

描述统计

一：集中趋势的测度

2.中位数。把一组数据按从小到大或从大到小的顺序进行排列，位置居中的数值叫做中位数，用Me表示：

3.众数：指一组数据中出现次数（频数）最多的变量值。不适用于定量数据。

【注意1】均值主要适用于数值型数据，但不适用于分类和顺序数据。

【注意2】均值容易受到极端值的影响，极端值会使得均值向极大值或极小值方向倾斜，使得均值对数据组的代表性减弱。

【注意1】中位数是一个位置代表值，主要用于顺序数据和数值型数据，但不适用于分类数据。

【注意2】中位数的优点：不受极端值的影响，抗干扰性强，尤其适用于偏斜分布的数值型数据。

【注意】适用于描述分类数据和顺序数据的集中趋势。在定量数据中，可能出现多众数和无众数的情况，因此众数不适于描述定量数据的集中位置。

		
	适用	特点
均值	定量变量	- 利用全部信息，受极端值影响 - 如有明显极端值，则代表性差
中位数	顺序变量定量变量	- 不受极端值影响 - 适用于分布不对称的数据 - 没有充分利用全部信息，稳定性优于众数，差于均值
众数	分类变量顺序变量	- 没有充分利用全部信息，不受极端值影响 - 适用于分布明显呈偏态的数据 - 可能不唯一

离散程度：反映的是数据之间的差异程度。集中趋势的测度值对一组数据的代表程度，取决于该组数据的离散水平。

数据的离散程度越大，集中趋势的测度值对该组数据的代表性就越差；离散程度越小，其代表性就越好。

（一）方差和标准差

(1) 方差：数据组中各数值与其均值离差平方的平均数。

(2) 方差越小, 说明数据值与均值的平均距离越小, 均值的代表性越好。

(1) 标准差: 方差的平方根

(2) 标准差与方差只适用于数值型数据, 它们对极端值很敏感。

二：离散程度的测度

离散系数消除了测度单位和观测值水平不同的影响

(二) 离散系数

1.离散系数：变异系数或标准差系数。

2.标准差与均值的比值:

3.适用：不同类别数据离散程度的比较

1.偏度：指数据分布的偏斜方向和程度，描述的是数据分布对称程度。

2.偏态系数：测度数据分布偏度的指标。取决于离差三次方的平均数与标准差三次方的比值。

(一) 偏态系数 (SK)

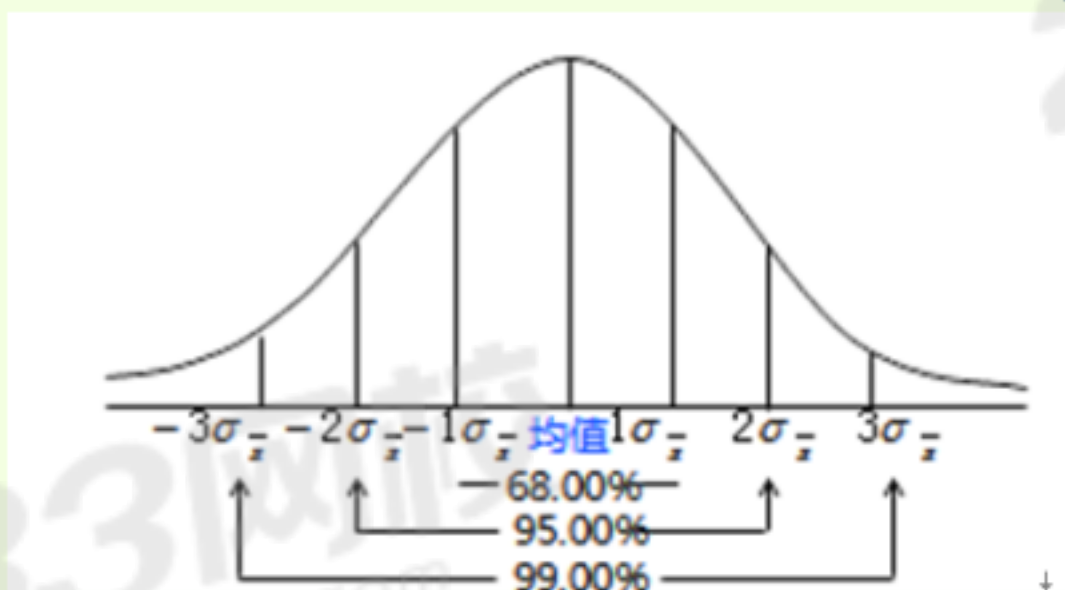
(1) $SK=0$: 数据分布是对称的;

(2) $SK > 0$: 数据分布右偏;

0~0.5 (轻度) 0.5~1 (中度) 大于1 (严重)

(3) $SK < 0$: 数据分布左偏;

0~ - 0.5 (轻度) - 0.5~ - 1 (中度) 大于 - 1 (严重)



(二) 标准分数

1.按相关的程度：完全相关、不完全相关和不相关。

2.按相关的方向：正相关和负相关。

3.按相关的形式：线性相关和非线性相关。

（一）相关关系分类

(二) 两变量的散点图：两个变量间的关系可以用散点图来展示。

1. Pearson相关系数：度量的是两个变量之间的线性相关关系

$r=0$ 只表示两个变量之间不存在线性相关关系, 并不说明变量之间没有任何关系

2. Pearson相关系数的取值范围: $-1 \leq r \leq 1$ 。

(1) $0 < r \leq 1$: 变量X和Y之间存在正线性相关关系;

(2) $-1 \leq r < 0$: 变量X和Y之间存在负线性相关关系;

(3) $r=1$: 变量X和Y之间为完全正线性相关, 即变量Y的取值完全依赖于变量X;

(4) $r = -1$: 变量X和Y之间为完全负线性相关, 即变量Y的取值完全依赖于变量X。

3.相关程度:

(1) $|r| \geq 0.8$: 高度相关;

(2) $0.5 \leq |r| < 0.8$: 中度相关;

(3) $0.3 \leq |r| < 0.5$: 低度相关;

(4) $|r| < 0.3$: 说明两个变量之间的相关程度极弱, 可视为无线性相关关系。