제2장 자료의 구조

2.1 서론

2.2 자료의 종류

(1) 질적 자료 (qualitative data) [범주형 자료 (categorical data)]

① 명목척도자료 (nominal scaling data)
: 단지 구분하기 위한 부호로 표시된 자료
[예] 김씨: 1, 이씨: 2, 박씨: 3

② 서수척도자료 (ordinal scaling data)
: 자료들 사이의 크기를 비교하여 내림차순 또는 오름차순으로 숫자(순위)를 부여한 것으로 실제값보다는 순서를 나타낸 자료
[예 A^+: 4.5, A: 4.0, B^+: 3.5, B: 3.0, C^+: 2.5, C: 2.0, D^+: 1.5, D: 1.0, F: 0.0

(2) 양적 자료 (quantitative[measurement] data)

① 구간척도자료 (Interval scaling data)
: 자료들 사이의 크기가 의미를 가지는 자료.
자료가 나타내는 숫자 자체만 보지 말고 숫자가 나타내는 의미를 보아야 한다.
[예] 성적, 온도

② 비율척도자료 (Ratio scaling data)
: 절대 0점이 있어서 비율로 이야기할 수 있는 자료
[예] 몸무게 - A는 B보다 몸무게가 두 배가 된다.

❑ 자료집합의 구성

① 변수 (variable): 이산형 (discrete) vs. 연속형 (continuous)

② 관측치 (observation)

2.3 자료구조의 표현과 요약

2.4 표와 그림을 이용한 자료의 표현

2.4.1 질적 자료 (범주형 자료)의 표현

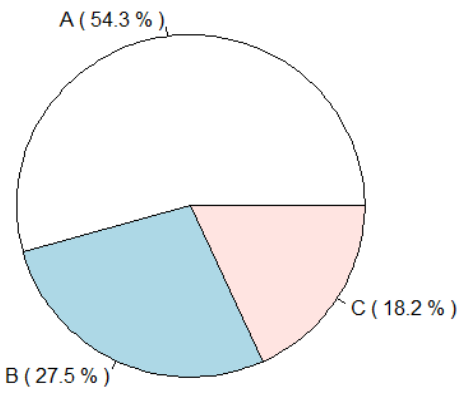
(1) 도수분포표 (frequency table): $\text{각 범주의 상대도수} = \frac{\text{해당 범주의 도수}}{\text{자료 전체의 개수}}$

[예제 2.1] 어느 대학교의 학생회장 선거에서 세 명의 학생(A, B, C)이 입후보하여 투표를 실시한 결과 후보 A는 1,520표를, 후보 B는 770표를, 후보 C는 510표를 얻었다. 결과를 도수분포표로 나타내어라.

후보자	도수	상대도수
A	1520	1520 / 2800 = 0.543
B	770	770 / 2800 = 0.275
C	510	510 / 2800 = 0.182
계	2800	1

(2) 파이 차트 (pie chart)

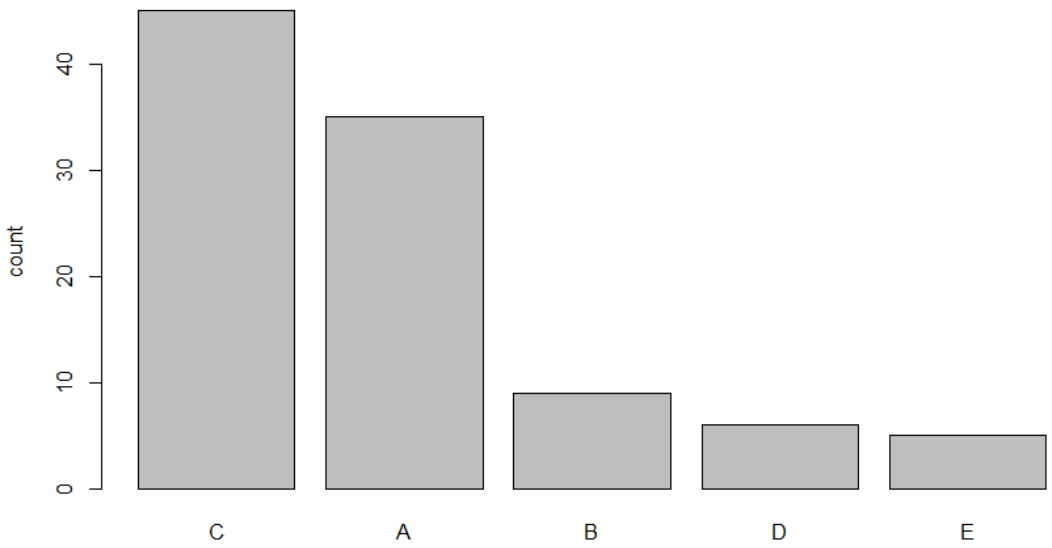
[예제 2.2] [예제 2.1]의 자료에 대한 파이 차트를 작성하라.



(3) 파레토 그림 (Pareto diagram)

[예제 2.3] 어느 생산 공장에서 생산되는 제품의 불량률을 낮추기 위하여 불량의 원인이 되는 요소를 제거하기로 하였다. 불량의 원인을 조사한 결과 다섯 가지의 요소(A, B, C, D, E)로 나타났다. 품질개선비용을 고려하여 이 중에서 가장 영향이 큰 요소 두 가지를 선택하여 중점적으로 개선하고자 한다. 전문가를 포함한 생산 관련자 100명에게 다섯 요소 중 가장 문제가 되는 요소를 적어내게 하여 정리하고 다음의 도수분포표를 작성하였다. 파레토 그림을 그려 개선 대상 요소를 선택하라.

요소	도수
A	35
B	9
C	45
D	6
E	5
계	100



2.4.2 이산형 자료의 표현

(1) 도수분포표

: 누적도수 = 해당 값까지의 도수의 값

$$\text{누적상대도수} = \frac{\text{누적도수}}{\text{전체 자료의 총 개수}}$$

[예제 2.4] 30쪽 분량으로 이루어진 어느 책에서 한 쪽당 발생하는 오자 개수의 분포에 대하여 조사를 하였다. 각 쪽에서 발견된 오자의 수는 다음과 같다. 오자의 수에 대한 (누적도수와 누적상대도수를 포함한) 도수분포표를 작성하고 임의의 한 쪽을 선택하였을 때 오자가 2개 이하일 확률을 구하라.

1 1 1 3 0 0 1 1 1 0 2 2 0 0 0
1 2 1 2 0 0 1 6 4 3 3 1 2 4 0

【풀이】

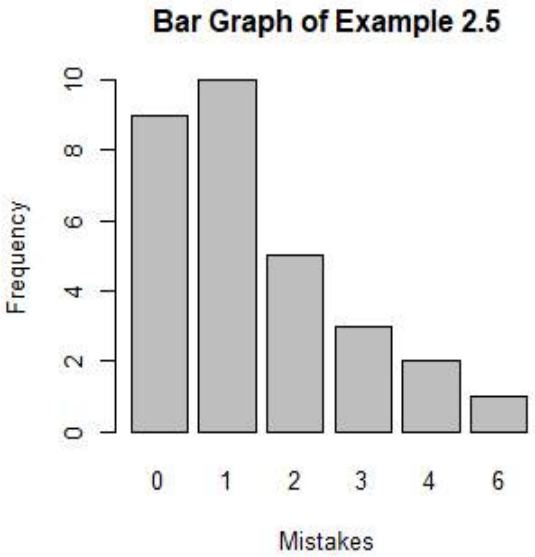
| 표 2-2 | 오자의 개수에 대한 도수분포표

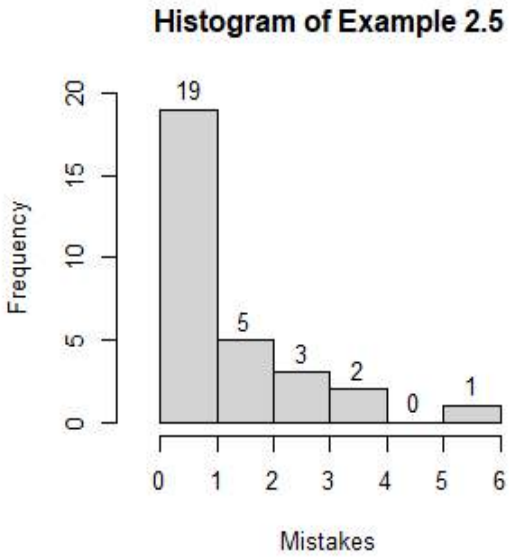
오자의 개수	도 수	상대도수	누적도수	누적상대도수
0	9	0.300	9	0.300
1	10	0.333	19	0.633
2	5	0.167	24	0.800
3	3	0.100	27	0.900
4	2	0.067	29	0.967
5	0	0.000	29	0.967
6	1	0.033	30	1.000
계	30	1.000		

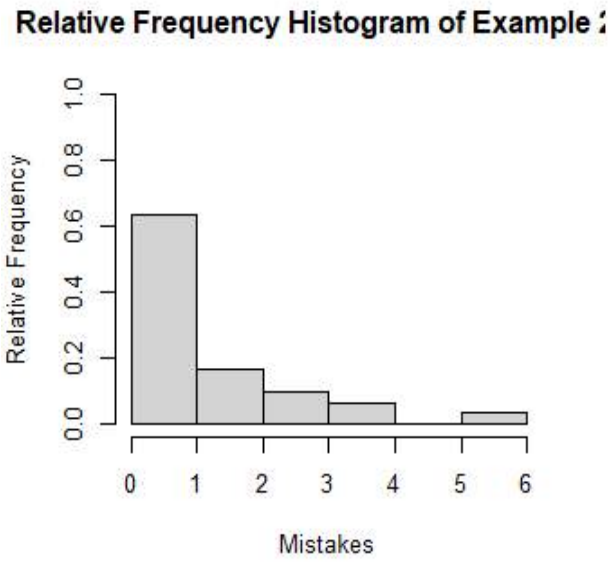
(2) 막대그림 (bar graph)과 히스토그램 (histogram)

[예제 2.5] [예제 2.4]의 오자 자료에 대한 막대그림과 히스토그램을 그려라.

【풀이】







2.4.3 연속형 자료의 표현

(1) 구간에 의한 도수분포표 (frequency table on interval)

[예제 2.6] 대학생들의 평소 운동시간을 조사하기로 하였다. 40명의 학생을 임의 선택하여 각 학생에 대하여 100일간 조사하였다. 각 학생의 일일 평균 운동시간이 다음과 같이 계산되었다(단위: 분). 구간에 의한 도수분포표를 작성하라.

90.12	97.29	100.89	14.39	26.94	27.57	35.84	69.48
73.02	89.86	102.71	122.77	75.18	49.70	49.93	50.97
76.34	40.31	1.70	10.20	12.14	14.10	41.85	42.44
48.01	48.32	52.22	52.42	52.53	55.39	62.30	64.90
67.14	68.65	79.41	82.54	84.20	84.98	87.42	87.78

【풀이】

표 2-3 운동시간에 대한 도수분포표

일일 평균운동시간(분)	도 수	상대도수
0 ~ 10	1	0.025
10 ~ 30	6	0.150
30 ~ 50	8	0.200
50 ~ 70	10	0.250
70 ~ 90	10a	0.250
90 ~ 110	4	0.100
110 ~ 130	1	0.025
계	40	1.00

(2) 점도표 (dot diagram)와 히스토그램 (histogram)

❑ (상대도수)히스토그램의 직사각형의 높이 $= \dfrac{\text{상대도수}}{\text{구간의 폭}}$

❑ 히스토그램에서 직사각형의 면적의 합은 1이다.

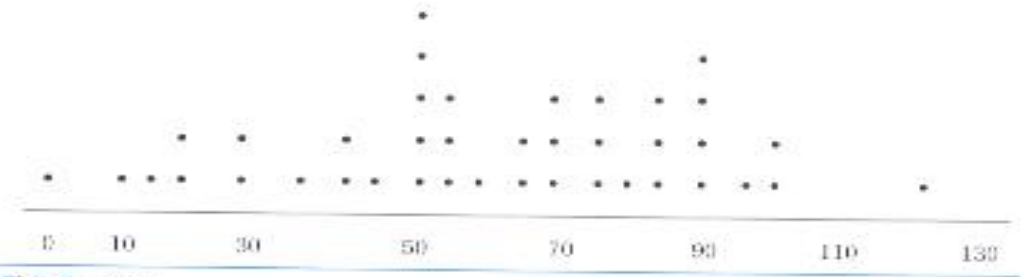
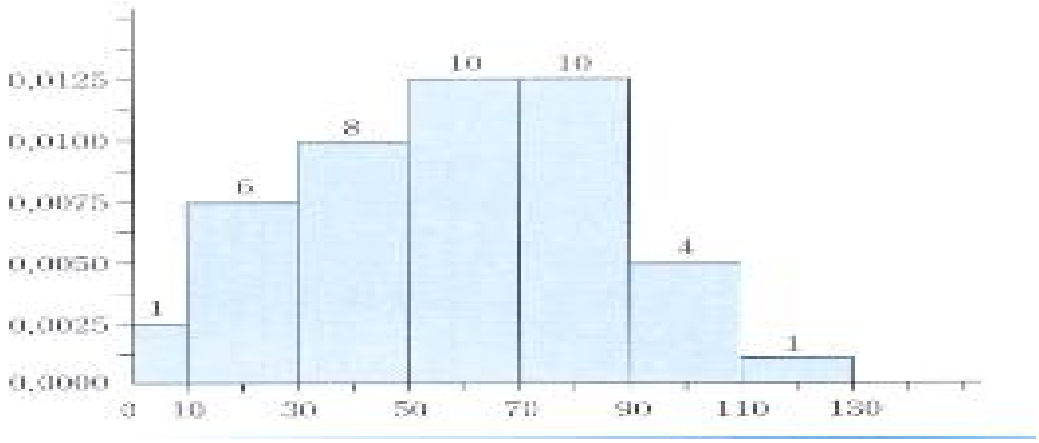

0
10
30
50
70
90
110
130

그림 2-5 | 점도표


0,0125
0,0100
0,0075
0,0050
0,0025
0,0000
1
6
8
10
10
4
1
0
10
30
50
70
90
110
130

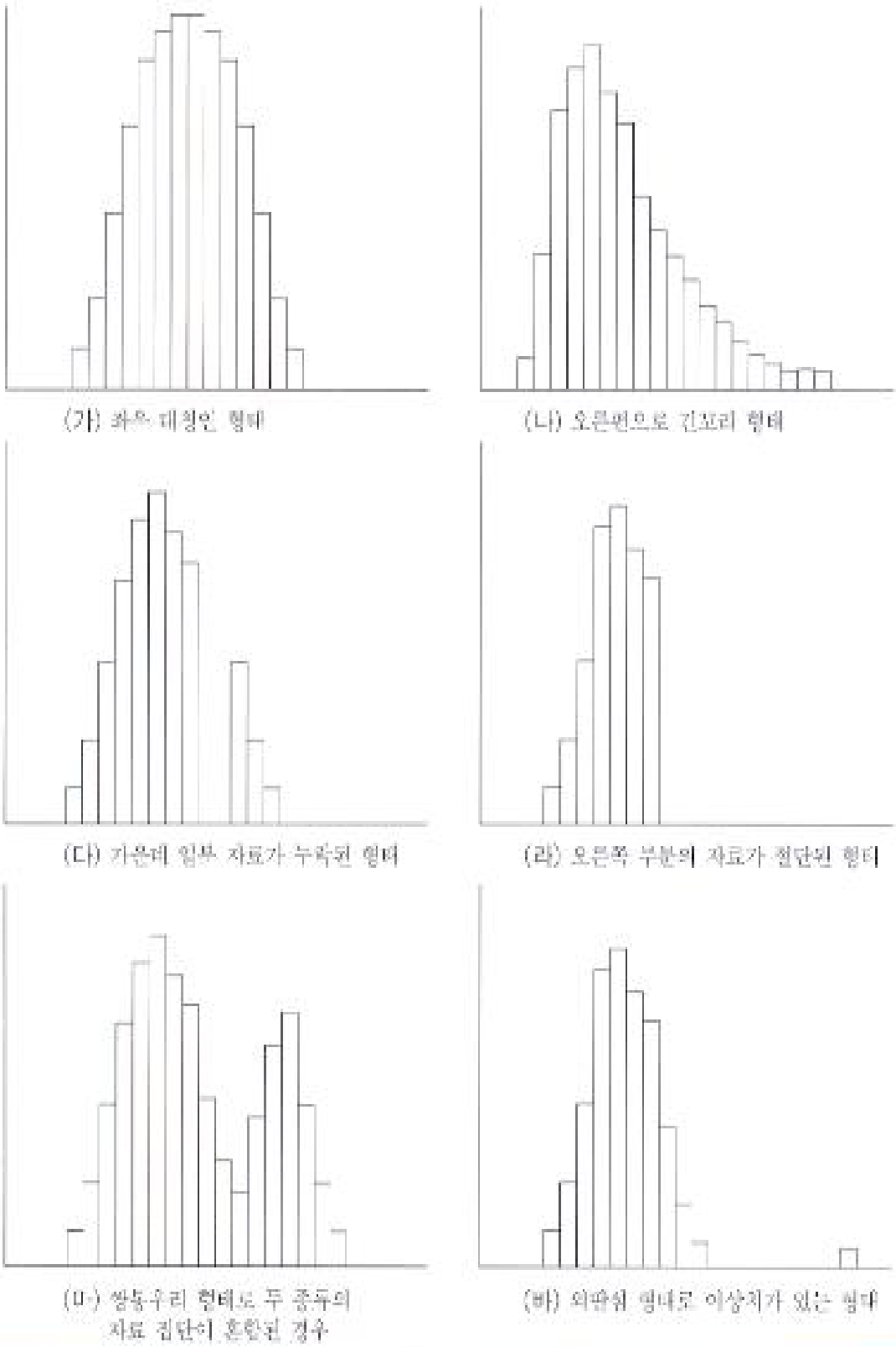


그림 2-7 | 여러 형태의 히스토그램

(3) 줄기와 잎 그림 (stem-and-leaf plot)

[작성법]

① 관측값을 보고 앞 단위와 뒷 단위를 정한다.

② 앞 단위를 줄기로 하여 순서대로 세로로 배열하고 그 옆에 수직선을 그린다.

③ 뒷 단위를 잎으로 하여 해당하는 관측값을 앞 단위 오른쪽에 가로로 기입한다.

④ 각 줄기에서 앞 부분의 값을 작은 숫자가 왼쪽에 오도록 크기 순서로 재배열한다.

[예제] 다음의 자료는 어느 대학 통계학과 신입생 51명의 키를 센티미터 단위로 기록한 것이다. 이 자료에 대한 줄기와 잎 그림을 그려라.

181	161	170	160	158	169	162	179	183	178	171	177	163
158	160	160	158	174	160	163	167	165	163	173	178	170
167	177	176	170	152	158	160	160	159	180	169	162	178
173	173	171	171	170	160	167	168	166	164	174	180	

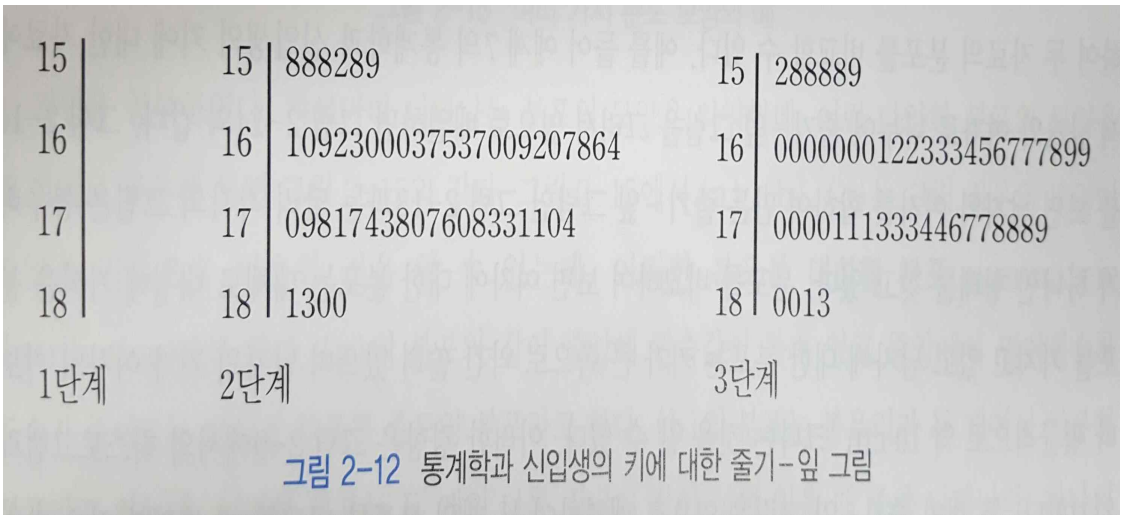


그림 2-12 통계학과 신입생의 키에 대한 줄기-잎 그림

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

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

남자	181	170	179	183	178	171	177	174	163	167	163	173	178	170
	167	177	176	180	169	178	173	173	171	171	170	174	180	
여자	161	160	158	169	162	163	158	160	160	158	160	165	170	152
	158	160	160	159	162	160	167	168	166	164				

2.5 자료구조의 중심을 나타내는 척도

(1) 평균 (mean[average])

① 정의: $x_1, x_2, \cdots, x_n$의 평균. $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

② 단점: 모든 관측값이 반영되므로 표본평균은 극단적으로 아주 큰 값이나 아주 작은 값에 영향을 많이 받는다.

(2) 중앙값 (median)

① 정의: 자료들을 크기 순으로 정렬하였을 때 순서에 따라 가장 가운데 있는 값

② 중앙값 구하는 방법

□ 자료의 개수(n)가 홀수: $\frac{n+1}{2}$번째 관측값

□ 자료의 개수(n)가 짝수: $\frac{n}{2}$번째 관측값과 $\frac{n}{2}+1$번째 관측값 사이의 중간값 또는 평균

③ 장점: 관측값들의 변화에 민감하지 않고 특히 아주 큰 관측값이나 아주 작은 관측값에 영향을 받지 않는다.

[예제 2.7] 크기 순으로 정의된 다음의 표본에서 표본평균과 표본중앙값을 구하라. 그리고 끝값인 15가 조사단위를 잘못하여 150으로 바뀌면 어떻게 되겠는가?

1, 3, 4, 6, 6, 7, 8, 8, 9, 10, 15

【풀이】

□ 평균: $\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{77}{11}=7$

□ 중앙값: $n=11$(홀수), 중앙값은 $\frac{n+1}{2}=\frac{11+1}{2}=6$ 번째 관측값$=7$

→ 이 경우 평균과 중앙값은 같은 값인 7을 갖는다.

만약 끝값이 15에서 150으로 변경되면

□ 평균: $\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{212}{11}=19.2727$

□ 중앙값: $n=11$(홀수), 중앙값은 $\frac{n+1}{2}=\frac{11+1}{2}=6$ 번째 관측값$=7$

→ 이 경우 평균은 이상치(150)의 영향을 많아 7에서 19.2727로 변경되지만, 중앙값은 상기 경우와 같은 7을 갖는다.

(3) 절사평균 (trimmed mean)

① $(100\times\alpha)\%$ 절사평균의 정의: 자료를 크기 순으로 나열하고 자료의 아래쪽 $(100\times\alpha)\%$와 위쪽 $(100\times\alpha)\%$를 버린 나머지 자료들의 평균

② $(100\times\alpha)\%$ 절사평균 구하는 방법: $\bar{x}_{\alpha}=\dfrac{x_{([na]+1)}+\cdots+x_{(n-[na])}}{n-2[na]}$

③ 장점

▫ 아주 큰 관측값이나 아주 작은 관측값에 영향을 받지 않는다.

▫ 평균과 중앙값의 성질을 모두 갖고 있다.

[예제] 10만 가구가 살고 있는 어떤 마을의 1년 소득을 조사한 자료(단위: 만원)가 다음과 같을 때, 15% 절사평균을 구하라.

950 1,050 10,310 760 1,470 1,530 1,170 1,240 1,090 1,020

【풀이】

재정렬: 760 950 1,020 1,050 1,090 1,170 1,240 1,470 1,530 10,310

$15\%\rightarrow\alpha=0.15,\ n=10$

$$\bar{x}_{0.15}=\frac{x_{([na]+1)}+\cdots+x_{(n-[na])}}{n-2[na]}$$

$$=\frac{x_{([10\times0.15]+1)}+\cdots+x_{(10-[10\times0.15])}}{10-2[10\times0.15]}=\frac{x_{(1+1)}+\cdots+x_{(10-1)}}{10-(2\times1)}=\frac{x_{(2)}+\cdots+x_{(9)}}{8}$$

$$=\frac{9520}{8}=1190$$

2.6 자료구조의 퍼짐을 나타내는 척도

(1) 편차 (deviation)

① 정의: 표본평균을 중심위치 측도로 사용할 때 각 관측값과 평균의 차이

② 공식: $x_i-\bar{x}\qquad(i=1,\cdots,n)$

③ 성질: $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=0$ → '퍼진 정도의 측도'로 부적합

(2) 분산 (variance)

① 공식: $s^2=\dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n-1}$ 여기서 $n-1$은 자유도(degree of freedom)

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

여기서

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

(3) 표준편차 (standard deviation): $s = \sqrt{s^2}$

[예제 2.9] [예제 2.7]의 자료에 대한 표본분산과 표본표준편차를 구하라.

【풀이】

	자료											합계
X	1	3	4	6	6	7	8	8	9	10	15	77
X^2	1	9	16	36	36	49	64	64	81	100	225	681

$$\text{표본분산} = s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right]$$

$$= \frac{1}{10}\left[681 - \frac{77^2}{11}\right] = 14.2$$

$$\text{표본표준편차} = s = \sqrt{s^2} = \sqrt{14.2} = 3.7683$$

(3) 표본범위 (sample range)

① 정의: 관측값에서 가장 큰 값과 가장 작은 값의 차이

② 공식: 범위 = max − min

③ 장점: 간편하게 구할 수 있고 해석이 용이하다.

④ 단점: 양 끝점에 의해서만 결정되기 때문에 중간에 위치한 관측값들이 어떻게 퍼져 있는가 하는 것은 전혀 고려되지 않는다. 특히 극단적으로 큰 값이나 작은 값이 있는 경우 그 관측값이 미치는 영향이 매우 클 수 있다.

(4) 백분위수 (percentile)

① 정의: 전체를 백 부분으로 나누어 각 경계선에 해당하는 값

② 예: 제25백분위수 = Q_1, 제50백분위수 = Q_2(중앙값), 제75백분위수 = Q_3

[예제] 서울의 한 전철역에서 인천의 한 전철역까지 소요되는 시간을 기록한 자료가 다음과 같다. (단위: 분) 제 50 백분위수인 중앙값과 제 20 백분위수를 구하라.

42 40 38 37 43 39 78 38 45 44 40 38 41 35 31 44

【풀이】
상기 관측값을 순서대로 재배열하면

31 35 37 38 38 38 39 40 40 41 42 43 44 44 45 78

$n = 16$ (짝수)

① $p = 0.5 \rightarrow np = 16 \times 0.5 = 8 \rightarrow$ 제50백분위수$= \dfrac{x_{(8)} + x_{(9)}}{2} = \dfrac{40 + 40}{2} = 40$

② $p = 0.2 \rightarrow np = 16 \times 0.2 = 3.2$ (정수 아님) $\rightarrow m = 3 + 1 = 4$
$\rightarrow$ 제20백분위수$= x_{(m)} = x_{(4)} = 38$

(5) 사분위수 (quartile)
① 정의: 전체를 네 부분으로 나누어 각 경계선에 해당하는 값
② 구성: 제25백분위수$= Q_1$, 제50백분위수$= Q_2$(중앙값), 제75백분위수$= Q_3$

(6) 표본사분위수범위 (sample interquartile range; IQR) $= Q_3 - Q_1$
① 정의: 제3사분위수와 제1사분위수 사이의 거리
② 공식: $Q_3 - Q_1$
③ 장점: 극단값에 영향을 받지 않고, 한쪽으로 치우친 분포에서 극단값을 제외한 퍼진 정도를 알려고 할 사용된다.
④ 단점: 사분위수범위에 대한 이론적 추론이 어렵기 때문에 분산이나 표준편차만큼 퍼진 정도의 측도로 많이 쓰이지는 않는다.

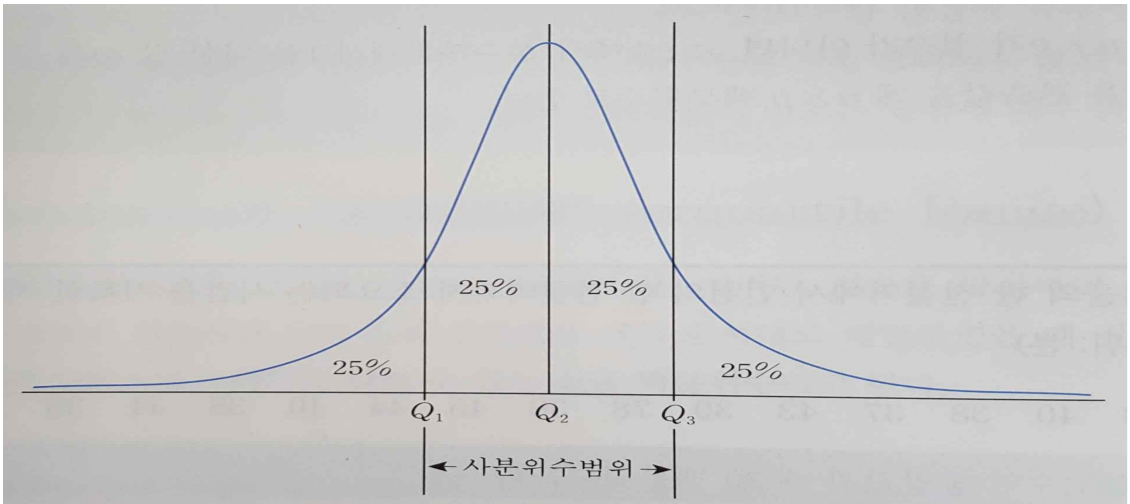


그림 3-4 사분위수의 위치와 사분위수범위

[예제 2.10] [예제 2.7] 자료에 대한 표본사분위수범위를 구하라.

【풀이】

$$1,\ 3,\ 4,\ 6,\ 6,\ 7,\ 8,\ 8,\ 9,\ 10,\ 15$$

$$Q_1 = 4,\ Q_3 = 9$$

$$\text{표본사분위수범위} = Q_3 - Q_1 = 9 - 4 = 5$$

(7) 상자(수염)그림 (box plot, box-and-whisker plot)

① 정의: 자료로부터 얻은 다섯 가지 요약수치인 최솟값, Q_1, Q_2, Q_3, 최댓값을 가지고 그림을 그린 것

② 상자(수염)그림 작성과정

□ 사분위수(Q_1, Q_2, Q_3)을 결정한다.

□ Q_1과 Q_3를 네모난 상자로 연결하고, 중앙값(Q_2)의 위치에 수직선을 긋는다.

□ $IQR = Q_3 - Q_1$을 계산한다.

□ 상자 양 끝에서 $1.5 \times IQR$ 크기의 범위를 경계로 하여, 이 범위에 포함되는 최솟값과 최댓값을 Q_1과 Q_3로부터 각각 선으로 연결한다.

□ 양 경계를 벗어나는 자료값들을 *로 표시하고, 이 점들을 이상점이라고 한다.

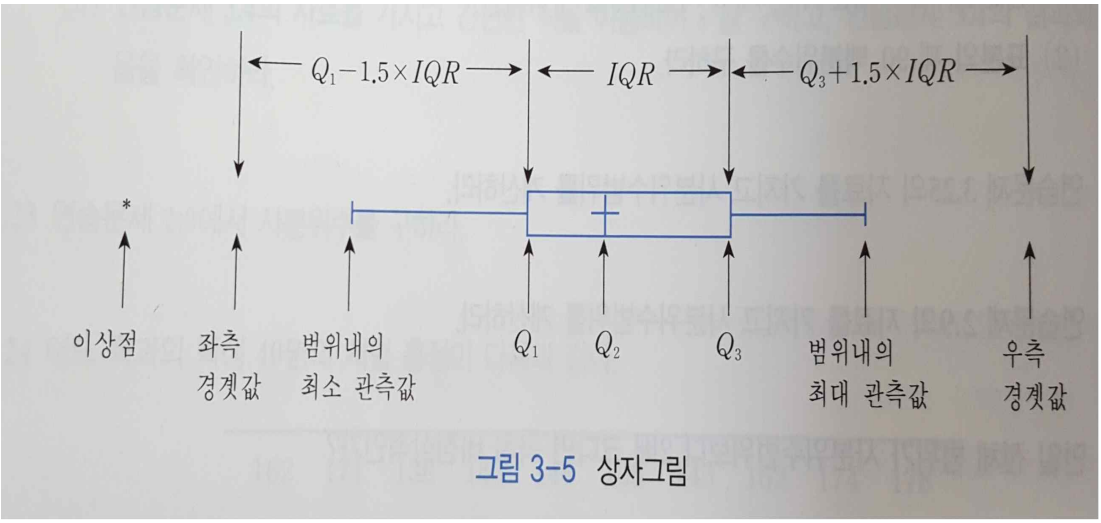


그림 3-5 상자그림

[예제] 어떤 교차로에서 교통소음 정도를 측정한 값이 아래와 같다. 측정자료를 이용하여 상자(수염)그림을 그려라.

55.9	63.8	57.2	59.8	65.7	62.7	60.8	51.3	61.8	56.0
66.9	56.8	66.2	64.6	59.5	63.1	60.6	62.0	59.4	67.2
63.6	60.5	66.8	61.8	64.8	55.8	55.7	77.1	62.1	61.0
58.9	60.0	66.9	61.7	60.3	51.5	67.0	60.2	56.2	59.4
67.9	64.9	55.7	61.4	62.6	56.4	56.4	69.4	57.6	63.8

【풀이】

① 상기 관측값들을 작은 값에서 큰 값 순으로 재정렬하기

② 사분위수 구하기, $n=50$

□ Q_1: $p=0.25 \rightarrow np=50\times0.25=12.5$(정수아님) $\rightarrow m=12+1=13$

$\rightarrow Q_1=x_{(m)}=x_{(13)}=57.6$

□ Q_2: $p=0.5 \rightarrow np=50\times0.5=25$(정수)

$\rightarrow Q_2=\dfrac{x_{(25)}+x_{(26)}}{2}=\dfrac{60.8+61.0}{2}=60.9$

□ Q_3: $p=0.75 \rightarrow np=50\times0.75=37.5$(정수아님) $\rightarrow m=37+1=38$

$\rightarrow Q_3=x_{(m)}=x_{(38)}=63.8$

③ 사분위수범위 구하기: $IQR=Q_3-Q_1=63.8-57.6=6.2$

④ 좌측[우측] 경계값 구하기

▫ 좌측 경계값 $= Q_1 - (1.5 \times IQR) = 57.6 - (1.5 \times 6.2) = 48.3$

▫ 우측 경계값 $= Q_3 + (1.5 \times IQR) = 63.8 + (1.5 \times 6.2) = 73.1$

⑤ 관측값들 중에서 좌우측 경계값 밖에 위치한 이상점 구하기

→ 48.3보다 작거나 73.1보다 큰 값 구하기 → 77.1

⑥ 상자(수염)그림 그리기

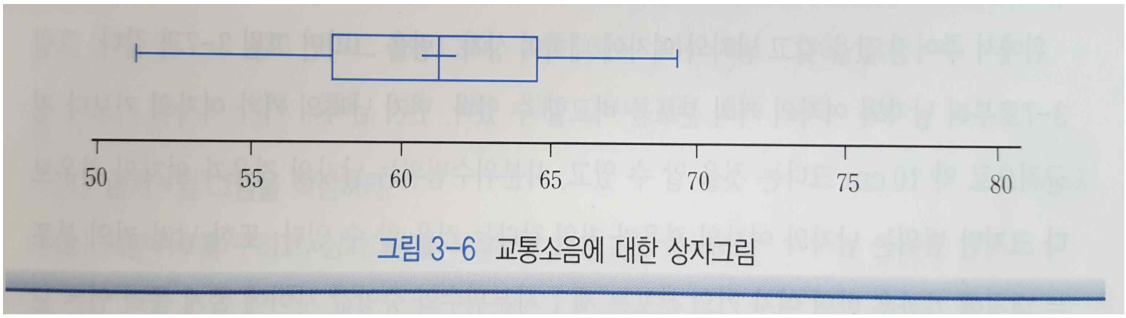


그림 3-6 교통소음에 대한 상자그림

(8) 다섯숫자 요약 (5-number summary)

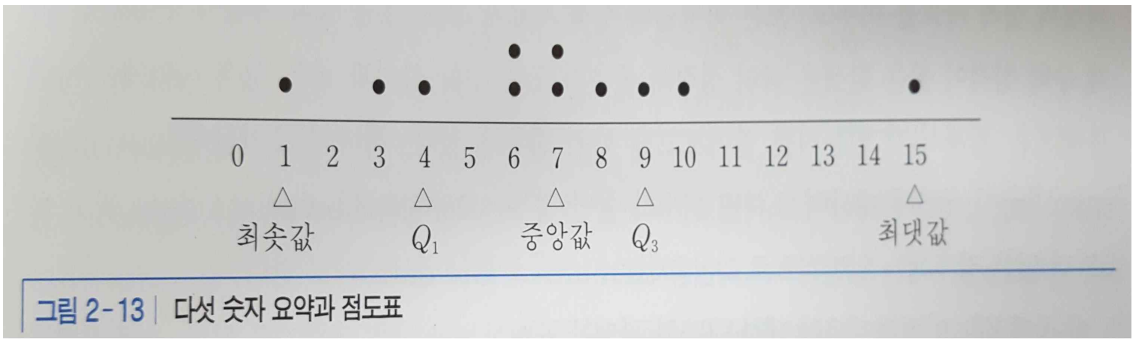


그림 2-13 | 다섯 숫자 요약과 점도표

2.7 이변량자료 (bivariate data)

(1) 단변량자료 (univariate data): 하나의 변수에 대한 자료
(2) 이변량자료 (bivariate data)
: 하나의 조사단위에 대하여 조사된 변수가 둘인 경우
(3) 다변량자료 (multivariate data)

2.8 이변량 범주형 자료의 구조와 표현

: 결합도수분포표 (joint frequency table[cross-table])
분할표 (contingency table)
cf) 양적자료: 산점도 (scatter plot)

[예제 2.11] 통계학개론을 수강하는 400명의 학생들에게 시험을 본 후 문제수준에 대하여 조사하였다. 남녀별로 구분 정리하여 다음 분할표를 얻었다.

	A	B	C
1	112	36	28
2	84	68	72

1: 남자, 2: 여자
A: 어렵다, B: 보통이다, C: 쉽다

	A	B	C	계
1	112 (0.28)	36 (0.09)	28 (0.07)	176 (0.44)
2	84 (0.21)	68 (0.17)	72 (0.18)	224 (0.56)
계	196 (0.49)	104 (0.26)	100 (0.25)	400 (1.00)

2.9 이변량 양적 자료의 표현

2.9.1 산점도 (scatter plot)

[예제 2.12] 이변량 양적 자료를 조사하여 다음의 15쌍의 관측치들을 얻었다. 관측치들의 산점도를 그려라.

(2, 4) (3, 7) (4, 6) (5, 8) (6, 10) (7, 9) (8, 13) (9, 11)
(4, 5) (6, 7) (3, 4) (7, 10) (2, 3) (5, 7) (8, 11)

【풀이】

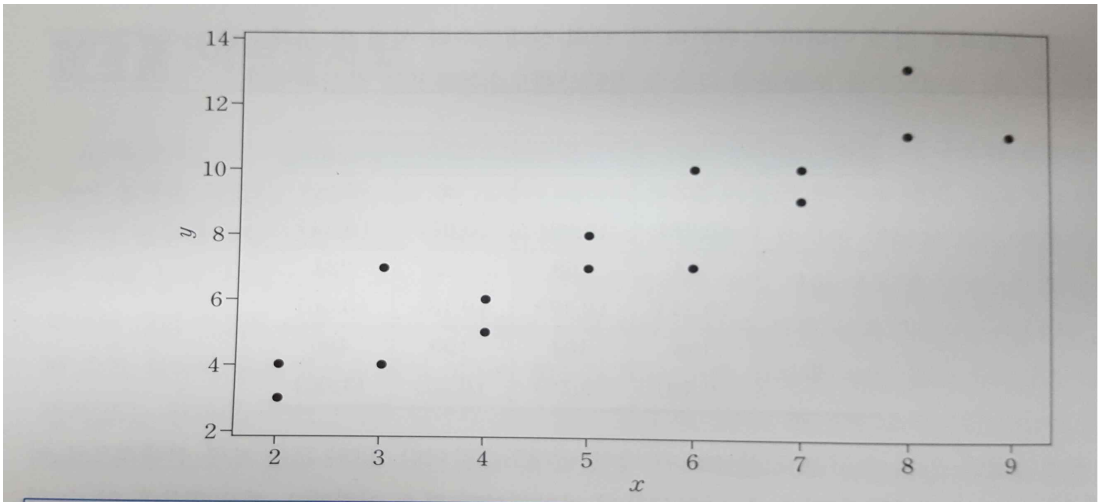


그림 2-14 | 산점도

2.9.2 선형관계의 척도로서의 상관계수

(1) 상관계수

n개의 자료쌍 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$에 대하여

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

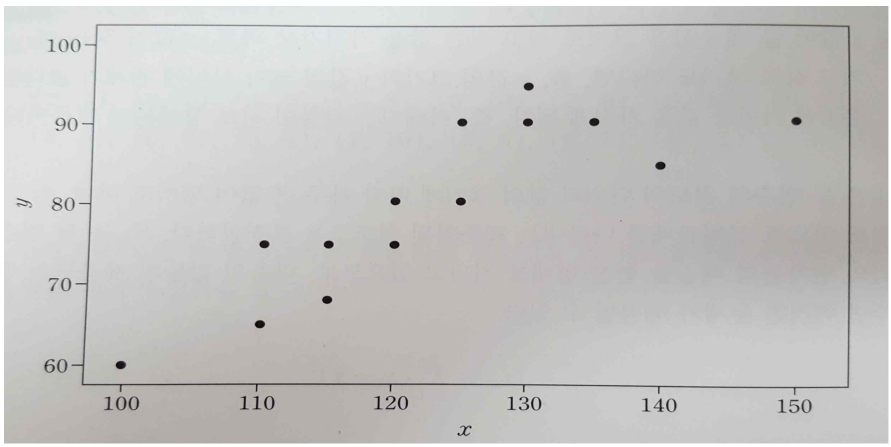
$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

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

[예제 2.13] 중학교 1학년 학생들에 대하여 IQ와 성적의 관계를 알아보기 위해 15명의 학생을 임의추출하여 다음의 자료를 얻었다. 산점도를 그리고 상관계수를 구하라.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
IQ	110	130	125	120	115	120	125	130	150	140	100	110	115	120	135
성적	75	90	80	80	70	75	90	95	90	85	60	65	75	75	90

【풀이】



학생	$X(IQ)$	Y(성적)	X^2	Y^2	XY
1	110	75	12,100	5,625	8,250
2	130	90	16,900	8,100	11,700
3	125	80	15,625	6,400	10,000
4	120	80	14,400	6,400	9,600
5	115	70	13,225	4,900	8,050
6	120	75	14,400	5,625	9,000
7	125	90	15,625	8,100	11,250
8	130	95	16,900	9,025	12,350
9	150	90	22,500	8,100	13,500
10	140	85	19,600	7,225	11,900
11	100	60	10,000	3,600	6,000
12	110	65	12,100	4,225	7,150
13	115	75	13,225	5,625	8,625
14	120	75	14,400	5,625	9,000
15	135	90	18,225	8,100	12,150
계	1,845	1,195	229,225	96,675	148,525

$$S_{xx}=\sum_{i=1}^{15}x_i^2-15\bar{x}^2=229{,}225-(1{,}845\times1{,}845)/15=2{,}290.00$$

$$S_{yy}=\sum_{i=1}^{15}y_i^2-15\bar{y}^2=96{,}675-(1{,}195\times1{,}195)/15=1{,}473.34$$

$$S_{xy}=\sum_{i=1}^{15}x_iy_i-15\bar{x}\,\bar{y}=148{,}525-(1{,}845\times1{,}195)/15=1{,}540.00$$

이므로 상관계수는 $r=\dfrac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}=0.8384$가 된다. ■

(2) 상관계수의 성질

① $-1 \leq r \leq 1$

② 산점도의 형태

❑ $r > 0$이면 작은 쪽에서 큰 쪽으로 올라가는 형태를 취하고 1에 가까워질수록 기울기가 양인 직선에 가까워진다.

❑ $r < 0$이면 큰 쪽에서 작은 쪽으로 내려가는 형태를 취하고 -1에 가까워질수록 기울기가 음인 직선에 가까워진다.

❑ r이 0에 가까워질수록 원형으로 넓게 퍼지는 형태가 된다.