# 第四章 数据分布特征

#一、集中趋势的计量

集中趋势反映的是数据中各数据所具有的共同趋势，即资料中各数据聚集的位置

1. 算数平均值$\bar{X}$
   - 简单算数平均值

     $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

     算术平均数的统计含义：算术平均数是同质总体各数据偶然性、随机性特征互相抵消后的稳定数值。反映数据集中的特征。
     算数平均值的性质一：任一组资料中，各项数值与其均值之差(离差)的代数和为0

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

     算数平均值的性质二：任一组资料中，各数据与均值离差的平方和最小

     $$\sum_{i=1}^{n}(x_i - \bar{X})^2 \leq \sum_{i=1}^{n}(x_i - A)$$

     其中`A`为任意数
     算数平均值的性质三：对于多组资料合并后总体均值就是每一组的权重乘以对应均值的累加和，其实就是加权算数平均值
   - 加权算数平均值
     如果数据是分组资料，经过次数分配，由于各组的次数不同，要用次数作为加权平均数。加权平均数的计算公式

     $$\bar{X} = \frac{\bar{X}_1 f_1 + \bar{X}_2 f_2 + \ldots + \bar{X}_k f_k}{f_1 + f_2 + \ldots + f_k}$$

     其中$\bar{X}_i$表示第`i`组的组中值，$f_i$表示第`i`组的次数
2. 中位数`Md`
   将数据按照变量值从小到大的顺序排列，处于中点位置的数值是中位数
   中位数的确定方法
   - 对于未分组的数据
     如果数据个数为奇数，则处于(n+1)/2位置的标志值是中位数
     如果数据个数为偶数，则处于n/2、n\2+1的两个标志值的平均数为中位数
   - 对于分组数据
     如果是单项分组数据，如按照分类标签进行分组的数据，如果确定中位数在该组之后，该组的标签就是对应就是中位数
     如果是组距分组数据，中位数是位于频数n/2位置的数

向上累计

$$Md = L_i + \frac{n/2 - F_{i-1}}{f_i}(U_i - L_i)$$

向下累计

$$Md = U_i + \frac{n/2 - F_{i+1}}{f_i}(U_i - L_i)$$

$L_i$是中位数所在组的下限， $f_i$是中位数所在组的次数。 $F_{i-1}$是中位数所在组的前一组的累积次数； $U_i - L_i$是中位数所在组的组距 =上限-下限
中位数的性质：

- 优点：中位数是位置平均数，不受极端值的影响，是较为稳健的集中趋势测度量指标
- 不足：中位数确定时只和中间位置有关，不考虑其他数值的大小，缺乏敏感性，不适合进行代数运算

3. 众数$M_0$
众数是一组资料中出现次数最多的数值，反映了数据集中的程度
   - 对于未分组的资料：众数就是出现次数最多的变量值
   - 对于分组的数据：在等距分组的情况下，频数最多的组是众数组，在该组中确定众数
   假设众数在第`i`组，则

$$M_0 = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})}d_i$$

$$M_0 = U_i - \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})}d_i$$

   $L_i$是众数所在组的下限， $U_i$是众数所在组的上限； $f_i$是众数所在组的次数，$f_{i-1}$是变量值小于众数组那个相邻组的次数， $f_{i+1}$是变量值大于众数组那个相邻组的次数，
   $d_i = U_i - L_i$是众数所在组的组距 =上限-下限
   众数的性质：
   - 优点：与中位数一样，性质简单明了，不受极端值的影响，适用于有较多数值项某一数值集中的情况，特别当资料是按照品质标志进行分组的时候，使用众数比较恰当
   - 不足：没有利用所有的观测值，缺少敏感性，不适合进一步进行代数运算，容易受到分组和样本的影响，众数可能不存在或者不止一个

4. 平均值，中位数，众数三者之间的关系
   对称分布$\bar{X} = Md = M_0$
   正偏态分布(右)$\bar{X} > Md > M_0$
   负偏态分布(左)$\bar{X} < Md < M_0$
   均值是数据分布的平衡点或重心，中位数把这个划分为两半，众数位于分布的顶端

$$\bar{X} - M_0 = 3(\bar{X} - Md)$$

$$\bar{X} = \frac{3Md - M_0}{2}$$

$$Md = \frac{M_0 - 2\bar{X}}{3}$$

5. 集中趋势的其他测度量
   - 分位数：四分位数、十分位数、百分位数
     四分位数：将资料按照大小顺序排序后，分成四等分，得到三个分割点，每个分割点的数值为四分位数，用$Q_1$，$Q_2$，$Q_3$表示。相似的有十分位数和百分位数等等
     分位数的计算：
     - 将资料按照大小顺序排列
     - 求出分位数所在的位置
     - 如果i是整数，所求的分位数就是该位置上的数值；如果i是非整数，则取i与i+1两个数值的平均数来计算分位数
     - 如果资料是分组数据，则各分位数可以按照下列公式计算：

       $$K_i = L_i + \frac{iN/K - F_{i-1}}{f_i}d_i$$

       其中：$K_i$表示第i个分位数，$L_i$表示第i个分位数所在组的下限，N表示数据总个数，$F_{i-1}$表示第i个K分位数所在组的前一组的累积次数，$f_i$表示第i个K分位数所在组的次数，$d_i = U_i - L_i$是第i个K分位数所在组的组距
     - 四分位数的位置确定方法：
       - $Q_L$位置$= \frac{n}{4}$， $Q_U$位置$= \frac{3n}{4}$
       - $Q_{位置} = \frac{[\frac{n+1}{2}]+1}{2}$
       - $Q_L = \frac{n+3}{4}$，$Q_U = \frac{3n+1}{4}$
         三种方法计算的四分位数不完全相同。但对 他们的解释是一样的，即排序数据中，至少25%的数 据小于等于$Q_L$，至少75%的数据小于等于$Q_U$
   - 几何平均数：

     $$M_g = (x_1 x_2 \dots x_N)^{\frac{1}{N}}$$

     用于计算平均比率或者平均速度。包括：
     对比例进行平均
     测定生产或经济变量的时间序列平均增长率
     对于平均通货膨胀率，银行平均年利率使用公式为

     $$\bar{R} = (x_1 x_2 \dots x_N)^{\frac{1}{N}} - 1$$

   - 调和平均值：
     调和平均值是观察值倒数之和的平均值的倒数，也叫做倒数平均数

     $$M_H = \frac{1}{\frac{\sum_{i=1}^{N}\frac{1}{x_i}}{N}}$$

     计算的是相对指标的平均值，如平均价格，平均成本，平均劳动生产率等等
   - 算数平均值、几何平均值、调和平均值三个关系：
     $M_H \le M_g \le \bar{X}$

#二、离中趋势的计算

离中趋势表示变量值的差异或者离散程度，常用的指标有：极差、方差、标准差、四分位差等，它们也被称为变异指标

1. 极差
   极差也称为全距，是一组数据的最大值和最小值的差

$$R = X_{max} - X_{min}$$

   极差越小，数据变动范围越小，平均数的代表性越高
2. 平均差
   平均差是指数据值与均值之差的绝对值的算数平均值，用符号$A.D.$表示

$$A.D. = \frac{\sum_{i=1}^{n} |x_i - \bar{X}|}{n}$$

3. 方差与标准差
   总体方差是观察值与其均值离差平方和的均值

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

   总体方差的另外一种表达方式

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

   总体标准差是总体方差的算术平方根

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

   样本方差与样本标准差

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

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

   当样本数据个数足够大时，样本方差与总体方差很接近
4. Chebishev定理与经验法则
   异常值判断，$3\sigma$准则，与算术平均数的偏差超过3个标准差的数据称为高度异常的数据
   标准化处理：将观察值与算术平均数相减再与标准差相除

$$Z_i = \frac{x_i - \bar{X}}{\sigma}$$

5. 相对离中趋势——变异系数
   变异系数又称为离散系数，是标准差与均值的比值

$$C.V. = \frac{\sigma}{\mu}$$

   表示数据相对离散程度的测度
   消除了水平高低和计量单位的影响
   用于对不同组别的离散程度进行比较
   对于使用均值和标准差不能判断的数据可以进一步使用变异系数进行比较
6. 离中趋势的其他测度
   - 四份位差

$$Q.D. = Q_3 - Q_1$$

   - 异众比率
     异众比率表示非众数值的次数之和站总次数的比重

$$V_{M_o} = \frac{n - f_{M_o}}{n}$$

     异众比率越大说明众数的代表性越低
   - 平均差系数

$$V_{AD} = \frac{A.D.}{\bar{X}}$$

---

#三、数据的分布形状

1. 偏斜度：
   偏斜度是对数据分布在平均数两侧的便宜方向和偏移程度的描述
   - Pearson偏态系数
     偏态系数以均值与中位数的差，除以标准差来衡量偏斜程度

$$Sk = \frac{3(\bar{X} - Md)}{\sigma}$$

$$Sk = \frac{\bar{X} - M_o}{\sigma}$$

     当对称分布的时候Sk=0，当Sk>0时候，分布是右偏(正偏)的，当Sk<0，分布是左偏(负偏)的
   - 矩法求偏态系数
     使用中心距来衡量分布的偏度

$$Sk = \frac{m_3}{\sigma^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^3}{\sigma^3}$$

样本偏斜度

$$Sk = \frac{n}{(n-1)(n-2)} * \sum(\frac{x_i - \bar{X}}{S})^3$$

2. 峰度系数

峰度是变量分布的曲线尖峭程度

$$K = \frac{m_4}{\sigma^4} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^4}{\sigma^4 n}$$

峰度系数K=3称为常态峰，K>3是尖峰，K<3是扁峰

样本峰度

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n}(\frac{x_i - \bar{X}}{S})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

该结果是与0比较，正态分布的K=0