第四章 数据概括性的度量

4.1 集中趋势的度量

众数 M_o 是一组数据中出现次数最多的变量值， 只有在数据量较大的情况下才有意义。众数不受极端数据影响，反应集中趋势。

中位数 M_e 是中间位置的变量值

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

四分位数 $Q_L=x_{(\frac{n}{4})},Q_M=M_e,Q_U=x_{(\frac{3n}{4})}$ ，如果落在 0.5 则两侧取平均，0.25 或 0.75 则两侧加权平均。

样本平均数 $\bar{x}$，总体平均 μ，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 。如果将样本分为 k 组，每组组中值 M_i，频数 f_i，则样本平均

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}M_if_i\tag{2}$$

> $$\forall a\in\mathbb{R},\ \sum_{i=1}^{n}(x_i-a)^2\geq\sum_{i=1}^{n}(x_i-\bar{x})^2\tag{3}$$

> ### 特殊平均数
>
> 几何平均数 G 用于计算数据的平均比率。例如平均收益率。
>
> $$G=\sqrt[n]{\prod_{i=1}^{n}x_i}\tag{4}$$
>
> 调和平均数 H 用于强调较小值的影响
>
> $$H=\sum_{i=1}^{n}\frac{n}{1/x_i}\tag{5}$$

众数、中位数和平均数

- 对称分布：$M_o=M_e=\bar{x}$
- 左偏分布：$\bar{x}<M_e<M_o$，右偏分布 $M_o<M_e<\bar{x}$

4.2 离散程度的度量

4.2.1 分类数据

异众比率 V_r 是指非众数组占总频数的比率，用于衡量众数的代表性。

$$V_r = \frac{(\sum f_i) - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i} \tag{6}$$

4.2.2 顺序数据

四分位差 $Q_d = Q_U - Q_L$ ，反映中间 50% 的离散程度，说明数据的集中程度。

4.2.3 数值型数据

极差 $R = \max(x_i) - \min(x_i)$

平均差（绝对离差）M_d

$$M_d = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| & \text{未分组数据} \\ \frac{1}{n}\sum_{i=1}^{k}|M_i - \bar{x}|f_i & \text{分组数据} \end{cases} \tag{7}$$

样本方差 s^2 ，标准差 s ；总体方差 σ^2，标准差 σ，未分组数据的样本方差和总体方差的表达式如下。

$$\begin{gathered} s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \\ s = \sqrt{s^2}, \quad \sigma = \sqrt{\sigma^2} \end{gathered} \tag{8}$$

样本方差的分母是 $n-1$ 是由于样本的自由度不是 n，被平均数限制了一个自由度。

如果是分组数据的话，计算公式为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(M_i - \bar{x})f_i \tag{9}$$

相对位置度量

(1) 标准分数由标准化得到 $z_i = (x_i - \bar{x})/s$

(2) 经验法则（对称数据）

(3) 切比雪夫不等式 （任意分布）

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2} \tag{10}$$

在 ± 3 个标准差之外的数据在统计学上被称为“离群点”（outlier）。

	经验法则	切比雪夫不等式
± 1	约 68%	\
± 2	约 95%	$\geq 75\%$

z_i	经验法则	切比雪夫不等式
± 3	约 99%	$\geq 89\%$
± 4	\	$\geq 94\%$

4.2.4 相对离散程度：离散系数

与方差和标准差相比，离散系数消除了变量值水平对离散程度测量的影响。离散系数（变异系数）$v_s = s/\bar{x}$

4.3 偏态与峰态的度量

偏态系数 SK 描绘数据的偏斜方向和程度。

$$\begin{aligned} SK &= \frac{n\sum(x_i-\bar{x})^3}{(n-1)(n-2)s^3} \\ &= \frac{\sum_{i=1}^k (M_i-\bar{x})^3 f_i}{ns^3} \qquad \text{分组数据} \\ &= \frac{n}{(n-1)(n-2)}\sum(z_i-\bar{z})^3 \end{aligned} \tag{11}$$

- 数据集中在左边叫“右偏”，因为有立方，因此极端值会更大地影响偏度系数。
  - $$SK = \frac{n}{(n-1)(n-2)}\frac{1}{s^3}\left(\sum_{x_i<\bar{x}}(x_i-\bar{x})^3 + \sum_{x_i>\bar{x}}(x_i-\bar{x})^3\right) \tag{12}$$
- SK 越接近于 0，则偏斜程度越小。
  - $|SK| > 1$ 则数据是高度偏态分布；
  - $0.5 < |SK| < 1$ 则中等偏态分布；
  - $SK > 0$ 为右偏，$SK < 0$ 为左偏。

峰度系数 K 描绘数据相对于正态分布的偏离情况。

$$\begin{aligned} K &= \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \\ &= \frac{\sum_{i=1}^k (M_i-\bar{x})^4 f_i}{ns} - 3 \end{aligned} \tag{13}$$

- 从图像比较峰度系数不能使用原始图像，须进行标准化。
- 当数据来源正态分布时 $K \simeq 0$，峰度系数越大图像的尾部越厚。

对于分组数据来说，组中值就是分组的中间。例如在 $200 \sim 300$ 的组中组中值就是 250。