数学知识

误差与残差

误差:即观测值与真实值的偏离;
残差:观测值与拟合值的偏离.
误 差分为两类：系统误差与随机误差。其中，系统误差与测量方案有关，通过改进测量方案可以避免系统误差。随机误差与观测者，测量工具，被观测物体的性质有关，只能尽量减小，却不能避免。
残差——与预测有关，残差大小可以衡量预测的准确性。残差越大表示预测越不准确。残差与数据本身的分布特性，回归方程的选择有关。
误差： 所有不同样本集的均值的均值,与真实总体均值的偏离.由于真实总体均值通常无法获取或观测到,因此通常是假设总体为某一分布类型,则有 N 个估算的均值; 表征的是观测/测量的精确度;
误差大,由异常值引起.表明数据可能有严重的测量错误;或者所选模型不合适,;
残差： 某样本的均值与所有样本集均值的均值， 的偏离； 表征取样的合理性,即该样本是否具代表意义;
残差大,表明样本不具代表性,也有可能由特征值引起.

等差等比数列

数列	通项公式	对称公式	求和公式
等差数列	$a_n=a_1+(n-1)d$	$a_m+a_n=a_i+a_j$，其中 $m+n=i+j$	(1)一般求和：$S_n=\frac{n(a_1+a_n)}{2}=na_1+\frac{1}{2}n(n-1)d$ (2)中项求和：$S_n=\begin{cases} na_{\frac{n+1}{2}}， & n\text{为奇数} \\ \frac{n}{2}(a_{\frac{n}{2}}+a_{\frac{n}{2}+1})， & n\text{为偶数} \end{cases}$
等比数列	$a_n=a_1\cdot q^{n-1}$	$a_m\cdot a_n=a_i\cdot a_j$，其中 $m+n=i+j$	$S_n=\begin{cases} \frac{a_1(1-q^n)}{1-q}， & q\neq 1 \\ na_1， & q=1 \end{cases}$
平方数列	$a_n=n^2$		$S_n=\frac{1}{6}n(n+1)(2n+1)$
立方数列	$a_n=n^3$		$S_n=[\frac{1}{2}n(n+1)]^2$

	等差数列	等比数列
定义	$a_{n+1}-a_n=d$	$\frac{a_{n+1}}{a_n}=q$
通项公式	$a_n=a_1+(n-1)d$	$a_n=a_1q^{n-1}$
前 n 项和公式	$S_n=na_1+\frac{n(n-1)}{2}d=\frac{n(a_1+a_n)}{2}$	①$q=1$时，$S_n=na_1$ ②$q\neq1$时，$S_n=\frac{a_1(1-q^n)}{1-q}=\frac{a_1-a_nq}{1-q}$
中项公式	a,A,b 成等差数列$\Leftrightarrow 2A=a+b$	a,G,b 成等比数列$\Rightarrow G^2=ab$
判定	$a_{n+1}-a_n=d$ $\Leftrightarrow 2a_{n+1}=a_n+a_{n+2}$ $\Leftrightarrow a_n=kn+b$ $\Leftrightarrow S_n=An^2+Bn$	$\frac{a_{n+1}}{a_n}=q$ $\Leftrightarrow a_{n+1}^2=a_n\cdot a_{n+2}(a_n\neq0)$ $\Leftrightarrow a_n=c\cdot q^{kn+b}(c,\ q\neq0)$ $\Leftrightarrow S_n=A+B\cdot q^n(q\neq0,A+B=0)$
性质	①$a_n=a_m+(n-m)d$	①$a_n=a_m\cdot q^{n-m}$
	②$m+n=p+q$时，$a_m+a_n=a_p+a_q$	②$m+n=p+q$时，$a_m\cdot a_n=a_p\cdot a_q$
	③$S_n,S_{2n}-S_n,S_{3n}-S_{2n},S_{4n}-S_{3n},\cdots\cdots$ 成等差数列	③$S_n,S_{2n}-S_n,S_{3n}-S_{2n},S_{4n}-S_{3n},\cdots\cdots$ 成等比数列 （$q\neq-1$）
	④ $a_n=\begin{cases}S_1 & ,n=1\\ S_n-S_{n-1} & ,n\geq2\end{cases}$	

做题

试纸

1000 瓶溶液，1 瓶是化学试剂，其他是清水。
化学试剂遇到试纸后 20min 后才会变色，1H，使用多少试纸才能把找出化学试剂。

1. 20 分钟 100，20 分钟 10，20 分钟 1 个，30 试纸。
2. 上述方法中前面一次清水使用的还可以再使用，于是需要 11 张。

Edwin Xu

3. 当做二进制。9 张？也不对

应该是有几种状态就需要几张???
4 种状态，4 张？
4^4< 1000
5 张？
4^5>1000

状态数^最小张数>试纸数

还需要理解

先看下面这个猪的题：
https://www.zhihu.com/question/60227816/answer/1274071217

1000 桶水，其中一桶有毒，猪喝毒水后会在 15 分钟内死去，想用一个小时找到这桶毒水，至少需要几头猪？
限制条件：
1.一滴毒水足以导致一头猪的死亡。死亡时间为 15 分钟内不确定的某个时间点。
2.其死亡只是毒水导致的，不会有其他因素导致死亡。
3.猪的承水量无穷大，且假设饮一桶花费时间为零。

一只猪在一个小时内会有几种状态？
1. 在第 0 分钟的时候喝了一桶水以后，第 15 分钟死去。
2. 第 15 分钟依然活着，喝了一桶水以后，第 30 分钟死去。
3. 第 30 分钟依然活着，喝了一桶水以后，第 45 分钟死去。
4. 第 45 分钟依然活着，喝了一桶水以后，第 60 分钟死去。
5. 第 45 分钟依然活着，喝了一桶水以后，第 60 分钟依然活着

1 只猪 1 个小时以后会有 5 种状态

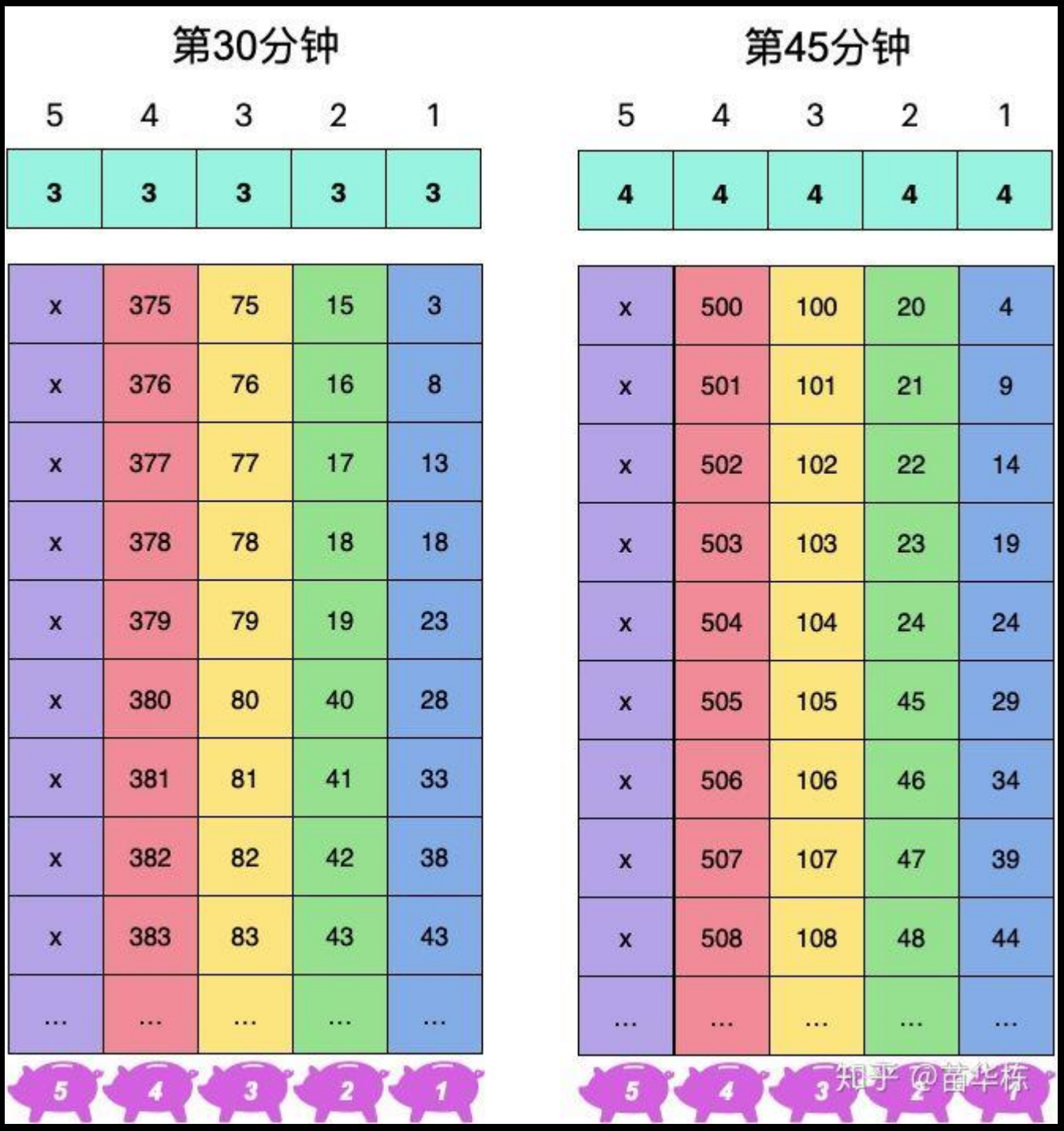
可以利用猪的 5 种状态构造一个 5 进制编码

第0分钟

5	4	3	2	1
1	1	1	1	1
625	125	25	5	1
626	126	26	6	6
627	127	27	7	11
628	128	28	8	16
629	129	29	9	21
630	130	30	30	26
631	131	31	31	31
632	132	32	32	36
633	133	33	33	41
…	…	…	…	…

第15分钟

5	4	3	2	1
2	2	2	2	2
x	250	50	10	2
x	251	51	11	7
x	252	52	12	12
x	253	53	13	17
x	254	54	14	22
x	255	55	35	27
x	256	56	36	32
x	257	57	37	37
x	258	58	38	42
…	…	…	…	…



如果 1 号猪第 30 分钟死了，2 号猪第 15 分钟死了，3 号猪第 45 分钟死了，4，5 号都活到了最后。则毒水对应的 5 进制编码是

$$0 \times 5^4 + 0 \times 5^3 + 3 \times 5^2 + 1 \times 5^1 + 2 \times 5^0 = 82$$

也就是第 82 桶水有毒

基数

Edwin Xu

简单来说，基数（cardinality，也译作势），是指**一个集合中不同元素的个数**。例如看下面的集合：这个集合有 9 个元素，但是 2 和 3 各出现了两次，因此不重复的元素为 1,2,3,4,5,9,7，所以这个集合的基数是 7。

基于 B 树的基数计数

基于 bitmap 的基数计数

为了克服 B 树不能高效合并的问题，一种替代方案是使用 bitmap 表示集合。也就是使用一个很长的 bit 数组表示集合，将 bit 位顺序编号，bit 为 1 表示此编号在集合中，为 0 表示不在集合中。例如“00100110”表示集合 {2，5，6}。bitmap 中 1 的数量就是这个集合的基数。

显然，与 B 树不同 bitmap 可以高效的进行合并，只需进行按位或（or）运算就可以，而位运算在计算机中的运算效率是很高的。但是 bitmap 方式也有自己的问题，就是内存使用问题。

很容易发现，bitmap 的长度与集合中元素个数无关，而是与基数的上限有关。例如在上面的例子中，假如要计算上限为 1 亿的基数，则需要 12.5M 字节的 bitmap，十个链接就需要 125M。关键在于，这个内存使用与集合元素数量无关，即使一个链接仅仅有一个 1UV，也要为其分配 12.5M 字节。

由此可见，虽然 bitmap 方式易于合并，却由于内存使用问题而无法广泛用于大数据场景。