

【统计简单学】

第二单元 常用统计量

授课教师：唐丽英 教授

新竹交通大学
工业工程与管理学系

第二单元 内容大纲

- 第一部份：连续型数据特征值之计算
 1. 集中趋势指标- 平均数、中位数与众数
 2. 分散趋势指标- 全距、变异数与标准差
 3. 偏态系数
 4. 峰度系数
- 第二部份：数据特征值之应用
 1. 经验法则
 2. 盒须图

第一部份：连续型数据特征值之计算

连续型数据特征值之计算

- 连续型原始数据(raw data)之特征值包括：
 1. 集中趋势 (**Central Tendency of Location**)指标
 2. 分散趋势 (**Dispersion**)指标
 3. 偏态 (**Skewness**)指标
 4. 峰度 (**Kurtosis**)指标

1.集中趋势指标

- 集中趋势
 - 「集中趋势」是指一组数据往其中央点位置集中的趋势。
- 常用的集中趋势指标
 - 平均数(mean)、中位数(median)、众数(mode)。

1) 平均数

- 群体平均数： $\mu = \frac{\sum x_i}{N}$
- 样本平均数： $\bar{X} = \frac{\sum x}{n}$

其中 N 表群体大小， n 表样本大小。

1.集中趋势指标

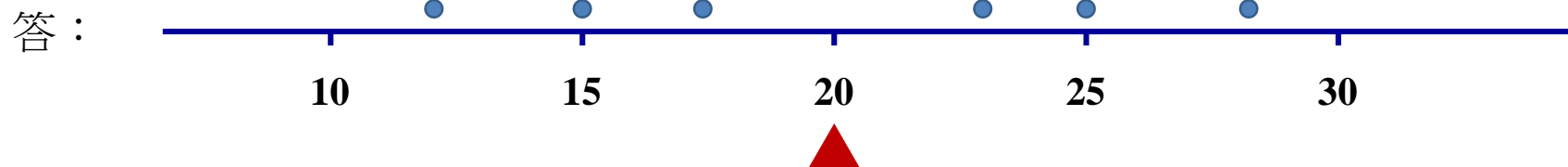
- **例1**：请找出下列群体数据之平均数：**0, 7, 3, 9, -2, 4, 6**。

答：
$$\mu = \frac{0+7+3+9-2+4+6}{7} = 3.857$$

- **例2**：请找出下列样本数据之平均数：**25, 12, 23, 28, 17, 15**。

答：
$$\bar{X} = \frac{25+12+23+28+17+15}{6} = 20$$

- **例3**：将例2之资料绘入下面之点图中，其平均数即为数据之「平衡点」。



1.集中趋势指标


2) 中位数

- 将一组数据由小至大排序后，位置在**最中间的数值**称为中位数。
 - 群体中位数： η \longrightarrow 读音：**eta**
 - 样本中位数： \tilde{x} \longrightarrow 读音：**X tilde**
- 找中位数之方法：
 - ① 当 n =奇数， \tilde{x} = 排序第 $(n+1)/2$ 位之数值。
 - ② 当 n =偶数， \tilde{x} = 排序第 $(n/2)$ 位及第 $(n/2)+1$ 位的两数值之平均数。

1.集中趋势指标


- 例4：请找出下列样本数据之中位数：9, 2, 7, 11, 14。

答：将样本数据排序后：2, 7, 9, 11, 14


 $\tilde{X} = 9$

- 例5：请找出下列样本数据之中位数：9, 2, 7, 11, 14, 6。

答：将样本数据排序后：2, 6, 7, 9, 11, 14


 $\tilde{X} = \frac{7+9}{2} = 8$

1.集中趋势指标

3) 众数

— 在一组数据中，出现次数最多的数值称为众数。

- 例6：请找出下列样本数据之众数：**3, 3, 2, 1, 4, 2, 3**。

答：众数=3。

- 例7：请找出下列样本数据之众数：**3, 1, 4, 2**。

答：众数=无。

1.集中趋势指标

- 何时用平均数？何时用中位数或众数？
 - 平均数对离群值非常敏感，而中位数或众数则对离群值较不敏感。
 - 当资料中有离群值时，建议使用中位数或众数，否则，使用平均数。

1.集中趋势指标

- **例8**：请找出下列样本数据之平均数、中位数与众数：
1, 3, 4, 6, 6, 9, 13。

答： 平均数 $\bar{X} = \frac{1+3+4+6+6+9+13}{7} = 6$ ，中位数 = 6 ，众数 = 6

- **例9**：若此组数据最后一笔资料改成**70**：**1, 3, 4, 6, 6, 9, 70**，请重新计算平均数、中位数与众数。

答： 平均数 $\bar{X} = \frac{1+3+4+6+6+9+70}{7} = 14.14$ ，中位数 = 6 ，众数 = 6

1.集中趋势指标

- **例10**：设有甲、乙两个学校大学生，其每月花费如下表（以新台币计）。请问哪一个学校的学生有较高的月花费？

	甲	乙
平均数	\$4,750	\$5,450
中位数	\$4,450	\$3,200

- 答：甲校的学生有较高之月花费。因其平均数和中位数非常接近，表示了该校有许多学生之月花费至少超过该校学生月花费之平均数或中位数。然而，乙校有一半的学生月花费是低于\$3,200元，而只有少数学生之月花费非常高。

2.分散趋势指标

- 分散趋势
 - 是表示一组数据分散的趋势。
- 常用的分散趋势指标
 - 全距 (Range)
 - 变异数(Variance)
 - 标准差 (Standard Deviation)

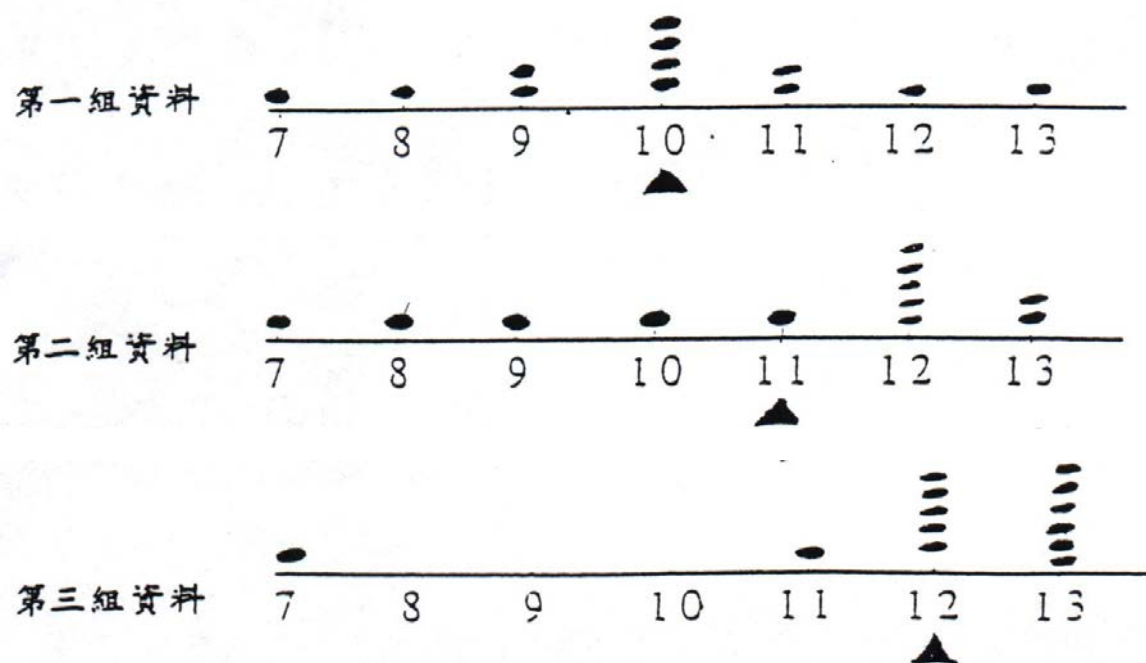
2.分散趋势指标

1) 全距

- 全距是用来衡量一组数据分散程度最简单的指标。
- 公式： $R = \text{最大值} - \text{最小值}$
- 用全距之缺点
 - 当一组数据中出现 离群值 或样本数很大时，全距并非一个很好的衡量数据分散程度的指标，因其无法解释最小值与最大值之间数据分散的情形。

2.分散趋势指标

- 例11：以下三组数据有相同之全距，不同之分布。



2.分散趋势指标

2) 变异数

- 群体变异数： $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N}$

- 样本变异数： $S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$

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

$$= \frac{\sum_{i=1}^n x_i^2 - 2\left(\frac{\sum_{i=1}^n x_i}{n}\right)\sum_{i=1}^n x_i + n\left(\frac{\sum_{i=1}^n x_i}{n}\right)^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - 2\frac{(\sum_{i=1}^n x_i)^2}{n} + \frac{(\sum_{i=1}^n x_i)^2}{n}}{n-1} = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n-1} = \frac{\text{平方和} - \frac{(\text{和的平方})}{\text{数据总和}}}{\text{数据总和} - 1}$$

计算用公式

2.分散趋势指标

3) 标准偏差

- 群体标准差： $\sigma = \sqrt{\sigma^2}$
- 样本标准差： $s = \sqrt{s^2}$

2.分散趋势指标

- 例12：请找出下列样本数据之平均数、变异数及标准差：
5, 8, 1, 2, 4。

x_i	5	8	1	2	4	$\sum_{i=1}^5 x_i = 20$
x_i^2	25	64	1	4	16	$\sum_{i=1}^5 x_i^2 = 110$

样本平均数 $\bar{X} = \sum_{i=1}^5 x_i / 5 = 20 / 5 = 4$

样本变异数 $S^2 = \frac{\sum_{i=1}^5 x_i^2 - (\sum_{i=1}^5 x_i)^2 / 5}{5-1} = \frac{110 - (20^2) / 5}{5-1} = 7.5$

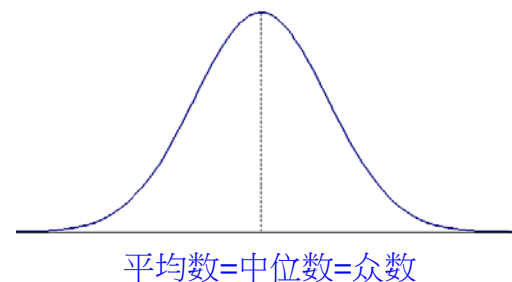
样本标准偏差 $S\sqrt{S^2} = \sqrt{7.5}$
=

3. 偏态系数

- 偏态
 - 是用来说明一组数据分布的形态。

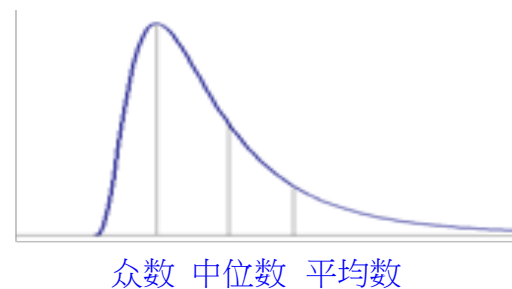
- 单峰分布有三种形态之偏态

1) 对称：平均数 ≡ 中位数



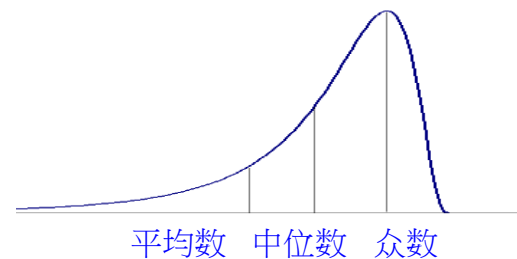
对称

2) 右偏，正偏：平均数 > 中位数



右偏

3) 左偏，负偏：平均数 < 中位数



左偏

3. 偏态系数

- 样本偏态系数之公式如下：

$$g_1 = \frac{M_3}{\left(S \sqrt{(n-1)/n} \right)^3} \quad \text{其中, } M_3 = \left[\sum_{i=1}^n (x_i - \bar{x})^3 \right] / n$$

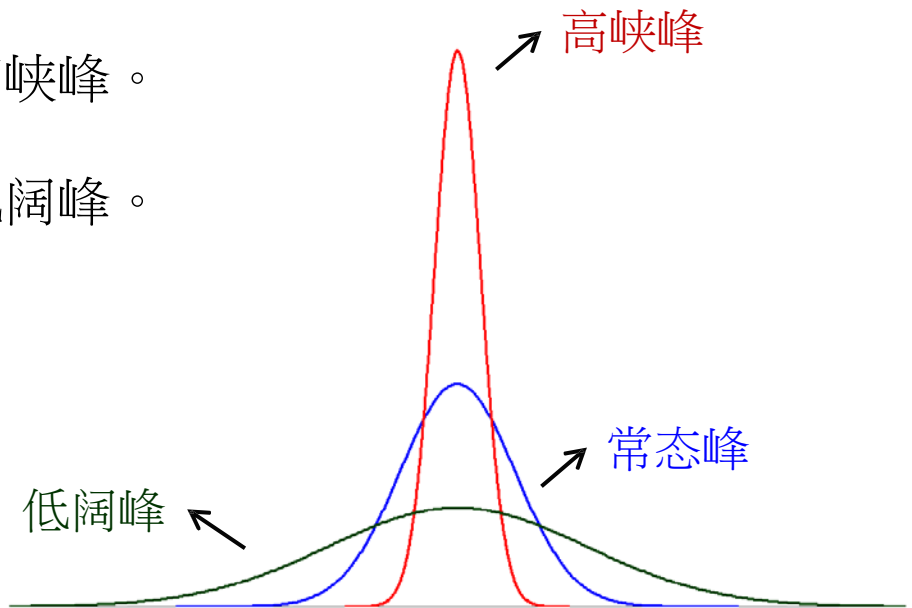
- 1) 偏态系数 = 0，表示样本分布呈对称。
- 2) 偏态系数 > 0，表示样本分布呈右偏。
- 3) 偏态系数 < 0，表示样本分布呈左偏。

4. 峰度系数

- 样本峰度系数之公式如下：

$$g_2 = \frac{M_4}{\left(s\sqrt{(n-1)/n}\right)^4} - 3 \quad \text{其中, } M_4 = \left[\sum_{i=1}^n (x_i - \bar{x})^4 \right] / n$$

- 1) 峰度系数 = 0，表示数据呈常态峰。
- 2) 峰度系数 > 0，表示资料呈高峡峰。
- 3) 峰度系数 < 0，表示资料呈低阔峰。



第一部份：连续型数据特征值之计算

【统计软件范例】

统计软件范例

例

- 某品牌手机经销站负责人为了要了解该手机在大台北地区销售的远景，特别检查最近六十天的销售记录，得销售量如下表所示

23	60	79	32	57	70	52	70	82	36
80	77	81	95	41	65	92	85	55	76
52	10	64	75	78	25	80	98	81	67
41	71	83	54	64	72	88	62	74	43
60	78	89	76	84	48	84	90	15	79
34	67	17	82	69	74	63	80	85	61

试求该品牌手机在大台北地区销售量之平均数、中位数、众数、全距、变异数、标准偏差、偏态系数及峰度系数。

统计软件范例 – Excel 报表

例

The screenshot displays the Microsoft Excel interface. The 'Data' ribbon is active, showing options for '取得外部資料' (Get External Data), '連線' (Connections), and '排序與篩選' (Sort & Filter). Below the ribbon, a spreadsheet is visible with the following data in column A:

	A	B	C	D	E	F	G	H
1	銷售量							
2	23							
3	80							
4	52							
5	41							
6	60							
7	34							
8	60							
9	77							
10	10							

The 'Data Analysis' task pane is open, showing a list of analysis tools. The tool '敘述統計' (Descriptive Statistics) is selected. The task pane includes buttons for '確定' (OK), '取消' (Cancel), and '說明(H)' (Help).

统计软体范例 – Excel 报表

例

	A	B
1	銷售量	
2		
3	平均數	65.41667
4	標準誤	2.725615
5	中間值	70.5
6	眾數	80
7	標準差	21.11252
8	變異數	445.7387
9	峰度	0.237671
10	偏態	-0.94704
11	範圍	88
12	最小值	10
13	最大值	98
14	總和	3925
15	個數	60

→ 中位数

→ 全距

统计软体范例 – Minitab 报表

例

The screenshot shows the Minitab software interface. The main window displays a worksheet titled 'Minitab - Untitled - [Worksheet 1 ***]' with a menu bar (File, Edit, Data, Calc, Stat, Graph, Editor, Tools, Window, Help) and a toolbar. The worksheet contains a column labeled 'C1' with the header '销售量' (Sales Volume) and 16 rows of data. Two dialog boxes are open: 'Display Descriptive Statistics' and 'Descriptive Statistics - Statistics'.

	C1
	销售量
1	23
2	80
3	52
4	41
5	60
6	34
7	60
8	77
9	10
10	71
11	78
12	67
13	79
14	81
15	64
16	83

Display Descriptive Statistics

Variables: 销售量

Descriptive Statistics - Statistics

- ☒ Mean
- ☐ SE of mean
- ☒ Standard deviation
- ☐ Variance
- ☐ Coefficient of variation
- ☒ First quartile
- ☒ Median
- ☒ Third quartile
- ☐ Interquartile range
- ☒ Mode
- ☐ Trimmed mean
- ☐ Sum
- ☒ Minimum
- ☒ Maximum
- ☒ Range
- ☐ Sum of squares
- ☒ Skewness
- ☒ Kurtosis
- ☐ MSSD
- ☐ N nonmissing
- ☐ N missing
- ☒ N total
- ☐ Cumulative N
- ☐ Percent
- ☐ Cumulative percent

Buttons: Help, OK, Cancel

统计软体范例 – Minitab 报表

例

Minitab 报表

Descriptive Statistics: 销售量

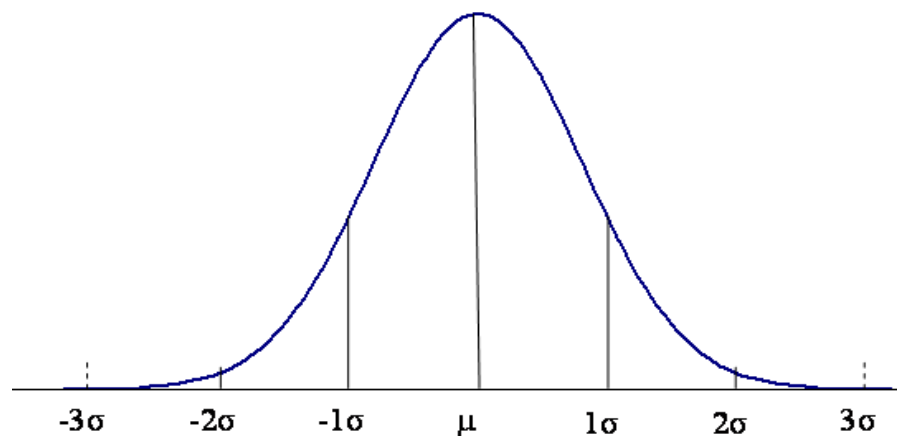
	样本数	平均数	标准差	变异数	最小值	中位数	最大值	众数	N for
Variable	N	Mean	StDev	Variance	Minimum	Median	Maximum	Mode	Mode
销售量	60	65.42	21.11	445.74	10.00	70.50	98.00	80	3

	偏态系数	峰度系数
Variable	Skewness	Kurtosis
销售量	-0.95	0.24

第二部份：数据特征值之应用

经验法则（又称**68%—95%—99.73%**法则）

- 利用经验法则（**The Empirical Rule**）可以决定数据分布之情形。
- 经验法则：若数据呈钟形分布，则约有
 - 1) 68.26% 的数据在 $\mu \pm \sigma$ 范围内
 - 2) 95.44% 的数据在 $\mu \pm 2\sigma$ 范围内
 - 3) 99.73% 的数据在 $\mu \pm 3\sigma$ 范围内



经验法则

- **例13**：一家半导体厂经理想要研究员工完成某项制程的时间。经理于是随机挑选了**40**位员工作测试，得到平均数**12.8**分钟与标准差**1.7**分钟。请以经验法则来描述此样本资料。

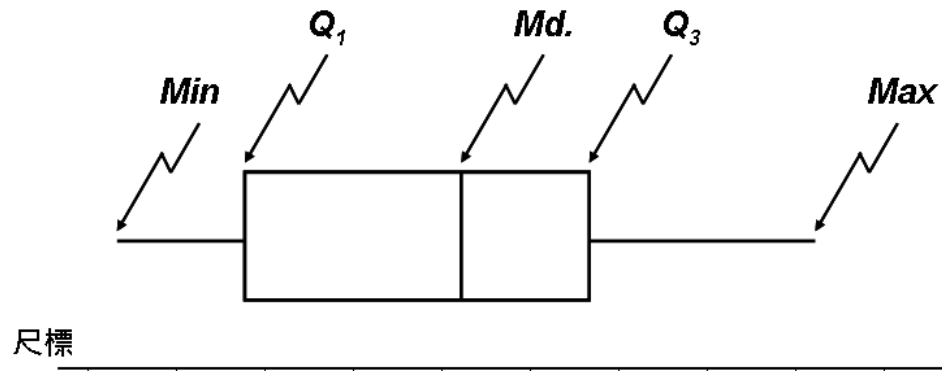
【解】： $n=40$, $\bar{X} = 12.8$, $S=1.7$

- 大约有68.26%的员工完成该制程的时间是介于 **$12.8 \pm 1.7 = (11.1, 14.5)$** 分钟。
- 大约有95.44%的员工完成该制程的时间是介于 **$12.8 \pm 2 \times 1.7 = (9.4, 16.2)$** 分钟。
- 大约有99.73%的员工完成该制程的时间是介于 **$12.8 \pm 3 \times 1.7 = (7.7, 17.9)$** 分钟。

盒须图

- 盒须图（**Box-Whisker Plot**）

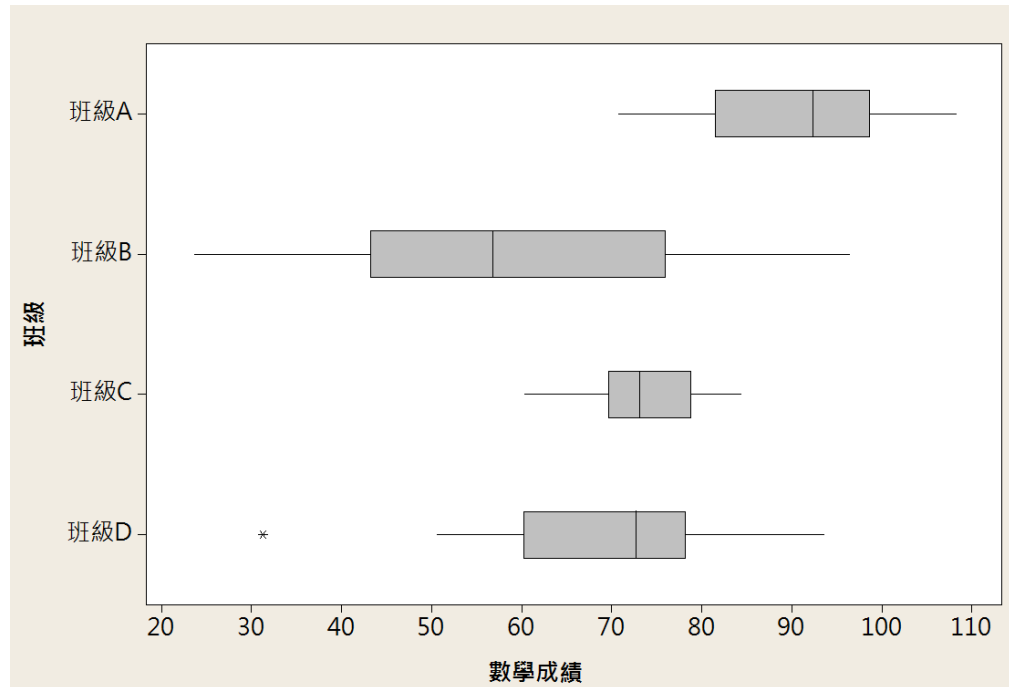
- 是资料的一种图形展示法。此图可同时显示资料之集中趋势、分散趋势、偏态、最小值、最大值等。此图又称「五指标摘要图」(five-number summary plot)



- Q_1 ：第一四分位数或第25百分位数。
- Q_2 ：第二四分位数或中位数(Md.)。
- Q_3 ：第三四分位数或第75百分位数。

盒须图

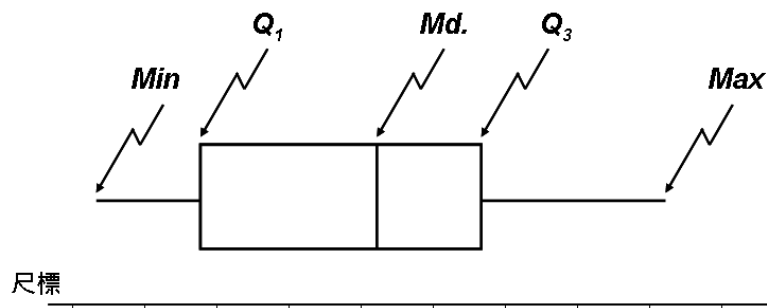
- 盒须图之主要功用
 - 1) 可有效的找出资料之主要特征值。
 - 2) 可同时比较数组资料。



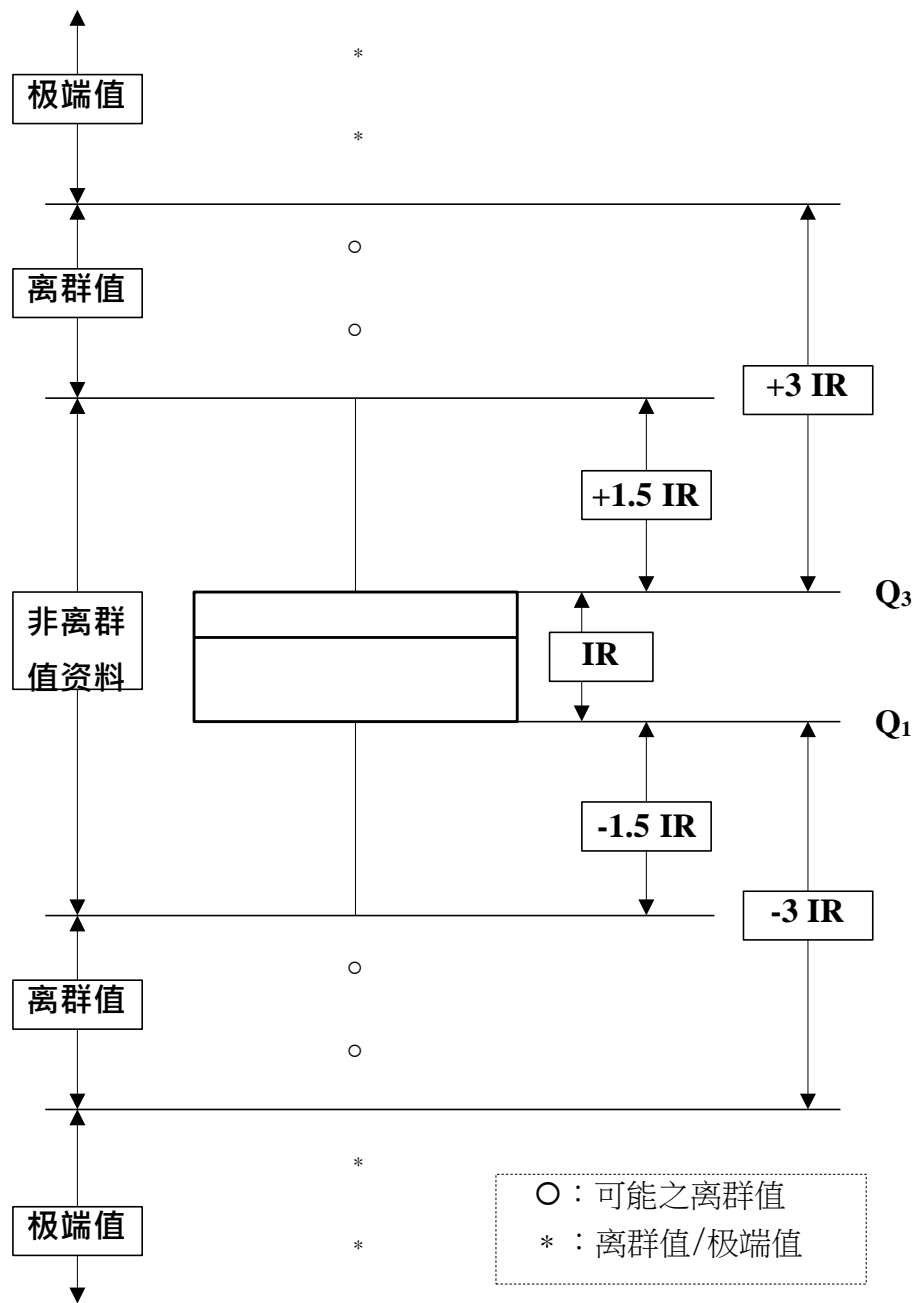
- 3) 可办认出离群值。

盒须图

- 何谓离群值（**Outliers**）？
 - 离群值是远大于或远小于同一笔数据中之其它值之数据。
- 如何利用盒须图辨认出离群值？
 - 1) 超过盒须图之盒 $1.5 \times (Q_3 - Q_1)$ 至 $3 \times (Q_3 - Q_1)$ 距离内之值可当作「可能之离群值」。
 - 2) 超过盒须图之盒 $3 \times (Q_3 - Q_1)$ 距离外之值可当作「非常可能之离群值（或极端值）」。



注：中四分位距 (Interquartile Range, IR) = $Q_3 - Q_1$ = 第75百分位数 - 第25百分位数



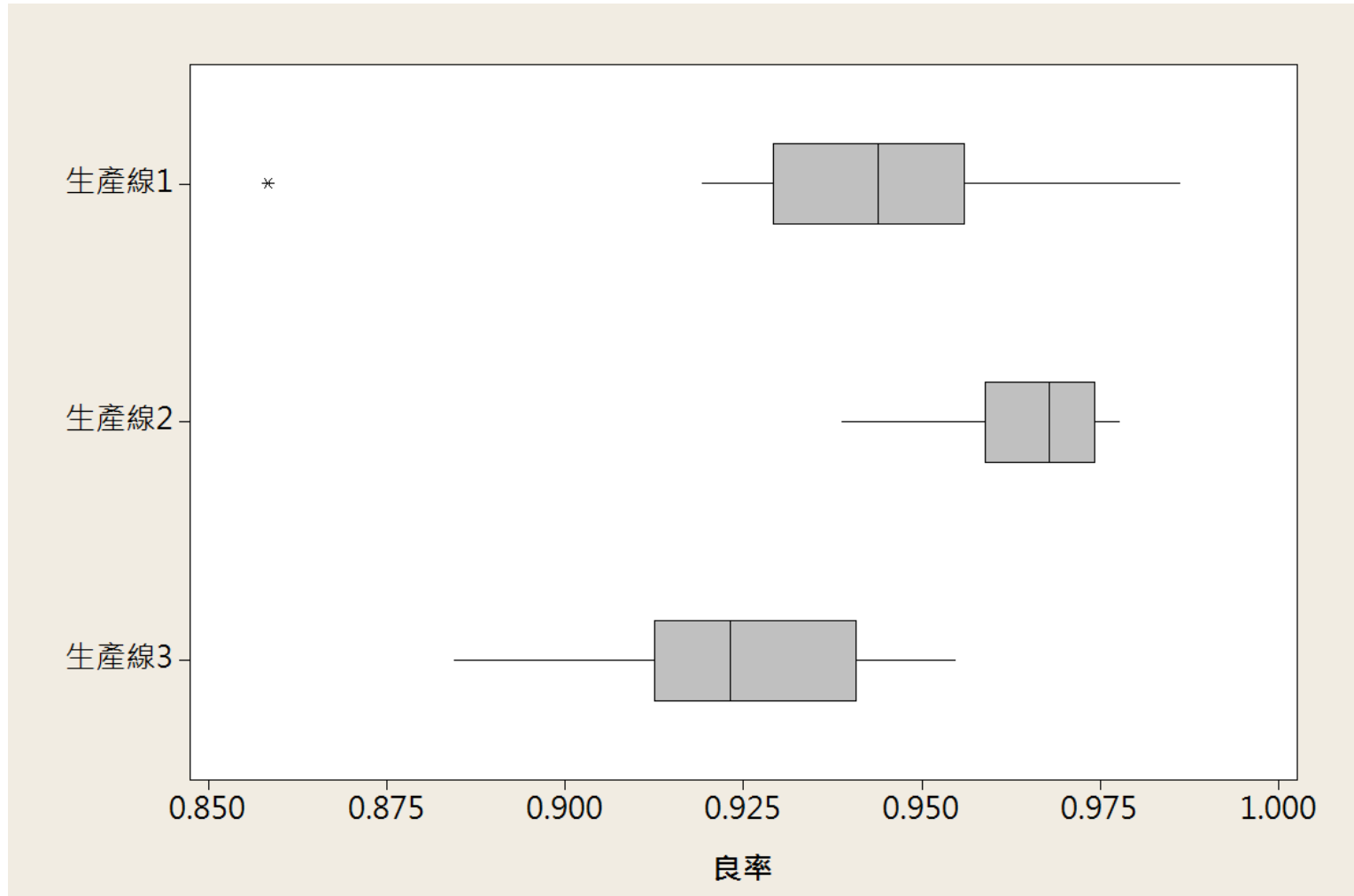
盒须图

- **例14**：下列资料为三条生产线的良率，请依资料绘制盒须图。

生产线1	生产线2	生产线3
0.99	0.98	0.91
0.86	0.97	0.91
0.95	0.94	0.92
0.94	0.97	0.93
0.93	0.97	0.91
0.96	0.96	0.94
0.95	0.94	0.92
0.92	0.97	0.95
0.96	0.98	0.88
0.94	0.97	0.95

盒须图

- 例14：三条生产线良率之盒须图绘制如下。



本单元结束

第二单元 简单回顾

简单回顾

- 连续型数据之特征值：
 - 集中趋势指标
 - 平均数、中位数及众数
 - 分散趋势指标
 - 全距、变异数及标准差
 - 偏态系数
 - 峰度系数
- 数据特征值之应用：
 - 经验法则
 - 盒须图