第 2 章　描述性统计—统计数据的整理

- 2.1　统计数据的整理
- 2.2　分布集中趋势的测度
- 2.3　分布离散程度的测度
- 2.4　分布偏态与峰态的测度

2.1 统计数据的整理

统计数据分组--组距分组

- 将变量值的一个区间作为一组
- 适合于连续变量
- 适合于变量值较多的情况
- 需要遵循“不重不漏”的原则
- 可采用等距分组，也可采用不等距分组

组距分组
（步骤）

1. 确定组数：组数的确定应以能够显示数据的分布特征和规律为目的
2. 确定组距：组距(Class Width)是一个组的上限与下限之差，可根据全部数据的最大值和最小值及所分的组数来确定，即
   - 组距＝(最大值 - 最小值)÷ 组数
3. 统计出各组的频数并整理成频数分布表

组距分组

- 1. 下限(low limit)：一个组的最小值
- 2. 上限(upper limit)：一个组的最大值
- 3. 组距(class width)：上限与下限之差
- 4. 组中值(class midpoint)：下限与上限之间的中点值

$$\text{组中值} = \frac{\text{下限值}+\text{上限值}}{2}$$

直方图
(histogram)

1. 用矩形的宽度和高度来表示频数分布的图形，实际上是用矩形的*面积*来表示各组的频数分布
2. 在直角坐标中，用横轴表示数据分组，纵轴表示频数或频率，各组与相应的频数就形成了一个矩形，即直方图
3. 直方图下的总面积等于1

【例】某班50名学生概率考试成绩如下：

75 65 80 81 92 63 77 79 54 98

85 72 66 84 83 60 82 78 64 90

81 78 76 86 68 76 73 71 88 87

65 57 46 89 78 66 87 79 84 78

96 88 67 38 67 75 83 82 68 85

试编出次数分配表并作出其直方图（**histogram**）.

由于 $x_{(n)}-x_{(1)}+1=98-38+1=61$ 取 63

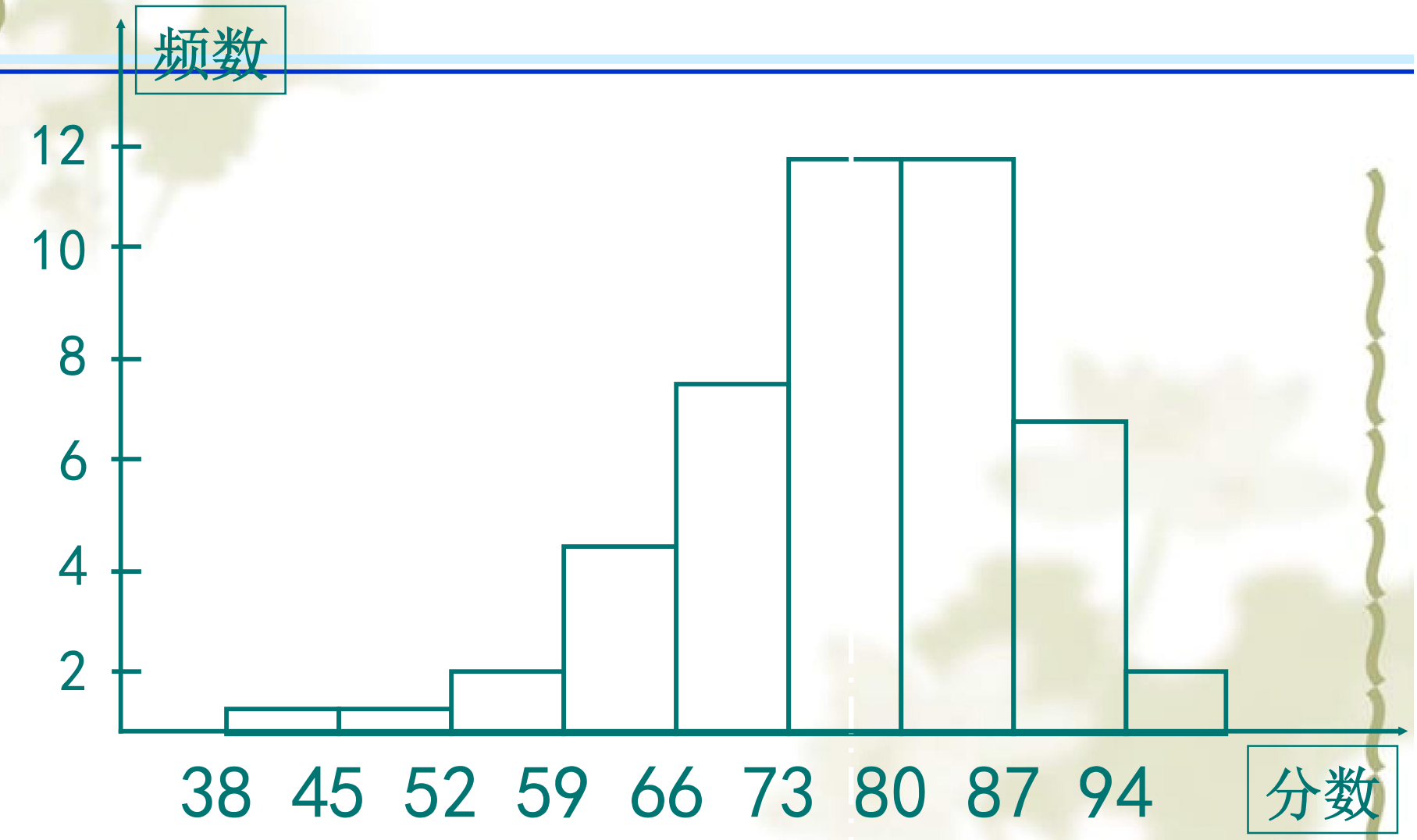
故分 9 组，每组组距为7，于是可得

次 数
分布表 ——→ 次数
直方图

成绩次数分布表

组号	组限	组中值	频数	分组和
1	38~44	41	1	41
2	45~51	48	1	48
3	52~58	55	2	110
4	59~65	62	5	310
5	66~72	69	8	552
6	73~79	76	12	912
7	80~86	83	12	996
8	87~93	90	7	630
9	94~99	97	2	194
合计		75.86	50	3793
平均值				

次数直方图

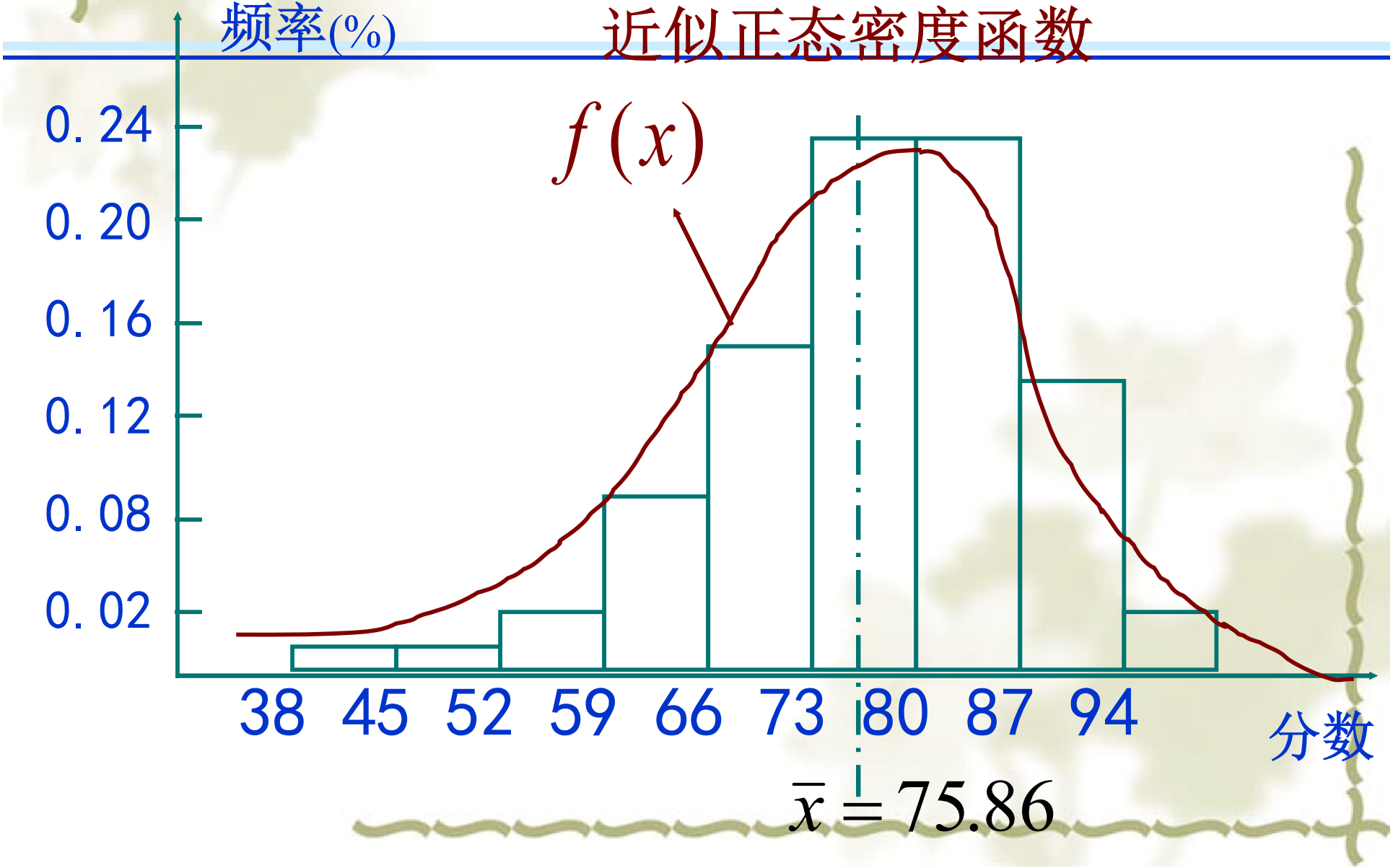


$\bar{x} = 75.86$

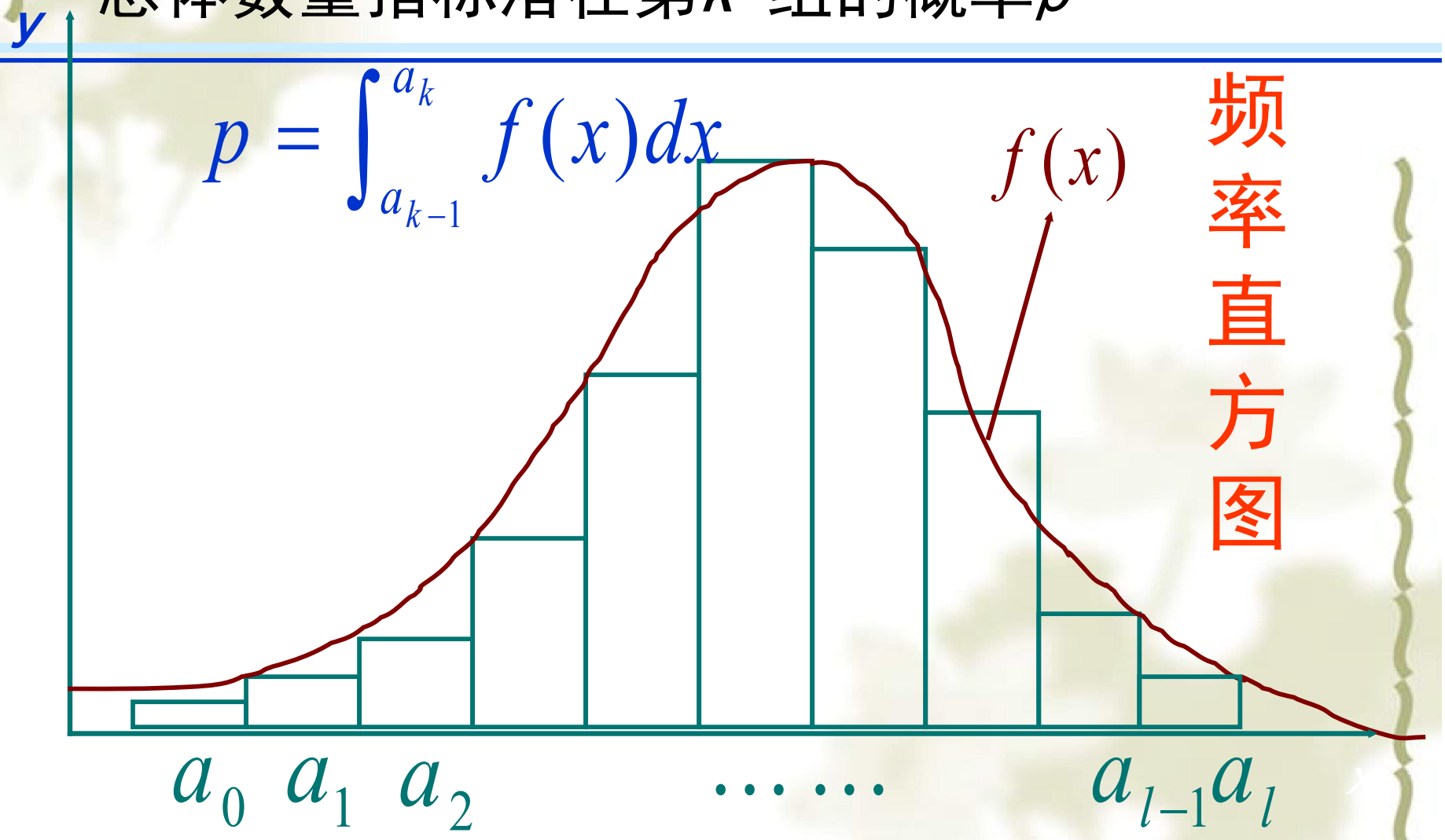
频率分配表

组号	组限	组中值	频数	频率	累积频率
1	38~44	41	1	0.02	0.02
2	45~51	48	1	0.02	0.04
3	52~58	55	2	0.04	0.08
4	59~65	62	5	0.1	0.18
5	66~72	69	8	0.16	0.34
6	73~79	76	12	0.24	0.58
7	80~86	83	12	0.24	0.82
8	87~93	90	7	0.14	0.96
9	94~99	97	2	0.04	1
合计			50	1	

频率直方图


频率(%)
近似正态密度函数
$f(x)$
0. 24
0. 20
0. 16
0. 12
0. 08
0. 02
38 45 52 59 66 73 80 87 94
分数
$\bar{x}=75.86$

总体数量指标落在第k 组的概率p



2.2 分布集中趋势的测度

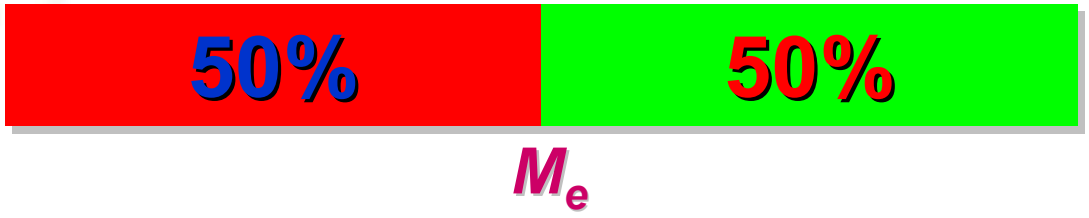
- 一、众数
- 二、中位数
- 三、分位数
- 四、均值
- 五、几何平均数
- 六、切尾均值
- 七、众数、中位数和均值的比较

众数
(mode)

1. 一组数据中出现次数最多的变量值
2. 适合于数据量较多时使用
3. 不受极端值的影响
4. 一组数据可能没有众数或有几个众数
5. 主要用于分类数据，也可用于顺序数据和数值型数据

中位数
(median)

1. 排序后处于中间位置上的值



2. 不受极端值的影响
3. 主要用于顺序数据，也可用数值型数据，但不能用于分类数据
4. 各变量值与中位数的离差绝对值之和最小，即

$$\sum_{i=1}^{n} \left| x_i - M_e \right| = \min$$

中位数
(位置的确定)

原始数据进行排序：

$$\text{中位数位置} = \frac{n+1}{2}$$

分组数据：

$$M_e \approx L + \frac{\frac{N}{2} - S_{m-1}}{f_m} \times i$$

$N/2$中位数位置L中位数所在组胡下组限

S_{m-1}中位数所在组以下各组的累积次数

f_m中位数所在组的次数

i中位数所在组的组距

数值型数据的中位数
(奇数个数据)

- 【例】 9个家庭的人均月收入数据
- 原始数据: 1500 750 780 1080 850 960 2000 1250 1630
- 排　　序: 750 780 850 960 1080 1250 1500 1630 2000
- 位　　置: 1 2 3 4 5 6 7 8 9

$$位置 = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

中位数 = 1080

数值型数据的中位数
(偶数个数据)

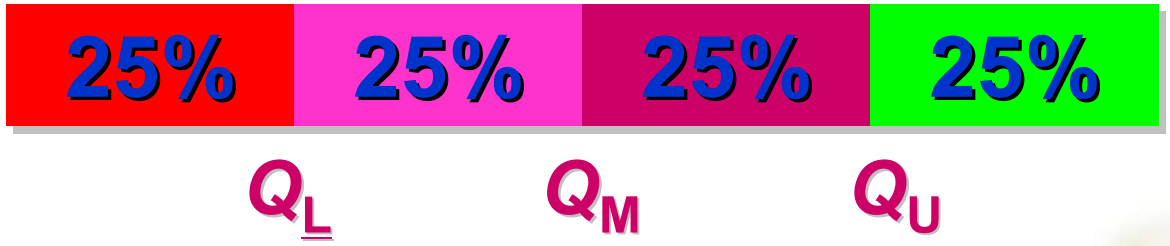
- 【例】：10个家庭的人均月收入数据
- 排　　序：660 750 780 850 960 1080 1250 1500 1630 2000
- 位　　置：1 2 3 4 5 6 7 8 9 10

$$位置=\frac{n+1}{2}=\frac{10+1}{2}=5.5$$

$$中位数=\frac{960+1080}{2}=1020$$

四分位数(quartile)

1. 排序后处于25%和75%位置上的值



2. 不受极端值的影响

3. 主要用于顺序数据，也可用于数值型数据，但不能用于分类数据

四分位数(位置的确定)

原始数据排序为顺序数据：

$$\begin{cases} Q_L \text{位置} = \dfrac{n+1}{4} \\ Q_U \text{位置} = \dfrac{3(n+1)}{4} \end{cases}$$