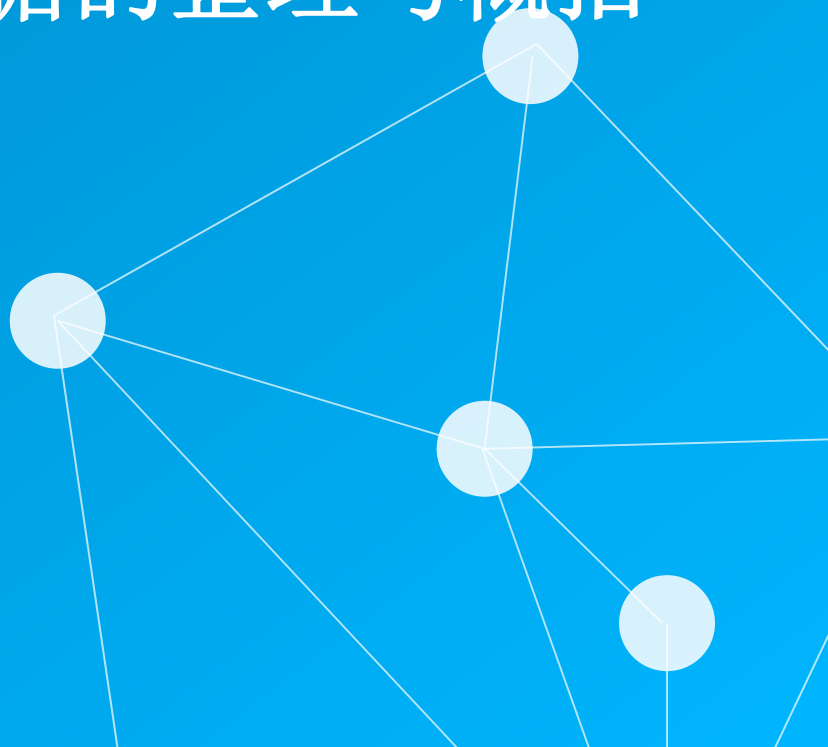
DATA
STRUCTURES
USING C++

02 CHAPTER

数据的整理与概括



第2章
数据的整理与概括

2.1 频数分布表

2.2 检查表

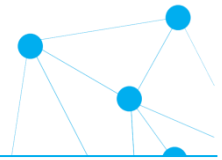
2.3 히스토그램

2.4 각종 그래프

2.5 중심위치의 척도

2.6 산포의 척도

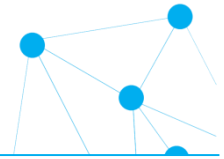
2.1 频数分布表



[例 2-1] R 基本包中的‘iris’数据的第二列‘花萼宽度’(Sepal.Width)的频数分布表

	代表值	频数	累计频数	相对频数	相对累计频数
(2, 2.2]	2.1	4	4	0.026666667	0.02666667
(2.2, 2.4]	2.3	7	11	0.046666667	0.07333333
(2.4, 2.6]	2.5	13	24	0.086666667	0.16000000
(2.6, 2.8]	2.7	23	47	0.153333333	0.31333333
(2.8, 3]	2.9	36	83	0.240000000	0.55333333
(3, 3.2]	3.1	24	107	0.160000000	0.71333333
(3.2, 3.4]	3.3	18	125	0.120000000	0.83333333
(3.4, 3.6]	3.5	10	135	0.066666667	0.90000000
(3.6, 3.8]	3.7	9	144	0.060000000	0.96000000
(3.8, 4]	3.9	3	147	0.020000000	0.98000000
(4, 4.2]	4.1	2	149	0.013333333	0.99333333
(4.2, 4.4]	4.3	1	150	0.006666667	1.00000000

2.1 频数分布表



- [例 2-2]* 规格 [5.00±1.00]Ω, 抵抗数据100个 [tab2-1.csv] 频数分布表 (15个 区间)

4.91	5.03	5.07	5.21	4.74	5.03	5.08	4.95	4.89	4.65
4.79	5.01	4.77	4.95	4.59	5.07	4.97	5.19	5.05	5.27
4.77	4.76	5.11	5.17	4.94	4.69	5.01	5.11	4.75	5.05
5.01	4.93	5.01	5.08	4.69	4.89	5.23	4.99	5.14	4.95
4.91	4.81	4.99	4.89	4.79	4.74	5.09	5.07	5.25	5.28
4.87	4.88	4.87	4.75	4.99	4.59	5.07	4.99	4.99	5.07
4.94	5.29	4.97	4.99	4.95	4.65	4.77	4.83	4.95	5.05
5.02	4.97	5.07	4.89	4.77	4.88	5.08	4.78	5.23	5.18
4.66	4.95	5.19	4.84	4.93	4.98	5.08	4.85	5.04	4.89
4.79	5.09	4.98	4.94	5.17	4.88	4.96	4.92	4.79	5.10

[表 2-2] 抵抗数据 ([表 2-1])的频数分布表

급	구간	대푯값	도수	누적도수	상대도수	상대누적도수
1	[4.575, 4.625]	4.60	2	2	0.02	0.02
2	[4.625, 4.675]	4.65	3	5	0.03	0.05
3	[4.675, 4.725]	4.70	2	7	0.02	0.07
4	[4.725, 4.775]	4.75	9	16	0.09	0.16
5	[4.775, 4.825]	4.80	6	22	0.06	0.22
6	[4.825, 4.875]	4.85	5	27	0.05	0.27
7	[4.875, 4.925]	4.90	11	38	0.11	0.38
8	[4.925, 4.975]	4.95	15	53	0.15	0.53
9	[4.975, 5.025]	5.00	13	66	0.13	0.66
10	[5.025, 5.075]	5.05	12	78	0.12	0.78
11	[5.075, 5.125]	5.10	9	87	0.09	0.87
12	[5.125, 5.175]	5.15	3	90	0.03	0.9
13	[5.175, 5.225]	5.20	4	94	0.04	0.94
14	[5.225, 5.275]	5.25	4	98	0.04	0.98
15	[5.275, 5.325]	5.30	2	100	0.02	1.00
			100		1.00	

2.2 检查表

- 2.2.1 计数表(tally sheet)

결점유형	5	10	15	20	25	30	35	도수
색상	𝍸	𝍸	𝍸					12
마무리	𝍸							1
인쇄	𝍸							2
복원력	𝍸	𝍸	𝍸	𝍸				16
얼룩	𝍸							1
선명도	𝍸	𝍸						8
흠집	𝍸	𝍸						10

2.2 检查表

- 2.2.2 列联表(contingency table)

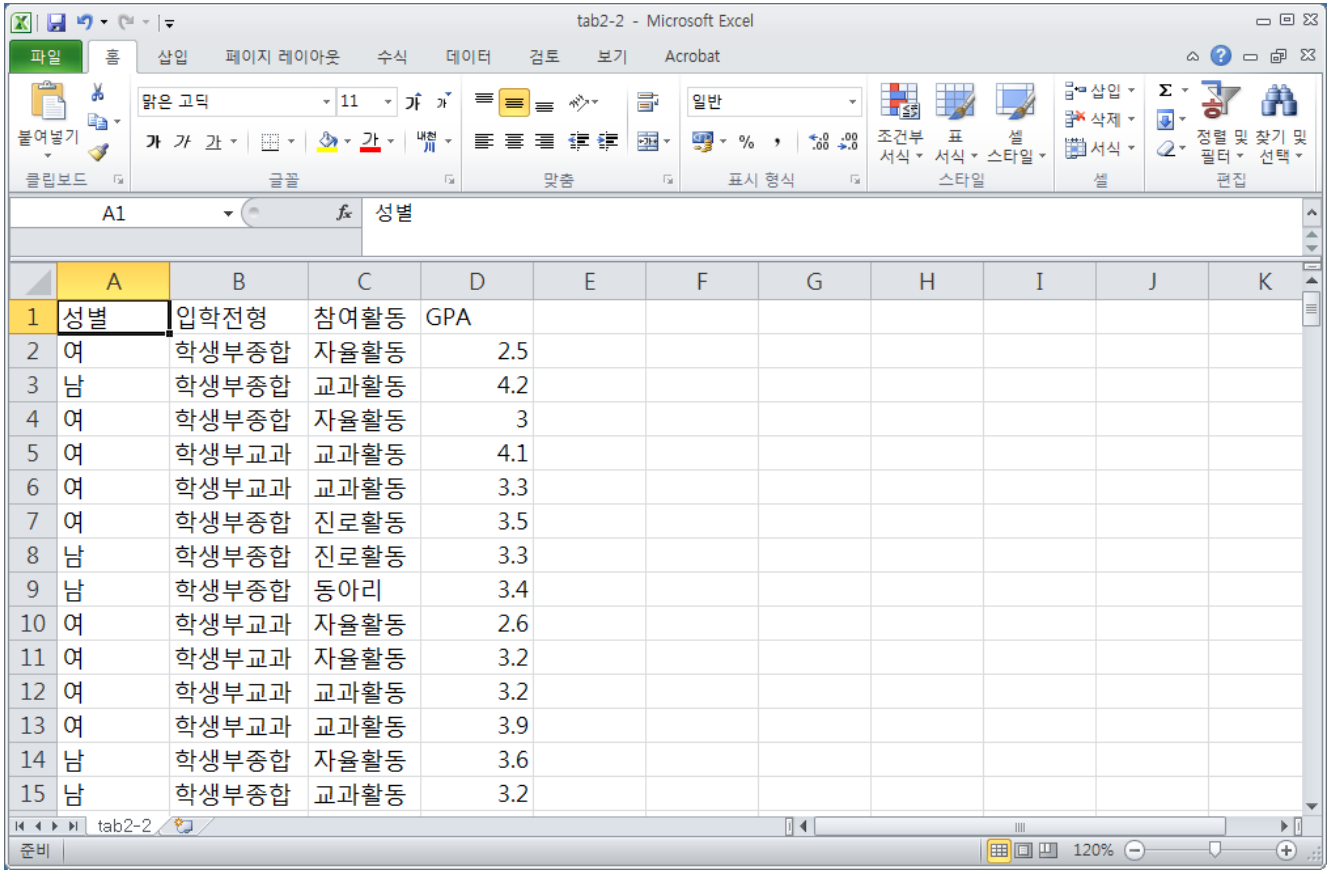
	교		대	
결점의 유형	낮	저녁	밤	합계
복원력	\	𝍸	𝍸	16
색상	\	𝍸 \	\	12
흠집	\	𝍸	\	10
선명도		\	𝍸 \	8
	5	23	18	46

[例 2-3] 对S大学A学部新生210名的调查数据 [tab2-2.csv]

(1) 新生的入学情况分布表

(2) 新生主要参与活动的分布表

(3) 新生的入学情况和参与活动联合分布表



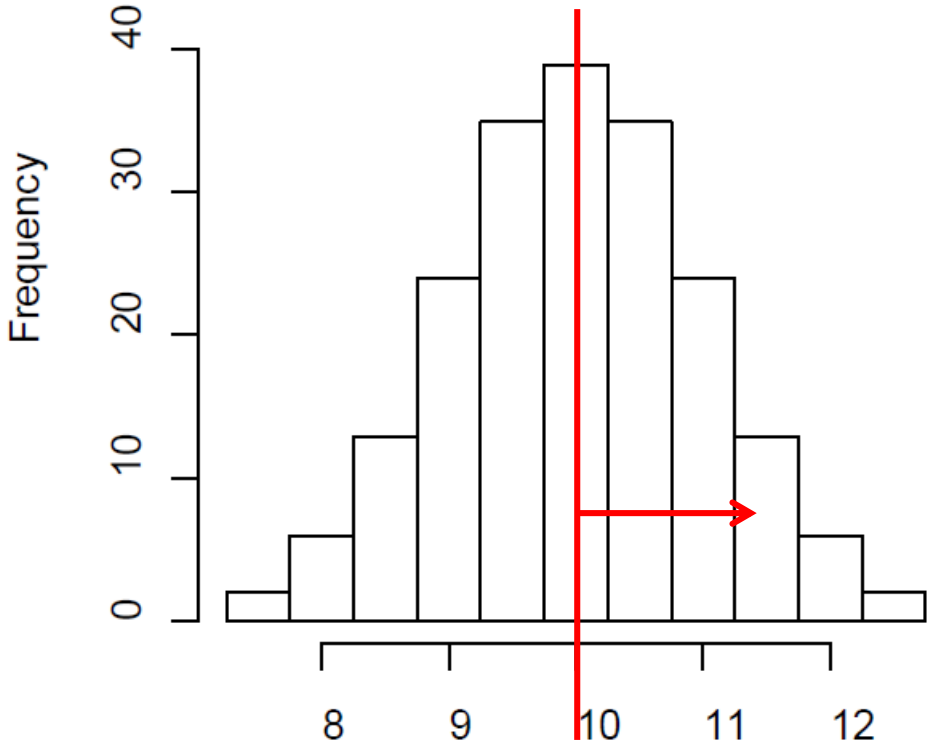
tab2-2 - Microsoft Excel

A1: 성별

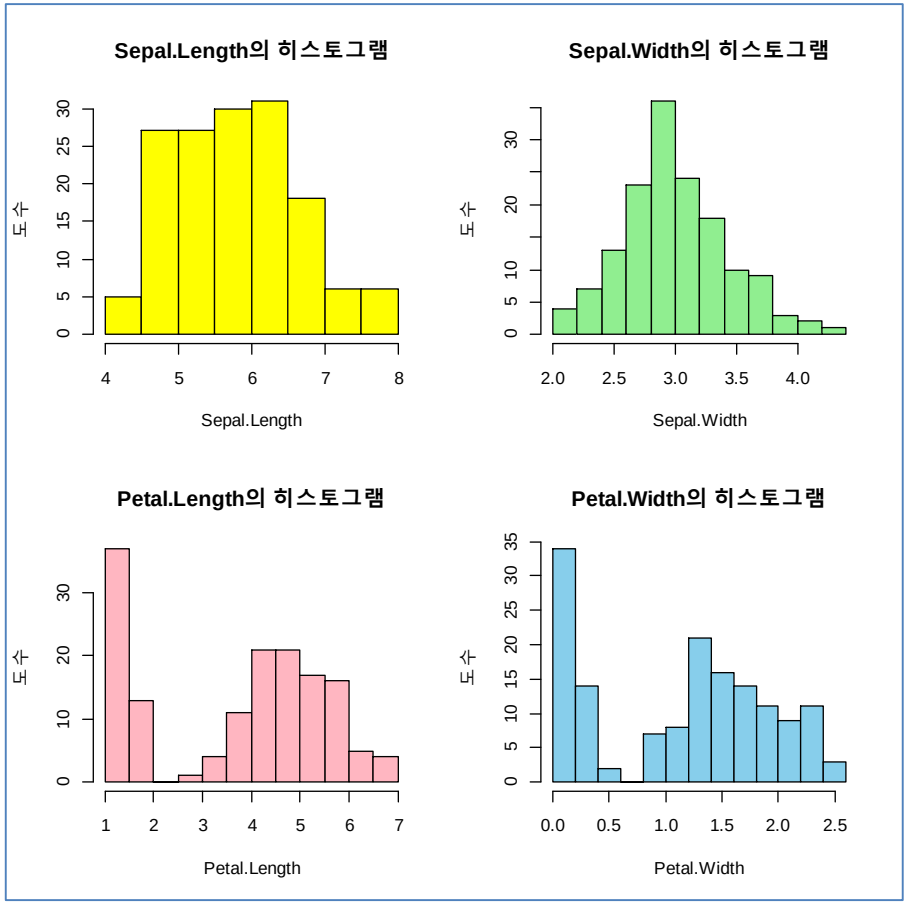
	A	B	C	D
1	성별	입학전형	참여활동	GPA
2	여	학생부종합	자율활동	2.5
3	남	학생부종합	교과활동	4.2
4	여	학생부종합	자율활동	3
5	여	학생부교과	교과활동	4.1
6	여	학생부교과	교과활동	3.3
7	여	학생부종합	진로활동	3.5
8	남	학생부종합	진로활동	3.3
9	남	학생부종합	동아리	3.4
10	여	학생부교과	자율활동	2.6
11	여	학생부교과	자율활동	3.2
12	여	학생부교과	교과활동	3.2
13	여	학생부교과	교과활동	3.9
14	남	학생부종합	자율활동	3.6
15	남	학생부종합	교과활동	3.2

2.3 直方图(histogram)

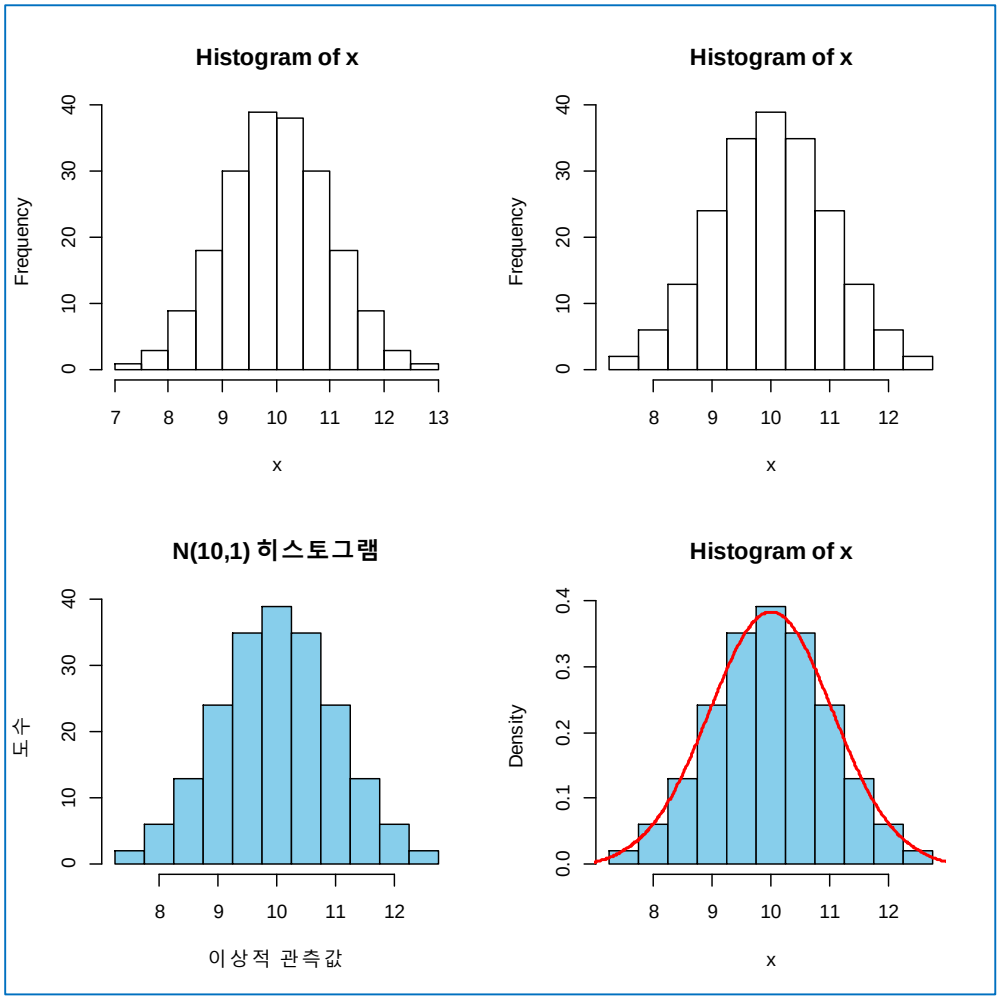
- 根据样本数据推测总体分布的特点
  ① 总体分布的形态(shape)
  ② 总体分布的中心位置(location)
  ③ 总体分布的散布(spread)



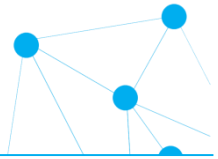
- [例 2-4] R基本包中的'iris'数据的第1列到第4列变量的直方图用一个画面显示



- [例 2-5]* 稳定分布 N(10,1)的分布形态



2.3 直方图(histogram)



- 不稳定(异常)分布

1. 孤岛型

   分布不稳定，混杂少量受影响的分布的情况

2. 双峰型

   分布具有两种特征，分有下属分布的情况

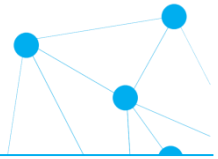
3. 缺口型

   因测量器本身的问题，部分区域的值未测量到的情况

4. 绝壁型

   全面检查后，除去一些临界值以下(以上)的产品的情况

2.3 直方图(histogram)



[例 2-6]* 不稳定分布直方图的正常分布 N(10, 1)

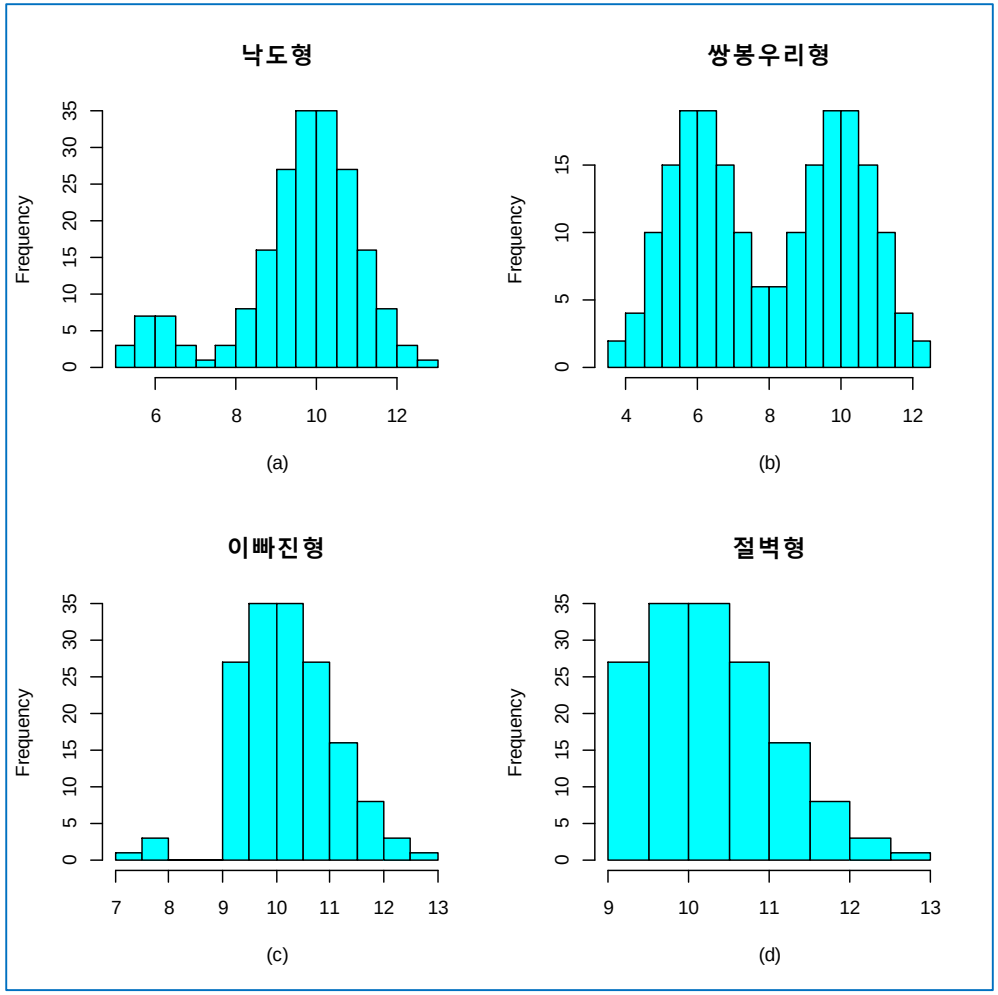
(a) 孤岛型：90%的N(10, 1)和10%的受影响的$N(6, 0.5^2)$

(b) 双峰型：50%的N(10, 1)和50%的N(6, 1)

(c) 缺口型：因测量器的问题，[8,9]区域的值未测量到

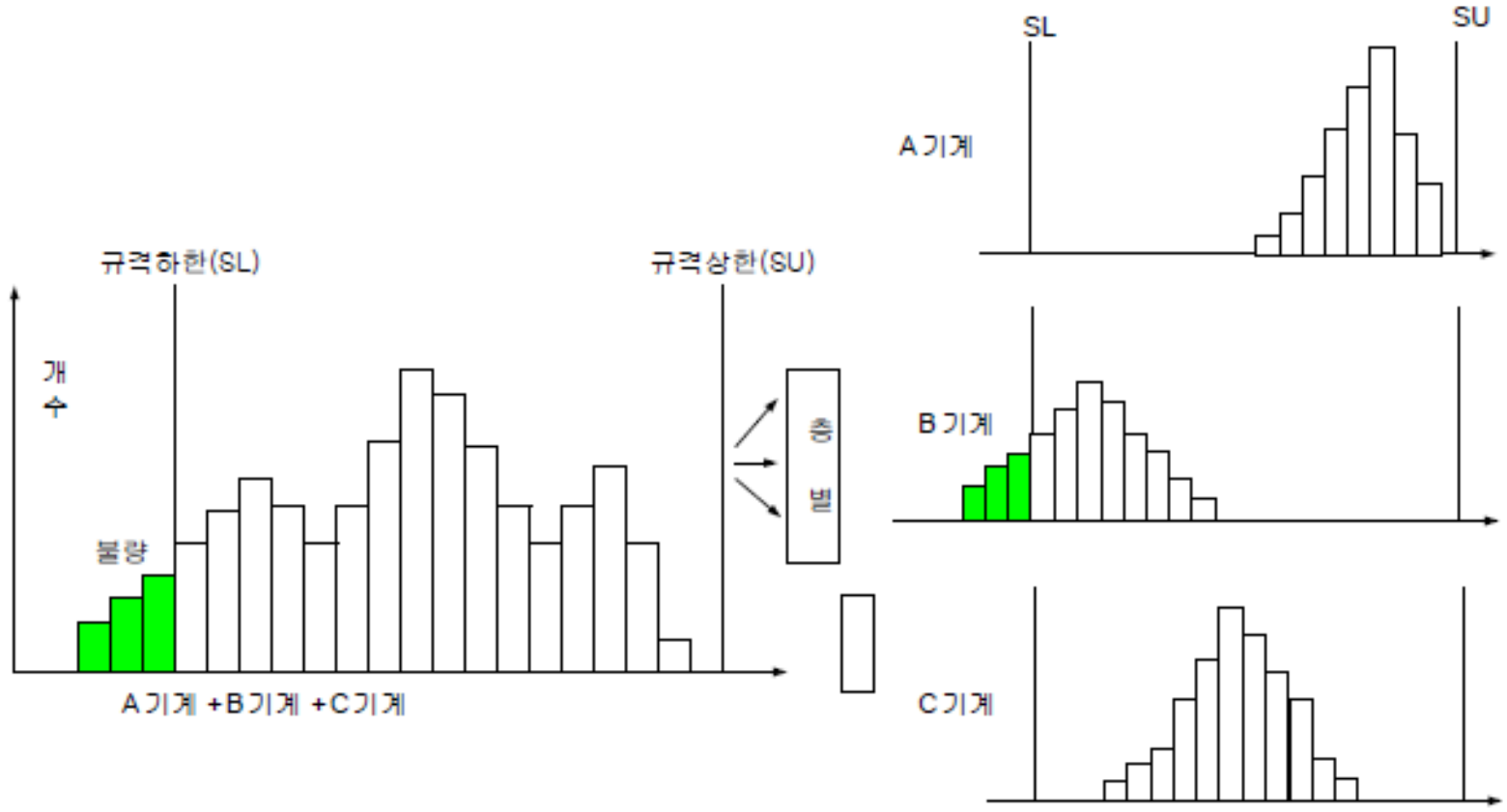
(d) 绝壁型：全面检查后，除去不足9.0的数据

[例 2-6]* 不稳分布的几种形态



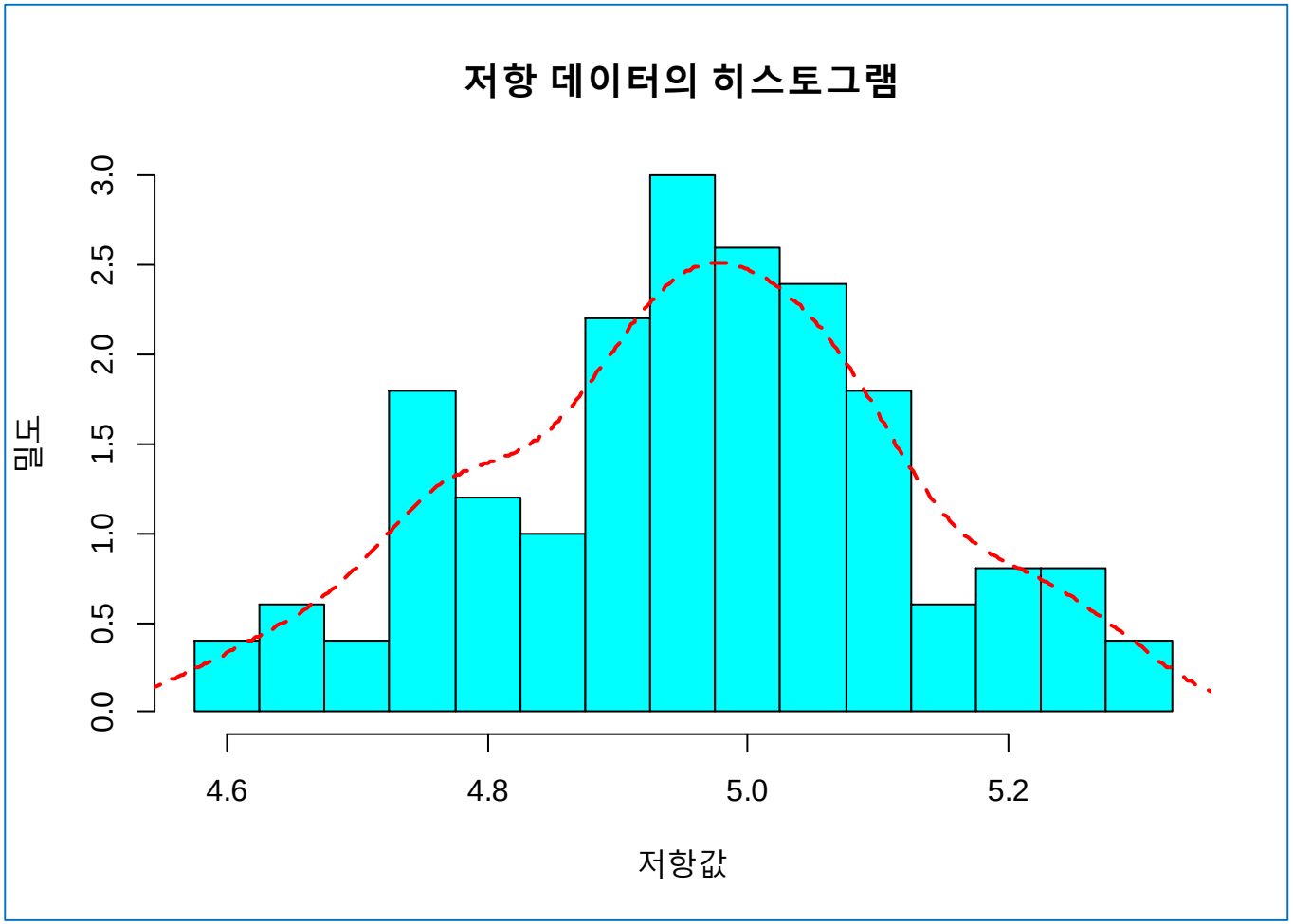
2.3 直方图(histogram)

- 分层(stratified)直方图

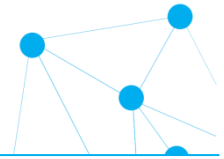


2.3 直方图(histogram)

[例 2-7] 抵抗数据直方图



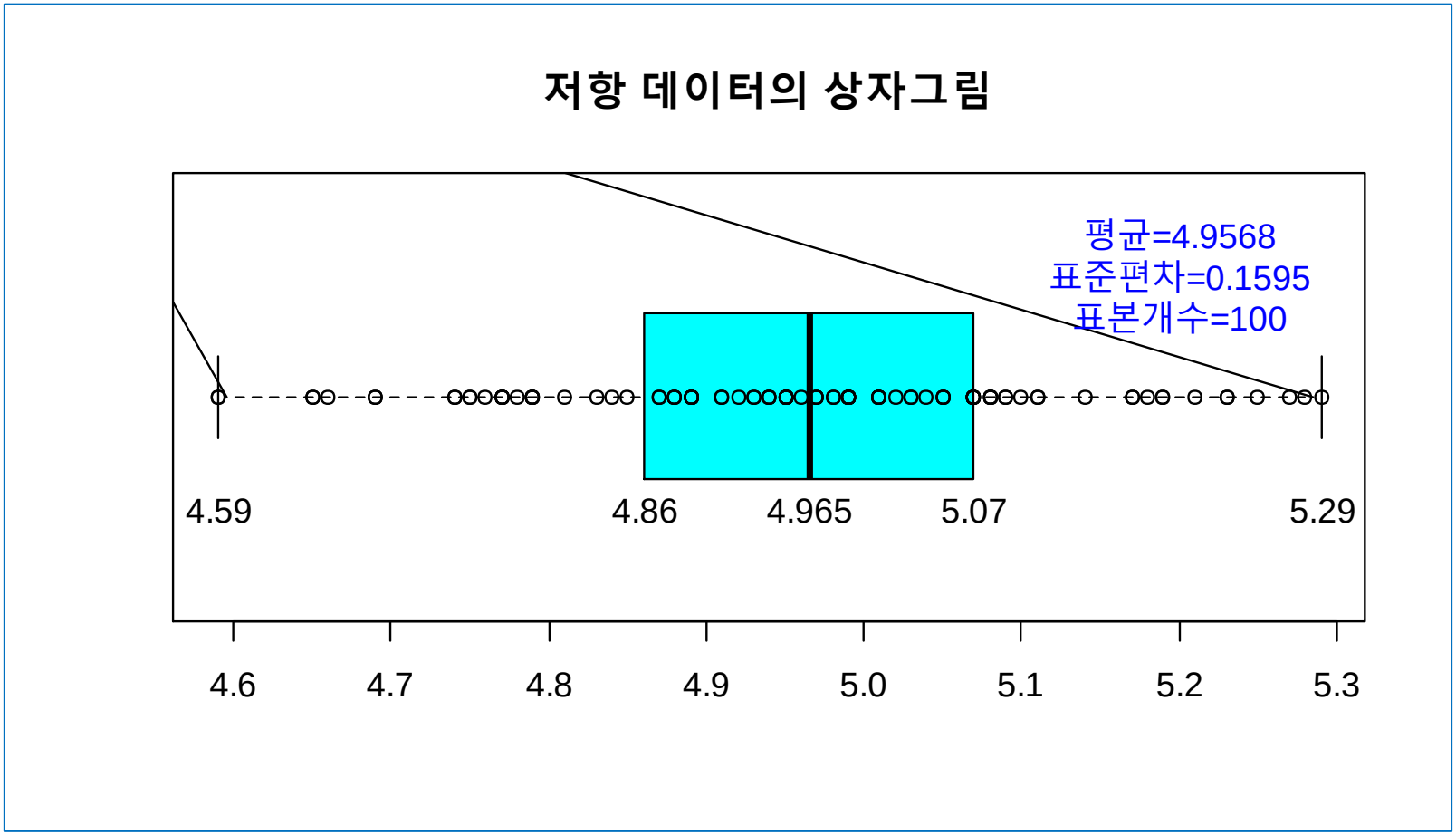
2.4.1 茎叶图(stem-and-leaf plot)



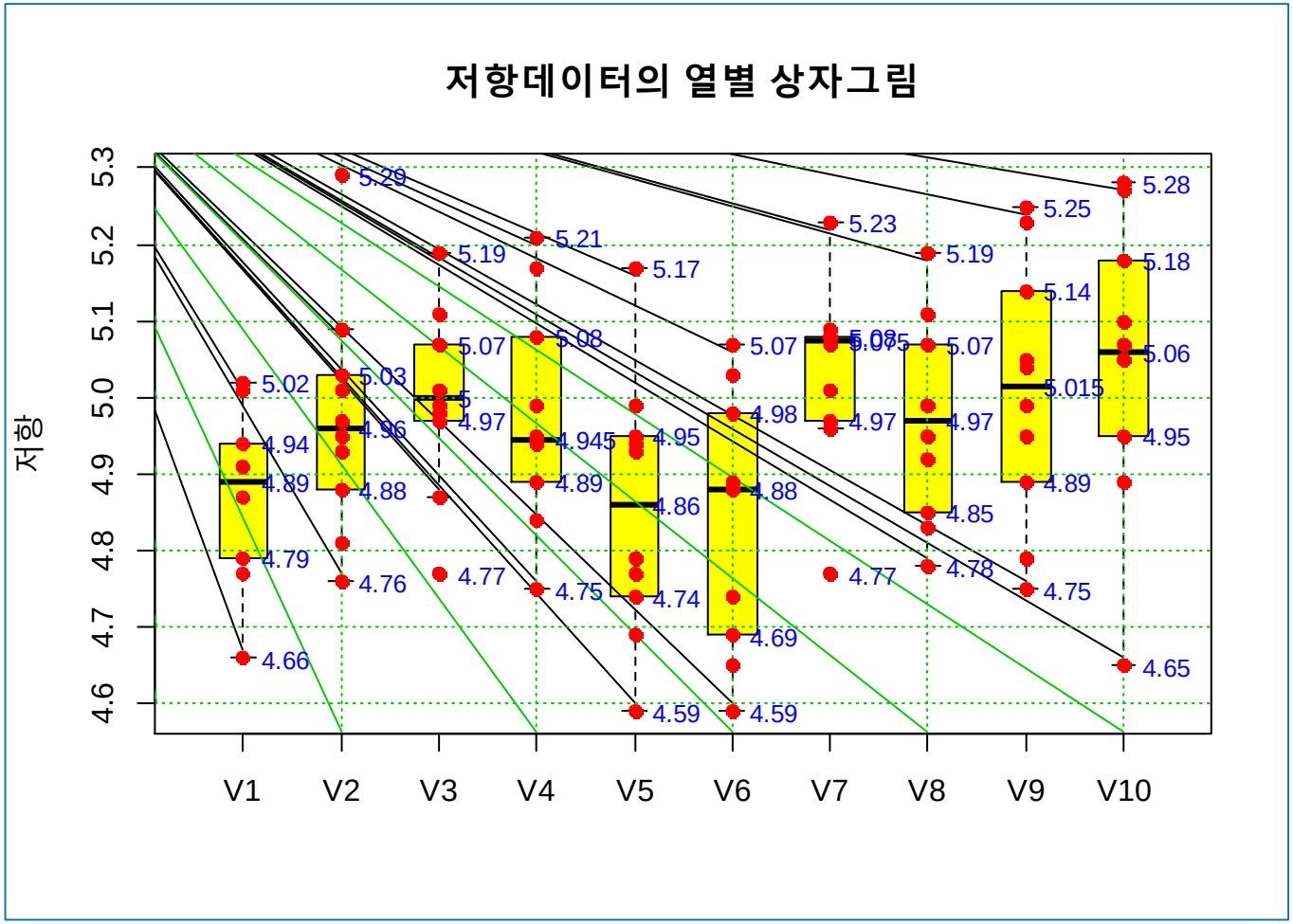
- [例 2-8] [表 2-1]的抵抗数据 → 茎叶图

```
# 抵抗数据茎叶图 ⇒ stem( ) 函数
stem(x)
 45  99
 46  55699
 47  44556777789999
 48  13457788899999
 49  1123344455555677788999999
 50  11112334555777777888899
 51  011477899
 52  1335789
```

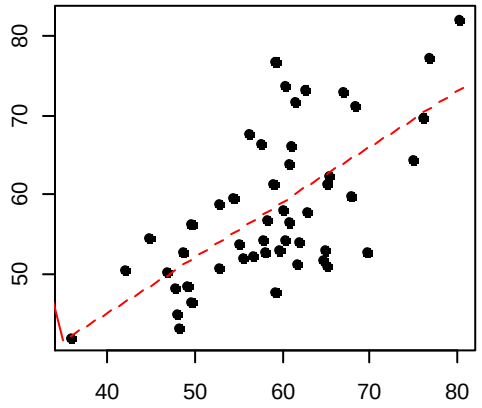
- [例 2-9] 抵抗数据箱线图


저항 데이터의 상자그림
평균=4.9568
표준편차=0.1595
표본개수=100
4.59
4.86
4.965
5.07
5.29
4.6
4.7
4.8
4.9
5.0
5.1
5.2
5.3

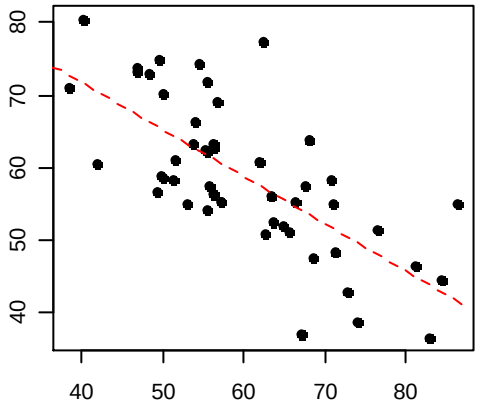
- [例 2-10] 抵抗数据各列箱线图



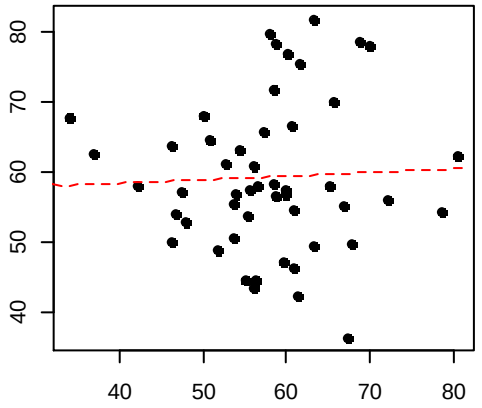
2.4.3 散点图(scatter diagram)



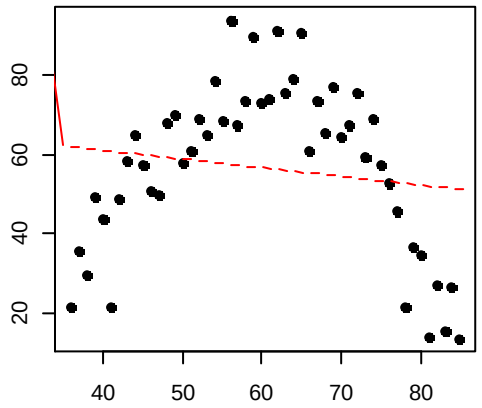
(a) 양의 상관



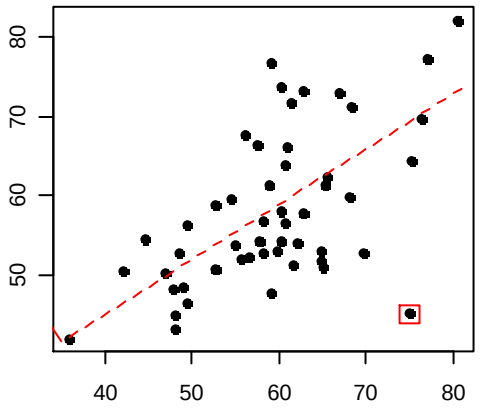
(b) 음의 상관



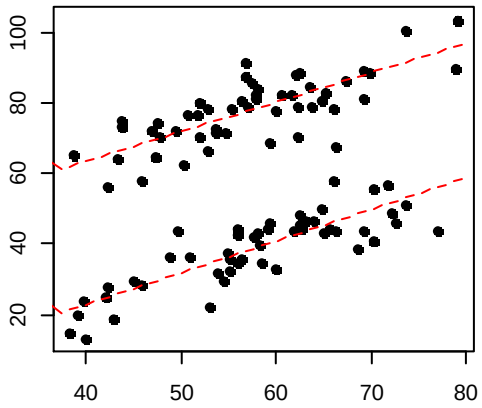
(c) 상관 희박



(d) 곡선 관계

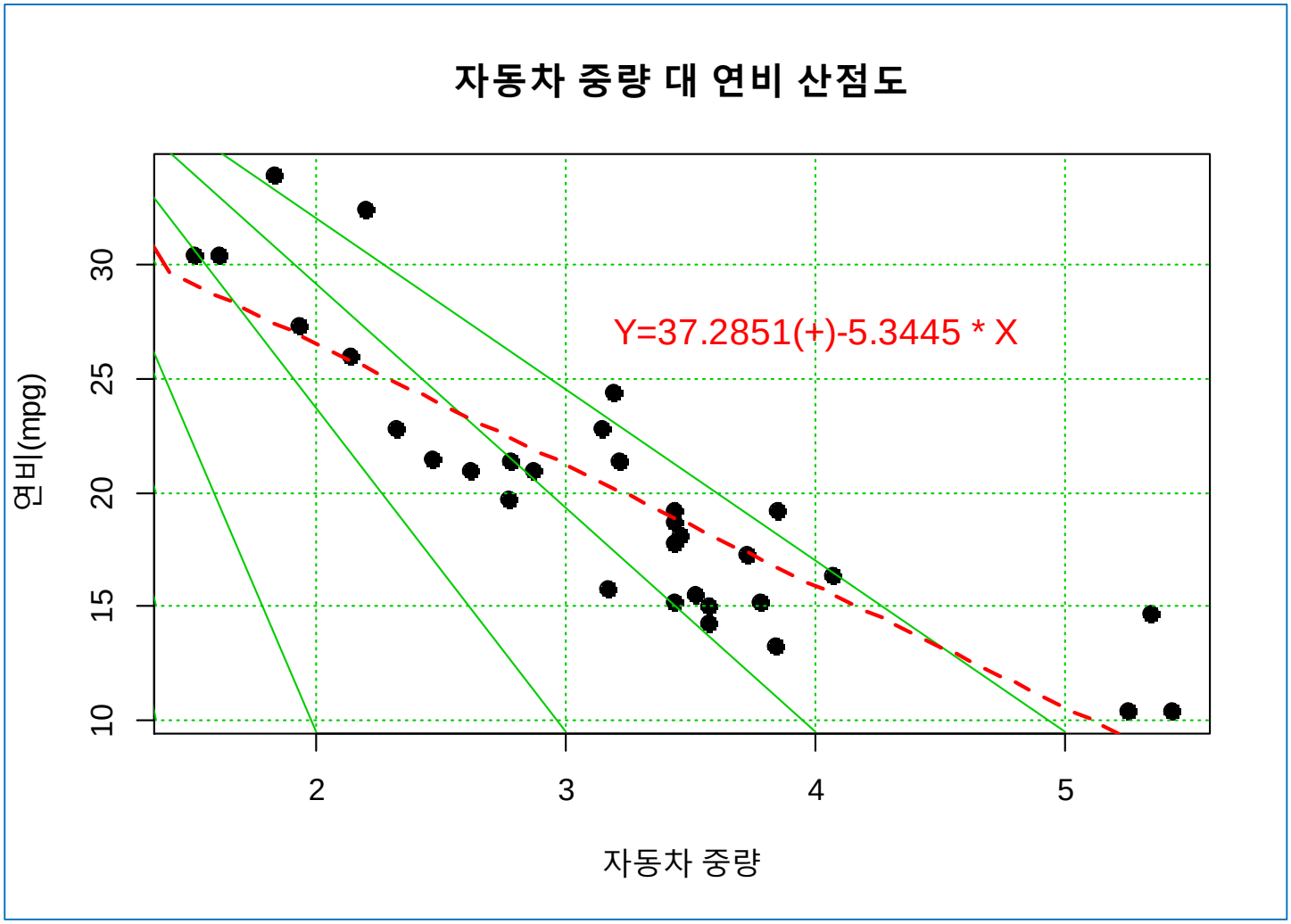


(e) 이상점

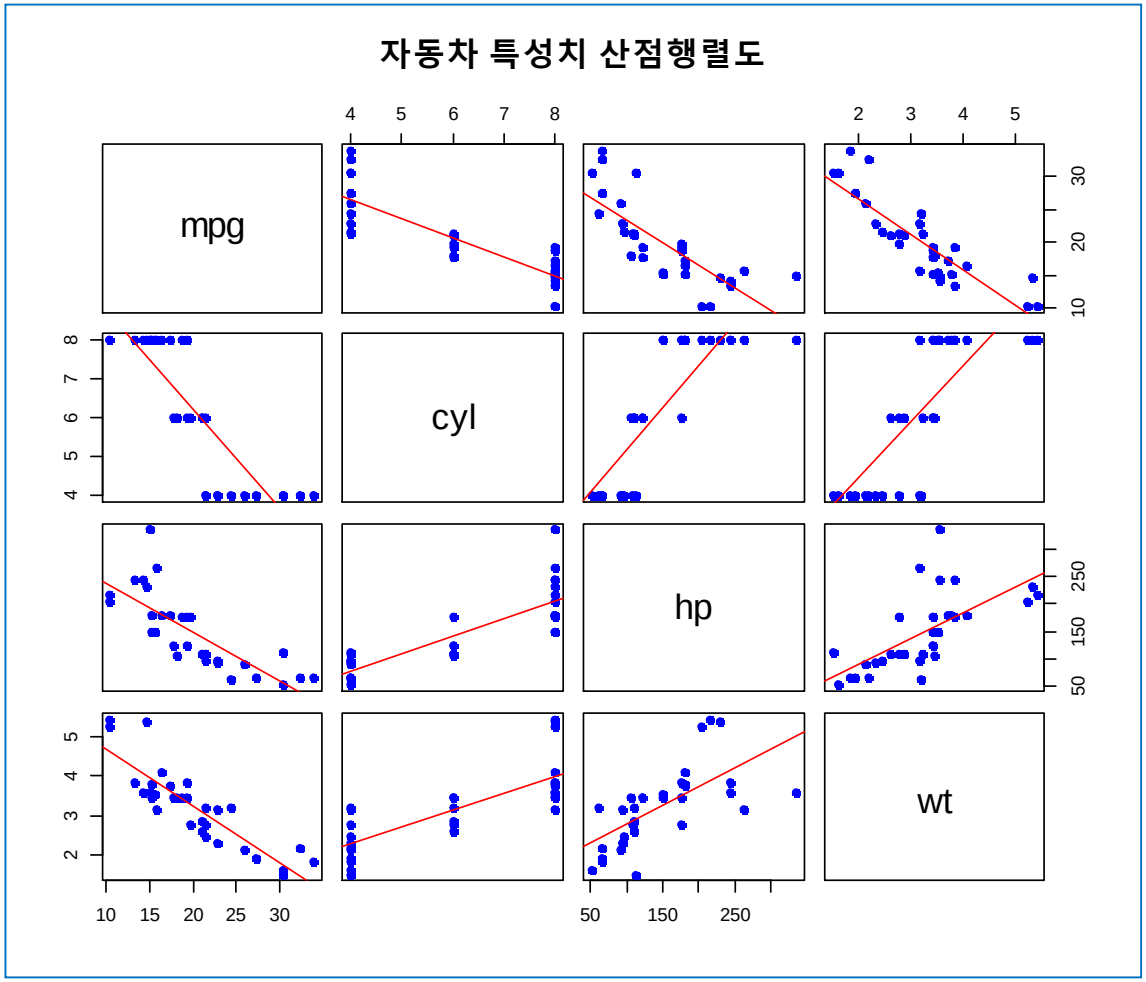


(f) 층별

- [例 2-11] ‘mtcars’数据中重量(wt)– 燃料费(mpg)散点图



[例 2-12] ‘mtcars’数据的散点矩阵图



2.5 中心位置的计量尺度

[例 2-13] 有关10个样本数据的中心位置的计量尺度

5	4	6	3	5	4	3	9	5	10

(1) 均值 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{54}{10} = 5.4$

(2) 中位数

3	3	4	4	5	5	5	6	9	10

$\Rightarrow \frac{5+5}{2} = 5$

(3) 众数

3	3	4	4	5	5	5	6	9	10

$\Rightarrow 5$

(4) 几何平均数 $\bar{x}_g = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = (9,720,000)^{1/10} \approx 4.998$

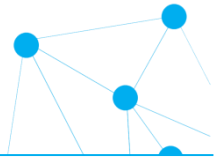
(5) 调和平均数 $\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \approx \frac{10}{2.1444} \approx 4.663$

(6) 切尾均值

3	4	4	5	5	5	6	9

$\Rightarrow \bar{x}_{0.1} = \frac{1}{8}\sum_{i=2}^{9} x_i = \frac{41}{8} = 5.125$

2.5 中心位置的计量尺度



- 中心位置代表值设定基准

① 利用定类尺度测量的数据，使用众数

② 如果分布对称且不存在异常点，使用样本均值

③ 如果分布不对称或存在异常值，使用中位数，并以样本均值作为参考值进行比较

④ 利用定位尺度测量的数据使用中位数

[例 2-14] 抵抗数据的中心位置的计量尺度

2.6 散布的计量尺度

① 样本方差 $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \dfrac{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 / n}{n-1}$

② 样本标准差 $s = \sqrt{s^2}$

③ 数据极差 $R = x_{\max} - x_{\min}$

④ 四分位数区间 $IQR = Q_3 - Q_1$

⑤ 变异系数 $CV = s / \bar{x}$

2.6 散布的计量尺度

[例 2-15] 关于10个样本数据散布的计量尺度

5	4	6	3	5	4	3	9	5	10

① 样本方差 $s^2 = \left\{ \sum_{i=1}^{10} x_i^2 - \frac{(\sum_{i=1}^{10} x_i)^2}{10} \right\} / (10-1) = \left(\frac{342 - 54^2/10}{9} \right) = 5.60$

② 样本标准差 $\Rightarrow s = \sqrt{5.60} \approx 2.366$

③ 数据极差 $R = x_{\max} - x_{\min} = 10 - 3 = 7$

3	3	4	4	5	5	5	6	9	10

④ 四分位数区间

$1 + 0.25 \times (10-1) = 3.25 \Rightarrow Q_1 = 4$

$1 + 0.75 \times 9 = 7.75 \Rightarrow Q_3 = 5 + (6-5) \times 0.75 = 5.75$

$\Rightarrow Q_3 - Q_1 = 1.75$

⑤ 变异系数

$\Rightarrow CV = \frac{s}{\bar{x}} \approx \frac{2.366}{5.4} \approx 0.438 \ (43.8\%)$

2.6 散布的计量尺度

[例 2-16] 抵抗数据散布的计量尺度

① 样本方差

$$s^2 = \left\{ \sum_{i=1}^{100} x_i^2 - \frac{(\sum_{i=1}^{100} x_i)^2}{100} \right\} / (100-1) = \left(\frac{2459.5046 - 495.68^2 / 100}{99} \right) \approx 0.0254$$

② 样本标准差 $\Rightarrow s = \sqrt{s^2} \approx 0.1595$

③ 数据极差 $R = x_{\max} - x_{\min} = 5.29 - 4.59 = 0.70$

④ 四分位数区间

$$1 + 0.25 \times (100-1) = 25.75,\ x_{(25)} = 4.85,\ x_{(26)} = 4.87$$

$$\Rightarrow Q_1 = 4.85 + 0.75 \times (4.87 - 4.85) = 4.865$$

$$1 + 0.75 \times 99 = 75.25,\ x_{(75)} = x_{(76)} = 5.07$$

$$\Rightarrow Q_3 = 5.07 \qquad \Rightarrow Q_3 - Q_1 = 5.07 - 4.865 = 0.205$$

⑤ 变异系数 $\Rightarrow CV = \frac{s}{\bar{x}} \approx \frac{0.1595}{4.9568} \approx 0.0322\ (3.22\%)$