大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

北京理工大学计算机学院　孙新

2019年1月

- **峰度 (kurtosis)也叫峰态**，统计学家Pearson于1905年首次提出，用于数据分布扁平程度的测度

- 峰态通常是与标准正态分布相比较而言的。
  - 如果以正态分布作为标准，不同分布的数据在均值附近的集中程度也不同。
  - 有的分布可能会显得“平坦”一些，有更多的数据分布在两侧。
  - 有的分布则看起来比较“尖锐”，数据更多地集中在均值附近。

1. 根据原始数据计算

$$K = \frac{n(n+1)\sum (x_i - \bar{x})^4 - 3\left[\sum (x_i - \bar{x})^2\right]^2 (n-1)}{(n-1)(n-2)(n-3)s^4}$$

2. 根据分组数据计算

$$K = \frac{\sum_{i=1}^{k} (M_i - \bar{x})^4 f_i}{ns^4} - 3$$

- 峰度是用来度量分布形状的指标。为正数意味着有更多的数据分布在两侧极端，为负意味着数据较多地集中在均值附近。

如果一组数据服从标准正态分布，则峰态系数的值等于0；若峰态系数的值明显不等于0，则表明分布比正态分布更平或更尖，通常称为平峰分布或尖峰分布

峰态系数=0扁平峰度适中
峰态系数<0为扁平分布或者平峰分布
峰态系数>0为尖峰分布

- 例题：峰态系数

某电脑公司销售量偏态及峰度计算表

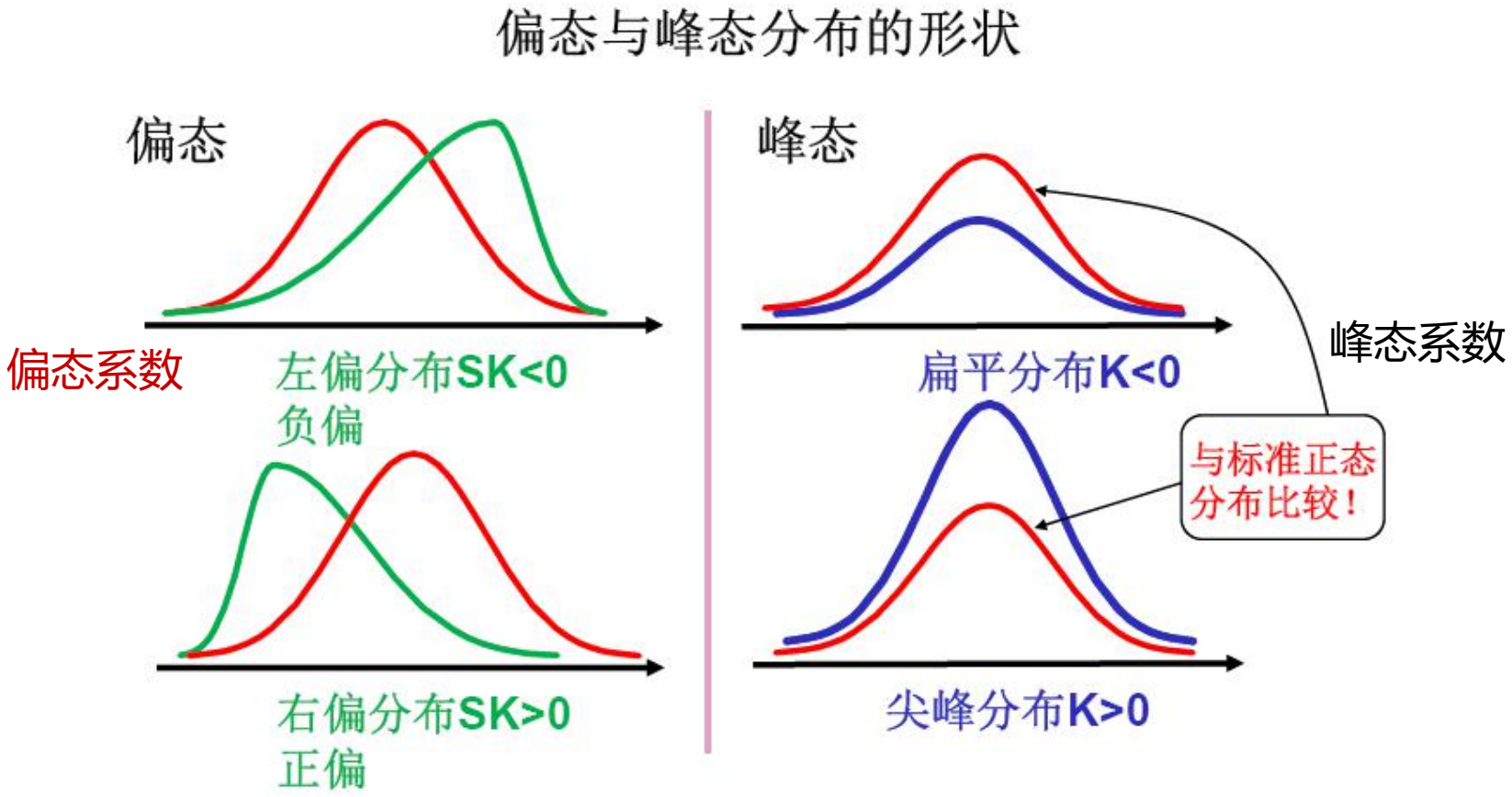
按销售量份组(台)	组中值(M_i)	频数 f_i	$(M_i-\bar{x})^3 f_i$	$(M_i-\bar{x})^4 f_i$
140 ~ 150	145	4	-256000	10240000
150 ~ 160	155	9	-243000	7290000
160 ~ 170	165	16	-128000	2560000
170 ~ 180	175	27	-27000	270000
180 ~190	185	20	0	0
190 ~200	195	17	17000	170000
200 ~210	205	10	80000	1600000
210 ~220	215	8	216000	6480000
220 ~ 230	225	4	256000	10240000
230 ~ 240	235	5	625000	31250000
合计	—	120	540000	70100000

- 例题：峰态系数

$$K = \frac{\sum_{i=1}^{k}(M_i - \bar{x})^4 f_i}{ns^4} - 3 = \frac{70100000}{120 \times (21.58)^4} - 3$$

$$= 2.694 - 3 = -0.306$$

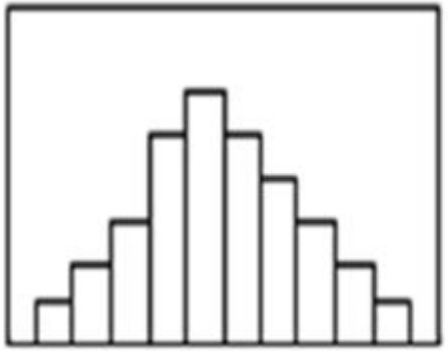
- 结论：偏态系数为负值，但与0的差异不大，说明电脑销售量为轻微扁平分布


偏态与峰态分布的形状
偏态
偏态系数
左偏分布SK<0
负偏
右偏分布SK>0
正偏
峰态
峰态系数
扁平分布K<0
与标准正态分布比较！
尖峰分布K>0

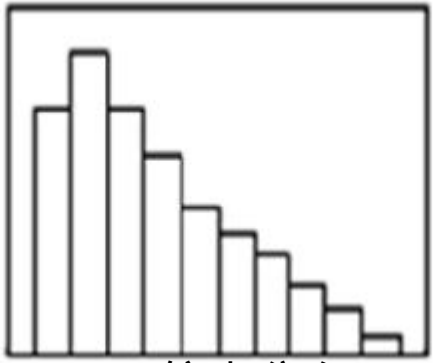
2.4 图形化分析方法-直方图

- 对于有限的数据，可以通过频率分布直方图来观察数据的分布，直方图是频数直方图的简称。
  - 它是用一系列宽度相等、高度不等的长方形表示数据的图。长方形的宽度表示数据范围的间隔，长方形的高度表示在给定间隔内的数据数。
- 作用：显示数据的分布特征，
- 直方图的适用场合：
  - 1. 数据是数值型时；
  - 2. 想弄清楚数据分布的形状；
  - 3. 确定一个过程的输出是否近乎符合正态分布

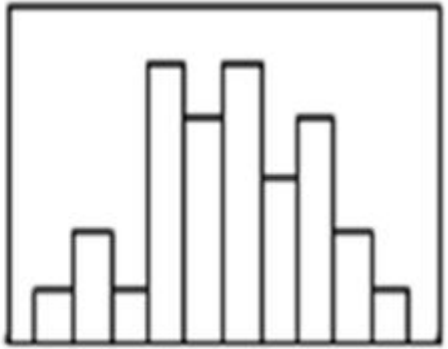
- 几种典型直方图形状



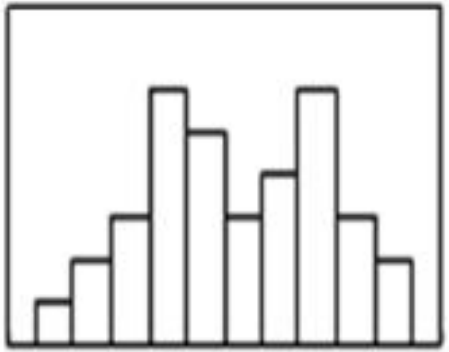
a）正态分布



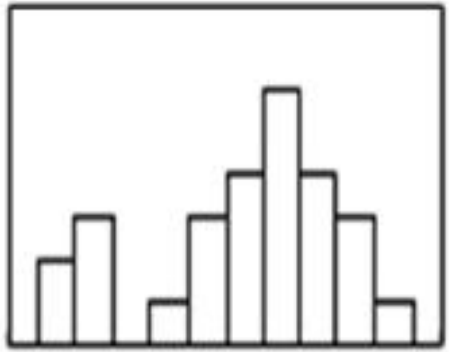
b）偏态分布



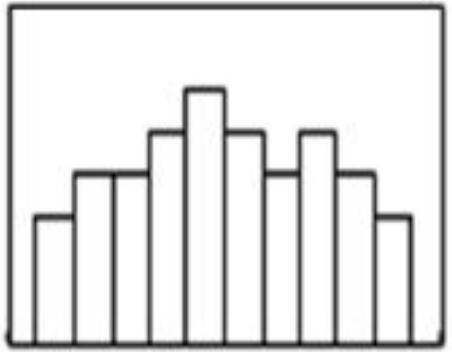
c）梳状/锯齿分布



d）双峰分布



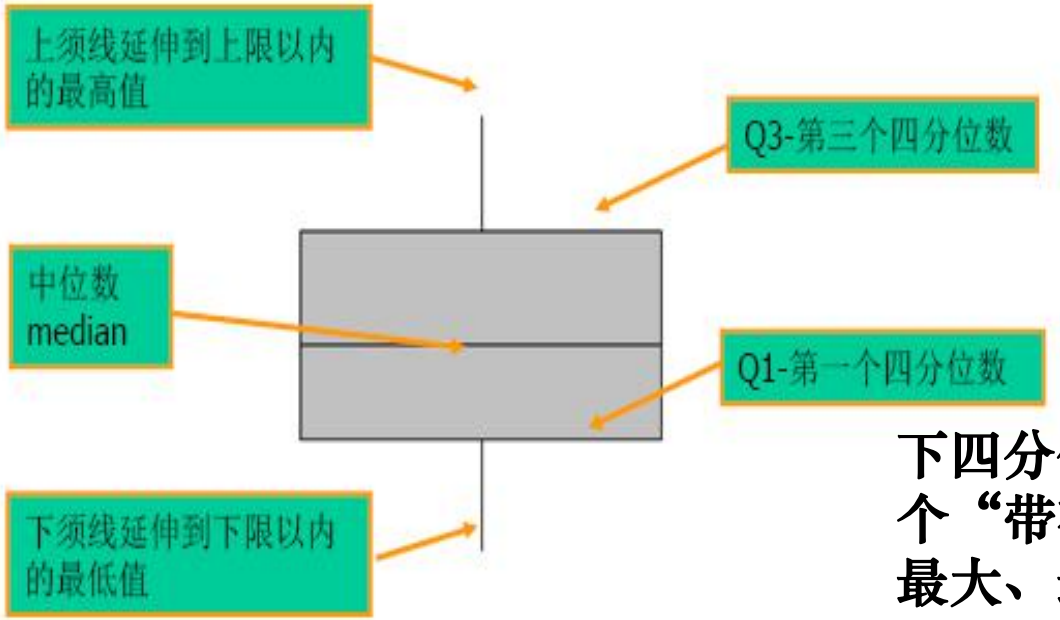
e）孤岛型分布



f）高原/平顶分布

2.4 图形化分析方法-箱形图

- 箱形图（Box-plot）又称为盒须图、盒式图或箱线图，是一种用作显示一组数据分散情况资料的统计图。

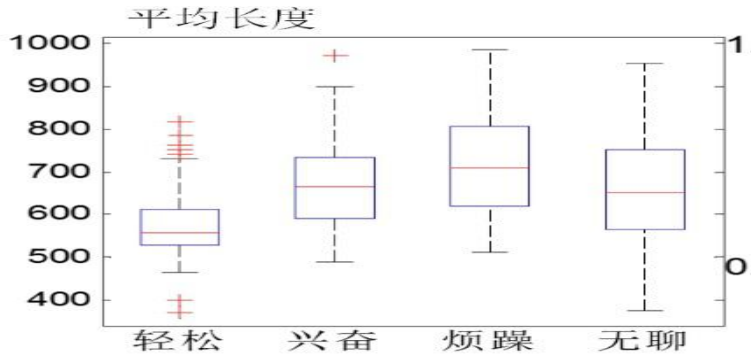


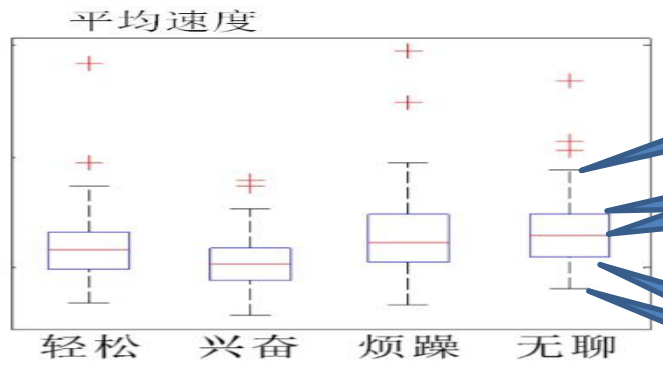
由五个数值点组成：最小值(*min*)，下四分位数(*Q1*)，中位数(*median*)，上四分位数(*Q3*)，最大值(*max*)

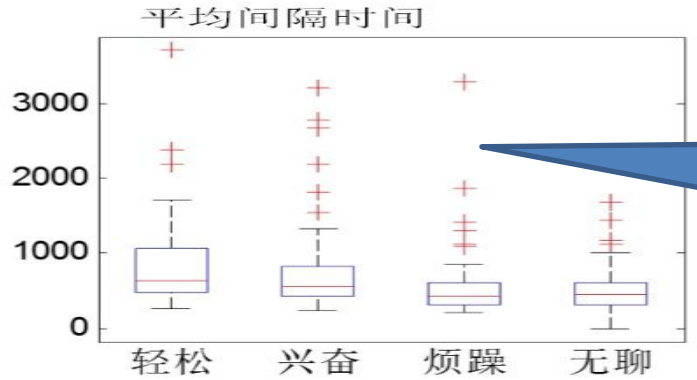
下四分位数、中位数、上四分位数组成一个“带有隔间的盒子” 上、下四分位数到最大、最小值之间建立一条延伸线

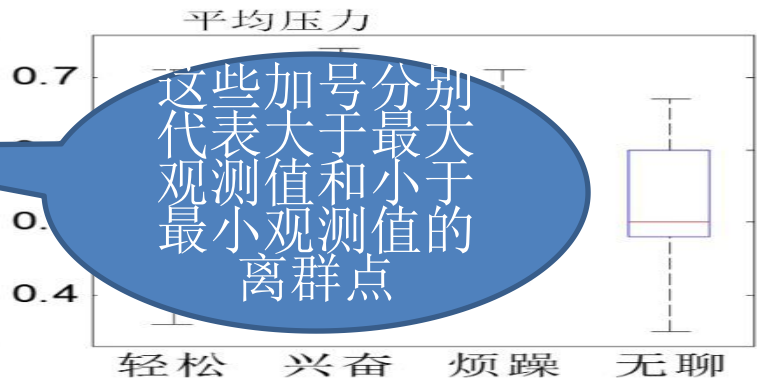
2.4 图形化分析方法-箱形图

2、统计数据分析方法









最大观测值

第三四分位数

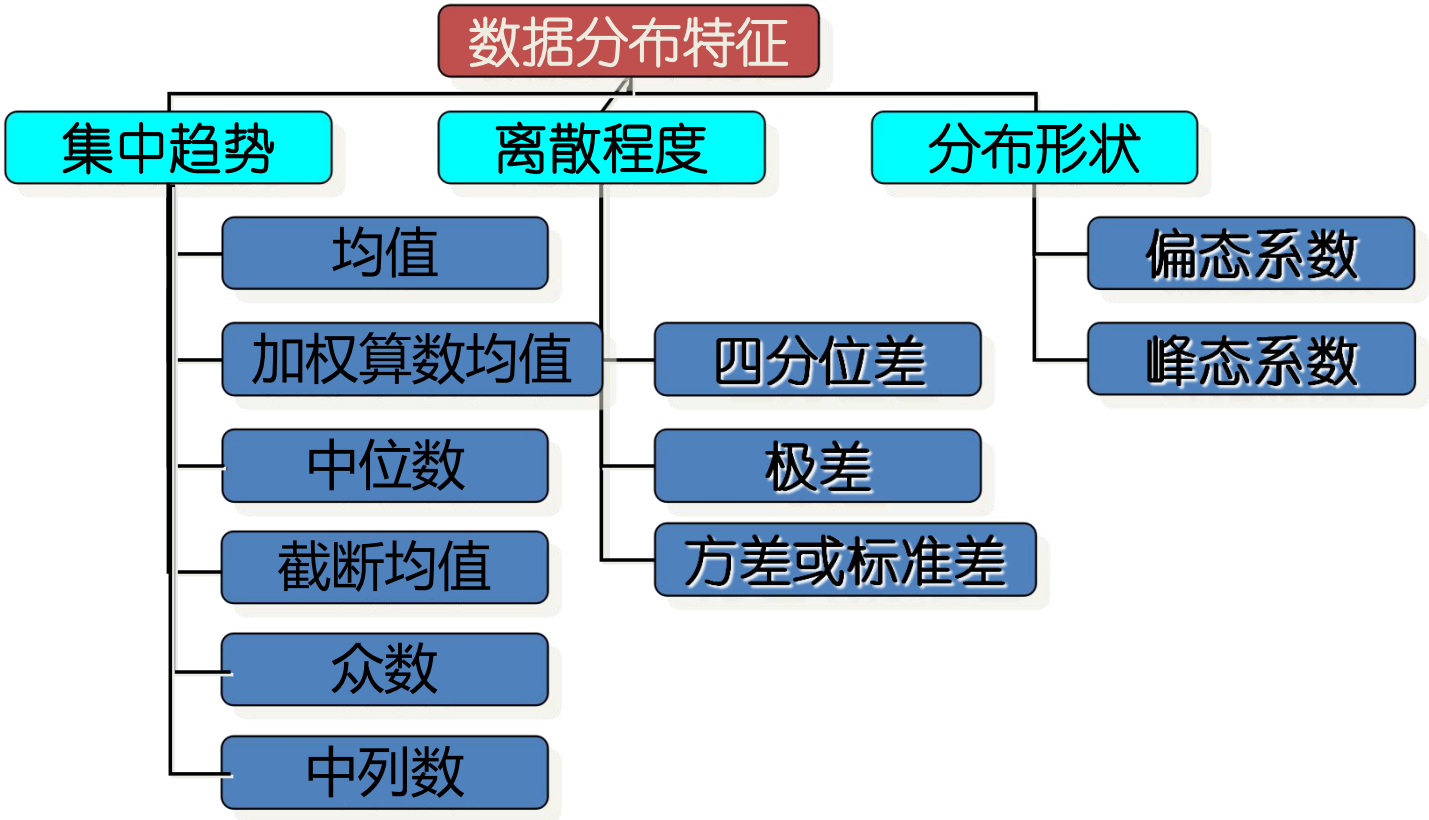
第一四分

最小观测值

这些加号分别代表大于最大观测值和小于最小观测值的离群点

小结：数据分布特征和描述统计量

2、统计数据分析方法



谢 谢

Thank you for your attention!