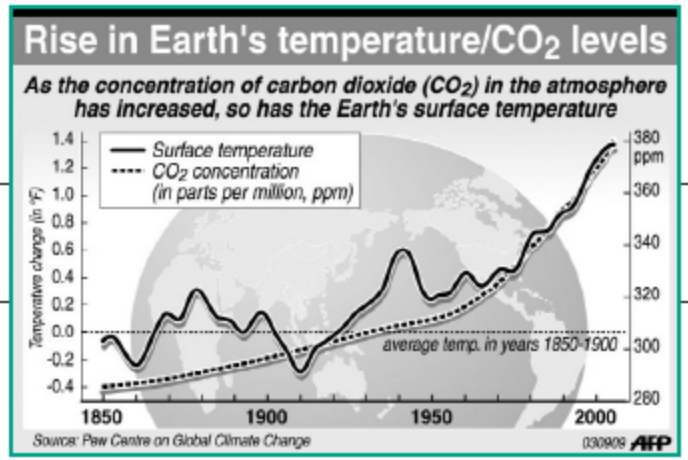
2



Data Summary and Presentation

LEARNING OBJECTIVES

After careful study of this chapter, you should be able to do the following:

1. Compute and interpret(解释) the sample mean(样本均值), sample variance(样本方差), sample standard deviation(样本标准差), sample median(样本中位数), and sample range(样本极差)….
2. Explain the concepts of sample mean, sample variance, population mean(总体均值), and population variance(总体方差).
3. Construct and interpret visual data displays(可视化数据表示), including the stem-and-leaf display(茎叶图), the histogram(直方图), and the box plot(箱线图) and understand how these graphical techniques are useful in uncovering(发现) and summarizing(汇总) patterns in data.
4. Explain how to use box plots and other data displays to visually compare two or more samples of data(使用箱线图和其他数据表示方式可视化地比较两个或者多个样本).
5. Know how to use simple time series plots(时间序列图) to visually display the important features of time-oriented data(基于时间的数据的重要特性).
6. Construct scatter plots and compute and interpret a sample correlation coefficient(相关系数).

2-1 Data Summary and Display

Sample Mean (样本均值)

If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample mean** is

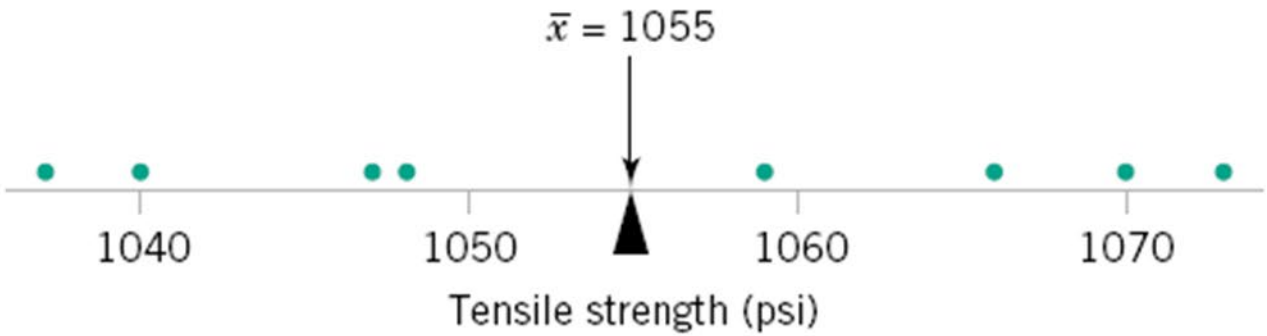
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

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

2-1 Data Summary and Display

EXAMPLE 2-1
O-Ring Strength:
Sample Mean

Figure 2-2 Dot diagram of O-ring tensile strength. The sample mean is shown as a balance point for a system of weights.



作为一个位置参数，样本均值的物理意义可以看作“平衡点”，即，每个观察/测代表放在x轴上的1磅重量，把一个支点放在$\bar{x}$的位置时，整个重量系统恰好平衡。

2-1 Data Summary and Display

Population Mean(总体均值)

For a finite population with N measurements/observations, the mean is

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

The sample mean(样本均值) is a reasonable estimate(合理估计) of the population mean(总体均值).

2-1 Data Summary and Display

Sample Variance and Sample Standard Deviation样本方差和样本标准差

If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, then the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \qquad (2\text{-}3)$$

The **sample standard deviation,** s, is the positive square root of the sample variance.

2-1 Data Summary and Display

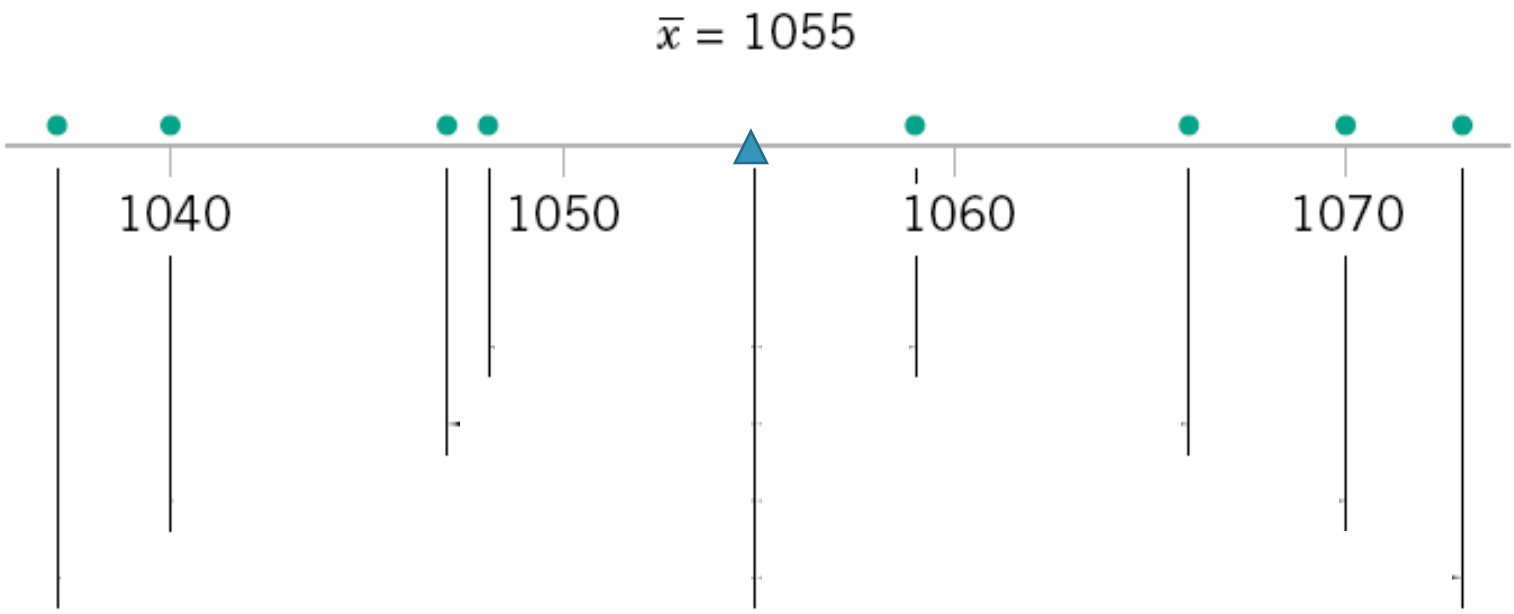


Figure 2-3 How the sample variance measures variability through the deviations $x_i - \bar{x}$.

样本方差如何通过偏差 $x_i - \bar{x}$ 测度变异性

2-1 Data Summary and Display

EXAMPLE 2-2

Table 2-1 Calculation of Terms for the Sample Variance and Sample Standard Deviation

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1048	−7	49
2	1059	4	16
3	1047	−8	64
4	1066	11	121
5	1040	−15	225
6	1070	15	225
7	1037	−18	324
8	1073	18	324
	8440	0.0	1348

2-1 Data Summary and Display

EXAMPLE 2-3
O-Ring Strength: Alternative Variance Calculation

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

$$\bar{x} =$$

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

The sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{8{,}905{,}548 - \frac{(8440)^2}{8}}{7} = \frac{1348}{7} = 192.57 \text{ psi}^2$$

The sample standard deviation is

$$s = \sqrt{192.57} = 13.9 \text{ psi}$$

2-1 Data Summary and Display

Population Variance总体方差

When the population is finite and consists of N values, we may define the population variance as

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

The sample variance is a reasonable estimate of the population variance.

2-1 Data Summary and Display

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

VS

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

Note that the divisor for the sample variance is the sample size minus 1, (n-1), whereas for the population variance it is the population size, N.

You can explain it in this way…

- **If we knew the true value of the population mean, μ, we could find the sample variance as the average squared deviation of the sample observations about μ. In practice, the value of μ is almost never known. → Find a substitute→$\bar{x}$**
- **However, the observations x_i tend to be closer to their average, $\bar{x}$, than to the population mean, μ.**
- **Therefore, to compensate for this we use n -1 as the divisor rather than n.**
- **If we used n as the divisor in the sample variance, we would obtain a measure of variability that is, on the average, consistently smaller than the true population variance σ^2.**

“自由度（degrees of freedom）”指的是，在一定的约束条件下，样本所能提供的独立的信息的个数。

2-1 Data Summary and Display

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

VS

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

Note that the divisor for the sample variance is the sample size minus 1, (n-1), whereas for the population variance it is the population size, N.

You can explain it in this way…

sample variance s^2 is based on n-1 degrees of freedom(自由度为n-1).
The term degrees of freedom results from the fact that the n deviations(偏差)

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$$

always sum to zero(和恒等于0), and so specifying the values of any n -1 of these quantities automatically determines the remaining one.(一旦知道其中n-1个值时，剩下的1个值就已经确定了)

2-2 Stem-and-Leaf Diagram茎叶图

A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set $x_1, x_2, \ldots, x_n$, where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps:

当观察值的数字至少有2位时，茎叶图就是表现数据集信息的好的方式。

Steps for Constructing a Stem-and-Leaf Diagram

1. Divide each number x_i into two parts: a **stem**, consisting of one or more of the leading digits, and a **leaf**, consisting of the remaining digit.
2. List the stem values in a vertical column.
3. Record the leaf for each observation beside its stem.
4. Write the units for stems and leaves on the display.

1. 把每一个数字分成两部分：茎，包含一位或者一位以上的主要数字；叶，包含余下位的数
2. 在垂直方向上列出茎的值
3. 在茎的旁边记录下每个观察的叶
4. 在图上写出茎和叶的单位

2-2 Stem-and-Leaf Diagram

EXAMPLE 2-4 Compressive Strength

- **To illustrate the construction of a stem-and-leaf diagram, consider the alloy compressive strength data(合金压力强度数据) in Table 2-2.**
- **We will select as stem values the numbers 7, 8, 9, . . . , 24.**

80个铝锂合金样品的压力强度

Table 2-2 Compressive Strength of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

也即把每个数字写成 $10k_1 + k_0$的形式，其中k_1是茎， k_0是叶

2-2 Stem-and-Leaf Diagram

EXAMPLE 2-4

Figure 2-4 Stem-and-leaf diagram for the compressive strength data in Table 2-2.

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

- **The resulting stem-and-leaf diagram is presented in Fig. 2-4.**
- **The last column in the diagram is a frequency count of the number of leaves associated with each stem（每个茎对应的叶的频数）.**

2-2 Stem-and-Leaf Diagram

Figure 2-5 Stem-and-leaf displays for Example 2-5.

Stem	Leaf
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

25个数据，茎太少，不能提供足够多信息，因此扩充…

L: 0, 1, 2, 3, 4
U: 5, 6, 7, 8, 9

z: 0, 1
t: 2, 3
f: 4, 5
s: 6, 7
e: 8, 9

茎太多，不好！

2-2 Stem-and-Leaf Diagram

Character Stem-and-Leaf Display

Stem-and-Leaf of Strength N = 80
Leaf Unit = 1.0

1	7	6
2	8	7
3	9	7
5	10	1 5
8	11	0 5 8
11	12	0 1 3
17	13	1 3 3 4 5 5
25	14	1 2 3 5 6 8 9 9
37	15	0 0 1 3 4 4 6 7 8 8 8 8
(10)	16	0 0 0 3 3 5 7 7 8 9
33	17	0 1 1 2 4 4 5 6 6 8
23	18	0 0 1 1 3 4 6
16	19	0 3 4 6 9 9
10	20	0 1 7 8
6	21	8
5	22	1 8 9
2	23	7
1	24	5

有序茎叶图

Figure 2-6 A stem-and-leaf diagram from Minitab.

2-2 Stem-and-Leaf Diagram

Table 2-3 Summary Statistics for the Compressive Strength Data from Minitab

Variable	N	Mean	Median	StDev	SE Mean
	80	162.66	161.50	33.77	3.78
	Min	Max	Q1	Q3	
	76.00	245.00	143.50	181.00	

- The **median(中位数)** is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even(偶数), the median is halfway between the two central values.
- The **range(极差)** is a measure of variability that can be easily computed from the ordered stem-and-leaf display. It is the maximum minus the minimum measurement. From Fig. 2-6 the range is 245 - 76 = 169.
- When an ordered set of data is divided into four equal parts, the division points are called **quartiles(四分位数)**. The first or lower quartile, q1, is a value that has approximately 25% of the observations below it and approximately 75% of the observations above it.

2-3 Histograms统计直方图

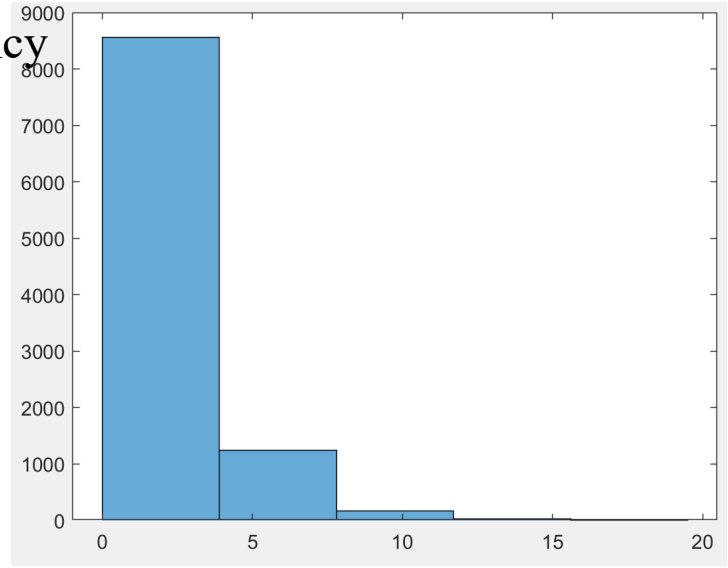
A **histogram** is a more compact(简洁的) summary of data than a stem-and-leaf diagram. To construct a histogram for continuous data, we must divide the range of the data into intervals, which are usually called **class intervals(分类间隔)**, **cells(组距)**, or **bins(箱距)**. If possible, the bins should be **of equal width** to enhance the **visual information**(直观信息) in the histogram.

- **5-20 bins(5-20组)**
- **The number of bins should increase with the number of observations *n***
- **Choosing the number of bins approximately equal to the square root of the number of observations *n* often works well in practice.***

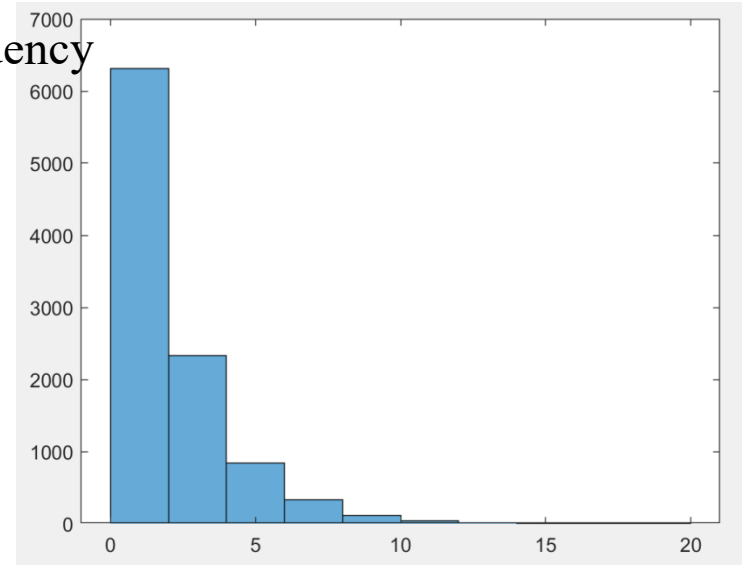
2-3 Histograms

Exponential distribution with $\lambda = 0.5$, *a sample with* 10000 *speciments*

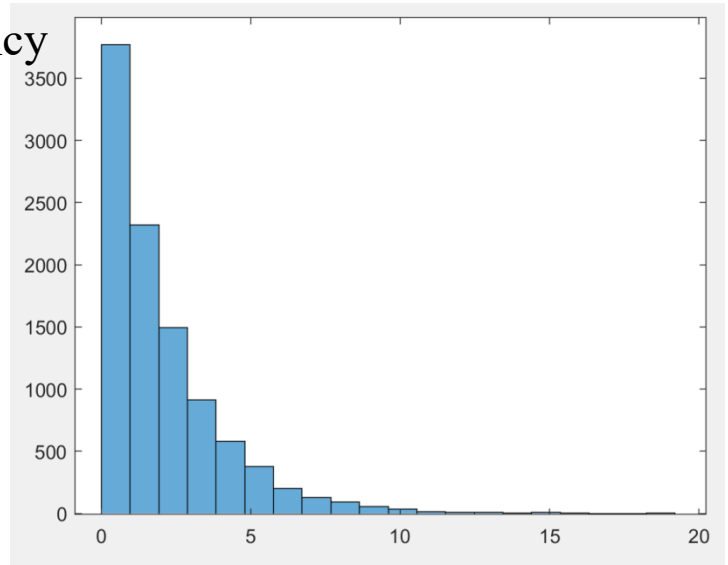
Frequency



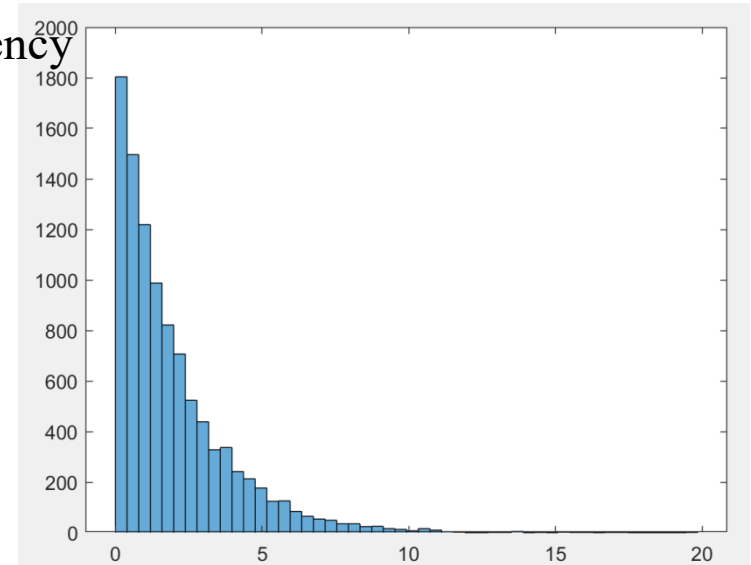
Frequency



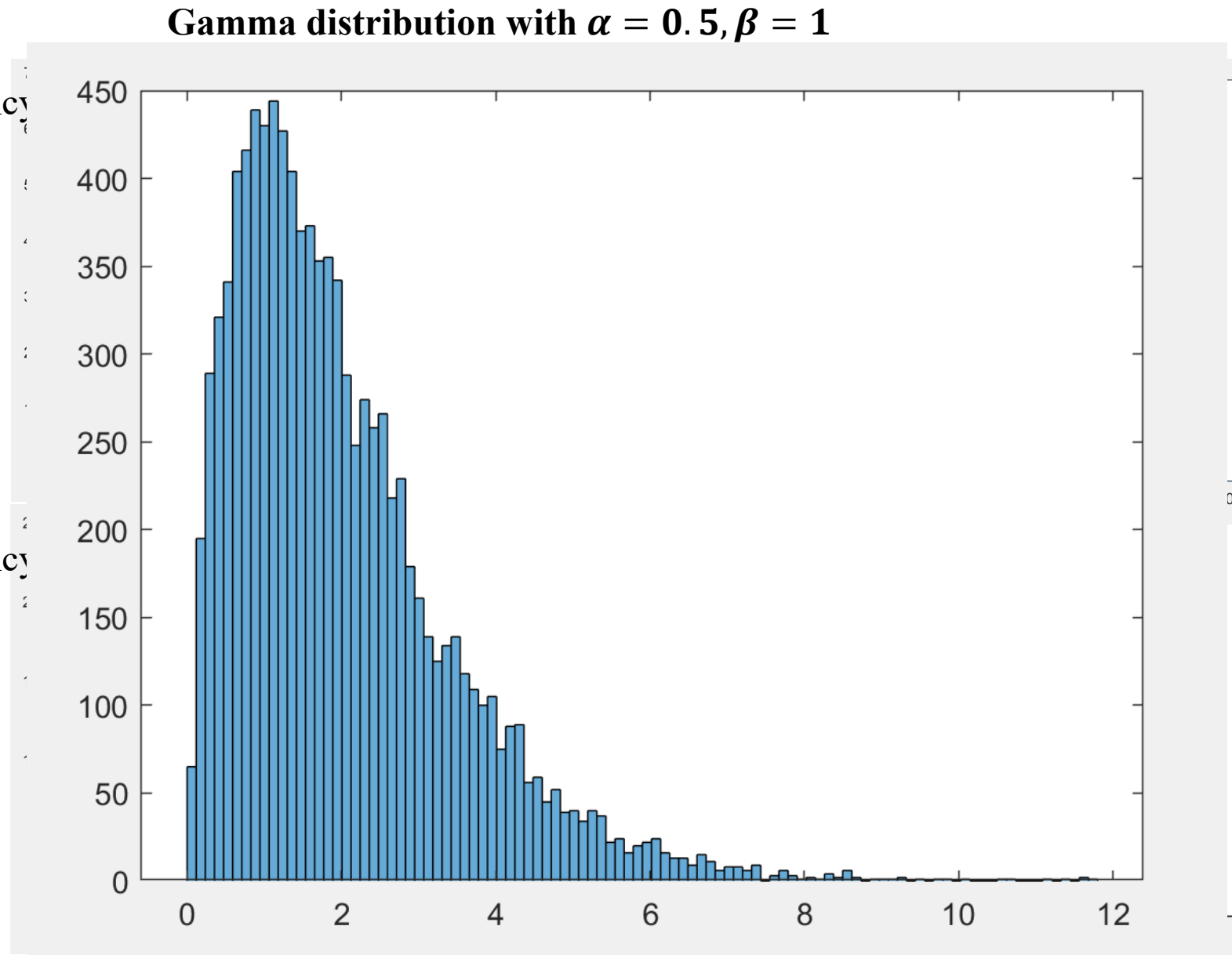
Frequency



Frequency



2-3 Histograms



2-3 Histograms

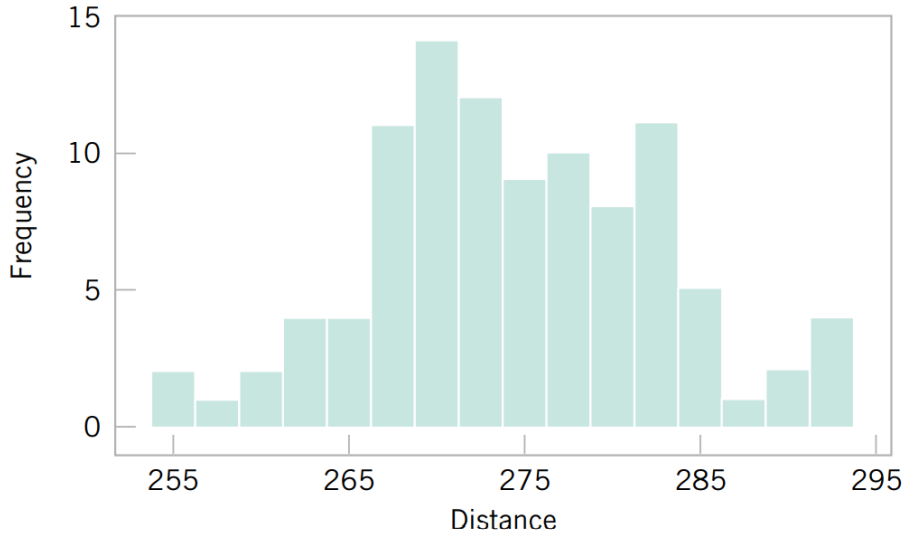


Figure 2-8 Minitab histogram with 16 bins for the golf ball distance data.

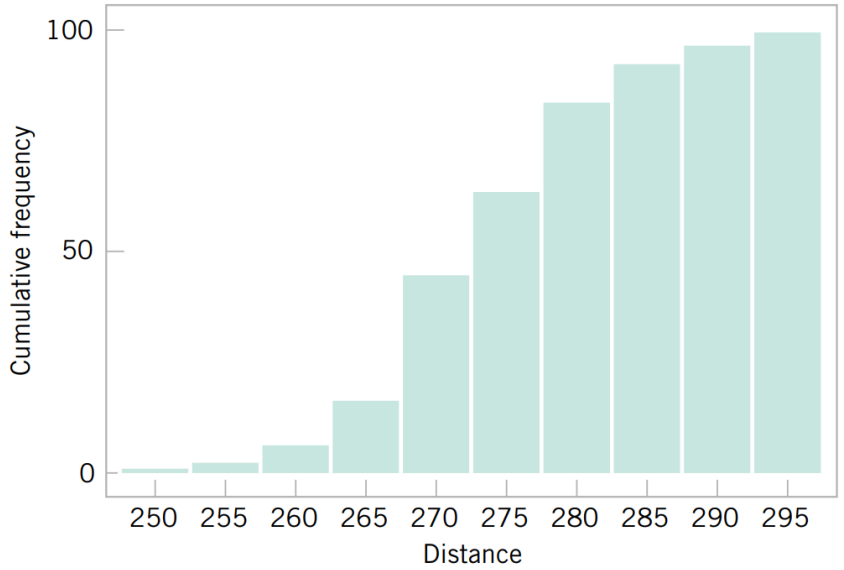


Figure 2-9 A cumulative frequency plot of the golf ball distance data from minitab.

累积频率图

$$F(x) = Pro\ \{X<x\}$$

2-3 Histograms

- An important variation(变体) of the histogram is the **Pareto chart(帕累托图)**.
- This chart is widely used in quality and process improvement studies where the data usually represent different types of defects, failure modes, or other categories of interest to the analyst.
- The categories are ordered so that the category with the largest number of frequencies is on the left, followed by the category with the second largest number of frequencies, and so forth.（相关的分类按照频数顺序从左向右依次排列）

2-3 Histograms

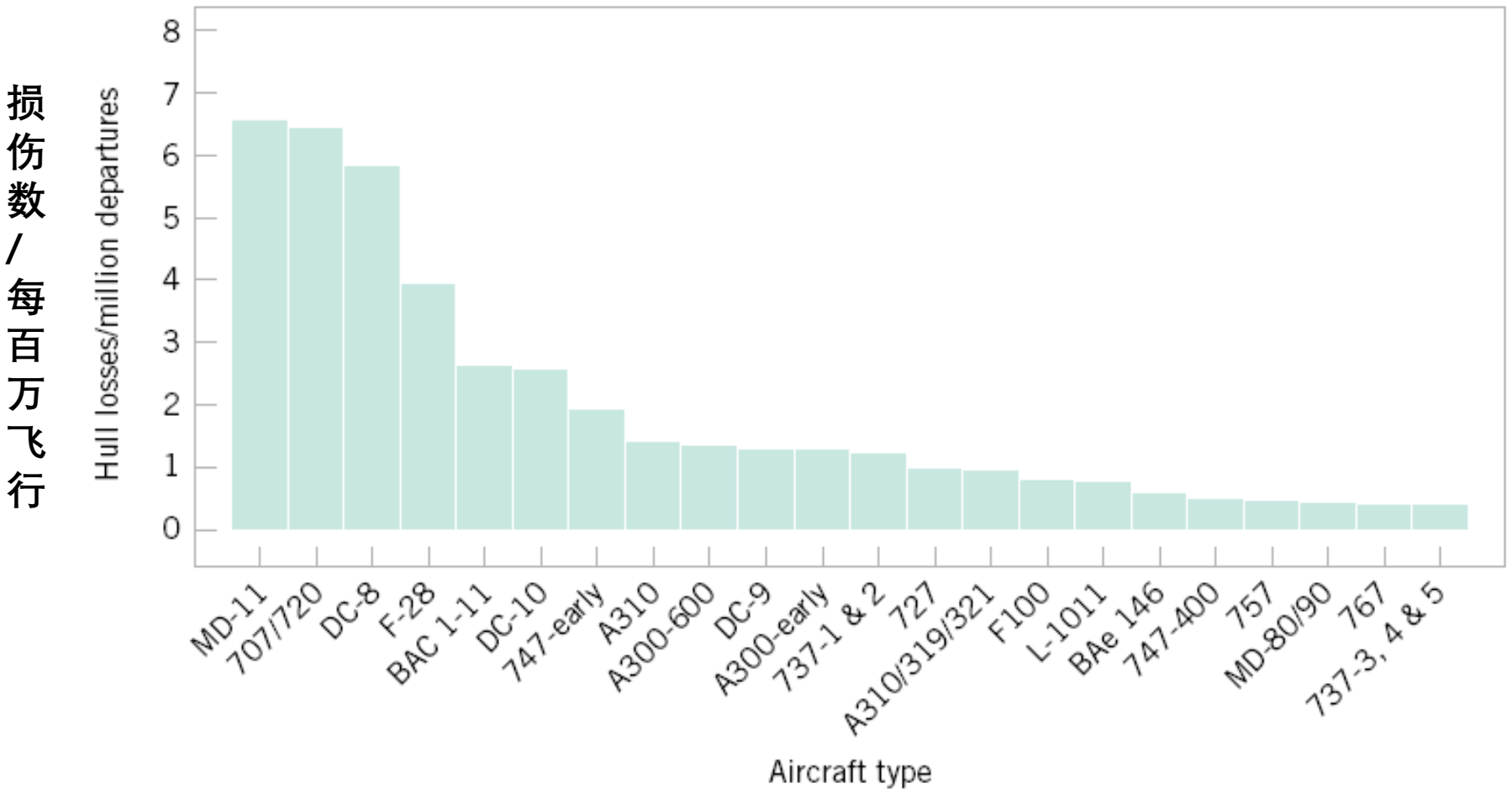


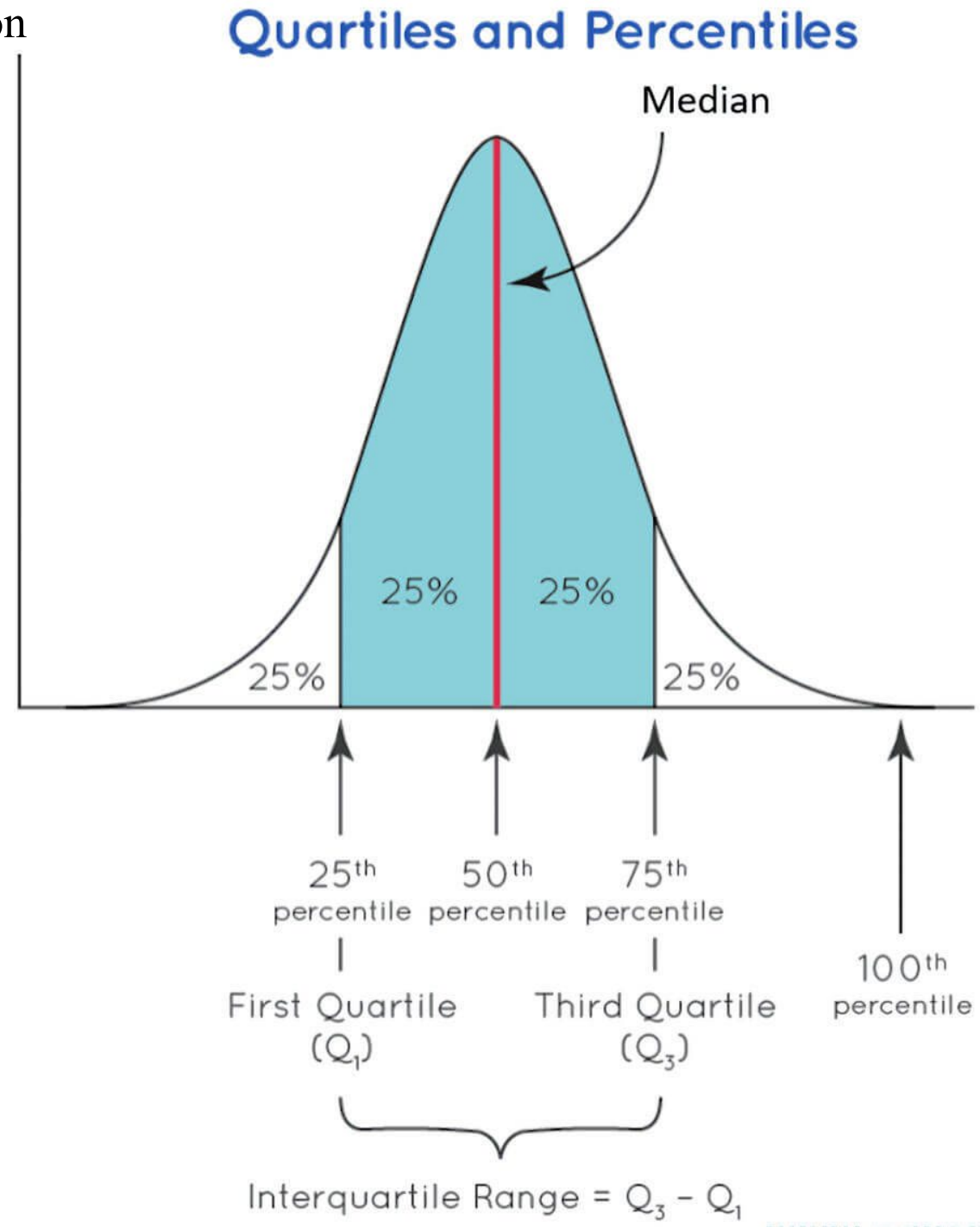
Figure 2-11 Pareto chart for the aircraft accident data.

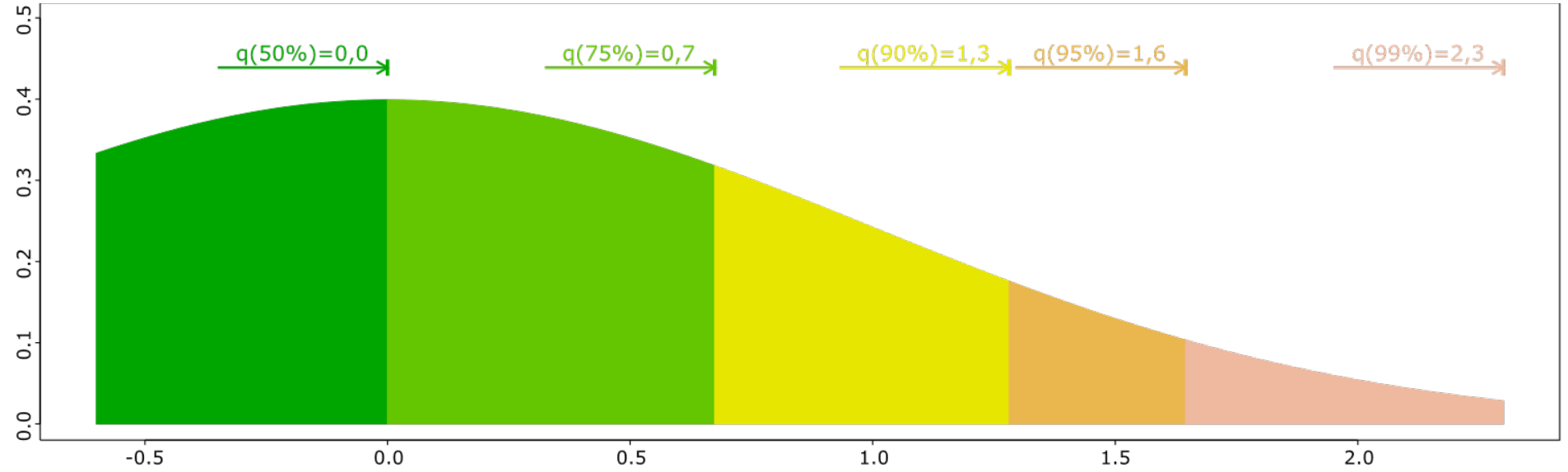
Question: Anything is missing in this case compared to the case of lottery fraud? Please discuss with your classmate for 2 minutes. Then I will let someone to talk about her/his opinion.

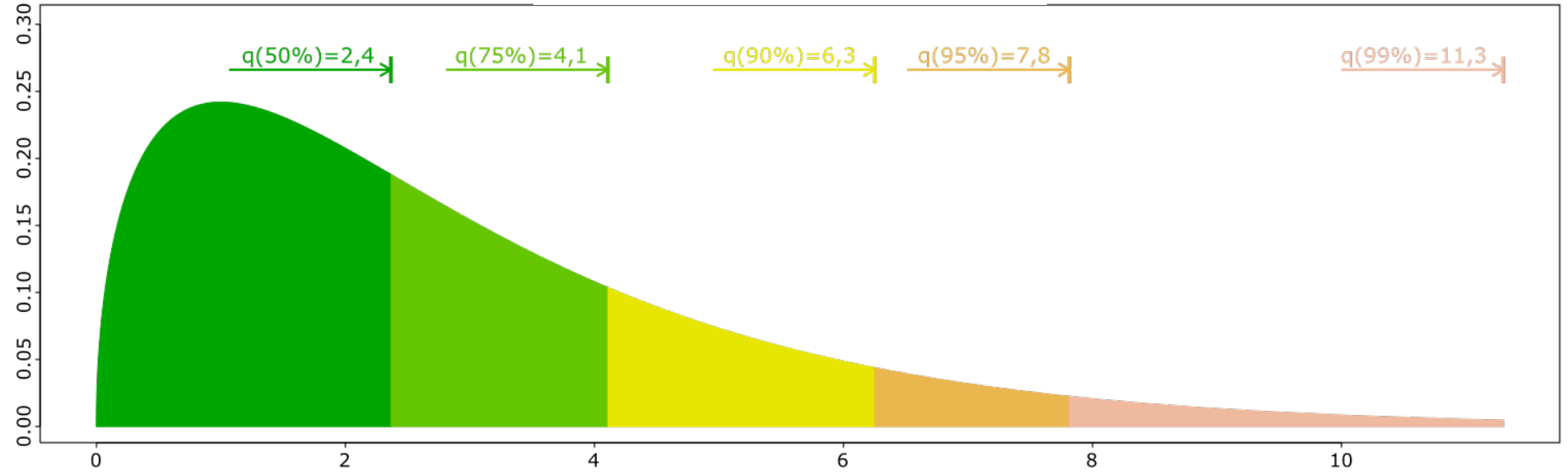
2-4 Box Plots


均线 MA5 111.65 MA10 112.72 MA20 113.84
前复权
340.50
340.50
216.75
93.00
93.00
2024-07
2024-08
2024-09
2024-10
2024-11
成交量16.11万手
2016-08
2018-04
2019-12
2021-07
2023-03
2024-11

Normal distribution




q(50%)=0,0
q(75%)=0,7
q(90%)=1,3
q(95%)=1,6
q(99%)=2,3
0.0
0.1
0.2
0.3
0.4
0.5
-0.5
0.0
0.5
1.0
1.5
2.0


q(50%)=2,4
q(75%)=4,1
q(90%)=6,3
q(95%)=7,8
q(99%)=11,3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10

2-4 Box Plots

Interquartile range(IQR，四分位极差) =75%分位数（Q3）- 25%分位数（Q1）

Second quartile = 50% quartile = 中位数 = median

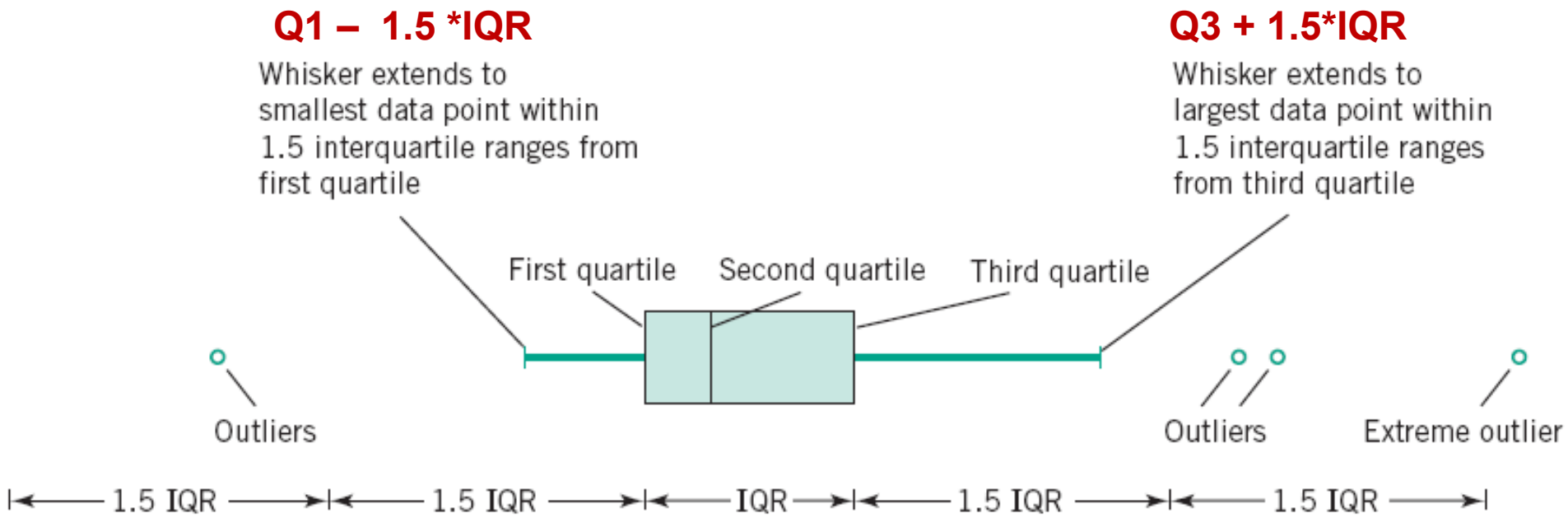
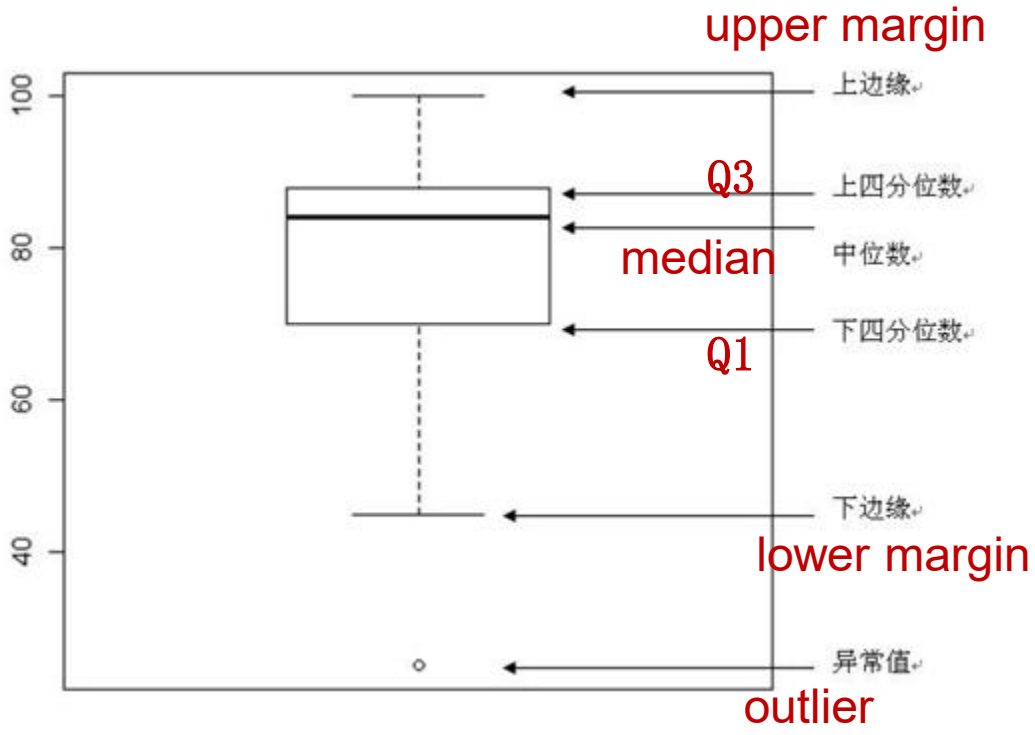


Figure 2-12
Description of a box plot.

2-4 Box Plots



箱盒图共由五个数值点构成，分别是最小观察值（下边缘），25%分位数（Q1），中位数，75%分位数（Q3），最大观察值（上边缘）。

- 中横线：中位数
- IQR：75%分位数（Q3）-25%分位数（Q1）
- 最小观察值（下边缘） = Q1 - 1.5 IQR
- 最大观察值 （上边缘）= Q3 + 1.5 IQR

The box plot consists of five numerical points: **minimum observation (lower edge), 25% quartile (Q1), median, 75% quartile (Q3), and maximum observation (upper edge).**

- Middle horizontal line: median
- IQR: 75% quartile (Q3) - 25% quartile (Q1)
- Minimum observed value (lower margin) = Q1 - 1.5*IQR
- Maximum observed value (upper margin) = Q3 + 1.5*IQR

2-4 Box Plots

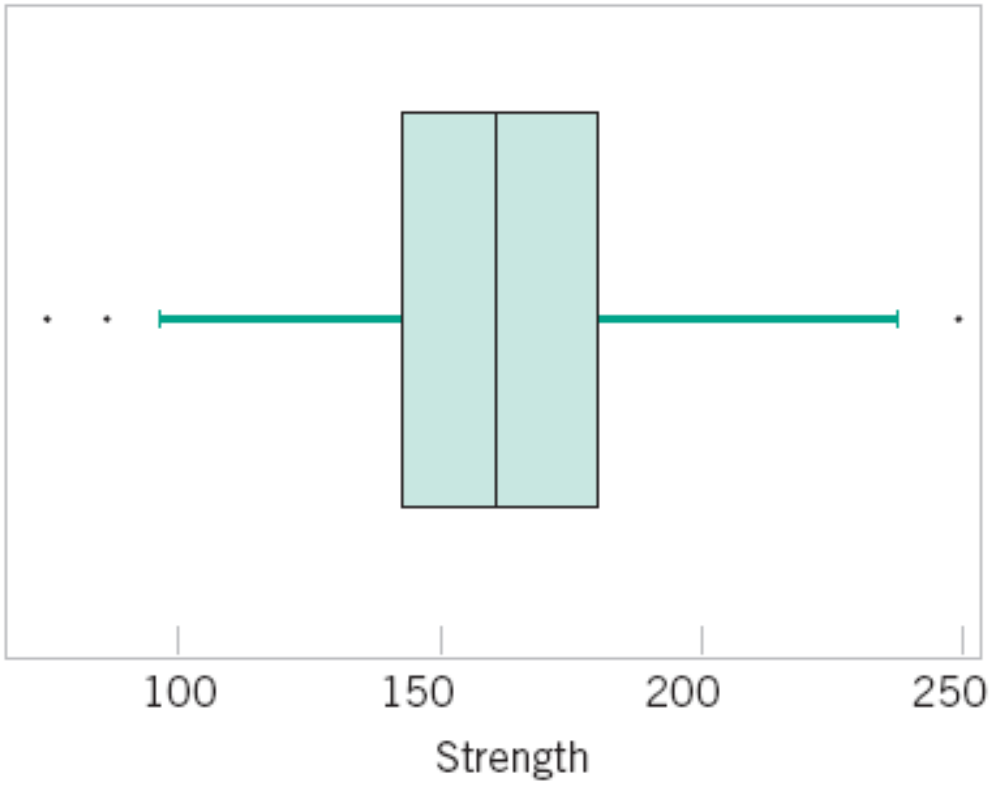


Figure 2-13 Box plot for compressive strength data in Table 2-2.

2-4 Box Plots

Original data

19	10	19	12	13	9	8	48	21

Step 1: Sort排序

1	2	3	4	5	6	7	8	9
8	9	10	12	13	19	19	21	48

Step 2: median中位数/ Second quartile第二个四分位数

1	2	3	4	5	6	7	8	9
8	9	10	12	13	19	19	21	48

Step 3：第一四分位数 Q1=(9+10)/2=9.5； 第三四分位数 Q3=(19+21)/2=20

1	2	3	4	5	6	7	8	9
8	9	10	12	13	19	19	21	48

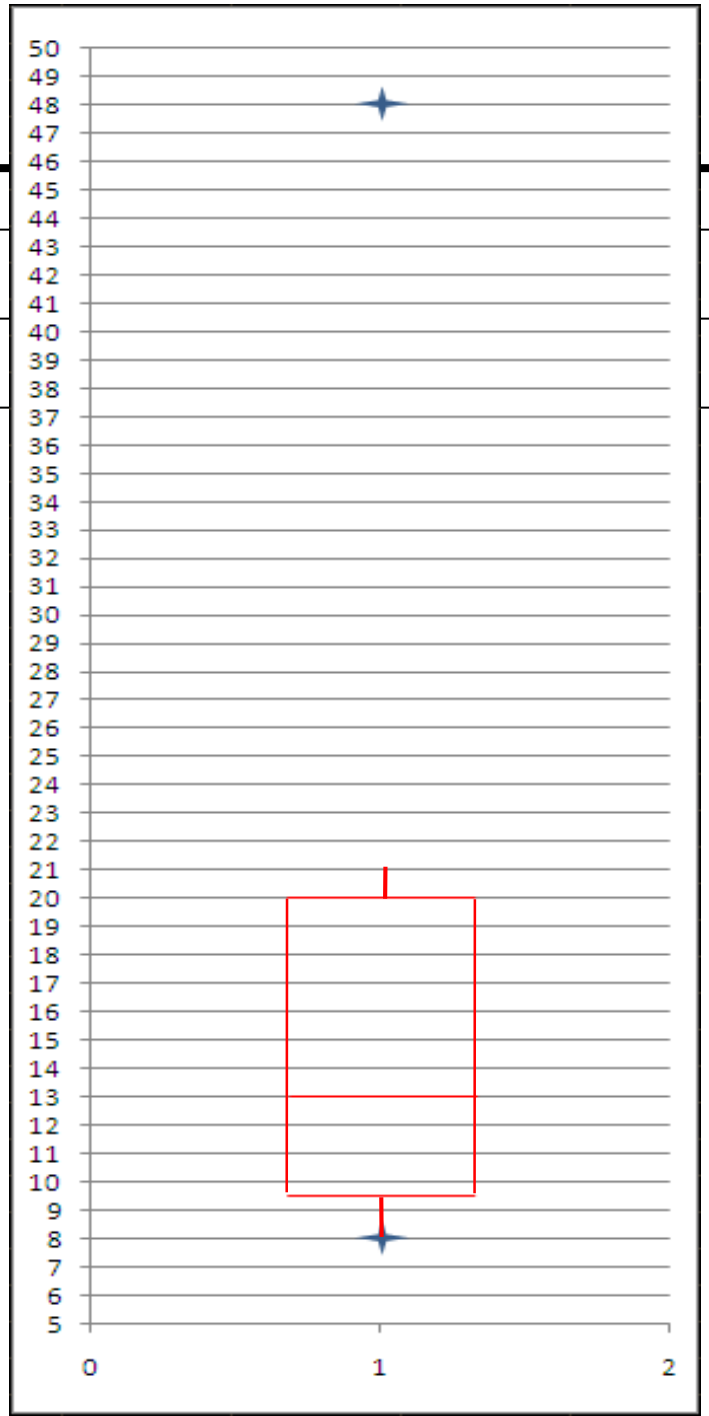
Step 4: **Interquartile range(IQR，四分位极差)**

- IQR= Q3-Q1=10.5
- 1.5*(Q3-Q1)=1.5*(20-9.5)=15.75
- Upper whisker (20+15.75)= 35.75
- Lower whisker (9.5-15.75)= -6.25

2-4 Box Plots

Min	8
Max	48
Q1	9.5
Q2	13
Q3	20

1	2						8	9
8	9						21	48



Upper whisker ∈ [20, 35.75]

Lower whisker ∈ [-6.25, 9.5]

2-4 Box Plots

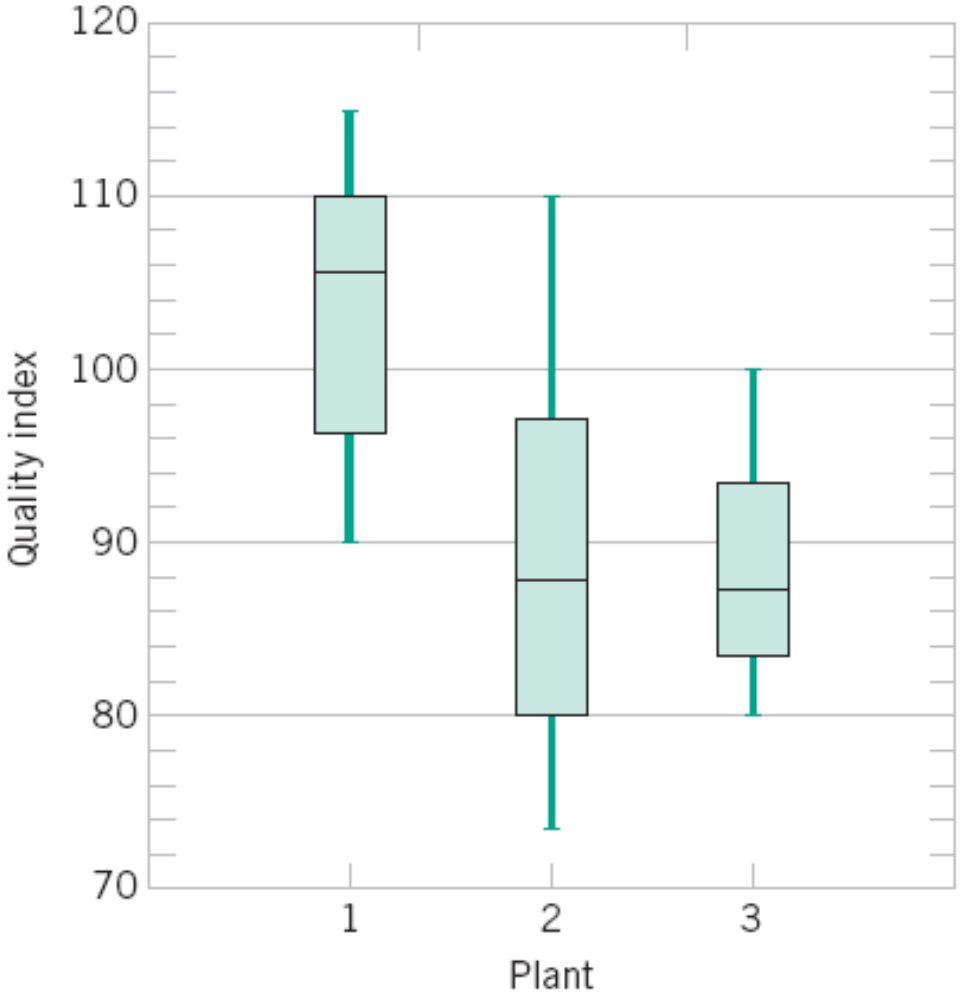


Figure 2-14 Comparative box plots of a quality index at three plants.

补充

- 众数(Mode)是在一组数据中, 出现次数最多的数据，是一组数据中的原数据，而不是相应的次数。
- 众数(Mode)是样本观测值在频数分布表中频数最多的那一组的组中值，主要应用于大面积普查研究之中。
- 一组数据中的众数不止一个。

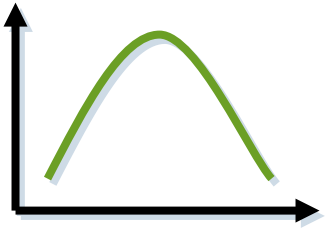
补充

➢无众数

原始数据: 10 5 9 12 6 8

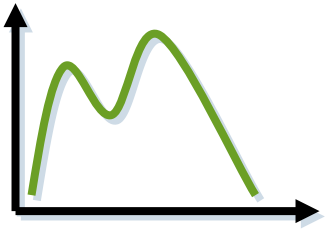
➢ 一个众数

原始数据: 6 5 9 8 5 5

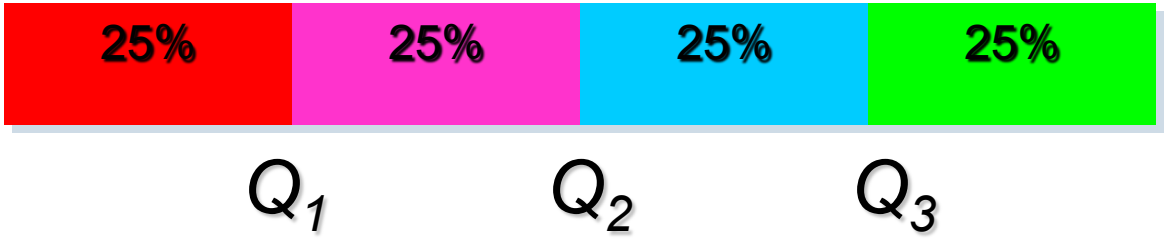


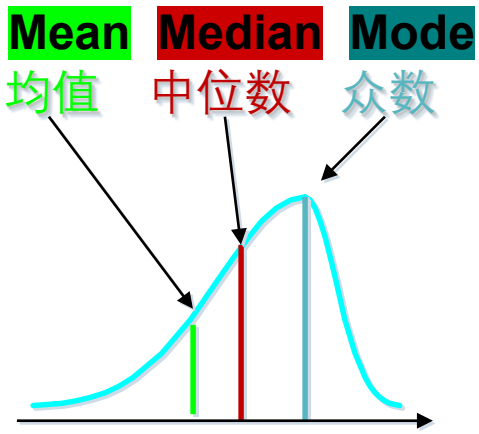
➢多于一个众数

原始数据: 25 28 28 36 42 42



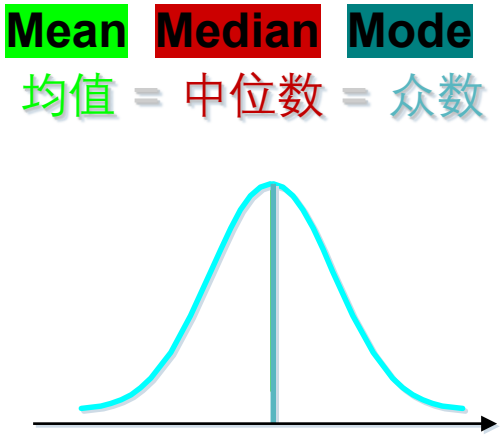
补充





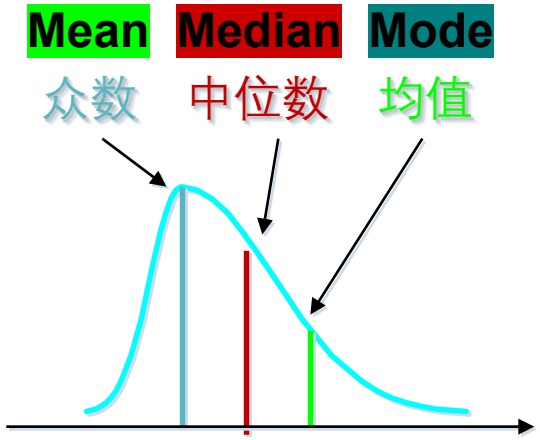
左偏分布

Left skewed distribution



对称分布

Symmetric distribution



右偏分布

Right skewed distribution

众数、中位数、平均数的特点和应用

1. **众数**
   - 不受极端值影响
   - 具有不惟一性
   - 数据分布偏斜程度较大时应用

2. **中位数**
   - 不受极端值影响
   - 数据分布偏斜程度较大时应用

3. **平均数**
   - 易受极端值影响
   - 数学性质优良
   - 数据对称分布或接近对称分布时应用

补充

几何平均数(geometric mean)

1. *n* 个变量值乘积的 *n* 次方根
2. 适用于对比率数据的平均
3. 主要用于计算平均增长率
4. 计算公式为
5. **相乘的各个比率或速度不为零或负值**

$$G_m = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

$$\lg G_m = \frac{1}{n}(\lg x_1 + \lg x_2 + \cdots + \lg x_n) = \frac{\sum_{i=1}^{n} \lg x_i}{n}$$

补充

几何平均数(geometric mean)(例题分析)

【例】一位投资者购持有一种股票，在2000、2001、2002和2003年收益率分别为4.5%、2.1%、25.5%、1.9%。计算该投资者在这四年内的平均收益率。

几何平均：

$$\overline{G} = \sqrt[4]{104.5\% \times 102.1\% \times 125.5\% \times 101.9\%} - 1$$
$$= 8.0787\%$$

算术平均：

$$\overline{G} = (4.5\% + 2.1\% + 25.5\% + 1.9\%) \div 4 = 8.5\%$$

补充

【例】某流水生产线有前后衔接的五道工序。某日各工序产品的合格率分别为95％、92％、90％、85％、80％，求整个流水生产线产品的平均合格率。

分析:

设最初投产100A个单位，则
第一道工序的合格品为100A×0.95；
第二道工序的合格品为（100A×0.95）×0.92；
……
第五道工序的合格品为
（100A×0.95×0.92×0.90×0.85）×0.80；

补充

因该流水线的最终合格品即为第五道工序的合格品， 故该流水线总的合格品应为

100A×0.95×0.92×0.90×0.85×0.80；

则该流水线产品总的合格率为：

$$\frac{\text{总合格品}}{\text{总产品}}=\frac{100\text{A}\times0.95\times0.92\times0.90\times0.85\times0.80}{100\text{A}}$$

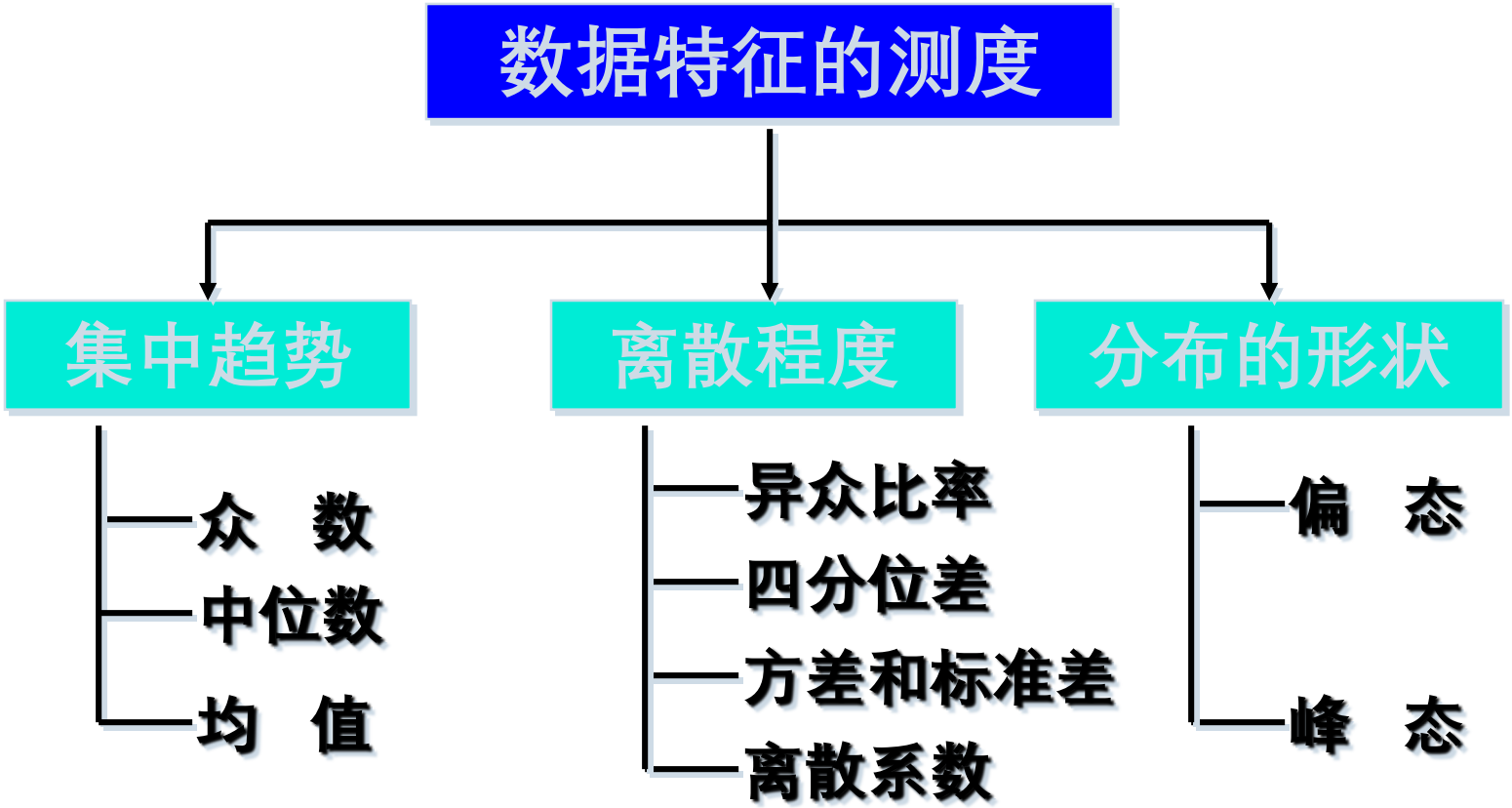
$$=0.95\times0.92\times0.90\times0.85\times0.80$$

即该流水线总的合格率等于各工序合格率的连乘积，符合几何平均数的适用条件，故需采用几何平均法计算。

$$\overline{X}_G=\sqrt[5]{0.95\times0.92\times0.90\times0.85\times0.80}$$

$$=\sqrt[5]{0.5349}=88.24\%$$

补充



2-5 Time Series Plots

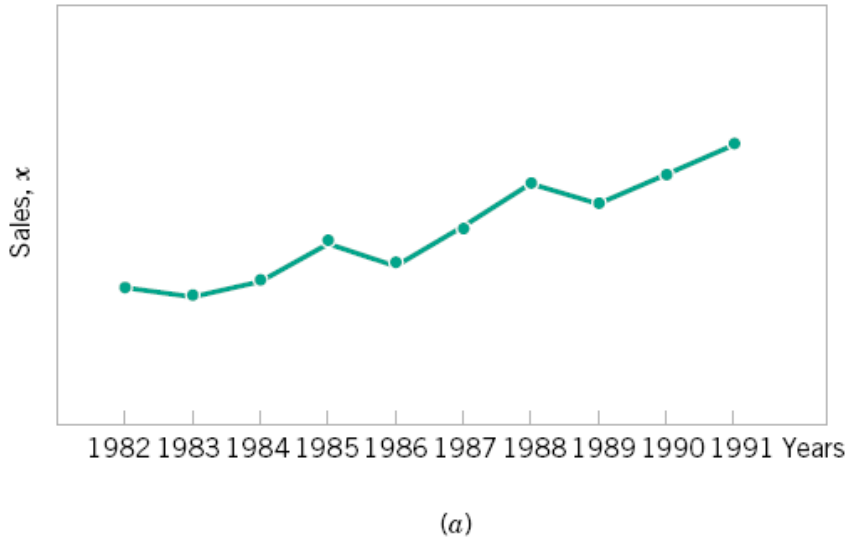
• A **time series** or **time sequence** is a data set in which the observations are recorded in the order in which they occur.

• A **time series plot** is a graph in which the vertical(垂直) axis denotes the observed value of the variable (say x) and the horizontal(水平) axis denotes the time (which could be minutes, days, years, etc.).

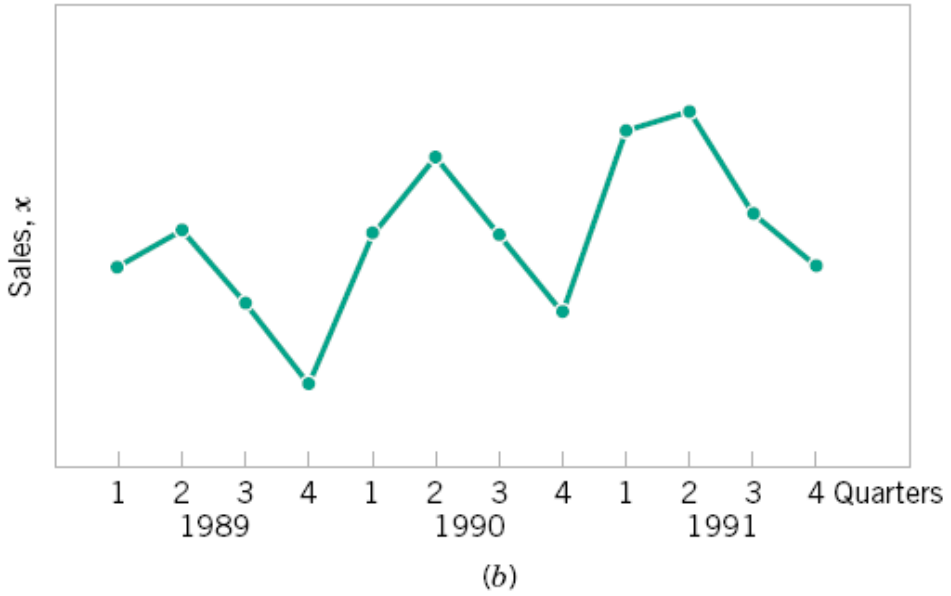
• When measurements are plotted as a time series, we often see

- **trends,**
- **cycles, or**
- **other broad features of the data(其它主要特征)**

2-5 Time Series Plots



(a)



(b)

Figure 2-15 Company sales by year (a) and by quarter (b).

2-5 Time Series Plots

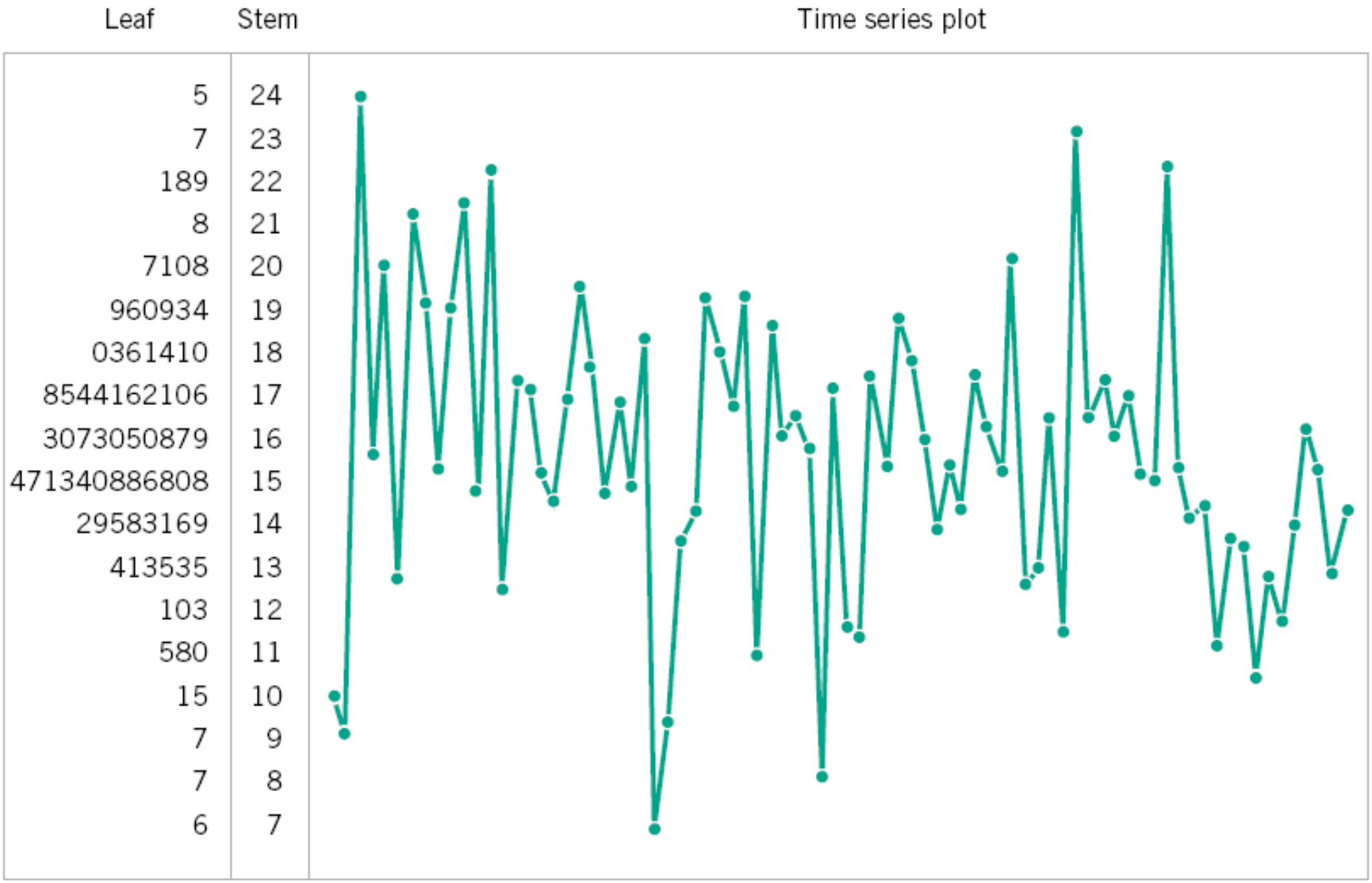


Figure 2-16 A digidot plot of the compressive strength data in Table 2-2.

2-5 Time Series Plots

Leaf	Stem
8	9e
6	9s
45	9f
2333	9t
0010000	9z
99998	8e
66677	8s
45	8f
23	8t
1	8z

Time series plot

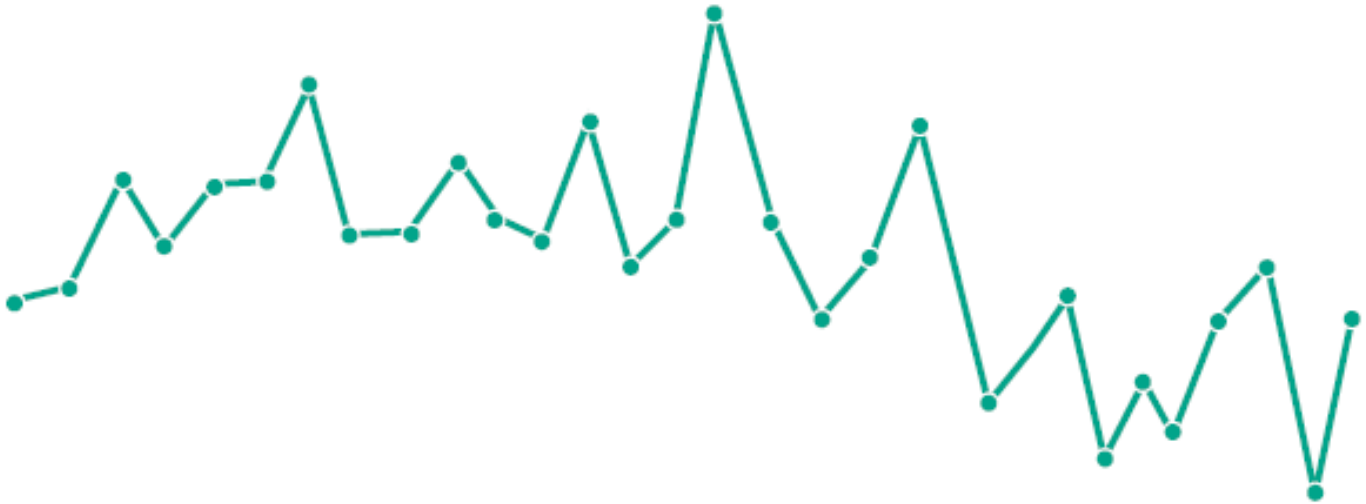


Figure 2-17 A digidot plot of chemical process concentration readings, observed hourly.

茎叶图与时间序列图结合：数点图

2-6 Multivariate Data

• The dot diagram, stem-and-leaf diagram, histogram, and box plot are descriptive displays for **univariate(单变量的)** data; that is, they **convey(传达，表现) descriptive information(描述信息)** about a single variable.

•Many engineering problems involve collecting and analyzing **multivariate data(多变量数据)**, or data on several different variables.

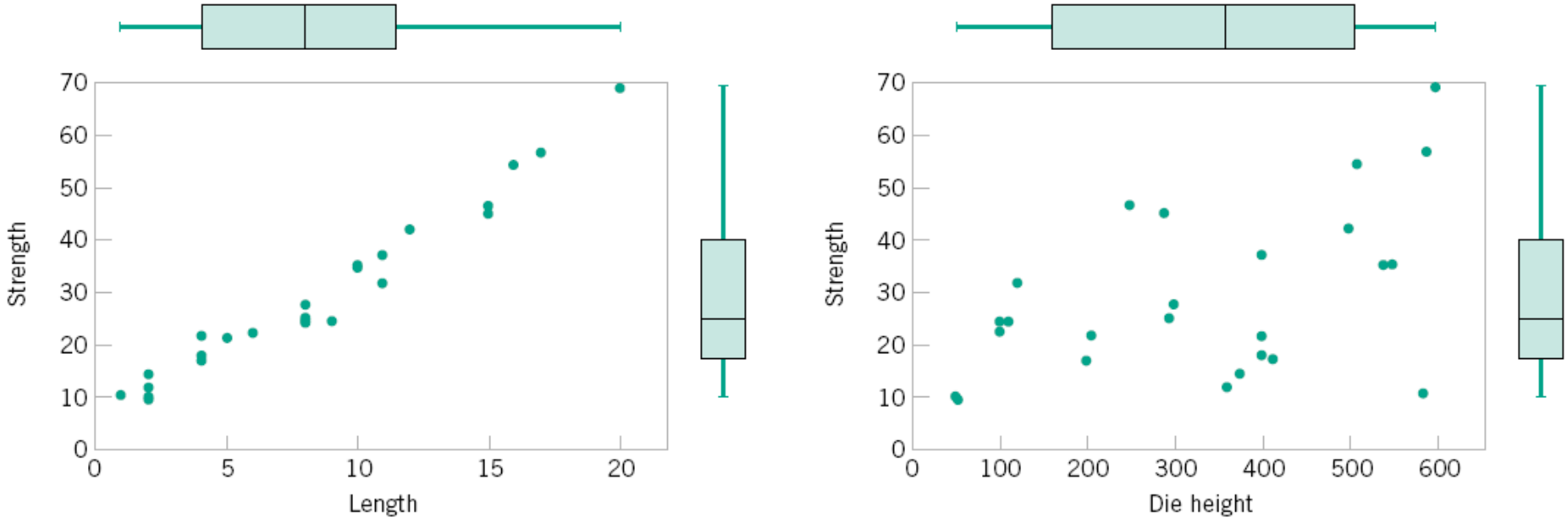
•In engineering studies involving multivariate data, often the objective is to determine the relationships among the variables or to build an empirical model.

2-6 Multivariate Data

Table 2-9 Wire Bond Data

Observation Number	Pull Strength, y	Wire Length, x_1	Die Height, x_2	Observation Number	Pull Strength, y	Wire Length, x_1	Die Height, x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

2-6 Multivariate Data



(*a*) Pull strength versus length

(*b*) Pull strength versus die height

Figure 2-19 Scatter diagrams and box plots for the wire bond pull strength data in Table 1-1. (*a*) Pull strength versus length. (*b*) Pull strength versus die height.

The strength of these relationships appears to be stronger for pull strength and length than it does for pull strength and die height.(拉拔强度和金属丝长度的关系好像要比拉拔强度与模子高度的关系要强一点)

2-6 Multivariate Data

Sample Correlation Coefficient样本相关系数

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

Given n pairs of data $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, the **sample correlation coefficient** r is defined by

$$r = \frac{S_{xy}}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \qquad (2\text{-}6)$$

with $-1 \le r \le +1$.

2-6 Multivariate Data

(a) r is near +1 (b) r is near −1 (c) r is near 0, y and x are unrelated (d) r is near 0, y and x are nonlinearly related

Figure 2-20 Scatter diagrams for different values of the sample correlation coefficient r. (a) r is near +1. (b) r is near −1. (c) r is near 0; y and x are unrelated. (d) r is near 0; y and x are nonlinearly related.

The sample correlation coefficient between wire bond pull strength and wire length is

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 2396 - \frac{(206)^2}{25} = 698.56$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 27179 - \frac{(725.82)^2}{25} = 6106.41$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 8008.5 - \frac{(206)(725.82)}{25} = 2027.74$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{2027.74}{\sqrt{(698.56)(6106.41)}} = 0.982$$

- Generally, we consider the correlation between two variables to be strong when 0.8≤r≤1, weak when 0≤r≤0.5, and moderate otherwise.
- Therefore, there is a strong correlation between the wire bond pull strength and wire length and a relatively weak to moderate (中等) correlation between pull strength and die height.

2-6 Multivariate Data

- There are several other useful graphical methods for displaying multivariate data.
- These data were collected during a sensory evaluation(感 官 评 估) experiment conducted by a scientist.
- The variables foam, scent, color, and residue (a measure of the extent of the cleaning ability) are descriptive properties evaluated on a 10-point scale.
- Quality is a measure of overall desirability of the shampoo, and it is the nominal response variable of interest to the experimenter.

Table 2-11 Data on Shampoo

Foam	Scent	Color	Residue	Region	Quality
6.3	5.3	4.8	3.1	1	91
4.4	4.9	3.5	3.9	1	87
3.9	5.3	4.8	4.7	1	82
5.1	4.2	3.1	3.6	1	83
5.6	5.1	5.5	5.1	1	83
4.6	4.7	5.1	4.1	1	84
4.8	4.8	4.8	3.3	1	90
6.5	4.5	4.3	5.2	1	84
8.7	4.3	3.9	2.9	1	97
8.3	3.9	4.7	3.9	1	93
5.1	4.3	4.5	3.6	1	82
3.3	5.4	4.3	3.6	1	84
5.9	5.7	7.2	4.1	2	87
7.7	6.6	6.7	5.6	2	80
7.1	4.4	5.8	4.1	2	84
5.5	5.6	5.6	4.4	2	84
6.3	5.4	4.8	4.6	2	82
4.3	5.5	5.5	4.1	2	79
4.6	4.1	4.3	3.1	2	81
3.4	5.0	3.4	3.4	2	83
6.4	5.4	6.6	4.8	2	81
5.5	5.3	5.3	3.8	2	84
4.7	4.1	5.0	3.7	2	83
4.1	4.0	4.1	4.0	2	80

2-6 Multivariate Data

	Foam	Scent	Color	Residue	Region
Scent	0.002				
Color	0.328	0.599			
Residue	0.193	0.500	0.524		
Region	−0.032	0.278	0.458	0.165	
Quality	0.512	−0.252	−0.194	−0.489	−0.507

2-6 Multivariate Data

	Foam	Scent	Color	Residue	Region
Scent	0.002				
Color	0.328	0.599			
Residue	0.193	0.500	0.524		
Region	−0.032	0.278	0.458	0.165	
Quality	0.512	−0.252	−0.194	−0.489	−0.507

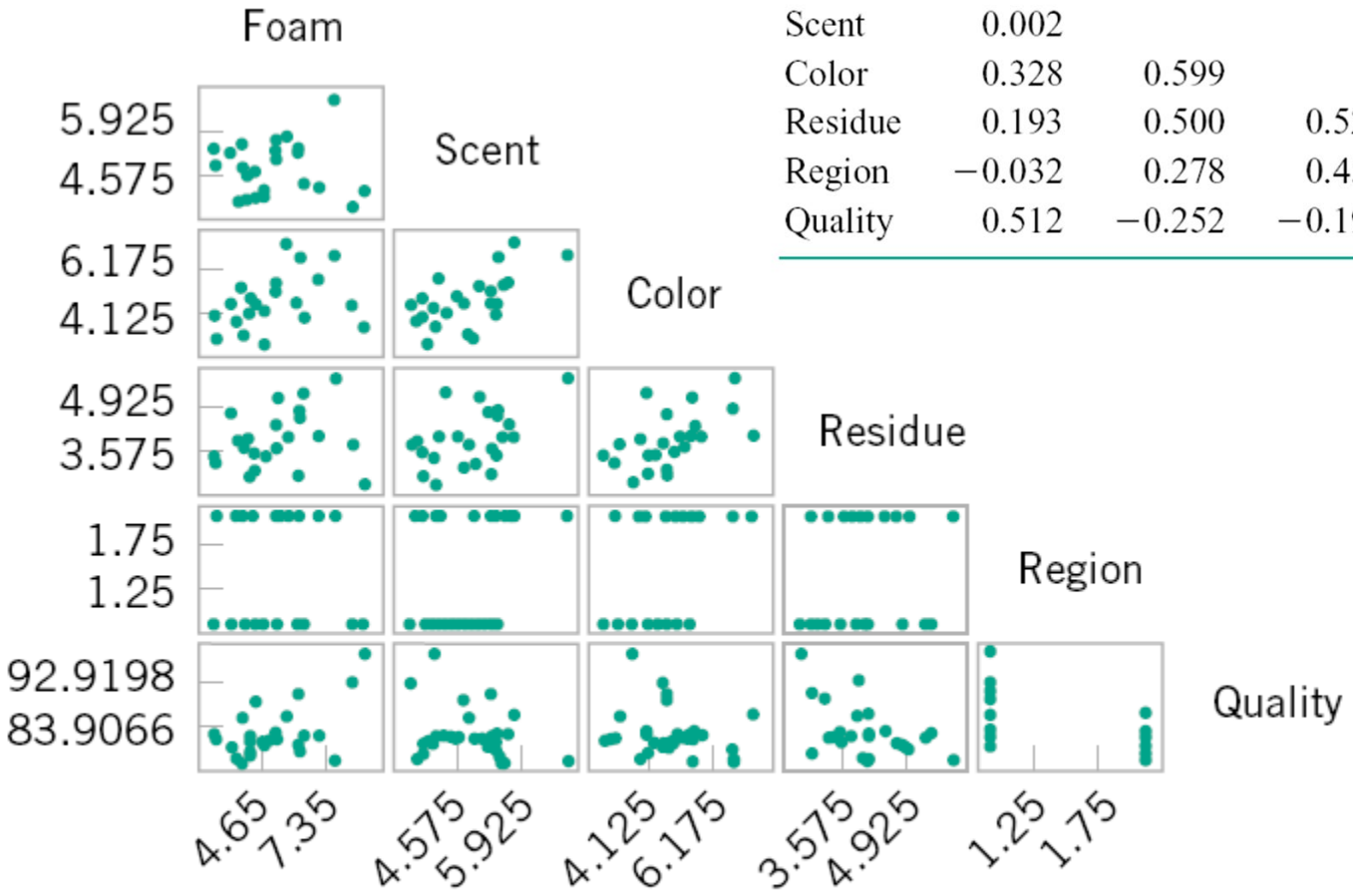


Figure 2-22 Matrix of scatter plots for the shampoo data in Table 2-11.

2-6 Multivariate Data

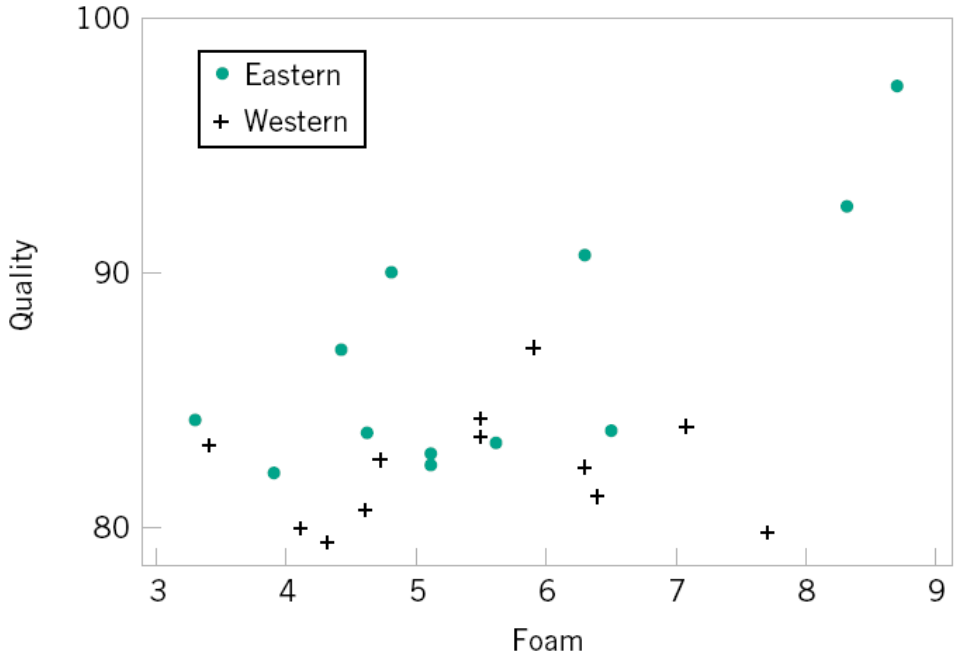


Figure 2-23 Scatter diagram of shampoo quality versus foam.

- Figure 2-23 is a scatter diagram of shampoo quality versus foam. In this scatter diagram we have used two different plotting symbols to identify the observations associated with the two different regions, thus allowing information about three variables to be displayed on a two-dimensional graph.
- The display in Fig. 2-23 reveals that the relationship between shampoo quality and foam may be different in the two regions. Another way to say this is that there may be an interaction between foam and region.

2-6 Multivariate Data

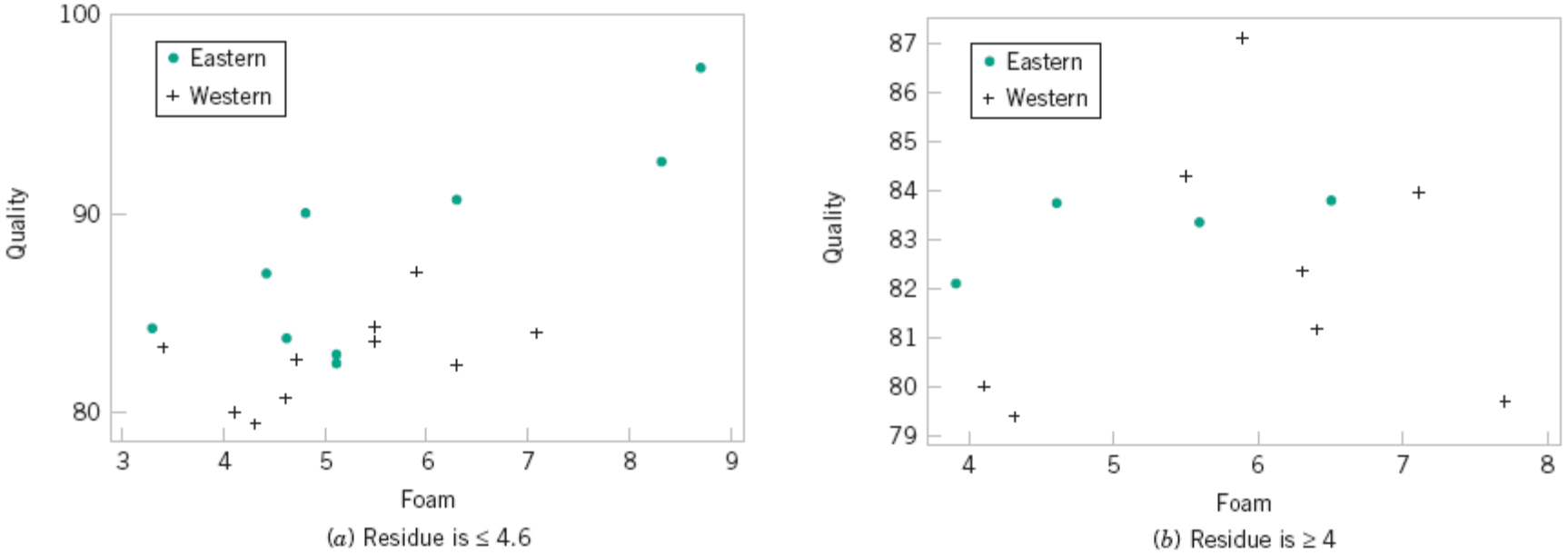
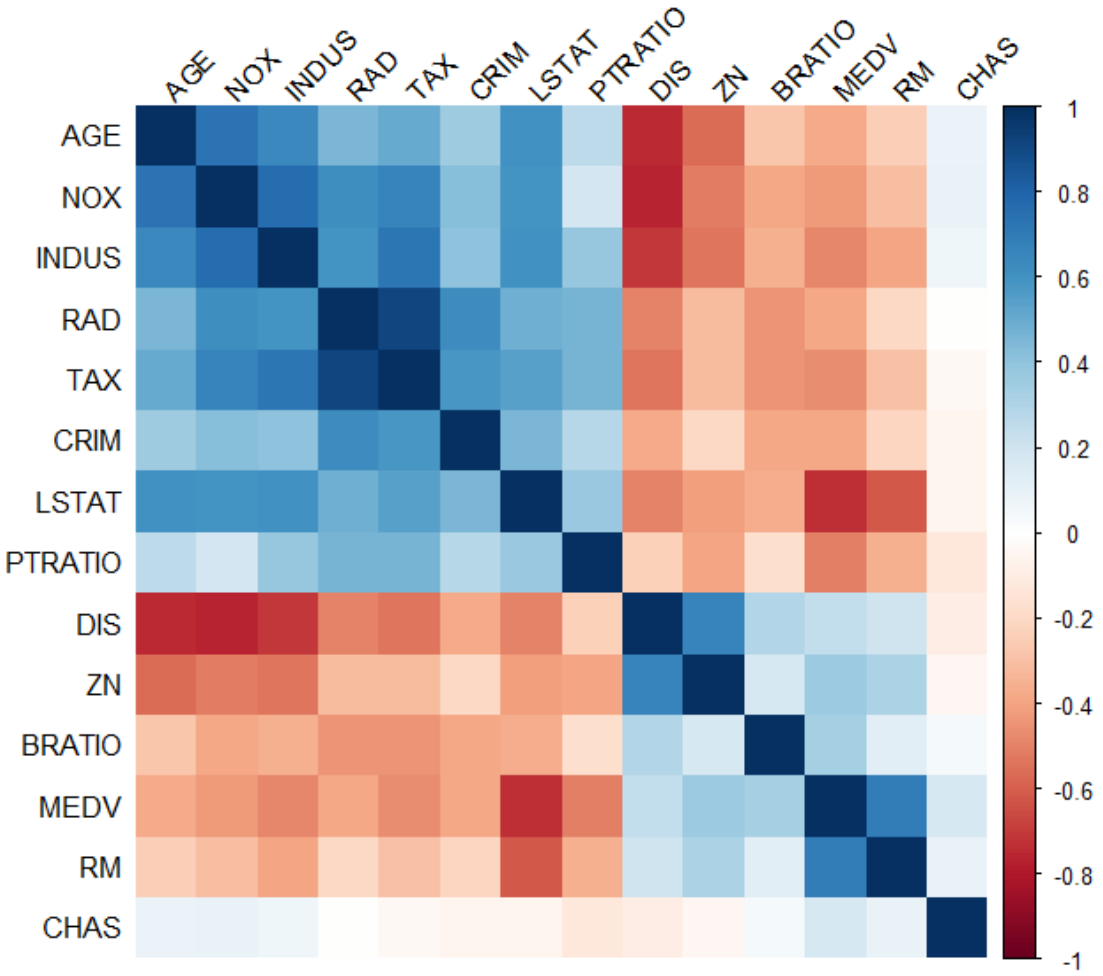


Figure 2-24 A coplot for the shampoo data. (*a*) Residue is ≤ 4.6. (*b*) Residue is ≥ 4.

- The variation of the scatter diagram in Fig. 2-23 works well when the third variable is discrete or categorical.
- When the third variable is continuous, a coplot may be useful. A **coplot(协同图)** for the shampoo quality data is shown in Fig. 2-24.
  - In this display, shampoo quality is plotted against foam, and as in Fig. 2-23, different plotting symbols are used to identify the two production regions.
  - The coplot indicates that the positive relationship between shampoo quality and foam is much stronger for the lower range of residue, perhaps indicating that too much residue doesn't always result in a good shampoo.

2-6 Multivariate Data-supplementary material



Heat map of correlation coefficient matrix using "corrplot" in R

使用R语言中的corrplot来绘制相关系数矩阵热图

使用Python中的Matplotlib和Seaborn库绘制相关性热力图

IMPORTANT TERMS AND CONCEPTS

Box plot
Degrees of freedom
Digidot plot
Dot diagram
Frequency
Histogram
Interquartile range, IQR
Marginal plot
Matrix of scatter plots
Median
Multivariate data
Ordered stem-and-leaf diagram
Pareto chart
Percentile
Population mean, μ
Population standard deviation, σ
Population variance, σ^2
Quartiles
Range
Relative frequency
Sample correlation coefficient
Sample mean, $\bar{x}$
Sample standard deviation, s
Sample variance, s^2
Scatter diagram
Stem-and-leaf diagram
Time sequence
Time series plot
Univariate data