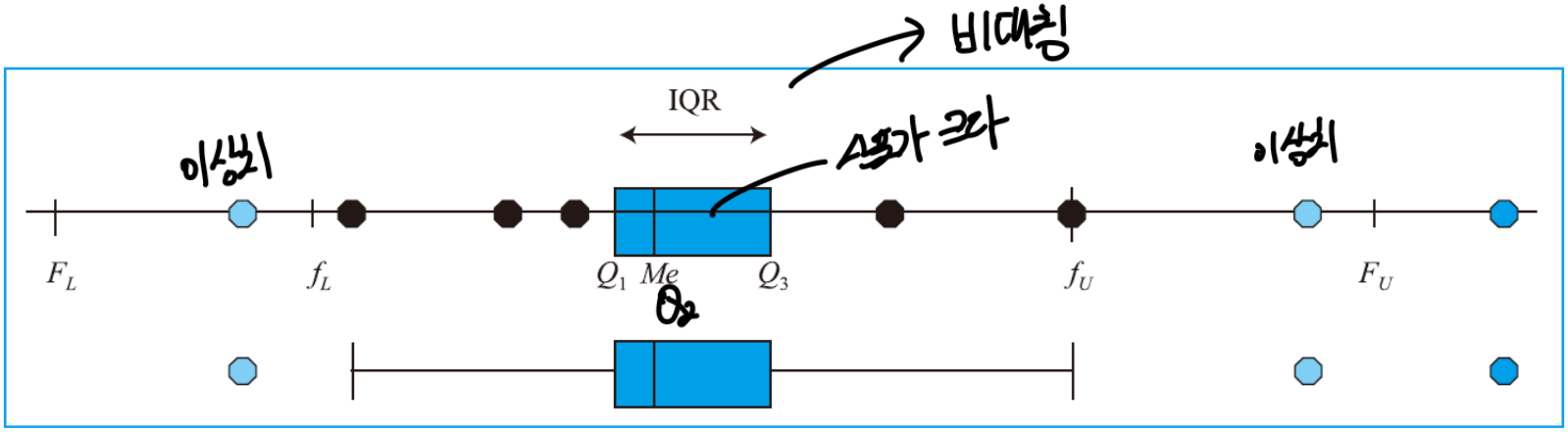
B회사 – 평균 : 7100, 표준편차 : 429, 변동계수 : $429/7100 = 0.06042$

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.



일반적인 상자그림의 형태

$F_L = Q_1 - 3IQR$

$F_U = Q_3 + 3IQR$

상자그림(box-plot) – 예제 20

- 예) 교차로의 소음 측정 자료 $(n = 50)$
  - Q_1: $np = 50 \times 0.25 = 12.5$ → 13번째 값 = 57.6
  - Q_2: $np = 50 \times 0.5 = 25$ → 25번째와 26번째의 평균 = 60.9
  - Q_3: $np = 50 \times 0.75 = 37.5$ → 38번째 값 = 63.8
  - IQR = 63.8 − 57.6 = 6.2
  - 경계값: $Q_1 - 1.5 \times IQR = 57.6 - 1.5 \times 6.2 = 48.3$
    $Q_3 + 1.5 \times IQR = 63.8 + 1.5 \times 6.2 = 73.1$
  - $Q_1 - 1.5 \times IQR$ 보다 큰 최솟값 = 51.3
  - $Q_3 + 1.5 \times IQR$ 보다 작은 최댓값 = 69.4

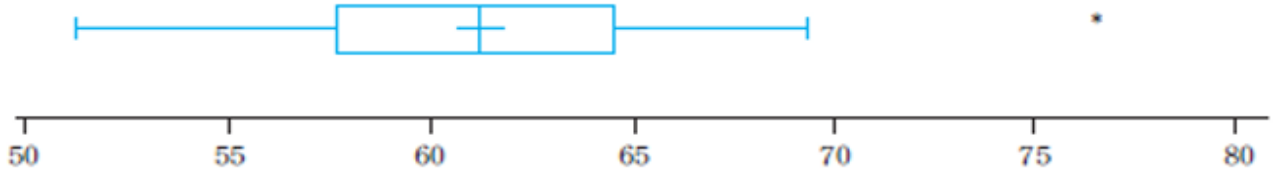


그림 6 교통소음에 대한 상자그림

편차이 ∝ 비대칭성
(Q1~Q2, Q2~Q3)

상자그림(box-plot) – 예제 21

- 상자그림은 자료 분포의 다양한 특성을 하나의 그림으로 나타낸 것 (중심위치, 퍼진 정도, 대칭성, 분포의 집중 정도, 이상점 등을 파악할 수 있다.)
  - 대칭적인 분포를 가진 자료의 상자그림은 어떤 모양일까?
  - 왼쪽으로 편중된 자료의 상자그림은 어떤 모양일까?
  - 오른쪽으로 편중된 자료의 상자그림은 어떤 모양일까?

- 예) 남학생과 여학생 키에 대한 상자그림

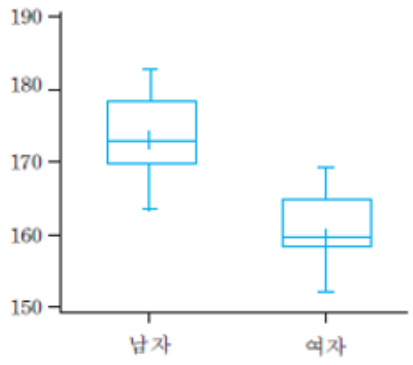


그림 7 상자그림을 통한 남자 신입생의 키와 여자 신입생의 키 비교

1. 남학생의 키는 여학생의 키보다 약 10cm 정도 크다.
2. 퍼진 정도는 남학생이 약간 더 크다.
3. 남학생의 키는 비대칭이 약하지만, 여학생의 키는 제1 사분위수와 중앙값 사이에 집중되어 비대칭이 심하다.