

各种平均数及其适用场景

2015年7月27日 星期一 下午10:30

一、简介

算术平均数、几何平均数、调和平均数、平方平均数。

详见[wiki](#)详细公式

二、适用场景、优缺点

1、算术平均数

优点：直观好算，解释性强

缺点：个体样本影响大，

例如北京市平均工资，应该用中位数，20%的样本值占了数值上的80%，（20%的人拥有了80%的财富）。

很多都可以用算术平均数的。

2、中位数

优点：个体样本基本没影响，只看偏序关系

缺点：会误以为是平均值，但是明明不是平均值

例如：北京市平均工资，应该用中位数

其实中位数是反映处于中等的样本的样本值，而不是样本值的平均、中点。

3、众数

简单粗暴、投票、少数服从多数

4、几何平均数

优点：对个体样本更敏感，乘法原则。适用于数据之间有高次的关系、一系列可以相乘的子指标。

缺点：对个体样本更敏感，比算术平均数还敏感，而且要求不能 ≤ 0

例如：

如果你有一个12*24*48公分的箱子，那么对应一个标准的立方体的平均是多少？ $(12*24*48)^{(1/3)}=24$ 公分。（注意这个是3次根号）

通货膨胀率计算，如果我们得到三年的通货膨胀率是1%，2%和10%，那么平均通货膨胀率 $= (1.01*1.02*1.10)^{(1/3)}=4.3\%$

我国房地产全国市场，上半年房价下降41%，而下半年上涨了44%，全年平均增长了1.5%吗？

5、调和平均数

有下面几种理解：

1，可以数据之间有 x^{-1} 关系的时候，而且在不知道分子的情况下用（因为分子可以约掉）

2, 这是一种对 x^{-1} 之后的结果做线性的算术平均的平均数

3, (见参考资料4) 是加权平均的变种, 内在假设频率分布反比于数据

例如: 往返路程中, 往程60km/h, 返程40km/h, 平均速度多少?

需要调和平均数, 这个不知道路程 $s(x^{-1}$ 的分子), 而且是对时间 t 求平均, 然后线性相加, 再求平均时间

例如: F值为什么是, 准确率和召回率的调和平均数?

因为要对准确率和召回率做平均, 其实 $P=TT/(TT+TF)$ $R=TT/(TT+FT)$, P 和 R 的区别是 TF 和 FT , 要对这个要素做平均, 这两个当时是同等地位的, 这两个有这样的此消彼涨的关系: 对 F 来说 TF 涨多少和 FT 涨多少效果是一样的。

例如: 有个和 F 值类似的 E 值, 在 F 值基础上进行了加权, 其实也是加在了 TF/FT 上另一种理解问题

$F1$ 值为什么用准确率和召回率的调和平均数, 有啥好处?

还不太清楚, 不过调和平均数其实是一个倒数平均数

我们对准确和召回做平均, 写出来式子其实是对:

$P=TT/(TT+TF)$ $R=TT/(TT+FT)$ 做平均, 可以理解为是在对 TF 和 FT 这两个因素做平均。但是这两个变量对最终效果 ($F1$) 的影响应该是什么关系的?。

首先, 应该是地位平等的;

其次, FT 和 TF 之间有什么样的此消彼长的关系, 也就是说 TF 涨多少相当于 FT 涨多少。在这里, 可以展开 P 和 R 的计算公式, 带入 $F1$ 的计算。发现 TF 合 FT 是线性的关系, 而且没有系数

$$(F1=2/(2+(FT+TF)/TT))$$

三、参考资料:

1, wiki详细公式 <https://zh.wikipedia.org/wiki/%E5%B9%B3%E5%9D%87%E6%95%B0%E4%B8%8D%E7%AD%89%E5%BC%8F>

2, 博客 很详细的对比和例子 非常推荐

<http://shenhaolaoshi.blog.sohu.com/150219505.html>

3, 高深一些 没啥用 <http://blog.sciencenet.cn/blog-2024-341374.html>

4, <http://www.guokr.com/question/484467/>