第 4 章　数据的概括性度量

内容提要

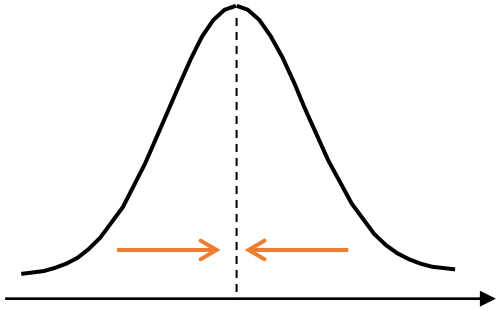
数据的3个分布特征

集中趋势

数据向其中心值靠拢或聚集的程度

度量

众数、中位数和四分位数、平均数

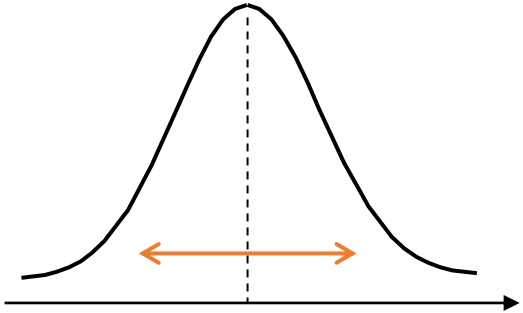


离散趋势

数据远离其中心值的趋势

度量

异众比率、四分位数差、方差和标准差、离散系数、标准分数

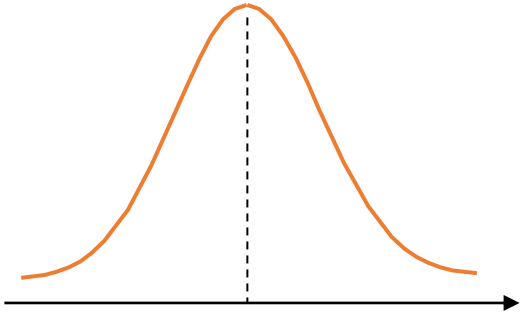


分布形状

数据分布的峰态与偏态

度量

偏度、峰度、J-B统计量



问题：上述3个特征描述的是哪种分布？为什么不是其他分布？

4.1 集中趋势的度量

- 众数
- 中位数和四分位数
- 平均数

众数

定义

众数(Mode)是一组数据中出现频数最多的数值。

表示： M_o

适合：无序分类数据

优点：不受极端值影响

说明

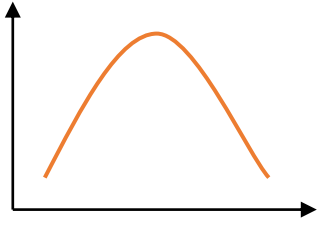
- 数据量较多时使用
- 一组数据可能没有众数，或者出现多个众数
- 也可用于有序分类和数值型数据

无众数

10 5 9 12 6 8

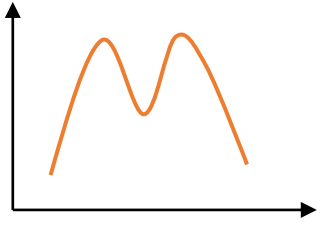
一个众数

6 5 9 8 5 5



多个众数

25 28 28 36 42 42



例子：无序分类数据的众数

不同品牌饮料的频数分布

饮料类型	频数	比例	百分比(%)
果汁	6	0.12	12
矿泉水	10	0.20	20
绿茶	11	0.22	22
其他	8	0.16	16
碳酸饮料	**15**	0.30	30
合计	50	1	100

这里的变量为“饮料类型”，属于**无序分类变量**，不同类型的饮料就是变量值

所调查的50人中，购买碳酸饮料的人数最多，为15人，占总被调查人数的30%，因此饮料类型的众数为“碳酸饮料”，即

M_o ＝**碳酸饮料**

例子：有序分类数据的众数

某城市家庭对住房状况的评价

回答类别	户数（户）	百分比（%）
非常不满意	24	8
不满意	**108**	36
一般	93	31
满意	45	15
非常满意	30	10
合计	300	100.0

这里的数据为**有序分类数据**，变量为“回答类别”

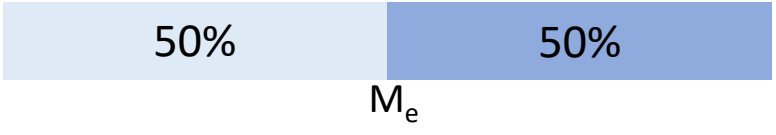
该城市中对住房表示不满意的户数最多，为108户，因此众数为“不满意”这一类别，即

$$M_o = \text{不满意}$$

中位数

定义

中位数(Median)是一组数据排序后处于中间位置的数值。



表示： M_e

适合：有序分类数据

优点：不受极端值影响

说明

- 常用于**收入分配**的研究
- 也可用于数值型数据

位置确定

$$\text{中位数位置} = \frac{n+1}{2}$$

数值确定

$$M_e = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n\text{为奇数} \\ \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\} & n\text{为偶数} \end{cases}$$

例子：有序分类数据的中位数

某城市家庭对住房状况的评价

回答类别	户数（户）	累计频数
非常不满意	24	24
不满意	108	132
一般	**93**	**225**
满意	45	270
非常满意	30	300
合计	300	-

中位数的位置为

(300+1)/2 = 150.5

从累计频数看，中位数在“一般”这一组别中

中位数为

M_e=一般

例子：数值型数据的中位数

例1：**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

排序与位置

750	780	850	960	**1080**	1250	1500	1630	2000
1	2	3	4	**5**	6	7	8	9

$$\text{中位数位置} = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

$$M_{\mathrm{e}}=1080$$

例2：**10个**家庭的人均月收入数据

原始数据

660 780 750 1080 850 960 2000 1250 1630 1500

排序与位置

660	750	780	850	**960**	**1080**	1250	1500	1630	2000
1	2	3	4	**5**	**6**	7	8	9	10

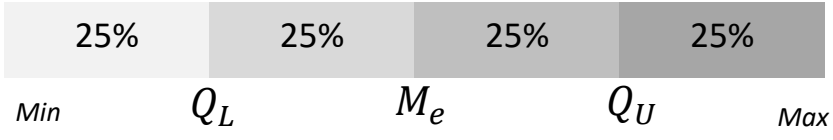
$$\text{中位数位置} = \frac{n+1}{2} = \frac{10+1}{2} = 5.5$$

$$M_{\mathrm{e}} = \frac{960+1080}{2} = 1020$$

四分位数

定义

四分位数(Quartile)是指一组数据排序后处于25%和75%位置的数值。



表示

- 下四分位数：Q_L 或 $Q_{25\%}$
- 上四分位数： Q_U或 $Q_{75\%}$

适合：有序分类数据

优点：不受极端值影响

位置确定

- **通常**
$$\begin{cases} Q_{\mathrm{L}}\text{位置} = \dfrac{n}{4} \\ Q_{\mathrm{U}}\text{位置} = \dfrac{3n}{4} \end{cases}$$
- **Excel**
$$\begin{cases} Q_{\mathrm{L}}\text{位置} = \dfrac{n+3}{4} \\ Q_{\mathrm{U}}\text{位置} = \dfrac{3n+1}{4} \end{cases}$$
- **SPSS**
$$\begin{cases} Q_{\mathrm{L}}\text{位置} = \dfrac{n+1}{4} \\ Q_{\mathrm{U}}\text{位置} = \dfrac{3(n+1)}{4} \end{cases}$$

数值确定：依据位置，比例分摊

例子：有序分类数据的四分位数

某城市家庭对住房状况的评价

回答类别	户数（户）	累计频数
非常不满意	24	24
不满意	**108**	**132**
一般	**93**	**225**
满意	45	270
非常满意	30	300
合计	300	-

Q_L位置= (300)/4 =75

Q_U位置 =(3×300)/4=225

从累计频数看，Q_L在“不满意”这一组别中；Q_U在“一般”这一组别中

四分位数为

Q_L = **不满意**

Q_U = **一般**

例子：数值型数据的四分位数

例1：9个家庭的人均月收入数据

原始数据	1500	750	780	1080	850	960	2000	1250	1630
排序后	750	**780**	**850**	960	1080	**1250**	**1500**	1630	2000
位置	1	**2**	**3**	4	5	**6**	**7**	8	9

Q_{L}位置 $= \frac{9}{4} = 2.25$

Q_{U}位置 $= \frac{3 \times 9}{4} = 6.75$

$Q_{\mathrm{L}} = 780 + (850 - 780) \times \mathbf{0.25}$
$= 797.5$

$Q_{\mathrm{U}} = 1250 + (1500 - 1250) \times \mathbf{0.75}$
$= 1437.5$

平均数

定义

平均数(Mean)，也称均值、算数平均数，它是一组数据相加后除以数据个数得到的数值。

重要性

集中趋势最常用的测度值

分类

- 简单平均数、加权平均数
- 总体平均数(μ)、样本平均数($\bar{x}$)

适合：数值型数据

缺点：容易受极端值影响

简单平均数 vs 加权平均数

简单平均数

某**未分组数据**为：x_1，x_2，…，x_n (总体数据x_N)

- **样本**平均数

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **总体**平均数

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

加权平均数

某**分组数据**

各组的组中值为：M_1，M_2，…，M_k

各组的频数为：f_1，f_2，…，f_k

- **样本**加权平均

$$\bar{x} = \frac{M_1 f_1 + M_2 f_2 + \cdots + M_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} M_i f_i}{n}$$

- **总体**加权平均

$$\mu = \frac{M_1 f_1 + M_2 f_2 + \cdots + M_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} M_i f_i}{N}$$

例子：加权平均数

某电脑公司销售量数据分组表

按销售量分组	组中值(M_i)	频数(f_i)	$M_i f_i$
140~150	145	4	580
150~160	155	9	1395
160~170	165	16	2640
170~180	175	27	4725
180~190	185	20	3700
190~200	195	17	3315
200~210	205	10	2050
210~220	215	8	1720
220~230	225	4	900
230~240	235	5	1175
合计	—	120	22200

$$\bar{x} = \frac{\sum_{i=1}^{k} M_i f_i}{n} = \frac{22200}{120} = 185$$

几何平均数

定义

几何平均数(Geometric mean)，是指n 个变量值乘积的 n 次方根。

表示： G_{m}

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

适合： 比率数据的平均

说明

- 常用于**平均增长率的计算**
- 特殊类型的**算术**平均数

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

几何平均数 vs 算术平均数

例子： 一位投资者购持有一种股票，连续4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**

几何平均数

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

$$= 8.0787\%$$

算术平均数

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

比较：众数、中位数和平均数

众数

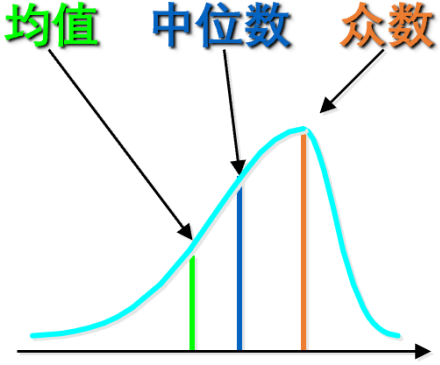
- 不受极端值影响
- 具有不惟一性
- 适合识别**峰值**

中位数

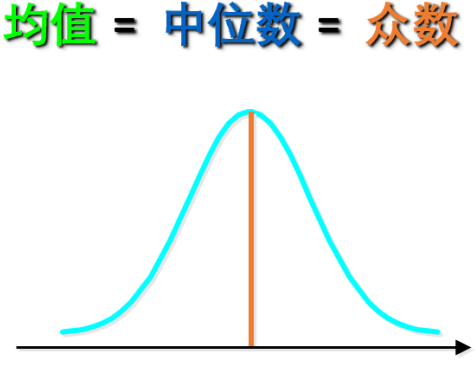
- 不受极端值影响
- 数据分布**偏斜**程度较大时应用

平均数

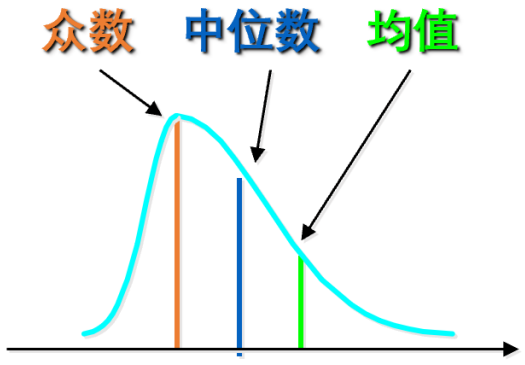
- 易受极端值影响
- 数据**对称分布**或接近对称分布时应用



左偏分布



对称分布



右偏分布

4.2 离散程度的度量

- 异众比率
- 四分位差
- 方差和标准差
- 离散系数
- 标准分数与经验法则

异众比率

定义

异众比率(Variation ratio)，是指非众数组的频数占总频数的比例。

表示：V_r

公式

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i}$$

应用：无序分类数据的离散程度测度

说明：

- 衡量**众数的代表性**
  - 异众比率越小，众数代表性越强，离散程度越小
  - 异众比率越大，众数代表性越弱，离散程度越大

例子：异众比率

不同品牌饮料的频数分布

饮料品牌	频数	比例	百分比(%)
果汁	6	0.12	12
矿泉水	10	0.20	20
绿茶	11	0.22	22
其他	8	0.16	16
碳酸饮料	**15**	0.30	30
合计	50	1	100

$$V_r = \frac{50 - 15}{50}$$

$$= 70\%$$

在所调查的50人当中，购买其他品牌饮料的人数占70%，异众比率比较大。

因此，用“碳酸饮料”代表消费者购买饮料品牌的状况，其代表性不是很好

四分位差

定义

四分位差(Quartile deviation)，也称四分位距、内距（Inter-quartile ranger），是一组数据75%位置上的四分位数与25%位置上的四分位数之差。

表示：*IQR*

公式

$$IQR=Q_U-Q_L$$

应用：有序分类数据的离散程度测度

说明：

- 不受极端值的影响
- 反映中间50%数据的离散程度
- 数值越大，离散程度越大

例子：四分位差

某城市家庭对住房状况的评价

回答类别	户数（户）	累计频数
非常不满意	24	24
不满意	**108**	**132**
一般	**93**	**225**
满意	45	270
非常满意	30	300
合计	300	-

假设：

非常不满意=1,不满意=2，一般=3, 满意=4, 非常满意=5

已知

Q_L = 不满意 = 2

Q_U = 一般 = 3

四分位差为

$IQR = Q_U - Q_L$

= 3 – 2 = 1

极差

定义

极差(Range)，也称全距，一组数据的最大值与最小值之差。

表示：*R*

公式

$$R = \max(x_i) - \min(x_i)$$

应用：数值型数据的离散程度测度

说明：

- 离散程度的最简单测度
- 容易受到极端值的影响

平均差

定义

平均差(Mean deviation)，也称平均绝对离差，是指各变量值与其平均数离差绝对值的平均数。

表示：M_d

公式

- 未分组数据

$$M_{\mathrm{d}} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- 分组数据

$$M_{\mathrm{d}} = \frac{\sum_{i=1}^{k} |M_i - \bar{x}| f_i}{n}$$

应用：**数值型数据**的离散程度测度

说明：数学性质较差，实际中应用较少

方差和标准差

定义

- 方差(Variance)：离差平方的平均数。
- 标准差(Standard deviation)：方差开方后的结果。

表示

- 总体方差和标准差：σ^2，σ
- 样本方差和标准差：s^2，s

应用：**数值型数据**的离散程度测度

说明：

- 方差(标准差)是应用**最广泛**的离散程度测度统计量
- 标准差具有**量纲**(与原始数据相同)

方差和标准差：公式

总体

- 未分组数据
  - 方差 $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
  - 标准差 $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$
- 分组数据
  - 方差 $\sigma^2 = \frac{\sum_{i=1}^{K}(M_i - \mu)^2 f_i}{N}$
  - 标准差 $\sigma = \sqrt{\frac{\sum_{i=1}^{K}(M_i - \mu)^2 f_i}{N}}$

样本

- **未分组数据**
  - **方差** $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - **标准差** $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- 分组数据
  - 方差 $s^2 = \frac{\sum_{i=1}^{k}(M_i - \bar{x})^2 f_i}{n-1}$
  - 标准差 $s = \sqrt{\frac{\sum_{i=1}^{k}(M_i - \bar{x})^2 f_i}{n-1}}$

问题：为什么样本方差和标准差除以n-1，而不是n？

自由度

定义

自由度(Degree of freedom)，是指数据个数与附加给独立的观测值的约束或限制的个数之差。

内涵：一组数据中可以**自由取值的个数**

例子

某样本有3个数值：x_1=2，x_2=4，x_3=9。当 $\bar{x}$ = 5 确定后，x_1，x_2和x_3有两个数据可以自由取值，另一个则不能自由取值，比如x_1=6，x_2=7，那么x_3则必然取2，而不能取其他值。

样本方差自由度的解释

- 样本均值的约束
- 无偏估计量

例子：样本标准差

某电脑公司销售量数据平均差计算表

按销售量分组	组中值(M_i)	频数(f_i)	$(M_i-\bar{x})^2$	$(M_i-\bar{x})^2 f_i$
140~150	145	4	40	160
150 ~ 160	155	9	30	270
160 ~170	165	16	20	320
170 ~180	175	27	10	270
180 ~ 190	185	20	0	0
190 ~ 200	195	17	10	170
200 ~ 210	205	10	20	200
210 ~220	215	8	30	240
220 ~230	225	4	40	160
230 ~240	235	5	50	250
合计	—	120	—	55400

$$s=\sqrt{\frac{\sum_{i=1}^{k}(M_i-\bar{x})^2 f_i}{n-1}}$$

$$=\sqrt{\frac{55400}{120-1}}=21.58(\text{台})$$

含义：每一天的销售量与平均数相比，平均相差21.58台。

离散系数

定义

离散系数(coefficient of variation),也称**变异系数**，它是一组数据的标准差与其相应的均值之比。

表示：*CV*

公式

$$CV = \frac{s}{\bar{x}}$$

应用：数值型数据离散程度的**相对**测度

说明：

- **相对离散程度**的测度
- 离散系数越大，相对离散程度越大；离散系数越小，相对离散程度与小
- 适合不同样本的**离散程度比较**

例子：离散系数

某管理局所属8家企业的产品销售数据

企业编号	产品销售额（万元）x_1	销售利润（万元）x_2
1	170	8.1
2	220	12.5
3	390	18.0
4	430	22.0
5	480	26.5
6	650	40.0
7	950	64.0
8	1000	69.0

$\bar{x}_1 = 536.25$(万元)

$s_1 = 309.19$(万元)

$$CV_1 = \frac{309.19}{536.25} = 0.577$$

$\bar{x}_2 = 32.5215$(万元)

$s_2 = 23.09$(万元)

$$CV_2 = \frac{23.09}{32.5215} = 0.710$$

结论：计算结果表明$CV_1 < CV_2$，说明产品销售额的离散程度小于销售利润的离散程度。

标准分数

定义

标准分数(coefficient of variation),是指某个数据与其平均数的离差除以标准差后的值。

表示：z_i

公式

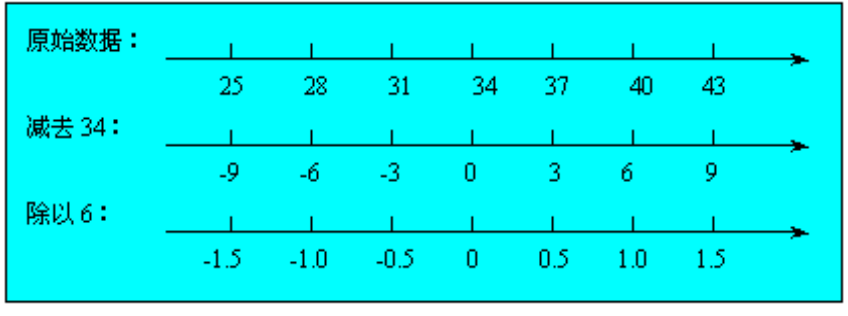
$$z_i = \frac{x_i - \bar{x}}{s}$$

应用

数值型数据的**相对位置的测度**

说明

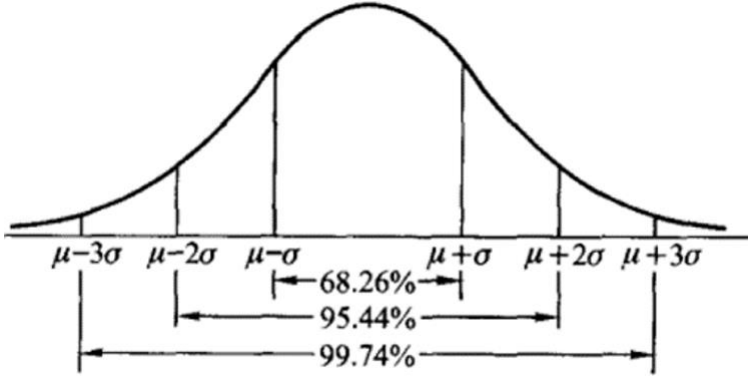
- 标准化会**消除量纲**，便于不同量纲的数据比较
- 标准化后的数据，容易判断**离群点(**outlier)，也称异常值
- 标准化是**线性变换**，不改变分布的形状



经验法则1：对称分布

对称分布

- 约有**68.26%**的数据在平均数加减**1个标准差**的范围之内;
- 约有**95.44%**的数据在平均数加减**2个标准差**的范围之内;
- 约有**99.74%**的数据在平均数加减**3个标准差**的范围之内;



定义

离群点(outlier)，也称异常值，是指**3个标准差之外**的数据点。

经验法则2：非对称分布

非对称分布

- 至少有**75%**的数据落在平均数加减**2个标准差**的范围之内；
- 至少有**89%**的数据落在平均数加减**3个标准差**的范围之内；
- 至少有**94%**的数据落在平均数加减**4个标准差**的范围之内。

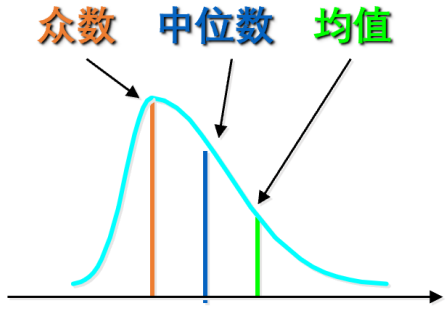
依据

如果一组数据是**非对称分布**，那么根据**切比雪夫不等式**，至少有$\left(1-\frac{1}{k^2}\right)$的数据落在平均数加减k个标准差的范围内。

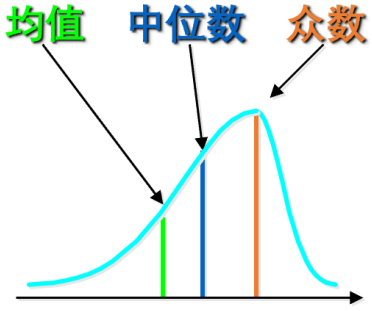
4.3 分布形状的度量

- 偏度
- 峰度
- Jarque-Bera统计量

偏度



右偏分布



左偏分布

定义

偏度(Skewness)，是指数据分布的不对称性。

测度

偏度系数(Coefficient of Skewness)

公式

$$SK = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

区间	含义	分布特点
S>1	高度偏态，右偏	右偏 （长的右拖尾）
$0.5 < S \leq 1$	中度偏态，右偏	
$0 < S \leq 0.5$	低度偏态，右偏	
S=0	无偏态	对称分布
$0 > S \geq -0.5$	低度偏态，左偏	左偏 （长的左拖尾）
$-0.5 > S \geq -1$	中度偏态，左偏	
S<-1	高度偏态，左偏	

峰度

定义

峰度(Kurtosis)，是指数据分布峰值的高低。

测度

峰度系数(Coefficient of Kurtosis)

公式

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

区间	分布特点	典型分布
K>0	尖峰分布 （数据分布更集中）	-
K=0	标准分布	标准正态分布
K<0	平峰分布 （数据分布更分散）	t 分布

Jarque-Bera统计量

公式

$$JB = \frac{S^2}{6/n} + \frac{(K-3)^2}{24/n}$$

其中n是观测数（或自由度）；S是样本偏度，K是样本峰度：

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}},$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{4/2}},$$

应用

分布服从**正态分布的综合测度**

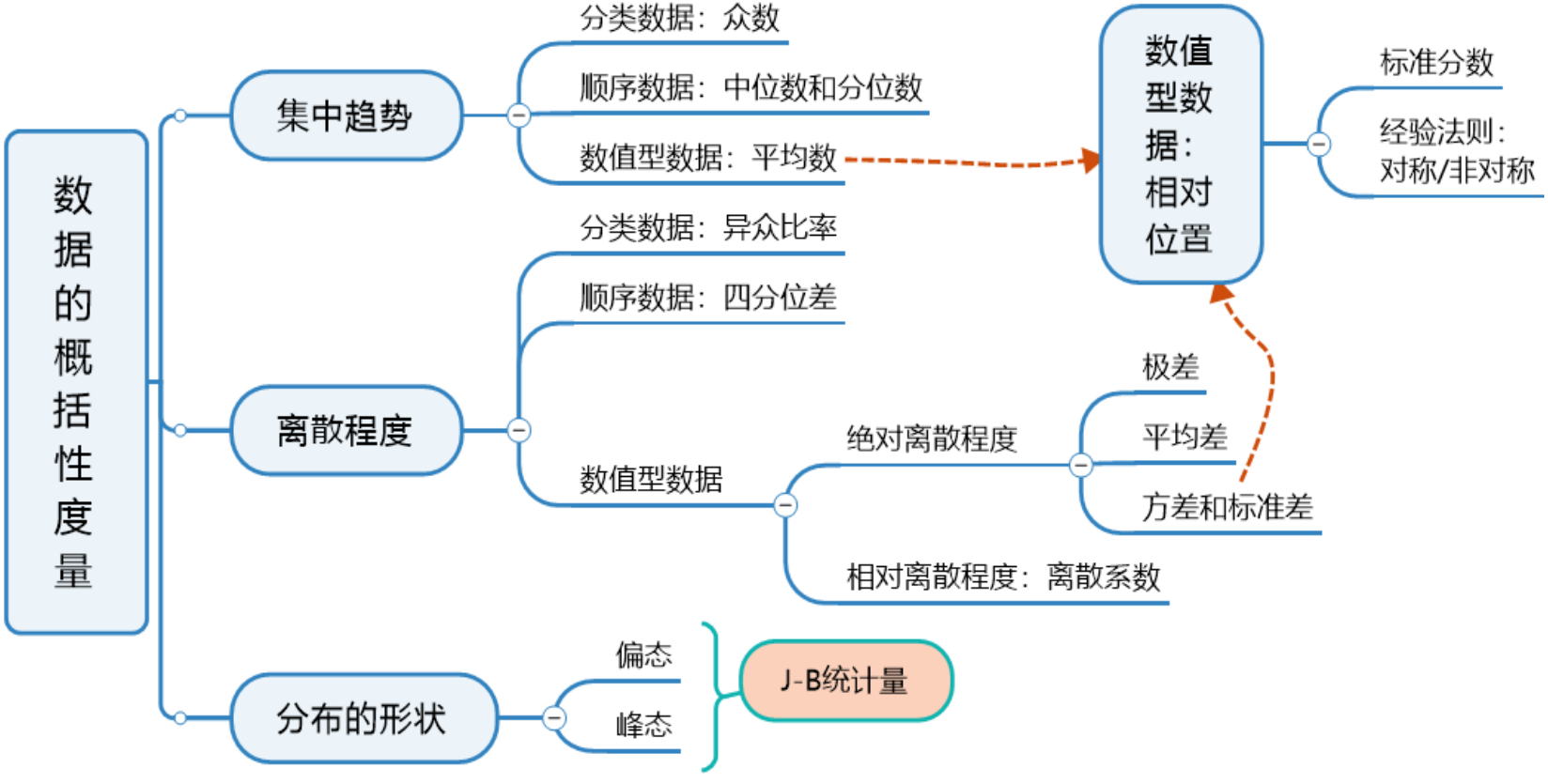
说明

- 正态分布：S=0, **K=3**

注意

这里的峰度公式与前面的公式写法不同，这里的K=3相当于前面的K=0。

本章小结


数据的概括性度量
集中趋势
分类数据：众数
顺序数据：中位数和分位数
数值型数据：平均数
离散程度
分类数据：异众比率
顺序数据：四分位差
数值型数据
绝对离散程度
极差
平均差
方差和标准差
相对离散程度：离散系数
数值型数据：相对位置
标准分数
经验法则：对称/非对称
分布的形状
偏态
峰态
J-B统计量