数值型数据的四分位数
(奇数个)

- 【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

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

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

数值型数据的四分位数
(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

$$Q_L\text{位置} = \frac{10+1}{4} = 2.75 \qquad Q_U\text{位置} = \frac{3\times(10+1)}{4} = 8.25$$

$$Q_L = 750 + 0.75\times(780-750) = 772.5$$

$$Q_U = 1500 + 0.25\times(1630-1500) = 1532.5$$

均值
(mean)

1. 集中趋势的最常用测度值
2. 一组数据的均衡点所在
3. 体现了数据的必然性特征
4. 易受极端值的影响
5. 用于数值型数据，不能用于分类数据和顺序数据

简单平均数
(simple mean)

设一组数据为：　x_1，x_2，…，x_n

总体平均数 $$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{n} x_i}{N}$$

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

加权平均数
(weighted mean)

设一组数据为： x_1， x_2， …， x_n

相应的频数为： f_1， f_2， …， f_k

总体平均数 $$\mu = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_K f_K}{f_1 + f_2 + \cdots + f_K} = \frac{\sum_{i=1}^{K} x_i f_i}{\sum_{i=1}^{K} f_i}$$

样本平均数 $$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{n} f_i}$$

加权平均数
(例题分析)

按加工数量分组	组中值 x_i	数量 f_i	$x_i f_i$
80-90	85	3	225
90-100	95	7	665
100-110	105	13	1365
110-120	115	5	575
120-130	125	2	250
合计	-	30	3110

$$\bar{x}=\frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}=\frac{3110}{30}=103.67(件)$$

平均数
(数学性质)

- 1. 各变量值与平均数的离差之和等于零

$$\sum_{i=1}^{n}(x_i-\bar{x})=0$$

2. 各变量值与平均数的离差平方和最小

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

几何平均数
(geometric mean)

1. n 个变量值乘积的 n 次方根
2. 适用于对比率数据的平均
3. 主要用于计算平均增长率
4. 计算公式为

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

5. 可看作是平均数的一种变形

$$\lg G_m = \frac{1}{n}(\lg x_1 + \lg x_2 + \cdots + \lg x_n) = \frac{\sum_{i=1}^{n} \lg x_i}{n}$$

几何平均数
(例题分析)

澳门从2002年-2012年的本地生产总值如下，试计算平均增长速度

年份	本地生產總值（百萬澳門元）	增长速度
2002	56,298.50	
2003	63,579.40	1.13
2004	82,294.00	1.29
2005	94,471.00	1.15
2006	116,570.50	1.23
2007	145,084.80	1.24
2008	166,265.10	1.15
2009	170,171.10	1.02
2010	226,940.50	1.33
2011	295,046.30	1.30
2012	348,216.40	1.18

$$\overline{G} = \sqrt[10]{1.13 \times 1.29 \times 1.15 \times \ldots\ldots 1.18}$$

$$= \sqrt[10]{6.185181} = 1.20\%$$

切尾均值
(trimmed Mean)

1. 去掉大小两端的若干数值后计算中间数据的均值
2. 在电视大奖赛、体育比赛及需要人们进行综合评价的比赛项目中已得到广泛应用
3. 计算公式为

$$\bar{x}_\alpha = \frac{x_{(n\alpha+1)} + x_{(n\alpha+2)} + \cdots + x_{(n-n\alpha)}}{n - 2 \times n\alpha}$$

n 表示观察值的个数； α 表示切尾系数，$0 \le \alpha < \frac{1}{2}$

切尾均值
(例题分析)

- 某歌唱比赛共有**11**名评委，对某位歌手的给分分别是：

$$x_1,\quad x_2,\quad x_3,\quad x_4,\quad x_5,\quad x_6,\quad x_7,\quad x_8,\quad x_9,\quad x_{10},\quad x_{11}$$

$$9.22, 9.25, 9.20, 9.30, 9.65, 9.30, 9.27, 9.20, 9.28, 9.25, 9.24$$

经整理得到顺序统计量值为

$$x_{(1)},\quad x_{(2)},\quad x_{(3)},\quad x_{(4)},\quad x_{(5)},\quad x_{(6)},\quad x_{(7)},\quad x_{(8)},\quad x_{(9)},\quad x_{(10)},\quad x_{(11)}$$

$$9.20,\ 9.20,\ 9.22,\ 9.24,\ 9.25,\ 9.25,\ 9.27,\ 9.28,\ 9.30,\ 9.30,\ 9.65$$

去掉一个最高分和一个最低分，取**1/11**

$$\bar{x}_{1/11} = \frac{x_{(11\times1/11+1)} + x_{(11\times1/11+2)} + \cdots + x_{(11-11\times1/11)}}{11-2\times11\times1/11}$$

$$= \frac{x_{(2)} + x_{(3)} + \cdots + x_{(10)}}{11-2}$$

$$= \frac{9.2 + 9.22 + \cdots + 9.3}{9} = 9.26$$

众数、中位数、平均数的比较

1. 众数
   - 不受极端值影响
   - 具有不惟一性
   - 数据分布偏斜程度较大时应用
2. 中位数
   - 不受极端值影响
   - 数据分布偏斜程度较大时应用
3. 平均数
   - 易受极端值影响
   - 数学性质优良
   - 数据对称分布或接近对称分布时应用

2.3 分布离散程度的测度

- 一、极差
- 二、内距
- 三、方差和标准差
- 四、离散系数

极差
(range)

1. 一组数据的最大值与最小值之差
2. 离散程度的最简单测度值
3. 易受极端值影响
4. 未考虑数据的分布
5. 计算公式为：

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

内距
(Inter-Quartile Range,IQR)

1. 也称四分位差
2. 上四分位数与下四分位数之差
   - 内 距$= Q_3 - Q_1$
3. 反映了中间**50%**数据的离散程度
4. 不受极端值的影响
5. 可用于衡量中位数的代表性

总体方差和标准差

(Population variance and Standard deviation)

方差的计算公式

未分组数据：

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

组距分组数据：

$$\sigma^2 = \frac{\sum_{i=1}^{K}(M_i-\mu)^2 f_i}{N}$$

标准差的计算公式

- 未分组数据：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

组距分组数据：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{K}(M_i-\mu)^2 f_i}{N}}$$

样本方差和标准差

(simple variance and standard deviation)

方差的计算公式

未分组数据：

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

组距分组数据：

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

标准差的计算公式

- 未分组数据：

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

组距分组数据：

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

离散系数
(coefficient of variation)

- 1.即标准差与其相应的均值之比
- 2.是对数据相对离散程度的测度
- 3.消除了数据水平高低和计量单位的影响
- 4. 用于对不同组别数据离散程度的比较
- 5. 计算公式为

$$v_\sigma = \frac{\sigma}{\mu} \qquad v_s = \frac{s}{\bar{x}}$$

离散系数
(例题分析)

某管理局抽查了所属的8家企业，其产品销售数据如表。试比较产品销售额与销售利润的离散程度

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

离散系数
(例题分析)

$$\bar{x}_1 = 536.25(\text{万元})$$

$$s_1 = 309.19(\text{万元})$$

$$v_1 = \frac{309.19}{536.25} = 0.577$$

$$\bar{x}_2 = 32.5215(\text{万元})$$

$$s_2 = 23.09(\text{万元})$$

$$v_2 = \frac{23.09}{32.5215} = 0.710$$

结论： 计算结果表明，$v_1<v_2$，说明产品销售额的离散程度小于销售利润的离散程度

2.4 分布偏态与峰态的测度

- 一、偏态及其测度
- 二、峰态及其测度

偏态
(skewness)

1. 统计学家**Pearson**于**1895**年首次提出
2. 是对数据分布偏斜程度的测度

- 偏态系数=0为对称分布
- 偏态系数> 0为右偏分布

4. 偏态系数< 0为左偏分布

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

峰态
(kurtosis)

1. 统计学家**Pearson**于**1905**年首次提出
2. 数据分布扁平程度的测度
3. 峰态系数=0扁平峰度适中
4. 峰态系数<0为扁平分布
5. 峰态系数>0为尖峰分布

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

例子

日产量（件）	人数（人）	组中值
50以下	11	45
50－60	13	55
60－70	70	65
70－80	120	75
80－90	50	85
90－100	30	95
100－110	5	105
110以上	1	115
合计	300	

$$\bar{x} = 75, \sigma = 12.27$$

$$m_3 = \frac{\sum (x - \bar{x})^3 f}{\sum f} = \frac{18000}{300} = 60$$

$$\sigma^3 = (12.27)^3 = 1847.28$$

$$SK = \frac{m_3}{\sigma^3} = \frac{60}{1847.28} = 0.032$$

- **SK〉0，轻微右偏分布**

$$m_4 = \frac{\sum (x - \bar{x})^4 f}{\sum f} = \frac{23600000}{300} = 78666.67$$

$$\sigma^4 = (12.27)^4 = 22666.18$$

$$K = \frac{m_4}{\sigma^4} - 3 = \frac{78666.67}{22666.18} - 3 = 3.47 - 3 = 0.47$$

- **K〉0，呈轻微尖峰分布**

运用EXCEL

- 某班**50**名学生概率考试成绩如下：
- 65 80 81 92 63 77 79 54 98
- 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

描述性统计

平均	75.8
标准误差	1.718566
中位数	78
众数	78
标准差	12.1521
方差	147.6735
峰度	1.064834
偏度	-0.84838
最小值	38
最大值	98
求和	3790
观测数	50