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

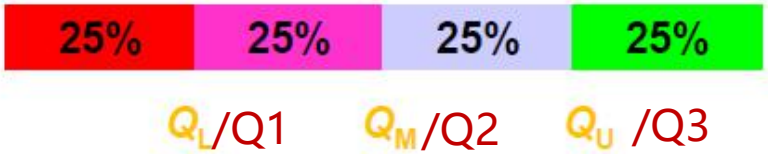
数据分析算法

北京理工大学计算机学院　孙新

2019年1月

- p分位数M_p
  - p分位数M_p 表示排在序列长度p（$0 \le p \le 1$）位置的数
  - 其中四分位数（p=0.25和p=0.75）最为常用

- 四分位数(quartile)
  - 排序后，处于25%和75%位置上的值



  - 不受极端值的影响
  - 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)

方法1：定义算法

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

方法2：较准确算法

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

例题：顺序数据的四分位数

甲城市家庭对住房状况评价的频数分布		
回答类别	甲城市	
	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	—

解：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

例题：数值型数据的四分位数

9个家庭的人均月收入数据(2种方法计算)

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

方法1

$$Q_L\text{位置}=\frac{9}{4}=2.25 \qquad Q_U\text{位置}=\frac{3\times 9}{4}=6.75$$

$$Q_L=780+(850-780)\times 0.25=797.5$$

$$Q_U=1250+(1500-1250)\times 0.75=1437.5$$

例题：数值型数据的四分位数

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

方法2

$$Q_L\text{位置}=\frac{9+1}{4}=2.5 \qquad Q_U\text{位置}=\frac{3(9+1)}{4}=7.5$$

$$Q_L=\frac{780+850}{2}=815 \qquad Q_U=\frac{1500+1630}{2}=1565$$

(4)截断均值： **指定** 0和100间的百分数p，丢弃高端和低端两端各（p/2）百分个数，然后用常规方法计算均值，所得到的结果即是截断均值

目的：**可以抵消少数极端值的影响，去掉最高和最低值的影响**

中位数是p=100%时的截断均值，

而均值是对应于p=0%时的截断均值

【例】计算{1,2,3,4,5,90} 的均值、中位数和p=40%截断均值

解：**均值= (1+2+3+4+5+90)/6=17.5；**

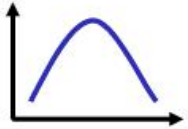
中位数=(3+4)/2=3.5；　p=40%截断均值=3.5

- 截断均值的计算方法：
- 1.p=40，则p/2=20,即从两端要除去总数据个数的20%个数
- 2.总数为6个，6*20%=1.2，即两边各除去1个数据
- 3.剩下中间的4个数据为：{2,3,4,5}
- 4.计算这四个数的均值为（2+3+4+5）/4=3.5

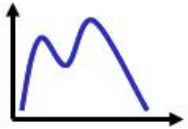
- (5) 众数(mode，也叫作峰)– M_0
  - 一组数据中出现次数最多的数值，叫做众数

无众数
原始数据: 10 5 9 12 6 8

一个众数
原始数据: 6 5 9 8 5 5



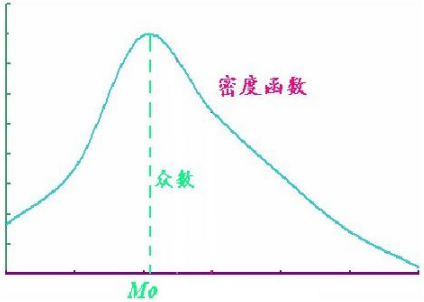
多于一个众数
原始数据: 25 28 28 36 42 42



- 若一个样品中只有一个众数/峰就叫单峰
- 若有两个或两个以上的峰就叫双峰或多峰。
- 其中最高的一个叫主峰,次高的叫次峰

- (5) 众数(mode，也叫作峰)– M_0
  - 通过定义我们知道，众数是在统计分布上具有明显集中趋势点的数值，代表数据的一般水平。
  - 适合于数据量较多时使用，不受极端值的影响
  - 不唯一：一组数据可能没有众数或有几个众数
  - 主要用于分类数据，也可用于顺序数据和数值型数据

众数是一组数据中出现次数最多的标志值。



例题：分类数据的众数

不同品牌饮料的频数分布			
饮料品牌	频数	比例	百分比(%)
可口可乐	15	0.30	30
旭日升冰茶	11	0.22	22
百事可乐	9	0.18	18
汇源果汁	6	0.12	12
露露	9	0.18	18
合计	50	1	100

解：这里的变量为“饮料品牌”，这是个分类变量，不同类型的饮料就是变量值

所调查的50人中，购买可口可乐的人数最多，为15人，占被调查总人数的30%，因此众数为“可口可乐”这一品牌，即

M_o=可口可乐

例题：顺序数据的众数

甲城市家庭对住房状况评价的频数分布		
回答类别	甲城市	
	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

解：这里的数据为顺序数据。变量为“回答类别”

甲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＝不满意

- (6) 中列数(midrange)
  - 在统计中指的是数据集里最大值和最小值的算术平均

- 例如： {1，3，7，9，0，3，5}
  - 它的中列数即为(0+9)/2=4.5

2.1数据的中心趋势度量

(1) 均值（平均数）
(2) 加权算数均值
(3) 中位数 四分位数
(4) 截断均值
(5) 众数
(6) 中列数

中位数：也叫中值，是倾斜数据集的最好度量方式

一组数据按从小到大(或从大到小)的顺序依次排列,处在中间位置的一个数(或最中间两个数据的平均数)

众数：

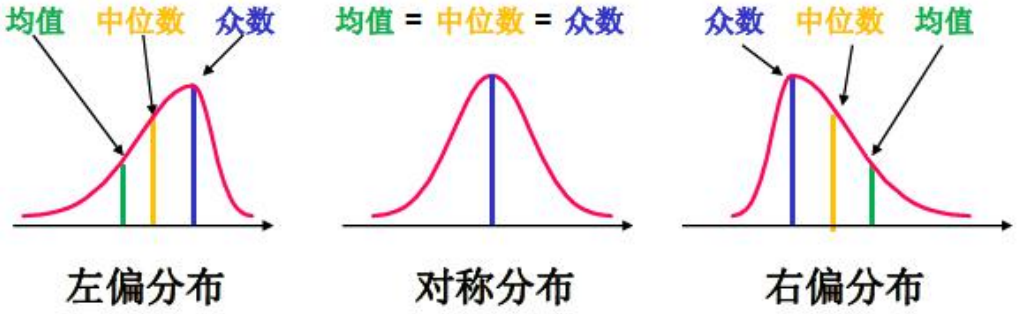
是在一组数据中，出现次数最多的数据

中列数： 数据集的最大和最小值的算术平均值

2、统计数据分析方法

众数、中位数、均值的特点

- 众数
  - 不受极端值影响
  - 缺点具有不惟一性，
  - 数据量较少时，不宜使用
  - 主要适合作为分类数据的集中趋势侧度值
- 中位数
  - 不受极端值影响
  - 数据分布偏斜程度较大时应用
  - 主要适合作为顺序数据的集中趋势侧度值



左偏分布:说明数据存在极小值,拉动平均值向极小值一方靠,而众数和中位数是位置代表值，不受极值的影响，三者关系表现为

$$\overline{x} < M_e < M_o$$

- 均值（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
  - 数据为偏态分布，特别是偏态程度较大时，中位数或众数代表性好