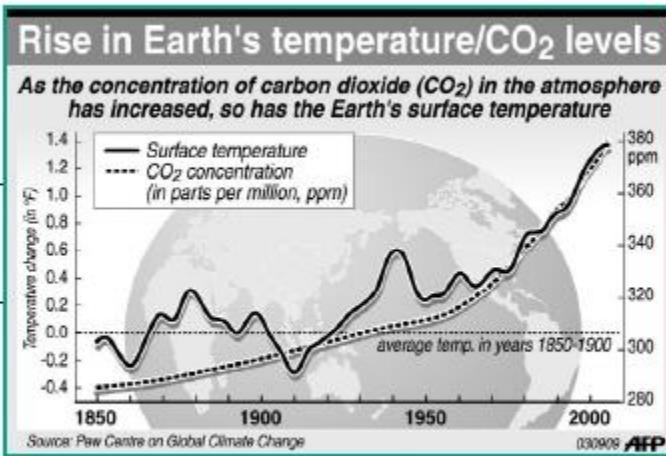


2



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, \dots, x_n , the **sample mean** is

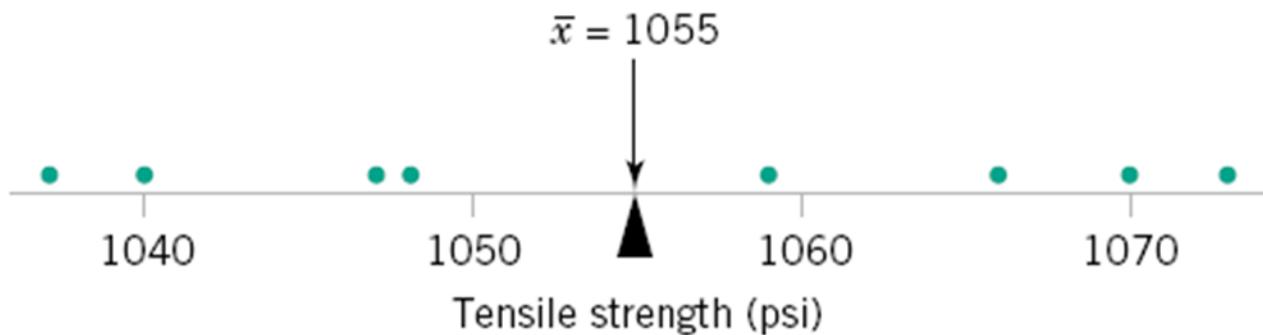
$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{\sum_{i=1}^n x_i}{n}\end{aligned}\tag{2-1}$$

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, \dots, x_n , then the **sample variance** is

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

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

2-1 Data Summary and Display

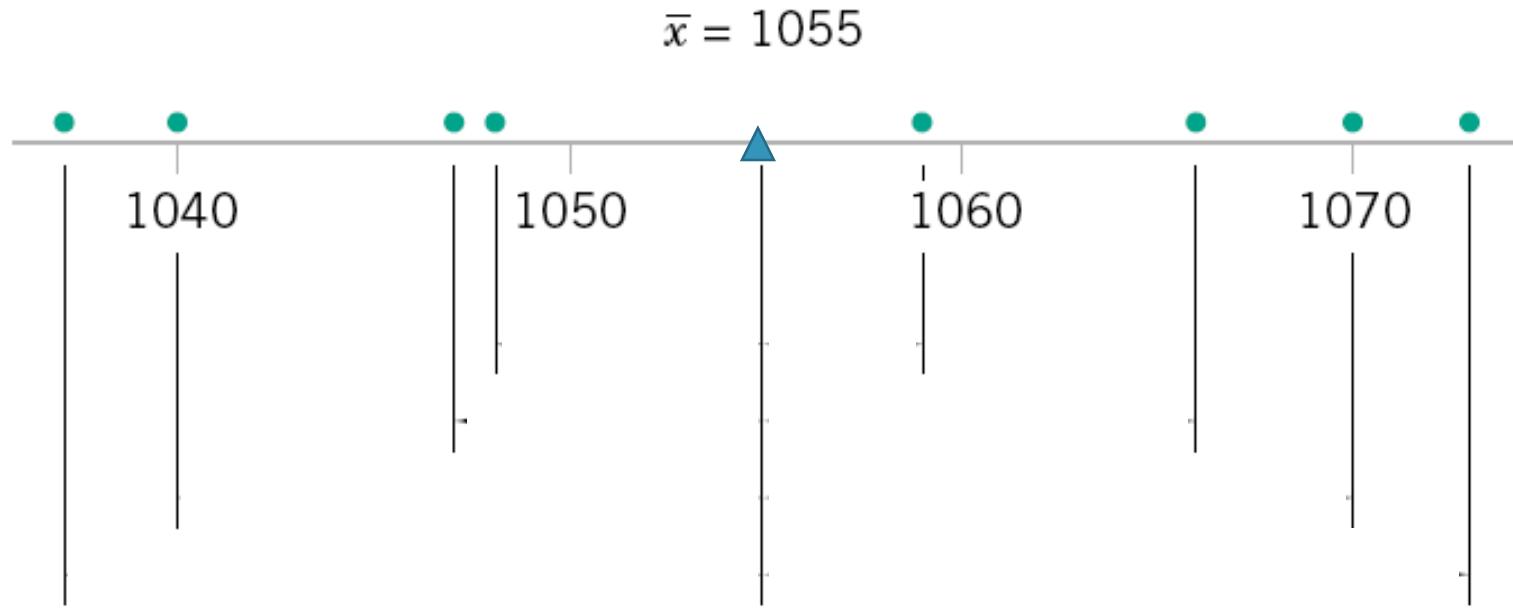


Figure 2-3 How the sample variance measures variability through the deviations $x_i - \bar{x}$.

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	<u>1073</u>	<u>18</u>	<u>324</u>
	8440	0.0	1348

2-1 Data Summary and Display

EXAMPLE 2-3

O-Ring Strength:

Alternative

Variance

Calculation

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

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)

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

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, \dots, 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	Stem	Leaf	Stem	Leaf
6	1 3 4 5 5 6	6L	1 3 4	6z	1
7	0 1 1 3 5 7 8 8 9	6U	5 5 6	6t	3
8	1 3 4 4 7 8 8	7L	0 1 1 3	6f	4 5 5
9	2 3 5	7U	5 7 8 8 9	6s	6
		8L	1 3 4 4	6e	
		8U	7 8 8	7z	0 1 1
		9L	2 3	7t	3
		9U	5	7f	5
				7s	7
				7e	8 8 9
				8z	1
				8t	3
				8f	4 4
				8s	7
				8e	8 8
				9z	
				9t	2 3
				9f	5
				9s	
				9e	

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
(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

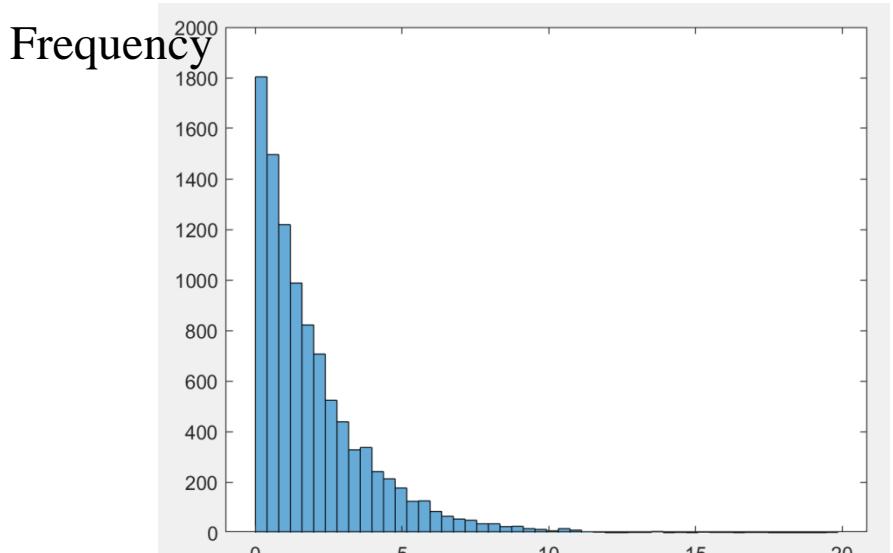
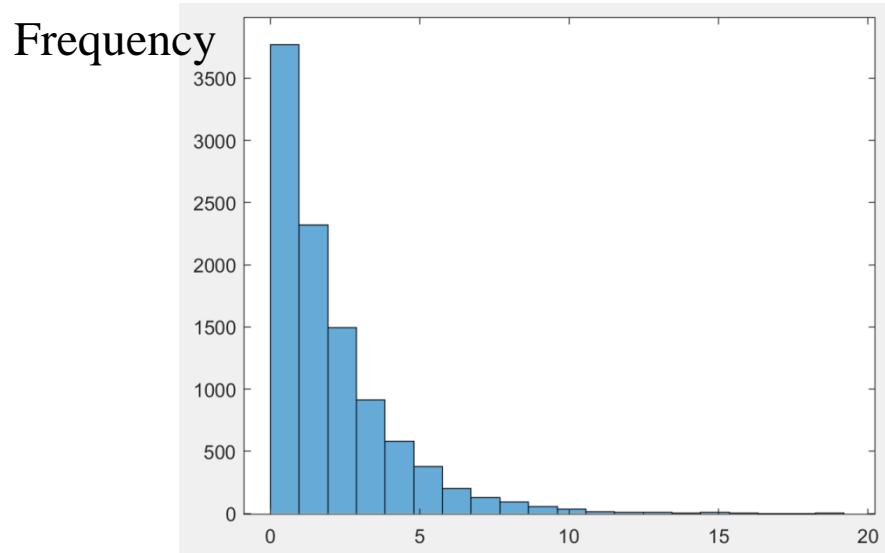
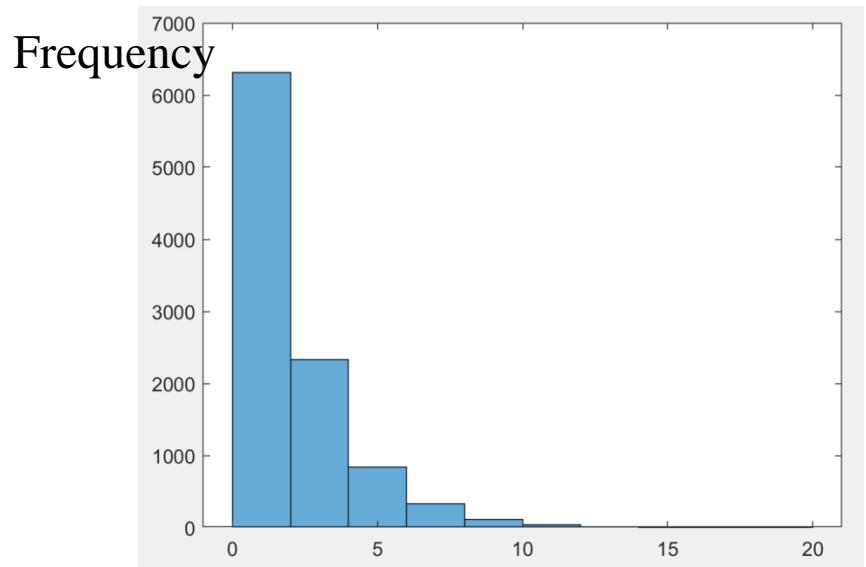
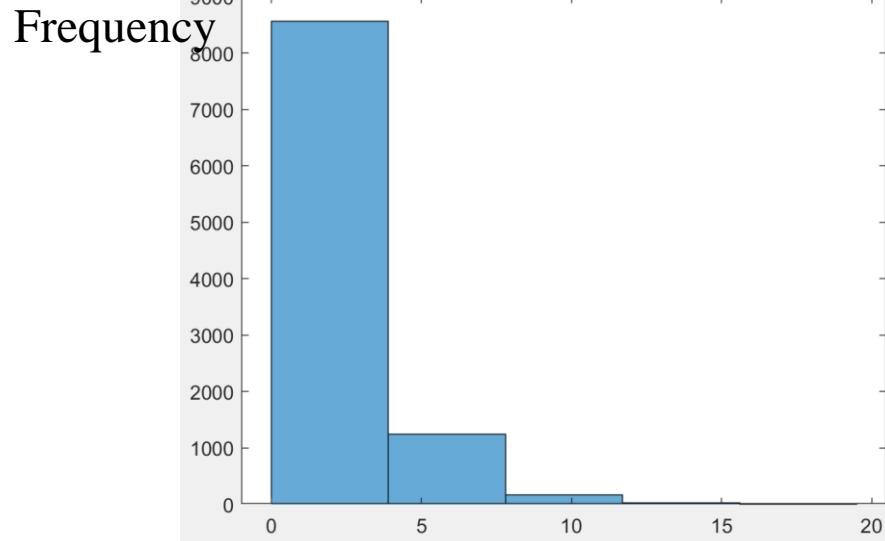
A **histogram** is a more compact(简洁的) summary of data than a stem-and-leaf diagram. To construct a histogram for continuous data, we must divide the range of the data into intervals, which are usually called **class intervals**(分类间隔), **cells**(组距), or **bins**(箱距). If possible, the bins should be **of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

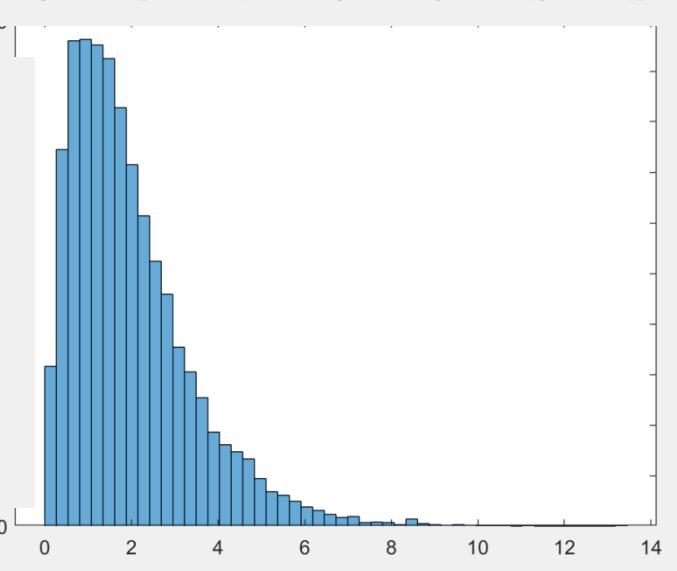
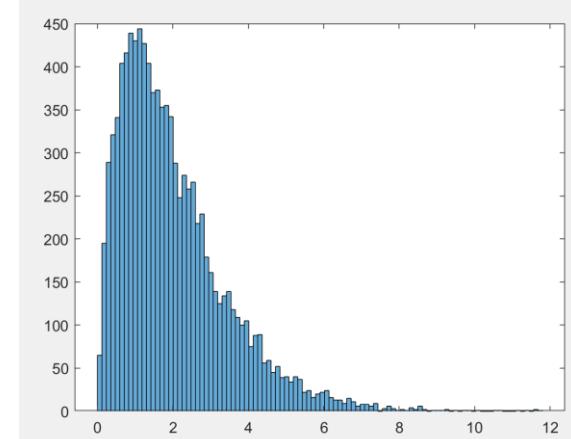
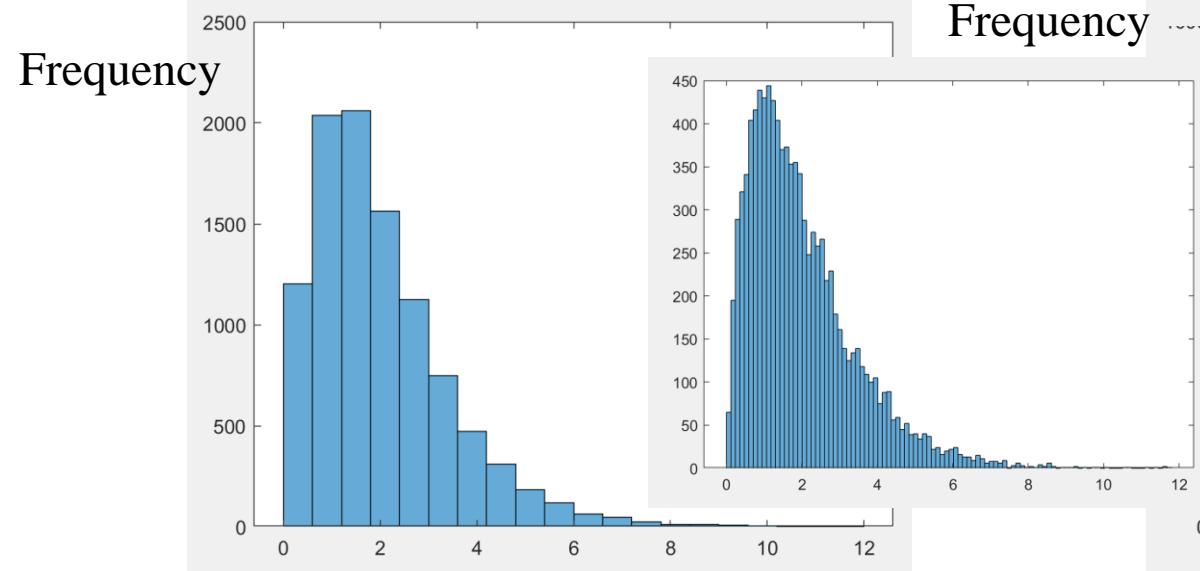
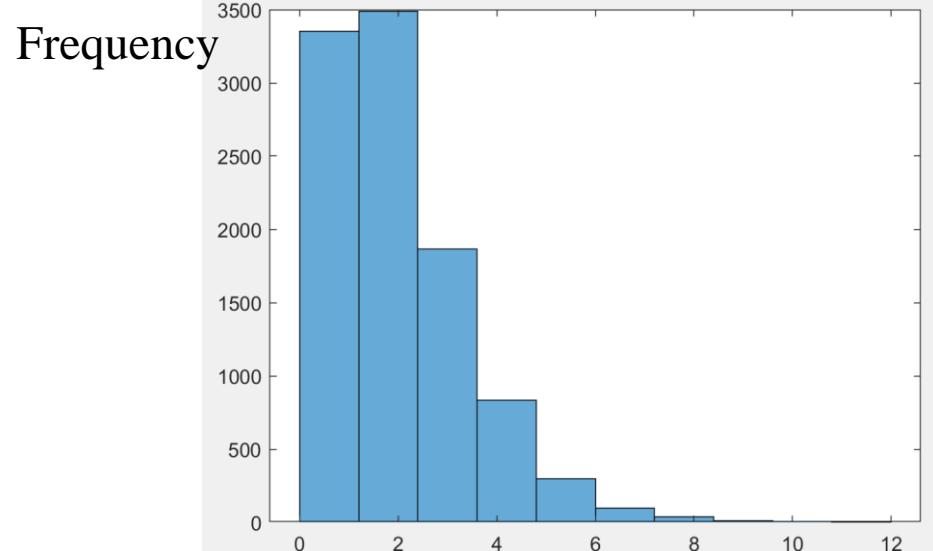
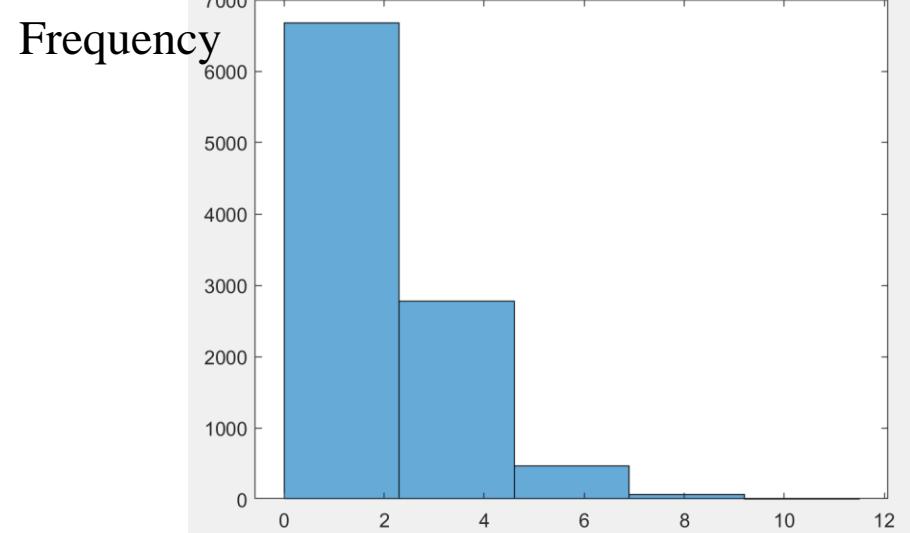
Exponential distribution with $\lambda = 0.5$,

a sample with 10000 specimens



2-3 Histograms

Gamma distribution with $\alpha = 0.5, \beta = 1$



2-3 Histograms

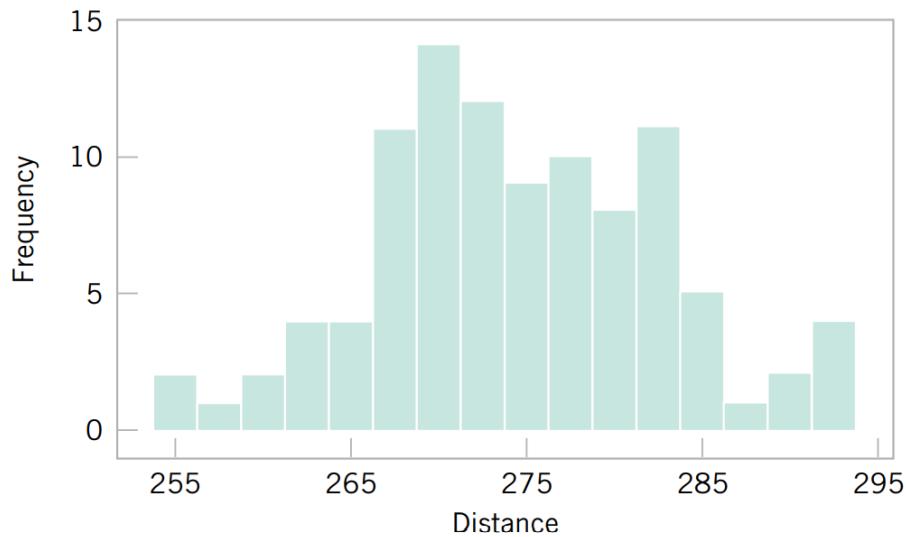


Figure 2-8 Minitab histogram with 16 bins for the golf ball distance data.

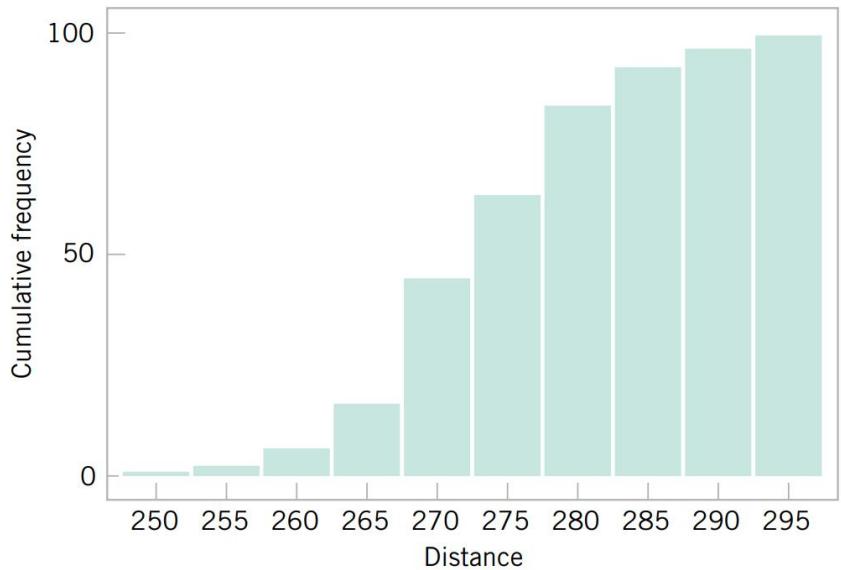


Figure 2-9 A cumulative frequency plot of the golf ball distance data from minitab.

累积频率图

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

10:45 生活周刊

沈阳你今天突然好冷

2025年10月中下旬
东北遭遇“断崖式降温”

沈阳降温 气温^{冷啊}，一下子从十几度
降到了有点太阳跟入冬了一样
降到零度以下

小羊
10-05

还是那场断崖式的气温下降
东北怎么就没有

春秋是在战国吗我
Swallows that didn't make it south froze to death in the deep autumn of the north.
"By this season, you shouldn't be able to see swallows anymore."

Say something...

这样不是更刺激吗?
天知道………… 网络素材
昨天我还就穿一件长袖
今天我都套上羽绒服了

真是一点过渡都没有😊

大连 你是直接入冬了吗?
过度都没了吗? 大连 你...

16

长春
你要冻死
我啊 😂😂😂



三联生活周刊 Author

Swallows that didn't make it south froze to death in the deep autumn of the north.

"By this season, you shouldn't be able to see swallows anymore."

On October 18, in Liaoyuan, Jilin, 49-year-old Zhang Weidong was driving on a road near the Yishan Reservoir dam when he saw a group of swallows quietly resting on the road. They "looked very cold, huddled together." Later, he searched online and saw that many people in Northeast China had also taken photos of swallows that had been stranded.

In mid-to-late October, Northeast China experienced a "plummeting temperature drop," with the temperature suddenly falling from the teens to below zero, disrupting the swallows' migration rhythm.

Liu Yu, an associate researcher at the Institute of Ecology of the Chinese Research Academy of Environmental Sciences, told this publication that climate change caused by human activities has, on one hand, led to global warming, and on the other hand, has continuously increased the frequency and intensity of extreme weather events in terrestrial regions. This makes situations like "swallows staying in the north" more likely to occur, which will further increase the risk of biodiversity loss.

#swallowsstayinthenorth #globalwarming #climate

Say something...





日照 气温 食物 内部的遗传机制

食物和内部的遗传机制



三联生活周刊

Follow

Swallows that didn't make it south froze to death in the deep autumn of the north.
"By this season, you shouldn't be able to see swallows anymore." On October 18,

Say something...

336 comment(s)



三联生活周刊 Author

Swallows that didn't make it south froze to death in the deep autumn of the north.

"By this season, you shouldn't be able to see swallows anymore."

On October 18, in Liaoyuan, Jilin, 49-year-old Zhang Weidong was driving on a road near the Yishan Reservoir dam when he saw a group of swallows quietly resting on the road. They "looked very cold, huddled together." Later, he searched online and saw that many people in Northeast China had also taken photos of swallows that had been stranded.

In mid-to-late October, Northeast China experienced a "plummeting temperature drop," with the temperature suddenly falling from the teens to below zero, disrupting the swallows' migration rhythm.

Liu Yu, an associate researcher at the Institute of Ecology of the Chinese Research Academy of Environmental Sciences, told this publication that climate change caused by human activities has, on one hand, led to global warming, and on the other hand, has continuously increased the frequency and intensity of extreme weather events in terrestrial regions. This makes situations like "swallows staying in the north" more likely to occur, which will further increase the risk of biodiversity loss.

#swallowsstayinthenorth #globalwarming #climate

Say something...



10:45 三联生活周刊

网络素材



所以说它就也没走



三联生活周刊

Follow

Swallows that didn't make it south froze to death in the deep autumn of the north
"By this season, you shouldn't be able to see swallows anymore." On October 18,

Say something...

336 comment(s)



三联生活周刊 Author

Swallows that didn't make it south froze to death in the deep autumn of the north.

"By this season, you shouldn't be able to see swallows anymore."

On October 18, in Liaoyuan, Jilin, 49-year-old Zhang Weidong was driving on a road near the Yishan Reservoir dam when he saw a group of swallows quietly resting on the road. They "looked very cold, huddled together." Later, he searched online and saw that many people in Northeast China had also taken photos of swallows that had been stranded.

In mid-to-late October, Northeast China experienced a "plummeting temperature drop," with the temperature suddenly falling from the teens to below zero, disrupting the swallows' migration rhythm.

Liu Yu, an associate researcher at the Institute of Ecology of the Chinese Research Academy of Environmental Sciences, told this publication that climate change caused by human activities has, on one hand, led to global warming, and on the other hand, has continuously increased the frequency and intensity of extreme weather events in terrestrial regions. This makes situations like "swallows staying in the north" more likely to occur, which will further increase the risk of biodiversity loss.

#swallowsstayinthenorth #globalwarming #climate

Say something...



10:45 三联生活周刊

网络素材



三联生活周刊

Follow

Swallows that didn't make it south froze to death in the deep autumn of the north.
"By this season, you shouldn't be able to see swallows anymore." On October 18,

Say something...

Say something...



336 comment(s)



三联生活周刊 Author

Swallows that didn't make it south froze to death in the deep autumn of the north.

"By this season, you shouldn't be able to see swallows anymore."

On October 18, in Liaoyuan, Jilin, 49-year-old Zhang Weidong was driving on a road near the Yishan Reservoir dam when he saw a group of swallows quietly resting on the road. They "looked very cold, huddled together." Later, he searched online and saw that many people in Northeast China had also taken photos of swallows that had been stranded.

In mid-to-late October, Northeast China experienced a "plummeting temperature drop," with the temperature suddenly falling from the teens to below zero, disrupting the swallows' migration rhythm.

Liu Yu, an associate researcher at the Institute of Ecology of the Chinese Research Academy of Environmental Sciences, told this publication that climate change caused by human activities has, on one hand, led to global warming, and on the other hand, has continuously increased the frequency and intensity of extreme weather events in terrestrial regions. This makes situations like "swallows staying in the north" more likely to occur, which will further increase the risk of biodiversity loss.

#swallowsstayinthenorth #globalwarming #climate

舍甫琴科



(吉国传奇斗士)

姓名：瓦伦蒂娜·舍甫琴科
别名：子弹
国籍：秘鲁
出生日期：1988年3月7日（37岁）
身高：165cm
体重：56.2kg（参赛体重）
毕业院校：吉尔吉斯国立艺术学院
运动项目：综合格斗（MMA）、泰拳
专业特点：

- 泰拳技术世界顶级
- 精准的站立打击
- 出色的防守反击能力
- 丰富的比赛经验

主要奖项：

- 8次世界泰拳联合会世界锦标赛冠军
- 2届世界武博运动会泰拳女子60公斤级冠军
- 2014年昆仑决9·木兰传奇女子60公斤级八人争霸赛冠军
- 8次卫冕UFC女子蝇量级冠军

经典语录：

"我为格斗而生"

张伟丽



(中国格斗女王)

姓名：张伟丽
国籍：中国
祖籍：河北省邯郸市
出生日期：1990年9月30日（35岁）
身高：163cm
体重：54kg（参赛体重）
毕业院校：邯郸少林智勇武术院
运动项目：综合格斗（MMA）
专业特点：

- 站立击打技术精湛
- 地面缠斗能力突出
- 体能充沛，抗击打能力强
- 比赛节奏控制出色

主要奖项：

- UFC女子草量级世界冠军
- 亚洲首位UFC世界冠军
- 昆仑决女子草量级冠军
- 韩国TOP FC女子草量级冠军
- 3次成功卫冕UFC金腰带

经典语录：

"我是冠军，不是暴君"

10:29

6G 4G

连播



张伟丽 (Following)

It's showtime. #zhangweilislifeisendless

♪ 张伟丽的原声 | Post now

Say something...

239

7,378

96

1,354

10:29

6G 4G



连播



张伟丽 (Following)

We all have a positive and negative self in our hearts. We should be grateful for the positive and upward "me" within us, always maintaining an upward posture, daring to venture and strive, and doing our utmost to reach our own highest goal. There is no failure or success, only process and success!

Say something...

317

31.8K

568

4,942

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

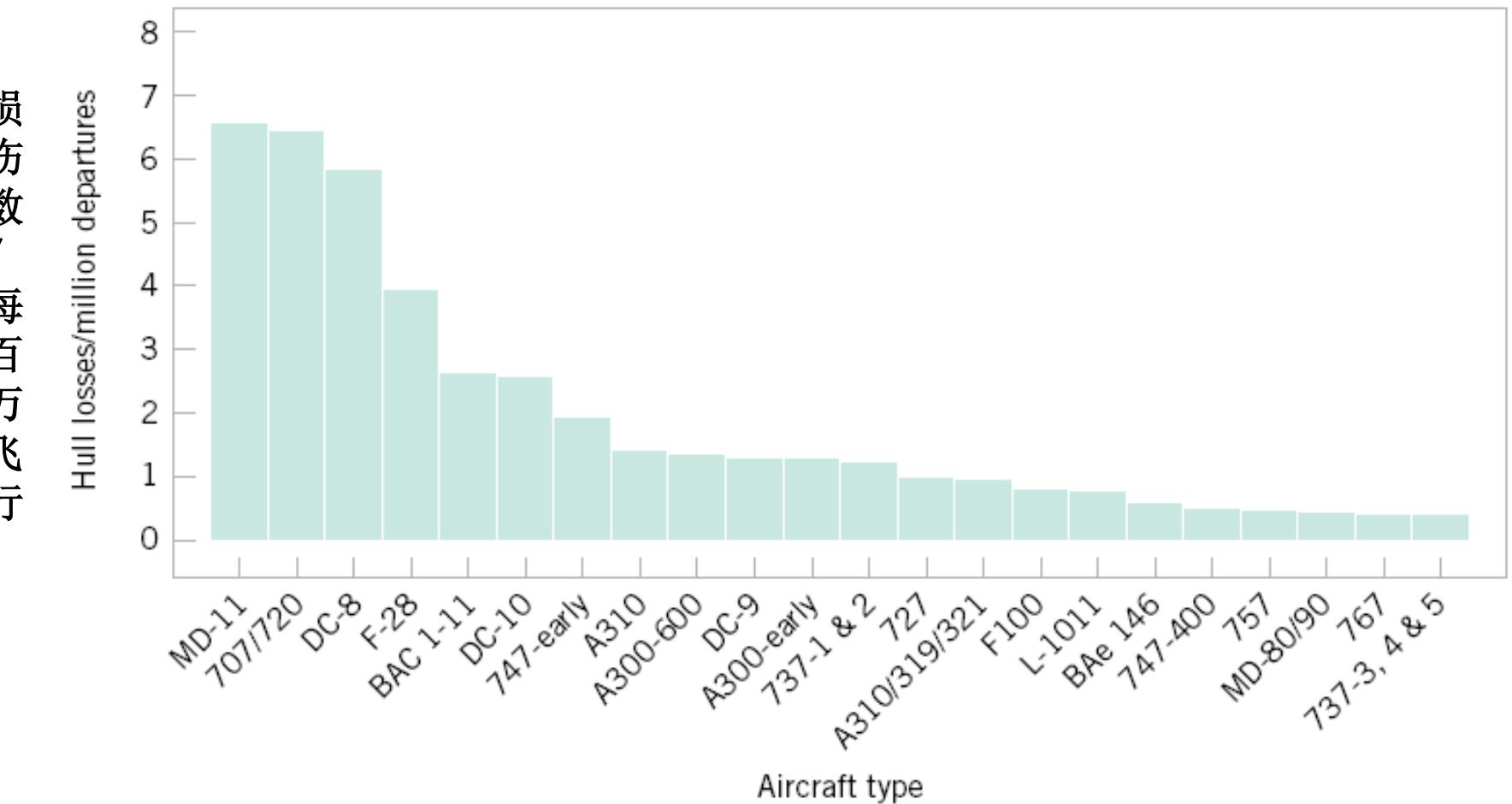


Figure 2-11 Pareto chart for the aircraft accident data.

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

- The **box plot** (箱线图) is a graphical display that simultaneously describes several important features of a data set, such as
 - **center**(中心),
 - **spread**(范围),
 - **departure from symmetry**(对称性的偏离),
- and identification of observations that lie unusually far from the bulk of the data(**Outlier**, 离群点).
 - **Whisker**(触须)
 - **Outlier**(离群点/异常值)
 - **Extreme outlier**(极端离群点)

Example: The impact of the US election on the stock market

2-4 Box Plots

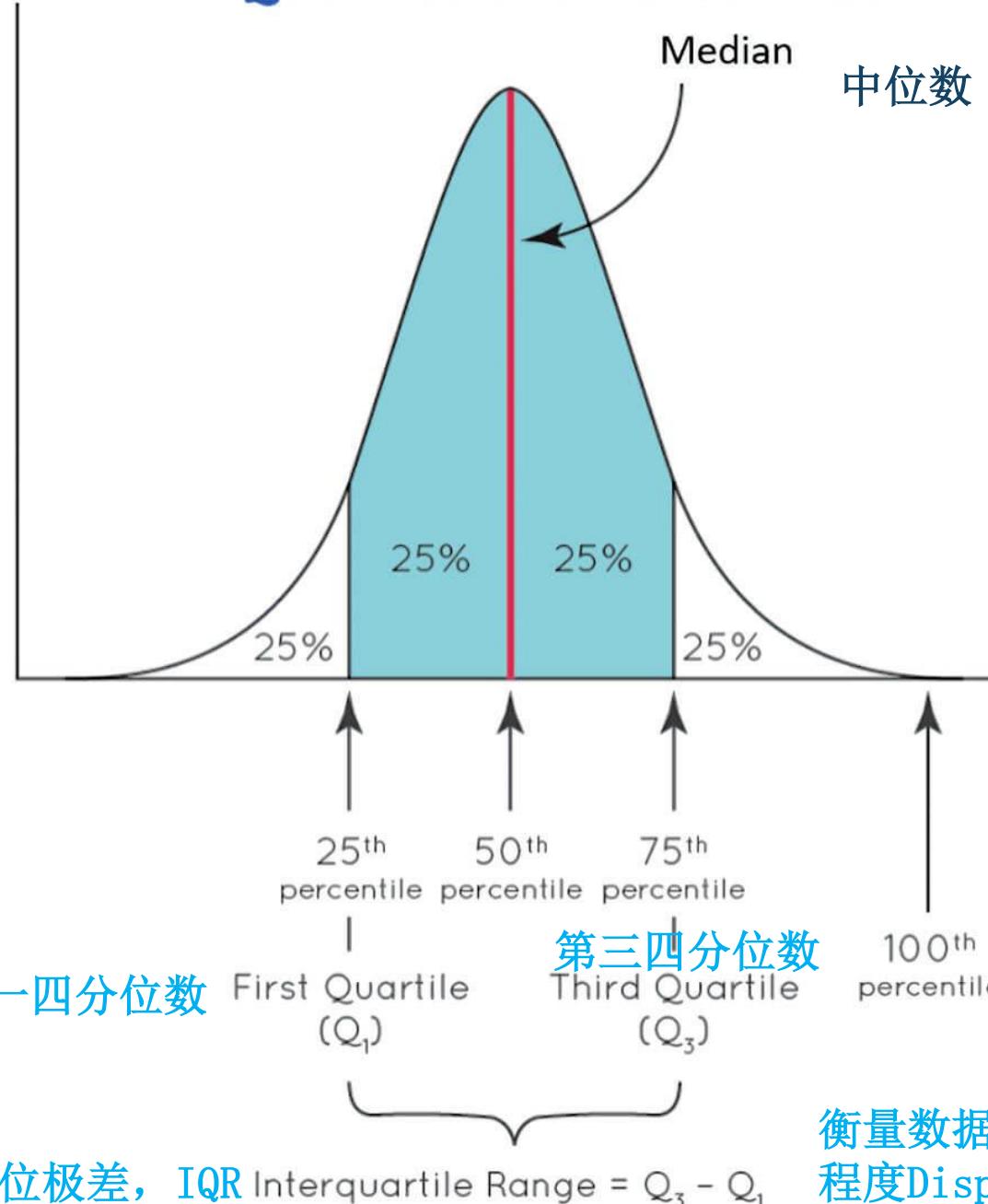
Example: The impact of the US election on the stock market



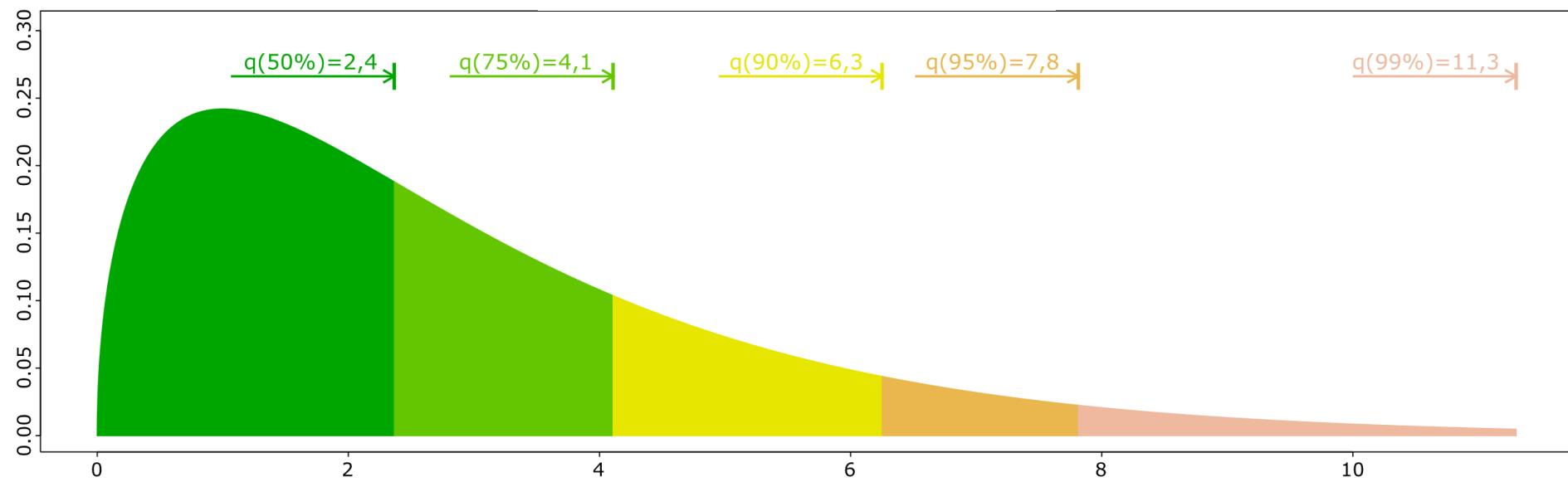
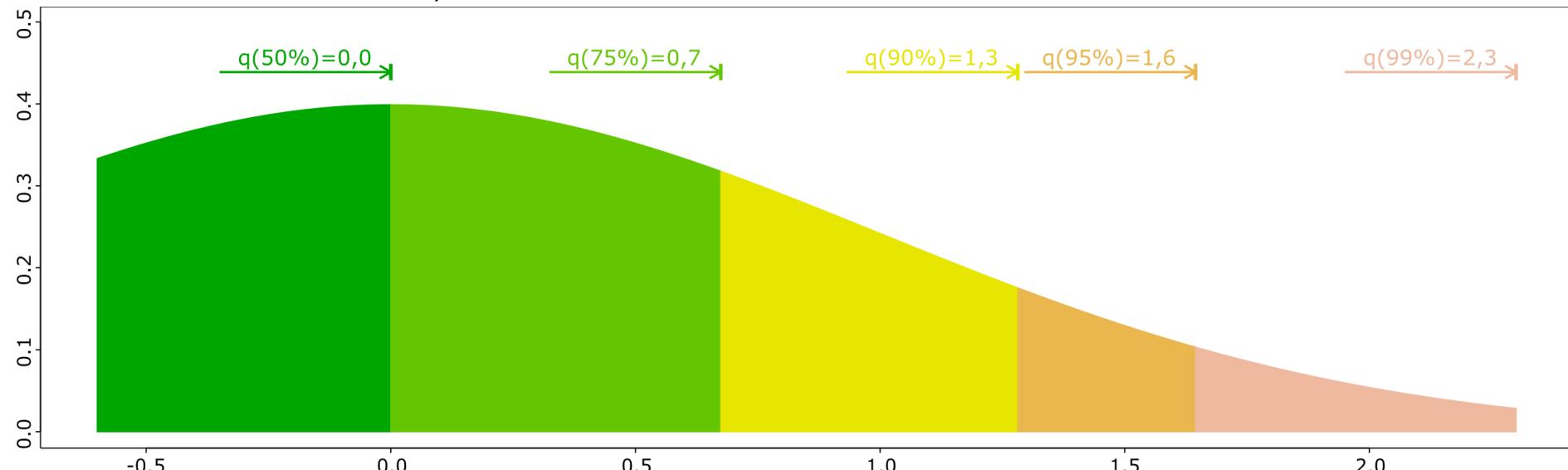
Normal distribution

Quartiles and Percentiles

四分位数与百分位数



Two examples of skewed distributions 偏态分布



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/percentiles = 中位数 = median

[Q1 – 1.5 *IQR, Q1]

Whisker extends to
smallest data point within
1.5 interquartile ranges from
first quartile

[Q3, Q3 + 1.5*IQR]

Whisker extends to
largest data point within
1.5 interquartile ranges
from third quartile

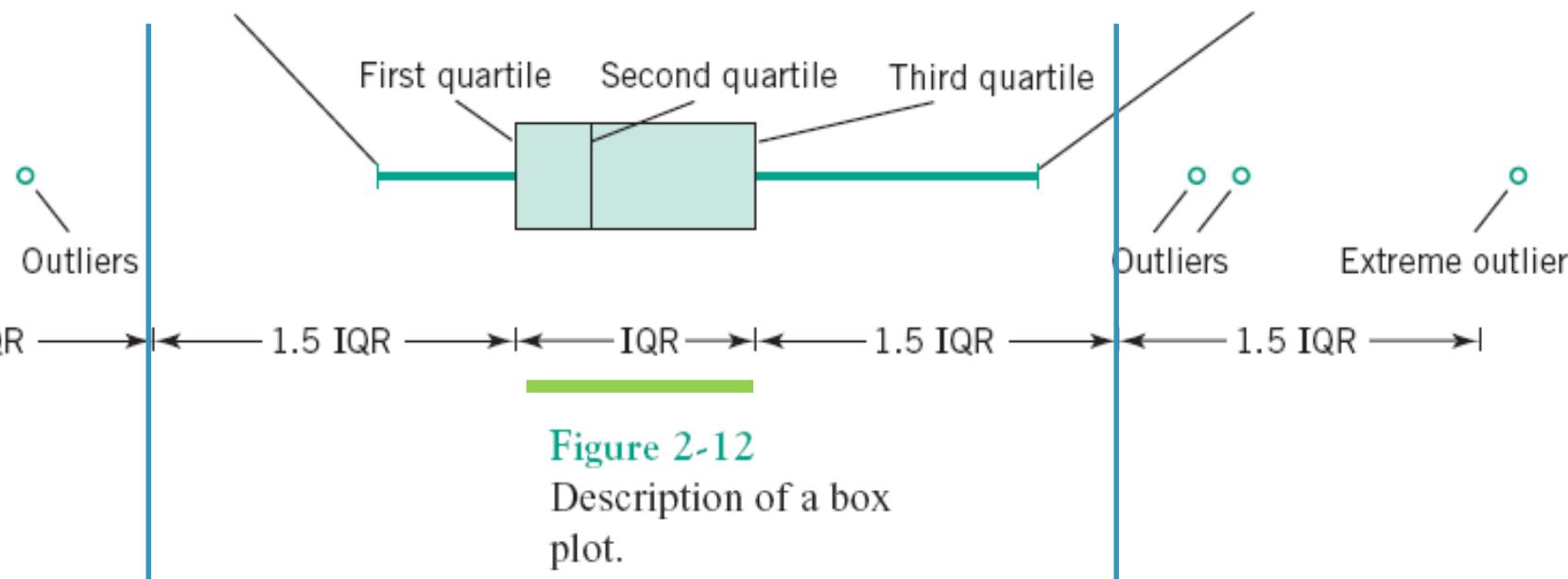
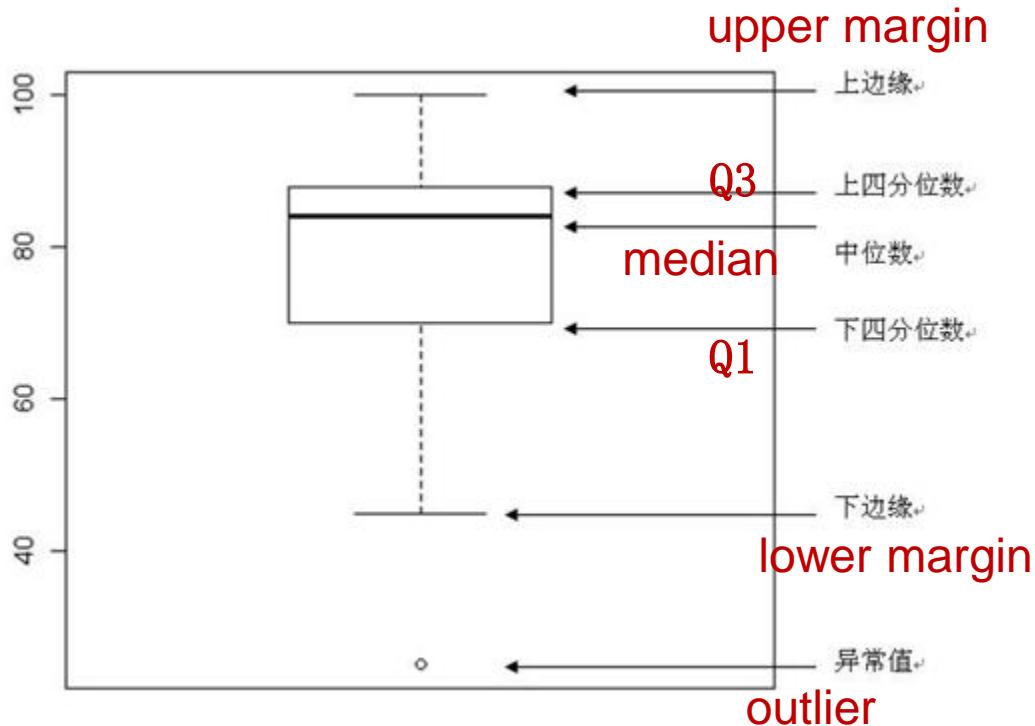


Figure 2-12
Description of a box
plot.

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） = $Q1 - 1.5 \text{ IQR}$
- 最大观察值（上边缘） = $Q3 + 1.5 \text{ IQR}$

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) = $Q1 - 1.5 * \text{IQR}$
- Maximum observed value (upper margin) = $Q3 + 1.5 * \text{IQR}$

2-4 Box Plots

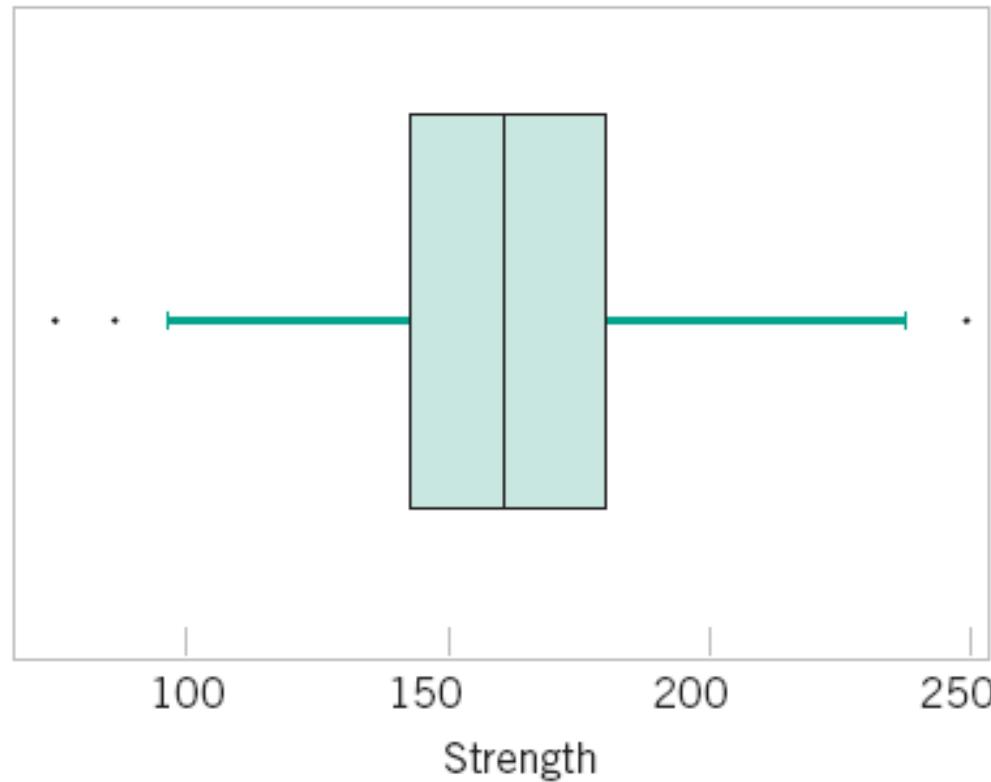


Figure 2-13 Box plot for compressive strength data in Table 2-2.

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 $Q3=(19+21)/2=20$

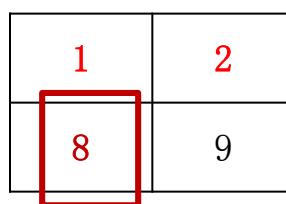
1	2	3	4	5	6	7	8	9
8	9	10	12	13	19	19	21	48

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

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

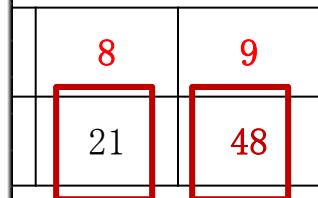
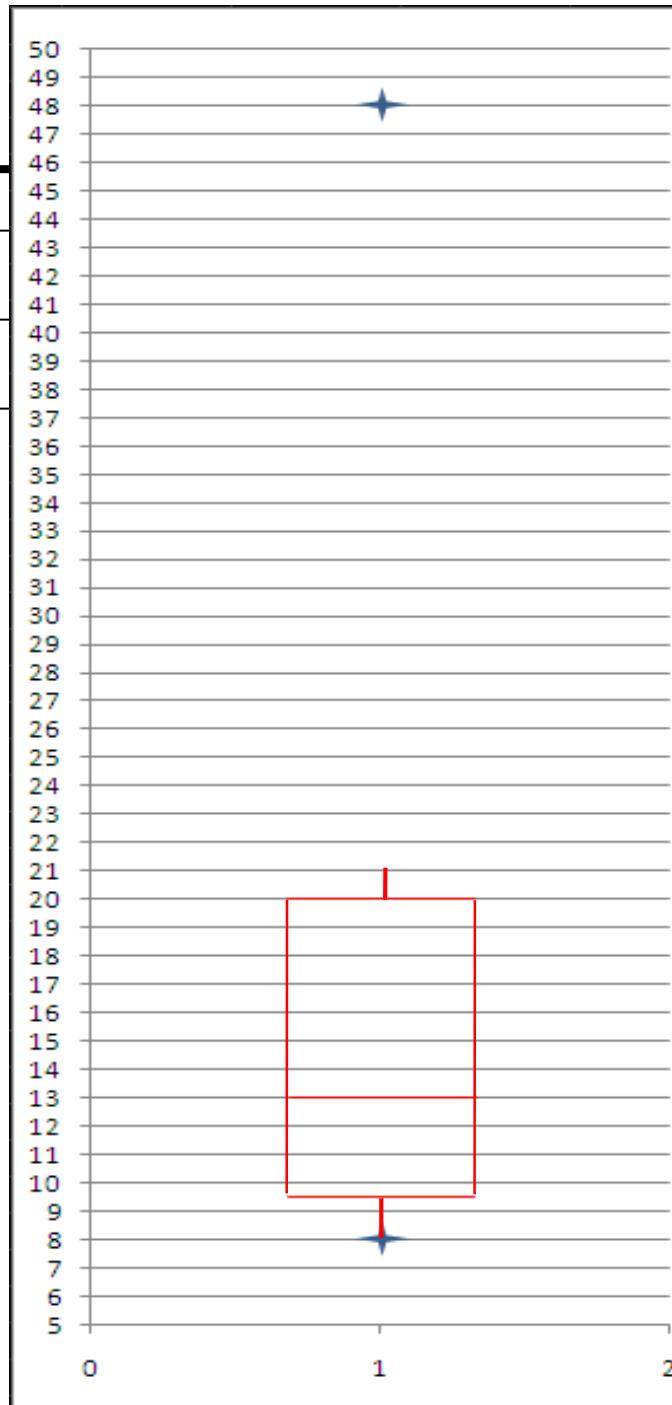
2-4 Box Plots

Min	8
Max	48
Q1	9.5
Q2	13
Q3	20



[Q3, Q3 + 1.5*IQR]
Upper whisker $\in [20, 35.75]$

[Q1 - 1.5 *IQR, Q1]
Lower whisker $\in [-6.25, 9.5]$



2-4 Box Plots

Purchasing Manager 采购经理 / Procurement Manager (偏战略采购、供应链场景)

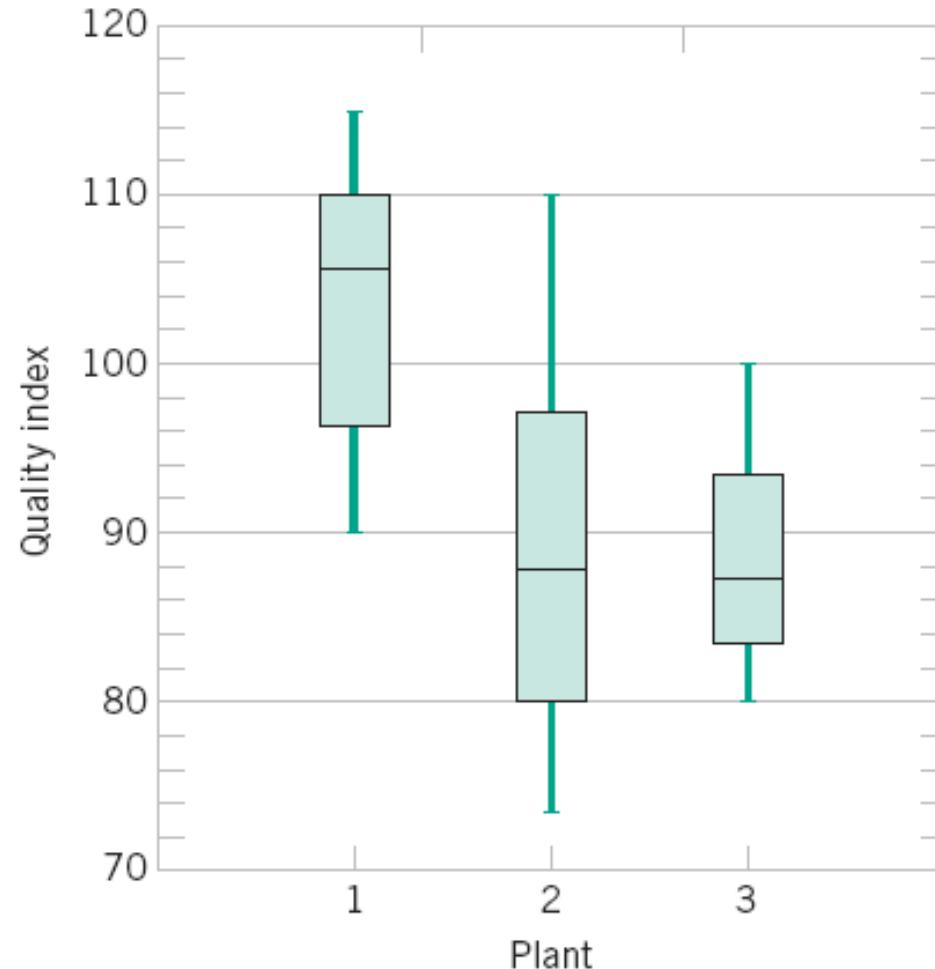


Figure 2-14 Comparative box plots of a quality index at three plants.

2-4 Box Plots

PMI (Purchasing Managers' Index)

- Definition: A globally recognized leading indicator of economic sentiment. It reflects the state of business activities in the manufacturing industry or service sector by surveying purchasing managers' opinions on key indicators such as production, new orders, inventory, supplier delivery, and employment.
- Key threshold: An index of ≥ 50 usually indicates economic expansion, while < 50 indicates economic contraction.
- 定义：全球通用的经济景气度先行指标，通过调查采购经理对生产、新订单、库存、供应商配送、就业等核心指标的看法，反映制造业或服务业的经营活动状况。
- 关键阈值：指数 ≥ 50 通常表示经济扩张， < 50 表示经济收缩。
- 常见细分：制造业 PMI (Manufacturing PMI) 、服务业 PMI (Services PMI) 、综合 PMI (Composite PMI) 。
- 示例：中国制造业采购经理指数连续 3 个月高于 50，显示工业生产持续回暖。China's Manufacturing Purchasing Managers' Index has been above 50 for three consecutive months, indicating sustained recovery in industrial production.

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:

6 8

10 5 9 12

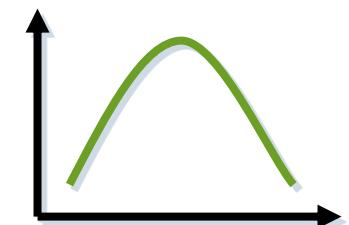


➤ 一个众数

原始数据:

8 5 5

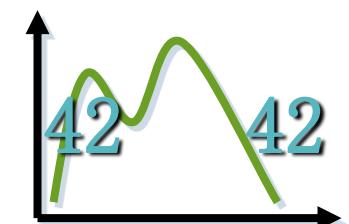
6 5 9



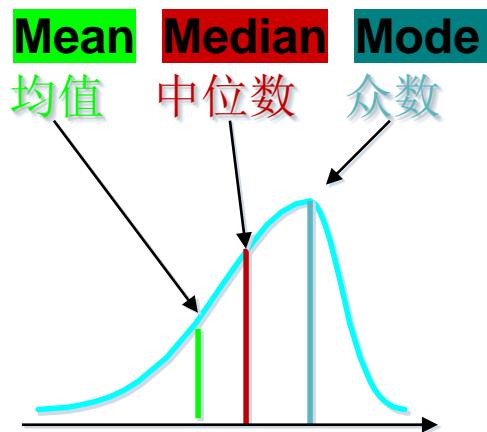
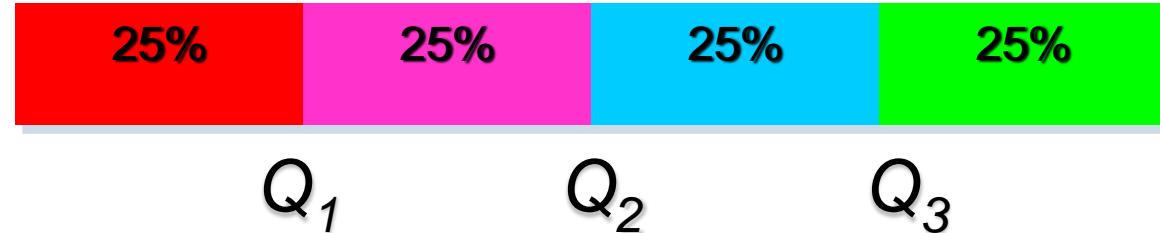
➤ 多于一个众数

原始数据:

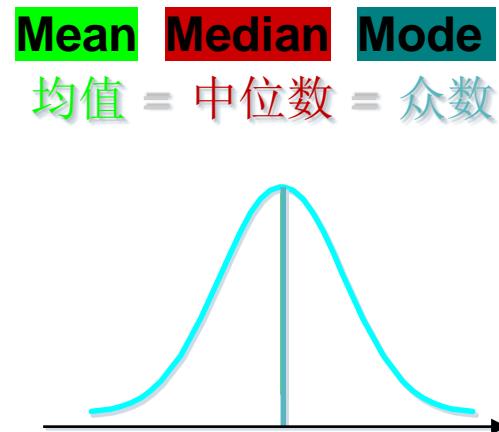
25 28 28 36



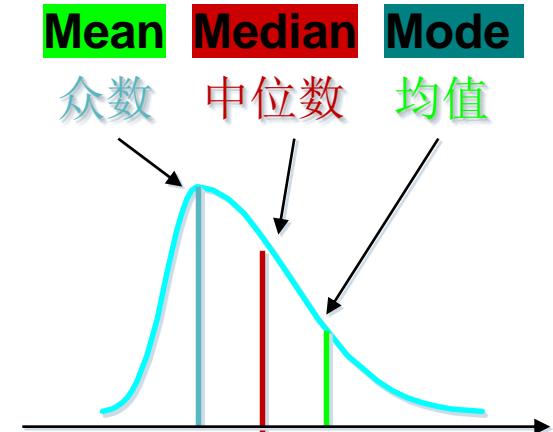
补充



左偏分布



对称分布



右偏分布

Left skewed distribution

Symmetric distribution

Right skewed distribution

补充

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：

$$\begin{aligned}\bar{G} &= \sqrt[4]{104.5\% \times 102.1\% \times 125.5\% \times 101.9\%} - 1 \\ &= 8.0787\%\end{aligned}$$

算术平均：

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

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 $100A \times 0.95$ ；
第二道工序的合格品为 $(100A \times 0.95) \times 0.92$ ；
.....
第五道工序的合格品为
 $(100A \times 0.95 \times 0.92 \times 0.90 \times 0.85) \times 0.80$ ；

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

$$100A \times 0.95 \times 0.92 \times 0.90 \times 0.85 \times 0.80;$$

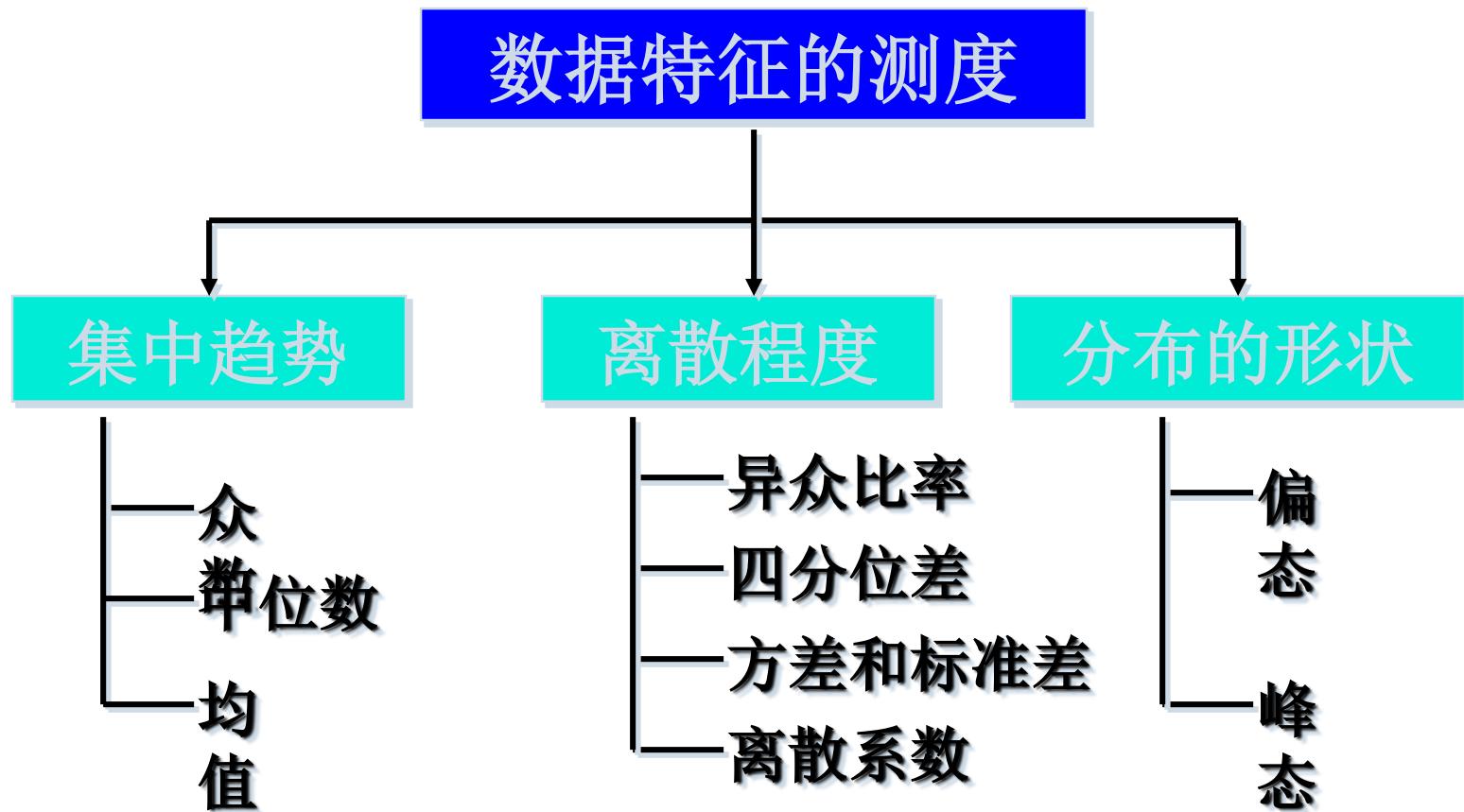
则该流水线产品总的合格率为：

$$\frac{\text{总合格品}}{\text{总产品}} = \frac{100A \times 0.95 \times 0.92 \times 0.90 \times 0.85 \times 0.80}{100A}$$
$$= 0.95 \times 0.92 \times 0.90 \times 0.85 \times 0.80$$

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。

$$\overline{X}_G = \sqrt[5]{0.95 \times 0.92 \times 0.90 \times 0.85 \times 0.80}$$
$$= \sqrt[5]{0.5349} = 88.24\%$$

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

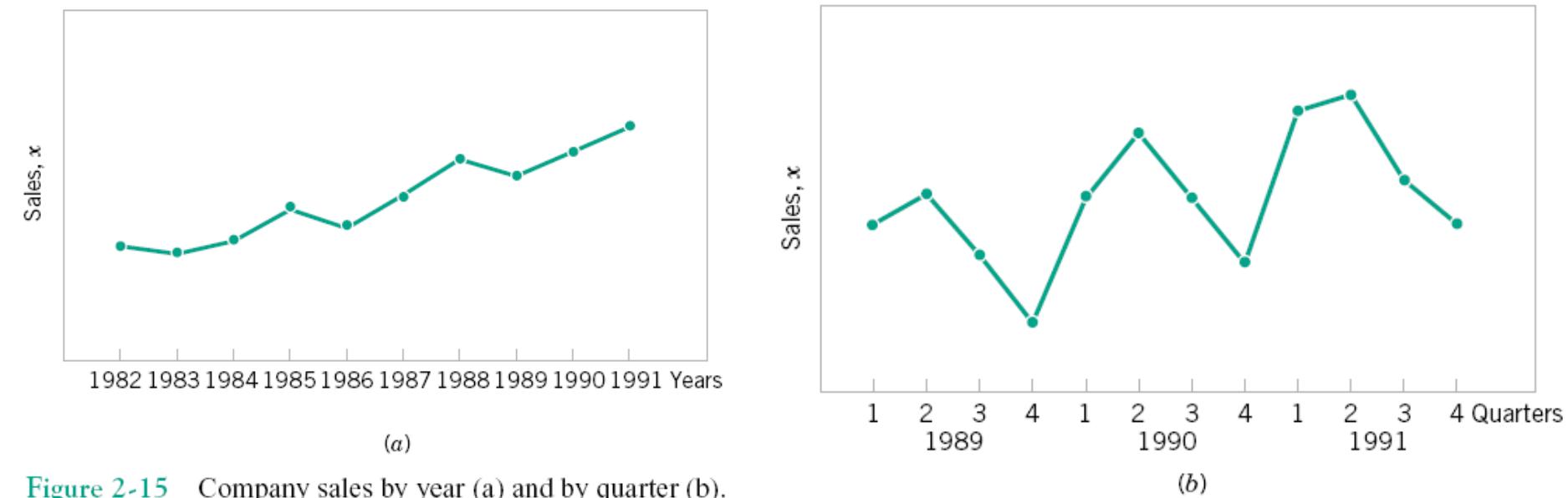


Figure 2-15 Company sales by year (a) and by quarter (b).

2-5 Time Series Plots

the alloy compressive strength data(合金压力强度数据) in Stem-and-Leaf Diagram

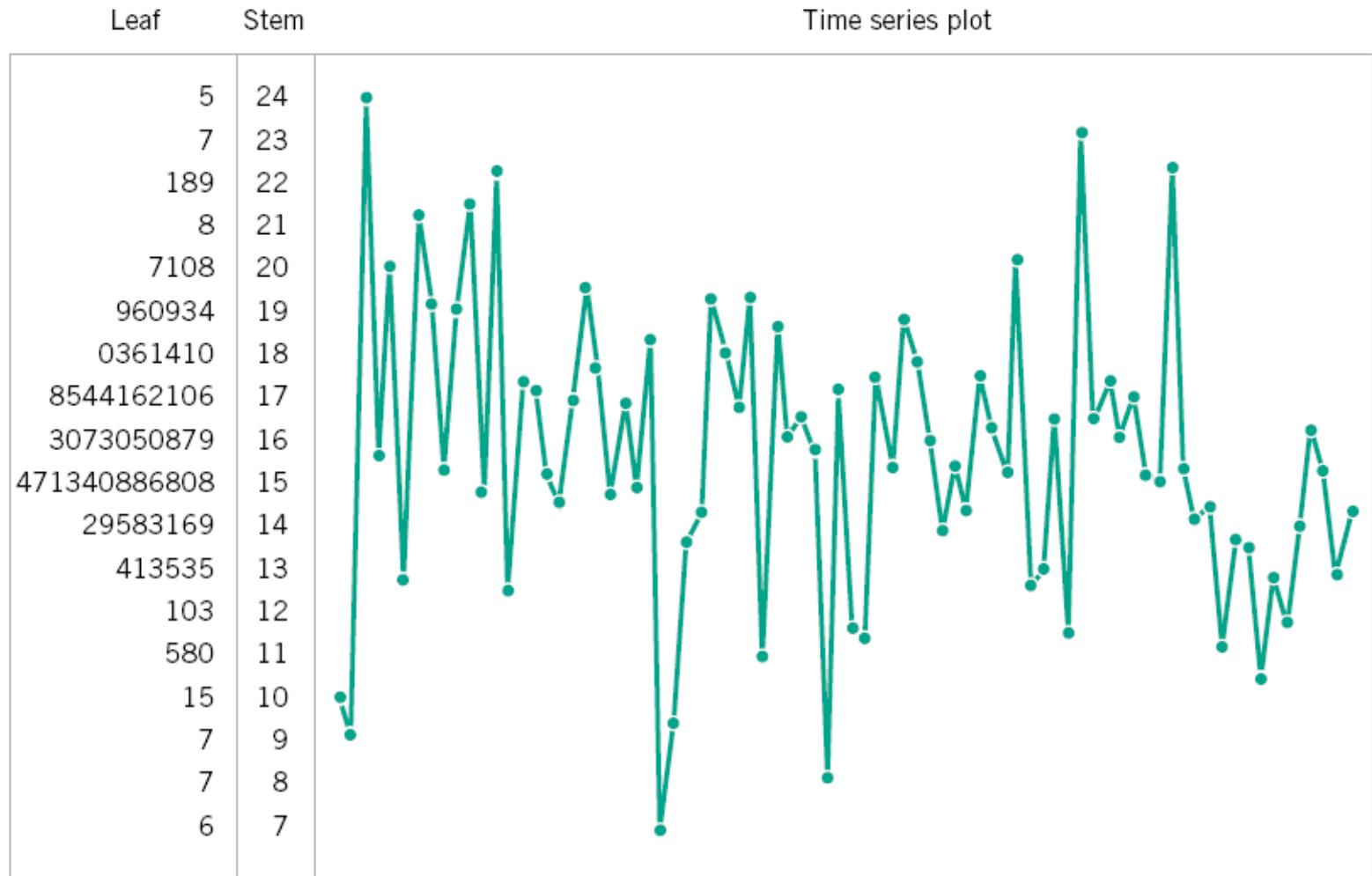


Figure 2-16 A digidot plot of the compressive strength data in Table 2-2.

2-5 Time Series Plots

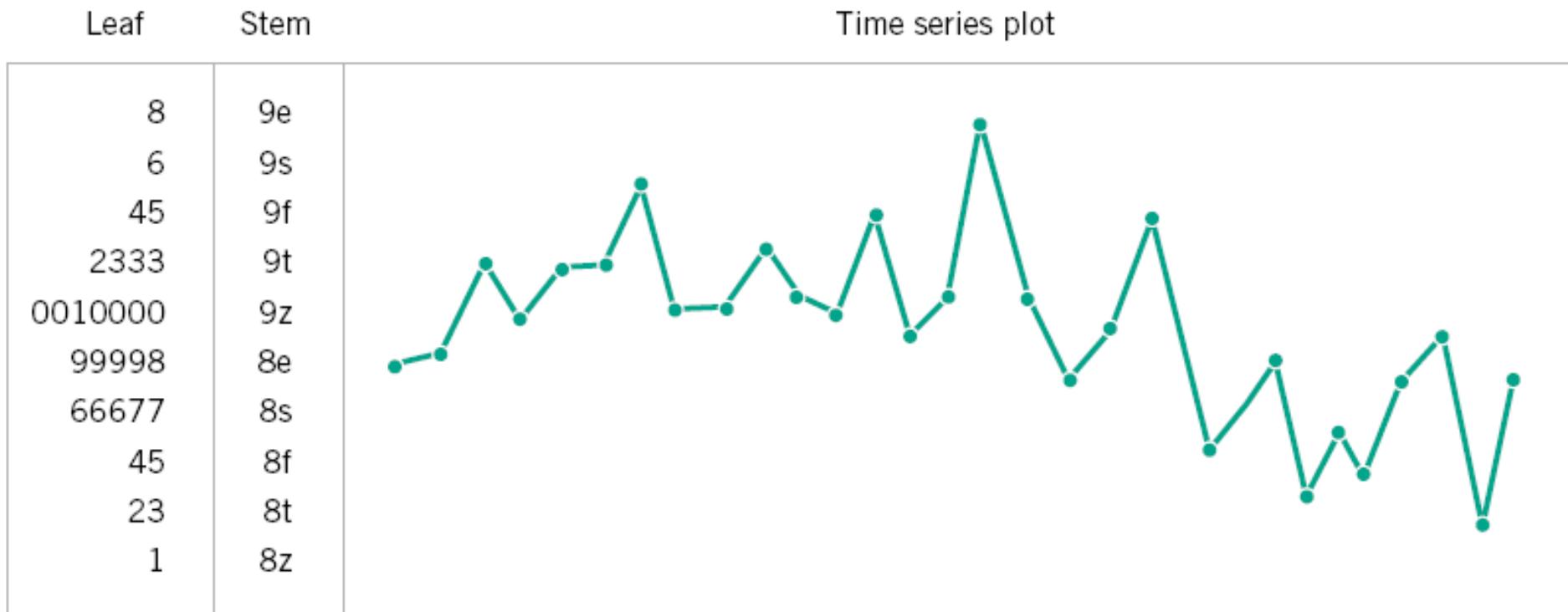


Figure 2-17 A digidot plot of chemical process concentration readings, observed hourly.

茎叶图与时间序列图结合：数点图

2-6 Multivariate Data

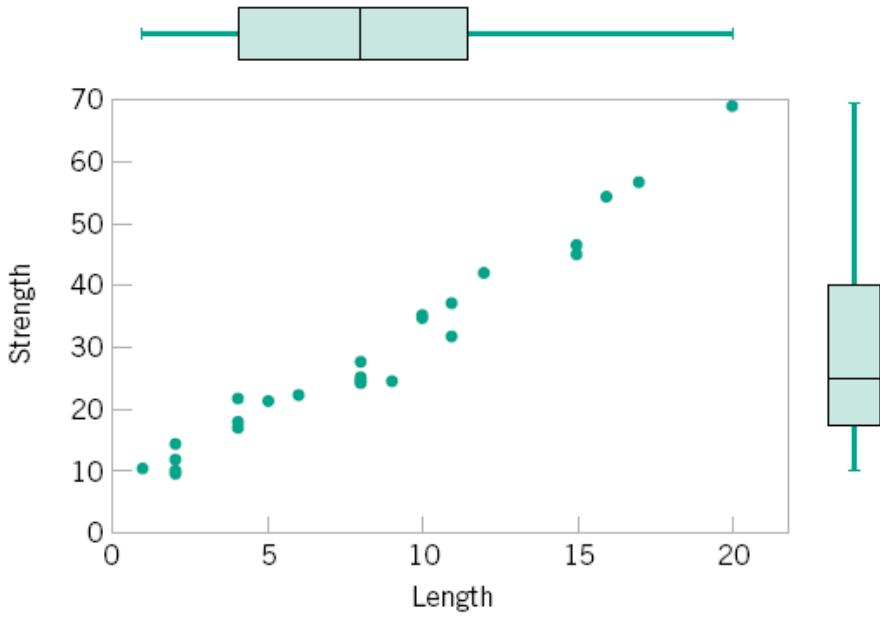
- The dot diagram, stem-and-leaf diagram, histogram, and box plot are descriptive displays for **univariate(单变量的)** data; that is, they **convey(传达, 表现)** **descriptive information(描述信息)** about a single variable.
- Many engineering problems involve collecting and analyzing **multivariate data(多变量数据)**, 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

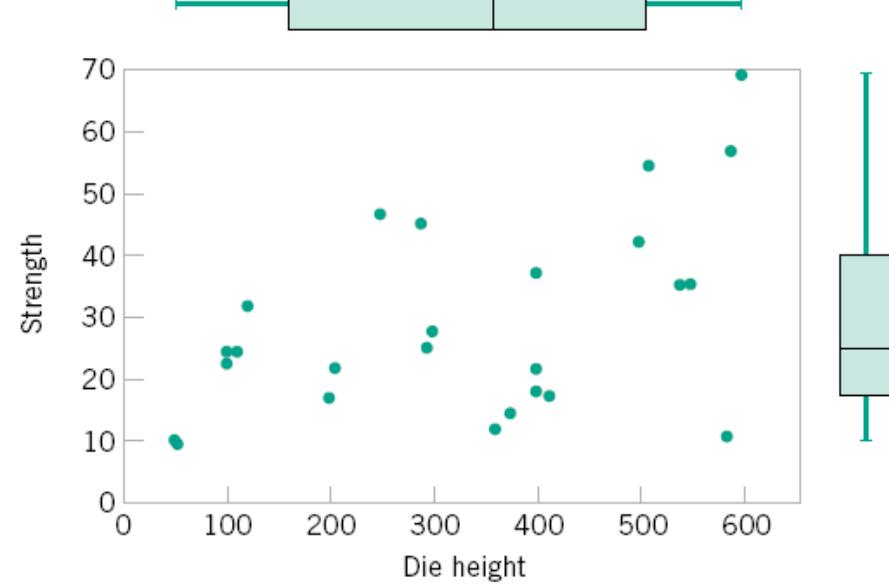
Table 2-9 Wire Bond Data

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

➤ Sample Covariance Formula 样本协方差公式

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

- 通过“每个观测值与均值的偏差乘积之和”，衡量两个变量偏差的协同变化（若 x 高于均值时 y 也高于均值，乘积为正，整体求和后反映正相关趋势）。

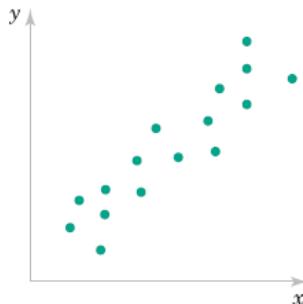
Given n pairs of data $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$, the **sample correlation coefficient** r is defined by

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

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

皮尔逊样本相关系数 (Pearson Sample Correlation Coefficient)

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 $0.8 \leq r \leq 1$, weak when $0 \leq r \leq 0.5$, 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.
泡沫、气味、颜色和残留物 (衡量清洁能力的指标) 是采用 10 分制评估的描述性属性
- Quality is a measure of **overall desirability** 整体满意度 of the shampoo, and it is the **nominal** 名义上的 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

- Calculate the Sample Correlation Coefficient 样本相关系数 by pairs of each factor/variable
- Fill the above correlation coefficients into the table in sequence

	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

- Could anyone tell me what the missing values in the above table are?

2-6 Multivariate Data

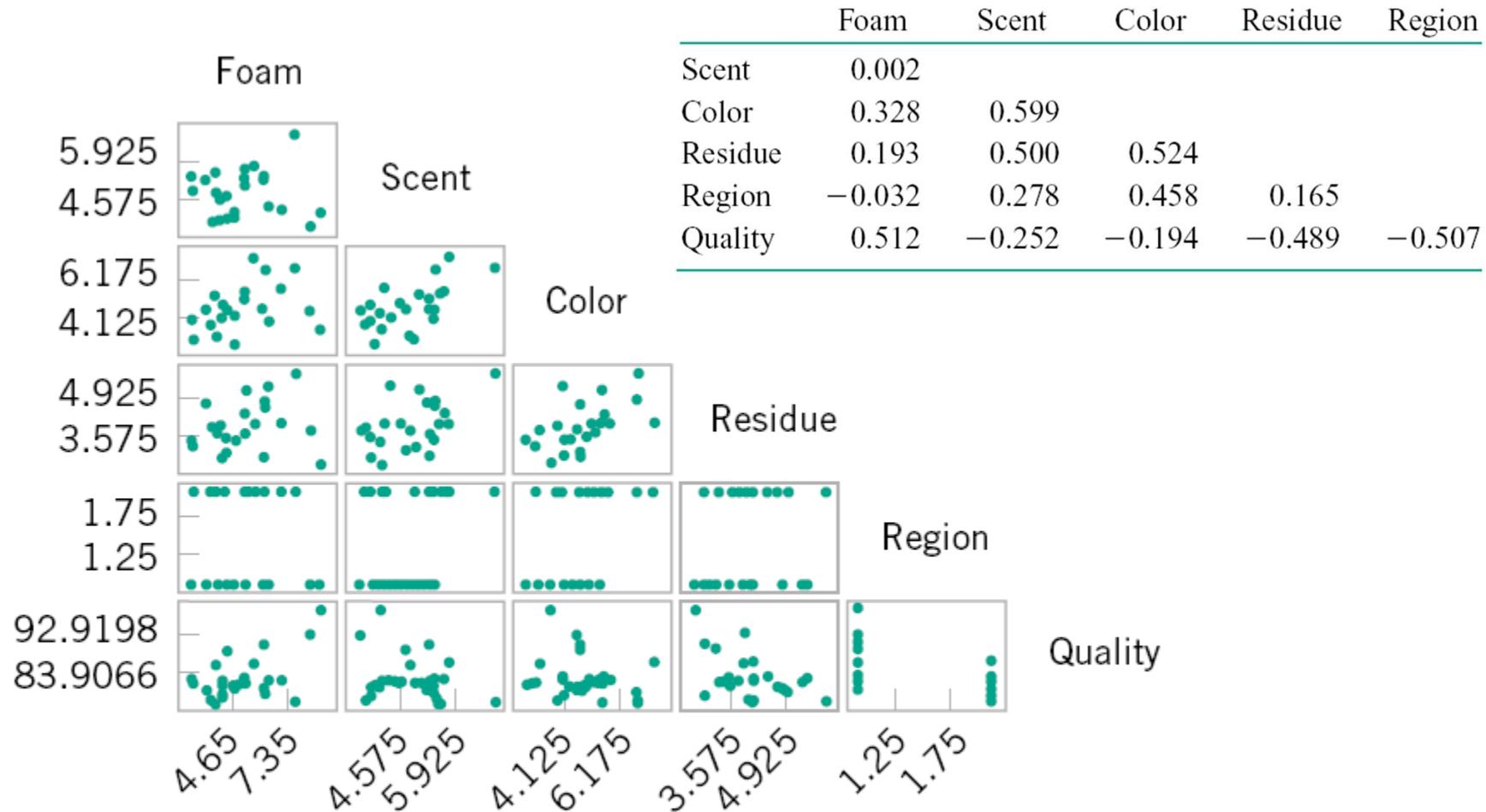


Figure 2-22 Matrix of scatter plots for the shampoo data in Table 2-11.

2-6 Multivariate Data

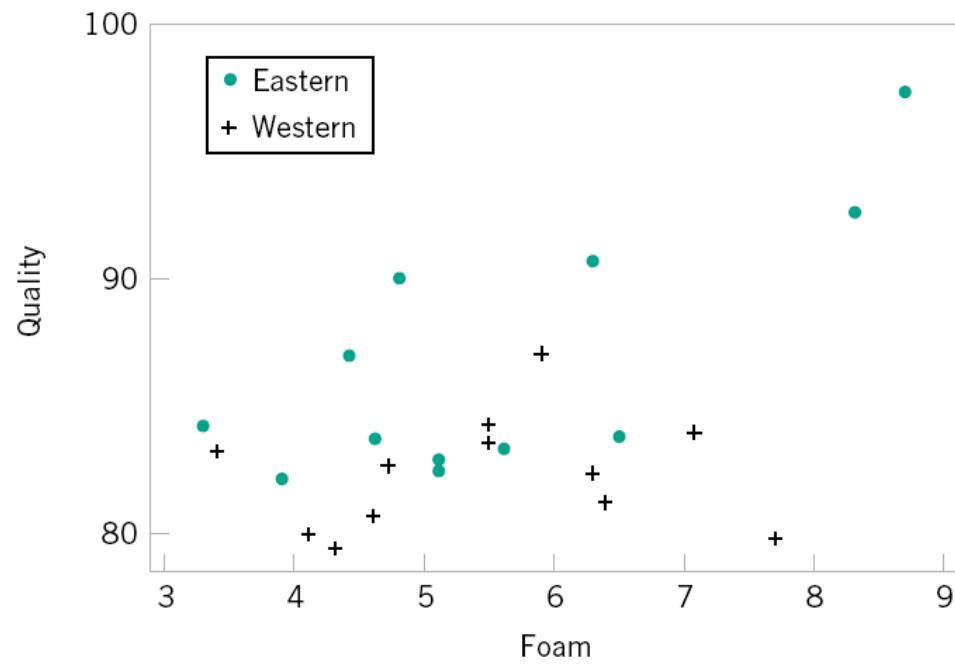
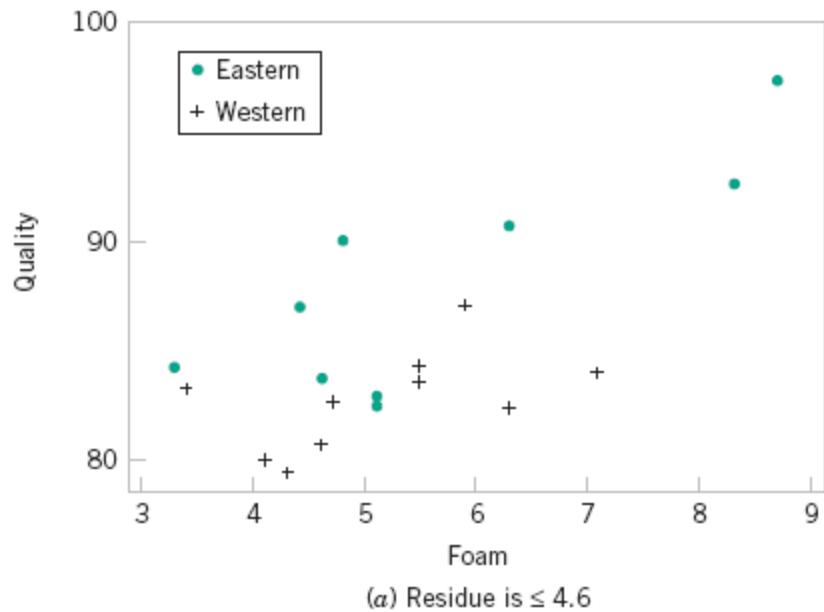


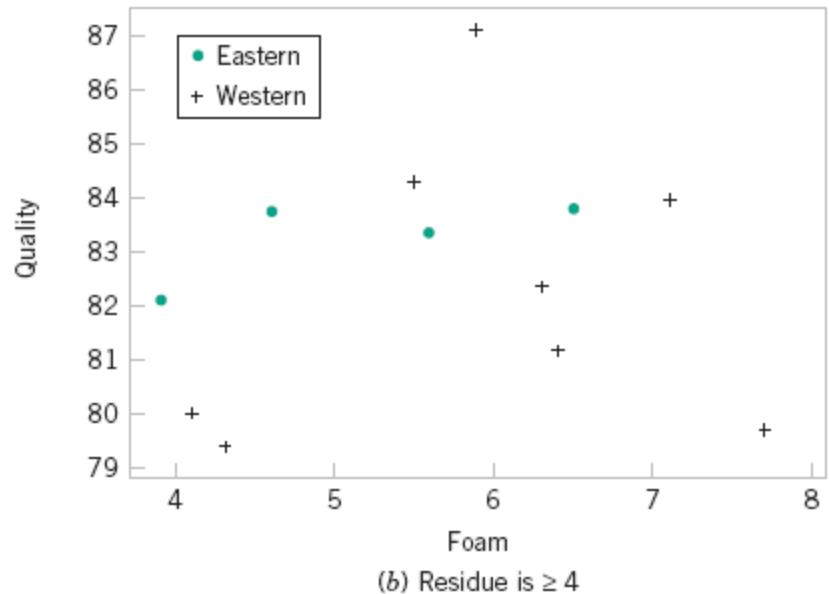
Figure 2-23 Scatter diagram of shampoo quality versus foam.

- 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



(a) Residue is ≤ 4.6

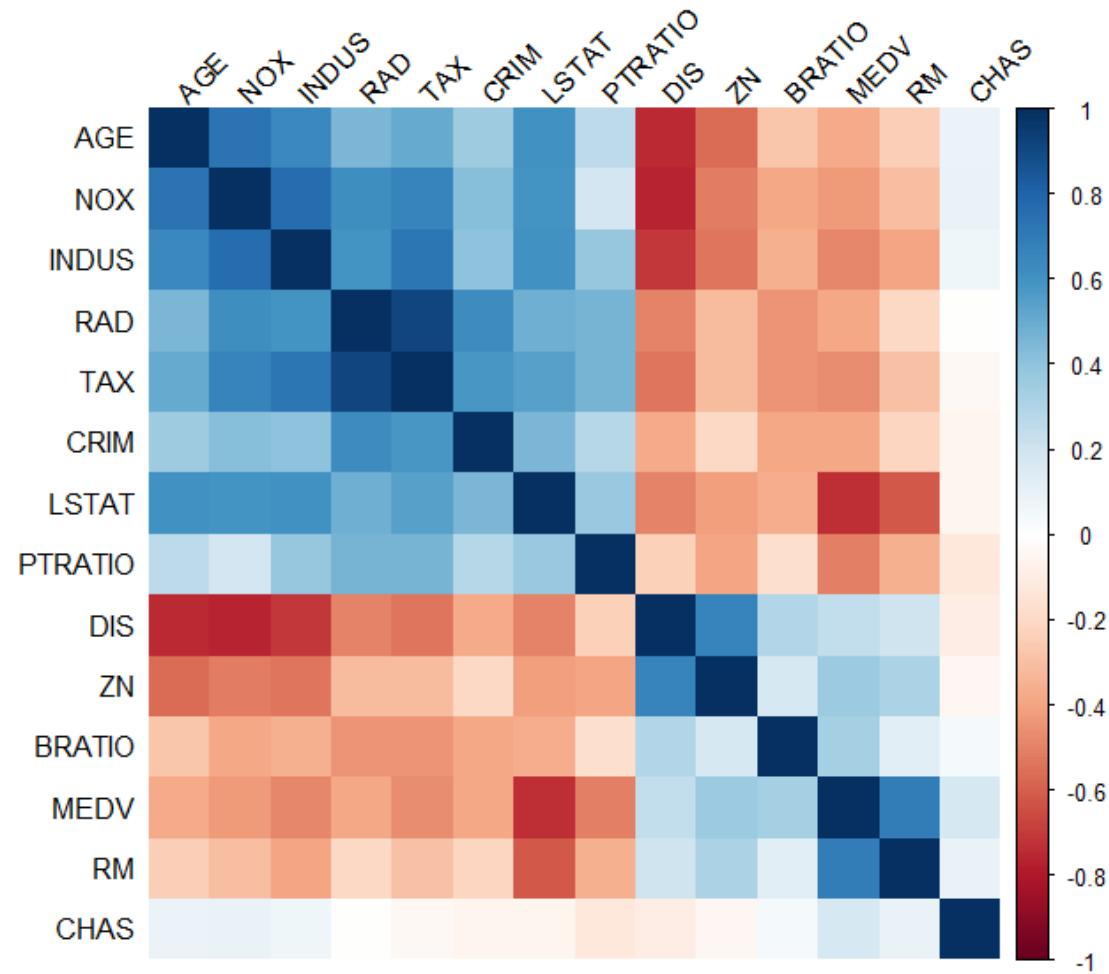


(b) Residue is ≥ 4

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

symmetric /symmetrical 对称的

departure from symmetry 对称性的偏离

Dispersion of data 数据的离散程度

Discrete 离散的

discrete variable 离散变量

skewed distributions 偏态分布