2-4 统计指标：集中趋势

版权©

作者：AI悦创

公众号：AI悦创

日期：2021年04月15日

期待你和我一起，用数据解析世界！

1. 集中趋势指标的特点

- 哪个营销渠道引流效果最佳?
- 什么岗位的薪酬水平最高?
- 哪个产品最受欢迎?
- ……

这个时候，我们就可以计算平均值 → 以便得到初步结论

2. 什么是集中趋势指标

- 用于体现数据一般水平的指标
- 最快速了解样本数据的概况
- 最常用的集中趋势指标就是平均值

3. 平均值

工号	薪酬
20200101	9000
20200102	12500
20200103	7500
20200104	8600
20200105	11000
20200106	9500
20200107	13500
20200108	14500
20200109	65000

= 所有数据相加 / 数据的个数

加和：151100

平均值：16789

很明显，这个平均值是出现异常的

原因是有异常值，有数值高于数据中的各个数据，大大的拉高了——我们平均值的水平

所以，这也是我们在数据预处理的时候 必做的一件事。

去除异常值：65000

工号	薪酬
20200101	9000
20200102	12500
20200103	7500
20200104	8600
20200105	11000
20200106	9500
20200107	13500
20200108	14500
20200109	65000

= 所有数据相加 / 数据的个数

加和：151100

平均值：16789　　差异：6026

去除异常值：10762

具有一定误导性，对异常数不敏感

所以单纯的计算平均值，是具有一定误导性的，因为：平均值对异常的数值是不敏感的。

4. 中位数

因为：单纯的计算平均值，是具有一定误导性的，因为：平均值对异常的数值是不敏感的。

这个时候我们就引出了：中位数

概念：按顺序排列后，居于中间位置的数

按顺序排列后，居于中间位置的数

奇数：位于(n+1)/2位置的数

偶数：最中间的两位数相加 / 2

更具有代表性

工号	薪酬
20200103	7500
20200104	8600
20200101	9000
20200106	9500
20200105	11000
20200102	12500
20200107	13500
20200108	14500
20200109	65000

5. 众数

概念：出现次数最多的数值

薪酬范围	出现次数
3001~4000	10
4001~5000	20
5001~6000	30
6001~7000	40
7001~8000	55
8001~9000	54
9001~10000	59
10001~20000	32
20001~30000	21

反应的是 局部特征、也就是：最密集，最常出现的，那个数据项，就是我们所称的：众数。

众数可以有多个

6. 章节回顾

7. 作业

1、计算平均值、中位数、众数

2、哪一组薪酬水平更高

薪酬组1	薪酬组2
8600	3900
8700	4500
9500	7300
10400	9500
11000	11500
11500	14500
12500	21000

基于本节课所学习的内容，给大家两组薪酬数据，计算这两组薪酬数据的平均值、中位数、众数，并根据平均值、中位数和众数来判断哪一组的薪酬水平更高，原因是什么?

你也可以把自己判断的理由和根据发表出，大家一起讨论～～

Tips：这两组薪酬数据的平均值是相等的哦，本节课作业的答案也会放到代码仓库去。

组1	组2
8600	3900
8700	4500
9500	7300
10400	9500
11000	11500
11500	14500
12500	21000
10314.286	10314.286

	平均值	中位数	众数
组1	10314.286	10400	所有
组2	10314.286	9500	所有
薪酬水平		1>2	

答案