天善智能
TIANSHAN SOFT

专注于商业智能BI和大数据的垂直社区平台

统计量的描述

Allen

课程目录

- 水平的度量
- 离散差异的度量
- 分布形状的度量
- 小结

天善智能 TIANSHAN SOFT

水平的度量—平均数

- 也称为均值，是常用的统计量之一
- 容易受到极端值的影响
- 设一组数据为 $x_1, x_2, \cdots, x_n$ ，那么平均数定义为 $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

天善智能 TIANSHAN SOFT

水平的度量—几何平均数

- n个变量值乘积的n次方根
- 主要用于计算平均增长率，适用于对比率数据的平均
- 设一组数据为 $x_1, x_2, \cdots, x_n$ ，那么几何平均数定义为 $\bar{x} = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$

水平的度量—中位数

- 排序后处于中间位置上的数据称为中位数

- 主要描述数据中心位置的数字特征

- 对于分布对称的数据，均值与中位数比较接近
  对于偏态分布的数据，均值与中位数偏差大

水平的度量—中位数

- 中位数不受极端值的影响，比较稳健
- 中位数记为

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

- 示例

水平的度量—百分位数

- 中位数的推广
- 若 $0<p<1$，则p分位点定义为 $m_p=\begin{cases} x_{([np]+1)} & np\text{不为整数} \\ \frac{1}{2}\left(x_{(np)}+x_{(np+1)}\right) & np\text{为整数} \end{cases}$
- 示例

天善智能 TIANSHAN SOFT

水平的度量—众数

- 一组数据中出现次数最多的变量值称为众数

- 不受极端值的影响，适合数据量较多的时候使用

- 一组数据中可能没有众数，也可能有多个众数

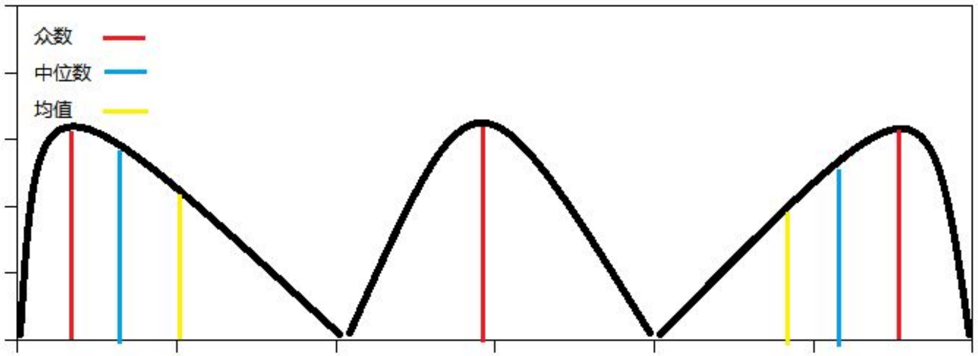
1,2,3,4,5,6,7,8,9,0

1,2,3,3,3,4,4,4,5,6,7,8,9,9,9

天善智能 TIANSHAN SOFT

水平的度量—总结

- 平均数：
  - 容易受到极端值的影响
  - 数据对称分布或接近对称分布时比较有代表性
- 中位数：
  - 不受极端值影响
  - 数据分布偏斜程度较大时有代表性



- 众数：
  - 不受极端值的影响
  - 数据分布偏斜程度较大且有明显峰值时比较有代表性

天善智能
TIANSHAN SOFT

离散程度的度量

- 问题：现有人均年收入为5000元的A、B两地，如果使用人均收入来衡量这两地的生活水平，那么这两地的生活水平一样高吗？

- 假设1：两地的95%以上的普通居民收入都持平在5000元

- 假设2：A地95%居民收入在5000元上下，B地有几个富翁，而其他很多收入都在2000元上下

天善智能 TIANSHAN SOFT

离散程度的度量—极差

- 一组数据中的最大值和最小值之差称为极差
- 未考虑数据的分布，易受极端值的影响
- 测量数据离散程度最简单的方法

A地最高人均收入为8000最低为3500
极差为：8000-3500

B地最高人均收入为50万最低为1500
极差为：500000-1500

天善智能
TIANSHAN SOFT

离散程度的度量—方差和标准差

- 反映各变量值和平均值之间的差异程度，反映数据取值分散性的一个度量
- 是反映数据离散程度最常用的度量值
- 样本方差的计算：$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

自由度指数据个数与附加给自由观测值的约束或限制个数之差

- 样本标准差的计算：$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

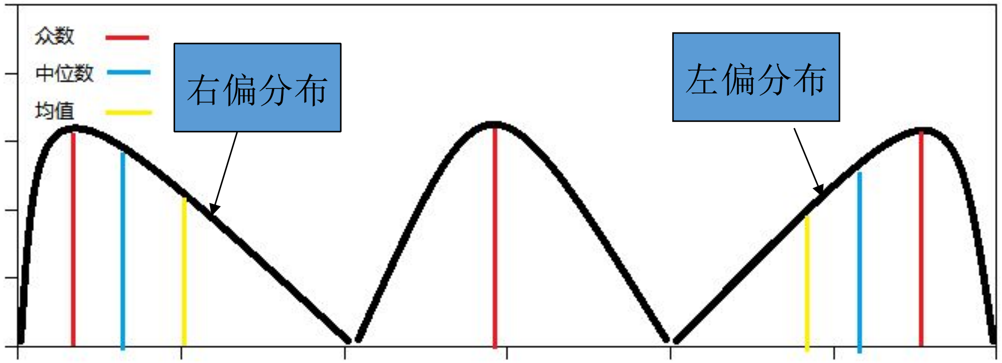
天善智能 TIANSHAN SOFT

分布形状的度量—偏度系数

- 反映数据分布的不对称性

- 偏度系数等于0为对称分布，大于0为右偏分布，小于0为左偏分布

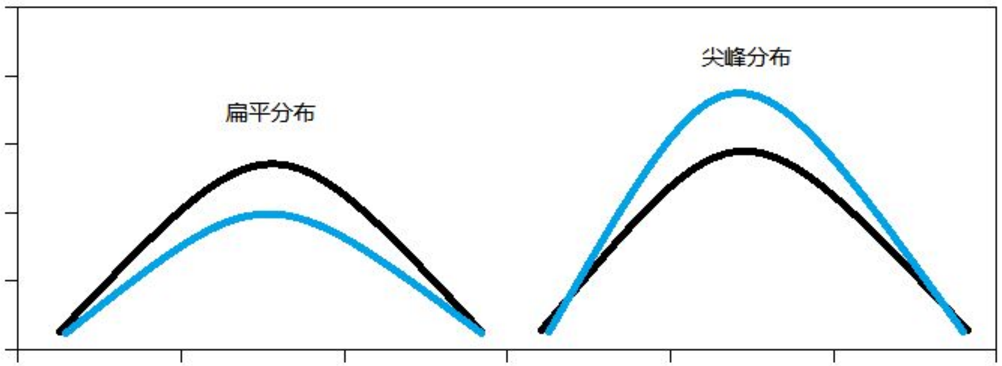
- 计算公式 $SK=\frac{n}{(n-1)(n-2)}\sum\left(\frac{x-\bar{x}}{s}\right)^3$



分布形状的度量—峰度系数

- 反映数据分布峰值的高低

- 峰度系数等于0时扁平峰度适中，峰度系数小于0时为扁平分布，峰度系数大于0时为尖峰分布



小结

- 水平的度量
- 离散差异的度量
- 分布形状的度量
- 小结