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

数据分析算法

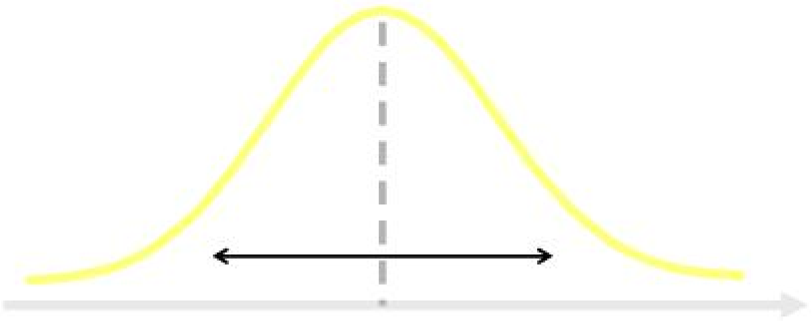
北京理工大学计算机学院　孙新

2019年1月

2.2数据的离散程度度量

离散程度也叫做离中趋势

- 它是数据分布的另一个重要特征
- 反映各变量值远离其中心值的程度(离散程度)
- 从另一个侧面说明了集中趋势测度值的代表程度
- 不同类型的数据有不同的离散程度测度值



- 离散程度的度量
  - （1）四分位差
  - （2）极差
  - （3）方差和标准差

2.2数据的离散程度度量

（1）四分位差(quartile deviation)是对顺序数据离散程度的测度，也称为四分位距（interquartile range, IQR）

四分位数

25%	25%	25%	25%

Q_L/Q1　Q_M/Q2　Q_U /Q3

- 四分位差定义为上四分位数与下四分位数之差，反映了中间50%数据的离散程度

$$Q_d = Q_U - Q_L$$ 或者 IQR = Q3 – Q1

- 四分位差不受极端值的影响，四分位差的数值越小，说明中间的数据越集中，其数值越大，说明中间的数据越分散

2.2数据的离散程度度量

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： 设非常不满意为1,不满意为2，一般为3，满意为 4,非常满意为5 。先计算:

上四分位数300*0.25=75

下四分位数300*0.75=225

Q_L = 不满意 = 2

Q_U = 一般 = 3

四分位差为

$$Q_d = Q_U - Q_L$$
$$= 3 - 2 = 1$$

2.2数据的离散程度度量

- 极差和方差是集值的散布度量，表明属性值是否散布很宽，或者是相对集中在当个点（如均值）附近。

（2）极差(range)：一组数据的最大值与最小值之差

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

- 极差是离散程度的最简单测度值
  - 易受极端值影响
  - 未考虑数据的分布
  - 特点：计算简便，直观易于理解。

(3)方差和标准差(variance and standard deviation)

- 方差s^2是用来反映这种数据分散程度的最常用的一种指标，反映了各变量值与均值的平均差异

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

- 方差的算术平方根被称为标准差s

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

几个概念的小结

方差：方差值越大说明该数据项波动越大

数值数据趋向于分散的程度

极差：最大值与最小值之差，极差忽略了数据内部差异而仅关注数据上下界的指标，

p分位数：百分之p的数据项位于或低于X_i

四分位数：

把所有数值由小到大排列并分成四等份，处于三个分割点位置的数值就是四分位数。

四分位差：

第三四分位数与第一四分位数的差距

- 【例】依据下面两组数据，分别计算均值、中位数和方差
  - 第一组：99个年收入10万的人和1个年收入1000万的人，
  - 第二组：60个年收入10万的人和40个年收入34.75万的人
- **解：** 第一组均值=(99*10+1000)/100=1990/100=19.9

  第二组均值=(60*10+40*34.75)/100=1990/100=19.9

  第一组中位数=第一组中位数=10

  第一组**方差9801**，

  第二组**方差148.5**

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

2、统计数据分析方法

- 2.1数据的中心趋势度量
  - 均值、加权算数均值、中位数、截断均值、众数、中列数
- 2.2数据的离散程度度量
  - 四分位差、极差、方差和标准差
- 2.3数据分布的度量
  - 偏态及其测度
  - 峰态及其测度

- 偏度（skewness）也称为偏态、偏态系数，
  - 是统计数据分布偏斜方向和程度的度量，
  - 是统计数据分布非对称程度的数字特征
  - 计算公式如下：

1. 根据原始数据计算 $$SK=\frac{n\sum(x_i-\bar{x})^3}{(n-1)(n-2)s^3}$$

2. 根据分组数据计算 $$SK=\frac{\sum_{i=1}^{k}(M_i-\bar{x})^3 f_i}{ns^3}$$

- 偏度是数据分布偏斜程度的测度
  - 偏度系数=0为对称分布
  - 偏度系数> 0为右偏分布
  - 偏度系数< 0为左偏分布
- 偏度系数大于1或小于-1，被称为高度偏态分布；
- 偏度系数在0.5～1或-1～-0.5之间，被认为是中等偏态分布；

- 偏度系数越接近0，数据分布相对比较对称。偏斜程度就越低
- 偏度大于0时，比均值更小的数据更多一些，反之则是比均值更大的数据较多

- 例题：偏度系数

某电脑公司销售量偏态及峰度计算表				
按销售量分组(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				
220				
230				
合计	—	120	540000	70100000

$$SK=\frac{\sum_{i=1}^{k}(M_i-\bar{x})^3 f_i}{ns^3}=\frac{\sum_{i=1}^{10}(M_i-185)^3 f_i}{120\times(21.58)^3}$$

结论：偏度系数为正值，但与0的差异不大，说明电脑销售量为轻微右偏分布，即销售量较少的天数占据多数，而销售量较多的天数则占少数